IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 1, 2020, accepted April 8, 2020, date of publication April 13, 2020, date of current version April 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2987767

Deep Convolutional Network for Stereo Depth Mapping in Binocular Endoscopy

XIONG-ZHI WANG[1,4], YUNFENG NIE[2], SHAO-PING LU[3], (Member, IEEE), AND JINGANG ZHANG[1]

[1]School of Future Technology, University of Chinese Academy of Sciences, Beijing 100039, China
[2]Brussel Photonics, Department of Applied Physics and Photoncis, Vrije Universiteit Brussel, 1050 Brussels, Belgium
[3]TKLNDST, CS, Nankai University, Tianjin 300071, China
[4]Department of Computer Science and Technology, Xidian University, Xi'an 710071, China

Corresponding author: Jingang Zhang (zhangjg@ucas.ac.cn)

This work was supported in part by the Joint Foundation Program of the Chinese Academy of Sciences for Equipment Pre-Feasibility Study under Grant 141A01011601, in part by the National Natural Science Foundation of China under Grant 61775219, Grant 61640422, and Grant 6177136, in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2019JM-557, and in part by the Equipment Research Program of the Chinese Academy of Sciences under Grant YJKYYQ20180039 and Grant Y70XA1HY.

ABSTRACT Depth mapping from binocular endoscopy images plays an important role in stereoscopic surgical treatment. Owing to the development of deep convolutional neural networks (CNNs), binocular depth estimation models have achieved many exciting results in the fields of autonomous driving and machine vision. However, the application of these methods to endoscopic imaging is greatly limited by the fact that binocular endoscopic images not only are rare, but also have unsatisfying features such as no texture, no ground truth, bad contrast, and high gloss. Aiming at solving the above-mentioned problems, we have built a precise gastrointestinal environment by the open-source software blender to simulate abundant binocular endoscopy data and proposed a 23-layer deep CNNs method to generate real-time stereo depth mapping. An efficient scale-invariant loss function is introduced in this paper to accommodate the characteristics of endoscope images, which improves the accuracy of achieved depth mapping results. Regarding the considerable training data for typical CNNs, our method requires only a few images (960×720 resolution) at 45 frames per second on an NVIDIA GTX 1080 GPU module, then the depth mapping information is generated in real-time with satisfactory accuracy. The effectiveness of the developed method is validated by comparing with state-of-the-art methods on processing the same datasets, demonstrating a faster and more accurate performance than other model frames.

INDEX TERMS Binocular endoscopes, deep convolutional neutral network, real-time evaluation, stereo depth mapping.

I. INTRODUCTION

With the continuous growth of public demands for minimal invasion and accurate operation in surgery, the concept of surgical navigation system with ''fine treatment'' and ''accurate surgery'' has become the tendency of future intelligent surgery development [1]. Surgical navigation system is the combination of surgery, computer technology, image processing, and stereoscopic vision to obtain the exact positioning of relevant lesions and dynamic movement orientations of surgical instruments, which assists doctors with binocular images in real-time diagnosis and treatment [2]–[4].

The associate editor coordinating the review of this manuscript and approving it for publication was Zhan-Li Sun.

Generally, stereo vision-based navigation systems need binocular depth difference (disparity) on top of conventional two-dimensional (2D) images, and the typical method is to register corresponding three-dimensional (3D) data(e.g. CT scans, etc.) into the intraoperative system in advance. After that, the registration algorithm executes a 3D reconstruction of the anatomical model from those preoperative data containing disparity information [5]. However, since the disparity information is not acquired in real-time, it causes deviations from the 3D reconstruction process.

The rapid progress in deep convolutional neural networks has been paving ways for stereo matching disparity estimation [4], [6]–[9], and more and more binocular disparity estimation methods based on this technique have emerged

in recent years. Although these methods have considerable advantages in machine vision and autonomous driving, their application in stereo endoscopic images is still facing certain challenges. To begin with, the training of neural networks relies greatly on considerable images, while binocular endoscopic data are difficult to obtain since there are few available relevant medical instruments. Even if the problem has been overcome, patient privacy and the absence of annotated experts can also hinder the establishment of binocular endoscopic image datasets. Secondly, current endoscopic images typically encounter issues such as uneven image texture, bad contrast and high gloss in certain portions which are caused by a proximal bright light source and a large field of view. Furthermore, irregular movement of human internal organs and limited space in operating endoscopic examinations have added the difficulty of applying those datasets to the training and the verification of the deep learning methods on the SceneFlow, KITTI2015 platform. Last but not least, the speed of current methods is generally limited to 1-2 frames per second (FPS), far from satisfying the requirements for real-time surgical navigation.

In this paper, we have trained a deep learning neural network for obtaining dense and high-accuracy stereo disparity mapping in an endoscopic environment for real-time surgeries. Previous work can be referred to as Google's StereoNet [10]. Given the truth that no corresponding binocular endoscopic data is available for neural network training, we have established a simulated 3D model to generate a large quantity of binocular endoscopic data within a very short time. Due to the special properties of the simulated dataset such as smoothness and limited texture, the L2 loss in the original model global optimization of disparity cannot well represent the image details. We have built a more effective loss function by including the spatial scale-invariant average squared error to further improve the quality of the evaluated disparity mapping function. After a series of testing, the proposed convolutional network can produce satisfactory results at a much faster speed of 45fps on an NVIDIA GTX 1080 GPU. To evaluate the performance of our neural network model, we have compared the simulation results from using other deep CNN frameworks or alike.

The contributions of this paper are summarized as follows. (1) The open-source modeling software(blender) is used to create a 3D gastrointestinal environment simulation model, which can generate numerous binocular endoscopic images with accurate disparity information in a short time. (2) A scale-invariant error loss function is proposed, providing a much more efficient evaluation of similarity for images in the endoscopic environment. (3) From the perspective of depth estimation in the endoscope environment, the proposed deep learning-based method is used for the first time to achieve an accurate real-time estimation of disparity. The simulation experiments demonstrate that our model can produce a full resolution disparity map in real-time processing. Besides, the training error of the network can reach as few as 0.41 pixels, and the processing time of a single image is 0.022 s.

