Unsupervised Cross-Spectral Stereo Matching by Learning to Synthesize

Mingyang Liang,[1,2*] **Xiaoyang Guo,**[3*] **Hongsheng Li,**[3] **Xiaogang Wang,**[3] **You Song**[1†]

[1]Beihang University, Beijing, China
[2]SenseTime Research
[3]The Chinese University of Hong Kong, Hong Kong, China
{liangmingyang,songyou}@buaa.edu.cn, {xyguo, hsli, xgwang}@ee.cuhk.edu.hk

Abstract

Unsupervised cross-spectral stereo matching aims at recovering disparity given cross-spectral image pairs without any depth or disparity supervision. The estimated depth provides additional information complementary to original images, which can be helpful for other vision tasks such as tracking, recognition and detection. However, there are large appearance variations between images from different spectral bands, which is a challenge for cross-spectral stereo matching. Existing deep unsupervised stereo matching methods are sensitive to the appearance variations and do not perform well on cross-spectral data. We propose a novel unsupervised cross-spectral stereo matching framework based on image-to-image translation. First, a style adaptation network transforms images across different spectral bands by cycle consistency and adversarial learning, during which appearance variations are minimized. Then, a stereo matching network is trained with image pairs from the same spectra using view reconstruction loss. At last, the estimated disparity is utilized to supervise the spectral translation network in an end-to-end way. Moreover, a novel style adaptation network F-cycleGAN is proposed to improve the robustness of spectral translation. Our method can tackle appearance variations and enhance the robustness of unsupervised cross-spectral stereo matching. Experimental results show that our method achieves good performance without using depth supervision or explicit semantic information.

Introduction

Multi-camera multi-spectral systems have become very common in many modern devices like Realsense, Kinect, and iPhoneX. Moreover, it has been proven that infrared images are very helpful in face recognition (Lezama, Qiu, and Sapiro 2017), detection (Xu et al. 2017), and scene parsing (St-Charles, Bilodeau, and Bergevin 2017).

Stereo matching is one of the most heavily investigated topics in computer vision (Scharstein and Szeliski 2002). Given a rectified image pair (I_l for the left image, I_r for the right image), stereo matching focuses on finding correspondence of each pixel between two images. If the right pixel $I_r(x-d, y)$ corresponds to the left pixel $I_l(x, y)$, then we

---

(a) Left VIS   (b) Right NIR

Clothes
Lights
Grass

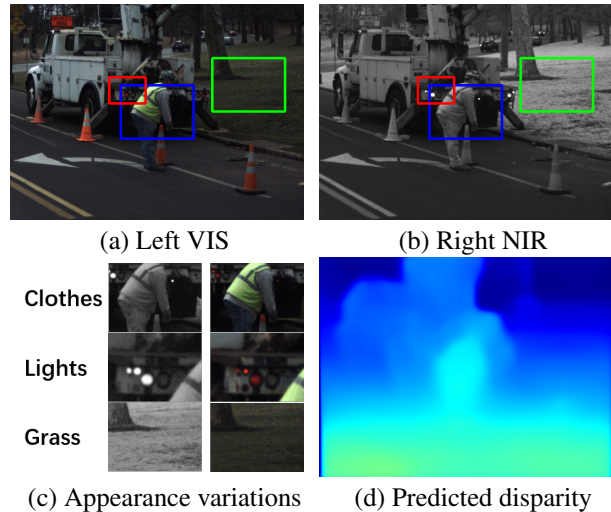(c) Appearance variations   (d) Predicted disparity

Figure 1: The appearance variations between cross-spectral image pairs. VIS and NIR represent visible and near-infrared images respectively. Almost invisible light sources in the VIS turn into a dazzling white in the NIR. Grassland also shows huge illumination difference between VIS and NIR. The goal of cross-spectral stereo matching is to overcome the appearance differences between different spectra and predict accurate disparity.

can define d as the disparity of the pixel $I_l(x, y)$. Moreover, if we know the camera's focal length f and the distance B between the two camera centers, the disparity can be converted into depth by fB/d.

Cross-spectral stereo matching is stereo matching for images from different spectra, for example the left image is a visible image and the right image is a near-infrared image in Fig. 1. The recovered depth provides additional information which is complementary to semantic features of individual spectrum. In addition, the estimated depth can help improve missing areas of depth images captured by depth sensors (e.g. reflection or transparent surfaces) (Chiu, Blanke, and Fritz 2011).

