A Depth Estimation Framework Based on Unsupervised Learning and Cross-Modal Translation

Jiafeng Shen, Kaiwei Wang*, Kailun Yang, Kaite Xiang, Lei Fei, Xinxin Hu, Huabing Li, Hao Chen
State Key Laboratory of Modern Optical Instrumentation, Zhejiang University, 38# Zheda Road,
Hangzhou 310027, China

ABSTRACT

In recent years, with the vigorous development of artificial intelligence and autonomous driving technology, the importance of scene perception technology is increasing. Unsupervised deep learning based methods have demonstrated a certain level of robustness and accuracy in some challenging scenes. By inferring depth from a single input image without any ground truth label, a lot of time and resources can be saved. However, unsupervised depth estimation has defects in robustness and accuracy under complex environment which could be improved by modifying network structure and incorporating other modal information. In this paper, we propose an unsupervised, monocular depth estimation network achieving high speed and accuracy, and a learning framework with our depth estimation network to improve depth performance by incorporating transformed images across different modalities. The depth estimator is an encoder-decoder network to generate the multi-scale dense depth map. The sub-pixel convolutional layer is adopted to obtain depth super-resolution by replacing the up-sample branches. The cross-modal depth estimation using near-infrared image and RGB image enhances the performance of depth estimation than pure RGB image. The training mode is to transfer both images to the same modality and then carry out super-resolved depth estimation for each stereo camera pair. Compared with the initial results of depth estimation using only RGB images, the experiment verifies that our depth estimation network with the cross-modal fusion system designed in this paper achieves better performance on public datasets and a multi-modal dataset collected by our stereo vision sensor.

Keywords: Unsupervised learning, monocular, depth estimation, cross-modal translation

1. INTRODUCTION

Two-dimensional image depth estimation, one of the basic problems of computer vision, has been studied for a long time. It has received extensive attention in the past, and has been applied in the fields of robotics, self-driving car, scene understanding and three-dimensional reconstruction. These applications usually use multiple-angle imaging of the same scene for depth estimation, such as stereo image pair, movement of single camera, and continuous frames for depth estimation. Although the depth estimation of many images has made great progress, monocular depth estimation that has the nature of low costs and few constraints, still needs to be improved in terms of robustness against real-world scenarios.

Monocular depth estimation technology, is a kind of method to predict the depth value of each pixel in an RGB image. It is very important to understand the scene geometry, but to give an accurate estimate of the depth of the image, is very challenging. This is because that it is an ill-posed problem due to the inherent ambiguity. Saxena et al.[3] first introduced to apply deep learning for monocular depth estimation, namely learning from the RGB image domain to the depth domain. Under this background, the multi-scale Convolutional Neural Networks (CNNs) have been proved effective to estimate the depth map. In particular, Godard et al.[7] proposed the use of the single-generation network to estimate disparity maps, and used the consistency constraint model with the learning process.

* Corresponding author. Email: wangkaiwei@zju.edu.cn.

In recent years, depth estimation approaches based on CNNs usually use an encoder-decoder architecture to predict dense depth maps. In these networks, repeated convolutions and pooling reduce the excessive output spatial resolution, which may be the bottleneck of high-resolution prediction. Therefore, many technologies, such as skip connections, multi-scale networks, multi-layer deconvolution networks, and super-resolution techniques, are used to synthesize higher-resolution depth maps.

In this paper, a method is adopted to take depth estimation as an image reconstruction problem in the training process. Moreover, our convolutional network does not need any ground truth value, but takes depth as an intermediate value, so as to learn to predict and correct the pixel level correspondence between image pairs. Network structure of this paper is based on the encoder-decoder structure. For the encoder stage, it can be regarded as two structures: one is the base dense feature extractor, while the other is a contextual information extractor. In the decoder stage, we introduce the decoder based on super-resolution residual pyramid and multi-scale depth map prediction. Our method preserves the hierarchical structure of the scene and can obtain accurate depth estimations for both large-scale scenes and small objects. In addition, our network achieves up to 30FPS on a standard GPU processor.

This paper also proposes a cross-modal translation framework to estimate depth maps by using other spectral information. We estimate the depth by using RGB images and Near Infra-Red (NIR) images, and prove that the modal translation is helpful for depth estimation under the condition of unsatisfactory illumination.

Our contributions are summarized as follows:

(1) We propose an unsupervised depth estimation network which achieves state-of-the-art performance evidenced by both accuracy and speed analysis.

(2) A cross-modal translation framework is proposed to estimate depth combining multiple spectra of information.

(3) The proposed method achieves state-of-the-art performance on the challenging KITTI dataset[9] and PittsStereo-RGBNIR dataset[17]. More importantly, the visual quality of recovered depth maps has been significantly improved, proving that the usage of other spectra of images is meaningful.

2. RELATED WORK

In recent years, there are many researches that focused on depth estimation from images or a single image[13]. CNNs have become the most successful techniques for depth estimation since neural network technology was widely used in visual tasks. State-of-the-arts approaches in leveraging both data and structural constraints mostly differ in the type of data and supervision used.

