An Alternative of LiDAR in Nighttime: Unsupervised Depth Estimation Based on Single Thermal Image

Yawen Lu, Guoyu Lu
Intelligent Vision and Sensing Lab, Rochester Institute of Technology

Abstract

Most existing autonomous driving vehicles and robots rely on active LiDAR sensors to detect the depth of the surrounding environment, which usually has limited resolution, and the emitted laser can be harmful to people and the environment. Current passive image-based depth estimation algorithms focus on color images from RGB sensors, which is not suitable for dark and night environment with limited lighting resource. In this paper, we propose a framework to estimate the scene depth directly from a single thermal image that can still observe the scene in the low lighting condition. We learn the thermal image depth estimation framework together with RGB cameras, which also mitigates the training condition due to the easy availability of RGB cameras. With the translated thermal images from color images from our generative adversarial network, our depth estimation method can explore the unique characteristics in thermal images through our novel contour and edge-aware constraints to obtain a stable and anti-artifact disparity. We apply the commonly available color cameras to navigate the learning process of thermal image depth estimation framework. With our approach, an accurate depth map can be predicted without any prior knowledge under various illumination conditions.

1. Introduction

Depth estimation is a fundamental task in computer vision and is essential for many real-world applications such as autonomous vehicles and UAV navigation. It has a close relationship with many other fields in 2D and 3D tasks such as semantic segmentation, 3D object recognition, visual odometry and SLAM, objection detection and scene reconstruction. Previous developments built on either synthetic [4] [43] or real-world datasets [8] [14], have contributed to the advancement of this filed. However, few efforts are put into the depth estimation from stereo pairs under nighttime scenarios. Traditional works such as in [21] [56] heavily depend on brightness and gradient constancy assumptions, which are not available at nighttime situations. Convolutional Neural Networks (CNN) have been successfully applied to estimate the scene depth from RGB images[10] [31] [32] [51] [29]. However, these methods mainly focus on supervised learning methods, which require a large amount of ground truth depth labels from either LiDAR or other depth sensors. Moreover, these labels are expensive to obtain and involve extensive human efforts in real-world applications. Therefore, research has been conducted to explore various constraints to realize semi-supervised [1] [35] [27] or unsupervised setting [13] [15] [55] to estimate the depth or disparity of a scene, which are mainly restricted in proper illumination conditions. Existing nighttime depth sensor mainly rely on LiDAR and other active illumination sensors, which usually have small resolution and may harm the environment due to the emitted laser beams.

In this paper, we explore to estimate the scene depth by a single thermal image to deal with the low illumination issue, especially night vision problems. To learn the single thermal image depth estimation, we first apply the disparity learned from the color image as the initialization for the training procedure of the thermal image depth estimation. As the thermal camera and color camera maintain a baseline distance, the disparity still retains errors, which we will rely on our neural network and loss constraints specifically designed for thermal-color consistency to eliminate. For multi-spectral images, the stereo image pair from left and right cameras has an obvious difference in spectral distribution and appearance variation, for which multi-spectral stereo matching is challenging. To solve this issue, we propose to utilize an image-to-image translation network to transform visible color images to long-wave infrared images to mitigate the obvious differences between them. Then the accurate disparity is estimated without any prior knowledge and supervision. Considering the characteristics of thermal images, we incorporate the illumination consistency and contrast information to make sure that the estimated disparity maps preserve more precise and stable intensity as the real thermal images.

To summarize, the key contributions are as follows: 1) We propose a novel single thermal image depth estimation framework, which utilizes visible color images to help train thermal image depth estimation under dark or night-

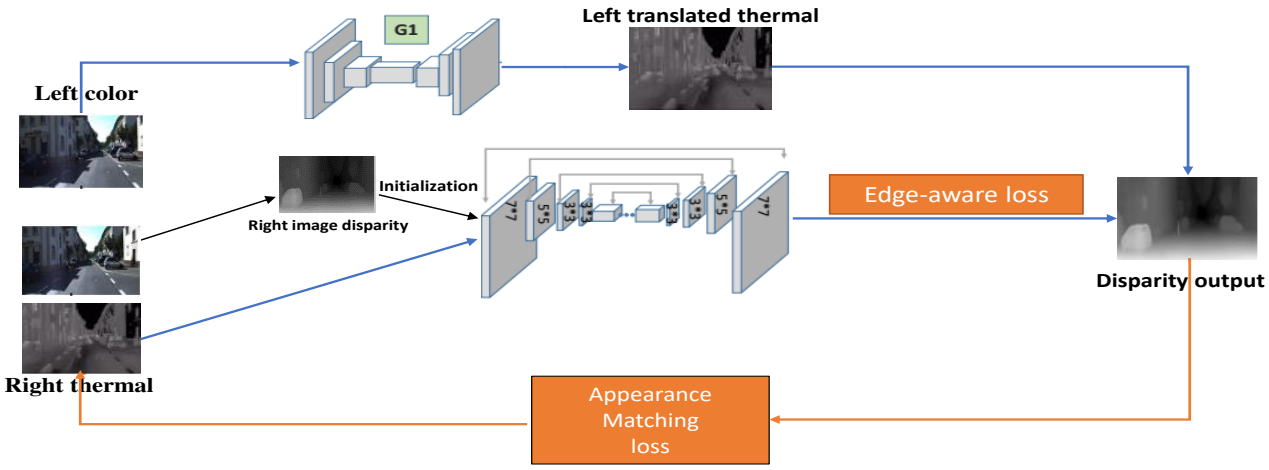

Figure 1: Basic single thermal image depth estimation framework. Translated thermal image from left color image is used together with the estimated disparity to recover the right thermal view. During the inference, only single thermal image input is required to generate the estimated disparity.

time environment; 2) The designed image translation network for RGB-thermal pair enables pixel matching between multi-spectral images that have apparent large discrepancy in both spectral distribution and visual variation; 3) The entire pipeline is trained under unsupervised learning setting, which does not require LiDAR or any other depth sensors for training; 4) The success of thermal depth estimation can provide an alternative solution of the active illumination LiDAR sensor for nighttime autonomous driving or other nighttime vision issues and enjoy the benefits of higher image resolution and less harm to human due to its passive sensor nature. Experimental results on both public multi-spectral dataset and our own collected dataset demonstrate that the proposed method outperforms other depth estimation methods for thermal images. This framework can be further extended to SfM, SLAM, and other computer vision tasks based on passive illumination sensors.

