Semi-supervised Learning via Improved Teacher-Student Network for Robust 3D Reconstruction of Stereo Endoscopic Image

Hongkuan Shi*
Zhiwei Wang*
shihk@hust.edu.cn
zwwang@hust.edu.cn
Huazhong University of Science and Technology

Jinxin Lv
Huazhong University of Science and Technology
jinx@hust.edu.cn

Yilang Wang
Huazhong University of Science and Technology
black_wang@hust.edu.cn

Peng Zhang
Huazhong University of Science and Technology
hustzp@hust.edu.cn

Fei Zhu
Huazhong University of Science and Technology
zhufei@hust.edu.cn

Qiang Li†
Huazhong University of Science and Technology
liqiang8@hust.edu.cn

ABSTRACT

3D reconstruction of stereo endoscope image, as an enabling technique for varied surgical systems, e.g., medical droids, navigations, etc., suffers from severe overfitting problems due to scarce labels. Semi-supervised learning based on Teacher-Student Network (TSN) is a potential solution, which utilizes a supervised teacher model trained on available labeled data to teach a student model on all images via assigning them pseudo labels. However, TSN often faces a dilemma: if given only few labeled endoscope images, the teacher model will be trained to be defective and induce high-noised pseudo labels, degrading the student model significantly. To solve this, we propose an improved TSN for a robust 3D reconstruction of stereo endoscope image. Specifically, two novel modules are introduced: 1) a semi-supervised teacher model based on adversarial learning to produce mostly correct pseudo labels by forcing a consistency in predictions for both labeled and unlabeled data, and 2) a confidence network to further filter out noisy pseudo labels by estimating a confidence for each prediction of the teacher model. By doing so, the student model is able to distill knowledge from more accurate and noiseless pseudo labels, thus achieving improved performance. Experimental results on two public datasets show that our improved TSN achieves a superior performance than the state-of-the-arts by reducing the averaged disparity error by at least 13.5%.

CCS CONCEPTS

• **Computing methodologies** → **Reconstruction**; • **Theory of computation** → **Semi-supervised learning**; • **Applied computing** → *Imaging*.

KEYWORDS

stereo matching; semi-supervised learning; teacher-student network; endoscopic image

ACM Reference Format:
Hongkuan Shi, Zhiwei Wang, Jinxin Lv, Yilang Wang, Peng Zhang, Fei Zhu, and Qiang Li. 2021. Semi-supervised Learning via Improved Teacher-Student Network for Robust 3D Reconstruction of Stereo Endoscopic Image. In *Proceedings of the 29th ACM Int'l Conference on Multimedia (MM '21), Oct. 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475527

1 INTRODUCTION

A modern Minimally Invasive Surgery (MIS) involves a wide variety of systems [1], for instance, medical droids/robotics, surgical navigation, virtual/augmented realities and so on. Among them, 3D reconstruction of stereo endoscope image is a crucial technique which enables the systems to fully understand surgical scenes for the follow-up actions [2].

Traditional methods [3–5] for 3D reconstruction based on a stereo image pair (left and right images) first match pixels between the left and right images, and then calculate spatial shifts of those matched pixels across the two view images as disparities, which can be used for computing depths with the intrinsic parameters of binocular camera given. The calculated disparity map is often very sparse and only has valid values for those locations which can find matching pixels in both view images. To make it denser, Stoyanov *et al.* [3] proposed to first identify several matched feature points as candidates, and then use a region growing method to propagate disparity values around the candidates. Penza *et al.* [4] proposed to refine the disparity image using super pixel segmentation. Chang *et al.* [5] introduced a dense stereo reconstruction approach using convex optimization with a cost volume. Despite their success in

*Co-first authors.
†Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China.
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475527

the past decade, these methods usually suffer from a significant degradation of performance if the scenes are too less or noisy information to find correct matched pixels, e.g., low-textured surface, specular reflection, depth discontinuity, etc.

Recently, methods [6–12] based on convolutional neural network (CNN) have shown their strengths that they can precisely recover 3D object locations/depths in complex scenes benefiting from the CNN's powerful capability of semantic reasoning. Their high performance significantly relies on a large amount of labeled data for supervised training. However, ground-truth disparity/depth map is usually scarce and hard to acquire, resulting in an overfitting problem consequently. The conflict becomes more intense when it comes to the medical data. For instance, in our case only 25 of total in-vivo stereo endoscope images have labels while thousands are infeasible to annotate.

To lower such label dependency in the medical domain, some methods [13, 14] proposed to first pre-train CNN models on some simulated or synthetic images whose labels are easy to obtain, and then adapted the pre-trained CNN models to the real scenarios via Transfer learning. Mahmood *et al.* [13] first built 3D virtual colon models, and then simulated many training images as well as their exported depth maps as ground-truths. A supervised depth estimator was trained on the fake data, and transferred to the real human data by domain adaptation. Similarly, Visentini-Scarzanella *et al.* [14] used an endoscopic video simulation of simple bronchial phantom models to train their network, and then converted video frames of complex phantom into frames similar to simple phantom by utilizing RGB rendering transcoder. Although some promising results have been shown in their cases, the conflict between limited data and data-driven CNN models is still not addressed by those methods. They just pass on the difficulties to another task of transfer learning that is also considerably challenging.

Fortunately, there is a family of CNNs [11, 12, 15, 16] can be trained requiring no label or only partial labels, a.k.a., unsupervised or semi-supervised CNNs. Godard *et al.* [11] proposed to train an unsupervised network for natural stereo images by maximizing a left-right disparity consistency instead of minimizing the pixel-wise loss which requires the ground-truth. Ye *et al.* [15] adopted a similar idea of employing an unsupervised Siamese network [11] but applied it to depth estimation of endoscopic images rather than natural ones. Li *et al.* [16] also utilized a disparity consistency of consecutive frames in a video to train an unsupervised depth and motion predictor based on monocular data. However, the disparity/depth maps predicted by those unsupervised CNNs are often untrustworthy for 3D reconstruction due to a lot of errors, but can be incorporated to enhance the supervised CNNs, which forms the basic idea of semi-supervised learning approaches.

