IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 6, 2021, accepted December 25, 2021, date of publication January 5, 2022, date of current version January 20, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3140435

Joint Disparity Estimation and Pseudo NIR Generation From Cross Spectral Image Pairs

ZENGHUI DUAN AND CHEOLKON JUNG, (Member, IEEE)
School of Electronic Engineering, Xidian University, Xian 710071, China
Corresponding author: Cheolkon Jung (zhengzk@xidian.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61872280 and Grant 62111540272.

ABSTRACT Cross spectral stereo matching is a challenging task due to different spectral properties causing unreliable results in correspondence estimation. In this paper, we propose joint disparity estimation and pseudo near infrared (NIR) generation from cross spectral image pairs. To bridge the spectral gap between paired images, we adopt differential map operations and non-local blocks to improve the local attention and global attention of the network. The proposed network is based on unsupervised learning that consists of one encoder and two decoders, which performs both spectral translation and disparity estimation. For cooperative learning, we use difference map operation to connect two decoders, thus improving the inference ability of the decoder in regions even with large spectral differences. Experimental results show that the proposed network achieves good performance in cross spectral stereo matching for unreliable regions such as shadows and glasses. Moreover, the proposed network generates pseudo NIR images nearly the same as the ground truth even in the regions with large spectral difference. Besides, we achieve real-time speed of 27 FPS for 582×429 image pairs on RTX 2060 6G GPU due to the low computational complexity.

INDEX TERMS Cross spectral, neural networks, disparity estimation, image translation, stereo matching.

I. INTRODUCTION

In recent years, self-driving cars [1], [2], robotics [3]–[6] and augmented reality [7]–[9] have gained much attention. The reconstruction and understanding of 3D scenes are the key for them. Therefore, multiple cameras and sensors are deployed in many applications such as Tesla assisted driving car and iPhone face unlocking. This environment provides an opportunity to realize depth estimation for stereo image pairs through multiple images. Depth estimation can be achieved by stereo matching. It finds disparity d at the corresponding pixel in two images, and then calculates depth by combining the camera focal length and baseline distance. In traditional methods, stereo matching usually needs four steps to get disparity: matching cost calculation, cost aggregation, disparity estimation and disparity refinement. Although they have achieved good performance in stereo matching [5], [10], they still have limitations in many aspects. For example, they are not appropriate for dealing with and textural regions with repetitive patterns. With the rapid development of deep learning, many researchers have introduced deep neural networks into vision tasks and achieved good performance [11]–[14].

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Sharif.

Some researchers have developed network architectures for stereo matching. Chang and Chen [15] replaced traditional four steps of stereo matching with neural networks to improve performance and runtime, named PSM Net. Mayer *et al.* [16] reported that only using the features of left and right views can train an end-to-end network to output disparity directly. Researchers found that there are some inherent defects in color image-based stereo matching.

In low light condition, color images are often obtained by increasing the exposure time, thus resulting in blurry and noise corrupted pictures. To deal with the problem in low light condition, Jeon *et al.* [18] used long exposure gray scale images and achieved a remarkable improvement in stereo matching. In low light condition, brightness information is more accurate than color information. Compared with RGB images, NIR images have more brightness information. Therefore, cross spectral stereo matching from RGB-NIR image pairs can provide a solution to them. However, due to the cross spectral difference between two images, it is difficult to find correspondence for disparity estimation. Fig. 1 shows a large spectral gap between Y channel and NIR image. The grass field in the Y channel is dark, but it is very bright in the NIR image. This is from different reflection degrees of visible and NIR spectra. The same phenomena appear

(a) Y channel (b) NIR image

(c) Spectral gap (d) Disparity map by ours

FIGURE 1. Spectral gap between Y channel and NIR image.

(a) Y channel (b) NIR image

(c) Disparity map by DMC [17] (d) Disparity map by ours

(e) Pseudo NIR image by STN [17] (f) Pseudo NIR image by ours

FIGURE 2. Results of disparity estimation and pseudo NIR generation. DMC: deep material aware cross spectral stereo matching. STN: spatial transform network. Compared with DMC [17], the proposed method estimates accurate disparity in shadows. Moreover, our pseudo NIR generation result is closer to the real NIR image in (b) than STN [17] (see the glass region).

in the areas such as lights and clothes (see Fig. 1(c)). The spectral inconsistency in some area causes errors in disparity estimation. However, the proposed method successfully deals with the spectral gap problem and produces an accurate disparity map (see Fig. 1(d)).

In this paper, we propose joint disparity estimation and pseudo NIR generation from cross spectral image pairs. Disparity estimation and pseudo NIR generation are closely related to each other, and better pseudo NIR image generation leads to better disparity estimation. Thus, we build an end-to-end network for joint disparity estimation and pseudo NIR generation based on unsupervised learning that consists of one encoder and two decoders. In decoding, we obtain a difference map to improve the capability of translating and estimating disparity in the regions with large spectral difference. First, we obtain the translation error by subtracting the warped NIR image from the generated pseudo-NIR image, and then take its absolute value to obtain the left difference image. We calculate the right difference map from the original NIR image and the distorted pseudo NIR image. Finally, we combine the difference map with the features obtained by the encoder and feed them into the decoder to obtain a pseudo NIR image and a stereo pair of disparity maps. Moreover, we add non-local blocks into the decoder to consider global attention. We use Y channel of RGB image and NIR image as input, and obtain disparity map and pseudo NIR image as output. The encoder extracts features of seven scales from the input image pair, which is connected with the two decoders by skip connection. The two decoders are the disparity estimation decoder (DED) and the pseudo NIR generation decoder (PNGD). DED and PNGD perform decoding simultaneously. They calculate the difference maps in the last three blocks of decoding and add them to the decoding feature. Since the difference map comes from the disparity map and pseudo NIR image obtained from the preceding decoding blocks, the two decoders cooperate with each other, and finally generate disparity maps (left and right) and pseudo NIR image, respectively. Fig. 2 shows the results of disparity estimation and pseudo NIR generation by the proposed method. The whole architecture of the proposed network is illustrated in Fig. 3.

