

# MonoTher-Depth: Enhancing Thermal Depth Estimation via Confidence-Aware Distillation

Xingxing Zuo, Nikhil Ranganathan, Connor Lee, Georgia Gkioxari, and Soon-Jo Chung

**Abstract**—Monocular depth estimation (MDE) from thermal images is a crucial technology for robotic systems operating in challenging conditions such as fog, smoke, and low light. The limited availability of labeled thermal data constrains the generalization capabilities of thermal MDE models compared to foundational RGB MDE models, which benefit from datasets of millions of images across diverse scenarios. To address this challenge, we introduce a novel pipeline that enhances thermal MDE through knowledge distillation from a versatile RGB MDE model. Our approach features a confidence-aware distillation method that utilizes the predicted confidence of the RGB MDE to selectively strengthen the thermal MDE model, capitalizing on the strengths of the RGB model while mitigating its weaknesses. Our method significantly improves the accuracy of the thermal MDE, independent of the availability of labeled depth supervision, and greatly expands its applicability to new scenarios. In our experiments on new scenarios without labeled depth, the proposed confidence-aware distillation method reduces the absolute relative error of thermal MDE by 22.88% compared to the baseline without distillation. The code will be available at: [https://github.com/ZuoJiaxing/monother\\_depth](https://github.com/ZuoJiaxing/monother_depth).

**Index Terms**—Deep Learning for Visual Perception, Range Sensing, Thermal Camera

## I. INTRODUCTION

DEPTH estimation is a fundamental problem in various applications, including autonomous driving, robotics, and mixed reality. Monocular depth estimation (MDE) from a single RGB camera is widely used and has seen significant progress recently [1]. Existing RGB MDE methods have achieved high accuracy, detailed fidelity, and great zero-shot generalization capabilities. A key factor contributing to the success of RGB MDE models is the abundance of existing RGB datasets with labeled depth available for training. For instance, methods such as UniDepth [2], Metric3D [3], and ZeroDepth [4] are trained on datasets containing 3M, 8M, and 17M images with labeled depth, respectively. DepthAnything [5] leverages 62M unlabeled images, on top of 1.5M labeled images, to achieve great generalization in diverse environments.

However, RGB cameras struggle in adverse visual conditions characterized by low light, fog, or smoke, which limit the performance of MDE in these scenarios. Thermal cameras, which capture long-wave infrared signals, can penetrate

atmospheric particles and provide reliable measurements in obscured and low-light conditions. Despite these advantages, thermal depth estimation has yet to be thoroughly explored. Specifically, thermal MDE inherits the challenges of RGB MDE, with added complexity due to typically low contrast, high signal-to-noise ratio, and a lack of texture and color information in thermal imagery.

Along with these challenges, thermal MDE must contend with the scarcity of labeled thermal datasets. This is in contrast to the abundance of datasets, both real and synthetic, for RGB MDE. In this work, we seek to enhance thermal MDE by taking advantage of off-the-shelf RGB MDE foundation models that have been pretrained on massive RGB datasets. In particular, we propose a novel framework that distills an RGB MDE model to thermal using a confidence-aware approach (Fig. 1). Unlike existing RGB-T works [6]–[8], our approach can work with RGB-T training image pairs that are not perfectly co-registered. We achieve this by adaptively guiding the distillation process using confidence derived from cross-modal features and depth consistency.

Our main contributions are summarized as follows:

- We introduce MonoTher-Depth, a novel semi-supervised distillation framework that distills an RGB MDE model to create a thermal MDE model.
- We propose confidence-aware distillation, based on cross-modal spatial consistency of feature spaces and depth estimates, to limit incorrect guidance and to eschew the need for co-registered RGB-T image pairs.
- We perform extensive validation and ablation studies on the MS<sup>2</sup> [9] and ViViD++ [10] datasets, demonstrating MonoTher-Depth’s effectiveness in enhancing thermal MDE learning with and without ground-truth depth supervision. In scenarios without labeled depth, our proposed confidence-aware distillation method reduces the absolute relative error of thermal MDE by 22.88% compared to the baseline without distillation.

## II. RELATED WORK

**RGB MDE.** Monocular depth estimation (MDE) from RGB images has made significant strides in recent years [2], [3], [11]–[18], demonstrating zero-shot capabilities across diverse image datasets. MiDaS [13] is a pioneering work that leverages a large collection of diverse datasets for training relative MDE model, showcasing a certain degree of zero-shot capability. Zoedepth [14] focuses on metric MDE with a dedicated metric bins module, pre-trained on 12 datasets using relative depth and fine-tuned on two datasets using metric depth. MetricDepth [3], [16] addresses the metric ambiguity in MDE

Manuscript received: 9 September, 2024; Revised 6 December, 2024; Accepted 6 January, 2025. This paper was recommended for publication by Editor Cesar Cadena upon evaluation of the Associate Editor and Reviewers’ comments. This work was supported by the Office of Naval Research. (Corresponding author: Xingxing Zuo.)

All authors are with California Institute of Technology, Pasadena, California, USA. E-mail:{zuo, nranganathan, clee, georgia, sjchung}@caltech.edu

Digital Object Identifier (DOI): see top of this page.

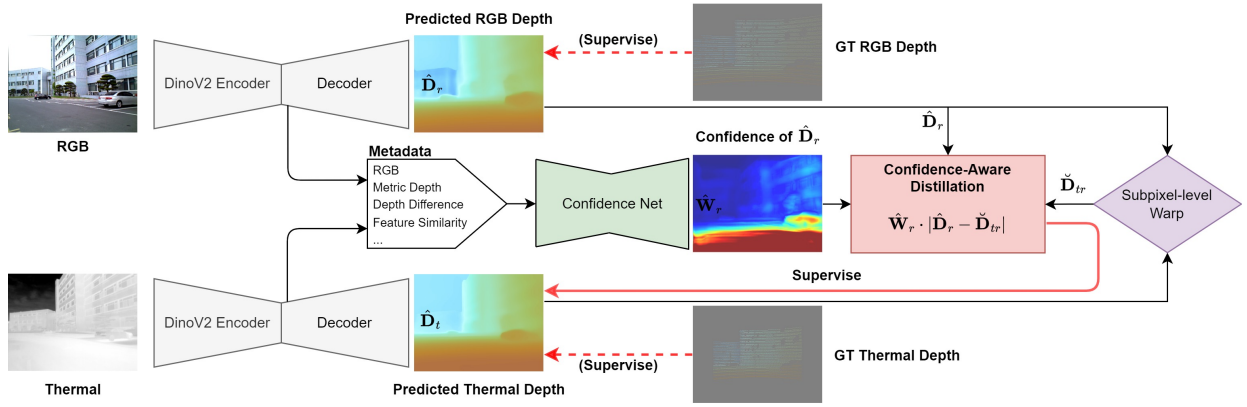


Fig. 1: **System architecture of MonoTher-Depth.** Our framework enhances the thermal MDE model by leveraging learned priors from an RGB model. To harness the strengths and mitigate the weaknesses of the RGB teacher model, we predict the confidence of its depth output  $\hat{W}_r$  using curated metadata that includes both thermal and RGB information. Whether ground-truth (GT) depth is available or not, our system improves thermal MDE through confidence-aware distillation by minimizing the confidence-weighted depth discrepancy between the predicted RGB depth  $\hat{D}_r$  and the wrapped thermal depth  $\check{D}_{tr}$ .

