Guoxia Xu

Improving Image Quality, Content, and Practicality: Knowledge-Oriented Methods for Information Enhancement

Thesis for the Degree of Philosophiae Doctor

Gjøvik, August 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Information Technology and Electrical Engineering
Department of Computer Science

ISBN 978-82-326-7250-9 (printed ver.)
ISBN 978-82-326-7249-3 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

IMT-report 2023:275

Doctoral theses at NTNU, 2023:275

Printed by NTNU Grafisk senter

DECLARATION

I, Guoxia Xu, declare that this thesis titled 'Improving Image Quality, Content, and Practicality: Knowledge-Oriented Methods for Information Enhancement', and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Guoxia Xu	Date

Contents

List of Figures x

Abstract 1

Acknowledgments 5

1 Introduction 7

1.1 Research Context 7

1.2 Research Motivation 9

1.3 Research Objectives 14

1.4 Research Questions 16

1.5 Listed of Included Publications 20

1.6 Thesis Structure 25

2 Literature Review 27

2.1 Image Quality Enhancement for Single-modal Information Enhancement (SIE) 27

2.1.1 The development of image quality enhancement technology 28

2.1.2 The detail description of Generative Adversarial Network (GAN) 30
2.2 Image Fusion for Multi-modal Information Enhancement (MIE) . 33
2.2.1 The development of image fusion technologies 34
2.2.2 The detailed description of computational method 36
2.3 Image Analysis for Task-driven Information Enhancement (TIE) . 38
2.3.1 The development of information enhancement for image segmentation 38
2.3.2 The development of information enhancement for object tracking 41
2.3.3 Revisited Temporal Response Based Method 43

3 Summary of Included Publications 45

3.1 Paper I: SSP-Net: A Siamese-based Structure-Preserving Generative Adversarial Network for Unpaired Medical Image Enhancement [176] 45
3.1.1 Abstract 45
3.1.2 Motivation 46
3.1.3 Methods 46
3.1.4 Result 47
3.2 Paper II: Disentangled Spatial-Transformation Guided GAN for Unpaired Medical Image Quality Enhancement 48
3.2.1 Abstract 48
3.2.2 Motivation 49
3.2.3 Methods 50
3.2.4 Result 51
3.3 Paper III: FCFusion: Fractal Component-wise Modeling with Group Sparsity for Medical Image Fusion 52
3.3.1 Abstract 52

3.3.2 Motivation . . . 52
3.3.3 Methods . . . 53
3.3.4 Result . . . 53
3.4 Paper IV: JADD-GAN: A Joint Attention Generative Adversarial Data Fusion Network for Object Detection and Tracking . . . 54
3.4.1 Abstract . . . 54
3.4.2 Motivation . . . 54
3.4.3 Methods . . . 55
3.4.4 Result . . . 56
3.5 Paper V: Multi-label Abdominal Image Segmentation with Partially Labeled Data: A Prototypical Consistent Learning Perspective 56
3.5.1 Abstract . . . 56
3.5.2 Motivation . . . 57
3.5.3 Methods . . . 57
3.5.4 Result . . . 58
3.6 Paper VI: Learning the Distribution-Based Temporal Knowledge with Low-Rank Response Reasoning for UAV Visual Tracking . . 59
3.6.1 Abstract . . . 59
3.6.2 Motivation . . . 60
3.6.3 Methods . . . 61
3.6.4 Result . . . 61

4 Discussion **63**

4.1 Unpaired Image Enhancement with Deep Neural Network . . . 63
4.1.1 Limitation Analysis . . . 66
4.2 Unsupervised Image Fusion via Optimization Learning and Deep Learning . . . 66
4.2.1 Limitation Analysis . . . 69

4.3 Leveraging the external information as weak supervision for high-level vision tasks. 70
4.3.1 Limitation Analysis 71
4.4 Additional Contributions 71
4.4.1 Data Fusion by Deep Learning 71
4.4.2 Temporal Information for Visual Tracking 72
4.4.3 The Feature Learning in Deep Learning 73

5 Conclusions and future perspectives **75**
5.1 Conclusion 75
5.2 Future Research Orientation 76

Bibliographie **77**

Appendices: **101**

List of Figures

1.1 The common discussions of low-level vision [40]. 9

1.2 The trends of publications covering the keyword Information Enhancement [1]. 11

1.3 The example of computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), and single photon emission computed tomography (SPECT). 12

1.4 The example of the degraded natural image and medical image. . . 13

1.5 A schematic overview of research problems, motivation, and contents 16

1.6 Relationship between research questions and research papers. . . . 21

2.1 Multi-modal medical images of the brain 28

2.2 The developments of Generative Adversarial Network (GAN) in [67]. 31

2.3 Examples of the cartoon and texture decomposition. 39

2.4 Examples of the partially labelled datase from [197]. 39

2.5 Examples of the Correlation filter tracker. 41

3.1 The framework of the proposed SSP-Net. 47

3.2 The proposed DSSGAN framework consists of two groups of GANs. 51

3.3 The image fusion process of our FCFusion model. 52

3.4 The architecture of the proposed JADD-GAN. 55

3.5 The framework of the proposed method. 58

3.6 Our proposed method adopts low-rank temporal response constraint and group feature selection to improve the stability of the correlation filter. 60

4.1 Examples of unpaired LQ images and HQ images of corneal confocal microscopy. 64

4.2 The sparse coding process of our FCFusion model. 67

Acronyms

ASR Adaptive Sparse Representation. 35

CNN Convolutional Neural Network. 30, 40

CS-MCA Convolutional Sparse Representation Morphological Component Analysis. 35, 37, 38

CSC Convolutional Sparse Coding. 38

CSR Convolutional Sparse Representation. 35, 37

CT Computed Tomography. 24, 27, 34, 38, 40

CycleGAN Cycle-Consistent Generative Adversarial Network. 29, 32, 33

DCF Discriminative Correlation Filter. 42, 43

DWT Discrete Wavelet Transformation. 36

EnlightenGAN Deep Light Enhancement Generative Adversarial Network. 30

GAN Generative Adversarial Network. vi, ix, 29–32, 64, 75

GFP Green Fluorescent Protein. 33

HIF Heterogeneous Image Fusion. 33

ISO International Organizations Standardization. 11

MCA Morphological Component Analysis. 35, 37

MIE Multi-modal Information Enhancement. vi, 14, 15, 18, 33, 35, 37, 63, 75

MOTs Multiple Organs and Tumors. 24, 40, 41

MRI Magnetic Resonance Imaging. 27, 30, 33, 34

MSE Mean Squared Error. 30

NSCT Non-subsampling Contour Transformation. 36

NSST Non-subsampling Shearlet Transformation. 36

OCR Optical Character Recognition. 8

OMP Orthogonal Matching Pursuit. 37

PC Phase Contrast. 33

PET Positron Emission Computed Tomography. 27, 33, 34

RP Research Paper. 21

RQ Research Question. 16

SIE Single-modal Information Enhancement. v, 14–17, 27, 29, 31, 63, 75

SOMP Simultaneous Orthogonal Matching Pursuit. 37

SPECT Single-photon Emission Computed Tomography. 33, 34

SR Sparse Representation. 35–37

StillGAN Structure and Illumination Constrained Generative Adversarial Network. 29

TIE Task-driven Information Enhancement. vi, 14, 15, 19, 38, 39, 41, 43, 63, 75

UAV Unmanned Aerial Vehicle. 42

VGG VGG Deep Convolutional Networks. 30

Abstract

With the all-around change caused by artificial intelligence (AI) technology, the need for perception in real scenes is increasingly urgent. As an essential medium to perceive the world, the integrity and richness of information often determine the performance of the algorithm. However, in the process of image capture, there are usually a variety of uncontrollable physical factors, resulting in image information degeneration and missing, thus affecting the acquisition and utilization of information in the application process of high-level visual tasks. The concept of information enhancement is becoming increasingly important in today's world, as we are constantly bombarded with vast amounts of information from multiple sources. It is not enough to have access to information. We need to be able to extract the most relevant and useful information from this abundance of data. This is where information enhancement comes in, as it aims to improve the quality and relevance of information through various techniques.

The present thesis examines that the core theme is ***information enhancement*** and a knowledge-oriented systematic framework of enhancement for various applications is proposed, which includes single-modal information enhancement (SIE), multi-modal information enhancement (MIE), and task-driven information enhancement (TIE). First, SIE investigates the single modal medical image quality enhancement techniques for improved visual inspection with the downstream application. Second, MIE aims to integrate multiple modal medical images by feature-level fusion for further understanding and complementing information enhancement. Finally, TIE extends the generality of enhancement perspective that embeds into high-level visual tasks.

Single-modal information enhancement (SIE) is focused on enhancing the quality and clarity of information from a single source or modality. The goal of SIE is

to improve the accuracy and usefulness of the information within a single modality, which can be especially important in situations where the data quality is low or the signal-to-noise ratio is poor. To achieve the research goal of this thesis, a framework of image information enhancement for two knowledge-oriented methods based on generative adversarial neural networks (GAN) is proposed.

Multi-modal information enhancement (MIE) involves integrating and enhancing information from multiple sources or modalities. This can include combining images to improve understanding or merging data from multiple sensors to provide a complete picture. The goal of MIE is to increase the amount and quality of information available by integrating multiple sources, thereby enabling better decision-making and understanding. To investigate the feasibility of MIE, one optimization-based fusion method and one deep learning GAN-based method are proposed and verified by medical image segmentation task and object detection/tracking.

Task-driven information enhancement (TIE) focuses on enhancing information specifically for a particular application. For this topic, two methods are proposed to verify how to embed knowledge-oriented information enhancement in high-level vision applications, including medical image segmentation and object detection/tracking.

The research conducted as part of the Ph.D. thesis on information enhancement will likely involve developing and evaluating various knowledge-oriented methods for implementing these information enhancement techniques. This may include analyzing image data from various modalities and developing algorithms to improve the quality, content, and relevance practicability of the information contained within these sources.

Overall, the Ph.D. thesis on information enhancement is to contribute to our understanding of how to extract the most useful and relevant information from the vast amount of data, to advance our knowledge of critical principles and techniques.

致我的爸爸徐厚洪，妈妈周礼娟，女朋友张瑞和姐姐徐欣欣。

Acknowledgments

Since starting my research in 2015, I have come to believe that pursuing academic research and education has truly transformed my life. It has changed my fate in a profound way, and I cannot help but feel fortunate. Looking back on my journey, there are so many people who have supported and helped me, for whom I am immensely grateful.

First and foremost, I express my profound gratitude towards my distinguished supervisors, Hao Wang, Marius Pedersen, Hu Zhu, and Meng Zhao. Their erudition and patience in acquainting me with the research domains of computer vision, discussing related work directions, and exhorting me to explore have been invaluable. Their fervor for scientific inquiry is truly inspiring and has engendered within me a constant state of curiosity and optimism.

I also extend my thanks to my esteemed colleagues at NTNU, including Di Wu, Yushan Pan, Xu Chen, and Mengtao Sun, for their unwavering support and for their probing inquiries about my research. Additionally, as an interdisciplinary study, I received considerable aid from my counterparts Lizhen Deng, Tiayu Yan, and Chunming He at NJUPT, who provided me with exposure to a new field of deep learning and medical image analysis.

Most significantly, I express my deepest love and appreciation to my family: my parents, Houhong Xu and Lijuan Zhou, my girlfriend, Rui Zhang, and my sister, Xinxin Xu. Their constant presence, unwavering love, and encouragement have been instrumental in making this journey as meaningful as it has been.

Finally, I feel profoundly fortunate to have been a member of the research group led by my supervisor, Hao Wang, during these extraordinary years of my thesis. To all the members from different universities, I extend my sincerest thanks for the

exhilarating research ideas and discussions that have expanded my perspective and deepened my comprehension of computer science.

Gjøik, May 2023 Guoxia Xu

Chapter 1

Introduction

This chapter aims to provide the details of this thesis's context, motivation, and research questions. In addition, it outlines the related research publications and their interrelationship with the research questions. Finally, the structure of this thesis is given.

1.1 Research Context

Vision is an innate ability for both humans and animals. We effortlessly perceive and understand the world around us without any conscious training. However, for machines, comprehending images is a formidable challenge. Computer vision is a field of study that aims to teach machines how to "see." Through continuous iteration with a clearly defined objective function, modern computer technology can complete various complicated tasks such as image and video classification, target tracking and detection, instance segmentation, and key point detection. Despite the remarkable progress made in the field of computer vision, the cost of training a single complex task is high, data collection is difficult, and the number of interdisciplinary image understanding tasks is overwhelming. The training process requires a clearly defined objective function to complete a complex task. At the same time, human vision is characterized by the ability to perform an extensive array of tasks without explicit instruction.

Nowadays, Artificial Intelligence (AI) has emerged as a burgeoning field worldwide, encompassing a vast array of definitions. Some researchers [151] define AI as "a system that thinks like humans," while others describe it as "a system that acts like humans." However, there is no universally accepted formal definition of AI, as it may vary depending on the environment in which it is applied. Despite this

ambiguity, AI plays a significant role in many aspects of computing technology.

To study this automagic process of vision, there is one main methodological distinction [1] to specify the position of scientific findings concerning research questions in the total field of vision research:

The total field of vision research may be divided into several subfields:

- Low-level Vision — which concerns extracting image properties from the retinal image.
- High-level Vision — which concerns the everyday functionality of perceptual organizations.

In particular, a discernible difference exists between AI with high-level vision and low-level vision. High-level vision has made remarkable progress recently, and its application is increasingly prevalent in various fields, such as face recognition and autonomous driving. Conversely, low-level vision remains a challenging area of research, with limited progress made thus far. Nevertheless, the field of AI is continuously evolving. With ongoing research, further advancements will likely be made in high-level and low-level vision, leading to even more impressive applications.

We acknowledge the tremendous power of machine learning in high-level vision tasks, as it can be trained with only enough data and information. However, giving more attention to the quality and source of the data and potential issues that may arise is crucial, particularly in low-level vision tasks. Understanding low-level vision can greatly aid high-level vision tasks.

For instance, in the past, when performing Optical Character Recognition (OCR) [8], it was necessary to identify text in images. However, the data available for machine learning tasks was limited, covering only a portion of the data. This resulted in undesired data alterations, such as changes in illumination, gradient, and destruction, which were unrelated to the actual words in the image.

Although it is possible to input these data alterations into the machine learning model directly, it is more efficient to remove them beforehand, especially when data is insufficient. Suppose someone has a thorough understanding of low-level vision Figure 1.1. In that case, they can consider removing these alterations before feeding the data into the machine learning model, resulting in more efficient machine learning that relies less on data.-

[1] The details can be found in https://ppw.kuleuven.be/apps/research/petervanderhelm/doc/visionintro.html

Computer Vision (CV) encompasses various facets, including signal acquisition, processing, analysis, and final comprehension, and is closely related to numerous disciplines, such as machine learning, optics, and computer graphics. Moreover, it is intertwined with our comprehension of the physical world. Nonetheless, due to the constraints of sensors, our comprehension often originates from two-dimensional or one-dimensional projections instead of full three-dimensional data. Image enhancement and reconstruction are fundamental research problems in im-

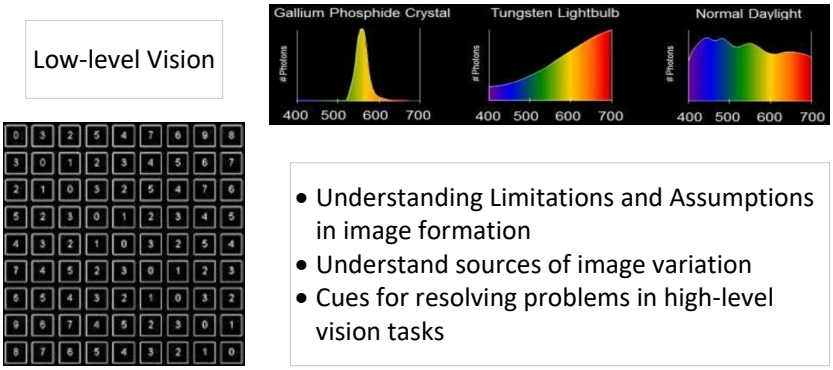

Figure 1.1: The common discussions of low-level vision [40].

age processing, which have been extensively studied and explored by scholars in related fields for a considerable time. The primary causes of image degradation are complex imaging equipment and unfavorable external imaging environments, such as system noise, camera shake, poor weather, and other factors. These can result in images of varying degrees of degradation, manifesting as noisy images, blurry images, and images polluted by rain and fog, among others. Enhancing and reconstructing degraded images to obtain a clear image with good visual quality is an incredibly challenging problem. As the foundational processing steps in many real-world vision systems, image enhancement and reconstruction aim to improve image quality and provide reliable information for subsequent visual decision-making.

1.2 Research Motivation

In this thesis, we would like to introduce the two concepts first. High-content images contain significant information or meaningful content, such as detailed images of cells, tissues, or complex scenes. These images often require sophisticated analysis techniques, such as image segmentation and feature extraction, to extract relevant information from the data.

On the other hand, low-content images contain relatively little information or meaningful content that typically has minimal information or details. Examples of low-content images include simple shapes, abstract patterns, or blurred backgrounds. These images do not typically require sophisticated analysis techniques and can be easily processed.

These terms [161] are commonly used in computer vision, image analysis, and scientific research, and their definitions can be found in various academic papers and textbooks related to these fields. Overall, the definitions of high-content and low-content images are widely accepted and used in various fields, and their specific applications and interpretations may vary depending on the context.

Image information enhancement refers to enhancing high-content images from low-content images, an important class of image processing techniques in low-level computer vision and image processing. Figure 1.2 shows that the citations and publications covering the keyword Information Enhancement, which is for papers published in the given year, not the number of citations in that year [2]. The trends are estimated by the number of publications of the given keyword(s) and show an increasing tendency.

It covers a wide range of real-world AI applications, such as medical imaging [45], [123], natural image [210], [148] amongst others. Other than improving perceptual image quality, it also helps to improve other computer vision tasks[146], [36], [135]. High-content images are desired urgently in the above application areas, such as intelligent surveillance, medical imaging, and remote sensing. To obtain images with higher content, a logical approach would be to upgrade the hardware (e.g., the imaging system).

Although recent years have witnessed the obvious progress of imaging devices and techniques, this kind of approach has two main limitations:

- It is inflexible and costly because the demand for practical applications is constantly changing;
- It can be used only to capture new high-content images but not to enhance the existing low-content images.

Compared to the hardware upgrade-based 'hard' solution, the signal processing-based 'soft' image content enhancement is more flexible and economical. With the image information enhancement techniques that reconstruct a higher content

[2]The search is based on major keywords *"Information Enhancement"* highlighting the main theme of this thesis: https://exaly.com/trends/Information-Enhancement/1970-2020.

output from the low content observation, we can obtain images with high content beyond the limit of imaging systems, thereby improving visual quality and benefiting the subsequent analysis and understanding tasks.

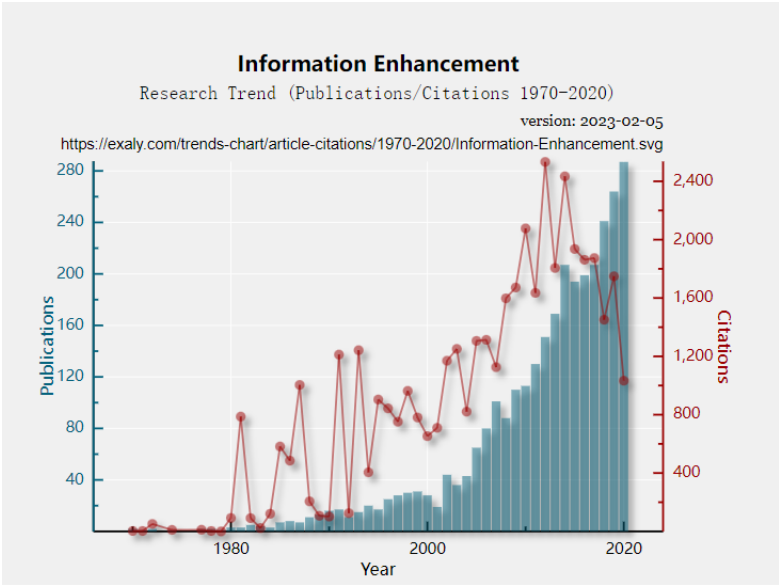

Figure 1.2: The trends of publications covering the keyword Information Enhancement [1].

The main goal of information enhancement is to solve the limitations of imaging equipment hardware, image processing software, and dataset collection environment since the sensors of a single type or setting can not fully characterize the imaging scene.

For example, visible light images usually contain rich texture details [46], but they are vulnerable to the impact of extreme environment and occlusion and lose the objects in the scene. On the contrary, some specific sensors [14] can effectively highlight prominent targets such as pedestrians and vehicles by capturing the radiation information emitted by objects but lack the ability to provide detailed descriptions. In addition, sensors with different International Organizations Standardization (ISO) [145] and exposure time can only capture scene information within their dynamic range, and inevitably lose information beyond the dynamic range.

In addition to the visible light field, there are situations in the medical image field where relevant information needs to be enhanced. In the medical field, imaging technology is mainly divided into using electromagnetic and acoustic energy ima-

ging [17]. The use of acoustic energy imaging refers to using different propagation speeds of ultrasound in different media to achieve real-time imaging directly.

Medical imaging technology has developed rapidly. In addition to Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and Single Photon Emission Computed Tomography (SPECT), there are also conventional imaging technologies such as X-ray and ultrasound. Taking the brain medical image, which can be checked in Figure 1.3 as an example, the image obtained by CT can provide rich anatomical details and can clearly distinguish the skull, brain parenchyma, cerebrospinal fluid, and non-pathological calcification areas in the brain; MRI can display abundant physiological and biochemical information, including nerves, blood vessels and soft tissues in the brain; PET/SPECT images can reflect the metabolism of markers in normal and diseased tissues and the blood flow signals of the brain. Due to the existence of equipment,

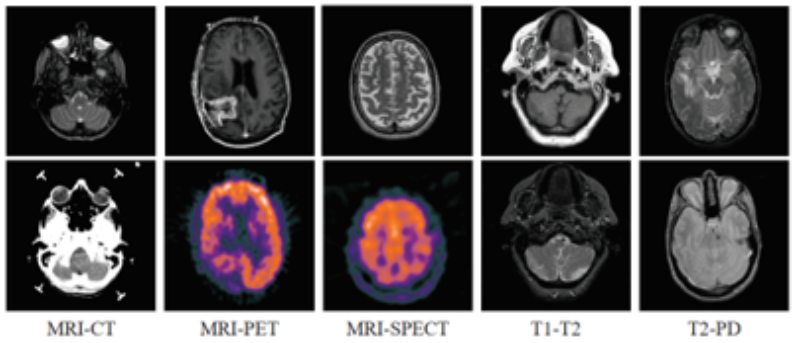

Figure 1.3: The example of computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), and single photon emission computed tomography (SPECT).

environment, operators, and other factors, low-quality images present significant challenges and lead to a poor explanation of the predictions. Image is often taken under sub-optimal lighting conditions, under the influence of backlit, uneven light, and dim light, due to inevitable environmental and/or technical constraints such as insufficient illumination and limited exposure time, which can be checked in Figure 1.4. Such images suffer from compromised aesthetic quality and unsatisfactory transmission of information for high-level tasks. This problem is generally challenging and inherently ill-posed since there are not always solved exactly.

For example, the inevitable different dye compositions in staining, illumination variants [74] in scanning technologies, and image artifacts such as noise and blurring would limit the prediction accuracy. Not only that, but other optical-based medical scenarios will also occur to the appearance variants. Endoscopic ima-

ging [115] suffers from non-uniform illumination due to the directory of the light source. Furthermore, different imaging technologies [131] have their own advantages, limitations, and clear scope. Various enhancement techniques are required to achieve these clear sensing situations to challenge these imaging problems. Information enhancement is the basic fundamental component for the development of AI applications.