Compared with existing methods, the main contributions of this paper are as follows:

- We propose an end-to-end network for cross spectral stereo matching that consists of one encoder and two decoders. We use difference maps and non-local blocks at the decoders to deal with the cross spectral difference between paired images.
- We add non-local blocks in the front-end of the decoder to consider global attention, and use difference maps in the back-end of the decoder to consider local attention. Thus, the proposed network overcomes large spectral gap between paired images while making the pseudo NIR image similar to the real NIR image.
- We build two decoders in the last three blocks: disparity estimation decoder (DED) and pseudo NIR generation decoder (PNGD). The decoders are cooperated with each other, i.e. a cooperative network, thus producing

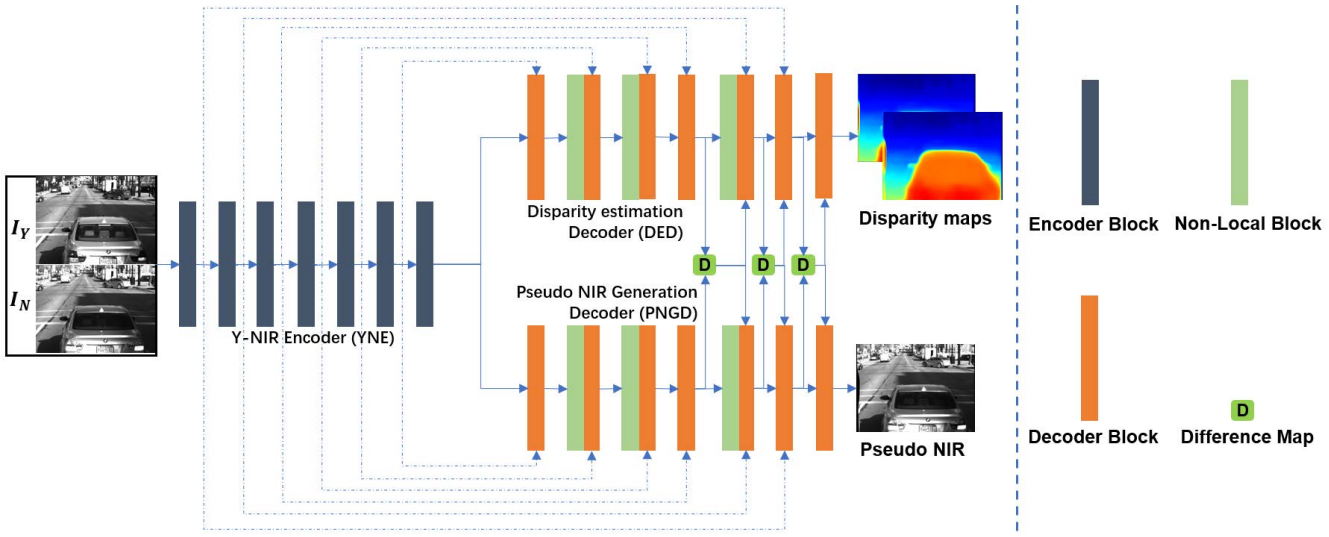


FIGURE 3. Whole architecture of the proposed network for cross spectral stereo matching. The proposed network consists of one encoder and two decoders. We use Y channel of RGB image and NIR image as the input, and we perform disparity estimation and pseudo NIR image generation concurrently by the two decoders: disparity estimation decoder (DED) and pseudo NIR generation decoder (PNGD). DED and PNGD cooperate with each other through difference maps. We use the same network structure for DED and PNGD, but utilize different loss functions for them to generate pseudo NIR images and estimate left and right disparity maps. The loss function for generating pseudo NIR image is Eq. (8), while the loss function of disparity estimation is Eq. (12).

accurate disparity estimation and pseudo NIR generation results.

II. RELATED WORK

A. CROSS-SPECTRAL STEREO MATCHING

The key of cross spectral stereo matching is to estimate correspondence between different spectra. Some methods focus on the design of a dense feature descriptor for the correspondence between pixels. Heo *et al.* [19] proposed a similarity measure which is robust to illumination and color changes for stereo matching under different radiation conditions. Shen *et al.* [20] designed the descriptor, named robust selective normalized cross correlation (RSNCC), to establish dense pixel correspondence between input images. Kim *et al.* [21] defined a dense adaptive self-correlation descriptor based on the observation that the self similarity in the image is not sensitive to modal changes. Jeong *et al.* [22] proposed a method based on CycleGAN to extract the hidden features for stereo matching. Other methods are spectral translation that translates the image pairs of different spectra into the same spectral pair. Chiu *et al.* [23] approximated infrared (IR) as the weighted fusion of R, G and B channels, and obtained the disparity map by global optimization. Jeon *et al.* [18] used linear channel combination to convert RGB images into monochromatic images, and then estimated the disparity map based on brightness constancy and edge similarity. Zhi *et al.* [17] also proposed a spectral translation network based on channel weighting to generate pseudo NIR image. They provided an unsupervised CNN framework to predict the disparity map. Lin *et al.* [24] used a kernel prediction network (KPN) to sum the NIR images with spatial weighting, and generated pseudo NIR images aligned with RGB images for face stereo matching in low light condition.

B. UNSUPERVISED DEPTH ESTIMATION

In recent years, many unsupervised methods have been used for optical flow and disparity estimation. Based on Taylor expansion, Garg *et al.* [25] proposed an unsupervised convolution neural network (CNN) for single view depth prediction that was trained from coarse to fine. Xie *et al.* [12] designed a Deep3D network to minimize pixel reconstruction loss and generate the right view image. Godard *et al.* [26] proposed a training loss based on epipolar geometry constraints that enforced disparity consistency produced by left and right images. Zhou *et al.* [27] started from a randomly initialized network, used left and right check to guide training, and updated network parameters iteratively. Yang *et al.* [28] conducted semantic feature embedding and regularized semantic cues as the loss term to improve learning disparity. Zhong *et al.* [29] used image warping errors as the loss function to drive the training process, and showed the effectiveness in KITTI and Middlebury stereo benchmark datasets. Zhang *et al.* [30] proposed ActiveStereoNet whose reconstruction loss was robust to noise, texture free region and illumination. ActiveStreoNet treated the oversmoothing problem in unsupervised disparity estimation, and preserved edges effectively by dealing with occlusion. They have achieved good performance in stereo matching between paired color images, but they are not suitable for cross spectral matching.