Learning based stereo: Most stereo matching estimation algorithms[2,21] calculate the similarity between each pixel in the first image and that in the second image. It is typically required to match stereo image pairs[4] online, and it is common to transform the problem of visual estimation into a one-dimensional search problem for each pixel. With the development of neural networks, some studies have shown that better stereo matching results can be achieved by transforming matching into supervised learning problems instead of using traditional hand-defined similarity functions. Mayer et al.[1] introduced a fully convolutional deep network called DispNet to learn the matching function. It utilizes convolution to calculate the corresponding relationship between the two images in training, and predicts the disparity for every pixel by minimizing a regression training loss.

Supervised Monocular Depth Estimation: Monocular depth estimation means that only one image is used as input to output the completed depth map during the test. Saxena et al.[3] proposed a monocular depth technology, introducing a model called Make3D to estimate the 3D position and direction of objects. Eigen et al.[25,26] showed it is possible to produce pixel depth estimation using a two-scale deep network trained on raw pixel images and their corresponding depth values by minimizing a scale invariant loss. Since then, a lot of methods based on the technology have been built, such as the CRFs[22] to improve accuracy, more appropriate usage of loss function based on the network structure and improvement of the loss function. But these monocular supervised depth estimation methods still rely on training pairs with high quality, pixel alignment, and true ground depth as well. One way to avoid capturing the ground truth of the depth, is to use composite data from the simulator and replace the data collection/annotation issues with domain adaptation and virtual scene creation issues. We also perform monocular depth estimation, but merely leverage binocular color images to train our network with no need for the ground-truth depth.

Unsupervised Monocular Depth Estimation: Unsupervised depth estimation[15] does not require any labels. It is expensive to obtain a large number of ground truth depths or disparity maps, and usually requires a data collection platform similar to the target task. Unsupervised training method has been proved to be a promising direction recently to circumvent this limitation. The unsupervised training method is usually focused on the training method design, such as synthesizing disparity map figure as a proxy task. The relevant work is to explore different architectures generally, such as using the shared network to simultaneously perform depth estimation and semantic segmentation[18,20,30,34]. In contrast, Garg et al.[24] put forward a kind of classical unsupervised depth estimation method based on image synthesis. Godard et al.[7] put forward that images of different view can be forward and reverse remodeled. There is also some work under the condition without any label, but jointly learning the depth and ego-motion from monocular videos[16,19]. Some adversarial learning strategies[27] are also employed to improve the quality by using image generation technology[28] to obtain better depth estimation.

Cross-Modal translation: Cross-modal translation, also known as cross-spectral translation[5,12,31,33], is simply to use different spectral information[32] to supplement the current spectrum, which has a wide range of applications in computer vision and image processing. Infrared, near-infrared, polarized images and other auxiliary RGB images are commonly used for face recognition, unmanned driving assistance, etc. RGB-NIR image pairs are usually used for shadow detection, scene recognition and scene analysis. Near-infrared images are also favorable to enhance color images and remove fog. A current trend is the application of multi-camera and multi-spectrum, and cross-spectrum is becoming more and more important. For example, Zhi et al.[17] utilize stereo matching and disparity estimation of corrected spectral pairs to complete vision tasks such as detection and tracking. In this paper, we employ modal translation network on NIR and RGB image pair, utilize the structural consistency of NIR images to depth estimation, which can be regarded as transferring unique features of NIR to RGB images to help monocular depth estimation which is demonstrated by our experiments.

3. METHODOLOGY

This section describes our monocular image Depth Prediction Network (DPN) and Cross-modal Translation Network (CTN). Our depth estimation network consists of two parts: an encoder for multi-scale feature extraction and a residual pyramid decoder. CTN is a generative adversarial network for content translation.

3.1 Monocular Depth Network Architecture:

Our depth estimation network based on the current popular DispNet framework[1,10]. Following the works in this line, we make similar changes to the encoder-decoder network, add skip connections between activation blocks of multi-scale extractor to save context information, and add a special module to expand the receptive field and save more details. Then the the decoder is composed of several projection recovery units, gradually improve the resolution of the characteristics. In order to improve the detail-capability of the network, we connect the bottom feature of the encoder with the top projection feature of the same resolution. The architecture is shown in Figure 1.

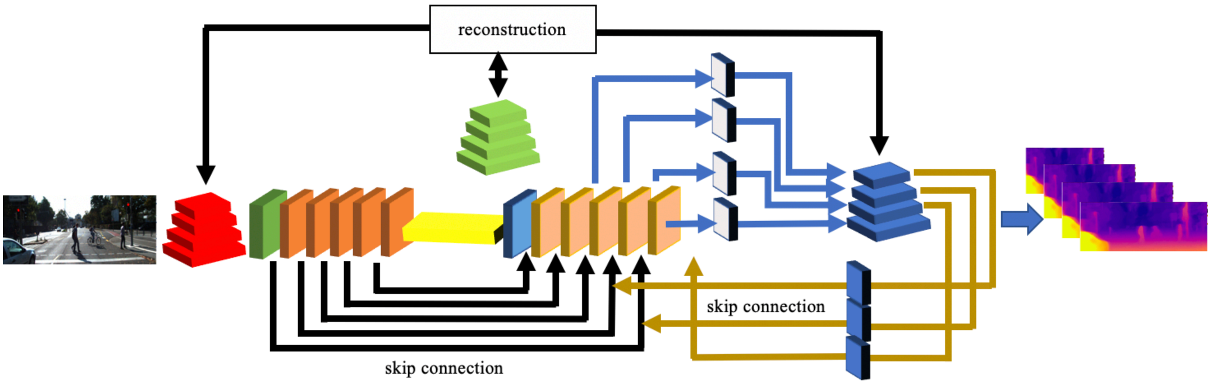

Figure 1. Monocular Depth Network Architecture.

