Doctoral theses at NTNU, 2022:191

Daniel Groos

Convolutional networks for video-based infant movement analysis

Towards objective prognosis of cerebral palsy from infant spontaneous movements

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of Philosophiae Doctor
Faculty of Medicine and Health Sciences
Department of Neuromedicine and Movement Science

NTNU
Norwegian University of Science and Technology

Daniel Groos

Convolutional networks for video-based infant movement analysis

Towards objective prognosis of cerebral palsy from infant spontaneous movements

Thesis for the Degree of Philosophiae Doctor

Trondheim, June 2022

Norwegian University of Science and Technology
Faculty of Medicine and Health Sciences
Department of Neuromedicine and Movement Science

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Medicine and Health Sciences
Department of Neuromedicine and Movement Science

ISBN 978-82-326-5582-3 (printed ver.)
ISBN 978-82-326-6906-6 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2022:191

Printed by NTNU Grafisk senter

Sammendrag

Cerebral parese (CP) er en samlebetegnelse på motoriske funksjonsforstyrrelser grunnet skade på hjernen tidlig i barnets utvikling. Det er særlig spedbarn med medisinske risikofaktorer, som for eksempel for tidlig fødsel, pustebesvær og infeksjoner, som står i fare for å utvikle CP. CP har innvirkning på barnets holdning og motorikk, men gir også andre utfordringer og komplikasjoner. Som følge av manglende tidlige symptomer blir ofte ikke diagnosen satt før 1-2 års alder. Tidlig gjenkjenning av CP hos spedbarn er viktig for å kunne starte målrettet behandling, forebygge komplikasjoner og redusere bekymring hos foreldre.

Undersøkelse av spedbarnets spontane bevegelser med metoden General Movement Assessment (GMA) kan indikere om et barn har CP allerede før 5 måneders alder. GMA utføres ved observasjon av et spedbarns spontane bevegelser i en video. Ettersom dette avhenger av tilgang til erfarne og trenede observatører er denne undersøkelsen ikke tilgjengelig for alle. Maskinlæringsbasert CP-prediksjon har blitt utforsket som et alternativ til GMA, men foreløpig har man ikke lyktes med å lokalisere de spontane bevegelsene til et spedbarn i en video på en presis måte. Samtidig er man avhengig av menneskelige eksperter for å kunne velge ut relevante egenskaper i spedbarnsbevegelsene og for å utvikle prediksjonsmodeller.

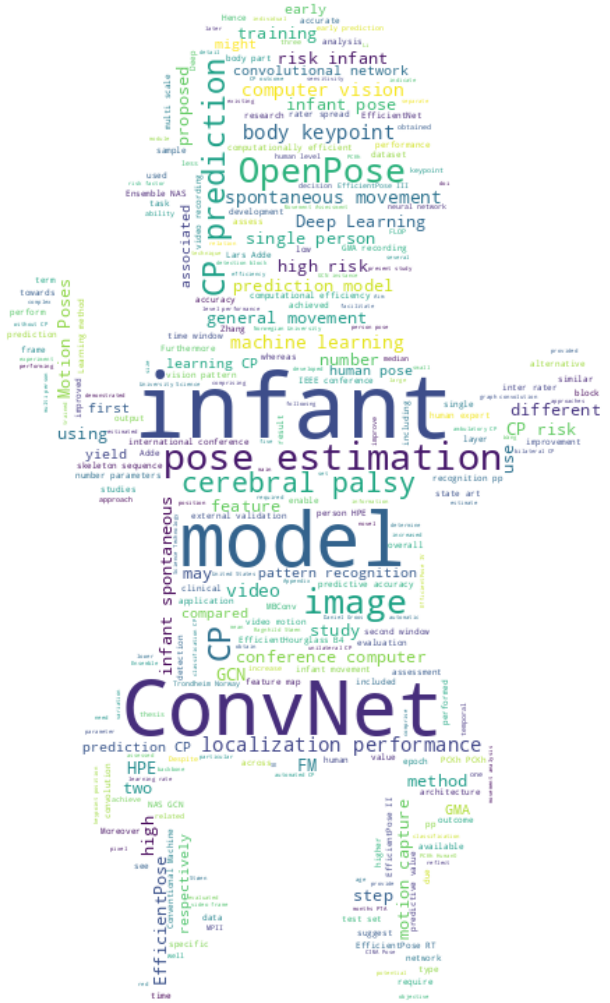
Konvolusjonelle nettverk kan tilpasse seg komplekse oppgaver gjennom automatisk utvelgelse av relevante egenskaper ved bruk av tilpassede nettverksarkitekturer. Formålet med denne avhandlingen var å undersøke presisjonen og beregningseffektiviteten til bildebaserte konvolusjonelle nettverk (ConvNets) for lokalisering av spedbarns spontane bevegelser i videoopptak, og å evaluere nøyaktigheten til grafbaserte konvolusjonelle nettverk (GCNs) for prediksjon av CP.

Resultatene fra dette doktorgradsarbeidet viser at ConvNets er i stand til å lokalisere spedbarnsbevegelser i video like godt som det et menneske gjør samtidig som videoen prosesseres i sanntid. En GCN-basert prediksjonsmodell for CP kan videre oppnå like god nøyaktighet som det kliniske eksperter gjør ved bruk av GMA ved 3 måneders alder. Prediksjonsmodellen har også svært god evne til å forutsi gående eller ikke-gående funksjon hos barn med CP og å skille mellom spedbarn som utvikler ensidig og tosidig lammelse.

Denne avhandlingen viser at konvolusjonelle nettverk kan brukes til videobasert bevegelsesanalyse av spedbarn for nøyaktig automatisk prediksjon av CP. Tidlig og objektiv gjenkjenning av CP hos spedbarn med medisinske risikofaktorer kan inspirere til utvikling av maskinlæringsbasert klinisk beslutningsstøtte og oppmuntre til videre forskning i grenseflaten mellom moderne medisinsk teknologi og klinisk ekspertkunnskap.

Kandidat: Daniel Groos
Institutt: Institutt for nevromedisin og bevegelsesvitenskap
Veiledere: Espen A.F. Ihlen, Lars Adde, Heri Ramampiaro

Ovennevnte avhandling er funnet verdig til å forsvares offentlig for graden PhD i medisinsk teknologi.
Disputas finner sted i auditorium MTA, Fred Kavli-bygget, NTNU.
Tirsdag 14. juni 2022 kl. 12:15.


infant
model
ConvNet
pose estimation
cerebral palsy
prediction
CP
OpenPose
computer vision
body keypoint
risk infant
convolutional network
infant pose
spontaneous movement
Deep Learning
single person
high risk
prediction model
general movement
machine learning
image
video
pattern recognition
localization performance
FM
HPE
GCN
EfficientPose
Motion Poses
early
training
might
proposed
associated
Ensemble NAS
learning CP
number
human pose
first
using
yield
feature
may
study
conference computer
method
GMA
motion capture
step
high
two
respectively
different
CP risk

Acknowledgments

In 2018, I set forth on this PhD journey at the Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology (NTNU) as a continuation of the Master's project conducted at the Department of Computer Science, NTNU. The PhD project was funded by the Faculty of Medicine and Health Sciences, NTNU.

Associate Professor Espen A.F. Ihlen took me under his wing and has my deepest gratitude, for devoting countless hours of guidance as main supervisor and for being a constant source of inspiration during the different stages of this journey. I appreciate that you have actively contributed to the development of convolutional networks and shared intriguing reflections on all types of architectural building blocks and deep learning techniques, from Squeeze-and-Excitation and pooling layers to data augmentation and weight initialization. Thank you for introducing me to the world of research in a friendly knowledge-sharing environment and for providing opportunities beyond the PhD. It has been a real pleasure learning from you and I look forward to more technical discussions in the time to come.

Dr. Lars Adde, the physiotherapist who became father of machine learning-based prediction of cerebral palsy from videoes, introduced me to this meaningful research topic in 2017. During the last five years, you have been the greatest of support as co-supervisor and brought me a step closer to understanding the real nature of interdisciplinary research. Thank you for your endless commitment to this research and for expanding my perspective through numerous conversations.

Professor Heri Ramampiaro, the man of many hats, allowed me in his group of Master students in 2017 and gave me the opportunity to embark on this exciting adventure. Apart from supervising Kristian and me during the Master's and providing guidance as co-supervisor during the PhD, I am grateful for the unique experiences I have today thanks to you, like the surreal opportunity to present the results of the Master's project to the Prime Minister.

The many years of effort spent by the co-authors and other contributors in recruiting infants into clinical studies and collecting and organizing considerable amounts of data, has made this research possible. I also want to express my appreciation to the co-authors for all your support in preparing and reviewing the papers of this thesis. A special thanks to Professor Ragnhild Støen for improving the manuscripts and for your invaluable reflections, especially on the difficult medical questions.

During the course of this PhD, I have been privileged to co-supervise and collaborate with Master students at the Department of Computer Science. The various machine learning approaches for video-based motion capture and prediction of cerebral palsy that have been explored in these Master projects have provided important insights for navigating this dynamic, evergrowing discipline. I would like to highlight the contribution of Andreas Haukeland and Sindre Aubert in discovering graph-based convolutional networks as an alternative for skeleton-based prediction of cerebral palsy, a seed that grew into the third paper of this thesis.

I am also thankful for the many experiences outside the core of my PhD. These years have been a great practice in communicating heavy concepts like deep learning to various audiences, from curious 10-year-old pupils at Forskningstorget to domain experts at international conferences. The PhD has also taught me about innovation and commercialization processes through collaboration with NTNU Technology Transfer Office, which I am grateful for. I also appreciate the opportunity to take part in the development of clinically usable software integrating methods proposed in the PhD. It has also been a great privilege to explore the potential of the proposed methods in other movement science applications, and in particular the collaboration with SenTIF and Olympiatoppen on video-based kinematic analysis of elite ski jumpers has been a motivating experience.

This PhD has far from solely been an academic journey. During my time at Øya and Gløshaugen, I have highly enjoyed the company of many wonderful and inspiring colleagues in GeMS and NTNU Neo at Øya and DART at Gløshaugen. Although there are many more people who deserve to be mentioned, Ailan, Arnhild, Astrid, Deepika, Hassan, Martin, Shweta, Sofia, and Tárik are few of whom I have been lucky to walk this path with and that I have shared precious moments with, also outside office hours.

I would also like to express my gratitude to some of my dear friends who have deliberately or undeliberately motivated me during these years of study. Tobias: All these experiences we have together would be a book of its own. Sulalit: You are the elder brother I did not have, I cherish your wisdom and life advice. Markus: I look forward to our very special celebration. Max: I have really enjoyed our "coffee breaks" over Teams. Ørjan, Erik, and Henrik: Our competition has been the perfect sidequest to PhD writing, be prepared for the upcoming season.

My parents Jan Erik and Ellen, words can not describe my appreciation for your endless love and support, even 750 kilometers away. You have travelled to Trondheim a dozen times and together we have shared incredible adventures in Norway's 11 counties. My sister Charlotte, her husband Bjørn Olav, and their kids also have a special place in my heart. Thank you for always inviting us over when the opportunity is there. My dearest Patricia, this journey would not have been completed without your exceptional help at home and patience from day to night. You are the love of my life and I look forward to growing old with you.

Summary

Cerebral palsy (CP) is the most common physical disability in childhood, with a particularly high prevalence in infants with medical risk factors (i.e., high-risk infants), like preterm birth. CP is caused by injury to the developing brain which affects a child's movement and posture but also involve associated impairments and complications. The lack of early pathological signs of CP, typically delays the diagnosis until 12 to 24 months of age. However, early detection of CP is necessary to improve function through targeted intervention.

The quality of spontaneous movements of infants has evolved as an accurate marker for CP before 5 months of age. The qualitative General Movement Assessment (GMA) enables early prediction of CP from infant spontaneous movements in a video. However, the dependency on highly experienced human GMA experts questions its scalability. Machine learning-based CP prediction has attempted to replicate the predictive accuracy of GMA, but currently lack precise motion capture of infant spontaneous movements in videos and require human expert involvement in selecting movement features and designing prediction models.

Convolutional networks have ability to adapt to complex tasks through automatic feature extraction with dedicated network architectures. In this thesis, we investigate the localization performance and computational efficiency of image-based convolutional networks (ConvNets) in video-based motion capture of infant spontaneous movements, and the predictive accuracy of graph-based convolutional networks (GCNs) for prediction of CP.

Results show that video-based motion capture harnessing ConvNets can approach human-level localization performance with real-time processing speeds. Moreover, a prediction model for CP utilizing GCNs can achieve predictive accuracy non-inferior to the clinically recommended human expert-based GMA in high-risk infants at 3 months age. Such a prediction model can also distinguish infants with ambulatory CP from non-ambulatory CP and infants with unilateral CP from bilateral CP.

This thesis demonstrates the potential of convolutional networks in video-based infant movement analysis. The knowledge acquired may pave the way for early, objective detection of CP in high-risk infants, encourage implementation of machine learning-based clinical decision support, and inspire future research to discover fruitful collaborations between contemporary medical technology and clinical expert knowledge.

List of papers

This thesis comprises the following papers:

Paper I

"EfficientPose: Scalable single-person pose estimation"

Daniel Groos, Heri Ramampiaro, Espen A.F. Ihlen

Applied Intelligence

2021; 51:2518-2533

Paper II

"Towards human-level performance on automatic pose estimation of infant spontaneous movements"

Daniel Groos, Lars Adde, Ragnhild Støen, Heri Ramampiaro, Espen A.F. Ihlen

Computerized Medical Imaging and Graphics

2022; 95:102012

Paper III

"Development and external validation of deep learning-based early prediction of cerebral palsy from spontaneous movements in high-risk infants"

Daniel Groos[†], Lars Adde[†], Sindre Aubert, Lynn Boswell, Raye-Ann deRegnier, Toril Fjørtoft, Deborah Gaebler-Spira, Andreas Haukeland, Marianne Loennecken, Michael Msall, Unn Inger Möinichen, Aurelie Pascal, Colleen Peyton, Heri Ramampiaro, Michael D. Schreiber, Inger Elisabeth Silberg, Nils Thomas Songstad, Niranjan Thomas, Christine Van den Broeck, Gunn Kristin Øberg, Espen A.F. Ihlen[‡], Ragnhild Støen[‡]

Submitted for publication

2022

[†] Daniel Groos and Lars Adde contributed equally as co-first authors

[‡] Espen A.F. Ihlen and Ragnhild Støen contributed equally as co-last authors

Abbreviations

AUC	Area under the receiver operating characteristic curve
CAM	Class activation mapping
ConvNet	Image-based convolutional network
CP	Cerebral palsy
C_{SD}	Standard deviation of the centroid of motion
FLOPs	Floating-point operations
FMs	Fidgety movements
FPS	Frames per second
GA	Gestational age
GCN	Graph-based convolutional network
GMA	General Movement Assessment
GMFCS	Gross Motor Function Classification System
GMs	General movements
HINE	Hammersmith Infant Neurological Examination
HPE	Human pose estimation
MBConv	Mobile inverted bottleneck convolution
MRI	Magnetic resonance imaging
NAS	Neural architecture search
NPV	Negative predictive value
PCK	Percentage of correct keypoints
PMA	Post-menstrual age
PPV	Positive predictive value
PTA	Post-term age
Q_{mean}	Mean of quantity of motion
ROC	Receiver operating characteristic
SCPE	Surveillance of Cerebral Palsy in Europe
SE	Squeeze-and-excitation
SGD	Stochastic gradient descent

Contents

1 Introduction **1**
1.1 Cerebral palsy 1
1.2 Early medical prediction of CP 2
1.3 Convolutional networks for machine learning-based CP prediction . 4
1.3.1 Video-based motion capture 8
1.3.2 Prediction model 9
2 Aims of the thesis **11**
3 Methods **13**
3.1 Research overview 13
3.2 Preliminaries 14
3.2.1 Convolution 14
3.2.2 EfficientNets 16
3.2.3 Graph convolution 19
3.3 Study I: Single-person pose estimation 20
3.3.1 MPII Human Pose Dataset 20
3.3.2 Experimental approach 20
3.3.3 EfficientPose and EfficientHourglass 20
3.3.4 Evaluation 23
3.4 Study II: Infant pose estimation 24
3.4.1 In-Motion Poses 24
3.4.2 Experimental approach 27
3.4.3 Comparative analysis 27
3.4.4 Evaluation 28
3.5 Study III: Prediction model for CP 29
3.5.1 Participants 29
3.5.2 Experimental approach 31
3.5.3 Ensemble-NAS-GCN 33
3.5.4 Statistical analysis 35
4 Summary of results **37**
4.1 Study I 37
4.2 Study II 38
4.3 Study III 39
5 Discussion **43**
5.1 Video-based motion capture 46

5.1.1 Localization performance 46
5.1.2 Computational efficiency 49
5.2 Prediction model 51
5.2.1 Classification and uncertainty 51
5.2.2 Explanations 55
5.3 Other considerations and avenues for future research 57
5.3.1 Field of use and transfer validity 57
5.3.2 One-step versus two-step approach 58
5.3.3 Relation to automated GMA 59
5.3.4 Multimodality convolutional network-based CP prediction . 60
6 Conclusion 61
A Neural architecture search for graph-based convolutional networks . 77
A.1 Search space 77
A.2 *K*-Best Search 79
A.3 Ensemble-NAS-GCN 82
B Single-person pose estimation 85
C Infant pose estimation 87
D Automated CP prediction 89

Chapter 1

Introduction

1.1 Cerebral palsy

Advancements in neonatal care have increased the survival rates of infants with medical risk factors, like preterm birth [1]. Every year, 15 million infants are born preterm, and the number keeps rising [2]. Extremely preterm birth (i.e., gestational age (GA) of less than 28 weeks) and other medical risk factors (e.g., neonatal encephalopathy and intraventricular hemorrhages) during the neonatal period pose a great risk for brain injuries and neurodevelopmental disabilities [3].

Cerebral palsy (CP) is the most common physical disability in childhood, with an overall prevalence of around two per 1 000 live births [4, 5] and 11% in extremely preterm infants [4]. CP is defined as "*a group of permanent disorders of the development of movement and posture, causing activity limitation, that are attributed to non-progressive disturbances that occurred in the developing fetal or infant brain. The motor disorders of cerebral palsy are often accompanied by disturbances of sensation, perception, cognition, communication, and behaviour, by epilepsy, and by secondary musculoskeletal problems*" [6]. Hence, CP is an umbrella term for permanent motor disorders, often involving associated impairments and complications, due to abnormal brain development in an early phase of life.

Although CP mirrors early disturbances to the infant brain, it takes developmental time for pathological signs of CP to emerge [7]. Accordingly, diagnosis of CP is typically performed between 12 and 24 months of age based on a constellation of clinical and neurological signs [8]. However, early detection of CP is necessary to optimize function and improve quality of life [8]. Especially, targeted intervention during the first two years of life, when the plasticity of the brain is at its highest [9], promotes improved motor and cognitive outcomes in children with CP [10]. Moreover, early detection may improve access to community services reducing further complications and reassure parents of healthy infants [11, 12]. Hence, there is a call for techniques accurately predicting CP from early markers [13].

1.2 Early medical prediction of CP

Before 5 months post-term age (PTA), prominent markers for CP are abnormalities in neonatal neuroimaging, infant spontaneous movements (i.e., movements that infants perform spontaneously without any external stimulation), and abnormal muscle tone, reflexes, and reactions [8, 14–16]. These markers have in common that they are age-specific, and hence require assessment techniques adapted to the particular developmental period [7]. At term-equivalent age, brain abnormalities may be identified using magnetic resonance imaging (MRI) [17]. The predictive validity for CP from infant spontaneous movements is strongest at 2-5 months PTA, using Prechtl's General Movement Assessment (GMA) [18, 19]. Neurological function (e.g., muscle tone, reflexes, and reactions) can be assessed with Hammersmith Infant Neurological Examination (HINE) at 3-18 months PTA using age-dependent cut-off values for prediction of CP [16, 20]. Recent guidelines recommend the use of MRI, GMA, and HINE to give an interim diagnosis of high risk of CP before 5 months PTA [8]. However, in infants with known medical risk factors, GMA at 2-5 months PTA has demonstrated the best predictive accuracy for later CP [7, 8].

GMA focuses on distinct infant spontaneous movements that "involve the whole body in a variable sequence of arm, leg, neck and trunk movements" [18], termed by Prechtl et al. [21] as general movements (GMs). GMs take different forms from 2 months post-menstrual age (PMA) [22] to 5 months PTA (Figure 1.1) [23], and particularly the type of GMs, called fidgety movements (FMs) [23], appearing between 2 and 5 months PTA is indicative of healthy motor development [24, 25]. FMs are characterized as "small movements of moderate speed and variable acceleration, of neck, trunk and limbs, in all directions" [18]. Whereas the presence of FMs indicates typical brain development during the major transition of neural functions around 3 months PTA [7, 26], a lack of FMs (i.e., absent FMs) during this period is often intertwined with abnormal brain development. This indicates neurological disorders [27, 28], especially CP [8, 29–31].

GMA qualitatively assesses FMs from a video recording of an infant in supine position [18]. This makes it a promising assessment which can be easily performed in clinic, even in low-resource settings, with a single video camera or smartphone [33–35], and that is non-invasive unlike MRI and HINE. FMs are perceived by trained human observers with visual gestalt perception [18]. This implies an advanced form of pattern recognition that pays attention to the overall character of whole-body movements rather than individual movements of certain body parts. As described by Lorenz [36]: "*Gestalt perception is able to take into account a greater number of individual details and more relationships between these than in any rational computation.*"

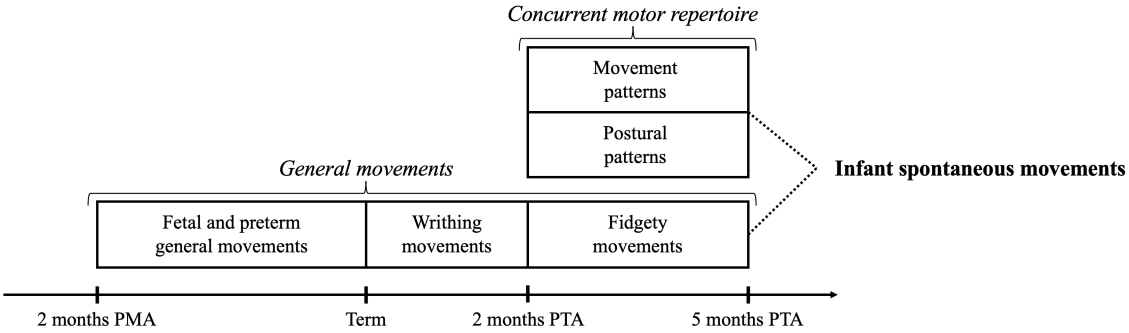


Figure 1.1: From 2 months PMA to 5 months PTA, general movements (GMs) take different forms; fetal and preterm GMs before term, writhing movements until 2 months PTA, and fidgety movements (FMs) between 2 and 5 months PTA [18]. During the FMs period, there are also other movement patterns (e.g., kicking, circular arm movements, and hand-to-hand contact) and postural patterns (e.g., head centered, body symmetry, and extended legs), occurring together with FMs [18, 23, 32]. These patterns, referred to as the concurrent motor repertoire, in combination with FMs constitute the infant spontaneous movements at this age.

Consequently, complex gestalt perception is attained through repetitive observation and learning from experience [36]. As a result, GMA requires extensive training and years of experience for high inter-observer reliability [37], which limit the number of certified GMA observers that are available [38], and, in turn, hamper the widespread clinical use of GMA. Furthermore, the subjective and qualitative nature of gestalt perception in GMA hinders the quantification of FMs, which makes them a questionable marker for CP [38]. Moreover, GMA during FMs age, unlike the Assessment of Motor Repertoire [18], does not take into account other movement patterns and postural patterns of infant spontaneous movements which occur together with FMs (i.e., concurrent motor repertoire in Figure 1.1 in GMA recordings [18, 23], and that also often appear atypical in infants with CP [32, 39]. Hence, there is a need for more objective and scalable techniques to support GMA for early prediction of CP.

1.3 Convolutional networks for machine learning-based CP prediction

Recent advancements in computer vision and machine learning, with the advent of deep learning, have provided automated solutions to many challenging tasks related to analysis of images and videos, which preserves the non-invasive character of GMA with no dependency to body-worn markers, sensors, or specialized laboratory equipment. In particular, the kind of deep learning called image-based convolutional network (ConvNet) has enabled automated analysis that is approaching or even surpassing human performance on specific visual tasks [40]. ConvNets achieve human-level performance by harnessing representation learning to detect complex task-specific features without the need for human expert involvement [41]. The resulting features obtained by ConvNets could represent complex gestalts similar to the modern computers Lorenz [36] foresaw when he stated that "*Gestalt perception can uncover an unsuspected regularity, whereas the rational abstraction process is absolutely incapable of doing so. With the exception of some very modern computers, which are able to superimpose a large number of curves and to derive a principle operating in them all, we have no means ... which is able to discover inherent principles.*"

Thus, ConvNets could be considered analogous to gestalt perception in humans. In particular, the ability of ConvNets to find intricate global features in data, through feature hierarchies of increasing abstraction level, could remind of the way humans detect meaningful patterns using gestalt perception to emphasize the overall picture rather than individual details. To obtain an analogue to gestalt perception in humans, ConvNets have drawn inspiration from the human visual cortex [41]. By employing convolution, neurons in the early layers of a ConvNet have a small receptive field and detect local features in an image (e.g., edges). This reminds of neurons in the primary visual cortex [42, 43]. The stacking of several layers increases the receptive fields of neurons in later ConvNet layers, like those in higher-order visual cortices, ultimately enabling detection of meaningful global features in images [41]. Furthermore, the principle of convolution and stacking of layers in convolutional networks to extract global features have proven promising beyond visual data and could be used to analyze complex whole-body movements [44].

In infant movement analysis, convolutional networks may be used to automatically quantify FMs and other complex movement patterns of spontaneous movements, and perform objective early prediction of CP, as a promising alternative to the subjective and qualitative gestalt perception in GMA. Former machine learning methods for CP prediction from GMA recordings at FMs age have either focused only on absent FMs as a surrogate outcome for CP (i.e., automated GMA), or all patterns of infant spontaneous movements related to CP outcome (i.e., automated CP prediction) [45–47], as summarized in Table 1.1 and 1.2, respectively. These have followed the two-step approach in Figure 1.2. In the first step, the infant spontaneous movements in a GMA recording are quantitatively represented as body keypoint positions in each video frame (i.e., video-based motion capture). Thereafter, in the second step (i.e., prediction model), features are extracted from the movements of body keypoints and analyzed by a classification method to perform prediction of outcome.

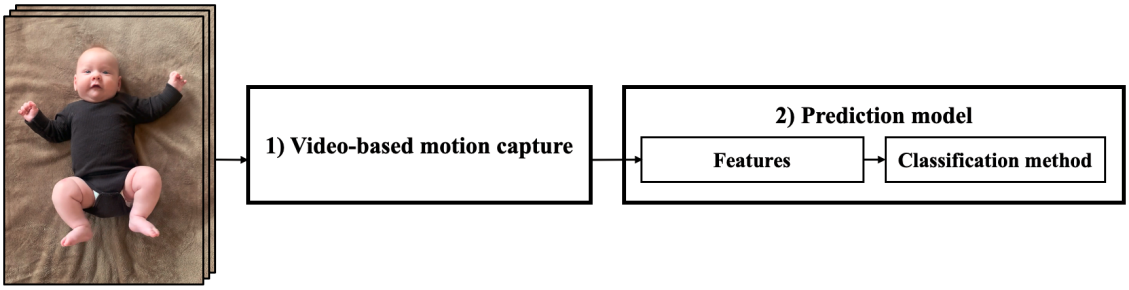


Figure 1.2: The two-step approach for machine learning-based CP prediction consists of 1) video-based motion capture quantitatively representing the infant spontaneous movements in a GMA recording, and 2) prediction model with selection of features and classification method to predict outcome.

Table 1.1: Previous studies on automated GMA from videos at 2-5 months PTA.

Year	Author	Sample size (absent FMs)	Video-based motion capture	Features	Classification method	Evaluation method	Predictive values		
							SE	SP	AUC
2009	Adde et al. [48]	137 (27)[a]	Frame differencing	C_{SD}	CML: C_{SD} cut-off	NR	81%	70%	0.83
2013	Adde et al. [49]	52 (9)	Frame differencing	C_{SD}	CML: C_{SD} cut-off	Leave-one-out CV	67%	79%	0.83
2017	Støen et al. [50]	241 (58[c])[b]	Frame differencing	C_{SD}	CML: C_{SD} cut-off	NR	80%	53%	0.73
2018	Orlandi et al. [51]	127 (29)	Optical flow	9 global features	CML: Random forest	Leave-one-out CV	31%	94%	0.83
2019	Schmidt et al. [52]	∼272 (NR)[d]	No motion capture	ConvNet features	DL: LSTM	20% test set	51%	27%	NR
2019	McCay et al. [53]	12 (4)	OpenPose	HOJD2D	CML: LDA	Leave-one-out CV	100%	100%	1.00
2021	Nguyen-Thai et al. [54]	235 (35[c])	OpenPose	Raw sequence	DL: GCN	5-fold CV	80%	62%	0.82
2021	McCay et al. [55]	12 (4)	OpenPose	HOJO2D+HOJD2D	CML: Ensemble classifier	Leave-one-out CV	100%	100%	1.00
2021	Sakkos et al. [56]	25 (6)	OpenPose	Raw sequence	DL: ConvNet+LSTM	Leave-one-out CV	83%	95%	NR

Abbreviations: SE, sensitivity; SP, specificity; AUC, area under the receiver operating characteristic curve; NR, not reported; CV, cross validation; C_{SD}, centroid of motion standard deviation; HOJD2D, histogram of joint displacement 2D; HOJO2D, histogram of joint orientation 2D; CML, conventional machine learning; DL, deep learning; LSTM, long short-term memory network; LDA, linear discriminant analysis; GCN, graph-based convolutional network; ConvNet, image-based convolutional network.

[a] The sample size was reported as the number of available videos, whereas the total number of infants was 82.

[b] The sample size was reported as the number of available videos, whereas the total number of infants was 150.

[c] Absent FMs also included sporadic FMs.

[d] The sample size was estimated from the reported number of 2 445 time windows of 20 seconds from 3-minute videos.

Table 1.2: Previous studies on automated CP prediction from videos at 2-5 months PTA.

Year	Author	Sample size (CP)	Video-based motion capture	Features	Classification method	Evaluation method	Predictive values		
							SE	SP	AUC
2010	Adde et al. [57]	30 (13)	Frame differencing	C_{SD}, Q_{mean}, Q_{SD}	CML: Linear classifier	NR	85%	88%	0.88
2012	Stahl et al. [58]	82 (15)	Optical flow	Relative frequencies	CML: SVM	10-fold CV	85%	96%	NR
2013	Adde et al. [49]	52 (9)	Frame differencing	C_{SD}	CML: C_{SD} cut-off	Leave-one-out CV	67%	77%	0.82
2014	Rahmati et al. [59]	78 (14)	Optical flow	A, P, CC	CML: SVM	Leave-one-out CV	50%	95%	NR
2015	Rahmati et al. [60]	78 (14)	Optical flow	2 376 FFT features	CML: PLSR	Leave-one-out CV	92%	87%	NR
2016	Rahmati et al. [61]	78 (14)	Optical flow	2 376 FFT features	CML: PLSR	Leave-one-out CV	86%	92%	0.93
2018	Orlandi et al. [51]	127 (16)	Optical flow	9 global features	CML: Random forest	Leave-one-out CV	44%	99%	0.82
2019	Ihlen et al. [62]	377 (41)	Optical flow	990 MEMD+HHT features	CML: PLSR+LDA	6-fold CV	93%	82%	0.87

Abbreviations: SE, sensitivity; SP, specificity; AUC, area under the receiver operating characteristic curve; NR, not reported; CV, cross validation; C_{SD}, centroid of motion standard deviation; Q_{mean}, quantity of motion mean; Q_{SD}, quantity of motion standard deviation; A, area out of standard deviation from moving average; P, periodicity; CC, correlation coefficient; FFT, fast Fourier transform; MEMD, multivariate empirical mode decomposition; HHT, Hilbert-Huang transform; CML, conventional machine learning; SVM, support vector machine; PLSR, partial least squares regression; LDA, linear discriminant analysis.

1.3.1 Video-based motion capture

Frame differencing

Whereas early studies on machine learning-based CP prediction, relied on the simple frame differencing technique for motion capture in GMA recordings (step 1, Figure 1.2) [48–50, 57], recent approaches have utilized more advanced video-based motion capture technologies. Since frame differencing pays equal attention to all pixel changes that appear between video frames, it is highly prone to disturbances (e.g., varying lighting conditions and background distractions) not associated with the infant movements.

Optical flow

The use of optical flow, first by Stahl et al. [58] and later in various studies [51, 59–62], enables grouping of pixel changes that relate to each other, and hence better distinguishes infant movements from noise and irrelevant video information. Furthermore, since most movements happen in limbs (i.e., arms and legs) and these often move independently of each other, optical flow can be used to divide the infant body into a set of segments and analyze movements of different limbs separately [63]. Despite this progress, optical flow is limited to a coarse segmentation of the body into a few body segments (e.g., head, arms, legs, and trunk). Moreover, optical flow collapses in case of occluded body parts, which introduced a need for regular manual segmentation to avoid losing movements of certain body parts [63].

Image-based convolutional networks

These limitations were addressed by the use of pose estimation, in particular the ConvNet-based framework called OpenPose [64]. This enabled fully automated estimation of an infant skeleton (i.e., infant pose estimation) in each video frame of a GMA recording, containing positional information of 18 predefined body keypoints [53–56]. Accordingly, the infant movements in a video can be represented as a sequence of detailed infant skeletons. However, the OpenPose method, trained and validated on images of adults [64], is not suited for the anatomical proportions of infants [46], which differ significantly from those of adults [65]. Only the recent study by Nguyen-Thai et al. [54] has utilized a modification of OpenPose adapted for infants [66]. However, the ability of OpenPose to precisely estimate body keypoint positions in videos containing a single infant, with limited computational budget available, is still questionable, due to its architectural components developed for multi-person human pose estimation (HPE) [64]. Hence, there is a need for developing ConvNets for single-person HPE, and perform retraining on infant images representative of the variation in GMA recordings, to systematically investigate the localization performance and computational efficiency of ConvNets in relation to OpenPose on infant pose estimation.

1.3.2 Prediction model

Conventional machine learning

The prediction models (step 2, Figure 1.2) of initial studies on machine learning-based CP prediction harnessed expert-based simple global features, such as the standard deviation of the centroid of motion (C_{SD}) or mean of quantity of motion (Q_{mean}), in combination with a single cut-off value or linear classifier to distinguish between infants with and without CP [48–50, 57]. It was believed that these features could cover important aspects about pathological movements. For example, it was suggested that higher C_{SD} could reflect a monotonous and stereotyped movement pattern in infants with absent FMs [48]. Despite the promising predictive accuracy of C_{SD} reported by Adde et al. [48], in a larger sample of infants Støen et al. [50] displayed more modest results and suggested that improved accuracy could be achieved by including additional features, such as frequency of limb movements.

Rahmati et al. [60, 61] proposed the use of the fast Fourier transform to extract frequency components associated with the movements of individual body parts, resulting in a total set of 2 376 features. Ihlen et al. [62] extended upon this by using multivariate empirical mode decomposition and Hilbert-Huang transform to capture dynamics of body part movements in the time-frequency domain. The larger feature sets introduced a need for classification methods that could reduce the dimensionality of provided features by retaining only components explaining most of the variance between classes. For this purpose, conventional machine learning with partial least squares regression was used to compress a feature set into a few latent variables [60–62]. However, although the proposed features in these studies express CP-related movements, as reflected by high predictive values for automated CP prediction (see Table 1.2), these features were also selected based on human assumptions, which does not guarantee that all movement information relevant for CP prediction is retained.

Graph-based convolutional networks

The ability of convolutional networks to extract complex task-relevant features, without human involvement, suggests potential for automatically discovering quantitative movement features directly related to CP outcome. This implies an unbiased search for infant spontaneous movements, whether FMs or patterns of the concurrent motor repertoire, that discriminate between infants with and without CP. Furthermore, the high capacity (i.e., number of parameters) of convolutional networks, compared to conventional machine learning methods, may yield improved predictive values for CP. Graph-based convolutional network (GCN) represent an interesting alternative towards achieving this, by detecting complex movement gestalts through explicit modeling of infant skeleton sequences, obtained by video-based motion capture, as spatiotemporal graphs. The alternating spatial and temporal operations in GCNs, which combines analysis of postural features across body keypoints within a single video frame and movement features of body keypoints across video frames, respectively [44], could enable detection of complex spatiotemporal features of coordinative whole-body movements (e.g., FMs) related to the prediction of CP. A few studies conducted simultaneously to ours have harnessed convolutional networks, including GCNs, in infant movement analysis to automatically quantify and classify FMs from infant skeleton sequences [54, 56]. However, no studies have thus far used convolutional networks for CP prediction by also including the concurrent motor repertoire of infant spontaneous movements.

Chapter 2

Aims of the thesis

The overall aim of this thesis was to propose convolutional networks for analysis of infant spontaneous movements in videos to achieve early objective prediction of CP. Two specific aims were pursued:

Aim I

Develop and validate image-based convolutional networks (ConvNets) to obtain motion capture precisely and efficiently estimating positions of infant body parts in video recordings.

Aim II

Develop and validate graph-based convolutional networks (GCNs) to obtain prediction model for CP in high-risk infants from spontaneous movements at 3 months age.

Chapter 3

Methods

3.1 Research overview

The overview of the research conducted in this thesis is presented in Figure 3.1. The research was divided into two main endeavors. The first was concerned with Aim I of the thesis, namely the development of ConvNets for motion capture of infants in video recordings. To achieve this, two specific subgoals were pursued. Aim Ia was to address the lack of demonstrated computational efficiency and localization performance among existing ConvNets for single-person HPE by proposing novel ConvNets. This formed the basis for Study I, comprising Paper I. Subsequently, Aim Ib was to retrain and evaluate the developed ConvNets for single-person HPE on a novel dataset of infant images to obtain feasible ConvNets for infant pose estimation in video-based motion capture. This was carried out in Study II and associated results presented in Paper II. The second main endeavor of the thesis involved the development and validation of a GCN-based prediction model for CP to address Aim II of the thesis. More specifically, we investigated how GCNs could harness the spatiotemporal graph structure of infant skeletons, obtained with motion capture in GMA recordings, to detect whole-body spontaneous movement features relevant for the prediction of CP. Study III included the development and associated validation of the prediction model, which was described in Paper III.

In the remainder of this chapter, we will first formally introduce the principle of convolution and relevant works on ConvNets and GCNs. Thereafter, we will explain our approach for developing and validating convolutional networks related to the specific studies of the present thesis, namely single-person pose estimation (Study I), infant pose estimation (Study II), and prediction model for CP (Study III).

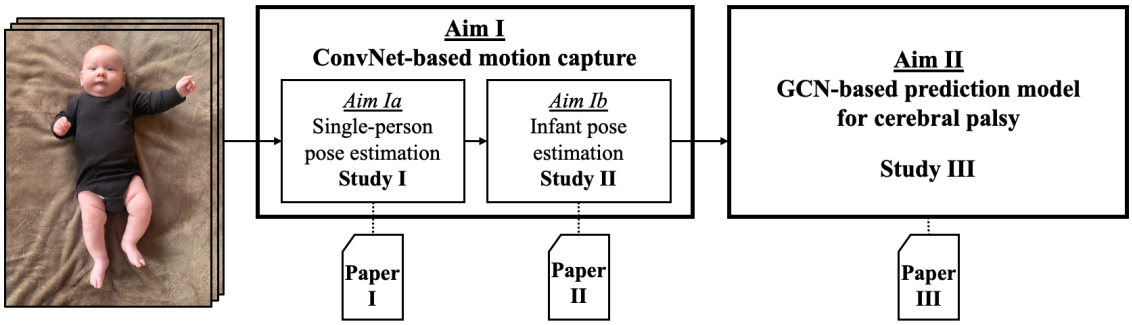


Figure 3.1: The thesis conducted research for developing convolutional networks for infant movement analysis for the purpose of automated CP prediction from videos of infant spontaneous movements at 3 months age. This included three studies: Study I with development of ConvNets for single-person HPE to address Aim Ia, which comprised Paper I, Study II with retraining and evaluation of the proposed single-person HPE ConvNets on infant pose estimation to target Aim Ib, which was described in Paper II, and Study III with development and evaluation of GCN-based prediction model for CP to achieve Aim II, which resulted in Paper III.

3.2 Preliminaries

3.2.1 Convolution

The protagonist of the methods developed in this thesis is the convolution operation. Based on some specific input X, for example an image, convolution performs a local calculation at each position (i, j) in the image (i.e., pixel), which takes into account not only the value of the associated pixel but also the values of neighboring pixels:

$$F(i,j) = (X * W)(i,j) = \sum_{v=0}^{V} \sum_{h=0}^{H} X(i+v, j+h) W(v,h) \tag{3.1}$$

As defined in Equation 3.1 and depicted by Figure 3.2, a weighted average $F(i, j)$ is computed, where a weight matrix W, called a kernel, determines the size of the neighborhood (i.e., kernel size) in vertical (V) and horizontal (H) direction. This is referred to as $V \times H$ convolution. The weights of the kernel specify which pattern (i.e., feature) in the input it should emphasize. In this way, convolution acts as a local feature detector, yielding a feature map F expressing in which locations of the image the feature has highest response. By employing multiple kernels of different weights, a single convolutional layer generates several feature maps, associated with different features (e.g., vertical and horizontal lines in Figure 3.2). The number of feature maps is referred to as the width of the ConvNet. To determine which specific features should be detected by a convolutional layer for a certain task, kernel weights are tuned through training on task-specific data. As illustrated by Figure 3.3, when convolutional layers are stacked on top

of each other in a ConvNet, more global features are detected due to increased receptive field, which facilitates automatic learning of complex patterns in data. The number of convolutional layers in a ConvNet is often referred to as the depth of the network. In 2012, Krizhevsky et al. [67] demonstrated the success of ConvNets on image classification, and few years later families of deeper ConvNets, like VGG [68] and ResNets [69], substantially improved upon this. However, the increasing depth and high complexity of these ConvNets do not align well with requirements for computational efficiency in real-world applications [70].

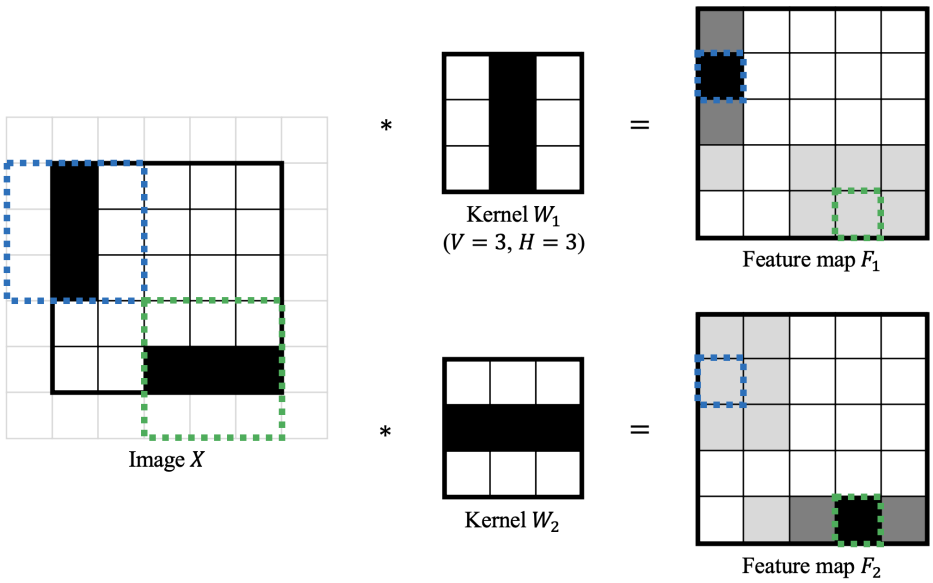


Figure 3.2: The convolution operation applied to an image X, where kernels W_1 and W_2, with vertical (V) and horizontal (H) kernel size of 3, detect specific features in the image (i.e., vertical line and horizontal line, respectively), yielding feature maps F_1 and F_2. To achieve resolution of feature maps consistent with the input image, zero padding in each direction is initially applied to the image, as depicted by white pixels with grey borders. The pixel with the highest value in F_1 (i.e., the black pixel with blue dotted border) represents the location in the image where the vertical line feature is detected by W_1, as reflected by the blue dotted region of the padded image. Similarly, the location associated with the green region in the padded image yields highest value in F_2 (i.e., the black pixel with green dotted border), as reflected by the presence of the horizontal line feature. Dark gray or light gray pixels in F_1 and F_2 represent regions in the padded image containing only parts of the features associated with W_1 and W_2, respectively.

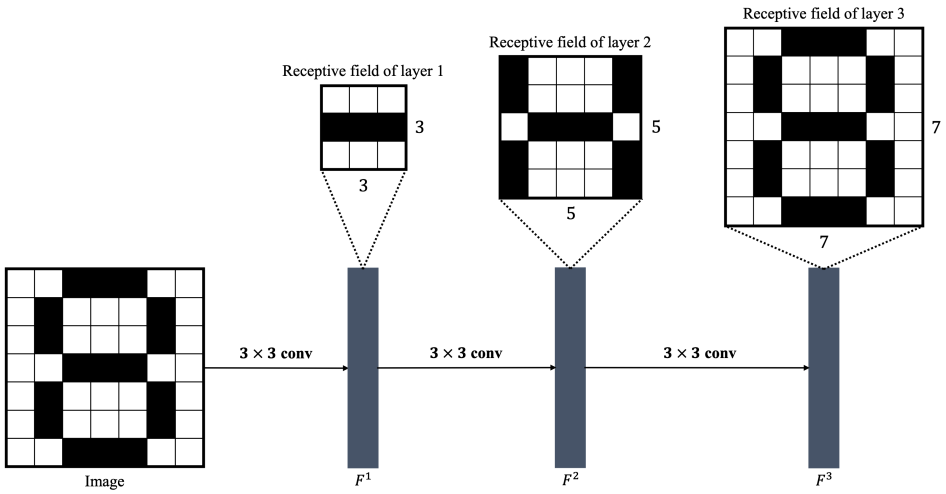


Figure 3.3: A ConvNet with several consecutive convolutional layers (e.g., 3 × 3 convolutions) gradually increases the size of the receptive field, from small receptive field in the feature maps of layer 1 (i.e., F^1) permitting detection of only simple features in local image regions (e.g., a horizontal line), to larger receptive fields in feature maps of later layers (i.e., F^2 and F^3 for layer 2 and layer 3, respectively), which eventually cover the whole image and accordingly enable detection of global features (e.g., the number eight in layer 3).

3.2.2 EfficientNets

Howard et al. [70] proposed MobileNet, employing a more computationally efficient type of convolutional layer, called depthwise separable convolution, which modifies the basic convolution in VGG (Figure 3.4a) and bottleneck convolution in ResNets (Figure 3.4b). As depicted by Figure 3.4c, the depthwise separable convolution first performs a depthwise convolution, with each kernel operating on a separate channel (i.e., feature map) in the input. The depthwise convolution is followed by a 1 × 1 (i.e., pointwise) convolution to integrate information across channels. MobileNet achieved similar accuracy to VGG-16 on the ImageNet benchmark for image classification, while reducing the number of parameters in the ConvNet by 33 times, from 138 to 4.2 million, and the number of floating-point operations (FLOPs) by 27 times, from 15 to 0.6 billion. Despite this improvement in computational efficiency, the MobileNet architecture was manually designed based on human heuristics. Considering the infinite number of possible configurations of a ConvNet, this makes it practically impossible for a human to find the single architecture providing the optimal balance between accuracy and computational efficiency.

Tan et al. [71] partitioned a ConvNet into blocks of consecutive layers and harnessed neural architecture search (NAS) on a novel search space to automatically determine the components of each block, by simultaneously optimizing ImageNet accuracy and inference latency on the CPU of a Pixel 1 phone. MnasNet was selected from a search through 8 000 different ConvNet architectures, and outperformed MobileNet both in terms of accuracy and computational efficiency [71]. The blocks of MnasNet extensively employ mobile inverted bottleneck convolution (MBConv) [72] with squeeze-and-excitation (SE) [73]. As illustrated by Figure 3.4d, MBConv is an extension of depthwise separable convolution with capacity to detect more fine-grained features than regular depthwise separable convolution by increasing the number of feature maps using computationally efficient 1×1 convolution, typically by three or six times, referred to as MBConv3 and MBConv6, respectively. SE employs channel-wise attention to perform recalibration of the features generated by the depthwise convolution by using global information to ensure informative features are emphasized. Moreover, in MnasNet MBConvs are accompanied by residual connections, which improve propagation of information across layers [69].

As a generalization of the NAS behind MnasNet, Tan and Le [74] performed optimization with the hardware-agnostic measure of FLOPs rather than device-specific inference latency. This resulted in the development of EfficientNet-B0, comprising only 0.4 billion FLOPs [74]. Despite the very low complexity, EfficientNet-B0 surpassed the accuracy of ResNet-50 [69, 74]. Consequently, Tan and Le investigated whether further improvement in accuracy was possible by scaling up EfficientNet-B0. To properly balance the different dimensions of the ConvNet, namely depth, width, and image resolution, compound scaling was proposed [74]. Compound scaling determined optimal scaling coefficients for depth, width, and resolution, denoted α, β, and γ, respectively, with the constraint that $\alpha \cdot \beta^2 \cdot \gamma^2 = 2$. In other words, by employing scaling coefficients on EfficientNet-B0, a more accurate EfficientNet-B1 with twice as many FLOPs as EfficientNet-B0 was obtained. By assuming the relationship between scaling coefficients holds for more complex models, scaling of EfficientNet-B1 resulted in EfficientNet-B2, and so forth. From this compound scaling, a family of eight EfficientNets was developed, from EfficientNet-B0, the most computationally efficient, to EfficientNet-B7, the most accurate, serving various computational budgets and accuracy requirements. Furthermore, for the most lightweight EfficientNets (EfficientNet-B0 through EfficientNet-B4) to run efficiently on edge devices, EfficientNet-Lite models were developed [75]. Recently, Song et al. [76, 77] demonstrated that advancements related to ConvNets, in general, and EfficientNets, in particular, are transferable to GCNs, by first extending ResNet into ResGCN [76], and thereafter developing EfficientGCNs combining compound scaling with consistent use of MBConvs [77].