However, the cross-spectral stereo matching is still a challenging task especially without depth supervision (Zhi et al. 2018), because there are great illumination differences be-
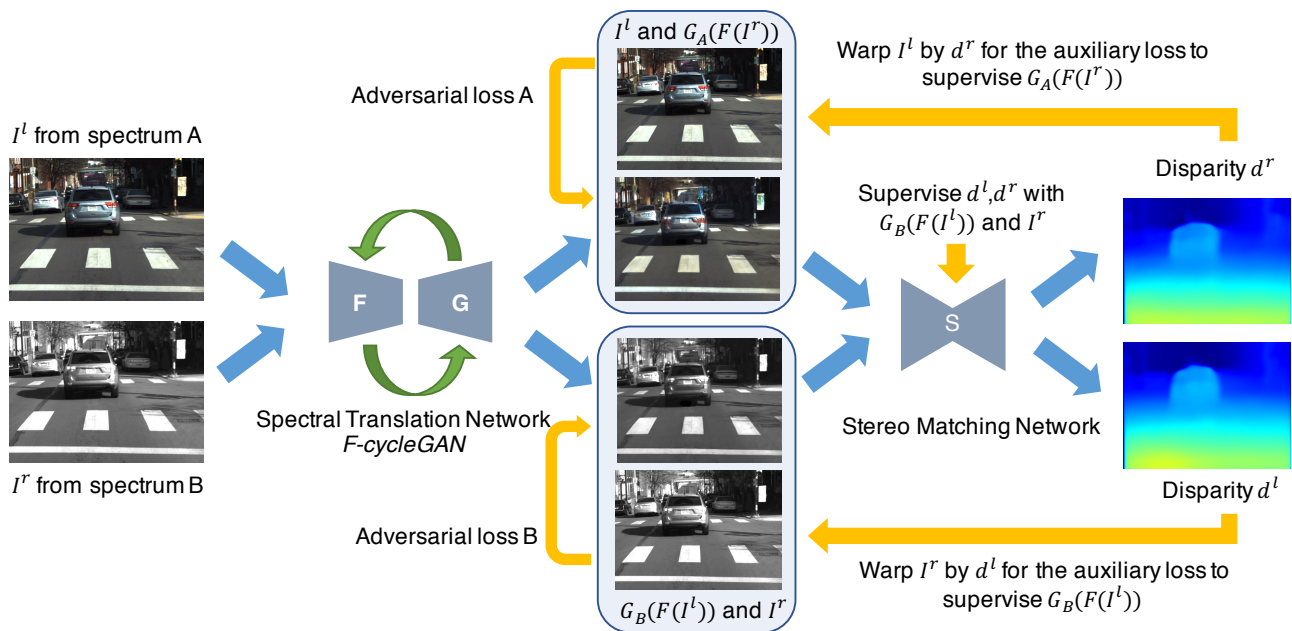
Figure 2: The structure of our proposed cross-spectral stereo matching framework. First, we input $\{I^l, I^r\}$ into the spectral translation network (STN) to get $G_B(F(I^l))$ and $G_A(F(l^r))$. The **F** network maps the images into hidden feature space X and **G** networks (G_A and G_B) convert feature X back into images of spectrum A and B. The stereo matching network (SMN) takes $I^{l,r}$, $G_B(F(I^l))$ and $G_A(F(l^r))$ to predict disparities $\{d^l, d^r\}$. By well-designed loss functions, we can train the STN and SMN jointly without depth ground-truth, extra-data or human-intervention.

tween images of different spectra. The translation between different spectra is quite complex and hard to accurately describe with a simple linear transformation. Figure 1 shows an example of cross-spectral (visible and near-infrared) stereo matching.

The key of traditional cross-spectral stereo matching is to design robust descriptors or features between the two modalities, such as ANCC (Heo, Lee, and Lee 2011) and DASC (Kim et al. 2015). However, these traditional methods are still not robust enough for transparent objects and large illumination variations. Zhi et al. (Zhi et al. 2018) proposed deep material-aware cross-spectral stereo matching, which tried to tackle the problem with deep neural networks and unsupervised learning. However, this method suffers from severe limitations: (i) The method requires additional semantic annotations to obtain auxiliary material information. (ii) The loss function is manually designed for different materials, which limits its applications to other scenarios.

To tackle the above problems, in this paper we employ image-to-image translation to assist cross-spectral stereo matching, and our full framework is shown in Figure 2. By regarding the difference between different spectral images as the different distributions, we explore the possibility of applying image-to-image translation methods to assist unsupervised cross-spectral stereo matching. We use two networks to transform images across different spectral bands and estimate disparity respectively. The first network is a spectral translation network (STN), which transforms images by cycle consistency and adversarial learning. The second network is a stereo matching network (SMN), which is trained with the image pairs transformed to the same spectrum by the spectral translation network. Then, we use the disparity predicted by the SMN to supervise the spectral translation network again. A novel share-encoder spectral translation network F-cycleGAN is employed to make the whole framework more robust.

Our contributions are as follows:
- We proposed a novel framework for cross-spectral stereo matching, which iteratively optimizes spectral translation network and stereo matching network.
- The proposed F-cycleGAN based on image-to-image translation and adversarial learning improves the robustness of image transformation.
- Our method surpasses state-of-the-art methods on cross-spectral stereo matching without depth supervision and extra human intervention.