3.2 Multi-scale feature extractor

Inspired by Godard et al.[7], given an input image, we use scale-pyramid technology to process the input image. We resize left and right images to a smaller size, to 1/2,1/4,1/8 of the original image, but we only input the left image into the feature extraction network, and the right image pyramid is used for image reconstruction. We can extract the large-scale structure and local details at the same time. We extract global features by using a convolution with a stride of 7, then generate multi-scale feature by using the down-sample group, where the down-sample group is composed by multiple Feature Residual Group (FRG) as seen in Figure 2. The size of feature map reduce double after a FRG[8]. With the feature extractor network keep deeper, the size of the feature map keeps smaller. After that, the size of the feature map reaches to 1/32 of the input image. Finally, a Spatial and Channel Attention Module[29] is used to enlarge the receptive field and the boost representation power.

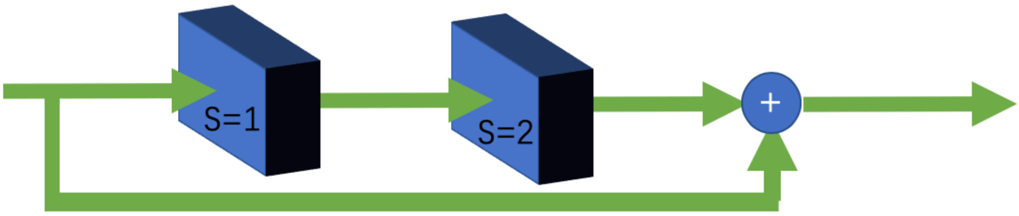

Figure 2. Feature Residual Group.

Our encoder body is a feature extraction module similar to the structure of ResNet18[14]. The detailed description of FRG is shown in Figure 3, describe it as feature residual block (FRB), S is stride of continuous convolution, but it replaces the two convolutions of 3×3 with two 1-dimensional convolution, 1×3 and 3×1, on the basic block, which not only reduces many training parameters and increases the speed of the forward pass, but also facilitates feature extraction in the direction of gradient to extract disparity. Another difference is that in order to save more detail features, we use skip-connections on the two blocks, or add a convolution, batch normalization and activation function Relu to the network in our experiment.

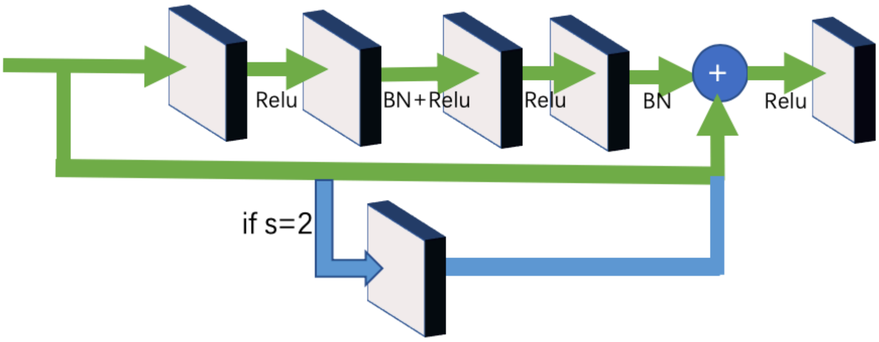

Figure 1. Feature Residual Block.

3.3 Residual Pyramid Decoder and Reconstruction

Inspired by single image super resolution, our encoder uses sub-pixel convolution layer[11] to recover resolution, which replaced the conventional bilinear upsampling. But the sub-pixel block has not been used in the beginning. The input map first passes through two up-conv blocks, where the up-conv is composed by a nearest neighbor and a convolution. Then, the size of feature map has been reduced to 1/8 of the original image, while the channel number increased to 8 times of the original image. Because we found that when the feature map is too small, the information of the feature map is concentrated on the channel, the sub-pixel representation is not a very lightweight way, which will cause the loss of edge information and the burden of parameter for channel number grows quickly. Our up-sampling scheme is to restore the image to the original resolution after 2 up-conv blocks and 3 sub-pixel blocks. Finally, we recover the multi-scale disparity pyramid generated by the disparity generation convolution block, which contains left and right disparity, so we can reconstruct the left and right view image with them. Through image reconstruction technology[6], we can compare the similarity with the pyramid of the right view. Thereby, we can obtain the more accurate disparity estimation through this cycle training.

3.4 Depth Estimation Network Training Loss

In order to train our depth estimation architecture, we define the total loss as a sum of three parts, and we define a loss C(s) at each output of deferent size. The total loss C are the sum of the C(s)(s=1,2,3,4).

$$\mathrm{C}=\sum_{s=1}^{4} C(s) \tag{1}$$

$$\mathrm{C(s)}= L_{ap}(l,r) + L_{ds}(l,r) + L_{lr}(l,r) \tag{2}$$

where L_{ap} is an image reconstruction loss, L_{ap} is a disparity smoothness loss and L_{ap} is a left-right disparity consistency loss. Each loss contains two parts, both the left and right components. Next, we present each component of our loss in detail.

