photonics

MDPI

Article

Thermal Image-to-LiDAR Depth Transformation via Pretrained Visual Model and Two-Stage Depth Refinement

HeeJeong Yoo [1] and Hoon Yoo [2,*]

[1] Department of Computer Science, Graduate School, Sangmyung University, Seoul 03016, Republic of Korea; yhjj2000@naver.com
[2] Department of Computer Science, Sangmyung University, Seoul 03016, Republic of Korea
* Correspondence: hunie@smu.ac.kr

Abstract

LiDAR sensors provide reliable physical distance measurements using laser signals, enabling accurate acquisition of 3D information for various optical systems. However, they are costly, require significant weight and space, and their reliability and accuracy degrade under adverse environmental and weather conditions. In contrast, thermal cameras operating in the infrared spectrum can capture stable visual information even in challenging scenarios such as nighttime, low-light, and rain. However, they cannot directly provide the physical 3D depth information that LiDAR offers. To design efficient optical systems, there is a growing need for techniques that transform thermal image data into LiDAR-like depth information. While deep learning models can theoretically learn direct mappings between thermal and LiDAR modalities, the scarcity of acquiring paired thermal–LiDAR datasets and the difficulty of acquiring them make this task challenging. In this paper, we propose a thermal image-to-LiDAR depth transformation framework. Our method leverages large-scale pretrained visual models for depth estimation to generate initial depth predictions from thermal inputs. Since pretrained RGB-based models face a modality gap when applied to thermal data, we introduce a two-stage depth refinement. Stage 1 corrects global scale inconsistencies, and Stage 2 refines local structural details. Experiments on the MS2 dataset demonstrate that the proposed framework consistently improves the initial DepthPro outputs across day, night, and rainy conditions. Both quantitative metrics and qualitative comparisons show that RGB-pretrained depth predictions can provide useful structural cues for thermal depth estimation when their global scale and local structural errors are explicitly refined.

Keywords: depth refinement; monocular depth estimation; thermal depth estimation; thermal imaging

Check for updates

Received: 28 May 2026
Revised: 11 July 2026
Accepted: 14 July 2026
Published: 21 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Thermal images are relatively less sensitive to illumination changes and can provide stable visual information even in conditions where visible-light-based perception systems are vulnerable, such as night-time, low-light, and rainy environments. Owing to these properties, thermal sensors have been widely utilized in various applications, including autonomous driving, robotics, and surveillance, where reliable depth estimation is particularly important for 3D perception tasks such as scene understanding, obstacle avoidance, and distance estimation. Recently, with the introduction of datasets and learning-based

methods for thermal depth estimation, the feasibility of estimating depth using only thermal images has gradually been demonstrated [1]. Nevertheless, thermal depth estimation remains more challenging than its RGB-based counterpart [1].

These challenges mainly arise from the inherent characteristics of the thermal domain. Since thermal images differ fundamentally from RGB images in both imaging principles and signal distributions, representations learned from visible-light images are not easily transferable to thermal data. In addition, thermal images provide limited texture and appearance cues, which restrict the reliable recovery of structural information required for depth estimation. To alleviate these limitations, thermal-specific self-supervised learning methods have been proposed [2], and multimodal approaches leveraging multi-spectral information as auxiliary input have also been reported [3]. However, such methods tend to rely on additional modalities, sensor alignment, or thermal-specific training procedures.

Meanwhile, in the field of RGB-based monocular depth estimation (MDE), foundation models trained on large-scale data have advanced rapidly in recent years. Recent RGB-based monocular depth estimation models, including DepthAnythingV2, have demonstrated strong generalization ability and fine-grained depth reconstruction performance [4]. Among these models, DepthPro has been reported as a representative MDE model that simultaneously achieves high resolution, sharp boundaries, and fast inference [5]. These advances suggest that RGB-based depth foundation models can provide useful initial predictions even for thermal inputs.

However, directly applying RGB-based monocular depth estimation (MDE) models to thermal images presents clear limitations. While the initial predictions preserve relative depth relationships to some extent, they still suffer from global scale inconsistency and local structural errors. Moreover, such models cannot accurately provide physical LiDAR depth information when applied to thermal imagery. Although learning-based models can be trained using paired thermal–LiDAR data, acquiring such paired data remains challenging, and the modality gap between thermal imagery and depth supervision still makes accurate thermal depth estimation difficult.

To address these challenges, this paper focuses on the analysis and refinement of RGB-pretrained DepthPro outputs for thermal inputs. Rather than aiming to replace fully supervised thermal-specific MDE models, this study investigates how the systematic errors of an RGB-pretrained depth model can be interpreted and reduced in the thermal domain. As illustrated in Figure 1, we propose a two-stage refinement framework built on top of the initial DepthPro predictions. In Stage 1, global scale correction is performed on the initial DepthPro predictions, followed by Stage 2, which refines the remaining local structural errors. This coarse-to-fine correction strategy enables thermal-domain-adaptive depth refinement while preserving the relative depth structure contained in the initial predictions. Experimental results show that the proposed framework consistently improves the initial DepthPro predictions across day, night, and rainy conditions.

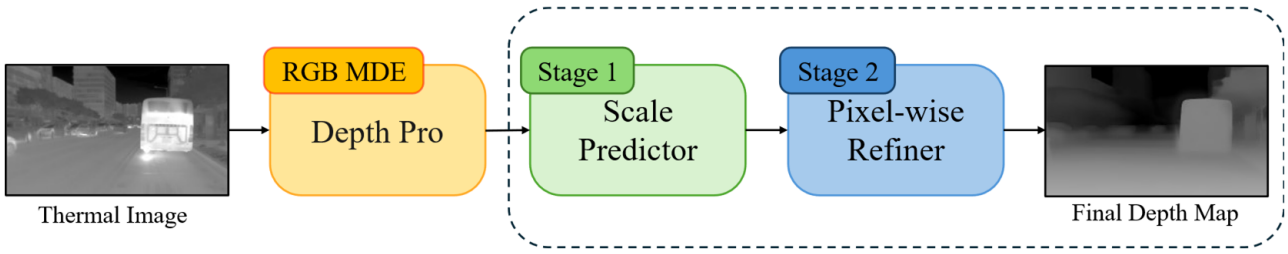


Figure 1. Overview of the proposed 2-stage thermal depth refinement framework. The dashed box indicates the main components in the proposed framework.

