2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Generalizable Thermal-based Depth Estimation via Pre-trained Visual Foundation Model

Ruoyu Fan[1], Wang Zhao[1], Matthieu Lin[1], Qi Wang[1,2], Yong-Jin Liu[1,*], *Senior Member, IEEE* and Wenping Wang[3], *Fellow, IEEE*

***Abstract*— Depth estimation is a crucial task in computer vision, applicable to various domains such as 3D reconstruction, robotics, and autonomous driving. In particular, thermal-based depth estimation has unique advantages, including night-time vision. However, the existing depth estimation method remains challenging in robust generalization due to limited data resources and spectral differences between thermal and RGB images. In this paper, we present a self-supervised approach to enhance thermal-based depth estimation by leveraging pre-trained visual models initially designed for RGB data. In detail, we design a novel two-stage training strategy, incorporating Low-rank Adapters and Convolutional Adapters, which not only significantly improves accuracy and robustness but also enables impressive zero-shot generalization capabilities. Our method outperforms existing thermal-based depth estimation models, opening new possibilities for cross-modal applications in computer vision and robotics research.**

I. Introduction

Depth Estimation is a fundamental task in computer vision with a wide array of applications ranging from 3D reconstruction and robotics to autonomous driving [1]–[4]. Over the years, researchers have explored various depth estimation methods leveraging various sensors, including RGB cameras, depth cameras, LiDAR, etc. Among these, thermal image-based methods have garnered increased attention due to their unique capabilities, such as nighttime vision and relatively dense resolution. With these advantages of thermal images, we can realize all-day depth estimation for outdoor scenes, showing considerable application potential in the future.

Consequently, researchers gradually focus on thermal-based depth estimation. MTN [7] designed a depth estimation network trained with monocular thermal and stereo RGB images through RGB-thermal spectral transfer. By leveraging the geometric constraints of depth and camera relative pose, Shin *et al.* [8] recently introduced self-supervised depth-pose learning into thermal-based depth estimation, easing the need for costly calibrated LiDAR points as groundtruth.

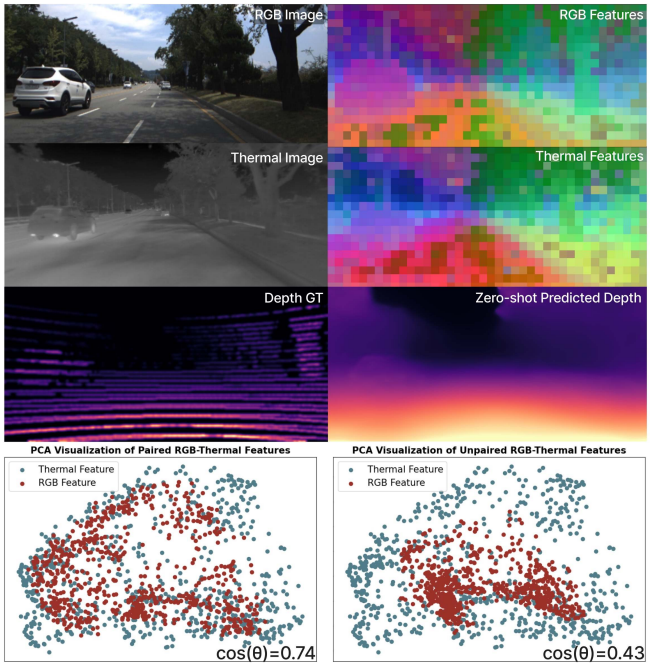


Fig. 1: Visualized Features from RGB and Thermal Images. We train a depth estimation network using Dinov2 [5] features on a thermal dataset and assess its generalizability the zero-shot dataset MS2 [6]. The scatterplots at the bottom visually depict the features of paired and unpaired RGB-thermal image pairs. The $\cos(\theta)$ denotes cosine similarity. Notably, paired images sharing the same scene semantic information exhibit more relevant distributions.

Following this line of research, Shin *et al.* [6] constructed the thermal-depth dataset and trained networks to predict depth from a single thermal image. More work [9]–[11] improved the performance through advanced network architectures, additional training regularizers, etc.

Despite the huge progress made in thermal-based depth estimation, few studies have addressed the challenge of generalizable thermal-based depth estimation. On the one hand, this is essential for real-world thermal perception but is non-trivial due to inherent data challenges in thermal images, including substantial noise, limited visual texture, and varying spectral capture frequencies. On the other hand, compared to RGB images, which contain rich appearance and semantic cues, predicting depth from monocular thermal images necessitates learning robust data priors. Unfortunately, most existing thermal datasets consist of fewer than 100,000 thermal images and are captured by various Long Wave Infrared (LWIR) sensors. As a result, it poses a

This work was partially supported by the Natural Science Foundation of China (Project Number 62332019, U2336214, 62162008), Beijing Natural Science Foundation (L222008), 2022 Special Funds of Zhongguancun Science City's Key Core Technologies, and Beijing Hospitals Authority Clinical Medicine Development of special funding support (ZLRK202330).