Image reconstruction loss: This loss encourages the reconstructed image to appear similar to the training input. A combination of a L1 loss and structural similarity measure (SSIM) term as our photometric image reconstruction cost L_{ap}, which encodes the quality of the reconstructed image $\widetilde{I}$ with the input image *I*, where N is the numbers of pixels:

$$L_{ap} = \frac{1}{N}\sum_{i,j} \alpha \, \frac{1-SSIM\left(I_{i,j},\widetilde{I}_{i,j}\right)}{2} + (1-\alpha)\left|I_{i,j} - \widetilde{I}_{i,j}\right| \tag{3}$$

Following previous research, we use a SSIM with 3×3 block filters and set $\alpha = 0.85$.

Disparity smoothness loss: This loss enforces smooth disparities, encourages the depth discontinuities to be locally smooth with an L1 penalty on the disparity gradients, and the gradients are weighted by an edge-aware term from the image domain:

$$L_{ds} = \frac{1}{N}\sum_{i,j}\left|\partial_x d_{ij}\right| e^{-\left\|\partial_x I_{ij}\right\|} + \left|\partial_y d_{ij}\right| e^{-\left\|\partial_y I_{ij}\right\|} \tag{4}$$

Left-right disparity consistency: This loss encourages the predicted left and right disparities to be consistent for producing more accurate disparity maps. In our training, we only input left image and output both the left and right disparity images, which inevitably leads to inconsistent disparities. Accordingly, we introduce an LI left-right disparity consistency penalty to overcome it. Like all the other loss, this loss is evaluated at all the output scale for left and right. For example, L_{lr}^{l} is depicted in the following equation:

$$L_{lr}^{l} = \frac{1}{N}\sum_{i,j} \left|d_{ij}^{l} - d_{ij+d_{ij}^{l}}^{r}\right| \tag{5}$$

3.5 Cross-modal Translation

Cross-modal translation is used when our RGB image is not ideal due to the external environment, and other spectral image cannot be directly applied to our depth estimation network. Inspired by current style transfer network, we choose Generative Adversarial Network (GAN) to achieve the transformation of local content and style of multi spectra. We use the framework is CycleGAN network, a kind of mirror symmetry GAN network that is widely used to implement domain adaptation, by reducing the difference of the image of two different scenarios. The CycleGAN network needs two datasets to train which do not need to be aligned, reducing the burden of data collection. The image after style transfer can be used to in our depth estimation network.

4. EXPERIMENTAL RESULTS

In this section, we describe the dataset, the implementation details and evaluations of the network on various training/testing configurations. Our method achieves the state-of-the-art results in terms of the speed and accuracy trade-off. As the standard in this field, we assess the performance of our depth estimation techniques following the protocol by Eigen et al.[25,26]. Data extracted from the KITTI dataset and the stereo images are trained, without any supervision. Sparse LiDAR measurement data are adopted as ground truth values to evaluate the performance of monocular depth estimation technology. Moreover, the depth estimation of pseudo RGB images after content migration is carried out. The experiment shows that our framework has achieved favorable results.

4.1 Dataset

KITTI: we train and evaluate our network on the KITTI dataset[9], which is composed of two HD cameras and Teledyne 64-line LiDARs. It includes semi-dense annotations, RGB images, and sparse depth maps. We use the training protocols proposed by Eigen et al.[25,26], and compare the result with previous methods on the standard KITTI disparity estimation benchmark. To be specific, we leverage the KITTI Eigen split that contains 22600 training, 888 validation, and 697 test stereo image pairs, whose main image size is 1242×375 pixels, we use the metrics described by Eigen to evaluate the estimated result, where metrics include: Abs Rel, Sq Rel, RMSE and RMSE log.

Cityscapes: Cityscapes dataset[23] consists of stereo pairs of about 50 cities in Germany, which are taken from vehicles traveling under different weather conditions. It consists of 22,973 stereo pairs with a size of 2048×1024 pixels. It is mainly used for transfer training, not for validation and testing.

PittsStereo-RGBNIR Dataset: The dataset is captured by 56mm baseline RGB camera and near infrared cameras installed on the vehicle, alternating among short, middle and long exposures. Only the middle exposures are utilized in our experiments, which collected nearly 1 million stereo images for 1164 ×858 pixels. We take a dataset subset including 2000 pairs, 2500 for training, and 500 for test, covering sunlight, cloudy and dark conditions of campus road, highway, housing, etc. It does not provide depth truth, so we can only make a qualitative statement. The example of the dataset is shown in Figure 4. The right image is RGB image, and the left image is NIR image. We can observe that the reflective part in RGB image, a big difficulty for depth estimation which could not be observed in NIR image obviously.

