Unsupervised Learning of Depth Estimation from Imperfect Rectified Stereo Laparoscopic Images

Huoling Luo, Congcong Wang, Xingguang Duan, Hao Liu, Ping Wang, Qingmao Hu, Fucang Jia*

Asterisk indicates corresponding author.

H. Luo, Q. Hu, and F. Jia* are with Research Lab for Medical Imaging and Digital Surgery, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, and Shenzhen College of Advanced Technology; University of Chinese Academy of Sciences, Shenzhen, China (correspondence e-mail: fc.jia@siat.ac.cn).

C. Wang is with School of Computer Science and Engineering, Tianjin University of Technology, Tianjin, China, and Department of Computer Science; Norwegian University of Science and Technology, Norway.

X. Duan is with Advanced Innovation Centre for Intelligent Robots & Systems, Beijing Institute of Technology, Beijing, China.

H. Liu is with State Key Lab for Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China.

P. Wang is with Department of Hepatobiliary Surgery, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China.

Abstract— *Objective*: Learning-based methods have achieved remarkable performances on depth estimation. However, the premise of most self-learning and unsupervised learning methods is built on rigorous, geometrically-aligned stereo rectification. The performances of these methods degrade when the rectification is not accurate. Therefore, we explore an approach for unsupervised depth estimation from stereo images that can handle imperfect camera parameters. *Methods*: We propose an unsupervised deep convolutional network that takes rectified stereo image pairs as input and outputs corresponding dense disparity maps. First, a new vertical correction module is designed for predicting a correction map to compensate for the imperfect geometry alignment. Second, the left and right images, which are reconstructed based on the input image pair and corresponding disparities as well as the vertical correction maps, are regarded as the outputs of the generative term of the generative adversarial network (GAN). Then, the discriminator term of the GAN is used to distinguish the reconstructed images from the original inputs to force the generator to output increasingly realistic images. In addition, a residual mask is introduced to exclude pixels that conflict with the appearance of the original image in the loss calculation. *Results*: The proposed model is validated on the publicly available Stereo Correspondence and Reconstruction of Endoscopic Data (SCARED) dataset and the average MAE is 3.054 mm. *Conclusion*: Our model can effectively handle imperfect rectified stereo images for depth estimation.

Keywords— unsupervised learning, depth estimation, stereo matching, laparoscopic image, imperfect rectified stereo images

1. Introduction

Depth estimation from color images, i.e., predicting their per-pixel absolute distances to the camera, is an important topic in the field of computer vision. It has many applications in practice, such as autonomously driving vehicles, robot-assisted surgery, and augmented reality (AR). For AR-based computer-assisted laparoscopic surgery [1], estimating the depth of the surgical site's surface is a critical step. It can serve as a foundation for reconstructing the intraoperative organ surface, and then, the surface

can be used for registration with the preoperative models derived from computed tomography or magnetic resonance image scans.

With the help of two eyes, humans perform well with regard to perceiving the exact positions of objects. Similarly, a stereoscopic vision system can be used to estimate the depth of a scene depending on a stereo pair of images, with the prerequisite that it must find a sufficient number of matched features between the left and right images, *i.e.*, stereo matching. The goal of stereo matching is to compute the disparity d per pixel in the reference image (usually referred to as the left image), where disparity is defined as the horizontal displacement between the corresponding pixels on one horizontal scan line of the rectified stereo image.

Traditional stereo matching algorithms struggle to find enough feature points in textureless, repetitive or highly reflective regions. In contrast, deep learning methods have been approved for learning powerful representations directly from raw color images in many application fields. However, vast amounts of corresponding per-pixel ground truth data are required for training supervised-based learning methods [2, 3]. Special devices, such as light detection and ranging (LiDAR) systems or Microsoft Kinect for real-world scenarios, are required to collect the ground truth data. Whereas, it is not trivial to obtain such data in laparoscopic settings, and the sizes of depth sensors hinder them from reaching the abdominal cavity to collect clinical datasets from real scenes for training. To alleviate this contradiction, some researchers use real laparoscopic images to synthesize ground truth depth data for training purposes [4]. To date, there is no existing human laparoscopic dataset that has associated ground truth depth or disparity datasets for training depth prediction models.

Nevertheless, this dilemma also promotes the development of self- or unsupervised learning-based methods for depth prediction. One family of unsupervised methods treats depth estimation as an image reconstruction problem during the training process. A premise of these methods is that a rectified pair of stereo images is required for feeding the network during the training stage, and this can guarantee that the corresponding pixels lie on the same horizontal scan-line, *i.e.*, the epipolar constraint. Given a pair of rectified stereo color images I_l and I_r, once the disparity d_l of a particular pixel located at (i, j) in the left image is available, its corresponding pixel position in the right image can also be calculated as $(i - d_l, j)$; thus, based on the right color image and the disparity map of the left image, a reconstructed image I'_l can be synthesized and vice versa. By minimizing the reconstructed and original color images, the disparity image can be predicted. Furthermore, according to the camera focal length f and the baseline b between the stereo rig cameras, the depth D can be calculated readily according to the following equation: $D = bf/d$.

However, many state-of-the-art (SOTA) methods struggle to process imperfect rectified stereo images; these difficulties are caused by suboptimal or faulty camera calibration, where the intrinsic and extrinsic parameters may be incorrectly estimated. This is very common when calibrating stereo cameras, as the calibration algorithm [5] is not guaranteed to run without any errors. A pictorial description of an imperfect rectified stereo image is shown in Fig. 1. Under the premise of a perfectly rectified stereo image pair (Fig. 1 (a)), the matching between I_{l1} and I_{r1} (or I_{l2} and I_{r2}) is constrained by a horizontal scan line. The distance d_1 between I_{l1} and point I'_{r1}, which is the location of I_{r1} mapped to the left view, is defined as the disparity of point I_{l1}. When the camera calibration parameters are imperfect (Fig. 1 (b)), the corresponding pixel of one particular pixel in the stereo image pair might not be located on the same horizontal scan-line, namely, it might result in a vertical shift. Therefore, the image reconstruction step of the unsupervised learning model training does not make sense under this situation.

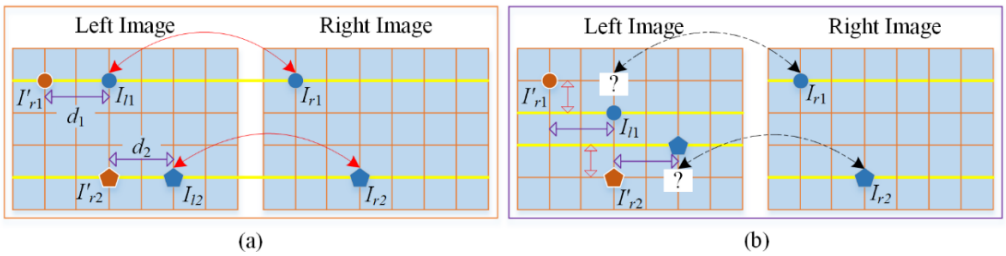

Fig. 1 Illustration of the alignment of corresponding pixels in rectified stereo images: (a) perfect rectified stereo image pair and (b) imperfect rectified stereo image pair

In summary, the main contributions of this paper are as follows:

- Without ground truth labels, we propose an unsupervised depth estimation method for stereo endoscopic images that achieves SOTA performances. The reconstructed and original images are approached within an adversarial learning framework to produce a real appearance, which can estimate accurate disparities indirectly.
- A vertical correction module is proposed to solve the issue of imperfect rectified stereo images, which is a highly critical problem for endoscopic surgery due to the high resolution of these stereo images. To the best of our knowledge, this is the first work in the literature to tackle this problem for learning- based methods in an end-to-end manner. Moreover, superior performance is obtained on imperfect rectified images by the proposed method compared with the performances of existing methods.
- According to the global and local residual of the reconstructed image, a residual mask is introduced for excluding pixels that conflict with the appearance of the original image to prevent contamination in the loss calculation.

2. Related work

For minimally invasive surgery, stereo laparoscopes have been introduced to provide surgeons with a 3D view of the surgical site, and this acts as the fundamental input of stereo algorithms to recover the 3D geometry of the site without using any external devices. We review previous works focusing on depth prediction with deep learning models and their applications in the area of laparoscopic or endoscopic images. These methods can be partially grouped into traditional stereo matching algorithms, supervised and unsupervised learning approaches.