C. IMAGE TO IMAGE TRANSLATION

Image to image translation is the transformation between image domains such as style transfer [14], [31], colorization [32], [33] and decolorization [34]. Song *et al.* [35] achieved outstanding decolorization performance by adaptively calculating the weighted sum of each channel between

color images. Isola *et al.* [36] used a conditional confrontation network as a general solution for image translation, and implemented style conversion between various image pairs. Zhu *et al.* [37] used CycleGAN which introduced the loss of cyclic consistency to achieve image translation of unpaired images. Because CycleGAN consumes too much memory, Zhi *et al.* [17] proposed a spectral translation network (STN) to generate the weights of RGB three channels and estimate the pseudo NIR image.

III. PROPOSED METHOD

As illustrated in Fig. 3, the proposed method can be divided into three parts: backbone of one encoder and two decoders (1E2Ds), Non-Local block (NLB) [38] and difference map operation (DMO). NLB is a kind of self attention mechanism that generates attention on important features, while DMO makes the decoder improve the translation and disparity estimation capability for the regions with large spectral gap.

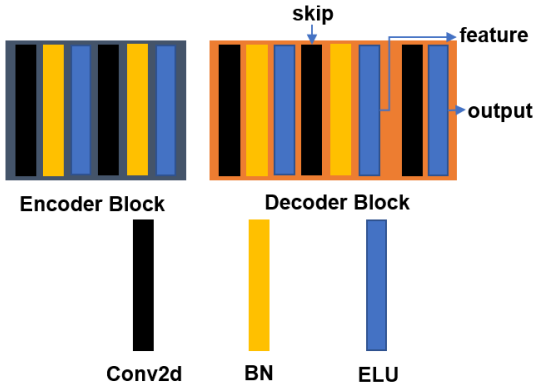


FIGURE 4. Encoder block and decoder block. To output the results, the decoder includes one convolution layer and one ELU activation layer.

A. BACKBONE OF ONE ENCODER AND TWO DECODERS

In recent years, a number of studies on deep learning are based on the network structure of encoder and decoder. This is because the encoder can extract both shallow structural features and deep semantic features, while the decoder can fuse the features of different scales to improve the accuracy. Mayer *et al.* [16] and Zhi *et al.* [17] proved that stereo matching can be realized by an end-to-end network. Inspired by them, we propose a network architecture based on one encoder and two decoders, called 1E2Ds. The encoder and decoder contain seven encoding blocks and seven decoding blocks, respectively, and realize the same size for output and input. As shown in Fig. 4, both the encoder block and the decoder block are composed of two ''convolution layer-batch normalization [39]-ELU activation [40]''. We add one ''convolution layer-ELU activation [40]'' into the decoder block as the output layer. For the disparity estimation, we use the encoder (YNE) to extract the left and right features F_Y, F_N based on Y-NIR stereo pair I_Y, I_N. Then, we use the features as the input of the disparity estimation decoder (DED) to predict the left and right disparity d^l, d^r. When we translate Y into NIR, we assume that there is a proportional relationship F between NIR and Y, thus the pseudo NIR is expressed as follows:

$$I_{pN} = F(p)I_Y(p) \tag{1}$$

where p represents each pixel position. As shown in Eq. (1), the pseudo NIR generation decoder (PNGD) predicts the scale matrix F. Different from the previous work, we put the features of NIR and Y into PNGD to generate F. After that, we obtain the predicted pseudo NIR image by Eq. (1).

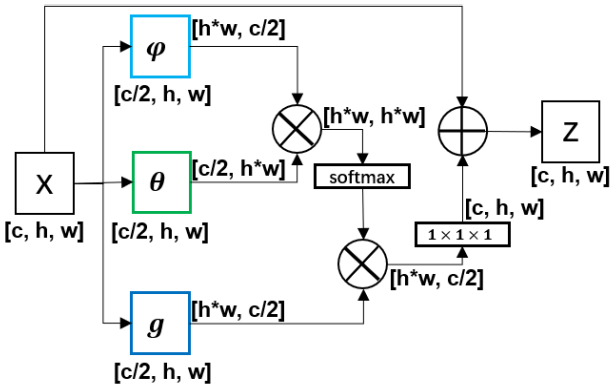


FIGURE 5. Flow of the non-local block (NLB). x is the input feature, z is the output feature, "⊗" is the matrix multiplication, and "⊕" is the matrix addition. "ϕ", "θ" and "g" are convolution layers with kernel 1 × 1 × 1.

B. NON-LOCAL BLOCK

Wang *et al.* [38] proposed non-local neural networks to provide a generic non-local operation in deep neural networks as follows:

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j) \tag{2}$$

where i and j denote the position of features, similarity $f(x_i, x_j) = e^{(\theta(x_i)^T \phi(x_j))}$, and coefficient $C(x) = \sum_{\forall j} f(x_i, x_j)$. For a given i, $\frac{1}{C(x)} f(x_i, x_j)$ performs the softmax computation along dimension j. Thus, we get:

$$y = softmax(\theta(x)^T \phi(x)) g(x) \tag{3}$$

The output z from the input x is as follows:

$$z = x + W_z \cdot y \tag{4}$$

where W_z is a weight matrix. By calculating the interaction between any two positions, the remote dependency can be captured directly without being limited to adjacent points, which is equivalent of constructing a convolution kernel with the same size as the feature map, thus more information can be retained. Fig. 5 shows the flow of NLB. Firstly, the input feature is convoluted by 1×1 to compress the number of channels, and then ϕ, θ and g features are obtained through three convolution layers. Based on reshape operation, the dimensions of the three features are transformed, and then matrix

multiplication is performed to obtain the self-correlation in the features. Then, softmax operation is performed on the self-correlation feature to get the weight [0, 1] for input features. Here, we get the self attention coefficient. Next, the attention coefficient is multiplied back to g, and expanded to dimension $C \times H \times W$. Finally, the residual of the input feature map x is added to the output. As shown in Fig. 3, we add NLB into the third, fifth and sixth layers of the decoder. The local blocks in the fifth and sixth layers are mainly used to enhance the attention of the network in semantic features, and the local blocks in the third layer are used to enhance the attention of the network in structural features.