II. RELATED WORK

In this section, the fundamentals of traditional methods to evaluate disparity, the relevant aspects of deep learning networks and their applications in medical image processing are introduced.

A. TRADITIONAL DISPARITY ESTIMATION

Traditional disparity estimation methods are based on feature matching between left and right images, and a typical stereo matching algorithm consists of four steps [11]: (1) Calculate the matching cost of each image patch within a disparity range; (2) Smooth the cost tensor obtained in the previous step through the aggregation method; (3) Estimate disparity by finding the lowest cost; (4) Refine the disparity by introducing a global smoothing function. According to the constraint range of the algorithm, it can be divided into local matching algorithm and global matching algorithm. Local matching method mainly studies the different strategies of matching cost and neighbor pixel aggregation [12]–[15], a simple Winner-Takes-All (WTA) selection strategy is usually used, but the Signal-Noise-Ratio (SNR) is increased by aggregation support window matching cost. Global algorithms include graph cutting, belief propagation, dynamic programming, particle swarm optimization, etc [16]–[19]. This kind of algorithm calculates the matching cost by pixels on the entire image, establishes a global energy function including data items and smoothing items, and calculate the optimal disparity value by minimizing the global energy function. HirschmÃijller *et al.* proposed a semi-global stereo matching algorithm based on the advantages of local and global algorithms. The algorithm adopts a global framework and uses an efficient one-dimensional path aggregation method to replace the two-dimensional minimization algorithm in the global algorithm.

B. DEEP LEARNING NETWORK

The latest research on deep learning for disparity estimation focuses on how to accurately calculate the matching cost and how to post-optimize the disparity map. Zbontar and Lecun [9], [20] proposed using the deep features of a neural network to calculate the matching cost for the first time, and they designed a deep Siamese convolution network to predict the similarity of blocks and calculate the matching cost. Luo [21] *et al.* accelerated the calculation of Siamese network matching cost by associating unitary features. PSMNet [22] refined the costs by implementing the spatial pyramid pool module. GwcNet [23] extended PSMNet with grouping related cost quantities and improved the 3D stacked hourglass network. In terms of post-processing of the disparity map, SGM-Net [24] made use of a neural network to predict the penalty parameters of SGM instead of manual adjustment or better results. L-ResMatch [25] solved mismatch by calculating the reflection confidence of the disparity map. GC-Net [26] combined environmental information to adjust the matching cost volume which further reduced the

mismatch of fuzzy areas. GA-Net [27] used a semi-global aggregation layer and a local boot aggregation layer to refine the fine structure. Liang [28] *et al.* included additional information such as semantic features to optimize the disparity map. All these methods have contributed to the field of computer vision, however, when applied to the endoscopic environment, they have faced challenges as mentioned in the introduction section.

C. MEDICAL IMAGE PROCESSING

A few works have been done in terms of using deep learning networks for estimating monocular endoscopic image depth, while the application to binocular images is blank mainly due to the absence of a reliable dataset. Faisal Mahmood [29] *et al.* combined deep convolutional neural networks and conditional random fields to estimate the depth of the monocular image. Nadeem and Kaufman [30] proposed a dictionary-based approach for depth estimation of colonoscopic monocular images. Reiter [31] *et al.* proposed a small network for three-dimensional reconstruction of the endoscopic sinus surgery. Xiongtong [32] *et al.* proposed a self-monitoring method to train convolutional neural networks to perform dense depth estimation from monocular endoscope data. Anita [33] *et al.* trained generative countermeasure network pix2pix to estimate depth from monocular endoscope images.

III. MODEL

A. MEDICAL SIMULATION DATA SYNTHESIS

Medical data has the characteristics of low relevance, incomplete records, and often contains a lot of personal privacy. As for binocular endoscopic data, it is even more difficult since few available instruments can deliver such data. In this paper, Blender, an open-source 3D animation production software, is used for building a precise 3D model of human internal organs, and Cycles rendering is used to improve the visualization effect.

The gastrointestinal model consists of stomach, intestine and gas pipelines. The establishment of the stomach model belongs to polygon modeling, which can be completed by using the mesh sphere for mesh segmentation, deformation, merging and chamfering. The establishment of intestinal and organ canal models belongs to curve modeling. Firstly, the Bessel curve is used for path fitting. Then, the mesh cube is used for extruding. Finally, the mesh segmentation, deformation, and chamfer are used to complete the modeling. The model can be selected different diffuse BSDF distributions for various materials to fulfill the whole gastrointestinal model, which is later supplemented with HD images of real gastrointestinal for UV editing and texture mapping. So far, the establishment of the entire model has been completed, as shown in 1.

Training and test datasets are generated using python scripts. We utilized the Bessel polynomial to fit the motion of the camera. After that, path constraints and the light source are added to the camera. The camera can be set as a multiview

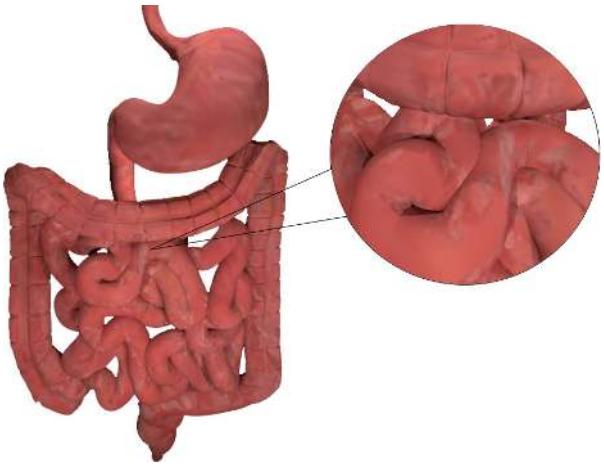


FIGURE 1. Human gastrointestinal simulation model and local enlarged view.