Traditional stereo matching algorithms: Traditional algorithms for searching for connections between pairs of stereo images can be divided into local and global approaches. Stoyanov et al. [6] presented a semi-dense reconstruction method for robotic-assisted surgery by first identifying a set of candidate feature matches as seeds. Zero-mean normalized cross correlation (ZNCC) was used as a dissimilarity measure to effectively cope with poor illumination and low numbers of texture regions. Then, disparity information was propagated around the seeds to reconstruct a semi-dense surface. The initial feature point matching results can influence the performance of this approach. Therefore, to maintain reliable matches, in [7], Bernhardt et al. adopted three strict confidence criteria to find optimal correspondence points and discard the outliers in each pair of images. Although the most reliable matches are kept in this method, for homogeneous surfaces, this method may not find enough matching points. To improve the block matching approach's matching percentage, Penza et al. [8] proposed two methods that follow the traditional approach of the block matching algorithm and census transform and then refined a disparity image using the simple linear iterative cluster (SLIC) algorithm. Although these

methods yield robust results, they rely on the fact that they find corresponding pixels in stereo images, and this assumption is vulnerable to failing when matching points in homogeneous surfaces.

In contrast, global methods have been proposed for dense stereo matching. For example, Chang et al. [9] introduced a dense stereo reconstruction approach by constructing a cost volume and performing convex optimization to solve a Huber-L model. The results demonstrated that the staircasing effect can be overcome to reconstruct a smooth surface in a surgical scene. Similarly, also from the point of global optimization, Wang et al. [10] developed a variational disparity estimation method for minimizing a global energy function over a stereo laparoscopic entire image to reconstruct the surface of a patient's liver. By defining a reasonable cost function with a data term, which is given by the original and gradient images, and a regularization term to constrain the disparity, their method achieved promising results on laparoscopic images. Nevertheless, the consumption of computing resource by their global method is remarkable, especially when processing high-resolution images, as this hinders the method from achieving real-time performance. Moreover, multi-view stereo methods, such as simultaneous localization and mapping (SLAM) [11, 12], can be used to reconstruct 3D structures in real time and estimate the poses of cameras. We refer the reader to the comprehensive review papers [13, 14] regarding 3D surface reconstruction in laparoscopic surgery with traditional stereo imaging for more information.

Supervised learning methods: With the success of deep learning, recent work has progressed in terms of merging learning-based approaches into the traditional stereo algorithm to achieve improved performances. Zbontar et al. [15] presented a Siamese convolutional neural network (CNN) to extract patch-wise features that are then processed by the classic step of computing the stereo matching cost. Then, a series of postprocessing steps are performed to refine the disparity map. Following the work of Eigen et al. [2], many researchers have trained networks using per-pixel ground truth depth data and developed approaches for solving the stereo matching problem end-to-end without any postprocessing [3, 16]. However, as mentioned in the introduction, it is very difficult to obtain enough ground truth data for training a supervised depth estimation model, especially in surgical applications.

Unsupervised learning methods: Recovering depth information from color images in an unsupervised or self-supervised learning manner is an attractive approach. In general, unsupervised estimation approaches can be divided into two different types, including direct and indirect disparity estimation. Deep stereo matching models [17, 18] have been proposed to estimate disparity maps directly. A survey by Scharstein and Szeliski [19] summarized that stereo matching algorithms include four steps: matching cost computation, cost aggregation, disparity computation and disparity refinement. Among these steps, the core task of a stereo algorithm is computing the corresponding pixels in pairs of images.

Different from direct approaches, indirect methods treat disparity (or depth) estimation as an image reconstruction problem. Garg et al. [20] proposed an innovative approach by training a model with pairs of rectified stereo images, where the supervised signal comes from the photometric error between the original input and the reconstructed image generated by the predicted disparity map of the network. Godard et al. [21] followed the same idea of image reconstruction for depth estimation. They solved the problem of [20], in which they could not fully differentiate between images during image reconstruction using bilinear sampling. Moreover, they introduced left-right disparity checking during the training stage and adopted DispNet [22] as the backbone network to achieve SOTA performance. Ye et al. [23] proposed a self-learning framework for surgical scene depth estimation using the same training process as in [20, 21]. Due to the absence of ground truth data, they adopted the structural similarity index (SSI) as the metric for evaluating the accuracy of the predicted disparity map. Recently, Liu et al. [24] presented a self-supervised approach to train a CNN model from monocular endoscopy data, in which they took the

results from the multi-view stereo method as supervised signals to train the depth estimation model. Similarly, Luo et al. [25] fused results from the traditional stereo method into the training stage of their model, in which the traditional stereo method was used to generate proxy disparity labels while their erroneous predictions were removed via a confidence measure. Then, these generated proxy disparity labels acted as auxiliary supervised signals to train the depth estimation model. To promote the precision of depth estimation, Mahjourian et al. [26] introduced validity masks to avoid penalizing areas that could not be seen in both views. Godard et al. [27] proposed a novel auto-masking approach to ignore pixels in training frames where no relative camera motion occurs.

In recent years, generative adversarial networks (GANs) [28] have been employed for unsupervised depth prediction and have shown promising performance [29]. The GAN framework was first proposed by Goodfellow et al. [28] and was based on the idea of training two subnetworks, a generator and a discriminator, to compete with each other simultaneously in a game. Since then, GANs have been applied to various generation tasks, one of which is the generation of laparoscopic/endoscopic images [4, 30] for certain target tasks. Pilzer et al. [29] integrated GANs into an indirect unsupervised depth prediction method and presented an architecture consisting of two generative subnetworks jointly trained with adversarial learning for the purpose of estimating disparity maps. However, similar to other indirect unsupervised depth prediction approaches, their approach requires rectified stereo images as input to constrain the search for corresponding pixels to the same horizontal scan-line, videlicet, it fails to estimate reasonable disparities using stereo images with imperfect rectification. To solve the problem of imperfectly rectified stereo images, Nguyen et al. [31] presented a set of modified matching cost functions, such as the absolute difference (AD), squared difference (SD) and ZNCC, for stereo matching methods. While they attempted to address the imperfect rectification problem in a traditional manner, we propose a trainable vertical correction module and insert it into a GAN framework for unsupervised disparity estimation; this is an end-to-end, simple yet effective and efficient deep learning-based method.

3. Proposed methods

In this section, the problem to be solved is first introduced, and then an overview of the proposed approach is presented. Next, we introduce the proposed vertical correction module, which addresses imperfect rectified stereo images, the residual mask, which excludes high bias regions during training, and the loss function of the network.

3.1 Problem Formulation

The aim of this work is to predict the disparity maps of surgical sites from stereo laparoscopic image pairs. Given a pair of rectified stereo images (I_l, I_r), our goal is to learn a function f_θ that can predict the per-pixel disparity $d = f_\theta(I_l, I_r)$, where θ denotes a set of parameters.

Most supervised methods attempt to estimate f_θ such that the estimated disparity d is as close to the ground truth disparity $\hat{d}$ as possible. In other words, the loss function $L(I_l, I_r) = l(f_\theta(I_l, I_r), \hat{d})$ is minimized, where l is a measure of the distance between the estimated disparity d and the ground truth disparity $\hat{d}$. To supervise the training of the network without knowing $\hat{d}$, an image reconstruction-based self-supervision strategy is adopted.

Once the disparity d_l of the left image is estimated, applying it to the right image would enable us to reconstruct a new left image $\hat{I}_l = R(I_r, d_l)$, where R indicates the reconstruction function. Similarly, the reconstructed right image can also be obtained by $\hat{I}_r = R(I_l, d_r)$. If the estimated disparity d is close to the ground truth $\hat{d}$, the discrepancies between the reconstructed image $\hat{I}_l/\hat{I}_r$ and the original image I_l/I_r can be minimized. In this work, for our research problem (see Fig. 1 (b)) in which the

rectified stereo image pair may have vertical pixel shifts, we define a new variable V to address the vertical shift of every pixel. Therefore, the reconstruction is redefined as $\hat{I}_l = R(I_r, d_l, V_l)$ and $\hat{I}_r = R(I_l, d_r, V_r)$, where V_l and V_r denote the vertical shifts of the left and right images, respectively. Finally, the loss function of our problem is defined based on the reconstruction residual between the reconstructed image $(\hat{I}_l, \hat{I}_r)$ and the original image (I_l, I_r), which can be written as $L(I_l, I_r) = l(\hat{I}_l, \hat{I}_r, I_l, I_r)$ and is utilized as a supervised signal for optimizing the network to generate an accurate disparity.

In addition, to further improve the accuracy of the disparity, a GAN framework is adopted. In this way, the combined disparity estimation and image reconstruction process is regarded as a generative network G that takes a rectified stereo image pair as input and outputs the reconstructed images. The reconstructed images and the corresponding reference images are set as the input of the discriminator. The discriminator can help to distinguish the reconstructed image from the real image, and so it can further supervise the network to generate reconstructed images that are as real as possible.