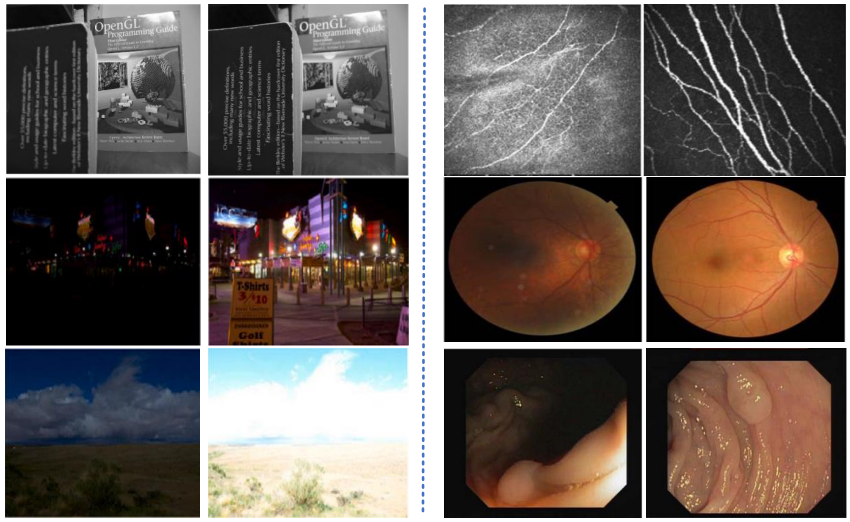

Figure 1.4: The example of the degraded natural image and medical image.

Considering the current real situation, this is a complicated task because no general unified theory is available for computational imaging to guide the enhancement of information. Furthermore, the lack of a quantitative standard for selecting the best criterion for information enhancement adds to the complexity of this task. Information enhancement of an image is the process of improving the quality or making it more visually appealing to some extent. The selection criterion for image enhancement depends on the specific application and the goals of the image enhancement. Here are some common selection criteria for image enhancement:

- **Image Quality:** The image quality refers to the resolution, brightness, contrast, sharpness, color balance, and other visual aspects of the image. The selection criterion for image enhancement should aim to improve the overall image quality to make it more visually appealing.
- **Image Content:** The selection criterion for image enhancement can depend on the image's content. For example, images with text may require different enhancement techniques than images with natural landscapes or portraits.

In some cases, the image content can determine the detail required in the enhancement process.

- **Downstream Application:** The selection criterion for image enhancement should also consider the image's intended use. For example, an image used for scientific analysis may require different enhancement techniques than one used for marketing purposes.

Meanwhile, it is usually difficult to measure whether the quality of the enhanced information is improved and to what extent. Even having multiple parameters controlling the enhancement output makes this situation even worse. Traditional image enhancement technologies [2][51] have been proposed to improve downstream application tasks. However, the dependence of these models on manually designed models has a greater impact on the results in complex optical medical scenarios.

In recent years, many methods [100, 159] introduced deep learning into image enhancement technology and improved the model's performance by learning direct mapping based on pairing learning. However, most of these fully supervised models need to train the image pairs strictly, which leads to the model's performance being greatly affected by the dataset. Not only information enhancement focuses on image quality, but it also relies on the content of the image description.

From [107], we notice three indicators to generalize the objective enhancement criteria:

- Make the enhancement processes intelligent.
- Select a suitable transform.
- Select the optimal parameters.

It becomes necessary to develop information enhancement that can be used as an effective standard when it is used for pre-processing and followed by other image-processing steps. Developing information enhancement technologies is helpful in supporting the low-level research community.

1.3 Research Objectives

The present work is a systematic framework of information enhancement around several applications. The main scientific output includes Single-modal Information Enhancement (SIE), Multi-modal Information Enhancement (MIE), and Task-driven Information Enhancement (TIE) for the current enhancement research community. Hence, we formulate the main objective as follows:

This research aims to develop methods to improve the visual appearance and content of several types of images. In order to reach the main goal, the research work is divided into three sub-objectives as defined below:

— **Objective 1:** *To create advanced research techniques and practical uses, construct a technical basis for improving Single-modal Information Enhancement (SIE) through the creation of an unpaired deep learning training framework.*

Medical image quality is the key judgment basis of clinical diagnosis and treatment in healthcare. However, low-quality medical images can not provide a large amount of precise information on human tissues or organs beyond the reach of human eyes. Furthermore, a network's generalization ability may be restricted due to a shortage of training data, making it difficult to perform effectively. Especially in real-world scenarios, it is hard to collect sufficient good registration datasets. Most methods only focus on the global appearance to enhance the low quality from the guidance of high-quality data in the same feature extraction network to achieve unpaired learning. The information entanglement caused by ignoring specific image characteristics between different quality images leads to the degradation of enhancement performance. Therefore, this objective focuses on developing a data-driven neural network framework with a new perspective on feature extraction and different comparative assessment study on state-of-the-art enhancement methods.

— **Objective 2:** *To propose a fusion guided framework for Multi-modal Information Enhancement (MIE), considering multiple optimization-driven constraints, data characteristics, and theoretical improvement.*

To deal with multi-modal information interactions, various strategies have been developed with success [184],[75],[163]. Some works [85] have gradually sought decomposition methods for feature extraction, but they ignored the explicit redundancy removal of the whole framework. The core of this objective is to propose a constraint-based optimization learning and data-driven framework for multi-modal image fusion-guided information enhancement. Medical and natural images are used in experiments to verify the performance of the proposed methods.

— **Objective 3:** *Investigating different knowledge extraction strategies and learning mechanisms to benefit the high-level visual tasks and further extend the Task-driven Information Enhancement (TIE) for high-level visual tasks.*

In this objective, we highlight the inherent concept limited by image quality and image fusion of information enhancement. Our primary emphasis is incorporating task-specific knowledge as external information to supervise high-level visual tasks such as medical image segmentation and visual object tracking. Our research objective is not limited to improving existing representation learning for neural networks from the perspective of information enhancement. Still, we also explore various forms of external information as weak supervision to inform network design.

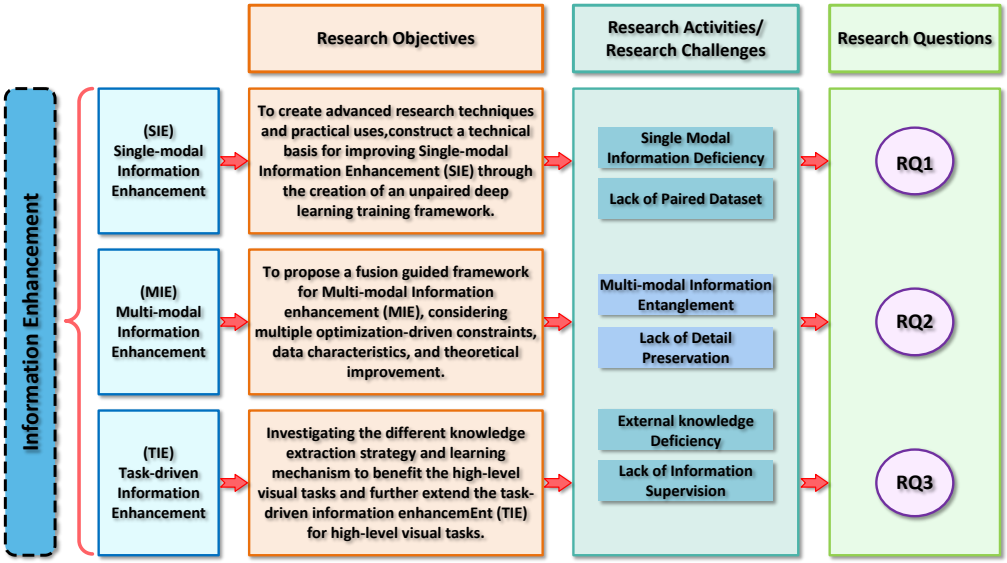

Figure 1.5: A schematic overview of research problems, motivation, and contents

1.4 Research Questions

Three main Research Question (RQ)s were formulated to lead the research activities to accomplish the research objectives. Figure 1.5 describes the research activities with the corresponding research questions. Building on these stated goals described above, this thesis will raise several of the following research questions:

— **RQ 1: What are training mechanisms of deep neural network in an end-to-end learning manner that can solve the problems of poor image quality and lack of paired data in Single-modal Information Enhancement (SIE)?**

The first research question stimulates investigating the existing problems of the training mechanism in Single-modal Information Enhancement (SIE), including domain gap, unpaired dataset, degraded image quality, and feature extraction technologies. This question attempts to recognize opportunities, prospects, and limitations of low-level vision in the medical image area. While traditional manual-engineering works about medical image enhancement have their research topics and corresponding methodologies, with the supporting deep learning powerful feature capability. The novel networks and training mechanism can be introduced into the requirements of real collected medical image data, which are difficult to acquire paired high-quality data in real-world scenarios.

Research question 1 can be further elaborated into more detailed research questions to establish an understanding of practical problems.

— **RQ 1-1:** Is it possible to use deep learning for an unpaired dataset for image enhancement?

  Most fully supervised models need to train the image pairs strictly, which leads to the performance of the model and is greatly affected by the dataset. Furthermore, medical images are restricted by different tissues and lesions, which makes it difficult to collect complete low-quality/high-quality data pairs. Therefore, in recent years, an unpaired data-learning model has become an important research topic in medical image enhancement. This research question gives the framework from different network structures to solve how to maintain structural information in medical image enhancement. This research question is answered in the research paper I and paper II (RP I and RP II; listed in Section 1.5).

— **RQ 1-2:** How to solve the domain gap problem in the unpaired dataset for image enhancement?

  In the current works, it is too complex to recognize the specific information between high-quality and low-quality, only relying on the same encoder network. Thus, it is prone to over-fitting and leads to the entanglement problem in high-dimensional feature representation in terms of the *domain gap, training scale, different structure, texture, and illumination intensity*. This research question is answered in the research paper II.

— **RQ 2: With respect to data fusion objective function and optimization constraint, how the technology of multi-modal fusion can achieve Multi-modal Information Enhancement (MIE) and how to solve information entanglement and lack of detail preservation?**

For the assessment of specific scenarios, different imaging modalities usually provide different information on the target structure. In actual practice, the data with different physical imaging principles have always been used together to obtain a more comprehensive view of specific information reorganization. On the other hand, although the data from different modalities have dramatic differences in appearance, some techniques are always required to aid in the following high-level visual tasks. This research question discusses the multi-modal sources that would lead to mixed mutual information from multiple modalities resulting in sub-optimal accuracy for each modality. To ensure accuracy and clarity, removing any redundant information from a given piece of content is important. This can be achieved by including all the necessary details while avoiding repetition or duplication of information. This approach is particularly useful when presenting complex or technical information, as it allows the data fusion method to focus on the most important points without being distracted by unnecessary or repetitive information. Thus, this is the first question should be how to evaluate the problem of feature extraction and redundancy. Furthermore, how to achieve edge and texture preservation between multi-modal images in a more general perspective?

— **RQ 2-1:** How to solve feature extraction and redundancy problem for multi-modal image information fusion enhancement?

How to preserve important biological information efficiently has become a main challenge. A single modal source provides limited information and cannot meet the requirements of patient verification, disease diagnosis, health monitoring, surgery, and radiation therapy. Fusing different sources into one image is necessary to obtain complementary information to assist aided-diagnosis frameworks. Feature extraction-driven methods have demonstrated that this explicit fusion guideline (extensions of feature extraction capacity only) leads to several issues for medical image fusion. The major problem of these algorithms is that the information is over-completed. Thus, the main solution is adapting the appropriate balance between feature extraction

and redundancy removal. This thesis answers this research question in RP III.

— **RQ 2-2:** How to achieve edge and texture preservation between multi-modal images?

This research question answers how to preserve information by optimization and deep learning from both perspectives. The multi-level feature information and optimization-driven constraints are considered to achieve our goal. This thesis answers this research question in RP III and RP IV.

— **RQ 3: To what extent the external knowledge information and discovering what information enhancement of Task-driven Information Enhancement (TIE) benefiting for high-level vision task decision-making support.**

After building the content and quality information enhancement model, the following research questions are about how we analyze the detail and task-driven information enhancement for high-level visual tasks. To generalize and diagnose the practical applications, we must examine the data characteristics, knowledge generalization, knowledge information distillation, and further precise evaluation models. Research question 3 will be distributed to four precise research sub-questions to analyze the detail task-driven information enhancement from different perspectives and objectives for high-level visual tasks. Two interesting topics are selected to verify the effectiveness of information enhancement in a task-driven manner, including segmentation of multiple organs and tumors on partially labeled medical image datasets and video-based visual object tracking. This research question is answered in RP V and RP VI.

— **RQ 3-1:** What external knowledge information can facilitate analysis of partially labeled medical image segmentation?

To analyze the task of segmenting multiple organs and tumors on a partially labeled medical image dataset, we first investigate the knowledge of conditioning class label information within network training as a manner of information supervision technology. It can also provide an

interesting perspective for solving multi-modal high-level visual tasks. This research question is addressed in RP V.

— **RQ 3-2:** Based on the conditioning class label information, what other method for partially labeled data and how to bring specific information to guide the single feature extraction network to gain the discrepancy between tasks?

According to the information enhancement theory, intra-domain interference of simultaneously segmenting organs and tumors raises a new research question. The question becomes, what other information can be used to solve this interference? What method can be guided to mitigate the discrepancy between tasks? This research question is addressed in RP V.

— **RQ 3-3:** How to embed the temporal transition information for object tracking?

In practice, the online visual tracking task only relies on the current frame information and is short of other information supervision. To improve performance, selecting the proper temporal information is a good choice to mitigate online tracker degeneration. This research question provides a new perspective to introduce how to generalize the data characteristics and exploit the data characteristics based on the information enhancement tool for high-level visual tasks. This research question is addressed in RP VI.

1.5 Listed of Included Publications

Based on the development of information enhancement, this core theme at different dimensions of this thesis, it is a field of what can be applied for research that draws on different data subjects, such integrative studies are important for identifying concrete research requirements and contributions in each research topic and further leads to the advancement of computer science and computer vision development. This thesis was conducted within a core theme around image and information academic framework that involves different public benchmark datasets for experiment verification.

This leads to a number of Research Paper (RP)s on different research topics. This section lists the six research papers included in this thesis, published in international journals or international conference proceedings. Figure 1.6 shows the relationship between research questions and the included RPs. The extended descriptions of the connections can be found in Chapter 3.

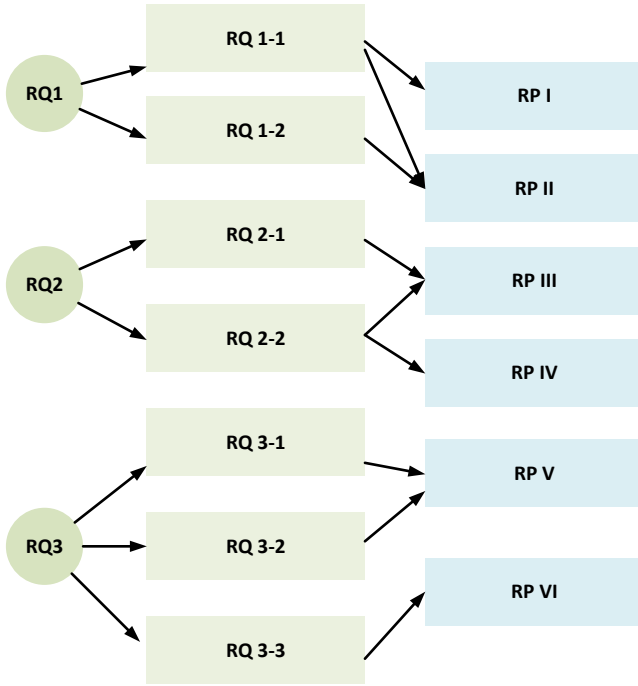

Figure 1.6: Relationship between research questions and research papers.

With the objective of establishing a cross-topic foundation to understand information enhancement in single-modal, multi-modal, and task-driven processes, this thesis contributes to image processing, and computer vision. The thesis is based on a collection of several papers. The list and the main contributions of the work of the papers are given below.

1. RP I [176]:

   Guoxia Xu, Hao Wang, Marius Pedersen, Meng Zhao, Hu Zhu: SSP-Net: A Siamese-based Structure-Preserving Generative Adversarial Network for Unpaired Medical Image Enhancement, IEEE/ACM Transactions on Computational Biology And Bioinformatics, 2023.

   Contribution: In this paper, a dual input mechanism image enhancement

method based on Siamese structure (SSP-Net) is proposed, which takes into account the structure of target highlight (texture enhancement) and background balance (consistent background contrast) from unpaired low-quality and high-quality medical images. Furthermore, the proposed method introduces the mechanism of the generative adversarial network to achieve structure-preserving enhancement by jointly iterating adversarial learning. Experiments comprehensively illustrate the performance in unpaired image enhancement of the proposed SSP-Net compared with other state-of-the-art techniques.

2. RP II:

   Guoxia Xu, Hao Wang, Hu Zhu, Marius Pedersen: Disentangled Spatial-Transformation Guided GAN for Unpaired Medical Image Quality Enhancement, Pending Submission, 2022.

   Contribution: Though cycle-consistent generative adversarial network (CycleGAN) has achieved great progress with unsupervised framework to deal with the unpaired medical image enhancement problem, most CycleGANs only focus on the global appearance to perceive the low-quality and high-quality data in a same feature extraction network to achieve unpaired learning. The information entanglement caused by ignoring the specific image characteristics between different quality images leads to the degradation of enhancement performance. In this paper, a new perspective of disentangled extraction based on CycleGAN (DSSGAN) is introduced, which can provide pixel-level supervision to preserve texture and detail information for image quality enhancement. As a result, we propose a disentangled generator structure to better enhance images of different quality perceptually. Furthermore, to model the spatial transformation in unpaired learning, the spatial transformation module is used to capture the spatial and structural features as supervised information between high-quality and low-quality images, thus fusing with the encoded information to capture the more accurate feature information. The experimental results in three datasets show that our proposed method has promising performance compared to other state-of-the-art algorithms.

3. RP III [175]

   Guoxia Xu, Xiaoxue Deng, Xiaokang Zhou, Marius Pedersen, Lucia Cimmino, Hao Wang: FCFusion: Fractal Component-wise Modeling with Group Sparsity for Medical Image Fusion, IEEE Transactions on Industrial Informatics, 18(12), 9141-9150, 2022.

   Contribution: Multimodal image fusion is the process of combing relevant

biological information that can be used for automated industrial applications. In this paper, we present a novel framework combining fractal constraint with group sparsity to achieve optimal fusion quality. Firstly, we adopt the idea of patch division and component-wise separation to perceive the fractal characteristics across multi-modality sources. Then, to preserve the spatial information against the redundancy of component-entanglement, group sparsity is proposed. A dual variable weighting rule is inherently embedded to mitigate the overfitting across the component penalty. Furthermore, the Alternating Direction Method of Multipliers (ADMM) is conducted for the proposed model optimization. The experiments show that our model has a better performance in quantitative visual quality and qualitative evaluation analysis. Finally, a real segmentation application of PET/CT image fusion proves the effectiveness of our algorithm.

4. RP IV [179]

   Guoxia Xu, Hao Wang, Meng Zhao, Hu Zhu: JADD-GAN: A Joint Attention Generative Adversarial Data Fusion Network for Object Detection and Tracking, the 20th IEEE International Conference on Smart City(SmartCity-2022), 2022.

   Contribution: Image fusion is the fusion of images captured by different sensors to generate a single image with enhanced information, and fusion technology, as one of the important branches in the field of information fusion, mainly realizes the processing of multi-source image information. However, many commonly used fusion methods usually ignore the visual naturalness and information fidelity of the fused images and lack emphasis on the salient information, which makes the fused images unsuitable for human visual perception. To address these shortcomings of existing methods, in this paper, we propose the Joint Attention and Dual Discriminator Generative Adversarial Data Fusion Network JADD-GAN. In the generator module, to increase the extraction of multi-level information by the network, we first adopt a dual encoder structure and give information fusion in the decoder part. Secondly, different discriminators are used for infrared and visible images in order to highlight the thermal radiation information and key textures. The effectiveness of the method is verified by experiments on four datasets, and the results show that the method can effectively highlight the thermal radiation information and key texture details of the fused images, fully demonstrating its great potential and performance.

5. RP V [177]

   Guoxia Xu, Hao Wang, Meng Zhao, Marius Pedersen, Hu Zhu: Multi-

label Abdominal Image Segmentation with Partially Labeled Data: A Prototypical Consistent Learning Perspective, The 7th IEEE Cyber Science and Technology Congress (CyberSciTech 2022), 2022.

Contribution: Recently, accurate automatic Computed Tomography (CT) segmentation of organs and tumors has the potential to facilitate clinical diagnosis and therapy. However, the automatic segmentation of Multiple Organs and Tumors (MOTs) is a complex task since they present variability in the partially labeled data due to limited manpower and resources. The most prevalent techniques are committed to proposing a unified framework for the multi-task segmentation problem while suffering from the domain gap and discrepancy caused by the imbalance of data distribution. To handle the aforementioned imbalance challenges, we introduce a novel prototype assignment strategy as weak enhancement information for a compact intra-class feature representation. Moreover, an exponential-based probability regularization term is proposed to avoid the inter-class imbalance problem caused by forcing the network to provide a consistent prototype label for adjacent features. Experiments comprehensively illustrate the performance of the proposed method compared with other state-of-the-art (SOTA) approaches both qualitatively and quantitatively.

6. RP VI [178]

Guoxia Xu, Hao Wang, Meng Zhao, Marius Pedersen, Hu Zhu: Learning the Distribution-Based Temporal Knowledge with Low-Rank Response Reasoning for UAV Visual Tracking, IEEE Transactions on Intelligent Transportation Systems, IEEE, 2022.

Contribution: The constraint-based correlation filter has shown good performance in object tracking, which has gained a lot of popularity in many intelligence transportation applications. In this work, a distribution-based temporal knowledge-driven method is proposed to leverage the temporal translation property in visual tracking. Instead of focusing on traditional issues in the correlation filter, we provide a new method of learning parametric distribution on temporal knowledge by Wasserstein distance which is successfully embedded to solve the problem of temporal degeneration in the learning process of tracking. Furthermore, we approximate optimal response reasoning with low-rank constraint over response consistency. Furthermore, the proposed method is solved by a simple iterative scheme with alternating direction multiplication ADMM algorithm. We demonstrate superior tracking performance in several public standard tracking benchmarks compared with state-of-the-art algorithms.

1.6 Thesis Structure

This thesis has eleven chapters which are divided into two parts. Part I of the thesis gives an overview of the research work, and Part II presents the included research papers.

Part I: Introductory Chapters

Chapter 1: (this chapter) introduces an overview of the thesis and consists of sections on the research context, motivation, research objectives, research questions, and the list of publications.

Chapter 2: presents a comprehensive and necessary application background, scientific foundation, and related work of the research subject areas.

Chapter 3: presents an extended summary of the included papers published in peer-reviewed internationally recognized conferences and journals. Each paper follows an IMR format: Introduction, Methodology, and Result. Full research papers are provided in Part II of this thesis.

Chapter 4: highlights and reflects upon the main contributions of this research.

Chapter 5: gives the conclusion of the research work, which includes research discussions, followed by some future research orientations and an epilogue.

Part II: Research Papers

Chapters 6-11 include the six research papers that constitute the main part of this thesis. The papers are presented in the same sequence as in Section 1.5.