Related Work

Unsupervised Depth Estimation

Garg et al. (Garg et al. 2016) first proposed to use warping-based view synthesis to learn disparity in an unsupervised way. The right image is first warped to the left view using disparity. Then, the absolute difference between the warped image and the left image, also called reconstruction error or photometric loss, is minimized to supervise disparity predictions. Godard et al. (Godard, Mac Aodha, and Brostow 2017) extended this idea by incorporating left-right consistency into the unsupervised loss. Zhou et al. (Zhou et al.

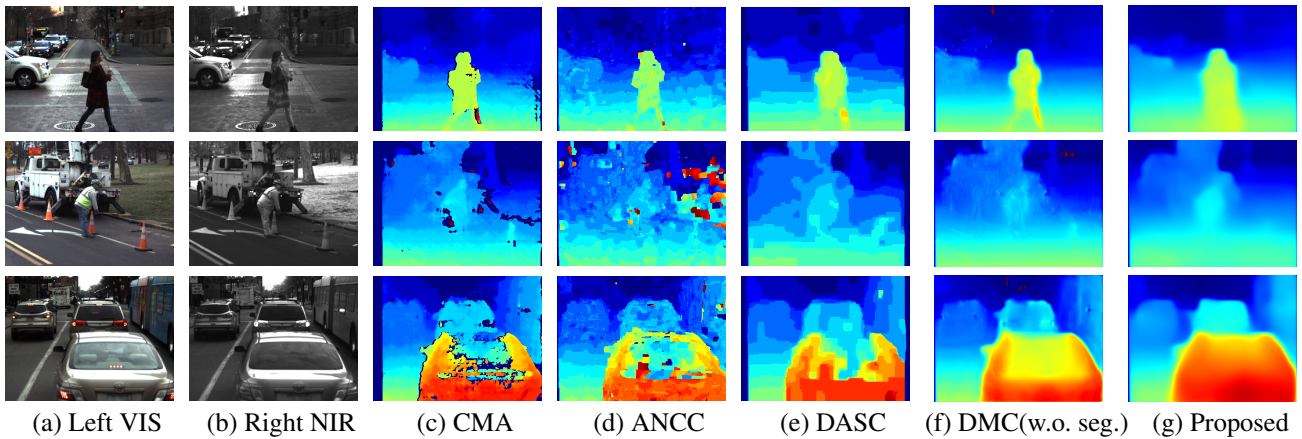(a) Left VIS   (b) Right NIR   (c) CMA   (d) ANCC   (e) DASC   (f) DMC(w.o. seg.)   (g) Proposed

Figure 3: Qualitative results on the evaluation dataset. The results of CMA , ANCC, DASC, DMC(w.o. seg.) are extracted from (Zhi et al. 2018). The proposed method performs well on the challenging materials like clothes (row 1,2), vegetation (row 2), lights (row 3). And comparing with others, the disparities predicted by the proposed method are cleaner and more reasonable.

2017b) proposed a framework which simultaneously predicted depth and frame-to-frame relative camera pose, which was trained with photometric loss using consecutive frames from videos. Zhou et al. (Zhou et al. 2017a) iteratively train a stereo network by filtering reliable predictions with left-right consistency check. However, unsupervised methods based on photometric loss often fail to predict accurate disparity for cross-spectral images due to the appearance differences.

Cross-spectral Stereo Matching

A series of robust matching costs were designed for radiometric variations. Mutual information (MI) measure (Egnal 2000) was extended by incorporating prior probabilities and 2D match surface (Fookes et al. 2004). Heo et al. (Heo, Lee, and Lee 2011) used color formation model explicitly and proposed Adaptive Normalized Cross-Correlation (ANCC) to tackle illumination changes and camera parameter differences. Local self-similarity (LSS) (Torabi and Bilodeau 2011) used window-based self similarity descriptor to do dense correspondence measure for thermal-visible videos. Pinggera1 et al. (Pinggera1, Breckon, and Bischof 2012) showed that dense gradient features based on HOG achieved better performance than MI and LSS descriptors. Aguilera et al. (Aguilera, Sappa, and Toledo 2015) proposed a feature descriptor for matching features points with nonlinear intensity variations. Kim et al. (Kim et al. 2015) proposed Dense adaptive self-correlation descriptor (DASC) by improving LSS descriptor with random receptive field pooling.

Another track of works tried to improve the quality of depth captured by RGBD cameras (Zhu et al. 2008; Chiu, Blanke, and Fritz 2011; de La Garanderie and Breckon 2014). Chiu et al. (Chiu, Blanke, and Fritz 2011) proposed a cross-modal adaptation method for cross-spectral stereo matching and fused predictions with depth captured with Kinect.

For deep learning methods, Aguilera et al. (Aguilera et al. 2016) learned a similarity measurement of cross-spectral image patches, which is a potential way to learn matching

cost for multi-spectrum images. Zhi et al. (Zhi et al. 2018) utilized deep segmentation maps to improve robustness of cross-spectral stereo matching, while the method required extra semantic annotations and manually designed losses for different materials, which made it hard to apply to other scenes.