2. Related work

Existing feature matching based 3D reconstruction approaches rely heavily on the accurate feature extraction and matching, which cannot always be guaranteed in natural scenes, such as low illumination, low textures, occlusions, dark and foggy environments. Most recent image-based depth estimation methods applying deep neural networks mainly rely on ground truth labels to train the models, which generate promising results [10]. Liu et al. [31, 32], Xu et al. [51] and Li et al. [30] proposed to deal with the depth estimation problem based on deep CNN and Conditional Random Field (CRF) learning. A fully convolutional residual architecture is proposed to predict the depth information given an RGB image in [29, 9, 50]. Semi-supervised approaches are proposed in [27, 46, 37, 52] to train CNN by using both supervised and unsupervised learning clues. Unsupervised learning networks [15, 13, 37] are also proposed to estimate the depth based on intensity consistency between input image pairs. However, the methods mentioned above mainly target at visible images, which contain extensive textures and may not be suitable for images without vast textures. For images in different modalities, they cannot estimate the disparity based on pixel intensity consistency. To deal with this issue, we add an image translation branch in our framework to translate between visible and thermal images, which can support to build correspondences between two different modalities' images.

As for image generation, GAN [17, 5] has been used to generate a new type of images. Deep convolutional generative networks [41] generate images and and classify them simultaneously. As a specific kind of image generation, image translation transforms the source domain images to the target domain images. "U-Net" architecture [42] has been applied in Pix2Pix [23] as the generator and a convolutional neural network classifier as the discriminator. Pix2PixHD [49], as an extension of [23], incorporates multi-scale generator and discriminator to translate images with more accurate detailed information. The approaches mentioned above require paired images for training. He et al. [19], CycleGAN [58], DiscoGAN [26], and DualGAN [53] were proposed to train the translation network using unpaired data in a CycleGAN. Our focus is on translating the visible images to thermal images, for which we design a network to explore the property that facilitates to explore the relationship between color images and thermal images.

Nighttime vision, however, is still a relatively unexplored topic as a result of its numerous challenges. Thermal images reflecting the temperature of the object surface have been applied to detect pedestrians or animals [36, 2, 7, 11] in the nighttime and rescue assignments [12]. Various features [38, 44, 33, 34] from thermal images have been studied and fused to improve the thermal image classification

accuracy. Depth sensors (e.g., Kinect and Intel Realsense) have been fused with thermal sensors to generate 3D thermal models [48, 3, 18, 6], which are limited in small scale due to the depth sensors' valid range limitation. The thermal image has been applied to estimate the disparity between two color images to estimate the depth [25, 48] under the assumption one color camera has exactly the same disparity of thermal image. As in normal situations, two cameras usually maintain a baseline distance, the disparity between two cameras cannot be the same. To solve this issue, different from [25] that applies thermal image to estimate color image disparity, we first apply color stereo images to estimate the disparity as the initialization of the thermal image disparity, which can be more accurate. We further use the translated thermal image from color image to enhance the accuracy of thermal image disparity from its initialization. In this process, we build our own neural network and various loss constraints specially targeting at thermal-color appearance and illumination consistency to obtain more accurate and robust depth estimation for thermal images.

3. Single Thermal Image Depth Estimation

Thermal images are lack of textures, for which most commonly used features in computer vision tasks cannot be extracted and matched on thermal images. As we train the single thermal image depth estimation framework in an unsupervised learning scheme, no depth labels are obtained from depth sensors or human measurement. The ideal scenario is to apply two thermal cameras and apply our CNN model to train the single image depth estimation network. However, thermal cameras with sufficient sensitivity and resolution are much more expensive than regular color cameras. It is our objective to train the system with the minimum cost without using an extra thermal camera which is not in the use of real applications. Therefore, we apply two color cameras to learn the single image depth estimation system together with the thermal camera, which can significantly reduce the training requirement and cost due to the much lower price of color cameras. To estimate the accurate scene depth from a single thermal image, three main steps are involved in the entire framework. The first step is to apply the stereo color cameras to estimate a disparity map for the color camera (e.g., right color camera) close to the thermal camera. As the thermal camera is physically close to the right color camera, the disparity map can be used as the initialization for the thermal camera relative to the left color camera. Once we have the disparity initialization, the second step is to translate the left color image to a thermal image. The translation network is under a dual CycleGAN designed for translating color images to thermal images, with loss constraints fitting the properties of color and thermal images. Once we obtained the translated thermal images, we can refine the thermal image disparity map initialized from the right color camera. The disparity map is updated through a CNN model. Although we can obtain a thermal image from a color image through image translation, it may still maintain difference from the real thermal image, for which we do not directly estimate the disparity from translated left thermal image and real right thermal image. However, with a roughly correct disparity for initialization, we can accurately converge to the disparity correctly reflects the pixel motion between left translated thermal image and right real thermal image. As we know the baseline of the cameras and focal length, the real depth can be estimated through triangulation. In the testing phase, we estimate the disparity based on a single thermal image. We set the same baseline for testing (in real application) as the training phase, which provides the correct correlation to the disparity estimated from the single image (see Fig. 1).