Chapter 2

Literature Review

2.1 Image Quality Enhancement for Single-modal Information Enhancement (SIE)

With the ongoing advancements in medical imaging technology, a variety of medical imaging modalities have become increasingly prevalent in clinical disease diagnosis, surgical assistance, and health monitoring. Among the most common types of medical images are Magnetic Resonance Imaging (MRI) [152], Positron Emission Computed Tomography (PET) [157], and Computed Tomography (CT) images [18], etc.

Different modes of medical images reflect different body structure information due to variations in imaging methods. MRI-T1 image (Figure 2.1(a)) offers clear and precise anatomical structure information similar to the clinical anatomical map. On the other hand, in MRI-T2 images (Figure 2.1(b)), lesion information is more prominent, providing a more intuitive view of the lesion as compared to normal tissues. Positron Emission Computed Tomography (PET) images (Figure 2.1 (c)) provide a wealth of information on human metabolism and blood flow, but their image resolution is comparatively lower. Computed Tomography (CT) images (Figure 2.1(d)) vividly display skeletal information, but they do not offer soft tissue information. It is essential to consider that medical imaging generates radiation and the image clarity increases with the radiation level and duration. Hence, it is crucial to minimize radiation exposure while ensuring the best image quality.

Medical image quality enhancement typically involves various tasks, including medical image super-resolution, multi-modal medical image fusion, and medical image denoising. Due to its convenience, speed, and cost-effectiveness, this field

has gained significant attention [43, 13, 65, 37]. However, to reduce collection time and minimize radiation dose for patients, the resolution of many medical images is limited, resulting in low-quality images. Additionally, medical images with a single mode cannot fully capture lesion information, and low-resolution images are often inadequate to describe the overall details of a lesion. Fully and clearly describing lesion information is crucial for subsequent medical processes, making medical image enhancement technology vital to bridge the gap between clinical requirements and imaging technology limitations. Thus, processing medical images using enhancement techniques can help doctors make more accurate diagnoses and provide better care for patients.

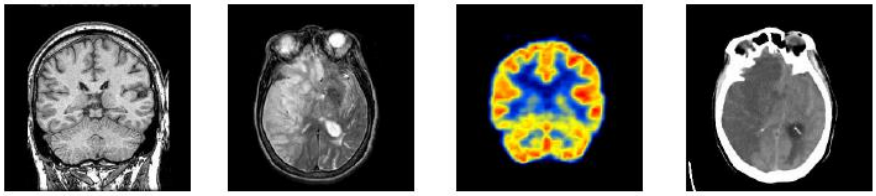

Figure 2.1: Multi-modal medical images of the brain

2.1.1 The development of image quality enhancement technology

Many excellent methods have been proposed for various image enhancement tasks in recent years. Histogram-modified local contrast enhancement was proposed in [160] to adjust the levels of contrast enhancement, which gave the resultant image a strong contrast and improved the local details present in the original image for a more relevant interpretation. He et al. [51] proposed a simple but effective image prior-dark channel prior to removing haze from a single input image. Using this prior to the haze imaging model, researchers can directly estimate the thickness of the haze and recover a high-quality haze-free image. However, like many traditional methods, this method requires a lot of data training and manual parameter setting, which is difficult to implement in practical applications. [52] presented an explicit image filter, which is effective in detail enhancement and fog removal. [203] proposed a 2D/3D symmetric filter to solve the problem of automatic blood vessel detection. However, most of these methods treat the foreground and background indiscriminately, resulting in poor fidelity of the image structure and loss of detailed information.

In practice, obtaining complete pairwise training data for many deep-learning tasks

can be challenging. Still, despite this, numerous breakthroughs in deep learning have been achieved in various research areas. For instance, Liu et al. [109] proposed a simple and effective method for removing image rain based on unpaired learning by analyzing the features of rain maps. The algorithm comprises a semi-supervised learning component and a knowledge distillation component. The semi-supervised portion utilizes a layer separation principle to estimate and reconstruct rain maps. In contrast, a rain direction regularizer is introduced to restrict the estimation network during the semi-supervised learning phase. Meanwhile, Lore et al. [116] accomplished the objectives of low-light enhancement and denoising by training a depth overlay sparse denoising autoencoder with dimming and denoising images. In addition, a novel training scheme was introduced in [79] to overcome dataset dependency in the Noise2Noise (N2N) model [80], which requires paired noisy images. The proposed scheme in [129] consists of a two-stage approach, including self-supervised learning and knowledge distillation, for learning a blind image denoising network from an unpaired set of clean and noisy images. However, this method may not perform well on real noise, which can be more complex than the pixel-independent noise used in their experiments. Moreover, due to the challenges in collecting medical datasets, learning from unpaired data with different imaging modes has become an increasingly popular research direction. The adversarial learning mode of the Generative Adversarial Network (GAN) structure has been applied to unpaired learning, and the Cycle-Consistent Generative Adversarial Network (CycleGAN) model has made a significant breakthrough in this field. This model proposes a framework to capture the unique characteristics of one image domain and translate them into another. However, the Cycle-Consistent Generative Adversarial Network (CycleGAN) method is limited by its separated learning modules for realizing the datasets, and it is challenging to maintain image quality by only restricting the background. To address these issues, Ma et al. [123] introduced the Structure and Illumination Constrained Generative Adversarial Network (StillGAN), a novel generalized bi-directional Generative Adversarial Network (GAN), to improve the quality of medical images. However, it is complex for this model to recognize high and low-quality images and generate images based on a fixed weight perception, which results in a coupling problem between low and high-quality images. Furthermore, the encoded data in Structure and Illumination Constrained Generative Adversarial Network (StillGAN) can lead to serious problems such as image structure confusion and lack of important information, especially when dealing with significant differences in location and shape. Fu et al. [42] presented a novel unsupervised shimmer image enhancement network based on generative adversarial networks and trained with unpaired shimmer and normal light images. The network addresses the issues of color bias and overexposure in shimmer image improvement by constructing

an illumination-aware attention module and introducing a novel identity-invariant loss. However, the challenge of preserving texture detail and harmonizing the background remains. In addition to image enhancement for medical images, unpaired image enhancement technology has also been applied to multi-mode image processing, as demonstrated in [7, 173]. [173] proposed a novel cross-modality image synthesis method that trains on unpaired data, enhancing synthesized images' quality. Despite these advancements, maintaining texture detail and harmonizing the background continue to pose difficulties in this field.

In other image enhancement tasks, several effective image enhancement methods have been proposed for various tasks. For instance, Upadhyay et al. [167] proposed an uncertainty-aware Generative Adversarial Network (GAN) for robust Magnetic Resonance Imaging (MRI) image enhancement, which utilized an adaptive loss function to reduce noise and improve robustness. Additionally, Deep Light Enhancement Generative Adversarial Network (EnlightenGAN) [69] employed the VGG Deep Convolutional Networks (VGG)-based perceptual loss principle to maintain the correlation between low-light input images and normal output images. Despite these efforts, the current image enhancement algorithms face spatial limitations and other challenges, as discussed previously.

2.1.2 The detail description of Generative Adversarial Network (GAN)

Convolutional Neural Network (CNN) in deep learning is a kind of feedforward neural network that is based on the theory of multi-layer perceptron in machine learning and contains convolution operation with sufficient depth structure [194]. Due to the similarity between convolution operation and filter processing in image processing and the powerful feature extraction capability of multi-layer convolution, CNN can replace manual to realize the complexity in traditional image processing, feature extraction pre-processing operation[201].

In the working process of neural networks, it is also necessary to define a loss function to measure the deviation between the output result of the current model and the expected value according to the task handled. The error of the current output can be calculated according to the loss function. Then the network model can obtain the gradient direction of the parameters of each neuron through multistage derivative and adjust the parameters according to the gradient direction. After iteration, the model can get a smaller loss error, that is, closer to the task target. In the field of image processing, Mean Squared Error (MSE) loss is a relatively common loss function, defined as:

$$-\qquad (2.1)$$

where n is the total number of pixels, and represent the pixel value predicted by the model and the expected output pixel value, respectively.

To leverage the advantages of neural networks, a good strategy for information enhancement is to use a generative adversarial network. The Generative Adversarial Network (GAN) refers to the creation of artificial instances from the dataset, which maximize the retention of the characteristics of the original dataset. According to these principles of countermeasures training, many GAN modeling frameworks have been generated. Bowles et al. [15] described GAN as a method of "unlocking" additional information from a dataset. As a generative modeling framework, GAN is significantly ahead of other similar models in terms of computing speed and generation quality.

As the most widely used model in the image generation model, the Generative Adversarial Network (GAN) was first proposed by Ian Goodfellow in 2014. The model comprises two parts: generator (G) and discriminator (D). The generator network is trained to generate new data samples that are similar to a given training dataset, while the discriminator network is trained to distinguish the generated samples from real samples in the training dataset.

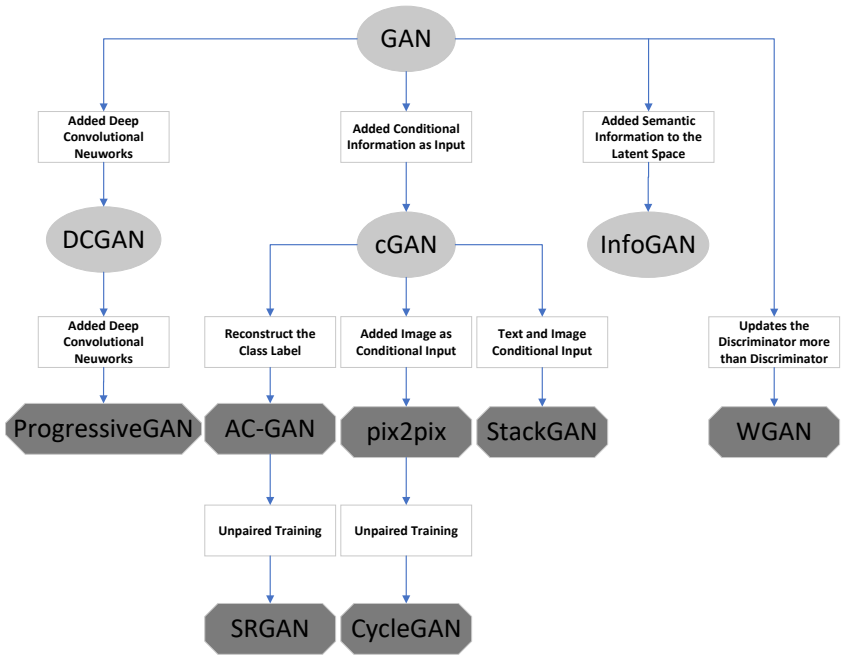

Figure 2.2: The developments of Generative Adversarial Network (GAN) in [67].

During training, the generator network generates new samples and the discriminator network evaluates them. If the discriminator network determines that the samples are fake, the generator network adjusts its parameters to generate more realistic samples, and the process repeats. Over time, the generator network gets better at generating realistic samples, while the discriminator network gets better at detecting fake samples, leading to an equilibrium where the generator network generates high-quality samples that are difficult to distinguish from real samples. The optimized objective of Generative Adversarial Network (GAN) can be described as the following equation:

(2.2)

In the process of training, the discriminator network D functions as a secondary classifier, improving its ability to distinguish between real and generated images with each update. The objective is to correctly label the two types of data and establish an accurate decision boundary between them. The generator network G is updated to produce images that can also be classified as real by the discriminator network, which results in generated images that approach the decision boundary and real images. Through iterative updates, the generated images progressively become more realistic, making it increasingly challenging for the discriminator network to differentiate between real and generated images. This process allows the generator network to fit the real data with a high level of fidelity.

Cycle-Consistent Generative Adversarial Network (CycleGAN) is a variant of Generative Adversarial Network (GAN) specifically designed for image-to-image translation task. It was introduced in 2017 by Jun-Yan Zhu et al. In Cycle-Consistent Generative Adversarial Network (CycleGAN), two generator-discriminator pairs are trained simultaneously, with each pair transforming an image from one domain to another. For example, one generator-discriminator pair might transform a photo of a horse into a painting of a horse, while the other pair transforms the painting back into a photo. The idea behind Cycle-Consistent Generative Adversarial Network (CycleGAN) is to capture the underlying mapping between the two domains and use it to translate images from one domain to another. In Cycle-Consistent Generative Adversarial Network (CycleGAN), and are two different image representations, and the Cycle-Consistent Generative Adversarial Network (CycleGAN) learns the translation and simultaneously. Different from "pix2pix" [63], training data in Cycle-Consistent Generative Adversarial Network (CycleGAN) is unpaired. Thus, they introduce Cycle Consistency to enforce forward-backward consistency which can be considered

as "pseudo" pairs of training data. With the Cycle Consistency, the loss function of Cycle-Consistent Generative Adversarial Network (CycleGAN) is defined as $\mathcal{L}$:

$$\mathcal{L} \qquad \mathcal{L} \qquad \mathcal{L} \tag{2.3}$$

Cycle-Consistent Generative Adversarial Network (CycleGAN)s utilize a loss function to train the generators to generate high-quality images and the discriminator to differentiate between real and generated images. In addition, an extra loss function is applied to guide the generator in preserving the input image content while performing the intended transformation. The outcome is a model capable of visually appealing image translations between different domains while preserving the input image content. The details can be checked in Figure 2.2.

2.2 Image Fusion for Multi-modal Information Enhancement (MIE)

The purpose of the Heterogeneous Image Fusion (HIF) task is to combine the global image and the detailed image obtained from different imaging sensors to generate robust and informative fused (high content) images, which can simultaneously keep the pixel intensity from the global images and the texture information from the detailed images. For the diversity of imaging sensors, Heterogeneous Image Fusion (HIF) is a wide range of topics in natural, medical, and biological image fusion tasks, e.g., infrared and visible image fusion (IVF) [180] [104], Positron Emission Computed Tomography (PET) and Magnetic Resonance Imaging (MRI) image fusion [191] [202], Single-photon Emission Computed Tomography (SPECT) and Magnetic Resonance Imaging (MRI) image fusion [56], and Green Fluorescent Protein (GFP) and Phase Contrast (PC) image fusion [165]. By integrating salient information from source images, the fused image contains more comprehensive information and thus has a better performance in downstream tasks.

As medical image devices continue to develop at a fast pace, the secure storage of medical image data has become crucial for efficient real industrial applications, such as biometric verification and clinical diagnosis. Sensor fusion [23] [10] has rapidly developed for verification and treatments. Schemes such as fingerprint verification [141] and biometric-based efficient medical image watermarking [5] have been successfully applied to real applications. Medical images and biometric data are inseparable, as seen in the use of Magnetic Resonance Imaging (MRI) for biological fingerprint in patient verification [166]. However, efficiently preserving important biological information has become a major challenge. Single modal sources provide limited information and cannot meet the requirements of patient verification, disease diagnosis, health monitoring, surgery, and radiation

therapy [31]. Therefore, it is important to keep data secure and enhance its usage in operative environments. Heterogeneous data has also attracted a lot of attention in many real applications. Fusing different sources into one image is necessary to obtain complementary information to assist aided-diagnosis frameworks [93]. For example, Magnetic Resonance Imaging (MRI) can provide high spatial resolution and depict soft tissue definition [189], while Computed Tomography (CT) captures hard tissue information with little distortion [66]. Positron Emission Computed Tomography (PET) can be used for the staging of uterine cervical cancer [76] and pancreatic cancer detection [150]. Single-photon Emission Computed Tomography (SPECT) reveals clinically significant changes in metabolism. With the growing appeal of image fusion, various fusion algorithms have been developed [188]. Image fusion aims to leverage the dominant features from multi-modal sources and synthesize the *common-unique* content together.

In past research, numerous methods have been proposed to achieve the fusion of heterogeneous images, including multi-scale transform-based methods [90], sparse representation-based methods [103], subspace-based methods [137], saliency-based methods, and hybrid methods [89]. With the emergence of deep learning, data fusion techniques can leverage the adaptive feature fusion from source images and well-designed loss functions [88] [121] [140] [122] [110]. However, current deep learning-based data fusion methods mainly rely on image reconstruction-based methods, which generally follow a similar architecture: feature extraction, feature fusion, and image reconstruction. Recently, autoencoder-based, convolution neural network-based, and generative adversarial network-based methods have been introduced, which benefit from the increasingly mature network architecture. However, the specific network structure of these methods is only applicable in specific scenarios, such as the unmanned driving method proposed in [204]. Therefore, further research is needed to explore the potential of deep learning-based data fusion in diverse scenarios.

2.2.1 The development of image fusion technologies

With the fierce development of computer vision and the fast-growing demand for application requirements, various heterogeneous image fusion methods were proposed. Simply, they can be classified into several categories, including multi-scale transform-based[90, 39], sparse representation-based [103, 72, 35], neural network-based [86, 140], subspace-based [137], saliency-based, hybrid-based [89], and other methods [119].

In this part, we will simply introduce the main content of those methods. Multi-scale transform-based methods have obtained the most effective usage rate in image fusion, which assumes that source images can be decomposed as sub-images

at different scales. Li et al.[90] proposed a method combining two-scale decomposition and weighted average technique. The large scale is the base layer of intensity transformation, and the small scale is the detail layer that captures information. In[39], the proposed algorithm adopted different fusion strategies for different subbands and can adaptively find the balance point between the grayscale image and diverse backgrounds in the image fusion process. Sparse representation-based methods [103, 72] focus on learning an over-complete dictionary with the guidance of massive high-quality image pairs. Since the data can be represented with a linear combination of elements sparse in the over-complete dictionary, it is a reasonable way to get satisfactory results.

To deal with the multi-modal information interactions, various strategies have developed with success from many perspectives. The Sparse Representation (SR) has attracted much attention from the natural sparsity of signals in medical image fusion. The aim is to learn the sparse coefficients based on a pre-trained dictionary with intrinsic features to approximate fine details. Thus, the SR with multi-scale transformation [113] and Sparse Representation (SR) with pulse coupled neural network [190] were constructed for mitigating the fixed feature extraction capacity. Later, Wang et al. [114] and Liu et al. [111] respectively conducted the fusion of medical images based on Adaptive Sparse Representation (ASR) and Convolutional Sparse Representation (CSR). The Convolutional Sparse Representation (CSR) model overcomes the shortcomings of limited ability in detail preservation and high sensitivity to misregistration of the SR model. In contrast, Jiang et al. [70] proposed a novel multi-component SR-based fusion method via Morphological Component Analysis (MCA) [158], which can obtain the sparse representations of cartoon and texture components of each source image. This component separation process can significantly improve the flexibility for designing more effective fusion strategies. Sadly, it would bring a significant amount of noise in. The Convolutional Sparse Representation Morphological Component Analysis (CS-MCA) model [112] integrated the advantages of Morphological Component Analysis (MCA) and Convolutional Sparse Representation (CSR), achieving a multi-component and global sparse representation of the source image. Joint SR proposed in [75] formed the dictionary from various modalities. Furthermore, the group sparsity representation [92], low rank prior [84] [85], and various Sparse Representation (SR) extensions have been proposed to address the medical image fusion problem. However, the over-smoothed issues inevitably would lead to color distortion and weaken the root of Sparse Representation (SR) based medical image fusion. Our analysis indicates that feature extraction with redundancy removal for preserving fine details remains a critical challenge.

In traditional spatial and transform domain methods, image decomposition, and

reconstruction are the two main processes designed. However, as previously summarized, feature extraction can also be conducted through multi-scale transformation or multi-scale geometric analysis and through several transformation-based and pyramid-structured methods. For instance, image decomposition lets the Discrete Wavelet Transformation (DWT) separate high-frequency from low-frequency information. Nonetheless, these decomposition methods suffer from directional feature distortion and shift invariance properties. To solve the degeneration of multiple scale features, popular representative modeling examples include the non-subsampling paradigm, such as the Non-subsampling Contour Transformation (NSCT) [12] and Non-subsampling Shearlet Transformation (NSST) [191]. In addition, hybrid schemes have also been investigated to enhance the feature extraction process, such as different extensions of the pulse-coupled neural network [32] and several improvements [191].

Neural network-based methods [86, 140] aimed to mimic the human brain's processing of neural information with their strong adaptability and anti-noise capacity. Subspace-based methods [137] aimed to generate low-dimensional subspaces from high-dimensional source images, considering that redundant features can interfere with feature extraction. Proper subspace-based methods can speed up image processing speed and accuracy. Saliency-based methods focus more on the essential objects or pixels rather than their edges or neighbors, which can improve the visual quality of the fused results by highlighting the intensities of the salient objects. Hybrid-based methods [89] focus on combining the advantages of the previously mentioned methods to improve the performance of the fusion strategy. Other fusion strategies, such as the method based on total variation [119], can also bring inspiration to the field of infrared and visible image fusion.

2.2.2 The detailed description of computational method

- Since the pioneering work of [184], a number of Sparse Representation (SR) based image fusion methods have been proposed. Typically, Sparse Representation (SR) based fusion is performed patch-wise by dividing the source image into several patches of the same size and applying image fusion at the patch level. A sliding window with a fixed number of pixels is used to select image patches to reduce block artifacts and improve the robustness of erroneous registrations. Consequently, the source image is divided into overlapping patches, and the standard sparse coding model [126] is independently applied to each patch. Mathematically, the applied SR model can be expressed as:

(2.4)

where $\mathcal{R}$ means a stacked vector version of an image patch of size $\bar{\ }$ $\_$. $\mathcal{R}$ means an over-complete dictionary. $\mathcal{R}$ is the sparse vector to be calculated, and the sparsity is measured by its -norm, which counts the number of non-zero entries. represents the tolerance of reconstruction error. The Orthogonal Matching Pursuit (OMP) algorithm is employed to solve this optimization problem. Later, [185] applied the Simultaneous Orthogonal Matching Pursuit (SOMP) algorithm to improve the fusion method, which can ensure that identical dictionary atoms decompose the source image patches at the same location. The target image is finally obtained for these patch-based SR methods by aggregating all reconstructed patches and averaging the overlapping pixels.

- Definition for the Cartoon and Texture Components. Here we provide a brief explanation of the cartoon and texture components. The theory was initially put forward from [125] and states that an image can be decomposed into a cartoon image and a texture map. The cartoon image captures the salient features, including piecewise smooth changes in illumination and edges, while the texture map provides detailed texture information within regions bounded by these edges. However, in some cases, such as during dictionary filter learning, the source medical images may not be well decomposed into the cartoon and texture components. This phenomenon is known as component entanglement. For example, as shown in Figure 2.3, the texture map obtained by Convolutional Sparse Representation Morphological Component Analysis (CS-MCA) suffers from component entanglement with the cartoon image, resulting in poor-quality information. In contrast, our method can derive a texture map rich in texture information and complements the cartoon image.

- means the Fourier transform of . and separately represent the transpose and conjugate transpose of . means convolution such as . means matrix dot product. means matrix multiplication and is an linear projection operator. is the vectorization of . means the Kronecker product. We also present the usage of the subscripts here. For a three-dimensional matrix like $\mathcal{R}$, unfold the matrix in the first dimension. Then means the -th matrix in the sequence, namely . The subscript has the same meaning as . In the beginning, we also add the subscript and to distinguish different components.

- Component-wise Fusion Method Revisited. Combining the standard Sparse Representation (SR)[184], Morphological Component Analysis (MCA)[68] and Convolutional Sparse Representation (CSR) model[108], the Convo-

lutional Sparse Representation Morphological Component Analysis (CS-MCA) model in [112]is defined as

–

(2.5)

where $\mathcal{R}$ is an entire image. and denote two sets of dictionary filters for the SR of the cartoon and texture components, respectively. and are the corresponding sparse coefficient maps. The dimensions of the dictionary filters are $\mathcal{R}$ and the coefficient maps are $\mathcal{R}$. are model coefficients. It can be seen from (2.5) that the proposed model promotes SR-based image fusion by combining two different methods. We can find from Figure 2.3. Source image:(a1-a2). Fused image obtained by CS-MCA: (fused image: b1; the texture map: c1; the cartoon image: d1). Fused image obtained by our proposed method: (fused image: b2; the texture map: c2; the cartoon image: d2). The CS-MCA model also combines Convolutional Sparse Coding (CSC) [196] to simultaneously implement a global sparse representation of two components and source images to overcome the shortcomings of the previous two methods.

