IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 26, 2021, accepted February 1, 2021, date of publication February 23, 2021, date of current version March 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3061495

TSFE-Net: Two-Stream Feature Extraction Networks for Active Stereo Matching

HAOJIE ZENG[1], BIN WANG[1], XIAOPING ZHOU[1], XIAOJING SUN[1], LONGXIANG HUANG[2], QIAN ZHANG[1], AND YANG WANG[1]

[1]College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai 200234, China
[2]Shenzhen Guangjian Technology Company Ltd., Shanghai 200135, China

Corresponding authors: Bin Wang (binwang@shnu.edu.cn) and Xiaoping Zhou (zxpshnu@163.com)

This work was supported by the Program of Shanghai Normal University (No. C-9000-20-309119) and the Shanghai Capacity Building Projects in Local Institutions under Grant 19070502900.

ABSTRACT In this paper, we propose TSFE-Net, two stream feature extraction networks for active stereo matching. First, we perform extra local contrast normalization (LCN) for dataset due to dependency between speckle intensity and distance. Second, we construct two stream feature extraction layers which consist of convolutional layers and deconvolutional layers in different scales to simultaneously learn the features of the original images and LCN images and aggregate context information to form the left and right features. Third, we convert the obtained depth map into disparity map in virtue of camera parameters to construct a supervised learning model. The TSFE-Net not only solves illumination effects between speckle intensity and distance but also reserves details of the original image. Our dataset are captured by RealSense D435 camera. We research extensive quantitative and qualitative evaluations based on a series of scenes, and achieve the end point error (EPE) accuracy of 0.335 on the TITAN XP platform only for valid pixel. The assessment results show that our network has the ability of real-time deep reconstruction for active pattern.

INDEX TERMS Active stereo matching, convolutional neural network, depth reconstruction, two stream feature extraction.

I. INTRODUCTION

Depth reconstruction technology is popular in computer vision, which is essential to virtual reality, augmented reality [1], and the fields of vehicle automated driving.

Depth reconstruction systems are divided into passive stereo systems and active stereo systems. The passive stereo system directly captures the target scene without additional light source. On the contrary, the active stereo system needs to actively project the light source to the target scene. By selecting the correct sensing wavelength, the camera captures the combination of active illumination and passive light. The texture of the target scene is enhanced by active light projector, which can solve a variety of real-world problems, such as textureless areas (e.g., slant wall and smooth object etc.), and thin structures. There are some applications based on active stereo system, such as Time of Flight (TOF) [2], Structured Light (SL) [3], and binocular stereo matching. TOF measures the distance based on the round-trip time of the emitted light between the object and the receiver. SL calculates the modulation pattern of the target scene based on the triangulation principle to obtain depth information. Binocular stereo matching calculates all corresponding pixels of object between two images. Our research is based on active binocular stereo matching.

Binocular stereo matching technology computes the disparity for all pixels in a pair of rectified images. Disparity is the horizontal displacement between a pair of corresponding pixels on the left and right images. For example, for the pixel whose position is (x,y) in the left image, and its corresponding point is found at (x-d,y) in the right image, where d is the disparity of this pixel. The depth is calculated by f*b/d, where f is the camera's focal length and b is the baseline distance between two camera centers. Traditional binocular stereo matching algorithms can be divided into local stereo matching algorithms [4]–[7], semi-global stereo matching algorithms (SGBM) [8]–[10], and global stereo matching algorithms [11]–[14] depending on the different matching strategies. Local methods aggregate matching costs with neighboring pixels and usually utilize the winner-take-all(WTA) strategy to choose the optimal disparity. For example, typical local stereo matching algorithms include block

The associate editor coordinating the review of this manuscript and approving it for publication was Andrea F. Abate.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

matching (BM) [15] algorithm. The global algorithms calculate the global energy function and minimize it to find the optimal disparity, such as dynamic programming (DP) [16] and belief propagation (BP) [17]. Unfortunately, these algorithms are sensitive to noise during the calculation process. TOF and SL are highly adaptable to the scene for day and night. However, TOF has motion artifacts and multipath interference. SL methods are susceptible to environmental lighting and interference with multiple devices. Recently, the deep architecture [18]–[21] based on the convolutional neural networks(CNNs) has become popular. It extracts information of images and aggregates contextual information by fusing spatial and channel information of local receptive fields. Although aforementioned methods have made significant progress, they are only suitable for deep reconstruction of scene with strong texture instead of textureless region (e.g., slant wall). Active binocular reconstruction accurately calculate depth of textureless areas in virtue of speckle information and scene itself information. For instance, when estimating the depth of the slant wall, it is difficult to complete stereo matching of corresponding pixels because pixels with different depths have close gray level information or RGB information. However, the existence of speckle information makes the pixel information of different depths various to complete the matching process of two images. Active Stereo Net [22], which is the first proposed deep learning solution for active stereo systems. Active Stereo Net can reconstruct depth of textureless scene based on active illumination.

In this paper, we propose two stream feature extraction networks for active stereo matching(TSFE-Net) based on end-to-end deep learning approach. It extends recent work on self-supervised active stereo net [22] and supervised passive stereo network [23] to achieve active binocular reconstruction. Our network solves reconstruction issue in textureless region. The feature extraction layer of cascade network [23] is modified to improve the performance and reduce computation. Besides, two stream feature extraction layers are constructed to learn the features of both original images and LCN images which decrease fading of active stereo patters with distance [22] and reserve details of original images. The main work of our paper can be summarized as follows:

1.Our TSFE-Net simultaneously learns the features of the original images and LCN images based on weighted local contrast normalization with two stream feature extraction layers.