Among the existing semi-supervised methods, Teacher-Student Network (TSN) [17–24] handles the situation that there are far more unlabeled data than labeled data, and thus presents a potential solution for 3D reconstruction of stereo endoscope image. TSN typically trains a teacher model in a supervised manner using labeled data, and then assigns numerous unlabeled data pseudo labels. Based on an assumption that among those pseudo labels correct predictions are dominant and consistent while errors are divergent, a student model can distil useful knowledge from plenty of training images assisted by pseudo labels, yielding more robust and generalized performance than the teacher model. However, the assumption why TSN works hardly hold if directly applying TSN on stereo endoscope images because the labeled data is too few to train a reliable teacher model, which will produce corrupted pseudo labels consequently and thus only teach out a worse student model.

To solve this dilemma, in this paper we proposed an improved TSN by introducing two novel modules, i.e., a semi-supervised teacher model and a confidence network, to not only guarantee more accurate pseudo labels but also suppress possible induced errors for teaching a student model. Specifically, the semi-supervised teacher model consists of a disparity estimation network (i.e., DEnet) and a Discriminator. For each input stereo image pair (i.e., left and right images), DEnet predicts a disparity map which is aligned with the left image, and meanwhile the Discriminator measures a distance, (e.g., Wasserstein distance [25]), between distributions of teacher-predicted disparity maps and ground-truth ones. If the input stereo image pair has labels, a pixel-wise loss is minimized, otherwise, the distribution distance derived by the Discriminator is minimized instead. Thanks to the assistance of numerous unlabeled data, the proposed semi-supervised teacher model is expected to avoid overfitting and improve the reliability of pseudo labels comparing to a supervised teacher model. In addition, a confidence network (i.e., Confnet) is further trained to estimate a confidence map for each teacher-predicted disparity map, where possible wrong predictions can be suppressed by low confidence weights. By distilling knowledge from more accurate and error-suppressed pseudo labels, a student model with the same architecture as the teacher model can be taught to achieve a better performance. To the best of our knowledge, our work is the first try that exploiting a semi-supervised learning via improving TSN for disparity estimation of stereo endoscope image.

In summary, our main contributions are listed:

1) We propose a semi-supervised teacher model based on adversarial learning which can avoid overfitting by taking advantage of both labeled and unlabeled data, and thus produce reliable pseudo labels.

2) We train a confidence network for identifying possible errors to make teacher-predicted pseudo labels further denoised. Combining both semi-supervised teacher model and confidence network, an improved TSN method can be developed to teach out a better student model for a robust disparity estimation even though a few labels are available.

3) The experimental results on a public endoscopic dataset demonstrate the effectiveness of the semi-supervised teacher model and the confidence network. The taught-out student model based on more accurate and denoised pseudo labels achieves a superior performance than the state-of-the-arts [7, 9] by reducing the averaged disparity error by at least 13.5% (from 0.89 pixels to 0.77 pixels in a 1024×1280 image).

2 METHOD

Figure 1 gives an overall architecture of our proposed improved TSN which consists of three major components, i.e., a semi-supervised teacher model, a confidence network, and a student model. In this section, we will detail each component, and describe the unsupervised and supervised training strategies respectively.

