Supplementary Material
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

1. Training Requirements of Previous Methods

Due to the weak self-supervisory signal of the thermal image, most previous works [6, 8, 10] utilize auxiliary self-supervision source (*i.e.*, RGB images) to train a depth estimation network.

Kim *et al*. [6] exploited spatial image reconstruction with paired stereo RGB images and an 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 [2]. We itemize the specifications and requirements of their proposed method.

- Estimate a depth map from a given thermal image.
- RGB stereo images for the spatial image reconstruction loss.
- Two RGB cameras, one thermal camera, and a beam splitter, which is used for the camera coordinate alignment.
- Synchronized image acquisition.

Lu *et al*. [8] also needs a specialized hardware system that consists of very closely located RGB stereo and thermal stereo cameras. They exploit an image translation network to synthesize a thermal-like left image from a left RGB image. After that, the spatial reconstruction loss between the thermal-like left and real right thermal images is used to train the depth network. The depth network takes a right thermal image and initial depth map estimated from stereo RGB images to estimate a depth map. We itemize the specifications and requirements of their proposed method.

- Estimate a depth map from a given thermal image and initial depth map.
- Left RGB image and right thermal image for the spatial image reconstruction loss.
- Stereo RGB images are used to estimate initial depth map via a stereo matching algorithm.
- Two RGB cameras and one thermal camera, which is closely located with the one RGB camera.
- Synchronized image acquisition.

Shin *et al*. [10] utilizes a temporal reconstruction loss with paired RGB-thermal images to train single-view depth and multiple-view pose networks. They utilize temporal image reconstruction loss of RGB and thermal sequences with a depth and pose estimated from thermal images. For this purpose, they proposed a forward depth and pose warping module that translates the coordinate system of depth map and relative camera pose from thermal image plane to the RGB image plane in a differentiable way. In this forward warping process, they need an extrinsic matrix between RGB and thermal cameras. We itemize the specifications and requirements of their proposed method.

- Estimate a depth map and relative camera pose from a given thermal image sequence.
- Temporal RGB and thermal image sequences for the temporal image reconstruction loss.
- One RGB camera and one thermal camera.
- Synchronized image acquisition / Extrinsic calibration between RGB and thermal cameras.

Compared to the previous methods, our proposed method doesn't requires any extra constraints such as specialized hardware (*i.e.*, beam splitter) [6], fixed sensor positioning [8], extrinsic calibration [10], and synchronized image acquisition [6, 8, 10]. The proposed method only requires unpaired RGB and thermal image sequences to train monocular depth network of thermal image. For this purpose, the proposed training framework effectively exploits both self-supervised learning of unpaired multi-spectral images and feature-level adversarial adaptation between multi-spectral images.

2. ViViD Dataset

We utilize ViViD benckmark dataset [7] to evaluate our proposed method. ViViD dataset [7] 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. Depending on the strength of the movement, they define $robust$ as a slow-motion sequence that doesn't contain any dynamic movement, $aggressive$ as a fast-motion sequence that contains lots of dynamic movement, and $unstable$ as a mixture of $robust$ and $aggressive$. Also, they define $global$ and day as a well-lit lighting condition with a complete light system, $local$ as a relatively weak light condition with a few light sources, and $dark$ and $night$ as no external light source. However, there is some weak light source in the outdoor night scene.

Shin *et al.* [10] split the dataset into the in/outdoor training-and-testing subsets to train and validate a monocular depth network. Their indoor training set consists of five sequences with well-light conditions; indoor-robust-global, indoor-robust-local, indoor-aggressive-global, indoor-unstable-global, and indoor-unstable-local. They make two indoor testing sets; one is a well-lit test set, and the other one is a bad-light (*i.e.*, dark) test set. The indoor well-lit testing set consists of two sequences; indoor-robust-varying and well-lit images of indoor-aggressive-local. The indoor bad-light testing set consists of three sequences; indoor-robust-dark, indoor-aggressive-dark, and indoor-unstable-dark.

The outdoor division is also similar; they divide the outdoor set into well-light and bad-light conditions. The outdoor training set contains the day sequences; outdoor-robust-day1 and outdoor-robust-day2. Outdoor testing set contain the night sequences; outdoor-robust-night1 and outdoor-robust-night2. The total data samples of each training and testing dataset are as follows. The indoor training set consists of 2,124 RGB and thermal image pairs. The indoor bad-light testing set consists of 1,201 pairs, and the well-lit testing set consists of 478 pairs. The outdoor training set consists of 2,213 pairs. The outdoor testing set consists of 2,019 pairs.