2.The feature extraction layers use convolution and deconvolution of different scales and dimensions to extract feature information and aggregate context information in the feature extraction layers.

3.We independently build the dataset that include different scenes(indoor scenes and outdoor scenes) at different illumination intensity with RealSense D435 camera.

4.We convert the obtained depth map from RealSense D435 camera into disparity map in virtue of camera parameters to construct an end to end supervised learning model.

II. RELATED WORK

Depth estimation methods are categorized in traditional methods and CNN methods based on deep learning.

A. TRADITIONAL METHODS

Traditional methods which are suitable for passive and active patterns include block matching(BM) [15], semi-global matching(SGBM) [10], BP [17], and graph cuts(GC) [24]. BM method refers to find correspondence in target image for all pixels of reference image within a window when the pixel's energy function is the minimum. SGBM method calculates cost volume from 16 or 8 directions to smooth cost with neighborhood constraints, similar to the dynamic programming(DP) [16] method. The DP method minimizes global energy by finding the minimum cost of pixel. GC method divides images to several non-overlapping patches based on gray information or color information, texture information and other features to make these patches are similar in the same block and show obvious differences between different blocks.

B. DEEP LEARNING METHODS

Convolutional networks have been proven very successful for a variety of recognition tasks, such as image classification, and face detection. According to the survey by Scharstein *et al.* [25], a typical stereo matching algorithm contains four steps: matching cost calculation, matching cost aggregation, disparity calculation, and disparity refinement. The end to end learning methods based on deep learning can be divided into passive stereo matching algorithms and active stereo matching algorithms.

In passive stereo matching systems, Zbontar and LeCun [26] firstly put forward using neural networks to compute matching costs. Mayer *et al.* [18] demonstrate the first scene flow estimation with a convolutional network. Kendall *et al.* present GC-Net [20] which construct a 3D cost volume and use a differentiable soft argmin operation to figure out the best matching disparity values from the cost volume and achieve sub-pixel accuracy without any additional post-processing or regularization. Following GC-Net, Chang and Chen [27] propose a pyramid pooling module to exploit global context information and introduce a stacked 3D hourglass networks to regularize cost volume. Khamis *et al.* [28] use a siamese network to extract features from the left and right image, and hierarchically recover high-frequency details through a learned upsampling function combining original image. Guo *et al.* [29] propose the group-wise correlation stereo network (GWC-Net) to construct the cost volume by group-wise correlation. The left features and the right features are divided into groups along the channel dimension, and correlation maps are computed among each group to obtain multiple matching cost proposals packed into a cost volume. Gu *et al.* [23] introduce a efficient cost volume formulation complementary in memory and time which can narrow the depth (or disparity) range of each stage by the prediction from the previous stage.

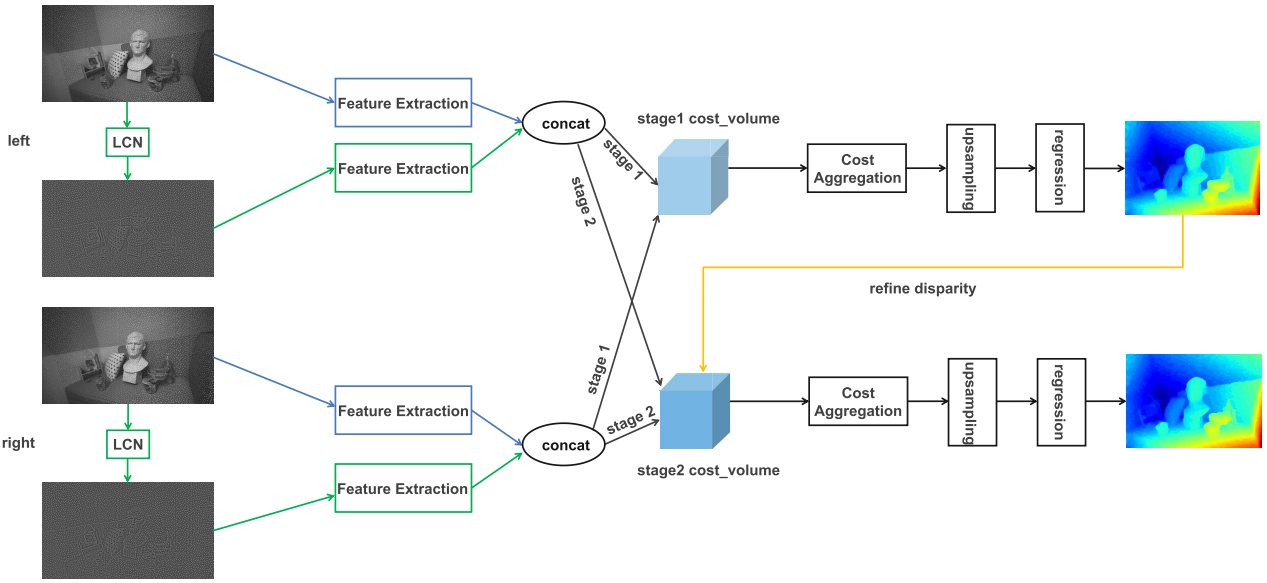


FIGURE 1. The pipeline of the network.

