Self-supervised Monocular Depth Estimation from Thermal Images via Adversarial Multi-spectral Adaptation

Ukcheol Shin Kwanyong Park Byeong-Uk Lee Kyunghyun Lee In So Kweon
Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Korea
{shinwc159, pkyong7, byeonguk.lee, kyunghyun.lee, iskweon77}@kaist.ac.kr

Abstract

Recently, thermal image based 3D understanding is gradually attracting attention for an illumination condition agnostic machine vision. However, the difficulty of the thermal image lies in insufficient training supervision due to its low-contrast and textureless properties. Also, introducing additional modality requires further constraints such as complicated multi-sensor calibration and synchronized data acquisition. To leverage additional modality information without such constraints, we propose a novel training framework that consists of self-supervised learning of unpaired multi-spectral images and feature-level adversarial adaptation. In the training stage, we utilize unpaired RGB/thermal video and partially shared network architecture consisting of modality-specific feature extractors and modality-independent decoder. Through the shared network design, the depth decoder can leverage the self-supervised signal of the unpaired RGB images. Feature-level adversarial adaptation minimizes the gap between RGB and thermal features and eventually makes the thermal encoder extract representative and informative features. Based on the proposed method, the trained depth network shows outperformed results than previous state-of-the-art methods.

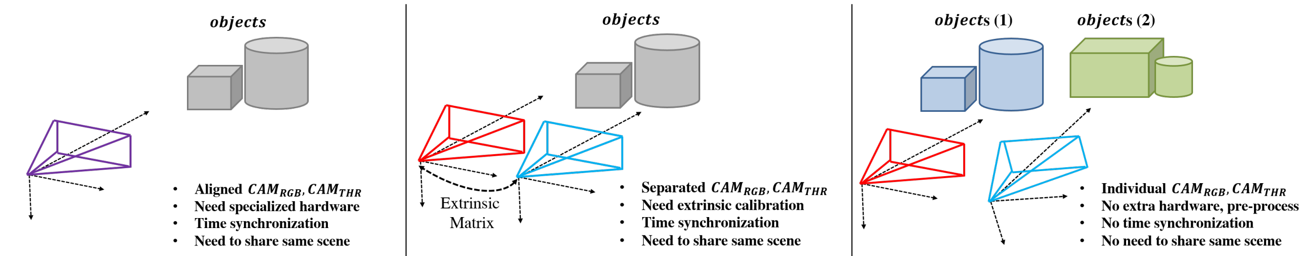
1. Introduction

Self-supervised learning of 3D understanding tasks such as depth, pose, and scene flow estimation [45, 44, 35, 11, 3, 15, 16, 24] have been researched to reduce the burden of expensive and careful ground-truth data creation process. Also, recent self-supervised learning research for depth and pose estimation [43, 13, 3] almost reached comparable performance with supervised baselines [9, 28, 1]. However, most studies have been researched on the RGB image domain. Therefore, these works show critical vulnerability and performance drop according to illumination and weather conditions, such as in low-lighted, cloudy, rainy, and foggy, and snowy scenes.

Long-wave infrared camera, also known as a thermal imaging camera, maintain consistent image quality because a thermal camera is less affected by weather and lighting condition changes. In addition, since it has sufficient image resolution, dense machine perceptions, such as dense semantic segmentation [38, 39] and depth estimation [27, 37], are also possible. Therefore, thermal image based 3D vision applications for a robust robot vision [7, 20, 37, 27] are gradually attracting attention recently. However, the difficulty of thermal image lies in its image properties. Thermal image tend to have low contrast and low texture information, which are the most fundamental sources in previous self-supervised depth and pose estimation approaches.

To takle the issue of thermal properties, the previous self-supervised depth estimation methods for thermal image [20, 37, 27] exploits RGB color images. Kim *et al*. [20] and Lu *et al*. [27] utilizes spatial image reconstruction with paired stereo RGB images and stereo RGB-thermal images. For this purpose, they need a specialized sensor system that consists of stereo RGB and one thermal cameras that shares the same principal axis with a beam splitter, or that consists of very closely located stereo RGB and stereo thermal cameras (Fig. 1-(a)). Shin *et al*. [37] use temporal image reconstruction with paired RGB-thermal images. Based on the method, they bring a performance improvement in the thermal image based depth estimation task. However, the method also inherited the above-mentioned multi-sensor problems such as complicated multi-sensor calibration and synchronized data acquisition (Fig. 1-(b)).

To address the thermal properties and multi-sensor problems, in this paper, we propose a novel training framework that combines self-supervised learning of unpaired multi-spectral images and feature-level adversarial adaptation for monocular depth estimation of thermal image. The proposed method effectively leverages additional modality information without requiring any extra constraint, such as specialized hardware, multi-sensor calibration process, and sensor synchronization compared to the previous methods [20, 27, 37] (Fig. 1-(c)).



(a) Strict spatio-temporal constraints [20, 27] (b) Relieved spatio-temporal constraints [37] (c) No spatio-temporal constraints (Ours)

Figure 1: **Required constraints on RGB-Thermal Training Data.** Previous self-supervised depth estimation methods for thermal image utilized RGB images in a training stage as an auxiliary self-supervision source. However, for this purpose, the previous methods [20, 27, 37] requires specialized hardware setup to built accurately aligned RGB-thermal image pair [20, 27], difficult multi-sensor extrinsic calibration process [37], and time synchronization between RGB and thermal streams [20, 27, 37]. On the other hand, our proposed method fully resolves the constraints between RGB and thermal images.