2.3 Image Analysis for Task-driven Information Enhancement (TIE)

2.3.1 The development of information enhancement for image segmentation

In this chapter, we discuss two topics of task-specific information enhancement object segmentation and object tracking, which include the core theme of "Enhancement" and how to embed the enhancement knowledge into the learning process. Firstly, we discuss the topic of segmentation. CT is a valuable tool for providing pathological information about dense structures in the human body, including bones and organs. It plays a critical role in diagnosing and treating renal tumors. However, the automatic clinical diagnosis of tumors remains challenging due to their various shapes, sizes, and fuzzy textures. Liver tumor segmentation technology [136], for example, uses re-segmentation to separate the liver and other organs from abdominal CT and obtain tumor images. This can help doctors accurately

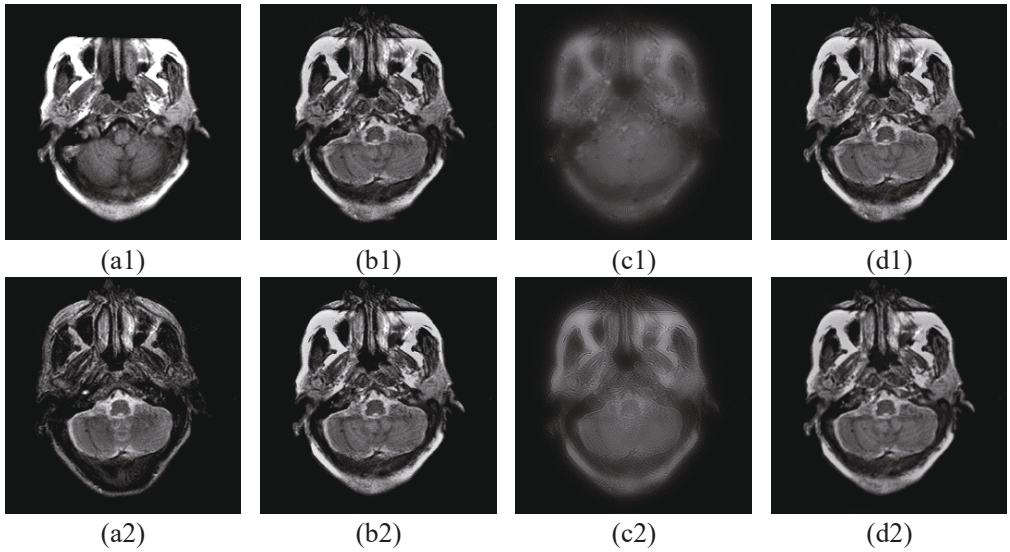

Figure 2.3: Examples of the cartoon and texture decomposition.

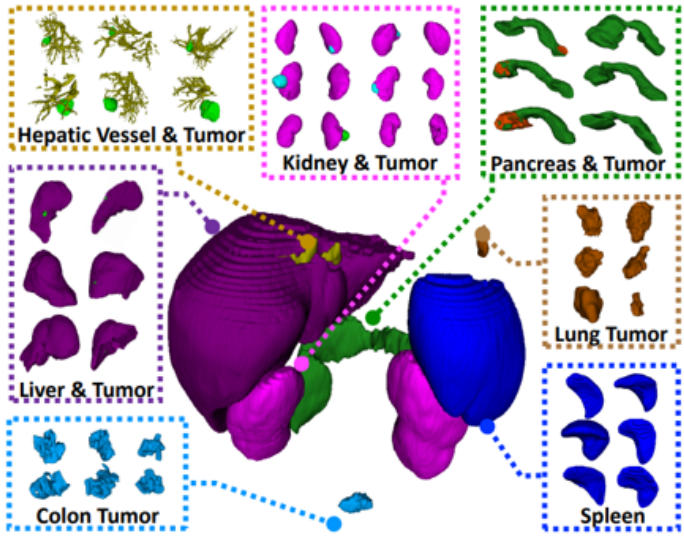

Figure 2.4: Examples of the partially labelled datase from [197].

evaluate the development of primary or secondary tumors and quickly formulate treatment plans [170, 174, 117]. Deep learning has revolutionized automatic data-driven medical image segmentation [153, 154, 199], benefiting from the powerful ability of neural network models to fit data. For instance, Andriy et al. proposed a segmentation network composed of an encoder-decoder structure that can reliably and effectively segment kidneys and kidney tumors from abdominal 3D CT scan in the arterial phase [132]. However, many segmentation methods can only segment specific organs and tumors, making them difficult to adopt for other segmentation tasks and wasting computing resources. Manual labeling of multiple organ medical images is limited by manpower and resources, making it intractable to obtain a self-contained dataset for training. In real scenarios, only one kind of tissue or organ is typically labeled, resulting in a partially labeled dataset. Most benchmark datasets are dedicated to specific organs, so the segmentation model becomes inefficient and inflexible. Consequently, the segmentation task of multiple organs and tumors on partially labeled datasets has become a crucial issue in computer-aided diagnosis [60].

Traditional model-based segmentation techniques enjoy the theoretical guarantee of the segmentation process while suffering from the fixed operator and the non-adaptive segmented rules [49] [50]. Many methods based on deep CNN had been proposed for MOTs segmentation [54, 95, 209]. Most methods have trained multiple independent networks for different targets (like only for the liver or kidney). The deep learning framework NNU-Net proposed by Fabian et al. [62] can independently make critical decisions required to convert the basic architecture to different data sets and segmented tasks without manual adjustment. A cascade trainable segmentation model proposed by Yu et al. [192] captured the global and local appearance information from crossbar patches. Zhang et al. [198] proposed a lightweight hybrid convolutional network segmentation method for liver and tumor within CT volume, using the codec structure and depth and space-time separation (DSTS) technology, which effectively reduced the complexity of the model. While these methods take various measures to reduce the complexity, the computational complexity of their models is still a challenge and can not be ignored. Fang et al. [38] proposed a new training strategy, which enabled the multi-scale depth neural network to be trained on multiple partially labeled datasets through a shared encoder and significantly reduced the computational complexity of the model. Chen et al. [20] and Shi et al. [155] adopted a similar multi-head network to solve this multiple partially dataset problem. Although these methods achieve impressive performance, they are short of dealing with new tasks. Most methods [38, 20, 155] [205] only relied on a shared backbone network to realize the common knowledge of several inputs and multiple output headers for different segmentation tasks. However, the mode assumption of multi-class segmentation

of partially labeled data may mislead some unlabeled organs as the background. The work of Zhang et al. [197] was a single input head network and a single output head segmentation head. Specifically, it adopted a dynamic segmentation head to solve the problem of partial labeling and can simultaneously segment organs and tumors to overcome the above problem. However, some inevitable noise from the dataset and network training based on conditioning class label information can influence the performance of feature extraction and organ segmentation. This problem explicitly results from the intra-domain interference of MOTs task.

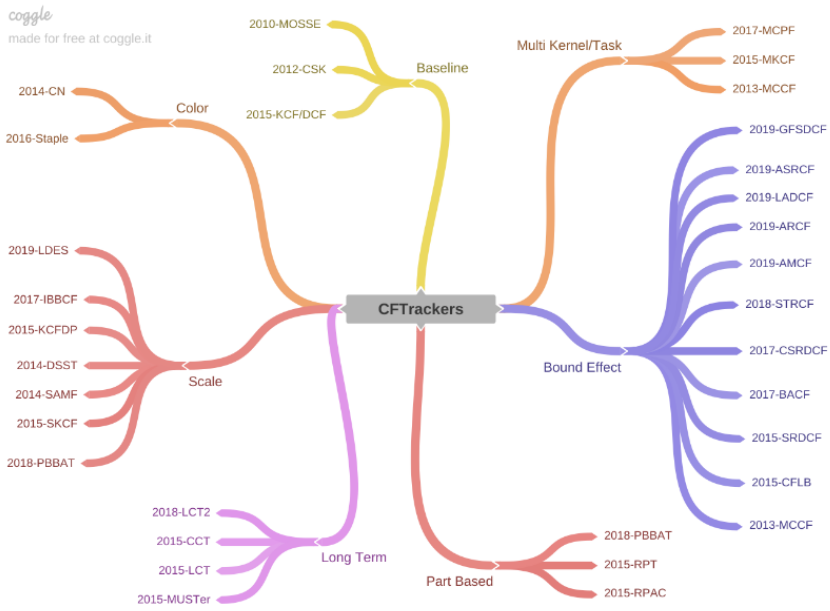

Figure 2.5: Examples of the Correlation filter tracker.

2.3.2 The development of information enhancement for object tracking

Secondly, we discuss the topic of tracking, the development[1] can be checked in Figure 2.5. Object tracking in the traffic domain is how to adapt to fast changes in the target's appearance. Even if the initial frame of an unknown scene is given, the central performance of predicting the target state of each frame will be limited by several appearance variants. Moreover, because the traditional target tracking background is fixed, the tracker only considers the problem of the target itself in the tracking process. However, due to the movement of traffic vehicles, the tracking process needs to take into account the complex scene variants and unpredicted

[1] https://github.com/HonglinChu/CFTrackers

interference. The characteristic of target tracking is that both the target and the background are in motion, which is hard to solve the difficulty of target tracking in traffic scenarios and achieve an excellent, intelligent traffic video monitoring effect. In addition, target tracking has brought more significant challenges due to mechanical vibration, object motion, target occlusion, and background clutters [41].

Benefiting from its easy implementation and fast prediction of Discriminative Correlation Filter (DCF), DCF has attracted a lot of attention in Unmanned Aerial Vehicle (UAV) tracking. Until now, there are three main research directions in DCF tracking: *spatial regularization, temporal smoothing,* and *robust feature representation*. To solve the first problem, spatial regularization DCF: SRDCF[7] is proposed based on the spatial penalty. This work also inspired other research work on spatial regularization[30, 82]. In [199], they offered a new DCF tracker by suppressing the constraint of spatial boundary effect with spatial feature selection. Moreover, the spatial reliability enhanced DCF[41] [118] had proposed to indicate the reliability of background. However, these methods do not adaptively depress the background and consider the temporal information. To solve the second problem, a temporal regularization is introduced by [82] and [99] to realize the joint spatial-temporal solution and obtain better performance. For the third question, with the development of robustness image feature extraction method on deep neural network, the performance of DCF-based trackers have significantly improved performance and solved the problem to some extent.

Recently, to combine temporal information, some latest models used a transformer to combine spatial and temporal information. STARK [183] had not used any proposals, anchors, and post-processing steps (such as cosine window or bounding box regression), which greatly simplified the visual tracking model. [21] developed a feature fusion network based on a self-context augmentation module with self-attention and a cross-feature augmentation module with cross-attention. Compared with correlation-based feature fusion, self-attention-based methods adaptively focus on useful information, such as edges and similar objects, and establish associations between distant features, enabling the tracker to obtain better classification and regression results. However, the response of redundant information in the global response will affect the accuracy. AutoTrack[99] automatically updated the hyper-parameters to accommodate the change of each frame with the global response. To achieve better performance, the spatial constraints with content-aware [47] and bilateral regression ranking model [206] and other different hybrid response mining [81] [61] based methods had been proposed.

While online learning of tracking has made good progress, there are still many problems in the temporal-based tracking framework. These existing methods only

discover the reliability of spatial or temporal or background or response; the reliability of the temporal knowledge transfer should also be investigated to avoid temporal degeneration. In existing temporal knowledge transfer based on the DCF tracking framework, euclidean distance is commonly used to measure the similarity of the targets of the two adjacent filters within a closed appearance [99, 181, 82].

Here, we recall a new concept about online temporal learning in visual tracking (*probability measurement*). This problem is unnoticed by the above methods. Does it also raise some questions: *what can we measure in online learning: probabilistic temporal fitting or direct temporal interpolation?* In a temporal-based framework, most methods assume that the target context between two frames is a component with minor changes. The change of two adjacent target distributions can be kept only by interpolation. However, this is not easy to appear in reality. In DCF tracking, it is evident that there is no such assumption that the tracked target has apparent occlusion or deformation. The noise drifting in the temporal domain will inevitably lead to tracking distortion. A more reasonable solution is to replace the measurement or transformation here. Regardless of temporal regularization or response mining methods, they are all looking for a transformation such that the representation of the updated frame is matched with historical information. The well-known transformation class can be expressed in [98], in which the Kullback Leibler divergence was used in a deep neural network for visual tracking [28]. However, no closed-form solution can express the similarity measurement. Moreover, the distorted appearance of the tracking in DCF challenges the spatial or temporal-based DCF methods. The above discussion motivates us to mitigate the problem of overfitting and omit the impact of unpredicted appearance. Fortunately, the Wasserstein distance with a common Lagrangian formulation alleviates the need for a common space. In [200], they proposed a novel approach to learning domain invariant feature representations. Wasserstein generative adversarial network (GAN) [6] learned a more reasonable and efficient approximation method and cured the main training problem of GAN.

2.3.3 Revisited Temporal Response Based Method

We first review the basic algorithm, which is conducted on the temporal response-based method [99]. Their objective function for joint optimization of the filter, as well as temporal regularization term, can be written as:

– –

(2.6)

— –

where is the extracted feature in frame , denotes number of channel, is the desired Gaussian-shaped response. respectively denote the filter of the -th channel trained in the -th and (-1)-th frame, indicates the convolution operator. The parameter is the local response, denoted as the reference parameter, , and optimized temporal regularization weighting parameter. The parameter represents the global response for automatic spatial regularization calculated by Equ. (2.7).

(2.7)

is used to crop the central part of the filter where the object is located. is a constant to adjust the weight of local response variations, and is inherited from spatial-temporal regularized correlation filter [83] to mitigate boundary effects, is the local variation vector.

Although many temporal response mining methods [99, 181, 82] have achieved good performance, the redundant information in the global response will lead to errors in the update of . Here, we adopt low-rank processing to avoid the interference of irrelevant information. In addition, when the target is deformed, the influence of temporal noise drift makes the target tracking inaccurate and will continue to affect the follow-up tracking.

Chapter 3

Summary of Included Publications

This chapter will summarize the published papers included in this thesis. These papers are published in peer-reviewed professional and academic international venues in medical image processing, computer vision, and intelligent algorithms. It consists of six research papers. Each paper is elaborated following an MFR format: Motivation, Formulation, and Result. Full versions of the research papers are given in Part II of this thesis.

3.1 Paper I: SSP-Net: A Siamese-based Structure-Preserving Generative Adversarial Network for Unpaired Medical Image Enhancement [176]

3.1.1 Abstract

Recently, unpaired medical image enhancement is one of the important topics in medical research. Although deep learning-based methods have achieved remarkable success in medical image enhancement, such methods face the challenge of low-quality training sets and the lack of a large amount of data for paired training data. In this paper, a dual input mechanism image enhancement method based on a Siamese structure (SSP-Net) is proposed, which takes into account the structure of target highlight (texture enhancement) and background balance (consistent background contrast) from unpaired low-quality and high-quality medical images. Furthermore, the proposed method introduces the mechanism of the generative adversarial network to achieve structure-preserving enhancement by jointly iterating

adversarial learning. Experiments comprehensively illustrate the performance in unpaired image enhancement of the proposed SSP-Net compared with other state-of-the-art techniques.

3.1.2 Motivation

Siamese networks [96] was originally proposed to deal with the classification problem, and it adopted the two-channel network with shared weights to measure the local distribution of the network. Not only is the classification label information considered, but also the local spatial distribution information between samples is achieved. It definitely helps for a small sample size task for classification performance [77]. Li et al. [94] presented a discriminative self-attentive recurrent generative adversarial network, based on a recurrent GAN architecture, to address the super-resolution issue of natural images. This network used unpaired samples to train both degraded and reconstructed networks. Contextual data was collected using a self-attentive method to reduce detail degradation. Bertinetto et al. [11] introduced a Siamese network into a target tracking task, which greatly improved the sample limitations of online learning over traditional tracking methods. All these work well to demonstrate the strong feature generalization capacity of a siamese network structure, which can solve the situation with mismatched training data.

Here, we tailor the same property to the medical image enhancement task. In this study, we employ a Siamese structure for training unpaired data, allowing the network to learn important features from high-quality (HQ) images and preserving the structural details of low-quality (LQ) images, such as target highlight (texture enhancement) and background balance (consistent background contrast). To learn the texture-preserving representation and enhance the visual quality, we introduce two generators for learning a random pair (LQ and HQ). Our network structure is similar to the work in [58] which is used for face hallucination. Instead of relying on faulty paired labeling and the same identity, our random input pair aims to learn a common and salient distribution between LQ and HQ medical images, as these prerequisites are not necessary for our input pair. The HQ image can be treated as prior knowledge to guide the enhancement of the LQ image. Overall, our model combines the siamese structure with a GAN, using shared weights to produce high-quality enhancement results in an adversarial manner, which ensures the robustness of the model to texture blur, structure weakening, and background noise.

3.1.3 Methods

In this paper, we propose a Siamese-based structure-preserving network, named SSP-Net, for corneal confocal microscopy image enhancement to handle the chal-

lenge of the deficiency of the paired LQ images and the HQ images. In this section, we introduce the details of our proposed SSP-Net, including an overview of the Siamese-based generative network structure and loss function.

The framework of the proposed SSP-Net is shown in Fig. 3.1, where the unpaired LQ images and the HQ images are simultaneously used as input into a Siamese-based GAN and the corresponding enhanced outputs are discriminated by a discriminator with the original LQ and HQ input in an adversarial manner. Our SSP-Net adopts two generators, which share the same structure and

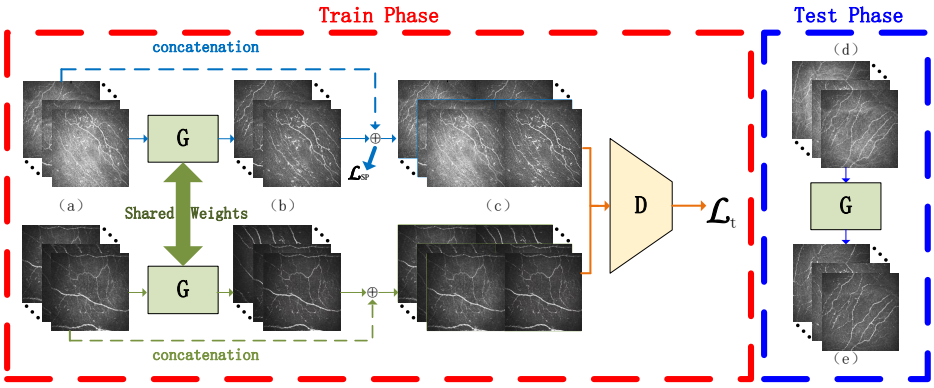

Figure 3.1: The framework of the proposed SSP-Net.

parameters, and one discriminator, where the discriminator is implemented by a PatchGAN [63] structure, which discriminates one image by its patch rather than the whole image, with a local illumination-sensitive constraint. The backbone of the generator is a U-Net-like structure, where there exist skip connections between the encoder and the decoder. Furthermore, the U-Net-like encoder-decoder module in the generator is not a symmetrical structure, where the backbone of the encoder module is a modified ResNet50 [53], where the fully-connected (FC) layer is removed and the max-pooling operation is replaced by a stride convolution for a more powerful texture preservation ability. Moreover, the decoder module is composed of five convolution layers in case of local illumination variations. The discriminator is composed of three convolution layers with the kernel size of , whose initialization follows a normal distribution with the parameter of .

3.1.4 Result

All the training images are initially resized by with a series of data augmentation strategies [123], e.g., random flipping, and color space transformation. Adam optimizer is applied to train the network with momentum terms

for the generator and for the discriminator, whose learning rate is initialized as and linearly warmed up to . When reaching 500 iterations, the learning rate will be halved every 15 epochs. The batch size and the related parameters , , are set as 4, 0.6, 0.4, and 5.1. All the experiments are implemented in PyTorch [138] on two RTX2080TI GPUs.

Four no-reference evaluation metrics are provided to comprehensively illustrate the superiority of our proposed SSP-Net, which are entropy, average gradient (AvG) [64], natural image quality evaluator (NIQE) [127], and perception-based image quality evaluator (PIQE) [169]. Entropy and AvG measure the amount of information and detail textures contained in the enhanced images, respectively, where a larger value indicates a better performance. Nevertheless, NIQE and PIQE correspond to the natural quality and perception-based feature quality modeled by the multivariate Gaussian model, where a lower score means better-reconstructed quality. We employ the area under the ROC curve (AUC) [59], accuracy (ACC) [59], sensitivity (SEN) [44], specificity (SPE) [44], false discovery rate (FDR) [24], dice coefficient (Dice) [73], G-Mean score (G-Mean) [9], and Kappa coefficient [78] as the evaluation metrics for segmentation task. A higher value denotes better performance.

Our SSP-NET model achieved the best results on all metrics, especially AvG and PIQE. The ideal result of AvG and Entropy shows that our structure-maintaining network accomplishes the goal of image enhancement well, extracts the detailed features of the original, and highlights the performance, making the enhanced image extremely fidelity to the texture structure. Compared with the StillGAN method with a non paired learning model, our values of NIQE and PIQE are very low, which ensures the reconstruction performance of SSP-Net.

3.2 Paper II: Disentangled Spatial-Transformation Guided GAN for Unpaired Medical Image Quality Enhancement

3.2.1 Abstract

Deep learning-based medical image enhancement has received significant research attention recently. Most of the existing methods fall into the supervised learning framework by synthetic training data. However, they are short of generalizing whole medical image visual appearances due to the gap between the simulation and practicability. Though cycle-consistent generative adversarial network (CycleGAN) has achieved great progress with an unsupervised framework to deal with the unpaired medical image enhancement problem, most CycleGANs only focus on the global appearance to perceive the low-quality and high-quality data in the same feature extraction network to achieve unpaired learning. The information

entanglement caused by here ignoring the specific image characteristics between different quality images leads to the degradation of enhancement performance. This paper introduces a new perspective of disentangled extraction based on CycleGAN (DSSGAN), which can provide pixel-level supervision to preserve texture and detail information for image quality enhancement. As a result, we propose a disentangled generator structure to enhance images of different qualities better perceptually. Furthermore, to model the spatial transformation in unpaired learning, the spatial transformation module is used to capture the spatial and structural features as supervised information between high-quality and low-quality images, thus fusing with the encoded information to capture the more accurate feature information. The experimental results in three datasets show that our proposed method has promising performance compared to other state-of-the-art algorithms.

3.2.2 Motivation

Due to the encoder-decoder structure adopted by the traditional generator [208], the unpaired input images of network training are randomly shuffled from low-domain and high-domain image sources. Moreover, for medical images, it is difficult to obtain such large-scale low/high-quality image pairs in real scenarios for training. Therefore, it is too complex to recognize the specific information between high quality and low quality by only relying on the same encoder of the generator. Thus, it is prone to over-fitting and leads to the entanglement problem in high-dimensional feature representation in terms of the *training scale, different structure, texture, and illumination intensity*. Concerning two novel problems here for the unpaired image enhancement task:

- Q.A: What information is extracted between low-quality and high-quality domains?
- Q.B: How to properly investigate information from high quality to low quality?

Referring to the Q.A., the most relevant medical image enhancement works treat this problem as image-to-image translation [106]. They lead to learning the shared representation with low-level image properties, such as texture or cartoon. Recently, one method is to learn the relations between the two domains with independent autoencoders for the two domains, but the existing methods always generate image-to-image translation-based enhancement according to shared weight perception [105]. Lin et al. [102] proposed the image-level disentanglement and instance-level disentanglement to learn domain-invariant representation for generalizable object detection. Motivated by [142], the simple two-pathway encoder and a single decoder for image content transfer. We propose the disentanglement

representation framework to preserve the informative features for medical image enhancement.