Image-to-image Translation

Image-to-image translation converts images from one modality to another, such as style transfer (Gatys, Ecker, and Bethge 2016; Johnson, Alahi, and Fei-Fei 2016), colorization (Cheng, Yang, and Sheng 2015), sketch to image (Zhu et al. 2016; Sangkloy et al. 2017). For image translation, training with only L1 loss results in predictions lack of local semantic details. Johnson et al. (Johnson, Alahi, and Fei-Fei 2016) combined per-pixel loss with perceptual loss to train a fast feed-forward network for image transformation. Isola et al. (Isola et al. 2017) proposed a general framework for image-to-image translation using conditional adversarial networks. High-resolution synthetic images can be generated by applying multi-scale structure and novel adversarial loss (Wang et al. 2018). Later, Zhu et al. (Zhu et al. 2017) proposed CycleGAN to translate image styles with unaligned images from different domains.

We utilize the method of (Zhu et al. 2017) to convert images across different spectra, which is a basis of our proposed framework.

Method

In this section, we provide a detailed description of each part of the proposed method. Our network can be divided into two parts, the spectral translation network (STN) and stereo matching network (SMN). STN is responsible for minimizing the differences between domains, and SMN is responsible for predicting the disparity.
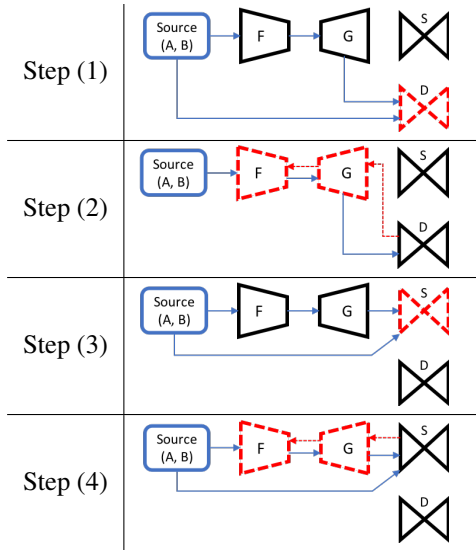
Figure 4: The steps of the iterative optimization. Here we omit subscript of spectrum cause the processing of the two modalities is identical. The F, G, S, D represent the F network, the generator network, the stereo matching network, and the adversarial discriminator. The *solid arrows* indicate the directions of data flow during the forward pass, while the *dotted arrows* represent the directions of gradient flow during the backward pass. The **red dotted** blocks are updated during the corresponding step and the **black solid** blocks are frozen.

Spectral Translation Network

The goal of the STN is to minimize the appearance variations between different spectra and provide the supervision information to the SMN. To achieve the goal, we proposed a novel style adaptation network F-cycleGAN as STN.

Given any image I_A of spectral A and image I_B of spectral B, we regard I_A and I_B sampled from two distributions A and B. We can define three mapping functions,

$$F : I_{A,B} \rightarrow X_{A,B},$$
$$G_A : X_{A,B} \rightarrow I_A,$$
$$G_B : X_{A,B} \rightarrow I_B,$$

where F encodes image $I_{A,B}$ to a unified feature space X. G_A and G_B are generators which convert features back into images in spectrum A and B respectively. In our implementation, we take the encoder and the decoder of the generator network in CycleGAN (Zhu et al. 2017) as the structure of our F network and $G_{A/B}$ networks.

The network F, G_A and G_B are supervised by adversarial losses (Goodfellow et al. 2014) and cycle-reconstruction loss. The adversarial loss is given by two discriminator networks D_A and D_B, which try to differentiate real and fake A or B images. We define $G_B(X_A)$ as I_B^{fake}, $G_A(X_B)$ as I_A^{fake}, $G_A(X_A)$ as I_A^{rec}, and $G_B(X_B)$ as I_B^{rec}.

The discriminator D_A aims to distinguish between I_A^{fake} and I_A. To train D_A, a classification loss $L_D^{adv,A}$ is used to classify I_A^{fake} and I_A. The loss for training discriminators is thus defined by

$$L_D^{adv} = L_D^{adv,A} + L_D^{adv,B}. \tag{1}$$

For generator networks, the loss can be mainly divided into two parts, adversarial loss and cycle consistency loss. The adversarial loss aims at fooling the discriminator networks and is given by

$$L_G^{adv} = L_G^{adv,A} + L_G^{adv,B}, \tag{2}$$

where $L_G^{adv,A}$ and $L_G^{adv,B}$ are achieved by maximizing the classification errors of discriminators D_A and D_B (details in (Zhu et al. 2017)). The cycle consistency loss is,

$$L_G^{cyc} = \frac{1}{N} \sum_{\mathbf{p} \in \Omega} \|I_A^{cyc}(\mathbf{p}) - I_A(\mathbf{p})\| \\ + \|I_B^{cyc}(\mathbf{p}) - I_B(\mathbf{p})\|, \tag{3}$$