In active stereo matching systems, [30] present Ultra Stereo Net which uses an unsupervised greedy optimization scheme to learn discriminative features for estimating correspondences in infrared images. Zhang *et al.* [22] introduce a self-supervised active stereo matching and use local contrast normalization (LCN). Besides, they propose a window-based loss aggregation with adaptive weights for each pixel to increase its discriminability and reduce the effect of local minima in the cost function. [31] improves the Siamese network by combining pyramid-pooling structure with the squeeze-and-excitation network. In [32], a pair of left and right images with abundant structured light information is adopted to acquire a coarse depth map by unsupervised CNNs which is used for phase unwrapping and phase matching to obtain accurate depth. However, simply feeding the LCN image into the model ignores the details of the original image. Compared to the above methods, our model consider original feature of input images and reduce dependency between brightness and distance.

III. METHODOLOGY

Our model is illustrated in Figure 1. The input of network is a pair of rectified images with active illumination captured from RealSense D435 camera. The network consists of four parts: extracting original image feature and LCN image feature after local contrast normalization, constructing cost volume, aggregating the cost volume, and refining the disparity map. It will be described in the following sections.

A. FEATURE EXTRACTION

Feature extraction is done by a series of convolutional and deconvolutional layers with different scales in the feature extraction layer. The purpose is to generate simple and reliable feature representation of pixel-wise from the input images which is useful for matching cost volume.

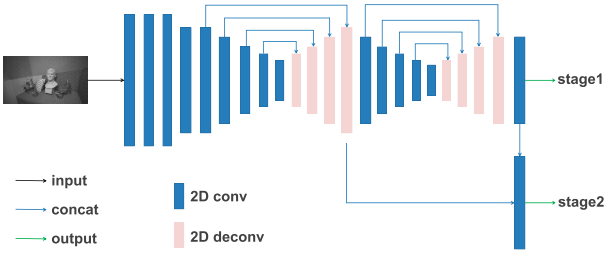


FIGURE 2. Feature extraction layer. The sub-network includes two stages, and each stage involves symmetric convolution and deconvolution to learn feature. The feature extraction part outputs feature representation of different scale and different resolution. Stage1 represents 1/4 resolution of original size. Stage2 represents 1/2 resolution of original size.

Specifically, given a pair of pixel l, r, we attempt to learn a set of deep features F_l and F_r. [22] computes the local mean μ and standard deviation σ in a small 9×9 patch to normalize the current pixel intensity. LCN is used to remove illumination effects between speckle intensity and scene distance. However, LCN ignores some details of original image due to normalization in patch. Therefore, we exploit two stream feature extraction layers to respectively extract features of original image and LCN image. The weights of every feature extraction layer are shared for left input and right input. The impressive results are shown in Table 3. In the feature extraction layer, 2D convolution layers are used to conduct down-sampling image. In order to avoid losing information, our model uses 2D deconvolution layers combining features of the previous layer to expand receptive field, as is illustrated in Figure 2. The layer outputs two stages of feature representation with $1/2$ resolution and $1/4$ resolution.

B. COST VOLUME FORMULATION

As shown in Figure 3, we construct a cost volume by concatenating features F_l and F_r of left images and right images.

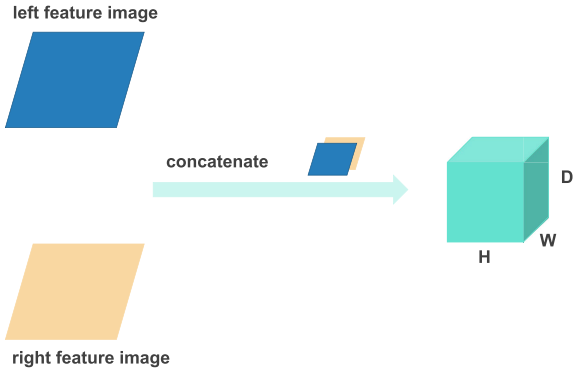


FIGURE 3. The formulation of cost volume. The size is *channel* $*$ *D* $*$ *H* $*$ *W*, where channel is the feature channels of feature maps,D is the disparity range, H and W is the spatial resolution of feature maps respectively.

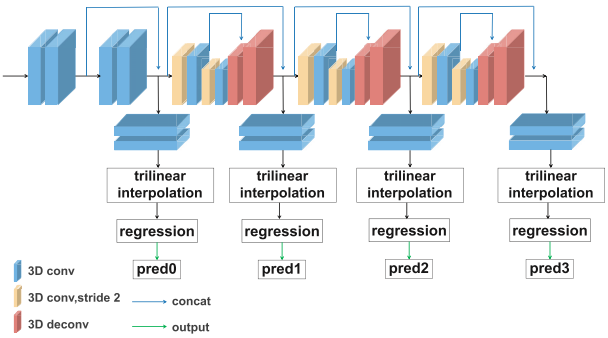


FIGURE 4. Cost aggregation layer which is based on the cost aggregation layer of the cascade network.

As mentioned in [23], we construct the two stage cost volume corresponding to the feature extraction layer. Differently from [23],the disparity range of the stage1 covers the entire disparity(maximum disparity is 144) range of the input scene. Besides, we denote the disparity interval at the first stage as 3 and the disparity interval sets to 1 at the second stage.