by transforming all training data into a canonical camera space with a fixed focal length. UniDepth [2] introduces a pseudo-spherical output space representation that effectively disentangles the camera parameters from the depth estimation process. Marigold [17] harnesses the rich priors captured in recent generative diffusion models to achieve generalizable and accurate MDE. DepthAnything [5] pretrains the MDE model with labeled data and then utilizes large-scale unlabeled data to learn robust representations, enhancing zero-shot capability and overall robustness.

**Thermal MDE.** Thermal MDE is significantly less explored in the existing literature, with only a limited number of datasets available for thermal MDE. Most existing thermal MDE methods are self-supervised [19], [20], relying on reconstructing thermal image sequences by predicted depth and poses between images. Some approaches also leverage the knowledge from RGB MDE models or fuse RGB information to enhance thermal MDE performance. Xu *et al.* [21] proposed fusing the predictions of an RGB MDE model and a thermal MDE model to obtain refined depth estimation in both day and night scenarios. Shin *et al.* [22] and Guo *et al.* [23] proposed unsupervised multi-spectrum stereo methods for thermal MDE, which supervise the thermal MDE model using self-reconstruction consistency loss calculated from the predicted depth and poses of sequential thermal and/or RGB images in videos. Shin *et al.* [20] adapted an RGB MDE model to the thermal domain in a self-supervised manner. They used impaired RGB and thermal videos to regress image poses and depth, reconstructing each image sequence within its respective domain, while enforcing that the RGB and thermal encoders produce indistinguishable feature maps through an adversarial loss. In contrast, our proposed confidence-aware distillation method does not require video sequences for training. Instead, it operates directly on the predicted depth, eliminating the need for complex discriminators operating on high-dimensional feature spaces.

**Knowledge Distillation for MDE.** Knowledge distillation was initially proposed for image recognition tasks [24], but

a few methods have also utilized it for MDE. Several studies [25], [26] have explored supervising an MDE model using predictions from a stereo matching network. Pilzer *et al.* [27] proposed enhancing the MDE model by refining the predicted depth based on the cycle inconsistency of left-right stereo image pairs. Aleotti *et al.* [28] introduced a method to distill a complex MDE model trained on a large-scale dataset into a lightweight model suitable for deployment on handheld devices. Poggi *et al.* [29] proposed a self-teaching strategy to quantify the uncertainty in self-supervised MDE trained from sequential video frames. URDC-Depth [30] cross-distills Transformer and CNN-based MDE networks by feeding the same image into both networks and fusing their depth predictions based on predicted uncertainty. Shi *et al.* [31] fused MDE results from multiple video frames to build a 3D mesh via TSDF-Fusion, using rendered depth from the mesh to fine-tune the MDE network. DepthAnything [5] trained a teacher MDE model with labeled depth and then distilled the trained model into a student model using a mix of labeled and massive unlabeled data. During distillation, strongly perturbed images are fed into the student model, while the unperturbed image is fed into the teacher model, allowing the student model to learn robustness to open-world images. In contrast to the methods mentioned above, our MonoTher-Depth focuses on cross-modality distillation from RGB to thermal. More importantly, our distillation process is confidence-aware, selectively absorbing the strengths of the teacher model while mitigating its weaknesses, and does not require sequential video frames for training.

### III. METHODOLOGY

#### A. Problem Setup

To overcome the challenges caused by a lack of co-registered RGB-T training datasets for thermal MDE, we distill a pretrained RGB MDE model into a thermal MDE model. Our training method requires RGB-T images with overlapping field-of-views and calibrated extrinsics, but does not need strict co-registration. The RGB model is not needed for inference.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

While the RGB depth model performs well on average, its performance may vary across different image regions and conditions. To avoid transferring potentially incorrect behaviors to the thermal model, we model the confidence of RGB predictions and use it to adaptively steer the thermal model during training. This confidence is generated by a dedicated neural network, which is trained when ground-truth depth is available. When ground-truth depth is not available, our method leverages the frozen RGB model and confidence network to modulate the distillation loss of the thermal MDE model.

### B. Metric MDE Network

We adopt the metric version of DepthAnything as both our RGB and thermal MDE models [5]. This model uses DinoV2 [32] for feature extraction, a DPT decoder [13] for relative depth prediction, and a metric bins module [14] to produce metric depth predictions. For our RGB teacher model, we use the pretrained DepthAnything model, which was trained on 63.5M images, showing exceptional performance across multiple MDE benchmarks.

The thermal MDE model utilizes the same architecture, but incorporates an additional preprocessing step that normalizes each raw 16-bit thermal image to the 2<sup>nd</sup> and 98<sup>th</sup> percentiles [33], [34]. While some learning-based thermal works [19], [33]–[35] also use Contrast Limited Adaptive Histogram Equalization (CLAHE), we did not find benefits for our MDE purpose.

### C. Sub-pixel Warp of Thermal-RGB Depth

In order for the thermal MDE model to learn from the RGB MDE model, a precise spatial mapping between thermal and RGB pixels is essential. We compute this using predicted depth maps and known camera intrinsics and extrinsics. Given the predicted depth from the RGB MDE model,  $\hat{\mathbf{D}}_r$ , we transform it into the thermal image plane as follows:

$$\hat{\mathbf{D}}_{rt}, \hat{\mathbf{u}}_{rt} = \pi(\mathbf{T}_r^t \hat{\mathbf{D}}_r \pi^{-1}(\mathbf{u}_r, \mathbf{K}_r), \mathbf{K}_t) \quad (1)$$

where  $\pi(\mathbf{x}, \mathbf{K})$  denotes the camera projection function that projects points  $\mathbf{x}$  into the image plane with camera intrinsic matrix  $\mathbf{K}$ . The inverse projection function  $\pi^{-1}(\mathbf{u}, \mathbf{K})$  maps image pixels  $\mathbf{u}$  back into the 3D unit plane. The transformation matrix  $\mathbf{T}_r^t$  represents the 6-DoF extrinsic transformation from the RGB camera to the thermal camera. In this equation, we omit the homogeneous conversion of vectors for simplicity. Using the transformation function in (1), we obtain the pixel correspondences  $\mathbf{u}_r$  in the RGB image and  $\hat{\mathbf{u}}_{rt}$  in the thermal image. The depth  $\hat{\mathbf{D}}_{rt}$  represents the RGB depth in the thermal image’s coordinate frame.