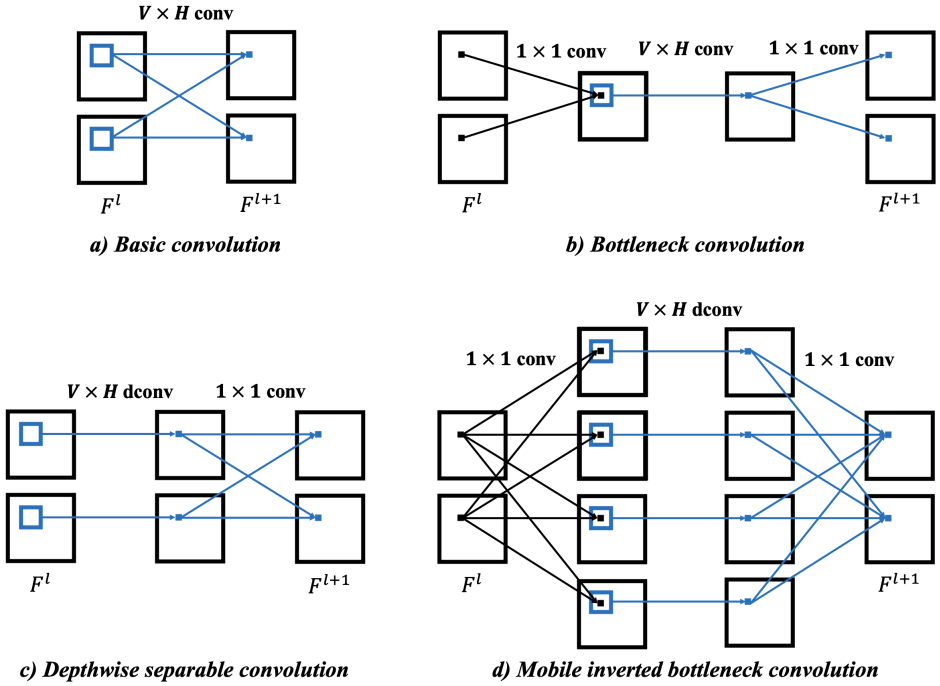


Figure 3.4: Four prominent types of convolutional layers operating on feature maps of layer l (i.e., F^l) to generate feature maps of layer $l+1$ (i.e., F^{l+1}). a) Basic convolution applies convolution once, with kernel size $V \times H$, where each generated feature map is computed from all feature maps of layer l. b) Bottleneck convolution first reduces the number of feature maps with 1×1 convolution (e.g., from two to one), before applying basic $V \times H$ convolution, followed by another 1×1 convolution restoring the number of feature maps. c) Depthwise separable convolution performs depthwise convolution (i.e., dconv), by $V \times H$ kernels operating on feature maps in layer l separately, followed by 1×1 convolution to integrate information across feature maps, yielding F^{l+1}. d) Mobile inverted bottleneck convolution first applies 1×1 convolution to increase the number of feature maps (e.g., from two to four), before performing $V \times H$ depthwise convolution, and another 1×1 convolution to restore the number of feature maps. To emphasize the differences between the four convolutional layers, other operations that are commonly included in convolutional layers, such as batch normalization and nonlinear activations, are omitted in the visualization.

3.2.3 Graph convolution

The main distinction between ConvNets and GCNs is the introduction of graph convolution. Figure 3.5 depicts the principle of the graph convolution operation. Graph convolution harnesses the structure of a graph G, like the infant skeleton model in Figure 3.5, to analyze spatial dynamics of features of neighboring nodes (i.e., body keypoints) using adjacency matrices. More specifically, based on some input X (e.g., biomechanical properties like positions and velocities), 1×1 convolution first increases the number of feature maps by N times ($N = 3$ in Figure 3.5). Subsequently, an equal number of the resulting features are processed by N parallel branches, where each branch performs multiplication with a distinct adjacency matrix A. The sum of the N matrix multiplications yields feature maps F^1 of layer 1. Adjacency matrices determine the type of spatial dynamics that are analyzed. For example, Figure 3.5 illustrates disentangled aggregation graph convolution [78] where an adjacency matrix $A_{(k)}$ defines the k-hop neighbors of body keypoints in the infant skeleton model, as well as the identity matrix (i.e., $A_{(0)}$). We refer the reader to Yan et al. [44] for further description of graph convolution.

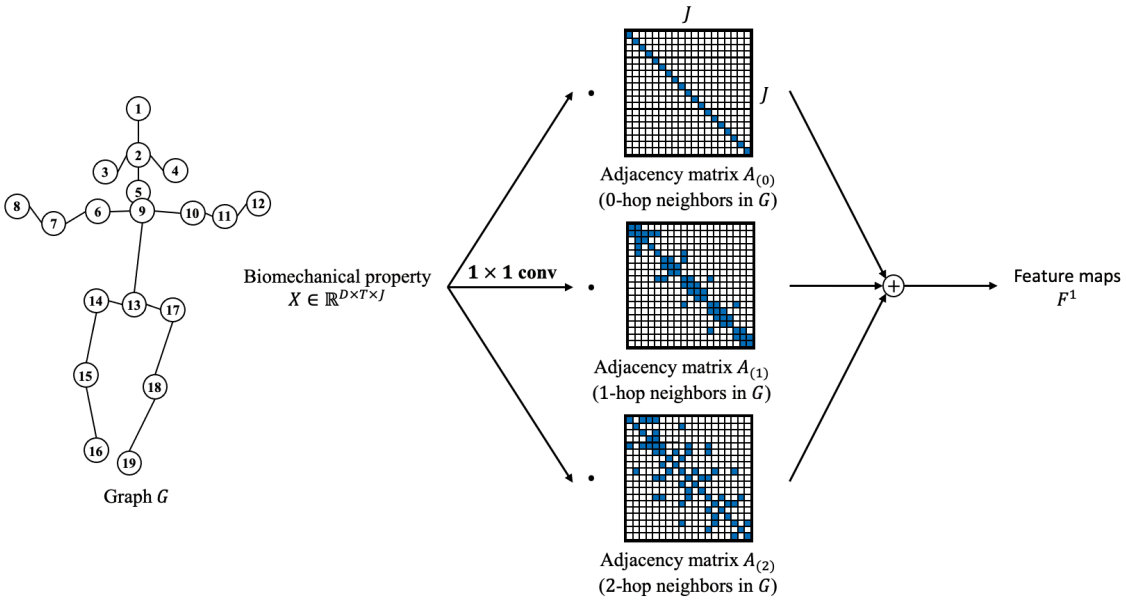


Figure 3.5: Graph convolution analyzes spatial dynamics of biomechanical properties X of body keypoints in graph G through 1×1 convolution followed by multiplication with a set of adjacency matrices $A_{(k)}$ (i.e., k-hop neighbors in G). The resulting products of the matrix multiplications are aggregated into feature maps of layer 1 (i.e., F^1).

3.3 Study I: Single-person pose estimation

3.3.1 MPII Human Pose Dataset

To develop and evaluate ConvNets for single-person HPE, experiments were conducted on the single-person benchmark of the openly available MPII Human Pose Dataset [79], referred to as MPII. MPII comprises images and associated annotations (x and y coordinates) of 16 body keypoints (i.e., head top, upper neck, shoulders, elbows, wrists, upper chest, right/mid/left pelvis, knees, and ankles) of mainly healthy adults in over 800 different outdoor and indoor activities extracted from public YouTube videos. The benchmark contains 28 880 (80%) images for training and validation, as well as a test set of 7 247 (20%) separate images for official evaluation.

3.3.2 Experimental approach

The training and validation portion of the MPII dataset were randomly split into datasets for training, 26 379 (91%) images, and validation, 2 501 (9%) images, while ensuring all frames (i.e., images) of a single video were placed into one of these datasets. ConvNets were proposed by combining transfer learning of state-of-the-art ConvNets (i.e., backbones) on ImageNet with novel architectural components targeting HPE (see Section 3.3.3 for more details on the proposed ConvNets). By employing supervised learning with stochastic gradient descent (SGD) on the training set, the ConvNets were tuned for HPE to optimize predictions of body keypoint locations relative to target coordinates. To avoid overfitting, data augmentation with horizontal flipping, rotation ($+/-$ 45 degrees), and scaling ($0.75-1.25$) was performed. Hyperparameters of the ConvNets and the training procedure (e.g., learning rate, batch size, and number of epochs) were determined from localization performance on the validation set. All experiments were carried out on an NVIDIA Tesla V100 GPU. Further details on the optimization procedure are described in Appendix B of Paper I.

3.3.3 EfficientPose and EfficientHourglass

EfficientPose

To address the limitations of OpenPose [64] regarding single-person HPE, we proposed EfficientPose, which modified several components of the ConvNet architecture of OpenPose, as depicted by Figure 3.6a and b. First, EfficientPose processes in separate branches two different resolutions of the input image (step 1, Figure 3.6b). A high-level branch, operating on a high-resolution image, has capacity to detect fine-grained features in the input image (e.g., small variations in nearby pixels), whereas a low-level branch, associated with a low-resolution image (i.e., half the height and width of the high-resolution image), detects less detailed image features. Second, the VGG-19 [68] backbone of OpenPose was replaced by the more accurate and computationally efficient EfficientNet [74] backbones (step 2,

Figure 3.6b). To enable detection of high-level semantic information from the fine-grained features in the high-level branch, the initial three blocks of an EfficientNet, with a resolution matching the input image, were included. For the low-level branch, the first two blocks of a lower-scale EfficientNet model were used. To facilitate detection of generic visual features, the EfficientNet backbones were initialized with pretrained ImageNet weights. Third, motivated by the success of multi-scale feature extraction in HPE [80, 81], the high-level and low-level features of the two EfficientNet backbones were concatenated into cross-resolution features (step 3, Figure 3.6b). This enables selective emphasis of image features of different abstraction levels, to yield an effective multi-scale feature extractor. Fourth, we modified the computationally expensive detection stage of OpenPose by reducing the number of detection passes from six to three while replacing basic 3×3 convolutions in the DenseNet inspired detection blocks [82] with more efficient E-MBConv6 (i.e., an adaptation of MBConv6 with a fixed number of feature maps in the depthwise convolution which employs E-swish activation [83]), yielding Mobile DenseNets (step 4, Figure 3.6b). Fifth, EfficientPose includes a stack of three 4×4 transposed convolutions [84] to improve the level of detail in the low-resolution heatmaps (i.e., output) of the final detection pass through upscaling with bilinear interpolation (step 5, Figure 3.6b).

In accordance with the compound scaling proposed by Tan and Le [74], the architecture of EfficientPose was scalable. By employing the scaling coefficients of EfficientNet regarding depth, width, and image resolution, five different variants of EfficientPose were developed. The models EfficientPose I, II, III, and IV gradually increase the input resolution, from 256×256 to 600×600 pixels, as well as the depth and width of the backbones and detection blocks, in relation to the most computationally efficient model, EfficientPose RT. EfficientPose RT comprises a single-resolution model matching the scale of EfficientNet-B0 with input resolution of 224×224 pixels. Apart from the five main EfficientPose variants presented in Paper I, this thesis also includes three additional EfficientPose models, EfficientPose RT Lite, I Lite, and II Lite, targeting deployment on edge devices. In EfficientPose Lite, the backbones of their original counterparts (i.e., EfficientPose RT, I, and II) were substituted with matching EfficientNet-Lite backbones [75], while omitting the low-level branch, cross-resolution feature extraction, and SE modules, and E-swish activations were replaced by ReLU6 [85]. Implementations of the EfficientPose models in common deep learning frameworks are made publicly available at `https://github.com/daniegr/EfficientPose`. See Paper I for technical details on EfficientPose.

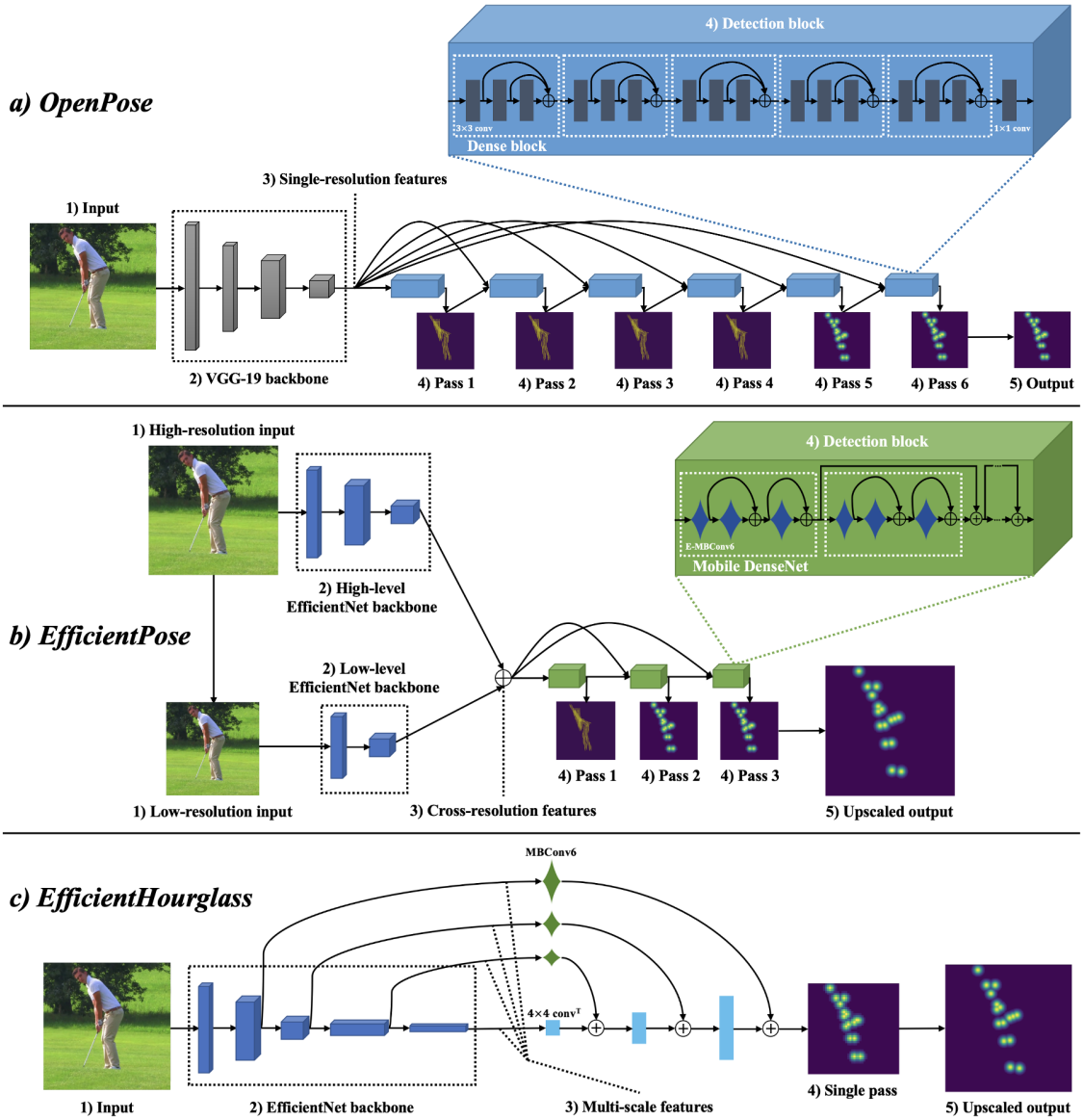


Figure 3.6: An overview of the architectural differences between a) OpenPose, b) EfficientPose, and c) EfficientHourglass. a) OpenPose utilizes 1) a single input image, 2) VGG-19 backbone, 3) single-resolution features, 4) six detection passes of Dense blocks with basic 3×3 convolutions, and 5) low-resolution output. b) EfficientPose modifies OpenPose by harnessing 1) both high-resolution and low-resolution input images, 2) high-level and low-level EfficientNet backbones, 3) cross-resolution features, 4) three detection passes of Mobile DenseNets with E-MBConv6, and 5) high-resolution upscaled output. c) EfficientHourglass employs 1) a single high-resolution input image, 2) EfficientNet backbone, 3) multi-scale features, which are upscaled with 4×4 transposed convolutions (i.e., conv^T), 4) a single detection pass, and 5) high-resolution upscaled output. Extension of Fig. 1 and 2 of Paper I.

EfficientHourglass

In addition to EfficientPose, we also proposed a ConvNet architecture motivated by the design of state-of-the-art ConvNets on the single-person MPII benchmark [86–88]. More specifically, the multi-scale hourglass architecture of Newell et al. [80] was modified into EfficientHourglass, by exploiting more extensively the generic visual features of an ImageNet-pretrained EfficientNet backbone related to an input image (step 1 and 2, Figure 3.6c). Local image features of high spatial resolution from the initial blocks of the EfficienNet were combined with global image features of lower spatial resolution in later blocks to construct a multi-scale feature extractor integrating features from four different scales (step 3, Figure 3.6c). In contrast to the original multi-scale hourglass based on bottleneck convolutions, EfficientHourglass consistently employs MBConvs with integrated SE and residual connection, which reduces the computational complexity. In the present thesis, EfficientHourglass is paired with an EfficientNet-B4 backbone, yielding EfficientHourglass B4. To facilitate estimation of body keypoint positions with sufficient level of detail, the proposed variant uses a high input resolution of 608×608 pixels, as opposed to 368×368 pixels with OpenPose. Moreover, similarly to EfficientPose, the network output is upscaled, using bilinear interpolation with two 4×4 transposed convolutions, but EfficientHourglass only requires a single detection pass (step 4 and 5, Figure 3.6c). Due to most network parameters originating from the pretrained EfficientNet-B4 backbone, EfficientHourglass B4 was trained with a standardized procedure using the Adam optimizer with a learning rate of 0.001 for 100 epochs, instead of the optimization procedure described in Section 3.3.2 specialized for ConvNets with detection blocks of mostly randomized weights (e.g., OpenPose and EfficientPose).

3.3.4 Evaluation

The localization performance of the EfficientPose models, EfficientHourglass B4, and the existing OpenPose ConvNet were evaluated on single-person MPII, in terms of the percentage of predictions of body keypoint locations within a fraction τ of the head size from annotated positions (i.e., ($PCK_h@\tau$) in Fig. 4b of Paper I). The ability of the ConvNets for coarse localization was measured by $PCK_h@0.5$, setting the threshold τ at 50% of the head size. On the other hand, $PCK_h@0.1$ assessed fine localization performance, reflected by a smaller acceptable level of error (i.e., 10% of the head size). A comparison of $PCK_h@0.5$ and $PCK_h@0.1$ of all ConvNets were performed on the MPII validation set, by employing multi-scale testing as commonly done in HPE benchmarking [81, 89]. Furthermore, predictions of EfficientPose RT, EfficientPose IV, EfficientHourglass B4, and OpenPose on the MPII test set were formally submitted to yield official evaluation of ConvNets in relation to other state-of-the-art methods for single-person HPE. The computational efficiency of ConvNets were measured in terms of computational complexity and model capacity, with FLOPs and number of parameters, respectively.

3.4 Study II: Infant pose estimation

3.4.1 In-Motion Poses

Video database

To adapt ConvNets for HPE to the anatomical proportions of 3-month-old infants and recordings setups of GMA to perform high-precision infant pose estimation, we harnessed a large international database of 1 424 recordings at 9-18 weeks PTA following GMA standards [18]. The database comprised video recordings of 1-9 minutes of infants with different medical risk factors (e.g., high-risk infants and typically developing infants) from standardized and less standardized setups at hospital as well as home-based smartphone recordings [31, 33]. These were collected through research initiatives in Norway, India, United States, Turkey, Belgium, Denmark, and Great Britain between September 2001 and September 2018. The use of videos for machine learning-based CP prediction was approved by the regional committee for medical and health research ethics in Norway, under reference numbers 2011/1811 and 2017/913, with written parental consent obtained before inclusion.

Datasets

From these videos, 20 000 video frames were extracted to compose the In-Motion Poses dataset (see Figure 3.7a for a selection of representative images). To ensure that all recording setups were sufficiently represented, a fixed portion of frames from each setup was included, 40% and 20% from standardized and less standardized hospital recordings, respectively, and 40% from home-based smartphone recordings. Within each setup, 80% of frames were randomly selected with an equal number of frames from each video. The remaining 20% of frames were manually selected to cover infant poses that are normal but less frequently occurring, based on the following criteria: 1) legs moving towards upper body, 2) overlap of body parts, and 3) crossing of body parts. See Table 3.1 for an overview of the number of videos and associated frames included in In-Motion Poses from each country and recording setup. Following a similar data split as in MPII, the total of 20 000 images were divided into 80% for training and validation, with 14 483 (72%) and 1 493 (8%) images in the training set and validation set, respectively, and 4 024 (20%) into test set. To ensure strict evaluation, the frames of a single video was represented in only one of these sets.

a) In-Motion Poses *b) Annotated skeleton model*

Figure 3.7: a) A selection of representative images in In-Motion Poses from standardized hospital recordings (top row), less standardized hospital recordings (middle row), and home-based smartphone recordings (bottom row). b) Annotated body keypoints and associated skeleton model. Adapted from Fig. 1 and 7 of Paper II.

Table 3.1: Contents of In-Motion Poses in terms of number of videos and video frames for each country and recording setup (i.e., standardized hospital recording, less standardized hospital recording, and home-based smartphone recording).

Setup	Country	Number of videos	Number of frames
Standardized hospital recording	India	418 (29.4%)	3 037 (15.2%)
	Norway	309 (21.7%)	2 153 (10.8%)
	United States	281 (19.7%)	2 114 (10.6%)
	Turkey	62 (4.4%)	404 (2.0%)
	Belgium	39 (2.8%)	292 (1.5%)
Less standardized hospital recording	Norway	137 (9.6%)	2 503 (12.5%)
	Turkey	62 (4.4%)	1 128 (5.6%)
	Great Britain	19 (1.3%)	359 (1.8%)
	Belgium	1 (0.1%)	9 (0.1%)
Home-based smartphone recording	Belgium	49 (3.4%)	4 100 (20.5%)
	Denmark	31 (2.2%)	2 622 (13.1%)
	Norway	16 (1.1%)	1 279 (6.4%)
Total		*1 424*	*20 000*

Table 3.2: The set of 19 body keypoints in In-Motion Poses and their associated definitions. Adapted from Table 4 of Paper II.

#	Body keypoint	Definition
1	Head top	Top of the forehead
2	Nose	Tip of the nose
3	Right ear	Center of the right ear
4	Left ear	Center of the left ear
5	Upper neck	Center of the larynx
6	Right shoulder	Center of the right shoulder joint
7	Right elbow	Center of the right elbow joint
8	Right wrist	Center of the right wrist joint
9	Upper chest	Midway between body keypoints 6 and 10
10	Left shoulder	Center of the left shoulder joint
11	Left elbow	Center of the left elbow joint
12	Left wrist	Center of the left wrist joint
13	Mid pelvis	Midway between body keypoints 14 and 17
14	Right pelvis	Right spina iliaca anterior superior
15	Right knee	Center of the right knee joint
16	Right ankle	Center of the right ankle joint
17	Left pelvis	Left spina iliaca anterior superior
18	Left knee	Center of the left knee joint
19	Left ankle	Center of the left ankle joint

Annotation

To enable training and evaluation of ConvNets on In-Motion Poses, the 20 000 images were annotated by 10 trained humans, to yield ground truths for x and y coordinates of body keypoint positions. A group of human movement scientists and clinical physiotherapists agreed upon a set of 19 body keypoints (i.e., head top, nose, ears, upper neck, shoulders, elbows, wrists, upper chest, right/mid/left pelvis, knees, and ankles), defined in Table 3.2, comprising an infant skeleton model (see Figure 3.7b). Based on these definitions, body keypoints were labelled using an adapted version of a separate annotation software [90]. Furthermore, by comparing annotation consistency of humans across 100 randomly selected frames, we computed the inter-rater annotation spread in terms of mean Euclidean distance to the average human annotation.

3.4.2 Experimental approach

The images and associated annotations of In-Motion Poses were harnessed to adapt MPII-trained ConvNets to infant pose estimation. A standardized supervised learning procedure, using the Adam optimizer with a learning rate of 0.001 for 100 epochs, performed fine-tuning of ConvNets on the training set while monitoring progress on the validation set. Data augmentation with random horizontal flipping, rotation ($+/-$ 45 degrees), and scaling ($0.75 - 1.25$) of images was applied. Models were trained on an NVIDIA GTX 1080 Ti or an NVIDIA Quadro RTX 8000, depending on the amount of GPU memory required.

3.4.3 Comparative analysis

The determine the feasibility of ConvNets for infant pose estimation, a comparison of the performance of nine different alternatives was conducted. First, the official version of the state-of-the-art method OpenPose [64, 91] was tested without fine-tuning to yield baseline performance of infant pose estimation. Thereafter, OpenPose was fine-tuned on In-Motion Poses to determine the effect of training on infant images. We also fine-tuned and analyzed another lightweight ConvNet inspired by OpenPose, called CIMA-Pose, which has previously achieved promising localization performance for infant pose estimation on standardized hospital recordings [90]. CIMA-Pose replaced the VGG-19 backbone of OpenPose with the first two blocks of a DenseNet-121 [82], while reducing the number of detection passes from six to two and making each detection block more computationally efficient by employing dilated convolutions [92]. For more details on the CIMA-Pose ConvNet, we advise the reader to consult the original paper [90]. Lastly, fine-tuned versions of the five main EfficientPose models (i.e., EfficientPose RT and I-IV) and EfficientHourglass B4 were included in the comparative analysis.

3.4.4 Evaluation

The localization performance of the ConvNets for infant pose estimation was evaluated on the test set of In-Motion Poses and was measured in terms of mean error (ME) as well as $PCK_h@\tau$ with different levels of error (i.e., thresholds) τ relative to head length l (i.e., distance between the body keypoints of head top and upper neck), as displayed by Figure 3.8. For coarse evaluation, $PCK_h@1.0$, $PCK_h@0.5$, and $PCK_h@0.3$ were computed, whereas $PCK_h@0.2$ and $PCK_h@0.1$ performed fine-grained evaluation. Furthermore, from the inter-rater spread of body keypoint$b$, denoted H_b, a separate threshold ${H_b}^{0.95}$ was proposed to define $PCK_h@Human^{0.95}$, expressing model performance relative to the performance of humans. More specifically, ${H_b}^{0.95}$ reflects the 95th percentile of the inter-rater spread, and hence $PCK_h@Human^{0.95} = 95\%$ indicates human-level performance. The formal definition of the proposed $PCK_h@Human^{0.95}$ metric is provided in Paper II. $PCK_h@Human^{0.95}$ was estimated for the best performing ConvNet in each model family. To evaluate the computational efficiency of ConvNets, the number of parameters and FLOPs were computed. Moreover, the inference latency of ConvNets in milliseconds and processing speed in frames per second (FPS) on an NVIDIA GTX 1080 Ti consumer GPU estimated the run-time performance of the ConvNets.

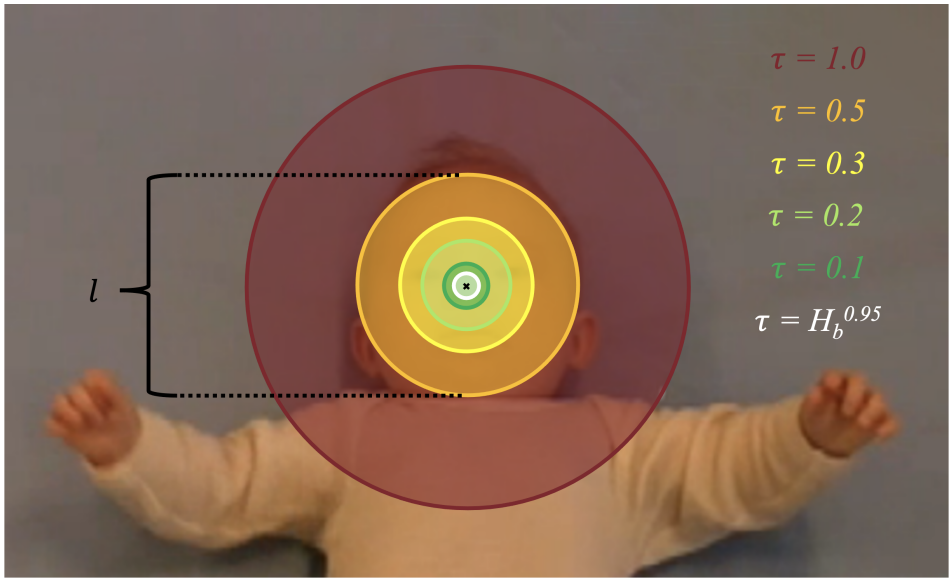


Figure 3.8: The percentage of predictions within τl distance from the ground truth location (i.e., $PCK_h@\tau$) with separate thresholds τ for coarse evaluation ($PCK_h@1.0$, $PCK_h@0.5$, and $PCK_h@0.3$), fine-grained evaluation ($PCK_h@0.2$ and $PCK_h@0.1$), and inter-rater spread ${H_b}^{0.95}$ of body keypoint b (e.g., nose) for evaluating localization performance relative to human-level performance ($PCK_h@Human^{0.95}$). Extension of Fig. 3 in Paper II.

3.5 Study III: Prediction model for CP

3.5.1 Participants

Infant sample

To develop and validate a GCN-based prediction model for CP, we harnessed a sample of 557 high-risk infants (i.e., infants with medical risk factors for CP), who were prospectively enrolled between September 2001 and October 2018 in previous studies from our group [31, 57, 93, 94]. More specifically, this included 248 and 190 infants with heterogeneous high-risk factors (e.g., very/extremely low GA/birth weight, neurological abnormalities, and congenital heart disease) from United States and Norway, respectively, as well as 82 infants with neonatal encephalopathy from India and 37 infants with perinatal stroke from Belgium.

All high-risk infants had been video recorded in a standardized setup during the FMs period at 9-18 weeks PTA following Prechtl's GMA standards [18], and associated GMA classifications had been performed by two experienced observers. Videos were classified based on temporal organization of FMs, including continual FMs (i.e., FMs occur frequently), intermittent FMs (i.e., FMs occur regularly, but less frequently compared to continual FMs), sporadic FMs (i.e., FMs occur only sporadically), and absent FMs (i.e., no FMs) [95], and FMs that appeared exaggerated were classified as abnormal FMs [18].

Moreover, the included high-risk infants had been considered for a diagnosis of CP after 12 months PTA, by a pediatrician following the decision tree of the Surveillance of Cerebral Palsy in Europe (SCPE) [96]. Based on characteristics of symptoms, infants with CP were further divided into different subtypes, namely spastic unilateral CP, spastic bilateral CP, dyskinetic CP, and ataxic CP [96]. The severity of CP, in terms of level of motor function, was determined by the Gross Motor Function Classification System (GMFCS) [97], from level I (i.e., the mildest form of CP) to level V (i.e., the most severe form of CP). GMFCS I, II, and III represent infants with ability to walk (i.e., ambulatory CP), whereas GMFCS IV and V constitute non-ambulatory CP. Specific clinical characteristics of high-risk infants are presented in eAppendix 2 of Paper III.

For each participating infant, a single video recording following GMA standards was included for further analysis. Ethical approval for development and evaluation of machine learning-based CP prediction from video recordings and CP outcomes of infants was provided by the regional committee for medical and health research ethics in Norway under reference number 2011/1811, and parental consent was obtained before inclusion.

Datasets

Datasets for method development (i.e., training and internal validation) and external validation were composed of high-risk infants stratified on center (step 1, Figure 3.9) and CP subtype (step 2, Figure 3.9). For each class (i.e., stratum) of infants, 418 (75%) infants were randomly placed into dataset for method development (blue path in step 3, Figure 3.9), whereas the remaining 139 (25%) comprised test set for external validation (red path). To enable 7-fold cross-validation for assessing internal validity, the 418 infants were randomly placed into seven distinct internal validation folds, comprising nine infants with CP and 50 or 51 infants without CP, based on a similar procedure for stratification on center and CP subtype.

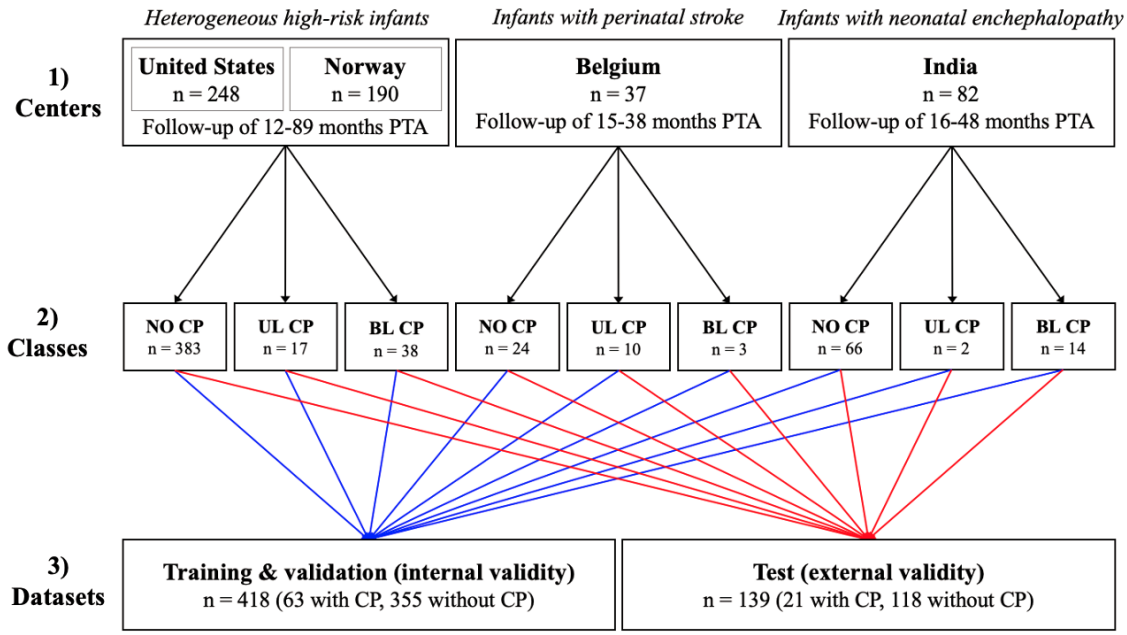


Figure 3.9: Randomization of high-risk infants into dataset for development (training and internal validation) and test set for external validation, from stratification on center and CP subtype (i.e., NO CP for infants without CP and UL CP and BL CP for infants with unilateral CP and bilateral CP, respectively). Adapted from Figure 1 of Paper III.

3.5.2 Experimental approach

Motion capture and pre-processing

The video recordings and associated binary CP outcomes (i.e., classification into CP or no CP) of infants in the method development dataset were utilized to construct a GCN-based CP prediction model. First, the video-based motion capture from Study II was used to estimate, for each frame in a video, an infant skeleton of x and y coordinates of the 19 body keypoints in In-Motion Poses. The infant skeletons of all frames together represented the infant spontaneous movements in a video as a spatiotemporal skeleton sequence. Subsequently, the skeleton sequence was pre-processed, with resampling to 30 Hz, temporal smoothing with 5-point median filter, and standardization of coordinates by centralizing on the median mid pelvis position and normalizing by two times the trunk length of the infant (i.e., median distance from upper chest to mid pelvis). To train and validate GCNs for binary CP classification each skeleton sequence was divided into 5 second windows, deemed to be the minimum time required by a GMA observer to determine whether FMs are present [98], while harnessing the binary CP outcome of the respective infant as a noisy label of the time window. During training, each infant of a particular class (i.e., CP or no CP) had an equal number of time windows, which were randomly selected from different parts of the skeleton sequence. Furthermore, to ensure proper optimization of GCNs, despite low prevalence of CP (15%) compared to no CP (85%), infants with CP had five times more time windows compared to infants without CP.

Optimization procedure

To determine the appropriate optimization procedure for training GCNs with randomized weights (i.e., no pretraining), a simple hyperparameter search was conducted on one internal validation fold (i.e., val1) with the commonly applied ST-GCN model [44]. In particular, we explored type of weight initialization (including LeCun initialization [99], forward and backward cases of He normal initialization [40], and Mean Var initialization [100]), numbers of time windows presented in each epoch of training per non-CP skeleton sequence, settings for data augmentation (i.e., rotation, scaling, and translation), type of optimizer (Adam or SGD), learning rate, and batch size. From this search, a suitable configuration comprised backward case of He normal for weight initialization, 12 time windows per non-CP skeleton sequence for each training epoch, data augmentation with $+/-$ 45 degrees rotation, $0.7-1.3$ scaling, and $+/-$ 0.3 translation, and SGD optimizer with learning rate of $5 \cdot 10^{-4}$ and batch size of 32. Subsequently, we performed 7-fold cross-validation of ST-GCN with the proposed optimization procedure for 200 epochs, yielding baseline performance for GCN-based CP prediction. The results in Table A.5 (Appendix A) suggest that two of the internal validation folds, val2 and val7, were particularly easy and hard to optimize against, respectively.

Neural architecture search

With the results of ST-GCN and the proposed pre-processing and optimization procedure in mind, we turned to the development of novel GCNs for CP prediction. To ensure that GCNs were particularly targeted towards CP prediction from 5 second windows, we proposed an automatic search (i.e., NAS) exploring different configurations of architectures across a variety of conventional and contemporary components of GCNs and ConvNets. More specifically, based on an overall architectural design (Figure 3.10), inspired by Song et al. [76, 77] to predict CP with confidence c from biomechanical properties (i.e., positions, velocities, and bones), 20 architectural choices were investigated. This comprised a search space of more than four billion possible GCN architectures of varying complexity and computational efficiency, with a minimum of five thousand parameters and 0.01 billion FLOPs and a maximum similar to ST-GCN (i.e., three million parameters and six billion FLOPs). The architectural choices included number of network blocks (i.e., depth), number of feature maps (i.e., width), type of convolution (basic, bottleneck, or MBConv), kernel size, activation function, SE, and residual connection. Furthermore, GCN-specific design choices were taken into account, like type of graph convolution (spatial configuration [44] or disentangled aggregation [78]) to integrate information of neighboring body keypoints through adjacency matrices (see Section 3.2.3 for an introduction to graph convolution) and attention mechanism (i.e., no attention or attention on channels, frames, or body keypoints [76]). To yield high-performing GCNs, we developed and employed a novel search strategy, called K-Best Search (see Algorithm 1 in Appendix A). For efficient convergence on an NVIDIA Tesla V100 GPU, K-Best Search exploits architectural choices of GCNs achieving high area under the receiver operating characteristic curve (AUC) across infants in the combined sample of the easy and hard validation folds (i.e., val2 and val7, respectively). The remaining five internal validation folds were used for supervised training of candidate architectures during K-Best Search. K-Best Search was repeated 10 times to yield 10 promising GCNs for CP prediction. Subsequently, 7-fold cross validation, with training for 200 epochs, was performed for each model to yield a total of 70 GCN instances. For further details on the NAS procedure and the obtained models, we refer the reader to Appendix A.

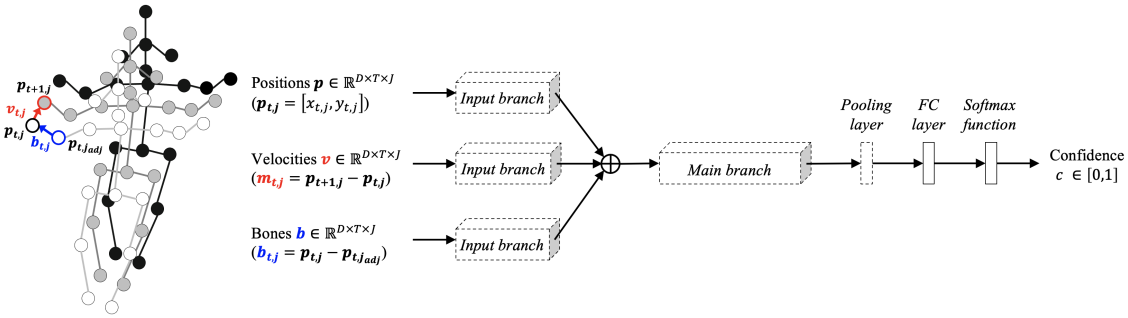


Figure 3.10: The proposed GCNs process biomechanical properties of 5 second windows, including positions p, velocities (i.e., change in position) v, and bones (i.e., distance from the neighboring body keypoint) b, through parallel input branches, main branch, pooling layer, fully connected (FC) layer, and softmax function to yield confidence c from 0.0 to 1.0 for the risk of CP. The number of time steps $T = 150$ (i.e., 30 Hz), the number of body keypoints $J = 19$, and the number of spatial dimensions $D = 2$. Adapted from eFigure 1 of Paper III.

3.5.3 Ensemble-NAS-GCN

Feature extraction

Figure 3.11 provides an overview of the proposed method for prediction of CP. First, the skeleton sequence, obtained from a single GMA recording by harnessing the video-based motion capture proposed in Study II (step 1 and 2, Figure 3.11), is divided into 5 second windows, with 2.5 seconds overlap between each time window. Each 5 second window is thereafter processed individually by all 70 GCN instances obtained with NAS. In a similar manner to how ConvNets operate on images by gradually increasing the receptive field in vertical and horizontal dimensions (see Figure 3.3), GCNs utilize increased receptive field in spatial and temporal dimensions of time windows. Initial GCN layers detect features of movements of neighboring body keypoints over few time steps, whereas later layers generate feature maps related to complex whole-body movements across 5 second windows (step 3, Figure 3.11). Furthermore, GCNs with dissimilar configurations (e.g., kernel size or graph convolution type) differ in how features are extracted from the biomechanical properties, and thus have different starting points for the prediction of CP.

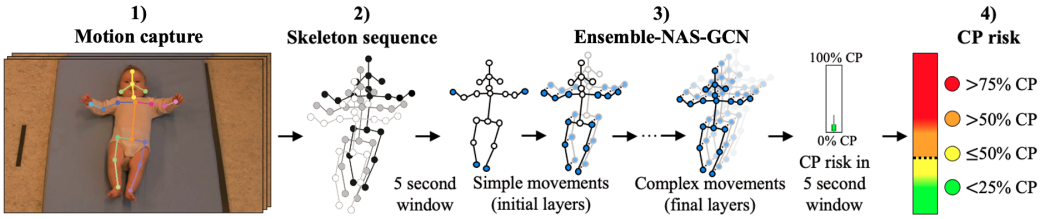


Figure 3.11: The overall procedure of the proposed method for prediction of CP. From left: The motion capture proposed in Study II is used to estimate positions of 19 body keypoints in all video frames (step 1) to comprise a skeleton sequence of infant spontaneous movements (step 2). Each 5 second window of the skeleton sequence is processed by 70 GCN instances obtained with NAS (i.e., Ensemble-NAS-GCN) to extract features, from simple movements in initial layers to complex whole-body movements in later layers, and compute CP risk in the time window as the median of the 70 predictions (step 3). Subsequently, the overall CP risk in the video is estimated as median CP risk across all 5 second windows (step 4), to yield final classification of CP based on a fixed decision threshold (dashed line). CP classification is color coded based on the agreement across GCN instances, with green and yellow as certain and uncertain decision of no CP, and orange and red as uncertain and certain decision of CP. Adapted from Figure 2 of Paper III.

Ensemble modeling

Accordingly, Ensemble-NAS-GCN was constructed as an ensemble model comprising of the 70 GCN instances, each employing a slightly different process to distinguish infants with CP from infants without CP based on the same 5 second window. The median of the 70 individual time window predictions estimates the CP risk in the 5 second window. Furthermore, the boxplot of the 70 predictions is used to determine the uncertainty of the ensemble through color coding (see boxplot in step 3, Figure 3.11), based on the agreement across the ensemble in the number of GCN instances predicting presence of CP (i.e., predictions above a fixed decision threshold). Green ($< 25.0\%$ predict CP) and yellow ($\leq 50.0\%$ predict CP) yield certain and uncertain ensemble prediction of no CP, whereas orange ($> 50.0\%$ predict CP) and red ($> 75.0\%$ predict CP) yield uncertain and certain prediction of CP.

CP risk and classification uncertainty

The final score for CP risk of Ensemble-NAS-GCN in the provided video is computed as the median CP risk across all 5 second windows of the skeleton sequence. An infant is classified into CP if the overall CP risk exceeds a fixed decision threshold, and otherwise Ensemble-NAS-GCN classifies the infant as not having CP. As depicted by step 4 in Figure 3.11, classification into CP is of high certainty (red) if $> 75.0\%$ and of low certainty (orange) if $> 50.0\%$ of the GCN instances agree on the decision of CP, and classification into no CP of low certainty (yellow) if $\leq 50.0\%$ and of high certainty (green) if $< 25.0\%$ classified the infant with CP.

Temporal and spatial explanations

Although GCNs, like other deep learning methods, are black box models, their decisions can be made more transparent through various techniques. In the present thesis, we have developed methods for temporally and spatially explaining the estimated CP risks in 5 second windows. Figure 3.12a demonstrates that box plots associated with each 5 second window can operate as temporal explanations informing about which specific time windows are associated with high (red) or low (green) CP risk. Furthermore, to obtain spatial information about which specific body parts are involved in the movements Ensemble-NAS-GCN associates with CP or no CP, we have used class activation mapping (CAM) [101]. By adopting the implementation of Song et al. [77] targeting GCNs, CAM computes the contribution of each individual body keypoint towards the prediction of CP, with red and green in Figure 3.12b indicating high and low contributions, respectively. The median contribution of each body keypoint across all time windows of a skeleton sequence yields an overall CAM of the respective infant.

3.5.4 Statistical analysis

The predictive accuracy of Ensemble-NAS-GCN for prediction of CP in the test set, comprising high-risk infants for external validation (described in Section 3.5.1), was compared with observational GMA [18], as well as the state-of-the-art conventional machine learning method for automated CP prediction (i.e., CIMA model proposed by Ihlen et al. [62]) and ST-GCN [44]. To ensure fair comparisons, the sensitivity of the methods was fixed at the level of GMA. Apart from sensitivity, we provided measures of specificity, positive and negative predictive value (PPV and NPV), and accuracy, including exact 95% confidence intervals using Clopper-Pearson, as well as receiver operating characteristic (ROC) curves and associated AUC. Furthermore, we analyzed the ability of Ensemble-NAS-GCN to differentiate between infants with ambulatory CP (i.e., GMFCS I, II, or III) and non-ambulatory CP (i.e., GMFCS IV or V), and unilateral CP and bilateral CP, based on differences in CP risks assessed with Wilcoxon rank sum test and P values below 0.05 were considered statistically significant. Finally, we assessed the robustness of the spatial explanation method CAM on the test set for external validation by computing the Spearman's rank correlation between the overall CP risk and the mean value of the overall CAM of infants.

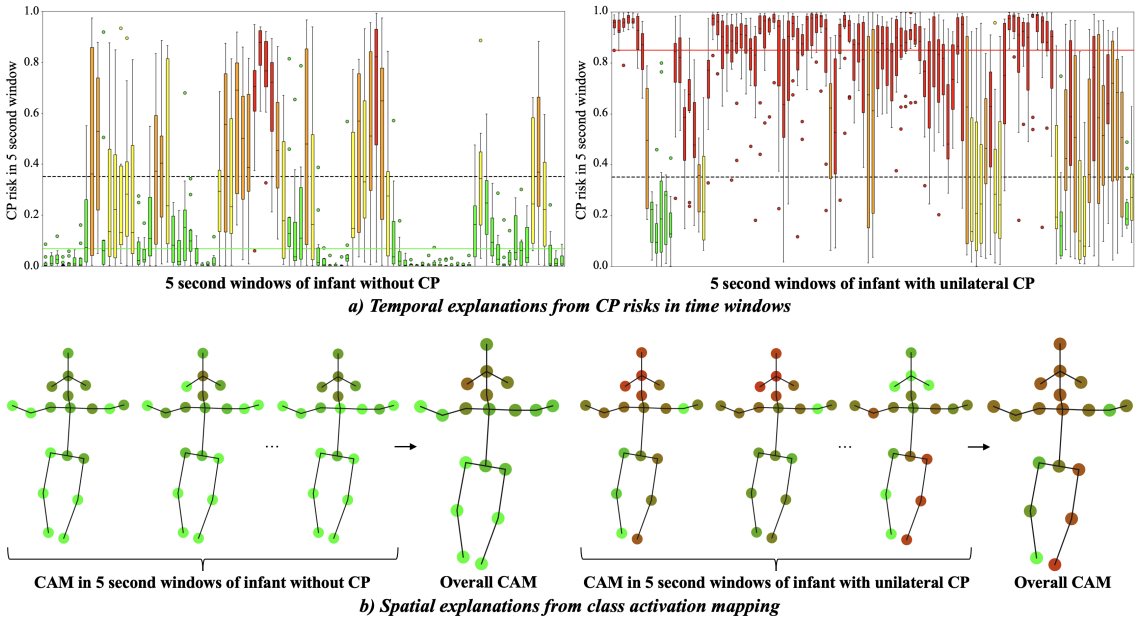


Figure 3.12: Temporal and spatial explanations associated with spontaneous movements in GMA recordings of one infant without CP and one infant with unilateral CP. a) The boxplots of CP risks in specific 5 second windows constitute temporal explanations. The dashed horizontal line reflects the decision threshold, whereas the green (left panel) and red (right panel) horizontal lines indicate the overall CP risks across all 5 second windows of the respective infants. b) Class activation mapping (CAM) enables spatial explanations by computing the contribution of each specific body keypoint towards prediction of CP in 5 second windows, and overall CAM is obtained through median aggregation across time windows. Red color in CAM indicates high contribution, suggesting that Ensemble-NAS-GCN associates the movements of a body keypoint with the presence of CP (e.g., for the body keypoints of the left leg in the overall CAM of the infant with unilateral CP), whereas green color reflects low contribution towards prediction of CP.