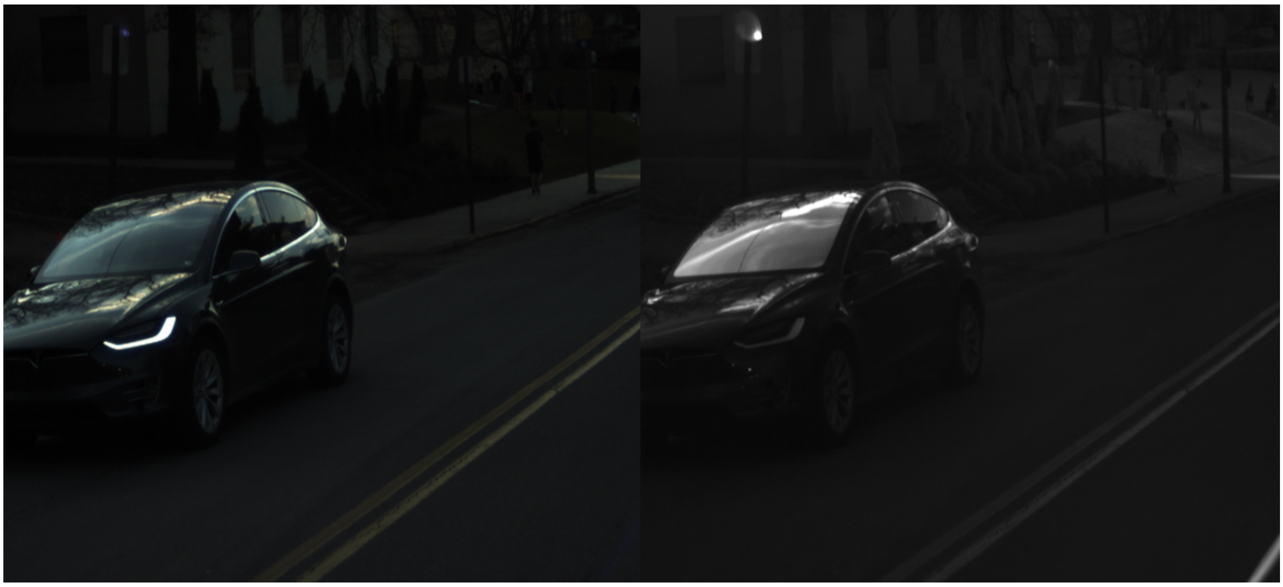

Figure 4. An image pair of PittsStereo-RGBNIR Dataset. The left image is RGB image, the right image is NIR image.

4.2 Implementation details

The network is implemented in TensorFlow. When training monocular estimation network, we used 30 hours to train 23000 pairs of images for 200 epochs on a 2080Ti GPU. For 512×256 images, the inference speed is very fast, above 30 frames per second, including the transmission times to and from the GPU. We used Adam optimizer algorithm to train the network setting the following parameter values for all training runs: Lambda 1 = 0.1, lambda 2 = 0.01, alpha = 0.85. The learning rate and batch size (4 or 8) are estimated via hyper-parameter search. The initial learning rate is 5e-4. We augment images before inputting to the network using the random horizontal flipping, random rotation, the random contrast and brightness with 50% of chance, and train our network on a random crop into resolution about 512×256 for KITTI and Cityscapes datasets. We employ a 2080TI GPU to train translation network on 2500 image pairs for 20 hours.

Post-processing: Post-processing step for output is necessary to reduce the effects of stereo occlusions on disparity ramps produced by the left side of the image and of the occluders. In the test time, there are two disparity maps, one is d generated by input image l, the other is the d'generated by l', the horizontally flipped image of l . We combine two disparity maps to output the final result by assigning the first 5% on the left of image d, and the last 5% on the right to disparities image d', and the central pixel of the final result is the average of two disparity maps. It achieves a higher precision, but doubles the amount of time. We indicate such results using **pp** in result tables.

4.3 Performance comparisons

Quantitative Evaluation: Following previous study, we used the four indexes, including Abs Rel, Sq Rel, RMSE, RMSE log and accuracy with three thresholds, to quantitatively assess the depth of our estimated performance. Table 1 shows the results of our depth estimation network and recent approaches.

Table 1. Results on KITTI 2015 which using the split of Eigen et al. K is the KITTI dataset; CS is the Cityscapes. The results of Liu et al.[35] were a mixture of left and right disparity images, rather than relying solely on the left monocular image. All the other results were taken from their papers. The red metric: lower is better; the green metric: higher is better.

Method	**Supervised**	**Dataset**	**Abs Rel**	**Sq Rel**	**RMSE**	**RMSE log**	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Saxena et al.	Yes	K	0.280	-	8.734	-	0.601	0.820	0.916
Eigen et al.	Yes	K	0.214	1.605	6.563	0.292	0.673	0.884	0.957
Liu et al.	Yes	K	0.201	1.584	6.471	0.273	0.680	0.898	0.967
Train set mean	No	K	0.361	4.826	8.102	0.377	0.638	0.804	0.894
Zhou et al.	No	K	0.208	1.768	6.856	0.283	0.678	0.885	0.973
Godard et al.	No	K	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Ours	No	K	0.130	1.301	6.174	0.225	0.824	0.933	0.975
Ours+pp	No	K	0.125	1.172	5.89	0.216	0.859	**0.983**	**0.977**
Ours	No	CS+K	0.124	1.066	5.312	0.218	0.847	0.942	0.973
Ours+pp	No	CS+k	**0.114**	**0.988**	**4.945**	**0.206**	**0.861**	0.949	0.976

It can be seen from Table 1 that our estimation network achieves good results for both unsupervised and supervised monocular depth estimation, and proves that post-processing is helpful to improve the accuracy.

