Deep Depth Estimation from Thermal Image

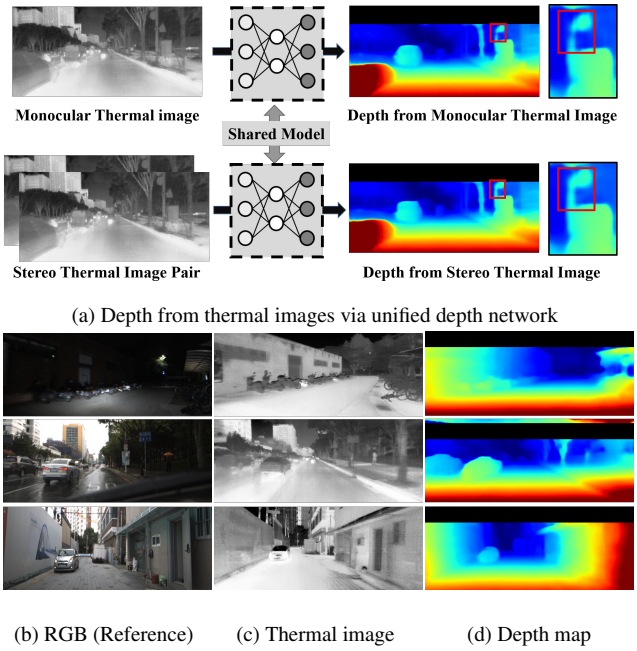
Ukcheol Shin
KAIST
shinwc159@gmail.com

Jinsun Park
Pusan National University
jspark@pusan.ac.kr

In So Kweon
KAIST
iskweon77@kaist.ac.kr

Abstract

Robust and accurate geometric understanding against adverse weather conditions is one top prioritized conditions to achieve a high-level autonomy of self-driving cars. However, autonomous driving algorithms relying on the visible spectrum band are easily impacted by weather and lighting conditions. A long-wave infrared camera, also known as a thermal imaging camera, is a potential rescue to achieve high-level robustness. However, the missing necessities are the well-established large-scale dataset and public benchmark results. To this end, in this paper, we first built a large-scale Multi-Spectral Stereo (MS^2) dataset, including stereo RGB, stereo NIR, stereo thermal, and stereo LiDAR data along with GNSS/IMU information. The collected dataset provides about 195K synchronized data pairs taken from city, residential, road, campus, and suburban areas in the morning, daytime, and nighttime under clear-sky, cloudy, and rainy conditions. Secondly, we conduct an exhaustive validation process of monocular and stereo depth estimation algorithms designed on visible spectrum bands to benchmark their performance in the thermal image domain. Lastly, we propose a unified depth network that effectively bridges monocular depth and stereo depth tasks from a conditional random field approach perspective. Our dataset and source code are available at https://github.com/UkcheolShin/MS2-MultiSpectralStereoDataset.



(a) Depth from thermal images via unified depth network

(b) RGB (Reference) (c) Thermal image (d) Depth map

Figure 1. **Depth from thermal images in various environments.** Our proposed network can estimate both monocular and stereo depth maps regardless of given a single or stereo thermal image via unified network architecture. Furthermore, depth estimation results from thermal images show high-level reliability and robustness under day-light, low-light, and rainy conditions.

1. Introduction

Recently, a number of researches have been conducted for accurate and robust geometric understanding in self-driving cars based on the widely-used benchmark datasets, such as KITTI [15], DDAD [17], and nuScenes [4]. Modern computer vision algorithms deploy a deep neural network and data-driven machine learning technique to achieve high-level accuracy, which needs large-scale datasets. However, from the perspective of robustness in real-world, the algorithms mostly rely on visible spectrum images and are easily degenerated by weather and lighting conditions.

Therefore, recent works have actively investigated alternative sensors such as Near-Infrared (NIR) cameras [39], LiDARs [16, 51], radars [14, 32], and long-wave infrared (LWIR) cameras [35, 45] to achieve reliable and robust geometric understanding in adverse conditions. Among these alternative and complementary sensors, LWIR camera (*i.e.*, thermal camera) has become more popular because of its competitive price, adverse weather robustness, and unique modality information (*i.e.*, temperature). Therefore, various thermal image based computer vision solutions [3, 21–23, 27, 35, 45, 47–50] to achieve high-level robustness have been actively attracting attention recently.

Table 1. **Comprehensive comparison of multi-modal datasets**. Compared to previous datasets [6, 8, 24, 25, 28, 53], the proposed Multi-Spectral Stereo (MS^2) dataset provides about 195K synchronized and rectified multi-spectral stereo sensor data (*i.e.*, RGB, NIR, thermal, LiDAR, and GNSS/IMU data) covering diverse locations (*e.g.*, city, campus, residential, road, and suburban), times (*e.g.*, morning, daytime, and nighttime), and weathers (*e.g.*, clear-sky, cloudy, and rainy).

Dataset	Year	Environment	Platform	Total # of Data Pairs	LiDAR	IMU	RGB		NIR		Thermal			Weather		
							Mono	Stereo	Mono	Stereo	Mono	Stereo	RAW	Daytime	Nighttime	Rain
CATS [53]	2017	In/Outdoor	Handheld	1.4K	✓	✓	✓	✓	×	×	✓	✓	✓	✓	✓	×
KAIST [6]	2018	Outdoor	Vehicle	Unknown	✓	✓	✓	✓	×	×	✓	×	✓	✓	✓	×
ViViD [25]	2019	In/Outdoor	Handheld	5.3K/4.3K	✓	✓	✓	×	×	×	✓	×	✓	✓	✓	×
MultiSpectralMotion [8]	2021	In/Outdoor	Handheld	121K/27.3K	✓	✓	✓	×	✓	×	✓	×	✓	✓	✓	×
ViViD++ [24]	2022	Outdoor	Vehicle	56K	✓	✓	✓	×	×	×	✓	×	✓	✓	✓	×
OdomBeyondVision [28]	2022	Indoor	Handheld/ UGV/UAV	71K/ 117K/21K	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	×
Ours	2022	Outdoor	Vehicle	195K	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

However, the missing necessities are the well-established large-scale dataset and public benchmark results. The publicly available datasets for autonomous driving are overwhelmingly composed of the visible spectrum band (*i.e.*, RGB images), but it very rarely includes other spectrum bands, such as the NIR band and LWIR band. Especially, despite the advantage of the LWIR band, just a few LWIR datasets have been recently released. However, these datasets are indoor oriented [8, 25, 28], small scale [25, 53], publicly unavailable [6], or limited sensor diversity [6, 24]. Therefore, the necessity is getting increase to design a large-scale multi-sensor driving dataset to investigate the feasibility and challenges associated with an autonomous driving perception system from multi-spectral sensors.