Similarly, we can transform the predicted thermal depth  $\hat{\mathbf{D}}_t$  into the RGB image plane using:

$$\hat{\mathbf{D}}_{tr}, \mathbf{u}_{tr} = \pi(\mathbf{T}_t^r \hat{\mathbf{D}}_t \pi^{-1}(\mathbf{u}_t, \mathbf{K}_t), \mathbf{K}_r) \quad (2)$$

To supervise the thermal MDE using the RGB model, we calculate the warped thermal depth corresponding to the RGB image, denoted as  $\check{\mathbf{D}}_{tr}$ . To achieve sub-pixel accuracy in the depth warp process, we sample the depth values of  $\hat{\mathbf{D}}_{tr}$  at the sub-pixel locations  $\hat{\mathbf{u}}_{tr}$  using bilinear interpolation:

$$\check{\mathbf{D}}_{tr} = f_{\text{bilinear}}(\hat{\mathbf{u}}_{tr}, \hat{\mathbf{D}}_{tr}) \quad (3)$$

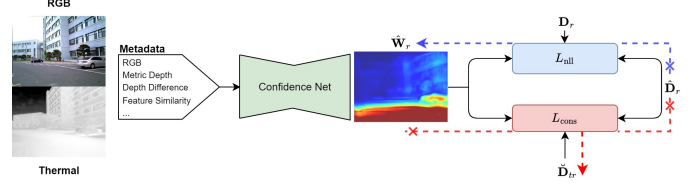


Fig. 2: **Pipeline of the confidence-aware distillation.** The predicted confidence  $\hat{\mathbf{W}}_r$  of the RGB depth  $\hat{\mathbf{D}}_r$  plays a key role in both the negative log-likelihood loss  $L_{nll}$  (5) and the consistency loss  $L_{con}$  (6). The  $L_{nll}$  loss propagates gradients back to the confidence network, while the  $L_{con}$  loss propagates gradients to the warped thermal depth  $\check{\mathbf{D}}_{tr}$ . Gradient flow is stopped along all other paths.

### D. Confidence-Aware Model Distillation

We leverage the versatile RGB MDE model to instruct the thermal MDE model at the depth output level by minimizing the discrepancy between the predicted RGB depth  $\hat{\mathbf{D}}_r$  and the warped thermal depth  $\check{\mathbf{D}}_{tr}$ . Since the reliability of RGB depth predictions may vary across different image regions, weighting all pixel-wise depth discrepancies equally during training can lead to poor results. To address this, we incorporate the confidence of the RGB MDE model into the distillation process.

We design a U-Net to predict the confidence  $\hat{\mathbf{W}}_r$  of the RGB MDE output  $\hat{\mathbf{D}}_r$ . To achieve this, we input RGB-aligned metadata into this network, leveraging the pixel correspondences  $\mathbf{u}_r$  and  $\hat{\mathbf{u}}_{rt}$  between the RGB and thermal depths. This metadata includes: (I) the cosine distance between the RGB feature map and its corresponding thermal feature map, (II) the sampled cosine distance between the thermal feature map and its corresponding RGB feature map, (III) the  $L_1$  distance  $|\hat{\mathbf{D}}_r - \check{\mathbf{D}}_{tr}|$ , (IV) the warped thermal depth  $\check{\mathbf{D}}_{tr}$ , (V) the RGB depth  $\hat{\mathbf{D}}_r$ , and (VI) the RGB image  $\mathbf{I}_r$ . Notably, all these curated metadata components are subpixel-aligned with the RGB image to facilitate precise confidence map prediction.

It is straightforward to obtain metadata components (III)–(VI). To calculate component (I), we extract the RGB and thermal feature embeddings from the last layer of the metric bins module, denoted as  $\mathbf{F}_r$  and  $\mathbf{F}_t$ , respectively. The cosine distance between these feature maps is computed as follows:

$$\mathbf{S}_r = \langle \mathbf{F}_r, f_{\text{bilinear}}(\hat{\mathbf{u}}_{rt}, \mathbf{F}_t) \rangle \quad (4)$$

where  $\langle \mathbf{A}, \mathbf{B} \rangle$  represents the element-wise cosine distance between feature maps  $\mathbf{A}$  and  $\mathbf{B}$ . In this context,  $f_{\text{bilinear}}(\mathbf{u}, \mathbf{F})$  refers to the bilinear interpolation sampling of  $\mathbf{F}$  at pixel locations  $\mathbf{u}$  unless otherwise specified. Similarly, we calculate the cosine distance  $\mathbf{S}_t$ . Since  $\mathbf{S}_t$  is aligned with the thermal image, we perform bilinear interpolation again to obtain component (II), denoted as  $\mathbf{S}_{tr} = f_{\text{bilinear}}(\hat{\mathbf{u}}_{tr}, \mathbf{S}_t)$ .

The confidence network is based on a standard U-Net architecture, specifically tailored for confidence estimation. The encoder consists of four downsampling layers, each reducing the spatial dimensions by half while increasing the number of feature channels. The decoder mirrors the encoder’s structure, employing four upsampling layers that progressively integrate

**IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.**

higher-level features with corresponding lower-level features from the encoder via skip connections. This process gradually restores the spatial resolution of the input metadata. The confidence  $\hat{\mathbf{W}}_r$  is finally predicted by a convolutional layer followed by a Sigmoid activation function.

Training of this confidence predictor is done only when ground-truth RGB depth  $\mathbf{D}_r$  is available (Fig. 2). To do this, we minimize the negative log-likelihood of a Laplacian distribution:

$$L_{\text{null}} = \frac{1}{N} \sum_i \hat{W}_r^i \cdot |\text{sg}(\hat{D}_r^i) - D_r^i| - \beta \log(\hat{W}_r^i) \quad (5)$$

where  $\text{sg}(\cdot)$  denotes the stop-gradient operation to prevent backpropagation through the ground truth. The ground-truth RGB depth is often sparse in outdoor scenarios, and  $i$  indexes all the  $N$  pixel locations where ground-truth RGB depth is available.