The main contributions of this paper are as follows:

1. We analyze the error characteristics of the initial predictions produced by directly applying DepthPro to the thermal domain and interpret them from the perspectives of global scale inconsistency and local structural errors.
2. We propose a two-stage refinement framework for RGB-pretrained DepthPro outputs, combining percentile-clipped thermal input normalization, global scale prediction, and local structural refinement.
3. We experimentally demonstrate that the proposed framework consistently improves the initial DepthPro predictions across day, night, and rainy conditions.

2. Related Work

2.1. Thermal Depth Estimation

Research on thermal depth estimation has generally evolved along two main directions. The first direction estimates depth using only thermal images. For example, the MS2 benchmark has been used to investigate depth estimation from monocular and stereo thermal images [6], providing benchmark results for representative thermal depth estimation models. In addition, a self-supervised method incorporating thermal image mapping was proposed to alleviate the difficulty of self-supervision caused by the low contrast and ambiguous boundaries of thermal images [2]. These studies demonstrate that meaningful depth estimation is possible using thermal images alone, even without additional auxiliary modalities.

The second direction leverages additional modalities such as RGB images. In this category, prior work trained a monocular thermal depth estimation network using disparity information obtained from RGB cameras together with RGB-to-thermal image translation results [7]. Other studies have proposed methods that reduce the feature-level domain gap between RGB and thermal images through unpaired RGB–thermal data and adversarial multi-spectral adaptation [8]. Although these approaches can compensate for the limitations of thermal-only input, they also increase dependence on additional sensors and modalities.

2.2. RGB-Based Monocular Depth Estimation Models

Recent studies on RGB-based monocular depth estimation have significantly improved generalization performance beyond specific scenes or sensors through large-scale training. Early work mainly focused on metric depth estimation tailored to individual datasets. Subsequently, MiDaS suggested a different direction by jointly training on multiple datasets with diverse characteristics, thereby achieving strong zero-shot cross-dataset generalization performance [8]. Building on this trend, later studies such as Metric3D, UniDepth, and DepthAnything further advanced toward stronger generalization performance or zero-shot metric depth estimation capability [9–11].

Within this line of research, DepthPro was proposed as a foundation model for zero-shot metric monocular depth estimation [5]. DepthPro can estimate a metric depth map with absolute scale even when metadata such as camera intrinsics are not provided, and it was designed to generate high-resolution depth maps with short inference times while preserving high-frequency structural details and sharp boundaries [5]. To this end, DepthPro employs a multi-scale ViT-based architecture, a training strategy combining real and synthetic data, and a focal length estimation module [5]. However, when such large-scale pretrained models are directly applied to domains with different imaging characteristics, such as thermal images, performance degradation may occur.

2.3. Domain Gap and Adaptation

Monocular depth estimation is sensitive to domain discrepancies between training and inference data, and distribution shifts such as synthetic-to-real transfer or day-to-night environmental changes can be major causes of performance degradation. To mitigate this problem, prior studies have proposed synthetic-to-real adaptation using image style transfer [12], geometry-aware symmetric domain adaptation [13], adversarial feature adaptation for nighttime images [14], and domain-separation-based self-supervised depth estimation for all-day environments [15]. These studies mainly focus on adapting the model to a new domain during training or reducing representational discrepancies across domains through feature alignment.

In contrast, this study focuses on refining the initial predictions obtained by applying an RGB-based pretrained monocular depth estimation model to thermal images without performing explicit domain adaptation. To this end, we design a two-stage refinement framework that progressively corrects the errors contained in the initial predictions.

3. Proposed Method

The goal of this study is to correct the errors that arise when DepthPro is directly applied to thermal inputs. To this end, we propose a 2-stage refinement framework. Stage 1 performs scene-level global scale correction on the initial predictions, and Stage 2 further refines the remaining local structural errors in a pixel-wise manner.

We adopt the model DepthPro as the initial depth generator because its output characteristics match the requirements of the proposed framework. Our framework requires an initial predictor to satisfy two conditions. First, since Stage 1 applies a single global multiplicative correction, the initial prediction should provide continuous depth values in metric form. Second, since Stage 2 corrects local structural errors, the initial prediction should preserve object boundaries and thin-structure cues to a reasonable degree. DepthPro performs zero-shot metric depth prediction and, through its multi-scale ViT-based architecture, produces high-resolution outputs with sharp boundary preservation, satisfying both conditions.

Figure 2 presents the overall architecture of the proposed framework, and the details of each stage are described below.

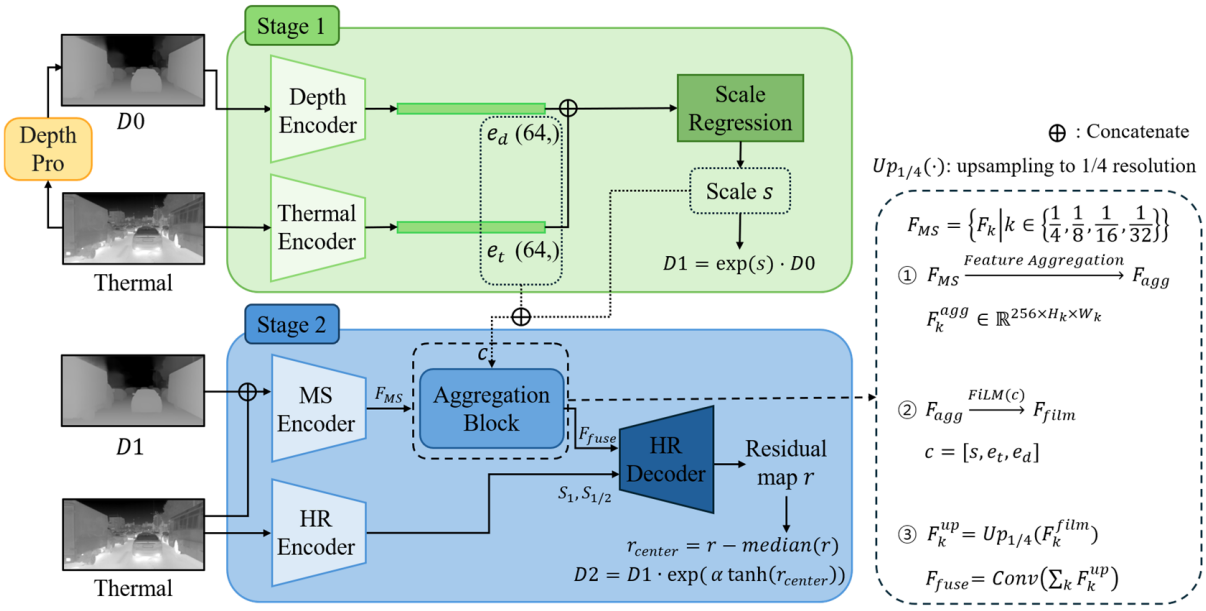


Figure 2. Overall architecture of the proposed 2-stage thermal depth refinement network.

