Different Modal Stereo: Simultaneous Estimation of Stereo Image Disparity and Modality Translation

Ryota Tanaka, Fumihiko Sakaue and Jun Sato

Nagoya Institute of Technology, Gokiso Showa, Naogya, 466-8555, Japan

Keywords: Multi-modal Images, Stereo Reconstruction, Deep Neural Networks, Modal Translation.

Abstract: We propose a stereo matching method from the different modal image pairs. In this method, input images are taken in different viewpoints by different modal cameras, e.g., an RGB camera and an IR camera. Our proposed method estimates the disparity of the two images and translates the modality of the input images to different modality simultaneously. To achieve this simultaneous estimation, we utilize two networks, i.e., a disparity estimation method from a single image and modality translation method. Both methods are based on the neural networks, and then we train the network simultaneously. In this training, we focus on several consistencies between the different modal images. By these consistencies, two kinds of networks are effectively trained. Furthermore, we utilize image synthesis optimization on conditional GAN, and the optimization provides quite good results. Several experimental results by open databases show that the proposed method can estimate disparity and translate the modality even if the modalities of the input image pair are different.

1 INTRODUCTION

3D reconstruction using a stereo camera system is one of the most important techniques in the field of computer vision, and then, various kinds of methods are studied extensively. In these methods, feature points such as SIFT are detected at first. After that, a corresponding point is determined in the other image. These techniques are based on the assumption that image feature points in the image pair are similar to each other. Thus, in the stereo camera systems, the same modality, representatively RGB, image pair is used.

On the other hand, various modal cameras, as shown in Fig.1 are often used for specific purposes in recent years. For example, IR (infrared) cameras are used for such as surveillance systems and driving assist systems at night with IR light sources. Since the human eyes can not observe IR light, these systems do not disturb human visual systems. Also, FIR (far infrared) thermal camera is often utilized for night surveillance since most of the humans and animals emit FIR light by body temperature. In these surveillance systems, not only these unique cameras but also conventional RGB cameras are combinedly used since RGB images are familiar to the human visual systems. In addition, traditional computer vision techniques are for the RGB images, and thus the RGB images are required to apply these techniques.

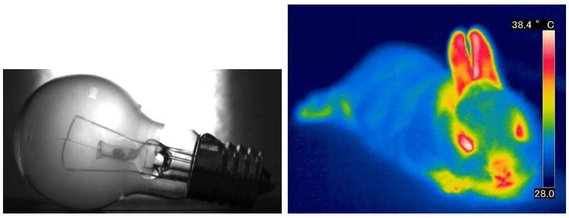
Furthermore, these systems often require 3D information in the scene. However, conventional stereo camera systems require the same modality cameras, and then additional cameras are necessary to construct the stereo system in existing techniques.

In this paper, we propose a stereo matching method from different modal image pairs. Our proposed method is based on two techniques. One of them is the modality translation using a neural network. Although different modal image pairs taken in the same viewpoint by the different modal cameras are required to train the deep neural networks in existing methods, our method achieves training of the network using image set taken from different viewpoints. The other one is a disparity estimation from a single image. In this technique, different modal images assist the estimation of disparities. By using these techniques, we achieve a stereo matching and modality translation simultaneously from a different modal image pair.

2 RELATED WORKS

Recently, techniques for multi-modal imaging are widely studied. Especially, multi-spectral imaging is one of the most important techniques because of its wide applications. In general, light spectral informa-

Tanaka, R., Sakaue, F. and Sato, J.
Different Modal Stereo: Simultaneous Estimation of Stereo Image Disparity and Modality Translation.
DOI: 10.5220/0009180205540560
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 4: VISAPP*, pages 554-560
ISBN: 978-989-758-402-2; ISSN: 2184-4321



(a) IR image (b) Thermal image

Figure 1: Different modal images taken by special cameras.

tion is taken by a special multispectral camera with a long capturing time. Thus, the imaging by this camera cannot be applied to dynamic scenes. To avoid the problem, Kiku et al.(Kiku et al., 2014; Monno et al., 2014) propose a special camera that equipped a special Bayer filter. The special Bayer pattern in this method filter different band light pixel by pixel, and then, several spectral information can be captured simultaneously by a single camera. Since the captured spectral information by the filter is sparse, the information is interpolated to dense information by image demosaicing technique using a guide image. This kind of special Bayer pattern can be applied to not only multi-spectral imaging but also HDR imaging, IR imaging and so on(Raskar et al., 2006; Levin et al., 2008; Fergus et al., 2006). Although these methods can capture multi-modal information from dynamic scenes, special and expensive camera systems are required.