Qualitative Evaluation: We compare our method with other state-of-the-arts methods in Figure 5. It can be seen that our method in predicting the depth of the figure on the vision is better than other schemes. Small objects in the distance and near large objects can be accurately perceived. The preservation of the multiple object structure is excellent, such as the car in the distance in the first line, the completeness of the structure of the truck in the second line, and the aggregation of multiple traffic elements in line 5. In addition, our method has a good balance between speed and accuracy, and can be applied to devices which need higher real-time performance.

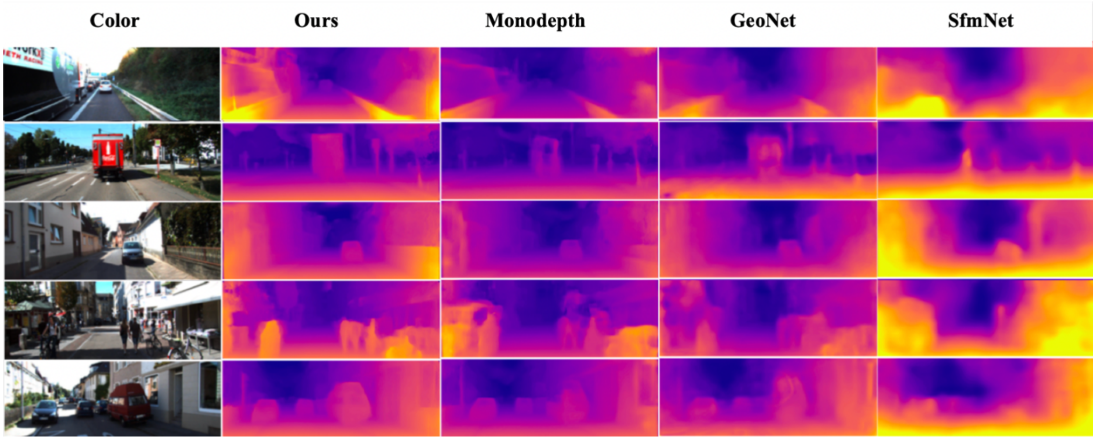

Figure 5. Some examples of the results and comparisons with others. From left to right: color test image, our depth output, Monodepth, GeoNet and SfmNet depth output.

4.4 Generalizing to other datasets

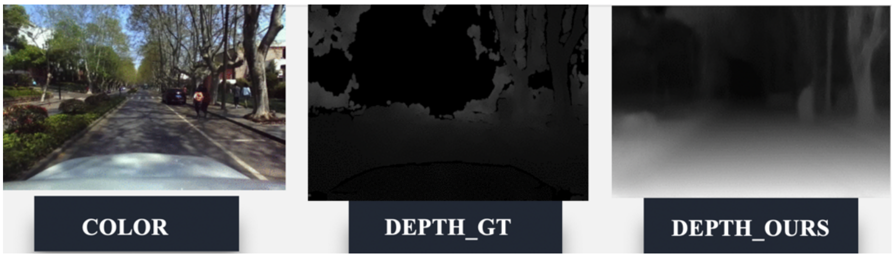

Figure 6. Comparison of visual results of depth estimation on the ZJU dataset. The DEPTH_GT is the ground truth depth map. The DEPTH_OURS is generated by our depth estimation network.

In this section, we further illustrate the generalization of our model to other datasets. We use the model trained on Cityscapes and KITTI, tested on the ZJU dataset. The ZJU dataset was collected by our laboratory, including left, right views and depth map. The depth map was obtained by binocular stereo matching. We collected about 5K pairs for depth estimation.

Table 2. We used our ground truth depth and the corresponding confidence map to evaluate the performance of the depth map calculated by our model, and used the same four indexes for evaluation. 0 means that we use a ground truth depth map, without excluding any information, and 90 means that we calculate the accuracy only for points whose confidence are above 90.

Confidence	Abs Rel	Sq Rel	RMSE	RMSE log
0	0.4078	7.0083	15.281	0.610
30	0.3793	6.3485	15.137	0.586
50	**0.2981**	**5.6736**	15.134	**0.416**
70	0.3043	5.7518	**14.892**	0.472
90	0.3139	5.9783	15.082	0.494

From Figure 8, the generalization of our depth estimate network is fine, but in the area of high confidence degree, the accuracy does not get a great progress compared with lower degree. We believe that our depth map is too smooth, because of the smooth of the loss function, and cannot be used for very high accuracy of scene.

4.5 Cross-modal image depth estimation

In Figure 8, the top left is NIR image, the bottom left is RGB image, and the right image is pseudo RGB image. It can be clearly found that there are obvious spectral differences between NIR and RGB images. The red box in the Figure 8 indicates that the structure of distant objects becomes very blurred in the RGB image when the light is blurry, but there is no such bad effect in the NIR image. The transferred image perfectly integrates these structure features and can be used for monocular depth estimation.