where N is the number of pixels, I_A^{cyc} means $G_A(F(I_B^{fake}))$, I_B^{cyc} means $G_B(F(I_A^{fake}))$, and Ω is the pixel coordinate space. To guarantee the network F maps the images to the same hidden semantic feature space, and prevent the STN from learning disparity, a auxiliary reconstruction loss is introduced to supervise the network:

$$L_G^{rec} = \frac{1}{N} \sum_{\mathbf{p} \in \Omega} \|I_A^{rec}(\mathbf{p}) - I_A(\mathbf{p})\| \\ + \|I_B^{rec}(\mathbf{p}) - I_B(\mathbf{p})\|. \tag{4}$$

Then the final loss for the image transformation network F, G_A, G_B and the adversarial discriminator are given by

$$L_G = \lambda_c L_G^{cyc} + \lambda_r L_G^{rec} + \lambda_a L_G^{adv} \tag{5}$$
$$L_D = \lambda_d L_D^{adv} \tag{6}$$

To make the expressions clearer, all the intermediate outputs are summarized as follows,

$$I_A \xrightarrow{F} X_A \xrightarrow{G_B} I_B^{fake} \xrightarrow{F} X_B^{fake} \xrightarrow{G_A} I_A^{cyc} \tag{7}$$
$$I_B \xrightarrow{F} X_B \xrightarrow{G_A} I_A^{fake} \xrightarrow{F} X_A^{fake} \xrightarrow{G_B} I_B^{cyc} \tag{8}$$
$$X_A \xrightarrow{G_A} I_A^{rec}, X_B \xrightarrow{G_B} I_B^{rec} \tag{9}$$

Cross-spectral Stereo Matching Network

Dispnet (Mayer et al. 2016), which takes concatenated images as input to directly regress disparities, is adopted as the SMN to predict disparity maps d^l, d^r for left and right images. Given rectified cross-spectral image pair I_{ori}^l, I_{ori}^r, without loss of generality, we assume the spectrum of I_{ori}^l as spectrum A and I_{ori}^r as spectrum B. STN is applied to transform the cross-spectral images to the same modality. After that, we concatenate $\{I_{ori}^l, G_B(F(I_{ori}^l))\}$ as I^l and $\{G_A(F(I_{ori}^r)), I_{ori}^r\}$ as I^r to get the image pair in the same modality, which can be used as the input to the stereo matching network and for the cross-spectral unsupervised loss which will be discussed in the following section.

| (a) Left VIS | (b) Left fake NIR | (c) Right NIR | (d) Right fake VIS | (e) Disparity |

Figure 5: Qualitative results of our proposed approach

It should be emphasized that we block the gradients from network inputs back into STN for training stability. It should be noted that the forms of the image used for input and supervision are not required to be identical which will be discussed in benchmark results section. We apply the training loss from (Godard, Mac Aodha, and Brostow 2017) which includes appearance matching loss L_S^{ap}, disparity smoothness loss L_S^{ds}, and left-right disparity consistency loss L_S^{lr}. We only show the left terms, since the right can be derived similarly.

Based on the left disparity d^l, we can get reconstructed left image $\tilde{I}^l$ from I^r with the warping operator ω, which can be described as

$$\tilde{I}^l = \omega(I^r, d^l) \Longleftrightarrow \tilde{I}^l_{x,y} = I^r_{x+d^l_{x,y},y}. \qquad (10)$$

Since the disparity value might be a float number, $I^r_{x+d^l_{x,y},y}$ is bilinearly sampled at the pixel $(x+d^l_{x,y}, y)$. For simplicity, we use a mask to stop calculating the gradients for the pixels which are unable to be warped (e.g. pixels out of bound).

The appearance matching loss L_S^{ap} encourages the reconstructed image to appear similar to the original image by comparing structure and intensity. We let $\delta(I_1, I_2)$ be the structural similarity function (Wang et al. 2004) and $\tilde{I}^l$ be the reconstruction of I^l from I^r. The appearance matching loss can be described as

$$L_S^{ap,l} = \frac{1}{N}\sum_{\mathbf{p}\in\Omega}\alpha\frac{1-\delta(I^l, \tilde{I}^l)(\mathbf{p})}{2} + \\ (1-\alpha)\left\|I^l(\mathbf{p}) - \tilde{I}^l(\mathbf{p})\right\|, \qquad (11)$$

where α denotes the weight coefficient for the structural dissimilarity function and L1 reconstruction loss. The loss L_S^{ds} enforces the disparity smoothness,

$$L_S^{ds,l} = \frac{1}{N}\sum_{\mathbf{p}\in\Omega}(|\partial_x d^l|\, e^{-\|\partial_x I^l\|} + |\partial_y d^l|\, e^{-\|\partial_y I^l\|})(\mathbf{p}), \qquad (12)$$

where ∂d^l and ∂I^l means the gradients of d^l and I^l. The loss $L_S^{lr,l}$ regularizes the consistency of the left disparity and the right disparity,

$$L_S^{lr,l} = \frac{1}{N}\sum_{\mathbf{p}\in\Omega}|d^l(\mathbf{p}) - \omega(d^r, d^l)(\mathbf{p})|. \qquad (13)$$