The other necessity is thoroughly validating vision applications on the LWIR band. Estimating a depth map from monocular and stereo images is one fundamental task for geometric understanding. Despite numerous recent studies in depth estimation, these works have mainly focused on depth estimation using RGB images. However, thermal images, which typically have lower resolution, less texture, and more noise than RGB images, could pose a challenge for stereo-matching algorithms. This means that the performances of these previous works in thermal image domains are uncertain and may not be guaranteed.

To this end, in this paper, we provide a large-scale multi-spectral dataset along with exhaustive experimental results and a new perspective of depth unification to encourage active research of various geometry algorithms from multi-spectral data to achieve high-level performance, reliability, and robustness against hostile conditions. Our contributions can be summarized as follows:

- We provide a large-scale Multi-Spectral Stereo (MS^2) dataset, including stereo RGB, stereo NIR, stereo thermal, and stereo LiDAR data along with GNSS/IMU data. Our dataset provides about 195K synchronized data pairs taken from city, residential, road, campus, and suburban areas in the morning, daytime, and nighttime under clear-sky, cloudy, and rainy conditions.
- We perform exhaustive validation and investigate that monocular and stereo depth estimation algorithms originally designed for visible spectral bands work reasonably in thermal spectral bands.
- We propose a unified depth network that bridges monocular depth and stereo depth estimation tasks from the perspective of a conditional random field approach.

2. Related Work

2.1. Thermal Image Dataset for 3D Vision

A well-established large-scale dataset is the most fundamental and top priority for modern deep neural network training. For the visible spectrum band, numerous large-scale datasets have been proposed such as KITTI [15], DDAD [17], Cityscape [7], Oxford [36], and nuScenes [4] datasets. On the other hand, the InfraRed (IR) spectrum band (*e.g.*, near-IR, short-wave IR, long-wave IR) is very rarely included in just a few datasets in a limited form despite its superior environmental robustness.

The comprehensive comparison is shown in Tab. 1. Most datasets are insufficient to investigate the feasibility of geometric and semantic understanding from multi-spectrum image sensors under diverse outdoor driving scenarios. More specifically, these datasets are indoor oriented [8, 25, 28], small scale [25, 53], publicly unavailable [6], limited sensor diversity [6, 24], limited weather condition [6, 24, 25], or missing RAW thermal data [53].

2.2. Depth From Visible Spectrum Band

Monocular Depth Estimation (MDE) has high-level universality because it estimates depth map from a single image. There have been numerous mainstream methods formulating depth estimation as per-pixel regression [26, 41, 42, 56] by directly estimating per-pixel depth value through a neural network, per-pixel classification [12, 13] by discretizing continuous depth range into discrete intervals, and classification-and-regression problems [2, 29].

However, MDE is an ill-posed problem; a single 2D image can be generated from an infinite number of distinct 3D scenes. Therefore, the estimated monocular depth map is inherently scale-ambiguous, has low generalization performance, and provides lower performance than depth estimation from multi-view images.

Stereo Depth Estimation (SDE) can estimate metric-scale depth map by utilizing a known camera baseline and disparity map from a rectified stereo image pair. Existing stereo matching networks can be categorized into 3D cost volume [30, 37, 52, 55] and 4D cost volume based methods [5, 18, 20, 43, 54]. The former one estimates a single-channel cost volume (*e.g.*, D×H×W) by measuring the similarity between left and right features. Then, they aggregate the contextual information via 2D convolution. These methods have high memory and computational efficiency, yet the encoded volume loses large content information leading to unsatisfactory accuracy.

The latter one builts multiple-channel cost volume (*e.g.*, D×C×H×W) by concatenating two left-right feature volumes [5,20], correlation-volume and left-right features [18], or attention-added features [54]. Then, they aggregate the 4D cost volume with 3D convolution layers. Current state-of-the-art models are mostly based on this method. However, this demands high memory consumption and cubic computational complexity that is expensive to deploy in a real-world application. The SDE task yields significant performance gains compared to the MDE task, yet the SDE task is still struggling to find accurate corresponding points in inherently ill-posed regions such as occlusion areas, repeated patterns, textureless regions, and reflective surfaces.

2.3. Depth From Thermal Spectrum Band

Thermal spectrum band has high-level robustness against various adverse weather and lighting conditions, such as rain, fog, dust, haze, and low-light conditions. However, due to the absence of a large-scale dataset, most previous studies for geometric understanding [3, 10, 21, 38, 47] are conducted on their own testbed. Also, most works focus to utilizes a thermal camera along with other heterogeneous sensors for the target geometric task rather than focusing on the thermal camera itself.

For the geometric understanding task that utilizes a deep neural network, a few researches [22, 35, 44–46] have been proposed recently. Most studies focus on the self-supervised depth estimation from thermal images with auxiliary modality guidance, such as aligned-and-paired RGB images [22], style transfer network [35], and paired RGB images [45]. Unlike the previous studies, in this paper, we target a supervised depth estimation from a single and stereo thermal image that has not yet been actively explored.

Table 2. **Sensor specification for the multi-spectral stereo system**. Our sensor system consists of RGB, NIR, thermal, and LiDAR stereo system along with a GNSS/IMU module. The data from RGB, NIR, and thermal stereo system was taken at 15 fps with synchronized signals. Lidar stereo data were taken at 10 fps.