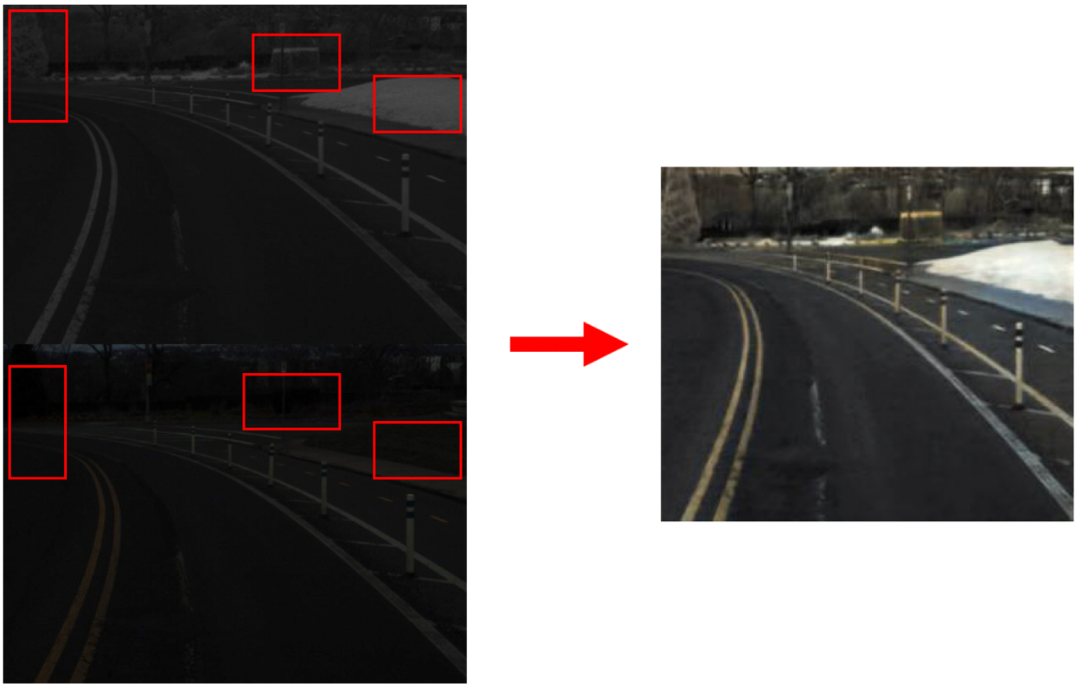

Figure 8. The performance of cross-modal translation. The top left image is a NIR image, the bottom left image is a RGB image, and the left image is the pseudo RGB image, which is generated from the NIR-to-RGB translation network.

In Figure 9, We can observe that the top right image of estimated by RGB directly, due to the reflection of car windows, the corresponding depth boundary is blurred, and depth of the structure of objects cannot be well estimated. The bottom right image is the depth estimation result conducted by the transferred NIR image, which obviously gets a better depth estimation at the original reflective place, and the boundary structure of the object is not lost. We believe that the depth estimation effect of the pseudo-color image and color image after content transfer is obviously better than that of the image after content transfer, indicating that our method is conducive to solving the problem of spectral limitation.

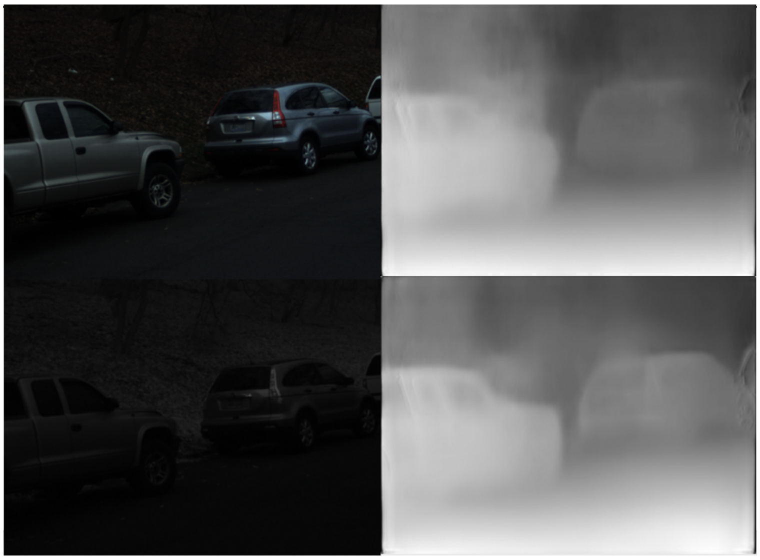

Figure 9. The comparison of the depth estimation. The top left image is RGB image; the bottom image is NIR image; the top right image is the depth map produce by RGB image; while the bottom right image is the depth map produce by the pseudo RGB image.

5. CONCLUSION

In this paper, we propose an unsupervised monocular depth estimation network with a cross-modal image translation framework. We use the accessible binocular images to estimate depth by indirectly obtaining disparity through image reconstruction technology. The network fully considered different hierarchies of global and detailed information, and our loss function improved the coherence of each predicted depth map from certain camera view, and achieved a more accurate result in KITTI dataset and more balance speed and accuracy than many fully supervised models. We also show that our model can be generalized to other datasets and obtain a robust depth map.