Stereo matching of different modal images is also studied since the stereo matching is one of the essential problems in the field of computer vision. Zbontar et al.(Zbontar and LeCun, 2016) propose a stereo matching method based on similarity learning using CNN. Although the method can be applied multi-modal image pair as well as the same modal image pair, the method may not work well when the modalities of input images are much different, such as thermal and RGB(Treible et al., 2017).

In our approach, we do not require correspondences between different modal images to train the network. This is large advantage because our method require just stereo camera pair in different modals to train the network. In addition, our method can apply to greatly different modal image pairs such as RGB images and thermal images. Thus, applicable field can become wider. Furthermore, our method estimates not only the disparity of the image pair but also modality translated images from each viewpoint. Therefore, a multi-modal image from a single viewpoint can be virtually obtained from the different viewpoints camera pair without any special devices. In the following sections, the detail of our proposed method is explained.

3 NETWORKS FOR DISPARITY ESTIMATION AND MODALITY TRANSLATION

3.1 Disparity Estimation from Single Image

As described in the previous section, our method utilizes two methods using deep neural networks. In this section, we summarize these methods before a detail explanation of our proposed method.

We first explain the disparity estimation from a single image(Luo et al., 2018). In this method, disparity maps are predicted from the texture information of the input images. As larger A large advantage of the method is that this method does not require correct disparity data to train the disparity estimation network. This method requires only a set of stereo image pairs to train the network. Let I_r and I_l denote the image taken by the right camera and the left camera, respectively. The disparity map $V(i,j)$ shows a disparity at point (i,j) in the left image. In this case, the left image can be translated to the right image $\tilde{I}_r$ as follows:

$$\tilde{I}_r(i,j) = I_l((i,j)+V(i,j)) \tag{1}$$

When the disparity map V is correct, the estimated right image close to the real right image I_r. Therefore, the disparity map can be evaluated as follows:

$$\varepsilon_V = \|\tilde{I}_r - I_r\| \tag{2}$$

By minimizing the ε_V, an appropriate disparity map can be estimated.

To estimate the disparity map V efficiently, not a direct disparity map V but probability maps $V_d(I_l)$ is estimated from the image. The probability map $V^d(i,j)$ denote a probability that the disparity at (i,j) is d. When the $I_l^{(d)}$ denotes the shifted image $I((i,j)+d)$, the viewpoint changed image $\tilde{I}_l^d$ can be estimated as follows:

$$\tilde{I}'_r = \sum_d I_l^{(d)} V^d(I_l) \tag{3}$$

This equation compute expected values of each point. The loss function can be defined as follows:

$$\varepsilon'_V = \|\tilde{I}'_r - I_r\| \tag{4}$$

By minimizing the loss ε'_V, an appropriate estimation function of the V^d can be estimated. Details of the network architecture and this method are in their paper(Luo et al., 2018).

Note that although the method does not require the correct disparity, modalities of the images should be the same since the technique uses the similarity of the image pair.

3.2 Modality Translation using Conditional GAN

We next consider the modality translation of the input image. In this translation, we utilize a framework of the conditional GAN to synthesize different modal image. Especially, pix2pix(Isola et al., 2017) provides effective image translation in a framework of conditional GAN (cGAN), and we utilize the pix2pix in our research.

The pix2pix is one of the representative networks to synthesize different characteristic images from the input image. Several image translation examples, such as an edge image to a color image, are reported based on the framework. In this method, a generator network and a discriminator network is used. The generator network synthesizes the translated image, and the discriminator network decides the image synthesized by the generator is valid or not. By training these two adversarial networks simultaneously, the generator accomplishes pretty useful image synthesis.