[1]R. Fan, W. Zhao, M. Lin and Y.-J. Liu are with the Department of Computer Science and Technology, Tsinghua University, China {fry21, zhao-w19, lyh21}@mails.tsinghua.edu.cn, liuyongjin@tsinghua.edu.cn

[2] Q. Wang is with the State Key Laboratory of Public Big Data, Guizhou University, Guiyang, China qiwang@gzu.edu.cn

[3]W. Wang is with the Department of Computer Science and Engineering, Texas A&M University, USA wenping@tamu.edu

*Corresponding Author

significant hurdle for training a generalizable thermal-based depth estimation network, as models trained on one dataset struggle to generalize to others and the open world. To address this challenge, our approach centers on harnessing the extensive knowledge encoded in a large-scale pre-trained visual foundation model to enhance its estimation and generalization capabilities. Specifically, we leverage a large-scale pre-trained model as an addtional source of data priors to enhance generalization capabilities and effectively fine-tune it in a cross-spectral manner, without sacrificing generalization. Our inspiration draws from recent breakthroughs in visual foundation models, such as [12]–[14], and their demonstrated effectiveness in diverse downstream tasks. Through our experiments, we unveil an intriguing finding: the large-scale pre-trained RGB encoders can also extract meaningful visual features even from distinct spectral domains, including thermal imagery (as shown in Fig. 1). This observation underscores the uptapped potential of utilizing RGB foundation models, pre-trained on billion-scale images, to assist cross-spectral tasks.

In particular, we develop a self-supervised thermal-based depth estimation system based on the RGB foundation model backbones, and propose an two-stage strategy to fine-tune it. The first stage focuses on training the depth output head and pose estimation network while maintaining the pre-trained backbone in a frozen state. This first stage enhances generalization by leveraging zero-shot features from the backbone. During the second stage, we proceed to fine-tune the backbone utilizing advanced techniques such as the Low-rank Adapter [15] (LoRA) and convolutional adapter. This refinement process enhances depth estimation without compromising generalization. The combined effect of this two-stage strategy culminates in the development of a precise, resilient, and adaptable monocular depth estimation system based on thermal imaging. Extensive experiments on several thermal-based depth benchmarks validate the state-of-the-art performance of our method and its significantly improved generalization abilities compared to baseline approaches.

We summarize our contributions as follows:

1) We present a novel insight into the cross-spectral generalization potential of pre-trained visual RGB foundation models, which we then leverage in the domain of self-supervised thermal-based depth estimation.
2) We introduce an effective two-stage training strategy with self-supervised guidance to adapt RGB foundation models for thermal-based depth estimation.
3) Our system demonstrates state-of-the-art performance on the training dataset ViViD [16], and showcases superior zero-shot generalization capabilities on the ViVid++ [17] and MS2 [6] datasets. Real-world experiments with captured data further validate the practical utility of our proposed system.

II. Related Work

A. Depth Estimation with Thermal Images

Previous methods primarily explore depth estimation with thermal images in adverse weather and lighting conditions. MTN [7] proposed a multi-task learning architecture for thermal image depth estimation, demanding stereo RGB image and monocular thermal image during training. Lu et al. [9] followed this setting and exploited a cross-spectral translation network to train a single-view depth network. Shin et al. [8] proposed multi-spectral temporal consistency loss to train a monocular depth estimation network with the supervision of visual image reconstruction loss. Recently, Shin et al. [18] pointed out the challenge of thermal-based depth estimation and proposed an image-enhancing method that outperforms other thermal-only depth estimation methods.

Nevertheless, a common limitation among these methods is their heavy reliance on limited data resources, leading to potential overfitting issues that constrain their ability to generalize beyond their training datasets. To address this limitation, we introduce the incorporation of generalizable visual prior knowledge, offering a potential solution to enhance the model's extending capability.

B. Self-supervised Depth Estimation

Self-supervised learning for depth and ego-motion is to estimate geometric metrics without expensive ground-truth labels. SfM-learner [19] trains the depth and pose estimation networks with the temporal image reconstruction loss. This method suffers from brightness inconsistency, non-Labertain surface, and moving objects. Therefore, various pixel masking techniques have been proposed for the robustness of reconstruction loss, such as explainability mask [20], depth inconsistency mask [21], [22] and flow mask [23], [24]. Monodepth2 [25] proposed minimum reprojection loss and auto-masking loss, which handle the occlusions and pixels keeping relatively stationary with the camera. In line with the self-supervised depth estimation paradigm, we adopt a similar approach, leveraging a pre-trained backbone to tackle the depth estimation task using thermal images.