Referring to the Q.B, the usage of the extracted information is usually followed the guided methodology in a multi-layer encoder-decoder manner. However, the encoded data is significantly different due to the unpaired data and degrades the spatial information in medical images. As a result, the model will produce aliasing and chaos when encoding the information of different images, leading to blurred image structure, and disordered illumination distribution [195, 156]. Actually, unpaired input is not prone to achieve optimal in the traditional CycleGAN proved by [123]. The image-to-image translation assumption is inevitably affected by illumination, noises, and other variants. Nevertheless, the existing deep learning image-to-image translation-based enhancement techniques avoid the difficulties of ideal medical image data collection. Due to the effects of heterogeneity and complex illumination conditions of natural images, the content of medical images is always homogeneous. Rather than the existing natural images unsupervised enhancement techniques[133], we prefer to use the above good property to solve the medical image enhancement. Moreover, we observe that the spatial homogeneity similarity across different quality medical images. Thus, we specially design a different module to perceive this specific information between high-quality and low-quality for medical image enhancement.

3.2.3 Methods

The framework of the proposed DSSGAN is shown in Fig.3.2. The main framework of our model is the CycleGAN structure, which contains two groups of GAN, namely Generator1/Discriminator1 (G1/D1) and Generator2/Discriminator2 (G2/D2). Where G1/D1 deals with the mapping and conversion of low-quality images to high-quality images, G2/D2 is the opposite.

Take G1/D1 as an example (G2/D2 is similar), our generator consists of three modules: encoder, STN, and decoder. The input low-quality image ( ) generates the corresponding predicted high-quality image ( ) through G1 and identifies the true and false through D1. Then, the generated image enters G2 to generate the corresponding secondary predicted low-quality return image ( ). Thus, a set of image cycle consistency of L-H-L (low domain to high domain back to low domain) is formed. Specifically, the corresponding simple coding information of the input image is obtained through the encoder and fused with the supervision information of the space, structure, and other features of the input image extracted by the decoupled STN module to obtain adaptive high-dimensional features. Then, the corresponding generated image is obtained through the decoder module.

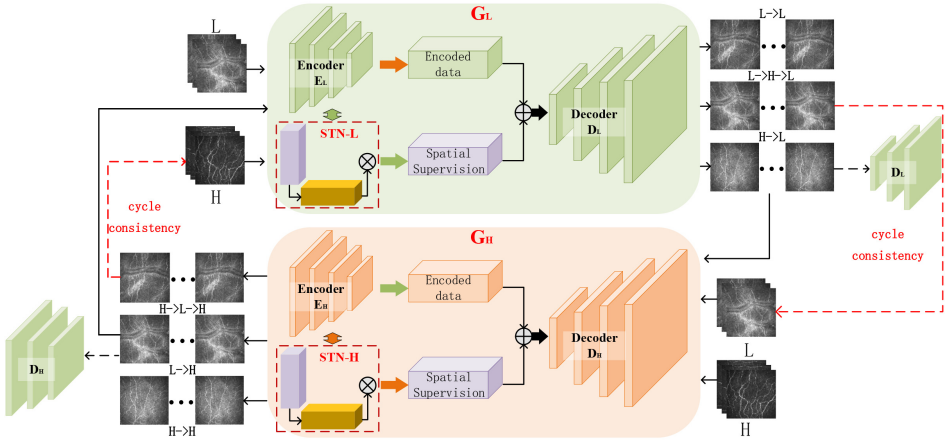

Figure 3.2: The proposed DSSGAN framework consists of two groups of GANs.

The generator with disentangled structure can differently learn and generate different types of images. At the same time, the matching of the traditional discriminator module and the generator module can form a complete process of confrontation learning. Furthermore, goes into G1 to generate in order to make G1 have a one-way generation from L to H.

3.2.4 Result

The public CVC-EndoSceneStrill dataset [168], the corneal confocal microscope (CCM) dataset[123], and Whole-slide images (WSI) from the Genome Data Sharing Data Portal [186] are used for the experiment. For evaluation, we adopt 5 no-reference evaluation metrics, namely Entropy [149], Average Gradient (AvG) [25], Natural Image Quality Evaluator (NIQE) [128], Perception-based Image Quality Evaluator (PIQE) [169], and Blind/Referenceless Image Spatial Quality Evaluator (Brisque) [143]. Five methods are selected in comparison with the proposed DSSGAN, which consist of two traditional methods DCP [51] and BM3D [26], and three deep learning methods EnlightenGAN [69], MSG [195] and StillGAN [123]. Moreover, the aforementioned learning-based methods are all retrained in the same dataset with the proposed DSSGAN. Our DSSGAN achieves the best results on all metrics except Entropy, especially PIQE, and NIQE.

3.3 Paper III: FCFusion: Fractal Component-wise Modeling with Group Sparsity for Medical Image Fusion

3.3.1 Abstract

Multimodal image fusion combes relevant biological information that can be used for automated industrial applications. This paper presents a novel framework combining fractal constraint with group sparsity to achieve optimal fusion quality. Firstly, we adopt the idea of patch division and component-wise separation to perceive the fractal characteristics across multi-modality sources. Then, group sparsity is proposed to preserve the spatial information against the redundancy of component entanglement. A dual variable weighting rule is inherently embedded to mitigate overfitting across the component penalty. Furthermore, the Alternating Direction Method of Multipliers (ADMM) is conducted for the proposed model optimization. Experiments show that our model performs better in quantitative visual quality and qualitative evaluation analysis. Finally, a real segmentation application of PET/CT image fusion proves the effectiveness of our algorithm.

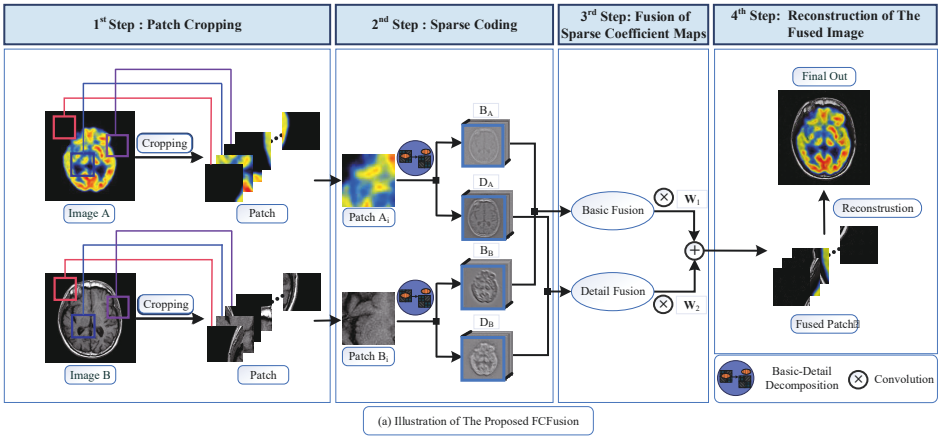

Figure 3.3: The image fusion process of our FCFusion model.

3.3.2 Motivation

As discussed earlier, feature extraction-driven methods have demonstrated that this explicit fusion guideline (extensions of feature extraction capacity only) leads to several issues for medical image fusion. The major problems of these algorithms are that the information is over-completed. Thus, adapting the appropriate balance between feature extraction and redundancy removal is the main solution. In this paper, the fusion quality is determined by the feature extraction and redundancy

removal in our proposed fractal component-wise prior and group sparsity model termed FCFusion. The main flowchart algorithm is shown in Figure 3.3. In general, we design the patch division with component-wise separation to perceive the fractal characteristics across the different components in multi-modality. Unlike traditional patch sparse representation (SR) based image fusion [184, 75] in sliding windows manner, they performed feature extraction directly for each patch. Our motivation relies on the proposed fractal constraint for feature extraction. To keep up with the redundancy removal for mitigating the over-smoothing problem, preserving the characteristic information by a group sparsity model is exploited in our proposed model. Unlike the model proposed in [92], to better promote detail preservation and remove redundancy, we use a fractal variable weighting coefficient strategy to select the features of each patch over the decomposed components. The saliency is reflected in the group-weighted sparse coefficient here for medical image fusion to achieve a few artifacts. Overall, our designed patch-level component feature extraction and group sparsity mainly focus on how to avoid over-smoothing from noise interference, color distortion, and artifacts.

3.3.3 Methods

It includes four steps: patch cropping, sparse coding, a fusion of sparse coefficient maps, and reconstruction. In step 2, the source images are divided into two components: basic B and detail D, and then they are processed in a parallel way. A new medical image fusion algorithm based on "fractals" is proposed, which intuitively imposes the patch-level component-wise separation to perceive the fractal characteristic across the different components in multi-modality sources. A new strategy of group sparsity for components is proposed to strengthen the detail preservation for medical image fusion, and the dual variable weighting is utilized to mitigate over-smoothing and remove redundancy for characterizing the detailed structure and fine components.

3.3.4 Result

Experiments are carried out by MATLAB R2016b on a computer with Dual-Core Intel Core i5 processor (1.8GHz) and 8GB 1600 MHz DDR3. For a fair comparison, for both basic B and detail D, we adopt the same fusion strategy as the convolutional sparse representation (CSR) based model[112]. We experimentally fix , , , , the patch size is and the iteration number is 3. We select seven metrics including Average Gradient (AG), Correlation Coefficient (CC), Entropy (EN), Mean Square Error (MSE), Root Mean Squared Error (RMSE), Mutual Information (MI), Spatial Frequency (SF) in [147]. We also calculate the average rank of these seven indicators, which is an F-rank. Our experiments use five different modalities of fusion (mag-

netic resonance imaging (MRI), positron emission tomography (PET), computed tomography (CT), and single photon emission computed tomography (SPECT) images), including MRI-CT, MRI-PET, MRI-SPECT, and T1-T2, T2-PD, where T1, T2, and PD are MRI images based on different weights. The source images used in the experiment are all from the Whole Brain Atlas[71] established by Harvard Medical School.

Our FCFusion model compares with seven existing medical image fusion methods, including NSST-PAPCNN[191], NSCT-PCDC[12], NSCT-RPCNN[32], GFF[91], CS-MCA[112], LP-SR[113] and IFCNN[202], where IFCNN is a deep learning based fusion method. Note that in the fusion of color images, the CS-MCA method is not included in the comparison. All parameters in these methods are set to the default values for unbiased comparison.

3.4 Paper IV: JADD-GAN: A Joint Attention Generative Adversarial Data Fusion Network for Object Detection and Tracking

3.4.1 Abstract

Image fusion is the fusion of images captured by different sensors to generate a single image with enhanced information, and fusion technology, as one of the important branches in the field of information fusion, mainly realizes the processing of multi-source image information. However, many commonly used fusion methods usually ignore the fused images' visual naturalness and information fidelity and lack emphasis on the salient information, making the fused images unsuitable for human visual perception. To address these shortcomings of existing methods, in this paper, we propose the Joint Attention and Dual Discriminator Generative Adversarial Data Fusion Network JADD-GAN. In the generator module, we first adopt a dual encoder structure and give information fusion in the decoder part to increase the extraction of multi-level information by the network. Second, different discriminators are used for infrared and visible images in order to highlight the thermal radiation information and key textures. The effectiveness of the method is verified by experiments on four datasets, and the results show that the method can effectively highlight the thermal radiation information and key texture details of the fused images, fully demonstrating its great potential and performance in solving the infrared and visible image fusion (IVF) problem.

3.4.2 Motivation

In order to increase the extraction of multi-level information by the network and to accurately reflect the salient features of the source images, in the generator part we use the trained network to double encode the infrared and visible images sep-

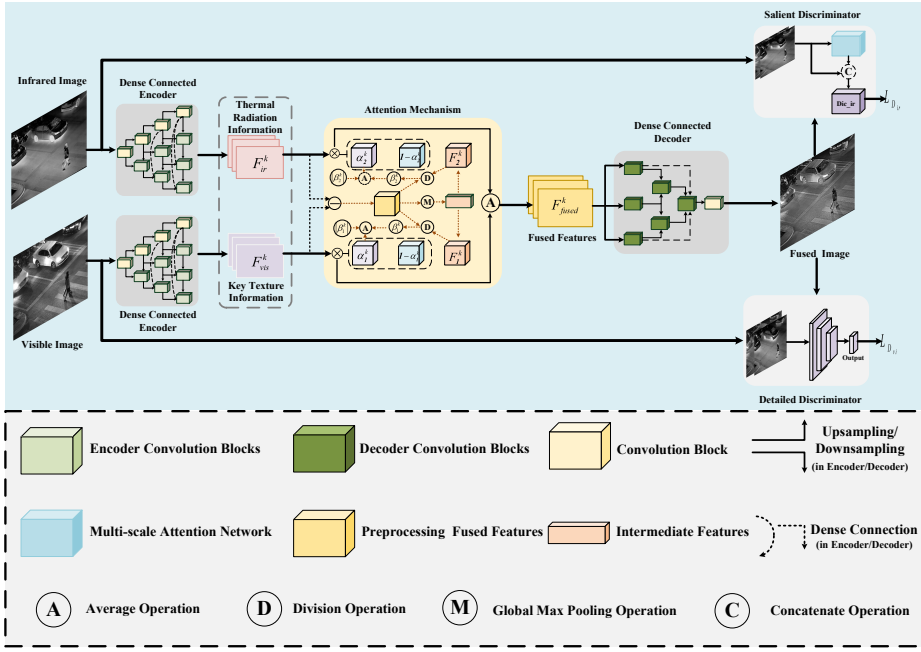

Figure 3.4: The architecture of the proposed JADD-GAN.

arately, use the attention model to combine these feature information and give information fusion in the decoder part. Different discriminators are used for infrared and visible images to better highlight thermal radiation information in infrared images and the detailed information and key textures in visible images by constructing different functions. Extensive experimental analyses based on four datasets are conducted to evaluate the performance of the proposed framework against the benchmark consisting of state-of-the-art IVF approaches.

3.4.3 Methods

Our proposed network architecture JADD-GAN which can be shown in Figure 3.4 has three parts: encoder sub-network, attention layer, and decoder sub-network. It can be found that our network extracts multi-scale depth features but does not deepen the network to some extent. A dual encoder is used in the encoding sub-network, and the decoding process gives partial fusion information. All intermediate features are fused by a dense jump-connected multiplexing layer. The goal of fusing infrared and visible images is to reconstruct a synthetic image with conspicuous targets and rich texture features. Selecting the right fusion strategy is essential, and in this work, we use the attention mechanism strategy. The human observation mechanism of external things is quite similar to the attention process.

When people watch external things, they typically do not look at things as a whole. but instead tend to selectively access particular key elements of the observed objects in accordance with their needs. Without adding additional computational and storage expense to the model, the attention mechanism enables the model to give varying weights to each input component and extract more crucial and significant information to improve judgments.

3.4.4 Result

We perform an experimental evaluation on four datasets (two for IVF, i.e., INO and RoadScene, and two for objection detection, i.e., and LLVIP). The 180/3500 multi-mode images were selected to train the fusion and detection tasks by random cropping and enhancement, respectively, cropping to 24k/151k blocks with pixels. Our training parameters were set as follows: batch size to 4, stage number N to 15. We set the number of channels of the size filter to 8, i.e. L = 8. The Adam optimizer is used to optimize the network with momentum terms (0.5, 0.99), and the learning rate is set to . All experiments are performed on an RTX3090TI GPU.

3.5 Paper V: Multi-label Abdominal Image Segmentation with Partially Labeled Data: A Prototypical Consistent Learning Perspective

3.5.1 Abstract

Recently, accurate automatic computed tomography (CT) segmentation of organs and tumors has the potential to facilitate clinical diagnosis and therapy. However, the automatic segmentation of multiple organs and tumors (MOTs) is a complex task, as they present variability in partially labeled data due to limited manpower and resources. The most prevalent techniques are committed to proposing a unified framework for the multi-task segmentation problem while suffering from the domain gap and discrepancy caused by the imbalance of data distribution. We introduce a novel prototype assignment strategy to handle the aforementioned imbalance challenges as weak enhancement information for a compact intra-class feature representation. Moreover, an exponential-based probability regularization term is proposed to avoid the inter-class imbalance problem caused by forcing the network to provide a consistent prototype label for adjacent features. Experiments comprehensively illustrate the performance of the proposed method compared with other state-of-the-art (SOTA) approaches both qualitatively and quantitatively.

3.5.2 Motivation

Traditional model-based segmentation techniques enjoy the theoretical guarantee of the segmentation process while suffering from the fixed operator and the non-adaptive segmented rules [49] [50] [207]. Many methods based on deep convolutional neural networks (CNN) have been proposed for MOTs segmentation [54, 95, 209]. Most methods have trained multiple independent networks for different targets (like only for the liver or kidney). The deep learning framework NNU-Net proposed by Fabian et al. [62] can independently make key decisions required to convert the basic architecture to different data sets and segmented tasks without manual adjustment. A cascade trainable segmentation model proposed by Yu et al. [192] captured the global and local appearance information from crossbar patches. Zhang et al. [198] proposed a lightweight hybrid convolutional network (LW-HCN) segmentation method for liver and tumor within CT volume, using the codec structure and depth and space-time separation (DSTS) technology, which effectively reduced the complexity of the model. While these methods take various measures to reduce the complexity, the computational complexity of their models is still a challenge and can not be ignored. Fang et al. [38] proposed a new training strategy, which enabled the multi-scale depth neural network to be trained on multiple partially labeled datasets through a shared encoder and significantly reduced the computational complexity of the model. Chen et al.[20] and Shi et al. [155] adopted a similar multi-head network to solve this multiple partially dataset problem. Although these methods achieve impressive performance, they are short of dealing with new tasks. Most methods [38, 20, 155] [205] only relied on a shared backbone network to realize the common knowledge of several inputs and multiple output headers for different segmentation tasks. However, the mode assumption of multi-class segmentation of partially labeled data may mislead some unlabeled organs as the background. The DoDNet proposed by Zhang et al. [197] was a single input head network and a single output head segmentation head. Specifically, it adopted a dynamic segmentation head to solve the problem of partial labeling and can simultaneously segment organs and tumors to overcome the above problem. However, some inevitable noise from the dataset and network training based on conditioning class label information can influence the performance of feature extraction and organ segmentation. This problem explicitly results from the intra-domain interference of MOTs task.

3.5.3 Methods

In order to solve the above problems, we introduce a novel prototype assignment strategy as a weak enhancement information to achieve a compact intra-class feature representation for our segmentation task. Firstly, to solve the problem of partially labeled data, the most important issue is to bring specific information to

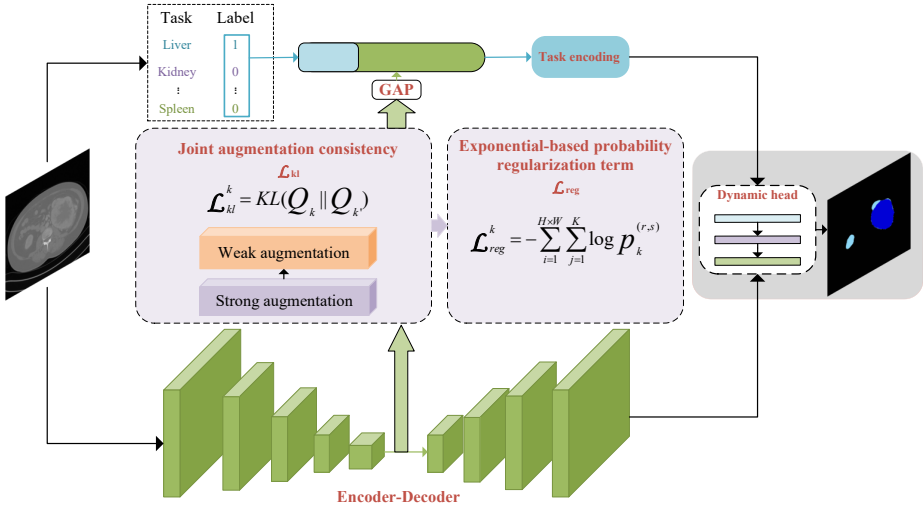

Figure 3.5: The framework of the proposed method.

guide the single feature extraction network to gain the discrepancy between tasks. However, the existing network ignores the problem of class-wise feature information to guide multiple-task learning by more compact intra-class feature representation. Thus, in this work, we exploit the implicit class-wise feature information by clustering learning to guide the learning of a partially labeled dataset segmentation task. Furthermore, we use an exponential-based probability regularization term, which encourages the output to be evenly distributed to different classes to overcome the problem of cluster degeneration like an empty cluster.

The network is motivated by [197] and consists of a feature extraction module, a prior knowledge extraction module, and a prior knowledge fusion module. Different from the existing method, we further investigate the works on how to exploit the capacity of prior knowledge. At first, to guide the single feature extraction module, the class one hot prior information is coded with middle representation . Figure 3.5, mainly includes 3 modules: a feature extraction head, a prior knowledge extraction module, and a prior knowledge fusion module. Kullback–Leibler (KL) divergence, i.e., joint augmentation consistency, and an exponential-based probability regularization term are introduced in the shared encoder-decoder.

3.5.4 Result

In this section, we compare the performance of the multi-organ and tumor segmentation (MOTS) dataset [197]. This dataset contains seven partially labeled sub-datasets involving the segmentation tasks of the kidney, liver, hepatic vessels,

pancreas, colon, lung, and spleen. A total of 1155 3D abdominal CT scans were collected from various clinical sites worldwide, including 920 for training and 235 for testing. Each scan is re-sliced to the same voxel size of . Five methods are selected in comparison with the proposed method, which consists of one multiple networks method, i.e., multi-nets, a multi-head networks approach, i.e., TAL [38], two single-network methods, i.e., Cond-NO and Cond-Dec [33], and a unified segmentation structure, i.e., DoDNet [197]. To ensure a fair comparison, we use the same encoder-decoder architecture for all methods.

Following the setting from [197], the processing of Hounsfield unit (HU) values is similar to the work in [197]: and linearly normalized them to . The weight standardization [144] is used for accelerating the training procedure. The stochastic gradient descent (SGD) algorithm with a momentum of 0.99 is adopted as the optimizer. The learning rate is initialized to 0.01 and decayed according to a polynomial policy , where the maximum epoch is set to 1000. The indicators for the experiment are the Dice similarity coefficient (Dice) and Hausdorff distance (HD)[162]. Dice and HD are commonly used for segmentation.

3.6 Paper VI: Learning the Distribution-Based Temporal Knowledge with Low-Rank Response Reasoning for UAV Visual Tracking

3.6.1 Abstract

In recent years, the constraint-based correlation filter has shown good performance in unmanned aerial vehicle (UAV) tracking, which has gained popularity in many intelligence transportation applications. This work proposes a distribution-based temporal knowledge-driven method to leverage the temporal translation property in UAV tracking. Instead of focusing on the traditional issues in the correlation filter, we provide a new method of learning parametric distribution on temporal knowledge by Wasserstein distance, which is successfully embedded to solve the problem of temporal degeneration in the learning process of tracking. Furthermore, we approximate optimal response reasoning with low-rank constraint over response consistency. Furthermore, the proposed method is solved by a simple iterative scheme with alternating direction multiplication ADMM algorithm. We demonstrate the superior tracking performance in several standard public UAV tracking benchmarks compared with state-of-the-art algorithms.