C. COST AGGREGATION

Our cost aggregation layer is based on the cost aggregation architectures of the Cascade Net [23], as is illustrated in Figure 4. The layer has three main hourglass networks. And outputs four predicted disparity maps and losses (pred0, pred1, pred2, and pred3). The loss function is described in Sect.III.D. Disparity regression defined in (1). The predicted disparity $\hat{d}$ is calculated via each disparity d and cost c_d based on soft argmin operation $\sigma(.)$.

$$\hat{d}_i = \sum_{d=0}^{D} d * \sigma(-c_d) \tag{1}$$

D. LOSS FUNCTION

We train our network in a fully supervised manner using groundtruth data from RealSense D435 camera. The loss function is calculated as the weighted summation of the four losses mentioned in [23], which is shown in (2):

$$L = \sum_{\rho=0}^{3} \sigma_\rho * loss_\rho \tag{2}$$

where $loss_\rho = \frac{1}{N} \times \sum_{i=1}^{N} smmoth(d_{gt}^i - \hat{d}_\rho^i)$, N is the sum of valid disparity value from RealSense D435 camera, d_{gt}^i is the disparity at valid pixel i, and $\hat{d}_\rho^i$ is the predicted disparity at the same pixel i, $loss_\rho(\rho = 0, 1, 2, 3)$ is output ρ from cost aggregation layer. As mentioned in [23], σ_ρ is the weight values which are set to be 0.5, 0.7, 1.0, 1.0.

IV. EXPERIMENTS

In this section, we evaluate our proposed two stream feature extraction network(TSFE-Net) on RealSense D435 dataset. The dataset and implementation details are described in Sect. IV.A. The accuracy and the performance of computational cost and time are showed in Sect. IV.B and Sect. IV.C. The results of ablation experiments are showed in Sect.IV.D.

A. DATASET AND SETUP

Our dataset provide 10530 training and 857 testing images of size 1280 $\times$ 720 with ground-truth depth maps, which are captured by the RealSense D435 camera. We used the camera to capture the left infrared image, right infrared image and depth value, as shown in Figure 5. We take the left and right infrared images as input, and convert the depth map into the disparity map as GroundTruth. We transform the depth map into disparity map by 641.44989 $\times$ 50 $\div$ *depth* (the focal length is 641.44989 pixel and the baseline is 50mm). The dataset contains a variety of scenes such as lab scene, office scene, kitchen scene, and bedroom scene. Our network is implemented with Pytorch. We use Kingma and Adam [33] optimizer, with $\beta_1 = 0.9$, $\beta_2 = 0.99$. In the process of feature extraction layer, the maximum downsampling resolution is a multiple of 64, so we need to crop the input image. To enhance the variety of scenarios, we train our model with a batch size of 4 on 2 Nvidia TITAN XP GPUs using 256 $\times$ 512 random crops from the train images. We test model with a batch size of 1 using 704 $\times$ 1024 fixed crops from the test images. According to the camera's depth range, the maximum of the disparity is set as 144. Before training, we normalize all images by subtracting their means and dividing their standard deviations, and use local contrast normalization to process input image with a window size of 11*11. For our RealSense D435 dataset, we train TSFE-Net for 20 epochs. In order to make the model converge, we verify that the dynamic learning rate has more advantages through experiments. The initial learning rate is set to be 0.001 for 6 epochs, and reduced by $\frac{4}{5}$ at epoch 7 and downscale by 2 after epoch 14.

The evaluation metrics include the end point error (EPE), the percentage of pixels which have greater than three pixel or 5% disparity error(D1-valid), the percentage of pixels with disparity error larger than 1 (Thres1), the percentage of pixels with disparity error larger than 2 (Thres2), the percentage of

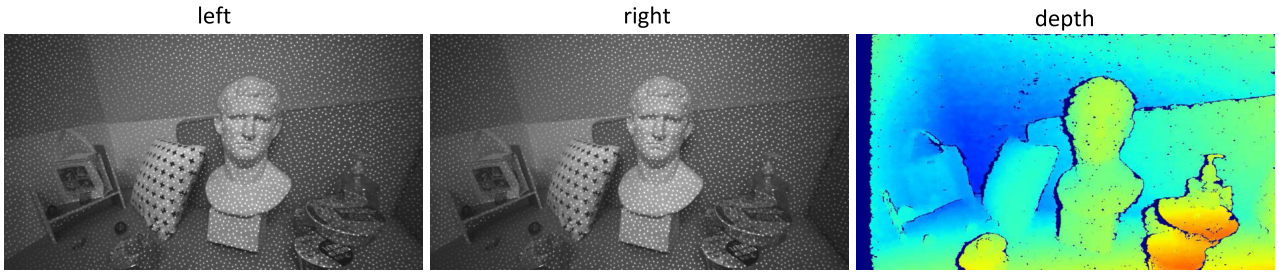


FIGURE 5. Sample images of dataset.

TABLE 1. Comparison of different models. Accuracy and speed metrics are used for evaluations on RealSense D435 dataset.

Model	EPE(px)	D1(%)	Thres1(%)	Thres2(%)	Thres3(%)	Memory(MiB)	Time(s)
PSM-Net [27]	0.971	2.0	4.9	2.7	2.2	7493	0.682
Stereo-Net [28]	2.22	4.9	8.9	5.6	4.9	1055	**0.05**
GWC-Net [29]	0.398	**0.9**	4.3	**1.6**	**1.1**	5465	0.452
TSFE-Net	**0.335**	1.4	**3.3**	2.0	1.5	4299	0.181