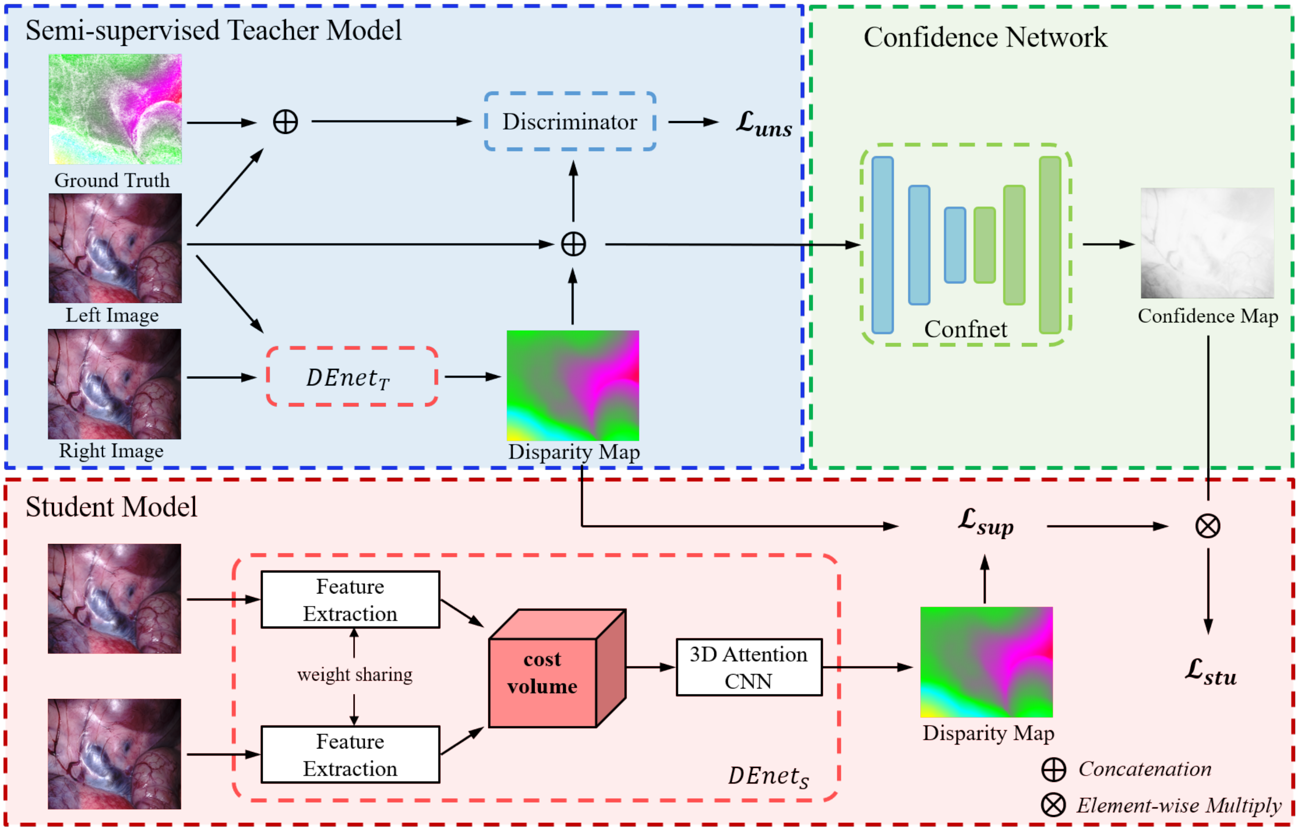


Figure 1: An overall architecture of our proposed improved Teacher-Student Network.

2.1 Semi-supervised Teacher Model for Reliable Pseudo Labels

The semi-supervised teacher model consists of a disparity estimation network (i.e., DEnet) and a Discriminator as shown in the top-left image of Figure 1. Note that the student model is also a DEnet with the same architecture as the teacher model but without weight sharing. To avoid misleading, we denote $DEnet_T$ and $DEnet_S$ to the teacher and student model respectively.

2.1.1 Disparity estimation network. Given a stereo image (I_l, I_r), where I_l and I_r represent the left and right color images with the size of H×W×3, our DEnet first utilizes a shared sub-network to extract two feature maps for I_l and I_r respectively, and then estimates a disparity map d aligned with the I_l based on the calculated cost volume as shown in the bottom image of Figure 1.

Although using the cost volume for disparity estimation is proposed by [8], we detail it to make this work self-included. Specifically, we employ a truncated ResNet-50 [26] to extract image features which down-samples the original image size four times. For each image, the extracted feature map F with the size of *H/4× W/4×32* is composed by concatenating and compressing the side outputs of last three layers.

The key idea of cost volume is to construct a disparity searching space by aligning F_l and F_r of the left and right images at different disparity-levels. To this end, we first determine the maximum searching disparity-level as S, and then horizontally translate F_l to the left for s pixels at the s-th disparity level, where $s = 1, 2, ..., S/4$. Theoretically, F_l should be translated along the epipolar line. Since the stereo image pairs are pre-calibrated, the epipolar line for each pixel is horizontal. The translated feature map $F_l@s$ at s-th disparity level is then concatenated with F_r along channel. The final cost volume is thus formed by combining these concatenated feature maps across all disparity levels, yielding a 4D volume with the size of *H/4× W/4×S/4×64.*

After that, we utilize a 3D Attention CNN to convert the cost volume to a soft-maxed voting space volume with the size of $H×W×S$, whose map at s-th channel indicates a probability map of disparities being s pixels. The disparity map d with the size of $H×W$ is finally obtained by calculating expected disparity levels based on those probability maps. The 3D Attention CNN consists of a three cascaded U-nets, each of which replaces its last encoder with the channel attention mechanism [27] to further enhance interendencies of cost volume across different disparity levels.

Typically, $DEnet_T$ is trained to minimizing the pixel-wise distance between its prediction d and the corresponding ground-truth $\hat{d}$, which, on the one hand, wastes the information carried by a plenty of unlabeled images without $\hat{d}$, and, on the other hand, makes the model overfit for the number of $\hat{d}$ is limited. For example, this study includes 25 stereo endoscope videos consisting of

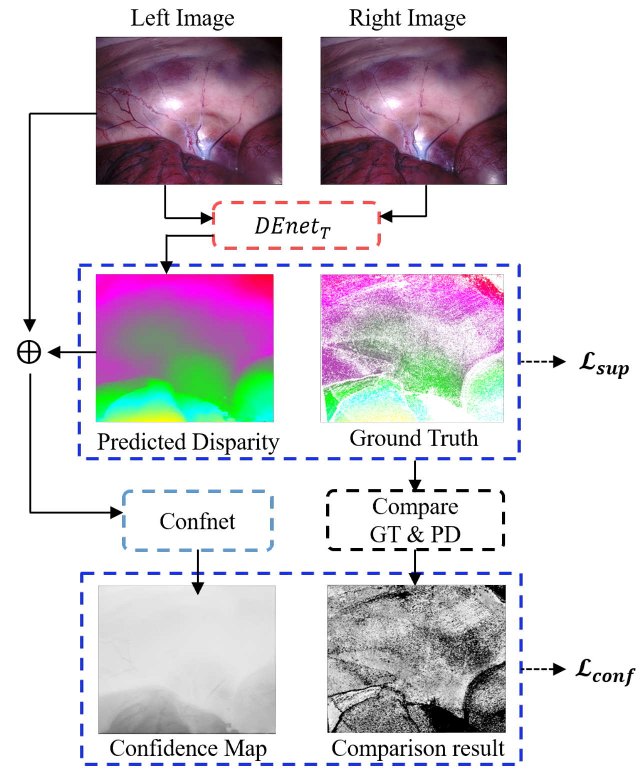


Figure 2: The supervised training for the teacher model and the confidence network.

thousands of frames, while only the first frame of each video has ground-truth $\hat{d}$. To avoid unnecessary waste and overfitting, we introduce a Discriminator to approximate the distribution distance between predictions and ground-truths. Minimizing the distance can derive helpful gradients for updating $DEnet_T$ even if feeding unlabeled images.