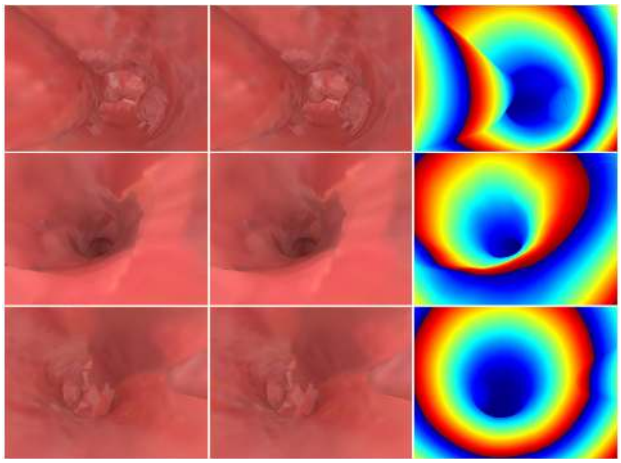


FIGURE 2. Examples of the simulation dataset, from left to right columns, are left camera view, right camera view, disparity image.

mode to generate the binocular view rendering. The stereo depth range depends on the camera parameters such as the focal length and the baseline. The disparity is calculated as (ref [11])

$$D = \frac{f^* b}{d}, \tag{1}$$

where f is the focal length of the camera, b is the baseline of the camera, d is the disparity, and D is the depth.

Binocular view, disparity map, and binocular video rendering are performed on a computer equipped with a 1080 module GPU. The focal length of the simulation camera is 15mm. The baseline is 1cm and the output mapping resolution is 960 * 720. The simulation results are shown in 2.

B. DISPARITY ESTIMATION FRAMEWORK

The disparity estimation of binocular images can be simply divided into two steps: image matching and disparity estimation. An important method of image matching is matching according to the feature information of the image. Our experiments demonstrate that low-resolution images can also

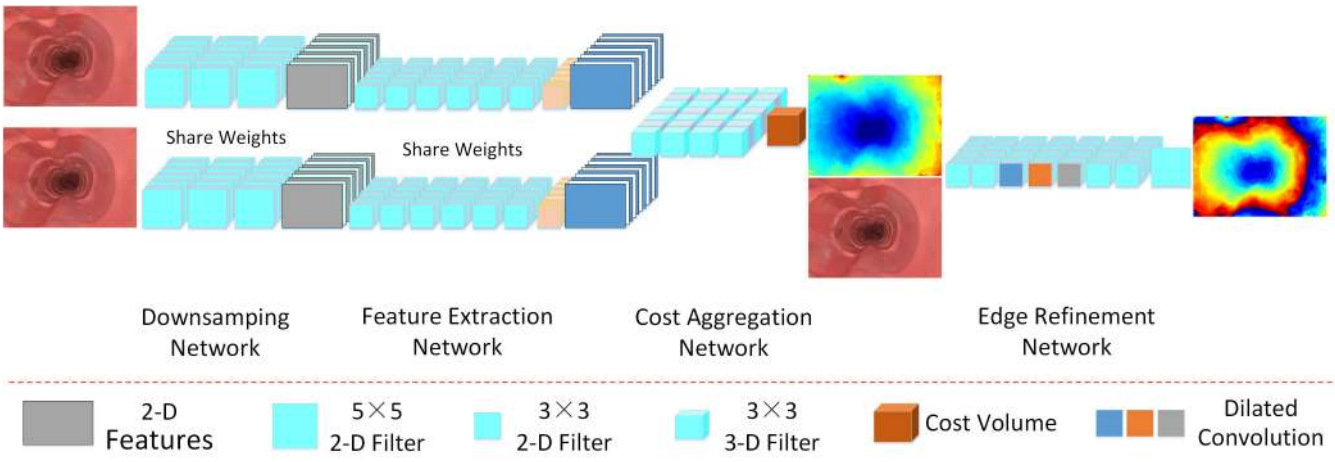


FIGURE 3. Network architecture. A pair of stereo images pass through the network for disparity prediction.

be adopted for binocular depth estimation tasks with acceptable precision to a certain extent. Compared with processing high-resolution images, processing low-resolution images can improve the efficiency of the algorithm. The traditional stereo matching method uses the optimal Euclidean distance between two feature vectors to judge whether it matches. In our model, the network learns an evaluation criterion according to the input feature vector, and then use this criterion to obtain the initial disparity.

Although the initial disparity value can be quickly estimated based on low-resolution images, the estimated disparity is rough and lacks a lot of detail. This paper uses the image color information to learn more details through edge enhancement network, and further improves the initial preliminary disparity value. Due to the special properties of the simulated dataset such as smoothness and limited texture, the model is trained by using the scale-invariant error loss function. The network structure model is shown in 3.

1) LIGHTWEIGHT NETWORK FEATURE EXTRACTION

We design a lightweight network to extract features through a low-resolution image, which can represent most of the feature information of the original image. Firstly, three convolutional layers of 5×5 are used to downsample high-resolution images with a step size of 2, such as the Downsampling Network part in Fig.3, where the gray part is the low-resolution representation of 32-channel images.

Siamese neural networks can learn a similarity measure from the image [34]. The input is mapped to the target space through a function $\phi_t(x)$, where t is the parameter. The purpose of Siamese neural networks is to find such a set of parameters as t, so that when X1 and X2 come from the same category, the loss of $L_t\ (X_1, X_2)$ is minimal (Eq.2),

$$L_t\ (X_1, X_2) = \|\phi_t\ (X_1) - \phi_t\ (X_2)\|\ . \tag{2}$$

Our model uses two Siamese neural networks [9] with shared weights to extract the features of the left and right binocular images. Deep network structures can handle the weak texture areas and small structures well and rapidly. This paper extracts low-resolution image features through 6 residual blocks, which is shown as Feature Extraction Network in Fig.3, where each residual block is composed of 3×3 convolution, batch regularization [35], and rectifying linear unit [36] operation. Finally, a 32-channel low-resolution image feature image (dark blue part in the figure) is output through an independent 2D convolution layer.