3.2 Network Architecture

An overview of the proposed network structure is depicted in Fig. 2. It can be roughly divided into two parts: a generative network and a discriminator network. Specifically, the network takes a rectified color stereo image pair I_l and I_r as input. The first stage is the performance of feature extraction on the input stereo images based on a ResNet [32] backbone network, and a feature map with 1/4 the size of the input image is then output. The two branches of the feature extraction network representing the left and right images share weights. The second stage is the construction of the cost volumes, and the previous left and right feature maps are first concatenated to form a five-dimensional cost volume for each view. The third stage is disparity estimation, during which serial three-dimensional convolutions are applied to the cost volumes to generate a coarsely-estimated initial disparity map. Moreover, this initial disparity map is then set as the input of a refinement module to acquire an accurate disparity map. The initial and refined disparity maps are fused with a 1x1 convolution operation to form the final disparity estimation.

However, as presented and discussed previously (Fig. 1), the corresponding pixels of the rectified stereo image pair may not lie on the same horizontal scan-line due to imperfect camera parameters. As a result, the disparity estimated from the third stage would tend toward an unreasonable matching under this situation (see the results of Fig. 6); thus, we introduce a vertical correction module (VCM) to tackle this issue.

Four different scales of the feature map from the feature extraction network for each camera view are set as inputs of the left and right VCM branches. Then, a subsequent two-dimensional convolution and deconvolution operation is applied on these multiscale feature maps to produce correction maps, V_l and V_r, for each view. The pixel value in the correction map represents the vertical shift distance to compensate for the nonhorizontal alignment effect caused by imperfect stereo rectification. To achieve a fine-grained disparity map, the left and right correction maps (V_l and V_r), the estimated disparity maps from the third stage (d_l and d_r), and the input stereo image pair (I_l, I_r) are sent to the fourth stage, where the images corresponding to the left and right views ($\hat{I}_l$, $\hat{I}_r$) can be reconstructed based on the definitions of disparity and bilinear interpolation. Finally, these reconstructed images ($\hat{I}_l$, $\hat{I}_r$), along with their corresponding original stereo images (I_l, I_r), are fed into a discriminator network $\mathfrak{D}$ to distinguish whether they are true or fake. In this way, the discriminator network can improve the disparity estimation stage and output a highly realistic disparity map.

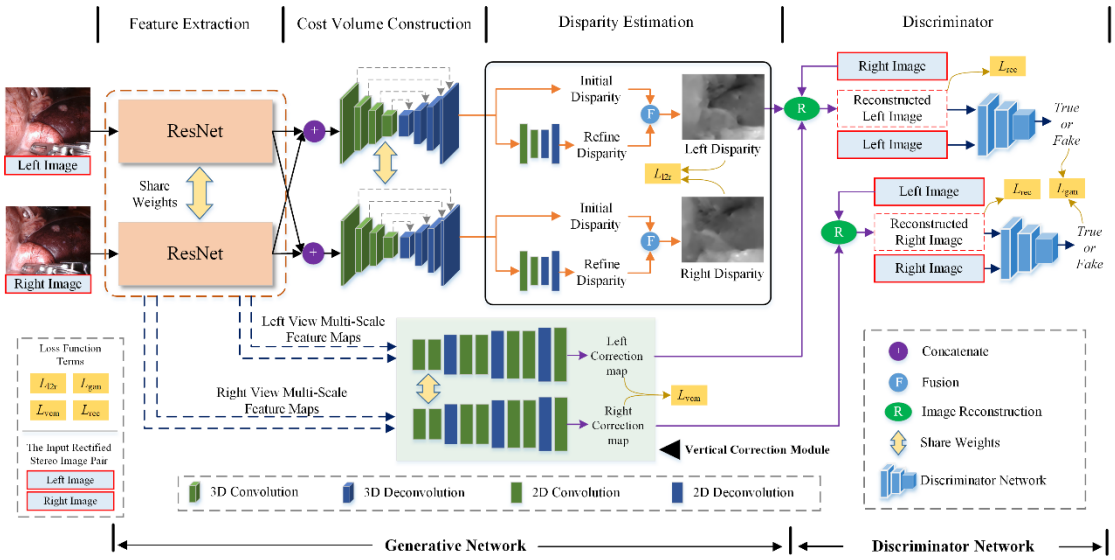

Fig. 2 Illustration of our proposed unsupervised depth estimation network. The disparity is first estimated from rectified stereo images. Then, the image reconstruction module is utilized to estimate synthetic left/right images, which are computed by the estimated right/left disparities, the proposed vertical correction module and the corresponding reference left/right images. The reconstruction-based losses and the discriminator are applied, acting as the supervision signal to train the network

3.3 Vertical Correction Module

Additional details regarding the vertical correction module are presented in this section.

Fig. 3 shows an example of an imperfect rectified stereo image pair from the SCARED [33] dataset. It can be seen distinctly that the same feature point is not located on the same horizontal scan line even when rectification is performed. Specifically, the corresponding pixels shift a certain distance in the vertical direction. This conflicts with the requirements of many SOTA unsupervised depth methods, and thus, those methods cannot estimate the disparity accurately.

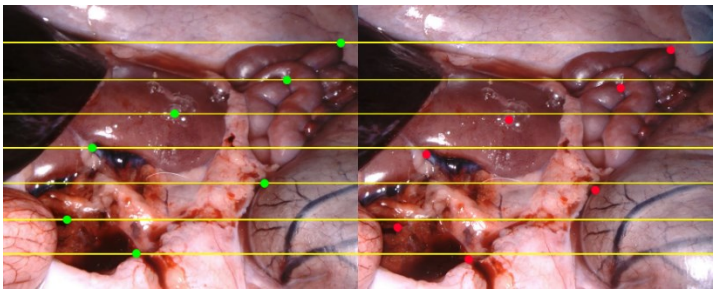

Fig. 3 An example of an imperfect rectified stereo image taken from the SCARED dataset. The yellow line represents the same scan line, and the dot markers are the same feature points (green in the left view and red in the right view)

Based on this observation, we propose a trainable vertical correction module to learn the shift distance of every pixel in the vertical direction, which can be used to reconstruct realistic images through bilinear resampling and drive the disparity prediction module to output an accurate disparity map. To fully utilize the feature information extracted from the input image, the feature extraction module outputs four different scales of the feature map for each view, and these are set as the input of the VCM. As shown in Fig. 4, first, we apply a convolution operation with a kernel of size 3×3 and concatenate feature maps with the same size. After the concatenation operation, convolution and batch normalization are used before the deconvolution operation. Finally, we employ a convolution with a sigmoid activation

function to output the correction map.

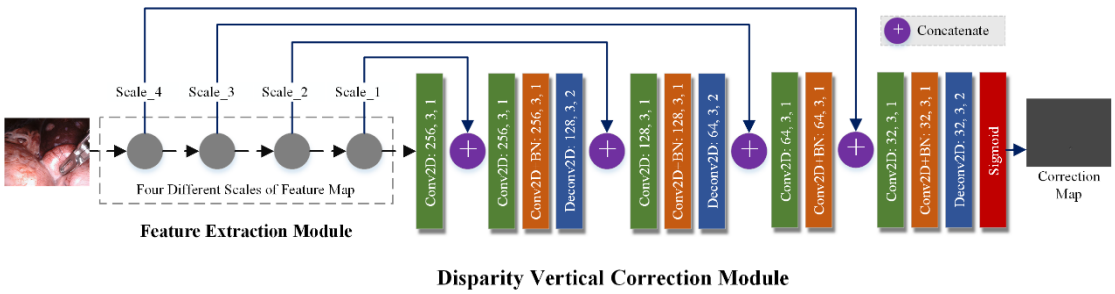

Fig. 4 Structure of the vertical correction module (VCM) for disparity. Different colors denote different operations, commented with the name of the operation, number of filters used, kernel size and stride

3.4 Residual Mask

To exclude the regions that are not visible due to parallax in both stereo views during training, we follow the same strategy as that in [26] to generate the validity mask. In [26], Mahjourian et al. validated that such pixels degrade the performance of the model if they participate in the loss calculation. Moreover, we propose residual masks based on the local and global residuals to ignore pixels for which their reconstruction residuals exceed a preset threshold.

Considering the original image (either the left image or the right image) I and the reconstructed image $\hat{I}$, for one particular pixel position (i, j), the local residual can be defined as $\delta_{ij} = I_{ij} - \hat{I}_{ij}$, and the global residual $\varphi = \frac{1}{\Omega}\sum_{(i,j)\in\Omega} |I_{ij} - \hat{I}_{ij}|$, where Ω denotes the whole region of the image. We assign a weight W_{ij} per pixel according to the local residual δ_{ij} and the global residual φ with $W_{ij} = \exp[-\delta_{ij}/(\varphi + \sigma)]$, where σ is a small constant value to prevent division by zero. The global residual φ is deterministic for a reconstructed image $\hat{I}$, so the per-pixel weight is inversely proportional to the local residual δ_{ij}; thus, we define a residual mask as:

$$M_{res\_ij} = \begin{cases} 0, & W_{ij} < \tau \\ 1, & W_{ij} \geq \tau \end{cases}, \qquad (1)$$

where τ is a preset constant value. We set $\tau = 0.95$ in our experiments. Furthermore, a validity mask M_{valid} can be obtained according to [35], so the final residual mask M per pixel is given by:

$$M = M_{valid} * M_{res} \quad (2)$$

An example of a validity mask [26] and our proposed residual mask is illustrated in Fig. 5, and the pixels in the white region of the mask are used to calculate the final loss.