Then the final loss for the SMN network is given by

$$L_{SMN} = \alpha_{ap}(L_S^{ap,l} + L_S^{ap,r}) + \alpha_{ds}(L_S^{ds,l} + L_S^{ds,r}) \\ + \alpha_{lr}(L_S^{lr,l} + L_S^{lr,r}). \qquad (14)$$

To further improve the performance, we introduce an auxiliary loss for the STN. First we can get the warped original images $I_{ori}^{\tilde{l}} = \omega(I_{ori}^r, d^l)$ and $I_{ori}^{\tilde{r}} = \omega(I_{ori}^l, d^r)$ with disparity prediction, then the auxiliary loss L_G^{aux} is defined by

$$
\begin{aligned}
L_G^{aux} = \alpha_{aux} \frac{1}{N} \sum_{\mathbf{p} \in \Omega} & \left\| G_B(F(I_{ori}^l))(\mathbf{p}) - I_{ori}^{\tilde{l}}(\mathbf{p}) \right\| \\
& + \left\| G_A(F(I_{ori}^r))(\mathbf{p}) - I_{ori}^{\tilde{r}}(\mathbf{p}) \right\|,
\end{aligned}
\tag{15}
$$

which attempts to tackle appearance variations and enhance the robustness of STN. There is a possibility that the reconstruction may encode both the disparity and spectral differences. We hold that by the cycle loss and reconstruction loss in Equ. 3 and Equ. 4, we can prevent the STN from learning disparity.

Iterative Optimization

We will introduce our iterative optimization approach in this section. All the losses required are presented in the Equ. 5, Equ. 6, Equ. 14, and Equ. 15. For simplicity, we omit subscript of spectrum for G_A, G_B, D_A, D_B because the optimization for the two modalities is identical.

Figure 4 shows the gradient flow across different network blocks. A randomly sampled cross-spectral image pair is provided to the entire system in each iteration. For the step (1), we train the D network by loss L_D from Equ. 6, which encourage the discriminator to distinguish between real and fake images. Then for the step (2), we train the F network and G network by loss L_G from Equ. 5. The stereo network S is trained in step (3) with the loss L_{SMN} from Equ. 14 by taking the translation results from G network as supervision. Finally, we use loss L_{aux} from Equ. 15 to train the F network and G network again for global optimization. The whole framework is first trained with several warmup epochs, using only step (1) and step (2), during which the stereo matching network is not trained. After the warmup stage, all four steps are used for further training.

Experiments

In this section, an evaluation of our method is performed on the benchmark dataset, and detailed analysis is given.

The network is trained on rectified cross-spectral stereo image pairs without any supervision in the form of ground truth disparity or depth. We evaluate on the PittsStereo-RGBNIR dataset proposed by (Zhi et al. 2018) which covers many material categories including lights, glass, glossy surfaces, vegetation, skin, clothing and bags. This dataset was captured by a visible (VIS) and near infrared (NIR) camera pairs. We define the left VIS as spectrum A and right NIR as spectrum B. The Left VIS consists of three spectral bands while the right NIR consists of only one band. For the simplicity of implementation, we convert NIR images into three channels.

Implementation Details

Architecture The G network and D network followed (Zhu et al. 2017) which has shown impressive results for image-to-image translation. The F network contains 4 residual blocks (He et al. 2016) and two stride-2 convolutions for down-sampling which is similar to the G network.

We used the DispNet (Mayer et al. 2016) as our stereo matching network SMN, and for the training stability, multi-scale predictions of SMN are applied following (Godard, Mac Aodha, and Brostow 2017). The weights of the STN were initialized from a Gaussian distribution with zero mean and 0.02 standard deviation, and the weights of the SMN were initialized with Kaiming initialization (He et al. 2015).

Parameters The SMN predicts the disparity directly instead of the ratio between disparity and image width. The disparity predictions are clamped to the range of zero to the image width. A scaling factor $\eta = 0.008$ is multiplied to the predictions for stable optimization.

The weights of the losses in STN are set to $\lambda_c = 10$, $\lambda_r = 5$, $\lambda_a = 1$, $\lambda_d = 1$, and the weights of losses in SMN are $\alpha_{ap} = 1$, $\alpha_{ds} = 0.2$, $\alpha_{lr} = 0.1$, $\alpha_{aux} = 20$. We use 5×5 window for calculating the structural similarity δ, and the α in Equ. 11 is set to 0.9.