An objective function of the cGAN can be expressed as follows:

$$\begin{aligned}\mathcal{L}_{cGAN}(G,D) \quad = \quad & \mathbb{E}_{x,y}[\log D(x,y)] \\ & +\mathbb{E}_{x,z}[\log(1-D(x,G(x,z))]\end{aligned} \tag{5}$$

where G is a generator and the generator synthesize the image from the noise z under the condition x . $D(x,y)$ is a discriminator and it provides a probability that y can be translated from y. In addition, the pix2pix uses a similarity between objective image y and a synthesized image $G(x,z)$ as follows:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y-G(x,z)\|] \tag{6}$$

From these, an optimized generator is computed as follows:

$$G^* = \arg\min_G \min_D \mathcal{L}_{cGAN}(G,D) + \lambda \mathcal{L}_{L1}(G) \tag{7}$$

Network architecture is in their paper(Isola et al., 2017).

Note that the pix2pix requires a set of corresponding image pairs. For example, when thermal images are translated to the RGB image, a thermal image set and an RGB image set taken the same scenes from the same viewpoints are required. In our research, the viewpoints of the cameras are different, and then, we cannot utilize the pix2pix directly in our method.

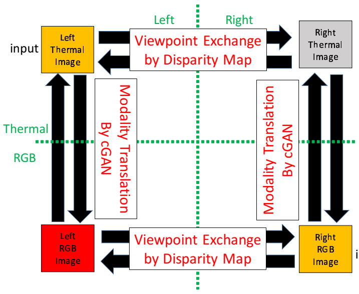

Left
Right
input
Left Thermal Image
Viewpoint Exchange by Disparity Map
Right Thermal Image
Thermal
RGB
Modality Translation By cGAN
Modality Translation By cGAN
Left RGB Image
Viewpoint Exchange by Disparity Map
Right RGB Image
input

Figure 2: Overview of our proposed network. All networks are trained simultaneously based on several loss functions.

4 SIMULTANEOUS ESTIMATION OF IMAGE DISPARITY AND MODALITY TRANSLATION

4.1 Overview

Let us explain our proposed method in this section. As described in **1**, we have two images that have different modalities and taken from different viewpoints. In this section, we consider the case when a thermal image I_l^T is taken from the left camera, and an RGB image I_r^C is taken from the right camera. From these images, we synthesize a viewpoint changed of modality translated images $\tilde{I}_r^C$ and $\tilde{I}_l^T$, and estimate a disparity map D.

Figure 2 shows an overview of our proposed method. As shown in this figure, we utilize a disparity estimation method shown in **3.1** and image translation method shown in **3.2**. However, these networks cannot be trained directly in our environment. Therefore, we define several losses to train the networks and estimate results. We explain the losses in the following sections.

4.2 cGAN Loss

We first define a cGAN loss. This loss evaluates that the translated image is valid or not by discriminators for cGAN. Let $V_l(I_l)$ and $V_r(I_r)$ denote viewpoint changing by the disparity map D. And let $G^C(I^C,z)$ and $G^T(I^T,z')$ denote modality translation from RGB images to thermal images and thermal images to RGB images. Discriminators $D(I_l^C,I_r^T)$ and $D(I_r^T,I_l^C)$ compute probability that the images I_l^C and I_r^T are correspond to the images I_r^T and I_l^C, respectively. By using the functions, The cGAN loss is defined as follows:

$$\begin{aligned}\mathcal{L}_c^{lr1} \quad = \quad & \mathbb{E}[\log D(I_l^T,I_r^C)] \\ & +\mathbb{E}[\log(1-D(G^T(V(I_l^T),z)))] \\ & +\mathbb{E}[\|I_r^C-G^T(V(I_l^T),z)\|_1]\end{aligned} \tag{8}$$

$$\begin{aligned} \mathcal{L}_c^{lr2} &= \mathbb{E}[\log D(I_l^T, I_r^C)] \\ &\quad +\mathbb{E}[\log(1 - D(V(G(I_l^T, z))))] \\ &\quad +\mathbb{E}[\|I_r^C - V(G(I_l^T, z))\|_1] \end{aligned} \tag{9}$$

In these losses, a left image is translated to a right image with the modality translation. In $\mathcal{L}^{lr1}$, the modality of the I_l^T is translated at first. After that, the viewpoint is changed from the left to the right. In $\mathcal{L}^{lr2}$, the order of the image transformation is reversed. As same as the left to right translation, losses for the right to left is defined as $\mathcal{L}^{rl1}$ and $\mathcal{L}^{rl2}$. These losses are very similar to the loss of the cGAN used in pix2pix because discriminator decides the translated image is valid or not based on the training images. A difference from the pix2pix is that the generator includes viewpoint exchange as well as modality translation.