To perform confidence-aware distillation, we minimize the confidence-weighted  $L_1$  loss:

$$L_{\text{cons}} = \frac{1}{M} \sum_j \text{sg}(\hat{W}_r^j) \cdot |\text{sg}(\hat{D}_r^j) - \check{D}_{\text{tr}}^j| \quad (6)$$

where  $j$  indexes all  $M$  pixel locations  $\check{D}_{\text{tr}}^j$  that fall within the RGB image after warping. To address issues due to occlusion (where the warped thermal depth may not have corresponding pixels in the RGB depth map) and imperfect ground-truth depth, we exclude the top 20% residuals in both (5) and (6). Lastly, we apply a mask for the consistency loss (6), considering only the pixel locations with the top 80% of feature similarity  $S_r$  (4).

### E. Implementation Details

Our distillation framework is versatile: it allows for replacement of the DepthAnything model, while maintaining compatibility with the confidence network that processes MDE model outputs. When ground-truth depth is available, we supervise the thermal and RGB MDE models using the SILOG loss [11], [14], [36]

$$L_{\text{silog}} = \sum_i \sqrt{\frac{1}{N} \sum_i (g^i)^2 - \frac{\lambda}{N^2} \left( \sum_i g^i \right)^2} \quad (7)$$

Here,  $g^i = \log \hat{D}^i - \log D^i$ , and  $N$  is the number of pixels with valid ground-truth depth in the image. We set  $\lambda = 0.15$  in our experiments.

In addition, we align depth and image discontinuities by regularizing the predicted depth  $\hat{\mathbf{D}}$  with a smoothness loss [37]

$$L_{\text{sm}}(\hat{\mathbf{D}}) = \nabla_x \hat{\mathbf{D}} \cdot \exp(-\nabla_x \mathbf{I}_r) + \nabla_y \hat{\mathbf{D}} \cdot \exp(-\nabla_y \mathbf{I}_r) \quad (8)$$

We only regularize the predicted RGB depth because applying this loss to thermal depth is potentially detrimental.

When ground-truth depth is available, we train the RGB and thermal MDE models, along with the confidence network, using the following combined loss:

$$L = L_{\text{silog}_r} + L_{\text{silog}_t} + \alpha \cdot L_{\text{cons}} + \beta \cdot L_{\text{null}} + \gamma \cdot L_{\text{sm}}(\hat{\mathbf{D}}_r) + \lambda \cdot L_{\text{sm}}(\hat{\mathbf{W}}_r) \quad (9)$$

where  $L_{\text{sm}}(\hat{\mathbf{W}}_r)$  denotes the smoothness loss on the predicted confidence map. We set  $\alpha = 0.2, \beta = 0.1, \gamma = 0.01, \lambda = 0.001$ .

When ground-truth depth is unavailable, we freeze the weights of a pretrained RGB MDE model and confidence network. We use the predicted RGB depth and confidence to train the thermal MDE model, applying only the depth consistency loss  $L_{\text{cons}}$  in this self-supervised fine-tuning process.

## IV. EXPERIMENTS

### A. Datasets

**MS<sup>2</sup>:** This is a multispectral stereo dataset [9] that provides synchronized thermal and RGB images, and projected LiDAR depth maps for benchmarking depth prediction. For MDE benchmarking, we use the left RGB images, left thermal images, and the LiDAR maps. We follow the official train/val/test splits. The train split consists of 7.6K image pairs, while the test split includes 2.3K, 2.3K, and 2.5K image pairs under *day*, *night*, and *rainy* conditions, respectively. Due to misalignment between the projected ground-truth LiDAR depth and the image—particularly at image edges—training directly with the provided LiDAR depth maps often produces blurred depth predictions. To mitigate this issue, we filter the LiDAR depth map using two strategies: (i) remove depths that exhibit significant pixel-intensity inconsistencies when back-projecting LiDAR points into both the left and right images, and (ii) remove depths that substantially deviate from those obtained via stereo matching [38]. While we utilize these filtered LiDAR depths during training to achieve sharper predicted depths, we continue to use the unfiltered, officially provided LiDAR depth for evaluation to ensure fairness and thoroughness. Although this setting yields crisper depth predictions, we find it slightly compromises certain evaluation metrics.

**ViViD++:** This dataset [10] provides data from RGB, thermal, and event cameras, as well as LiDAR. It captures both indoor and outdoor scenes. For our study, we only use the official outdoor splits. The outdoor training split, which consists of data collected during the day, is optionally used for self-supervised fine-tuning of our method (see Sec. III-E). The test split consists of nighttime data to show the generalization of our method.

### B. Training Details

We initialized the RGB and thermal MDE network encoders with pretrained weights from DepthAnything [5]. We randomly initialized the decoder, the metric bins module of the MDE models, and the confidence network. We used the AdamW optimizer with a learning rate of 8.5e-5 and a weight decay of 0.01. We applied brightness and contrast jitters for data augmentation. The networks are trained for 5 epochs on two Nvidia RTX6000 Ada GPUs, with an input image resolution of  $256 \times 640$  on MS<sup>2</sup> dataset, taking approximately 20 hours. For experiments on the ViViD++ dataset, the input image resolution is  $480 \times 640$ .

### C. Evaluation Protocols

We report standard metrics [9], [14], including absolute relative error (AbsRel), squared relative difference (SqRel),

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

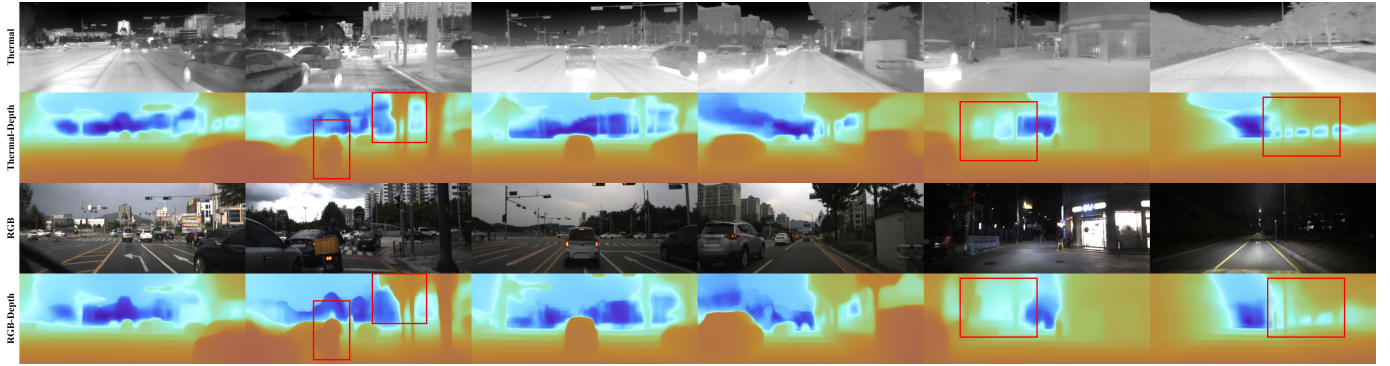


Fig. 3: **Monocular Depth Estimation on MS<sup>2</sup> dataset** [9]. Top to bottom: normalized thermal image, predicted thermal depth, RGB image, and predicted RGB depth. Red boxes highlight significant differences between the thermal and RGB depth predictions. Left to right: every two columns showcase the rainy, day, and night conditions, respectively.