Training and Testing The entire network contains about 54 million trainable parameters, of which 33 million parameters are in SMN. The dataset is split into two sets for training (40000 pairs) and testing (2000 pairs), which is the same as (Zhi et al. 2018). The STN and SMN are trained on 40000 cross-spectral image pairs with Adam optimizer (Kingma and Ba 2014) (batch size = 16 and learning rate = 0.0002).

For data augmentation, we flip the input images of STN horizontally with a 50% chance. Input images are resized into 512×384 for the entire network. We perform an instance normalization on the images provided to the SMN as input. The training process takes about 34 hours using 8 Nvidia TITAN Xp GPUs. The network is first trained with 15 warm-up epochs (with only step 1, 2, the SMN is not trained during this stage), and then trained with all 4 steps for 10 epochs.

For testing, the predicted disparity maps are bilinearly upsampled to the original size with disparity values multiplied with the horizontal scaling factor. 5030 sparse points on 2000 testing images are evaluated to compute the root mean square error.

Benchmark Results

For the sake of comparison, we choose the root mean square error (RMSE) as an indicator for our comparison, and we calculate the RMSE of each material category and obtain the average value as the final result **Mean**, following (Zhi et al. 2018). We have tested five network structure choices: *only SMN, STN+SMN, STN(F)+SMN, STN(F)+SMN(aux)(ori)*, and *STN(F)+SMN(aux)*.

For *only SMN*, STN is not employed and the cross-spectral image pairs are directly used as the unsupervised supervision. *STN+SMN* employs the original Cycle-GAN as the spectral translation network. For the *STN(F)* series, we use our proposed F-cycleGAN as the STN. *(aux)* represents using the auxiliary loss during training. The *STN(F) + SMN(aux)(ori)* means the original image pairs instead of

Table 1: Quantitative results. The RMSE of disparity for each material is evaluated. The RMSE results and execute times of CMA, ANCC, DASC, DMC(w.o. seg.), DMC(w. seg.) are extracted from (Zhi et al. 2018), where the DMC(w. seg.) means the method of (Zhi et al. 2018) with material-aware confidence. The proposed methods are tested on a single NVIDIA TITAN Xp GPU, which is the same as (Zhi et al. 2018). The network structure changes (row 7-10) lead to the improvement of performance.

Method	Common	Light	Glass	Glossy	Veg.	Skin	Clothing	Bag	Mean	Time(s)
CMA	1.60	5.17	2.55	3.86	4.42	3.39	6.42	4.63	4.00	227
ANCC	1.36	2.43	2.27	2.41	4.82	2.32	2.85	2.57	2.63	119
DASC	0.82	1.24	1.50	1.82	1.09	1.59	**0.80**	1.33	1.28	44.7
DMC(w.o. seg.)	**0.51**	1.08	1.05	1.57	0.69	**1.01**	1.22	0.90	1.00	0.02
DMC(w.seg.)	0.53	0.69	0.65	0.70	0.72	1.15	1.15	0.80	0.80	0.02
Only SMN	1.25	1.37	1.13	1.65	1.07	1.50	1.18	0.96	1.27	0.02
STN + SMN	1.13	1.55	1.05	1.52	0.89	1.23	1.14	0.98	1.18	0.04
STN(F) + SMN	1.24	1.02	0.92	1.32	0.79	1.10	1.03	0.92	1.04	0.04
STN(F) + SMN(aux)(ori)	0.75	0.86	**0.63**	1.05	0.81	1.16	0.99	**0.74**	0.87	0.02
STN(F) + SMN(aux)	0.68	**0.80**	0.67	**1.05**	**0.68**	1.04	0.98	0.80	**0.84**	0.04
Full Method	0.68	**0.80**	0.67	**1.05**	**0.68**	1.04	0.98	0.80	**0.84**	0.04

concatenated image pairs are used as inputs. All the methods with *STN* except *STN(F) + SMN(aux)(ori)* take concatenated original images and translated fake image pairs from STN as the inputs of SMN. We found that using only the NIR image and the fake NIR image in the unsupervised loss of SMN achieved better results, thus in all of our experiments, we employ only NIR images for the unsupervised supervision of SMN.

We compare the performance of the proposed method with other cross-spectral stereo matching methods like CMA(Chiu, Blanke, and Fritz 2011), ANCC(Heo, Lee, and Lee 2011), DASC(Kim et al. 2015), and DMC(Zhi et al. 2018). Table 1 presents the comparison with disparity RMSE and execution time. For DMC(Zhi et al. 2018), they incorporate material-aware confidence into the disparity prediction network, which requires semantic segmentation labels and manually defined loss for each kind of material. For fair comparison, we also list their results without the material-aware confidence.

On average, our approach outperforms other methods without extra human intervention. On lights, glass, glossy surface, and bag, our approach performs better than others. Table 1 also presents the changes in the results of our three comparative experiments, *STN+SMN*, *STN(F)+SMN*, and *STN(F)+SMN(aux)*. The results show that the F-cycleGAN and the framework for jointly training are able to improve the performance of unsupervised stereo matching. We also find that it is still hard to translate the appearance of clothing between VIS and NIR by the STN, possibly because the material of clothing is more variable than others, which leads to an unstable correspondence.