Specifically, we treat the left image and its teacher-predicted disparity map (I_l, d) as a fake pair, and $(I_l, \hat{d})$ as a real pair. Note that these two pairs do not have to come from the same video frame in order to include all unlabeled frames. Inspired by PatchGan [28], the Discriminator utilizes four convolutional layers sequentially to convert each image-disparity pair, i.e., (I_l, d) or $(I_l, \hat{d})$, to a score map, each entry of which indicates a chance of its receptive field belonging to a real pair. By doing so, $DEnet_T$ can be trained in both supervised and unsupervised manners.

2.1.2 Unsupervised Training. The dataset with total n stereo image pairs contains m pairs with ground-truth disparity maps ($m \ll n$). We denote the entire dataset and the labeled subset as $\mathbb{N}$ and $\mathbb{M}$ respectively for convenience of understanding. The Discriminator is usually trained to approximate a JS divergence (JSD) distance [29] between the real and fake pairs. However, the JSD distance derives no gradient if the Discriminator is well trained [25, 30], which is unhelpful for the convergence of $DEnet_T$. By comparison, Wasserstein distance [25, 30] can derives useful gradients to guide $DEnet_T$ in any case. Specifically, the Wasserstein distance $\mathcal{L}_{uns}$ is approximated as follows:

$$\begin{aligned}\mathcal{L}_{uns} = \max \Big\{ & \mathrm{E}_{\mathbb{M}} \left[D \left(I_l, \hat{d} \right) \right] \\ & - \mathrm{E}_{\mathbb{N}} \left[D \left(I_l, d \right) \right] - R_D \Big\}, \text{ w.r.t. Discriminator}\end{aligned} \tag{1}$$

Where $\mathrm{E}_{\mathbb{M}}$ means sampling real pairs from $\mathbb{M}$, and $\mathrm{E}_{\mathbb{N}}$ is sampling fake pairs from $\mathbb{N}$, $D(\cdot)$ is the averaged score of the score map predicted by the Discriminator, R_D is used for enforcing the 1-Lipschitz constraint of D, and details of it can refer to [25].

$DEnet_T$ has to minimize the approximated distance, a.k.a., to fool the Discriminator by confusing teacher-predicted disparity map d with $\hat{d}$. Therefore, the final objective of unsupervised learning for $DEnet_T$ is formulated as follows

$$\min \mathcal{L}_{uns}, \qquad \text{w.r.t. } DEnet \tag{2}$$

Note that, even though $\mathcal{L}_{\text{uns}}$ is approximated by incorporating both labeled and unlabeled data, minimizing it w.r.t. $DEnet_T$ only requires unlabeled ones if the Discriminator is fixed, having the teacher model being optimized in an unsupervised manner.

2.1.3 Supervised Training. As shown in Figure 2, we employ the smooth L1 loss as a pixel-wise distance between the prediction and the corresponding ground-truth $\hat{d}$ to train $DEnet$ in a supervised manner. The smooth L1 loss is computed as follows:

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{3}$$

Typically, the smooth L1 loss treats all pixels equally, which could be problematic for the long-tail phenomenon, i.e., a large number of pixels fall into a small depth range since doctors often operate the endoscope to observe nearby tissues or organs, making the model tend to predict small depth values. To deal with this, we propose to pay more attention to the distant depth regions, and use the depth-aware smooth L1 loss for $DEnet_T$ as follows:

$$\mathcal{L}_{sup}(d, \hat{d}) = \sum^{N} \frac{1}{\alpha_d} \cdot smooth_{L_1}(d - \hat{d}) \tag{4}$$

where N is the number of labeled pixels, α_d is defined as the normalized $\hat{d}$, and thus $1/\alpha_d$ can act as a depth aware attention map that makes the model focus more on distant depth regions. When the depth increases, the corresponding weight will also increase linearly. It should be noted that due to the characteristics of structured light measurement, the ground truth can not cover the whole image densely, leaving some vacant positions like the one in Figure 2. Therefore, these unlabeled pixels are excluded when calculating the $\mathcal{L}_{sup}$.

2.2 Confidence Estimation Network for Suppressing Possible Errors

The trained $DEnet_T$ processes all unlabeled stereo frames and yields pseudo labels for guiding the following student network. Although the reliability of those pseudo labels could be improved by the proposed semi-supervised teacher model, possible errors among them can still have a negative impact on the learning of student

Table 1: Comparison results for the ablation study. UIT indicates using how many unlabeled images for training, e.g., 0.5 means a half of image suppressed by its confidence map.

Methods	UIT	> 1px (%)	> 2px (%)	> 3px (%)	EPE (px)
Supervised $DEnet_T$	-	25.93±9.92	7.01±4.84	2.81±2.46	0.91±0.11
Semi-supervised $DEnet_T$	1000	24.75±9.22	6.51±4.11	2.59±2.15	0.88±0.10
$DEnet_S$*-Base*	1000	24.61±9.70	6.59±4.36	2.60±2.22	0.87±0.10
$DEnet_S$*-D*	1000	24.11±9.36	6.01±3.90	2.22±1.94	0.83±0.10
$DEnet_S$*-C*	883.4	23.11±10.13	5.85±3.98	2.16±1.94	0.81±0.12
$DEnet_S$*-C&D*	913.9	**22.32±9.66**	**5.51±3.72**	**2.04±1.87**	**0.77±0.10**

model. Therefore, we introduce a confidence network (i.e., Confnet) to estimate how accurate the prediction of $DEnet_T$ is.

As shown the top-right image of Figure 1, we develop an encoder-decoder architecture as our Confnet to convert a predicted disparity map concatenated with its original left image into a confidence map, $C = Confnet\ (I_l, d)$. The encoding blocks consists of 3 convolutional layers, followed by 3 decoding layers to restore the original resolution. Outputs of these layers are activated by LeakyReLU function. At last, a convolutional layer with a 3×3 kernal followed by a sigmoid function produces a confidence map with each entry ranging from 0 to 1.