4.3 Image Consistency Loss

We next define an image consistency loss. In this loss, we assume that the translated images should be the same even if exchanging order is different. Therefore, we define the loss as follows:

$$\mathcal{L}_i^{lr} = \mathbb{E}[\|V(G(I_r^T, z)) - G^T(V(I_r^T), z)\|_1] \tag{10}$$

$\mathcal{L}_i^{rl}$ for the right to left translation is also defined in this loss.

Furthermore, we assume that synthesized images in the same modality and the same viewpoint should be the same even if the original images are different. For example, $\tilde{I}_r^T$ from the I_l^T by viewpoint exchange and $\tilde{I}_r^T$ from the I_r^C by the modality translation should be the same under this assumption. Thus, the second image consistency loss is defined as follows:

$$\begin{aligned} \mathcal{L}_{i2} &= \mathbb{E}[\|V(I_l^T) - G^C(I_r^C, z)\|_1] \\ &\quad +\mathbb{E}[\|V(I_r^C) - G^C(I_l^T, z)\|_1] \end{aligned} \tag{11}$$

4.4 Cycle Loss

We next define a cycle loss. This loss is based on the loss of the cycle GAN(Zhu et al., 2017). This loss focus on the similarity between an original image and a synthesized image by a combination of translation and inversed translation. In the computation of this loss, an input image is translated to the different image at first, and back to the original domain by inversed translation. By this combined translation, the input image should be back to the original image if the translation is valid. That is, the loss can be defined as follows:

$$\begin{aligned} \mathcal{L}_c^T &= \mathbb{E}[\log(D^C(I_r^c))] \\ &\quad +\mathbb{E}[\log(1 - D^C(G^T(I_r^T, z)))] \\ &\quad +\mathbb{E}[\|I_l^T - G^C(G^T(I_l^T, z), z)\|_1] \end{aligned} \tag{12}$$

where D^C denotes a discriminator computing a probability that the image is the RGB image or not. This loss is combination of the original GAN and cycle image consistency.

4.5 Attention Loss

As described above, our networks for estimating the disparity map and modality translation are trained simultaneously using a training dataset that is taken in different modalities and different viewpoints. However, networks often cannot be trained appropriately. This is because the image translation networks G have a significant redundancy. Therefore, the network G often includes not only modality translation but also viewpoint changing even if the losses shown in the previous sections are minimized. If the networks G include the viewpoint exchange V, we cannot estimate the disparity appropriately.

To avoid this excessive image translation, we define the attention loss for modality translation. For the attention loss, we assume that an image region that includes essential information is the same, even if the modality of the image is changed. Therefore, we extract a degree of attention by using Grad-CAM(Selvaraju et al., 2017) and evaluate the image translation based on the attentions. The difference between the Grad-CAM result from each image is used as the attention loss. Therefore, attention loss is defined as follows:

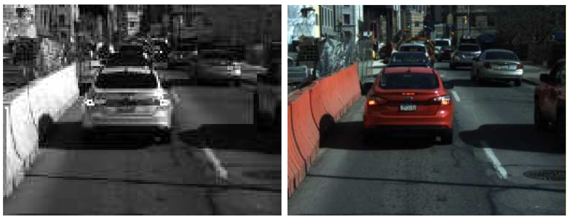
$$\begin{aligned} \mathcal{L}_{\dashv} &= \mathbb{E}[\|A(I_l^T) - A(V(I_r^C))\|_1] \\ &\quad \mathbb{E}[\|A(l_r^C) - A(V(I_l^T)\|_1] \end{aligned} \tag{13}$$

where A denote attention extraction by Grad-CAM. As this equation indicates, the attention loss does not include the image translation G, and thus the disparity exchange V and the image translation G can be separated by minimizing the attention loss.

By minimizing all losses mentioned above, disparity map estimation and image translation are accomplished simultaneously.

5 IMAGE ESTIMATION USING THE DEEP NEURAL NETWORKS

We last describe image synthesis techniques using the deep neural networks trained in the previous sections. In fact, the disparity map and modality translation results are not determined uniquely by the described networks. Because the generator G requires not only an input image but also noise in a latent space as input



(a) IR image (b) RGB image

Figure 3: Examples of input IR and RGB image pair.

data. Thus, the estimated result can changes according to the noise z in the latent space.