C. Adapter-based Tuning

Adapter-based tuning has applications in various domains. In NLP, Houlsby [26] introduced a bottleneck structure adapter module, adding a minimal number of trainable parameters to a pre-trained fixed model, achieving results comparable to fine-tuning the entire network. This concept has been extended to introduce fewer trainable parameters [27], reduce inference time [15], [28], and enhance performance [29], [30]. In computer vision, adapters have been proposed for image classification tasks [31], [32], and well-designed adapters have been applied in dense prediction tasks [33]–[35]. This paper leverage LoRA [15] as a key component of our adapters, underscoring the adaptability and versatility of this technique for optimizing and fine-tuning models.

III. Method

In Sec. III-A, we introduce the model architecture, including the pre-trained backbone, output head, and pose network, which forms the foundation of our system in stage one. Then, in Sec. III-B, we describe the adapter-based tuning strategy

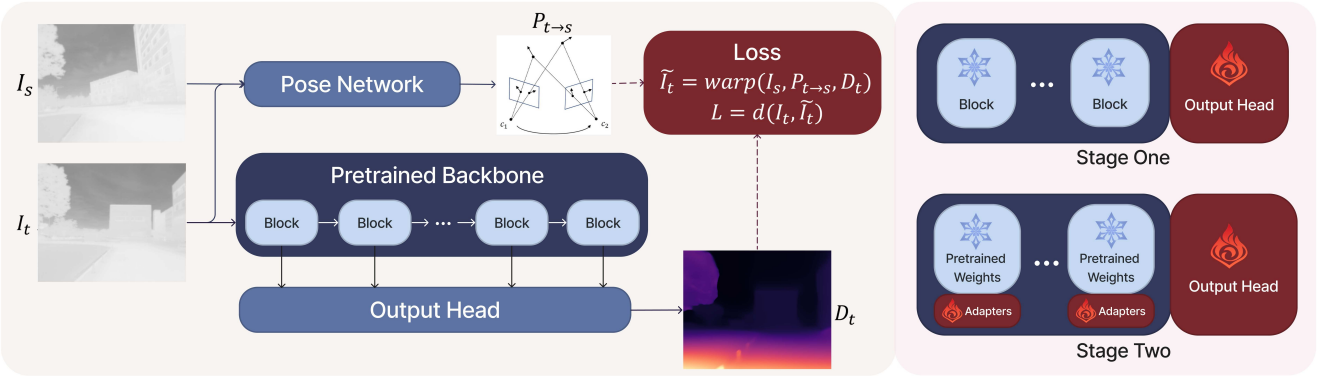


Fig. 2: Our Approach. **Left**: The complete pipeline of our method integrates a pre-trained visual backbone with an output module for monocular thermal-based depth estimation. The predicted depth, along with the estimated relative pose, drives reconstruction loss for network training. **Right**: Our two-stage training scheme starts with freezing the entire backbone and training the output head in stage one. Adapters refine transformer weights in stage two.

and its impact on our method, enhancing depth estimation in stage two. Finally, in Sec. III-C, we present detailed loss functions and masking strategies employed in our approach.

A. Stage One: Building a Foundation

To build a strong and efficient foundation model, we introduce the RGB-pretrained visual model Dinov2 [5], which is originally developed for processing visual RGB data. As illustrated in Sec. I, the Dinov2 has the capability to effectively extract meaningful features from thermal images. Through adaption and transfer training, we leverage this transformer-based architecture to capture the inherent patterns and structures in thermal data. This pre-trained architecture not only accelerates the converging speed but enhances the comprehensive ability of the depth network.

To further elaborate on the network architecture, we extend the capabilities of the transformer backbone for thermal-based depth estimation by incorporating an output head inspired by the Dense Prediction Vision Transformer (DPT) [36]. This output head is specifically trained to comprehend the features extracted from different blocks within the pre-trained transformer backbone and efficiently translate them into depth maps using transpose convolutional layers. In addition, we construct a pose network using CNN to estimate the relative pose between two adjacent thermal images.

Stage One. To preserve the prior knowledge embedded in the pre-trained model for representing thermal images, we freeze the transformer backbone while training the output head from scratch for depth estimation. Simultaneously, the pose network is trained to predict the 6-DoF relative pose, crucial for computing the reconstruction loss within the self-supervised training paradigm.

The initial stage imparts the depth network (pre-trained backbone + output head) with a foundational understanding of thermal images. Through transfer learning, the model acquires a fundamental generalization capability in thermal image depth estimation tasks. It predicts reasonable depth maps even in the absence of specific thermal dataset knowledge, showing its proficiency in zero-shot transfer scenarios.