Figure 2: Our camera system, which involves a thermal camera (the most left one) and two SLR cameras (second left one and the most right one). The most right thermal camera is to pair up with the other thermal camera to generate ground truth depth map, which is not used in training process.

3.1. Thermal Image Disparity Initialization

As thermal image cannot directly match the color image, we apply two color cameras to provide an initial disparity map for the thermal image. To realize this idea, we fix the thermal camera to be close to one of the color cameras. Therefore relative to the other side of the color camera, the adjacent color camera's disparity map can be similar to the thermal image. Our camera system is shown in Fig. 2. Two SLR color cameras have been fixed on the rig of the cart. The thermal camera is set to be close to the right color camera (left in the picture). The left thermal camera is to pair up with the right thermal camera to create ground truth depth map to evaluate our method, which is not required in the training process. For data collection, we use digital single-lens reflex camera Canon SL1 to capture color images and FLIR A615 to capture high-quality thermal images. The images from visible camera are further resized to the same size as thermal images. Sample images of the collected dataset are given in Fig. 3.

We use stereo reconstruction to estimate the disparity from the right color image to the left one. We calibrate the thermal and color cameras to know the rotation and translation matrices between them [45]. From the calibration matrix, we can also roughly crop the overlapping region between the right color camera and the thermal camera [48] so that the right color camera's disparity map can be close

Figure 3: Image samples from indoor and outdoor scenes of our collected dataset. The first frame in each row is the color image from left camera, the second and third frame are the right thermal and right color images respectively.

to the real thermal image's disparity. Though the disparity of the right color camera can be used to train the single thermal image depth estimation, the thermal camera and color camera still maintain a distance and the overlapping region is a rough estimation, for which we need to utilize a disparity map specialized for the thermal image to learn the depth estimation framework. Once we have the thermal image disparity initialization, we will rely on the translated thermal image and our CNN targeted at thermal images to obtain an accurate estimation of the disparity map.

3.2. Image Translation from Color to Thermal

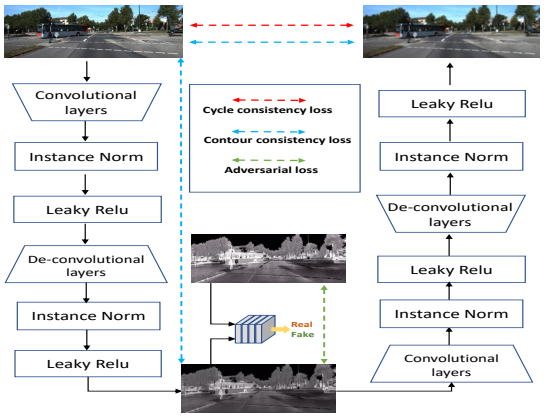

Figure 4: Our color to thermal image translation framework.

Once we have the rough estimation of the real thermal image disparity, we expect to have a thermal image as real as possible so that the disparity refinement can be accurate. To achieve this objective, we design a thermal image translation network targeting at color to thermal translation. To mitigate the training requirement, we apply a dual learning structure for unpaired data. So the color image does not necessarily have a corresponding thermal image capturing the same scene.

Under the learning structure, we explore the constraints for the neural network. We apply cycle consistency loss, cycle adversarial loss, and illumination consistency and contrast loss to constrain the network towards the accurate translation from color image to thermal image. The detailed image translation framework is shown in Fig. 4.

Inspired by the cycle structure in [58], we apply the cycle consistency as the primary structural constraint of the entire network. Though we are more interested in the translation relationship from color to thermal domain, accurate translation from thermal to color domain and maintaining the consistency between the original color image and the translated color image can also enhance the translation effect from color to thermal images. The output thermal image from the main generator can be the input of the auxiliary generator to output a translated color image. The cycle consistency is to enforce the input color image and the translated color image as close as possible. The loss constraint can be denoted as the following equation:

$$L_{cycle} = E_{c \in ori(c)}[\|Ga(Gm(c)) - c\|_1] \quad (1)$$

The cycle loss L_{cycle} is based on the expectation E of all the training samples distance between the generated color image $Ga(Gm(c))$ and the original color image c. The main generator Gm translates the original color image x to a thermal image $Gm(c)$, which is further translated to a color image $Ga(Gm(c))$ by the auxiliary generator Ga. Through enforcing the cycle consistency, we can improve the translation performance of both color to thermal modality and thermal to color modality.

To learn the transformation relationships between images in the source and target domains, we apply the adversarial loss to combine generator and discriminator pairs at the same time. Adversarial loss is for the discriminator to distinguish the real images and the generated images, which evaluates both generator and discriminator and can reduce the distribution variation of both types of images. Still, we focus on the translation from color image to thermal image, whose adversarial loss is as follows:

$$\begin{aligned} L_{GAN,C\to T} = {} & E_{t \in T_{ori(t)}}[1 - logD(t)] \\ & + E_{c \in C_{ori(c)}}[logD(G(c))] \end{aligned} \quad (2)$$

For the adversarial loss translating color image to thermal image $L_{GAN,C\to T}$, the discriminator D distinguishes whether the image is original (output 1) or generated (output 0). Once the input is thermal image t in the original thermal dataset T, the discriminator can reduce the loss by output 1 from $[1 - logD(t)]$. Similarly, if the input is a color image c in the original color dataset C, the discriminator is expected to output 0 for the generated thermal image $G(c)$.