Our contributions can be summarized as follows:

- We propose a self-supervised learning method of unpaired RGB-thermal images to provide self-supervisory signal and effectively transfer RGB domain knowledge to the thermal domain by exploiting depth decoder sharing, unpaired multi-spectral image reconstruction, and locally consistent thermal image scaling method.
- We propose an adversarial feature adaptation method to enhance a feature representation ability of the thermal image encoder by minimizing feature-space domain gap between RGB and thermal features.
- We demonstrate that the proposed method outperforms previous state-of-the-art approaches on the ViViD benchmark dataset [23] both quantitatively and qualitatively without requiring any extra constraints.

2. Related Works

2.1. Self-supervised Depth from Thermal Image

Recently, self-supervised depth estimation methods from thermal images are getting attention [20, 27, 37, 36] to leverage weather and lighting condition agnostic properties of the thermal image. However, the difficulty of a thermal image lies in its image properties, such as low contrast ratio and low texture information, which weakens the self-supervisory signal of the image reconstruction loss.

Therefore, most previous works [20, 27, 37] utilize auxiliary self-supervision source to train a depth estimation network. Kim *et al*. [20] exploited spatial image reconstruction with paired stereo RGB images and estimated depth map from a thermal image. For this purpose, they design a sensor system consisting of two RGB cameras, one thermal camera, and a beam splitter for the principal axis alignment of RGB-thermal cameras [7]. Lu *et al*. [27] also needs a specialized hardware system that has very closely located RGB stereo and thermal stereo camera. They exploit an image translation network to synthesize a thermal-like left image. After that, the spatial reconstruction loss between the thermal-like left and real right thermal images is used to train the depth network. Shin *et al*. [37] utilizes a temporal reconstruction loss with paired RGB-thermal images to train single-view depth and multiple-view pose networks.

These methods [20, 27, 37] bring a performance improvement by leveraging additional self-supervision sources. However, these methods require extra constraints such as a specialized image setup, complicated multi-sensor calibration, and synchronized data acquisition. On the other hand, our proposed method does not require any extra constraints by exploiting adversarial domain adaptation and self-supervised learning of unpaired RGB-thermal videos.

2.2. Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA) aims to transfer the knowledge from the labeled source domain to the unlabeled target domain. It has shown remarkable progress on many computer vision tasks such as image classification [41], semantic segmentation [40], and object detection [5]. A common strategy for UDA is to reduce the domain gap by constructing shared embedding space across both source and target domains. Under this goal, many works introduce adversarial training [14] and the main difference among them is in which the embedding space is shared (*e.g.* image-level [29, 31, 46, 29, 6, 18, 12], feature-level [41, 5, 18, 32], and prediction-level [40, 4, 26, 30, 21, 25]). However, most works still target the scenario from label-rich domain to unlabeled domain in RGB modality.

Apart from the previous works, we investigate the cross-modality transfer learning setup, viewing each modality as an independent domain. In addition, instead of expensive annotations, we leverage self-supervised learning of depth and pose estimation on both domains. Thus, our network is trained in a fully unsupervised manner.

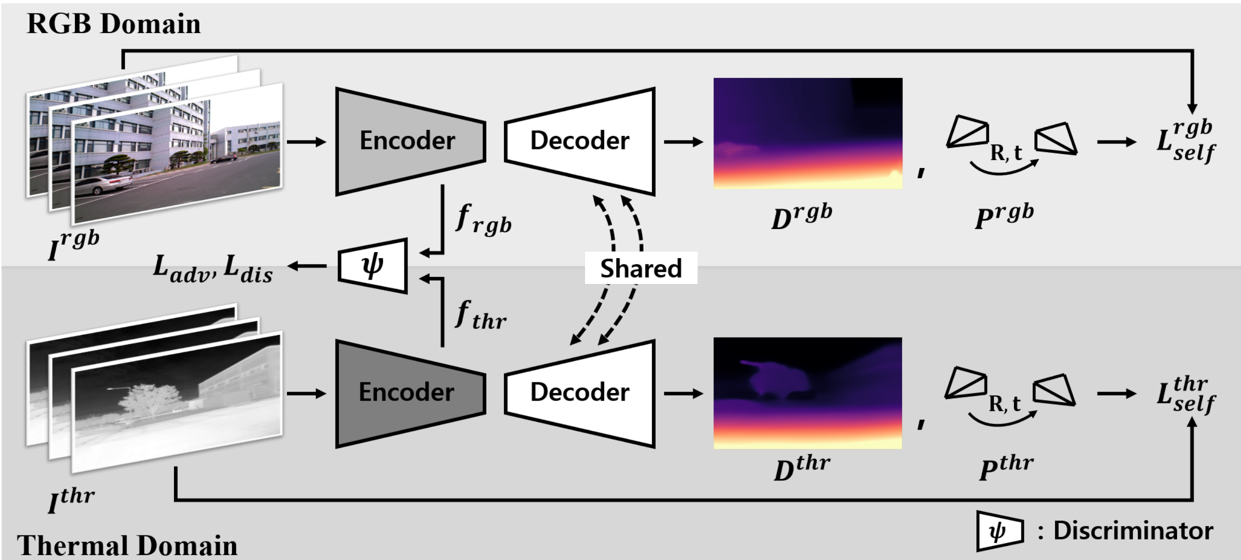


Figure 2: **Overall pipeline of our proposed training framework**. The overall architecture of our framework consists of two domain-specific encoders (E_{thr} and E_{rgb}), a domain-shared decoder, and discriminator ψ. Given unpaired RGB and thermal images, the networks estimate depths (D^{rgb} and D^{thr}) and relative poses (P^{rgb} and P^{thr}) on each image domain. After that, the networks are trained with a self-supervised loss L_{self} by reconstructing each image sequence. At the same time, feature-level domain adaptation explicitly guides the thermal extractor to encompass representative feature extraction ability via adversarial loss L_{adv} between RGB and thermal feature maps (f_{rgb} and f_{thr}).