B. Stage Two: Enhancing Thermal Adaptability

In the second stage of our method, we take a step towards enhancing our model's feature extraction capabilities for thermal images, by jointly fine-tuning the pre-trained visual backbone in the training. However, traditional fine-tuning of large backbone is very costful and may decrease the generalization ability of the backbone due to the limited training data. Furthermore, due to the substantial noise and low-texture regions often present in thermal images, the self-supervised training is potentially ill-posed and may affect the integrity of the original pre-trained backbone weights.

To address these challenges and maintain our model's robustness, we opt for an adapter-based tuning approach. Specifically, inspired by the discovery that adapter-based tuning consistently outperforms traditional fine-tuning in low-resource and cross-lingual tasks [37], [38]. We explore the use of adapters in the second stage, reinforced by experimental results that highlight their effectiveness in enhancing the model's adaptability to thermal data.

More precisely, we integrate both LoRA [15] and convolutional adapters into our tuning pipeline. Specifically, we apply LoRA to W_q and W_v in the attention module of each block in the vision transformer [39]. Our convolutional adapter consists of three layers: a 3x3 convolution layer that reduces the channel dimension from C to the hidden dimension h, a GeLU layer [40] for activation, and a final 3x3 convolution layer that restores the original number of channels. Importantly, we reshape the flattened image features (excluding the [cls] token) from a size of $\frac{HW}{P^2} \times C$ to $C \times \frac{H}{P} \times \frac{W}{P}$ before passing them through these layers. This reorganization process transforms the image tokens into a 2D format, allowing us to incorporate convolutional inductive bias into the vision transformer architecture. The processed features are then flattened and added as a residual to the input. By positioning the convolutional adapters at the end of each block, we emphasize their crucial role in enhancing the model's ability to comprehend thermal images.

Stage Two. We train these adapters, in conjunction with the output head and pose estimation network trained, while

the pre-trained weights of the backbone remain frozen. This process enhances the model's depth estimation capabilities on the training dataset. Compared to training the model for its generalizable ability in stage one, this step tailors the model to the specific characteristics of the dataset at hand, while preserving its generalization prowess for thermal images. Overall, the two-stage approach strikes a balance, empowering the model with adaptability to diverse thermal scenarios without compromising its overall thermal-based depth estimation performance.

C. Self-supervised Losses

Our self-supervised training approach follows a methodology similar to other monocular depth estimation methods and consists of the following components:

1) Image Reconstruction Loss: With the depth map D_t estimated by depth network and relative pose $T_{t\to s}$ estimated by pose network from the input thermal images I_t, I_s, the warped image $\tilde{I}_t$ is synthesized in the projective geometry manner [19]. The thermal image reconstruction loss L_{rec} is calculated by measuring the L1 difference and the Structural Similarity Index Map (SSIM) [41] between I_t and $\tilde{I}_t$:

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

where α indicates the factor scale. The reconstruction loss plays the main role in the model optimization.

2) Smoothness Loss: The smoothness loss L_{sm} [25] promotes edge-aware smoothness by comparing the input image to the output depth, which effectively refines the depth predictions in low-texture regions.

$$L_{sm} = \sum(|\partial_x D_t| \cdot e^{\partial_x I_t} + |\partial_y D_t| \cdot e^{\partial_y I_t}). \quad (2)$$

3) Geometry Consistency Loss: The geometry consistency loss L_{gc} [21] minimizes the relative difference of warped depth $\tilde{D}_t$ and D_t, to have the same scale-consistency geometry structure. It is defined as:

$$L_{gc} = \frac{1}{|V|}\sum_{p\in V} D_{diff}(p),\ D_{diff} = \frac{|\tilde{D}_t - D_t|}{\tilde{D}_t + D_t},$$

where $\tilde{D}_t$ is computed by warping D_s with the estimated D_t and $T_{t\to s}$, V indicates the set of valid pixels projected from D_s to D_t, $|V|$ denotes the number of pixels in V.

With the scale factors λ_{sm}, λ_{gc}, the overall training loss for depth and pose estimation networks in two stages is defined as follows:

$$L_{total} = L_{rec} + \lambda_{sm}L_{sm} + \lambda_{gc}L_{gc}. \quad (3)$$

4) Masks and Image Enhancement: Besides the above losses, we utilize several methods to improve training. An auto-mask mechanism [25] is integrated to exclude stationary pixels in vehicle movement scenarios. A geometrically inconsistent pixel mask [21] is introduced to disregard moving objects and occluded regions. Temporal Mapping [18] is applied to equalize and rearrange pixel values in thermal image sequences as a superior self-supervised signal in stage two. Overall, the combination of these self-supervised training components ensures that our model acquires a robust depth estimation capability for thermal images despite the inherent challenges presented by such data.