Besides the cross-domain cycle consistency and domain adversarial constraints, spatial constraints can also be applied to train the network. As thermal and color images are two quite different modalities, the translated spectral values may also be quite different. We therefore propose to enhance the spectral similarity between the original and the translated images. More specifically, we expect the translated and the original images to be highly correlated and their average illumination strength to be close for each local patch. At the same time, we expect the illuminance contrast

within a patch between generated and original images to be similar as well, which represents the illuminance distribution and can be evaluated by the intensity variance. Our illuminance loss combines correlation relationship, average illuminance coefficient, and a contrast term together into consideration defined as follows:

$$L_{local\_lx}(x, x') = \frac{\sigma_{x,x'}}{\sigma_x \sigma_{x'}} \cdot \frac{2\bar{x}\bar{x'}}{(\bar{x}^2 + \bar{x'}^2)} \cdot \frac{2\sigma_x \sigma_{x'}}{\sigma_x^2 + \sigma_{x'}^2} \quad (3)$$

where x and x' are the original image and translated image in the same modality. The first term is the correlation relationship between the original data and predicted data. The second term is to reduce the variation of average illuminance between x and x'. The third term is to measure the intensity contrast to guarantee that they are in similar distribution. By minimizing their structural difference in illuminance, correlation, and contrast in the images from statistical characteristics, the quality of the generated image can be improved. As we desire to improve the image quality of each local region, we choose to scan the entire image by sliding a 5-pixel dimensional square window through the entire image with a moving step size as 2. Assuming there are M steps in this process, the whole image illuminance loss is provided by:

$$L_{lx} = \frac{1}{M} \sum_{i=1}^{M} (1 - L_{local\_lx}^{i}) \quad (4)$$

which averages all the local region illuminance consistency and contrast.

3.3. Neural Network Structure for Single Thermal Image Depth Estimation

Once we obtain the left and right thermal images, we target at estimating the left and right disparity maps simultaneously once provided a pair of images from the calibrated stereo cameras. The problem can be defined as follows: given every image pair from the stereo camera, our model constructs two CNN networks to predict their corresponding disparity maps respectively. As a result, the corresponding predicted output disparity maps can be used to reconstruct original input images from the stereo pair. Applying the basic idea from [15], our model further introduces a two CNN block scheme to extract geometry features and learns a regression pixel-wise output both from the left and right image with encoder-decoder structure. After extracting feature representations from the encoder, the decoder network further constrains the output disparity to produce the warped image $\widetilde{I^w}$ ($\widetilde{I^l}$ or $\widetilde{I^r}$) by moving pixels from the original input image along the epipolar line using bilinear sampling [24] in the surrounding positions. The relationship between the warped image $\widetilde{I^w}$, predicted disparity map $\widetilde{d}$, and the original input image I can be expressed as:

$$\widetilde{I^w} = I(x + \widetilde{d}) \quad (5)$$

Then the reconstructed image can be expressed as:

$$\widetilde{I}(p) = \sum_{i \in (t,b), j \in (r,l)} w_{ij} \widetilde{I^w}(p_{ij}) \quad (6)$$

The final predicted depth map is able to be calculated based on the stereo triangulation $\widetilde{D} = fB/\widetilde{d}$. Here, f is the focal length of the camera and B is the known baseline of the stereo cameras. After obtaining the initial depth prediction for a pyramid of multiple resolutions, we can refine the depth map based on our reconstruction errors and edge consistency constraints to produce more accurate results. We design loss constraints from both spectral and spatial perspectives to guide the training process.

3.3.1 Appearance Matching

Each of the CNN networks in our single image depth estimation branch produces a corresponding disparity map $\widetilde{d}$, and the reconstructed image is generated by the other original image together with its estimated disparity map. Appearance matching loss is introduced to enforce the estimated reconstructed images to be the same as the original images at the corresponding pixels. By combining the Structural Similarity Index Metric (SSIM) structure [15] and the introduced Charbonnier loss factor, the appearance matching loss for the right and left images is defined as:

$$\begin{aligned} L_{match} = \frac{1}{N} \sum_{ij} \frac{a}{2} \tilde{p}(1 - SSIM(I_{ij}, \tilde{I}_{ij})) \\ + (1 - a)\tilde{p}(||I_{ij} - \tilde{I}_{ij}||_1) \end{aligned} \quad (7)$$

where $\|\cdot\|_1$ is used to represent L_1 norm operator which calculates the mean absolute value. I_{ij} represents the original input left and right images, and $\tilde{I}_{ij}$ represents our reconstructed left and right images. α is a constant parameter and here we choose 0.85 as its value. The value for SSIM ranges from 0 and 1, where 1 means that they can realize a perfect matching.

3.3.2 Edge-aware Smoothness

Edge-aware smoothness is introduced as a penalty term to further prevent small divergent depth values from occluded regions and low-textured areas. It attempts to minimize depth-related Laplacian of Gaussian (LoG) filter whose each element is weighted by the corresponding image gradient, described as follows:

$$L_{edge} = w \left\| \frac{|\bigtriangledown^2 (D \otimes G)|}{\|\bigtriangledown I\|} \right\|_1 \times \frac{1}{\|D\|} \quad (8)$$

where $\bigtriangledown$ and $\bigtriangledown^2$ refer to the gradient and Laplacian operator respectively. G represents a 5×5 Gaussian kernel. D and I correspond to the predicted disparity map and input image respectively. w is the weight parameter and is

set to 0.25 in our network. Through scaling the first term by dividing the mean disparity value, the output can be normalized. Based on the edge-aware loss, we can present small depth changes due to the noise, which is especially helpful for thermal images, as the thermal image representing the object surface temperature usually maintains uniform and contains noise due to the sensor sensitivity.