Chapter 4

Summary of results

4.1 Study I

The aim in Study I of the thesis was to develop highly precise and computationally efficient ConvNets for single-person pose estimation. For this purpose, the EfficientPose and EfficientHourglass models were proposed. The most precise variant in each model family, EfficientPose IV and EfficientHourglass B4, outperformed the commonly applied OpenPose ConvNet, with 2.4% and 12.6 — 13.5% increase in coarse and fine localization performance, respectively, while improving computational efficiency with 1.4 — 4.0 times fewer parameters and 2.2 — 6.0 times fewer FLOPs (Table B.1 in Appendix B). ConvNets with further improved computational efficiency were also able to display localization performance comparable to OpenPose. EfficientPose RT increased the fine localization performance of OpenPose by 1.7%, despite 56 and 184 times improvement in number of parameters and FLOPs, respectively. This study demonstrated the ability of low-complexity ConvNets to perform precise localization of body keypoints in video recordings of a single individual.

4.2 Study II

Based on the ConvNets proposed in Study I, the aim of Study II was to achieve motion capture of infants in GMA recordings, with high localization performance and computational efficiency. By constructing the In-Motion Poses dataset from an international, large-scale database of GMA videos of infants aged 9-18 weeks PTA, ConvNets were retrained for infant pose estimation. As depicted by Figure 4.1, the two most precise ConvNets, EfficientHourglass B4 and EfficientPose III, demonstrated superior localization performance compared to OpenPose by approaching human-level performance. See Appendix D of Paper II at ScienceDirect for localization performance in an external GMA recording. Furthermore, these ConvNets achieved processing speeds approximating real-time performance (i.e., 30 FPS) on a consumer GPU. The most computationally efficient ConvNet, EfficientPose RT, surpassed the localization performance of OpenPose, while operating at 198 FPS (see Figure 4.1 and Table C.1 in Appendix C). Overall, this study achieved video-based motion capture that is capable of localizing, with high precision, body keypoints of infants in GMA recordings, while operating efficiently.

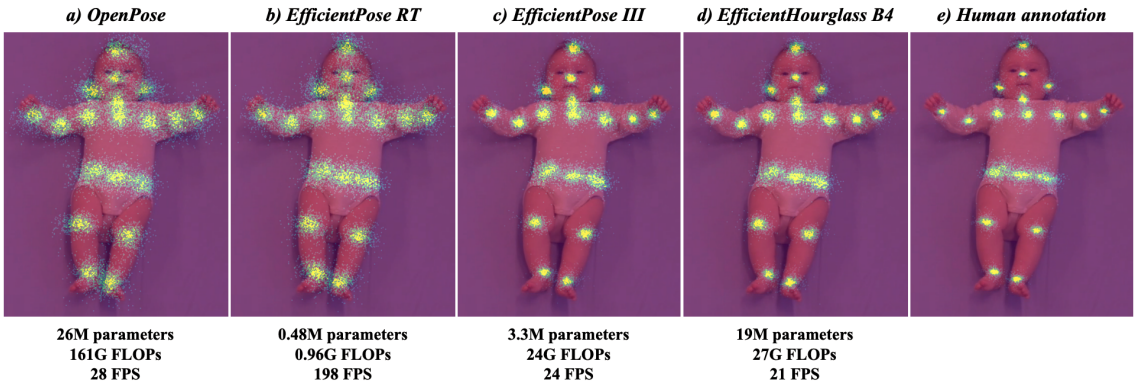


Figure 4.1: The distributions of prediction errors and computational efficiency (i.e., number of parameters in millions (M), number of FLOPs in billions (G), and processing speed in FPS) of a) OpenPose, in relation to the most computationally efficient ConvNet b) EfficientPose RT, and the most precise ConvNets c) EfficientPose III and d) EfficientHourglass B4, and e) inter-rater spread of human annotators (i.e., human-level performance). Extension of Fig. 4 in Paper II.

4.3 Study III

In Study III, the spontaneous movements of high-risk infants in GMA recordings were quantified by the motion capture from Study II, with the aim to develop and validate a GCN-based prediction model for CP. With fixed level of sensitivity (see dashed horizontal line in Figure 4.2 and Table D.1 in Appendix D), the proposed prediction model, Ensemble-NAS-GCN, displayed significantly improved specificity (94.1%, 95% CI: [88.2%, 97.6%]) compared to the state-of-the-art conventional machine learning method (i.e., CIMA model; 72.9%, 95% CI: [63.9%, 80.7%], $P < 0.001$) and the existing GCN-based method (i.e., ST-GCN; 83.9%, 95% CI: [76.0%, 90.0%], $P = 0.002$), and non-inferior specificity compared to the clinically recommended human expert-based GMA (88.7%, 95% CI: [81.5%, 93.8%], $P = 0.079$). Among the high-risk infants with and without CP, 66.7% and 88.1% were classified with high certainty, respectively (see Figure 4.3a and b). Moreover, Ensemble-NAS-GCN had higher CP risk in infants with non-ambulatory CP compared to ambulatory CP ($P = 0.007$), and for infants with bilateral CP compared to unilateral CP ($P = 0.029$). The sum of contributions of individual body keypoints, obtained with CAM, was very strongly correlated with CP risk (i.e., Spearman's rank correlation coefficient of 0.946), and hence CAM may be used to provide spatial explanations associated with decision of CP or no CP. This study demonstrated the potential of an GCN-based ensemble model obtained by NAS for automated CP prediction.

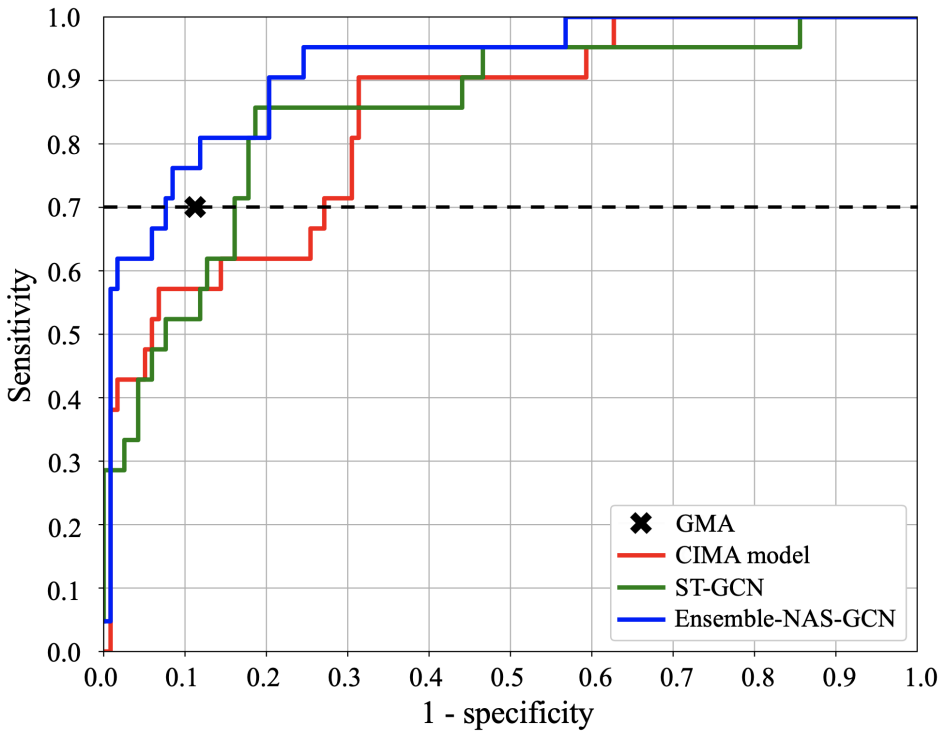


Figure 4.2: ROC curves of CIMA model, ST-GCN, and the proposed Ensemble-NAS-GCN, in relation to the sensitivity (i.e., dashed horizontal line) and specificity of GMA, on the test set of high-risk infants. Extension of eFigure 2 in Paper III.

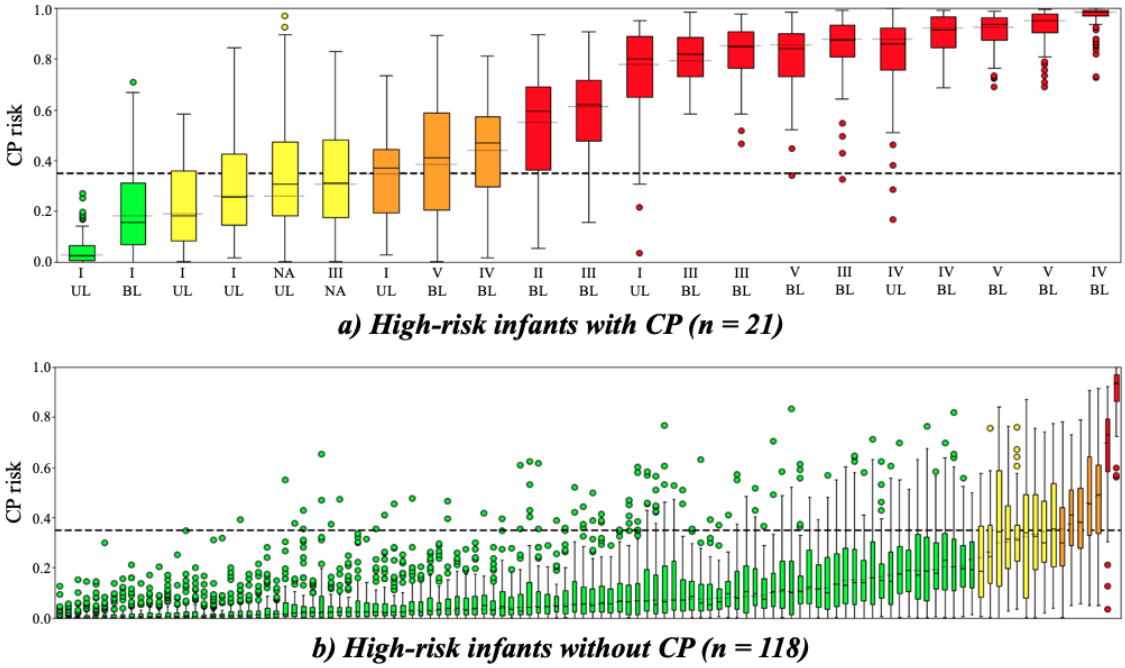


Figure 4.3: Distribution of CP risks and classification uncertainties (i.e., red and green for high certainty on decision of CP and no CP, respectively, and orange and yellow for uncertain classification of CP and no CP) of a) high-risk infants with CP in the test set, where the x-axis displays GMFCS level (i.e., I-V) and CP subtype (UL and BL for unilateral CP and bilateral CP, respectively, and NA for not available) and b) high-risk infants without CP in the test set. Adapted from Figure 3 of Paper III.

Chapter 5

Discussion

The early detection of CP from movement markers during infancy can initiate early follow-up and interventions of infants to optimize function and improve quality of life when brain plasticity is high. The qualitative assessment of infant spontaneous movements with GMA at 2-5 months PTA has currently reported the highest predictive accuracy for CP in high-risk infants. The overall objective of the present thesis was to develop convolutional networks for analysis of infant spontaneous movements in video recordings for the purpose of early objective prediction of CP.

This thesis revealed that machine learning-based CP prediction harnessing convolutional networks for video-based infant movement analysis at 3 months PTA may achieve predictive accuracy non-inferior to GMA in external validation on a representative sample of high-risk infants. Although recent studies on machine learning-based CP prediction have approached the performance of GMA using less conservative evaluation methods, like cross-validation, similar results have not previously been seen in external validation [47].

The high external validity of the proposed machine learning-based CP prediction indicates potential for use in infants from different countries with various medical risk factors. Furthermore, the assessment is performed automatically from a single video recording. This may facilitate widespread clinical adoption of non-invasive, objective screening for CP in high-risk infants, which is aligned with United Nations Sustainable Development Goals 3 and 10 [102] to "*ensure healthy lives and promote well-being for all at all ages*" and "*reduce inequality within and among countries*". The ability of the proposed CP prediction model to differentiate between infants with ambulatory and non-ambulatory CP and unilateral and bilateral CP may enable early prognosis to support physiotherapists in designing personalized interventions to better function and improve efficacy of follow-up. Furthermore, the provided color coding for classification uncertainty may aid health professionals in interpreting a decision of the machine learning-based CP prediction. High classification uncertainty could indicate a clinical choice for a follow-up consultancy or recommend the use of alternative assessment techniques, like GMA or HINE. Hence, machine learning-based CP prediction may supplement existing

expert-based assessments in clinical practice.

A possible future clinical service implementation of the machine learning-based CP prediction is outlined by Figure 5.1. Based on a GMA recording that is taken by a parent at home or a clinician in the hospital and uploaded into the electronic health record system, the video-based motion capture (step 1, Figure 5.1) could run automatically at a hospital computer to localize infant body keypoints. Subsequently, the automated prediction model (step 2, Figure 5.1) may harness this quantitative movement information to estimate CP risk and uncertainty of classification, along with temporal and spatial explanations, which are provided to physiotherapists and pediatricians through an interactive clinical dashboard.

In the remainder of this chapter, we describe strengths and limitations regarding the proposed video-based motion capture and prediction model for CP, as well as other considerations and avenues for future research.

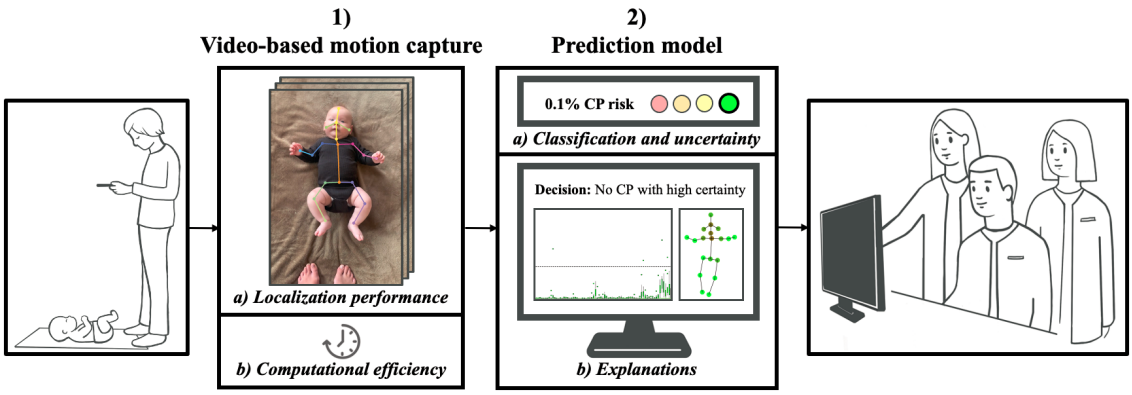


Figure 5.1: Emmet is born at 27 weeks GA and the preterm birth puts him at increased risk of CP. He is referred to a follow-up program after discharge from hospital and his parents are informed that an early prediction of CP can be performed at 3 months PTA. At the time of assessment, Emmet's mother Charlotte uses a smartphone application at home to record a 3-minute video of Emmet's spontaneous movements, following GMA standards, and to upload the video safely into the electronic health record system. Physiotherapist Lars accesses the electronic health record system at a hospital computer, selects the video of Emmet, and initiates the machine learning-based CP prediction. First, the video-based motion capture (step 1) localizes the body keypoints of Emmet across the video frames. The video-based motion capture has a) high localization performance (i.e., correctness in position estimates), which enables accurate quantification of the movement patterns and postural patterns of Emmet needed for prediction of CP, and b) high computational efficiency, causing the video to be fully processed when Lars gets hold of his colleagues, pediatrician Ragnhild and physiotherapist Toril, 3 minutes later. Within this time, the less computationally expensive operation, the prediction model for CP (step 2) has also completed. From the quantitative movement information extracted by the video-based motion capture, the prediction model a) estimates the CP risk of Emmet to 0.1%, recommending a classification into no CP with low uncertainty (green color code). The recommendation is provided to physiotherapists Lars and Toril and pediatrician Ragnhild through b) an interactive clinical dashboard which also includes the temporal and spatial explanations behind the decision. The temporal explanations enable the health care personnel to examine the specific time windows of the video where the prediction model detects spontaneous movement patterns related to no CP, whereas the spatial explanations suggest which of Emmet's body parts are involved in this decision. The health care personnel agree with the machine learning method on the decision of no CP. Physiotherapist Lars reaches out to Emmet's mother Charlotte and books an appointment to reassure Charlotte that there is no reason to believe that Emmet is at risk of CP and that they can come to the hospital at 6 months PTA to verify his motor development.

5.1 Video-based motion capture

To address Aim I of the thesis, namely to obtain ConvNets feasible for video-based motion capture of infants, we hypothesized that ConvNet architectures of high localization performance and computational efficiency in single-person HPE would appropriately transfer to infant pose estimation. High localization performance and computational efficiency are important in a clinical application because this ensures that the infant spontaneous movements are correctly represented while avoiding unnecessary delay for the available health care personnel in Figure 5.1.

5.1.1 Localization performance

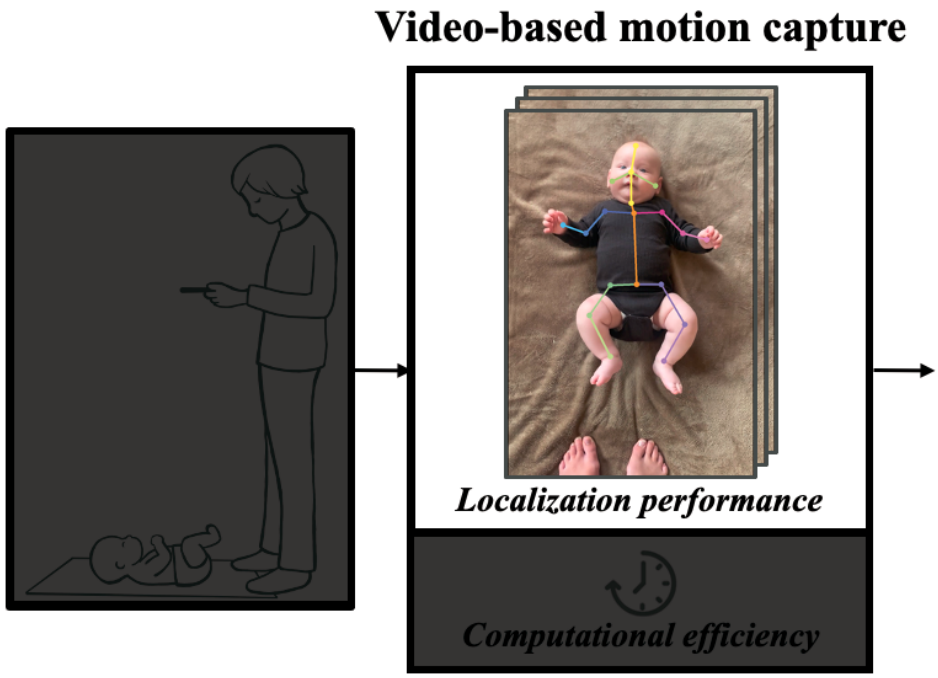


Figure 5.2: The localization performance of video-based motion capture reflects its ability to correctly estimate the positions of body keypoints.

The localization performance of video-based motion capture determines how precisely body keypoints are localized by the ConvNet. In particular, fine localization performance measures the ability to represent movements of small amplitude, like FMs, which could contain important discriminative information for prediction of CP. Figure 5.2 depicts that the skeleton model of the infant Emmet can be precisely estimated by the use of the proposed video-based motion capture. Study I and II confirmed improved fine localization performance of EfficientPose and EfficientHourglass compared to commonly applied OpenPose ConvNet, on single-person HPE and infant pose estimation, respectively.

The OpenPose architecture was originally designed not only to localize body keypoints of different humans in a variety of activities, but simultaneously to tackle the presence of several person instances (i.e., multi-person HPE) [64]. This was reflected in the high capacity of the ConvNet, which might also have contributed to high robustness against coarse prediction errors, like misses and inversions [103], as suggested by higher coarse localization performance of OpenPose on single-person HPE compared to lower-scale EfficientPose models (e.g., EfficientPose RT). However, the low spatial resolution in inputs and outputs of the OpenPose ConvNet restricts the fine localization performance, especially for persons occupying a smaller portion of the video frame [104]. The higher spatial resolutions of EfficientPose and EfficientHourglass promoted increased correctness in estimated body keypoint positions, with a general tendency on single-person HPE and infant pose estimation that higher spatial resolution increased the fine localization performance.

Nevertheless, it should be emphasized that high spatial resolution alone is not sufficient to obtain highly precise ConvNets for HPE. As indicated by EfficientPose RT achieving higher fine localization performance than OpenPose despite lower spatial resolution, appropriate composition of architectural components, capacity, and compound scaling of ConvNets across depth, width, and resolution are of high importance. The present thesis proposed multi-scale feature extraction with state-of-the-art EfficientNet backbones [74]. Sun et al. [81] has previously reported improved fine localization performance with lower spatial resolution by employing multi-scale ConvNet. This might suggest that multi-scale feature extraction partly overcomes the challenge with precise localization of body keypoints of small-scale persons. The use of compound scaling in EfficientPose and EfficientHourglass might have enabled further improvements in localization performance, e.g., by carefully balancing the required number of features (i.e., ConvNet width) in accordance with the complexity of features (i.e., ConvNet depth).

However, the adoption of compound scaling coefficients, optimized for image classification with EfficientNet [74], in EfficientPose and EfficientHourglass might be suboptimal for HPE. Although existing compound scaling coefficients may provide a good starting point, as reflected by the success of EfficientDet for object detection [105], task-specific compound scaling could yield improved performance [106]. Further studies could therefore systematically assess the appropriate compound scaling of ConvNet dimensions in single-person HPE and infant pose estimation. This could be performed by first using multi-objective NAS to arrive at a computationally efficient baseline network with decent localization performance, similar to EfficientNet-B0 for image classification. Subsequently, under the assumption of a doubling in FLOPs, optimal compound scaling coefficients could be found from a small grid search, which eventually yield ConvNets of higher scale. Furthermore, to attempt transferring the knowledge of high-scale ConvNets with high localization performance into a ConvNet of higher computational efficiency, knowledge distillation [107] could be employed, an approach that has previously shown promising results in single-person HPE [108].

Future research should also establish the required level of localization performance of video-based motion capture for infant movement analysis. This might vary between applications, and hence should be examined more thoroughly for each application separately. In automated CP prediction, we might conduct an experiment to indicate whether the high localization performance of the proposed video-based motion capture is necessary, or if a lower level of localization performance would be sufficient. A simple approach would be to simulate different magnitudes of Gaussian noise in normalized body keypoint positions to assess associated changes in estimated CP risk. However, it might be worth considering that lower-scale ConvNets could have different biases in localization behavior compared to highly precise ConvNets. Hence, a more informative experiment would be to estimate body keypoint positions with several ConvNets, for example EfficientPose RT, I, and III, of increasing localization performance on infant pose estimation, to determine the relationship between localization performance and differences in estimated CP risk.

The effect of statistical techniques for post-processing ConvNet predictions could also be investigated. Temporal smoothing of body keypoint trajectories with median filter [109] might diminish certain inaccuracies of lower-scale ConvNets (e.g., EfficientPose RT) in frame-by-frame infant pose estimation, which might make these models promising alternatives for computationally efficient video-based motion capture. Moreover, soft-argmax localization of body keypoints [110] might further improve the fine localization performance of ConvNets.

Despite these potentials for further improvements, the high fine localization performance of higher-scale EfficientPose models and EfficientHourglass B4 in comparison to OpenPose could enable more fine-grained analysis of movement kinematics, such as detection of small amplitude movements (e.g., FMs), due to the reduction in jitter (i.e., fine prediction errors) [103]. The high correctness in body keypoint positions, in turn, may facilitate that an accurate prediction of CP in the clinical service implementation in Figure 5.1 is achieved by the spontaneous movements of the infant Emmet being precisely represented.

5.1.2 Computational efficiency

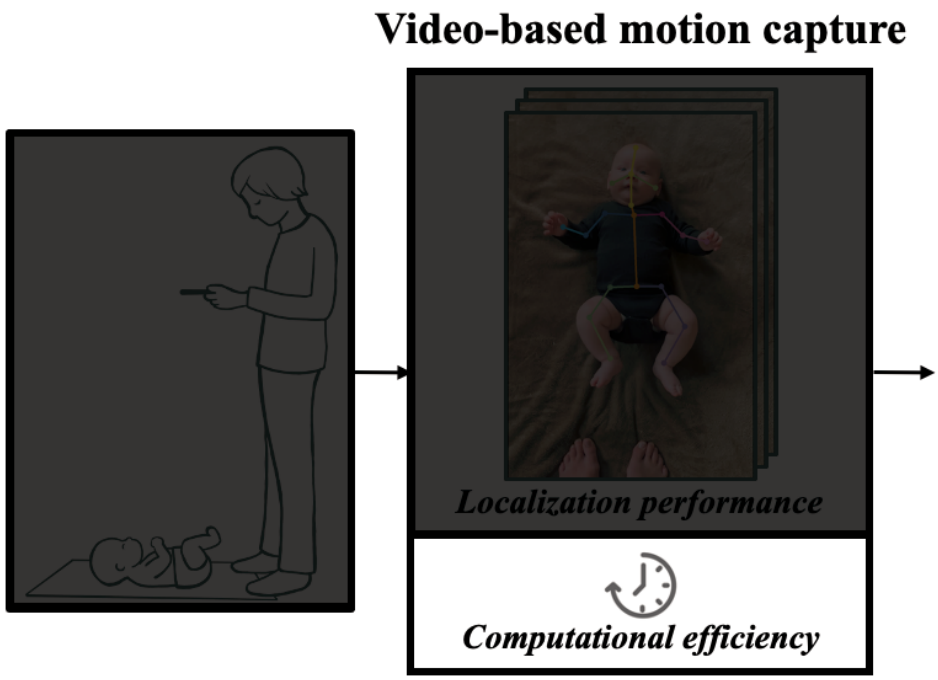


Figure 5.3: The computational efficiency of video-based motion capture comprises the computational complexity (number of FLOPs), model capacity (number of parameters), and run-time performance (inference latency and processing speed).

To obtain feasible video-based motion capture for clinical use, apart from high localization performance, adequate computational efficiency should be emphasized. The computational efficiency of ConvNets, in terms of computational complexity (i.e., number of FLOPs), model capacity (i.e., number of parameters), and run-time performance (i.e., inference latency and processing speed), dictates the computational resources and time required to perform video-based motion capture. In Figure 5.3, the clock reflects upon that high computational efficiency is necessary to avoid wasting valuable time of health care personnel in waiting for a response from the clinical decision support system. Study I and II demonstrated more efficient use of FLOPs and parameters in EfficientPose and EfficientHourglass in relation to OpenPose.

Despite the demonstrated reduction in computational complexity and model capacity of EfficientPose and EfficientHourglass compared to OpenPose, these improvements were not proportional to increase in processing speed. As reported by Li et al. [111], contemporary hardware accelerators (e.g., GPUs and TPUs) are dependent on operational intensity of ConvNets (i.e., FLOPs per memory byte) to maximize processing speed. Consequently, the lower operational intensity of MB-

Convs, extensively employed in EfficientPose and EfficientHourglass, in relation to basic convolutions, utilized in OpenPose, constrains the processing speeds that can be achieved [111].

However, the current implementation of ConvNets could be made more efficient using techniques for compressing ConvNets with minimal loss of localization performance, like parameter pruning and quantization [112]. For example, if we assume a similar improvement in processing speed as achieved by EfficientNet-Lite0 with integer-only post-training quantization [75], EfficientPose III would process a 3-minute video recording in 2 minutes on a consumer GPU, whereas EfficientPose RT would only require 15 seconds. This might eventually enable real-time decentralized processing of video recordings with high-precision motion capture on smartphones, eliminating the need for specialized hardware while preserving patient privacy.

Apart from determining the time consumption of the video-based motion capture in clinical use, the computational efficiency could impact the necessary time and data required to perform training of ConvNets. In this thesis, the EfficientPose and EfficientHourglass ConvNets required less extensive training and less available training data than OpenPose to reach a certain level of localization performance. Like EfficientPose and EfficientHourglass, OpenPose learnt anatomical proportions of infants in supine position with fine-tuning on infant images from GMA recordings, as reflected by improved localization performance on infant pose estimation. This aligns well with previous results by Chambers et al. [66] and confirms the hypothesis of Sciortino et al. [65] that retraining is necessary for ConvNets to adapt to infant pose estimation. However, EfficientPose and EfficientHourglass better utilized human annotations, as demonstrated by prediction errors resembling the inter-rater spread of human annotators, a trait not displayed by OpenPose. The CIMA-Pose ConvNet [90], a more computationally efficient version of OpenPose, also achieved improved localization performance on infant pose estimation compared to OpenPose over fewer epochs of training. This suggests that ConvNets of higher computational efficiency, in terms of number of parameters and FLOPs, may be beneficial for rapid convergence and better utilization of training data on infant pose estimation.

The high computational efficiency of the proposed video-based motion capture is therefore important both to yield efficient training and to achieve high run-time performance in clinical practice. The former enables high localization performance, but also social impact due to reduced energy use and carbon emissions related to tuning of the ConvNet. The latter ensures that the health care personnel in Figure 5.1 perceive the clinical support system as responsive and do not experience any prolonged wait, which would have been the case with lower run-time performance (e.g., a processing speed of 5 FPS had required the video-based motion capture to spend 18 minutes to process the 3-minute video).

5.2 Prediction model

Aim II of the thesis, to obtain a GCN-based prediction model for CP from spontaneous movements of high-risk infants, was approached through ensemble modeling and automatic architecture search (i.e., NAS) while emphasizing the robustness and trustworthiness of the CP prediction model. The robustness and trustworthiness of machine learning-based CP prediction are important in a clinical service implementation to ensure that the classification into CP or no CP of a high-risk infant (e.g., Emmet in Figure 5.1) is likely to be correct and that it is properly documented such that the clinical team can verify the decision and suggest the appropriate follow-up care. We refer to the robustness of the prediction model as the ability to perform accurate classification while indicating the associated uncertainty of classification, whereas trustworthiness relates to the credibility of the support evidence (i.e., explanations).

5.2.1 Classification and uncertainty

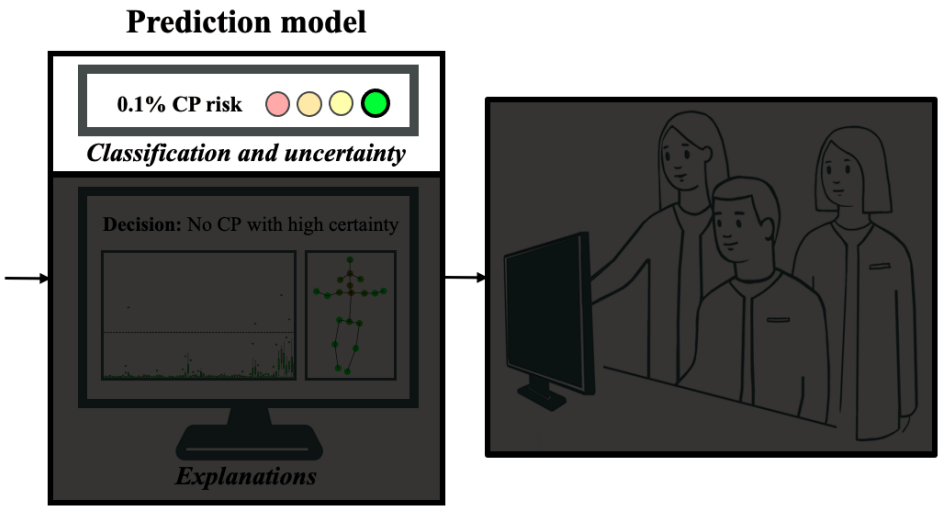


Figure 5.4: The classification and uncertainty of prediction model for CP express the decision of determining whether the infant spontaneous movements relate to CP or no CP and the associated level of confidence of this decision.

The classification represents the recommendation of the prediction model for decision of CP or no CP, whereas the uncertainty indicates the confidence level of the prediction model in making this recommendation. In Figure 5.4, the low estimated CP risk of 0.1% across the ensemble of GCN instances in the prediction model suggests to the health care personnel that the infant Emmet should be classified as not having CP. Furthermore, green color code reflects a low classification uncertainty, with more than 75% of GCN instances agreeing on the decision of no CP, which communicates to the health care personnel that the prediction model has strong evidence for the classification into no CP. In Study III, the robustness of Ensemble-NAS-GCN in high-risk infants was confirmed with improved predictive values for automated CP prediction compared to the state-of-the-art conventional machine learning method (i.e., CIMA model) and the existing GCN-based method (i.e., ST-GCN), and non-inferior performance to the clinically recommended GMA.

The improvement of Ensemble-NAS-GCN, in relation to CIMA model, could reflect upon the ability of GCNs to operate directly on biomechanical properties of raw skeleton sequences. This contrasts to the human expert-based extraction of a small set of relevant movement features, such as C_{SD} [48–50, 57] or frequency components [60–62], in conventional machine learning methods like CIMA model. The convolution-based automatic feature extraction performs higher-order analysis of complex relationships in whole-body movements to identify abstract features associated with the presence or absence of CP. This could mimic how gestalt perception emphasizes global patterns over individual details [36]. Hence, the demonstrated success of this seamless integration of automatic extraction of abstract features with prediction of outcome suggests that GCNs could represent a solution to the question of Silva et al. [46] regarding "*if and how human gestalt perception can be appropriately emulated by artificial intelligence*".

The use of ensemble modeling in Ensemble-NAS-GCN improved the predictive accuracy of classification by reducing the generalization error of individual model instances through "wisdom of the crowd" [113]. Ensemble modeling was also harnessed by CIMA model and few other methods for conventional machine learning-based CP prediction [51, 55, 62]. However, in contrast to the 70 GCN instances in Ensemble-NAS-GCN, CIMA model only comprises six model instances. This might limit the accuracy improvement of ensemble aggregation by most classification problems requiring at least tens of individual instances to reach convergence [114]. Furthermore, a higher level of classification uncertainty in CIMA model would be expected due to small ensemble size [115].

Nevertheless, it remains to establish the optimal ensemble configuration for automated CP prediction. Further research could examine the number of GCN instances in Ensemble-NAS-GCN, e.g., from 1 to 100, to assess differences both in terms of predictive accuracy and classification uncertainty. This might indicate whether the predictive accuracy saturates at a certain number of GCN instances and express how many GCN instances are required for negligible sampling error when computing classification uncertainty. Future studies could also explore alternative strategies for aggregating predictions of individual GCN instances. The

use of median aggregation in the present thesis yields equal contribution to all GCN instances in all time windows despite that different GCN instances could pay attention to different patterns of infant spontaneous movements. Improved predictive accuracy might be achieved by giving higher attribution to GCN instances considered active (i.e., of high or low predicted CP risk). Similarly, all time windows across a skeleton sequence might not contain discriminative movements. Accordingly, techniques like temporal attention-based aggregation, previously used in machine learning-based CP prediction by Nguyen-Thai et al. [54], could differentiate the attribution of time windows.

Apart from ensemble modeling, the use of NAS in Ensemble-NAS-GCN enabled improved predictive values for CP compared to the existing GCN-based method ST-GCN [44], commonly applied for tasks involving spatiotemporal skeleton sequences (e.g., human action recognition). This was despite that each of the different GCNs obtained by NAS had lower capacity (i.e., number of parameters) and complexity (i.e., number of FLOPs) than ST-GCN (see Table A.5 in Appendix A). This might suggest that automated CP prediction, approached as a binary classification task, does not demand as many degrees of freedom in GCN architectures compared to human action recognition. Alternatively, the use of other architectural components (e.g., type of convolutional layers, SE, activation functions, and attention mechanisms) and their configurations (e.g., kernel size, bottleneck factor, and SE ratio), and NAS to determine the optimal order of operations in a GCN, might have enabled more efficient use of parameters by constructing GCN architectures specialized towards detecting movement features related to CP and no CP.

However, the different design choices in Ensemble-NAS-GCN should be more extensively examined in future studies. In particular, the overall architectural design utilized in GCNs, inspired by the efforts of Song et al. [76, 77] in human action recognition from 3D skeleton sequences, could be investigated. The differences to the use of 2D sequences for automated CP prediction questions the validity of this architectural design. An important consideration to assess is the individual contributions of the included biomechanical properties (i.e., positions, velocities, and bones) and whether additional biomechanical properties should be incorporated. Further studies could also investigate alternatives to the fixed GCN structure (i.e., parallel input branches, feature concatenation, main branch, pooling layer, and fully-connected layer), and examine the specific architectural choices constraining GCN variants explored by NAS. It could also be established whether there are time window lengths that yield improvements compared to a time window length of 5 seconds. For example, we could assess a selection of different time window lengths, from 1 second to 30 seconds, by running NAS five times for each of the time window lengths and retain the time window length that achieves candidate GCNs with highest predictive accuracy. Such an experiment might indicate whether 5 seconds are sufficient to capture CP-related movement patterns. Alternatively, longer time windows might be preferred by including low-frequency movement patterns, but also more postural patterns and movement pat-

terns within a single time window could enable GCNs to analyze the co-existence of different patterns in infants with and without CP.

The proposed NAS strategy was designed to pursue GCN architectures reaching high predictive values within few epochs of training from restricted amount of available data. Accordingly, Ensemble-NAS-GCN was able to achieve performance non-inferior to observational GMA, despite each GCN instance being trained on a small number of skeleton sequences from associated infants, including only 54 skeleton sequences from infants with CP. In contrast, Yan et al. [44] trained ST-GCN on 240 000 skeleton sequences to perform human action recognition [116]. In yet other applications, convolutional networks may even require large-scale pretraining on millions or billions of data points to maximize accuracy [117]. The high external validity of Ensemble-NAS-GCN on the separate test set suggests that the training data appropriately approximates the natural variation among high-risk infants. Furthermore, the various spontaneous movement patterns contained in few skeleton sequences might have sufficiently covered the distribution of movements related to CP and no CP, diminishing the need for a large-scale training set or separate pretraining. Moreover, the splitting of skeleton sequences into 5 second windows, each comprising fewer distinct patterns of infant spontaneous movements, and hyperparameter search to determine an appropriate optimization procedure, including settings for data augmentation and weight initialization, might have facilitated robust convergence of GCNs.

Nevertheless, GCNs might further increase predictive accuracy for classification of CP with access to a larger training sample. This could involve large-scale collaboration with international research groups to utilize the proposed video-based motion capture on retrospective databases of GMA recordings with associated CP outcomes, to obtain a larger number of skeleton sequences. This might be a cumbersome process which requires the handling of ethical and privacy concerns of data sharing. Therefore, alternative technical solutions to increase the amount and heterogeneity of training data could be to generate synthetic skeleton sequences from existing training data, e.g., by the use of Gaussian process in CSGN [118], or by developing strategies for semi-supervised learning, like SESAR [119], to also harness skeleton sequences of GMA recordings without CP outcomes.

The present thesis has showcased the versatility of NAS to uncover the appropriate compositions of GCNs by harnessing infant spontaneous movements and CP outcomes from a limited number of infants. Furthermore, ensemble modeling has enabled accurate classification and estimation of classification uncertainty. This could be used in screening of CP in high-risk infants to enable health care personnel to quickly identify infants, like Emmet in Figure 5.1, displaying spontaneous movements indicating typical motor development. Moreover, classification into CP with low uncertainty could initiate targeted intervention from 3 months PTA to optimize function, whereas uncertain classifications may advise health care personnel to perform GMA or HINE to complement machine learning-based CP prediction.

5.2.2 Explanations

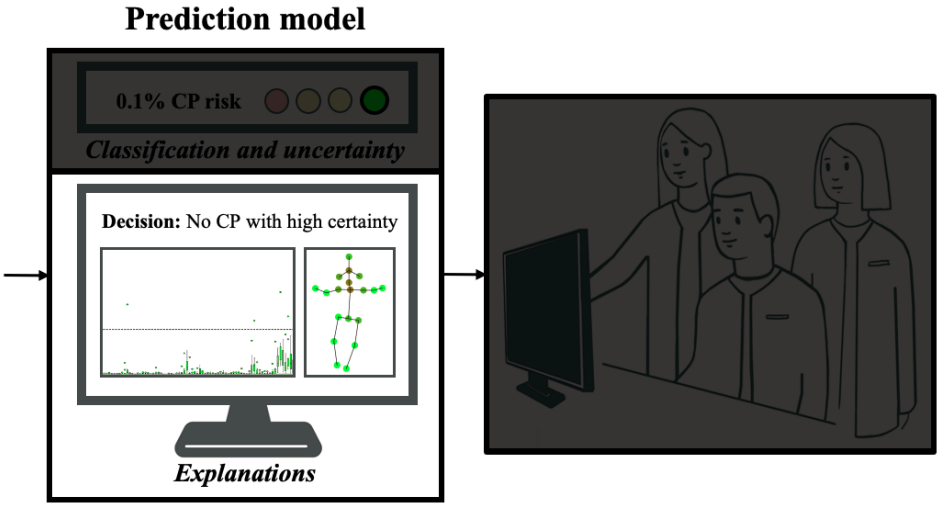


Figure 5.5: The explanations of prediction model for CP represents the supporting evidence associated with a classification into CP or no CP.

The explanations of the machine learning-based CP prediction comprise supporting evidence to verify and improve the understanding of a classification, to establish the trustworthiness of the prediction. In a clinical service implementation, explanations can be presented through a clinical dashboard, as illustrated in Figure 5.5, for health care personnel to assess the video periods where the spontaneous movements of the infant Emmet are associated with low CP risk (temporal aspect) and which body parts that are involved in these movements (spatial aspect). Such temporal and spatial explanations for the decisions of Ensemble-NAS-GCN in classifying high-risk infants into CP or no CP were proposed in Study III.

Although the present thesis did not assess the usefulness of these explanations, further studies could examine how such information might yield insights into the patterns of infant spontaneous movements Ensemble-NAS-GCN associates with CP and no CP. Systematic investigation of the connection between predictions of high and low CP risk and associated temporal and spatial explanations could determine hypotheses for potential biomarkers of CP in infant spontaneous movements. This might potentially elicit quantitative evidence for absent FMs as an important marker for CP, or establish the relevance of patterns of the concurrent motor repertoire for prediction of CP.

As a starting point towards achieving this, temporal explanations could be exploited to narrow down the investigation to video periods (e.g., 30 seconds) in the test set that yield high predictive certainty for CP or no CP. Subsequently, a qualitative experiment could include an equal number of video periods of high and low estimated CP risk which are displayed to GMA observers, blinded to the predictions of Ensemble-NAS-GCN, to assess the presence of FMs in these video periods. The portions of video periods with high and low CP risk that contain FMs indicate whether or not FMs are deemed discriminative for CP by Ensemble-NAS-GCN. Alternatively, an extended experiment could ask observers familiar with the Assessment of Motor Repertoire [18] to systematize the presence of different movement patterns and postural patterns of the concurrent motor repertoire in video periods of high and low estimated CP risk. This could suggest which specific infant movement patterns that are considered by Ensemble-NAS-GCN as related to CP or no CP.

These findings could thereafter be correlated with spatial explanations, obtained with CAM, to assess if the body keypoints involved in the infant spontaneous movement patterns of high and low CP risk receive a contribution towards prediction of CP that is high and low, respectively. A high correlation would strengthen the hypothesis that the movement patterns or postural patterns observed by the human experts are the same patterns of infant spontaneous movements as emphasized by Ensemble-NAS-GCN in prediction of CP. Moreover, this would further verify the robustness of CAM in providing accurate spatial explanations, which could be useful for health care personnel in interpreting decisions of the prediction model. On the contrary, a lack of correlation could suggest that the patterns considered discriminative by Ensemble-NAS-GCN differ from the patterns detected by human experts. However, it might also indicate that regular CAM is not suitable and should be replaced with related alternatives (e.g., Grad-CAM [120] or Grad-CAM++ [121]) or other explanation techniques, like attention weights [54].

Although further research is required to fully exploit and refine the explanations provided by the proposed machine learning method, this might eventually make these black box models more transparent. Accordingly, clinical end users of a decision support system might better interpret the decision of the prediction model in Figure 5.1 for classification of Emmet into no CP to establish trust in machine learning-based CP prediction.

5.3 Other considerations and avenues for future research

5.3.1 Field of use and transfer validity

In this thesis, the ConvNets for video-based motion capture and the GCNs in the prediction model for CP were trained on samples of infants recruited in previous studies. Although this included infants of various medical risk factors, these infants originated from a limited number of countries. Furthermore, the infants representing a certain country were mostly restricted to a specific medical risk factor (e.g., perinatal stroke for infants from Belgium and neonatal encephalopathy for infants from India). It is uncertain whether these aspects might affect the localization performance of video-based motion capture and predictive accuracy of prediction model for CP in GMA recordings of high-risk infants from other countries or of different medical risk factors. However, in general, convolutional networks require proper validation on relevant data in use cases outside the domain of the training data [41]. This is an important ingredient for understanding the knowledge limitations of the proposed methods and a step towards obtaining pre-market specification enabling use in medical devices. For example, Paper II reported decreased localization performance of ConvNets on synthetic GMA recordings. This was an indication that the proposed video-based motion capture should only be applied to video recordings following GMA standards in a real-world setting. Nevertheless, ConvNets might adapt to synthetic GMA recordings through retraining on an extended training set of In-Motion Poses which includes synthetic images and associated body keypoint annotations from MINI-RGBD [122].

Another use case that requires further validation is the robustness of Ensemble-NAS-GCN in GMA recordings captured by parents at home using smartphones, like the scenario in the possible clinical service implementation in Figure 5.1. Due to the lack of home-based smartphone recordings with available CP outcomes, training of Ensemble-NAS-GCN for CP prediction was restricted to standardized and less standardized hospital recordings where the camera was placed on a stand. However, home-based videos recorded with hand-held smartphones have been reported as feasible in observational GMA [33, 35, 123], which could indicate a potential for use in machine learning-based CP prediction. Nevertheless, there might be systematic differences in skeleton sequences caused by variations in camera angle and movement artefacts due to hand-held smartphone, but it is uncertain whether this alters the behavior of Ensemble-NAS-GCN in estimating CP risk. To assess this, future research could employ Ensemble-NAS-GCN on the home-based smartphone recordings of high-risk infants collected by Adde et al. [33]. A similar distribution of classification uncertainties across these infants compared to high-risk infants in Study III might suggest appropriate behavior. However, subsequent studies are required to verify this by determining the predictive values of Ensemble-NAS-GCN in home-based smartphone recordings from collected CP outcomes.

It would also be valuable to investigate the predictive accuracy of Ensemble-NAS-GCN in different groups of high-risk infants with and without CP. The low number of infants with CP in the test set in Study III limited the possibility of doing subgroup analysis to assess differences in predictive values of Ensemble-NAS-GCN in infants with different medical risk factors (e.g., premature birth, perinatal stroke, and neonatal encephalopathy) and in infants with spastic, dyskinetic, and ataxic CP. Moreover, future research should determine whether the predictive accuracy of Ensemble-NAS-GCN can be replicated in other samples of high-risk infants and verify that the behavior of the prediction model is consistent across infants from different countries.

5.3.2 One-step versus two-step approach

Another consideration which is worth examining in future research is whether the predominant two-step approach for machine learning-based CP prediction, comprising separate steps for video-based motion capture and prediction model, is optimal. The quantification of movement information into skeleton sequences by video-based motion capture could remove traits in GMA recordings relevant for the prediction of CP. It could for example be valuable to allow prediction models analyze infants' finger postures and facial expressions, since it has been demonstrated that infants with CP often have atypical variability of finger postures and absent or atypical tongue movements [32]. Schmidt et al. [52] attempted to perform automated GMA directly from raw video frames, omitting the video-based motion capture. However, the initial results of this attempt were not very promising, which we suspect was due to challenges regarding handling the large amount of visual information in video frames, including the presence of irrelevant video information (e.g., background noise, skin color, and clothing) potentially misleading the prediction model. Predictions directly from video also require proper addressing to avoid decisions from unjust grounds. In particular, it could be necessary with an initial step to remove identifiable information in GMA recordings, such as race and gender, e.g., by the use of the SMIL infant model developed by Hesse et al. [124].

5.3.3 Relation to automated GMA

The present thesis approached machine learning-based CP prediction as automated CP prediction. Although this enabled GCNs to independently determine the discriminative value of different patterns of infant spontaneous movements for prediction of CP, the alternative approach of automated GMA has other benefits. In contrast to automated CP prediction, automated GMA does not consider other movement patterns than FMs, which could make predictions in automated GMA more interpretable, by directly indicating the presence or absence of FMs. Moreover, the exclusive focus on FMs might simplify the training of machine learning systems, where methods for automated CP prediction might struggle discovering discriminative cues in infant spontaneous movements or possibly be confused by dissimilar pathological movements in infants with different subtypes of CP. Furthermore, with automated CP prediction, the use of a single noisy label of CP or no CP across a GMA recording further challenges the detection of CP-related movements due to infants with CP also displaying normal patterns of infant spontaneous movements [32].