3.1. Stage 1: Global Scale Correction

The objective of Stage 1 is to correct the scene-level global scale inconsistency in the initial prediction $D0$ produced by DepthPro. To achieve this, Stage 1 estimates a global log-scale factor s from the thermal image and the initial prediction. The predicted log-scale factor is applied to the initial prediction as follows to generate the corrected depth map $D1$:

$$D1 = \exp(s) \cdot D0 \tag{1}$$

This global correction allows the subsequent local refinement in Stage 2 to be performed on a more stable input. The multiplicative log-scale formulation was adopted because it adjusts the overall depth scale while preserving the relative depth structure of the initial prediction. In addition, applying the exponential function to the predicted log-scale factor guarantees a positive scale factor, thereby maintaining physically valid positive depth values.

First, the input thermal image is normalized to the range of [0, 1] using percentile normalization based on the 1st and 99th percentiles of the pixel-value distribution in each frame and is then fed into the thermal encoder. The encoder consists of four convolutional blocks followed by global average pooling and produces a 64-dimensional feature vector e_t. The initial prediction $D0$ is first clipped at 80 m according to the evaluation range of the MS2 benchmark and then converted to the log domain. It is subsequently fed into the Depth Encoder, which has the same architecture as the Thermal Encoder, to produce another 64-dimensional feature vector e_d. The two encoders are designed as lightweight convolutional encoders because Stage 1 aims to estimate only the scene-level scale factor rather than dense pixel-wise corrections. The 64-dimensional embeddings provide compact image-level and depth-level representations while keeping the Scale Regression module simple. Local structural correction is therefore left to Stage 2, which performs pixel-wise refinement using multi-scale and high-resolution features.

Finally, the two embedding vectors e_t and e_d are concatenated and passed to the Scale Regression module, implemented as an MLP with two hidden layers, to predict the global scale s. In this way, Stage 1 estimates the scene-level global scale inconsistency in the initial prediction and generates the corrected prediction $D1$ for the subsequent refinement stage.

3.2. Stage 2: Pixel-Wise Refinement

The objective of Stage 2 is to refine the remaining local structural errors using the globally corrected prediction $D1$ obtained from Stage 1. The inputs to Stage 2 consist of the thermal image, the corrected prediction $D1$, the embedding vectors e_t and e_d extracted in Stage 1, and the global log-scale factor s. The thermal image is replicated into three channels, and $D1$ is transformed into a normalized log-depth map. These are concatenated to form a four-channel input composed of a three-channel thermal image and a one-channel log-depth map.

This four-channel input is fed into the MS encoder, whose backbone is ConvNeXtV2-Tiny [16]. ConvNeXtV2-Tiny is adopted as a compact hierarchical backbone that provides multi-scale features for contextual reasoning without substantially increasing the complexity of the Stage 2 refiner. The encoder extracts a set of multi-resolution features F_{MS} at 1/4, 1/8, 1/16, and 1/32 resolution from this four-channel input. In parallel, the thermal image is also processed by a separate HR encoder, which produces skip features S_1 and $S_{1/2}$ at the original and half resolutions, respectively, to preserve fine structural information. This dual-encoder design is motivated by the characteristics of thermal images, which often provide weak texture cues but still contain useful boundary information, requiring both multi-scale contextual reasoning and high-resolution structural preservation for effective refinement.

The multi-resolution features F_{MS} are then processed by the aggregation block, which consists of three steps shown in Figure 2: ① feature aggregation, ② feature-wise linear modulation (FiLM), and ③ multi-scale fusion. First, in step ①, features in F_{MS} are aggregated in a top-down manner to construct a 256-channel pyramid feature F_{agg}, where each scale component satisfies $F_k^{agg} \in \mathbb{R}^{256\times H_k\times W_k}$. The channels are unified to 256 across all scales so that the features can be integrated by element-wise summation in the fusion step. Next, in step ②, the condition vector $c = [s,\ e_t,\ e_d]$ is formed by concatenating the global log-scale factor and the Stage 1 embedding vectors and is injected into the aggregated features through FiLM to produce the conditioned features F_{film}. This conditioning allows the global scale information estimated in Stage 1 to guide the local refinement process in Stage 2. Finally, in step ③, each conditioned feature is upsampled to 1/4 resolution as $F_k^{up} = \mathrm{Up}_{1/4}\left(F_k^{film}\right)$, where $\mathrm{Up}_{1/4}(\cdot)$ denotes resizing to the 1/4-resolution feature size. The upsampled features are integrated by element-wise summation followed by a convolutional block, producing the fused feature $F_{fuse} = \mathrm{Conv}\left(\sum_k F_k^{up}\right)$ with 256 channels at 1/4 resolution.

The HR decoder uses the fused feature F_{fuse} obtained above together with the skip features S_1 and $S_{1/2}$, respectively, through skip connections. In this process, F_{fuse} is progressively upsampled and combined with the corresponding skip features, and a residual map r is finally predicted at the original resolution. Through this design, Stage 2 preserves the global consistency established in Stage 1 while more precisely correcting local structural errors using boundary and fine-grained cues contained in the thermal image.

The predicted residual map is incorporated not through direct additive correction but through multiplicative residual correction. First, for each input image, the median value of the predicted residual map is subtracted to obtain a zero-centered residual map r_{center}. The prediction $D2$ is then expressed as follows:

$$r_{center} = r - median(r) \tag{2}$$

$$D2 = D1 \cdot \exp(\alpha \tanh(r_{center})) \tag{3}$$

where α is a constant that controls the correction magnitude. In our experiments, α is set to 0.5 so that the multiplicative correction factor is bounded within $[\exp(-\alpha), \exp(\alpha)] \approx [0.61, 1.65]$, preventing Stage 2 from excessively altering the global scale established in Stage 1. The $\tanh(\cdot)$ function suppresses excessively large correction values, while median subtraction reduces the global bias of the correction, thereby encouraging Stage 2 to focus on local refinement.