Fig. 4: **Predicted confidence and depth error on MS<sup>2</sup> dataset** [9]. Left to right: depth error overlaid on the image, confidence overlaid on the image, and predicted RGB depth.

root mean square error (RMSE, in meter), RMSE logarithm (RMSElog), and the threshold accuracy  $\delta = \%$  of pixels s.t.  $\max(d_i/\hat{d}_i, \hat{d}_i/d_i) < 1.25^n$ ,  $n = \{1, 2, 3\}$ .

In outdoor scenarios with ground-truth depth from accumulated LiDAR scans, most pixels with valid ground-truth depth are close to the camera. Simple averaging of metrics over all valid pixels can skew results, as performance on nearby points may dominate. To address this, we report both unweighted metrics averaged over all valid pixels and weighted metrics [39], which calculate averages across depth bins. Each bin spans 5 meters, and the evaluation covers depths from 0 to 80 meters for both the MS<sup>2</sup> and ViViD++ datasets.

#### D. Evaluation of Metric Monocular Depth Estimation

**Evaluation Results on MS<sup>2</sup>.** Table I presents evaluation results for metric monocular depth estimation on the MS<sup>2</sup> dataset test split. In all tables throughout this paper, we use bold to highlight the best performance. All compared methods were trained or fine-tuned on thermal data of the MS<sup>2</sup> training split. In comparison to methods trained exclusively on thermal images [5], [12], [14], [36], [40], [41], our MonoTherDepth model performs the best across all metrics.

Our method outperforms both Zoedepth [14] and DepthAnything [5], which are closely related to our approach and trained with the thermal images and ground-truth thermal depth supervision. The improvement of our method is attributed to the effective knowledge distillation from the RGB MDE model during training. While the ground-truth thermal and RGB depth maps from LiDAR offer highly accurate depth information and potentially reduce the benefit of the RGB teacher, the results still highlight the substantial benefits of our confidence-aware distillation. Additionally,

Table II provides additional categorized evaluations under *Day*, *Night*, and *Rainy* conditions. Both unweighted and weighted metrics are presented. Figure 3 presents the MDE results from our method, highlighting the advantages of thermal MDE over RGB MDE in challenging scenarios. To give an impression of the predicted confidence of the RGB depth, we visualize the depth error and the confidence in Fig. 4. Interestingly, the regions with large depth errors (blue) coincide with the low-confidence areas (blue), indicating that the confidence map can inclusively reflect the depth error.

**Zero-Shot Generalization on ViViD++.** To evaluate zero-shot generalization, we assess various methods on the outdoor test split of the ViViD++ dataset. All networks are evaluated with weights trained on the MS<sup>2</sup> dataset. The results for unweighted metrics are presented in the top part of Table III, with qualitative results shown in Fig. 5. We report the zero-shot performance of both our RGB MDE model and thermal MDE model, denoted ‘Ours-ZS’. Since the test sequences are collected at night, the thermal MDE model significantly outperforms the RGB MDE model, achieving RMSE values of 4.440 and 5.903, respectively.

Additionally, we report the results of our method trained without RGB-to-thermal distillation, denoted as ‘Ours-NoDist-ZS’. It is evident that ‘Ours-ZS’ with distillation demonstrates superior performance and better generalization compared to ‘Ours-NoDist-ZS’, highlighting the effectiveness of our distillation approach.

**Self-Supervised Finetuning on ViViD++.** One of the key advantages of our framework is its ability to perform self-supervised fine-tuning of the thermal model using the RGB model, especially in new applications where ground-truth depth is not available. Before fine-tuning, we evaluated the zero-shot performance of our method on the outdoor training split (captured at daytime) of the ViViD++ dataset, with results shown in Table III. The RGB model demonstrates good generalization, achieving an RMSE of 4.199, which is much lower than the thermal model’s RMSE of 5.335. Therefore, the RGB model is able to teach the thermal model and improve its performance.

We fine-tuned our thermal MDE model using the RGB MDE model on the training split of the ViViD++ dataset

## IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE I: Quantitative evaluation of MDE results with various methods on the MS<sup>2</sup> dataset [9]. (Unweighted metrics)

Methods	Modality	Error ↓				Accuracy ↑		
		AbsRel	SqRel	RMSE	RMSElog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
DORN [12]	Ther	0.109	0.540	3.66	0.144	0.887	0.982	0.997
BTS [40]	Ther	0.086	0.380	3.163	0.117	0.926	0.990	0.998
Adabins [36]	Ther	0.088	0.377	3.152	0.119	0.924	0.990	0.998
NeWCRF [41]	Ther	0.080	0.331	2.937	0.109	0.937	0.993	0.999
ZoeDepth [14]	Ther	0.091	0.425	3.202	0.123	0.915	0.989	0.998
DepthAnything [5]	Ther	0.075	0.287	2.719	0.103	0.945	<b>0.995</b>	<b>0.999</b>
Ours	Ther	<b>0.072</b>	<b>0.275</b>	<b>2.677</b>	<b>0.100</b>	<b>0.949</b>	<b>0.995</b>	<b>0.999</b>

TABLE II: Detailed quantitative evaluations under ‘Day’, ‘Night’ and ‘Rainy’ conditions on the MS<sup>2</sup> dataset [9]. We show both the unweighted (UnW.) and weighted (W.) metrics.