3. Method

3.1. Method Overview

The proposed method aims to solve the weak self-supervision problem of thermal images by utilizing additional modality information without requiring multi-sensor calibration, synchronized data acquisition, and a specialized hardware setup. The ideas of the proposed method to utilize unpaired RGB and thermal images are shown in Fig. 2.

First, we designed a partially shared network architecture to propagate a self-supervised loss L^{rgb}_{self} of unpaired RGB images. Here, we consider modality-specific encoders because we observed the RGB and thermal images have a high appearance gap and data distribution differences. Through the shared network design, the depth decoder can leverage the self-supervised losses of both the unpaired RGB and thermal images (L^{rgb}_{self}, L^{thr}_{self}).

However, the thermal encoder E_{thr} still suffers from insufficient self-supervision since the loss L^{rgb}_{self} is not propagated to the thermal encoder. Therefore, secondly, we exploit a domain adaptation method in the feature space to provide an additional self-supervision and transfer the representative feature extraction ability of the RGB encoder E_{rgb} to the thermal encoder E_{thr}. As a result, the thermal encoder can learn to extract informative feature maps even from the low-textured thermal images. Based on the network design, self-supervised learning of unpaired RGB-thermal video, and feature-level adaptation, our proposed method effectively leverages additional modality information without relying on the multi-sensor calibration, synchronized data acquisition, and specialized hardware setup.

3.1.1 Training Objective

The proposed method utilizes unpaired RGB and thermal images in the training stage to leverage efficient self-supervisory signal of the RGB domain. Our proposed method mainly consists of two learning methods; self-supervised learning via unpaired RGB-Thermal images (L^{rgb}_{self} and L^{thr}_{self}) and feature-space domain adaptation via adversarial loss L_{adv} between RGB and thermal features. Our overall training loss to train single-view depth and multiple-view pose estimation network is as follows:

$$L_{total} = L^{rgb}_{self} + L^{thr}_{self} + \lambda_{adv} L_{adv}, \tag{1}$$

where L_{self} indicates the self-supervised learning loss and λ_{adv} is a scale factor for the adversarial loss L_{adv}. Self-supervised learning loss of RGB domain L^{rgb}_{self} propagates depth extraction knowledge via the shared depth decoder from the RGB source to the thermal target domain. Adversarial loss L_{adv} enhances the feature extraction ability of the thermal feature encoder E_{thr} by minimizing domain gap between RGB and thermal feature spaces. Note that the discriminator ψ is trained with the discriminator loss L_{dis}.

3.2. Adversarial Multi-spectral Feature Adaptation

Under the guidance of a self-supervised signal on both modalities, the shared depth decoder is trained in a domain invariant way so that both features, f_{thr} and f_{rgb}, are well decoded into the depth space. However, the thermal feature extractor E_{thr} still tends to extract less discriminative features compared to the RGB feature extractor E_{rgb}. Although RGB and thermal images have a large discrepancy in input distribution, their feature space should share strong spatial and local similarities according to depth of scene. Thus, we utilize this insight to transfer the knowledge from RGB to thermal domain via an adversarial alignment of their features.

3.2.1 Discriminator Loss

The discriminator ψ attempts to distinguish whether a given feature is generated from RGB or thermal domain. The competition between the feature extractor E_{thr} and the discriminator ψ helps the feature extractor E_{thr} to generate indistinguishable feature f_{thr} with RGB feature f_{rgb} from the thermal image. The loss function L_{Dis} to train the discriminator ψ is defined as follows:

$$L_{Dis} = L_{MSE}(\psi(f_{thr}), 0) + L_{MSE}(\psi(f_{rgb}), 1), \quad (2)$$

where $\psi(\cdot)$ denotes prediction result of the discriminator ψ, L_{MSE} is Mean Squared Error loss.

3.2.2 Adversarial Loss

The purpose of adversarial loss is to enhance the representation ability of the thermal extractor E_{thr} by minimizing domain gap between RGB feature f_{rgb} and thermal feature f_{thr}. This process is accomplished by the competition between the feature extractor E_{thr} and the discriminator ψ. Thermal feature extractor E_{thr} struggles to make the discriminator ψ misclassify the given thermal feature f_{thr} as belonging to the RGB feature space. The adversarial loss, which makes the feature extractor E_{thr} extracts an RGB domain like feature, is defined as follows :

$$L_{adv} = L_{MSE}(\psi(f_{thr}), 1), \quad (3)$$

3.3. Self-supervised Training

As shown in Fig. 2, the networks are trained in a self-supervised manner by reconstructing each spectrum image with intrinsic matrix, estimated depth map, and estimated relative camera pose. Even if a thermal image based reconstruction loss provides a weak self-supervisory signal, RGB image based loss signal is propagated to the shared depth decoder D_{sh} and leads to knowledge transfer from RGB to the thermal domain. Self-supervised training loss to train single-view depth and multiple-view pose estimation network is as follows:

$$L_{self} = L_{rec} + \lambda_{gc} L_{gc} + \lambda_{sm} L_{sm}, \quad (4)$$

where L_{rec} indicates image reconstruction loss, L_{gc} is geometric consistency loss, L_{sm} is edge-aware depth smoothness loss, and λ_{gc} and λ_{sm} are hyper parameters. In the following subsections, we use two consecutive images $[I_t, I_s]$ (*i.e.*, target and source images) for a concise explanation.