Although Equations (1) and (3) share a multiplicative correction form, they operate at different levels. Equation (1) applies a single image-level scale factor to correct scene-level global scale inconsistency, whereas Equation (3) applies bounded pixel-wise residual correction factors to refine the remaining local structural errors.

3.3. Loss Function

For the supervision of Stage 1, a ground-truth-based global log-scale target s^* was generated for each scene from the LiDAR ground-truth depth G and the initial prediction $D0$. First, the set of valid pixels Ω in G is defined as follows:

$$\Omega = \left\{ i \middle| G_i > 10^{-3},\ G_i < 80 \right\} \tag{4}$$

Here, Ω denotes the set of valid pixels used for depth supervision, and i denotes an individual pixel index in Ω. Then, the median of the log-depth difference between the

ground-truth depth and the initial prediction over the valid pixel region is used as the scene-level global log-scale target:

$$s^* = median_{i\in\Omega}(\log G_i - \log \max(D0_i, 10^{-3})) \quad (5)$$

Here, $\max(D0_i, 10^{-3})$ is introduced to ensure numerical stability in the logarithmic operation by imposing a minimum depth value. This target represents the scene-level global scale difference in the log domain, and the use of the median operator reduces the influence of outliers.

Stage 1 is trained to minimize the difference between the predicted global log-scale factor s and the ground-truth-based global log-scale target s^*. To this end, the following Smooth $L1$ loss is used:

$$L_{stage1} = SmoothL1(s, s^*) \quad (6)$$

During the training of Stage 2, a Smooth $L1$ loss is applied to the log-depth difference so that the final predicted depth $D2$ matches the ground truth G:

$$L_{depth} = \frac{1}{|\Omega|}\sum_{i\in\Omega} SmoothL1(\log D2_i, \log G_i) \quad (7)$$

In addition, to prevent unnecessary oscillation in the predicted correction map, edge-aware total variation regularization and edge-aware Laplacian regularization are applied to the correction map used in Stage 2, defined as $m = \alpha \tanh(r_{center})$. Each term is defined as follows:

$$L_{tv} = \frac{1}{|\Omega|}\sum_{i\in\Omega} \omega_i \left(|\partial_x m_i| + |\partial_y m_i|\right) \quad (8)$$

$$L_{lap} = \frac{1}{|\Omega|}\sum_{i\in\Omega} \omega_i |\Delta m_i| \quad (9)$$

Here, ∂_x and ∂_y denote first-order spatial derivatives along the horizontal and vertical directions in the image plane, respectively, and Δ denotes the discrete Laplacian operator applied to the correction map. In addition, ω_i denotes an edge-aware weight determined from the gradient magnitude of the thermal image. This design suppresses excessive smoothing around strong thermal boundaries while discouraging unnecessary oscillation of the correction map in relatively textureless regions.

Finally, the loss for Stage 2 is defined as follows:

$$L_{stage2} = L_{depth} + \lambda_{tv} L_{tv} + \lambda_{lap} L_{lap} \quad (10)$$

In our experiments, λ_{tv} and λ_{lap} were set to 0.005 and 0.1, respectively. These parameters were used as fixed auxiliary regularization weights to suppress unnecessary spatial fluctuations and local oscillations in the correction map, while keeping the log-depth reconstruction loss as the dominant training objective. In addition, the correction magnitude parameter α in Equation (3) was fixed at 0.5 to limit the pixel-wise correction magnitude. The same parameter values were used for all experiments.

4. Experiments and Results

4.1. Experimental Settings

4.1.1. MS2 Dataset

The MS2 dataset used in this study is a large-scale multi-sensor driving dataset that provides stereo RGB, stereo NIR, and stereo thermal images, together with LiDAR-based depth maps projected onto the image coordinate system of each modality. Following the

official benchmark protocol, we used the predefined train/validation/test split, consisting of 26K images for training and 4K images for validation. The test split is further divided into three subsets: 2.3K images for the day subset, 2.3K images for the night subset, and 2.5K images for the rain subset. For monocular depth estimation, the left camera image from each stereo pair was used. The thermal images and the projected depth maps have a resolution of 640 × 256.

4.1.2. Implementation Details

Stage 1 was trained for 20 epochs using the AdamW optimizer, with a batch size of 32, a learning rate of 2×10^{-4}, and a weight decay of 10^{-4}. The training was conducted on an NVIDIA RTX 2070 Super GPU, and the total training time was approximately 20 min.

Stage 2 was trained for 30 epochs using the AdamW optimizer, with a batch size of 16, an initial learning rate of 10^{-4}, and a weight decay of 10^{-4}. The learning rate was scheduled using a cosine annealing scheduler with a minimum learning rate of 10^{-6}. In addition, automatic mixed precision (AMP) training was adopted to improve computational efficiency and reduce GPU memory usage. Stage 2 training was performed on a single NVIDIA RTX 4090 GPU, and the total training time was approximately 22.5 h.

4.1.3. Evaluation Metrics

To evaluate depth estimation performance, we used five error metrics, namely mean absolute error (MAE), absolute relative error (AbsRel), squared relative error (SqRel), root mean square error (RMSE), and root mean square logarithmic error (RMSElog), together with three threshold-based accuracy metrics, δ_1, δ_2 and δ_3. For all error metrics, lower values indicate better performance, whereas for the accuracy metrics, higher values indicate better performance. The threshold accuracy δ_k is defined as follows:

$$\delta_k = \frac{1}{N}\left|\left\{ i \,\middle|\, max\left\{\frac{\hat{d}_i}{d_i}, \frac{d_i}{\hat{d}_i}\right\} < 1.25^k \right\}\right|, \quad k \in \{1,\ 2,\ 3\} \tag{11}$$

where N denotes the number of valid ground-truth pixels, and d_i and $\hat{d}_i$ represent the ground-truth depth and the predicted depth at pixel i, respectively. In addition, δ_k indicates the proportion of pixels whose predicted depth lies within a factor of 1.25^k of the ground-truth depth, and the allowable error range becomes wider as k increases.