Therefore, we optimize the noise z in the latent space to synthesize an appropriate image. In this optimization, the losses used for network training can be used because the losses represent consistencies of the estimated data. Therefore, we compute the optimized noise z^* which minimizes all the losses with fixed networks as follows:

$$z^* = \arg\min \sum_i \mathcal{L}_i \tag{14}$$

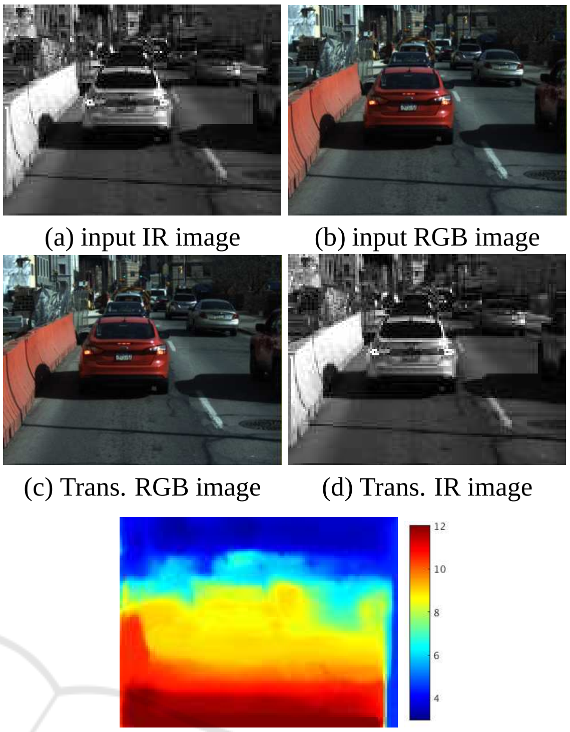
where $\mathcal{L}_i$ denote a loss explained in the previous section. From the optimized noise z^* and input images, a disparity map and modality translated images are synthesized. Thus, we accomplish the disparity estimation and the modality translation from the different modal and different viewpoint image pairs.

6 EXPERIMENTAL RESULTS

6.1 IR Image and RGB Image

In this section, we show several experimental results. We first show the results when the input image pair is an IR image and an RGB image. In this experiment, we utilized the PittsStereo-RGBNIR dataset(Zhi et al., 2018) for training and testing. This dataset includes stereo pairs of IR images and RGB images Size of these images is 582 × 429. These images were rectified, and then the epipolar line in the images are parallelized to the horizontal axis. By using the rectified images, a disparity map and modality translated images, i.e., from IR to RGB and RGB to IR, are estimated by our proposed method. Examples of rectified input images are shown in Fig.3

Figure 4 and 5 shows the estimated results by our proposed method. In this figure, images (a) and (b) shows input images, and (c) and (d) show modality translated results by our proposed method. In this figure, the viewpoints of the images in the same column are the same. The image (e) shows the estimated disparity. In the disparity image, colors in each pixel show the value of the disparity, and the values are colorized by the right color bar.



(a) input IR image (b) input RGB image

(c) Trans. RGB image (d) Trans. IR image

(e) Estimated disparity

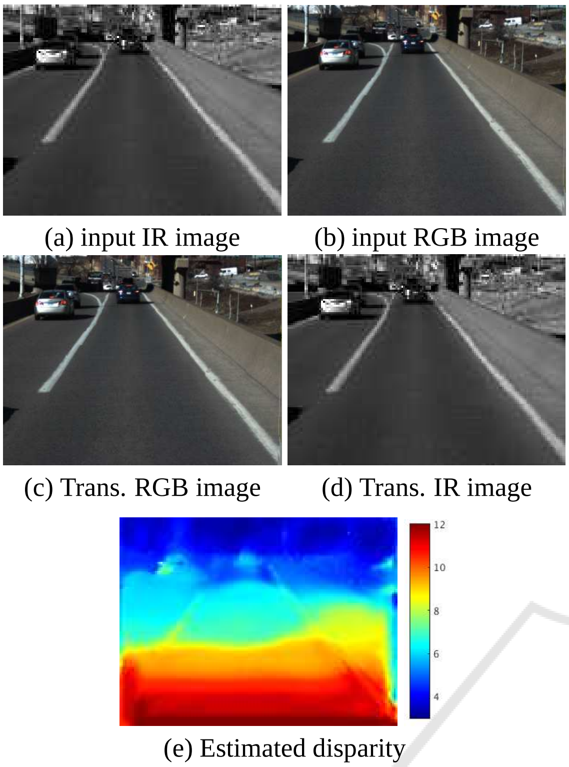
Figure 4: Input images and estimated results: (a) and (b) are input images, and (c) and (d) are translated results. The images in the same column were (virtually) taken in the same viewpoint. An image (e) shows estimated disparity by color indicated in the right color bar.