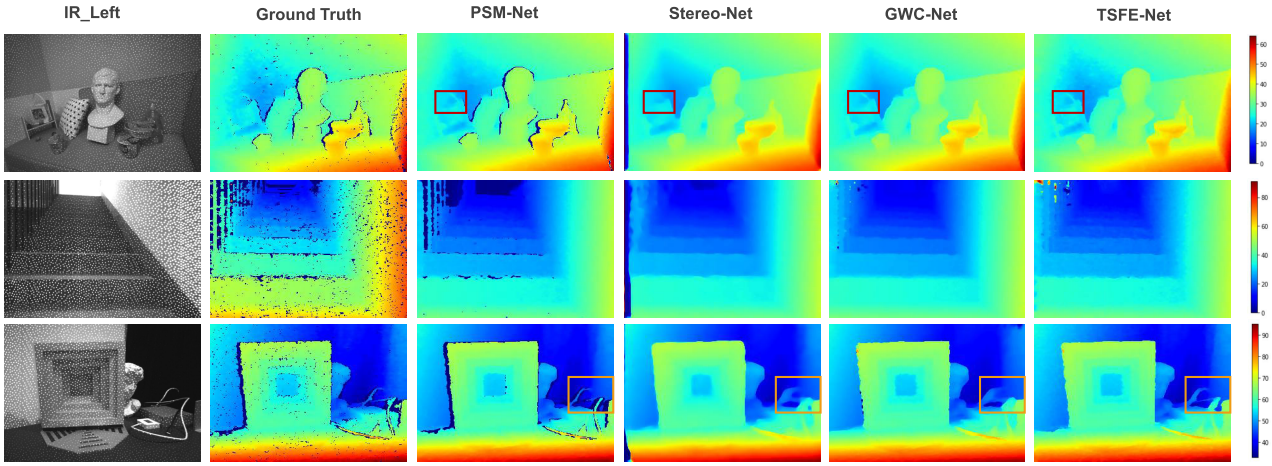


FIGURE 6. Experimental results on different models. The first column shows the left input image of stereo image pair. From second column to sixth column: the disparity maps obtained by RealSense D435 camera, PSM-Net, Stereo-Net, GWC-Net and our TSFE-Net.

pixels with disparity error larger than 3 (Thres3). We also record GPU memory and run-time for each model.

B. QUALITATIVE RESULTS

In this section, we compare our TSFE-Net with other state-of-the-art networks. In experiment, the Stereo-Net code downloads from [34] and we use 8X multi-model in experiment. As is shown in Table 1, our TSFE-Net runs at a faster speed and achieves better accuracy than PSM-Net. Although Stereo-Net is faster than our TSFE-Net in speed, TSFE-Net outperforms it by 1.885 pixel in average EPE. The group number is set to 40 for GWC-Net. Our model accuracy is close to GWC-Net, however, the run-time drops from 0.452 to 0.181 seconds. Figure 6 illustrates some examples of the disparity maps estimated by PSM-Net, Stereo-Net, and GWC-Net. Columns from left to right: left image of input, disparity maps of RealSenseD435 camera's output, PSM-Net [27]'s output, Stereo-Net [28]'s output, GWC-Net [29]'s output, and our network's output. Our TSFE-Net yields more robust results, particularly in thin texture regions. As is indicated in the red rectangle in Figure 6, TSFE-Net reconstructs clear edge of bag. Besides, it has more accurate reconstruction for wire and other models have the conglutination phenomenon in the orange rectangle.

In Figure 7, we carried out experimental analysis of the low texture and plain colors region. We use standard measuring instrument with eight squares in the laboratory. The color of each square is same but varies in depth by 1mm. The experimental results show that our model is more hierarchical and smooth. Camera's output has many invalid value. The output of PSM-Net has a lot of mutation values. The result of GWC-Net is uneven on the surface of these squares.

As shown in Table 2, we study the effects of our feature extraction layer with convolution and deconvolution by comparing with Cascade-Net(our baseline model) based on feature extraction layer of PSM-Net. From the first to the second stage for Cascade-Net, the number of disparity hypothesis is 48, 24, and the corresponding disparity interval is set to

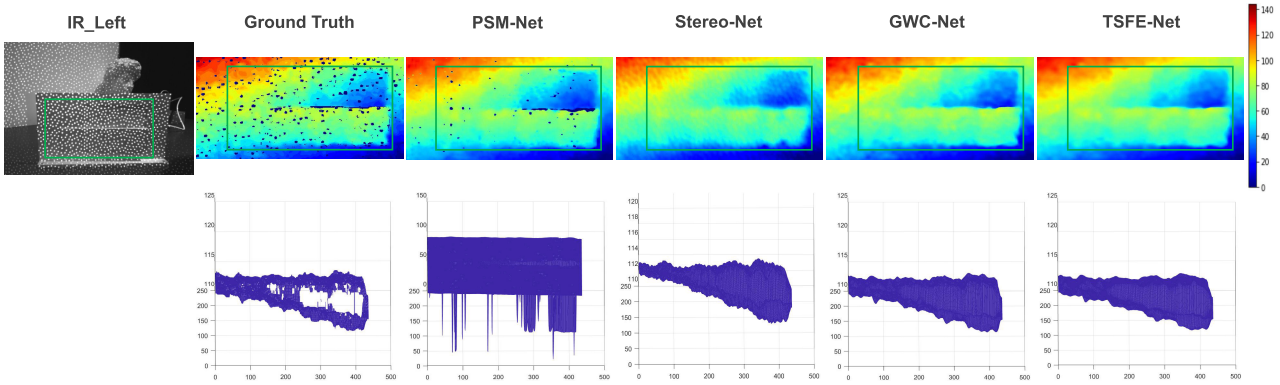


FIGURE 7. The experimental results of standard measurement instrument on different models.

TABLE 2. Comparison of our feature extraction layer and Cascade-Net based on PSM-Net feature extraction layer. Our feature extraction layer is implemented based on convolution and deconvolution. Accuracy and speed metrics are used for evaluations on RealSense D435 dataset.