3. Evaluation Metric

We utilized the depth evaluation metrics commonly used to measure the accuracy and error of depth estimation results [11, 3, 9]. Also, we applied NYU v2 [11] and KITTI [4] evaluation settings for indoor and outdoor evaluation set, respectively. The depth evaluation metric measure the difference between the ground-truth depth D_{gt} and the predicted depth from our network D_{pred} on the valid pixel set V of D_{gt}. Since the monocular depth network estimate a relative scale depth, the scale of D_{gt} is used to recover the scale of D_{pred} before measuring the differences.

4. Further Discussion of Experimental Results

4.1. Adversarial Multi-spectral Domain Adaptation

Here, we further describe the effect of feature-level domain adaptation between multi-spectral images. As described in the main paper, we have two options to provide self-supervision via domain adaptation; prediction-level domain adaptation (*i.e.*, depth map) and feature-level domain adaptation (*i.e.*, feature vector).

4.1.1 Prediction-level Adversarial Domain Adaptation

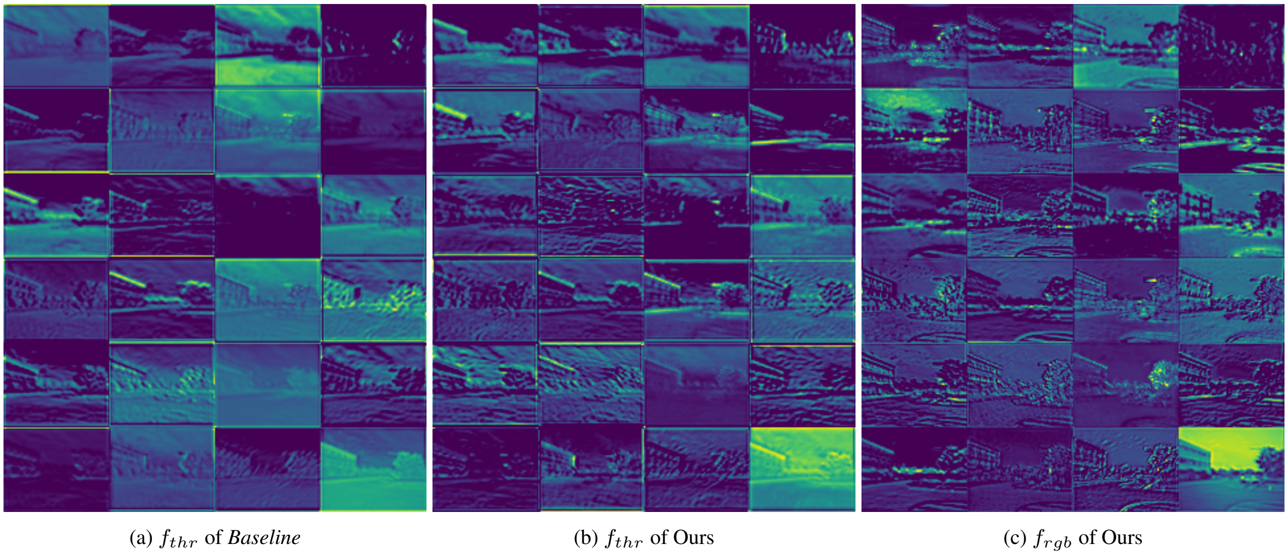
Prediction-level domain adaptation minimizes the domain gap between the depth maps predicted from RGB and thermal images instead of feature maps. Therefore, the losses (L_{adv} and L_{dis}) take depth maps of each modality as an input instead of feature maps and are propagated to the depth decoder and feature encoder. However, the prediction-level domain adaptation leads to marginal performance improvement compared to feature-level adaptation (See Table2 of main manuscript). We believe this is because, during the back-propagating process, the adversarial loss is getting weakened and not enough to train the thermal encoder. Also, the depth decoder is less affected by the prediction-level adversarial loss since the decoder already leverages sufficient training signal from self-supervised losses of unpaired RGB-thermal videos.

4.1.2 Feature-level Adversarial Domain Adaptation

On the other hand, feature-level domain adaptation (*i.e.*, (3) of Table2) 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. The feature map visualization also supports this result as shown in Fig. 1.

Baseline model trained with thermal image reconstruction loss (See Table2) tends to extract homogeneous and monotonous feature maps, compared to the RGB image feature f_{rgb}. Thermal image contains almost the same structure information compared to an RGB image. However, its contrast between structures is too weak to provide enough self-supervision signal. Therefore, the weak self-supervision signal leads to a weak feature representation ability.