2) MATCHING COST CALCULATION

Firstly, we subtract the feature vector values of left and right binocular images to get the initial matching cost. A 3D convolution can be considered not only texture content information but also structural geometry information [37] because it considers three dimensions of height, width, and disparity. In this paper, four 3D convolution layers of 3×3 is used to learn a minimum matching cost measurement criterion, such as the Cost Aggregation Network part in Fig.3. Finally, a matching cost $C_i(d)$ (grey part in the figure) within the allowed disparity range is obtained. The traditional algorithm needs to calculate the optimal matching cost of this disparity range according to

$$d_i = \arg\min_d C_i(d), \tag{3}$$

where i is the disparity range, ranging from 1 to the maximum disparity D. When disparity i is d, the optimal matching cost is obtained, and the optimal disparity is d_i.

Since the matching cost $C_i(d)$ of a deep neural network is a high-dimensional vector, directly calculating its minimum value is non-differentiable, which causes the network to fail to transfer parameters. According to the suggestion of Kendall *et al.* [27], we calculated the optimal matching cost of disparity range through weighted regression function as

$$d_i = \sum_{d=1}^{D} d \cdot \frac{\exp\left(-C_i(d)\right)}{\sum_{d'} \exp\left(-C_i\left(d'\right)\right)}. \tag{4}$$

If a disparity d can minimize $C_i(d)$, it will be recovered by weighted averaging. At this point, the preliminary single-channel disparity estimation has been completed.

3) DISPARITY OPTIMIZATION

The disparity map obtained after coarse-grained depth estimation cannot satisfy the precision requirement, therefore the sub-pixel optimization technique is needed. The traditional SGM algorithm uses the quadratic curve interpolation method to obtain sub-pixel precision. To suppress noise, a median filter or bilateral filter is used for post-processing.

In this paper, we use the bilinear interpolation method to up-sample the low-resolution depth map. Moreover, we combine the depth map with the RGB binocular image, and use the color information of the RGB image to learn the edge details, as shown in the Edge Refinement Network of Fig.3.

Firstly, the network splices the preliminary disparity image and RGB image, and obtains the tensor of 32 channels through a 2D convolution layer of 3×3. Then, the 32-channel tensor passes through 6 residual blocks, in which the operation of dilation convolution is used in each residual block to expand the sensing field [38], and dilation is set as (1, 2, 4, 8, 1, 1). Finally, the tensor passes through a 2D 3 * 3 convolutional layer to obtain an edge-enhanced disparity effect.

4) IMPROVED LOSS FUNCTION

Due to the special properties of the simulated dataset such as smoothness and limited texture, the prediction of scene depth on a global scale is fuzziness. Much of the fuzziness can be explained by the degree to which average depth is predicted. We use scale-invariant error to measure the relationship between points in the scene, regardless of the absolute value of the absolute global scale (L2 loss). We then define an average squared error function with a constant log space scale [39].

$$L\left(y, y^*\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\log y_i - \log y_i^* + \alpha\left(y, y^*\right)\right)^2, \quad (5)$$

$$\alpha\left(y, y^*\right) = \frac{1}{n}\sum_{i}\left(\log y_i^* - \log y_i\right), \quad (6)$$

where y is the predicted disparity and y^* is the real disparity. $\alpha\left(y, y^*\right)$ is the average value of (y, y^*) interpolation in logarithmic space, so the scale-invariant error can be regarded as the error after averaging.

The model training loss function is defined as follows:

$$L\left(y, y^*\right) = \frac{1}{n}\sum_{i} d_i^2 - \frac{\lambda}{n^2}\left(\sum_{i} d_i\right)^2, \quad (7)$$

$$d_i = \log y_i - \log y_i^*, \quad (8)$$

where d_i is the disparity value of (y, y^*) at the i pixel. When λ is 1, there is a scaling relationship between the predicted value and the real value, and the loss is a scale-invariant error. When λ is 0, the predicted value is the same as the real value, and the loss is L2 loss. The model in this paper takes λ as 0.5, which contains both absolute information and relative information, making the image look more realistic.

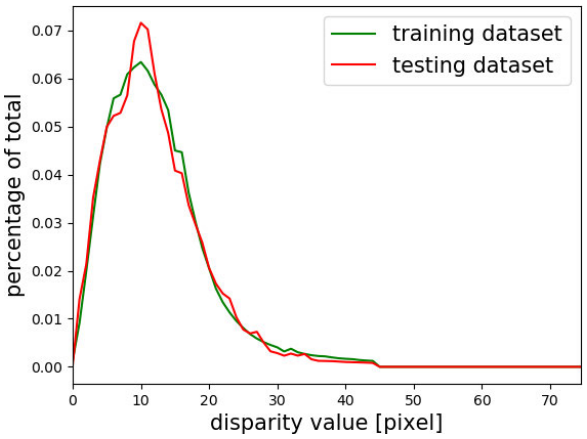


FIGURE 4. Disparity distribution of dataset.