All experiments followed the MS2 benchmark protocol. Only pixels with ground-truth depth values satisfying $10^{-3} < d_i < 80$ m were regarded as valid. The predicted depth values $\hat{d}_i$ were clipped to the same range before evaluation.

4.2. Analysis of Thermal Input and Depth Pro Output Characteristics

4.2.1. Performance Comparison According to Input Normalization Method

To feed thermal images into DepthPro, two input normalization methods were compared. The first was the RAW method, in which the full range (0–16,383) of the 14-bit thermal images provided by the MS2 dataset was linearly normalized to the range of 0–1, then converted again to the 8-bit range of 0–255, replicated into three channels, and used as input. The second was the percentile-clipped normalization method, denoted as PCTL. In this method, only the 1st–99th percentile range of the intensity values in each image was used instead of the full range. The selected range was then normalized to 0–1, converted to the 8-bit range, replicated into three channels, and fed into DepthPro. Table 1 presents the evaluation results obtained when the test split of the dataset was used as input to DepthPro using these two methods.

Table 1. Comparison of two input normalization methods.

	MAE	AbsRel	SqRel	RMSE	RMSElog	δ_1	δ_2	δ_3
RAW	7.303	0.365	4.132	10.126	0.530	0.349	0.618	0.766
PCTL	4.833	0.251	2.382	7.293	0.307	0.573	0.873	0.953

The experimental results show that the PCTL method achieved overall better depth estimation performance than the RAW method. Since the RAW method applies the same fixed range to all images, it is directly affected by scene-dependent differences in thermal distributions and extreme values. In contrast, the PCTL method readjusts the effective range according to the distribution of each image and reduces the influence of outliers. As a result, the contrast of structurally informative regions within the image can be emphasized more stably, allowing DepthPro to utilize relative depth patterns more effectively. Therefore, the PCTL method was used as the input normalization strategy for DepthPro in all subsequent experiments.

4.2.2. Analysis of Relative Depth Preservation in Depth Pro Outputs

To examine whether the outputs of DepthPro preserve the relative depth structure of a scene even when the absolute depth values themselves are not accurate, a rank-based evaluation was conducted using Spearman's rank correlation coefficient, Kendall's tau, and PDA, as shown in Table 2. For each frame, up to 20,000 valid pixels were sampled, and 30,000 pixel pairs were then randomly drawn from these samples to compute Kendall's tau and PDA. For PDA, the margin was set to 0.1, so that only pixel pairs with a depth difference greater than 10% in the ground truth were regarded as valid comparison pairs.

Table 2. Rank-based evaluation of relative depth preservation in DepthPro outputs.

	Spearman	Kendall Tau	PDA
PCTL	0.936	0.827	0.898

Spearman's rank correlation coefficient measures the overall agreement in rank ordering between the ground-truth and predicted depth values, where a value closer to 1 indicates that the two depth maps share a similar global ranking structure. Kendall's tau evaluates how consistently the relative front–back ordering between arbitrary pixel pairs is preserved, thereby more directly reflecting pairwise rank consistency. PDA denotes the proportion of pixel pairs with sufficiently large depth differences in the ground truth for which the prediction correctly identifies the relative ordering, and a higher value indicates more accurate preservation of structural depth relationships.

As shown in Table 2, the average results were 0.936 for Spearman's rank correlation, 0.827 for Kendall's tau, and 0.898 for PDA. These results indicate that although the outputs of Depth Pro differ from the ground truth in terms of absolute scale, they preserve a substantial portion of the relative depth ordering and structural relationships within a scene. In particular, the high values of both Spearman's rank correlation and Kendall's tau suggest that not only the global ranking structure but also the pairwise relative ordering is preserved with reasonable stability. In addition, the high PDA value indicates that the model maintains the relative ordering well in regions with meaningful depth differences. These findings support the feasibility of the subsequent Stage 1 global scale correction and Stage 2 pixel-wise refinement.

4.2.3. Analysis of the Effect of Image-Wise Global Scale Alignment

To examine whether global scale inconsistency is one of the major causes of performance degradation in the outputs of DepthPro, a preliminary analysis was conducted by computing the image-wise optimal scale between the ground-truth depth and the initial prediction and then applying it to the initial prediction. The scale was computed in the same manner as the ground-truth-based global log-scale target s^*, which is used only for Stage 1 supervision during training, as defined in Equation (5). As a result, most test metrics were substantially improved even with simple global alignment alone. This suggests that global scale inconsistency is one of the major sources of error in the initial outputs of DepthPro.

However, even after such alignment, the evaluation metrics were still not sufficiently high, indicating that global correction alone is not enough to fully resolve the remaining errors. Table 3 quantitatively compares the original DepthPro output with the result of this GT-based image-wise scale alignment. In Table 3, "Scale" denotes the result obtained by applying this GT-based image-wise optimal scale to the initial DepthPro prediction, and it is used only for preliminary analysis rather than as the inference output of Stage 1. Based on these observations, we designed a 2-stage refinement framework that combines Stage 1 for global scale correction and Stage 2 for refining the remaining local structural errors.

Table 3. Effect of image-wise global scale alignment.

	MAE	**AbsRel**	**SqRel**	**RMSE**	**RMSElog**	δ_1	δ_2	δ_3
PCTL	4.833	0.251	2.382	7.293	0.307	0.573	0.873	0.953
Scale	3.190	0.143	1.314	5.776	0.207	0.808	0.945	0.980

4.3. Ablation Study of the Proposed Framework

This section analyzes the role of the proposed two-stage refinement strategy in the proposed framework. Specifically, we compare four settings: the original DepthPro output with percentile-based post-processing, the result obtained by applying only Stage 1 global scale alignment, the result obtained by applying only Stage 2 pixel-wise refinement, and the full framework using both Stage 1 and Stage 2. This ablation is designed to evaluate the individual effects of global scale alignment and pixel-wise residual refinement and to examine whether the explicit Stage 1 correction is necessary before applying Stage 2. The results are reported for the day, night, and rain splits, and "Avg" denotes the average over the three splits. The best value for each metric is highlighted in bold.