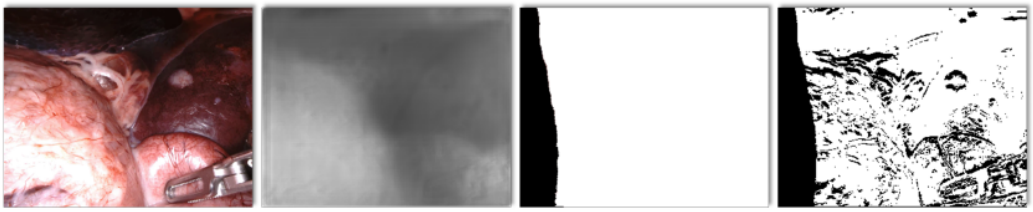

Fig. 5 Example of a validity mask and our proposed residual mask: (a) the rectified left view image, (b) the predicted disparity map of our model, (c) validity mask by [26] and (d) our proposed residual mask

3.5 Loss Function

We introduce the definition of the loss function in this section. The loss function used for our unsupervised depth estimation network consists of four different terms.

Image reconstruction term. The network tries to learn to estimate the disparity map for one view,

and based on this disparity, it generates an image by sampling pixels from the opposite stereo image. Therefore, the more accurate the disparity map is, the more similar the original image and reconstructed image are. Under this hypothesis, the image reconstruction appearance loss term can be defined with the L_1 norm as $L_{rec1} = \|I_l - \hat{I}_l\| + \|I_r - \hat{I}_r\|$, where $\| \|$ is the L_1 norm operation. Furthermore, we adopt the loop consistency constraint [34] for the image photometric loss, *i.e.*, given an image, we can reconstruct its two versions by using the estimated disparity map and the opposite stereo image. Specifically, the reconstructed left image $\hat{I}_l$ is generated by sampling the right image I_r with disparity d_l. On the other hand, based on the left image I_l and the right disparity map d_r, the right image $\hat{I}_r$ can be synthesized. Then, we sample the synthesized right image $\hat{I}_r$ with the disparity map d_l, and a new version of the left image $\hat{I'}_l$ can also be constructed. Thus, the loop consistency term is given as $L_{rec2} = \|I_l - \hat{I'}_l\|$. Therefore, the image reconstruction term is defined as:

$$L_{rec} = L_{rec1} + L_{rec2}. \qquad (3)$$

Left-right disparity consistency term. Similiar to the loop consistency constraint, we can synthesize a new version of the left disparity map d'_l by sampling the right disparity map d_r with the estimated disparity map d_l. Following the same approach, the new disparity map d'_r matched to the right image can also be generated. Thus, the left-right disparity consistency term is defined as:

$$L_{l2r} = \|d_l - d'_l\| + \|d_r - d'_r\|. \qquad (4)$$

Vertical shift constraint term. Due to the presence of imperfect rectified stereo images, we introduce the VCM to learn the per-pixel vertical shift. For one particular pixel, the shift direction of the left view is reversed to that of the right view (see Fig. 1). According to this observation and denoting the consistency check as L_{vcm}, we define the vertical shift constraint loss as:

$$L_{vcm} = \|\|V_l| - |V_r|\|, \qquad (5)$$

where V_l and V_r denote the left and right vertical shift maps, respectively.

Generative adversarial term. To benefit from the advantage of the adversarial learning strategy and to generate a highly realistic image, the reconstructed image $\hat{I}_l$ and the true left image I_l are fed into a generative subnetwork $\mathfrak{D}_l$ to discriminate whether $\hat{I}_l$ is fake or true, and the subnetwork outputs a scalar value. The same is true for $\hat{I}_r$ and I_r. Following the formulation of [28], we formulate the generative adversarial term as follows:

$$L_{gan} = \mathbb{E}_{I_l \sim p(I_l)}[\log \mathfrak{D}_l(I_l)] + \mathbb{E}_{I_r \sim p(I_r)}\left[\log\left(1 - \mathfrak{D}_l(\hat{I}_r)\right)\right] +$$

$$\mathbb{E}_{I_r \sim p(I_r)}[\log \mathfrak{D}_r(I_r)] + \mathbb{E}_{I_l \sim p(l)}[\log\,(1 - \mathfrak{D}_r(\hat{I}_l))], \qquad (6)$$

where the cross-entropy loss is adopted to measure the expectations of I_l and I_r against the distributions of $p(I_l)$ and $p(I_r)$, respectively.

The full loss functions. For the image reconstruction, left-right disparity consistency and vertical shift constraint loss terms, we multiply each of them by the residual mask to exclude invalid or high residual pixels and prevent them from contaminating the loss calculation. Thus, L_{rec}, L_{l2r} and L_{vcm} are rewritten as follows:

$$L_{rec} = \|M_l \cdot (I_l - \hat{I}_l)\| + \|M_r \cdot (I_r - \hat{I}_r)\| + \|M_l \cdot (I_l - \hat{I'}_l)\| \quad (7)$$

$$L_{l2r} = \|M_l \cdot (d_l - d'_l)\| + \|M_r \cdot (d_r - d'_r)\| \quad (8)$$

$$L_{vcm} = \|M_l \cdot M_r \cdot (|V_l| - |V_r|)\|. \qquad (9)$$

The full loss function can be written as:

$$L = \alpha_1 L_{rec} + \alpha_2 L_{l2r} + \alpha_3 L_{vcm} + \alpha_4 L_{gan}, \qquad (10)$$

where $\alpha_i, i = 1, \ldots 4$ is the weight for balancing different loss terms.

4. Experiments

Based on efforts by the Intuitive Surgical Inc., the stereo correspondence and reconstruction of endoscopic data (SCARED) dataset [33] was created for depth estimation from laparoscopic images. The SCARED contains images of the abdominal anatomy of in vivo pigs, which were obtained by using a da Vinci Xi endoscopy and a projector. There are nine sub-datasets from different pigs, including seven sub-datasets for training and two for evaluating the performance of the proposed method. Within each sub-dataset, there are five keyframes: the associated camera calibration file and camera transformation matrix of every frame relative to its initial camera position, as well as the video file and point clouds seen by the left camera and the right camera, respectively.

We use the OpenCV library (https://opencv.org) to extract frames from the video file and divide the original video frame into left and right views (the top half for the left view and the bottom half for the right view) with a 1280×1024 pixel resolution for the monocular frame, and rectification is also performed during the process. The experiments are performed based on the SCARED dataset. We downsample the image to 1/4 of its original size, 320×256 pixels, for training.

4.1 Implementation Details

Our model is implemented based on [18] using TensorFlow [35]. A ResNet-50 backbone network is adopted for feature extraction, and the left image and right image feature extraction branches share weights. Furthermore, four different scales of the feature map from the feature extraction backbone network are set as the inputs of the VCM. The left and right feature maps from the feature extraction module are concatenated to form a five-dimensional cost volume in the form of [batch_size, max_disparity_number, height, width, feature_map_number], with one cost volume for each view. In our experiments, we set batch_size = 2, max_disparity_number = 128 and feature_map_number = 128, and the image size [height, width] for the cost volumes is resized to 1/4 of the size of the input training sample.

Then, a three-dimensional convolution is applied to these two cost volumes to output the initial estimated disparity map. This initial disparity map is then fed into a refinement network to obtain an accurate disparity map. The initial and refined disparity maps are fused with a 1×1 convolution to achieve the final disparity map. For the discriminators $\mathfrak{D}_l$ and $\mathfrak{D}_r$, we employ the same network structure as that in [29], which has five consecutive convolutional operations followed by batch normalization. Each convolution is performed with a kernel size of 3, a stride size of 2 and a padding size of 1. The implementation of the vertical correction module is shown in Fig. 4. All the convolution operations in the VCM have the same kernel size, stride size and padding size. We set a vertical shift constant (in our experiments, this is set as 5) for the last convolution, which uses the sigmoid function as the activation function, to control the correction range.

For the reconstruction and warping operations, a bilinear sampler is used as [21], and we modify the bilinear sampler by combining it with the vertical shift. The parameters of the employed loss function are $\alpha_1 = 1.0$, $\alpha_2 = 0.001$, $\alpha_3 = 0.1$, and $\alpha_4 = 0.0001$.

4.2 Experimental Setup

We train our model with a standard training procedure by initializing the network with random weights and training it for 5 epochs. Three types of dataset splits are performed to verify the performance of our proposed model and those of other methods. One split is done to obtain a pure imperfect dataset, one split follows the instructions of the MICCAI 2019 SCARED challenge, and the third split is a random split of the dataset.