Metrics	TestSet	Method	Error ↓				Accuracy ↑		
			AbsREL	SqRel	RMSE	RMSElog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
UnW.	Day	DepthAnything [5]	0.063	0.222	2.477	0.091	0.959	0.996	<b>0.999</b>
		Zoedepth [14]	0.078	0.342	2.979	0.110	0.932	0.992	<b>0.999</b>
		Ours	<b>0.059</b>	<b>0.210</b>	<b>2.420</b>	<b>0.087</b>	<b>0.965</b>	<b>0.997</b>	<b>0.999</b>
	Night	DepthAnything [5]	0.077	0.276	2.565	0.102	0.944	<b>0.996</b>	<b>1.000</b>
		Zoedepth [14]	0.087	0.344	2.827	0.114	0.927	0.993	0.999
		Ours	<b>0.075</b>	<b>0.268</b>	<b>2.541</b>	<b>0.101</b>	<b>0.945</b>	<b>0.996</b>	<b>1.000</b>
	Rainy	DepthAnything [5]	0.085	0.358	3.085	0.115	0.932	0.992	<b>0.999</b>
		Zoedepth [14]	0.107	0.577	3.754	0.143	0.888	0.983	0.996
		Ours	<b>0.080</b>	<b>0.342</b>	<b>3.041</b>	<b>0.111</b>	<b>0.939</b>	<b>0.993</b>	<b>0.999</b>
W.	Day	DepthAnything [5]	0.083	0.553	4.181	0.104	0.935	0.995	0.999
		Zoedepth [14]	0.100	0.804	5.015	0.124	0.894	0.989	0.999
		Ours	<b>0.079</b>	<b>0.538</b>	<b>4.126</b>	<b>0.099</b>	<b>0.942</b>	<b>0.996</b>	<b>1.000</b>
	Night	DepthAnything [5]	0.091	0.603	4.081	0.107	0.924	0.994	<b>1.000</b>
		Zoedepth [14]	0.100	0.713	4.483	0.117	0.907	0.992	0.999
		Ours	<b>0.088</b>	<b>0.597</b>	<b>4.075</b>	<b>0.105</b>	<b>0.925</b>	<b>0.995</b>	<b>1.000</b>
	Rainy	DepthAnything [5]	0.100	0.724	4.768	0.122	0.905	0.991	<b>0.999</b>
		Zoedepth [14]	0.122	1.055	5.750	0.148	0.844	0.983	0.997
		Ours	<b>0.096</b>	<b>0.715</b>	<b>4.755</b>	<b>0.118</b>	<b>0.910</b>	<b>0.992</b>	<b>0.999</b>

TABLE III: Generalization and self-supervised finetuning test on ViViD++ outdoor dataset [42]. ‘ZS’ means ‘Zero Shot’, and ‘SSFT’ means ‘Self supervised finetuning.’ Unweighted metrics are shown.

Split	Method	Modality	Error ↓				Accuracy ↑		
			AbsRel	SqRel	RMSE	RMSElog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Test	ZoeDepth [14]	Ther	0.174	1.163	5.633	0.227	0.685	0.944	0.989
	DepthAnything [5]	Ther	0.166	0.937	4.929	0.204	0.719	0.973	0.994
	Ours-NoDist-ZS	Ther	0.154	0.798	4.540	0.187	0.758	0.983	<b>0.997</b>
	Ours-ZS	RGB	0.198	2.037	5.903	0.241	0.710	0.940	0.978
	Ours-ZS	Ther	0.153	0.801	4.440	0.184	0.768	0.984	<b>0.997</b>
	Ours-SSFT	Ther	<b>0.118</b>	<b>0.593</b>	<b>4.006</b>	<b>0.147</b>	<b>0.897</b>	<b>0.988</b>	<b>0.997</b>
Training	Ours-ZS	Ther	0.164	1.042	5.335	0.203	0.718	0.971	0.996
	Ours-ZS	RGB	0.124	0.641	4.199	0.153	0.883	0.988	0.997

through our proposed confidence-aware distillation. Notably, no ground-truth information was used during the self-supervised fine-tuning process. After fine-tuning, the performance of the thermal MDE model significantly improved, as denoted by ‘Ours-SSFT’ in Table III. Compared to the zero-shot thermal model ‘Ours-ZS’, ‘Ours-SSFT’ shows a substantial reduction in AbsRel from 0.153 to 0.118 (a 22.88% decrease) and an improvement in threshold accuracy ( $\delta < 1.25$ ) from 76.8% to 89.7%. The predicted depth from ‘Ours-ZS’ and ‘Ours-SSFT’ and their error maps are visualized in Fig. 5. These results show that our confidence-aware distillation method effectively enhances thermal MDE performance in new scenarios, which is crucial for deploying thermal MDE models in novel scenarios.

### E. Ablation Study

We conducted an ablation study on the MS<sup>2</sup> dataset to evaluate the impact of various design choices in our framework. The following configurations were examined: (1) No Distillation: This configuration trains the thermal and RGB MDE models together using their respective ground-truth depth maps, without any distillation step. (2) No Confidence Network: In this setup, we perform distillation from the RGB model to the thermal model without incorporating the confidence network for weighting. (3) No Multi-Modal Confidence: Here, we remove all input components related to the thermal branch when predicting the confidence of the RGB MDE, using only RGB-related metadata in the confidence network.

The results of these configurations are summarized in Table IV, where they are denoted as ‘No Dist.’, ‘No Conf.’,