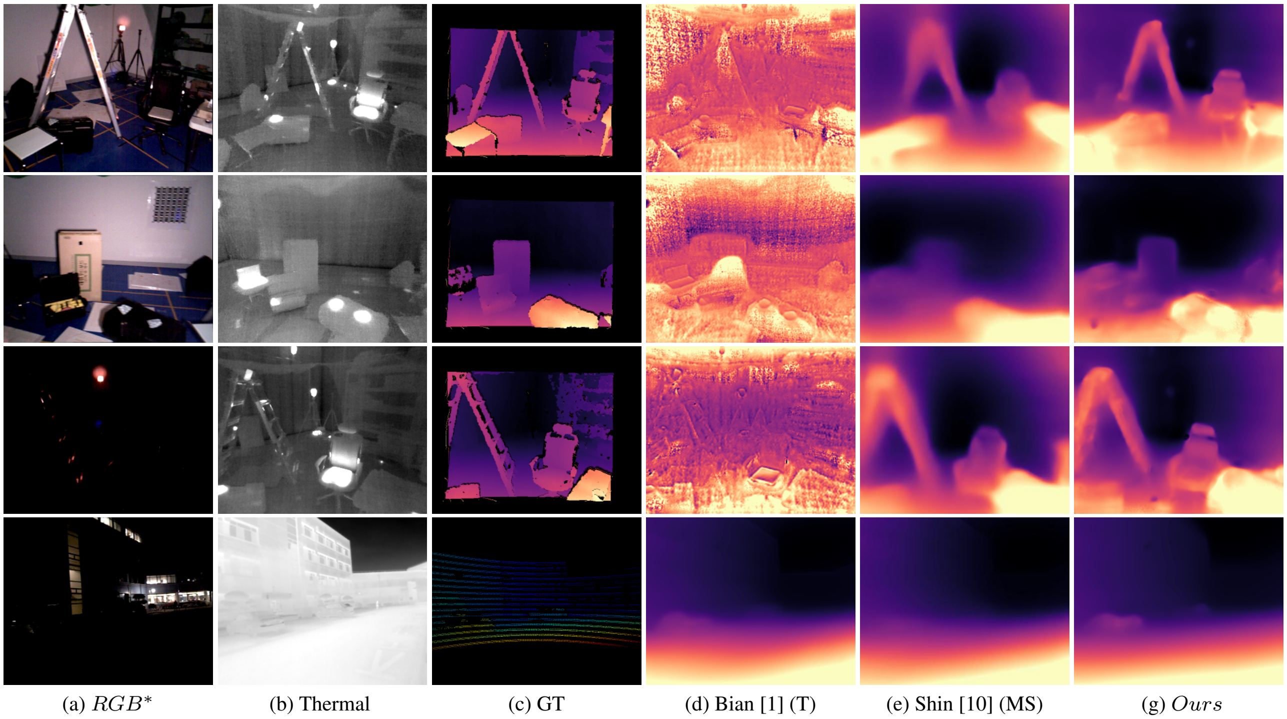
On the other hand, the proposed adversarial multi-spectral feature adaptation leads to extracting informative and high-textured feature maps similar to RGB feature maps, as shown in (b) and (c) of Fig. 1. Also, this representative feature map leads to the edge-preserved depth estimation results, as shown in Fig. 2.


(a) f_{thr} of *Baseline*

(b) f_{thr} of Ours

(c) f_{rgb} of Ours

Figure 1: **Qualitative comparison of the feature maps (f_{thr} and f_{rgb}).** From right to left, we visualizes half channel of 2^{nd} scale feature map for Baseline model (a), thermal encoder of $Ours$ (b), and RGB encoder of $Ours$ (c). Without our proposed training method, the thermal encoder tends to extract homogeneous and monotonous feature maps. On the other hand, the proposed method allows the thermal encoder to extract information and high-textured feature maps similar to RGB feature maps via adversarial feature adaptation.


(a) RGB^{*}

(b) Thermal

(c) GT

(d) Bian [1] (T)

(e) Shin [10] (MS)

(g) $Ours$

Figure 2: **Qualitative comparison of depth estimation results on ViViD dataset [7]**. 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.

4.2. Self-supervised Learning of Thermal Videos

As shown in Fig. 2, Bian *et al*. [1], which is trained with thermal video only, shows un-delightful results especially in the indoor scenario. We found this is caused by the high-level noise of indoor thermal images. Differ from the outdoor scene, indoor scene generally has a tiny temperature range (*i.e*., most objects have a similar temperature). Therefore, in the process of raw thermal image visualization, the sensor noise is also amplified, as shown in the indoor thermal images of Fig. 2.

In order to train the network based on these kinds of images, we need to filter out or handle the noisy pixels for accurate image reconstruction loss. However, we found the per-pixel auto-mask scheme used in Bian *et al*. [1] highly affected by the high-level noise and hard to filter out the noisy pixel. On the other hand, the original auto-mask implementation [5] that calculates the mask based on the combination of SSIM and L1 loss shows more reliable results. Therefore, we follow the original auto-mask implementation [5] to handle the noisy indoor thermal images.

References

[1] 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.

[2] 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.

[3] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 4796–4803. IEEE, 2018.

[10] 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.

[11] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012.