Pure imperfect dataset split (denoted as split_1) We check the SCARED dataset thoroughly and find that subdatasets 4 and 5 are imperfect rectified stereo images. The distribution of the training

sample is shown in Table 1. We use dataset 4 and keyframe 1 to keyframe 3 in dataset 5 as the training sample and keyframe 4 and keyframe 5 in dataset 5 as the testing sample. There are 5341 stereo image pairs for training and 413 stereo pairs for testing in *split_1*.

Table 1 Training sample distribution of pure imperfect rectified stereo images in SCARED

	K1	K2	K3	K4	K5
D4	729	541	408	349	1
D5	198	1674	1441	**412**	**1**

D denotes dataset, K denotes keyframe

MICCAI 2019 SCARED sub-challenge dataset split (denoted as split_2) In this dataset split, we follow the requirements of the MICCAI 2019 Stereo Correspondence and Reconstruction of Endoscopic Data Sub-Challenge, which is part of the Endoscopic Vision Challenge. Sub-datasets 1 to 7 are used for training, sub-datasets 8 and 9 are the testing samples, and there are 22,985 stereo pairs in the training data and 5,925 stereo pairs in the testing data.

Random dataset split (denoted as split_3) In this split, we randomly select a portion of samples from dataset 1 to dataset 7 from the SCARED as the testing dataset (5%) and the rest as the training dataset (95%). Similiar to *split_1*, we compare our method to the four previously introduced SOTA unsupervised methods (see Fig. 9) in this split.

4.3 Evaluation Metrics

To evaluate the performance of the proposed method, we compare it with four SOTA unsupervised methods [18, 21, 29, 36]. The evaluation is implemented on all three data splits. In addition, we compare our method with the three top-performing methods, which can represent the SOTA performances for stereo matching on endoscopic data.

Three popular measures are utilized to quantitatively evaluate the reconstruction results. The point clouds of each camera frame are saved in the tiff file format for the SCARED dataset. The ground truth of the depth value for every pixel can be retrieved from these files. We adopt the mean absolute error (MAE) as the evaluation metric for accuracy, and this is the same approach as in the SCARED sub-challenge. The MAE is defined as in equation (11).

$$\mathrm{MAE} = \frac{1}{N}\sum_{(i,j)\in\Omega} |\hat{z}_{i,j} - z_{i,j}| \quad , \qquad (11)$$

where $\hat{z}$ is the ground truth depth, z is the predicted value, and N is the number of valid pixels in region Ω. We also use the root mean square error (RMSE) and the point-to-point mean absolute distance error (MADE) for the accuracy evaluation of three-dimensional reconstructed points, both of which are defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i,j\in\Omega}\left(\hat{z}_{i,j} - z_{i,j}\right)^2}, \qquad (12)$$

$$\mathrm{MADE} = \frac{1}{N}\sum_{(i,j)\in\Omega}\sqrt{\left[\hat{x}_{i,j} - x_{i,j}\right]^2 + \left[\hat{y}_{i,j} - y_{i,j}\right]^2 + \left[\hat{z}_{i,j} - z_{i,j}\right]^2}. \qquad (13)$$

4.4 Experimental Results

In this section, the qualitative and quantitative experimental results, which include four parts, are demonstrated and analyzed. First, we compare our proposed method to other unsupervised methods on an imperfect dataset split to illustrate the performance of the proposed vertical correction module. Then, we compare our method to the SOTA methods on the SCARED challenge split. Third, we also compare our method with other unsupervised methods on a random split of the dataset. Finally, we conduct

ablation studies on the network to analyze the effectiveness of the GAN framework and the proposed VCM.

Results on Imperfect Datasets (split_1) Table 2 illustrates the quantitative results of different unsupervised depth estimation methods based on the training process with pure imperfect rectified image pairs. Our model can dramatically improve upon the performances of the other methods under imperfect rectified stereo image pairs. For example, a 27% improvement is achieved on keyframe 4 over Godard et al.'s method [21] in terms of MAE. In addition, the qualitative results in Fig. 6 clearly show that when training based on the imperfect dataset split, the reconstruction-based indirect method cannot produce meaningful estimation results. By introducing a vertical correction module and adversarial learning, our proposed method can estimate smoother and more accurate disparities than those of other methods.

Table 2 Quantitative results based on the pure imperfect training/testing dataset split

		Godard [21]	Wong [36]	Pilzer [29]	Skinner [18]	**Ours**
D5	K4	7.92	12.34	8.94	10.20	**5.82**
	K5	2.88	6.21	6.00	4.59	**1.54**
MAE (mm)		5.40	9.28	7.47	7.40	**3.68**

To further investigate the estimated depth accuracy for a particular pixel, we resample a portion of the grid in one testing frame and plot the depth values for comparison. Fig. 7 shows the predicted depth values and ground truths of the sampled pixels. Intuitively, the proposed method demonstrates superior performance.

Results on the MICCAI 2019 SCARED Sub-Challenge (split_2) To further evaluate the performance of our proposed method, we follow the instructions of the MICCAI 2019 SCARED sub-challenge and compare the proposed method with the top 3 existing methods in this challenge. Table 3 reports the results over different keyframes in the SCARED datasets. The proposed unsupervised method achieves comparable performance to those of the SOTA methods, and small differences in MAE are observed. The visual results predicted from test datasets of our method are presented in Fig. 8. For visualization purposes, we convert the estimated disparity map to a point cloud based on the given camera parameters, and this indicates promising performance.

Table 3 Results of our proposed method and the baseline method based on the requirements of MICCAI 2019 Stereo Correspondence and Reconstruction of Endoscopic Data sub-challenge

	D8 (test dataset 1)					D9 (test dataset 2)					MAE
	K1	K2	K3	K4	K5	K1	K2	K3	K4	K5	
Trevor Zeffiro	7.91	2.97	1.71	2.52	2.91	5.39	1.67	4.34	3.18	2.79	**3.54**
J.C. Rosenthal	8.25	3.36	2.21	2.03	1.33	8.26	2.29	7.04	2.22	0.42	3.74
Congcong Wang	6.30	2.15	3.41	3.86	4.80	6.57	2.56	6.72	4.34	1.19	4.19
Dimitris Psychogyios 1	7.73	2.07	1.94	2.63	0.62	4.85	0.65	1.62	0.77	0.41	2.33
Dimitris Psychogyios 2	7.41	2.03	1.92	2.75	0.65	4.78	1.19	3.34	1.82	0.36	2.63
Sebastian Schmid	7.61	2.41	1.84	2.48	0.99	4.33	1.10	3.65	1.69	0.48	2.66
Godard et al [21]	39.00	21.86	21.06	28.05	14.78	23.70	7.99	14.44	2.67	10.24	18.38
Pilzer et al [29]	19.77	30.07	30.29	16.85	57.74	31.00	20.87	13.54	22.59	73.43	31.62
Wong et al [36]	39.53	24.85	20.20	25.24	12.37	29.33	11.95	18.60	4.40	13.99	20.05
Skinner et al [18]	8.95	2.83	2.41	2.46	2.61	6.10	0.99	3.08	0.90	1.27	3.16
Ours	8.62	2.69	2.36	2.29	2.51	6.06	0.95	2.97	0.86	1.23	**3.05**