Sensor	Model	Frame Rate	Characteristics
RGB camera	PointGrey BlackFly-S BFS-U3-51S5C	Max 75 fps	2448×2048 pixel Global Shutter
	Kowa LM5JC10M		82.2° (H) × 66.5°(V) FoV
NIR camera	Intel RealSense D435i	Max 90 fps	1280 × 720 pixel Global Shutter 69° (H) × 42°(V) FoV
Thermal camera	FLIR A65C	Max 30 fps	640×512 pixel 45° (H) × 37°(V) FoV Uncooled VOX microbolometer 16-bit Raw data
LiDAR	Velodyne VLP-16	Max 20 fps	Accuracy: ± 3 cm Measurement range : 100m 360° (H), ±15° (V) FoV
GNSS/IMU	LORD Microstrain 3DM-GX5-45	10/100 Hz	Position, Velocity, Attitude, Acceleration, etc.

3. Multi-Spectral Stereo (MS^2) Dataset

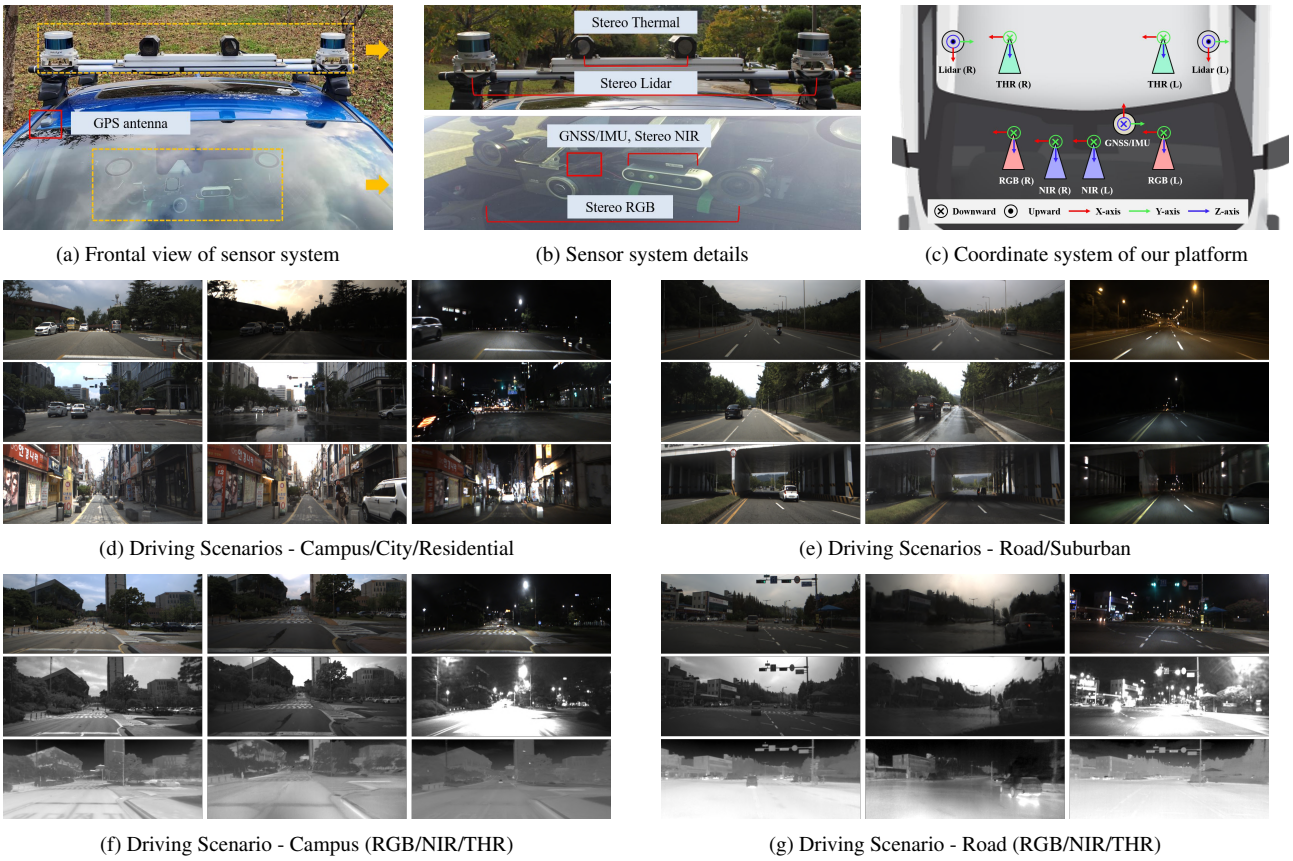
3.1. Multi-Spectral Stereo Sensor System

Despite the well-known advantages of the long-wave infrared camera (*i.e.*, thermal camera) [9, 19, 57], the absence of a large-scale dataset still interrupts the development and investigation of condition-agnostic autonomous driving perception systems from the thermal spectrum domain. To this end, we designed a data collection platform that consists of RGB, NIR, thermal, and LiDAR stereo system along with a GNSS/IMU module, as shown in Fig. 2-(a),(b), and (c). Each sensor specification is described in Tab. 2.

Accurate time-synchronization is one important prerequisite for various geometric tasks with multiple sensors, such as depth estimation, odometry, 3D detection, and 3D reconstruction. Therefore, we synchronize RGB and NIR stereo cameras via an external synchronizer. Thermal stereo cameras are synchronized with the sync signal of the left thermal camera. Also, a software trigger is used to synchronize the two systems at the start time of each data acquisition. Please refer to the supplementary material for more details on calibration and sensor system configuration.