IV. Experiments

A. Implentation Details

To validate our proposed method, we trained our networks on the ViViD [16] outdoor dataset, with 2 sequences as a training set (2225 images) and 2 sequences as a testing set (2023 images). For the experiment of zero-shot dataset transfer, we used MS2 [6] and ViViD++ [17] to evaluate the generalizing ability of our method.

We used Dinov2 [5] with pre-trained ViT-Base/14 architecture as our fixed feature extractor. The networks are trained on a single A100 GPU with 80 GB memory for 400 epochs on stage one and 400 epochs on stage two using the Adam optimizer [42], which takes about 24 hours to train our networks in total. The learning rate is set to $1e^{-4}$. Our code is implemented using PyTorch Library [43]. We utilize random crop and horizontal flip augmentations for input thermal images. The resolution of input thermal images is set to 256×320. The rank of LoRA is set to 4. The reduced dimension of the convolutional adapter h is set to 4. The scale factors λ_{sm} and λ_{gc} are set to 0.1 and 0.5 seperately. The parameter α is set to 0.85.

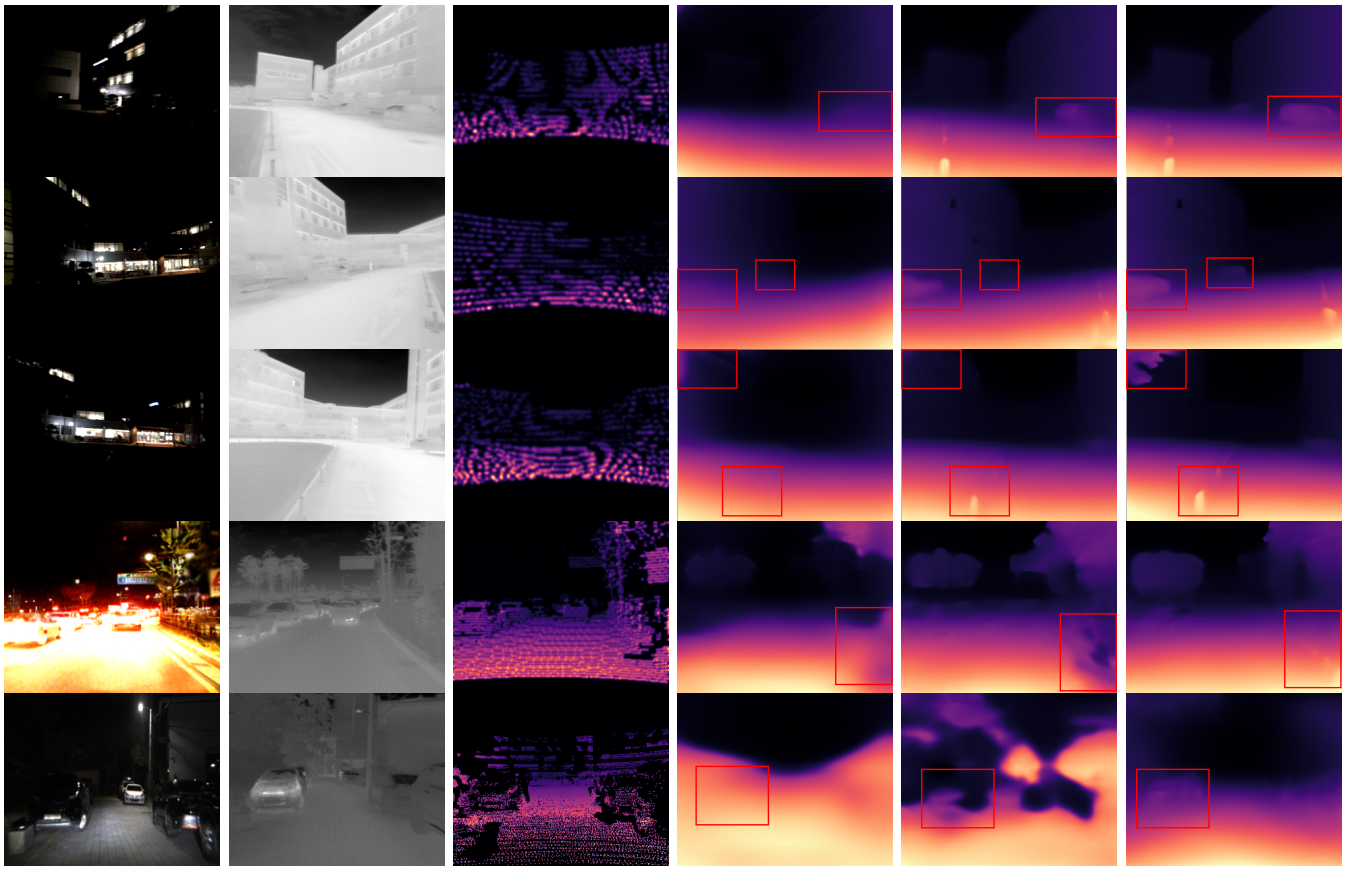
B. Self-supervised Estimation Results

In our experimental evaluation, we benchmark our model against state-of-the-art unsupervised depth estimation methods using outdoor ViViD thermal image sequences. The models are trained on day-time sequences and tested on night-time low-light sequences with different inputs and supervising signals. Specifically, we compared our method to three state-of-the-art approaches: Bian et al. [21], which is designed for RGB sequences, and two other methods specifically tailored for thermal sequences [8], [18]. The comparison results are presented in Table I. The comprehensive results provide a quantitative assessment, showing the superior performance of our methods. Additionally, we offer qualitative comparison results in Fig. 3.

These results emphasize the enhanced accuracy of our thermal-based depth estimation method compared to state-of-the-art methods, demonstrating the effectiveness of leveraging prior knowledge from pre-trained visual models. Our method consistently improves from stage one to stage two, underscoring the continuous refinement of thermal-based depth estimation in our two-stage strategy.

C. Pose Estimation Results

As a side product of our method, the pose network estimates relative pose to track thermal image sequences. We evaluate the proposed pose network against ORB-SLAM2 [45] (designed for RGB) and Shin [18] on two night-time test sequences with Absolute Trajectory Error (ATE) and Relative Error (RE) using thermal images, as shown in Table III. Our method shows a comparable result to Shin *et al.* [18].