C. DIFFERENCE MAP OPERATION

Based on the careful observation of the previous spectral translation, we find that it is difficult to translate categories of clothing, grass, and glass in the spectral translation. There are still some errors between the generated pseudo NIR image and the real NIR image. To get good pseudo NIR image generation for the regions with large spectral gap, we estimate the spectral translation error by combining the disparity estimation results from the DED network. We use it to connect the last three decoding blocks between two decoders (see Fig. 3). As shown in Fig. 6 is the flow of difference map operation (DMO). First, we need to warp the input NIR image I_N to get the real warped NIR image I_{wN} aligned with the pseudo NIR image I_{pN}. This process can be described as follows:

$$I_{wN} = \omega(I_N, d^l) \tag{5}$$

where ω is the warping operator, which can be expressed as follows:

$$I^l_{x,y} = I^r_{x+d^l_{x,y},y} \tag{6}$$

Then, we take the absolute value of the difference between I_{pN} and I_{wN} as follows:

$$I^l_D = |I_{pN} - I_{wN}| \tag{7}$$

For PNGD, we feedback the left difference map to the network. For DED, since we generate left and right disparities, we feedback the left and right difference maps together for the disparity estimation on unreliable regions.

D. LOSS FUNCTION

Due to the lack of datasets for training, unsupervised learning is suitable for cross spectral stereo matching. Since it is hard to generate the ground truth disparity maps, we use three terms of left-right consistency L_{LRC}, structural similarity L_a, and smoothness L_s in the loss function. L_{LRC} measures consistency between left and right disparity maps, while L_a and L_s are estimated from the pseudo NIR and warped NIR images. Therefore, we use two decoders: One decoder outputs left and right disparity maps, while the other outputs pseudo NIR images.

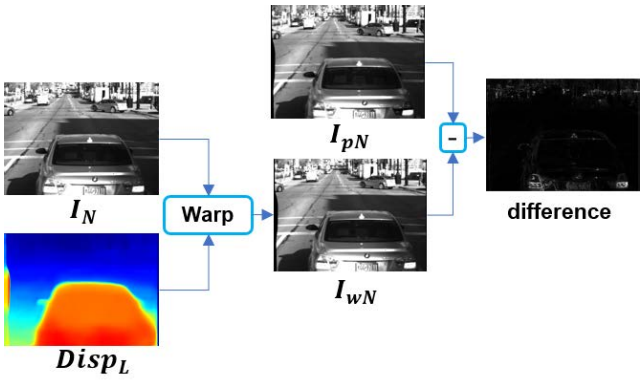


FIGURE 6. Flow of the difference map operation (DMO). "Warp" is the warping operator. "-" represents matrix subtraction operation. We take the absolute value after the "-" operation to get the difference map.

1) DED LOSS

DED predicts left-right disparities d^l, d^r based on the encoder features. The DED loss consists of three terms: a left and right consistency (LRC) term L_{LRC}, a material aware alignment term L_a and a material aware smoothness term L_s. Zhi *et al.* [17] reported that material aware alignment loss $L_{a,m}$ and smoothness loss and $L_{s,m}$ effectively improve the disparity estimation of unreliable regions. Therefore, the total loss function in this paper follows the Zhi *et al.*'s work [17] as follows:

$$\begin{aligned} L_{DED} = {} & \lambda_c L_{LRC} \\ & + \lambda_{a,m} \sum_{m \in \mathcal{M}} \mu_m (L^l_{a,m} + L^r_{a,m}) \\ & + \lambda_{s,m} \sum_{m \in \mathcal{M}} \mu_m (L^l_{s,m} + L^r_{s,m}) \end{aligned} \tag{8}$$

where $\mathcal{M}$ ='light', 'glass', 'glossy', 'vegetation', 'skin', 'clothing', 'bag', 'common'; μ_m denotes segmentation information; λ_c, $\lambda_{a,m}$ and $\lambda_{s,m}$ are weights of terms which are empirically set in training.

The LRC term L_{LRC} describes the consistency of left and right disparity maps as follows:

$$\begin{aligned} L_{LRC} = {} & \frac{1}{N} \sum_{p \in \Omega} \left(|d^l(p) - \omega(d^r, -d^l)(p)| \right) \\ & + \frac{1}{N} \sum_{p \in \Omega} \left(|d^r(p) - \omega(d^l, d^r)(p)| \right) \end{aligned} \tag{9}$$

where N is the number of pixels in one image, and Ω is the pixel coordinate space.

The material aware alignment term L_a compares intensity and structure between the aligned NIR and pseudo NIR images [17]. We use the structural dissimilarity function $\delta(I_1, I_2)$ [41] for the term as follows:

$$L^l_{a,m} = \frac{1}{N} \sum_{p \in \Omega} \left(\alpha \delta(I_{pN}, I_{wN})(p) + (1-\alpha) |I_{pN}(p) - I_{wN}(p)| \right) \tag{10}$$

where α is set to be 0.85 as suggested by Godard *et al.* [42].

The material aware smoothness term L_s encourages disparities to be locally smooth and edge-aware as follows [17]:

$$L^l_{s,m} = \frac{1}{N}\sum_{p\in\Omega}((|\partial_x d^l|e^{-\|\partial_x I_Y\|} + |\partial_y d^l|e^{-\|\partial_y I_Y\|})(p)) \quad (11)$$

where ∂I is the gradient of input image I and ∂d is the gradient of disparity. For $L_{a,m}$ and $L_{s,m}$, we only provide the loss function in the left Y channel.

2) PNGD LOSS

The PNGD network predicts the scale matrix F between the Y channel and real NIR image. Then, the pseudo NIR is generated by Eq. (1). To make the generated pseudo NIR image closer to the real NIR image, we use the consistency between the left and right views as the PNGD loss as follows:

$$L_{PNGD} = |I_{pN} - I_{wN}| + |I_{wpN} - I_N| \quad (12)$$

where I_{pN} is the pseudo NIR image, I_{wN} is the warped NIR image, I_{wpN} is the warped pseudo NIR image, and I_N is the input NIR image.