3.2. Data Collection

We collect multi-spectral stereo data (*i.e.*, stereo RGB, NIR, thermal, and LiDAR data) along with GNSS/IMU data under various locations, lighting conditions, and weather conditions. Specifically, we obtain the synchronized multi-spectral data from campus, city, residential area, suburban area, and multiple road environments. Also, we provide various time diversities (*e.g.*, morning, daytime, and nighttime) and weather diversities (*e.g.*, clear-sky, cloudy, and rainy) for each representative location (Fig. 2-(d) and (e)).



(a) Frontal view of sensor system (b) Sensor system details (c) Coordinate system of our platform

(d) Driving Scenarios - Campus/City/Residential (e) Driving Scenarios - Road/Suburban

(f) Driving Scenario - Campus (RGB/NIR/THR) (g) Driving Scenario - Road (RGB/NIR/THR)

Figure 2. **Overview of our proposed Multi-Spectral Stereo (MS^2) outdoor driving dataset.** We designed a data collection platform that consists of RGB, NIR, thermal, and LiDAR stereo system along with a GPS/IMU module (*i.e.* (a),(b),(c)). The collected dataset are taken under locations of campus, city, residential area, road, and suburban with various time slots (morning, day, and night) and weather conditions (clear-sky, cloudy, and rainy) (*i.e.* (d) and (e)). According to the surrounding conditions, each spectrum sensor shows different aspects, advantages, and disadvantages induced by their sensor characteristics (*i.e.*, (f) and (g)). Further examples and details are described in the supplementary material.

This aims to investigate and evaluate the generalization and domain gap handling abilities of a deep neural network. It also targets to explore the possibility of multi-sensor complementation and the characteristics of each sensor under various conditions (Fig. 2-(f) and (g)). Compared to previous datasets [6, 8, 24, 25, 28, 53], the proposed dataset provides about 195K synchronized and rectified multi-spectral data pairs (*i.e.*, RGB, NIR, thermal, LiDAR, and GNSS/IMU data) covering diverse locations, times, weathers, and sensors.

3.3. Multi-Spectral Stereo (MS^2) Depth Dataset

Ground-Truth Generation Process. To create a dense Ground-Truth (GT) depth map, we accumulated 10 successive stereo LiDAR data by utilizing interpolated odometry information from GNSS/IMU sensor in a similar way to KITTI dataset [15]. Specifically, we calculate every pose information of each sensor's time stamp by interpolating GNSS/IMU sensor data. Afterward, we aggregate 10 successive stereo LiDAR data for each target thermal image via transformation matrices between consecutive data and refine the aggregated point cloud via the Iterative Closest Point (ICP) algorithm [1]. Then, the refined and aggregated 3D point cloud is projected to the thermal image plane to get the final semi-dense depth map.

Training Set Configuration. From the MS^2 dataset, we periodically sampled the thermal images and filter out the static vehicle movement to make training, validation, and evaluation splits for the learning of monocular and stereo depth networks. We utilize 26K data pairs for training, 4K pairs for validation, and 5.8K, 6.8K, and 5.2K pairs for evaluation of daytime, nighttime, and rainy conditions. We make the training set splits have almost zero overlap in time, weather, and location diversity. The split details can be found in the supplementary material.

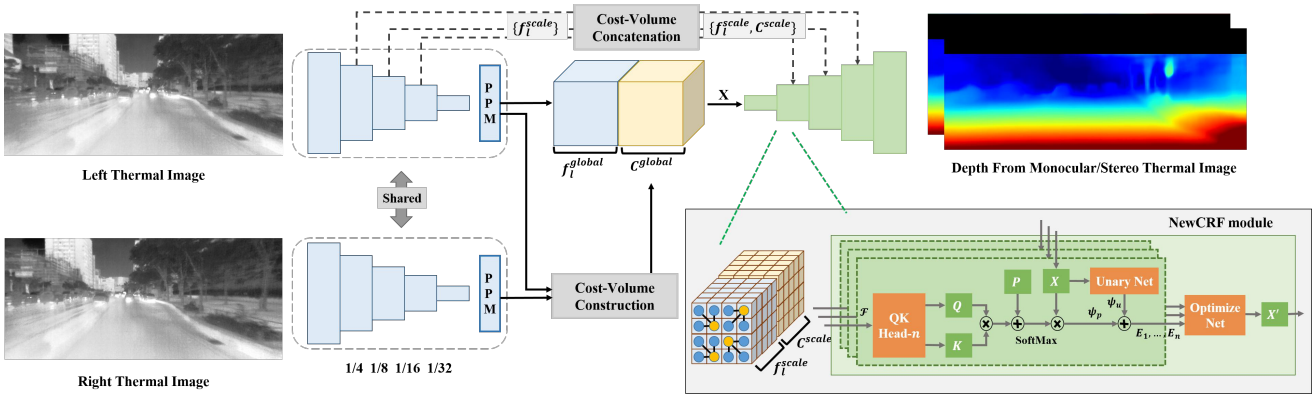


Figure 3. **Overall pipeline of our proposed depth estimation network.** We design a single network that can estimate both monocular and stereo depth maps from given a single or stereo thermal image. We bridge monocular depth and stereo depth estimation by regarding the cost-volume as additional information for Neural Window Conditional Random Field (NeWCRF) block [56]. Initially, the network extracts multi-scale feature maps via Swin-Transformer backbone model [31] and aggregates the global contextual information via Pyramid Pooling Module (PPM) head [58]. If the right thermal image is available, the network generates each scale of single-channel cost-volume (*i.e.*, $D^{scale} \times H^{scale} \times W^{scale}$) based on feature similarity of the left-right features. If only the left image is available, the network utilizes zero-filled cost-volume. The depth maps are estimated from the multi-scale concatenated features via NeWCRF blocks [56].

4. Depth Estimation from Thermal Image

4.1. Bridging Monocular and Stereo Depth Estimation

In this section, we connect the Monocular Depth Estimation (MDE) and Stereo Depth Estimation (SDE) tasks via the Conditional Random Field(CRF) perspective. MDE network has the advantage of high-level universality that doesn't need extra constrain such as pre-rectification, extrinsic matrix information, and additional images. However, MDE networks suffer from inherent scale ambiguity and generalization issues. On the other hand, SDE networks provide an accurate metric-scale depth map by finding horizontal correspondences between rectified left-and-right images. But, the SDE network is hard to provide a reliable depth map in the ill-posed regions such as occlusion areas, repeated patterns, textureless regions, and reflective surfaces.