The cross-modal framework provides a way to transfer images from other spectra into images that our depth estimation network can forward propagation, while retaining unique features of the spectrum to help the depth estimation break through the limitations of spectrum. In the future work, we can further improve the accuracy and speed of unsupervised monocular depth estimation, and focus on applications to other deep learning task using multi spectra of information.

REFERENCES

[1] N. Mayer, E. Ilg, P. Ha ̈usser, 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 CVPR, 2016.

[2] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms.” 2002.

[3] A. Saxena, M. Sun, and A. Y. Ng, “Make3d: Learning 3d scene structure from a single still image,” PAMI, 2009.

[4] W. Luo, A. Schwing, and R. Urtasun, “Efficient deep learning for stereo matching, in CVPR,” 2016. 2.

[5] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros A.A, “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks,” in 2017 IEEE International Conference on Computer Vision (ICCV) 2017.

[6] Peng, X. Saenko, K, “Synthetic to real adaptation with generative correlation alignment networks,” 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018, pp. 1982–1991.

[7] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in CVPR, 2017.

[8] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al., “Tensorflow: Large-scale machine learning on heterogeneous distributed systems,” arXiv preprint arXiv:1603.04467,2016.

[9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, “Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR),” 2013.

[10] E. Shelhamer, J. Long, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in PAMI, 2016.

[11] W. Shi, J. Caballero, F. Huszar, and at al., “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in CVPR 2016, pages 1874–1883, 2016.

[12] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, Eli Shechtman, “Toward Multimodal Image-to-Image Translation,” in NIPS, 2017.

[13] W. Zhuo, M. Salzmann, X. He, and M. Liu, “Indoor scene structure analysis for single image depth estimation,” in CVPR, 2015.

[14] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, “Deeper depth prediction with fully convolutional residual networks,” in 3DV, 2016.

[15] J. N. Kundu, P. K. Uppala, A. Pahuja, and R. V. Babu, “Adadepth: Unsupervised content congruent adaptation for depth estimation,” in CVPR, 2018.

[16] R. Mahjourian, M. Wicke, and A. Angelova, “Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints,” in CVPR, 2018.

[17] Tiancheng Zhi, Bernardo R. Pires, Martial HebertandSrinivasa G, “Deep Material-aware Cross-spectral Stereo Matching,” in Narasimhan IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2018.

[18] M. Cordts, M. Omran, S. Ramos, T.Rehfeld, M.Enzweiler, R.Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in CVPR, 2016.

[19] Zhichao Yin and Jianping Shi, “Geonet: Unsupervised learning of dense depth, optical flow and camera pose,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[20] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia, “Segstereo: Exploiting semantic information for disparity estimation,” in 15th European Conference on Computer Vision (ECCV), 2018.

[21] Jure Zbontar and Yann LeCun, “Stereo matching by training a convolutional neural network to compare image patches,” Journal of Machine Learning Research, 17(1-32): 2, 2016.

[22] B.Li, C.Shen, Y.Dai, A.vandenHengel, and M.He, “Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs," in CVPR, 2015.

[23] M.Cordts, M.Omran, S.Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in CVPR, 2016.

[24] R. Garg, G. Carneiro, and I. Reid, “Unsupervised CNN for single view depth estimation: Geometry to the rescue,” in ECCV, 2016.

[25] D. Eigen and R. Fergus, “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture,” in ICCV, 2015.

[26] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in NIPS, 2014.

[27] Andrea Pilzer, Dan Xu, Mihai Marian Puscas, Elisa Ricci, Nicu Sebe, “Unsupervised Adversarial Depth Estimation using Cycled Generative Networks,” in 3DV, 2018.

[28] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,”

[29] Lei Zhang, Weihai Chen∗, Chao Hu, “S&CNet: A Enhanced Coarse-to-fine Framework For Monocular Depth Completion,” arXiv preprint arXiv:1907.06071, 2019.

[30] K. Yang, L.M. Bergasa, E. Romera, K. Wang, “Robustifying Semantic Cognition of Traversability across Wearable RGB-Depth Cameras,” Applied Optics, 58(12), pp. 3141-3155, 2019.

[31] N. Long, K. Wang, R. Cheng, W. Hu, K. Yang, “Unifying Obstacle Detection, Recognition and Fusion Based on Millimeter Wave Radar and RGB-Depth Sensors for the Visually Impaired,” Review of Scientific Instruments, 2019.

[32] K. Yang, L.M. Bergasa, E. Romera, X. Huang, K. Wang, “Predicting polarization beyond semantics for wearable robotics,” in IEEE-RAS International Conference on Humanoid Robots (Humanoids 2018), Beijing, China, November 2018.

[33] R. Cheng, K. Wang, J. Bai, Z. Xu, “OpenMPR: Recognize places using multimodal data for people with visual impairments," in Measurement Science and Technology, 2019.

[34] K. Yang, X. Hu, L.M. Bergasa, X. Huang, D. Sun, K. Wang, “Can we PASS beyond the Field of View? Panoramic Annular Semantic Segmentation for Real-World Surrounding Perception,” in IEEE Intelligent Vehicles Symposium (IV2019), Paris, France, June 2019.

[35] F. Liu, C. Shen, G. Lin, and I. Reid, “Learning depth from single monocular images using deep convolutional neural fields,” in PAMI, 2015.