For training Confnet, we first determine how accurate the predicted d is. Specifically, for the i-th predicted disparity in d we consider it as an accurate one and assign it a label 1 if $\left|d_i - \hat{d}_i\right| < 3$ pixels, or assign it a label 0 otherwise. By doing so, we can obtain a ground-truth map $\hat{C}$ for each Confnet-estimated confidence map C. And then we can train Confnet by minimizing the difference between C and $\hat{C}$:

$$\mathcal{L}_{conf} = \sum^{N} (\hat{C}\log(C) + (1-\hat{C})\log(1-C)) \tag{5}$$

where N is the number of labeled pixels.

2.3 Student Model For Distilling Knowledges

The bottom image of Figure 1 shows how to utilize pseudo labels and their corresponding confidence maps to train a student model. Specifically, given a pseudo label d and its corresponding confidence map C, the student model $DEnet_S$ is optimized based on an error-suppressed Smooth L1 loss by using C as a weight map. Note that for those labeled data, we can still use their ground-truths for training the student model based on depth-aware smooth L1 loss in Eq. 4. The final loss can be formulated as follows:

$$\mathcal{L}_{stu} = \begin{cases} \mathcal{L}_{sup}\left(d_s, \hat{d}\right) & \text{if } (I_l, I_r) \in \mathbb{M} \\ C \cdot \mathcal{L}_{sup}\left(d_s, DEnet_T\left(I_l, I_r\right)\right) & \text{otherwise} \end{cases} \tag{6}$$

where d_s is a predicted disparity map by the student model $d_s = DEnet_S\left(I_l, I_r\right)$.

In addition, we mask out the values of those predictions on the reflective regions when calculating $\mathcal{L}_{stu}$ to avoid potential negative affects on the training. Specifically, we discard those pixels of an image if their saturation value is less than 0.1 and intensity value is greater than 0.9. Moreover, an auxiliary loss proposed in [11] is employed to encourage a smooth predicted disparity map. Same procedures are also performed to the teacher model.

3 EXPERIMENTS

3.1 Datasets and Evaluation metrics

Stereo Correspondence and Reconstruction of Endoscopic Data (SCARED) is a medical scene dataset [31]. This dataset contains binocular images of fresh porcine cadaver abdominal anatomy. There are 7 sub-datasets obtained from 7 pigs, each of which contains 5 keyframes. The keyframes consist of binocular keyframe images, sequence video frames with the size of 1024×1280 and the corresponding depth. The video frames depth is interpolated from the keyframe depth according to the transformation matrix relative to the keyframe camera position.

We find that among 7 sub-datasets, only 5 sub-datasets can be used as the labeled images (5x5=25 keyframes in total); other two sub-datasets have wrong calibration parameters and cannot be used as training datasets.

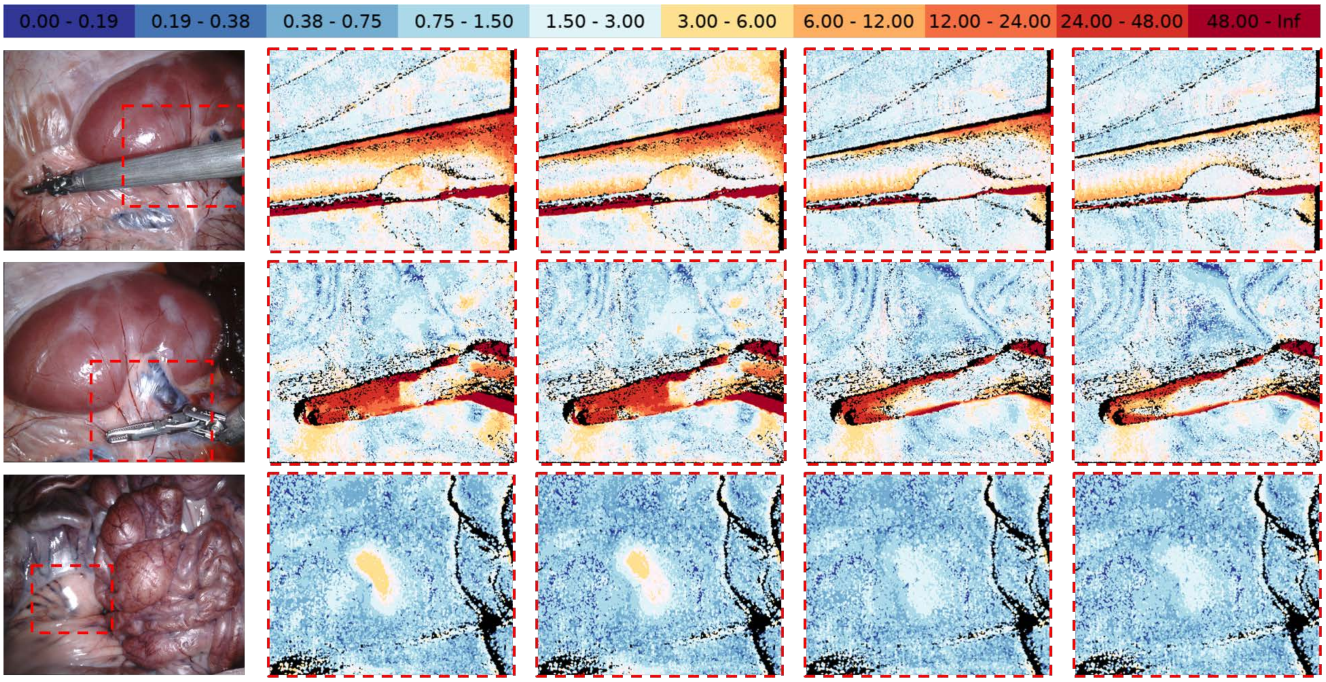
Scene Flow datasets are synthetic stereo datasets [10], including Flyingthings3D, Driving and Monkaa. The datasets provide 35,454 training and 4,370 testing images with the size of 960×540 as well as accurate ground truth disparity maps. It is the most commonly used dataset to train a deep network from scratch.