3) TOTAL LOSS

The proposed network is an end-to-end network, and the total loss function is the sum of two decoders' losses as follows:

$$L_{total} = L_{DED} + L_{PNGD} \quad (13)$$

IV. EXPERIMENTS

A. DATASET

We use Pitts-Stereo RGB-NIR dataset proposed by [17] for training and testing. The dataset includes a variety of scenes such as campus roads, highways, downtown, parks and residential areas with a variety of weather and light conditions: sunny, cloudy and dark conditions. In addition, Zhi *et al.* [17] designed 8 material labels for common, lights, glass, glossy surfaces, vegetation, skin, clothing, and bags with the pretrained Deeplab [43]. We split the dataset into two parts: the training set contains 40,000 RGB-NIR image pairs, while the testing set contains 2,000 RGB-NIR image pairs.

B. IMPLEMENTATION DETAILS

1) PARAMETERS

The ratio of disparity to image width is predicted by DED network, and the ratio of Y to NIR is predicted by PNGD network. The weights of the losses in DED are set to $\lambda_c = 2$, $\lambda_{a,m} = 1$ for all materials. When calculating the smoothing loss of lights, glass and grosses, set $\lambda_{s,m}$ to 3,000, 1,000 and 80, respectively. For other materials, we set $\lambda_{s,m} = 25$.

2) TRAINING AND TESTING

We implement the proposed network by PyTorch framework [44] and train it using a DGX station with NVIDIA Tesla V100 GPU. The parameter setting is as follows. The weights are initialized by Kaiming initialization [45] and optimized by Adam optimizer. We set the Adam weight decay to 0.0001 and the batch size to 32. The learning rate is 0.0005 decreasing by half every 12 epochs. The model is trained 48 epochs, which takes 22 hours to finish.

C. COMPARISONS WITH THE STATE-OF-THE-ART METHODS

1) DISPARITY ESTIMATION

We compare the proposed network with state-of-the-art methods: PSM Net [15], dense adaptive self-correlation descriptor (DASC) [21] and deep material aware-cross spectral stereo matching (DMC) [17]. Fig. 7 shows visual comparison among them. DASC shows good matching results in clothing and shadow areas, but achieves worse performance in glass. This is because there exists a big spectral gap in glass, which makes feature extraction difficult. The overall performance of DMC is good, but it is easy to have mismatch in shadow areas and nearby objects causing errors in disparity estimation. PSM Net achieves the worst matching performance, which is suitable for stereo matching in single modal image pairs. The proposed method outperforms the others in most cases and performs better than DMC in shadow areas. However, the proposed method performs slightly worse than DMC for matching distant objects.

2) PSEUDO NIR GENERATION

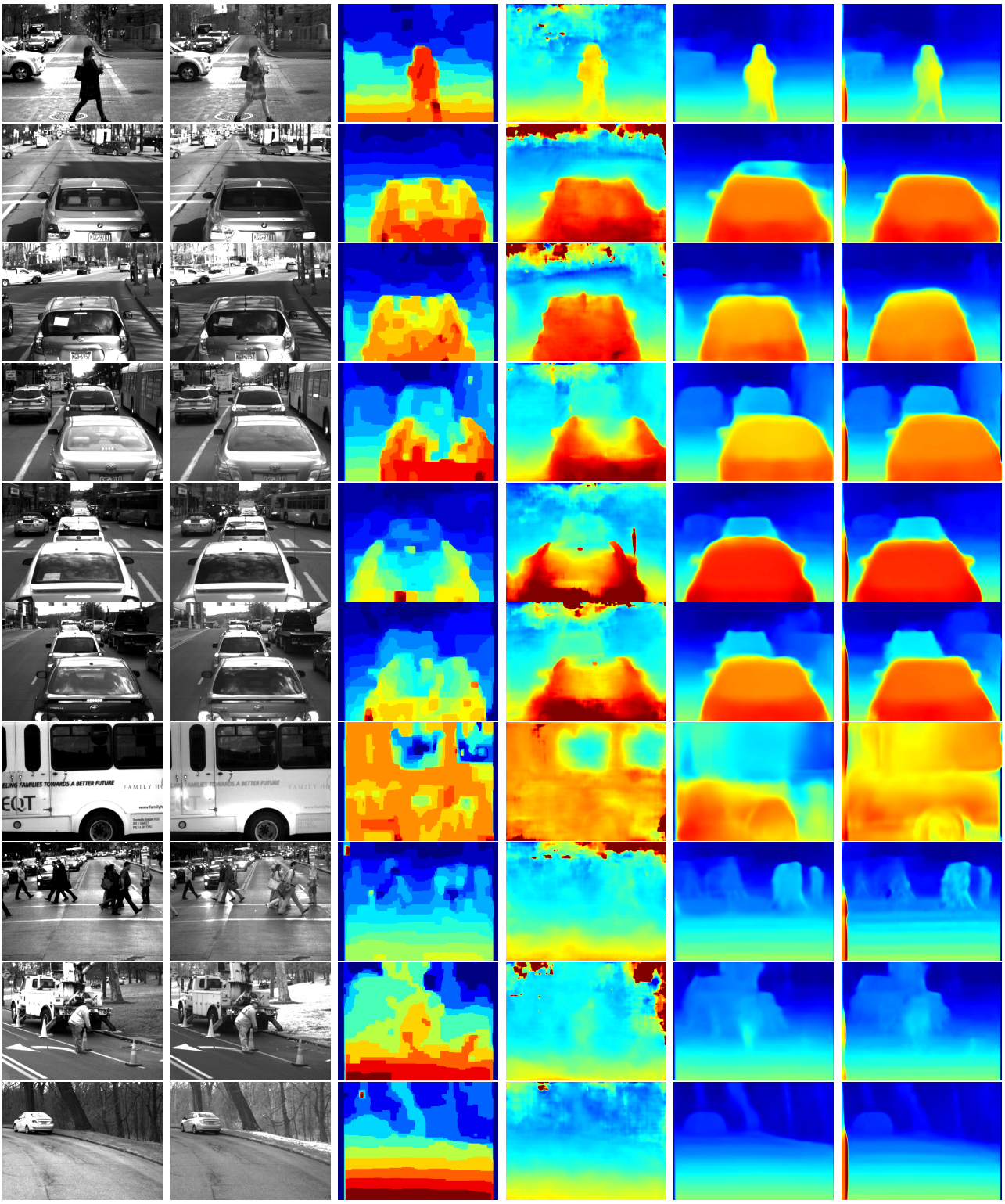
We compare the proposed method with state-of-the-art methods for image-to-image translation: CycleGAN [37], FUCC [22], and the spectral translation network (STN) [17]. CycleGAN [37] has been widely used in other image-to-image translation fields, while FUCC [22] is an unpaired image training method based on hidden features. Fig. 8 shows visual comparison among them in pseudo NIR generation. It can be seen that the results of CycleGAN are not natural. FUCC performs better than that of CycleGAN, but its brightness is dark, thus reducing the visual quality. STN seems to greatly improved, but the translation of clothes is still not accurate. The proposed method looks closer to the real NIR image with black holes in the left regions caused by disparity.