Model	EPE(px)	D1(%)	Thres1(%)	Thres2(%)	Thres3(%)	Memory(MiB)	Time(s)
Cascade-Net [23]	0.439	1.62	4.33	2.29	1.71	4603	0.205
1layer_raw	**0.352**	**1.40**	**3.50**	**2.10**	**1.50**	**4017**	**0.149**

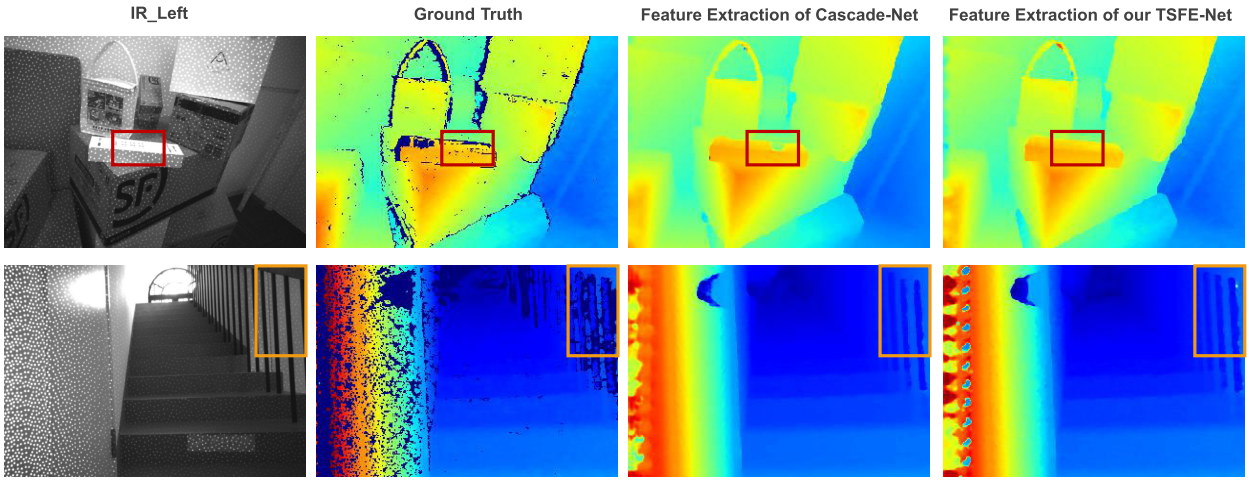


FIGURE 8. Comparison between two different feature extraction layers. From left to right: left image of input, disparity maps of RealSenseD435 camera's output, Cascade-Net's output, and our network's output.

3 and 1 due to the depth field of RealSense D435 camera. The other settings of the experiment are same except for the feature extraction layer. And the spatial resolution of feature maps is set to 1/4, 1/2 and gradually increases to 1 of original input size. For these two networks, we only input original image(1layer_raw). The experimental results show the network based on our feature extraction layer improves the model's accuracy by 0.09 pixel in average EPE and model's speed by 0.06 seconds in time. In Figure 8, columns from left to right: left image of input, disparity maps of RealSenseD435 camera's output, Cascade-Net [23]'s output based on PSM-Net feature extraction, and ours output. The visualization results indicate our feature extraction layer outperforms Cascade-Net especially in thin texture region. For example, our output has continuous edge of box on the top row, however camera's and Cascade-Net's edge are irregular. On the bottom row, our output is closer to the real input in the edge of stairs. And Cascade-Net's top edge losses details of stairs.

C. QUANTITATIVE RESULTS

We evaluate the model on a larger number of test images to prove the effectiveness of the proposed method. Test scenarios include bedrooms, staircases, standard measurement props, and so on. Our method achieves 0.335 pixel in average EPE. It is worth noticing that, in the bedroom scene, our towelling is accurately predicted. In the fourth column, we also correctly output the stair rails and our prediction result is better than RealSense D435 camera. More visualization results are shown in Figure 9. Test scenarios include

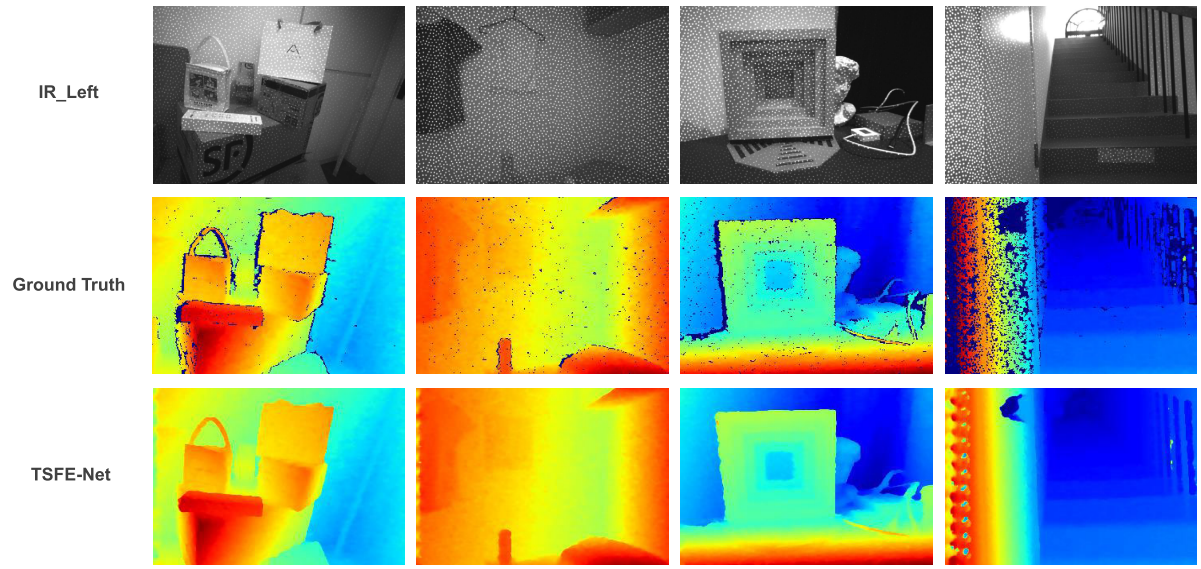


FIGURE 9. Disparity visualization results on the test dataset. From top to bottom: left image of input, disparity maps of RealSenseD435 camera's output, disparity maps of our TSFE-Net's output.