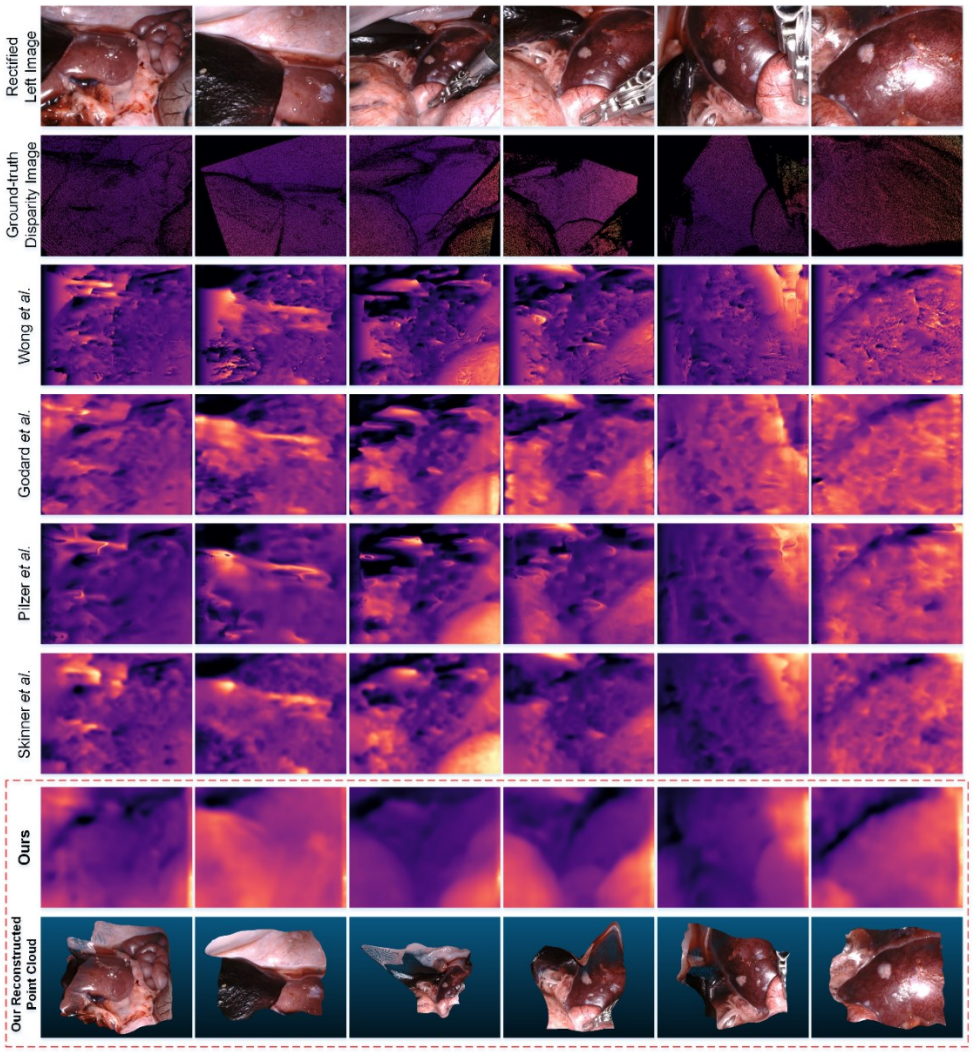

Fig. 6 Visual comparison between other unsupervised methods and ours. For pure imperfect rectified stereo images, our results have the smoothest and most reasonable surfaces

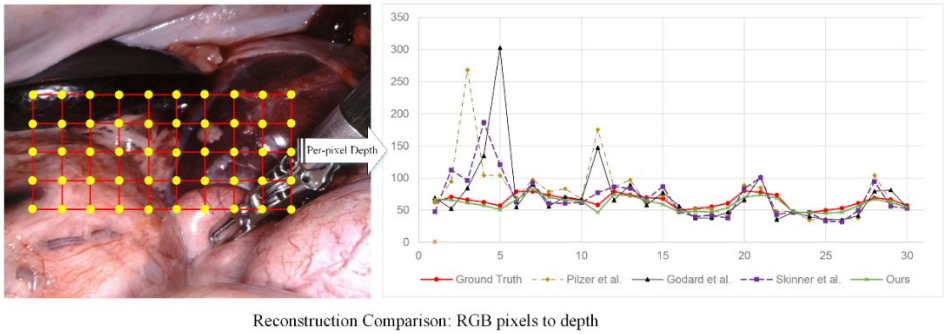

Fig. 7 Comparison of per-pixel predicted depth values. The points where the depth values are zero or NaN in the ground truth are excluded from the sample position (the red line represents the ground truth)

Results on the Random Dataset Split (split_3) Table 4 lists the quantitative results on dataset *split_3* obtained by different unsupervised learning methods. It can be found that the proposed method achieves the best performance. The absolute improvements of the proposed method over the second-best method are 7%, 7% and 5% in terms of the MAE, RMSE, and MADE, respectively. Some reconstruction

results of each method are illustrated in Fig. 9. The performances of the other methods are degraded for challenge cases, as highlighted by the red rectangle. The proposed method can still perform well under those challenging cases, thereby indicating that the robustness of this method is higher than that of the other methods.

Table 4 Prediction errors on *split_3* obtained by the unsupervised methods

Method	MAE (mean±STD)	RMSE (mean±STD)	MADE (mean±STD)
Wong et al.	6.24±6.69	11.38±9.13	8.24±6.55
Godard et al.	5.46±6.05	10.23±8.69	7.42±5.94
Pilzer et al.	5.97±6.74	10.33±9.49	7.92±6.60
Skinner et al.	5.76±6.63	9.96±9.34	7.71±6.54
Ours	**5.05±4.47**	**9.21±5.46**	**7.06±4.91**

Ablation Studies To investigate the different network modules proposed in our method, we perform an ablation study by adding each module to the baseline network and evaluating them on the *split_1* and *split_2* data splits. The performances of different combinations of the proposed modules are shown in Table 5 and Table 6.

From Table 5, we can observe that the GAN and VCM improve the results by 27% and 29%, respectively, over that of the baseline method. A combination of the GAN and VCM increases the MAE score by 50%, and this confirms the effectiveness and benefits of the proposed method and the proposed VCM for imperfect rectified datasets. From Table 6, we can observe that the GAN framework, the VCM module and their combination can slightly improve upon the baseline methods, with absolute improvements of 2.6%, 2.8%, and 3%, respectively. Noticeably, the proposed VCM is designed for imperfect rectified datasets.

Table 5 Ablation study on the *split_1* data split

		Baseline	+GAN	+VCM	+GAN/VCM
D 5	K 4	10.02	8.12	8.10	**5.82**
	K 5	4.59	2.48	2.29	**1.54**
MAE (mm)		7.31	5.30	5.20	**3.68**

Table 6 Ablation study on the *split_2* data split

		Baseline	+GAN	+VCM	+GAN/VCM
D 8	K_0	8.95	8.68	8.85	8.62
	K_1	2.83	2.72	2.75	2.69
	K_2	2.41	2.34	2.26	2.36
	K_3	2.46	2.31	2.28	2.29
	K_4	2.61	2.67	2.48	2.51
D 9	K_0	6.10	6.04	5.98	6.06
	K_1	0.99	0.96	0.95	0.95
	K_2	3.08	2.97	3.10	2.97
	K_3	0.90	0.87	0.88	0.86
	K_4	1.27	1.22	1.19	1.23
MAE (mm)		3.16	3.08	3.07	**3.05**

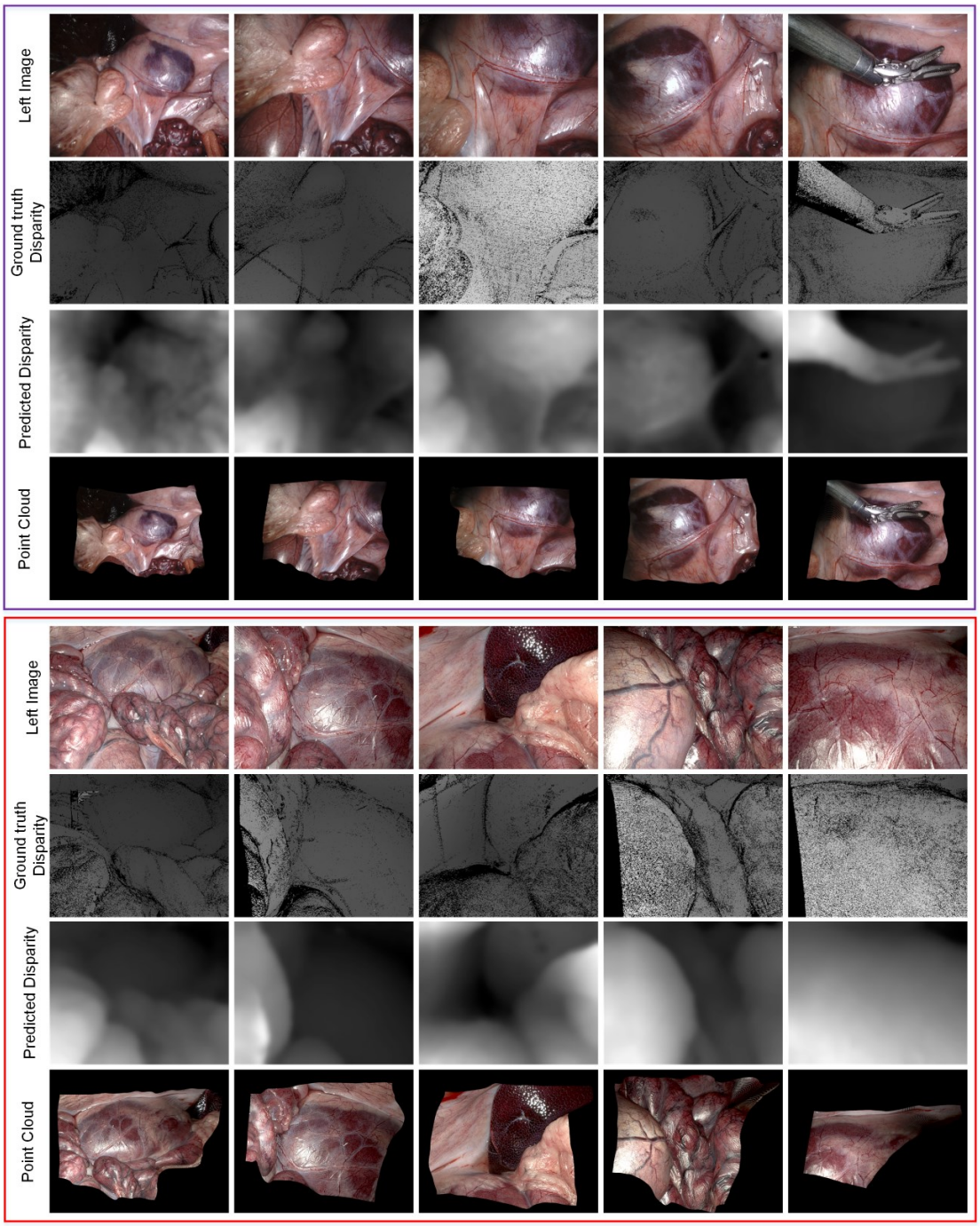

Fig. 8 Qualitative results of the compared unsupervised depth estimation methods and ours