On the other hand, a similar issue might occur in automated GMA, where a single GMA recording is commonly assigned an overall label of present FMs or absent FMs [47], despite FMs most often being present in limited parts of a video, reflecting upon the temporal organization of FMs [18]. Moreover, whereas automated CP prediction is based on the long-term CP outcome, diagnosed by a pediatrician blinded to the spontaneous movements in a GMA recording, methods for automated GMA learn to predict FMs based on the subjective gestalt perception of GMA observers. Taking into account the variance among GMA observers, and in particular the lower reliability of observers with limited experience [37], we could therefore expect machine learning methods for automated GMA to show a similar behavior and hence be constrained by the performance of the respective observer. Nevertheless, further studies should assess the feasibility of both approaches, and determine if automated GMA and automated CP prediction could potentially complement each other.

5.3.4 Multimodality convolutional network-based CP prediction

Whereas the present thesis was limited to CP prediction from infant spontaneous movements in GMA recordings, the importance of other prominent markers for CP and their associated assessment techniques should not be neglected. Hence, further efforts could explore the use of multimodality prediction models that extends the proposed CP prediction by incorporating analysis of neonatal MRI scans to assess brain abnormalities. A unified CP prediction framework might extract features for each modality separately with suitable techniques, like ConvNets (e.g., U-Net [125]) for MRI scans, and the proposed combination of ConvNets and GCNs for infant spontaneous movements in GMA recordings, followed by concatenation into multimodality features from which a classification of CP and no CP could be performed. This could potentially improve the predictive accuracy of early CP prediction while providing clinically meaningful explanations connecting cues of brain development and spontaneous movements of infants.

Chapter 6

Conclusion

This thesis approached the challenge of early prediction of CP through video-based infant movement analysis harnessing convolutional networks. The infant spontaneous movements in a video recording at 3 months PTA were localized using highly precise and computationally efficient ConvNet-based motion capture. A GCN-based prediction model, in turn, analyzed this movement information to provide an objective prediction of CP. The predictive accuracy of this non-invasive, automated assessment in high-risk infants was non-inferior to the clinically recommended GMA, while possessing the ability to differentiate ambulatory CP from non-ambulatory CP and unilateral CP from bilateral CP. Although the proposed solution is yet to demonstrate feasibility in clinical practice and transfer validity in new samples of high-risk infants, these findings portray a future where convolutional networks may play an important role in next-generation clinical decision support.

Bibliography

[1] Pierre-Yves Ancel, François Goffinet, Pierre Kuhn, Bruno Langer, Jacqueline Matis, Xavier Hernandorena, Pierre Chabanier, Laurence Joly-Pedespan, Bénédicte Lecomte, Françoise Vendittelli, et al. “Survival and morbidity of preterm children born at 22 through 34 weeks’ gestation in France in 2011: results of the EPIPAGE-2 cohort study.” In: *JAMA pediatrics* 169.3 (2015), pp. 230–238.

[2] Hannah Blencowe, Simon Cousens, Doris Chou, Mikkel Oestergaard, Lale Say, Ann-Beth Moller, Mary Kinney, and Joy Lawn. “Born too soon: the global epidemiology of 15 million preterm births.” In: *Reproductive health* 10.1 (2013), pp. 1–14.

[3] Betty Vohr, Linda L. Wright, M. Hack, G. Aylward, and D. Hirtz. “Follow-up care of high-risk infants.” In: *PEDIATRICS-SPRINGFIELD-* 114.2 (2004).

[4] Maryam Oskoui, Franzina Coutinho, Jonathan Dykeman, Nathalie Jette, and Tamara Pringsheim. “An update on the prevalence of cerebral palsy: a systematic review and meta-analysis.” In: *Developmental Medicine & Child Neurology* 55.6 (2013), pp. 509–519.

[5] Sandra Julsen Hollung, Torstein Vik, Stian Lydersen, Inger Johanne Bakken, and Guro L. Andersen. “Decreasing prevalence and severity of cerebral palsy in Norway among children born 1999 to 2010 concomitant with improvements in perinatal health.” In: *european journal of paediatric neurology* 22.5 (2018), pp. 814–821.

[6] Peter Rosenbaum, Nigel Paneth, Alan Leviton, Murray Goldstein, Martin Bax, Diane Damiano, Bernard Dan, and Bo Jacobsson. “A report: the definition and classification of cerebral palsy April 2006.” In: *Dev Med Child Neurol* 109 (2007), pp. 8–14.

[7] Mijna Hadders-Algra. “Early Diagnostics and Early Intervention in Neurodevelopmental Disorders—Age-Dependent Challenges and Opportunities.” In: *Journal of clinical medicine* 10.4 (2021), p. 861.

[8] Iona Novak, Cathy Morgan, Lars Adde, James Blackman, Roslyn N. Boyd, Janice Brunstrom-Hernandez, Giovanni Cioni, Diane Damiano, Johanna Darrah, Ann-Christin Eliasson, et al. “Early, accurate diagnosis and early

intervention in cerebral palsy: advances in diagnosis and treatment." In: *JAMA pediatrics* 171.9 (2017), pp. 897–907.

[9] Mijna Hadders-Algra. "General movements: a window for early identification of children at high risk for developmental disorders." In: *The Journal of pediatrics* 145.2 (2004), S12–S18.

[10] Catherine Morgan, Linda Fetters, Lars Adde, Nadia Badawi, Ada Bancale, Roslyn N. Boyd, Olena Chorna, Giovanni Cioni, Diane L. Damiano, Johanna Darrah, et al. "Early intervention for children aged 0 to 2 years with or at high risk of cerebral palsy: international clinical practice guideline based on systematic reviews." In: *JAMA pediatrics* 175.8 (2021), pp. 846–858.

[11] Gillian Baird, Helen McConachie, and David Scrutton. "Parents' perceptions of disclosure of the diagnosis of cerebral palsy." In: *Archives of disease in childhood* 83.6 (2000), pp. 475–480.

[12] Katherine Guttmann, John Flibotte, and Sara B. DeMauro. "Parental perspectives on diagnosis and prognosis of neonatal intensive care unit graduates with cerebral palsy." In: *The Journal of Pediatrics* 203 (2018), pp. 156–162.

[13] Sarah McIntyre, Cathy Morgan, Karen Walker, and Iona Novak. "Cerebral palsy–don't delay." In: *Developmental disabilities research reviews* 17.2 (2011), pp. 114–129.

[14] Christa Einspieler, Peter B. Marschik, Arend F. Bos, Fabrizio Ferrari, Giovanni Cioni, and Heinz F.R. Prechtl. "Early markers for cerebral palsy: insights from the assessment of general movements." In: *Future Neurology* 7.6 (2012), pp. 709–717.

[15] Lianne J. Woodward, Peter J. Anderson, Nicola C. Austin, Kelly Howard, and Terrie E. Inder. "Neonatal MRI to predict neurodevelopmental outcomes in preterm infants." In: *New England Journal of Medicine* 355.7 (2006), pp. 685–694.

[16] Domenico M. Romeo, Daniela Ricci, Claudia Brogna, and Eugenio Mercuri. "Use of the Hammersmith Infant Neurological Examination in infants with cerebral palsy: a critical review of the literature." In: *Developmental Medicine & Child Neurology* 58.3 (2016), pp. 240–245.

[17] Janneke van't Hooft, Johanna H. van der Lee, Brent C. Opmeer, Cornelieke S.H. Aarnoudse-Moens, Arnold G.E. Leenders, Ben Willem J. Mol, and Timo R. de Haan. "Predicting developmental outcomes in premature infants by term equivalent MRI: systematic review and meta-analysis." In: *Systematic reviews* 4.1 (2015), pp. 1–10.

[18] Christa Einspieler, Heinz F.R. Prechtl, Arend Bos, Fabrizio Ferrari, and Giovanni Cioni. *Prechtl's method on the qualitative assessment of general movements in preterm, term and young infants*. Vol. 167. Mac Keith Press London, 2004.

[19] Amanda K.L. Kwong, Tara L. Fitzgerald, Lex W. Doyle, Jeanie L.Y. Cheong, and Alicia J. Spittle. "Predictive validity of spontaneous early infant movement for later cerebral palsy: a systematic review." In: *Developmental Medicine & Child Neurology* 60.5 (2018), pp. 480–489.

[20] Leena Haataja, Eugenio Mercuri, Rivka Regev, Frances Cowan, Mary Rutherford, Victor Dubowitz, and Lilly Dubowitz. "Optimality score for the neurologic examination of the infant at 12 and 18 months of age." In: *The Journal of pediatrics* 135.2 (1999), pp. 153–161.

[21] Heinz F.R. Prechtl, J.W. Fargel, Hans-Martin Weinmann, and H.H. Bakker. "Postures, motility and respiration of low-risk pre-term infants." In: *Developmental Medicine & Child Neurology* 21.1 (1979), pp. 3–27.

[22] Johanna I.P. De Vries, Gerard H.A. Visser, and Heinz F.R. Prechtl. "The emergence of fetal behaviour. I. Qualitative aspects." In: *Early human development* 7.4 (1982), pp. 301–322.

[23] Brian Hopkins and Heinz F.R. Prechtl. "A qualitative approach to the development of movements during early infancy." In: *Continuity of neural functions from prenatal to postnatal life* (1984), pp. 179–197.

[24] Christa Einspieler, Robert Peharz, and Peter B. Marschik. "Fidgety movements–tiny in appearance, but huge in impact." In: *Jornal de Pediatria* 92 (2016), pp. 64–70.

[25] Heinz F.R. Prechtl, Christa Einspieler, Giovanni Cioni, Arend F. Bos, Fabizi Ferrari, and Dieter Sontheimer. "An early marker for neurological deficits after perinatal brain lesions." In: *The Lancet* 349.9062 (1997), pp. 1361–1363.

[26] Mijna Hadders-Algra. "Neural substrate and clinical significance of general movements: an update." In: *Developmental Medicine & Child Neurology* 60.1 (2018), pp. 39–46.

[27] Christa Einspieler, Alison M. Kerr, and Heinz F.R. Prechtl. "Is the early development of girls with Rett disorder really normal?" In: *Pediatric Research* 57.5 (2005), pp. 696–700.

[28] Christa Einspieler, Jeff Sigafoos, Katrin D. Bartl-Pokorny, Rebecca Landa, Peter B. Marschik, and Sven Bölte. "Highlighting the first 5 months of life: General movements in infants later diagnosed with autism spectrum disorder or Rett syndrome." In: *Research in Autism Spectrum Disorders* 8.3 (2014), pp. 286–291.

[29] Christa Einspieler and Heinz F.R. Prechtl. “Prechtl’s assessment of general movements: a diagnostic tool for the functional assessment of the young nervous system.” In: *Mental retardation and developmental disabilities research reviews* 11.1 (2005), pp. 61–67.

[30] Heinz F.R. Prechtl. “State of the art of a new functional assessment of the young nervous system. An early predictor of cerebral palsy.” In: *Early human development* 50.1 (1997), pp. 1–11.

[31] Ragnhild Støen, Lynn Boswell, Raye-Ann DeRegnier, Toril Fjørtoft, Deborah Gaebler-Spira, Espen A.F. Ihlen, Cathrine Labori, Marianne Loennecken, Michael Msall, Unn Inger Möinichen, et al. “The predictive accuracy of the general movement assessment for cerebral palsy: a prospective, observational study of high-risk infants in a clinical follow-up setting.” In: *Journal of clinical medicine* 8.11 (2019), p. 1790.

[32] Christa Einspieler, Arend F. Bos, Magdalena Krieber-Tomantschger, Elsa Alvarado, Vanessa M. Barbosa, Natascia Bertoncelli, Marlette Burger, Olena Chorna, Sabrina Del Secco, Raye-Ann DeRegnier, et al. “Cerebral palsy: early markers of clinical phenotype and functional outcome.” In: *Journal of clinical medicine* 8.10 (2019), p. 1616.

[33] Lars Adde, Annemette Brown, Christine Van Den Broeck, Kris DeCoen, Beate Horsberg Eriksen, Toril Fjørtoft, Daniel Groos, Espen A.F. Ihlen, Siril Osland, Aurelie Pascal, et al. “In-Motion-App for remote General Movement Assessment: a multi-site observational study.” In: *BMJ open* 11.3 (2021), e042147.

[34] Amanda K.L. Kwong, Abbey L. Eeles, Joy E. Olsen, Jeanie L.Y. Cheong, Lex W. Doyle, and Alicia J. Spittle. “The Baby Moves smartphone app for general movements assessment: Engagement amongst extremely preterm and term-born infants in a state-wide geographical study.” In: *Journal of paediatrics and child health* 55.5 (2019), pp. 548–554.

[35] Katarina A. Svensson, Maria Örtqvist, Arend F. Bos, Ann-Christin Eliasson, and Heléne E.K. Sundelin. “Usability and inter-rater reliability of the NeuroMotion app: A tool in General Movements Assessments.” In: *European Journal of Paediatric Neurology* 33 (2021), pp. 29–35.

[36] Konrad Lorenz. “Gestalt-wahrnehmung als Quelle wissenschaftlicher Erkenntnis.” In: *Zeitschrift für experimentelle und angewandte Psychologie* (1959). English translation: Konrad Lorenz. "Gestalt perception as a source of scientific knowledge." In: *Studies in animal and human behaviour* 2 (1971), pp. 281–322.

[37] Colleen Peyton, Aurelie Pascal, Lynn Boswell, Raye-Ann DeRegnier, Toril Fjørtoft, Ragnhild Støen, and Lars Adde. “Inter-observer reliability using the General Movement Assessment is influenced by rater experience.” In: *Early Human Development* 161 (2021), p. 105436.

[38] Nathalie Maitre. “Skepticism, cerebral palsy, and the General Movements Assessment.” In: *Developmental Medicine & Child Neurology* 60.5 (2018), pp. 438–438.

[39] Christa Einspieler, Peter B. Marschik, and Heinz F.R. Prechtl. “Human motor behavior: Prenatal origin and early postnatal development.” In: *Zeitschrift für Psychologie/Journal of Psychology* 216.3 (2008), pp. 147–153.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification.” In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[41] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT press, 2016.

[42] Tai Sing Lee, David Mumford, Richard Romero, and Victor A.F. Lamme. “The role of the primary visual cortex in higher level vision.” In: *Vision research* 38.15-16 (1998), pp. 2429–2454.

[43] David Marr and Tomaso Poggio. “A computational theory of human stereo vision.” In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 204.1156 (1979), pp. 301–328.

[44] Sijie Yan, Yuanjun Xiong, and Dahua Lin. “Spatial temporal graph convolutional networks for skeleton-based action recognition.” In: *Thirty-second AAAI conference on artificial intelligence*. 2018.

[45] Christian B. Redd, Mohan Karunanithi, Roslyn N. Boyd, and Lee A. Barber. “Technology-assisted quantification of movement to predict infants at high risk of motor disability: A systematic review.” In: *Research in Developmental Disabilities* 118 (2021), p. 104071.

[46] Nelson Silva, Dajie Zhang, Tomas Kulvicius, Alexander Gail, Carla Barreiros, Stefanie Lindstaedt, Marc Kraft, Sven Bölte, Luise Poustka, Karin Nielsen-Saines, et al. “The future of General Movement Assessment: The role of computer vision and machine learning–A scoping review.” In: *Research in developmental disabilities* 110 (2021), p. 103854.

[47] Muhammad Tausif Irshad, Muhammad Adeel Nisar, Philip Gouverneur, Marion Rapp, and Marcin Grzegorzek. “AI approaches towards Prechtl’s assessment of general movements: A systematic literature review.” In: *Sensors* 20.18 (2020), p. 5321.

[48] Lars Adde, Jorunn L. Helbostad, Alexander Refsum Jensenius, Gunnar Taraldsen, and Ragnhild Støen. “Using computer-based video analysis in the study of fidgety movements.” In: *Early human development* 85.9 (2009), pp. 541–547.

[49] Lars Adde, Jorunn Helbostad, Alexander R. Jensenius, Mette Langaas, and Ragnhild Støen. “Identification of fidgety movements and prediction of CP by the use of computer-based video analysis is more accurate when based on two video recordings.” In: *Physiotherapy theory and practice* 29.6 (2013), pp. 469–475.

[50] Ragnhild Støen, Nils Thomas Songstad, Inger Elisabeth Silberg, Toril Fjørtoft, Alexander Refsum Jensenius, and Lars Adde. “Computer-based video analysis identifies infants with absence of fidgety movements.” In: *Pediatric Research* 82.4 (2017), pp. 665–670.

[51] Silvia Orlandi, Kamini Raghuram, Corinna R. Smith, David Mansueto, Paige Church, Vibhuti Shah, Maureen Luther, and Tom Chau. “Detection of atypical and typical infant movements using computer-based video analysis.” In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2018, pp. 3598–3601.

[52] William Schmidt, Matthew Regan, Micheal Fahey, and Andrew Paplinski. “General movement assessment by machine learning: Why is it so difficult.” In: *J. Med. Artif. Intell* 2 (2019).

[53] Kevin D. McCay, Edmond S.L. Ho, Claire Marcroft, and Nicholas D. Embleton. “Establishing pose based features using histograms for the detection of abnormal infant movements.” In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2019, pp. 5469–5472.

[54] Binh Nguyen-Thai, Vuong Le, Catherine Morgan, Nadia Badawi, Truyen Tran, and Svetha Venkatesh. “A spatio-temporal attention-based model for infant movement assessment from videos.” In: *IEEE journal of biomedical and health informatics* 25.10 (2021), pp. 3911–3920.

[55] Kevin D. McCay, Edmond S.L. Ho, Dimitrios Sakkos, Wai Lok Woo, Claire Marcroft, Patricia Dulson, and Nicholas D. Embleton. “Towards explainable abnormal infant movements identification: A body-part based prediction and visualisation framework.” In: *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE. 2021, pp. 1–4.

[56] Dimitrios Sakkos, Kevin D. Mccay, Claire Marcroft, Nicholas D. Embleton, Samiran Chattopadhyay, and Edmond S.L. Ho. “Identification of abnormal movements in infants: A deep neural network for body part-based prediction of cerebral palsy.” In: *IEEE Access* 9 (2021), pp. 94281–94292.

[57] Lars Adde, Jorunn L. Helbostad, Alexander R. Jensenius, Gunnar Taraldsen, Kristine H. Grunewaldt, and Ragnhild StØen. “Early prediction of cerebral palsy by computer-based video analysis of general movements: a feasibility study.” In: *Developmental Medicine & Child Neurology* 52.8 (2010), pp. 773–778.

[58] Annette Stahl, Christian Schellewald, Øyvind Stavdahl, Ole Morten Aamo, Lars Adde, and Harald Kirkerod. “An optical flow-based method to predict infantile cerebral palsy.” In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 20.4 (2012), pp. 605–614.

[59] Hodjat Rahmati, Ole Morten Aamo, Øyvind Stavdahl, Ralf Dragon, and Lars Adde. “Video-based early cerebral palsy prediction using motion segmentation.” In: *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2014, pp. 3779–3783.

[60] Hodjat Rahmati, Harald Martens, Ole Morten Aamo, Øyvind Stavdahl, Ragnhild Støen, and Lars Adde. “Frequency-based features for early cerebral palsy prediction.” In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2015, pp. 5187–5190.

[61] Hodjat Rahmati, Harald Martens, Ole Morten Aamo, Øyvind Stavdahl, Ragnhild Støen, and Lars Adde. “Frequency analysis and feature reduction method for prediction of cerebral palsy in young infants.” In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 24.11 (2016), pp. 1225–1234.

[62] Espen A.F. Ihlen, Ragnhild Støen, Lynn Boswell, Raye-Ann de Regnier, Toril Fjørtoft, Deborah Gaebler-Spira, Cathrine Labori, Marianne C. Loennecken, Michael E. Msall, Unn I. Möinichen, et al. “Machine learning of infant spontaneous movements for the early prediction of cerebral palsy: A multi-site cohort study.” In: *Journal of Clinical Medicine* 9.1 (2020), p. 5.

[63] Hodjat Rahmati, Ralf Dragon, Ole Morten Aamo, Luc van Gool, and Lars Adde. “Motion segmentation with weak labeling priors.” In: *German Conference on Pattern Recognition*. Springer. 2014, pp. 159–171.

[64] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. “Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7291–7299.

[65] Giuseppa Sciortino, Giovanni Maria Farinella, Sebastiano Battiato, Marco Leo, and Cosimo Distante. “On the Estimation of Children’s Poses.” In: *International conference on image analysis and processing*. Springer. 2017, pp. 410–421.

[66] Claire Chambers, Nidhi Seethapathi, Rachit Saluja, Helen Loeb, Samuel R. Pierce, Daniel K. Bogen, Laura Prosser, Michelle J. Johnson, and Konrad P. Kording. “Computer vision to automatically assess infant neuromotor risk.” In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28.11 (2020), pp. 2431–2442.

[67] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks.” In: *Advances in neural information processing systems* 25 (2012).

[68] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition.” In: *arXiv preprint arXiv:1409.1556* (2014).

[69] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[70] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.” In: *arXiv preprint arXiv:1704.04861* (2017).

[71] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. “MnasNet: Platform-Aware Neural Architecture Search for Mobile.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2820–2828.

[72] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. “MobileNetV2: Inverted Residuals and Linear Bottlenecks.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4510–4520.

[73] Jie Hu, Li Shen, and Gang Sun. “Squeeze-and-Excitation Networks.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7132–7141.

[74] Mingxing Tan and Quoc Le. “EfficientNet: Rethinking model scaling for convolutional neural networks.” In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114.

[75] Renjie Liu. “Higher accuracy on vision models with EfficientNet-Lite.” In: *TensorFlow Blog* (2020).

[76] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. “Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition.” In: *proceedings of the 28th ACM international conference on multimedia*. 2020, pp. 1625–1633.

[77] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. “Constructing stronger and faster baselines for skeleton-based action recognition.” In: *arXiv preprint arXiv:2106.15125* (2021).

[78] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. “Disentangling and unifying graph convolutions for skeleton-based action recognition.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 143–152.

[79] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. “2D human pose estimation: New benchmark and state of the art analysis.” In: *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*. 2014, pp. 3686–3693.

[80] Alejandro Newell, Kaiyu Yang, and Jia Deng. “Stacked hourglass networks for human pose estimation.” In: *European conference on computer vision*. Springer. 2016, pp. 483–499.

[81] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. “Deep high-resolution representation learning for human pose estimation.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5693–5703.

[82] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. “Densely connected convolutional networks.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[83] Eric Alcaide. “E-swish: Adjusting Activations to Different Network Depths.” In: *arXiv preprint arXiv:1801.07145* (2018).

[84] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[85] Alex Krizhevsky and Geoff Hinton. “Convolutional Deep Belief Networks on CIFAR-10.” In: *Unpublished manuscript* 40.7 (2010), pp. 1–9.

[86] Hong Zhang, Hao Ouyang, Shu Liu, Xiaojuan Qi, Xiaoyong Shen, Ruigang Yang, and Jiaya Jia. “Human pose estimation with spatial contextual information.” In: *arXiv preprint arXiv:1901.01760* (2019).

[87] Wei Tang, Pei Yu, and Ying Wu. “Deeply learned compositional models for human pose estimation.” In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 190–206.

[88] Lipeng Ke, Ming-Ching Chang, Honggang Qi, and Siwei Lyu. “Multi-scale structure-aware network for human pose estimation.” In: *Proceedings of the european conference on computer vision (ECCV)*. 2018, pp. 713–728.

[89] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. “Learning feature pyramids for human pose estimation.” In: *proceedings of the IEEE international conference on computer vision*. 2017, pp. 1281–1290.

[90] Daniel Groos and Kristian Aurlien. “Infant Body Part Tracking in Videos Using Deep Learning–Facilitating Early Detection of Cerebral Palsy.” MA thesis. NTNU, 2018.

[91] OpenPose. *Real-time multi-person keypoint detection library for body, face, hands, and foot estimation*. `https://github.com/CMU-Perceptual-Computing-Lab/openpose`. Accessed on: 30 May 2021.

[92] Fisher Yu and Vladlen Koltun. “Multi-scale context aggregation by dilated convolutions.” In: *arXiv preprint arXiv:1511.07122* (2015).

[93] Aurelie Pascal, Paul Govaert, Els Ortibus, Gunnar Naulaers, Adde Lars, Torill Fjørtoft, Ann Oostra, Aleksandra Zecic, Filip Cools, Eva Cloet, et al. "Motor outcome after perinatal stroke and early prediction of unilateral spastic cerebral palsy." In: *European Journal of Paediatric Neurology* 29 (2020), pp. 54–61.

[94] Karoline Aker, Niranjan Thomas, Lars Adde, Beena Koshy, Miriam Martinez-Biarge, Ingeborg Nakken, Caroline S. Padankatti, and Ragnhild Støen. "Prediction of outcome from MRI and general movements assessment after hypoxic-ischaemic encephalopathy in low-income and middle-income countries: data from a randomised controlled trial." In: *Archives of Disease in Childhood-Fetal and Neonatal Edition* 107.1 (2022), pp. 32–38.

[95] Jasmin Dibiasi and Christa Einspieler. "Can spontaneous movements be modulated by visual and acoustic stimulation in 3-month-old infants?" In: *Early human development* 68.1 (2002), pp. 27–37.

[96] Christine Cans. "Surveillance of cerebral palsy in Europe: a collaboration of cerebral palsy surveys and registers." In: *Developmental Medicine & Child Neurology* 42.12 (2000), pp. 816–824.

[97] Robert Palisano, Peter Rosenbaum, Stephen Walter, Dianne Russell, Ellen Wood, and Barbara Galuppi. "Gross motor function classification system for cerebral palsy." In: *Dev Med Child Neurol* 39.4 (1997), pp. 214–23.

[98] Simon Reich, Dajie Zhang, Tomas Kulvicius, Sven Bölte, Karin Nielsen-Saines, Florian B. Pokorny, Robert Peharz, Luise Poustka, Florentin Wörgötter, Christa Einspieler, et al. "Novel AI driven approach to classify infant motor functions." In: *Scientific Reports* 11.1 (2021), pp. 1–13.

[99] Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. "Efficient backprop." In: *Neural networks: Tricks of the trade*. Springer, 2012, pp. 9–48.

[100] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. "Transfusion: Understanding transfer learning for medical imaging." In: *Advances in neural information processing systems* 32 (2019).

[101] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning deep features for discriminative localization." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2921–2929.

[102] United Nations. *The 17 goals*. `https://sdgs.un.org/goals`. Accessed on: 18 January 2022.

[103] Matteo Ruggero Ronchi and Pietro Perona. "Benchmarking and error diagnosis in multi-instance pose estimation." In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 369–378.

[104] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. “HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 5386–5395.

[105] Mingxing Tan, Ruoming Pang, and Quoc V. Le. “EfficientDet: Scalable and efficient object detection.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10781–10790.

[106] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. “Scaled-YOLOv4: Scaling cross stage partial network.” In: *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*. 2021, pp. 13029–13038.

[107] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. “Distilling the knowledge in a neural network.” In: *arXiv preprint arXiv:1503.02531* 2.7 (2015).

[108] Feng Zhang, Xiatian Zhu, and Mao Ye. “Fast human pose estimation.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3517–3526.

[109] John W. Tukey. *Exploratory data analysis*. Vol. 2. Reading, MA, 1977.

[110] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. “End-to-end training of deep visuomotor policies.” In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 1334–1373.

[111] Sheng Li, Mingxing Tan, Ruoming Pang, Andrew Li, Liqun Cheng, Quoc V. Le, and Norman P. Jouppi. “Searching for fast model families on datacenter accelerators.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 8085–8095.

[112] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. “A survey of model compression and acceleration for deep neural networks.” In: *arXiv preprint arXiv:1710.09282* (2017).

[113] Vijay Kotu and Bala Deshpande. *Predictive analytics and data mining: concepts and practice with RapidMiner*. Morgan Kaufmann, 2014.

[114] Daniel Hernández-Lobato, Gonzalo Martınez-Muñoz, and Alberto Suárez. “How large should ensembles of classifiers be?” In: *Pattern Recognition* 46.5 (2013), pp. 1323–1336.

[115] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and scalable predictive uncertainty estimation using deep ensembles.” In: *Advances in neural information processing systems* 30 (2017).

[116] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. “The kinetics human action video dataset.” In: *arXiv preprint arXiv:1705.06950* (2017).

[117] Zihang Dai, Hanxiao Liu, Quoc Le, and Mingxing Tan. “CoAtNet: Marrying convolution and attention for all data sizes.” In: *Advances in Neural Information Processing Systems* 34 (2021).

[118] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. “Convolutional sequence generation for skeleton-based action synthesis.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 4394–4402.

[119] Jingyuan Li and Eli Shlizerman. “Sparse semi-supervised action recognition with active learning.” In: *arXiv preprint arXiv:2012.01740* (2020).

[120] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. “Grad-CAM: Visual explanations from deep networks via gradient-based localization.” In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[121] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. “Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks.” In: *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2018, pp. 839–847.

[122] Nikolas Hesse, Christoph Bodensteiner, Michael Arens, Ulrich G. Hofmann, Raphael Weinberger, and Andreas S. Schroeder. “Computer vision for medical infant motion analysis: State of the art and RGB-D data set.” In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 2018.

[123] Amanda K.L. Kwong, Abbey L. Eeles, Joy E. Olsen, Diana Zannino, Timothy Kariotis, and Alicia J. Spittle. “Instructional guides for filming infant movements at home are effective for the General Movements Assessment.” In: *Journal of paediatrics and child health* (2021).

[124] Nikolas Hesse, Sergi Pujades, Javier Romero, Michael J. Black, Christoph Bodensteiner, Michael Arens, Ulrich G. Hofmann, Uta Tacke, Mijna Hadders-Algra, Raphael Weinberger, et al. “Learning an infant body model from RGB-D data for accurate full body motion analysis.” In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 792–800.

[125] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional networks for biomedical image segmentation.” In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[126] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. “Searching for activation functions.” In: *arXiv preprint arXiv:1710.05941* (2017).

[127] Hanna Mazzawi, Xavi Gonzalvo, Aleks Kracun, Prashant Sridhar, Niranjan Subrahmanya, Ignacio Lopez-Moreno, Hyun-Jin Park, and Patrick Violette. “Improving Keyword Spotting and Language Identification via Neural Architecture Search at Scale.” In: *Interspeech*. 2019, pp. 1278–1282.

[128] Asaf Noy, Niv Nayman, Tal Ridnik, Nadav Zamir, Sivan Doveh, Itamar Friedman, Raja Giryes, and Lihi Zelnik. “ASAP: Architecture search, anneal and prune.” In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 493–503.

Appendix A

Neural architecture search for graph-based convolutional networks

The architectures of GCNs for CP prediction from skeleton sequences were determined through automatic NAS based on the following search space and search strategy.

A.1 Search space

The search space, specifying the degrees of freedom of the GCN, comprised 20 architectural choices, yielding more than four billion possible novel GCN architectures. The choices included the number of stacked modules of alternating spatial graph convolutions and temporal convolutions in input branches and main branch (i.e., network depth), the width (i.e., number of channels) of modules, block types (i.e., types of convolutions), specification of graph convolution, type of residual connection and SE, number of parallel scales in temporal convolutions, type of attention mechanism, and so forth. All architectural choices and available alternatives are summarized in Table A.1, where it is also indicated which architectural building blocks are affected by the choice (i.e., input branches, main branch, or pooling layer) or the architecture overall (i.e., general properties).

Table A.1: The search space of 20 architectural choices with 2-4 possible alternatives.

	Architectural choice	**Alternatives**
Input branches	No. modules of input branches	1, 2, 3
	Width of input branches	6, 8, 10, 12
	Block type in initial module[a]	Basic, Bottleneck, MBConv
	Residual type in initial module[a]	None, Block [76], Module [76], Dense [76]
	No. temporal scales in input branches	1, 2, 3, Linear[b]
Main branch	No. levels of main branch	1, 2
	No. modules of main branch levels	1, 2, 3
	Width of first level of main branch[c]	6, 8, 10, 12
	No. temporal scales in main branch	1, 2, 3, Linear[b]
Pooling layer	Pooling layer type	Global average, Spatial average
General properties	Graph convolution type	Spatial configuration [44], DA 2[d] [78], DA 4[d] [78], DA 4+2[d] [78]
	Block type[e]	Basic, Bottleneck, MBConv
	Bottleneck factor	2, 4
	Residual type[e]	None, Block [76], Module [76], Dense [76]
	SE type	None, Inner [73], Outer [73], Both [73]
	SE ratio	2, 4
	SE ratio type	Relative, Absolute
	Attention type	None, Channel [76], Frame [76], Joint [76]
	Nonlinearity type	ReLU, Swish [126]
	Temporal kernel size	3, 5, 7, 9

[a] The initial module is the first module of input branches.
[b] Linear scaling indicates that number of temporal scales increases by one for each module.
[c] For the second level of the main branch, the width is doubled, while also reducing the time dimension by a factor of 2.
[d] Graph convolutions with disentangled aggregation (DA) have different number of hops in neighborhood (i.e., 2 or 4), where 4+2 yields separate number of hops in input modules and main module, with 4 and 2, respectively.
[e] There is a separate architecture choice associated with the initial module.

A.2 *K*-Best Search

To navigate the vast number of configurations in the architectural search space, a computationally efficient NAS strategy, K-Best Search, was developed, along with an effective estimation of performance of proposed architectures on CP prediction. The proposed NAS procedure is outlined in Algorithm 1.

Algorithm 1: K-Best Search

Input: Population size K, Architectural choices C, Performance threshold θ, Start temperature T_0, End temperature T_∞, Temperature drop τ, Training samples S_{train}, Validation samples S_{val}

Init: Population $P \leftarrow \emptyset$, Temperature $T \leftarrow T_0$, Unsuccessful trials $U \leftarrow 0$

while $|P| < K$ **do**
- $candidate \leftarrow$ `RandomAlternatives`(C)
- $performance(candidate) \leftarrow$ `TrainEval`$(candidate, S_{train}, S_{val})$
- **if** $performance(candidate) \geq \theta$ **then**
  - $P \leftarrow P \cup candidate$

while $U < K$ **do**
- $candidate \leftarrow$ `GoodAlternatives`(P, T)
- $performance(candidate) \leftarrow$ `TrainEval`$(candidate, S_{train}, S_{val})$
- **if** $performance(candidate) > \min_{x \in P} performance(x)$ **then**
  - $U \leftarrow 0$
  - $P \leftarrow (P \setminus \operatorname{argmin}_{x \in P} performance(x)) \cup candidate$
- **else**
  - $U \leftarrow U + 1$
  - **if** $U = K$ **and** $T > T_\infty$ **then**
    - $U \leftarrow 0$
    - $T \leftarrow T - \tau$

return $\operatorname{argmax}_{x \in P} performance(x)$

Table A.2: Different values of K in K-Best Search.

	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$
AUC S_{val}	0.937	0.939	0.935	0.949	0.940	0.946
No. trials	10	22	25	49	65	78

Inspired by the efforts of Mazzawi et al. [127], K-Best Search greedily utilizes the K best-performing GCN architectures to construct architectural candidates containing alternatives of architectural choices present in the population of K architectures. Moreover, motivated by Noy et al. [128], annealing was employed, ensuring high degree of exploration of different alternatives in the early phase of the search (i.e., high temperature), while gradually cooling down to ensure higher degree of exploitation of best-performing alternatives in the later phase of the search. More specifically, $GoodAlternatives$ composes an architecture $candidate$ from architectural choices $c \in C$ with probability $Prob$ of selecting an alternative $a \in A$ being controlled by its ranking R in the current population (e.g., if alternative X is present in the 1st and 2nd best-performing candidates in the population with $K = 5$, whereas alternative Y occurs in the 3rd, 4th, and 5th best-performing candidates, then $R(X) = 5 + 4 = 9$ and $R(Y) = 3 + 2 + 1 = 6$), along with temperature T determining the importance of ranking:

$$Prob(a) = \frac{exp(\frac{R(a)}{T})}{\sum_{a' \in A} exp(\frac{R(a')}{T})} \tag{A.1}$$

The performance of candidate architectures was estimated by $TrainEval$ based on the AUC of the candidate across subjects in the two validation folds that ST-GCN [44] performed best and worst at, val2 and val7 according to Table A.5, comprising $S_{val} = S_{val2} \cup S_{val7}$, yielding a proxy for the accuracy of 7-fold cross-validation accuracy. The remaining infants of the training and validation dataset ($S_{train} = S_{val1} \cup S_{val3} \cup S_{val4} \cup S_{val5} \cup S_{val6}$) were used for supervised training of the proposed architecture for 100 epochs. An early stopping scheme was employed to leave out candidates not converging towards high AUC, with checkpoints at epochs 10, 20, 30, 40, 50, 60, 70, 80, and 90, with associated AUC criteria of 0.75, 0.775, 0.8, 0.825, 0.85, 0.875, 0.9, 0.925, and 0.95. Start temperature T_0, end temperature T_∞, and temperature drop τ were set to 10, 1, and 3, respectively, whereas the performance threshold θ for inclusion in the initial population was AUC ≥ 0.9. The appropriate value of K, providing trade-off between the achieved AUC and number of candidate architectures (i.e., trials), was determined as 5 based on the saturation of AUC observed at $K > 5$ in Table A.2. Table A.3 displays that K-Best Search outperformed Random Search across 675 trials (10 iterations of K-Best Search with $K = 5$), both with respect to the number of trials surpassing each AUC checkpoint (e.g., 335 (49.6%) trials of AUC ≥ 0.9 compared to 144 (21.3%) for Random Search), median AUC, and highest AUC overall.

Table A.3: Performance comparison of *K*-Best Search and Random Search across 675 trials.

	No. trials with AUC S_{val}										Median AUC (IQR)	Highest AUC
	< 0.75	≥ 0.75	≥ 0.775	≥ 0.8	≥ 0.825	≥ 0.85	≥ 0.875	≥ 0.9	≥ 0.925	≥ 0.95		
Random Search	385 (57.0%)	290 (43.0%)	279 (41.3%)	266 (39.4%)	256 (37.9%)	242 (35.9%)	214 (31.7%)	144 (21.3%)	32 (4.7%)	0 (0.0%)	0.720 ([0.640, 0.894])	0.944
***K*-Best Search**	236 (35.0%)	439 (65.0%)	428 (63.4%)	423 (62.7%)	417 (61.8%)	407 (60.3%)	390 (57.8%)	335 (49.6%)	158 (23.4%)	4 (0.6%)	0.900 ([0.695, 0.925])	0.956

A.3 Ensemble-NAS-GCN

The K-Best Search was repeated 10 times to yield 10 novel GCNs suited for the task of CP prediction (Table A.4). The performance of each of the GCNs, as well as the ST-GCN baseline, in 7-fold cross-validation is provided in Table A.5. Moreover, similar to Mazzawi et al. [127], intra-model ensembles combined the seven instances of each GCN, to yield performance on the test set. The 10 intra-model ensembles, originating from K-Best Search, were merged into the ensemble of 70 GCN instances composing Ensemble-NAS-GCN.

Table A.4: Characteristics of architectures obtained by K-Best Search. Extension of eTable 4 in Paper III.

Architectural choice	***K*-Best Search no.**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
No. modules of input branches	3	2	3	3	2	3	1	2	3	2
Width of input branches	10	10	12	10	10	12	8	6	6	12
Block type in initial module	Bottleneck	Basic	Basic	Basic	Bottleneck	Basic	Basic	MBConv	Bottleneck	Basic
Residual type in initial module	None	Dense	None	Block	Dense	Dense	Module	Block	Dense	Dense
No. temporal scales in input branches	1	3	2	2	3	1	3	2	1	2
No. levels of main branch	1	1	1	1	2	2	2	2	2	1
No. modules of main branch levels	1	3	2	1	1	3	3	2	1	3
Width of first level of main branch	12	12	8	8	6	10	12	12	12	10
No. temporal scales in main branch	1	2	2	Linear	3	Linear	3	1	Linear	3
Pooling layer type	Global	Global	Global	Spatial	Global	Global	Spatial	Global	Spatial	Spatial
Graph convolution type	DA 2	DA 4+2	SC	DA 4	DA 4	DA 2	DA 2	DA 4	DA 2	SC
Block type	Basic	MBConv	Basic	Basic	Basic	Bottleneck	Basic	Basic	Basic	Basic
Bottleneck factor	4	2	2	4	2	4	4	4	4	4
Residual type	None	Block	Module	Dense	None	Block	None	Dense	None	None
SE type	None	Outer	Inner	None	Outer	None	None	Outer	Outer	None
SE ratio	-	4	2	-	2	-	-	4	4	-
SE ratio type	-	Absolute	Absolute	-	Absolute	-	-	Absolute	Absolute	-
Attention type	Channel	Channel	None	Channel	Channel	Channel	None	None	Channel	Channe
Nonlinearity type	ReLU	Swish	ReLU	Swish	ReLU	Swish	Swish	ReLU	ReLU	Swish
Temporal kernel size	9	7	7	7	7	3	9	5	9	7
AUC S_{val}	**0.949**	**0.942**	**0.938**	**0.943**	**0.937**	**0.956**	**0.953**	**0.953**	**0.932**	**0.947**
No. trials	**49**	**60**	**85**	**65**	**68**	**72**	**50**	**85**	**60**	**81**

Abbreviations: DA, disentangled aggregation; SC, spatial configuration.

Table A.5: AUC of the proposed GCNs on validation folds 1-7 (i.e., val1-val7) and test set in relation to ST-GCN.

	AUC								Computational efficiency	
	Val1	**Val2**	**Val3**	**Val4**	**Val5**	**Val6**	**Val7**	**Test**[a]	**Parameters**	**GFLOPs**
ST-GCN [44]	0.904	0.989	0.887	0.887	0.864	0.933	0.747	0.851	3 075 740	6.21
***K*-Best Search no. 1**	0.941	1.000	0.909	0.950	0.953	0.944	0.882	0.921	233 362	0.76
***K*-Best Search no. 2**	0.928	0.991	0.922	0.913	0.960	0.878	0.869	0.917	75 764	0.26
***K*-Best Search no. 3**	0.950	1.000	0.930	0.939	0.967	0.947	0.876	0.903	174 767	0.58
***K*-Best Search no. 4**	0.939	1.000	0.904	0.965	0.927	0.902	0.902	0.906	430 812	1.62
***K*-Best Search no. 5**	0.954	0.998	0.924	0.937	0.958	0.900	0.867	0.921	107 233	0.32
***K*-Best Search no. 6**	0.928	1.000	0.898	0.963	0.882	0.891	0.895	0.931	236 180	0.44
***K*-Best Search no. 7**	0.956	1.000	0.928	0.963	0.964	0.878	0.906	0.918	1 691 288	2.48
***K*-Best Search no. 8**	0.913	1.000	0.911	0.952	0.936	0.884	0.906	0.904	400 962	0.60
***K*-Best Search no. 9**	0.928	1.000	0.926	0.967	0.931	0.911	0.900	0.929	554 652	0.63
***K*-Best Search no. 10**	0.954	1.000	0.930	0.946	0.927	0.898	0.891	0.916	559 899	1.81

[a] AUC on the test set was evaluated based on the predictions of the intra-model ensemble.

Appendix B

Single-person pose estimation

Table B.1: Localization performance of ConvNets for single-person pose estimation on the validation and test sets of MPII [79] in PCK_h@0.5 and PCK_h@0.1, and computational efficiency in number of parameters and billions of floating-point operations (GFLOPs). Extension of Table 3 and 4 in Paper I.

ConvNet	Localization performance				Computational efficiency	
	MPII validation		MPII test			
	PCK_h@0.5	PCK_h@0.1	PCK_h@0.5	PCK_h@0.1	Parameters	GFLOPs
OpenPose [64]	87.60	22.76	88.78	22.49	25 939 664	160.36
EfficientPose RT	82.88	23.56	84.78	24.17	460 284	0.87
EfficientPose I	85.18	26.49	-	-	719 736	1.67
EfficientPose II	88.18	30.17	-	-	1 732 944	7.70
EfficientPose III	89.51	30.90	-	-	3 229 772	23.35
EfficientPose IV	89.75	35.63	91.18	35.97	6 560 938	72.89
EfficientPose RT Lite	80.58	23.08	-	-	404 608	0.86
EfficientPose I Lite	83.65	27.71	-	-	591 632	1.54
EfficientPose II Lite	87.13	30.81	-	-	1 464 960	7.25
EfficientHourglass B4	90.22	34.81	91.2	35.12	18 696 471	26.62

Appendix C

Infant pose estimation

Table C.1: Localization performance of ConvNets for infant pose estimation on the test set of In-Motion Poses, in terms of mean error (ME) and different thresholds τ of $PCK_h@\tau$, including 1.0, 0.5, 0.3, 0.2, and 0.1, as well as ${H_b}^{0.95}$ for $PCK_h@Human^{0.95}$, and computational efficiency in number of parameters, billions of floating-point operations (GFLOPs), inference latency in milliseconds, and processing speed in frames per second (FPS). Adapted from Table 1 and 2 of Paper II.

ConvNet	Localization performance							Computational efficiency			
	ME	**1.0**	**0.5**	**0.3**	**0.2**	**0.1**	**$H_b^{0.95}$**	**Parameters**	**GFLOPs**	**Latency**	**Speed**
OpenPose library [91]	0.1432	96.99	95.51	90.90	81.49	49.66	-	-	-	62.33	16.04
OpenPose [64]	0.1087	99.94	99.61	97.65	90.40	54.89	62.03	26 011 743	161.08	35.21	28.40
CIMA-Pose [90]	0.0988	99.98	99.83	98.74	93.09	59.69	66.58	2 380 495	15.65	11.49	87.03
EfficientPose RT	0.1022	99.96	99.69	98.15	92.15	58.71	-	481 336	0.96	5.06	197.54
EfficientPose I	0.0974	99.98	99.83	98.81	93.68	60.78	-	743 476	1.79	7.05	141.91
EfficientPose II	0.0969	99.97	99.84	98.54	92.41	62.25	-	1 759 372	7.94	19.38	51.61
EfficientPose III	0.0731	99.99	99.94	99.54	97.57	78.21	81.59	3 258 888	23.78	41.92	23.86
EfficientPose IV	0.0834	99.98	99.93	99.45	96.77	71.10	-	6 595 430	73.62	96.48	10.37
EfficientHourglass B4	0.0681	99.99	99.95	99.56	97.67	81.11	86.71	18 699 936	27.01	47.01	21.27

Appendix D

Automated CP prediction

Table D.1: Predictive values for CP of GMA, CIMA model, ST-GCN, and Ensemble-NAS-GCN, and associated 95% confidence intervals, in the test set of 139 high-risk infants (21 with CP and 118 without CP) for external validation, from true and false positives (TP and FP) and true and false negatives (TN and FN), given a fixed sensitivity of 70.0%. Extension of Table 1 and eFigure 2 of Paper III.

Method	TP	FP	TN	FN	Sensitivity	Specificity	PPV	NPV	Accuracy	AUC
GMA [18]	14	13	102	6	70.0 [45.7, 88.1]	88.7 [81.5, 93.8]	51.9 [32.0, 71.3]	94.4 [88.3, 97.9]	85.9 [78.9, 91.3]	-
CIMA model [62]	15	32	86	6	71.4 [47.8, 88.7]	72.9 [63.9, 80.7]	31.9 [19.1, 47.1]	93.5 [86.3, 97.6]	72.7 [64.5, 79.9]	0.839
ST-GCN [44]	15	19	99	6	71.4 [47.8, 88.7]	83.9 [76.0, 90.0]	44.1 [27.2, 62.1]	94.3 [88.0, 97.9]	82.0 [74.6, 88.0]	0.851
Ensemble-NAS-GCN	15	7	111	6	71.4 [47.8, 88.7]	94.1 [88.2, 97.6]	68.2 [45.1, 86.1]	94.9 [89.2, 98.1]	90.6 [84.5, 94.9]	0.921

Paper I

Applied Intelligence (2021) 51:2518–2533
https://doi.org/10.1007/s10489-020-01918-7

EfficientPose: Scalable single-person pose estimation

Daniel Groos[1] · **Heri Ramampiaro**[2] · **Espen AF Ihlen**[1]

Accepted: 28 August 2020 / Published online: 6 November 2020

Abstract

Single-person human pose estimation facilitates markerless movement analysis in sports, as well as in clinical applications. Still, state-of-the-art models for human pose estimation generally do not meet the requirements of real-life applications. The proliferation of deep learning techniques has resulted in the development of many advanced approaches. However, with the progresses in the field, more complex and inefficient models have also been introduced, which have caused tremendous increases in computational demands. To cope with these complexity and inefficiency challenges, we propose a novel convolutional neural network architecture, called EfficientPose, which exploits recently proposed EfficientNets in order to deliver efficient and scalable single-person pose estimation. EfficientPose is a family of models harnessing an effective multi-scale feature extractor and computationally efficient detection blocks using mobile inverted bottleneck convolutions, while at the same time ensuring that the precision of the pose configurations is still improved. Due to its low complexity and efficiency, EfficientPose enables real-world applications on edge devices by limiting the memory footprint and computational cost. The results from our experiments, using the challenging MPII single-person benchmark, show that the proposed EfficientPose models substantially outperform the widely-used OpenPose model both in terms of accuracy and computational efficiency. In particular, our top-performing model achieves state-of-the-art accuracy on single-person MPII, with low-complexity ConvNets.

Keywords Human pose estimation · Model scalability · High precision · Computational efficiency · Openly available

1 Introduction

Single-person human pose estimation (HPE) refers to the computer vision task of localizing human skeletal keypoints of a person from an image or video frames. Single-person HPE has many real-world applications, ranging from outdoor activity recognition and computer animation to clinical assessments of motor repertoire and skill practice among professional athletes. The proliferation of deep convolutional neural networks (ConvNets) has advanced HPE and further widen its application areas. ConvNet-based HPE with its increasingly complex network structures, combined with transfer learning, is a very challenging task. However, the availability of high-performing ImageNet [9] backbones, together with large tailor-made datasets, such as MPII for 2D pose estimation [1], has facilitated the development of new improved methods to address the challenges.