Evaluation metrics We conducted leave-one-out experiments on the 5 usable sub-datasets. The end-point disparity error (EPE) and percentages of 1-px, 2-px, 3-px disparity outliers on 5 sub-datasets are averaged and reported. EPE is defined as the mean average disparity error in pixel unit. The outliers are defined as the pixels whose disparity errors are larger than the threshold (1-px, 2-px, 3-px). The evaluation results are also reported on Scene Flow datasets.

3.2 Training Details

For training, 20 labeled images were selected as trainset and another 5 as testset. No subject was cross-used in the trainset and testset. We performed data augmentation on the labeled images, including horizontal flipping, random gamma and brightness shifts. Another 1000 unlabeled images were selected from the video frames at intervals.

We trained the semi-supervised teacher network and the confidence network jointly for 100 epochs. The Discriminator and the teacher model $DEnet_T$ were trained alternately to play the min-max game like other GAN-based approaches [25]. Gradients derived



(a) Supervised $DEnet_T$ **(b) Semi-supervised** $DEnet_T$ **(c)** $DEnet_S$*-C* **(d)** $DEnet_S$*-C&D*

Figure 3: Error maps of ablation study on SCARED images. The left panel shows the left input image of stereo image pair. For each input image, the error maps obtained by different settings of our model are illustrated. Different colors indicate absolute distances between predicted disparity and ground truth.

from L_{conf} were used to update Confnet only. The student model was trained for 100 epochs after the semi-supervised teacher model was well-trained.

Our model were implemented using PyTorch on NVIDIA GeForce Titan-RTX GPU with batch size of 8. All models were trained with Adam optimizer [32] (β_1=0.9, β_2=0.999). The initial learning rate of DEnet was set to 0.001 and decayed by 0.5 every 20 epochs. The learning rate of Confnet was 0.0001. It’s worth noting that we pretrained a *DEnet* for 10 epochs using the Sceneflow dataset following PSMNet [7].

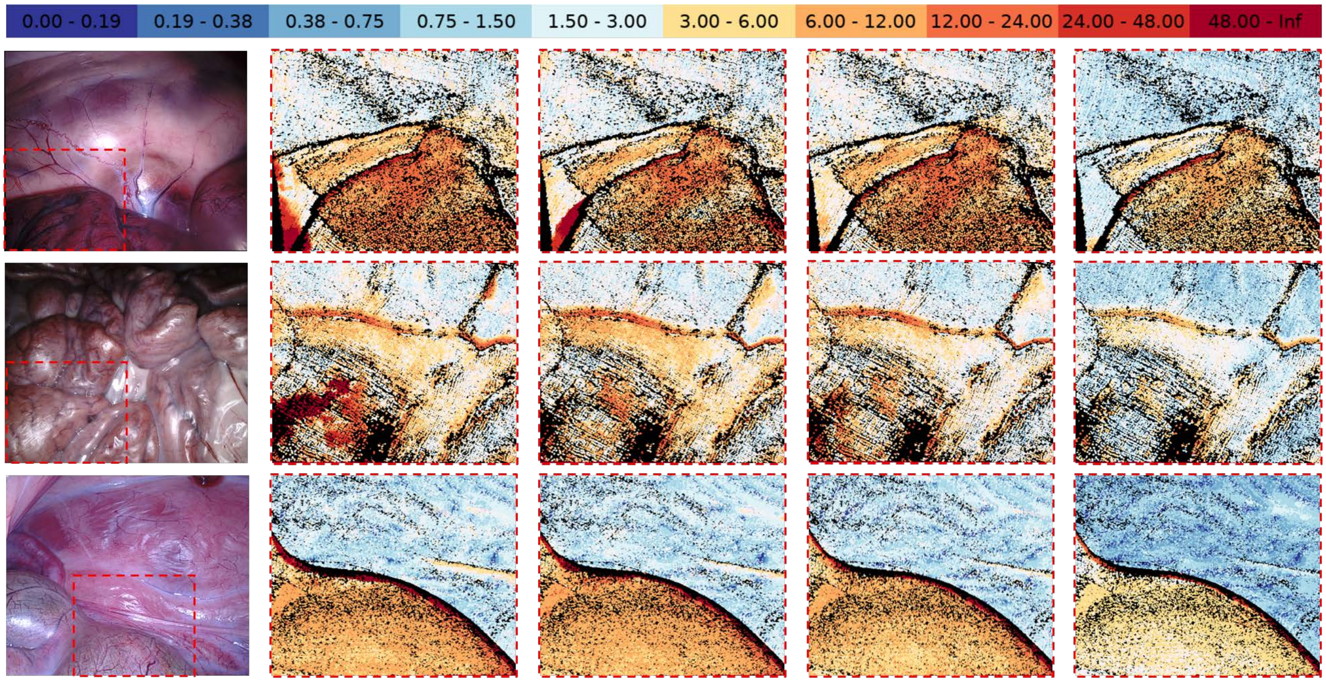
3.3 Ablation Study

In order to exam the effectiveness of our proposed improve TSN, we train two teacher models in a supervised and semi-supervised manner respectively, i.e., Supervised $DEnet_T$ and Semi-supervised $DEnet_T$. Four versions of student model are compared: 1) $DEnet_S$*-Base* trained based on those assigned by Supervised $DEnet_T$, 2) $DEnet_S$*-D* trained based on pseudo labels assigned by Semi-supervised $DEnet_T$, 3) $DEnet_S$*-C* trained based on error-suppressed pseudo labels by both Supervised $DEnet_T$ and its Confnet, and 4) $DEnet_S$*-C&D* trained on error-suppressed ones by both Semi-supervised $DEnet_T$ and its Confnet. Comparison results of these student and teacher models are shown in Table 1.