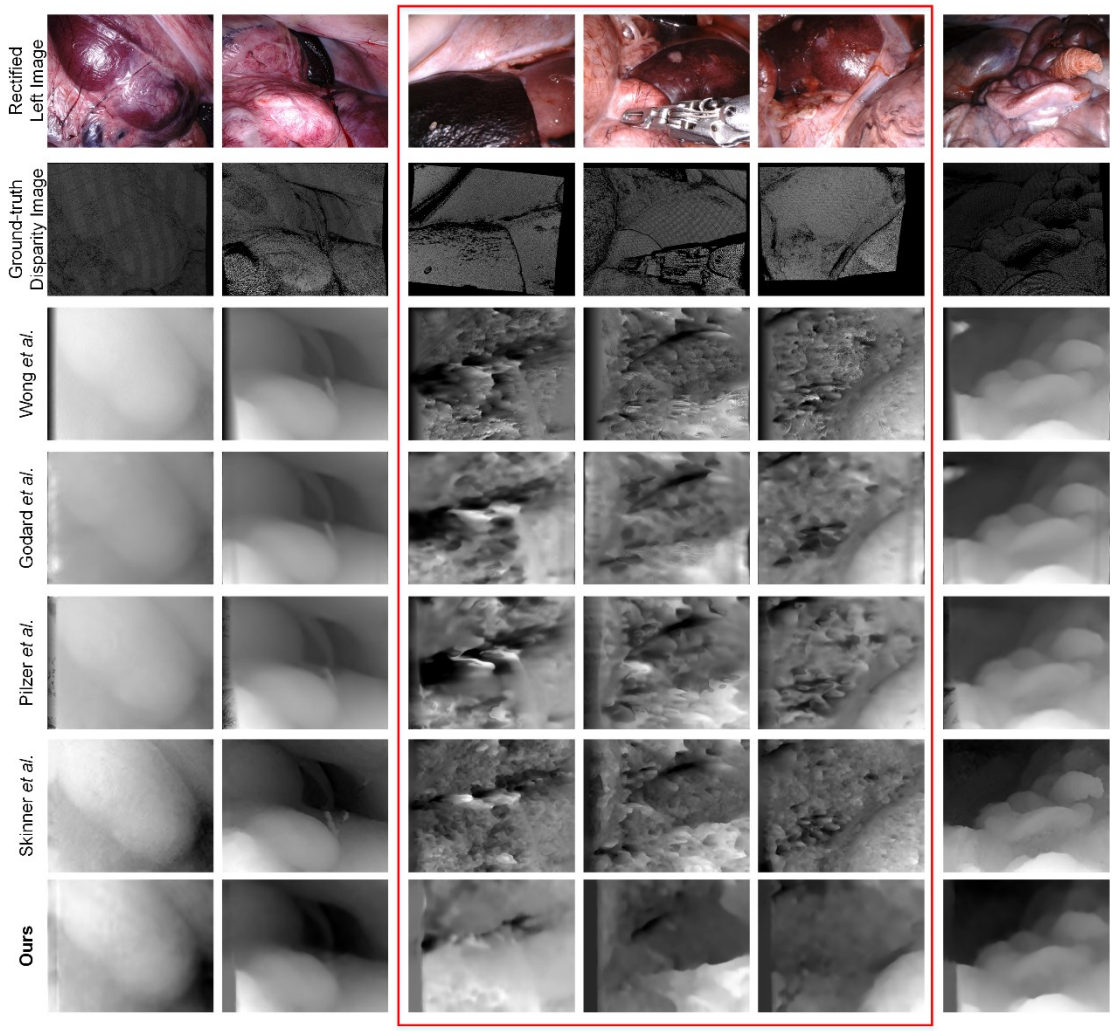

Fig. 9 Qualitative results of compared unsupervised depth estimation methods and ours

5. Discussion

The comparison methods [21, 29, 36] in Table 2 are a family of unsupervised networks for depth prediction that share the same style, and they require rectified stereo images for training. The common premise is that the same feature point must be located on the same horizontal scan line, and this is a similar condition to that of the stereo method in [18]. However, it is obvious that the *split_1* dataset (see Fig. 3 for more information) cannot meet this requirement. Without this precondition, the results in Fig. 6 show that the estimated disparity maps are rough, and these methods cannot find the matched points in the same scan line; this is caused by imperfect rectification resulting from the use of imperfect camera parameters. Thanks to our proposed VCM- and GAN-based training scheme, our model's results are smoother and more reasonable than those of other methods. The quantitative results in Table 2 also confirm the validity of our proposed model, which can dramatically improve upon the performances of the baseline models under imperfect rectified stereo image pairs.

It is worth mentioning that, the results denoted with the dotted box in Table 3 were submitted after the MICCAI 2019 challenge day. Dimitris Psychogyios 1 uses Deep Pruner [37] model and pretrained on Sceneflow dataset [22], Dimitris Psychogyios 2 makes use of Hierarchical deep Stereo Matching network (HSM) [38], these two methods both ignored the interpolation frames as well as dataset 4 and 5 due to imperfect camera calibration. Sebastian Schimid takes the method similar to pwc-net [39] and

gwc-net [40]. These three results submitted after challenge day are supervised-based learning methods. The top-ranked method on challenge day in Table 3 is built on PSMNet [3], which is also a supervised deep learning architecture, while our method can be trained in an unsupervised manner. This is a very promising method, as mentioned in the previous section. It is very difficult to obtain a per-pixel ground truth dataset to train supervised learning methods in the field of laparoscopic or endoscopic surgery, but our method provides a new option for training without ground truth data, and it can achieve the same performance as that of the SOTA supervised method.

In summary, the proposed method obtains remarkable improvements over the performances of the baseline methods on imperfect rectified images (*split_1*). In addition, thanks to the trainable vertical correction module and adversarial learning scheme, even when the training dataset is mixed with imperfectly and perfectly rectified stereo images (*split_2* and *split_3*), our proposed model can slightly improve upon the performances of the baseline methods for general datasets. Moreover, the extensive ablation study conducted in the above subsection clearly demonstrates that when the VCM and adversarial learning are combined, our method can achieve the best performance. All the obtained quantitative and qualitative experimental results confirm the effectiveness and benefits of the proposed method.

In our experiments (Table 3), we notice that the SOTA monocular methods (Wong et al. [36], Godard et al. [21] and Pilzer et al. [29]) cannot predict depth values at the correct scale when applying the models to different scene datasets. Without other auxiliary methods to confirm the scale, the monocular methods have difficulty recovering the surface of the surgical site correctly at an accurate depth scale. In addition, in Table 3, the MAE for keyframe 1 of every dataset is obviously larger than those of the other keyframes. After checking the datasets frame by frame, we find that the focal length of the camera is changed when collecting these datasets, while the estimated depth is recovered according to the given fixed-depth focal length, and this is the main reason for the observed performance degradation. It is a very common operation to adjust the focal length of the camera during laparoscopic or endoscopic surgery. Unfortunately, none of the participants in the MICCAI 2019 SCARED sub-challenge can address this issue correctly. How can the correct depth be inferred under various focal lengths? This will also be part of our future work.

While the proposed approach achieves promising results, there are several limitations that require further investigation in future work. First, the proposed method cannot properly address the edges of different tissues or organs and retains some holes in these regions. Second, it cannot recover the depth correctly when the camera is zoomed in or zoomed out. It requires the camera parameters to remain unchanged after calibration when converting the disparity to depth using the offline calibrated focal length. Possible idea for overcoming those limitations could be introducing an edge-aware refinement term into the framework to improve the prediction of the edge region. For the second problem, while keeping the focal length of the camera fixed is not realistic, one possible solution could be introducing an object of known size into the scene. Third, as mentioned in [38], as the camera calibration is a homography transformation, the imperfect alignment may include vertical and horizontal shift, while we only consider the vertical correction, future work will also focus on this problem. In addition, involving interframe, inter-video geometric constraints are another potential approach. Thus, for future work, we plan to investigate these two limitations. In addition, an exploration of the interframe temporal information for the purpose of depth estimation could also be studied.