TABLE 3. Comparison of ablation experiment results.

Model	EPE(px)	D1(%)	Thres1(%)	Thres2(%)	Thres3(%)
1layer_raw	0.352	**1.4**	3.5	2.1	**1.5**
1layer_LCN	0.403	1.7	4.0	2.3	1.8
1layer_{raw,LCN}	0.409	1.9	3.5	2.3	2.0
TSFE-Net	**0.335**	**1.4**	**3.3**	**2.0**	**1.5**

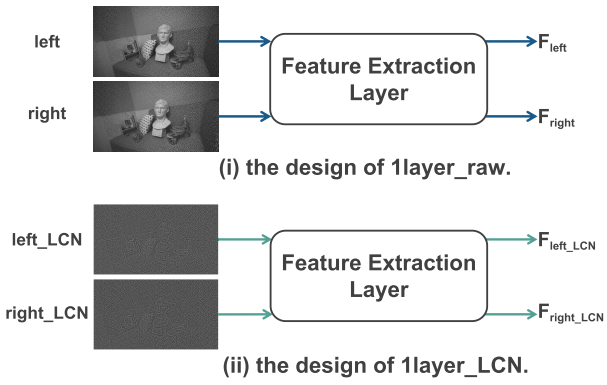


FIGURE 10. The design of 1layer_raw and 1layer_LCN.

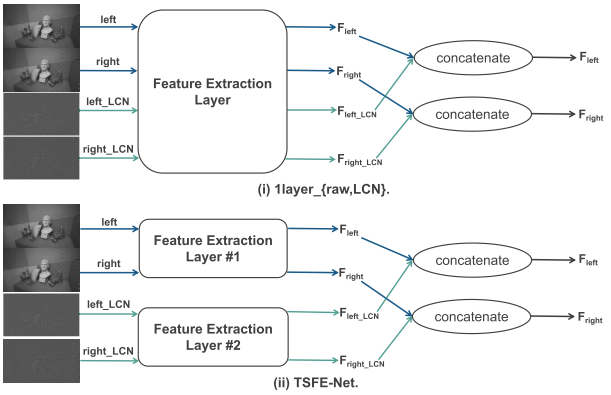


FIGURE 11. The design of 1layer_{raw,LCN} and TSFE-Net.

office, bedrooms, standard measurement props, and staircases. From top to bottom, input left images, disparity maps of RealSense D435 camera's output, and TSFE-Net' output.

D. ABLATION EXPERIMENTS

In order to analyze the effect of the two-stream feature extraction layer designed in our network, we conduct experiments with different input and procession. Our ablation experiments include: (1) feeding the original image into the network with only one feature extraction layer(1layer_raw), as shown in the Figure 10(i); (2) feeding LCN image after local contrast normalization process into the network with only one feature extraction layer(1layer_LCN), as shown in the Figure 10(ii); (3) feeding original image and LCN image into the network using one and the same feature extraction(1layer_{raw,LCN}), as shown in the Figure 11(i); (4) feeding original image and LCN image into different two-stream feature extraction layers(TSFE-Net), as shown in the Figure 11(ii). As is shown in Table 3, the evaluation error of TSFE-Net is minimal. The results show that the design of two-stream feature extraction layer is effective and accurate.

V. CONCLUSION

In this work, we present two-stream feature extraction networks(TSFE-Net), the depth estimation method based on deep learning for active stereo systems. Firstly, we use

two stream feature extraction layers to learn information of original image and LCN image which solve illumination effects and reserve original details. Secondly, we improve the Cascade-Net's feature extraction layer to reduce computation. Experimental results show that our network has better accuracy and faster speed based on convolutional layer and the deconvolutional layer than Cascade-Net. Thirdly, we construct a supervised network for dataset captured by RealSense D435 camera. Our experiments show that our TSFE-Net can predict depth with a subpixel precision and has real-time output with a runtime cost of 6 frame using a NVIDIA TITAN Xp for 704*1024 input image. In the following research, we will continue to explore new methods to improve the edge and accuracy of the output.

REFERENCES

[1] S. O. Escolano, C. Rhemann, S. R. Fanello, D. Kim, and S. Izadi, "Holoportation: Virtual 3D teleportation in real-time," in *Proc. 29th Annu. Symp. User Interface Softw. Technol.*, Oct. 2016, pp. 741–754.

[2] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "High-quality scanning using time-of-flight depth superresolution," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2008, pp. 1–7.

[3] M. A. A. Neil, R. Juskaitis, and T. Wilson, "Method of obtaining optical sectioning by using structured light in a conventional microscope," *Opt. Lett.*, vol. 22, no. 24, p. 1905, 1997.

[4] K.-J. Yoon and I.-S. Kweon, "Locally adaptive support-weight approach for visual correspondence search," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 924–931.

[5] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," in *Proc. CVPR*, Jun. 2011, pp. 504–511.

[6] 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.