By comparing the results of Supervised $DEnet_T$ and Semi-supervised $DEnet_T$, we can find that using all unlabeled images indeed improves the quality of predictions of teacher model, which promises a more reliable pseudo labels for teaching the student model. Also, as can be seen in Table 1, the student models are consistently improved comparing with their corresponding teacher models, which implies that semi-supervised learning based on teacher-student network is an effective solution to our task for those unlabeled data carries useful information which was ignored by previous works.

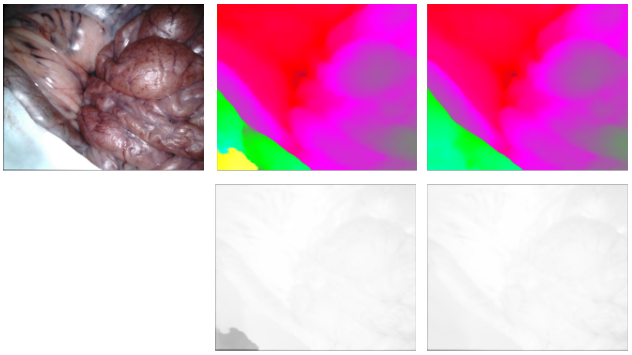
Comparison results between $DEnet_S$*-Base* and $DEnet_S$*-D* demonstrate that distilling knowledge from more accurate pseudo labels make the teach-out student model more robust, which well justify our motivations, i.e., more reliable pseudo labels are provided, more effective a student model can be taught. Interestingly, we find that $DEnet_S$*-C* can achieve a comparable performance to $DEnet_S$*-D* with more than 10% wrong predictions suppressed (from 1000 to 883.4), which verifies the key role of our proposed Confnet to improve TSN. It is worth noted that after error suppressing, the number of unlabeled images for training is increased from 883.4 to 913.9 since Semi-supervised $DEnet_T$ can provide less noisy predictions, reducing a data waste possibly induced by introducing confidence maps. At last, $DEnet_S$*-C&D* achieves the lowest error among all versions of student model, well demonstrating the effectiveness our proposed semi-supervised teacher model and confidence network for improving the quality of knowledge for teaching.

Figure 3 visually shows the results of compared four versions of student model. These visualized error maps for each student model is obtained by computing the absolute distance between the predicted disparity and the ground truth. Areas with large errors



(a) GANet **(b) PSMNet** **(c) Ji et al. [33]** **(d) Ours**

Figure 4: Error maps of predicted disparity on SCARED images. The left panel shows the left input image of stereo image pair. For each input image, the error maps obtained by various methods are illustrated. Different colors indicate absolute distances between predicted disparity and ground truth.



(a) Supervised $DEnet_T$ **(b) Semi-supervised** $DEnet_T$

Figure 5: Prediction of unlabeled SCARED images by the teacher model. The first row and the second row are the predicted disparity maps and corresponding confidence maps respectively. Darker areas in the confidence map represent lower confidence.

are zoomed in and indicated by red dashed rectangles. As can be seen, $DEnet_S$*-C&D* achieves the best performance especially in areas with reflection and depth discontinuity.

Figure 5 illustrates the predicted disparity map and corresponding confidence map of unlabeled images by Supervised $DEnet_T$ and Semi-supervised $DEnet_T$. The visualized confidence maps demonstrate that the Confnet can localize the potential error effectively and suppressed it with lower weights.

3.4 Comparison with the State-of-the-arts

We compared with several state-of-the-art methods including SGM [33], GAnet [9], PSMnet [7] and Ji *et al.* [34]'s work. SGM[1] is a famous traditional stereo matching method, and was tuned to achieve the best performance by setting aggregation paths Dir = 8, smoothness penalties P1 = 7 and P2 = 100. For GAnet and PSMnet, we made a direct usage of the source codes and tuned their hyper-parameters. For Ji *et al.* [34]'s method, We strictly followed the paper for reimplementation and adopted the same hyper-parameters as ours presented in Sec. 3.2. Ji *et al.* [34] adopted a similar idea with us, which utilized two discriminators to train the depth estimation network in a semi-supervised manner. In contrast, we do not directly output predictions of the semi-supervised depth estimator, but use them to lead a better student model with possible errors suppressed.

3.4.1 SCARED dataset. The comparison results on the SCARED dataset are listed in Table 2, from which we can have three key observations. First, by comparing the first three rows, we can find that the CNN-based methods achieve better performances than SGM, demonstrating a powerful reasoning capability of CNNs. Second, the comparison results from the 2nd to 4th rows verify that

[1] Available in Opencv.

Table 2: Performance comparison with SCARED test sets

Methods	>1px (%)	>2px (%)	>3px (%)	EPE (px)
SGM [33]	30.04±11.52	8.67±6.22	3.71±3.13	1.25±0.14
GANet [9]	25.36±11.16	7.04±4.97	2.96±2.46	1.01±0.20
PSMNet [7]	26.53±10.16	7.25±4.91	2.88±2.46	0.92±0.12
Ji *et al.* [34]	25.51±9.72	6.70±4.69	2.62±2.33	0.89±0.11
Ours	**22.32±9.66**	**5.51±3.72**	**2.04±1.87**	**0.77±0.10**

Table 3: Performance comparison with Scene Flow test sets

Methods	>1px (%)	>2px (%)	>3px (%)	EPE (px)
GANet [9]	10.03	5.44	3.96	0.93
PSMNet [7]	10.40	5.69	4.18	0.95
DEnet	**9.73**	**5.28**	**3.90**	**0.91**

the ignored information carried by those unlabeled images can help boosting the performance of disparity estimation significantly. Third, our method surpass all other methods including a semi-supervised model based on adversarial learning [34], which well demonstrates the superiority of our improved TSN. Comparing with them, the EPE value is reduced by 22.77%, 14.29%, 13.48% respectively.