3.3.1 Image Reconstruction Loss

As shown in Fig. 2, the depth and pose networks estimate a depth map D_t and relative camera pose $P_{t \to s}$ from a consecutive images I_t, I_s. After that, a synthesized image $\tilde{I}_t$ is generated with the source image I_s, target depth map D_t, and relative pose $P_{t \to s}$ in the inverse warping manner [45]. The image reconstruction loss, which consists of L1 difference and Structural Similarity Index Map (SSIM) [42], is calculated by measuring the difference between the synthesized and original target images, as follows:

$$L_{pe}(I_t, \tilde{I}_t) = \frac{\gamma}{2}(1 - SSIM(I_t, \tilde{I}_t)) + (1 - \gamma)||I_t - \tilde{I}_t||_1, \quad (5)$$

where γ indicates scale factor between SSIM and L1 loss.

3.3.2 Locally Consistent Thermal Image Scaling

As shown in Fig. 3, a typical thermal camera generates a relative scale thermal image in a built-in pipeline [8]. The camera convert a RAW thermal image into a scaled thermal image by normalizing the RAW image with its min and max values. Therefore, as the temperature distribution within a scene change, the overall contrast of the scaled thermal image also change. Furthermore, too high- or low- temperature objects lead to a zero-contrast image like indoor images.

Therefore, we propose a locally consistent thermal image scaling method to preserve a temporal consistency and increase image details for the image reconstruction process. The proposed scaling method is formulated as follows :

$$I^T_{t,t-1,t+1} = clamp\left(\frac{I^{T,raw}_{t,t-1,t+1} - \tau_{min}}{\tau_{max} - \tau_{min}}, \tau_{min}, \tau_{max}\right), \quad (6)$$

where the local min-max values (τ_{min}, τ_{max}) are defined as $\tau_{min} = \frac{1}{|N|}\sum_{n=1}^{N} percent(I^{T,raw}_n, \sigma)$ and $\tau_{max} = \frac{1}{|N|}\sum_{n=1}^{N} percent(I^{T,raw}_n, 1-\sigma)$. The local min-max values are adaptively decided by averaging over σ-th and $(1-\sigma)$-th percentile values of each image. We utilize the percentile values to exclude too high- and low- temperature

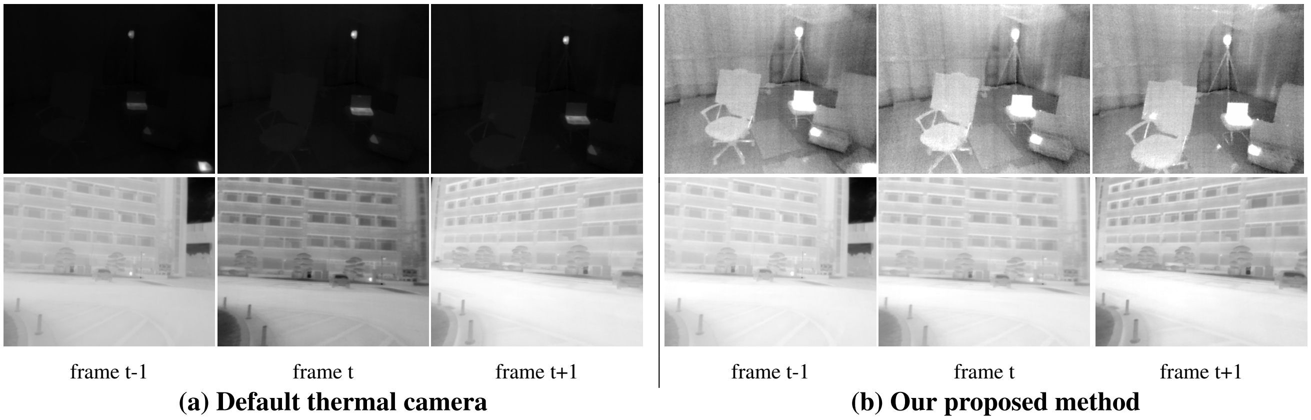


Figure 3: **Locally consistent thermal image scaling**. Typical thermal camera produce a relative scale thermal image in a default setting (a). Therefore, as the temperature distribution within a scene change, the overall contrast of the thermal image also change (Bottom left images). In addition, too high- or low- temperature measurement leads zero-contrast image like indoor images (Top left images).

observation. After that the local min-max values are used to generate locally consistent scaled thermal images. The $clamp(\cdot)$ function clamps a value between an upper and lower bound. We use a RAW thermal image as a network input.The locally consistent scaled images are used for the reconstruction and smoothness loss calculation.

3.3.3 Smoothness Loss

As the image reconstruction loss usually does not provide informative self-supervision in low-texture and homogeneous regions, we regularize the estimated depth map to have smooth property by adding edge-aware smoothness loss L_{sm} [11].

$$L_{sm} = \sum_{p} |\nabla D_t| \cdot e^{-|\nabla I_t|}, \tag{7}$$

where ∇ is first differential operator along spatial direction.

3.3.4 Geometric Consistency Loss

Geometric consistency loss L_{gc} [2] regularizes the estimated depth maps (D_t, D_s) to have scale-consistent 3D structure by minimizing geometric inconsistency. The geometry consistency loss L_{gc} and inconsistency map D_{diff} are defined as follows :

$$L_{gc} = \frac{1}{|V_p|} \sum_{p \in V_p} D_{diff}, \quad D_{diff} = \frac{|\tilde{D}_t - D_t^{'}|}{\tilde{D}_t + D_t^{'}}, \tag{8}$$

where $\tilde{D}_t$ is the synthesized depth map by warping the source depth map D_s and relative pose $P_{t \to s}$. $D_t^{'}$ is the interpolated depth map of D_t to share the same coordinate with the synthesized depth map $\tilde{D}_t$.