[7] J. Lu, H. Yang, D. Min, and M. N. Do, "Patch match filter: Efficient edge-aware filtering meets randomized search for fast correspondence field estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1854–1861.

[8] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.

[9] M. Bleyer and M. Gelautz, "Simple but effective tree structures for dynamic programming-based stereo matching," in *Proc. 3rd Int. Conf. Comput. Vis. Theory Appl.*, 2008, pp. 415–422.

[10] 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)*, Jun. 2005, pp. 807–814.

[11] 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.

[12] F. Besse, C. Rother, A. Fitzgibbon, and J. Kautz, "PMBP: PatchMatch belief propagation for correspondence field estimation," *Int. J. Comput. Vis.*, vol. 110, no. 1, pp. 2–13, Oct. 2014.

[13] Y. Zhang and T. Chen, "Efficient inference for fully-connected CRFs with stationarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 582–589.

[14] Y. Li, D. Min, M. S. Brown, M. N. Do, and J. Lu, "SPM-BP: Sped-up PatchMatch belief propagation for continuous MRFs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4006–4014.

[15] V. E. Seferidis, "General approach to block-matching motion estimation," *Opt. Eng.*, vol. 32, no. 7, pp. 1464–1474, 1993.

[16] H. Sakoe, *Dynamic Programming Algorithm Optimization for Spoken Word Recognition*. San Mateo, CA, USA: Morgan Kaufmann, 1990.

[17] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Generalized belief propagation," in *Proc. Adv. Neural Inf. Process. Syst.*, Denver, CO, USA, vol. 13, 2001, pp. 689–695.

[18] 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.

[19] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2462–2470.

[20] 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.

[21] J. Pang, W. Sun, J. S. Ren, C. Yang, and Q. Yan, "Cascade residual learning: A two-stage convolutional neural network for stereo matching," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 887–895.

[22] Y. Zhang, S. Khamis, C. Rhemann, J. Valentin, A. Kowdle, V. Tankovich, M. Schoenberg, S. Izadi, T. Funkhouser, and S. Fanello, "Activestereonet: End-to-end self-supervised learning for active stereo systems," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 784–801.

[23] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2495–2504.

[24] N. Xu, N. Ahuja, and R. Bansal, "Object segmentation using graph cuts based active contours," *Comput. Vis. Image Understand.*, vol. 107, no. 3, pp. 210–224, Sep. 2007.

[25] 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)*, 2001, pp. 7–42.

[26] 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.

[27] 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.

[28] 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. (ECCV)*, 2018, pp. 573–590.

[29] 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.

[30] S. R. Fanello, J. Valentin, C. Rhemann, A. Kowdle, V. Tankovich, P. Davidson, and S. Izadi, "UltraStereo: Efficient learning-based matching for active stereo systems," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6535–6544.

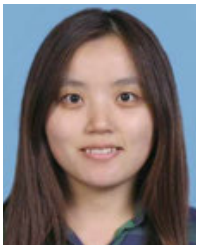
[31] Q. Du, R. Liu, B. Guan, Y. Pan, and S. Sun, "Stereo-matching network for structured light," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 164–168, Jan. 2019.

[32] F. Li, Q. Li, T. Zhang, Y. Niu, and G. Shi, "Depth acquisition with the combination of structured light and deep learning stereo matching," *Signal Process., Image Commun.*, vol. 75, pp. 111–117, Jul. 2019.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[34] X. Li. (2018). *Stereonet: Guided Hierarchical Refinement for Real-Time Edge-Aware Depth Prediction*. [Online]. Available: https://github.com/meteorshowers/StereoNet-ActiveStereoNet

HAOJIE ZENG received the B.S. degree from the Department of Physical Science and Information Technology, Liaocheng University, China, in 2014. She is currently a Graduate Student with Shanghai Normal University, China. Her research interest includes deep reconstruction, including passive stereo matching and active stereo matching.

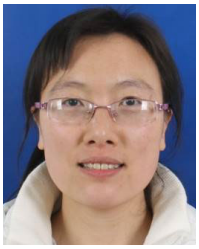


BIN WANG received the Ph.D. degree from the Department of Automation, Shanghai Jiao Tong University, China, in 2014. She is currently an Associate Professor with Shanghai Normal University, China. Her research interests include computer vision, machine learning, image processing, and urban computing.

XIAOPING ZHOU received the Ph.D. degree in information and communication engineering from the University of Shanghai, Shanghai, China, in 2011. From 2011 to 2013, he was a Postdoctoral Fellow with the Communication Laboratory, Shanghai Jiao Tong University, China. He is currently a Full Professor with Shanghai Normal University, Shanghai. His current research interests include mobile communication systems, image processing, parameter estimation, and electrostatic discharge.

XIAOJING SUN is currently a Graduate Student with Shanghai Normal University, China. Her research interest includes computer vision.

LONGXIANG HUANG received the Ph.D. degree from the Department of Automation, Shanghai Jiao Tong University, China, in 2017. He joined Shenzhen Guangjian Technology Company Ltd., as an Algorithm Leader, in April 2018. His research interests include 3D reconstruction and robot navigation.



QIAN ZHANG received the Ph.D. degree from Shanghai University, China. She is currently an Associate Professor with Shanghai Normal University, China. Her research interest includes video processing.

YANG WANG received the Ph.D. degree from the Chinese Academy of Sciences (CAS). He joined the College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, as an Assistant Professor, in 2017. He served as a Technical Advisor in public data security for the government. He has published several research articles in reputed journals. His current research interests include big data, compressive sensing, next-generation optical processors, electric system modeling, and performance analysis.