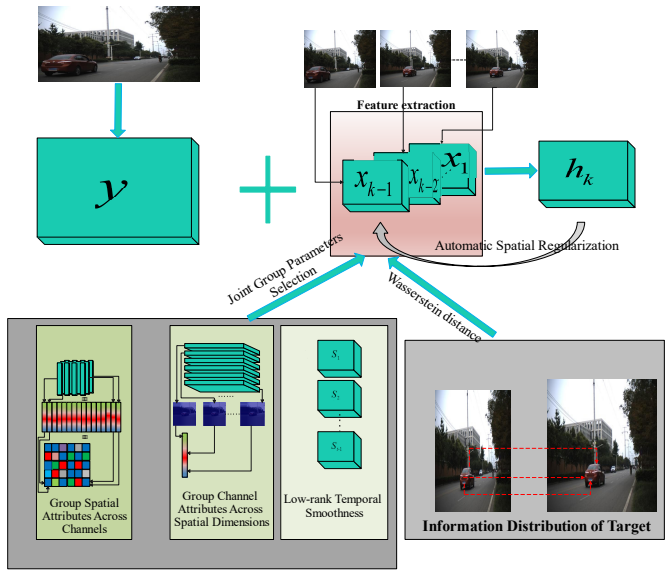

Figure 3.6: Our proposed method adopts low-rank temporal response constraint and group feature selection to improve the stability of the correlation filter.

3.6.2 Motivation

Benefiting from its easy implementation and fast prediction of the discriminative correlation filter (DCF), DCF has attracted attention in UAV tracking. Recently, to combine temporal information, some of the latest models used a transformer to combine spatial and temporal information. STARK [183] had not used any proposals, anchors, and post-processing steps (such as cosine window or bounding box regression), which greatly simplified the visual tracking model. While online learning of tracking has made good progress, there are still many problems in the temporal-based tracking framework. These existing methods only discover the reliability of spatial or temporal or background or response, the reliability of the temporal knowledge transfer is also deserved to investigate to avoid temporal degeneration. In existing temporal knowledge transfer based on the DCF tracking framework, the Euclidean distance is commonly used to measure the similarity of the targets of the two adjacent filters within a closed appearance [99, 181, 82]. Here, we recall a new concept about online temporal learning in visual tracking (*probability measurement*). This problem is not noticed by the above methods and raises some questions: *what can we measure in online learning: probabilistic temporal fitting or direct temporal interpolation?*

3.6.3 Methods

With the development of target tracking, research on low rank has made great progress and achieved good results. He et al. [55] had been successfully used in object tracking by exploiting low-rank constraints to capture the underlying structure of candidate particles. To mitigate this issue, we further investigate the low-rank reasoning over the temporal response. Therefore, we propose a new model (ATGT) as shown in Figure 3.6. The main contributions of our ATGT method include:

A novel Wasserstein distance regularization method for measuring the temporal transition is proposed. By adaptively incorporating the probability temporal fitting manner, the filter is able to mitigate the problem of temporal degeneration. Differently from inducing the representation, the low-rank constraint is conducted on the temporal response to achieve beyond response consistency for improving tracking robustness and overcoming the appearance variants. The iterative process is solved using the ADMM algorithm. A comprehensive evaluation of ATGT, including UVA123@10fps, DTB70, OTB100, UAVDT-M, and UAVDT-S. The results demonstrate the advantages of the ATGT, as well as its advantages over the most advanced trackers.

3.6.4 Result

We evaluate the performance of our ATGT and other trackers on six benchmark datasets, including DTB70[154], UAVDT-S[34], OTB100[172], UAVDT-M[34], and UVA123@10fps[130]

For quantitative comparison, we use the precision plot[172] and the success plot[172]. The precision plot illustrates the percentage of frames whose tracked locations are within the given threshold distance to the ground truth. A representative precision score with a threshold equal to 20 pixels is used to rank the trackers.

The results are compared with 11 state-of-the-art trackers with both HOG feature-based trackers and deep-based trackers, i.e, KCF[164], DSST [29], SAMF[97], SRDCF[7], STRCF[82], ECO-HC (with gray-scale)[27], AutoTrackC[99], GFSDCF[181], ARCF-HC[154], HOG-LR, LADCF[182], ARCF-H[154]. We evaluate our tracker on the dataset DTB70. The result shows the precision and success plots of all trackers. Among the existing methods, our proposed method has the best performance with scores of 0.492 and 0.714 on precision and success plot.

Chapter 4

Discussion

In the previous chapter, the articles included in this dissertation have been summarized concerning motivation, methodology, and results. In this chapter, the findings in each article are discussed with respect to the three initial research objectives and overall contributions to computational techniques in computer vision. The discussion is grouped into three parts corresponding to these research objectives. The first part is related to the work on Image Quality Enhancement for Single-modal Information Enhancement (SIE) (**RP I** and **RP II**), the second part is related to the Image Fusion Enhancement for Multi-modal Information Enhancement (MIE) (**RP III** and **RP VI**), and the last part is related to the Image Analysis for Task-driven Information Enhancement (TIE) (**RP V** and **RP IV**).

4.1 Unpaired Image Enhancement with Deep Neural Network

From an application perspective, RP I and RP II address the same main issue and test in many benchmark datasets which had been published in high-level journals and collect a few parts of the experimental dataset by ourselves. The main goal of these two works is not a denoising-like method for enhancing the quality, it is a method for adjusting the illumination and contrast for human eye perception and matching the corresponding high-level task for verification.

There have been an increasing number of studies on image quality improvement utilizing the unpaired learning approach in recent years thanks to the works of the generative adversarial network (GAN). [22] transformed the input image into the enhanced image based on a bidirectional GAN utilizing an image intensifier. An unpaired GAN model was proposed in [208] that aimed to solve the task with unpaired training data. It learned a map by converting images from the source

domain to the target domain and combined with inverse mapping to enhance images by using cyclic consistency loss. A model for converting MR images to CT images was proposed in [3] using paired-unpaired unsupervised attention-guided GANs (uagGANs). Based on paired datasets for pre-training and initialization, the uagGAN model was then retrained on unpaired datasets using a cascade method. In order to produce fine-structured images, pairwise pre-training was used to combine the Wasserstein GAN adversarial loss function with two new non-adversarial losses. In [101], they suggested a colorization network based on the CycleGAN model with a combination of the perceptual loss function and full variational loss function, in order to secure color medical images and enhance the quality of synthetic images while using unpaired training image data. The model in [69] also was constructed without low-light/normal light image pairs and can well handle various real-world test images. However, the biggest problem with these methods is that they focus on the global constraints of appearance and consistency and have poor performance in local detail learning. **RP I** proposes a Siamese-based struc-

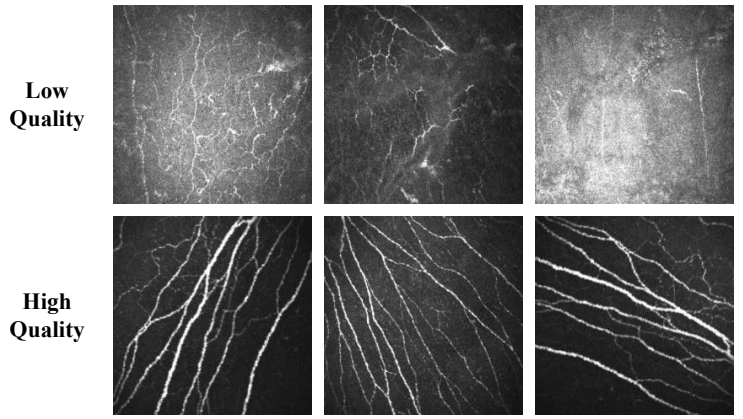

Figure 4.1: Examples of unpaired LQ images and HQ images of corneal confocal microscopy.

ture with dual inputs, low-quality (LQ) images, and high-quality (HQ) images, allowing the network to learn salient features from the HQ images while maintaining structural information in the LQ images. The proposed SSP-Net can generate high-quality enhanced results in an adversarial manner, which ensures the robustness of the SSP-Net to blurred textures, weakened structure, and background noise. The methodologies of **RP I** are based on Generative Adversarial Network (GAN), and the testing dataset shown in Figure 4.1 gives an example of three unpaired sets of LQ and HQ images, where the first row is LQ and the second row is HQ. The first and second LQ images represented challenges of uneven lighting distri-

bution and blurred texture detail, respectively, while the third LQ image contained both degradations. In contrast to these LQ images, the second row of HQ images is visually superior, with uniform illumination and clear structural detail, which is critical for many medical imaging applications: segmentation, computer-aided diagnosis, etc. The first and second LQ images represent the challenges of heterogeneous illumination distribution and blurred texture detail, respectively. The third LQ image is incorporated with the two aforementioned degradations, which is a more common phenomenon in LQ images.

Although **RP I** makes the progressive contribution to enhancing quality. Motivated by the multi-source domain gap, we develop the algorithm that is explored in the follow-up work **RP II**. Concerning two novel problems here for the unpaired image enhancement task:

- Q.A: What information is extracted between low-quality and high-quality domains?
- Q.B: How to properly investigate information from high quality to low quality?

Referring to Q.A, the most relevant medical image enhancement works treat this problem as image-to-image translation [106]. They lead to learning the shared representation with low-level image properties, such as texture or cartoon. Recently, one method has been to learn the relations between the two domains with independent autoencoders for the two domains, but existing methods always generate image-to-image translation-based enhancement according to shared weight perception[105]. Lin et al.[102] proposed image-level disentanglement and instance-level disentanglement to learn domain-invariant representation for generalizable object detection. Motivated by [142], the simple two-pathway encoder and a single decoder for image content transfer. We propose the disentanglement representation framework to preserve informative features for medical image enhancement.

Referring to Q.B, the usage of the extracted information is usually followed by the guided methodology in a multi-layer encoder-decoder manner. However, the encoded data is significantly different due to the unpaired data and degrades the spatial information in the medical image. As a result, the model will produce aliasing and chaos when encoding the information of different images, leading to blurred image structure, and disordered illumination distribution [195, 156]. In fact, unpaired input is not prone to achieve optimal in traditional CycleGAN as proved by [123]. The image-to-image translation assumption is inevitably affected by illumination, noises, and other variants. Nevertheless, the existing deep

learning image-to-image translation-based enhancement techniques avoid the difficulties of ideal medical image data collection. Due to the effects of heterogeneity and complex illumination conditions of natural images, the content of medical images is always homogeneous. Rather than the existing natural images unsupervised enhancement techniques[133], we prefer to use the above good property to solve medical image enhancement. Moreover, we observe the spatial homogeneity similarity across different quality medical images. Thus, we specially design a different module to perceive this specific information between high quality and low quality for medical image enhancement.

The **RP II** can be treated as the extension of **RP I**. And, from **RP II**, the experimental dataset uses three scenarios: corneal confocal microscope (CCM) dataset[123], public CVC-EndoSceneStrill dataset [168] and Whole-slide images (WSI) from the Genome Data Sharing Data Portal [186]. It is noticeable that these articles are generalized to any low-light application and do not depend on image source and calibration.

4.1.1 Limitation Analysis

The main limitation of **RP II** and **RP I** is the use of a small-scale testing dataset. Despite being evaluated on three benchmark datasets, the algorithms are yet to undergo large-scale verification. The testing dataset's small size could lead to overfitting, where the algorithm performs well on the specific data it was trained on but fails to generalize to other datasets. It is crucial to emphasize that the developed techniques require efficient verification in various scenarios and to be tested in real-world cases.

4.2 Unsupervised Image Fusion via Optimization Learning and Deep Learning

Our objective is not solely confined to enhancing image quality but also encompasses a thorough investigation into the fusion of images derived from multiple modalities.

From **RP III**, we concentrate on the fact that the structure of medical multimodality is a heterogenous system with self-similarity characteristics in nonlocal region [48]. Specifically, inspired by the fractal analysis, the fractal dimension represents the degree of self-similarity, and fractal features are extracted via a multimodal statistical distribution. Fractal analysis has been proposed for medical image diagnosis [4] and [171] for accurately identifying the spatial distribution and density. Thus, in this paper, the medical image fusion problem is tailored to quantify structural heterogeneity and balance the degree of heterogeneity. Here,

the notion of fractal analysis in this paper reflects that the small-scale structures of a fractal set resemble large-scale structures in an irregular region of interest in a medical image [134]. Thus, it intuitively gives inspiration for how to perceive image fusion. The motivation is similar to the work in [124] that proposed the fractal differential to enhance image details like edges and textures. By the construction of fractal constraint on convolution sparse representation-based fusion framework, we achieve the implicit perception of the fractal characteristic of components.

To improve fusion, the fractal constraint method is applied to retain image structural information to form a large group of patches. At times, the proposed fractal constraint is similar to the non-local constraint [139, 187]. They specifically highlight the importance of self-similarity to achieve more representative features. However, in medical image fusion, the sources used for fusion belong to different modalities. Compared to the method in [16, 139, 187], the non-local method is mainly used to remove noise by using the similarity in each group patch after which the image is divided. For our proposed method in **RP III**, the non-local-like fractal constraint is retained in the feature extraction on the separated components. Simultaneously, we avoid the intermediate group matching procedure [19] for the reason that block matching of the patch comes from different positions of the source image.

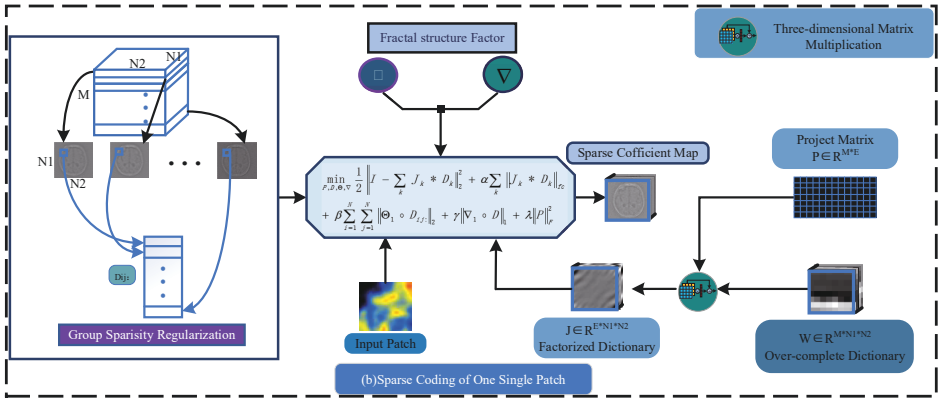

Figure 4.2: The sparse coding process of our FCFusion model.

As discussed earlier, feature extraction-driven methods have demonstrated that this explicit fusion guideline (extensions of feature extraction capacity only) leads to several issues for medical image fusion. The major problems of these algorithms are that the information is over-completed. Thus, the main solution is adapting the appropriate balance between feature extraction and redundancy removal. To realize our proposed *feature extraction* and *redundancy removal* to overcome the

over-smoothing issue. Figure 4.2 shows the schematic of the proposed fractal component-wise model, consisting of three parts: CSR, fractal component-wise constraint, and group sparsity. We first propose a fractal component-wise model and intuitively embed the group sparsity and its group variable weight techniques. Then we put the optimization process of the proposed method.

In **RP III**, we did not analyze the natural image fusion. It is understood in the literature that natural image has developed many deep-driven methods. The detailed analysis of image fusion quality would require a more general subjective experiment. In general, the contribution of **RP III** is toward a better methodology to guide the further method using the two core ideas: *feature extraction* and *redundancy removal* as a golden baseline for further research works.

Moreover, the feature extraction and redundancy removal ideas are combined into **RP IV**. In this work, the image fusion is not limited to the medical case, and it is moved to the natural image case for object detection and tracking.

The development of deep learning has driven great progress in image fusion, and the powerful feature extraction and reconstruction capabilities of neural networks make the fusion results promising. Building on the works of **RP III**, we design the deep network for the image fusion task in **RP IV**. The most challenging aspect of designing deep learning fusion models is based on the unsupervised learning truth to produce fused images. To solve this issue, Li et al. [87] presented the RFN-Nest, a new end-to-end fusion network design that trains the residual fusion network using both a new detail-preserving loss function and a feature-enhancing loss function. A two-stage training procedure is used to learn the fusion model, which makes up for the drawbacks of conventional approaches by being learnable. The work in [120], an image fusion network with a feature extraction network that can selectively extract salient target characteristics from infrared pictures and background texture features of visible images and perform salient target recognition and crucial information fusion implicitly. To solve the shortcomings of the training dataset, an unsupervised end-to-end learning system [57] may provide enough benchmark training datasets using visible and infrared frames. Additionally, a strong hybrid loss function that combines a modified structural similarity (M-SSIM) metric and a total variation (TV) is utilized to compensate for the lack of labeled datasets. Thermal radiation and texture details can be adaptively merged, and noise interference can be reduced, by constructing an unsupervised learning process.

Therefore, in **RP IV**, our main contributions include three aspects: Following the feature extraction idea in **RP III**, to increase the extraction of multi-level information by the network, and to accurately reflect the salient features of the source

images. In the generator part, we use the trained network to double encode the infrared and visible images separately, use the attention model to combine these feature information and give information fusion in the decoder part. For the redundancy removal idea, we use the different discriminators that are used for infrared and visible images to better highlight the thermal radiation information in infrared images and the detailed information and key textures in visible images by constructing different functions. Extensive experimental analyses based on four datasets are conducted to evaluate the performance of the proposed framework against the benchmark consisting of state-of-the-art fusion approaches.

But it is noticeable that the learning mechanism of fusion is in an unsupervised manner. Thus, it is trained to extract features from the input images and combine them in a meaningful way to generate the output image. Overall, deep learning neural networks for image fusion are powerful tools for integrating information from multiple sources to produce a more complete and informative image. By training these networks on large datasets, they can learn to perform complex image fusion tasks with high accuracy, making them valuable tools in a variety of fields.

4.2.1 Limitation Analysis

In **RP III**, the main limitation is the computational cost. In general, fusion processing often requires significant computational resources as they involve computing the objective function and its first and second derivatives. And to ensure obtaining the global optimal solution during the optimization process, one common strategy is to increase the search space of the optimization algorithm. However, it inevitably brings additional computational complexity.

Furthermore, while the existing deep neural networks have shown great promise for image fusion, there are several limitations that must be considered in **RP IV**. The first limitation is training data for image fusion, this can be a challenge, especially in cases where obtaining annotated data is difficult or time-consuming. Even though we have many testing datasets, there is no ground truth for the fusion task itself. The trained network is prone to overfitting, this can result in the network memorizing the training data instead of learning the underlying patterns, leading to poor generalization performance on new, unseen data.

Despite these limitations, based on these existing limitations, it also indicates our future development direction for deep neural networks for image fusion.

4.3 Leveraging the external information as weak supervision for high-level vision tasks.

In **RP V**, we explore external information to enhance current feature extraction. We introduce a novel prototype assignment strategy as weak enhancement information to achieve a compact intra-class feature representation for our segmentation task. Firstly, to solve the problem of partially labeled data, the most important issue is to bring specific information to guide the single feature extraction network to gain the discrepancy between tasks. However, the existing network ignores the problem of class-wise feature information to guide multiple-task learning by more compact intra-class feature representation. Thus, in this work, we exploit implicit class-wise feature information by clustering learning to guide the learning of a partially labeled dataset segmentation task. Furthermore, we use an exponential-based probability regularization term [193], which encourages the output to be evenly distributed to different classes to overcome the problem of cluster degeneration like an empty cluster.

For **RP VI**, actually, the distorted appearance in visual tracking challenges the spatial or temporal-based DCF methods. The above discussion motivates us to mitigate the problem of overfitting and omit the impact of unpredicted appearance. Fortunately, the Wasserstein distance with a common Lagrangian formulation and alleviates the need for a common space. In [200], they proposed a novel approach to learning domain invariant feature representations. Wasserstein generative adversarial network (GAN) [6] learned a more reasonable and efficient approximation method and cured the main training problem of GAN. Thus, we leverage a probability temporal fitting method motivated by the Wasserstein distance. To improve the anti-noise performance of the tracking, we use the Wasserstein distance to measure the similarity of the filter distribution instead of the previous linear interpolation method for the estimation of the temporal filter.

To mitigate this issue, we further investigate the low-rank reasoning over the temporal response. A novel Wasserstein distance regularization method for measuring the temporal transition is proposed. By adaptively incorporating the probability temporal fitting manner, the filter is enabled to mitigate the problem of temporal degeneration. Unlike inducing the representation, the low-rank constraint is conducted on the temporal response to achieve beyond response consistency for improving tracking robustness and overcoming the appearance variants. The iterative process is solved by the ADMM algorithm. A comprehensive evaluation of ATGT, including UVA123@10fps, DTB70, OTB100, UAVDT-M, and UAVDT-S. The results demonstrate the advantages of the ATGT, as well as its advantages over the most advanced trackers.

4.3.1 Limitation Analysis

In **RP V**, the dataset used in the experiment is very heavy and lacks of the lightweight architecture of multiple organs and tumors segmentation.

Discriminative Correlation Filter (DCF) is a popular method for visual object tracking, which has achieved good performance in terms of accuracy and efficiency. However, it also has several limitations in **RP VI**, including:

Requirement for Regular Sampling: the proposed method relies on a regular sampling of the search region, which may cause suboptimal results when the target moves in an irregular manner.

Lack of Online Model Update: the proposed method does not have a mechanism for updating the model online, which may result in a decrease in tracking performance in a long-term tracking process over time as the target appearance changes.

4.4 Additional Contributions

In addition to the computational analysis of medical images and natural image information enhancement, other works related to the computer vision domain were finished during the research development.

4.4.1 Data Fusion by Deep Learning

Background

Multimodal sentiment analysis of social media has attracted increasing attention. Its core idea is to discover a heuristic fusion strategy to analyze the sentiment orientations over heterogeneous multimodal sources from a learned compact multimodal representation. Unfortunately, existing multimodal fusion techniques not only struggle to achieve entire heterogeneous data interaction but are also unable to dynamically assess the quality of various modal data to determine predictability.

Methods

To address the above issues, the first one is that we present a novel profound tensor evidence fusion network for multimodal sentiment classification termed DTEF. Firstly, we propose a common view evaluation network that uses an extended short-term memory network (LSTM) and a tensor-based neural network to extract rich inter-modal and intra-modal information. Then, we propose a unique time cue evaluation network that takes advantage of the temporal granularity associated with numerous pattern sequences. To make reliable decisions, we finally incorporate uncertainty through the trusted fusion layer, which improves the accuracy and robustness of sentimental classification. Our model is validated using the CMU-

MOSEI and CMU-MOSI datasets, and the experimental findings demonstrate the superior performance of the proposed network in terms of accuracy compared with the state-of-the-art methods.

Outcome

There are two related publications, as shown below.

1. Zongyang Wang, **Guoxia Xu**, Xiaokang Zhou, Kim, J. Y., Hu Zhu, Lizhen Deng. Deep Tensor Evidence Fusion Network for Sentiment Classification. IEEE Transactions on Computational Social Systems, 2022.

   Contribution: Problem formulation, Experiment Design and Development, Paper Writing.

4.4.2 Temporal Information for Visual Tracking

Background Many discriminative correlation filters (DCF) based methods have successfully leveraged the guidance for solving two problems (i.e., the boundary effect and temporal filtering degradation) as a model before visual tracking. Regardless of the specific content of the tracking algorithms, the intuitive motivation of these methods is to control the degeneration of the updating loss of the objective function with a structural framework. However, while these methods rely primarily on various explicit prior regularization items, they always ignore the loss from the data fidelity term. Furthermore, these trackers only adopt first-order data-fitting information and have difficulty maintaining robust tracking in unconstrained scenarios, especially in the case of complex appearance variations.

Methods

To address the above issues, we propose a bilateral weighted regression ranking model with a spatial-temporal correlation filter, namely, BWRR. Here, we resort to two procedures for solving the above problems. First, BWRR introduces a bilateral constraint into the data fidelity term to control the filter learning data term's loss of rows and columns. The weighted matrices could impose an adaptive penalty for significant data loss during learning to avoid the tracking offset and model degradation problems. Second, the data of the updated weighted matrices is not directly applied to the calculation of the filter during each iteration. Instead, a new weighted product matrix is obtained by ranking and numerical transformation for updating the filter. We show that the proposed model converts the original correlation filter regression problem into a regression-with-ranking problem, thus avoiding the problem of positive and negative sample imbalance. Overall, the BWRR model is approximated as a linear equality constraint problem, which is iteratively

solved by the alternating direction method of multipliers(ADMM). Qualitative and quantitative evaluations demonstrate the effectiveness and superiority of our proposed approach through extensive and quantitative experiments on the OTB, VOT, and UAV datasets.