Hence, the full objective function that optimizes the network parameters becomes $L_{total} = \lambda_{cycle}L_{cycle} + \lambda_{GAN,C\to T}L_{GAN,C\to T} + \lambda_{lx}L_{lx} + \lambda_{match}L_{match} + \lambda_{edge}L_{edge}$, where λ_{cycle}, $\lambda_{GAN,C\to T}$, λ_{lx}, λ_{match} and λ_{edge} represent weights for different loss constraints and are set as 10, 1, 0.5, 1, 0.5 respectively.

4. Experimental Results

4.1. Data Collection

The proposed framework is trained on a rectified multispectral stereo dataset which contains 10000 image pairs covering different scenes including campus, indoor and outdoor halls. The depth maps obtained from the color RGB images by stereo reconstruction are transformed to the thermal image. We use the thermal image stereo reconstruction result as the ground truth to verify the single thermal image depth estimation framework. Before applying stereo reconstruction [47] on thermal images, we also use the color stereo disparity to limit the correspondence search range, which can enhance the stereo reconstruction speed and accuracy relying on color stereo reconstruction result. The second thermal camera is used to create the ground truth disparity. However, our training procedure does not require the second thermal camera. The thermal stereo setting is mainly for ground truth depth map generation.

As there is no available public datasets for stereo multispectral applications sharing the same setting and purpose as our collected dataset, we further verify our method on popular KITTI benchmarks for autonomous driving. We applied trained model to translate raw color images into long-infrared thermal domain, and then train our depth estimation pipeline for single thermal image on it. As the thermal image and right color image have the same camera extrinsic in this case, we apply the rotation and translation operation on the translated right thermal images based on our thermal-to-right-color camera calibration matrix to reassembly our camera setting. 2000 testing images are randomly chosen to obtain the visual comparison with other methods.

4.2. Implementation Details

We train the network on a GTX 1080Ti GPU with 11G memory with input image size of 640×480. The Adam optimizer is applied with β_1=0.9 and β_2=0.99. The learning rate starts with $1e^{-4}$ and gradually decreases to the half of the original rate in every 20 epochs. Horizontal flip is utilized to augment the training data to prevent potential overfitting. The weights of the depth estimation network and image translation network are initialized with Kaiming initialization method [20] with a batch size of 4. The network is trained for 50 epochs.

4.3. Comparison with State-of-the-art methods

Method	Abs Rel	Sq Rel	RMSE	RMSE log	δ <1.25	δ <1.25^2	δ <1.25^3
SfMLearner [57]	0.24	2.12	10.43	2.31	0.19	0.36	0.53
Geonet [54]	0.22	1.93	11.20	2.23	0.22	0.40	0.61
Monodepth [15]	0.19	1.38	8.11	1.64	0.45	0.66	0.77
3-Net [40]	0.19	1.27	7.91	1.52	0.48	0.69	0.81
PydNet [39]	0.22	1.51	7.98	1.54	0.46	0.66	0.80
Lai et.al [28]	0.31	2.69	13.58	3.21	0.16	0.27	0.43
Monodepth2 [16]	**0.17**	1.32	7.87	1.46	0.49	0.67	0.78
Ours	**0.17**	**1.15**	**7.29**	**1.37**	**0.54**	**0.72**	**0.83**

Table 1: Comparison results when evaluating on the testing set of our collected dataset. Each image is associated with a ground truth depth. The proposed method is compared with the state-of-the-art monocular depth estimation techniques [57] [54] and stereo techniques [15] [40] [39] [28][16] [28]. The first four columns represent the error metrics(lower means better). The last three columns are the accuracy metrics(higher means better).

The performance of our proposed approach is evaluated with other recent methods using multiple standard evaluation metrics, including absolute relative difference (Abs Rel), squared relative difference (Sq Rel), Root Mean Square Error (RMSE), RMSE log, and the maximum ratio δ between the predicted depth and the ground truth depth to be within the threshold of 1.25, 1.25^2, and 1.25^3. For the error metrics of absolute relative difference, squared relative difference, RMSE, and RMSE log, the lower value means the better performance, and for accuracy fitting depth ratio threshold, the higher number indicates more images has small difference compared with ground truth, which represents better performance.

We perform a quantitative analysis for the accuracy and error of the proposed method with respect to the other state-of-the-art methods based on the single thermal image as input. As can observed in Table. 1, our proposed model for thermal image depth estimation provides the best depth estimation results based on all the evaluation metrics compared to other state-of-the-art methods for single image depth estimation trained by both monocular video sequence [57] [54] and stereo cameras [15] [40] [39] [28][16], especially a large improvement in the accuracy metrics, which demonstrates the superior effectiveness of our method in single thermal image depth estimation. For single image depth estimation based on monocular video sequence, we train the network based on the original thermal images. For single image depth estimation methods for comparison, we apply the original thermal image and the translated left thermal image from left color image to compose the thermal image pair to train the network, which enforces the training data to be in the same modality for a fair comparison.