Figure 4 illustrates error maps of these CNN-based methods and ours. Areas with large errors are zoomed in and indicated by red dashed rectangles. The results qualitatively show the robustness of our method than other methods at the edge of the organ and the flat area.

3.4.2 Scene Flow dataset. To make this work consistent with others, we also evaluate the performance of the supervised DEnet separately on Scene Flow test sets. Comparison results in Table 3 show that the supervised DEnet achieves a slightly better performance than GANet and PSMNet by reducing EPE by around 3%. A further 10% improvement can be expected if introducing extra images with no need to label them.

4 CONCLUSION

In this paper, we proposed an improved Teacher-Student Network to robustly estimate disparity maps for stereo endoscopic images. We proposed a semi-supervised teacher model based on adversarial learning to guarantee more reliable and less noisy pseudo labels. Also, we introduced a confidence network to estimate the reliability of those pseudo labels, and further suppress possible errors. With more accurate and denoised pseudo labels, the student model is taught out to have better performance of disparity estimation even for those regions hard to predict, e.g., flat, reflective surfaces. The experimental results on public endoscopic dataset (SCARED) demonstrate a superior performance of our improved TSN comparing with other state-of-the-arts including both supervised and semi-supervised methods. This work can enable preoperative registration and VR/AR technology applications in varied surgical systems, and the future work includes exploring more advanced techniques, e.g., self-supervised learning, to further improve the reliability of pseudo labels.

ACKNOWLEDGMENTS

This research was funded by the Fundamental Research Funds for the Central Universities (Grant No. 2021XXJS033), Science Fund for Creative Research Group of China (Grant No. 61721092) and Director Fund of WNLO.

REFERENCES

[1] Danail Stoyanov, George P Mylonas, Mirna Lerotic, Adrian J Chung, and Guang-Zhong Yang. Intra-operative visualizations: Perceptual fidelity and human factors. *Journal of Display Technology*, 4(4):491–501, 2008.

[2] Russell H Taylor, Arianna Menciassi, Gabor Fichtinger, Paolo Fiorini, and Paolo Dario. Medical robotics and computer-integrated surgery. *Springer handbook of robotics*, pages 1657–1684, 2016.

[3] Danail Stoyanov, Marco Visentini Scarzanella, Philip Pratt, and Guang-Zhong Yang. Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 275–282. Springer, 2010.

[4] Veronica Penza, Jesús Ortiz, Leonardo S Mattos, Antonello Forgione, and Elena De Momi. Dense soft tissue 3d reconstruction refined with super-pixel segmentation for robotic abdominal surgery. *International journal of computer assisted radiology and surgery*, 11(2):197–206, 2016.

[5] Ping-Lin Chang, Danail Stoyanov, Andrew J Davison, et al. Real-time dense stereo reconstruction using convex optimisation with a cost-volume for image-guided robotic surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 42–49. Springer, 2013.

[6] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[7] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018.

[8] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017.

[9] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019.

[10] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016.

[11] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[12] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019.

[13] Faisal Mahmood and Nicholas J Durr. Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. *Medical image analysis*, 48:230–243, 2018.

[14] Marco Visentini-Scarzanella, Takamasa Sugiura, Toshimitsu Kaneko, and Shinichiro Koto. Deep monocular 3d reconstruction for assisted navigation in bronchoscopy. *International journal of computer assisted radiology and surgery*, 12(7):1089–1099, 2017.

[15] Menglong Ye, Edward Johns, Ankur Handa, Lin Zhang, Philip Pratt, and Guang-Zhong Yang. Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. *arXiv preprint arXiv:1705.08260*, 2017.

[16] Ling Li, Xiaojian Li, Shanlin Yang, Shuai Ding, Alireza Jolfaei, and Xi Zheng. Unsupervised learning-based continuous depth and motion estimation with monocular endoscopy for virtual reality minimally invasive surgery. *IEEE Transactions on Industrial Informatics*, 2020.

[17] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[18] Avital Oliver, Augustus Odena, Colin Raffel, Ekin D Cubuk, and Ian J Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *arXiv preprint arXiv:1804.09170*, 2018.

[19] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017.

[20] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3227–3237, 2020.

[21] Andrea Pilzer, Stephane Lathuiliere, Nicu Sebe, and Elisa Ricci. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9768–9777, 2019.

[22] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, and Pheng-Ann Heng. Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. *arXiv preprint arXiv:1808.03887*, 2018.

[23] Wenhui Cui, Yanlin Liu, Yuxing Li, Menghao Guo, Yiming Li, Xiuli Li, Tianle Wang, Xiangzhu Zeng, and Chuyang Ye. Semi-supervised brain lesion segmentation with an adapted mean teacher model. In *International Conference on Information Processing in Medical Imaging*, pages 554–565. Springer, 2019.

[24] Christian S Perone and Julien Cohen-Adad. Deep semi-supervised segmentation with weight-averaged consistency targets. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 12–19. Springer, 2018.

[25] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[26] 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.

[27] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[28] 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.

[29] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[30] Zhiwei Wang, Yi Lin, Kwang-Ting Tim Cheng, and Xin Yang. Semi-supervised mp-mri data synthesis with stitchlayer and auxiliary distance maximization. *Medical image analysis*, 59:101565, 2020.

[31] Max Allan, Jonathan Mcleod, Cong Cong Wang, Jean Claude Rosenthal, Ke Xue Fu, Trevor Zeffiro, Wenyao Xia, Zhu Zhanshi, Huoling Luo, Xiran Zhang, et al. Stereo correspondence and reconstruction of endoscopic data challenge. *arXiv preprint arXiv:2101.01133*, 2021.

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[33] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007.

[34] Rongrong Ji, Ke Li, Yan Wang, Xiaoshuai Sun, Feng Guo, Xiaowei Guo, Yongjian Wu, Feiyue Huang, and Jiebo Luo. Semi-supervised adversarial monocular depth estimation. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2410–2422, 2019.