Visualization Results

Figure 5 presents the visualized results of the proposed method which suggests that the proposed approach is able to handle the illumination variations between different spectra. Comparing to other unsupervised methods in Figure 3, our method provides cleaner and more reasonable disparity predictions.

Conclusion

We have presented an unsupervised cross-spectral stereo matching method which can be trained in an end-to-end way without extra data or excessive human intervention. We propose F-cycleGAN based on the work of the (Zhu et al. 2017) as STN, which is able to minimize the appearance variations between different spectra without the loss of geometric information and improve the robustness of the stereo matching network SMN. Our experimental results show that our method outperforms other state-of-the-art methods. Our approach can be directly applied to other spectra, such as short-wave infrared or medium-wave infrared images.

In the future, we expect to further enhance the capabilities of the STN network for subtle visual differences. The structural similarity loss in the unsupervised loss of SMN, which is illumination sensitive, could also be improved to better supervise the stereo matching network.

References

Aguilera, C. A.; Aguilera, F. J.; Sappa, A. D.; Aguilera, C.; and Toledo, R. 2016. Learning cross-spectral similarity measures with deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1–9.

Aguilera, C. A.; Sappa, A. D.; and Toledo, R. 2015. Lghd: A feature descriptor for matching across non-linear intensity variations. In *Image Processing (ICIP), 2015 IEEE International Conference on*, 178–181. IEEE.

Cheng, Z.; Yang, Q.; and Sheng, B. 2015. Deep colorization. In *Proceedings of the IEEE International Conference on Computer Vision*, 415–423.

Chiu, W.-C.; Blanke, U.; and Fritz, M. 2011. Improving the kinect by cross-modal stereo. In *BMVC*, volume 1, 3. Citeseer.

de La Garanderie, G. P., and Breckon, T. P. 2014. Improved depth recovery in consumer depth cameras via disparity space fusion within cross-spectral stereo. In *BMVC*.

Egnal, G. 2000. Mutual information as a stereo correspondence measure.

Fookes, C.; Maeder, A.; Sridharan, S.; and Cook, J. 2004. Multi-spectral stereo image matching using mutual information. In *3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on*, 961–968. IEEE.

Garg, R.; BG, V. K.; Carneiro, G.; and Reid, I. 2016. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, 740–756. Springer.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2414–2423.

Godard, C.; Mac Aodha, O.; and Brostow, G. J. 2017. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, 7.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heo, Y. S.; Lee, K. M.; and Lee, S. U. 2011. Robust stereo matching using adaptive normalized cross-correlation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(4):807–822.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. *arXiv preprint*.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 694–711. Springer.

Kim, S.; Min, D.; Ham, B.; Ryu, S.; Do, M. N.; and Sohn, K. 2015. Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2103–2112.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lezama, J.; Qiu, Q.; and Sapiro, G. 2017. Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6807–6816. IEEE.

Mayer, N.; Ilg, E.; Hausser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; and Brox, T. 2016. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4040–4048.

Pinggera1, P.; Breckon, T.; and Bischof, H. 2012. On cross-spectral stereo matching using dense gradient features. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*.

Sangkloy, P.; Lu, J.; Fang, C.; Yu, F.; and Hays, J. 2017. Scribbler: Controlling deep image synthesis with sketch and color. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2.

Scharstein, D., and Szeliski, R. 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision* 47(1-3):7–42.

St-Charles, P.; Bilodeau, G.; and Bergevin, R. 2017. Mutual foreground segmentation with multispectral stereo pairs. In *Proc. Int. Conference on Computer Vision Workshops*, 375–384.

Torabi, A., and Bilodeau, G.-A. 2011. Local self-similarity as a dense stereo correspondence measure for themal-visible video registration. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, 61–67. IEEE.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13(4):600–612.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 5.

Xu, D.; Ouyang, W.; Ricci, E.; Wang, X.; and Sebe, N. 2017. Learning cross-modal deep representations for robust pedestrian detection. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Zhi, T.; Pires, B. R.; Hebert, M.; and Narasimhan, S. G. 2018. Deep material-aware cross-spectral stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1916–1925.

Zhou, C.; Zhang, H.; Shen, X.; and Jia, J. 2017a. Unsupervised learning of stereo matching. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2.

Zhou, T.; Brown, M.; Snavely, N.; and Lowe, D. G. 2017b. Unsupervised learning of depth and ego-motion from video. In *CVPR*, volume 2, 7.

Zhu, J.; Wang, L.; Yang, R.; and Davis, J. 2008. Fusion of time-of-flight depth and stereo for high accuracy depth maps. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8. IEEE.

Zhu, J.-Y.; Krähenbühl, P.; Shechtman, E.; and Efros, A. A. 2016. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*, 597–613. Springer.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*.