Besides quantitative analysis, Fig. 5 provides a visual comparison with different existing methods on the collected dataset. Our model can generate superior depth estimation output. Compared with [15] [40] [39] [16], the proposed

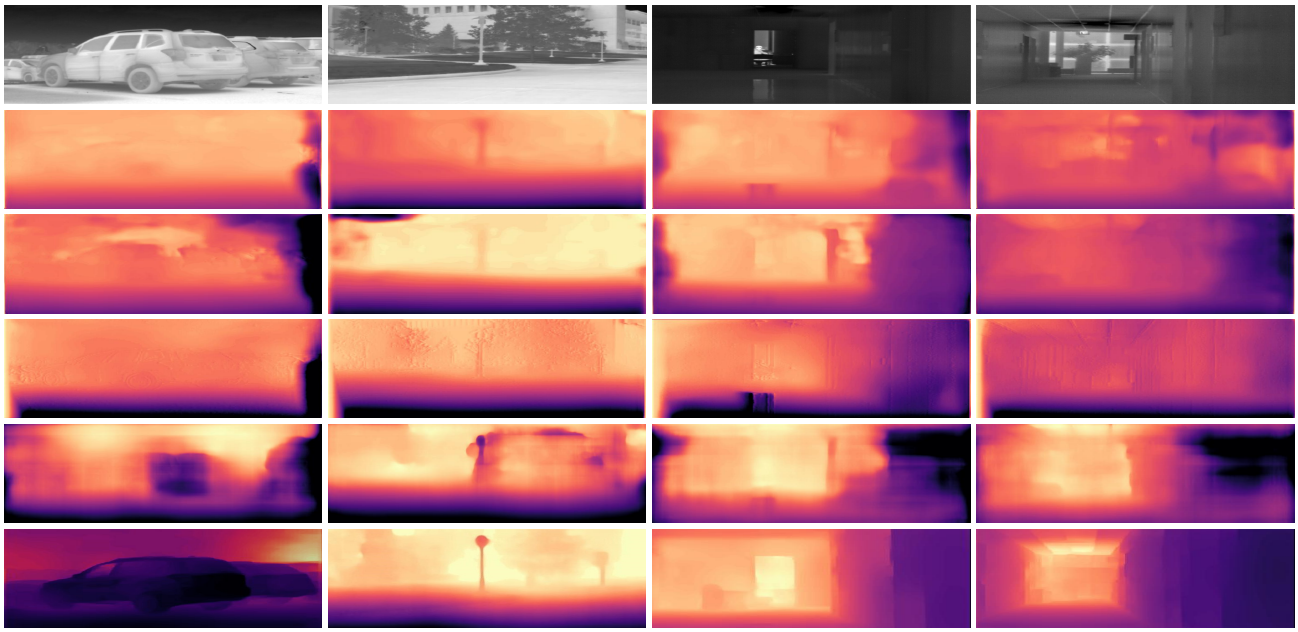

Figure 5: Estimated depth maps of our method compared with other state-of-the-art methods trained on the collected dataset. First row: input thermal images; Second row: predicted depth maps from [15]; Third row: predicted depth maps from [40]; Fourth row: predicted depth maps from [39]; Fifth row: predicted depth maps from [16]; Sixth row: our predicted depth maps.

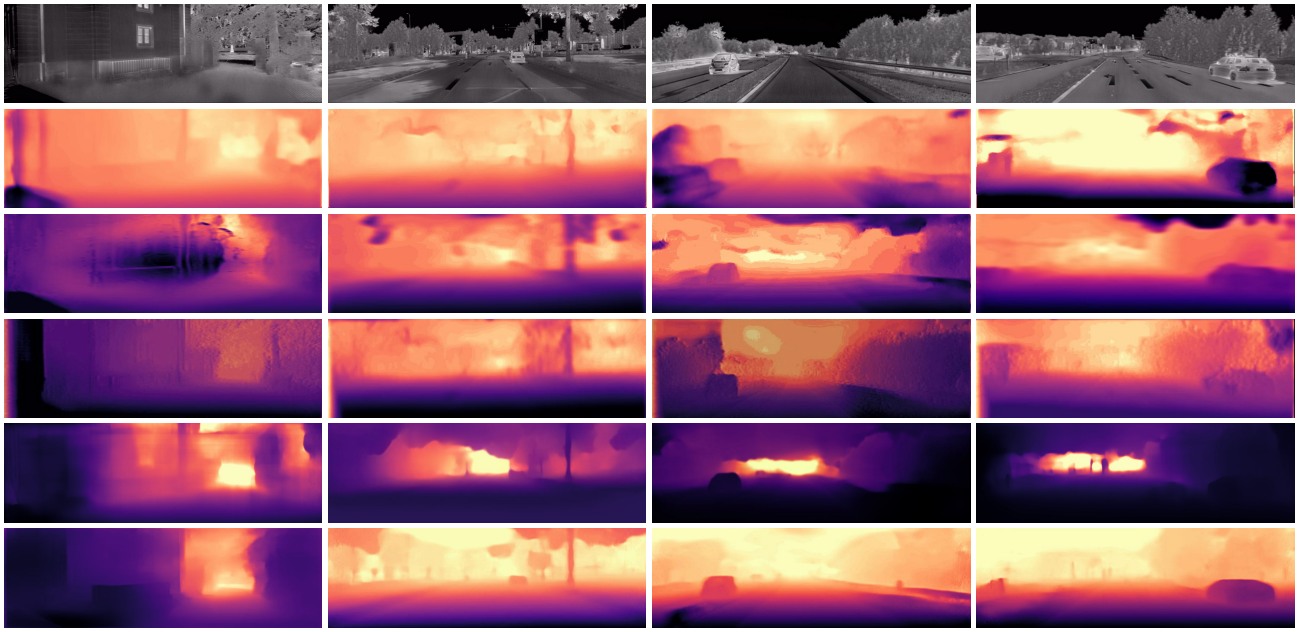

Figure 6: Estimated depth maps of our method compared with other state-of-the-art methods on KITTI dataset. First row: input thermal images; Second row: predicted depth maps from [15]; Third row: predicted depth maps from [40]; Fourth row: predicted depth maps from [39]; Fifth row: predicted depth maps from [16]; Sixth row: our predicted depth maps.

method is able to get rid of many wrong predictions such as in the sky and on the walls. Overall, the estimated depth maps based on thermal images prove the effectiveness of our proposed method on the single thermal image depth estimation tasks for both indoor and outdoor scenes.