Table 4 presents the quantitative results of the ablation study. Compared with the original DepthPro output, applying Stage 1 alone substantially improves depth accuracy. In terms of average performance, AbsRel is reduced from 0.251 to 0.157, RMSE decreases from 7.293 to 6.052, and $\boldsymbol{\delta_1}$ improves from 0.573 to 0.774. These results indicate that the initial DepthPro predictions contain a significant global metric scale mismatch when directly applied to thermal images, and that Stage 1 effectively alleviates this error through image-wise global scale alignment.

The Stage 2-only setting also improves the original DepthPro output. Its average AbsRel decreases from 0.251 to 0.167, RMSE decreases from 7.293 to 5.178, and δ_1 increases from 0.573 to 0.765. This demonstrates that pixel-wise residual refinement is effective for correcting local depth errors. However, compared with Stage 1 alone, the Stage 2-only setting shows slightly worse AbsRel and $\boldsymbol{\delta_1}$, although it achieves a lower RMSE. This suggests that Stage 2 can reduce large local errors, but it does not fully replace the explicit global scale correction performed by Stage 1.

Table 4. Ablation study of the proposed two-stage refinement framework.

	Split	MAE	AbsRel	SqRel	RMSE	RMSElog	δ_1	δ_2	δ_3
Depth Pro (PCTL)	Day	4.172	0.222	1.869	6.663	0.253	0.627	0.920	0.982
	Night	4.758	0.268	2.567	7.211	0.309	0.556	0.862	0.951
	Rain	5.518	0.263	2.690	7.955	0.356	0.538	0.838	0.927
	Avg	4.833	0.251	2.382	7.293	0.307	0.573	0.873	0.953
Ours (Stage 1)	Day	2.929	0.129	1.150	5.525	0.188	0.834	0.960	0.988
	Night	3.400	0.168	1.493	5.901	0.220	0.760	0.936	0.982
	Rain	4.114	0.173	1.643	6.680	0.254	0.732	0.911	0.964
	Avg	3.497	0.157	1.433	6.052	0.222	0.774	0.935	0.977
Ours (Stage 2)	Day	2.626	0.139	0.849	4.547	0.179	0.814	0.968	0.994
	Night	3.037	0.169	1.196	4.870	0.215	0.768	0.938	0.976
	Rain	3.972	0.190	1.662	6.047	0.266	0.715	0.911	0.960
	Avg	3.231	0.167	1.246	5.178	0.221	0.765	0.938	0.976
Ours (Stage 1 + 2)	Day	**2.085**	**0.099**	**0.547**	**3.756**	**0.139**	**0.891**	**0.982**	**0.997**
	Night	**2.310**	**0.119**	**0.662**	**3.864**	**0.156**	**0.860**	**0.974**	**0.995**
	Rain	**3.249**	**0.144**	**1.034**	**5.114**	**0.199**	**0.793**	**0.950**	**0.985**
	Avg	**2.566**	**0.121**	**0.755**	**4.268**	**0.166**	**0.847**	**0.968**	**0.992**

The full Stage 1 + 2 framework achieves the best overall performance among the proposed variants. In terms of average performance, AbsRel is further reduced to 0.121, RMSE decreases to 4.268, and δ_1 increases to 0.847. Compared with Stage 1 alone, adding Stage 2 reduces AbsRel from 0.157 to 0.121 and RMSE from 6.052 to 4.268. Compared with Stage 2 alone, applying Stage 1 before Stage 2 further improves AbsRel from 0.167 to 0.121, RMSE from 5.178 to 4.268, and δ_1 from 0.765 to 0.847. These results show that Stage 1 and Stage 2 play complementary roles: Stage 1 first corrects sample-dependent global scale mismatch, while Stage 2 further refines the remaining pixel-wise residual errors and local structural inconsistencies.

To further clarify the role of Stage 1, we analyze the distribution of the image-wise multiplicative scale factors predicted by Stage 1. The scale factors reported in Table 5 are not log-scale values but the actual multiplicative factors applied to the initial DepthPro depth map in meters before Stage 2 refinement.

Table 5. Distribution of scale factors predicted by Stage 1.

Split	Mean	Median	P5	P95
Day	0.955	0.910	0.738	1.305
Night	1.036	0.918	0.721	1.440
Rain	1.141	0.932	0.757	1.579
Avg	1.047	0.918	0.733	1.425

As shown in Table 5, although the overall mean scale factor is 1.047, this value alone does not fully represent the image-wise variation in the required scale correction. The median scale factor over all test images is 0.918, and the P5–P95 range is 0.733–1.425. Similar trends are observed in each split, with P5–P95 ranges of 0.738–1.305, 0.721–1.440, and 0.757–1.579 for the day, night, and rain splits, respectively. These results indicate that the scale mismatch is not a single constant bias, but varies depending on the input image and imaging condition. Therefore, Stage 1 does not simply apply a nearly constant 1.047 correction; rather, it predicts an image-wise global scale factor before the local refinement stage.

Overall, the ablation results demonstrate that both stages are necessary for progressively adapting DepthPro predictions to thermal imagery. Stage 1 improves global metric consistency by correcting image-dependent scale mismatch, whereas Stage 2 further enhances the predictions by refining local depth structures. The full two-stage framework achieves the most stable and accurate results, supporting the effectiveness of explicitly decomposing the refinement process into global scale alignment and pixel-wise residual refinement.

4.4. Performance Comparison with Existing Methods

In this section, we compare the proposed method with four MDE models reported in the MS2 benchmark [6], namely DORN [17], AdaBins [18], BTS [19], and NeWCRFs [20], through both quantitative and qualitative evaluation. Table 6 presents the test metrics of the four models and the results obtained by applying both Stage 1 and Stage 2. For each split, the best value for each metric is highlighted in bold.

Table 6. Quantitative comparison with existing methods.