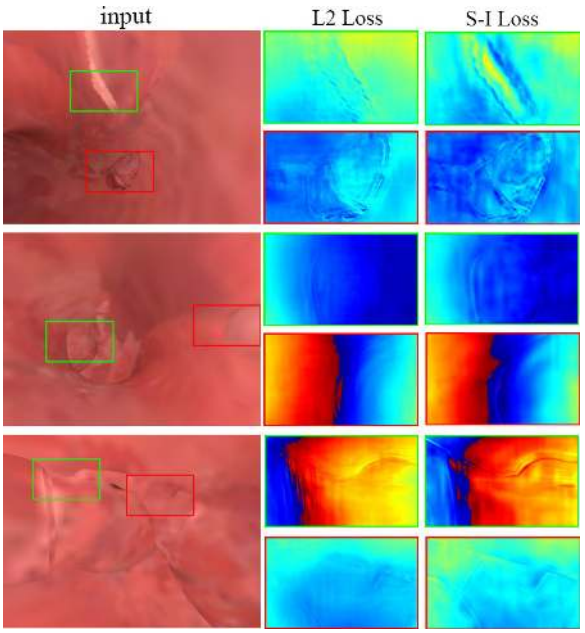


FIGURE 5. Effect after improving loss; L2 Loss: Loss of mean square error; S-L Loss: Loss of scale-invariant error.

IV. EXPERIMENTS AND DISCUSSIONS

In this paper, an efficient spatial scale-invariant mean square error loss function is designed based on StereoNet [10] network structure. The dataset used in our experiment is generated by blender, which contains the left and right binocular views, and the left and right binocular disparity maps. The camera parameters are set as described above. We generated 800 training sets and 100 test sets. The disparity distribution of the used dataset is such as Fig.4, and the maximum disparity value in the dataset is 69.5 pixels.

The deep learning model uses the PyTorch framework to conduct training and test on a single 1080 GPU. The model uses the RMSprop optimizer, and the initial learning rate is 1e-3. On the simulation dataset, the learning rate adjustment

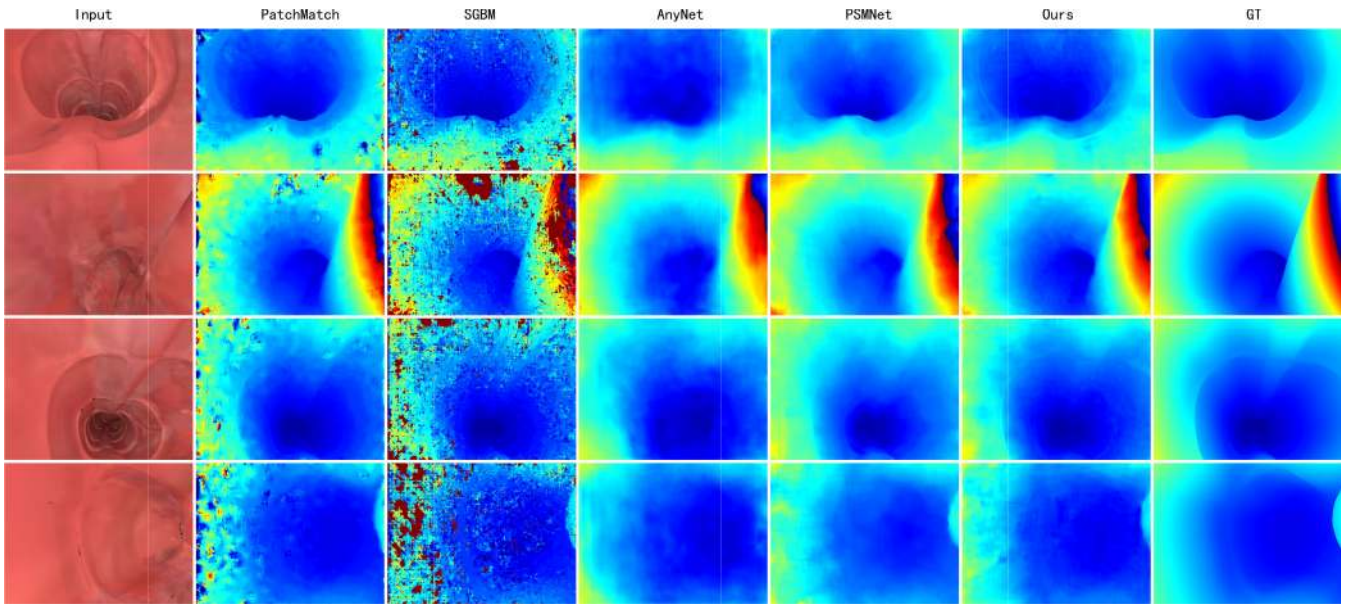


FIGURE 6. Disparity images of different methods are compared.

TABLE 1. Performance comparison of different depth estimation methods.

Method	KPE(%)			EPE(pixel)	RunTime(s)
	1pixel	3pixel	5pixel		
Ours	7.6925	1.0315	0.4143	0.4145	0.0218
AnyNetWithSPN	15.1508	2.5635	1.0286	0.6460	0.0422
AnyNetNoSPN	19.0698	2.9734	1.1392	0.7253	0.0361
PatchMatch	18.3242	5.8759	3.2821	0.9127	14.2051
SGBM	49.7855	22.1498	15.5069	5.9222	0.1802
PSMNet	8.2529	1.3141	0.5149	1.0899	0.2661

strategy is $lr * \lambda^n$. The training process lasts for a total of 16 epochs. Each epoch consists of the initial estimation of disparity and the optimization stage of disparity. The maximum disparity value of the model is set as 192. Dataset training took a total of 1.2 hours, and all results are verified on average on 100 random datasets. The result after improving loss is as shown in Fig.5.