The second one designs a new method by introducing a second-order data-fitting term to the DCF; we propose a second-order spatial–temporal correlation filter (SSCF) learning model. Specifically, the SSCF tracker incorporates both the first-order and second-order data-fitting terms into the DCF framework, making the learned correlation filter more discriminative. Meanwhile, the spatial–temporal regularization was integrated to develop a robust model for tracking complex appearance variations. Finally, extensive experiments were conducted on the benchmark databases.

Outcome

There are two related publications, as shown below.

1. Hu Zhu, Hao Peng, **Guoxia Xu**, Lizhen Deng, Yueying Cheng, and Aiguo Song. Bilateral weighted regression ranking model with spatial-temporal correlation filter for visual tracking. IEEE Transactions on Multimedia 24: 2098-2111, 2021.

   Contribution: Problem formulation, System and Experiment Design, paper review.

2. Yu-Feng Yu, Long Chen, Haoyang He, Jianhui Liu, Weipeng Zhang, **Guoxia Xu**, Second-Order Spatial-Temporal Correlation Filters for Visual Tracking. Mathematics, 2022.

   Contribution: Problem formulation, System and Experiment Design, paper review.

4.4.3 The Feature Learning in Deep Learning

Background Traffic sign classification plays a vital role in autonomous vehicles for its powerful capability in information representation. However, the low-quality data of traffic signs captured by in-vehicle cameras often inevitably bring inherent challenges to the one-shot classification task. Apart from the problem of data degradation, learning-based classification techniques of real traffic signs also come across the challenges of intra-class and inter-class data imbalance from the training data.

Methods To overcome the problems above, we propose an end-to-end degradation robust deep model, PcGAN, to classify traffic signs using few-shot learning. The proposed PcGAN models the joint distribution between the degraded traffic signal data and the corresponding prototypes from both degradation removal and generation perspectives by two alternating optimized modules, which ensures the generalization of the learned embedding of latent space for novel tasks. A multi-task loss function is designed to improve the robustness of PcGAN. Numerous experiments comprehensively demonstrate that the accuracy of our proposed PcGAN is improved by 5% compared with other state-of-the-art (SOTA) approaches in few-shot classification.

Outcome

There are two related publications, as shown below.

1. Lizhen Deng, Chunming He, **Guoxia Xu**, Hu Zhu, Hao Wang, PcGAN: A Noise Robust Conditional Generative Adversarial Network for One Shot Learning, IEEE Transactions on Intelligent Transportation Systems, 2022.

   Contribution: Problem formulation, System and Experiment Design, paper writing.

Chapter 5

Conclusions and future perspectives

In this chapter, we will provide an overall conclusion of the thesis and give a perspective for future work.

5.1 Conclusion

This Ph.D. thesis aims at researching information enhancement with special emphasis on three types Single-modal Information Enhancement (SIE), Multi-modal Information Enhancement (MIE), and Task-driven Information Enhancement (TIE). The topic of the Ph.D. is multidisciplinary research of imaging problems consisting of medical images and natural images with respect to existing information deficiency issues of imaging technologies. The dissertation provides an analysis of several public benchmark datasets and proposes several state-of-the-art computational techniques to improve visual quality, content supplement, and high-level visual tasks.

Methodological perspective: The research articles in this project contribute to image enhancement, image fusion, automated multiple organs, and tumor segmentation/object tracking algorithms. Image enhancement problems are challenging as they depend on acquisition hardware and lighting conditions. This Ph.D. project proposes two unpaired medical image enhancement algorithms based on Generative Adversarial Network (GAN) methodology for solving several imaging issues in medical scenarios. Experimental results show that the proposed method provides better visualization of the desired image attributes. Moving to image fusion, we proposed an image fusion algorithm based on the fractal idea, which intuitively

imposes the patch-level component-wise separation to perceive the fractal characteristic across the different components in multi-modality sources. Furthermore, we design a deep learning-based method for the image fusion task. To increase the extraction of multi-level information by the network and to accurately reflect the salient features of the source images, in the generator part, we use the trained network to double encode the source images separately, use the attention model to combine these feature information and give information fusion in the decoder part. In the later period of this Ph.D. project, we are focusing on bringing the information enhancement perspective to high-level tasks. We investigate a novel prototype assignment strategy as weak enhancement information to achieve a compact intra-class feature representation for multiple organ segmentation problems based on a partially labeled dataset. Furthermore, a novel Wasserstein distance regularization method for measuring the temporal transition is proposed to exploit the information enhancement idea in temporal video object tracking. By adaptively incorporating the probability temporal fitting manner, the filter is enabled to mitigate the problem of temporal degeneration.

Application perspective: This thesis identifies the challenges and opportunities in information enhancement from three levels and gives our understanding of quality, content, and tasks. The methods developed during this project could flourish in the existing information enhancement research community. Furthermore, enhancement ideas can be further used to inspire other researchers to design their works, and enhancement can be used for different levels, architectures, and cases.

5.2 Future Research Orientation

We have identified many research challenges that can be addressed in future work. They are grouped according to their focus on the different high-level pipelines.

We found that information enhancement can be mainly about a plug-and-play countermeasure strategy, which is to generate more salient images to improve the 'task processor' ability to extract discriminative information. Intuitively, it may also enhance the performance of the enhancement algorithm in low-level on real-world images. The previous methods are to generate a more realistic degraded image through degeneration, form a pseudo-image pair, and then use such pseudo-image pairs to train an enhancement network. However, these two steps are separate and not end-to-end. In addition, such images do not provide targeted degradation for the weakness of the enhancement network and thus do not fully strengthen the resilience of the enhancement network to complex degradation. We can further use the confrontation strategy to bridge these two areas. In phase 1, we fix the intensifier and optimize the degeneration. The purpose is to make the degeneration generate degenerate images that can make the enhancer fail as much as possible

(how to evaluate the failure has not been decided, such as PSNR, SSIM, etc.). In phase 2, we fix the degeneration, optimize the enhancer, and force the enhancer to restore such degenerate images perfectly. By alternating these two steps, we hope that the degeneration can generate targeted, degraded images to improve the recovery ability of the intensifier for complex degradation so as better to improve the performance of the intensifier on real-world images.

Different scenarios have different cases, and our target is how to design a unified case to handle these questions from different perspectives. Color deviation of many nature images; Low-light images may also have deviations, and foggy images (equivalent to low-light images) can be treated similarly to low-light images, which will also have corresponding recovery problems. This has brought difficulties to the network: these tasks are extreme, and the deep learning network based on data statistics is difficult to learn. Perhaps under this consideration, we must guide a non-pairing high-quality image when restoring a low-quality image. Since HQ pictures do not have such extreme degradation, HQ can provide real-time correction assistance. In addition, this guidance is also conducive to our unified enhancement because low-quality maps have their differences, while high-quality maps have their common advantages. Such explicit guidance can strengthen the stability and generalization of our enhanced network.

Bibliography

[1] Research trend for information enhancement. https://exaly.com/trends/Information-Enhancement, February 2023. Date visited: 02/03/23.

[2] Mohammad Abdullah-Al-Wadud, Md Hasanul Kabir, M Ali Akber Dewan, and Oksam Chae. A dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, 53(2):593–600, 2007.

[3] Alaa Abu-Srhan, Israa Almallahi, Mohammad AM Abushariah, Waleed Mahafza, and Omar S Al-Kadi. Paired-unpaired unsupervised attention guided gan with transfer learning for bidirectional brain mr-ct synthesis. *Computers in Biology and Medicine*, 136:104763, 2021.

[4] Omar S Al-Kadi, Daniel YF Chung, Constantin C Coussios, and J Alison Noble. Heterogeneous tissue characterization using ultrasound: a comparison of fractal analysis backscatter models on liver tumors. *Ultrasound in medicine & biology*, 42(7):1612–1626, 2016.

[5] Puvvadi Aparna and Polurie Venkata Vijay Kishore. Biometric-based efficient medical image watermarking in e-healthcare application. *IET Image Processing*, 13(3):421–428, 2019.

[6] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. PMLR, 2017.

[7] Nikhilanand Arya and Sriparna Saha. Multi-modal classification for human breast cancer prognosis prediction: Proposal of deep-learning based stacked

ensemble model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 1–1, 2020.

[8] Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283, 2022.

[9] Yuri Sousa Aurelio, Gustavo Matheus de Almeida, Cristiano Leite de Castro, and Antonio Padua Braga. Learning from imbalanced data sets with weighted cross-entropy function. *Neural processing letters*, 50(2):1937–1949, 2019.

[10] Emrah Benli, Richard Lee Spidalieri, and Yuichi Motai. Thermal multisensor fusion for collaborative robotics. *IEEE Transactions on Industrial Informatics*, 15(7):3784–3795, Jul. 2019.

[11] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016.

[12] Gaurav Bhatnagar, QM Jonathan Wu, and Zheng Liu. Directive contrast based multimodal medical image fusion in nsct domain. *IEEE transactions on multimedia*, 15(5):1014–1024, 2013.

[13] Gaurav Bhatnagar, QM Jonathan Wu, and Zheng Liu. Human visual system inspired multi-modal medical image fusion framework. *Expert Systems with Applications*, 40(5):1708–1720, 2013.

[14] Sujoy Kumar Biswas and Peyman Milanfar. Linear support tensor machine with lsk channels: Pedestrian detection in thermal infrared images. *IEEE Transactions on Image Processing*, 26(9):4229–4242, 2017.

[15] Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger Gunn, Alexander Hammers, David Alexander Dickie, Maria Valdés Hernández, Joanna Wardlaw, and Daniel Rueckert. Gan augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*, 2018.

[16] Antoni Buades, Bartomeu Coll, and J. M. Morel. A non-local algorithm for image denoising. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005.

[17] Abraham Chandy et al. A review on iot based medical imaging technology for healthcare applications. *Journal of Innovative Image Processing (JIIP)*, 1(01):51–60, 2019.

[18] Lihong Chang, Xiangchu Feng, Xiaolong Zhu, Rui Zhang, Ruiqiang He, and Chen Xu. Ct and mri image fusion based on multiscale decomposition method and hybrid approach. *IET Image Processing*, 13(1):83–88, 2019.

[19] Yi Chang, Luxin Yan, and Sheng Zhong. Hyper-laplacian regularized unidirectional low-rank tensor recovery for multispectral image denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4260–4268, 2017.

[20] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

[21] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8126–8135, 2021.

[22] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6306–6314, 2018.

[23] Hyeyeon Choi, Jong Pil Yun, Bum Jun Kim, Hyeonah Jang, and Sang Woo Kim. Attention-based multimodal image feature fusion module for transmission line detection. *IEEE Transactions on Industrial Informatics*, 2022.

[24] Justin R Chumbley and Karl J Friston. False discovery rate revisited: Fdr and topological inference using gaussian random fields. *Neuroimage*, 44(1):62–70, 2009.

[25] Guangmang Cui, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Optics Communications*, 341:199–209, 2015.

[26] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising with block-matching and 3d filtering. In *Image Processing: Algorithms and Systems, Neural Networks, and Machine Learning*, volume 6064, page 606414. International Society for Optics and Photonics, 2006.

[27] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Eco: Efficient convolution operators for tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6638–6646, 2017.

[28] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[29] Martin Danelljan, Gustav Häger, Fahad Khan, and Michael Felsberg. Accurate scale estimation for robust visual tracking. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–11. Bmva Press, 2014.

[30] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[31] Mohhammad Daneshzand, Reza A Zoroofi, and Miad Faezipour. Mr image assisted drug delivery in respiratory tract and trachea tissues based on an enhanced level set method. In *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education*, pages 1–7, 2014.

[32] Sudeb Das and Malay Kumar Kundu. A neuro-fuzzy approach for medical image fusion. *IEEE transactions on biomedical engineering*, 60(12):3347–3353, 2013.

[33] Konstantin Dmitriev and Arie E Kaufman. Learning multi-class segmentations from single-class datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9501–9511, 2019.

[34] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 370–386, 2018.

[35] Jiao Du, Weisheng Li, and Bin Xiao. Anatomical-functional image fusion by information of interest in local laplacian filtering domain. *IEEE Transactions on Image Processing*, 26(12):5855–5866, 2017.

[36] Shan Du and Rabab K Ward. Adaptive region-based image enhancement method for robust face recognition under variable illumination conditions. *IEEE transactions on circuits and systems for video technology*, 20(9):1165–1175, 2010.

[37] Nastaran Emaminejad, Wei Qian, Yubao Guan, Maxine Tan, Yuchen Qiu, Hong Liu, and Bin Zheng. Fusion of quantitative image and genomic biomarkers to improve prognosis assessment of early stage lung cancer patients. *IEEE Transactions on Biomedical Engineering*, 63(5):1034–1043, 2015.

[38] Xi Fang and Pingkun Yan. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. *IEEE Transactions on Medical Imaging*, 39(11):3619–3629, 2020.

[39] Peng Feng, Jing Wang, Biao Wei, and Deling Mi. A fusion algorithm for gfp image and phase contrast image of arabidopsis cell based on sfl-contourlet transform. *Computational and mathematical methods in medicine*, 2013, 2013.

[40] William T Freeman and Egon C Pasztor. Learning low-level vision. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1182–1189. IEEE, 1999.

[41] Changhong Fu, Jin Jin, Fangqiang Ding, Yiming Li, and Geng Lu. Spatial reliability enhanced correlation filter: An efficient approach for real-time uav tracking. *IEEE Transactions on Multimedia*, pages 1–1, 2021. doi=10.1109/TMM.2021.3118891.

[42] Ying Fu, Yang Hong, Linwei Chen, and Shaodi You. Le-gan: unsupervised low-light image enhancement network using attention module and identity invariant loss. *Knowledge-Based Systems*, 240:108010, 2022.

[43] Shruti Garg, K Ushah Kiran, Ram Mohan, and US Tiwary. Multilevel medical image fusion using segmented image by level set evolution with region competition. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 7680–7683. IEEE, 2006.

[44] Tessa SS Genders, Sandra Spronk, Theo Stijnen, Ewout W Steyerberg, Emmanuel Lesaffre, and MG Myriam Hunink. Methods for calculating sensitivity and specificity of clustered data: a tutorial. *Radiology*, 265(3):910–916, 2012.

[45] Hayit Greenspan. Super-resolution in medical imaging. *The computer journal*, 52(1):43–63, 2009.

[46] Kai Guo, Shuai Wu, and Yong Xu. Face recognition using both visible light image and near-infrared image and a deep network. *CAAI Transactions on Intelligence Technology*, 2(1):39–47, 2017.

[47] Ruize Han, Wei Feng, and Song Wang. Fast learning of spatially regularized and content aware correlation filter for visual tracking. *IEEE Transactions on Image Processing*, 29:7128–7140, 2020.

[48] Mehdi Hassan, Safdar Ali, Hani Alquhayz, and Khushbakht Safdar. Developing intelligent medical image modality classification system using deep transfer learning and lda. *Scientific reports*, 10(1):1–14, 2020.

[49] Chunming He, Xiaobo Wang, Lizhen Deng, and Guoxia Xu. Image threshold segmentation based on glle histogram. In *2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pages 410–415. IEEE, 2019.

[50] Chunming He, Lei Xu, Guosheng Lu, and Lizhen Deng. Glle entropic threshold segmentation based on fuzzy entropy. *Journal of Nanjing University of Information Science and Technology*, 11(6):757–763, 2019.

[51] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010.

[52] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1397–1409, 2012.

[53] 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*, pages 770–778, 2016.

[54] Kelei He, Xiaohuan Cao, Yinghuan Shi, Dong Nie, Yang Gao, and Dinggang Shen. Pelvic organ segmentation using distinctive curve guided fully convolutional networks. *IEEE transactions on medical imaging*, 38(2):585–595, 2018.

[55] Yujie He, Min Li, Jinli Zhang, and Junping Yao. Infrared target tracking based on robust low-rank sparse learning. *IEEE Geoscience and Remote Sensing Letters*, 13(2):232–236, 2015.

[56] Haithem Hermessi, Olfa Mourali, and Ezzeddine Zagrouba. Multimodal medical image fusion review: Theoretical background and recent advances. *Signal Processing*, 183:108036, 2021.

[57] Ruichao Hou, Dongming Zhou, Rencan Nie, Dong Liu, Lei Xiong, Yanbu Guo, and Chuanbo Yu. Vif-net: an unsupervised framework for infrared and visible image fusion. *IEEE Transactions on Computational Imaging*, 6:640–651, 2020.

[58] Chih-Chung Hsu, Chia-Wen Lin, Weng-Tai Su, and Gene Cheung. Sigan: Siamese generative adversarial network for identity-preserving face hallucination. *IEEE Transactions on Image Processing*, 28(12):6225–6236, 2019.

[59] Jin Huang and Charles X Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310, 2005.

[60] Rui Huang, Yuanjie Zheng, Zhiqiang Hu, Shaoting Zhang, and Hongsheng Li. Multi-organ segmentation via co-training weight-averaged models from few-organ datasets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 146–155. Springer, 2020.

[61] Ziyuan Huang, Changhong Fu, Yiming Li, Fuling Lin, and Peng Lu. Learning aberrance repressed correlation filters for real-time uav tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2891–2900, 2019.

[62] Fabian Isensee, Paul F Jäger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. Automated design of deep learning methods for biomedical image segmentation. *arXiv preprint arXiv:1904.08128*, 2019.

[63] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[64] P Jagalingam and Arkal Vittal Hegde. A review of quality metrics for fused image. *Aquatic Procedia*, 4:133–142, 2015.

[65] Alex Pappachen James and Belur V Dasarathy. Medical image fusion: A survey of the state of the art. *Information fusion*, 19:4–19, 2014.

[66] A.P. James and B.V. Dasarathy. Medical image fusion: A survey of the state of the art. *Information Fusion*, 19:4–19, 2014.

[67] Jiwoong J Jeong, Amara Tariq, Tobiloba Adejumo, Hari Trivedi, Judy W Gichoya, and Imon Banerjee. Systematic review of generative adversarial

networks (gans) for medical image classification and segmentation. *Journal of Digital Imaging*, 35(2):137–152, 2022.

[68] Y. Jiang and M. Wang. Image fusion with morphological component analysis. *Information Fusion*, 18(1):107–118, 2014.

[69] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021.

[70] Yong Jiang and Minghui Wang. Image fusion with morphological component analysis. *Information Fusion*, 18:107–118, 2014.

[71] J. A. Johnson and J. A. Becker. The whole brain altas. *Online*, 1997. doi=http://www.med.harvard.edu/aanlib/.

[72] Hyungjoo Jung, Youngjung Kim, Hyunsung Jang, Namkoo Ha, and Kwanghoon Sohn. Unsupervised deep image fusion with structure tensor representations. *IEEE Transactions on Image Processing*, 29:3845–3858, 2020.

[73] Donggoo Kang, Sangwoo Park, and Joonki Paik. Sdban: Salient object detection using bilateral attention network with dice coefficient loss. *IEEE Access*, 8:104357–104370, 2020.

[74] Neel Kanwal, Fernando Pérez-Bueno, Arne Schmidt, Kjersti Engan, and Rafael Molina. The devil is in the details: Whole slide image acquisition and processing for artifacts detection, color variation, and data augmentation: A review. *IEEE Access*, 10:58821–58844, 2022.

[75] Minjae Kim, David K Han, and Hanseok Ko. Joint patch clustering-based dictionary learning for multimodal image fusion. *Information Fusion*, 27:198–214, 2016.

[76] Kazuhiro Kitajima, Yuko Suenaga, Yoshiko Ueno, Tomonori Kanda, Tetsuo Maeda, Masashi Deguchi, Yasuhiko Ebina, Hideto Yamada, Satoru Takahashi, and Kazuro Sugimura. Fusion of pet and mri for staging of uterine cervical cancer: comparison with contrast-enhanced 18f-fdg pet/ct and pelvic mri. *Clinical Imaging*, 38(4):464–469, 2014.

[77] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *International Conference on Machine Learning*, volume 2. Lille, 2015.

[78] Helena C Kraemer. Kappa coefficient. *Wiley StatsRef: statistics reference online*, pages 1–4, 2014.

[79] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void - learning denoising from single noisy images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2124–2132, 2019.

[80] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*, 2018.

[81] Fan Li, Changhong Fu, Fuling Lin, Yiming Li, and Peng Lu. Training-set distillation for real-time uav object tracking. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9715–9721. IEEE, 2020.

[82] Feng Li, Cheng Tian, Wangmeng Zuo, Lei Zhang, and Ming-Hsuan Yang. Learning spatial-temporal regularized correlation filters for visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4904–4913, 2018.

[83] Feng Li, Cheng Tian, Wangmeng Zuo, Lei Zhang, and Ming-Hsuan Yang. Learning spatial-temporal regularized correlation filters for visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4904–4913, 2018.

[84] Huafeng Li, Xiaoge He, Dapeng Tao, Yuanyan Tang, and Ruxin Wang. Joint medical image fusion, denoising and enhancement via discriminative low-rank sparse dictionaries learning. *Pattern Recognition*, 79:130–146, 2018.

[85] Huafeng Li, Yitang Wang, Zhao Yang, Ruxin Wang, Xiang Li, and Dapeng Tao. Discriminative dictionary learning-based multiple component decomposition for detail-preserving noisy image fusion. *IEEE Transactions on Instrumentation and Measurement*, 69(4):1082–1102, 2019.

[86] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019.

[87] Hui Li, Xiao-Jun Wu, and Tariq Durrani. Nestfuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. *IEEE Transactions on Instrumentation and Measurement*, 69(12):9645–9656, 2020.

[88] Hui Li, Xiao-Jun Wu, and Josef Kittler. Infrared and visible image fusion using a deep learning framework. In *2018 24th international conference on pattern recognition (ICPR)*, pages 2705–2710. IEEE, 2018.

[89] Hui Li, Xiao-Jun Wu, and Josef Kittler. Mdlatlrr: A novel decomposition method for infrared and visible image fusion. *IEEE Transactions on Image Processing*, 29:4733–4746, 2020.

[90] Shutao Li, Xudong Kang, and Jianwen Hu. Image fusion with guided filtering. *IEEE Transactions on Image processing*, 22(7):2864–2875, 2013.

[91] Shutao Li, Xudong Kang, and Jianwen Hu. Image fusion with guided filtering. *IEEE Transactions on Image processing*, 22(7):2864–2875, 2013.

[92] Shutao Li, Haitao Yin, and Leyuan Fang. Group-sparse representation with dictionary learning for medical image denoising and fusion. *IEEE Transactions on biomedical engineering*, 59(12):3450–3459, 2012.

[93] Xiang Li, Yuchen Jiang, Minglei Li, and Shen Yin. Lightweight attention convolutional neural network for retinal vessel image segmentation. *IEEE Transactions on Industrial Informatics*, 17(3):1958–1967, 2020.

[94] Xiaoguang Li, Ning Dong, Jianglu Huang, Li Zhuo, and Jiafeng Li. A discriminative self-attention cycle gan for face super-resolution and recognition. *IET Image Processing*, 15(11):2614–2628, 2021.

[95] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.

[96] Ya Li, Xinmei Tian, Xu Shen, and Dacheng Tao. Classification and representation joint learning via deep networks. In *International Joint Conference on Artificial Intelligence*, volume 2017, page 67, 2017.

[97] Yang Li and Jianke Zhu. A scale adaptive kernel correlation filter tracker with feature integration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 254–265. Springer, 2014.