	Split	MAE	AbsRel	SqRel	RMSE	RMSElog	δ_1	δ_2	δ_3
DORN	Day	1.929	0.101	0.443	3.388	0.135	0.909	0.987	0.998
	Night	1.887	0.099	0.401	3.152	0.129	0.914	0.989	0.998
	Rain	2.688	0.127	0.757	4.377	0.167	0.842	0.971	0.994
	Avg	2.182	0.109	0.540	3.660	0.144	0.887	0.982	0.996
AdaBins	Day	1.638	0.079	0.327	3.003	0.111	0.937	0.992	**0.999**
	Night	1.577	0.079	0.273	2.679	0.105	0.947	0.995	0.999
	Rain	2.254	0.104	0.518	3.726	0.139	0.891	0.983	0.997
	Avg	1.834	0.088	0.377	3.153	0.119	0.924	0.990	0.998
BTS	Day	1.599	0.075	0.314	2.945	0.107	0.943	0.992	0.998
	Night	1.604	0.079	0.282	2.755	0.106	0.943	0.995	0.999
	Rain	2.254	0.101	0.531	3.740	0.137	0.896	0.983	0.997
	Avg	1.831	0.086	0.380	3.163	0.117	0.927	0.990	0.998
NeWCRFs	Day	**1.497**	**0.071**	**0.271**	**2.717**	**0.099**	**0.951**	**0.994**	**0.999**
	Night	**1.494**	**0.074**	**0.248**	**2.544**	**0.099**	**0.952**	**0.996**	**1.000**
	Rain	**2.111**	**0.095**	**0.464**	**3.503**	**0.127**	**0.909**	**0.988**	**0.998**
	Avg	**1.712**	**0.081**	**0.331**	**2.937**	**0.109**	**0.937**	**0.992**	**0.999**
Ours (Stage 1 + 2)	Day	2.085	0.099	0.547	3.756	0.139	0.891	0.982	0.997
	Night	2.310	0.119	0.662	3.864	0.156	0.860	0.974	0.995
	Rain	3.249	0.144	1.034	5.114	0.199	0.793	0.950	0.985
	Avg	2.566	0.121	0.755	4.268	0.166	0.847	0.968	0.992

According to the quantitative results, NeWCRFs achieves the best overall performance among the compared methods, and the proposed method does not outperform the thermal-specific baselines in terms of overall accuracy. For example, NeWCRFs achieves the best average AbsRel of 0.081, whereas the proposed method achieves an average AbsRel of 0.121. However, the baseline results were obtained from state-of-the-art models that were directly trained on thermal–LiDAR data pairs. In contrast, the results of the proposed method are obtained using a completely different approach: post-processing the outputs of an RGB-based pretrained model. The fact that our approach achieves comparable objective performance to direct training methods, even without retraining DepthPro, is highly encouraging and demonstrates the potential of the proposed technique.

Here, it is important to note certain considerations when evaluating the results. When computing the metrics in Table 6, the ground-truth LiDAR depth data are inherently sparse and contains missing regions. In these areas, metrics such as AbsRel cannot be

reliably calculated, and thus the results in Table 6 were obtained using the masking strategy adopted in prior state-of-the-art methods. Consequently, the reported results do not represent evaluations across all pixels of the thermal-derived depth maps, highlighting the need for complementary approaches. Since arbitrarily modifying the ground truth of standard datasets lies beyond the scope of this study, incorporating visual evaluation or qualitative comparison is essential for a more comprehensive assessment.

Figures 3 and 4 present qualitative comparisons between the proposed method and the baseline methods. In both figures, (a) shows the input thermal image, and (b) shows the ground truth. (c)–(f) correspond to the results of the compared methods, (g) shows the output of DepthPro, (h) shows the result after applying Stage 1, (i) shows the result after applying only Stage 2, and (j) shows the result after applying both Stage 1 and Stage 2. For clearer qualitative comparison, all depth maps in the experimental figures are visualized using a shared color-mapped log-depth scale over 1–80 m, so that the same color represents the same depth value across all figures.

In Figure 3, the red-box region contains a human object. The baseline methods roughly capture the presence of the object, but fail to preserve the human shape clearly. In contrast, the final output of the proposed framework in (j) produces a more distinct human structure. A similar trend is observed in the blue-box region containing a motorcycle, where the baseline methods yield blurred or mixed representations, whereas the proposed method provides a clearer structural representation. In the yellow-box region, the baseline methods exhibit abnormally near depth responses, while the proposed method mitigates such artifacts. Comparing (g) and (h), Stage 1 adjusts the overall depth scale of the scene by correcting the global scale mismatch in the initial DepthPro prediction. The Stage 2-only result in (i) indicates that pixel-wise refinement alone is insufficient when the initial prediction still contains global scale inconsistency, often leading to less stable local structures. In contrast, the full Stage 1 + 2 result in (j) combines global scale alignment with local residual refinement, producing a more consistent depth prediction.

Figure 4 more clearly illustrates the effect of applying Stage 2 after Stage 1. In the red-box region corresponding to a traffic sign, the baseline methods generally capture the circular upper part, but fail to preserve the thin pole structure accurately. In particular, (d) fails to represent the sign structure, while (c), (e), and (f) produce results where thin structures such as the pole are either thickened or exhibit discontinuities, failing to preserve their precise shape and continuity. The result after Stage 1 in (h) maintains the upper part of the sign, but the pole remains unclear and appears closer than in the ground truth. The Stage 2-only result in (i) shows an abnormal depth response in the background region containing trees, indicating that pixel-wise refinement alone can be unstable when global scale alignment is not performed beforehand. In contrast, the result in (j), after applying Stage 2, shows that the depth structure in this region is corrected toward the ground truth, and the pole is more clearly represented. Furthermore, in the yellow-box region, the baseline methods exhibit abnormal depth patterns in far regions, whereas the proposed method alleviates such errors. This observation suggests that applying Stage 2 after global scale correction contributes to further correcting the local structural errors that remain after Stage 1.