Besides, we compare our model with different disparity estimation methods. AnyNet [37] network, PatchMatch [12] algorithm, SGBM [17], and PSMNet [22] methods are implemented respectively. The experimental results are shown in Fig.6. The experiment uses two common metrics: (1) Endpoint-error (EPE): Average Euclidean distance between estimated disparity and ground-truth, the smaller the test result, the better. (2) K-pixel-error (KPE): Endpoint error exceeds the percentage of k pixels, the smaller the test result, the better. We calculate the KPE, EPE and single image running time of each method as shown in Tab.1

It can be seen from Fig.6 that the disparity map produced by our method has the best result and is closest to the real disparity map. The disparity maps produced by AnyNet and PSMNet network are slightly fuzzy, the disparity map produced by PatchMatch has a small amount of noise, and the disparity map produced by SGBM algorithm has a large amount of noise. From the Tab.1, the error percentage of each pixel of our model is smaller than that of the comparison methods, and the average EPE is also the smallest, which further indicates that our model has the highest accuracy and is closest to the real disparity map. The accuracy of SGBM algorithm is the lowest, which also explains the phenomenon that the disparity map generated by the algorithm has a large amount of noise. In terms of the running time of a single image, our model takes the least time(only 0.0218s), that is, 45FPS, while the running speed of the AnyNet network is only 27 FPS. So our model can achieve the purpose of real-time image processing.

The training result of the neural network model will be affected by some factors. In this learning network model, the training data volume, maximum initial disparity, and post-processing edge enhancement are mainly included. The following factors are analyzed and the generalization ability of the model is analyzed.

A. DATA VOLUME EFFECT

To find a more suitable amount of data corresponding to the model, we experimented on the influence of different amounts of data on the model. At the same time, the test set is added to the training set to test the generalization ability of the model. Average loss of two stages and test error analysis of different Numbers of datasets, the results are as shown in Fig.7 and Fig.8.

According to the loss curve, when the data amount of the training set is 400,800,900, the loss value reaches the minimum and tends to be stable with the increase of the iteration epochs. When the data amount is 100, the loss value

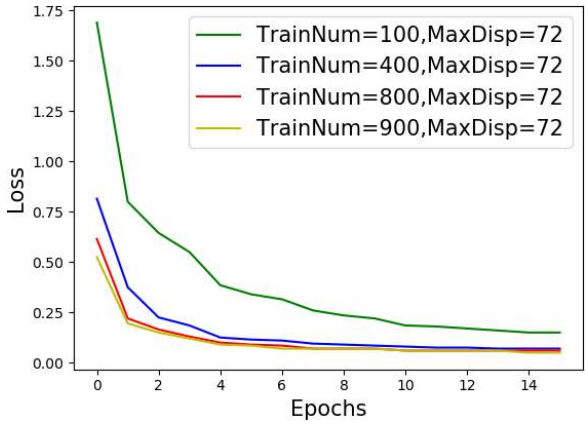


FIGURE 7. Loss curves of different data volume.

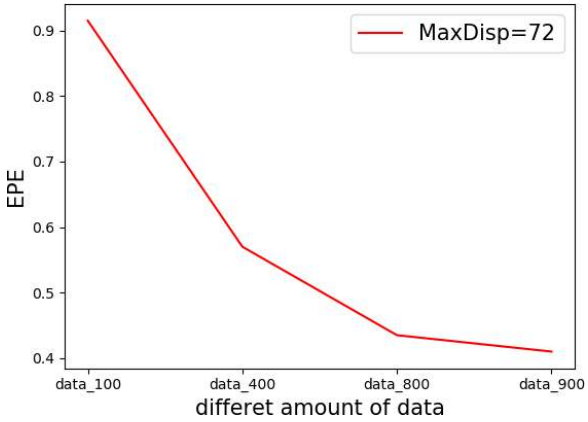


FIGURE 8. EPE curve of different data volume.

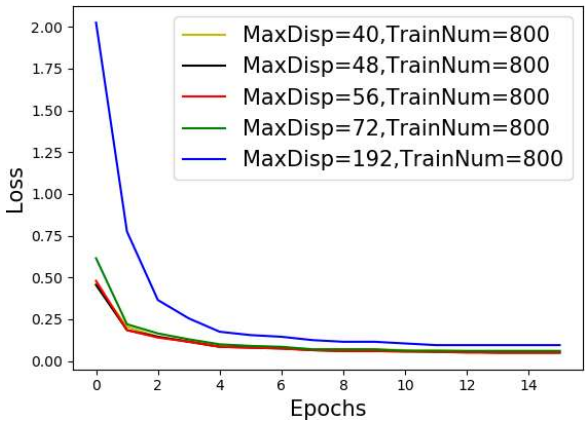


FIGURE 9. Loss curves of different maximum disparity.

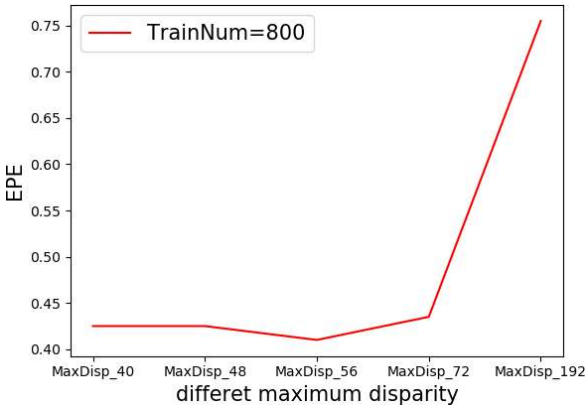


FIGURE 10. EPE curve of different maximum disparity.

of the model is much larger than other data amounts, and the loss curve is not yet stable. According to the error of the test set, it can be observed that the error is large when the training data amount is 100. With the increase of training data, the test error decreases gradually. It can be seen that our model performs well when the data volume is about 800, and the test error can be reduced after adding the training set into the test set. An appropriate increase in the amount of data can slightly fine-tune the accuracy of the model, which indicates that the model has a good generalization ability.

B. MAXIMUM DISPARITY EFFECT

The deep learning model needs an initial maximum disparity for learning, while some images in the simulation data set have a large deviation. According to the statistical results of disparity distribution in the dataset, the maximum disparity value in the dataset is 69.5. Since the model starts to sample the original resolution image three times, the convolution step is 2, so the initial maximum disparity value should be a multiple of 8. Considering the influence of different maximum initial disparity values on the model, the data volume of the fixed training set is 800, and the maximum disparity values are set as 192,72,56,48,40 for experimental analysis. The analysis results are shown in Fig.9 and Fig.10.