6. Conclusion

In this paper, we focus on developing an unsupervised learning framework for the depth estimation of endoscopic stereo image pairs that can handle imperfect rectified images. By introducing a vertical correction module and adversarial learning, our model can address imperfect rectified stereo images, even when the training sample contains a mixture of perfect and imperfect samples. Our model can also predict disparities more accurately and smoothly. Moreover, a residual mask is proposed to exclude outliers for improved loss computations. Our method can achieve comparable results to those of the SOTA supervised learning method.

Funding

This work was supported by grants from the NSFC Grant Program (62172401 and 12026602), the National Key R&D Program, China (Nos. 2019YFC0118100 and 2017YFC0110903), the Key-Area Research and Development Program of Guangdong Province, China (No. 2020B010165004), the Shenzhen Key Basic Science Program (No. JCYJ20180507182437217), and the Shenzhen Key Laboratory Program (ZDSYS201707271637577).

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Standard This article does not contain any studies with human participants performed by any of the authors.

References

[1] Luo H, Yin D, Zhang S, Xiao D, He B, Meng F, Zhang Y, Cai W, He S, Zhang W, Hu Q, Guo H, Liang S, Zhou S, Liu S, Sun L, Guo X, Fang C, Liu L, Jia F: Augmented reality navigation for liver resection with stereoscopic laparoscope. Comput Methods Programs Biomed 187:105099, 2020.

[2] Eigen D, Puhrsch C, Fergus R: Depth map prediction from a single image using a multi-scale deep network. NIPS, 2366-2374, 2014.

[3] Chang JR, Chen YS: Pyramid stereo matching network. CVPR, 5410-5418, 2018.

[4] Rau A, Edwards PE, Ahmad OF, Riordan P, Janatka M, Lovat B, Stoyanov D: Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. Int J Comput Assist Radiol Surg 14:1167-1176, 2019.

[5] Zhang Z: A flexible new technique for camera calibration. IEEE Trans Pattern Anal Mach Intell 22:1330-1334, 2000.

[6] Stoyanov D, Scarzanella MV, Pratt P, Yang GZ: Real-time stereo reconstruction in robotically assisted minimally invasive surgery. MICCAI, 275-282, 2010.

[7] Bernhardt S, Abi-Nahed J, Abugharbieh R: Robust dense endoscopic stereo reconstruction for minimally invasive surgery: MICCAI Medical Computer Vision (MCV), 254-262, 2012.

[8] Penza V, Ortiz J, Mattos LS, Forgione A, De Momi E: Dense soft tissue 3D reconstruction refined with super-pixel segmentation for robotic abdominal surgery. Int J Comput Assist Radiol Surg 11:197-206, 2016.

[9] Chang PL, Stoyanov D, Davison AJ: Real-time dense stereo reconstruction using convex optimisation with a cost-volume for image-guided robotic surgery. MICCAI, 42-49, 2013.

[10] Wang C, Cheikh FA, Kaaniche M, Elle OJ: Liver surface reconstruction for image guided surgery. Proc SPIE, 10576:105762H, 2018.

[11] Chen L, Tang W, John NW, Wan TR, Zhang JJ: SLAM-based dense surface reconstruction in

monocular Minimally Invasive Surgery and its application to Augmented Reality. Comput Methods Programs Biomed 158:135-146, 2018.
[12] Mahmoud N, Collins T, Hostettler A, Soler L, Doignon C, Montiel JMM: Live tracking and dense reconstruction for handheld monocular endoscopy. IEEE Trans Med Imaging 38:79-89, 2018.
[13] Maier-Hein L, Mountney P, Bartoli A, Elhawary H, Elson D, Groch A, Kolb A, Rodrigues M, Sorger J, Speidel S, Stoyanov D: Optical techniques for 3D surface reconstruction in computer-assisted laparoscopic surgery. Med Image Anal 17:974-996, 2013.
[14] Lin B, Sun Y, Qian X, Goldgof D, Gitlin R, You Y: Video-based 3D reconstruction, laparoscope localization and deformation recovery for abdominal minimally invasive surgery: a survey. Int J Med Robot 12:158-178, 2016.
[15] Žbontar J, LeCun Y: Stereo matching by training a convolutional neural network to compare image patches. J Mach Learn Res 17:2287-2318, 2016.
[16] Khamis S, Fanello S, Rhemann C, Kowdle A, Valentin J, Izadi S: Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. ECCV, 573-590, 2018.
[17] Zhou C, Zhang H, Shen X, Jia J: Unsupervised learning of stereo matching. ICCV, 1567-1575, 2017.
[18] Skinner KA, Zhang J, Olson EA, Johnson-Roberson M: Uwstereonet: Unsupervised learning for depth estimation and color correction of underwater stereo imagery. ICRA, 7947-7954, 2019.
[19] Scharstein D, Szeliski R: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int J Comput Vis 47:7-42, 2002.
[20] Garg R, Bg VK, Carneiro G, Reid I: Unsupervised cnn for single view depth estimation: Geometry to the rescue. ECCV, 740-756, 2016.
[21] Godard C, Mac Aodha O, Brostow GJ: Unsupervised monocular depth estimation with left-right consistency. CVPR, 270-279, 2017.
[22] Mayer N, Ilg E, Hausser P, Fischer P, Cremers D, Dosovitskiy A, Brox T: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. CVPR, 4040-4048, 2016.
[23] Ye M, Johns E, Handa A, Zhang L, Pratt P, Yang GZ: Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. arXiv preprint arXiv:1705.08260, 2017
[24] Liu X, Sinha A, Ishii M, Hager GD, Reiter A, Taylor RH, Unberath M: Dense depth estimation in monocular endoscopy with self-supervised learning methods. IEEE Trans Med Imaging 39:1438-1447, 2019.
[25] Luo H, Hu Q, Jia F: Details preserved unsupervised depth estimation by fusing traditional stereo knowledge from laparoscopic images. Healthc Technol Lett 6:154-158, 2019.
[26] Mahjourian R, Wicke M, Angelova A: Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. CVPR, 5667-5675, 2018.
[27] Godard C, Mac Aodha O, Firman M, Brostow GJ: Digging into self-supervised monocular depth estimation. ICCV, 3828-3838, 2019.
[28] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y: Generative adversarial networks. NIPS, 2672-2680, 2014.
[29] Pilzer A, Xu D, Puscas M, Ricci E, Sebe N: Unsupervised adversarial depth estimation using cycled generative networks. 3DV, 587-595, 2018.
[30] Pfeiffer M, Funke I, Robu MR, Bodenstedt S, Strenger L, Engelhardt S, Roß T, Clarkson MJ, Gurusamy K, Davidson BR, Maier-Hein L, Riediger C, Welsch T, Weitz J, Speidel S: Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image

translation. MICCAI, 119-127, 2019.
[31] Nguyen PH, Ahn CW: Stereo matching methods for imperfectly rectified stereo images. Symmetry 11:570, 2019.
[32] He K, Zhang X, Ren S, Sun J: Deep residual learning for image recognition. CVPR, 770-778, 2016.
[33] Allan M, Mcleod J, Wang CC, Rosenthal JC, Hu Z, Gard N, Eisert P, Fu KX, Zeffiro T, Xia W, Zhu Z, Luo H, Jia F, Zhang X, Li X, Sharan L, Kurmann T, Schmid S, Sznitman R, Psychogyios D, Azizian M, Stoyanov D, Maier-Hein L, Speidel S: Stereo Correspondence and Reconstruction of Endoscopic Data Challenge. arXiv preprint arXiv:2101.01133, 2021.
[34] Zhong Y, Dai Y, Li H: Self-supervised learning for stereo matching with self-improving ability. arXiv preprint arXiv:1709.00930, 2017.
[35] Abadi M, Barham P, Chen J, et al.: Tensorflow: A system for large-scale machine learning. In 12th USENIX Symposium on Operating Systems Design and Implementation, 265-283, 2016.
[36] Wong A, Soatto S: Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction: CVPR, 5644-5653, 2019.
[37] Duggal S, Wang S, Ma WC, Hu R, Urtasun R: Deeppruner: Learning efficient stereo matching via differentiable patchmatch. ICCV, 4384-4393, 2019.
[38] Yang G, Manela J, Happold M, Ramanan D: Hierarchical deep stereo matching on high-resolution images. CVPR, 5515-5524, 2019.
[39] Sun D, Yang X, Liu MY, Kautz J: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. CVPR, 8934-8943, 2018.
[40] Guo X, Yang K, Yang W, Wang X, Li H: Group-wise correlation stereo network. CVPR, 3273-3282, 2019.