Further visual comparison with various of most recent methods on KITTI dataset is shown in Fig. 6. Our model is capable to generate a reasonable prediction output with clear boundaries on objects. Compared with [15] [40] [39] [16], the proposed method is able to get rid of many wrong predictions such as in sky and on walls. Our estimated depth maps based on thermal images recover an accurate relative positions and object boundaries from given scenes, and prove that the effectiveness of our proposed method on single thermal image depth estimation.

4.4. Image Translation Evaluation

	PSNR	SSIM	COS	RMSE
MUNIT [22]	11.67	0.52	0.72	0.37
DualGAN [53]	9.40	0.34	0.75	0.39
Pix2PixHD [49]	16.91	0.72	0.83	0.34
Ours	18.04	0.77	0.89	0.24

Table 2: Translation results from color to IR on the collected datatset analyzed from PSNR, SSIM, COS similarity and RMSE.

Figure 7: Image samples translated from color visible domain to infrared radiation domain on our collected dataset. The first column: color image input. The second column: translated thermal image output.

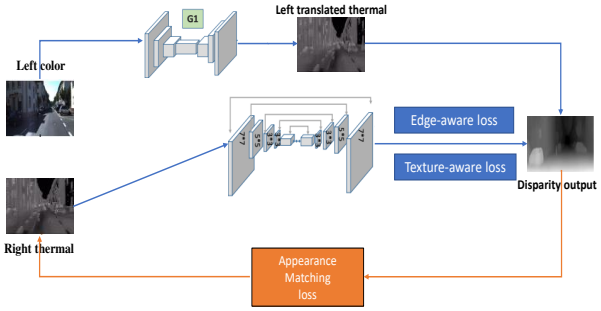

Figure 8: Ablation study to train the framework structure using one thermal image and one color image without initialization stage.

As part of our entire pipeline, image translation module from visible image to thermal image also plays a critical role. To evaluate the image translation effectiveness, we conduct three experiments to test the visual and quantitative performance on both KITTI and our collected dataset. First, we show the visual performance of our trained model when applying on KITTI dataset [14] to successfully translate the input color visible image to synthetic thermal image.

Then we provide both quantitative and visual performance of the proposed model on the new collected dataset. Table. 2 analyses commonly used four measurement metrics of image quality during the evaluation, which are Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM), COS Similarity (COS) and Root Mean Squared Error (RMSE). For the first three metrics, the one that with the highest value means that the translated output is the closest to the target ground truth. For the last metric, the lowest value means it shows the best performance on translation. It can be observed from Table. 2 that our results achieve the best performance on both of the four metrics, especially for RMSE. Visual performance of the translated thermal image results on our collected dataset is shown as Fig. 7, in comparison with the ground truth thermal image. The result demonstrates that the output from our translation framework is able to recover the subtle temperature contrast coinciding with the ground truth thermal image collected from our high-resolution thermal camera. This provides a critical support for our depth estimation pipeline compared with two different settings, as discussed in Section 4.5.

4.5. Ablation analysis

To prove the effectiveness of the proposed method that we first apply color stereo images to estimate an initial disparity as the thermal image disparity, we conduct an ablation study here with different framework setting to train our depth estimation model and show the visual results. We compare against two configurations as shown in Fig. 8. Compared with our framework, there is no initialization from stereo color images to help with the learning process. As a result, the depth estimation effect is weakened. The depth estimation result compared with the framework Fig. 1 is shown as Fig. 9.

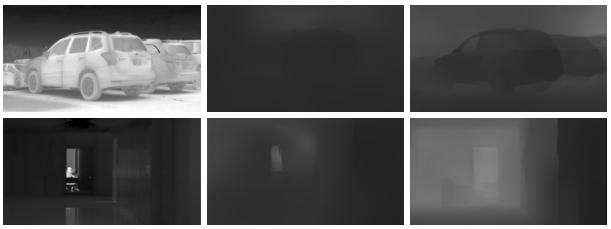

Figure 9: Comparison of the estimated results from pipeline of Fig. 8 and Fig. 1. From left to right: Input thermal image; Result from the framework as Fig. 8; Result from framework as Fig. 1

5. Conclusion

We propose an unsupervised single thermal image depth estimation framework to estimate scene depth in the environment with low illumination as an alternative of the active sensor LiDAR, which suffers low resolution and brings harm to the environment due to the emitted laser. To roughly obtain the disparity for the thermal image relative to another color camera, we apply two color cameras to provide a disparity initialization for the thermal camera. we further translate the color camera to the thermal domain through our color-to-thermal image translation network to refine the disparity. The network recovers the disparity with multiple designed constraints targeting at thermal images from both spectral and spatial perspectives. Experiments demonstrate the outstanding performance of our single thermal image depth estimation method, which enables extensive tasks (e.g., autonomous driving) that is mainly available in daytime to be also possible in nighttime.

References

[1] Ali Jahani Amiri, Shing Yan Loo, and Hong Zhang. Semi-supervised monocular depth estimation with left-right consistency using deep neural network. *arXiv preprint arXiv:1905.07542*, 2019.

[2] Massimo Bertozzi, Alberto Broggi, Paolo Grisleri, Thorsten Graf, and Michael Meinecke. Pedestrian detection in infrared images. In *IVS*, pages 662–667, 2003.

[3] Dorit Borrmann, Andreas Nüchter, Marija Dakulovic, Ivan Maurovic, Ivan Petrovic, Dinko Osmankovic, and Jasmin Velagic. The project thermalmapper-thermal 3d mapping of indoor environments for saving energy. In *SyRoCo*, pages 31–38, 2012.