[98] Yang Li, Jianke Zhu, Steven CH Hoi, Wenjie Song, Zhefeng Wang, and Hantang Liu. Robust estimation of similarity transformation for visual object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8666–8673, 2019.

[99] Yiming Li, Changhong Fu, Fangqiang Ding, Ziyuan Huang, and Geng Lu. Autotrack: Towards high-performance visual tracking for uav with automatic spatio-temporal regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11923–11932, 2020.

[100] Qiusheng Lian, Wenfeng Yan, Xiaohua Zhang, and Shuzhen Chen. Single image rain removal using image decomposition and a dense network. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1428–1437, 2019.

[101] Yihuai Liang, Dongho Lee, Yan Li, and Byeong-Seok Shin. Unpaired medical image colorization using generative adversarial network. *Multimedia Tools and Applications*, 81(19):26669–26683, 2022.

[102] Chuang Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhu Wang, and Jianfei Cai. Domain-invariant disentangled network for generalizable object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8771–8780, 2021.

[103] Guangcan Liu and Shuicheng Yan. Latent low-rank representation for subspace segmentation and feature extraction. In *2011 international conference on computer vision*, pages 1615–1622. IEEE, 2011.

[104] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[105] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017.

[106] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. *Advances in neural information processing systems*, 29, 2016.

[107] Xinwei Liu, Marius Pedersen, and Renfang Wang. Survey of natural image enhancement techniques: Classification, evaluation, challenges, and perspectives. *Digital Signal Processing*, page 103547, 2022.

[108] Y. Liu, X. Chen, R. Ward, and Z. Wang. Image fusion with convolutional sparse representation. *IEEE Signal Procession Letters*, 23(12):1882–1886, Dec. 2016.

[109] Yang Liu, Ziyu Yue, Jinshan Pan, and Zhixun Su. Unpaired learning for deep image deraining with rain direction regularizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4753–4761, 2021.

[110] Yu Liu, Xun Chen, Hu Peng, and Zengfu Wang. Multi-focus image fusion with a deep convolutional neural network. *Information Fusion*, 36:191–207, 2017.

[111] Yu Liu, Xun Chen, Rabab K. Ward, and Z. Jane Wang. Image fusion with convolutional sparse representation. *IEEE Signal Processing Letters*, PP(99):1–1, 2016.

[112] Yu Liu, Xun Chen, Rabab K Ward, and Z Jane Wang. Medical image fusion via convolutional sparsity based morphological component analysis. *IEEE Signal Processing Letters*, 26(3):485–489, 2019.

[113] Yu Liu, Shuping Liu, and Zengfu Wang. A general framework for image fusion based on multi-scale transform and sparse representation. *Information fusion*, 24:147–164, 2015.

[114] Yu Liu and Zengfu Wang. Simultaneous image fusion and denoising with adaptive sparse representation. *IET Image Processing*, 9(5):347–357, 2015.

[115] Mingzhu Long, Zhuo Li, Xiang Xie, Guolin Li, and Zhihua Wang. Adaptive image enhancement based on guide image and fraction-power transformation for wireless capsule endoscopy. *IEEE transactions on biomedical circuits and systems*, 12(5):993–1003, 2018.

[116] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61:650–662, 2017.

[117] Ying Lu, Chunming He, Yu-Feng Yu, Guoxia Xu, Hu Zhu, and Lizhen Deng. Vector co-occurrence morphological edge detection for colour image. *IET Image Processing*, 15(13):3063–3070, 2021.

[118] Alan Lukezic, Tomas Vojir, Luka Cehovin Zajc, Jiri Matas, and Matej Kristan. Discriminative correlation filter with channel and spatial reliability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6309–6318, 2017.

[119] Jiayi Ma, Chen Chen, Chang Li, and Jun Huang. Infrared and visible image fusion via gradient transfer and total variation minimization. *Information Fusion*, 31:100–109, 2016.

[120] Jiayi Ma, Linfeng Tang, Meilong Xu, Hao Zhang, and Guobao Xiao. Stdfusionnet: An infrared and visible image fusion network based on salient target detection. *IEEE Transactions on Instrumentation and Measurement*, 70:1–13, 2021.

[121] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020.

[122] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48:11–26, 2019.

[123] Yuhui Ma, Jiang Liu, Yonghuai Liu, Huazhu Fu, Yan Hu, Jun Cheng, Hong Qi, Yufei Wu, Jiong Zhang, and Yitian Zhao. Structure and illumination constrained gan for medical image enhancement. *IEEE Transactions on Medical Imaging*, 40(12):3955–3967, 2021.

[124] Benoît Mathieu, Pierre Melchior, Alain Oustaloup, and Ch Ceyral. Fractional differentiation for edge detection. *Signal Processing*, 83(11):2421–2432, 2003.

[125] F.G. Meyer, A .Z. Averbuch, and R.R. Coifman. Multilayered image representation: Application to image compression. *IEEE Transactions on Image Processing*, 11(9):1072–1080, Sept. 2002.

[126] Elad Michael. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer Publishing Company, Incorporated, 2010.

[127] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a completely blind image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[128] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.

[129] Nick Moran, Dan Schmidt, Yu Zhong, and Patrick Coady. Noisier2noise: Learning to denoise from unpaired noisy data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12064–12072, 2020.

[130] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 445–461. Springer, 2016.

[131] Ghulam Murtaza, Liyana Shuib, Ainuddin Wahid Abdul Wahab, Ghulam Mujtaba, Ghulam Mujtaba, Henry Friday Nweke, Mohammed Ali Algaradi, Fariha Zulfiqar, Ghulam Raza, and Nor Aniza Azmi. Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. *Artificial Intelligence Review*, 53:1655–1720, 2020.

[132] Andriy Myronenko and Ali Hatamizadeh. 3d kidneys and kidney tumor semantic segmentation using boundary-aware networks. *arXiv preprint arXiv:1909.06684*, 2019.

[133] Zhangkai Ni, Wenhan Yang, Shiqi Wang, Lin Ma, and Sam Kwong. Towards unsupervised deep image enhancement with generative adversarial network. *IEEE Transactions on Image Processing*, 29:9140–9151, 2020.

[134] Edward Oczeretko, Marta Borowska, Agnieszka Kitlas, Andrzej Borusiewicz, and Malgorzata Sobolewska-Siemieniuk. Fractal analysis of medical images in the irregular regions of interest. In *2008 8th IEEE International Conference on BioInformatics and BioEngineering*, pages 1–6. IEEE, 2008.

[135] Karen Panetta, Landry Kezebou, Victor Oludare, and Sos Agaian. Comprehensive underwater object tracking benchmark dataset and underwater image enhancement with gan. *IEEE Journal of Oceanic Engineering*, 47(1):59–75, 2021.

[136] Shuchao Pang, Anan Du, Mehmet A. Orgun, Zhenmei Yu, and Guanfeng Liu. Ctumorgan: a unified framework for automatic computed tomography tumor segmentation. *European journal of nuclear medicine and molecular imaging*, 2020.

[137] Simone Parisotto, Luca Calatroni, Aurélie Bugeau, Nicolas Papadakis, and Carola-Bibiane Schönlieb. Variational osmosis for non-linear image fusion. *IEEE Transactions on Image Processing*, 29:5507–5516, 2020.

[138] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning

library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[139] Yi Peng, Deyu Meng, Zongben Xu, Chenqiang Gao, Yi Yang, and Biao Zhang. Decomposable nonlocal tensor dictionary learning for multispectral image denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2949–2956, 2014.

[140] K Ram Prabhakar, V Sai Srikar, and R Venkatesh Babu. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. *2017 IEEE International Conference on Computer Vision*, pages 4724–4732, 2017.

[141] Salil Prabhakar and Anil K Jain. Decision-level fusion in fingerprint verification. *Pattern Recognition*, 35(4):861–874, 2002.

[142] Ori Press, Tomer Galanti, Sagie Benaim, and Lior Wolf. Emerging disentanglement in auto-encoder based unsupervised image content transfer. *arXiv preprint arXiv:2001.05017*, 2020.

[143] Ajinkya M. Pund, Shubham C. Anjankar, Ankush D. Kadu, and Anagha A. Wankhede. A spatial domain feature based approach for no reference image quality assessment of jpeg compressed images. In *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, pages 1–6, 2017.

[144] Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan Yuille. Weight standardization. arxiv e-prints, page. *arXiv preprint arXiv:1903.10520*, 2019.

[145] Tamer Rabie. Adaptive hybrid mean and median filtering of high-iso long-exposure sensor noise for digital photography. *Journal of Electronic Imaging*, 13(2):264–277, 2004.

[146] Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier, Muhammad Salman Khan, et al. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in biology and medicine*, 132:104319, 2021.

[147] P.V. Rama Raju, D. Sudha Rani, and G. Challaram. Comparison of medical image fusion methods using image quality metrics. In *Proc. 2018 Int. Conf. Commun., Comput. Internet Things (IC3IoT)*, pages 449–454, Chennai, IN, Feb. 2018.

[148] Wenqi Ren, Sifei Liu, Lin Ma, Qianqian Xu, Xiangyu Xu, Xiaochun Cao, Junping Du, and Ming-Hsuan Yang. Low-light image enhancement via a deep hybrid network. *IEEE Transactions on Image Processing*, 28(9):4364–4375, 2019.

[149] J Wesley Roberts, Jan A Van Aardt, and Fethi Babikker Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2(1):023522, 2008.

[150] J. Ruf, E. Lopez Hnninen, M. BHmig, I. Koch, T. Denecke, M. Plotkin, J. Langrehr, B. Wiedenmann, R. Felix, and H. Amthauer. Impact of fdg-pet/mri image fusion on the detection of pancreatic cancer. *Pancreatology*, 6(6):512–519, 2006.

[151] Stuart J Russell. *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010.

[152] Alexander W Sauter, Hans F Wehrl, Armin Kolb, Martin S Judenhofer, and Bernd J Pichler. Combined pet/mri: one step further in multimodality imaging. *Trends in molecular medicine*, 16(11):508–515, 2010.

[153] Oliver Schoppe, Chenchen Pan, Javier Coronel, Hongcheng Mai, Zhouyi Rong, Mihail Ivilinov Todorov, Annemarie Müskes, Fernando Navarro, Hongwei Li, Ali Ertürk, et al. Deep learning-enabled multi-organ segmentation in whole-body mouse scans. *Nature communications*, 11(1):1–14, 2020.

[154] H. Seo, C. Huang, M. Bassenne, R. Xiao, and L. Xing. Modified u-net (mu-net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in ct images. *IEEE Transactions on Medical Imaging*, 2019.

[155] Gonglei Shi, Li Xiao, Yang Chen, and S Kevin Zhou. Marginal loss and exclusion loss for partially supervised multi-organ segmentation. *Medical Image Analysis*, 70:101979, 2021.

[156] Munendra Singh, Ashish Verma, and Neeraj Sharma. Optimized multistable stochastic resonance for the enhancement of pituitary microadenoma in mri. *IEEE Journal of Biomedical and Health Informatics*, 22(3):862–873, 2018.

[157] Randal Slates, Simon Cherry, Abdel Boutefnouchet, Yiping Shao, M Dahlborn, and Keyvan Farahani. Design of a small animal mr compatible pet scanner. *IEEE Transactions on Nuclear Science*, 46(3):565–570, 1999.

[158] Jean Luc Starck, Michael Elad, and David Donoho. Redundant multiscale transforms and their application for morphological component separation. *Advances in Imaging & Electron Physics*, 132(04):287–348, 2004.

[159] Long Sun, Zhenbing Liu, Xiyan Sun, Licheng Liu, Rushi Lan, and Xiaonan Luo. Lightweight image super-resolution via weighted multi-scale residual network. *IEEE/CAA Journal of Automatica Sinica*, 8(7):1271–1280, 2021.

[160] M Sundaram, K Ramar, N Arumugam, and G Prabin. Histogram modified local contrast enhancement for mammogram images. *Applied soft computing*, 11(8):5809–5816, 2011.

[161] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022.

[162] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15(1):1–28, 2015.

[163] Wei Ren Tan, Chee Seng Chan, Pratheepan Yogarajah, and Joan Condell. A fusion approach for efficient human skin detection. *IEEE Transactions on Industrial Informatics*, 8(1):138–147, Feb. 2012.

[164] Ming Tang and Jiayi Feng. Multi-kernel correlation filter for visual tracking. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3038–3046, 2015.

[165] Wei Tang, Yu Liu, Juan Cheng, Chang Li, and Xun Chen. Green fluorescent protein and phase contrast image fusion via detail preserving cross network. *IEEE Transactions on Computational Imaging*, 7:584–597, 2021.

[166] Yasuyuki Ueda, Junji Morishita, Shohei Kudomi, and Katsuhiko Ueda. Usefulness of biological fingerprint in magnetic resonance imaging for patient verification. *Medical & biological engineering & computing*, 54(9):1341–1351, 2016.

[167] Uddeshya Upadhyay, Viswanath P Sudarshan, and Suyash P Awate. Uncertainty-aware gan with adaptive loss for robust mri image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3255–3264, 2021.

[168] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017, 2017.

[169] N Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Channappayya, and Swarup S Medasani. Blind image quality evaluation using perception based features. In *2015 Twenty First National Conference on Communications (NCC)*, pages 1–6. IEEE, 2015.

[170] Yan Wang, Yuyin Zhou, Wei Shen, Seyoun Park, Elliot K Fishman, and Alan L Yuille. Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. *Medical image analysis*, 55:88–102, 2019.

[171] Hiroki Watanabe, Koichi Hayano, Gaku Ohira, Shunsuke Imanishi, Toshiharu Hanaoka, Atsushi Hirata, Masayuki Kano, and Hisahiro Matsubara. Quantification of structural heterogeneity using fractal analysis of contrast-enhanced ct image to predict survival in gastric cancer patients. *Digestive Diseases and Sciences*, pages 1–6, 2020.

[172] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2411–2418, 2013.

[173] Lei Xiang, Yang Li, Weili Lin, Qian Wang, and Dinggang Shen. Unpaired deep cross-modality synthesis with fast training. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 155–164. Springer, 2018.

[174] Lingxi Xie, Qihang Yu, Yuyin Zhou, Yan Wang, Elliot K Fishman, and Alan L Yuille. Recurrent saliency transformation network for tiny target segmentation in abdominal ct scans. *IEEE transactions on medical imaging*, 39(2):514–525, 2019.

[175] Guoxia Xu, Xiaoxue Deng, Xiaokang Zhou, Marius Pedersen, Lucia Cimmino, and Hao Wang. Fcfusion: Fractal componentwise modeling with group sparsity for medical image fusion. *IEEE Transactions on Industrial Informatics*, 18(12):9141–9150, 2022.

[176] Guoxia Xu, Hao Wang, Marius Pedersen, Meng Zhao, and Hu Zhu. Sspnet: A siamese-based structure-preserving generative adversarial network for unpaired medical image enhancement. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 1–11, 2023.

[177] Guoxia Xu, Hao Wang, Marius Pedersen, Hu Zhu, and Meng Zhao. Multi-label abdominal image segmentation with partially labeled data: A prototypical consistent learning perspective. In *2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on*

Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), pages 1–7. IEEE, 2022.

[178] Guoxia Xu, Hao Wang, Meng Zhao, Marius Pedersen, and Hu Zhu. Learning the distribution-based temporal knowledge with low rank response reasoning for uav visual tracking. *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[179] Guoxia Xu, Hao Wang, Meng Zhao, and Hu Zhu. Jadd-gan: A joint attention generative adversarial data fusion network for object detection and tracking. In *2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, pages 1829–1836, 2022.

[180] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2020.

[181] Tianyang Xu, Zhen-Hua Feng, Xiao-Jun Wu, and Josef Kittler. Joint group feature selection and discriminative filter learning for robust visual object tracking. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 7950–7960, 2019.

[182] Tianyang Xu, Zhen-Hua Feng, Xiao-Jun Wu, and Josef Kittler. Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking. *IEEE Transactions on Image Processing*, 28(11):5596–5609, 2019.

[183] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10448–10457, 2021.

[184] Bin Yang and Shutao Li. Multifocus image fusion and restoration with sparse representation. *IEEE transactions on Instrumentation and Measurement*, 59(4):884–892, 2009.

[185] Bin Yang and Shutao Li. Pixel-level image fusion with simultaneous orthogonal matching pursuit. *Information fusion*, 13(1):10–19, 2012.

[186] Li Yang. P20-m nci cgems data portal: Sharing data for genome-wide association studies. *Journal of biomolecular techniques: JBT*, 18(1), 2007.

[187] Longshan Yang, Linlin Xu, Junhuan Peng, Yongze Song, Alexander Wong, and David A Clausi. Nonlocal band-weighted iterative spectral mixture model for hyperspectral imagery denoising. *IEEE Transactions on Geoscience and Remote Sensing*, 58(8):5588–5601, 2020.

[188] Ruizhen Yang, Bolun Du, Puhong Duan, Yunze He, Hongjin Wang, Yigang He, and Kai Zhang. Electromagnetic induction heating and image fusion of silicon photovoltaic cell electrothermography and electroluminescence. *IEEE Transactions on Industrial Informatics*, 16(7):4413–4422, Jul. 2020.

[189] Yong Yang, Yue Que, Shuying Huang, and Pan Lin. Multimodal sensor medical image fusion based on type-2 fuzzy logic in nsct domain. *IEEE Sensors Journal*, 16(10):3735–3745, 2016.

[190] Li Yin, Mingyao Zheng, Guanqiu Qi, Zhiqin Zhu, Fu Jin, and Jaesung Sim. A novel image fusion framework based on sparse representation and pulse coupled neural network. *IEEE Access*, 7:98290–98305, 2019.

[191] Ming Yin, Xiaoning Liu, Yu Liu, and Xun Chen. Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain. *IEEE Transactions on Instrumentation and Measurement*, 68(1):49–64, 2018.

[192] Qian Yu, Yinghuan Shi, Jinquan Sun, Yang Gao, Jianbing Zhu, and Yakang Dai. Crossbar-net: A novel convolutional neural network for kidney tumor segmentation in ct images. *IEEE Transactions on Image Processing*, 28(8):4060–4074, 2019.

[193] Yu-Feng Yu, Guoxia Xu, Ke-Kun Huang, Hu Zhu, Long Chen, and Hao Wang. Dual calibration mechanism based l₂, p-norm for graph matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(6):2343–2358, 2021.

[194] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

[195] Hang Zhang and Kristin Dana. Multi-style generative network for real-time transfer. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[196] He Zhang and Vishal M Patel. Convolutional sparse and low-rank coding-based image decomposition. *IEEE Transactions on Image Processing*, 27(5):2121–2133, 2017.

[197] Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1195–1204, 2021.

[198] Jianpeng Zhang, Yutong Xie, Pingping Zhang, Hao Chen, Yong Xia, and Chunhua Shen. Light-weight hybrid convolutional network for liver tumor segmentation. In *IJCAI*, volume 19, pages 4271–4277, 2019.

[199] Liang Zhang, Jiaming Zhang, Peiyi Shen, Guangming Zhu, Ping Li, Xiaoyuan Lu, Huan Zhang, Syed Afaq Shah, and Mohammed Bennamoun. Block level skip connections across cascaded v-net for multi-organ segmentation. *IEEE Transactions on Medical Imaging*, 39(9):2782–2793, 2020.

[200] Tianzhu Zhang, Kui Jia, Changsheng Xu, Yi Ma, and Narendra Ahuja. Partial occlusion handling for visual tracking via robust part matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1258–1265, 2014.

[201] Wei Zhang, Kazuyoshi Itoh, Jun Tanida, and Yoshiki Ichioka. Parallel distributed processing model with local space-invariant interconnections and its optical architecture. *Applied optics*, 29(32):4790–4797, 1990.

[202] Yu Zhang, Yu Liu, Peng Sun, Han Yan, Xiaolin Zhao, and Li Zhang. Ifcnn: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54:99–118, 2020.

[203] Yitian Zhao, Yalin Zheng, Yonghuai Liu, Yifan Zhao, Lingling Luo, Siyuan Yang, Tong Na, Yongtian Wang, and Jiang Liu. Automatic 2-d/3-d vessel enhancement in multiple modality images using a weighted symmetry filter. *IEEE transactions on medical imaging*, 37(2):438–450, 2017.

[204] Shan Zhong, Meng Wei, Shengrong Gong, Kaijian Xia, Yuchen Fu, Qiming Fu, and Hongsheng Yin. Behavior prediction for unmanned driving based on dual fusions of feature and decision. *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[205] Hu Zhu, Haopeng Ni, Shiming Liu, Guoxia Xu, and Lizhen Deng. Tnlrs: Target-aware non-local low-rank modeling with saliency filtering regular-

ization for infrared small target detection. *IEEE Transactions on Image Processing*, 29:9546–9558, 2020.

[206] Hu Zhu, Hao Peng, Guoxia Xu, Lizhen Deng, Yueying Cheng, and Aiguo Song. Bilateral weighted regression ranking model with spatial-temporal correlation filter for visual tracking. *IEEE Transactions on Multimedia*, 2021.

[207] Hu Zhu, Yiming Qiao, Guoxia Xu, Lizhen Deng, and Yu-Feng Yu. Dspnet: A lightweight dilated convolution neural networks for spectral deconvolution with self-paced learning. *IEEE Transactions on Industrial Informatics*, 16(12):7392–7401, 2019.

[208] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[209] Zhuotun Zhu, Yingda Xia, Lingxi Xie, Elliot K Fishman, and Alan L Yuille. Multi-scale coarse-to-fine segmentation for screening pancreatic ductal adenocarcinoma. In *International conference on medical image computing and computer-assisted intervention*, pages 3–12. Springer, 2019.

[210] Peixian Zhuang, Jiamin Wu, Fatih Porikli, and Chongyi Li. Underwater image enhancement with hyper-laplacian reflectance priors. *IEEE Transactions on Image Processing*, 31:5442–5455, 2022.

Published Research Papers

Research Paper I

Guoxia Xu, Hao Wang, Marius Pedersen, Meng Zhao, Hu Zhu: SSP-Net: A Siamese-based Structure-Preserving Generative Adversarial Network for Unpaired Medical Image Enhancement, IEEE/ACM Transactions on Computational Biology And Bioinformatics, 2023.

Research Paper II

Guoxia Xu, Hao Wang, Hu Zhu, Marius Pedersen: Disentangled Spatial-Transformation Guided GAN for Unpaired Medical Image Quality Enhancement, Pending Submission, 2022.

This paper is awaiting publicaton and is not included

Research Paper III

Guoxia Xu, Xiaoxue Deng, Xiaokang Zhou, Marius Pedersen, Lucia Cimmino, Hao Wang: FCFusion: Fractal Component-wise Modeling with Group Sparsity for Medical Image Fusion, IEEE Transactions on Industrial Informatics, 18(12), 9141-9150, 2022.

Research Paper IV

Guoxia Xu, Hao Wang, Meng Zhao, Hu Zhu: JADD-GAN: A Joint Attention Generative Adversarial Data Fusion Network for Object Detection and Tracking, the 20th IEEE International Conference on Smart City(SmartCity-2022), 2022.

Research Paper V

Guoxia Xu, Hao Wang, Meng Zhao, Marius Pedersen, Hu Zhu: Multi-label Abdominal Image Segmentation with Partially Labeled Data: A Prototypical Consistent Learning Perspective, The 7th IEEE Cyber Science and Technology Congress (CyberSciTech 2022), 2022.

Research Paper VI

Guoxia Xu, Hao Wang, Meng Zhao, Marius Pedersen, Hu Zhu: Learning the Distribution-Based Temporal Knowledge with Low Rank Response Reasoning for UAV Visual Tracking, IEEE Transactions on Intelligent Transportation Systems, IEEE, 2022.

ISBN 978-82-326-7250-9 (printed ver.)
ISBN 978-82-326-7249-3 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU
Norwegian University of
Science and Technology