3.3.5 Invalid Pixel Masking

We filtered out invalid reconstruction signals by checking depth consistency [3] and static pixel [11] as follows:

$$L_{rec} = \frac{1}{|V_p|} \sum_{V_p} M_{self} \cdot M_{auto} \cdot L_{pe}(I_t, \tilde{I}_t), \tag{9}$$

where self discovery mask M_{self} [3] excludes moving objects and occluded regions defined as $M_{self} = 1 - D_{diff}$, auto mask M_{auto} [11] excludes the static and low-texture pixels which remains the same between adjacent frames, defined as $M_{auto} = \Big[L_{pe}(I_t^{eh}, \tilde{I}_t^{eh}) < L_{pe}(I_t^{eh}, I_s^{eh})\Big]$, V_p stands for valid points that are successfully projected from I_s to the image plane of I_t, and $|V_p|$ defines the number of points in V_p. Lastly, $[\cdot]$ is the Iverson bracket.

4. Experimental Results

4.1. Implementation Details

4.1.1 Dataset

We utilize ViViD benckmark dataset [23] to evaluate our proposed method. ViViD dataset [23] provides various sensor data streams; a thermal camera, an RGB-D camera, an event camera, and Lidar information. Also, the dataset consists of 10 indoor sequences and 4 outdoor sequences. Each sequence is taken under different lighting and motion conditions. To train monocular depth network, We follow the dataset split used in Shin *et al*. [37]. The indoor training set consists of 5 well-lit image sequences, and the remaining sequences are divided into indoor well-lit and zero-light(dark) evaluation sets. The outdoor training set consists of 2 day-light sequences, and the remaining sequences are used for the outdoor night evaluation set.

4.1.2 Network Architecture

We utilize ResNet-18 backbone [17] as domain specific feature extractors, decoder part of DispResNet [35] as a domain shared depth decoder, PoseNet [35] as a pose decoder, and discriminator of PatchGAN [19] as a feature space discriminator ψ. The first layer of the thermal feature extractor is modified to take single-channel thermal image. The RGB domain networks are initialized with the KITTI dataset [10] pre-trained weights to leverage the large-scale dataset trained task-specific knowledge by following common UDA strategy.

4.1.3 Training Setup

We utilize the PyTorch library [33] to implement our proposed method. We trained a depth network for the 200 epochs on the single RTX Titan GPU with 24GB memory. We take about 12 hours to train the depth and pose networks with a batch size 8. During the training, we used a pose network as an auxiliary network to exploit self-supervised loss. The hyper-parameters for the loss function are set to as follows. The scale values (λ_{gc}, λ_{sm}, γ, and λ_{adv}) are set to 0.5, 0.1, 0.85, and $2e^{-5}$. The percentile value σ is set to 1%. The discriminator loss L_ψ is also multiplied with the scale factor $2e^{-5}$. We utilize three Adam optimizer [22] to train the depth, pose, and discriminator networks. Two optimizers are used for the depth and pose network of RGB branch and thermal branch networks. The other one is used for the discriminator network. The learning rates of RGB, thermal, and discriminator optimizers are set to $1e^{-6}$, $1e^{-4}$, and $1e^{-6}$. We utilize random crop and horizontal flip for the data augmentation of both RGB and thermal images.

4.2. Single-view Depth Estimation Results

We compare our proposed method with the state-of-the-art self-supervised depth networks [2, 3, 37] to validate the effeteness of our method. Note that we cannot reproduce the previous works [20, 27] because they don't release their source code and needs paired stereo RGB and thermal images with a specific condition. The supervised baselines, such as DispResNet [35] and Midas-v2 [34], provides an upper bound of the self-supervised learning network.

The experimental results are shown in Tab. 1 and Fig. 4. Overall, the RGB image based depth networks (*i.e.*, RGB input of Tab. 1) records high accuracy and low error score in the well-lit indoor evaluation set. However, the performance significantly decreases when sufficient lighting condition is not guaranteed, such as indoor dark and outdoor night evaluation sets. On the other hand, the thermal image based depth networks (*i.e.*, Thermal input of Tab. 1) show consistent depth estimation performance regardless of illumination condition.

However, as shown in Fig. 4, the self-supervised monocular depth networks for thermal images (*i.e.*, Bian *et al*. and Shin *et al*. (T)) show inaccurate depth estimation results, especially in the indoor scenario. Depending on the surrounding environments, self-supervised loss of thermal image shows different aspects. Thermal image of indoor scenes generally has high noise and low contrast and leads to training failure. On the other hand, thermal image of outdoor scenes has relatively high contrast and low noise. Therefore, it can generate enough self-supervisory loss to the networks (*i.e.*, Bian *et al*. and Shin *et al*. (T) in outdoor results). However, both outdoor and indoor thermal images still doesn't contain enough texture, color, and contrast information compared to RGB images.

Therefore, additional information such as RGB video can be a great rescue for thermal image based depth network. Shin *et al*. [37] exploits paired RGB images, extrinsic parameters, and forward warping module to propagate image reconstruction loss of RGB images to the thermal image depth network. However, this method requires additional extrinsic calibration and synchronized data acquisition processes that typically require high expertise. On the other hand, the proposed method is not restrained by these processes because it learns from unpaired RGB and thermal video. Also, despite the absence of these processes, the proposed learning method shows outperformed or comparable depth estimation performance in all evaluation sets. Furthermore, as shown in Fig. 4, our method demonstrates clean and sharp depth map results via adversarial feature adaptation, compared to the previous state-of-the-art self-supervised depth networks.