[4] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

[5] Richard Chen, Faisal Mahmood, Alan Yuille, and Nicholas J Durr. Rethinking monocular depth estimation with adversarial training. *arXiv preprint arXiv:1808.07528*, 2018.

[6] Y Cho and C Wang. 3d thermal modeling for existing buildings using hybrid lidar system. In *Computing in Civil Engineering (2011)*, pages 552–559. 2011.

[7] Grzegorz Cielniak, Tom Duckett, and Achim J Lilienthal. Data association and occlusion handling for vision-based people tracking by mobile robots. *RAS*, 58(5):435–443, 2010.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[9] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015.

[10] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, 2014.

[11] Antonio Fernández-Caballero, José Carlos Castillo, Javier Martínez-Cantos, and Rafael Martínez-Tomás. Optical flow or image subtraction in human detection from infrared camera on mobile robot. *RAS*, 58(12):1273–1281, 2010.

[12] A Garcia-Cerezo, A Mandow, JL Martinez, J Gómez-de Gabriel, J Morales, A Cruz, A Reina, and J Seron. Development of alacrane: A mobile robotic assistance for exploration and rescue missions. In *SSRR*, pages 1–6, 2007.

[13] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, pages 740–756, 2016.

[14] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[15] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

[16] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. *ICCV*, 2019.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[18] Youngjib Ham and Mani Golparvar-Fard. Rapid 3d energy performance modeling of existing buildings using thermal and digital imagery. In *Construction Research Congress 2012: Construction Challenges in a Flat World*, pages 991–1000, 2012.

[19] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *NIPS*, pages 820–828, 2016.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034, 2015.

[21] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 807–814. IEEE, 2005.

[22] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, pages 172–189, 2018.

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017.

[24] Max Jaderberg, Karen Simonyan, and Andrew Zisserman. Spatial transformer networks. In *NIPS*, pages 2017–2025, 2015.

[25] Namil Kim, Yukyung Choi, Soonmin Hwang, and In So Kweon. Multispectral transfer network: Unsu-

pervised depth estimation for all-day vision. In *AAAI*, 2018.

[26] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, pages 1857–1865, 2017.

[27] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *CVPR*, pages 6647–6655, 2017.

[28] Hsueh-Ying Lai, Yi-Hsuan Tsai, and Wei-Chen Chiu. Bridging stereo matching and optical flow via spatiotemporal correspondence. In *CVPR*, pages 1890–1899, 2019.

[29] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, pages 239–248, 2016.

[30] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1119–1127, 2015.

[31] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *CVPR*, 2015.

[32] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian D Reid. Learning depth from single monocular images using deep convolutional neural fields. *TPAMI*, 38(10):2024–2039, 2016.

[33] Guoyu Lu, Yan Yan, Li Ren, Philip Saponaro, Nicu Sebe, and Chandra Kambhamettu. Where am i in the dark: Exploring active transfer learning on the use of indoor localization based on thermal imaging. *Neurocomputing*, 173:83–92, 2016.

[34] Guoyu Lu, Huili Yu, and Chun Yuan. Getting rid of night: Thermal image classification based on feature fusion. In *ICPR*, pages 2827–2832, 2018.

[35] Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li, and Liang Lin. Single view stereo matching. In *CVPR*, pages 155–163, 2018.

[36] Harsh Nanda and Larry Davis. Probabilistic template based pedestrian detection in infrared videos. In *IVS*, pages 15–20, 2002.

[37] Jogendra Nath Kundu, Phani Krishna Uppala, Anuj Pahuja, and R Venkatesh Babu. Adadepth: Unsupervised content congruent adaptation for depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2656–2665, 2018.

[38] Bin Pan, Zhenwei Shi, and Xia Xu. Longwave infrared hyperspectral image classification via an ensemble method. *International Journal of Remote Sensing*, 38(22):6164–6178, 2017.

[39] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. Towards real-time unsupervised monocular depth estimation on cpu. In *IROS*, 2018.

[40] Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In *3DV*, pages 324–333, 2018.

[41] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015.

[43] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.

[44] Philip Saponaro, Scott Sorensen, Abhishek Kolagunda, and Chandra Kambhamettu. Material classification with thermal imagery. In *CVPR*, pages 4649–4656, 2015.

[45] Philip Saponaro, Scott Sorensen, Stephen Rhein, and Chandra Kambhamettu. Improving calibration of thermal stereo cameras using heated calibration board. In *ICIP*, pages 4718–4722, 2015.

[46] Nikolai Smolyanskiy, Alexey Kamenev, and Stan Birchfield. On the importance of stereo for accurate depth estimation: An efficient semi-supervised deep neural network approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1007–1015, 2018.

[47] Scott Sorensen, Philip Saponaro, Stephen Rhein, and Chandra Kambhamettu. Multimodal stereo vision for reconstruction in the presence of reflection. 2015.

[48] Stephen Vidas, Peyman Moghadam, and Michael Bosse. 3d thermal mapping of building interiors using an rgb-d and thermal camera. In *ICRA*, pages 2311–2318, 2013.

[49] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation

with conditional gans. In *CVPR*, pages 8798–8807, 2018.

[50] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *European Conference on Computer Vision*, pages 842–857. Springer, 2016.

[51] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *CVPR*, 2017.

[52] Nan Yang, Rui Wang, Jorg Stuckler, and Daniel Cremers. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 817–833, 2018.

[53] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, pages 2849–2857, 2017.

[54] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1983–1992, 2018.

[55] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *CVPR*, 2018.

[56] Chi Zhang, Zhiwei Li, Yanhua Cheng, Rui Cai, Hongyang Chao, and Yong Rui. Meshstereo: A global stereo model with mesh alignment regularization for view interpolation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2057–2065, 2015.

[57] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2017.

[58] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017.