In both results, the modality translated images are pretty good. Positions of the objects in the input image and the translated image are much similar. Furthermore, the translated images are very natural, and we could not discriminate which image was the translated image by ourselves. Also, the disparity maps gradually change according to the depth of the image. In these images, the upper region of the image has far depth, and the lower region has close depth. The estimated disparity represents the change of this disparity. Furthermore, the depth of the region, including the vehicle is different from the other region, and it indicates that the depth changes by the vehicle.

These results indicate that our proposed method can estimate the disparity maps and modality translated images from the IR and the RGB image pair.

6.2 Results from Thermal Image and RGB Image

We next show the results from thermal images and


(a) input IR image (b) input RGB image

(c) Trans. RGB image (d) Trans. IR image

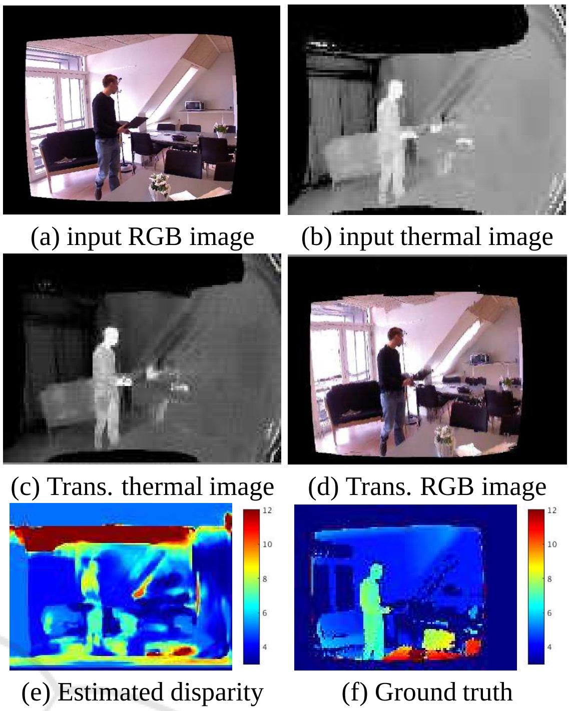
(e) Estimated disparity

Figure 5: The other input images and estimated results.

RGB images. In this experiment, we utilized VAP Trimodal people segmentation dataset(Treible et al., 2017). The dataset includes pair of thermal images and RGB images with a disparity of them. Size of images is 680×480. Examples of the input images are shown in the Fig.6(a), (b).

From these results, we estimated the modality translated images and the disparity maps. Figure 6 and 7 shows input images and estimated results. As same as the Fig 4 and 5 image (a) $\sim$ (e) shows input images and estimated results. The image (f) shows the ground truth of the disparity map.

In these results, modality translation makes clear images in both the thermal to RGB and RGB to thermal. The brightness of the estimated thermal images is slightly different from the input image. We consider that the difference was occurred by the difference of the intensity in each input image. However, the difference can be suppressed easily by ordinary image processing techniques such as brightness normalization. In the disparity image, several regions have different disparity from the ground truth. In this region, textures of the image are very slight, and it might occur open aperture problems. However, disparities could be estimated in textured regions and edges of the objects, and then, our method can estimate the disparity if the input image provides sufficient information.