They can complement each other by bridging two tasks and, at the same time, flexibly estimate depth maps from given monocular or stereo images, as shown in Fig. 3. To this end, we utilize the recently proposed MDE network, Neural Window FC-CRF (NeWCRF), to connect two tasks. Specifically, we regard the estimated cost volume as additional information for NeWCRF blocks. Therefore, when the right image is available, we add each cost volume of multi-scale left-and-right features to the left image feature F_L^{scale}. If only the left image is available, the network utilizes zero-filled cost volume.

4.2. Feature Extraction and Aggregation

We adopt Swin transformer [31] as our backbone network. The backbone network extract feature in four scale-level (*i.e.*, 1/4, 1/8, 1/16, and 1/32) from the given images. After that, the pyramid pooling module(PPM) [58] aggregates global context information with global average pooling of receptive fields 1, 2, 3, and 6 from the last scale-level. The features of remained scales are provided to each level of decoders via a skip-connected manner.

4.3. Cost Volume Construction

Most state-of-the-art stereo matching networks [5,18,54] utilize a 4D cost volume with 3D convolution layer to achieve higher performance. However, the 4D cost volume based method requires costly memory and computation consumptions. Also, the method makes it hard to associate monocular depth estimation in the network architecture by enforcing the utilization of both left-right feature maps always.

Therefore, we utilize correlation cost volume (*i.e.*, 3D cost volume) [30,37,52,55] that has a single-channel correlation map for each disparity level. The method loses some correlation information between left-right features, yet it can be easily associated with a monocular depth estimation network as additional information. The cost volume of each scale is estimated as follows:

$$C^{scale}(d,x,y) = \frac{1}{N_c} < f_l^{scale}(x,y), f_r^{scale}(x-d,y) >, \quad (1)$$

where $< \cdot, \cdot >$ is the inner product, N_c denotes the number of channels, and f_l^{scale} and f_r^{scale} are the feature map of

each scale. The cost volume of each scale is concatenated with the corresponding feature map of the left image f_l^{scale} to form skip-connection input $\mathcal{F}$ for the NeWCRF blocks.

4.4. Neural Window FC-CRF

NeWCRF [56] implements traditional CRF as the form of neural network in a computation efficient way by utilizing shifted window multi-head attention module [31]. Given the previous prediction result X and concatenated feature $\mathcal{F}$, the NeWCRF block estimate unary potential ψ_u and pairwise potential ψ_p via multi-head attention mechanism (*i.e.*, NeWCRF block of Fig. 3), as follows:

$$\psi_{p_u} = \theta_u(X), \sum_i \psi_{p_i} = \text{SoftMax}(Q \cdot K^T + P) \cdot X, \quad (2)$$

where θ_u is the parameter of a unary network and Q, K, P are query, key, and position embedding matrix of attention block. After that, the optimized net, which consists of two MLP layers, estimates the current stage result $X^{'}$. And the $X^{'}$ is regarded as X for the next NeWCRF block.

4.5. Disparity and Inverse Depth Prediction

The proposed network estimates four scale prediction results (*i.e.*, 1/4, 1/8, 1/16, and 1/32) from the last four NeWCRF blocks. When a single image is fed to the network, we regard the prediction results as an inverse depth map. For the stereo image pair, we regard the prediction results as a common disparity map. For the prediction features X of each scale, the network employs two convolution layers to get a single-channel (disparity/inverse depth) volume. After that, the volume is upsampled and converted into a probability volume by the softmax function along the disparity dimension. Finally, the predicted value is computed as follows:

$$D_{pred} = \sum_{k=0}^{D_{max}-1} k \cdot p_k, \quad (3)$$

where k denotes disparity level, p_k indicates the corresponding probability, and D_{max} is the maximum value of disparity range.

4.6. Loss Function

We utilize a multi-scale smooth L1 loss, that is commonly adopted in the SDE task, to train our network.

$$L_{sup} = \sum_{scale=0}^{3} \lambda_{scale} \cdot (\text{Smooth}_{L_1}(D^{scale}_{pred,mono}, D_{GT}) + \text{Smooth}_{L_1}(D^{scale}_{pred,stereo}, D_{GT})), \quad (4)$$

where λ indicates the coefficient for the prediction result of each scale, D_{GT} denotes the GT disparity map, and Smooth_{L_1} is the smooth L1 loss.

5. Experimental Results

5.1. Implementation Details

MDE and SDE Networks For the validation of various MDE and SDE networks designed for the visible spectrum band, we train and evaluate representative MDE and SDE networks on the proposed MS^2 dataset. Specifically, we adopt regression [26], classification [13], classification-and-regression [2], and modern transformer [56] based MDE networks (*i.e.*, BTS, DORN, AdaBins, and NeWCRF). Also, we employ 3D cost volume [55] and 4D cost volume [18, 54] based SDE networks (*i.e.*, AANet, GwcNet, and ACVNet). We utilize their official source code to implement each network architecture. All networks are initialized with ImageNet pretrained [11] or provided backbone model by following their original implementations [2, 13, 18, 26, 54–56]. We utilize the PyTorch library [40] to implement our proposed method and other comparison methods.

Optimizer and Data Augmentation All models are trained for 60 epochs on a single A6000 GPU with 48GB of memory. We utilize a batch size of 8 for all MDE model training and 4 for all SDE model training. For our method, we use a batch size of 6. We adopt AdamW optimizer [34] with an initial learning rate $1e^{-4}$ for all model training. Cosine Annealing Warm Restarts [33] is used as a learning rate scheduler. For the data augmentation, we apply random center crop-and-resize, brightness jitter, and contrast jitter for all model training. Horizontal flip is additionally applied to the MDE networks. We set the coefficients of multi-scale L1 loss λ_{scale} to 0.5, 0.5, 0.7, and 1.0. The maximum value of disparity range D_{max} is set to 192.