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

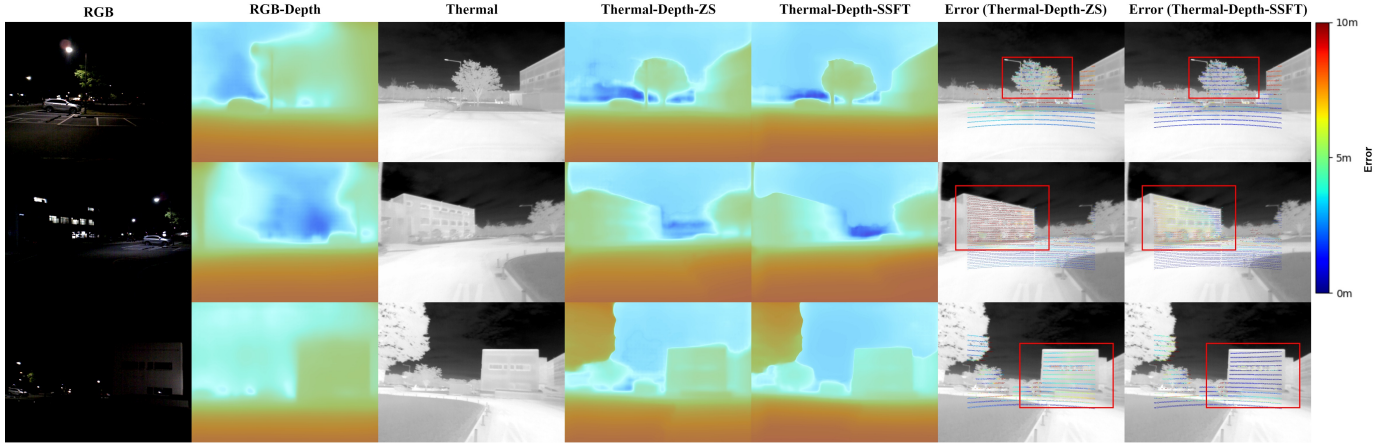


Fig. 5: **Monocular Depth Estimation on ViViD++ dataset [10].** From left to right: the RGB image, our predicted RGB depth, the normalized thermal image, our predicted thermal depth with zero-shot (Thermal-Depth-ZS), our predicted thermal depth after self-supervised fine-tuning (Thermal-Depth-SSFT), the depth error of Thermal-Depth-ZS, and the depth error of Thermal-Depth-SSFT. The red boxes highlight areas with a significant decrease in error after self-supervised fine-tuning.

TABLE IV: **The ablation study for our design choices.** The weighted metrics are shown.

Setup	RMSE↓	SqRel ↓	$\delta < 1.25 \uparrow$
Ours	4.330	0.619	0.925
No Dist.	4.359	0.626	0.923
No Conf.	4.469	0.664	0.919
No Mul. Conf.	4.345	0.625	0.924

and ‘No Mul. Conf.’, respectively. Our full method, which includes all design choices, achieves the best performance. Interestingly, the results reveal that distillation without the predicted confidence (‘No Conf.’) is detrimental to thermal MDE performance, even showing significantly worse results compared to the setup without any distillation. This underscores the importance of confidence-aware distillation.

#### F. Validation on Real Robots

We have conducted a series of qualitative zero-shot experiments at night on robotic hardware in a real-world setting, as shown in Fig. 6, where MonoTher-Depth is deployed in a zero-shot manner. The predicted depth image is projected into a point cloud, which is then processed with uniform downsampling and radial outlier rejection. Ground points are removed, and the resulting point cloud is flattened into a 2D plane. Finally, the points are clustered using the density-based spatial clustering (DBSCAN) algorithm [43] and converted into a map of convex polygons via the alphashape method [44]. The resulting 2D polygonal obstacle maps derived from thermal depths are shown on the far right of Fig. 6 and are very close to the 2D polygonal obstacle maps obtained from 3D LiDAR point clouds. These 2D polygonal maps from thermal depths are then used for obstacle avoidance along a predefined path to a waypoint, successfully achieving collision-free navigation. These experiments demonstrate the effectiveness of our proposed MonoTher-Depth in out-of-distribution settings for real-world robotic applications.

#### G. Discussion and Limitations

Our confidence-aware distillation method demonstrates impressive performance across scenarios with and without la-

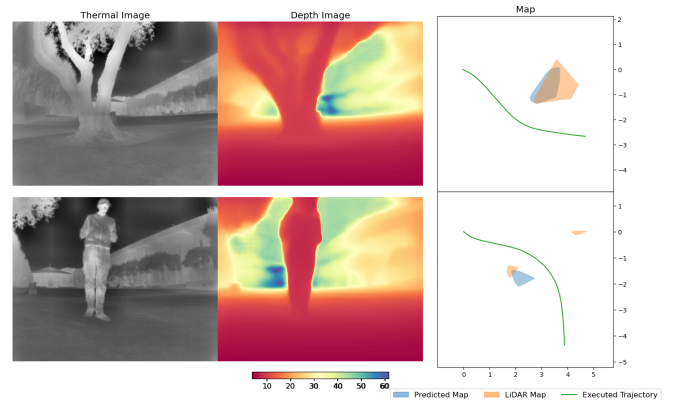


Fig. 6: **Zero-shot deployment of MonoTher-Depth** onto a robotic hardware platform, a Traxxas-Maxx car equipped with a FLIR ADK thermal camera, Velodyne Puck LiDAR, and Simply NUC Ruby for computation. Two example demonstrations, in which the robot is tasked with avoiding an obstacle (a tree in the top series, a person in the bottom series). The resulting polygonal obstacle map, along with a map derived from ground-truth LiDAR data, is shown in the far right.

beled depth supervision. However, there are inherent limitations that MDE models trained on outdoor datasets struggle with indoor scenarios due to significant domain gaps. This outdoor-to-indoor generalization issue is also shown in [5] and [14]. Similarly, our models trained on the MS<sup>2</sup> outdoor dataset exhibit poor zero-shot performance on the ViViD++ indoor scenarios [42]. The self-supervised fine-tuning using RGB-to-thermal distillation does not substantially improve performance in such cases, primarily due to the already poor predictions in indoor settings.

#### V. CONCLUSION

We present MonoTher-Depth, a thermal monocular depth estimation (MDE) method that incorporates knowledge distilled from large, foundational RGB MDE models. To prevent the thermal MDE model from being adversely affected by the RGB MDE model in challenging scenarios, we introduce a novel confidence-aware distillation method. This approach

**IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.**

adaptively adjusts the distillation strength based on the predicted confidence, utilizing both thermal and RGB information. By incorporating confidence-aware distillation, our thermal model achieves significant improvements in depth estimation accuracy, particularly in new scenarios where ground-truth depth supervision is unavailable.