TABLE 1. Ablation study on pseudo NIR image generation in terms of PSNR (unit: dB) and SSIM.

	Without both	Without NLB	Without DMO	Proposed
PSNR	26.0	25.26	25.48	27.51
SSIM	0.78	0.78	0.78	0.80

3) ABLATION STUDY

We perform ablation study on the network structure: "only using encoder-decoder structure", "not using self attention mechanism" and "not using difference map operation". Fig. 9 shows ablation study on nonlocal block (NLB) and difference map operation (DMO). Without NLB and DMO, the disparity estimation in clothing, shadow and reflective areas



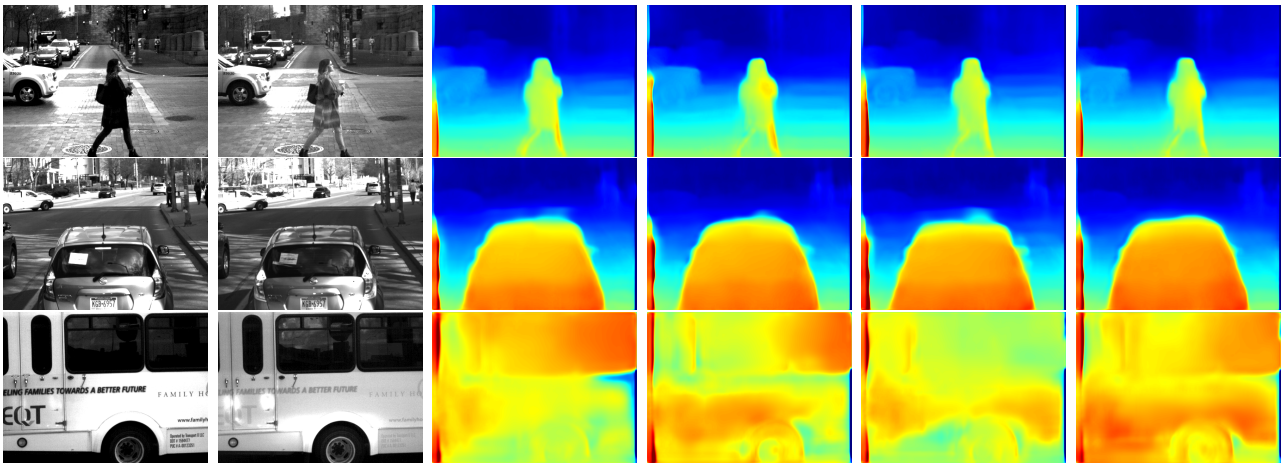
(a) Left Y (b) Right NIR (c) DASC [21] (d) PSM Net [15] (e) DMC [17] (f) Ours

FIGURE 7. Disparity Estimation qualitative comparison. Rows 1, 8, 9: Clothing. Rows 2, 3: Shadow. Rows 4, 5, 6: Glasses. Row 7: Near scene. Row 10: Distant scene.

is not accurate. When only DMO is added, the disparity estimation results in shadow and reflective areas become better. When only the local block is added, the disparity estimation results in clothing become better, while those in shadow and glass become worse. The full network structure improves the results in all aspects. Through the ablation experiment, it can

(a) Y (b) NIR (c) CycleGAN [37] (d) FUCC [22] (e) STN [17] (f) Ours

FIGURE 8. Visual comparison among different methods in pseudo NIR generation. From the first and second rows, the proposed method performs better than the others in the glass area. The third and fourth rows show that the proposed method is more accurate than the others in the pseudo NIR generation of clothes, while the fifth row shows the performance in the reflective area that our result is closest to the real NIR image (see the left side of cars).



(a) Y (b) NIR (c) No all (d) No NLB (e) No DMO (f) Full method

FIGURE 9. Ablation study on nonlocal block (NLB) and difference map operation (DMO). We compare the results in the regions with a big spectral gap such as clothing, shadow, and glass. The disparity estimation results in them are significantly improved by NLB and DMO. DMO effectively treats the spectral difference of shadow and reflection area, while NLB overcomes the spectral difference in clothing area.

be revealed that NLB mainly improves the disparity estimation in clothing and glass, while DMO pays more attention to the shadow and reflective areas. When the two blocks are used, the proposed method achieves trustworthy performance in disparity estimation method by taking both advantages. Moreover, we provide another ablation study on the pseudo NIR image generation in Table 1. We calculate PSNR and SSIM values on the pseudo NIR images. NLB and DMO increase 1.51dB in PSNR and 0.02 in SSIM. However, without both NLB and DMO, the PSNR performance is 26.0 dB,

which is better than those without NLB and without DMO. The results indicate that NLB and DMO are complementary. DMO concerns details in pseudo NIR images, while NLB concerns global information in them.

TABLE 2. PSNR (unit: dB) comparison among different methods for disparity estimation and pseudo NIR generation. The proposed method achieves the best performance.