An increasing trend in computer vision has driven towards more efficient models [11, 38, 46]. Recently, EfficientNet [47] was released as a scalable ConvNet architecture, setting benchmark record on ImageNet with a more computationally efficient architecture. However, within human pose estimation, there is still a lack of architectures that are both accurate and computationally efficient at the same time. In general, current state-of-the-art architectures are computationally expensive and highly complex, thus making them hard to replicate, cumbersome to optimize, and impractical to embed into real-world applications.

✉ Daniel Groos
daniel.groos@ntnu.no

Heri Ramampiaro
heri@ntnu.no

Espen AF Ihlen
espen.ihlen@ntnu.no

[1] Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, Trondheim, Norway

[2] Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway

The OpenPose network [6] (OpenPose for short) has been one of the most applied HPE methods in real-world applications. It is also the first open-source real-time system for HPE. OpenPose was originally developed for multi-person HPE, but has in recent years been frequently applied to various single-person applications within clinical research and sport sciences [15, 32, 34]. The main drawback with OpenPose is that the level of detail in keypoint estimates is limited due to its low-resolution outputs. This makes OpenPose less suitable for precision-demanding applications, such as elite sports and medical assessments, which all depend on high degree of precision in the assessment of movement kinematics. Moreover, by spending 160 billion floating-point operations (GFLOPs) per inference, OpenPose is considered highly inefficient. Despite these issues, OpenPose seems to remain a commonly applied network for single-person HPE performing markerless motion capture from which critical decisions are based upon [2, 56].

In this paper, we stress the lack of publicly available methods for single-person HPE that are both computationally efficient and effective in terms of estimation precision. To this end, we exploit recent advances in ConvNets and propose an improved approach called EfficientPose. Our main idea is to modify OpenPose into a family of scalable ConvNets for high-precision and computationally efficient single-person pose estimation from 2D images. To assess the performance of our approach, we perform two separate comparative studies. First, we evaluate the EfficientPose model by comparing it against the original OpenPose model on single-person HPE. Second, we compare it against the current state-of-the-art single-person HPE methods on the official MPII challenge, focusing on accuracy as a function of the number of parameters. The proposed EfficientPose models aim to elicit high computational efficiency, while bridging the gap in availability of high-precision HPE networks.

In summary, the main contributions of this paper are the following:

- We propose an improvement of OpenPose, called EfficientPose, that can overcome the shortcomings of the popular OpenPose network on single-person HPE with improved level of precision, rapid convergence during optimization, low number of parameters, and low computational cost.
- With EfficientPose, we suggest an approach providing scalable models that can suit various demands, enabling a trade-off between accuracy and efficiency across diverse application constraints and limited computational budgets.
- We propose a new way to incorporate mobile ConvNet components, which can address the need for computationally efficient architectures for HPE, thus facilitating real-time HPE on the edge.
- We perform an extensive comparative study to evaluate our approach. Our experimental results show that the proposed method achieves significantly higher efficiency and accuracy in comparison to the baseline method, OpenPose. In addition, compared to existing state-of-the-art methods, it achieves competitive results, with a much smaller number of parameters.

The remainder of this paper is organized as follows: Section 2 describes the architecture of OpenPose and highlights research which it can be improved from. Based on this, Section 3 presents our proposed ConvNet-based approach, EfficientPose. Section 4 describes our experiments and presents the results from comparing EfficientPose with OpenPose and other existing approaches. Section 5 discusses our findings and suggests potential future studies. Finally, Section 6 summarizes and concludes the paper.

For the sake of reproducibility, we will make the EfficientPose models available at https://github.com/daniegr/EfficientPose.

2 Related work

The proliferation of ConvNets for HPE following the success of DeepPose [54] has set the path for accurate HPE. With OpenPose, Cao et al. [6] made HPE available to the public. As depicted by Fig. 1, OpenPose comprises a multi-stage architecture performing a series of detection passes. Provided an input image of 368 × 368 pixels, OpenPose utilizes an ImageNet pretrained VGG-19 backbone [41] to extract basic features (step 1 in Fig. 1). The features are supplied to a DenseNet-inspired detection block (step 2) arranged as five dense blocks [23], each containing three 3 × 3 convolutions with PReLU activations [20]. The detection blocks are stacked in a sequence. First, four passes (step 3a-d in Fig. 1) of part affinity fields [7] map the associations between body keypoints. Subsequently, two detection passes (step 3e and 3f) predict keypoint heatmaps [53] to obtain refined keypoint coordinate estimates. In terms of level of detail in the keypoint coordinates, OpenPose is restricted by its output resolution of 46 × 46 pixels.

The OpenPose architecture can be improved by recent advancements in ConvNets, as follows: First, automated network architecture search has found backbones [47, 48, 62] that are more precise and efficient in image classification than VGG and ResNets [21, 41]. In particular, Tan and Le [47] proposed compound model scaling to balance the image resolution, width (number of network

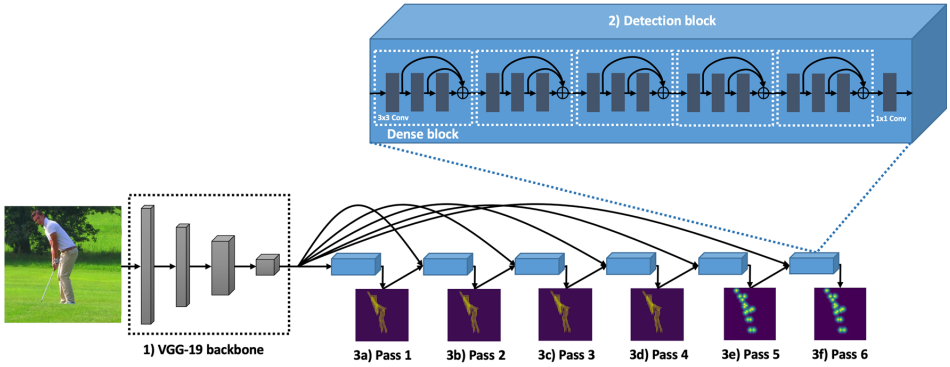


Fig. 1 OpenPose architecture utilizing 1) VGG-19 feature extractor, and 2) detection blocks performing 4+2 passes of estimating part affinity fields (3a-d) and confidence maps (3e and 3f)

channels), and depth (number of network layers). This resulted in scalable convolutional neural networks, called EfficientNets [47], with which the main goal was to provide lightweight models with a sensible trade-off between model complexity and accuracy across various computational budgets. For each model variant EfficientNet-Bϕ, from the most computationally efficient one being EfficientNet-B0 to the most accurate model, EfficientNet-B7 ($\phi \in [0, 7] \in \mathbb{Z}^{\geq}$), the total number of FLOPs increases by a factor of 2, given by

$$(\alpha \cdot \beta^2 \cdot \gamma^2)^{\phi} \approx 2^{\phi}. \tag{1}$$

Here, α, β and γ denote the coefficients for depth, width, and resolution, respectively, and are set as

$$\alpha = 1.2, \beta = 1.1, \gamma = 1.15. \tag{2}$$

Second, parallel multi-scale feature extraction has improved the precision levels in HPE [25, 33, 44, 57], emphasizing both high spatial resolution and low-scale semantics. However, existing multi-scale approaches in HPE are computationally expensive, both due to their large size and high computational requirements. For example, a typical multi-scale HPE approach has often a size of $16 - 58$ million parameters and requires $10 - 128$ GFLOPS [8, 33, 36, 44, 49, 57, 61]. To cope with this, we propose cross-resolution features, operating on high- and low-resolution input images, to integrate features from multiple abstraction levels with low overhead in network complexity and with high computational efficiency. Existing works on Siamese ConvNets have been promising in utilizing parallel network backbones [17, 18]. Third, mobile inverted bottleneck convolution (MBConv) [38] with built-in squeeze-and-excitation (SE) [22] and Swish activation [37] integrated in EfficientNets has proven more accurate in image classification tasks [47, 48] than regular convolutions [21, 23, 45], while substantially reducing the computational costs [47]. The efficiency of MBConv modules stem from the depthwise convolutions operating in a channel-wise manner [40]. With this approach, it is possible to reduce the computational cost by a factor proportional to the number of channels [48]. Hence, by replacing the regular 3×3 convolutions with up to 384 input channels in the detection blocks of OpenPose with MBConvs, we can obtain more computationally efficient detection blocks. Further, SE selectively emphasizes discriminative image features [22], which may reduce the required number of convolutions and detection passes by providing a global perspective on the estimation task at all times. Using MBConv with SE may have the potential to decrease the number of dense blocks in OpenPose. Fourth, transposed convolutions with bilinear kernel [30] scale up the low-resolution feature maps, thus enabling a higher level of detail in the output confidence maps.

By building upon the work of Tan and Le [47], we present a pool of scalable models for single-person HPE that is able to overcome the shortcomings of the commonly adopted OpenPose architecture. This enables trading off between accuracy and efficiency across different computational budgets in real-world applications. The main advantage of this is that we can use ConvNets that are small and computationally efficient enough to run on edge devices with little memory and low processing power, which is impossible with OpenPose.

3 The EfficientPose approach

In this section, we explain in details the EfficientPose approach. This includes a detailed description of the EfficientPose architecture in light of the OpenPose architecture, and a brief introduction to the proposed variants of EfficientPose.

3.1 Architecture

Figures 1 and 2 depict the architectures of OpenPose and EfficientPose, respectively. As can be observed in these two figures, although being based on OpenPose, the EfficientPose architecture is different from the OpenPose architecture in several aspects, including 1) both high and low-resolution input images, 2) scalable EfficientNet backbones, 3) cross-resolution features, 4) and 5) scalable Mobile DenseNet detection blocks in fewer detection passes, and 6) bilinear upscaling. For a more thorough component analysis of EfficientPose, see Appendix A.

The input of the network consists of high and low-resolution images (1a and 1b in Fig. 2). To get the low-resolution image, the high-resolution image is downsampled into half of its pixel height and width, through an initial average pooling layer.

The feature extractor of EfficientPose is composed of the initial blocks of EfficientNets [47] pretrained on ImageNet (step 2a and b in Fig. 2). High-level semantic information is obtained from the high-resolution image using the initial three blocks of a EfficientNet with $\phi \in [0, 7]$ (see (1)), outputting C feature maps (2a in Fig. 2). Low-level local information is extracted from the low-resolution image by the first two blocks of a lower-scale EfficientNet-backbone (2b in Fig. 2) in the range $\phi \in [0, 3]$. Table 1 provides an overview of the composition of EfficientNet backbones, from low-scale B0 to high-scale B7. The first block of EfficientNets utilizes the MBConvs shown in Fig. 3a and b, whereas the second and third blocks comprise the MBConv layers in Fig. 3c and d.

The features generated by the low-level and high-level EfficientNet backbones are concatenated to yield cross-resolution features (step 3 in Fig. 2). This enables the EfficientPose architecture to selectively emphasize important local factors from the image of interest and the overall structures that guide high-quality pose estimation. In this way, we enable an alternative simultaneous handling of different features at multiple abstraction levels.

From the extracted features, the desired keypoints are localized through an iterative detection process, where each detection pass performs supervised prediction of output maps. Each detection pass comprises a detection block and a single 1×1 convolution for output prediction. The detection blocks across all detection passes elicit the same basic architecture, comprising Mobile DenseNets (see step 4 in Fig. 2). Data from Mobile DenseNets are forwarded to subsequent layers of the detection block using residual connections. The Mobile DenseNet is inspired by DenseNets [23] supporting reuse of features, avoiding redundant layers, and MBConv with SE, thus enabling low memory footprint. In our adaptation of the MBConv operation ($E\text{-}MBConv6(K \times K, B, S)$ in Fig. 3e), we consistently utilize the highest performing combination from [46], i.e., a kernel size ($K \times K$) of 5×5 and an expansion ratio of 6. We also avoid downsampling (i.e., $S = 1$) and scale the width of Mobile DenseNets by outputting number of channels relative to the high-level

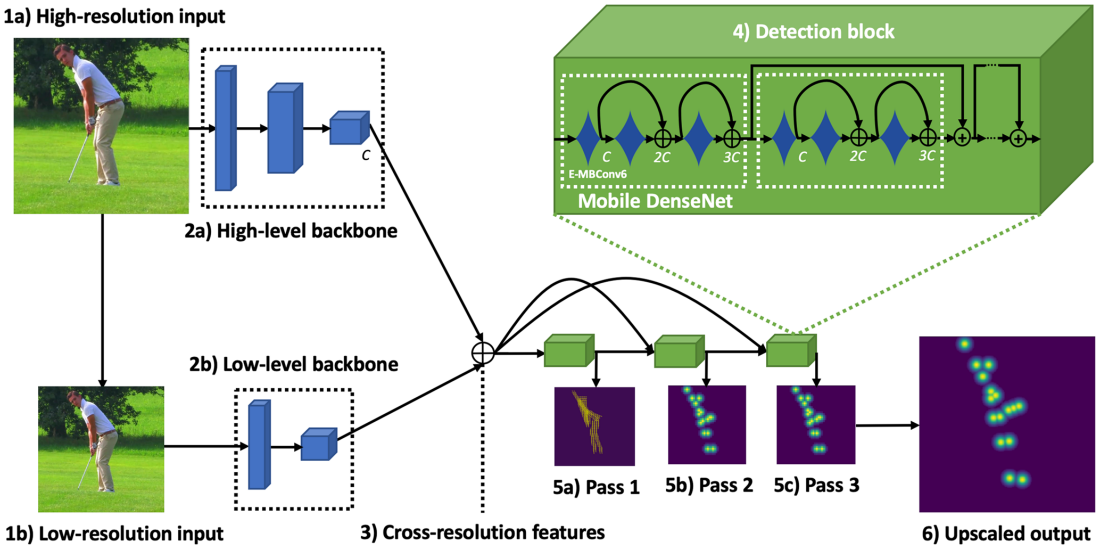


Fig. 2 Proposed architecture comprising 1a) high-resolution and 1b) low-resolution inputs, 2a) high-level and 2b) low-level EfficientNet backbones combined into 3) cross-resolution features, 4) Mobile DenseNet detection blocks, 1+2 passes for estimation of part affinity fields (5a) and confidence maps (5b and 5c), and 6) bilinear upscaling

Table 1 The architecture of the initial three blocks of relevant EfficientNet backbones

Block	B0	B1	B2	B3	B4	B5	B7
1	$Conv(3\times3,32,2)\,BN\,Swish$			$Conv(3\times3,40,2)\,BN\,Swish$	$Conv(3\times3,48,2)\,BN\,Swish$		$Conv(3\times3,64,2)\,BN\,Swish$
	$MBConv1(3\times3,16,1)$			$MBConv1(3\times3,24,1)$			$MBConv1(3\times3,32,1)$
	–	$MBConv1^*(3\times3,16,1)$		$MBConv1^*(3\times3,24,1)$		$\left[MBConv1^*(3\times3,24,1)\right]\times2$	$\left[MBConv1^*(3\times3,32,1)\right]\times3$
2	$MBConv6(3\times3,24,2)$			$MBConv6(3\times3,32,2)$		$MBConv6(3\times3,40,2)$	$MBConv6(3\times3,48,2)$
	$MBConv6^*(3\times3,24,1)$	$\left[MBConv6^*(3\times3,24,1)\right]\times2$		$\left[MBConv6^*(3\times3,32,1)\right]\times2$	$\left[MBConv6^*(3\times3,32,1)\right]\times3$	$\left[MBConv6^*(3\times3,40,1)\right]\times4$	$\left[MBConv6^*(3\times3,48,1)\right]\times6$
3	$MBConv6(5\times5,40,2)$		$MBConv6\ (5\times5,48,2)$		$MBConv6(5\times5,56,2)$	$MBConv6(5\times5,64,2)$	$MBConv6(5\times5,80,2)$
	$MBConv6^*(5\times5,40,1)$	$\left[MBConv6^*(5\times5,40,1)\right]\times2$	$\left[MBConv6^*(5\times5,48,1)\right]\times2$		$\left[MBConv6^*(5\times5,56,1)\right]\times3$	$\left[MBConv6^*(5\times5,64,1)\right]\times4$	$\left[MBConv6^*(5\times5,80,1)\right]\times6$
I	224×224	240×240	260×260	300×300	380×380	456×456	600×600
C	40		48		56	64	80
α^ϕ	$1.2^0=1.0$	$1.2^1=1.2$	$1.2^2=1.4$	$1.2^3=1.7$	$1.2^4=2.1$	$1.2^5=2.5$	$1.2^7=3.6$

For $Conv(K\times K,N,S)$, $K\times K$ denotes filter size, N is number of output feature maps, and S is stride. BN denotes batch normalization. I defines input size, corresponding with image resolution on ImageNet, whereas α^ϕ refers to the depth factor as determined by (1)

backbone ($B = C$). We modify the original $MBConv6$ operation by incorporating E-swish as activation function with β value of 1.25 [16]. This has a tendency to accelerate progression during training compared to the regular Swish activation [37]. We also adjust the first 1×1 convolution to generate a number of feature maps relative to the output feature maps B rather than the input channels M. This reduces the memory consumption and computational latency since $B \leq M$, with $C \leq M \leq 3C$. With each Mobile DenseNet consisting of three consecutive $E\text{-}MBConv6$ operations, the module outputs $3C$ feature maps.

EfficientPose performs detection in two rounds (step 5a-c in Fig. 2). First, the overall pose of the person is anticipated through a single pass of skeleton estimation (5a). This aims to facilitate the detection of feasible poses and to avoid confusion in case of several persons being present in an image. Skeleton estimation is performed utilizing part affinity fields as proposed in [7]. Following skeleton estimation, two detection passes are performed to estimate heatmaps for keypoints of interest. The former of these acts as a coarse detector (5b in Fig. 2), whereas the latter (5c in Fig. 2) refines localization to yield more accurate outputs.

Note that in OpenPose, the heatmaps of the final detection pass are constrained to a low spatial resolution, which are incapable of achieving the amount of details that are normally inherent in the high-resolution input [6]. To improve this limitation of OpenPose, a series of three transposed convolutions performing bilinear upsampling are added for $8\times$ upscaling of the low-resolution heatmaps (step 6 in Fig. 1). Thus, we project the low-resolution output onto a space of higher resolution in order to allow an increased level of detail. To achieve the proper level of interpolation while operating efficiently, each transposed convolution increases the map size by a factor of 2, using a stride of 2 with a 4×4 kernel.

3.2 Variants

Following the same principle as suggested in the original EfficientNet [47], we scale the EfficientPose network architecture by adjusting the three main dimensions, i.e., input resolution, network width, and network depth, using the coefficients of (2). The results from this scaling are five different architecture variants that are given in Table 2, referred to as EfficientPose I to IV and RT). As can be observed in this table, the input resolution, defined by the spatial dimensions of the image ($H\times W$), is scaled utilizing the high and low-level EfficientNet backbones that best match the resolution of high and low-resolution inputs (see Table 1). Here, the network width refers to the number of feature maps that are generated by each $E\text{-}MBConv6$. As described in Section 3.1, width scaling is achieved using the

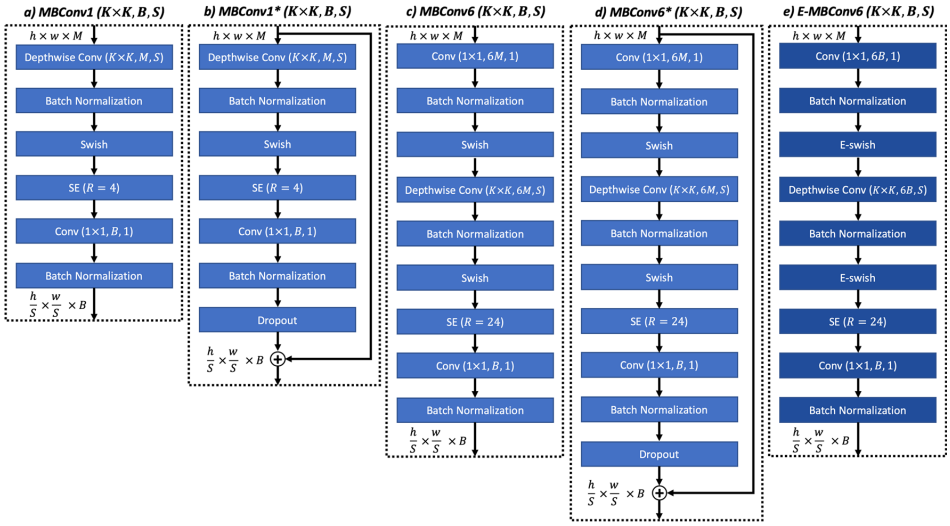


Fig. 3 The composition of MBConvs. From left: a-d) $MBConv(K \times K, B, S)$ in EfficientNets performs depthwise convolution with filter size $K \times K$ and stride S, and outputs B feature maps. $MBConv^*$ (b and d) extends regular MBConvs by including dropout layer and skip connection. e) $E\text{-}MBConv6(K \times K, B, S)$ in Mobile DenseNets adjusts $MBConv6$ with E-swish activation and number of feature maps in expansion phase as $6B$. All MBConvs take as input M feature maps with spatial height and width of h and w, respectively. R is the reduction ratio of SE

same width as the high-level backbone (i.e., C). The scaling of network depth is achieved in the number of Mobile DenseNets (i.e., $MD(C)$ in Table 2) in the detection blocks. Also, this ensures that receptive fields across different models and spatial resolutions have similar relative sizes. For each model variant, we select the number (D) of Mobile DenseNets that best approximates the original depth factor α^{ϕ} in the high-level EfficientNet backbone (Table 1). More specifically, the number of Mobile DenseNets are determined by (3), rounding to the closest integer. In addition to EfficientPose I to IV, the single-resolution model EfficientPose RT is formed to match the scale of the smallest EfficientNet model, providing HPE in extremely low latency applications.

$$D = \lfloor \alpha^{\phi} + 0.5 \rfloor \tag{3}$$

3.3 Summary of proposed framework

As can be inferred from the discussion above, the EfficientPose framework comprises a family of five ConvNets (i.e., EfficientPose I-IV and RT) that are constructed by compound scaling [47]. With this, EfficientPose exploits the advances in computationally efficient ConvNets for image recognition to

Table 2 Variants of EfficientPose obtained by scaling resolution, width, and depth

Stage	EfficientPose RT	EfficientPose I	EfficientPose II	EfficientPose III	EfficientPose IV
High-resolution input	224×224	256×256	368×368	480×480	600×600
High-level backbone	B0 (Block 1-3)	B2 (Block 1-3)	B4 (Block 1-3)	B5 (Block 1-3)	B7 (Block 1-3)
Low-resolution input	–	128×128	184×184	240×240	300×300
Low-level backbone	–	B0 (Block 1-2)	B0 (Block 1-2)	B1 (Block 1-2)	B3 (Block 1-2)
Detection block	$MD(40)$	$MD(48)$	$[MD(56)] \times 2$	$[MD(64)] \times 3$	$[MD(80)] \times 4$
Prediction pass 1	$Conv(1 \times 1, 2P, 1)$				
Prediction pass 2-3	$Conv(1 \times 1, Q, 1)$				
Upscaling	$[Conv^T(4 \times 4, Q, 2)] \times 3$				

Mobile DenseNets $MD(C)$ computes $3C$ feature maps. P and Q denotes the number of 2D part affinity fields and confidence maps, respectively. $Conv^T(K \times K, O, S)$ defines transposed convolutions with kernel size $K \times K$, output maps O, and stride S

Fig. 4 The MPII single-person pose estimation challenge. From left: a) 10 images from the MPII test set displaying some of the variation and difficulties inherent in this challenge. b) The evaluation metrics $PCK_h@50$ and $PCK_h@10$ define the average of predictions within τl distance ($l = 0.6d$) from the ground-truth location (e.g., left elbow), with τ being 50% and 10%, respectively

construct a scalable network architecture that is capable of performing single-person HPE across different computational constraints. More specifically, EfficientPose utilizes both high and low-resolution images to provide two separate viewpoints that are processed independently through high and low-level backbones, respectively. The resulting features are concatenated to produce cross-resolution features, enabling selective emphasis on global and local image information. The detection stage employs a scalable mobile detection block to perform detection in three passes. The first pass estimates person skeletons through part affinity fields [7] to yield feasible pose configurations. The second and third passes estimate keypoint locations with progressive improvement in precision. Finally, the low-resolution prediction of the third pass is scaled up through bilinear interpolation to further improve the precision level.

4 Experiments and results

4.1 Experimental setup

We evaluate EfficientPose and compare it with OpenPose on the single-person MPII dataset [1], containing images of mainly healthy adults in a wide range of different outdoor and indoor everyday activities and situations, such as sports, fitness exercises, housekeeping activities, and public events (Fig. 4a). All models are optimized on MPII using stochastic gradient descent (SGD) on the mean squared error (MSE) of the model predictions relative to the target coordinates. More specifically, we applied SGD with momentum and cyclical learning rates (see Appendix B for more information and further details on the optimization procedure). The learning rate is bounded according to the model-specific value of which it does not diverge during the first cycle (λ_{max}) and $\lambda_{min} = \frac{\lambda_{max}}{3000}$. The model backbones (i.e., VGG-19 for OpenPose, and EfficientNets for EfficientPose) are initialized with pretrained ImageNet weights, whereas the remaining layers employ random weight initialization. Supported by our experiments on training efficiency (see Appendix A), we train the models for 200 epochs, except for OpenPose, which requires a higher number of epochs to converge (see Fig. 5 and Table 5).

The training and validation portion of the dataset comprises 29K images, and by adopting a standard random split, we obtain 26K and 3K instances for training and validation, respectively. We augment the images during training using random horizontal flipping, scaling ($0.75 - 1.25$), and rotation ($+/- 45$ degrees). We utilize a batch size of 20, except for the high-resolutional EfficientPose III and IV, which both require a smaller batch size to fit into the GPU memory, 10 and 5, respectively. The experiments are carried out on an NVIDIA Tesla V100 GPU.

The evaluation of model accuracy is performed using the $PCK_h@\tau$ metric. $PCK_h@\tau$ is defined as the fraction of predictions residing within a distance τl from the ground truth location (see Fig. 4b). l is 60% of the diagonal d of the head bounding box, and τ the accepted percentage of misjudgment relative to l. $PCK_h@50$ is the standard performance metric for MPII but we also include the stricter $PCK_h@10$ metric for assessing models' ability to yield highly precise keypoint estimates. As commonly done in the field, the final model predictions are obtained by applying multi-scale testing procedure [44, 49, 57]. Due to the restriction in the number of attempts for official evaluation on MPII, we only used the test metrics on the OpenPose baseline, and the most efficient and most accurate models, EfficientPose RT and EfficientPose IV, respectively. To measure model efficiency, both FLOPs and number of parameters are supplied.

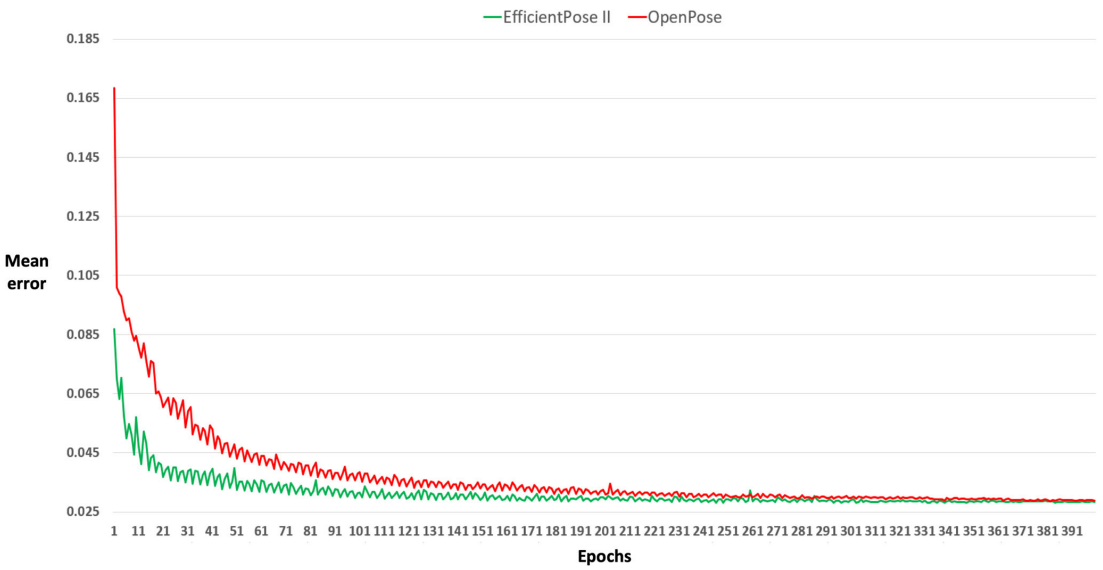


Fig. 5 The progression of the mean error of EfficientPose II and OpenPose on the MPII validation set during the course of training

4.2 Results

Table 3 shows the results of our experiments with OpenPose and EfficientPose on the MPII validation dataset. As can be observed in this table, EfficientPose consistently outperformed OpenPose with regards to efficiency, with $2.2 - 184\times$ reduction in FLOPs and $4 - 56\times$ fewer number of parameters. In addition to this, all the model variants of EfficientPose achieved better high-precision localization, with a $0.8 - 12.9\%$ gain in $PCK_h@10$ as compared to OpenPose. In terms of $PCK_h@50$, the high-end models, i.e., EfficientPose II-IV, managed to gain $0.6 - 2.2\%$ improvements against OpenPose. As Table 4 depicts, EfficientPose IV achieved state-of-the-art results (a mean $PCK_h@50$ of 91.2) on the official MPII test dataset for models with number of parameters of a size less than 10 million.

Compared to OpenPose, EffcientPose also exhibited rapid convergence during training. We optimized both approaches on similar input resolution, which defaults to 368×368 for OpenPose, corresponding to EfficientPose II. The training graph shown in Fig. 5 demonstrates that EfficientPose converges early, whereas OpenPose requires up to 400 epochs before achieving proper convergence. Nevertheless, OpenPose benefited from this prolonged training in terms of precision, with a 2.6% improvement in $PCK_h@50$ during the final 200 epochs, whereas EfficientPose II had a minor gain of 0.4% (see Table 5).

5 Discussion

In this section, we discuss several aspects of our findings and possible avenues for further research.

Table 3 Performance of EfficientPose compared to OpenPose on the MPII validation dataset, as evaluated by efficiency (number of parameters and FLOPs, and relative reduction in parameters and FLOPs compared to OpenPose) and accuracy (mean $PCK_h@50$ and mean $PCK_h@10$)

Model	Parameters	Parameter reduction	FLOPs	FLOP reduction	$PCK_h@50$	$PCK_h@10$
OpenPose [6]	$25.94M$	$1\times$	$160.36G$	$1\times$	87.60	22.76
EfficientPose RT	$0.46M$	$56\times$	$0.87G$	$184\times$	82.88	23.56
EfficientPose I	$0.72M$	$36\times$	$1.67G$	$96\times$	85.18	26.49
EfficientPose II	$1.73M$	$15\times$	$7.70G$	$21\times$	88.18	30.17
EfficientPose III	$3.23M$	$8.0\times$	$23.35G$	$6.9\times$	89.51	30.90
EfficientPose IV	$6.56M$	$4.0\times$	$72.89G$	$2.2\times$	89.75	35.63

Table 4 State-of-the-art results in $PCK_h@50$ (both for individual body parts and overall mean value) on the official MPII test dataset [1] compared to the number of parameters

Model	Parameters	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
Pishchulin et al., ICCV'13 [35]	–	74.3	49.0	40.8	32.1	36.5	34.4	35.2	44.1
Tompson et al., NIPS'14 [53]	–	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
Lifshitz et al., ECCV'16 [28]	76M	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
Tang et al., BMVC'18 [50]	10M	97.4	96.2	91.8	87.3	90.0	87.0	83.3	90.8
Newell et al., ECCV'16 [33]	26M	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Zhang et al., CVPR'19 [60]	3M	98.3	96.4	91.5	87.4	90.9	87.1	83.7	91.1
Bulat et al., FG'20 [5]	9M	98.5	96.4	91.5	87.2	90.7	86.9	83.6	91.1
Yang et al., ICCV'17 [57]	27M	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Tang et al., ECCV'18 [49]	16M	98.4	96.9	92.6	88.7	91.8	89.4	86.2	92.3
Sun et al., CVPR'19 [44]	29M	98.6	96.9	92.8	89.0	91.5	89.0	85.7	92.3
Zhang et al., arXiv'19 [61]	24M	98.6	97.0	92.8	88.8	91.7	89.8	86.6	92.5
OpenPose [6]	25.94M	97.7	94.7	89.5	84.7	88.4	83.6	79.3	88.8
EfficientPose RT	0.46M	97.0	93.3	85.0	79.2	85.9	77.0	71.0	84.8
EfficientPose IV	6.56M	98.2	96.0	91.7	87.9	90.3	87.5	83.9	91.2

5.1 Improvements over OpenPose

The precision of HPE methods is a key success factor for analyses of movement kinematics, like segment positions and joint angles, for assessment of sport performance in athletes, or motor disabilities in patients. Facilitated by cross-resolution features and upscaling of output (see Appendix A), EfficientPose achieved a higher precision than OpenPose [6], with a 57% relative improvement in $PCK_h@10$ on single-person MPII (Table 3). What this means is that the EfficientPose architecture is generally more suitable in performing precision-demanding single-person HPE applications, like medical assessments and elite sports, than OpenPose.

Another aspect to have in mind is that, for some applications (e.g., exercise games and baby monitors), we might be more interested in the latency of the system and its ability to respond quickly. Hence, the degree of correctness in keypoint predictions might be less crucial.

Table 5 Model accuracy on the MPII validation dataset in relation to the number of training epochs

Model	Epochs	$PCK_h@50$
OpenPose [6]	100	80.47
OpenPose [6]	200	85.00
OpenPose [6]	400	87.60
EfficientPose II	100	87.05
EfficientPose II	200	88.18
EfficientPose II	400	88.56

In such scenarios, with applications that demand high-speed predictions, the 460K parameter model, EfficientPose RT, consuming less than one GFLOP, would be suitable. Nevertheless, it still manages to provide higher precision level than current approaches in the high-speed regime, e.g., [5, 50]. Further, the scalability of EfficientPose enables flexibility in various situations and across different types of hardware, whereas OpenPose suffers from its large number of parameters and computational costs (FLOPs).

5.2 Strengths of the EfficientPose approach

The use of MBConv in HPE is to the best of our knowledge an unexplored research area. This has also been partly our main motivation for exploring the use of MBConv in our EfficientPose approach, recognizing its success in image classification [47]. Our experimental results showed that EfficientPose approached state-of-the-art performance on the single-person MPII benchmark despite a large reduction in the number of parameters (Table 4). This means that the parameter-efficient MBConvs provide value in HPE as with other computer vision tasks, such as image classification and object detection. This, in turns, makes MBConv a very suitable component for HPE networks. For this reason, it would be interesting to investigate the effect of combining it with other novel HPE architectures, such as Hourglass and HRNet [33, 44].

Further, the use of EfficientNet as a backbone, and the proposed cross-resolution feature extractor combining several EfficientNets for improved handling of basic features, are also interesting avenues to explore further. From the present study, it is reasonable to assume that EfficientNets could

replace commonly used backbones for HPE, such as VGG and ResNets, which would reduce the computational overheads associated with these approaches [21, 41]. Also, a cross-resolution feature extractor could be useful for precision-demanding applications by providing an improved performance on $PCK_h@10$ (Table 6).

We also observed that EfficientPose benefited from compound model scaling across resolution, width and depth. This benefit was reflected by the increasing improvements in $PCK_h@50$ and $PCK_h@10$ from EfficientPose RT through EfficientPose I to EfficientPose IV (Table 3). To conclude, we can exploit this to further examine scalable ConvNets for HPE, and thus obtain insights into appropriate sizes of HPE models (i.e., number of parameters), required number of FLOPs, and obtainable precision levels.

In this study, OpenPose and EfficientPose were optimized on the general-purpose MPII Human Pose Dataset. For many applications (e.g., action recognition and video surveillance) the variability in MPII may be sufficient for directly applying the models on real-world problems. Nonetheless, there are other particular scenarios that deviate from the setting addressed in this paper. The MPII dataset comprises mostly healthy adults in a variety of every day indoor and outdoor activities [1]. In less natural environments (e.g., movement science laboratories or hospital settings) and with humans of different anatomical proportions such as children and infants [39], careful consideration must be taken. This could include a need for fine-tuning of the MPII models on more specific datasets related to the problem at hand. As mentioned earlier, our experiments showed that EfficientPose was more easily trainable than OpenPose (Fig. 5 and Table 5). This trait of rapid convergence suggests that exploring the use of transfer learning on the EfficientPose models on other HPE data could provide interesting results.

5.3 Avenues for further research

The precision level of pose configurations provided by EfficientPose in the context of target applications is a topic considered beyond the scope of this paper and has for this reason been left for further studies. We can establish the validity of EfficientPose for robust single-person pose estimation already by examining whether the movement information supplied by the proposed framework is of sufficiently good quality for tackling challenging problems, such as complex human behavior recognition [12, 29]. To assess this, we could, for example, compare the precision level of the keypoint estimates supplied by EfficientPose with the movement information provided by body-worn movement sensors. Moreover, we could combine the proposed image-based EfficientPose models with body-worn sensors, such as inertial measurement unit (IMU) [27], or physiological signals, like electrical cardiac activity and electrical brain activity [14], to potentially achieve improved precision levels and an increased robustness. Our hypothesis is that using body-worn sensors or physiological instruments could be useful in situations where body parts are extensively occluded, such that camera-based recognition alone may not be sufficient for accurate pose estimation.

Another path for further study and validation is the capability of EfficientPose to perform multi-person HPE. The improved computational efficiency of EfficientPose compared to OpenPose has the potential to also benefit multi-person HPE. State-of-the-art methods for multi-person HPE are dominated by top-down approaches, which require computation that is normally proportional to the number of individuals in the image [13, 59]. In crowded scenes, top-down approaches are highly resource demanding. Similar to the original OpenPose [6], and few other recent works on multi-person HPE [19, 24], EfficientPose incorporates part affinity fields, which would enable the grouping of keypoints into persons, and thus allowing to perform multi-person HPE in a bottom-up manner. This would reduce the computational overhead into a single network inference per image, and hence yield more computationally efficient multi-person HPE.

Further, it would be interesting to explore the extension of the proposed framework to perform 3D pose estimation as part of our future research. In accordance with recent studies, 3D pose projection from 2D images can be achieved, either by employing geometric relationships between 2D keypoint positions and 3D human pose models [58], or by leveraging occlusion-robust pose-maps (ORPM) in combination with annotated 3D poses [3, 31].

The architecture of EfficientPose and the training process can be improved in several ways. First, the optimization procedure (see Appendix B) was developed for maximum $PCK_h@50$ accuracy on OpenPose, and simply reapplied to EfficientPose. Other optimization procedures might be more appropriate, including alternative optimizers (e.g., Adam [26] and RMSProp [52]), and other learning rate and sigma schedules.

Second, only the backbone of EfficientPose was pretrained on ImageNet. This could restrict the level of accuracy on HPE because large-scale pretraining not only supplies robust basic features but also higher-level semantics. Thus, it would be valuable to assess the effect of pretraining on model precision in HPE. We could, for example, pretrain the majority of ConvNet layers on ImageNet, and retrain these on HPE data.

Third, the proposed compound scaling of EfficientPose assumes that the scaling relationship between resolution, width, and depth, as defined by (2), is identical in HPE and image classification. However, the optimal compound

scaling coefficients might be different for HPE, where the precision level is more dependent on image resolution, than for image classification. Based on this, a topic for further studies could be to conduct neural architecture search across different combinations of resolution, width, and depth in order to determine the optimal combination of scaling coefficients for HPE. Regardless of the scaling coefficients, the scaling of detection blocks in EfficientPose could be improved. The block depth (i.e., number of Mobile DenseNets) slightly deviates from the original depth coefficient in EfficientNets based on the rigid nature of the Mobile DenseNets. A carefully designed detection block could address this challenge by providing more flexibility with regards to the number of layers and the receptive field size.

Fourth, the computational efficiency of EfficientPose could be further improved by the use of teacher-student network training (i.e., knowledge distillation) [4] to transfer knowledge from a high-scale EfficientPose teacher network to a low-scale EfficientPose student network. This technique has already shown promising results in HPE when paired with the stacked hourglass architecture [33, 60]. Sparse networks, network pruning, and weight quantization [11, 55] could also be included in the study to facilitate the development of more accurate and responsive real-life systems for HPE. Finally, for high performance inference and deployment on edge devices, further speed-up could be achieved by the use of specialized libraries such as NVIDIA TensorRT and TensorFlow Lite [10, 51].

In summary, EfficientPose tackles single-person HPE with an improved degree of precision compared to the commonly adopted OpenPose network [6]. In addition to this, the EfficientPose models have the ability to yield high performance with a large reduction in number of parameters and FLOPs. This has been achieved by exploiting the findings from contemporary research within image recognition on computationally efficient ConvNet components, most notably MBConvs and EfficientNets [38, 47]. Again, for the sake of reproducibility, we have made the EfficientPose models publicly available for other researchers to test and possibly further development.

6 Conclusion

In this work, we have stressed the need for a publicly accessible method for single-person HPE that suits the demands for both precision and efficiency across various applications and computational budgets. To this end, we have presented a novel method called EfficientPose, which is a scalable ConvNet architecture leveraging a computationally efficient multi-scale feature extractor, novel mobile detection blocks, skeleton estimation, and bilinear upscaling. In order to have model variants that are able to flexibly find a sensible trade-off between accuracy and efficiency, we have exploited model scalability in three dimensions: input resolution, network width, and network depth. Our experimental results have demonstrated that the proposed approach has the capability to offer computationally efficient models, allowing real-time inference on edge devices. At the same time, our framework offers flexibility to be scaled up to deliver more precise keypoint estimates than commonly used counterparts, at an order of magnitude less parameters and computational costs (FLOPs). Taking into account the efficiency and high precision level of our proposed framework, there is a reason to believe that EfficientPose will provide an important foundation for the next-generation markerless movement analysis.

In our future work, we plan to develop new techniques to further improve the model effectiveness, especially in terms of precision, by investigating optimal compound model scaling for HPE. Moreover, we will deploy EfficientPose on a range of applications to validate its applicability, as well as feasibility, in real-world scenarios.

Acknowledgements The research is funded by RSO funds from the Faculty of Medicine and Health Sciences at the Norwegian University of Science and Technology. The experiments were carried out utilizing computational resources provided by the Norwegian Open AI Lab.

Funding Open Access funding provided by NTNU Norwegian University of Science and Technology (incl St. Olavs Hospital - Trondheim University Hospital);.

Appendix A Ablation study

To determine the effect of different design choices in the EfficientPose architecture, we carried out component analysis.

Training efficiency

We assessed the number of training epochs to determine the appropriate duration of training, avoiding demanding

Table 6 Model accuracy on the MPII validation dataset in relation to the use of cross-resolution features

Model	Cross-resolution features	Parameters	FLOPs	$PCK_h@50$	$PCK_h@10$
EfficientPose I	✓	$0.72M$	$1.67G$	83.56	26.35
EfficientPose I		$0.68M$	$1.58G$	83.64	25.79
EfficientPose II	✓	$1.73M$	$7.70G$	87.05	29.87
EfficientPose II		$1.69M$	$7.50G$	86.93	29.16

optimization processes. Figure 5 suggests that the largest improvement in model accuracy occurs until around 200 epochs, after which training saturates. Table 5 supports this observation with less than 0.4% increase in $PCK_h@50$ with 400 epochs of training. From this, it was decided to perform the final optimization of the different variants of EfficientPose over 200 epochs. Table 5 also suggests that most of the learning progress occurs during the first 100 epochs. Hence, for the remainder of the ablation study 100 epochs were used to determine the effect of different design choices.

Cross-resolution features

The value of combining low-level local information with high-level semantic information through a cross-resolution feature extractor was evaluated by optimizing the model with and without the low-level backbone. Experiments were conducted on two different variants of the EfficientPose model. On coarse prediction ($PCK_h@50$) there is little to no gain in accuracy (Table 6), whereas for fine estimation ($PCK_h@10$) some improvement ($0.6 - 0.7\%$) is displayed taking into account the negligible cost of $1.02 - 1.06\times$ more parameters and $1.03 - 1.06\times$ increase in FLOPs.

Skeleton estimation

The effect of skeleton estimation through the approximation of part affinity fields was assessed by comparing the architecture with and without the single pass of skeleton estimation. Skeleton estimation yields improved accuracy with $1.3 - 2.4\%$ gain in $PCK_h@50$ and $0.2 - 1.4\%$ in $PCK_h@10$ (Table 7), while only introducing an overhead in number of parameters and computational cost of $1.3 - 1.4\times$ and $1.2 - 1.3\times$, respectively.

Number of detection passes

We also determined the appropriate comprehensiveness of detection, represented by number of detection passes. EfficientPose I and II were both optimized on three different variants (Table 8). Seemingly, the models benefit from intermediate supervision with a general trend of increased performance level in accordance with number of detection passes. The major benefit in performance is obtained by expanding from one to two passes of keypoint estimation, reflected by $1.6 - 1.7\%$ increase in $PCK_h@50$ and $1.8 - 1.9\%$ in $PCK_h@10$. In comparison, a third detection pass yields only $0.5 - 0.8\%$ relative improvement in $PCK_h@50$ compared to two passes, and no gain in $PCK_h@10$ while increasing number of parameters and computation by $1.3\times$ and $1.2\times$, respectively. From these findings, we decided a beneficial trade-off in accuracy and efficiency would be the use of two detection passes.

Upscaling

To assess the impact of upscaling, implemented as bilinear transposed convolutions, we compared the results of the two respective models. Table 9 reflects that upscaling yields improved precision on keypoint estimates by large gains of $9.2 - 12.3\%$ in $PCK_h@10$ and smaller improvements of $0.5 - 1.1\%$ on coarse detection ($PCK_h@50$). As a consequence of increased output resolution upscaling slightly increases number of FLOPs ($1.04 - 1.1\times$) with neglectable increase in number of parameters.

Appendix B Optimization procedure

Most state-of-the-art approaches for single-person pose estimation are extensively pretrained on ImageNet [44, 61],

Table 7 Model accuracy on the MPII validation dataset in relation to the use of skeleton estimation

Model	Skeleton estimation	Parameters	FLOPs	$PCK_h@50$	$PCK_h@10$
EfficientPose I	✓	$0.72M$	$1.67G$	83.56	26.35
EfficientPose I		$0.54M$	$1.37G$	81.13	25.00
EfficientPose II	✓	$1.73M$	$7.70G$	87.05	29.87
EfficientPose II		$1.27M$	$6.03G$	85.75	29.67

Table 8 Model accuracy on the MPII validation dataset in relation to the number of detection passes

Model	Detection passes	Parameters	FLOPs	$PCK_h@50$	$PCK_h@10$
EfficientPose I	1	$0.52M$	$1.33G$	81.85	24.51
EfficientPose I	2	$0.72M$	$1.67G$	83.56	26.35
EfficientPose I	3	$0.92M$	$2.02G$	84.35	26.42
EfficientPose II	1	$1.24M$	$5.92G$	85.42	28.01
EfficientPose II	2	$1.73M$	$7.70G$	87.05	29.87
EfficientPose II	3	$2.22M$	$9.49G$	87.55	29.61

Table 9 Model accuracy on the MPII validation dataset in relation to the use of upscaling

Model	Upscaling	Parameters	FLOPs	$PCK_h@50$	$PCK_h@10$
EfficientPose I	✓	$0.72M$	$1.67G$	83.56	26.35
EfficientPose I		$0.71M$	$1.52G$	82.42	14.02
EfficientPose II	✓	$1.73M$	$7.70G$	87.05	29.87
EfficientPose II		$1.73M$	$7.37G$	86.56	20.66

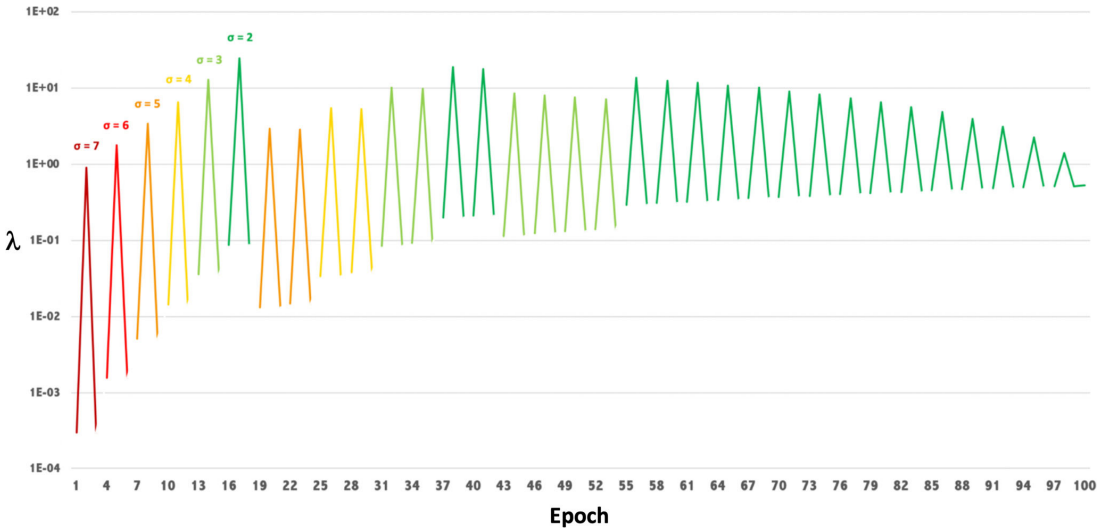


Fig. 6 Optimization scheme displaying learning rates λ and σ values corresponding to the training of EfficientPose II over 100 epochs

enabling rapid convergence for models when adapted to other tasks, such as HPE. In contrast to these approaches, few models, including OpenPose [6] and EfficientPose, only utilize the most basic pretrained features. This facilitates construction of more efficient network architectures but at the same time requires careful design of optimization procedures for convergence towards reasonable parameter values.