(a) RGB **(b)** Thermal **(c)** Groundtruth **(d)** Shin *et al.* [8] **(e)** Shin *et al.* [18] **(f)** Ours - Stage 1&2

Fig. 3: Qualitative Results. These figures compare our method with state-of-the-art (SOTA) approaches. The top three rows depict experiments on ViViD, while the bottom two rows display zero-shot depth evaluation results on ViViD++ and MS2, respectively. Our method consistently achieves higher accuracy, producing fine-grained depth maps during self-supervised training.

TABLE I: Self-Supervised Depth Estimation Results on ViViD. Our method surpasses state-of-the-art thermal-based estimation methods.

	Input	Supervision	Error↓				Accuracy↑		
			RMSE	RMSE log	Abs Rel	Sq Rel	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
SfM-Learner [21]	RGB	RGB	12.000	0.595	0.617	9.971	0.400	0.587	0.720
Shin *et al.* [8] (MS)	Thermal	Thermal	4.697	0.184	0.146	0.873	0.801	0.973	0.993
Shin *et al.* [18]	Thermal	R & T	4.132	0.150	0.109	**0.703**	0.887	0.980	**0.994**
Ours - Stage 1	Thermal	Thermal	4.407	0.165	0.121	0.863	0.872	0.976	0.992
Ours - Stage 1&2	Thermal	Thermal	**3.944**	**0.144**	**0.099**	0.741	**0.906**	**0.981**	**0.994**

TABLE II: Zero-Shot Cross-Dataset Estimation Results: Evaluation of models using raw thermal images with proper preprocessing. The brackets after the datasets denote the camera model for thermal images. The upper section of the table showcases methods trained on extensive RGB-Depth datasets, while the lower section presents models trained exclusively on ViViD data.

	Training Dataset		ViViD++ (A65) ↓				MS2(A65C) ↓			
	Input	Supervision	RMSE	RMSE log	Abs Rel	Sq Rel	RMSE	RMSE log	Abs Rel	Sq Rel
MiDaS-v21 [44]	RGB	Depth	10.588	0.403	0.289	3.651	9.449	0.326	0.237	2.641
DPT-Hybrid [36]	RGB	Depth	11.686	0.398	0.337	7.615	8.422	**0.250**	**0.202**	2.768
DPT-Large [36]	RGB	Depth	10.299	**0.369**	0.281	4.611	8.362	0.270	0.203	2.274
Shin *et al.* [8]	Thermal	R & T	10.979	0.420	0.377	5.751	15.031	0.614	0.577	11.598
Shin *et al.* [18]	Thermal	Thermal	10.245	0.412	0.330	4.267	13.342	0.529	0.521	9.039
Ours - Stage 1	Thermal	Thermal	10.499	0.395	**0.273**	**3.032**	8.356	0.299	0.230	2.292
Ours - Stage 1&2	Thermal	Thermal	**9.653**	0.387	0.303	3.464	**8.251**	0.295	0.228	**2.234**

D. Ablation Study

To assess the individual contributions of our proposed components, we conduct an ablation study on ViViD sequences and summarized the results in Table IV. Notably, omitting the pre-trained backbone rendered the model unable to generate reasonable depth estimates, which highlights the

TABLE III: Pose Estimation Results on ViViD. Our method demonstrates comparable results to the state-of-the-art methods.