5.2. Depth Estimation from Thermal Images

We provide the comprehensive comparison of representative MDE and SDE networks on our MS^2 depth dataset, as shown in Tab. 3. Also, the advantage of depth estimation from thermal images can be observed in Fig. 4.

Monocular Depth Estimation The performance tendency of MDE networks is generally preserved in the thermal spectrum domain, similar to KITTI depth benchmark results [15]. MDE networks with regression heads for depth map prediction (*i.e.*, BTS and NeWCRF) have clear advantages in error metrics over methods with classification heads by directly regressing precise depth values. On the other hand, the classification head (*i.e.*, DORN and Ours) achieve higher accuracy scores by explicitly binning depth range.

The proposed unified network (*i.e.*, Ours (Mono)) generally shows comparable results with the state-of-the-art MDE method by showing higher scores in accuracy metrics yet, lower metrics in some error metrics. We think the performance gap comes from the depth prediction head and loss function. All MDE networks utilize GT depth maps

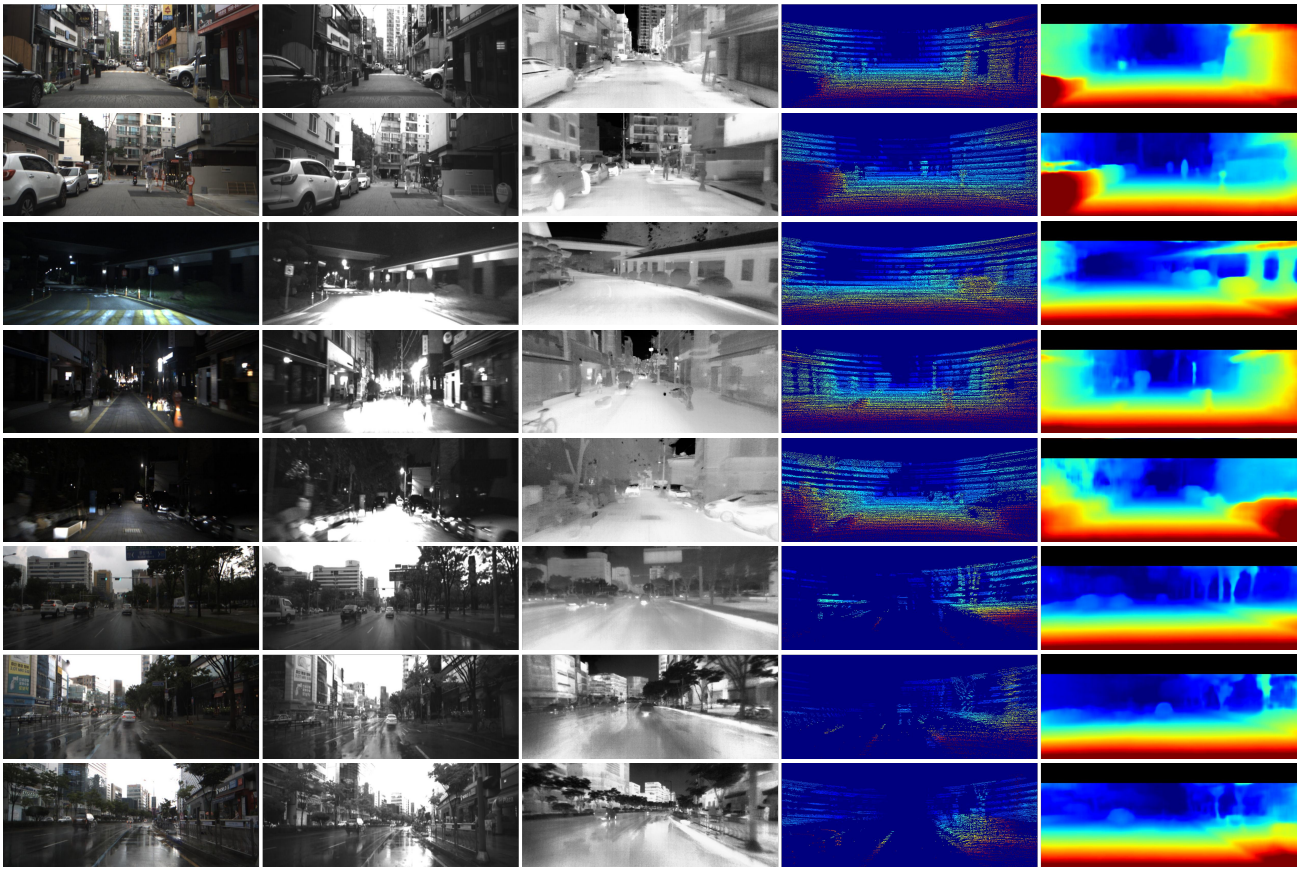
Table 3. **Quantitative comparison of depth estimation results on the proposed dataset**. We compare our network with state-of-the-art monocular and stereo depth estimation networks [2, 13, 18, 26, 54–56]. *Ours* shows comparable results in both monocular and stereo depth estimation results. Differing from the other networks, *Ours* has high-level practicality and flexibility in that it can flexibly estimate a depth map regardless of a single or stereo thermal image input. Reg and Cls indicate regression and classification heads for MDE task. The two types of SDE (*i.e.*, 3D and 4D CV) denote 3D and 4D cost volume, respectively. The best performance in each block is highlighted in **bold**.

(a) Monocular Depth Estimation Results on the Evaluation Set of Our MS^2 Depth Dataset.