Training of pose estimation models is complicated due to the intricate nature of output responses. Overall, optimization is performed in a conventional fashion by minimizing the MSE of the predicted output maps Y with respect to ground truth values $\hat{Y}$ across all output responses N.

The predicted output maps should ideally have higher values at the spatial locations corresponding to body part positions, while punishing predictions farther away from the correct location. As a result, the ground truth output maps must be carefully designed to enable proper convergence during training. We achieve this by progressively reducing the circumference from the true location that should be rewarded, defined by the σ parameter. Higher probabilities $T \in [0, 1]$ are assigned for positions P closer to the ground truth position G (4).

$$T_i = \exp\left(-\frac{\|P_i - G\|_2^2}{\sigma^2}\right) \tag{4}$$

The proposed optimization scheme (Fig. 6) incorporates a stepwise σ scheme, and utilizes SGD with momentum of 0.9 and a decaying triangular cyclical learning rate (CLR) policy [42]. The σ parameter is normalized according to the output resolution. As suggested by Smith and Topin [43], the large learning rates in CLR provides regularization in network optimization. This makes training more stable and may even increase training efficiency. This is valuable for network architectures, such as OpenPose and EfficientPose, less heavily concerned with pretraining (i.e., having larger portions of randomized weights). In our adoption of CLR, we utilize a cycle length of 3 epochs. The learning rate (λ) converges towards λ_∞ (5), where λ_{max} is the highest learning rate for which the model does not diverge during the first cycle and $\lambda_{min} = \frac{\lambda_{max}}{3000}$, whereas σ_0 and σ_∞ are the initial and final sigma values, respectively.

$$\lambda_\infty = 10^{\frac{\log(\lambda_{max})+\log(\lambda_{min})}{2}} \cdot 2^{\sigma_0-\sigma_\infty} \tag{5}$$

References

1. Andriluka M, Pishchulin L, Gehler P, Schiele B (2014) 2d human pose estimation: new benchmark and state of the art analysis. In: IEEE Conference on computer vision and pattern recognition (CVPR)
2. Barra P, Bisogni C, Nappi M, Freire-Obregón D, Castrillón-Santana M (2019) Gait analysis for gender classification in forensics. In: International conference on dependability in sensor, cloud, and big data systems and applications. Springer, New York, pp 180–190
3. Benzine A, Luvison B, Pham QC, Achard C (2020) Single-shot 3d multi-person pose estimation in complex images. Pattern Recognition p 107534
4. Buciluǎ C, Caruana R, Niculescu-Mizil A (2006) Model compression. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, pp 535–541
5. Bulat A, Kossaifi J, Tzimiropoulos G, Pantic M (2020) Toward fast and accurate human pose estimation via soft-gated skip connections. In: FG
6. Cao Z, Hidalgo Martinez G, Simon T, Wei S, Sheikh YA (2019) Openpose: realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence
7. Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7291–7299
8. Chu X, Yang W, Ouyang W, Ma C, Yuille AL, Wang X (2017) Multi-context attention for human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1831–1840
9. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on computer vision and pattern recognition, Ieee, pp 248–255
10. Developer N (2020) NVIDIA TensorRT (accessed February 23, 2020). https://developer.nvidia.com/tensorrt
11. Elsen E, Dukhan M, Gale T, Simonyan K (2020) Fast sparse convnets. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14629–14638
12. Fernando T, Denman S, Sridharan S, Fookes C (2018) Tracking by prediction: a deep generative model for mutli-person localisation and tracking. In: 2018 IEEE winter conference on applications of computer vision (WACV), IEEE, pp 1122–1132
13. Fieraru M, Khoreva A, Pishchulin L, Schiele B (2018) Learning to refine human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 205–214
14. Fiorini L, Mancioppi G, Semeraro F, Fujita H, Cavallo F (2020) Unsupervised emotional state classification through physiological parameters for social robotics applications. Knowl-Based Syst 190:105217
15. Firdaus NM, Rakun E (2019) Recognizing fingerspelling in sibi (sistem isyarat bahasa indonesia) using openpose and elliptical fourier descriptor. In: Proceedings of the international conference on advanced information science and system, pp 1–6
16. Gagana B, Athri HU, Natarajan S (2018) Activation function optimizations for capsule networks. In: 2018 international conference on advances in computing, communications and informatics (ICACCI), IEEE, pp 1172–1178
17. Gao P, Yuan R, Wang F, Xiao L, Fujita H, Zhang Y (2019) Siamese attentional keypoint network for high performance visual tracking. Knowledge-based Systems p 105448
18. Gao P, Zhang Q, Wang F, Xiao L, Fujita H, Zhang Y (2020) Learning reinforced attentional representation for end-to-end visual tracking. Inf Sci 517:52–67
19. Guan CZ (2019) Realtime multi-person 2d pose estimation using shufflenet. In: 2019 14Th international conference on computer science & education (ICCSE), IEEE, pp 17–21

20. He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision, pp 1026–1034
21. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
22. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
23. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
24. Huang Y, Shum HP, Ho ES, Aslam N (2020) High-speed multi-person pose estimation with deep feature transfer. Computer Vision and Image Understanding p 103010
25. Ke L, Chang MC, Qi H, Lyu S (2018) Multi-scale structure-aware network for human pose estimation. In: Proceedings of the european conference on computer vision (ECCV), pp 713–728
26. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: ICLR
27. Kundu AS, Mazumder O, Lenka PK, Bhaumik S (2018) Hand gesture recognition based omnidirectional wheelchair control using imu and emg sensors. J Intell Robot Syst 91(3-4):529–541
28. Lifshitz I, Fetaya E, Ullman S (2016) Human pose estimation using deep consensus voting. In: European conference on computer vision. Springer, New York, pp 246–260
29. Liu L, Wang S, Hu B, Qiong Q, Wen J, Rosenblum DS (2018) Learning structures of interval-based bayesian networks in probabilistic generative model for human complex activity recognition. Pattern Recogn 81:545–561
30. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440
31. Mehta D, Sotnychenko O, Mueller F, Xu W, Sridhar S, Pons-Moll G, Theobalt C (2018) Single-shot multi-person 3d pose estimation from monocular rgb. In: 2018 international conference on 3d vision (3DV), IEEE, pp 120–130
32. Nakai M, Tsunoda Y, Hayashi H, Murakoshi H (2018) Prediction of basketball free throw shooting by openpose. In: JSAI International symposium on artificial intelligence. Springer, New York, pp 435–446
33. Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In: European conference on computer vision. Springer, New York, pp 483–499
34. Noori FM, Wallace B, Uddin MZ, Torresen J (2019) A robust human activity recognition approach using openpose, motion features, and deep recurrent neural network. In: Scandinavian conference on image analysis. Springer, New York, pp 299–310
35. Pishchulin L, Andriluka M, Gehler P, Schiele B (2013) Poselet conditioned pictorial structures. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 588–595
36. Rafi U, Leibe B, Gall J, Kostrikov I (2016) An efficient convolutional network for human pose estimation. In: BMVC, vol 1, p 2
37. Ramachandran P, Zoph B, Le QV (2018) Searching for activation functions. In: 6th international conference on learning representations, ICLR 2018
38. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC (2018) Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4510–4520
39. Sciortino G, Farinella GM, Battiato S, Leo M, Distante C (2017) On the estimation of children's poses. In: International conference on image analysis and processing. Springer, New York, pp 410–421
40. Sifre L, Mallat S (2014) Rigid-motion scattering for image classification. Ph. D thesis
41. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: ICLR
42. Smith LN (2017) Cyclical learning rates for training neural networks. In: 2017 IEEE winter conference on applications of computer vision (WACV), IEEE, pp 464–472
43. Smith LN, Topin N (2019) Super-convergence: very fast training of neural networks using large learning rates. In: Artificial intelligence and machine learning for multi-domain operations applications, international society for optics and photonics, vol 11006, p 1100612
44. Sun K, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5693–5703
45. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence
46. Tan M, Chen B, Pang R, Vasudevan V, Sandler M, Howard A, Le QV (2019) Mnasnet: platform-aware neural architecture search for mobile. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2820–2828
47. Tan M, Le QV (2019) Efficientnet: rethinking model scaling for convolutional neural networks. In: ICML
48. Tan M, Le QV (2019) Mixconv: mixed depthwise convolutional kernels. In: BMVC
49. Tang W, Yu P, Wu Y (2018) Deeply learned compositional models for human pose estimation. In: Proceedings of the european conference on computer vision (ECCV), pp 190–206
50. Tang Z, Peng X, Geng S, Zhu Y, Metaxas DN (2018) Cu-net: coupled u-nets. In: BMVC
51. TensorFlow (2020) Deploy machine learning models on mobile and IoT devices (accessed February 23, 2020). https://www.tensorflow.org/lite
52. Tieleman T, Hinton G (2012) Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. COURSERA Neural Netw Mach Learn 4(2):26–31
53. Tompson JJ, Jain A, LeCun Y, Bregler C (2014) Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in neural information processing systems, pp 1799–1807
54. Toshev A, Szegedy C (2014) Deeppose: human pose estimation via deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1653–1660
55. Tung F, Mori G (2018) Clip-q: deep network compression learning by in-parallel pruning-quantization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7873–7882
56. Vitali A, Regazzoni D, Rizzi C, Maffioletti F (2019) A new approach for medical assessment of patient's injured shoulder. In: International design engineering technical conferences and computers and information in engineering conference. american society of mechanical engineers, vol 59179, p v001t02a049
57. Yang W, Li S, Ouyang W, Li H, Wang X (2017) Learning feature pyramids for human pose estimation. In: proceedings of the IEEE international conference on computer vision, pp 1281–1290
58. Yuan H, Li M, Hou J, Xiao J (2020) Single image-based head pose estimation with spherical parametrization and 3d morphing. Pattern Recognition p 107316

59. Zhang F, Zhu X, Dai H, Ye M, Zhu C (2020) Distribution-aware coordinate representation for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7093–7102
60. Zhang F, Zhu X, Ye M (2019) Fast human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3517–3526
61. Zhang H, Ouyang H, Liu S, Qi X, Shen X, Yang R, Jia J (2019) Human pose estimation with spatial contextual information. arXiv:1901.01760
62. Zoph B, Vasudevan V, Shlens J, Le QV (2018) Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8697–8710

Daniel Groos received the M.Sc. degree in Computer Science from the Norwegian University of Science and Technology (NTNU). He is currently PhD Research Fellow at the Department of Neuromedicine and Movement Science, NTNU, Trondheim, Norway. His research interests include deep learning-based computer vision, markerless human movement analysis, and medical technology.

Heri Ramampiaro is Professor at the Department of Computer Science, NTNU, Trondheim, Norway. He is Head of the Data and Artificial Intelligence (DART) research group and Deputy head of the department. Ramampiaro holds a PhD degree in Computer Science from NTNU, M.Sc. from Stavanger University College (now University of Stavanger), and B.Eng. from Aalesund College (now NTNU). His current main research interests include machine learning, data/text mining, information retrieval, and NLP.

Espen AF Ihlen is Associate Professor at the Department of Neuromedicine and Movement Science, NTNU, Trondheim, Norway. Ihlen holds a PhD degree in Clinical Medicine and M.Sc. in Human Movement Science from NTNU. His main interest is development of convolutional neural network architectures and machine learning for human movement analysis.

Paper II

Computerized Medical Imaging and Graphics 95 (2022) 102012

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Towards human-level performance on automatic pose estimation of infant spontaneous movements

Daniel Groos [a], Lars Adde [b,d], Ragnhild Støen [b,e], Heri Ramampiaro [c], Espen A.F. Ihlen [a,*]

[a] Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, Trondheim, Norway
[b] Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway
[c] Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway
[d] Clinic of Clinical Services, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway
[e] Department of Neonatology, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway

ARTICLE INFO

Keywords:
Computer-based risk assessment
Convolutional neural networks
Developmental disorders
Infant pose estimation
Markerless video-based analysis

ABSTRACT

Assessment of spontaneous movements can predict the long-term developmental disorders in high-risk infants. In order to develop algorithms for automated prediction of later disorders, highly precise localization of segments and joints by infant pose estimation is required. Four types of convolutional neural networks were trained and evaluated on a novel infant pose dataset, covering the large variation in 1424 videos from a clinical international community. The localization performance of the networks was evaluated as the deviation between the estimated keypoint positions and human expert annotations. The computational efficiency was also assessed to determine the feasibility of the neural networks in clinical practice. The best performing neural network had a similar localization error to the inter-rater spread of human expert annotations, while still operating efficiently. Overall, the results of our study show that pose estimation of infant spontaneous movements has a great potential to support research initiatives on early detection of developmental disorders in children with perinatal brain injuries by quantifying infant movements from video recordings with human-level performance.

1. Introduction

During the first months of life, spontaneous infant movements may indicate later developmental disorders, such as cerebral palsy (CP), Rett syndrome, and autism spectrum disorder (Novak et al., 2017; Einspieler et al., 2005, 2014). Early identification of infants at high risk for developmental disorders is essential in order to successfully select appropriate follow-up approaches, and is of greatest importance in research to evaluate early interventions (Støen et al., 2017). The expert-based observation of general movements (GMs) from video recordings, known as the general movement assessment (GMA) (Einspieler et al., 2004), has recently been recommended for clinical use in high-risk infants less than five months of age (Novak et al., 2017). It is especially the fidgety type of GMs, which typically occur between two and five months post-term age, that have shown to predict normal motor development with high accuracy (Einspieler et al., 2016). However, GMA is dependent on individual expert-based training and interpretations, requires time for video observation and analysis, and triggers a high demand for skilled observers if implemented in large-scale screening (Støen et al., 2017). As an evolving alternative to observational GMA, computer-based methods for objective and consistent risk-assessment are explored (Adde et al., 2010). This supports clinicians in diagnostics, ultimately identifying infants in need for early interventions and focused follow-up care.

Computer-based assessment of infant movements aggregates quantitative movement information from video recordings to yield estimates for the risk of later disorders, like CP (Ihlen et al., 2020). Hence, higher level of correctness in the representation of movement kinematics, such as segment positions and joint angles, facilitates optimal risk analysis. Fidgety movements are small movements of moderate speed and variable acceleration, of neck, trunk, and limbs, in all directions (Einspieler et al., 2004). Automated assessment of such movements requires precise localization of the body parts for proper computer-based risk analysis.

The widespread use of conventional video recordings to capture infant movements has established the need for markerless motion capture, which enables the extraction of movement information in an unobtrusive manner (Rahmati et al., 2015). This provides a low-cost alternative to sensor-based motion capture, which can be performed both at the

* Corresponding author.
E-mail address: espen.ihlen@ntnu.no (E.A.F. Ihlen).

https://doi.org/10.1016/j.compmedimag.2021.102012
Received 17 November 2020; Received in revised form 17 October 2021; Accepted 21 October 2021
Available online 26 November 2021

a) In-Motion Poses *b) Annotated skeleton model*

Fig. 1. a) A selection of video frames from In-Motion Poses, originating from standardized and less standardized hospital recordings (top and middle row, respectively), and videos captured from home by parents using the In-Motion smartphone application (Adde et al., 2021) (bottom row). Infant faces are blurred to ensure anonymity. b) The set of 19 body keypoints annotated in the images of In-Motion Poses.

clinic and at home (Adde et al., 2021). Markerless motion capture has the potential to make movement assessments more widely available and promotes worldwide collaboration in analysis of infant movements. Moreover, existing large-scale databases of infant recordings, collected by clinical GMA networks (Støen et al., 2019; Orlandi et al., 2018; Ferrari et al., 2019; Morgan et al., 2019; Kwong et al., 2019; Gima et al., 2019), can be exploited to yield more accurate computer-based methods for risk assessments.

Convolutional neural networks (ConvNets) have improved the techniques for extracting human movement information from conventional 2D videos (Toshev and Szegedy, 2014; Newell et al., 2016; Cao et al., 2019). State-of-the-art markerless motion capture tracks movements automatically through frame-by-frame pose estimation, where the ConvNets predict x and y coordinates of a predefined set of body keypoints, directly from the raw video frames (Andriluka et al., 2014). However, most existing human pose estimation (HPE) methods are targeted towards adults, which compared to infants, differ in anatomical proportions and distribution of body poses (Sciortino et al., 2017). Employed on infant images, the localization performance drops significantly, with 10% of the estimated body keypoint positions placed outside a head length distance from the annotated ground truth positions (i.e., 90% in the PCK_h@1.0 metric described in Section 2.3) (Sciortino et al., 2017). From this, Sciortino et al. (2017) conclude that there is a need to tune HPE ConvNets to the task of infant pose estimation.

Following along these lines, Chambers et al. (2020) retrain the openly available OpenPose network (Cao et al., 2019) by utilizing a dataset of 9039 manually annotated infant images. This improves infant pose estimation, reducing the mean error by 60% (Chambers et al., 2020). Despite this advance, a recent study carried out by our group found that OpenPose lacks the sufficient scaling of network depth, network width, and image resolution for optimal pose estimation (Groos et al., 2020b). Other alternatives to OpenPose, such as DeeperCut (Insafutdinov et al., 2016) used in DeepLabCut (Mathis et al., 2018), posses similar shortcomings as single-scale networks targeted towards multi-person pose estimation. Recent developments in HPE outperform OpenPose and variants by deploying novel multi-scale networks and by maintaining higher spatial resolution (Newell et al., 2016; Sun et al., 2019). OpenPose is also computationally inefficient, which makes it less convenient for real-world applications (Groos et al., 2020b). ConvNet model scaling addresses this challenge by providing trade-offs in localization performance and computational efficiency across various computational budgets (Groos et al., 2020b), better serving single-person applications.

The main objective of the present study is to obtain computationally efficient markerless pose estimation of the spontaneous movements of infants with a localization performance approaching that of human expert annotations. We exploit a large and heterogeneous infant pose dataset covering infant recordings from multiple sites across the world to conduct a comparative analysis of the localization performance and computational efficiency of eight different ConvNet models, including the commonly used OpenPose network. We compare the performance level of the ConvNets with the inter-rater spread of human expert annotations.

2. Materials and methods

In this section, we introduce In-Motion Poses, describe the ConvNet models included in the comparative study, and explain the various performance metrics used to evaluate the ConvNets.

2.1. In-Motion Poses

We developed a dataset comprising infant images with associated human annotations as the ground truth body keypoint positions. We used a large-scale database of 1424 recordings of 9–18 weeks post-term old infants to facilitate pose estimation of the spontaneous movements of infants in supine position across various recording setups. The videos were collected between 2001 and 2018 through different research projects on observational GMA, and all the recordings follow the standards for video-based GMA during the fidgety movement's period (i.e., infants wear a diaper or a onesie, are awake, alert, and content, are not disturbed or using pacifier, and are positioned in the center of a mattress or blanket with the whole body visible) (Einspieler and Prechtl, 2005). The resolution of videos varied from 576 × 720 to 1080 × 1920. The study was approved by the regional committee for medical and health research ethics in Norway, under reference numbers 2011/1811 and 2017/913 on 14 January 2019 and 9 October 2019, respectively. Written parental consent was obtained before inclusion.

From these recordings, we proposed a dataset of 20000 video frames. The dataset emphasizes the heterogeneity in spontaneous movements by including videos from 12 different sites from seven countries across the

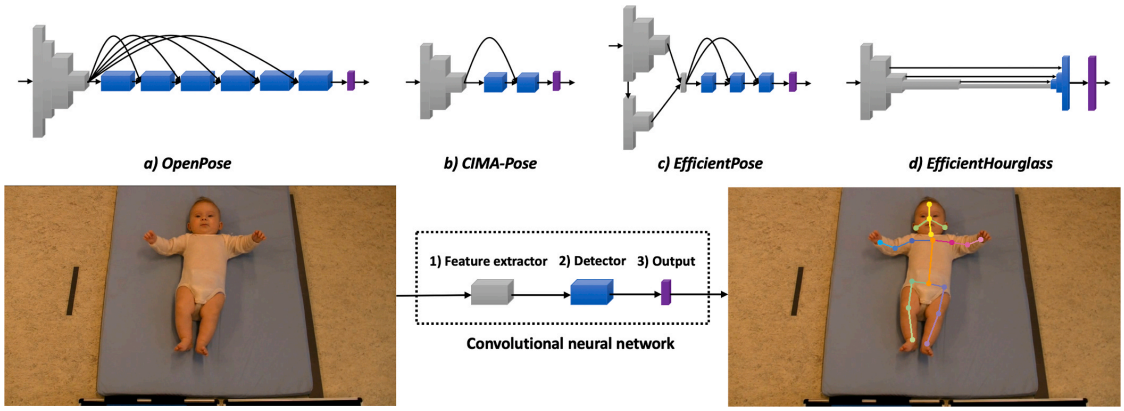


Fig. 2. ConvNets address infant pose estimation from video frames in a frame-by-frame manner by 1) extracting image features, 2) determining features relevant for detection, and 3) estimating infant keypoint positions. The height of the ConvNet blocks (i.e., feature extractor, detector, and output) indicates the block's spatial resolution in relation to the resolution of the input image.

globe (i.e., Norway, India, United States, Turkey, Belgium, Denmark, and Great Britain). The videos cover different groups of infants (e.g., typically developing infants, preterm infants, and other high-risk infants enrolled in hospital-based follow-up programs), and are recorded either by clinicians in a hospital setup or by parents using a smartphone application at home (Adde et al., 2021; Støen et al., 2019) (see Fig. 1a for examples from the dataset). To ensure all video variations were represented, 8000 (40%) frames originated from standardized hospital recordings, 8000 (40%) from home-based smartphone recordings, and the remaining 4000 (20%) from less standardized hospital videos. In each of these three subsets, 80% of the frames were randomly picked with an equal number of frames from each video. Moreover, to achieve proper variation of infant poses, the remaining 20% of frames cover infant poses that occur less frequently, and hence might be particularly challenging for an automatic pose estimator. These frames were manually selected from a random pool of 20000 separate frames (8000, 8000, and 4000 for each subset, respectively), with selection criteria including 1) legs moving towards upper body, 2) overlap of body parts, and 3) crossing of body parts. The resulting total of 20000 frames were split into training (14483 (72%)), validation (1493 (8%)), and test sets (4024 (20%)) in a common machine learning fashion. To mitigate bias and ensure objective evaluation, all frames of a single infant video were placed into one of these three sets.

For the ConvNet models to learn from the data in a supervised fashion, and to be able to validate and test the models, the infant images were annotated to produce the ground truth positions. As depicted by Fig. 1b, 19 distinct body keypoints (i.e., head top, nose, ears, upper neck, shoulders, elbows, wrists, upper chest, right/mid/left pelvis, knees, and ankles) comprised a skeleton model of the infant. The definitions of the body keypoints were agreed upon by a group of human movement scientists and clinical physiotherapists (see Appendix A for a complete overview). Using a separate software tool (Groos and Aurlien, 2018), 10 human expert annotators (two human movement scientists, two physiotherapists, and six engineers) estimated the x and y coordinates of body keypoints, through manual annotation. All body keypoints were annotated in all images regardless of their type of visibility (i.e., visible or occluded). This resulted in a total of 380000 human labels (i.e., 19 annotated keypoint positions for each of the 20000 frames). To measure the consistency between the experts, all annotators estimated the positions of body keypoints in the same sample of 100 randomly selected inter-rater frames. The frames were selected with a similar distribution across recording setups as the full dataset (i.e., 40% standardized, 40% home-based, and 20% less standardized). We computed the inter-rater annotation disagreement in terms of the mean inter-rater spread H of each body keypoint b. We calculated the mean distance of an annotation $(x_{b,i,j}, y_{b,i,j})$ of an individual expert j of a body keypoint's position in image i, to the average annotation $(\bar{x}_{b,i}, \bar{y}_{b,i})$, across the N (i.e., 10) experts for the S (i.e., 100) frames (see 1). H was normalized according to the head length of the infant in the image, defined as the distance from the top of the head to the upper neck (l_i).

$$H_b = \frac{1}{N \cdot S} \sum_{i=1}^{S} \sum_{j=1}^{N} \frac{\sqrt{(x_{b,i,j} - \bar{x}_{b,i})^2 + (y_{b,i,j} - \bar{y}_{b,i})^2}}{l_i} \quad (1)$$

2.2. Comparative analysis

By the use of the aforementioned dataset, we trained and evaluated a selection of ConvNet models for the task of infant pose estimation. First, the ConvNet of the state-of-the-art method for infant pose estimation, the OpenPose network (Cao et al., 2019; OpenPose, 2021) (see Fig. 2a for an architectural overview), was trained to yield baseline performance on In-Motion Poses, while also evaluating the official OpenPose library without fine-tuning[1] (OpenPose, 2021). Unless otherwise specified, OpenPose refers to OpenPose ConvNet fine-tuned on In-Motion Poses. Second, we trained a more computationally efficient approach inspired by OpenPose, named CIMA-Pose (see Fig. 2b), which has displayed promising results on infant pose estimation on videos from standardized clinical setups (Groos and Aurlien, 2018). CIMA-Pose comprises a ConvNet with low complexity, reflected by 2.4 million parameters compared to 26 million for OpenPose. OpenPose and CIMA-Pose operate on similar image input resolutions of 368×368 pixels[2]. Third, EfficientPose (Fig. 2c) comprises a family of scalable ConvNets demonstrating 57% improvement in high-precision pose estimation compared to OpenPose, despite significant reduction in computational cost (i.e., FLOPs) and number of parameters (Groos et al., 2020b). EfficientPose yields five model variants, EfficientPose RT and I-IV, obtained by the use of compound model scaling on input resolution, network width, and network depth. The computational requirements of

[1] The raw images in In-Motion Poses were downsampled and zero padded to square aspect ratio to achieve the input resolution of the ConvNets.

[2] The latest version of OpenPose (v1.7.0) was used with default settings maintained. Evaluation on In-Motion Poses was performed on the keypoints in the 25-keypoint body model that exist in In-Motion Poses (i.e., all keypoints except head top and upper neck).

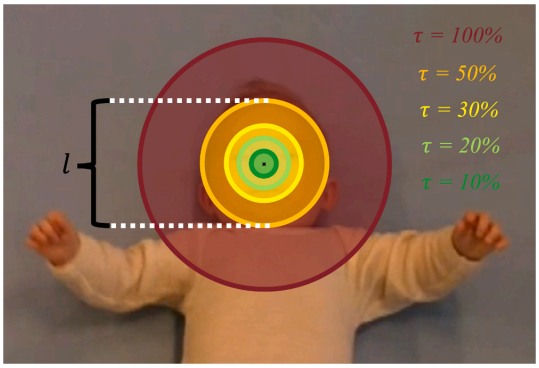


Fig. 3. $PCK_h@\tau$, the percentage of predictions within τl distance from the ground truth location (e.g., nose), is computed across five different thresholds τ (i.e., 100%, 50%, 30%, 20%, and 10%), evaluating the localization performance of a model, from coarse to fine.

EfficientPose span from less than one GFLOP to 74 GFLOPs, which is substantially less than the 161 GFLOPs of OpenPose. Fourth and finally, we optimized an EfficientHourglass model with EfficientNet-B4 backbone (i.e., EfficientHourglass B4) (Groos et al., 2020a), displayed in Fig. 2d. Inspired by the original multi-scale hourglass of Newell et al. (2016), EfficientHourglass performs parallel processing of image features at different scales, while conserving the level of detail (i.e., resolution) inherent in the input image. With an input resolution of 608 × 608, EfficientHourglass B4 maintains a resolution of at least 152 × 152 pixels throughout the stages of the network (i.e., feature extractor, detector, and output), compared to the consistent low resolution of 46 × 46 pixels in the detector and output of the single-scale OpenPose architecture (Cao et al., 2019; Groos et al., 2020a). For further details of the different ConvNets, the reader is referred to their original papers (Cao et al., 2019; Groos et al., 2020a, 2020b; Groos and Aurlien, 2018).

In the experiments, all models (except the underlying model of the official OpenPose library) were trained using a standardized optimization procedure. Pretraining on the general-purpose MPII HPE dataset (Andriluka et al., 2014) was performed, followed by fine-tuning on the training set of In-Motion Poses using the Adam optimizer for 100 epochs with a learning rate of 0.001. We applied data augmentation with random horizontal flipping, scaling (0.75–1.25), and rotation (+/−45 degrees). The optimization procedure was obtained through tuning of models on the validation set of In-Motion Poses.

2.3. Evaluation protocol and performance metrics

To evaluate the localization performance of the models included in the comparative analysis, positions of body keypoints were predicted on the separate test set of In-Motion Poses, comprising 4024 images. The retrained OpenPose, CIMA-Pose, EfficientPose, and EfficientHourglass were evaluated using the model outputs upscaled to input resolution with bilinear interpolation (e.g., three transposed convolutions, each with a stride of 2 and 4 × 4 kernel, performed 8 × upscaling in OpenPose, to increase the spatial resolution of outputs from 46 × 46 to 368 × 368), omitting the expensive multi-scale testing and flipping procedure commonly used for benchmarking HPE (Tang et al., 2018; Yang et al., 2017), whereas default post-processing was employed with the official version of OpenPose. Model localization performance was determined by comparing the model outputs to human annotations. The performance metrics included percentage of correct keypoints according to head size ($PCK_h@\tau$), normalized mean error (*ME*), and a proposed metric; percentage of correct keypoints according to human-level performance ($PCK_h@Human^{0.95}$). $PCK_h@\tau$ computes the fraction of keypoints within τl_i distance from the annotated position, where l_i is the infant head length of image i. To account for both model robustness and performance in high-precision pose estimation, we calculated measures of $PCK_h@\tau$ across various percentages τ of the head size (see Fig. 3). Coarse evaluation was performed with $PCK_h@1.0$, $PCK_h@0.5$, and $PCK_h@0.3$, and fine-grained evaluation by $PCK_h@0.2$ and $PCK_h@0.1$. Moreover, the *ME* measure reflects the average localization performance of model m on body part b in terms of the mean distance of a model's predictions to the ground truth locations:

$$ME_{m,b} = \frac{1}{S}\sum_{i=1}^{S} d_{m,b,i} \tag{2}$$

where $d_{m,b,i} = \frac{\sqrt{(x_{m,b,i}-\widehat{x}_{b,i})^2+(y_{m,b,i}-\widehat{y}_{b,i})^2}}{l_i}$ is the Euclidean distance from the estimated keypoint position $(x_{m,b,i}, y_{m,b,i})$ of model m to the human annotation $(\widehat{x}_{b,i}, \widehat{y}_{b,i})$, for keypoint b in image i of the test set. *ME* was normalized with respect to the head length l_i. To compare model performance against human-level performance, we introduce a metric, called $PCK_h@Human^{0.95}$. $PCK_h@Human^{0.95}$ defines the percentage of model predictions within the 95th percentile of the inter-rater spread of human experts:

$$PCK_h@Human^{0.95}{}_{m,b} = \frac{1}{S}\sum_{i=1}^{S} \delta(d_{m,b,i}) \tag{3}$$

$$\delta(d_{m,b,i}) = \left\{ \begin{array}{ll} 1, & \text{if } d_{m,b,i} \le H_b{}^{0.95} \\ 0, & \text{otherwise} \end{array} \right\} \tag{4}$$

Here, δ is a binary step function with threshold $H_b{}^{0.95}$ defining the 95th percentile of the inter-rater spread (where the mean inter-rater spread H_b is specified in Equation 1). In other words, $PCK_h@Human^{0.95}$ is equivalent to $PCK_h@\tau$ when $H_b{}^{0.95} = \tau$. Thus, $PCK_h@Human^{0.95} = 95\%$ reflects human-level performance. By utilizing the intraclass correlation coefficient (*ICC*) proposed by Fisher (1992), we also compared consistency (i.e., *ICC*(*C*, 1)) and agreement (i.e., *ICC*(*A*, 1)) between model localization error and inter-rater spread across body parts. The *ICC* values, and associated 95% confidence intervals, between the model *ME* and the inter-rater spread *H* of the human experts were calculated using a two-way model. Perfect agreement and consistency with inter-rater spread across body keypoints (i.e., *ICC*(*A*, 1) = *ICC*(*C*, 1) = 1) will suggest that a model displays human-level performance.

In addition to model localization performance, we evaluated the computational efficiency of the ConvNet models. We provide measures for model complexity (number of parameters), computational cost (FLOPs), and inference time (latency). The inference latency per image was estimated from model predictions on an NVIDIA GTX 1080 Ti GPU with TensorFlow 2.5, CUDA 11.0, and CUDNN 8.1. We used a batch size of 128 and computed the median latency in milliseconds over 10 computational runs.

2.4. Sample efficiency

To assess the amount of training data required for ConvNets to converge on the task of infant pose estimation, we carried out experiments with variation in the number of images in the training set, across a range of samples from no fine-tuning[3] to 100 images to the full training set of 14483 infant frames. To evaluate differences in sample efficiency between different ConvNet architectures, experiments were carried out for the most accurate ConvNet in each of the four model families. All experiments were performed over 100 epochs of training, and model

[3] When models were evaluated without fine-tuning, predictions were made only on the subset of 16 body keypoints that were available both in the MPII dataset and In-Motion Poses.

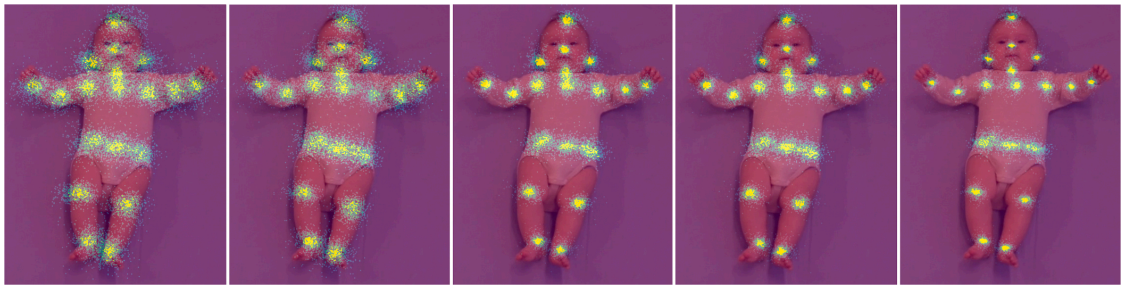
Table 1
The performance of the different ConvNets, pretrained on MPII (Andriluka et al., 2014) and fine-tuned on In-Motion Poses, as well as the official OpenPose library (OpenPose, 2021), in terms of localization performance on the test set of In-Motion Poses, and computational efficiency of the ConvNets from run-time experiments on an NVIDIA GTX 1080 Ti GPU.

		Localization performance						Computational efficiency		
Model	Resolution	@1.0*	@0.5*	@0.3*	@0.2*	@0.1*	ME	Parameters	FLOPs	Latency
OpenPose library	-	96.99%	95.51%	90.90%	81.49%	49.66%	0.1432**	-	-	62.33*** ms
OpenPose	368 × 368	99.94%	99.61%	97.65%	90.40%	54.89%	0.1087	26,011,743	161,077,013,640	35.21 ms
CIMA-Pose	368 × 368	99.98%	99.83%	98.74%	93.09%	59.69%	0.0988	2,380,495	15,645,092,494	11.49 ms
EfficientPose RT	224 × 224	99.96%	99.69%	98.15%	92.15%	58.71%	0.1022	481,336	955,490,248	5.06 ms
EfficientPose I	256 × 256	99.98%	99.83%	98.81%	93.68%	60.78%	0.0974	743,476	1,785,432,722	7.05 ms
EfficientPose II	368 × 368	99.97%	99.84%	98.54%	92.41%	62.25%	0.0969	1,759,372	7,944,292,598	19.38 ms
EfficientPose III	480 × 480	99.99%	99.94%	99.54%	97.57%	78.21%	0.0732	3,258,888	23,777,830,318	41.92 ms
EfficientPose IV	600 × 600	99.98%	99.93%	99.45%	96.77%	71.10%	0.0834	6,595,430	73,621,311,041	96.48 ms
EfficientHourglass B4	608 × 608	99.99%	99.95%	99.56%	97.67%	81.11%	0.0681	18,699,936	27,009,544,472	47.01 ms

* PCK_h@1.0, PCK_h@0.5, PCK_h@0.3, PCK_h@0.2, and PCK_h@0.1 are abbreviated as @1.0, @0.5, @0.3, @0.2, and @0.1, respectively.
** Keypoints in certain images, where the OpenPose library lack predictions due to not being confident, are excluded in computation of *ME*.
*** Latency estimate of the OpenPose library includes time required to pre-process images and perform default post-processing of ConvNet predictions.

Table 2
The localization performance of OpenPose, CIMA-Pose, EfficientPose III, and EfficientHourglass B4, all pretrained on MPII (Andriluka et al., 2014) and fine-tuned on In-Motion Poses, on the test set of In-Motion Poses, in relation to human-level performance (i.e., inter-rater spread *H*) across body parts *b*, as evaluated by the proposed PCK_h@*Human*$^{0.95}$ metric.

			PCK_h@Human$^{0.95}$			
b	H_b	$H_b^{0.95}$	OpenPose	CIMA-Pose	EfficientPose III	EfficientHourglass B4
Head top	0.0554	0.1158	60.39%	57.60%	81.59%	89.31%
Nose	0.0301	0.0574	32.03%	42.89%	74.48%	82.41%
Right ear	0.0603	0.1906	88.57%	92.40%	94.41%	92.00%
Left ear	0.0502	0.1364	73.31%	77.49%	88.54%	89.04%
Upper neck	0.0527	0.1212	80.67%	83.23%	88.77%	89.19%
Right shoulder	0.0531	0.1106	62.97%	73.14%	85.71%	86.63%
Right elbow	0.0429	0.0956	52.81%	71.00%	81.71%	86.73%
Right wrist	0.0386	0.0851	45.43%	60.93%	80.14%	82.60%
Upper chest	0.0643	0.1200	69.38%	72.44%	77.31%	79.42%
Left shoulder	0.0576	0.1204	63.25%	60.71%	88.07%	88.74%
Left elbow	0.0418	0.0959	48.19%	46.92%	82.50%	85.69%
Left wrist	0.0388	0.0901	48.83%	52.44%	79.08%	84.74%
Mid pelvis	0.0781	0.1587	82.75%	82.50%	86.43%	90.01%
Right pelvis	0.0812	0.1553	78.31%	80.89%	87.30%	88.72%
Right knee	0.0549	0.1119	66.58%	77.24%	86.63%	89.02%
Right ankle	0.0417	0.0902	51.07%	60.21%	75.47%	80.79%
Left pelvis	0.0828	0.1603	79.25%	77.53%	88.07%	90.31%
Left knee	0.0489	0.1049	49.06%	48.29%	88.22%	89.71%
Left ankle	0.0408	0.0861	45.75%	47.24%	75.70%	82.38%
All body parts	*0.0534*	*0.1161*	*62.03%*	*66.58%*	*81.59%*	*86.71%*



a) OpenPose ***b) CIMA-Pose*** ***c) EfficientPose III*** ***d) EfficientHourglass B4*** ***e) Human annotation***

Fig. 4. From left: a-d) The distribution of model prediction errors of the different ConvNets on 1000 randomly sampled frames (according to the distribution of standardized hospital recordings, home-based smartphone recordings, and less standardized hospital recordings) from the test set of In-Motion Poses across body parts, and e) the distribution of the inter-rater spread of the 10 human experts across 100 inter-rater frames (i.e., a total of 1000 annotations). The prediction errors are normalized according to the head size of the infant in the sample image.

Table 3
Absolute agreement and consistency (i.e., *ICC*(*A*, 1) and *ICC*(*C*, 1)) of ConvNets in relation to human expert inter-rater spread across body parts, with 95% confidence intervals in brackets.

	OpenPose	CIMA-Pose	EfficientPose III	EfficientHourglass B4
ICC(*A*, 1)	0.00 [−0.03, 0.07]	0.08 [−0.04, 0.32]	0.47 [−0.03, 0.84]	0.64 [−0.03, 0.91]
ICC(*C*, 1)	0.02 [−0.43, 0.46]	0.45 [0.01, 0.75]	0.94 [0.85, 0.98]	0.96 [0.91, 0.99]

performance in *ME*, PCK_h@0.5, and PCK_h@0.1 were calculated on the test set of In-Motion Poses. The smaller training samples were constructed by randomly selecting a subset of frames from the original training set, while maintaining the distribution of videos proposed in Section 2.1. Hence, the smaller samples and the full training set have similar variation in recording setups.

3. Results

Table 1 gives an overview of the performances of the eight different ConvNets, as well as the official version of OpenPose, on In-Motion Poses. In terms of localization performance, a 6–37% decrease in *ME* compared to the OpenPose baseline is achieved. This is supported by a higher robustness (i.e., gains in PCK_h@1.0, PCK_h@0.5, and PCK_h@0.3). In high-precision pose estimation, PCK_h@0.1 from 58.71% to 81.11% can be observed, compared to 54.89% and 49.66% for fine-tuned OpenPose and official OpenPose, respectively. With regards to computational efficiency, all models are smaller, with 1.4–54 times fewer parameters, and require less computation than OpenPose, i.e., 2.2–169 times less FLOPs. Moreover, the most computationally efficient ConvNet, EfficientPose RT, achieved run-time performance of 198 frames per second.

Table 2 displays the localization performance of the top-performing ConvNet of each model family. The most accurate model, EfficientHourglass B4, achieved an *ME* of 0.0681 compared to the average human inter-rater spread *H* of 0.0534. This equals an average percentage of human-level performance (i.e., $PCK_h@Human^{0.95}$) of 86.71%, compared to 62.03% for OpenPose. Fig. 4 shows a close resemblance between the spread of the human annotations and the estimates of EfficientPose III and EfficientHourglass B4 across body keypoints. This resemblance was supported by a significant consistency, *ICC*(*C*, 1), and high agreement, *ICC*(*A*, 1), between the spread of human expert annotations and the mean error of EfficientPose III and EfficientHourglass B4 (see Table 3). The lower *ICC*(*A*, 1) compared to *ICC*(*C*, 1) reflects a slightly higher *ME* for the ConvNet models compared to the inter-rater spread *H* of the human experts. A similar resemblance with human annotations was not achieved with OpenPose.

Fig. 5 illustrates that fine-tuning significantly improves localization performance of infant pose estimation compared to no fine-tuning (i.e., W/O). Moreover, all ConvNets benefit from increased training set size, especially in terms of the PCK_h@0.1 measure (Fig. 5c). However, whereas localization performance of OpenPose and CIMA-Pose saturates at sample sizes beyond 5000 images, EfficientPose III and EfficientHourglass B4 benefit from larger training sets. There is also a tendency that EfficientPose III and EfficientHourglass are more stable across dataset sizes, with a smaller difference in localization performance from 100 to 14483 images, compared to OpenPose and CIMA-Pose.

In Fig. 6, the localization performance of EfficientHourglass B4 is assessed qualitatively by providing model predictions on a selection of challenging images (i.e., less frequently occurring infant poses as described in Section 2.1) in the test set of In-Motion Poses.

4. Discussion

The main objective of the study was to obtain computationally efficient markerless infant pose estimation with a level of localization performance approaching that of human expert annotations. A comparative analysis has showed that performance levels comparable to

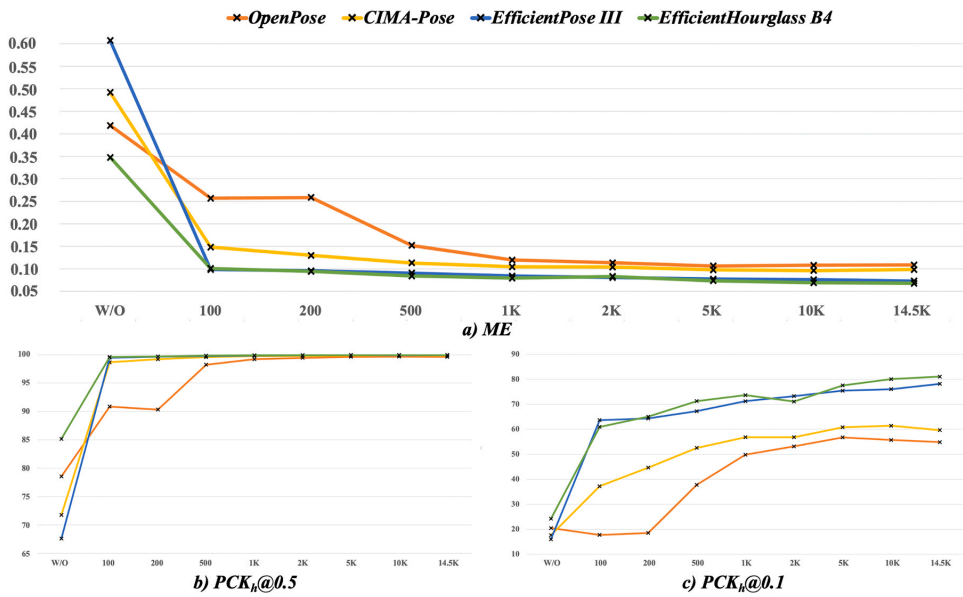


Fig. 5. Localization performance of OpenPose, CIMA-Pose, EfficientPose III, and EfficientHourglass B4, all pretrained on MPII (Andriluka et al., 2014), without fine-tuning (i.e., W/O) and with increasing amounts of data (from 100 to 14483 images) for fine-tuning on In-Motion Poses.

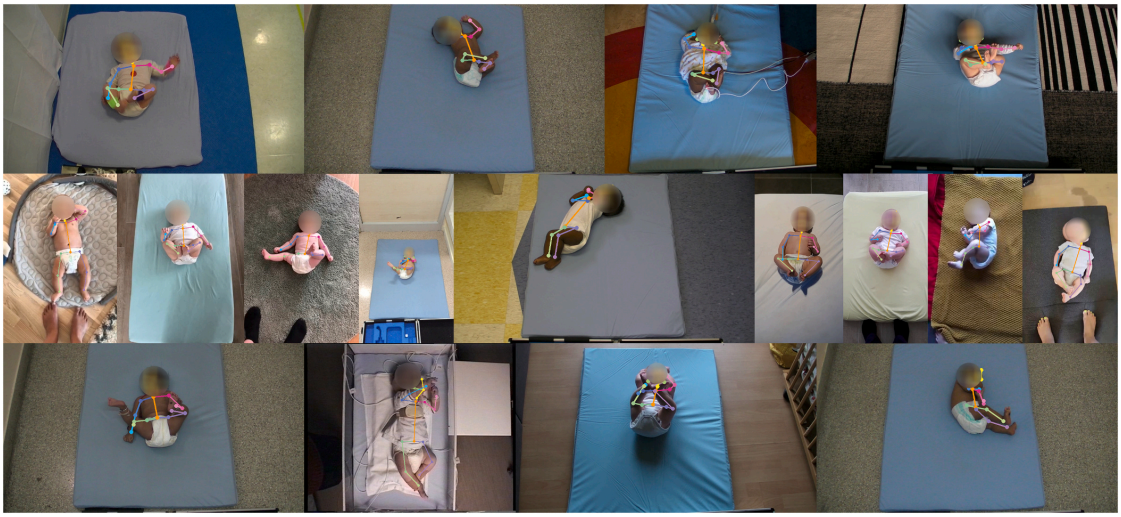


Fig. 6. Predictions of EfficientHourglass B4 on rare but normal infant poses in the test set of In-Motion Poses. The first and second row contain images where the model correctly predicted the position of body keypoints. The third row indicates cases where the model missed certain body keypoints (images from left to right: 1) right ankle, 2) head top and nose, 3) right elbow and right wrist, and 4) right wrist and left wrist). Infant faces are blurred to ensure anonymity.

human expert performance can be achieved, by utilizing contemporary ConvNets for HPE together with an extensive infant video database. This is reflected by $PCK_h@Human^{0.95}$ of the top-performing ConvNets approaching human-level performance, whereas the commonly applied OpenPose network does not reach similar level of localization performance.

4.1. Improving localization performance

The large improvement in localization performance compared to the state-of-the-art method OpenPose (Cao et al., 2019) is due to two main reasons. First, the hypothesis of Sciortino et al. (2017), that HPE ConvNets require fine-tuning on a selection of infant images to perform well on pose estimation of infants, is confirmed. The introduction of a large-scale infant pose dataset, In-Motion Poses, has improved the localization performance of OpenPose from 78.56% to 99.61% on PCK_h@0.5, as illustrated by Fig. 5b. Taking into account the error taxonomy of Ruggero Ronchi and Perona (2017), this indicates that the coarse localization errors, like the frequency of inversions (i.e., the predictions that appear at an incorrect body keypoint, such as misinterpretation of the left and right wrist) and misses (i.e., the erroneous localizations that are made without interfering with other keypoints), have been reduced. Despite the increased robustness with regards to coarse prediction errors, the optimal level of localization performance has not been reached. Further improvement of the ConvNets may be achieved by more systematically studying the cases where the models fall short, for example with substantial occlusion of body parts or specific body postures. Fig. 6 indicates that such scenarios exist. Accordingly, we could extend the existing dataset with images that target these situations to further improve model robustness through retraining. In a future perspective, it would also be valuable to assess if we could take into account the temporal information of a video to reduce prediction errors due to occlusion or rare body postures. Pose tracking that extends beyond frame-by-frame pose estimation may achieve this, but current progress in the field is restricted to processing a single pair of video frames with limited gap in time (Bertasius et al., 2019), which may not address cases of prolonged occlusion.