(a) input RGB image (b) input thermal image

(c) Trans. thermal image (d) Trans. RGB image

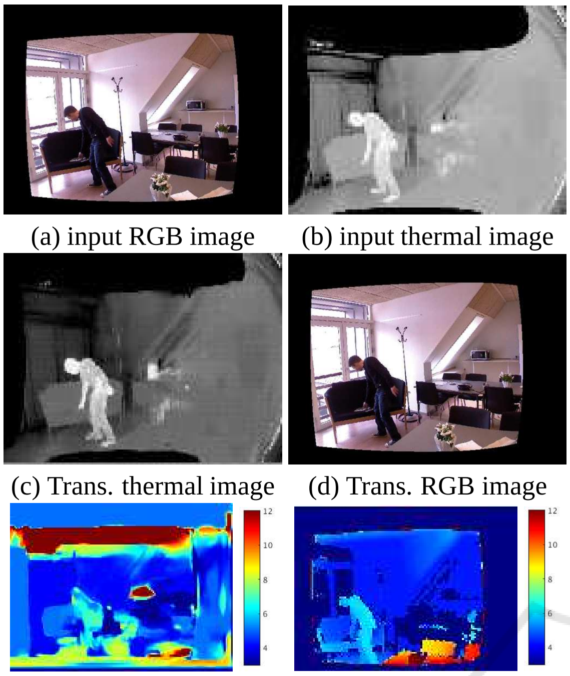
(e) Estimated disparity (f) Ground truth

Figure 6: Input images and estimated results: (a) and (b) are input images, and (c) and (d) are translated results. Images (e) and (f) shows estimated disparity and ground truth, respectively. RMSE between the estimated disparity and the ground truth was 3.22[pix].

RMSE of the two estimated results were 3.22[pix] and 3.25[pix]. These values are relatively small to the whole image disparity. These results indicate that our proposed method can accomplish simultaneous computation of modality translation and disparity estimation even if the modality of the input image pair is drastically different.

7 CONCLUSION

In this paper, we propose simultaneous computation of image modality translation and disparity estimation from the different modal images based on the neural networks. In this method, we utilize two kinds of networks. The first one is for estimating the disparity, and the second one is for modality translation. We define several losses to optimize the network simultaneously and appropriately. In addition, we propose the optimal image synthesis technique by minimizing the loss functions on the GAN. Several experimental results show that our proposed method can translate the modality of images and estimate disparity from the different modal image pairs. Notably, even if the modalities of the image pair are far from



(a) input RGB image (b) input thermal image

(c) Trans. thermal image (d) Trans. RGB image

(e) Estimated disparity (f) Ground truth

Figure 7: Input images and estimated results. RMSE between the estimated disparity and ground truth was 3.25[pix].

each other such as RGB and thermal, our method can estimate modality translated image and disparity. Although the qualitative evaluation of our proposed method is good, it is not enough because we cannot use a sufficient number of images, including ground truth. Thus, we construct the database for evaluation and evaluate our proposed method extensively in future work.

REFERENCES

Fergus, R., Singh, B., Hertzmann, A., Roweis, S. T., and Freeman, W. T. (2006). Removing camera shake from a single photograph. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 787–794. ACM.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. (2017). Image-to-image translation with conditional adversarial networks. In *Proc. CVPR2017*.

Kiku, D., Monno, Y., Tanaka, M., and Okutomi, M. (2014). Simultaneous capturing of rgb and additional band images using hybrid color filter array. In *Digital Photography X*, volume 9023, page 90230V. International Society for Optics and Photonics.

Levin, A., Sand, P., Cho, T. S., Durand, F., and Freeman, W. T. (2008). Motion-invariant photography. In *ACM Transactions on Graphics (TOG)*, volume 27. ACM.

Luo, Y., Ren, J., Lin, M., Pang, H., Sun, W., Li, H., and Lin, L. (2018). Single view stereo matching. In *Proc. CVPR2018*, pages 155 – 163.

Monno, Y., Kiku, D., Kikuchi, S., Tanaka, M., and Okutomi, M. (2014). Multispectral demosaicking with novel guide image generation and residual interpolation. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 645–649. IEEE.

Raskar, R., Agrawal, A., and Tumblin, J. (2006). Coded exposure photography: motion deblurring using fluttered shutter. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 795–804. ACM.

Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localizatio. In *Proc. ICCV2017*.

Treible, W., Saponaro, P., Sorensen, S., Kolagunda, A., ONeal, M., Phelan, B., Sherbondy, K., and Kambhamettu, C. (2017). Cats: A color and thermal stereo benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2961–2969.

Zbontar, J. and LeCun, Y. (2016). Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(1-32):2.

Zhi, T., Pires, B., Hebert, M., and Narasimhan, S. (2018). Deep material-aware cross-spectral stereo matching. In *Proc. CVPR2018*.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. ICCV2017*.