According to Fig.9 and Fig.10, the loss with different maximum disparity tends to be stable and reaches the minimum around the 12th iteration. When the maximum disparity is 192, the loss is slightly larger than in other situations. When the maximum disparity is 56, the loss is the minimum. The EPE curve shows that the maximum disparity is set to 192 with a great error, 72, 56, 48, and 40 with a little difference, but the test error is slightly smaller at 56. It can be seen that the setting of the maximum disparity of the model should be based on the maximum disparity of the dataset. Considering the partial deviation of the dataset, the setting of the maximum disparity should be slightly lower than the maximum disparity of the dataset.

C. EDGE ENHANCEMENT EFFECT

The recovery of low-resolution disparity to high-resolution disparity consists of two stages. In the first stage, the low-resolution image is directly sampled by bilinear interpolation up to the original resolution image size. In the second stage, RGB image color information is combined with edge enhancement network learning edge information to restore the original resolution image size, as shown in Fig.3. To measure the influence of direct sampling and edge enhancement on the model under different datasets, we fixed the maximum disparity as 56, and calculated the average loss and test error

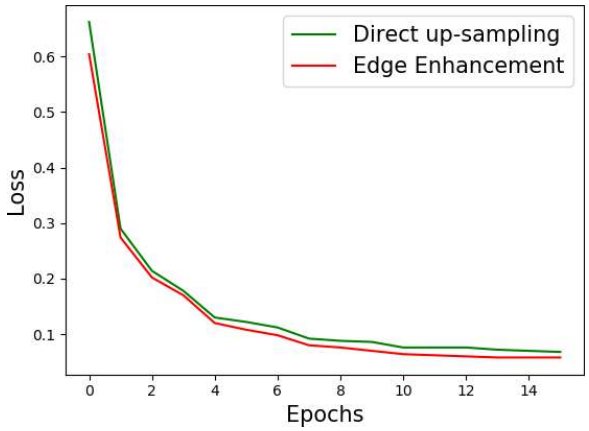


FIGURE 11. Whether the edge enhances the loss curve.

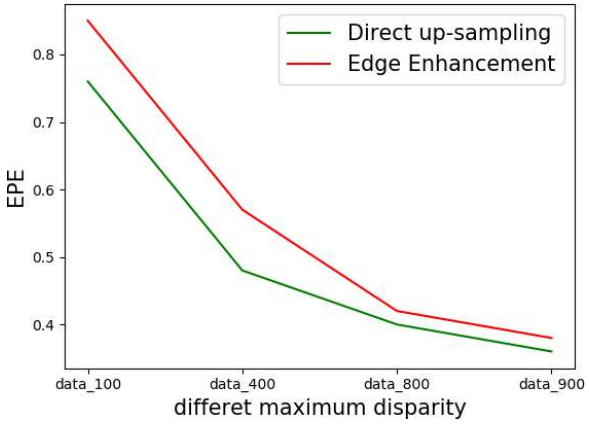


FIGURE 12. Whether the edge of different data volume enhances the error.

of the two methods under different datasets. The analysis results are as shown in Fig.11 and Fig.12.

The loss with edge enhancement is slightly less than the loss with direct up-sampling. According to the generation of results, when there are more contour images in the simulation data, the edge enhancement effect is obvious. However, when the image in the dataset is less contoured, the enhanced part only plays a smoothing role. According to the EPE curve, the test error after edge enhancement only increases slightly, The reason of this phenomenon is there are relatively few pictures with abundant edges in the dataset, which further indicates that our model can be well applied to the real datasets.

V. CONCLUSION

We have proposed a new CNN-based approach to realize the real-time depth mapping for endoscopy imaging by overcoming the problem of inadequate training data and slow speed. The established 3D simulation model has the advantages of easy access, high efficiency, low cost, relatively high precision with ground-truth data and independence of hardware difference, which can provide considerable training data for neural networks in general. Compared with state-of-the-art methods(AnyNet,PatchMatch,SGBM,PSMNet), our proposed deep learning method has the smallest EPE (0.4145 pixels) and the fastest running speed (0.0218s, 45fps). With the real-time image processing capacity, this method is suitable for surgical navigation that relies on on-site 3D positioning information. We have further analyzed the influencing factors of the proposed network model from the perspective of training data size, maximum initial disparity, and post-processing edge enhancement. The statistics show that our model performs well under a small amount of training data down to 1000 images, which is suitable for the medical imaging application as training data are typically insufficient. The maximum initial disparity should be determined according to the maximum disparity estimation of the applicable scene, and the performance of the model will be improved from the obvious data sets such as the edge contour.

Future work will focus on incorporating the developed method with more complicated medical environments (gloss, texturization, occlusion and etc.). Particularly, we have planned a cooperation with corresponding hospitals (Beijing Friendship Hospital, Capital Medical University) to test and improve our model by training under actual datasets.

ACKNOWLEDGMENT

Xiong-Zhi Wang and Yunfeng Nie contributed equally to this work.

REFERENCES

[1] R. H. Taylor, A. Menciassi, G. Fichtinger, P. Fiorini, and P. Dario, "Medical robotics and computer-integrated surgery," in *Springer handbook Robotics* Berlin, Germany: Springer, 2016, pp. 1657–1684.

[2] R. L. Galloway, "The process and development of image-guided procedures," *Annu. Rev. Biomed. Eng.*, vol. 3, no. 1, pp. 83–108, Aug. 2001.

[3] T. M. Peters, "Image-guidance for surgical procedures," *Phys. Med. Biol.*, vol. 51, no. 14, pp. R505–R540, Jul. 2006.

[4] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1851–1858.

[5] J. M. Fitzpatrick, "The role of registration in accurate surgical guidance," *Proc. Inst. Mech. Eng., H, J. Eng. Med.*, vol. 224, no. 5, pp. 607–622, May 2010.