Second, the large improvement in PCK_h@0.2, PCK_h@0.1, and $PCK_h@Human^{0.95}$ of CIMA-Pose, EfficientPose, and EfficientHourglass B4, compared to OpenPose, is due to a reduction in fine prediction errors. EfficientPose III, EfficientPose IV,[4] and EfficientHourglass B4 reduce fine prediction errors better than OpenPose by operating on increased input and output resolutions. The consistent high resolution of EfficientHourglass B4 seems to maximize this benefit by displaying the highest values of PCK_h@0.1 and $PCK_h@Human^{0.95}$. However, the increase of resolution comes at the cost of reduced computational efficiency, in terms of increased number of FLOPs and decreased latency (see Table 1). Thus, alternative methods for post-processing of ConvNet predictions (e.g., soft-argmax (Levine et al., 2016)), or post-processing of the frame-by-frame position estimates over consecutive frames by low-pass filters, such as median filtering (Tukey, 1977), might reduce fine prediction errors more effectively. However, this demands that the video has a sufficient sample rate (e.g., 60 fps). Furthermore, fine prediction errors may also be minimized by decreasing the spread in annotated keypoint positions. As illustrated in Fig. 4, the distributions of prediction errors of EfficientPose III and EfficientHourglass B4 across body parts resemble the inter-rater spread of the human experts (e.g., higher variation in the placement of the keypoints of the pelvis, compared to the nose keypoint). This indicates that contemporary ConvNets for HPE, when supplied with sufficient amounts of training data (see Fig. 5 for the effect of sample size), are able to maximize the benefit of human annotations. Hence, a hypothesis for further studies is that more precisely annotated keypoints will further eliminate fine prediction errors, by model error being highly correlated with the inter-rater spread of human experts (see Table 3). Consequently, lower variation in the annotation of the keypoints of the pelvis may improve the ability of the ConvNets to localize these keypoints with high localization performance. More consistent annotations between human experts, reflected by lower inter-rater spread, may be obtained by proposing more precise definitions of the keypoint positions, than those in Appendix A. This could be particularly valuable for body keypoints

[4] EfficientPose IV displayed lower localization performance than EfficientPose III on In-Motion Poses, due to small batch size during training, which was necessary for the model to fit into GPU memory. As demonstrated by Tables 1 and 5, EfficientPose IV performed better than EfficientPose III in case of similar batch sizes.

that currently have higher inter-rater spread (e.g., for the keypoint of the upper chest). Human expert annotations may also be supplemented or replaced by other methods, such as marker-based solutions and 3D motion capture systems. These approaches may also yield performance improvements beyond fine prediction errors, by providing more precise annotations of occluded keypoints than can be achieved with 2D videos. We suggest that studies on infant pose estimation, and HPE in general (e.g., on challenges such as MPII (Andriluka et al., 2014)), judge localization performance against metrics related to human-level performance, such as $PCK_h@Human^{0.95}$, to evaluate the progress on these tasks in relation to human-level performance.

4.2. Improving computational efficiency

Our comparative analysis has shown that a large model size (i.e., number of parameters) is not necessary for high-precision infant pose estimation. On similar input resolution, both OpenPose and CIMA-Pose were outperformed by the more computationally efficient low-complexity EfficientPose II model on $PCK_h@0.1$ (see Table 1). Instead, it appears that high-precision infant pose estimation can be obtained with a relatively small number of parameters. This is demonstrated by EfficientPose III displaying only 5.12% decrease in $PCK_h@Human^{0.95}$, compared to EfficientHourglass B4, despite having 5.7 times fewer parameters. Combining this observation with the influence of high input and output resolution on localization performance, we would suggest further studies to investigate the effect of high resolution with low-complexity ConvNets. This could potentially narrow the current gap in localization performance between computationally efficient ConvNets, such as EfficientPose RT, and high-precision counterparts that are less computationally efficient, like EfficientPose III and EfficientHourglass B4. It would also be of particular interest to systematically study the optimal trade-off between localization performance and computational efficiency, by carefully assessing the localization performance of ConvNets of various complexities across different image resolutions. Our study suggests that ConvNets developed for HPE can be simplified when transferred to the infant pose estimation domain. HPE targets more complex circumstances and environments (e.g., images of multiple persons, a wide range of different activities, individuals of varying age, and substantial occlusion), whereas infant pose estimation is concerned with a single, clearly visible infant in supine position according to the guidelines of GMA (Einspieler et al., 2004; Andriluka et al., 2014). Potential paths for reducing network complexity could be 1) a decrease in network width (i.e., number of feature maps), and 2) less extensive use of multi-scale ConvNet architectures. The former may more appropriately address the little diversity in infant videos compared to the far-reaching HPE task, whereas the latter takes into account the small variation in an infant's distance to the camera and anatomical proportions. Nevertheless, from studying the inference latency of the ConvNets, we observed processing speeds from 10 to 198 fps (Table 1) on an NVIDIA GTX 1080 Ti consumer GPU. Further speedups of the pool of models studied in this paper may be obtained by implementing the ConvNets in low-level code like C++ or CUDA. Thus, a three-minute video of infant spontaneous movements could potentially be processed by a high-precision pose estimator in less than three minutes, which is feasible for clinical use. Moreover, the efficiency of the ConvNets can be further enhanced by utilizing techniques for compressing models with minimal loss of localization performance. Quantization-aware training, knowledge distillation, model pruning, and sparse kernels are paths that are worth to investigate (TensorFlow, 2020; Bucilua et al., 2006; Tung and Mori, 2018; Elsen et al., 2020). By obtaining accelerated and compressed ConvNets, the automatic pose estimation have the potential to be deployed locally at smartphones in the clinic and at home. Thus, infant pose estimation will be more easily applicable, while preserving patient privacy through decentralized processing of infant recordings on local devices.

4.3. External validity

In previous studies on ConvNet-based markerless infant pose estimation from 2D videos, investigations have been restricted to small or synthetic samples of infant videos (Hesse et al., 2018; Chambers et al., 2020). Hence, the external validity of such approaches is debatable, since ConvNets require large amounts of realistic images across various settings related to the task at hand to perform well on pose estimation. In this study, we have utilized a large-scale international database of GMA certified video recordings to train the ConvNets. Subsequently, we have validated the models on a separate set of 284 infant videos from a diverse range of hospital and home-based setups (see Fig. 1a). The high resistance to coarse prediction errors of the evaluated ConvNets suggests that infant pose estimation promotes flexibility in application in real-world scenarios. This encompasses various settings (e.g., clinic, research center, and home), across different countries, and without depending on specific camera equipment. When assessing the transfer validity of the ConvNets fine-tuned on In-Motion Poses on the synthetic dataset proposed by Hesse et al. (2018), only the best performing ConvNet on In-Motion Poses, EfficientHourglass B4, outperformed the official version of the state-of-the-art method OpenPose and displayed an acceptable transfer by maintaining a high level of localization performance (Table 6 and Fig. 8). This could suggest that the high-capacity multi-scale feature extractor of EfficientHourglass B4, through pre-training on MPII (Andriluka et al., 2014) and fine-tuning on In-Motion Poses, has learnt features that generalize beyond the natural infant images of In-Motion Poses. On the contrary, the feature extractors of OpenPose, CIMA-Pose, and EfficientPose are of lower relative capacity and contain fewer abstraction levels (i.e., scales) compared to EfficientHourglass B4 (Fig. 2). Hence, these fine-tuned ConvNets might lack the ability for appropriate transfer beyond recording setups of In-Motion Poses (e.g., plain backgrounds, and natural lighting and shading). However, the consistent localization performance of the official OpenPose library (OpenPose, 2021) (Tables 1 and 6) suggests that training on a sufficiently heterogeneous and large-scale human pose dataset, such as COCO (Lin et al., 2014) of 250000 human poses from various contexts, may combat the lack of high-capacity and multi-scale feature extraction to yield better generalizability. Similar effects could be achieved by combining In-Motion Poses with synthetic or natural infant pose datasets covering the variation in recording setups we want ConvNets to be tuned towards. Nevertheless, we should take into consideration the overall model capacity (i.e., number of parameters), which for CIMA-Pose and EfficientPose might not be sufficient to achieve appropriate transfer from In-Motion Poses to synthetic infants. We could therefore investigate ConvNet compound scaling on infant pose estimation, to determine the appropriate scaling factors of input resolution, network width, and network depth. Further studies should also more thoroughly assess the external validity of the trained ConvNets on real-life infant recordings, to verify that the high level of localization performance demonstrated by the present study indeed can be reproduced. This involves assessing the robustness in operating on video recordings from different recording setups with large variations in aspects, such as video quality, background environment, camera angle, and lighting conditions. The infant pose estimators could also be validated across groups of infants with different age, size, skin color, clothing, and postural variability within datasets like In-Motion Poses. Moreover, the degree of localization performance of the ConvNets in relation to state-of-the-art marker-based motion capture systems could also be assessed (Vicon, 2020; Qualisys, 2020). It is worth stressing that it is unrealistic to expect flawless pose estimation in recording situations highly dissimilar to the settings the models have been trained and evaluated in. However, the models can be retrained on other video databases when keypoint annotations are available. It is also worth investigating if the predefined set of body keypoints is sufficient for performing relevant assessments of characteristics of infant spontaneous movements identified in clinical GMA. However, for applications

emphasizing movement kinematics of other body keypoints (e.g., rotation of hands and feet, and relative movements of fingers or toes), the proposed infant pose estimation can be extended through retraining of ConvNets on different annotated sets of keypoints.

In summary, with improved ConvNet architectures and an extensive database of infant video recordings, body keypoint positions can be estimated with human-level performance. This will enable capturing more subtle infant movements and postures, and, consequently, improve early detection of risk-related infant movement kinematics (Ihlen et al., 2020; Einspieler et al., 2019). These improved ConvNets will also facilitate the assessments of infant movement kinematics which require a high level of detail, like fidgety movements or postural patterns in specific parts of the body, such as side-to-side head movements and atypical head centering (Einspieler et al., 2019).

5. Conclusions

The present study represents a significant progress towards clinically feasible markerless pose estimation of infant movements between two to five months of post-term age. This has been achieved by combining state-of-the-art ConvNets for human pose estimation with a novel heterogenous infant dataset. Highly precise detection of body keypoints enables accurate localization of segments and joints, which may facilitate computer-based assessment of characteristics of infant spontaneous movements related to risk of developmental disorders. With no dependency to body-worn markers, sensors or other expensive laboratory equipment, the automatic infant pose estimation can handle videos both captured by parents at home and by physicians at a hospital clinic. In conclusion, this technology has the potential to facilitate further research initiatives on infant movement analysis and motivate national and worldwide collaborations.

CRediT authorship contribution statement

Daniel Groos: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Lars Adde:** Conceptualization, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Ragnhild Støen:** Conceptualization, Investigation, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Heri Ramampiaro:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration. **Espen A.F. Ihlen:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was possible only due to the unified In-Motion research initiative on computer-based assessment of infant spontaneous movements and prediction of cerebral palsy, resulting in the multi-site database of infant recordings. The authors would like to acknowledge the following key personnel and institutions contributing in collecting video recordings: Norway; Toril Larsson Fjørtoft at St. Olavs University Hospital, Inger Elisabeth Silberg at Oslo University Hospital, Nils Thomas Songstad at University Hospital of North Norway, Angelique Tiarks at Levanger Hospital, Henriette Paulsen at Vestfold Hospital Trust, India; Niranjan Thomas at Christian Medical College Vellore, United States; Colleen Peyton at University of Chicago Comer Children's Hospital, Raye-Ann de Regnier and Lynn Boswell at Ann & Robert H Lurie Children's Hospital of Chicago, Turkey; Akmer Mutlu at Hacettepe University, Belgium; Aurelie Pascal at Ghent University, Denmark; Annemette Brown at Nordsjællands Hospital Hillerød, Great Britain; Anna Basu at Newcastle upon Tyne Hospitals. This work was supported by the Liaison Committee between the Central Norway Regional Health Authority and the Norwegian University of Science and Technology under project number 90056100, the Joint Research Committee between St. Olavs University Hospital and the Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, the DeepInMotion project funded by the Research Council of Norway with grant number 327146, and RSO Funds from the Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology under project number 81115200.

Appendix A. Keypoint definitions

The set of 19 body keypoints along with their definitions (see Fig. 7 and Table 4) were agreed upon by an expert group of human movement scientists and infant physiotherapists. The body keypoints were selected to cover most effectively the many degrees of freedom in the infant movements, while at the same time being properly defined to facilitate consistent annotation across humans.

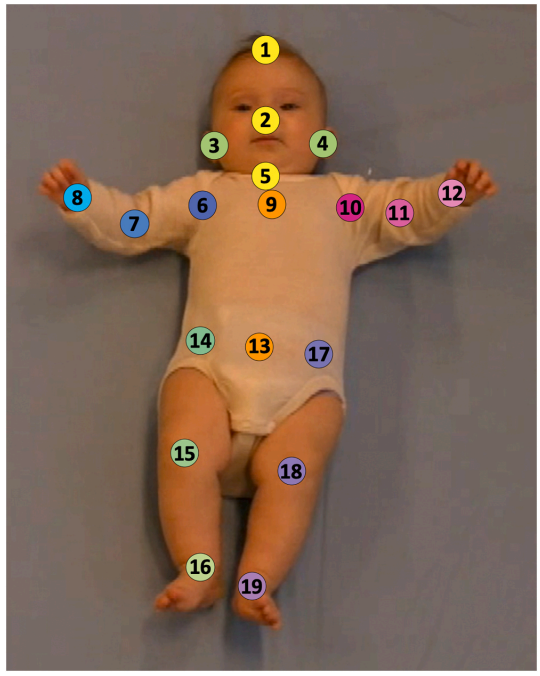


Fig. 7. The placements of the 19 different body keypoints on an infant.

Table 4
Definitions of body keypoints.

#	Body keypoint	Definition
1	Head top	Top of the forehead
2	Nose	Tip of the nose
3	Right ear	Center of the right ear
4	Left ear	Center of the left ear
5	Upper neck	Center of the larynx
6	Right shoulder	Center of the right shoulder joint
7	Right elbow	Center of the right elbow joint
8	Right wrist	Center of the right wrist joint
9	Upper chest	Midway between 6 and 10
10	Left shoulder	Center of the left shoulder joint
11	Left elbow	Center of the left elbow joint
12	Left wrist	Center of the left wrist joint
13	Mid pelvis	Midway between 14 and 17
14	Right pelvis	Right spina iliaca anterior superior
15	Right knee	Center of the right knee joint
16	Right ankle	Center of the right ankle joint
17	Left pelvis	Left spina iliaca anterior superior
18	Left knee	Center of the left knee joint
19	Left ankle	Center of the left ankle joint

Appendix B. Batch size inspection

We assessed the effect of fine-tuning the EfficientPose models on a reduced batch size of four images (i.e., the batch size of EfficientPose IV) to investigate possible performance degrade with EfficientPose IV due to inappropriate batch size. In comparison to Table 1, Table 5 displays performance degrade from training with reduced batch size, most evident in terms of high-precision localization, with 11.20–30.83% reduction in PCK_h@0.1.

Appendix C. Transfer validity

To investigate the transfer validity of the methods in our comparative analysis, we evaluated the localization performance of the models fine-tuned on In-Motion Poses, as well as the official OpenPose library, on the openly available MINI-RGBD dataset proposed by Hesse et al. (2018) (Table 6). The MINI-RGBD dataset comprises 12 synthetic infant video recordings of quite different nature than the recordings in In-Motion Poses. Localization performance, in terms of PCK_h@1.0, PCK_h@0.5, PCK_h@0.3, PCK_h@0.2, PCK_h@0.1, and *ME*, was measured on the subset of 12 body keypoints that are similar for MINI-RGBD and In-Motion Poses (i.e., nose, upper neck, shoulders, elbows, wrists, knees, and ankles). Since MINI-RGBD does not contain a keypoint for the top of the forehead, the head length of an infant was estimated as two times[5] the distance between the annotated keypoints of the nose and upper neck. This ensures that the evaluation metrics reflect a similar level of correctness as the metrics used with the evaluation on In-Motion Poses in Table 1.

Furthermore, for the most accurate ConvNet, namely EfficientHourglass B4, we conducted a qualitative experiment by estimating the locations of the 19 body keypoints in In-Motion Poses on a randomly selected frame in each of the 12 infant videos in the MINI-RGBD dataset (Fig. 8).

We also supply as Supplementary material frame-by-frame predictions of keypoint locations in a real, external infant recording for the best performing ConvNet in each model family, as well as by the use of the official version of OpenPose. The recording follows the standards for GMA (Einspieler and Prechtl, 2005), and has been recorded using the setup of the In-Motion App (Adde et al., 2021), which is similar to the home-based smartphone recordings in In-Motion Poses.

Table 5
The localization performance of EfficientPose RT and I-III on the test set of In-Motion Poses, when trained with the batch size of EfficientPose IV, followed by the performance difference in relation to the experiments in Table 1.

Model	PCK_h@ 1.0	PCK_h@ 0.5	PCK_h@ 0.3	PCK_h@ 0.2	PCK_h@ 0.1	ME
EfficientPose RT	99.80% (−0.16%)	99.32% (−0.37%)	92.93% (−5.22%)	72.50% (−19.65%)	27.88% (−30.83%)	0.1717 (0.0695)
EfficientPose I	99.94% (−0.04%)	99.66% (−0.17%)	97.22% (−1.59%)	85.42% (−8.26%)	38.38% (−22.40%)	0.1311 (0.0336)
EfficientPose II	99.98% (0.01%)	99.78% (−0.06%)	98.01% (−0.53%)	89.85% (−2.56%)	49.73% (−12.52%)	0.1137 (0.0168)
EfficientPose III	99.99% (0.00%)	99.94% (0.00%)	99.47% (−0.07%)	96.48% (−1.09%)	67.01% (−11.20%)	0.0884 (0.0152)

Table 6
The transfer validity of the different ConvNets, pretrained on MPII (Andriluka et al., 2014) and fine-tuned on In-Motion Poses, and the official OpenPose library (OpenPose, 2021), in terms of localization performance on the MINI-RGBD dataset (Hesse et al., 2018).

Model	PCK_h@1.0	PCK_h@0.5	PCK_h@0.3	PCK_h@0.2	PCK_h@0.1	ME
OpenPose library	98.35%	97.02%	94.47%	90.75%	73.80%	0.1030
OpenPose	88.59%	79.59%	71.77%	62.27%	38.41%	0.3926
CIMA-Pose	95.72%	88.99%	81.27%	71.83%	46.68%	0.2415
EfficientPose RT	94.98%	91.28%	86.91%	79.98%	53.83%	0.2135
EfficientPose I	93.13%	91.09%	88.16%	81.98%	56.19%	0.2772
EfficientPose II	92.49%	90.41%	87.41%	80.57%	54.60%	0.3263
EfficientPose III	83.79%	81.45%	79.60%	76.06%	58.56%	0.8559
EfficientPose IV	93.02%	91.15%	89.05%	86.14%	71.35%	0.2565
EfficientHourglass B4	99.81%	99.17%	97.52%	94.13%	75.86%	0.0845

[5] The head length of an infant (i.e., the distance from head top to upper neck) in In-Motion Poses was in average 1.98 times the distance from nose to upper neck.

Fig. 8. Predictions of keypoint locations of EfficientHourglass B4 in randomly selected frames from videos in the MINI-RGBD dataset (Hesse et al., 2018).

Appendix D. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.compmedimag.2021.102012.

References

Adde, L., Brown, A., Van den Broeck, C., De Coen, K., Horsberg Eriksen, B., Fjørtoft, T., Groos, D., Ihlen, E.A., Osland, S., Pascal, A., et al., 2021. The In-Motion-App for remote general movement assessment: a multi-site observational study. BMJ Open 11. https://doi.org/10.1136/bmjopen-2020-042147.

Adde, L., Helbostad, J.L., Jensenius, A.R., Taraldsen, G., Grunewaldt, K.H., Støen, R., 2010. Early prediction of cerebral palsy by computer-based video analysis of general movements: a feasibility study. Dev. Med. Child Neurol. 52, 773–778.

Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B., 2014. 2D human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, pp. 3686–3693.

Bertasius, G., Feichtenhofer, C., Tran, D., Shi, J., Torresani, L., 2019. Learning temporal pose estimation from sparsely-labeled videos. Adv. Neural Inf. Process. Syst. 3027–3038.

Bucilua, C., Caruana, R., Niculescu-Mizil, A., 2006. Model compression. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 535–541.

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A., 2019. OpenPose: realtime multi-person 2d pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell.

Chambers, C., Seethapathi, N., Saluja, R., Loeb, H., Pierce, S.R., Bogen, D.K., Prosser, L., Johnson, M.J., Kording, K.P., 2020. Computer vision to automatically assess infant neuromotor risk. IEEE Trans. Neural Syst. Rehabil. Eng. 28, 2431–2442.

Einspieler, C., Bos, A.F., Krieber-Tomantschger, M., Alvarado, E., Barbosa, V.M., Bertoncelli, N., Burger, M., Chorna, O., DelSecco, S., DeRegnier, R.A., et al., 2019. Cerebral palsy: early markers of clinical phenotype and functional outcome. J. Clin. Med. 8, 1616.

Einspieler, C., Kerr, A.M., Prechtl, H.F., 2005. Is the early development of girls with rett disorder really normal? Pediatr. Res. 57, 696–700.

Einspieler, C., Peharz, R., Marschik, P.B., 2016. Fidgety movements-tiny in appearance, but huge in impact. J. De. Pediatr. 92, S64–S70.
Einspieler, C., Prechtl, H.F., 2005. Prechtl's assessment of general movements: a diagnostic tool for the functional assessment of the young nervous system. Ment. Retard. Dev. Disabil. Res. Rev. 11, 61–67.
Einspieler, C., Prechtl, H.R., Bos, A., Ferrari, F., Cioni, G., 2004. Prechtl's Method on the Qualitative Assessment of General Movements in Preterm, Term and Young Infants. Mac Keith Press.
Einspieler, C., Sigafoos, J., Bartl-Pokorny, K.D., Landa, R., Marschik, P.B., Bölte, S., 2014. Highlighting the first 5 months of life: General movements in infants later diagnosed with autism spectrum disorder or rett syndrome. Res. Autism Spectr. Disord. 8, 286–291.
Elsen, E., Dukhan, M., Gale, T., Simonyan, K., 2020. Fast sparse ConvNets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14629–14638.
Ferrari, F., Plessi, C., Lucaccioni, L., Bertoncelli, N., Bedetti, L., Ori, L., Berardi, A., DellaCasa, E., Iughetti, L., D'Amico, R., 2019. Motor and postural patterns concomitant with general movements are associated with cerebral palsy at term and fidgety age in preterm infants. J. Clin. Med. 8, 1189.
Fisher, R.A., 1992. Statistical methods for research workers. In: Breakthroughs in Statistics. Springer, pp. 66–70.
Gima, H., Shimatani, K., Nakano, H., Watanabe, H., Taga, G., 2019. Evaluation of fidgety movements of infants based on gestalt perception reflects differences in limb movement trajectory curvature. Phys. Ther. 99, 701–710.
Groos, D., Adde, L., Ihlen, E.A., 2020a. Approaching human precision on automatic markerless tracking of human movements. Gait Posture 81, 117–118.
Groos, D., Aurlien, K., 2018. Infant body part tracking in videos using deep learning – facilitating early detection of cerebral palsy (Master's thesis), NTNU.
Groos, D., Ramampiaro, H., Ihlen, E.A., 2020b. Efficientpose: scalable single-person pose estimation. Appl. Intell. 51 (4), 2518–2533. https://doi.org/10.1007/s10489-020-01918-7.
Hesse, N., Bodensteiner, C., Arens, M., Hofmann, U.G., Weinberger, R., Sebastian Schroeder, A., 2018. Computer vision for medical infant motion analysis: state of the art and RGB-D data set. In: Proceedings of the European Conference on Computer Vision (ECCV).
Ihlen, E.A., Støen, R., Boswell, L., de Regnier, R.A., Fjørtoft, T., Gaebler-Spira, D., Labori, C., Loennecken, M.C., Msall, M.E., Möinichen, U.I., et al., 2020. Machine learning of infant spontaneous movements for the early prediction of cerebral palsy: a multi-site cohort study. J. Clin. Med. 9, 5.
Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B., 2016. DeeperCut: a deeper, stronger, and faster multi-person pose estimation model. In: European Conference on Computer Vision. Springer, pp. 34–50.
Kwong, A.K., Eeles, A.L., Olsen, J.E., Cheong, J.L., Doyle, L.W., Spittle, A.J., 2019. The baby moves smartphone app for general movements assessment: engagement amongst extremely preterm and term-born infants in a state-wide geographical study. J. Paediatr. Child Health 55, 548–554.
Levine, S., Finn, C., Darrell, T., Abbeel, P., 2016. End-to-end training of deep visuomotor policies. J. Mach. Learn. Res. 17, 1334–1373.
Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., 2014. Microsoft COCO: common objects in context. In: European Conference on Computer Vision. Springer, pp. 740–755.
Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., Bethge, M., 2018. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat. Neurosci. 21, 1281–1289.
Morgan, C., Romeo, D.M., Chorna, O., Novak, I., Galea, C., DelSecco, S., Guzzetta, A., 2019. The pooled diagnostic accuracy of neuroimaging, general movements, and neurological examination for diagnosing cerebral palsy early in high-risk infants: a case control study. J. Clin. Med. 8, 1879.
Newell, A., Yang, K., Deng, J., 2016. Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision. Springer, pp. 483–499.
Novak, I., Morgan, C., Adde, L., Blackman, J., Boyd, R.N., Brunstrom-Hernandez, J., Cioni, G., Damiano, D., Darrah, J., Eliasson, A.C., et al., 2017. Early, accurate diagnosis and early intervention in cerebral palsy: advances in diagnosis and treatment. JAMA Pediatr. 171, 897–907.
OpenPose, 2021. Real-time multi-person keypoint detection library for body, face, hands, and foot estimation. ⟨https://github.com/CMU-Perceptual-Computing-Lab/openpose⟩ (Accessed 30 May 2021).
Orlandi, S., Raghuram, K., Smith, C.R., Mansueto, D., Church, P., Shah, V., Luther, M., Chau, T., 2018. Detection of atypical and typical infant movements using computer-based video analysis. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp. 3598–3601.
Qualisys, 2020. Human biomechanics. ⟨https://www.qualisys.com/applications/human-biomechanics/⟩ (Accessed 17 July 2020).
Rahmati, H., Dragon, R., Aamo, O.M., Adde, L., Stavdahl, Ø., Van Gool, L., 2015. Weakly supervised motion segmentation with particle matching. Comput. Vis. Image Underst. 140, 30–42.
Ruggero Ronchi, M., Perona, P., 2017. Benchmarking and error diagnosis in multi-instance pose estimation. Proc. IEEE Int. Conf. Comput. Vis. 369–378.
Sciortino, G., Farinella, G.M., Battiato, S., Leo, M., Distante, C., 2017. On the estimation of children's poses. In: International Conference on Image Analysis and Processing. Springer, pp. 410–421.
Støen, R., Boswell, L., De Regnier, R.A., Fjørtoft, T., Gaebler-Spira, D., Ihlen, E.A., Labori, C., Loennecken, M., Msall, M., Möinichen, U.I., et al., 2019. The predictive accuracy of the general movement assessment for cerebral palsy: a prospective, observational study of high-risk infants in a clinical follow-up setting. J. Clin. Med. 8, 1790.
Støen, R., Songstad, N.T., Silberg, I.E., Fjørtoft, T., Jensenius, A.R., Adde, L., 2017. Computer-based video analysis identifies infants with absence of fidgety movements. Pediatr. Res. 82, 665–670.
Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5693–5703.
Tang, W., Yu, P., Wu, Y., 2018. Deeply learned compositional models for human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 190–206.
TensorFlow, 2020. Quantization aware training. https://www.tensorflow.org/model_optimization/guide/quantization/training (Accessed 17 April 2020).
Toshev, A., Szegedy, C., 2014. DeepPose: human pose estimation via deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1653–1660.
Tukey, J.W., 1977. Exploratory data analysis, 2, Reading, MA.
Tung, F., Mori, G., 2018. CLIP-Q: deep network compression learning by in-parallel pruning-quantization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7873–7882.
Vicon, 2020. The most trusted mocap ecosystem. https://www.vicon.com/applications/life-sciences/ (Accessed 17 July 2020).
Yang, W., Li, S., Ouyang, W., Li, H., Wang, X., 2017. Learning feature pyramids for human pose estimation. In: Proceedings of the IEEE international conference on computer vision, pp. 1281–1290.

Daniel Groos: Daniel Groos received the M.Sc. degree in Computer Science from the Norwegian University of Science and Technology (NTNU). He is currently PhD Research Fellow at the Department of Neuromedicine and Movement Science, NTNU, Trondheim, Norway. His research interests include deep learning-based computer vision, markerless human movement analysis, and medical technology.

Lars Adde: Lars Adde is Senior Researcher at the Department of Clinical and Molecular Medicine, NTNU, Trondheim, Norway and Physiotherapist at the Clinic of Clinical Services, Trondheim University Hospital, Trondheim, Norway. Adde holds a PhD degree in Clinical Medicine from NTNU, Master from NTNU, and Bachelor in Physiotherapy from Oslo University College (now Oslo Metropolitan University). He is Associate Director in six multisite studies predicting cerebral palsy and long-term motor outcomes by use of clinical and computerized assessment of spontaneous infant movements. His main research interest is movement analysis of young infants.

Ragnhild Støen: Ragnhild Støen is Professor at the Department of Clinical and Molecular Medicine, NTNU, Trondheim, Norway. Støen holds a PhD degree and Master M.D. from NTNU. She is Head of Neonatology at the Department of Neonatology, Trondheim University Hospital, Trondheim, Norway. Her particular interests are early neuroprotection after perinatal asphyxia, early detection of cerebral palsy in high-risk infants and ethical challenges related to infants born at the border of viability.

Heri Ramampiaro: Heri Ramampiaro is Professor at the Department of Computer Science, NTNU, Trondheim, Norway. He is Head of the Data and Artificial Intelligence (DART) research group and Head of the department. Ramampiaro holds a PhD degree in Computer Science from NTNU, M.Sc. from Stavanger University College (now University of Stavanger), and B.Eng. from Aalesund College (now NTNU). His current main research interests include machine learning, data/text mining, information retrieval, and NLP.

Espen A.F. Ihlen: Espen A. F. Ihlen is Associate Professor at the Department of Neuromedicine and Movement Science, NTNU, Trondheim, Norway. Ihlen holds a PhD degree in Clinical Medicine and M.Sc. in Human Movement Science from NTNU. His main interest is development of convolutional neural network architectures and machine learning for human movement analysis.

Paper III

Development and External Validation of Deep Learning-Based Early Prediction of Cerebral Palsy From Spontaneous Movements in High-Risk Infants

Authors

Daniel Groos, MS[1,†]
Lars Adde, PT, PhD[2,3,†]
Sindre Aubert, MS[4]
Lynn Boswell, PT, MS[5]
Raye-Ann deRegnier, MD[5,6]
Toril Fjørtoft, PT, PhD[2,3]
Deborah Gaebler-Spira, MD[6,7]
Andreas Haukeland, MS[4]
Marianne Loennecken, PT[8]
Michael Msall, MD[9,10]
Unn Inger Möinichen, PT, MS[8]
Aurelie Pascal, PT, PhD[11]
Colleen Peyton, PT, DPT[6,12]
Heri Ramampiaro, PhD[4]
Michael D. Schreiber, MD[12]
Inger Elisabeth Silberg, MD[8]
Nils Thomas Songstad, MD, PhD[13]
Niranjan Thomas, MD, PhD[14]
Christine Van den Broeck, PT, PhD[11]
Gunn Kristin Øberg, PT, PhD[8,15]
Espen A.F. Ihlen, PhD[1,‡]
Ragnhild Støen, MD, PhD[2,16,‡]

[1] Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, Trondheim, Norway
[2] Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway
[3] Clinic of Clinical Services, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway
[4] Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway
[5] Ann and Robert H Lurie Children's Hospital of Chicago, Chicago, Illinois, United States
[6] Northwestern University Feinberg School of Medicine, Chicago, Illinois, United States
[7] Shirley Ryan AbilityLab, Chicago, Illinois, United States
[8] Division of Paediatric and Adolescent Medicine, Oslo University Hospital, Oslo, Norway
[9] Section of Developmental and Behavioral Pediatrics, University of Chicago, Comer Children's Hospital, Chicago, Illinois, United States
[10] Kennedy Research Center on Neurodevelopmental Disabilities, University of Chicago, Comer Children's Hospital, Chicago, Illinois, United States
[11] Department of Rehabilitation Sciences and Physiotherapy, Ghent University, Ghent, Belgium
[12] Department of Pediatrics, University of Chicago, Comer Children's Hospital, Chicago, Illinois, United States
[13] Department of Pediatrics and Adolescent Medicine, University Hospital of North Norway, Tromsø, Norway
[14] Department of Neonatology, Christian Medical College Vellore, Vellore, Tamil Nadu, India
[15] Department of Health and Care Sciences, Faculty of Health Sciences, The Arctic University of Norway, Tromsø, Norway
[16] Department of Neonatology, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway

[†] Daniel Groos and Lars Adde contributed equally as co-first authors.
[‡] Espen A.F. Ihlen and Ragnhild Støen contributed equally as co-last authors.

Corresponding Author: Lars Adde, PT, PhD, Department of Molecular and Clinical Medicine, Norwegian University of Science and Technology, Olav Kyrres gate 11, 7030 Trondheim, Norway (lars.adde@ntnu.no, +4791897615).

Key Points

Question: What is the external validity of Deep Learning-based prediction of cerebral palsy from infant spontaneous movements at 3 months post-term age?

Findings: In this prognostic study of 557 infants, Deep Learning-based early prediction of cerebral palsy demonstrated sensitivity of 71%, specificity of 94%, positive predictive value of 68% and negative predictive value of 95%. Deep Learning-based cerebral palsy prognosis was associated with later functional level and subtype in children with cerebral palsy.

Meaning: Deep Learning-based assessments could support early detection of cerebral palsy in high-risk infants.

Abstract

Importance: Early identification of infants with cerebral palsy (CP) is essential for early intervention, yet existing clinical expert-based assessments do not enable widespread use, and Conventional Machine Learning alternatives lack validity on external samples.

Objective: To develop and assess the external validity of a novel Deep Learning-based prediction of CP from a video of spontaneous movements at 3 months corrected age (CA).

Design: Prognostic study on early prediction of CP in infants with increased risks of perinatal brain injury enrolled between 2001 and 2018 in previous studies.

Setting: 13 hospitals in United States, Norway, India, and Belgium.

Participants: 557 high-risk infants, with 418 (75.0%) randomized into sample for prognostic model development and 139 (25.0%) for external validation. We included all infants with a video from 7-18 weeks CA assessed with the General Movement Assessment (GMA), and who were evaluated for a diagnosis of CP after 12 months CA.

Main Outcomes and Measures: Deep Learning-based prediction of CP was performed automatically from a single video. The primary outcome was CP and associated functional level and subtype. We assessed sensitivity, specificity, positive and negative predictive values, and accuracy.

Results: Median CA at assessment was 12 (IQR: 11-13) weeks. Eighty-four (15.1%) infants were diagnosed with CP at mean 3.4 (SD: 1.7) years. On external validation, Deep Learning-based CP prediction displayed sensitivity of 71.4% (95% CI: 47.8%-88.7%) and specificity of 94.1% (95% CI: 88.2%-97.6%). Positive and negative predictive values were 68.2% (95% CI: 45.1%-86.1%) and 94.9% (95% CI: 89.2%-98.1%), respectively, and accuracy 90.6% (95% CI: 84.5%-94.9%). Corresponding sensitivity and specificity of GMA were 70.0% (95% CI: 45.7%-88.1%) and 88.7% (95% CI: 81.5%-93.8%), respectively. The automated prediction model had higher sensitivity in infants with non-ambulatory (100.0%; 95% CI: 63.1%-100.0%) compared to ambulatory CP (58.3%; 95% CI: 27.7%-84.8%; $P = .02$), and spastic bilateral (92.3%; 95% CI: 64.0%-99.8%) compared to spastic unilateral CP (42.9%; 95% CI: 9.9%-81.6%; $P < .001$).

Conclusions and Relevance: Deep Learning-based prediction of CP at 3 months CA provided predictive accuracy non-inferior to GMA on external validation. The study indicates possible avenues for utilizing Deep Learning-based software for objective, early detection of CP in clinical settings.

Introduction

Cerebral palsy (CP) is the most common physical disability in children, causing functional limitation and co-occurring impairments[1] (e.g., pain, musculoskeletal deformities, seizures, and communication and sleep disorders) due to injury to the developing brain[2]. CP is typically diagnosed between 12 and 24 months of age and associated classification of severity occurs even later in childhood[3,4]. Early identification of infants at high risk of CP is essential to provide targeted follow-up and interventions during infancy when neuroplasticity is high[5,6], improve access to community services for proactive management to minimize complications[7], and reassure parents of high-risk infants who are unlikely to develop CP[8].

The General Movement Assessment (GMA) is recommended as the most accurate clinical test for prognosis of CP in infants before 5 months[4,9], based on the absence of the fidgety (FMs) type of general movements (i.e., spontaneous movements involving the whole infant body)[10,11]. The GMA is based on observation of infants' general movements in video recordings by clinical experts. The method requires considerable training[12], and rater experience may influence GMA reliability[13]. These facts hamper widespread clinical use[14].

With advancements in the field of Artificial Intelligence, Machine Learning techniques have been developed as objective, low-cost alternatives to GMA[14-17]. Former Machine Learning techniques for tracking and classification of infant spontaneous movements generally aimed to predict CP by proposing restricted sets of manually selected movement features used in combination with conventional statistical methods (e.g., logistic regression and support vector machine)[18-22]. A recent study by our group demonstrated predictive values of such Conventional Machine Learning-based CP prediction, approaching the level of GMA[23]. Despite this progress, there are fundamental challenges yet to be addressed. The restricted set of manually selected movement features have an unknown relation to observational GMA, which questions the construct validity of Conventional Machine Learning techniques. External validation is consequently lacking due to small sample sizes and short duration of follow-up evaluations[17]. As a result, validation is performed using less conservative methods, including leave-one-out cross-validation, and by using absence of FMs as a surrogate outcome for CP[14,24].

A new field within Machine Learning, called Deep Learning, has enabled automatic detection of discriminative movement features through representation learning[25]. That means dynamically selecting features relevant for the task at hand without any human expert involvement. Accuracy of Deep Learning improves with increasing amounts of data (e.g., videos), and Deep Learning has capacity to detect features representing intricate relationships in data, like complex full-body general movements.

Our primary objective was to develop a Deep Learning-based early prediction of CP from infant spontaneous movements during the FMs period, and to perform external validation on a multicenter sample of high-risk infants. Our secondary objective was to compare the predictive accuracy with the clinically recommended GMA and Conventional Machine Learning for CP prediction, and to evaluate the ability of the Deep Learning-based prediction model to forecast functional level and CP subtype.

Methods

Participants

The sample comprised 557 high-risk infants prospectively enrolled between September 2001 and October 2018 in previously published studies from our group[22,26-28]. See eAppendix 1 for explanation of how these studies differ from the present study. Infants were included based on the following criteria: 1) available video recording following standards of Prechtl's GMA[29] during the FMs period at 9-18 (median: 12; IQR: 11-13) weeks corrected age (CA), 2) available GMA classifications of FMs, and 3) known CP status at 12 (median: 38; IQR: 23-46) months CA or older. Two infants with videos at 7 and 8 weeks CA were included for method development. Infants excluded due to missing video recording, GMA classification, or CP status are reported elsewhere[22,26-28]. The sample size was determined by the number of infants from previous studies with the aforementioned available data. See eAppendix 2 and eTable 1 in the Supplement for clinical characteristics of infants (including gestational age, birth weight, and infant sex). The present study was approved by the regional committee for medical and health research ethics (REC Central-Committee 4.2007.2327) in Norway and local Institutional Review Boards in United States, Belgium, and India. Written parental consent was obtained before inclusion.

Video of Infant Spontaneous Movements and Classification of General Movements

Infants were recorded in supine position during active wakefulness over 1-9 (median: 5) minutes, following GMA standards[29]. A conventional video camera (Sanyo VPC-HD2000 and Sony DCR-PC100E) at recording rate of 24-60 (median: 30) frames per seconds and video resolution of 576x720 to 1080x1920 (median: 720x1280) was used in a standardized setup comprising mattress and stationary overhead camera. If more than one recording was available, the one between 12 and 13 weeks CA was used.

GMA classifications from previous studies were used[22,26-28]. Two experienced observers (LA and TF) blinded to the medical history of the infants, classified the videos as normal (FMs sporadically, intermittently, or continuously present) or abnormal (absent FMs). Infants classified with exaggerated FMs, excessive in amplitude and speed, were a priori excluded from the analysis of GMA due to unpredictable outcomes of this category. In cases of disagreement between observers, videos were reassessed, and consensus was reached.

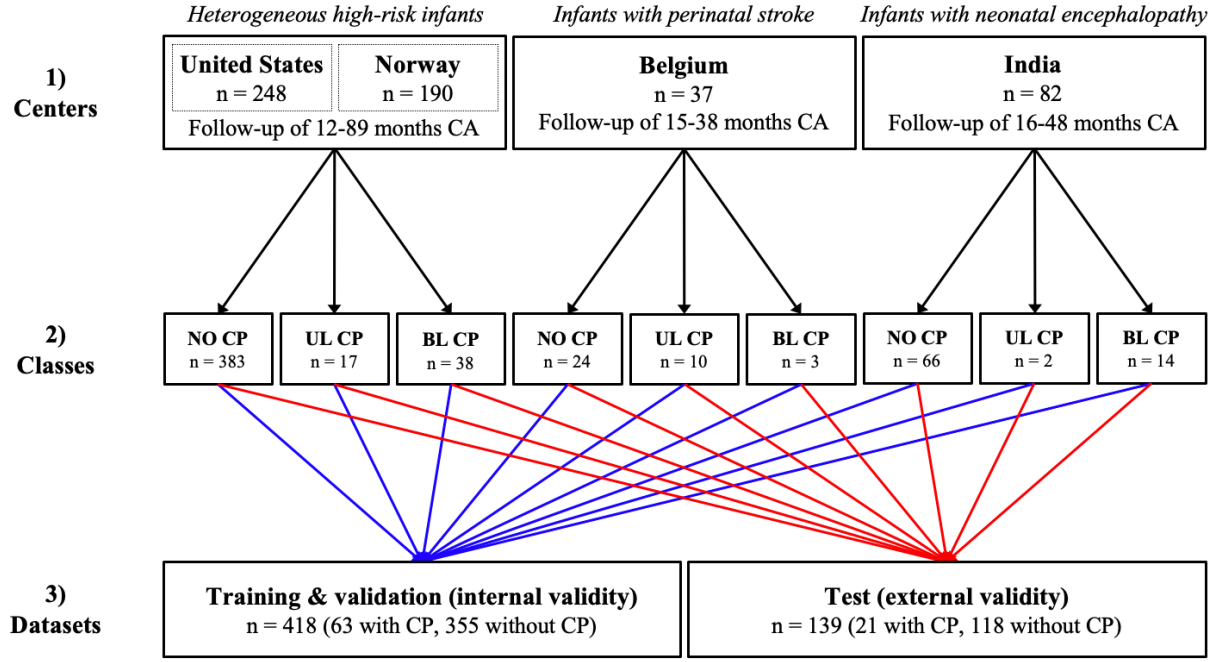
Cerebral Palsy Status, Subtype of Cerebral palsy, and Functional Level

The primary outcome of CP was diagnosed by a pediatrician following the decision tree of the Surveillance of Cerebral Palsy in Europe (SCPE)[30], which included the classification of CP subtypes into spastic unilateral, spastic bilateral, dyskinetic, and ataxic. Follow-up time differed between studies, ranging from 18 months to 5 years of age[22,26-28]. The Gross Motor Function Classification System (GMFCS)[3] was used to classify functional level into ambulatory CP (level I, II, and III) and non-ambulatory CP (level IV and V). Pediatricians responsible for CP diagnosis were blinded to GMA classifications.

Datasets for Method Development and External Validation

To achieve representative samples for method development, i.e., training and internal validation, and external validation, all high-risk infants were stratified into classes based on study and nationality (centers in step 1, Figure 1) and CP subtype (step 2, Figure 1). As shown in step 3 in Figure 1, 75.0% of infants of each class (blue path) were randomly assigned into dataset for method development (training and validation), and the remaining 25.0% (red path) into test set for external validation. The infants for method development were further divided into seven internal validation samples (i.e., folds), each comprising nine infants with CP and 50 or 51 infants without CP. This enabled 7-fold cross-validation for evaluating internal validity. The internal validation samples were constructed utilizing a similar procedure for stratification on center and CP subtype, as with the external test set in Figure 1.

Figure 1. Datasets for development and external validation

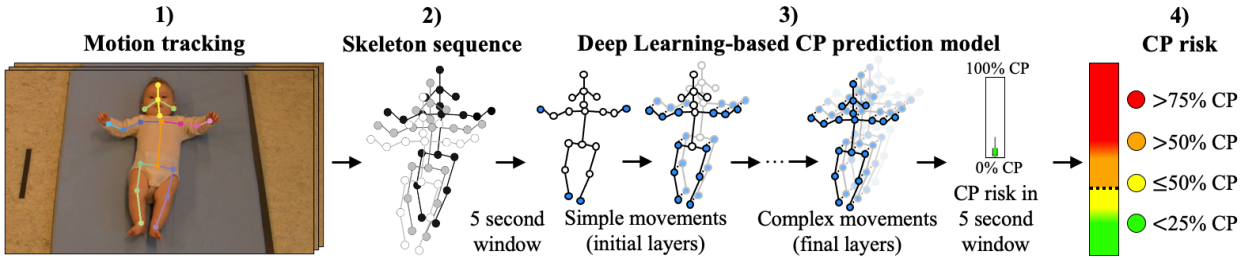


Infants diagnosed with CP where subtype was not available were placed into class for unilateral CP (UL CP) in case of GMFCS level I or II, and into bilateral CP (BL CP) for GMFCS level III, IV, or V. Infants with dyskinetic CP and ataxic CP were placed into BL CP.
Abbreviations: NO CP, without CP; UL CP, spastic unilateral CP; BL CP, spastic bilateral CP; CA, corrected age.

Deep Learning-Based CP Prediction

The overall concept of the Deep Learning-based CP prediction method is presented in Figure 2.

Figure 2. The steps involved in Deep Learning-based CP prediction



1) Motion Tracking

The raw infant video was processed by a motion tracker[31] localizing horizontal (x) and vertical (y) coordinates of 19 body keypoints (i.e., forehead, nose, ears, upper neck, shoulders, elbows, wrists, upper chest, right/left/mid pelvis, knees, and ankles) (step 1, Figure 2), creating infant skeletons. The motion tracker had previously been trained and validated on infant videos satisfying GMA standards, following the anatomical definitions of In-Motion Poses[32]. Further technical details on the motion tracker are described in the original papers by Groos et al.[31,32].

2) Skeleton Sequence

The infant skeletons of all video frames composed a spatiotemporal skeleton sequence (step 2, Figure 2), representing infant movements in the video. The skeleton sequence was divided into 5 second windows, which were processed by the Deep Learning-based prediction model to estimate CP risk in that particular time window.

3) Deep Learning-Based CP Prediction Model

Automatic detection of movement features: To automatically detect movement features related to CP outcome, a novel Deep Learning procedure was developed. As illustrated by step 3 in Figure 2, a Deep Learning model consists of multiple layers. The initial layers detect features of simple movements of a single limb or joint, whereas the subsequent layers detect features of complex whole-body movements. To prevent manual selection bias, the number of layers and type of computation in each layer was set by an automatic search for optimal Deep Learning models on the data for training and internal validation. The first 10 automatically selected models were defined as artificial experts and retrained on the seven internal validation samples (eTable 2 in the Supplement provides internal validation results). Each of the resulting 70 artificial expert instances utilized the biomechanical properties (position, velocity, and body segment length) in 5 second windows to detect complex whole-body movement features that distinguished infants with CP from infants without CP. See eAppendix 3, eFigure 1, eTable 3, and eTable 4 in the Supplement for details on the automatic search procedure and configurations of selected Deep Learning models.

Group of artificial experts and uncertainty of decision: Based on the obtained movement features in each of the 70 artificial expert instances, the CP risk was estimated on a continuous scale from low (0.0%) to high (100.0%). The median value of the 70 individual artificial expert predictions was used as CP risk in the 5 second window, with uncertainty of CP risk color coded based on the agreement across the 70 predictions. Green (0-17 (<25.0%) predict CP) and yellow (18-35 (≤50.0%) predict CP) represent certain and uncertain prediction of no CP, and orange (36-52 (>50.0%) predict CP) and red (53-70 (>75.0%) predict CP) represent uncertain and certain prediction of CP, respectively.

4) CP risk

The final score for CP risk in a total video was estimated as the median CP risk across all 5 second windows of the skeleton sequence (step 4, Figure 2). This score was used to classify an infant into CP or no CP based on a fixed decision threshold (see eAppendix 4, eFigure 2, and eTable 5 in the Supplement for different thresholds). Classification into CP was considered certain (red) if >75.0% and uncertain (orange) if >50.0% of the artificial expert instances classified as CP, and classification into no CP uncertain (yellow) if ≤50.0% and certain (green) if <25.0% classified as CP (step 4, Figure 2).

Conventional Machine Learning-Based CP Prediction

To enable fair comparison between the Deep Learning-based CP prediction and the Conventional Machine Learning method previously presented by our group[23], retraining of the Conventional Machine Learning method on the dataset of the present study was performed. See Ihlen et al.[23] for more details on the Conventional Machine Learning-based CP prediction.

Statistical Analysis

The sensitivity of the methods on the external validation was fixed a priori at the level of GMA in the present study to ensure fair comparisons. Clopper-Pearson was used to provide exact 95% confidence intervals of sensitivity, specificity, positive and negative predictive value (PPV and NPV), and accuracy, computed with the conf package in R (R Core Team) version 4.0. The association between CP risk and GMFCS level was assessed with difference in CP risk in infants with ambulatory (GMFCS I, II, or III) and non-ambulatory CP (GMFCS IV or IV) using 2-sided Wilcoxon rank sum test computed with SciPy in Python (Python Software Foundation) version 3.6. *P* values below 0.05 were considered statistically significant. Wilcoxon rank sum test was also used to assess difference in CP risk of infants with spastic unilateral and spastic bilateral CP.