	ATE ↓	RE ↓
ORB-SLAM2 [45]	0.1851 ± 0.1234	0.0292 ± 0.0138
Shin *et al.* [18]	$\mathbf{0.0527 \pm 0.0274}$	0.0279 ± 0.0132
Ours	0.0655 ± 0.0333	$\mathbf{0.0276 \pm 0.0131}$

TABLE IV: Ablation Study. Fine-tuning indicates training all the weights in the backbone during stage two. Our method achieves superior performance across these settings.

	RMSE ↓	RMSE log ↓	Abs Rel ↓	Sq Rel ↓
w/o pre-training	9.279	0.473	0.465	5.455
w/o stage two	4.407	0.165	0.121	0.863
w/o adapters	4.023	0.147	0.103	0.849
fine-tuning	4.980	0.199	0.153	1.041
Ours	**3.944**	**0.144**	**0.099**	**0.741**

foundational role played by the pre-trained backbone. Additionally, our two-stage training scheme and the adapter-based tuning significantly enhance the network's performance. As discussed in Sec.III-B, fine-tuning, due to the limited dataset scale and self-supervised paradigm, yielded suboptimal results. It's worth noting that the image reconstruction loss in self-supervised paradigm indirectly optimizes the weights for depth estimation, which may affect the pre-trained weights and their ability to represent thermal images.

E. Zero-shot Estimation Across Datasets

We extensively evaluate the generalization capabilities of our depth estimation network by subjecting it to rigorous testing on two challenging datasets: ViViD++ (8 sequences) and MS2 (20 sequences). These datasets capture data using diverse sensors, requiring our model to infer depth maps from raw LWIR imaging values with varying distributions. To assess the generalizable performance of our method, it is compared to two different types of methods: I. Networks designed for generalized depth estimation, trained on large-scale RGB-Depth datasets [29], [36], and II. Networks exclusively trained on thermal datasets [8], [18], shown in Table II. The qualitative results are shown in 3.

Our experiments reveal several key findings. Firstly, our experiments demonstrate the superior performance of our method in zero-shot cross-dataset transfer compared to category II, while achieving results comparable to category I. Moreover, when appropriate preprocessing is applied to thermal images, generalized RGB-Depth models exhibit reasonable performance in thermal-based depth estimation. However, due to the inherent spectral differences, RGB-based methods sometimes struggle to interpret thermal images effectively, as shown in Fig. 4.

In constrast, networks exclusively trained on thermal datasets perform less effectively in zero-shot evaluations, particularly in the case of the MS2 dataset. This disparity highlights the strength of our model, which is rooted in image representations derived from visual foundation models rather than specific image textures and patterns, allowing it to overcome challenges faced by other thermal models.

(a) MS2-Thermal **(b)** MiDaS-v21 [44] **(c)** Ours - Stage 1&2

Fig. 4: A failure case of MiDaS-v21 [44] for the generablized RGB-Depth model to interpret thermal images.

Fig. 5: Experimental Setup in the Real World: Includes a pan-tilt system, an RGB camera, and a thermal camera.

Furthermore, as discussed in Sec.III-C, during the initial stage, our model gains a general capability for thermal-based depth estimation by leveraging prior knowledge from visual images and a thermal-trained output head. The second stage, focused on refinement, underscores the robustness achieved in stage one, with minimal improvement in zero-shot tasks.

F. Real-world Thermal-based Depth Estimation

In addition to utilizing online datasets for experimentation, we applied our method to real-world depth estimation tasks using infrared thermal imaging, as illustrated in Fig. 5. The camera model we used is the DS-2TD2037T-10. We conducted capture sessions for all-day driving sequences in a campus setting, and the outcomes of these experiments are presented in our accompanying demo video.

V. Conclusion

In this paper, we have addressed the challenge of generalization in thermal-based depth estimation by introducing a pioneering approach. Leveraging the capabilities of pre-trained visual models, we demonstrate the potential to transfer knowledge from the visual to the thermal domain, bridging the spectral gap between these modalities. Our two-stage training strategy, augmented by the adapter modules, enhances the model's performance, particularly on the training dataset, while enabling zero-shot dataset transfer for thermal images. Notably, our method surpasses state-of-the-art thermal networks in terms of generalization and performance. This work represents a step forward in harnessing pre-trained models for thermal-based depth estimation, contributing to broader cross-modal applications and advancements in computer vision.

REFERENCES

[1] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4009–4018.

[2] J. Kopf, X. Rong, and J.-B. Huang, "Robust consistent video depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1611–1621.

[3] D. Wofk, F. Ma, T.-J. Yang, S. Karaman, and V. Sze, "Fastdepth: Fast monocular depth estimation on embedded systems," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 6101–6108.