4.3. Ablation Study

4.3.1 Self-supervised Learning of Unpaired RGB-Thermal Videos

We conduct ablation study about the self-supervised learning of unpaired RGB-thermal video, as shown in Tab. 2. For the baseline model (*i.e.*, *Baseline*), we trained depth networks with a self-supervised loss L^{thr}_{self} of thermal video only. After that, we design a network architecture that has modality specific encoders and shared depth decoder head to exploit unpaired RGB-thermal video. The model (1) are trained with the self-supervised losses of both RGB and thermal video (L^{rgb}_{self} and L^{thr}_{self}). The self-supervised learning of unpaired videos improves overall network performance by propagating of self-supervised loss of RGB video to the shared depth decoder. However, the thermal feature encoder cannot leverage the loss of RGB video and still suffer from lack of self-supervision signal.

Table 1: **Quantitative comparison of depth results on ViViD evaluation sets [23]**. We compare our network with state-of-the-art self-supervised depth networks [2, 3, 37]. Overall, $Ours$ shows outperformed and comparable results in all evaluation sets without requiring multi-sensor calibration and synchronized data acquisition. The best performance in each block is highlighted in **bold**.

(a) Depth estimation result on the ViViD indoor well-lit/zero-light testset.

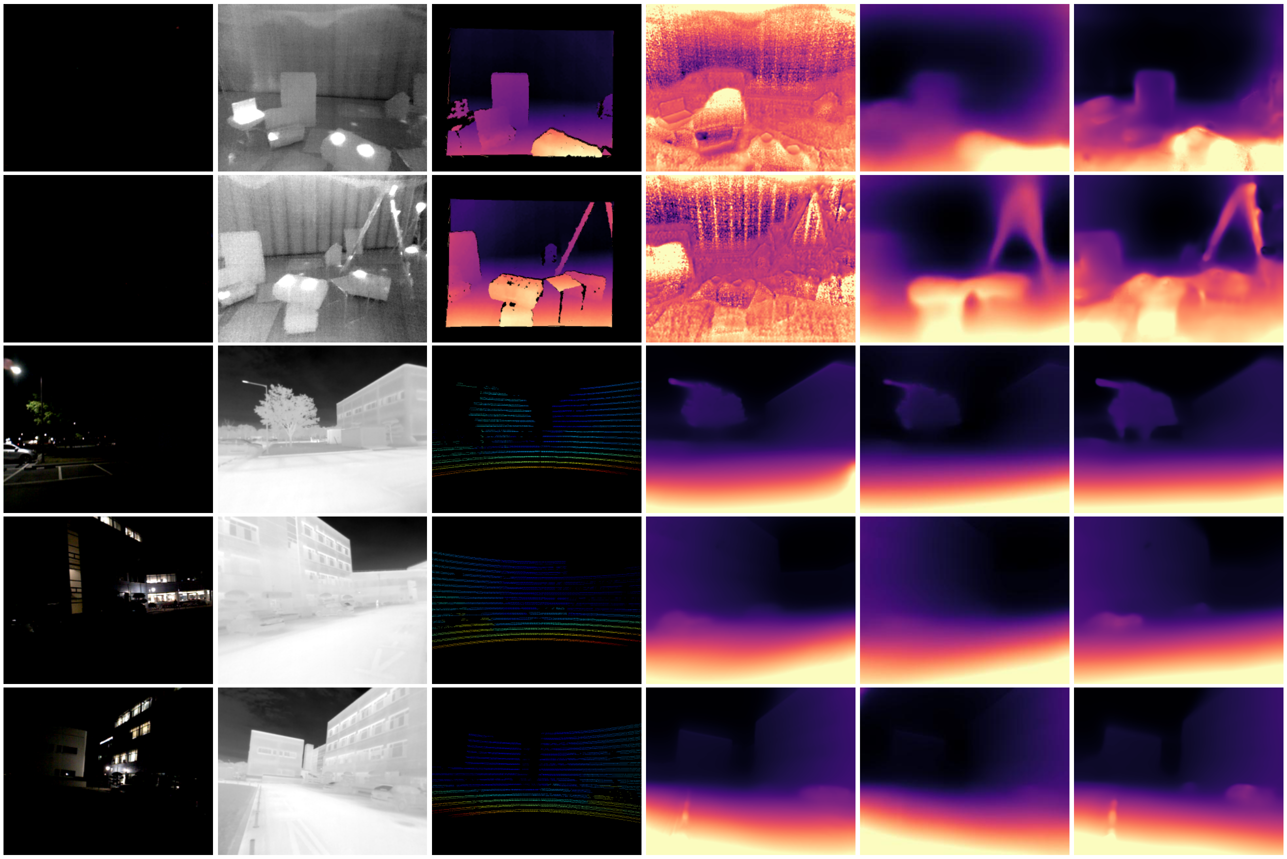
Scene	Methods	Input	Supervision	Cap	**Error ↓**				**Accuracy ↑**		
					AbsRel	SqRel	RMS	RMSlog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Indoor Well-lit	Midas-v2 (ResNext101) [34]	RGB	Depth	0-10m	0.194	0.348	0.370	0.210	0.928	0.979	0.991
	DispResNet (ResNet18)	Thermal	Depth	0-10m	0.117	0.097	0.462	0.170	0.869	0.960	0.991
	Midas-v2 (EfficientNet-Lite3)	Thermal	Depth	0-10m	0.062	0.044	0.282	0.107	0.946	0.983	0.995
	Midas-v2 (ResNext101)	Thermal	Depth	0-10m	**0.057**	**0.039**	**0.269**	**0.102**	**0.954**	**0.984**	**0.995**
	Bian *et al*. [2] (ver.NeurIPS)	RGB	RGB	0-10m	0.327	0.532	0.715	0.306	0.661	0.932	0.979
	Bian *et al*. [3] (ver.IJCV)	Thermal	Thermal	0-10m	0.274	0.317	0.897	0.316	0.544	0.840	0.969
	Shin *et al*. [37] (T)	Thermal	Thermal	0-10m	0.225	0.201	0.709	0.262	0.620	0.920	0.993
	Shin *et al*. [37] (MS)	Thermal	RGB&T	0-10m	**0.156**	**0.111**	**0.527**	**0.197**	0.783	**0.975**	**0.997**
	$Ours$	Thermal	RGB&T	0-10m	0.160	0.129	0.554	0.203	**0.793**	0.961	0.992
Indoor Dark	Midas-v2 (ResNext101) [34]	RGB	Depth	0-10m	0.351	0.545	0.766	0.327	0.624	0.875	0.976
	DispResNet (ResNet18)	Thermal	Depth	0-10m	0.124	0.094	0.466	0.174	0.854	0.963	0.992
	Midas-v2 (EfficientNet-Lite3)	Thermal	Depth	0-10m	0.060	0.036	0.273	0.105	0.950	0.985	0.996
	Midas-v2 (ResNext101)	Thermal	Depth	0-10m	**0.053**	**0.032**	**0.257**	**0.099**	**0.958**	**0.987**	**0.996**
	Bian *et al*. [2] (ver.NeurIPS)	RGB	RGB	0-10m	0.452	0.803	0.979	0.399	0.493	0.786	0.933
	Bian *et al*. [3] (ver.IJCV)	Thermal	Thermal	0-10m	0.277	0.311	0.866	0.318	0.540	0.833	0.967
	Shin *et al*. [37] (T)	Thermal	Thermal	0-10m	0.232	0.222	0.740	0.268	0.618	0.907	0.987
	Shin *et al*. [37] (MS)	Thermal	RGB&T	0-10m	0.166	0.129	0.566	0.207	0.768	0.967	0.994
	$Ours$	Thermal	RGB&T	0-10m	**0.160**	**0.124**	**0.547**	**0.202**	**0.789**	**0.969**	**0.994**