	CycleGAN [37]	FUCC [22]	STN [17]	Ours
DASC [21]	10.47	11.1	21.25	20.41
PSM Net [15]	10.26	10.66	17.33	18.19
DMC [17]	10.45	10.81	22.51	17.65
Ours	10.36	10.60	18.94	27.50

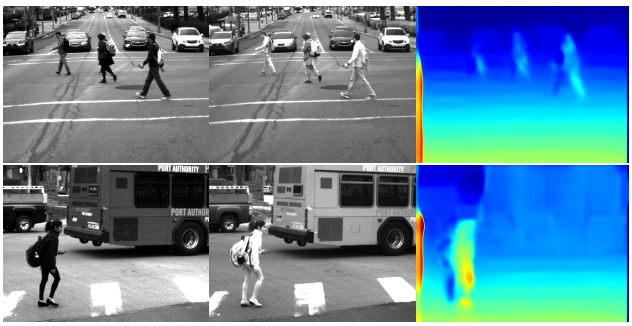
4) QUANTITATIVE MEASUREMENTS

For quantitative measurements, we use PSNR and SSIM as evaluation metrics for disparity estimation and pseudo NIR generation. Since it is difficult to generate the NIR image aligned with the Y channel, we use the warped NIR image as the reference for PSNR and SSIM evaluations. For a convincing experiments, we first use the disparity estimation results d^l to warp I_N, and get the warped NIR image I_wN by disparity estimation methods. Then, we measure PSNR and SSIM on the pseudo NIR images I_{pN} generated by various spectral translation networks. Tables 2 and 3 show PSNR and SSIM comparisons among different methods for disparity estimation and pseudo NIR generation, respectively. As shown in Table 2, we fix the disparity estimation method and then get the pseudo NIR generation results. When the disparity estimation method is DASC [21] or DMC [17], STN [17] achieves the best PSNR performance by 21.25dB and 22.51dB, respectively. When we use PSM Net [15] and the proposed method for disparity estimation, the proposed method achieves the best PSNR performance by 18.19dB and 27.50dB, respectively. Then, we fix the pseudo NIR generation method to get the disparity estimation results. When we use CybleGAN [37] and FUCC [22] for pseudo NIR generation, DASC [21] achieves the best PSNR performance by 10.47dB and 11.1dB, respectively. When we use STN [17] for pseudo NIR generation, DMC has the best PSNR performance by 22.51dB. On the whole, the proposed method achieves 27.50dB in PSNR performance and outperforms the others in both disparity estimation and pseudo NIR generation. Moreover, we provide SSIM comparison among different methods for disparity estimation and pseudo NIR generation in Table 3. SSIM represents structural similarity in images that reflects perceptual quality of the human visual system (HVS). As shown in the table, the proposed method achieves 0.80 in SSIM performance and outperforms the others in both disparity estimation and pseudo NIR generation.

Moreover, we provide quantitative measurements in terms of disparity RMSE and runtime (s/pair) in Table 4. We use DMC dataset [23] for experiments. To test PSM Net, DASC, DMC and the proposed method, we use a PC with Ryzen 5 3600 CPU and Tesla v100 32G GPU running Ubuntu 18.04 and Pytorch 1.4. Since DASC does not need GPU, it is tested on CPU. As shown in the table, the RMSE of PSM Net [15] is the largest, which indicates the worst performance. This is because PSM Net [15] is a disparity estimation method from a color image pair and is not suitable for cross-spectral images. DASC [21] is a traditional stereo matching method based on dense adaptive self-correlation descriptor. DASC performs the best on clothing, and its average RMSE reaches 1.28. It can only run on the CPU, which takes 33.92 s/pair. DMC [23] reaches 0.80, i.e. the best performance, which achieves good performance in almost all categories with the lowest runtime. The proposed method gets the best performance in the skin category with comparable runtime, which is worse than DMC in RMSE.

TABLE 3. SSIM comparison among different methods for disparity estimation and pseudo NIR generation. The proposed method generates pseudo NIR images the most similar to the real NIR images.

	CycleGAN [37]	FUCC [22]	STN [17]	Ours
DASC [21]	0.35	0.41	0.73	0.73
PSM Net [15]	0.34	0.41	0.58	0.56
DMC [17]	0.36	0.42	0.69	0.60
Ours	0.36	0.41	0.74	0.80



(a) Y (b) NIR (c) Ours

FIGURE 10. Failure cases by the proposed network. Since clothes have a big spectral gap between Y channel and NIR image, the proposed network suffers from disparity estimation for them.

5) FAILURE CASES

However, there are some failure cases in the proposed method. As shown in Fig. 10, the proposed method causes errors in the disparity estimation of clothes with a big spectral gap. The intensity of NIR images highly depends on the material property, but the intensity of Y channel highly depends on the visible light reflectivity. Thus, Y channel (dark) and NIR image (bright) is opposite in the intensity polarity for clothes, which makes the proposed method a wrong prediction. Since the proposed method generates pseudo NIR images based on the disparity estimation results, holes appear in the pseudo NIR images (see the left regions of our results in Fig. 8).

TABLE 4. Quantitative measurements in terms of disparity RMSE and runtime (s). For experiments, we use a PC with Ryzen 5 3600 CPU and Tesla v100 32G GPU running Ubuntu 18.04 and Pytorch 1.4. Since DASC does not need GPU, it is tested on CPU.

Method	Common	Light	Glass	Glossy	Vegetation	Skin	Clothing	Bag	Mean	time(s)
DASC [21]	0.82	1.24	1.50	1.82	1.09	1.59	0.80	1.33	1.28	33.92
PSM Net [15]	10.66	16.78	6.47	6.29	19.15	12.72	16.68	9.31	12.26	0.2
DMC [17]	0.53	0.69	0.65	0.70	0.72	1.15	1.15	0.80	0.80	0.003
Ours	0.61	1.10	0.93	1.24	0.95	1.14	1.27	0.85	1.01	0.009

V. CONCLUSION

In this paper, we have proposed joint disparity estimation and pseudo NIR generation from cross spectral image pairs. We have adopted NLB and DMO to bridge the spectral gap between Y channel and NIR image. The proposed network consists of one encoder (YNE) and two decoders (PNGD and DED) based on unsupervised learning. To facilitate cooperative learning between two decoders, we have proposed DMO to obtain the difference map between the generated pseudo NIR image and the warped NIR image. We have combined DMO with the encoded features and feed them into the decoders for joint disparity estimation and pseudo NIR generation. Moreover, we have introduced NLB into the decoder to capture global attention in the proposed network. Experimental results demonstrate that the proposed method achieves the state-of-the-art performance in both disparity estimation and pseudo NIR generation, i.e. 27.5dB in PSNR and 0.8 in SSIM. Our future work includes filling holes caused by pseudo NIR generation.

REFERENCES

[1] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8445–8453.