Methods	TestSet	**Error** ↓				**Accuracy** ↑			Type	
		AbsRel	SqRel	RMSE	RMSElog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Reg	Cls
DORN [13]	Day	0.144	1.288	5.483	0.230	0.856	0.941	0.970		✓
	Night	0.136	1.136	5.290	0.212	0.863	0.950	0.976		
	Rain	0.180	1.934	6.735	0.276	0.781	0.910	0.955		
	Avg	0.151	1.419	5.776	0.237	0.837	0.935	0.968		
BTS [26]	Day	0.122	0.905	4.923	0.198	0.857	0.951	0.980	✓	
	Night	0.114	0.798	4.701	0.184	0.870	0.959	0.984		
	Rain	0.157	1.395	6.053	0.243	0.791	0.926	0.969		
	Avg	0.129	1.008	5.169	0.206	0.843	0.947	0.978		
AdaBins [2]	Day	0.129	0.976	5.108	0.205	0.847	0.947	0.979	✓	✓
	Night	0.119	0.822	4.749	0.187	0.864	0.958	0.984		
	Rain	0.168	1.545	6.336	0.254	0.771	0.918	0.965		
	Avg	0.137	1.084	5.330	0.212	0.831	0.943	0.977		
NeWCRF [56]	Day	0.120	0.864	4.852	0.195	0.858	0.952	0.982	✓	
	Night	0.112	0.755	4.594	0.179	0.875	0.961	0.985		
	Rain	0.155	1.352	5.956	0.240	0.795	0.929	0.970		
	Avg	0.127	**0.965**	5.077	**0.202**	0.846	**0.949**	**0.980**		
Ours (Mono)	Day	0.115	0.983	4.895	0.201	0.882	0.952	0.977		✓
	Night	0.107	0.850	4.658	0.185	0.894	0.961	0.981		
	Rain	0.152	1.567	6.020	0.247	0.822	0.928	0.964		
	Avg	0.123	1.103	5.134	0.208	0.869	0.948	0.975		
Ours (Stereo)	Day	0.113	0.948	4.852	0.200	0.884	0.953	0.977		✓
	Night	0.105	0.811	4.584	0.183	0.896	0.961	0.981		
	Rain	0.149	1.499	5.940	0.245	0.826	0.929	0.965		
	Avg	**0.120**	1.057	**5.068**	0.207	**0.872**	**0.949**	0.975		

(b) Disparity Estimation Results on the Evaluation Set of Our MS^2 Depth Dataset.

Methods	TestSet	Lower is better					Type	
		EPE-all(px)	D1-all(%)	$>$ 1px(%)	$>$ 2px(%)	$>$ 3px(%)	3D CV	4D CV
GwcNet [18]	Day	0.905	5.5	19.2	8.4	5.5		✓
	Night	0.946	5.6	26.0	10.2	5.6		
	Rain	1.070	7.2	24.3	11.1	7.2		
	Avg	0.969	6.0	23.3	9.9	6.0		
AANet [55]	Day	0.939	5.8	20.2	8.8	5.8	✓	
	Night	0.995	6.1	27.9	11.1	6.1		
	Rain	1.091	7.5	25.3	11.6	7.5		
	Avg	1.005	6.4	24.7	10.5	6.4		
ACVNet [54]	Day	0.898	5.5	18.9	8.3	5.5		✓
	Night	0.943	5.5	25.9	10.1	5.5		
	Rain	1.056	7.2	23.6	10.9	7.2		
	Avg	**0.962**	6.0	**23.0**	9.8	6.0		
Ours (Mono)	Day	1.033	6.4	23.1	10.5	6.4	✓	
	Night	0.946	5.6	29.6	9.8	5.6		
	Rain	1.261	8.7	24.4	14.6	8.7		
	Avg	1.066	6.8	24.4	11.4	6.8		
Ours (Stereo)	Day	0.957	5.7	22.7	9.1	5.7	✓	
	Night	0.853	4.8	21.3	8.2	4.8		
	Rain	1.159	7.7	29.1	12.4	7.7		
	Avg	0.976	**5.9**	24.0	**9.7**	**5.9**		



(a) RGB (Reference Only) (b) NIR (Reference Only) (c) THR (d) GT disparity (e) Ours (stereo)

Figure 4. **Qualitative results of stereo disparity estimation on the MS^2 depth dataset**. Predicted disparity map from stereo thermal images shows high-level robust estimation results regardless of lighting and weather condition. However, inherent hardware noise and the absence of high-frequency information lead to blurry prediction results for specific regions such as the regions that have similar thermal radiation values (*i.e.*, temperature) and noisy areas generated by the sensor itself. We think multi-spectral modality fusion can achieve both robustness and reliability. Further results and comparisons with other MDE and SDE networks can be found in the supplementary material.

that can provide precise distance information. On the other hand, our network is trained with the disparity map that can be regarded as discretized distance information. Luckily, the performance gaps are narrowed down by utilizing the right image as additional guidance thanks to our unified architecture. Also, we believe an investigation of an effective form of prediction head for unified MDE and SDE tasks can boost the overall performance.

Stereo Depth Estimation Generally, the method utilizing 4D cost volume aggregation with 3D convolution layer (*i.e.*, GwcNet and ACVNet) provide precise disparity estimation results than 3D cost volume methods (*i.e.*, AANet and Ours (stereo)). However, the strict constraints of the architecture module and left-and-right images degenerate network flexibility. On the other hand, our proposed network has high-level practicality and flexibility by exploiting a single network for both monocular and stereo depth estimation. At the same time, the proposed network can provide comparable performance with 4D cost volume based methods.

6. Conclusion

In this paper, we built a large-scale Multi-Spectral Stereo (MS^2) dataset, including stereo RGB, stereo NIR, stereo thermal, and stereo LiDAR data along with GNSS/IMU information. Also, we conduct an exhaustive validation process of MDE and SDE algorithms whether they work well in the thermal spectrum band. Lastly, we propose a unified depth network that effectively bridges monocular depth and stereo depth tasks from a conditional random field perspective. We hope our paper encourage active research of various computer vision algorithm from multi-spectral data to achieve high-level performance, reliability, and robustness against challenging environments.