(b) Depth estimation result on the ViViD outdoor night testset.

Scene	Methods	Input	Supervision	Cap	**Error ↓**				**Accuracy ↑**		
					AbsRel	SqRel	RMS	RMSlog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Outdoor Night	Midas-v2 (ResNext101) [34]	RGB	Depth	0-80m	0.264	2.187	7.110	0.306	0.571	0.833	0.955
	DispResNet (ResNet18)	Thermal	Depth	0-80m	0.159	1.101	5.019	0.212	0.857	0.964	0.980
	Midas-v2 (EfficientNet-Lite3)	Thermal	Depth	0-80m	0.090	0.464	3.385	0.130	0.910	0.981	0.995
	Midas-v2 (ResNext101)	Thermal	Depth	0-80m	**0.078**	**0.369**	**3.014**	**0.118**	**0.933**	**0.988**	**0.996**
	Bian *et al*. [2] (ver.NeurIPS)	RGB	RGB	0-80m	0.617	9.971	12.000	0.595	0.400	0.587	0.720
	Bian *et al*. [3] (ver.IJCV)	Thermal	Thermal	0-10m	0.133	0.848	4.639	0.175	0.834	0.976	0.993
	Shin *et al*. [37] (T)	Thermal	Thermal	0-80m	0.157	1.179	5.802	0.211	0.750	0.948	0.985
	Shin *et al*. [37] (MS)	Thermal	RGB&T	0-80m	0.146	0.873	4.697	0.184	0.801	0.973	0.993
	$Ours$	Thermal	RGB&T	0-80m	**0.111**	**0.778**	**4.177**	**0.153**	**0.889**	**0.981**	**0.994**

Table 2: **Ablation study of the proposed method on ViViD outdoor evaluation set**. Our proposed method exploits two learning methods; self-supervised learning of unpaired multi-spectral videos and adversarial domain adaptation between multi-spectral features. We validate the effect of each component of our proposed method and another selectable option.

Model	Self Sup		Domain Adapt		**Error ↓**				**Accuracy ↑**		
	L^{thr}_{self}	L^{rgb}_{self}	$feat$	$pred$	AbsRel	SqRel	RMS	RMSlog	< 1.25	$< 1.25^2$	$< 1.25^3$
Baseline	✓				0.132	0.926	5.090	0.182	0.823	0.965	0.990
(1)	✓	✓			0.120	0.801	4.545	0.167	0.853	0.974	0.992
(2)	✓	✓		✓	0.118	0.802	4.561	0.165	0.862	0.977	0.993
(3)	✓	✓	✓(2^{nd})		**0.111**	**0.778**	**4.177**	**0.153**	**0.889**	**0.981**	**0.994**
(4)	✓	✓	✓(3^{rd})		0.137	0.986	5.029	0.185	0.820	0.964	0.990
(5)	✓	✓	✓(2^{nd})	✓	0.118	0.872	4.386	0.160	0.876	0.978	0.993
$Ours$	✓	✓	✓(2^{nd})		**0.111**	**0.778**	**4.177**	**0.153**	**0.889**	**0.981**	**0.994**



(a) RGB^{*} (b) Thermal (c) GT (d) Bian [3] (T) (e) Shin [37] (MS) (g) $Ours$

Figure 4: **Qualitative comparison of depth estimation results on ViViD dataset [23]**. Our method demonstrates clean and sharp depth map results via adversarial feature adaptation and self-supervised learning of unpaired multi-spectral video, compared to previous state-of-the-art self-supervised depth networks. *We visualize RGB images to show light conditions.