[4] M. Hu, S. Wang, B. Li, S. Ning, L. Fan, and X. Gong, "Penet: Towards precise and efficient image guided depth completion," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 656–13 662.

[5] 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," *arXiv preprint arXiv:2304.07193*, 2023.

[6] U. Shin, J. Park, and I. S. Kweon, "Deep depth estimation from thermal image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1043–1053.

[7] K. Namil *et al.*, "Multispectral transfer network: Unsupervised depth estimation for all-day vision," in *AAAI Conference on Artificial Intelligence,(Apr. 1, 2018)*, pp. 6983–6991.

[8] 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 Robotics and Automation Letters*, vol. 7, no. 2, pp. 1103–1110, 2021.

[9] Y. Lu and G. 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*, 2021, pp. 3833–3843.

[10] 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5798–5807.

[11] S. Yoon and J. Cho, "Deep multimodal detection in reduced visibility using thermal depth estimation for autonomous driving," *Sensors*, vol. 22, no. 14, p. 5084, 2022.

[12] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[13] H. Bao, L. Dong, and F. Wei, "BEiT: BERT pre-training of image transformers," 2021.

[14] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, "ibot: Image bert pre-training with online tokenizer," *International Conference on Learning Representations (ICLR)*, 2022.

[15] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[16] A. J. Lee, Y. Cho, S. Yoon, Y. Shin, and A. Kim, "Vivid: Vision for visibility dataset," in *IEEE Int. Conf. Robotics and Automation (ICRA) Workshop: Dataset Generation and Benchmarking of SLAM Algorithms for Robotics and VR/AR*, 2019.

[17] A. J. Lee, Y. Cho, Y.-s. Shin, A. Kim, and H. Myung, "Vivid++: Vision for visibility dataset," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6282–6289, 2022.

[18] 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 Robotics and Automation Letters*, vol. 7, no. 3, pp. 7771–7778, 2022.

[19] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1851–1858.

[20] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki, "Sfm-net: Learning of structure and motion from video," *arXiv preprint arXiv:1704.07804*, 2017.

[21] J. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M.-M. Cheng, and I. Reid, "Unsupervised scale-consistent depth and ego-motion learning from monocular video," *Advances in neural information processing systems*, vol. 32, 2019.

[22] J.-W. Bian, H. Zhan, N. Wang, Z. Li, L. Zhang, C. Shen, M.-M. Cheng, and I. Reid, "Unsupervised scale-consistent depth learning from video," *International Journal of Computer Vision*, vol. 129, no. 9, pp. 2548–2564, 2021.

[23] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1983–1992.

[24] R. Wang, S. M. Pizer, and J.-M. Frahm, "Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5555–5564.

[25] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3828–3838.

[26] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799.

[27] J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, and I. Gurevych, "Adapterfusion: Non-destructive task composition for transfer learning," in *16th Conference of the European Chapter of the Associationfor Computational Linguistics, EACL 2021*. Association for Computational Linguistics (ACL), 2021, pp. 487–503.

[28] A. Chavan, Z. Liu, D. Gupta, E. Xing, and Z. Shen, "One-for-all: Generalized lora for parameter-efficient fine-tuning," 2023.

[29] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, and G. Neubig, "Towards a unified view of parameter-efficient transfer learning," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=0RDcd5Axok

[30] Y. Zhang, K. Zhou, and Z. Liu, "Neural prompt search," 2022.

[31] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," *Advances in neural information processing systems*, vol. 30, 2017.

[32] S. Jie and Z.-H. Deng, "Convolutional bypasses are better vision transformer adapters," *arXiv preprint arXiv:2207.07039*, 2022.

[33] Y.-C. Liu, C.-Y. Ma, J. Tian, Z. He, and Z. Kira, "Polyhistor: Parameter-efficient multi-task adaptation for dense vision tasks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 889–36 901, 2022.

[34] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, "Vision transformer adapter for dense predictions," in *The Eleventh International Conference on Learning Representations*, 2022.

[35] D. Yin, Y. Yang, Z. Wang, H. Yu, K. Wei, and X. Sun, "1% vs 100%: Parameter-efficient low rank adapter for dense predictions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 116–20 126.

[36] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 179–12 188.

[37] R. He, L. Liu, H. Ye, Q. Tan, B. Ding, L. Cheng, J. Low, L. Bing, and L. Si, "On the effectiveness of adapter-based tuning for pretrained language model adaptation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 2208–2222.

[38] G. Chen, F. Liu, Z. Meng, and S. Liang, "Revisiting parameter-efficient tuning: Are we really there yet?" in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 2612–2626.

[39] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[40] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[41] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[43] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[44] 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 transactions on pattern analysis and machine intelligence*, vol. 44, no. 3, pp. 1623–1637, 2020.

[45] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE transactions on robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.