Acknowledgment This work was supported by PoliceLab 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 / 220122M0500]

References

[1] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie, 1992. 4

[2] 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, 6, 7

[3] Paulo Vinicius Koerich Borges and Stephen Vidas. Practical infrared visual odometry. *IEEE Transactions on Intelligent Transportation Systems*, 17(8):2205–2213, 2016. 1, 3

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 2

[5] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018. 3, 5

[6] 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. 2, 4

[7] 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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[8] Weichen Dai, Yu Zhang, Shenzhou Chen, Donglei Sun, and Da Kong. A multi-spectral dataset for evaluating motion estimation systems. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5560–5566. IEEE, 2021. 2, 4

[9] Kevser Irem Danaci and Erdem Akagunduz. A survey on infrared image and video sets. *arXiv preprint arXiv:2203.08581*, 2022. 3

[10] Jeff Delaune, Robert Hewitt, Laura Lytle, Cristina Sorice, Rohan Thakker, and Larry Matthies. Thermal-inertial odometry for autonomous flight throughout the night. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1122–1128. IEEE, 2019. 3

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[12] Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4738–4747, 2019. 2

[13] 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. 2, 6, 7

[14] Stefano Gasperini, Patrick Koch, Vinzenz Dallabetta, Nassir Navab, Benjamin Busam, and Federico Tombari. R4dyn: Exploring radar for self-supervised monocular depth estimation of dynamic scenes. In *2021 International Conference on 3D Vision (3DV)*, pages 751–760. IEEE, 2021. 1

[15] 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. 1, 2, 4, 6

[16] Vitor Guizilini, Rares Ambrus, Wolfram Burgard, and Adrien Gaidon. Sparse auxiliary networks for unified monocular depth prediction and completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11078–11088, 2021. 1

[17] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[18] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3273–3282, 2019. 3, 5, 6, 7

[19] Keli Huang, Botian Shi, Xiang Li, Xin Li, Siyuan Huang, and Yikang Li. Multi-modal sensor fusion for auto driving perception: A survey. *arXiv preprint arXiv:2202.02703*, 2022. 3

[20] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*, pages 66–75, 2017. 3

[21] Shehryar Khattak, Christos Papachristos, and Kostas Alexis. Keyframe-based thermal–inertial odometry. *Journal of Field Robotics*, 37(4):552–579, 2020. 1, 3

[22] 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. 1, 3

[23] 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. 1

[24] Alex Junho Lee, Younggun Cho, Young-sik Shin, Ayoung Kim, and Hyun Myung. Vivid++: Vision for visibility dataset. *IEEE Robotics and Automation Letters*, 7(3):6282–6289, 2022. 2, 4

[25] 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. 2, 4

[26] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 2, 6, 7

[27] Chenglong Li, Wei Xia, Yan Yan, Bin Luo, and Jin Tang. Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation. *arXiv preprint arXiv:1907.10303*, 2019. 1

[28] Peize Li, Kaiwen Cai, Muhamad Risqi U Saputra, Zhuangzhuang Dai, and Chris Xiaoxuan Lu. Odombeyondvision: An indoor multi-modal multi-platform odometry dataset beyond the visible spectrum. *arXiv preprint arXiv:2206.01589*, 2022. 2, 4

[29] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022. 2

[30] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2811–2820, 2018. 3, 5

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 5, 6

[32] Yunfei Long, Daniel Morris, Xiaoming Liu, Marcos Castro, Punarjay Chakravarty, and Praveen Narayanan. Radar-camera pixel depth association for depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12507–12516, 2021. 1

[33] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 6

[35] 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. 1, 3

[36] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017. 2

[37] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 3, 5

[38] Yasuto Nagase, Takahiro Kushida, Kenichiro Tanaka, Takuya Funatomi, and Yasuhiro Mukaigawa. Shape from thermal radiation: Passive ranging using multi-spectral lwir measurements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12661–12671, 2022. 3

[39] Jinsun Park, Yongseop Jeong, Kyungdon Joo, Donghyeon Cho, and In So Kweon. Adaptive cost volume fusion network for multi-modal depth estimation in changing environments. *IEEE Robotics and Automation Letters*, 2022. 1

[40] 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. 6

[41] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 2

[42] René 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. 2

[43] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13906–13915, 2021. 3

[44] 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 ego-motion. *IEEE Robotics and Automation Letters*, 7(3):7771–7778, 2022. 3

[45] 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. 1, 3

[46] Ukcheol Shin, Kwanyong Park, Byeong-Uk Lee, Kyunghyun Lee, and In So Kweon. Self-supervised monocular depth estimation from thermal images via adversarial multi-spectral adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5798–5807, 2023. 3

[47] Young-Sik Shin and Ayoung Kim. Sparse depth enhanced direct thermal-infrared slam beyond the visible spectrum. *IEEE Robotics and Automation Letters*, 4(3):2918–2925, 2019. 1, 3

[48] Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. *arXiv preprint arXiv:1909.10980*, 2019. 1

[49] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. 4(3):2576–2583, 2019. 1

[50] 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. 1

[51] Jie Tang, Fei-Peng Tian, Wei Feng, Jian Li, and Ping Tan. Learning guided convolutional network for depth completion. *IEEE Transactions on Image Processing*, 30:1116–1129, 2020. 1

[52] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 195–204, 2019. 3, 5

[53] Wayne Treible, Philip Saponaro, Scott Sorensen, Abhishek Kolagunda, Michael O'Neal, Brian Phelan, Kelly Sherbondy, and Chandra Kambhamettu. Cats: A color and thermal stereo benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2961–2969, 2017. 2, 4

[54] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12981–12990, 2022. 3, 5, 6, 7

[55] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1959–1968, 2020. 3, 5, 6, 7

[56] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3916–3925, 2022. 2, 5, 6, 7

[57] Yuxiao Zhang, Alexander Carballo, Hanting Yang, and Kazuya Takeda. Autonomous driving in adverse weather conditions: A survey. *arXiv preprint arXiv:2112.08936*, 2021. 3

[58] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 5