4.3.2 Feature-level Adversarial Domain Adaptation

We adopted the principal idea of domain adaptation to compensate for the insufficient self-supervision of the thermal encoder. There are two ways to leverage RGB domain information. We can provide self-supervision via prediction-level domain adaptation (*i.e.*, depth map) or feature-level domain adaptation (*i.e.*, feature vector). We found the domain adaptation in the first scale low-level feature (1^{st}) and high-level feature map (4^{th}) immediately converged to a trivial solution. It seems that this phenomenon occurs because too early low-level features or high-level features are too easy or difficult for the discriminator to distinguish at the beginning of training.

The prediction-level domain adaptation (2) leads to marginal performance improvement. On the other hand, feature-level domain adaptation (3) brings high performance boosting. We found the feature-level domain adaptation explicitly guides the thermal extractor to encompass representative feature extraction ability via adversarial loss between RGB and thermal features. Further analysis can be found in the supplementary material.

5. Conclusion

In this paper, we propose a novel training framework that combines self-supervised learning of unpaired multi-spectral images and adversarial multi-spectral feature adaptation for monocular depth estimation from thermal image. The proposed method aims to solve the weak self-supervision problem of thermal images by utilizing additional modality information without requiring multi-sensor calibration, synchronized data acquisition, and a specialized hardware setup. Based on the proposed method, the trained depth estimation network shows outperformed results than previous state-of-the-art networks.

Acknowledgment

This work was supported by Police-Lab 2.0 Program funded by the Ministry of Science and ICT(MSIT, Korea) and Korean National Police Agency(KNPA, Korea) [Project Name: AI System Development for a Image processing Based on Multi-Band(visible,NIR,LWIR) Fusion Sensing / Project Number: 220122M0500]

References

[1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021.

[2] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *Advances in neural information processing systems*, pages 35–45, 2019.

[3] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Zhichao Li, Le Zhang, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth learning from video. *International Journal of Computer Vision*, 129(9):2548–2564, 2021.

[4] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 1900–1909, 2019.

[5] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018.

[6] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1791–1800, 2019.

[7] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):934–948, 2018.

[8] Inc FLIR Systems. Users Manual FLIR Ax5 Series. [Online]. Available: `https://www.flir.com/globalassets/imported-assets/document/flir-ax5-usre-manual.pdf`, 2016.

[9] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018.

[10] 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.

[11] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019.

[12] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 2477–2486, 2019.

[13] Juan Luis Gonzalez and Munchurl Kim. Plade-net: Towards pixel-level accuracy for self-supervised single-view depth estimation with neural positional encoding and distilled matting loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6851–6860, 2021.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[15] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, 2020.

[16] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. In *ICLR*, 2020.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *Proc. of Int'l Conf. on Machine Learning (ICML)*, pages 1989–1998, 2018.

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[20] Namil Kim, Yukyung Choi, Soonmin Hwang, and In So Kweon. Multispectral transfer network: Unsupervised depth estimation for all-day vision. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[21] Yeong-Hyeon Kim, Ukcheol Shin, Jinsun Park, and In So Kweon. Ms-uda: Multi-spectral unsupervised domain adaptation for thermal image semantic segmentation. *IEEE Robotics and Automation Letters*, 6(4):6497–6504, 2021.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] Alex Junho Lee, Younggun Cho, Sungho Yoon, Youngsik Shin, and Ayoung Kim. ViViD : Vision for Visibility Dataset. In *ICRA Workshop on Dataset Generation and Benchmarking of SLAM Algorithms for Robotics and VR/AR*, Montreal, May. 2019. Best paper award.

[24] Seokju Lee, Francois Rameau, Sunghoon Im, and In So Kweon. Self-supervised monocular depth and motion learning in dynamic scenes: Semantic prior to rescue. *International Journal of Computer Vision*, 130(9):2265–2285, 2022.

[25] Taeyeop Lee, Byeong-Uk Lee, Inkyu Shin, Jaesung Choe, Ukcheol Shin, In So Kweon, and Kuk-Jin Yoon. Uda-cope: Unsupervised domain adaptation for category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14891–14900, 2022.

[26] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In

Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 6936–6945, 2019.

[27] Yawen Lu and Guoyu Lu. An alternative of lidar in nighttime: Unsupervised depth estimation based on single thermal image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3833–3843, 2021.

[28] Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li, and Liang Lin. Single view stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 155–163, 2018.

[29] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 4500–4509, 2018.

[30] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 3764–3773, 2020.

[31] Kwanyong Park, Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Preserving semantic and temporal consistency for unpaired video-to-video translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1248–1257, 2019.

[32] Kwanyong Park, Sanghyun Woo, Inkyu Shin, and In So Kweon. Discover, hallucinate, and adapt: Open compound domain adaptation for semantic segmentation. *Advances in Neural Information Processing Systems*, 33:10869–10880, 2020.

[33] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[34] Rene Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[35] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12240–12249, 2019.

[36] Ukcheol Shin, Kyunghyun Lee, Byeong-Uk Lee, and In So Kweon. Maximizing self-supervision from thermal image for effective self-supervised learning of depth and egomotion. *IEEE Robotics and Automation Letters*, 7(3):7771–7778, 2022.

[37] Ukcheol Shin, Kyunghyun Lee, Seokju Lee, and In So Kweon. Self-supervised depth and ego-motion estimation for monocular thermal video using multi-spectral consistency loss. *IEEE Robotics and Automation Letters*, 2021.

[38] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters (RAL)*, 4(3):2576–2583, 2019.

[39] Yuxiang Sun, Weixun Zuo, Peng Yun, Hengli Wang, and Ming Liu. Fuseseg: Semantic segmentation of urban scenes based on rgb and thermal data fusion. *IEEE Trans. on Automation Science and Engineering (TASE)*, 2020.

[40] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 7472–7481, 2018.

[41] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[42] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[43] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1164–1174, 2021.

[44] 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.

[45] 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.

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, pages 2223–2232, 2017.