[6] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. V. D. Smagt, D. Cremers, and T. Brox, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.

[7] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4040–4048.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, pp. 234–241, 2015.

[9] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *Proc. ICML*, vol. 17, 2016, p. 2.

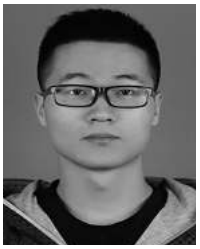
[10] S. Khamis, S. Fanello, C. Rhemann, A. Kowdle, J. Valentin, and S. Izadi, "Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 573–590.

[11] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," in *Proc. IEEE Workshop Stereo Multi-Baseline Vis. (SMBV)*, 2002, pp. 7–42.

[12] M. Bleyer, C. Rhemann, and C. Rother, "PatchMatch Stereo–Stereo matching with slanted support windows," in *Proc. Brit. Mach. Vis. Conf.*, 2011, pp. 1–11.

[13] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 504–511, Feb. 2013.

[14] X. Mei, X. Sun, W. Dong, H. Wang, and X. Zhang, "Segment-tree based cost aggregation for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 313–320.
[15] Q. Yang, "A non-local cost aggregation method for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1402–1409.
[16] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," *Int. J. Comput. Vis.*, vol. 70, no. 1, pp. 41–54, Oct. 2006.
[17] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2005, pp. 807–814.
[18] A. Klaus, M. Sormann, and K. Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, May 2006, pp. 15–18.
[19] J. Sun, N.-N. Zheng, and H.-Y. Shum, "Stereo matching using belief propagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 787–800, 2003.
[20] J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1592–1599.
[21] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5695–5703.
[22] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5410–5418.
[23] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3273–3282.
[24] A. Seki and M. Pollefeys, "SGM-nets: Semi-global matching with neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 231–240.
[25] A. Shaked and L. Wolf, "Improved stereo matching with constant highway networks and reflective confidence learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4641–4650.
[26] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 66–75.
[27] F. Zhang, V. Prisacariu, R. Yang, and P. H. S. Torr, "GA-net: Guided aggregation net for End-To-End stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 185–194.
[28] Z. Liang, Y. Feng, Y. Guo, H. Liu, W. Chen, L. Qiao, L. Zhou, and J. Zhang, "Learning for disparity estimation through feature constancy," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2811–2820.
[29] F. Mahmood, R. Chen, S. Sudarsky, D. Yu, and N. J. Durr, "Deep learning with cinematic rendering: Fine-tuning deep neural networks using photorealistic medical images," *Phys. Med. Biol.*, vol. 63, no. 18, 2018, Art. no. 185012.
[30] S. Nadeem and A. Kaufman, "Computer-aided detection of polyps in optical colonoscopy images," in *Proc. Med. Imag., Comput.-Aided Diagnosis*, Mar. 2016, Art. no. 978525.
[31] A. Reiter, S. Leonard, A. Sinha, M. Ishii, R. H. Taylor, and G. D. Hager, "Endoscopic-CT: Learning-based photometric reconstruction for endoscopic sinus surgery," in *Proc. Med. Imag., Image Process.*, Mar. 2016, Art. no. 978418.
[32] X. Liu, A. Sinha, M. Unberath, M. Ishii, G. D. Hager, R. H. Taylor, and A. Reiter, "Self-supervised learning for dense depth estimation in monocular endoscopy," in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Berlin, Germany: Springer, 2018, pp. 128–138.
[33] A. Rau, P. J. E. Edwards, O. F. Ahmad, P. Riordan, M. Janatka, L. B. Lovat, and D. Stoyanov, "Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 14, no. 7, pp. 1167–1176, Jul. 2019.
[34] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2005, pp. 539–546.
[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167
[36] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, vol. 30, 2013, p. 3.
[37] Y. Wang, Z. Lai, G. Huang, B. H. Wang, L. van der Maaten, M. Campbell, and K. Q. Weinberger, "Anytime stereo image depth estimation on mobile devices," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 5893–5900.
[38] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 472–480.
[39] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Adv. Neural Inform. Process. Syst.*, 2014, pp. 2366–2374.



XIONG-ZHI WANG was born in Weinan, Shaanxi, China, in 1995. He is currently pursuing the master's degree in computer technology with the Department of Computer Science and Technology, China Xidian University, Xi'an, China. He is also an exchange member of the School of Future Technology, University of Chinese Academy of Sciences, Beijing, China, and the College of Computer Science, Nankai University, China. His current research interests include computer vision and machine learning (especially deep learning).

YUNFENG NIE received the Ph.D. degree in optical engineering from Vrije Universiteit Brussel under EU's FP7 Marie Curie Programme 'ADOPSYS' with exchanges in LPI Company, Madrid, Spain, and the University of Jena, Germany. She has been a full-time Researcher with the Faculty of Engineering, VUB, since 2014. She has been very active in freeform optical design algorithms, biomedical photonics, imaging spectrometers, and computational imaging.

SHAO-PING LU (Member, IEEE) received the Ph.D. degree in computer science from Tsinghua University, China, in 2012. He worked as a Postdoctoral and a Senior Researcher with Vrije Universiteit Brussel (VUB), from 2013 to 2017. He is currently an Associate Professor with Nankai University, Tianjin, China. His research interests include intersection of visual computing, with particular focus on 2D&3D image and video processing, computational photography and representation, visual scene analysis, machine learning, and mathematical optimization.

JINGANG ZHANG is currently an Associate Professor with the University of Chinese Academy of Sciences (UCAS). He has presided over more than ten national and ministerial-level scientific research projects, such as the National Natural Science Foundation of China and the Joint Foundation Program of the Chinese Academy of Sciences for equipment pre-feasibility study. His research interests include image denosing, deblurring, and dehazing, image/video analysis and enhancement, and related high-level vision problems.