Results

Among 139 high-risk infants comprising the external validation, 21 (15.1%) had a diagnosis of CP (see eTable 1 in the Supplement for details). Predictive accuracies of the Deep Learning method, GMA and previously reported Conventional Machine Learning method are presented in Table 1. The Deep Learning method achieved higher accuracy than the Conventional Machine Learning method ($P < .001$), but no significant improvement compared to GMA ($P = .11$).

Table 1. Predictive values on external validation[a] given a fixed sensitivity of 70.0%[b]

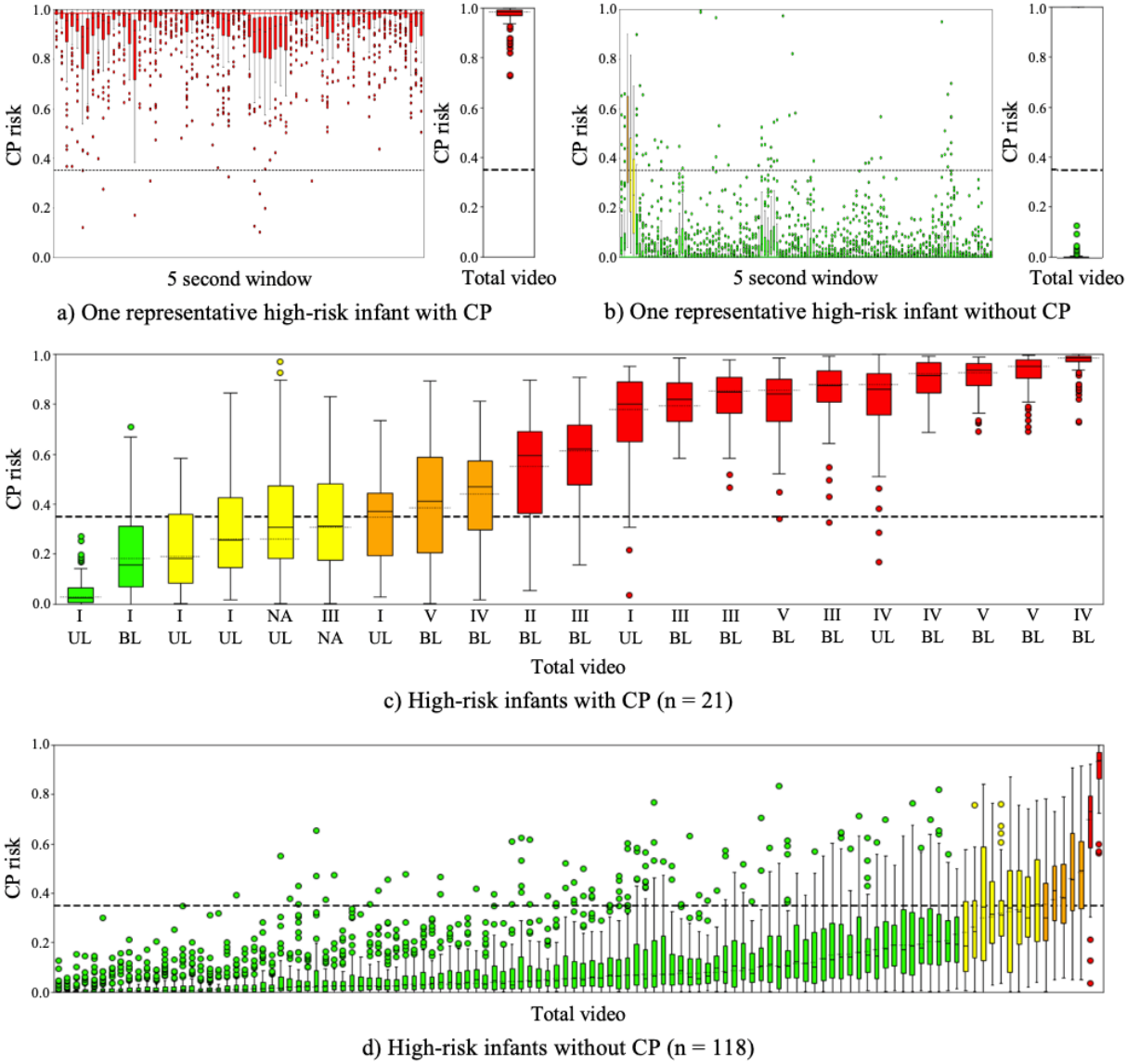
Method	TP	FP	TN	FN	Sensitivity %	Specificity %	PPV %	NPV %	Accuracy %
Deep Learning method	15	7	111	6	71.4 [47.8, 88.7]	94.1 [88.2, 97.6]	68.2 [45.1, 86.1]	94.9 [89.2, 98.1]	90.6 [84.5, 94.9]
General Movement Assessment	14	13	102	6	70.0 [45.7, 88.1]	88.7 [81.5, 93.8]	51.9 [32.0, 71.3]	94.4 [88.3, 97.9]	85.9 [78.9, 91.3]
Conventional Machine Learning method	15	32	86	6	71.4 [47.8, 88.7]	72.9 [63.9, 80.7]	31.9 [19.1, 47.1]	93.5 [86.3, 97.6]	72.7 [64.5, 79.9]

All values are provided in percentages, along with 95% confidence interval.
Abbreviations: TP, true positives; FP, false positives; TN, true negatives; FN, false negatives; PPV, positive predictive value; NPV, negative predictive value.
[a] The external validation comprised four infants (one with CP, three without CP) with exaggerated FMs (excluded by GMA), yielding three true negatives and one false negative both with Deep Learning-based and Conventional Machine Learning-based method.
[b] Sensitivity is fixed at the level of GMA.

Fourteen (66.7%) of the 21 infants with CP were classified with high certainty, including 12 (85.7%) true positives (i.e., red classifications in Figure 3c) and two (14.3%) false negatives (green). Moreover, 104 (88.1%) of the 118 infants without CP were classified with high certainty, 102 (98.1%) true negatives (green in Figure 3d) and two (1.9%) false positives (red). Figure 3a and b illustrate CP risks across 5 second windows for one infant with certain classification of CP and no CP, respectively.

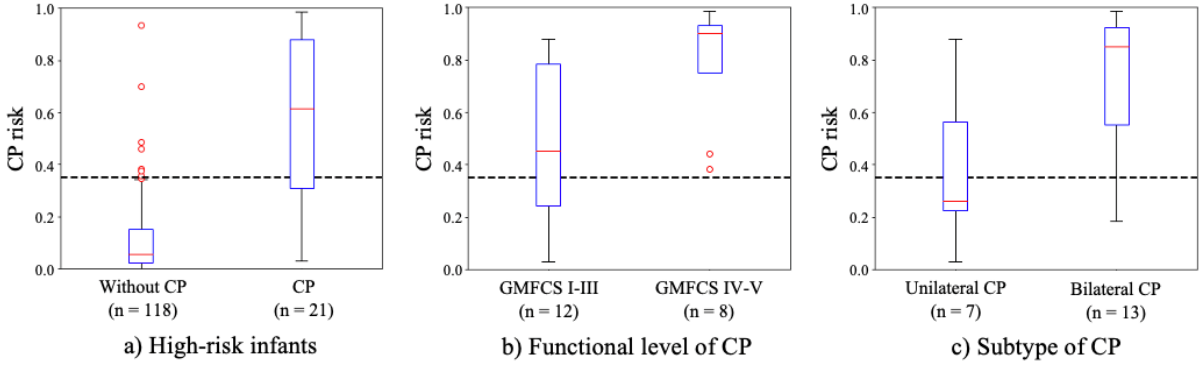
The Deep Learning-based CP prediction model had higher sensitivity (i.e., percentage of infants above decision threshold) in infants with non-ambulatory CP (100.0%; 95% CI: 63.1%-100.0%) compared to ambulatory CP (58.3%; 95% CI: 27.7%-84.8%; $P = .02$), and in infants with spastic bilateral CP (92.3%; 95% CI: 64.0%-99.8%) compared to spastic unilateral CP (42.9%; 95% CI: 9.9%-81.6%; $P < .001$), as depicted by Figure 3c. Figure 4 displays significantly higher estimated CP risk for non-ambulatory motor function (median: 0.90; IQR: 0.75-0.93) compared to ambulatory motor function (median: 0.45; IQR: 0.24-0.78; $P = .007$), and for spastic bilateral CP (median: 0.85; IQR: 0.55-0.92) compared to spastic unilateral CP (median: 0.26; IQR: 0.23-0.56; $P = .03$).

Figure 3. Predictions of infants in the external validation (n = 139)



From top: a) and b) CP risk in 5 second windows (left) and aggregated CP risk (right) across the total video of one representative high-risk infant in the external validation with and without CP, respectively, both classified correctly with high classification certainty. c) and d) Distribution of individual CP risks (y-axis) and boxplots of classification uncertainties of the 70 artificial expert predictions in the Deep Learning method among c) high-risk infants in the external validation with CP (n = 21) with x-axis displaying GMFCS level (i.e., I-V) and CP subtype at time of diagnosis and d) high-risk infants in external validation without CP (n = 118). The dashed horizontal line reflects the decision threshold. In box plots, black dashed and solid horizontal lines are aggregated CP risk and median CP risk across artificial experts, respectively, lower and upper edges are interquartile ranges, and whiskers the range or 1.5 times the interquartile range. Red and orange color coding represent certain and uncertain classification into CP, whereas green and yellow represent certain and uncertain classification into no CP.
Abbreviations: UL, spastic unilateral CP; BL, spastic bilateral CP; NA, not available.

Figure 4. Boxplots of CP risk of infants in the external validation (n = 139) of different outcomes



From left: distribution of CP risk across infants in the external validation a) with and without CP, b) with ambulatory versus non-ambulatory CP (GMFCS I-III and IV-V, respectively), and c) with spastic unilateral CP versus spastic bilateral CP. In box plots, red solid horizontal lines are median aggregated CP risk across infants, lower and upper edges are interquartile ranges, and whiskers the range or 1.5 times the interquartile range.

Discussion

In this study, a fully automated Deep Learning-based early CP prediction demonstrated predictive accuracy non-inferior to the clinically recommended GMA in an external multicenter sample of high-risk infants. This is a significant improvement compared to previously published Conventional Machine Learning-based CP prediction[23]. Furthermore, the prediction model differentiated between infants who developed ambulatory and non-ambulatory CP, as well as unilateral and bilateral CP.

The high external validity of the proposed Deep Learning method reflects the robustness of automated assessment of spontaneous movements in infants with various medical risk factors from different countries and with variation in video quality. This is a significant progress from previous studies, which either lack sufficient sample size or external validity[17]. Moreover, the method uses a single 5-minute video, which can easily be performed in a non-invasive manner in clinic or from home[33]. This suggests potential for widespread clinical adoption. The ability to flag certainty of predictions through a color-coding scheme and the flexibility to adjust the decision threshold in accordance with a preference of few false negatives or few false positives, underpin the potential clinical relevance of the prediction model. Furthermore, the capacity to differentiate between unilateral and bilateral CP, and ambulatory and non-ambulatory motor function may have important clinical implications. This may support decisions in early pediatric care by initiating targeted intervention to improve function, prevent complications, and enhance efficacy of follow-up.

The improved predictive accuracy of using Deep Learning for CP prediction in comparison to Conventional Machine Learning commonly applied in former studies[18-23], may be due to several reasons. First, Deep Learning has the capacity to find intricate relationships in the data through processing by several layers. This suggests that Deep Learning could handle the high complexity and variation in infant spontaneous movements. Second, manual selection of movement features, as required by Conventional Machine Learning[20,21] is eliminated by the automatic feature detection. This may also suggest flexibility in adapting the Deep Learning method to other infant samples and clinical outcomes. However, more studies are needed to identify which movement features the Deep Learning model selects as relevant for CP. In the present study, we have not investigated if the prediction model used features related to FMs, other movement and postural patterns in the early motor repertoire (e.g., kicking and body symmetry), or yet unidentified patterns of movement[34].

The non-inferior performance of the Deep Learning method compared to observational GMA may indicate an upcoming paradigm shift in early prediction of CP. A recent review by Silva et al.[14] highlighted that adoption of automated CP prediction in clinical practice is restricted by existing Machine Learning methods lacking the predictive accuracy of GMA. The feasibility of home-based smartphone recordings[33,35] and associated infant motion tracking[32] may be combined with the proposed Deep Learning-based CP prediction to obtain a fully automated system for clinical decision support.

The sensitivity of observational GMA was lower than reported in some previous studies[9,36], but similar to what has been found in other studies from outside the highly skilled academic settings[37,38]. A sensitivity below what is commonly reported[4] may, at least partly, be due to the classification of sporadic FMs as normal. This is in contrast to what is commonly taught in courses by the General Movements Trust, but it increases accuracy and PPV as shown in a previous study by our group[26]. Furthermore, in our study, a single assessment around 12 weeks CA may have contributed to lower sensitivity, compared to studies performing several assessments throughout the FMs period[39,40].

The present study included infants recruited from several sites based on a variety of risk factors for perinatal brain injury[22,26-28]. Despite the diverse set of risk factors and clinical characteristics of infants, the prevalence of CP in each diagnostic group matches numbers seen in literature[41-43]. This suggests that the results are generalizable to clinical follow-up programs for NICU graduates based on an increased risk of adverse neurodevelopment.

Limitations

Our study has several limitations. A separate dataset for method development limits the number of infants with CP for assessing external validity. This limits the possibility of doing subgroup analysis of CP subtypes and GMFCS levels. A minority of infants were assessed for CP before two years of age, and this may have resulted in a few infants with mild phenotypes not being identified. Short follow-up may also have led to less accurate GMFCS classification due to lower reliability below two years of age[44]. However, inaccurate GMFCS classification in a few infants is unlikely to change the general interpretation of results, since classification rarely changes from ambulatory CP to non-ambulatory CP, and vice versa[44]. The present study comprised videos recorded in a standardized setup, and hence the Deep Learning-based CP prediction requires validation in home-based smartphone recordings.

Conclusions

The proposed Deep Learning-based prediction model for CP based on spontaneous movements in a video taken at 3 months CA, demonstrated predictive accuracy non-inferior to observational GMA on external validation in a high-risk population. The predictive model also differentiated between infants with ambulatory and non-ambulatory CP and infants with unilateral and bilateral CP. A fully automated system for infant motion tracking and CP prediction may serve as an important decision support for clinicians caring for high-risk infants. Future research is needed to identify specific movement biomarkers related to CP outcome and to facilitate widespread clinical use.

Article Information

Author Contributions: Dr. Adde and Dr. Ihlen had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: Groos, Adde, deRegnier, Fjørtoft, Gaebler-Spira, Peyton, Schreiber, Silberg, Songstad, Thomas, Øberg, Ihlen, Støen
Acquisition, analysis, or interpretation of data: Groos, Adde, Aubert, Boswell, deRegnier, Fjørtoft, Gaebler-Spira, Haukeland, Loennecken, Msall, Möinichen, Pascal, Peyton, Ramampiaro, Silberg, Songstad, Thomas, Van den Broeck, Øberg, Ihlen, Støen
Drafting of the manuscript: Groos, Adde, Ihlen, Støen
Critical revision of the manuscript for important intellectual content: Groos, Adde, Aubert, Boswell, deRegnier, Fjørtoft, Gaebler-Spira, Haukeland, Loennecken, Msall, Möinichen, Pascal, Peyton, Ramampiaro, Schreiber, Silberg, Songstad, Thomas, Van den Broeck, Øberg, Ihlen, Støen
Statistical analysis: Groos, Adde, Ihlen, Støen
Obtained funding: Adde, deRegnier, Peyton, Ihlen, Støen
Administrative, technical, or material support: Adde, Boswell, deRegnier, Pascal, Peyton, Ramampiaro, Schreiber, Silberg, Songstad, Thomas, Van den Broeck, Ihlen, Støen
Supervision: Adde, Ramampiaro, Ihlen, Støen

Conflict of Interest Disclosures: None reported.

Funding/Support: The study was supported by the Liaison Committee between the Central Norway Regional Health Authority and the Norwegian University of Science and Technology (SO: 90056100), the Joint Research Committee between St. Olavs Hospital, Trondheim University Hospital and the Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, the Friends of Prentice, Chicago, United States and the Shaw research grant in nursing and allied health professions, Chicago, United States, and RSO funds from the Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology.

Role of the Funder/Sponsor: The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Additional Contributions: The authors thank the parents and children who participated in this study, and the neonatologists who helped recruiting the infants. We also thank Astrid Ustad and Laila Kristoffersen, both at the Norwegian University of Science and Technology, Norway, Randi Vågen from St. Olavs Hospital, Trondheim University Hospital, Norway, Per Gunnar Sandstrøm and Gunfrid V. Størvold at Nord-Trøndelag Hospital Trust, Levanger, Norway, Cathrine Labori at University Hospital of North Norway, and Annamarie Russow from Ann and Robert H Lurie Children's Hospital of Chicago, United States, for their support in performing video recordings and organizing and managing data files and analysis. No additional compensation, outside of their usual salary, was received for their contributions. Written parental consent was obtained for including the infant still image in Figure 2.

References

1. Novak I, Hines M, Goldsmith S, Barclay R. Clinical prognostic messages from a systematic review on cerebral palsy. *Pediatrics.* 2012;130(5):e1285-312. doi:10.1542/peds.2012-0924.

2. Rosenbaum P, Paneth N, Leviton A, et al. A report: the definition and classification of cerebral palsy April 2006. *Dev Med Child Neurol Suppl.* 2007;109:8-14.

3. Palisano R, Rosenbaum P, Walter S, Russell D, Wood E, Galuppi B. Gross motor function classification system for cerebral palsy. *Dev Med Child Neurol.* 1997;39(4):214-23. doi:10.1111/j.1469-8749.1997.tb07414.x.

4. Novak I, Morgan C, Adde L, et al. Early, accurate diagnosis and early intervention in cerebral palsy: advances in diagnosis and treatment. *JAMA Pediatr.* 2017;171(9):897-907. doi:10.1001/jamapediatrics.2017.1689.

5. Anderson V, Spencer-Smith M, Wood A. Do children really recover better? Neurobehavioural plasticity after early brain insult. *Brain.* 2011;134(Pt 8):2197-221. doi:10.1093/brain/awr103.

6. Morgan C, Fetters L, Adde L, et al. Early Intervention for Children Aged 0 to 2 Years With or at High Risk of Cerebral Palsy: International Clinical Practice Guideline Based on Systematic Reviews. *JAMA Pediatr.* 2021;175(8):846-858. doi:10.1001/jamapediatrics.2021.0878.

7. Baird G, McConachie H, Scrutton D. Parents' perceptions of disclosure of the diagnosis of cerebral palsy. *Arch Dis Child.* 2000;83(6):475-80. doi:10.1136/adc.83.6.475.

8. Guttmann K, Flibotte J, DeMauro SB. Parental perspectives on diagnosis and prognosis of neonatal intensive care unit graduates with cerebral palsy. *J Pediatr.* 2018;203:156-162. doi:10.1016/j.jpeds.2018.07.089.

9. Bosanquet M, Copeland L, Ware R, Boyd R. A systematic review of tests to predict cerebral palsy in young children. *Dev Med Child Neurol.* 2013;55(5):418-26. doi:10.1111/dmcn.12140.

10. Einspieler C, Yang H, Bartl-Pokorny KD, et al. Are sporadic fidgety movements as clinically relevant as is their absence? *Early Hum Dev.* 2015;91(4):247-52. doi:10.1016/j.earlhumdev.2015.02.003.

11. Hadders-Algra M. Neural substrate and clinical significance of general movements: an update. *Dev Med Child Neurol.* 2018;60(1):39-46. doi:10.1111/dmcn.13540.

12. Maitre N. Skepticism, cerebral palsy, and the General Movements Assessment. *Dev Med Child Neurol.* 2018;60(5):438. doi:10.1111/dmcn.13733.

13. Peyton C, Pascal A, Boswell L, et al. Inter-observer reliability using the General Movement Assessment is influenced by rater experience. *Early Hum Dev.* 2021;161:105436. doi:10.1016/j.earlhumdev.2021.105436.

14. Silva N., Zhang D, Kulvicius T, et al. The future of General Movement Assessment: The role of computer vision and machine learning - A scoping review. *Res Dev Disabil.* 2021;110:103854. doi:10.1016/j.ridd.2021.103854.

15. Marcroft C, Khan A, Embleton ND, Trenell M, Plötz T. Movement recognition technology as a method of assessing spontaneous general movements in high risk infants. *Front Neurol.* 2015;5:284. doi:10.3389/fneur.2014.00284.

16. Cabon S, Porée F, Simon A, Rosec O, Pladys P, Carrault G. Video and audio processing in paediatrics: a review. *Physiol Meas.* 2019;40(2):02TR02. doi:10.1088/1361-6579/ab0096.

17. Redd CB, Karunanithi M, Boyd RN, Barber LA. Technology-assisted quantification of movement to predict infants at high risk of motor disability: A systematic review. *Res Dev Disabil.* 2021;118:104071. doi:10.1016/j.ridd.2021.104071.

18. Stahl A, Schellewald C, Stavdahl Ø, Aamo OM, Adde L, Kirkerod H. An optical flow-based method to predict infantile cerebral palsy. *IEEE Trans Neural Syst Rehabil Eng.* 2012;20(4):605-14. doi: 10.1109/TNSRE.2012.2195030.

19. Adde L, Helbostad J, Jensenius AR, Langaas M, Støen R. Identification of fidgety movements and prediction of CP by the use of computer-based video analysis is more accurate when based on two video recordings. *Physiother Theory Pract.* 2013;29(6):469-75. doi:10.3109/09593985.2012.757404.

20. Rahmati H, Martens H, Aamo OM, Stavdahl Ø, Støen R, Adde L. Frequency analysis and feature reduction method for prediction of cerebral palsy in young infants. *IEEE Trans Neural Syst Rehabil Eng.* 2016;24(11):1225-1234. doi:10.1109/TNSRE.2016.2539390.

21. Orlandi S, Raghuram K, Smith CR, et al. Detection of atypical and typical infant movements using computer-based video analysis. *Annu Int Conf IEEE Eng Med Biol Soc*. 2018;3598-3601. doi:10.1109/EMBC.2018.8513078.

22. Adde L, Helbostad JL, Jensenius AR, Taraldsen G, Grunewaldt KH, Støen R. Early prediction of cerebral palsy by computer-based video analysis of general movements: a feasibility study. *Dev Med Child Neurol.* 2010;52(8):773-8. doi:10.1111/j.1469-8749.2010.03629.x.

23. Ihlen EA, Støen R, Boswell L, et al. Machine learning of infant spontaneous movements for the early prediction of cerebral palsy: A multi-site cohort study. *J Clin Med.* 2019;9(1):5. doi:10.3390/jcm9010005.

24. Irshad MT, Nisar MA, Gouverneur P, Rapp M, Grzegorzek M. AI approaches towards Prechtl's assessment of general movements: A systematic literature review. *Sensors (Basel).* 2020;20(18):5321. doi:10.3390/s20185321.

25. Goodfellow I, Bengio Y, Courville A. Deep learning. *MIT Press*. 2016.

26. Støen R, Boswell L, De Regnier RA, et al. The predictive accuracy of the general movement assessment for cerebral palsy: a prospective, observational study of high-risk infants in a clinical follow-up setting. *J Clin Med.* 2019;8(11):1790. doi:10.3390/jcm8111790.

27. Pascal A, Govaert P, Ortibus E, et al. Motor outcome after perinatal stroke and early prediction of unilateral spastic cerebral palsy. *Eur J Paediatr Neurol.* 2020;29:54-61. doi:10.1016/j.ejpn.2020.09.002.

28. Aker K, Thomas N, Adde L, et al. Prediction of outcome from MRI and general movements assessment after hypoxic-ischaemic encephalopathy in low-income and middle-income countries: data from a randomised controlled trial. *Arch Dis Child Fetal Neonatal Ed.* 2022;107(1):32-38. doi:10.1136/archdischild-2020-321309.

29. Einspieler C, Prechtl H, Bos A, Ferrari F, Cioni G. Prechtl's Method on the Qualitative Assessment of General Movements in Preterm, Term and Young Infants. *Mac Keith Press*. 2004;167.

30. Cans C. Surveillance of cerebral palsy in Europe: a collaboration of cerebral palsy surveys and registers. *Dev Med Child Neurol.* 2000;42(12):816-24. doi:10.1017/s0012162200001511.

31. Groos D, Ramampiaro H, Ihlen EA. EfficientPose: Scalable single-person pose estimation. *Appl Intell*. 2021;51(4):2518-2533. doi:10.1007/s10489-020-01918-7.

32. Groos D, Adde L, Støen R, Ramampiaro H, Ihlen EA. Towards human-level performance on automatic pose estimation of infant spontaneous movements. *Comput Med Imaging Graph.* 2022;95:102012. doi:10.1016/j.compmedimag.2021.102012.

33. Adde L, Brown A, Van Den Broeck C, et al. In-Motion-App for remote General Movement Assessment: a multi-site observational study. *BMJ Open.* 2021;11(3):e042147. doi:10.1136/bmjopen-2020-042147.

34. Einspieler C, Bos AF, Krieber-Tomantschger M, et al. Cerebral palsy: early markers of clinical phenotype and functional outcome. *J Clin Med.* 2019;8(10):1616. doi:10.3390/jcm8101616.

35. Kwong AK, Eeles AL, Olsen JE, Zannino D, Kariotis T, Spittle AJ. Instructional guides for filming infant movements at home are effective for the General Movements Assessment. *J Paediatr Child Health.* 2021. doi:10.1111/jpc.15838.

36. Kwong AK, Fitzgerald TL, Doyle LW, Cheong JL, Spittle AJ. Predictive validity of spontaneous early infant movement for later cerebral palsy: a systematic review. *Dev Med Child Neurol.* 2018;60(5):480-489. doi:10.1111/dmcn.13697.

37. Constantinou JC, Adamson-Macedo EN, Mirmiran M, Fleisher BE. Movement, imaging and neurobehavioral assessment as predictors of cerebral palsy in preterm infants. *J Perinatol.* 2007;27(4):225-9. doi:10.1038/sj.jp.7211664.

38. Datta AN, Furrer MA, Bernhardt I, et al. Fidgety movements in infants born very preterm: predictive value for cerebral palsy in a clinical multicentre setting. *Dev Med Child Neurol.* 2017;59(6):618-624. doi:10.1111/dmcn.13386.

39. Prechtl HF, Einspieler C, Cioni G, Bos AF, Ferrari F, Sontheimer D. An early marker for neurological deficits after perinatal brain lesions. *Lancet.* 1997;349(9062):1361-3. doi:10.1016/S0140-6736(96)10182-3.

40. Bruggink JL, Einspieler C, Butcher PR, Van Braeckel KN, Prechtl HF, Bos AF. The quality of the early motor repertoire in preterm infants predicts minor neurologic dysfunction at school age. *J Pediatr.* 2008;153(1):32-9. doi:10.1016/j.jpeds.2007.12.047.

41. Grunt S, Mazenauer L, Buerki SE, et al. Incidence and outcomes of symptomatic neonatal arterial ischemic stroke. *Pediatrics.* 2015;135(5):e1220-8. doi:10.1542/peds.2014-1520.

42. Zhang S, Li B, Zhang X, Zhu C, Wang X. Birth asphyxia is associated with increased risk of cerebral palsy: a meta-analysis. *Front Neurol.* 2020;11:704. doi:10.3389/fneur.2020.00704.

43. Hafström M, Källén K, Serenius F, et al. Cerebral palsy in extremely preterm infants. *Pediatrics.* 2018;141(1):e20171433. doi:10.1542/peds.2017-1433.

44. Gorter JW, Ketelaar M, Rosenbaum P, Helders PJ, Palisano R. Use of the GMFCS in infants with CP: the need for reclassification at age 2 years or older. *Dev Med Child Neurol.* 2009;51(1):46-52. doi:10.1111/j.1469-8749.2008.03117.x.

Supplementary Online Content

eAppendix 1. Related Previously Published Papers

The present study is related to five previously published papers:

Støen et al.[1]: This study (Støen et al.[1]) did not assess a Machine Learning-based CP prediction method but was a study of GMA and its predictive accuracy for CP. The present study harnessed video recordings, GMA classifications, and CP outcomes of the infant sample from Norway and United States collected by Støen et al.[1]

Adde et al.[2]: This study (Adde et al.[2]) evaluated a simple statistical method for Conventional Machine Learning-based CP prediction without assessing the external validity. The Machine Learning method used was entirely different from the method presented in the present study. The present study harnessed video recordings, GMA classifications, and CP outcomes of the infant sample from Norway collected by Adde et al.[2]

Pascal et al.[3]: This study (Pascal et al.[3]) did not assess a Machine Learning-based CP prediction method but assessed the prediction of CP using GMA. The present study harnessed video recordings, GMA classifications, and CP outcomes of the infant sample from Belgium collected by Pascal et al.[3]

Aker et al.[4]: This study (Aker et al.[4]) did not assess a Machine Learning-based CP prediction method but assessed CP prediction using GMA. The present study harnessed video recordings, GMA classifications, and CP outcomes of the infant sample from India collected by Aker et al.[4]

Ihlen et al.[5]: The present study and the study by Ihlen et al.[5] both harnessed the video recordings, GMA classifications, and CP outcomes of the infant sample from Norway and United States collected by Støen et al.[1] but the previous study of Ihlen et al.[5] evaluated a semiautomated Conventional Machine Learning method for CP prediction, in contrast to the fully automated Deep Learning method of the present study. The study of Ihlen et al.[5] neither assessed the external validity of the Conventional Machine Learning-based CP prediction.

eAppendix 2. Characteristics of Included Infants

The study sample comprised infants from four previous studies[1-4]. The first study[1] enrolled infants who were referred to neurodevelopmental follow-up at discharge from tertiary care NICUs at three sites in Norway and two sites in the United States, including a) preterm infants with a gestational age (GA) < 29 weeks and/or birthweight (BW) < 1000 g (n=119), b) infants with congenital heart disease (CHD) in need of cardiac surgery before 4 weeks of age (n=41), c) infants with neonatal arterial ischemic stroke (NAIS; n=9), d) infants with neonatal encephalopathy (NE; n=43), e) infants with other risk factors for adverse neurological development (e.g. congenital anomalies and/or chromosomal abnormalities with extended NICU stay beyond 10 weeks CA, neonatal seizures, central nervous system (CNS) abnormalities, abnormal neonatal imaging, CNS infection, severe hypoglycemia; n=80), and f) infants with GA < 31 weeks and/or BW < 1500 g, enrolled in a randomized controlled trial of two different doses of inhaled nitric oxide for neuroprotection, requiring oxygen at birth (NOVA2 trial; https://clinicaltrials.gov/ct2/show/NCT00515281; n=116). The second study[2] recruited infants from 4 sites in Norway, including a) preterm infants with GA < 28 weeks and/or BW < 1000 g (n=15), b) infants with NE (n=4), c) infants with NAIS (n=2), and d) infants with other risk factors for adverse neurological development (n=9). The third study[3] comprised infants with NAIS from six sites in Belgium (n=37). The fourth study[4] comprised infants admitted to a tertiary care NICU in South India with NE (n=82). See eTable 1 for descriptive statistics, including cerebral palsy (CP) status, of infants in datasets for training & validation (i.e., method development) and test set (i.e., external validation) and study affiliation.

eTable 1. Characteristics of Infants in Datasets

	Heterogeneous high-risk infants[1,2]		**Infants with perinatal stroke[3]**		**Infants with neonatal encephalopathy[4]**	
	Training & validation	*Test*	*Training & validation*	*Test*	*Training & validation*	*Test*
No. infants (%)	328 (78.5)	110 (79.1)	28 (6.7)	9 (6.5)	62 (14.8)	20 (14.4)
GA, mean (SD), w.	31.7 (6.2)	30.6 (5.9)	34.8 (5.4)	36.4 (5.1)	39.1 (1.4)	39.4 (1.2)
BW, mean (SD), g	1849 (1245)	1678 (1209)	2430 (1088)	2669 (1209)	2904 (512)	2948 (555)
Sex, no. (%)						
Male	181 (55.2)	57 (51.8)	17 (60.7)	5 (55.6)	39 (62.9)	11 (55.0)
Female	147 (44.8)	53 (48.2)	11 (39.3)	4 (44.4)	23 (37.1)	9 (45.0)
CA rec., mean (SD), w.	12.3 (1.3)	12.2 (1.4)	12.0 (1.6)	11.1 (0.8)	11.2 (1.8)	12.0 (1.7)
CA fol., mean (SD), m.	40.5 (15.7)	39.7 (17.3)	27.9 (5.3)	27.7 (6.2)	19.4 (4.3)	20.8 (6.8)
CP diagnosis, no. (%)	41 (12.5)	14 (12.7)	10 (35.7)	3 (33.3)	12 (19.4)	4 (20.0)
CP subtype, no. (%)						
Spastic unilateral	12 (29.3)	4 (28.6)	8 (80.0)	2 (66.7)	1 (8.3)	1 (25.0)
Spastic bilateral	22 (53.7)	9 (64.3)	2 (20.0)	1 (33.3)	9 (75.0)	3 (75.0)
Dyskinetic	5 (12.2)				1 (8.3)	
Ataxic	1 (2.4)					
Not available	1 (2.4)	1 (7.1)			1 (8.3)	
GMFCS, no. (%)						
I	13 (31.7)	2 (14.3)	3 (30.0)	2 (66.7)	2 (16.7)	2 (50.0)
II	5 (12.2)		4 (40.0)		1 (8.3)	1 (25.0)
III	3 (7.3)	5 (35.7)	1 (10.0)		1 (8.3)	
IV	8 (19.5)	3 (21.4)	1 (10.0)	1 (33.3)	2 (16.7)	
V	11 (26.8)	3 (21.4)			5 (41.7)	1 (25.0)
Not available	1 (2.4)	1 (7.1)	1 (10.0)		1 (8.3)	

Abbreviations: GA, gestational age; w., weeks; BW, birth weight; CA, corrected age; rec., recording; fol., follow-up; m., months; CP, cerebral palsy; GMFCS, Gross Motor Function Classification System.

eAppendix 3. Deep Learning-Based CP Prediction Model

In the following section, we describe stepwise how the Deep Learning-based prediction model was developed:

Step 1: The skeleton sequence was resampled to 30 Hz and a 5-point temporal median filter was applied to each skeletal coordinate time series. Subsequently, the skeleton sequence was centralized according to the median mid pelvis location and normalized by two times the trunk length of the infant (i.e., median distance from upper chest to mid pelvis).

Step 2: The skeletal sequence was divided into 5 second windows comprising $T\ (= 5\ s \cdot 30\ s^{-1} = 150)$ time steps of $J\ (= 19)$ body keypoints (i.e., joints), each with $D\ (= 2)$ coordinates (i.e., $x_{t,j}$ and $y_{t,j}$ for keypoint j at time step t). In each 5 second window, the infant skeletons were rotated spatially for vertical alignment of upper chest and mid pelvis in the first time step.

Step 3: Biomechanical properties, i.e., position $\boldsymbol{p_{t,v}}$ and velocity $\boldsymbol{v_{t,j}}$ of each skeletal keypoint and distance from the neighboring body keypoint $\boldsymbol{b_{t,j} = p_{t,j} - p_{t,j_{adj}}}$, were defined for each time step in a 5 second window and used as input variables to the Deep Learning-based CP prediction model.

Step 4: The processing of the input variables was performed by an ensemble of Graph Convolutional Networks (GCNs) (i.e., artificial expert instances) where the overall architecture is illustrated in eFigure 1. The configurations of the input branches, main branch, and pooling layer in eFigure 1 and their general properties were determined by *K*-Best Search[6] with the search space summarized in eTable 3. The performance of GCN architectures was evaluated by the area under the receiver operating characteristic curve (AUC) on internal validation folds of the dataset. All GCNs were optimized using He initialization[7], Stochastic Gradient Descent with learning rate of $5 \cdot 10^{-4}$ and Nesterov momentum of 0.9, and batch size of 32 on an NVIDIA Tesla V100 GPU. 5 second windows were randomly sampled from the skeleton sequence and data augmentation with scaling (0.7 – 1.3), rotation (+/- 45 degrees), and translation (+/- 0.3) was employed to avoid overfitting. The best performing GCN architectures from 10 iterations of *K*-Best Search (i.e., artificial experts), as summarized in eTable 4, were trained for 200 epochs on each of the seven folds in the cross-validation, comprising seven sets of model parameters associated with each of the GCN architectures.

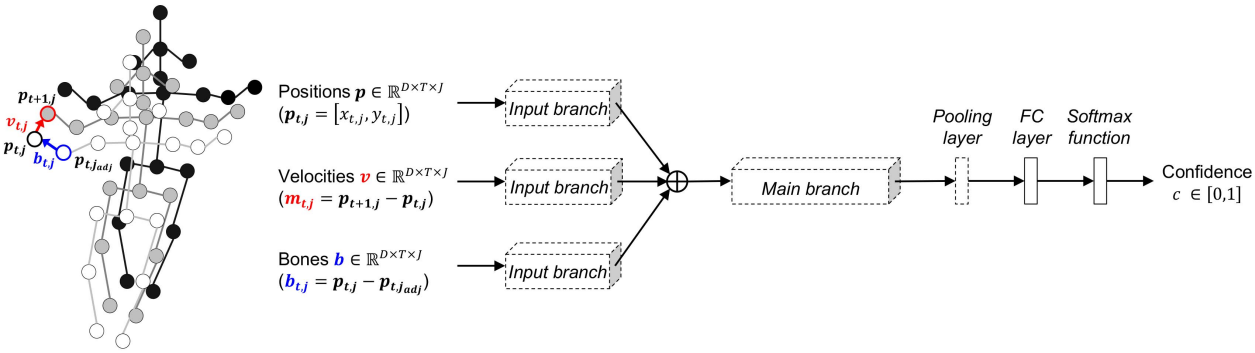
Step 5: The seven versions of the 10 obtained GCN models in eTable 4 constituted the 70 artificial expert instances. The artificial expert instances were utilized to yield CP predictions according to eFigure 1 on unseen skeleton sequences of the test set with 2.5 seconds overlap between each 5 second window.

eTable 2. Internal Validation of Deep Learning Method

TP	FP	TN	FN	Sensitivity %	Specificity %	PPV %	NPV %	Accuracy %
45	16	339	18	71.4 [58.7, 82.1]	95.5 [92.8, 97.4]	73.8 [60.9, 84.2]	95.0 [92.2, 97.0]	91.9 [88.8, 94.3]

The internal validity was evaluated using 7-fold cross-validation, with sensitivity fixed at the level of GMA (i.e., 70.0%). All values are provided in percentages, along with 95% confidence interval.
Abbreviations: TP, true positives; FP, false positives; TN, true negatives; FN, false negatives; PPV, positive predictive value; NPV, negative predictive value.

eFigure 1. Overall Architecture of Graph Convolutional Networks



From left: A GCN processes biomechanical properties, describing positions p, velocities (change in position) v, and bones (distance from the neighboring body keypoint) b, of a window of T time steps from a skeleton sequence, J body keypoints (joints), and D spatial dimensions, through parallel input branches, followed by a main branch, pooling layer, fully connected (FC) layer, and softmax function to yield confidence c about the risk of CP from 0.0 (no CP) to 1.0 (CP).

eTable 3. Search Space of 20 Architectural Choices

	Architectural choice	Alternatives
Input branches	No. modules of input branches	1, 2, 3
	Width of input branches	6, 8, 10, 12
	Block type in initial module[a]	Basic[8], Bottleneck[8], MBConv[9]
	Residual type in initial module[a]	None, Block[8], Module[8], Dense[8]
	No. temporal scales in input branches	1, 2, 3, Linear[b]
Main branch	No. levels of main branch	1, 2
	No. modules of main branch levels	1, 2, 3
	Width of first level of main branch[c]	6, 8, 10, 12
	No. temporal scales in main branch	1, 2, 3, Linear[b]
Pooling layer	Pooling layer type	Global average, Spatial average
General properties	Graph convolution type	Spatial configuration[10], DA $2^{11,d}$, DA $4^{11,d}$, DA 4+$2^{11,d}$
	Block type[e]	Basic[8], Bottleneck[8], MBConv[9]
	Bottleneck factor	2, 4
	Residual type[e]	None, Block[8], Module[8], Dense[8]
	SE type	None, Inner[12], Outer[12], Both[12]
	SE ratio	2, 4
	SE ratio type	Relative, Absolute
	Attention type	None, Channel[8], Frame[8], Joint[8]
	Nonlinearity type	ReLU[13], Swish[13]
	Temporal kernel size	3, 5, 7, 9

Abbreviations: MBConv, mobile inverted bottleneck convolution; DA, disentangled aggregation; SE, Squeeze-and-Excitation; ReLU, rectified linear unit.

[a] The initial module is the first module of input branches.

[b] Linear scaling indicates that number of temporal scales increases by one for each module.

[c] For the second level of the main branch, the width is doubled, while also reducing the time dimension by a factor of 2.

[d] Graph convolutions with disentangled aggregation have different number of hops in neighborhood (i.e., 2 or 4), where 4+2 yields separate number of hops in input modules and main module, with 4 and 2, respectively.

[e] There is a separate architecture choice associated with the initial module.

eTable 4. Characteristics of Architectures Obtained by *K*-Best Search

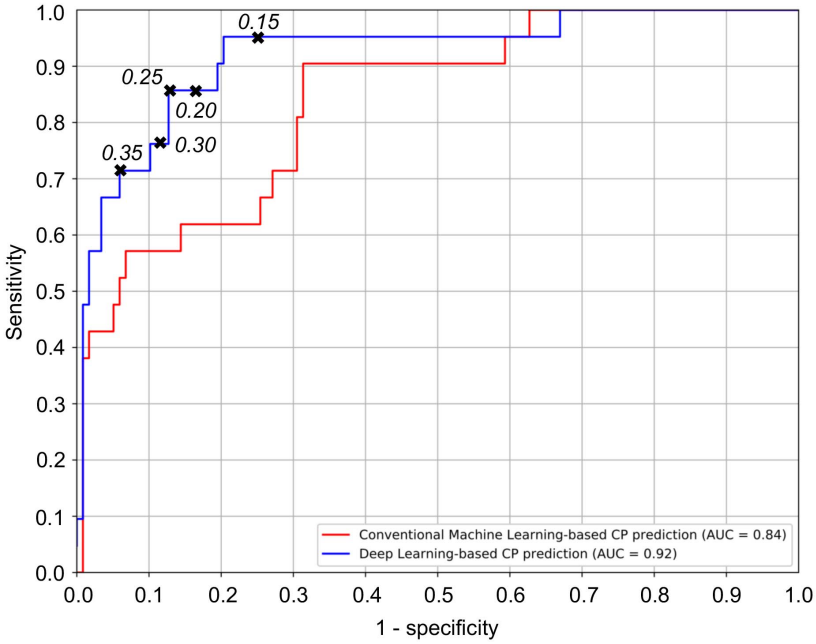
	***K*-Best Search**									
Architectural choice	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
No. modules of input br.	3	2	3	3	2	3	1	2	3	2
Width of input br.	10	10	12	10	10	12	8	6	6	12
Block type in initial mod.	Bottl.	Basic	Basic	Basic	Bottl.	Basic	Basic	MBC.	Bottl.	Basic
Residual type in initial mod.	None	Den.	None	Block	Den.	Den.	Mod.	Block	Den.	Den.
No. tmp. scales in input br.	1	3	2	2	3	1	3	2	1	2
No. levels of main br.	1	1	1	1	2	2	2	2	2	1
No. modules of main br. levels	1	3	2	1	1	3	3	2	1	3
Width of first level of main br.	12	12	8	8	6	10	12	12	12	10
No. tmp. scales in main br.	1	2	2	Lin.	3	Lin.	3	1	Lin.	3
Pooling layer type	Gl.	Gl.	Gl.	Sp.	Gl.	Gl.	Sp.	Gl.	Sp.	Sp.
Graph convolution type	DA 2	DA 4+2	SC	DA 4	DA 4	DA 2	DA 2	DA 4	DA 2	SC
Block type	Basic	MBC.	Basic	Basic	Basic	Bottl.	Basic	Basic	Basic	Basic
Bottl. factor	4	2	2	4	2	4	4	4	4	4
Residual type	None	Block	Mod.	Den.	None	Block	None	Den.	None	None
SE type	None	Outer	Inner	None	Outer	None	None	Outer	Outer	None
SE ratio	-	4	2	-	2	-	-	4	4	-
SE ratio type	-	Abs.	Abs.	-	Abs.	-	-	Abs.	Abs.	-
Attention type	Ch.	Ch.	None	Ch.	Ch.	Ch.	None	None	Ch.	Ch.
Nonlinearity type	ReLU	Sw.	ReLU	Sw.	ReLU	Sw.	Sw.	ReLU	ReLU	Sw.
Tmp. kernel size	9	7	7	7	7	3	9	5	9	7
AUC	0.949	0.942	0.938	0.943	0.937	0.956	0.953	0.953	0.932	0.947

Abbreviations: br., branch; bottl., bottleneck; MBC., mobile inverted bottleneck convolution; den., dense; mod., module; tmp., temporal; lin., linear; gl., global; sp., spatial; DA, disentangled aggregation; SC, spatial configuration; SE, Squeeze-and-Excitation; abs., absolute; ch., channel; ReLU, rectified linear unit; sw., swish; AUC, area under the receiver operating characteristic curve.

eAppendix 4. Decision Thresholds

From the receiver operating characteristic (ROC) curve in eFigure 2, we observe there are several feasible options for the choice of decision threshold in the Deep Learning-based CP prediction, depending on the preference of high sensitivity and few false negatives (e.g., decision threshold of 0.15), high specificity and few false positives (e.g., 0.35), or a compromise of the two (e.g., 0.25), with an overall area under the ROC curve of 0.92. In eTable 5, the distributions of uncertainty of predictions across different decision thresholds are also presented.

eFigure 2. ROC Curves on External Validation



Different decision thresholds (0.15, 0.20, 0.25, 0.30, and 0.35) of the Deep Learning method are marked with cross and text in italic.
Abbreviations: AUC, area under the receiver operating characteristic curve.

eTable 5. Predictive Values of Decision Thresholds on External Validation

Thres.	TP	FP	TN	FN	Sens. %	Spec. %	PPV %	NPV %	Acc. %	Infants with CP				Infants without CP			
0.15	20	30	88	1	95.2	74.6	40.0	98.9	77.7	17	3	0	1	70	18	18	12
0.20	18	20	98	3	85.7	83.1	47.4	97.0	83.5	14	4	2	1	78	20	10	10
0.25	18	15	103	3	85.7	87.3	54.6	97.2	87.1	13	5	2	1	87	16	9	6
0.30	16	14	104	5	76.2	88.1	53.3	95.4	86.3	12	4	4	1	95	9	10	4
0.35	15	7	111	6	71.4	94.1	68.2	94.9	90.6	12	3	4	2	102	9	5	2

Red and orange color coding represent certain and uncertain classification into CP, whereas green and yellow represent certain and uncertain classification into no CP.
Abbreviations: thres., decision threshold; TP, true positives; FP, false positives; TN, true negatives; FN, false negatives; sens., sensitivity; spec., specificity; PPV, positive predictive value; NPV, negative predictive value; acc., accuracy.

eReferences

1. Støen R, Boswell L, De Regnier RA, et al. The predictive accuracy of the general movement assessment for cerebral palsy: a prospective, observational study of high-risk infants in a clinical follow-up setting. *J Clin Med.* 2019;8(11):1790. doi:10.3390/jcm8111790.

2. Adde L, Helbostad JL, Jensenius AR, Taraldsen G, Grunewaldt KH, Støen R. Early prediction of cerebral palsy by computer-based video analysis of general movements: a feasibility study. *Dev Med Child Neurol.* 2010;52(8):773-8. doi:10.1111/j.1469-8749.2010.03629.x.

3. Pascal A, Govaert P, Ortibus E, et al. Motor outcome after perinatal stroke and early prediction of unilateral spastic cerebral palsy. *Eur J Paediatr Neurol.* 2020;29:54-61. doi:10.1016/j.ejpn.2020.09.002.

4. Aker K, Thomas N, Adde L, et al. Prediction of outcome from MRI and general movements assessment after hypoxic-ischaemic encephalopathy in low-income and middle-income countries: data from a randomised controlled trial. *Arch Dis Child Fetal Neonatal Ed.* 2022;107(1):32-38. doi:10.1136/archdischild-2020-321309.

5. Ihlen EA, Støen R, Boswell L, et al. Machine learning of infant spontaneous movements for the early prediction of cerebral palsy: A multi-site cohort study. *J Clin Med.* 2019;9(1):5. doi:10.3390/jcm9010005.

6. Groos D. Convolutional networks for video-based infant movement analysis: Towards objective prognosis of cerebral palsy from infant spontaneous movements. *NTNU*. 2022 (submitted thesis).

7. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE international conference on computer vision.* 2015;1026-1034. doi:10.1109/ICCV.2015.123.

8. Song YF, Zhang Z, Shan C, Wang L. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. *Proceedings of the 28th ACM International Conference on Multimedia.* 2020;1625-1633. doi:10.1145/3394171.3413802.

9. Tan M, Le Q. EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning, PMLR.* 2019;97:6105-6114.

10. Yan S, Xiong Y, Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. *Thirty-second AAAI conference on artificial intelligence*. 2018.

11. Liu Z, Zhang H, Chen Z, Wang Z, Ouyang W. Disentangling and unifying graph convolutions for skeleton-based action recognition. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 2020;143-152. doi:10.1109/CVPR42600.2020.00022.

12. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2018;7132-7141. doi:10.1109/CVPR.2018.00745.

13. Ramachandran P, Zoph B, Le QV. Searching for activation functions. *arXiv.* 2017;1710.05941.

ISBN 978-82-326-5582-3 (printed ver.)
ISBN 978-82-326-6906-6 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU

Norwegian University of
Science and Technology