[2] P. Li, X. Chen, and S. Shen, "Stereo R-CNN based 3D object detection for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7644–7652.

[3] M. Mancini, G. Costante, P. Valigi, and T. A. Ciarfuglia, "Fast robust monocular depth estimation for obstacle detection with fully convolutional networks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 4296–4303.

[4] M. Ye, E. Johns, A. Handa, L. Zhang, P. Pratt, and G.-Z. Yang, "Self-supervised Siamese learning on stereo image pairs for depth estimation in robotic surgery," 2017, *arXiv:1705.08260*.

[5] A. Saxena, J. Schulte, and A. Y. Ng, "Depth estimation using monocular and stereo cues," in *Proc. IJCAI*, vol. 7, 2007, pp. 2197–2203.

[6] K. Schmid, T. Tomic, F. Ruess, H. Hirschmuller, and M. Suppa, "Stereo vision based indoor/outdoor navigation for flying robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 3955–3962.

[7] N. Zenati and N. Zerhouni, "Dense stereo matching with application to augmented reality," in *Proc. IEEE Int. Conf. Signal Process. Commun.*, Nov. 2007, pp. 1503–1506.

[8] C. Nguyen, S. DiVerdi, A. Hertzmann, and F. Liu, "Depth conflict reduction for stereo VR video interfaces," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2018, pp. 1–9.

[9] H. Alhaija, S. K. Mustikovela, L. Mescheder, A. Geiger, and C. Rother, "Augmented reality meets computer vision: Efficient data generation for urban driving scenes," *Int. J. Comput. Vis.*, vol. 126, no. 9, pp. 961–972, 2018.

[10] 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.

[11] M. Zhu, Z. Feng, X. Zhou, R. Xiao, Y. Qi, and X. Zhang, "Specific emitter identification based on synchrosqueezing transform for civil radar," *Electronics*, vol. 9, no. 4, p. 658, 2020.

[12] J. Xie, R. B. Girshick, and A. Farhadi, "Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks," in *Proc. ECCV*, 2016, pp. 842–857.

[13] M. Zhu, Z. Feng, and X. Zhou, "A novel data-driven specific emitter identification feature based on machine cognition," *Electronics*, vol. 9, no. 8, p. 1308, Aug. 2020.

[14] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[15] 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.

[16] 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.

[17] T. Zhi, B. R. Pires, M. Hebert, and S. G. Narasimhan, "Deep material-aware cross-spectral stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1916–1925.

[18] H.-G. Jeon, J.-Y. Lee, S. Im, H. Ha, and I. S. Kweon, "Stereo matching with color and monochrome cameras in low-light conditions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4086–4094.

[19] Y. S. Heo, K. M. Lee, and S. U. Lee, "Joint depth map and color consistency estimation for stereo images with different illuminations and cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1094–1106, May 2013.

[20] X. Shen, L. Xu, Q. Zhang, and J. Jia, "Multi-modal and multi-spectral registration for natural images," in *Proc. ECCV*, 2014, pp. 309–324.

[21] S. Kim, D. Min, B. Ham, S. Ryu, M. N. Do, and K. Sohn, "DASC: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2103–2112.

[22] S. Jeong, S. Kim, K. Park, and K. Sohn, "Learning to find unpaired cross-spectral correspondences," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5394–5406, Nov. 2019.

[23] W.-C. Chiu, U. Blanke, and M. Fritz, "Improving the Kinect by cross-modal stereo," in *Proc. Brit. Mach. Vis. Conf.*, 2011.

[24] S. Lin, J. Zhang, J. Chen, Y. Wang, Y. Liu, and J. Ren, "Cross-spectral stereo matching for facial disparity estimation in the dark," *Comput. Vis. Image Understand.*, vol. 200, Nov. 2020, Art. no. 103046.

[25] R. Garg, B. V. Kumar, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Proc. ECCV*, 2016, pp. 740–756.

[26] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6602–6611.

[27] C. Zhou, H. Zhang, X. Shen, and J. Jia, "Unsupervised learning of stereo matching," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1576–1584.

[28] G. Yang, H. Zhao, J. Shi, Z. Deng, and J. Jia, "SegStereo: Exploiting semantic information for disparity estimation," in *Proc. ECCV*, 2018, pp. 636–651.

[29] Y. Zhong, Y. Dai, and H. Li, "Self-supervised learning for stereo matching with self-improving ability," 2017, *arXiv:1709.00930*.

[30] 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)*, Sep. 2018, pp. 784–801.

[31] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," 2016, *arXiv:1603.08155*.
[32] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. ECCV*, 2016, pp. 649–666.
[33] P. Vitoria, L. Raad, and C. Ballester, "ChromaGAN: Adversarial picture colorization with semantic class distribution," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2434–2443.
[34] X. Kang, P. Duan, S. Li, and J. A. Benediktsson, "Decolorization-based hyperspectral image visualization," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4346–4360, Aug. 2018.
[35] Y. Song, L. Bao, X. Xu, and Q. Yang, "Decolorization: Is RGB2gray() out?" in *Proc. SIGGRAPH Asia Tech. Briefs*, 2013, pp. 1–4.
[36] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.
[37] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.
[38] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
[39] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*.
[40] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*.
[41] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[42] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 270–279.
[43] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[44] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in Pytorch," in *Proc. NIPS Workshop*, 2017.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

ZENGHUI DUAN received the B.S. degree in electronic engineering from Xidian University, China, in 2019, where he is currently pursuing the M.S. degree. His research interests include image processing, machine learning, and video coding.

CHEOLKON JUNG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronic engineering from Sungkyunkwan University, Republic of Korea, in 1995, 1997, and 2002, respectively. He was a Research Staff Member with the Samsung Advanced Institute of Technology, Samsung Electronics, Republic of Korea, from 2002 to 2007. He was a Research Professor with the School of Information and Communication Engineering, Sungkyunkwan University, from 2007 to 2009. Since 2009, he has been with the School of Electronic Engineering, Xidian University, China, where he is currently a Full Professor and the Director of the Xidian Media Laboratory. His main research interests include image and video processing, computer vision, pattern recognition, machine learning, computational photography, video coding, virtual reality, information fusion, multimedia content analysis and management, and 3-D TV.