REFERENCES

- [1] U. Rajapaksha, F. Soheli, H. Laga, D. Diepeveen, and M. Bannamoun, "Deep learning-based depth estimation methods from monocular image and videos: A comprehensive survey," *ACM Computing Surveys*, 2024.
- [2] L. Piccinelli, Y.-H. Yang, C. Sakaridis, M. Segu, S. Li, L. Van Gool, and F. Yu, "Unidepth: Universal monocular metric depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 10 106–10 116.
- [3] W. Yin, C. Zhang, H. Chen, Z. Cai, G. Yu, K. Wang, X. Chen, and C. Shen, "Metric3d: Towards zero-shot metric 3d prediction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 9043–9053.
- [4] V. Guizilini, I. Vasiljevic, D. Chen, R. Ambrus, and A. Gaidon, "Towards zero-shot scale-aware monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 9233–9243.
- [5] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 10 371–10 381.
- [6] S. S. Shivakumar, N. Rodrigues, A. Zhou, I. D. Miller, V. Kumar, and C. J. Taylor, "PST900: RGB-Thermal calibration, dataset and segmentation network," in *Proc. IEEE Int. Conf. Rob. Autom.* IEEE, 2020, pp. 9441–9447.
- [7] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, 2020.
- [8] S. A. Deevi, C. Lee, L. Gan, S. Nagesh, G. Pandey, and S.-J. Chung, "Rgb-x object detection via scene-specific fusion modules," in *Proc. IEEE/CVF Winter Conf. on Appl. of Comput. Vis.*, 2024, pp. 7366–7375.
- [9] U. Shin, J. Park, and I. S. Kweon, "Deep depth estimation from thermal image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 1043–1053.
- [10] A. J. Lee, Y. Cho, Y.-s. Shin, A. Kim, and H. Myung, "Vivid++: Vision for visibility dataset," *IEEE Robot. and Auto. Lett.*, vol. 7, no. 3, pp. 6282–6289, 2022.
- [11] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Proc. Adv. Neural Inform. Process. Syst.*, vol. 27, 2014.
- [12] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 2002–2011.
- [13] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, 2020.
- [14] S. F. Bhat, R. Birkel, D. Wofk, P. Wonka, and M. Müller, "Zoodepth: Zero-shot transfer by combining relative and metric depth," *arXiv:2302.12288*, 2023.
- [15] C. Liu, S. Kumar, S. Gu, R. Timofte, and L. Van Gool, "Va-depthnet: A variational approach to single image depth prediction," in *Proc. Int. Conf. Learn. Represent.*, 2023.
- [16] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, K. Wang, G. Yu, C. Shen, and S. Shen, "Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation," *arXiv:2404.15506*, 2024.
- [17] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 9492–9502.
- [18] S. Shao, Z. Pei, X. Wu, Z. Liu, W. Chen, and Z. Li, "Iebins: Iterative elastic bins for monocular depth estimation," in *Proc. Adv. Neural Inform. Process. Syst.*, vol. 36, 2024.
- [19] U. Shin, K. Lee, B.-U. Lee, and I. S. Kweon, "Maximizing self-supervision from thermal image for effective self-supervised learning of depth and ego-motion," *IEEE Robot. and Auto. Lett.*, vol. 7, no. 3, pp. 7771–7778, 2022.
- [20] U. Shin, K. Park, B.-U. Lee, K. Lee, and I. S. Kweon, "Self-supervised monocular depth estimation from thermal images via adversarial multi-spectral adaptation," in *Proc. IEEE/CVF Winter Conf. on Appl. of Comput. Vis.*, 2023, pp. 5798–5807.
- [21] J. Xu, X. Liu, J. Jiang, K. Jiang, R. Li, K. Cheng, and X. Ji, "Unveiling the depths: A multi-modal fusion framework for challenging scenarios," *arXiv:2402.11826*, 2024.
- [22] U. Shin, K. Lee, S. Lee, and I. S. Kweon, "Self-supervised depth and ego-motion estimation for monocular thermal video using multi-spectral consistency loss," *IEEE Robot. and Auto. Lett.*, vol. 7, no. 2, pp. 1103–1110, 2021.
- [23] Y. Guo, H. Kong, and S. Gu, "Unsupervised multi-spectrum stereo depth estimation for all-day vision," *IEEE Trans. on Intell. Veh.*, 2023.
- [24] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv:1503.02531*, 2015.
- [25] X. Guo, H. Li, S. Yi, J. Ren, and X. Wang, "Learning monocular depth by distilling cross-domain stereo networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 484–500.
- [26] F. Tosi, F. Aleotti, M. Poggi, and S. Mattoccia, "Learning monocular depth estimation infusing traditional stereo knowledge," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 9799–9809.
- [27] A. Pilzer, S. Lathuiliere, N. Sebe, and E. Ricci, "Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 9768–9777.
- [28] F. Aleotti, G. Zaccaroni, L. Bartolomei, M. Poggi, F. Tosi, and S. Mattoccia, "Real-time single image depth perception in the wild with handheld devices," *Sensors*, vol. 21, no. 1, p. 15, 2020.
- [29] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia, "On the uncertainty of self-supervised monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 3227–3237.
- [30] S. Shao, Z. Pei, W. Chen, R. Li, Z. Liu, and Z. Li, "Urcdc-depth: Uncertainty rectified cross-distillation with outflip for monocular depth estimation," *IEEE Trans. Multimedia*, 2023.
- [31] X. Shi, G. Dikov, G. Reitmayr, T.-K. Kim, and M. Ghafoorian, "3d distillation: Improving self-supervised monocular depth estimation on reflective surfaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 9133–9143.
- [32] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *Trans. on Machine Learning Research*, 2023.
- [33] C. Lee, M. Anderson, N. Raganathan, X. Zuo, K. Do, G. Gkioxari, and S.-J. Chung, "Caltech aerial rgb-thermal dataset in the wild," in *Proc. Eur. Conf. Comput. Vis.*, 2024.
- [34] C. Lee, J. G. Frennert, L. Gan, M. Anderson, and S.-J. Chung, "Online self-supervised thermal water segmentation for aerial vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2023, pp. 7734–7741.
- [35] A. M. Reza, "Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement," *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 38, pp. 35–44, 2004.
- [36] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 4009–4018.
- [37] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3828–3838.
- [38] L. Lipson, Z. Teed, and J. Deng, "Raft-stereo: Multilevel recurrent field transforms for stereo matching," in *2021 International Conf. on 3D Vision (3DV)*. IEEE, 2021, pp. 218–227.
- [39] M. Vankadari, S. Hodgson, S. Shin, K. Zhou, A. Markham, and N. Trigoni, "Dusk till dawn: Self-supervised nighttime stereo depth estimation using visual foundation models," in *2024 IEEE International Conference on Robotics and Automation (Proc. IEEE Int. Conf. Rob. Autom.)*. IEEE, 2024, pp. 17 976–17 982.
- [40] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," *arXiv:1907.10326*, 2019.
- [41] W. Yuan, X. Gu, Z. Dai, S. Zhu, and P. Tan, "New crfs: Neural window fully-connected crfs for monocular depth estimation," *arXiv:2203.01502*, 2022.
- [42] A. J. Lee, Y. Cho, Y.-s. Shin, A. Kim, and H. Myung, "Vivid++: Vision for visibility dataset," *IEEE Robot. and Auto. Lett.*, vol. 7, no. 3, pp. 6282–6289, 2022.
- [43] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.
- [44] N. Akkiraju, H. Edelsbrunner, M. Facello, P. Fu, E. Mucke, and C. Varella, "Alpha shapes: definition and software," in *Proc. Intl. Comput. Geom. Software Workshop*, vol. 63, no. 66, 1995.