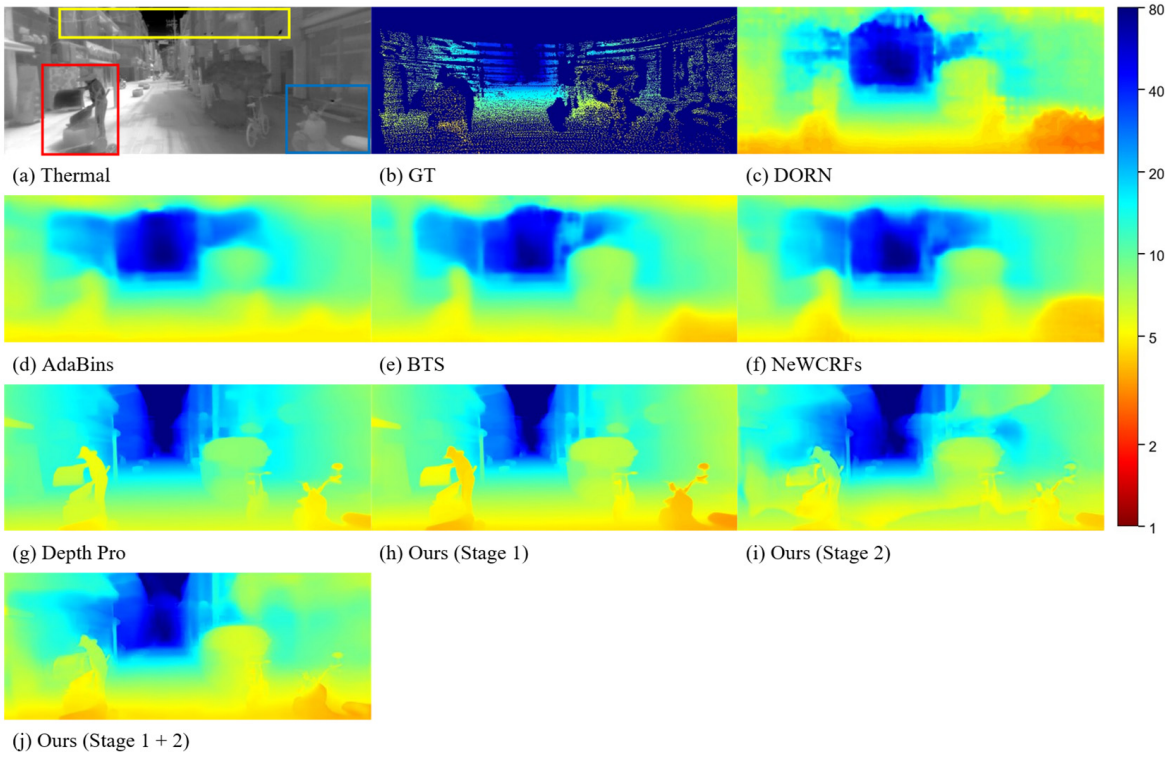


Figure 3. Qualitative comparison of existing methods and the proposed method (Example 1).

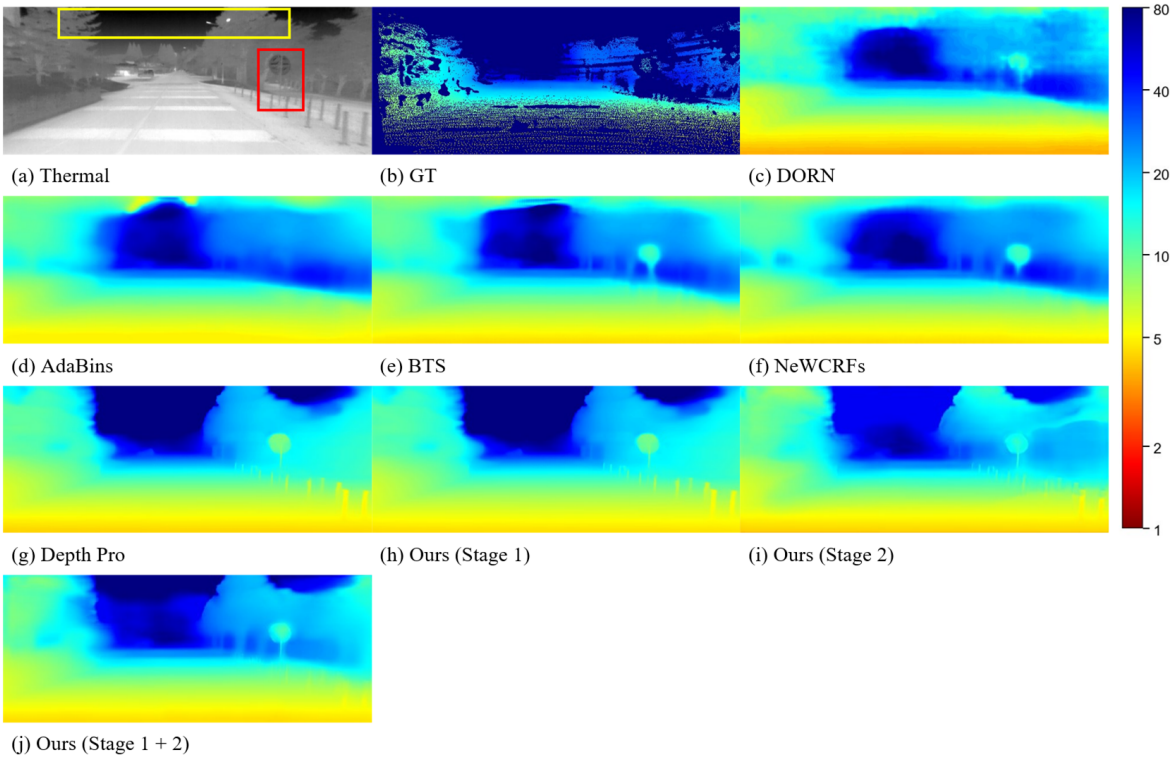


Figure 4. Qualitative comparison of existing methods and the proposed method (Example 2).

5. Conclusions

This study analyzed the limitations of directly applying an RGB-based monocular depth estimation model DepthPro to thermal inputs and proposed a 2-stage depth refinement framework to address these issues. Through empirical analysis, we found that the initial predictions of DepthPro do not provide accurate absolute depth values but preserve

the relative depth structure of the scene to a certain extent. We further identified that the major errors can be categorized into global scale inconsistency and local structural errors. Based on this observation, we designed a framework that combines global scale correction in Stage 1 with pixel-wise refinement in Stage 2. Experimental results on the MS2 dataset show that the proposed method consistently improves upon the original DepthPro outputs across all conditions. In terms of average performance, AbsRel was reduced from 0.251 to 0.157 after Stage 1 and further to 0.121 after Stage 2. RMSE was also reduced from 7.293 to 6.052 and further to 4.268, while the accuracy metric δ_1 improved from 0.573 to 0.774 and then to 0.847. Qualitative results also showed that the proposed method produces clearer object boundaries and more structurally consistent depth maps while alleviating abnormal depth responses in distant regions and boundary distortions observed in the initial predictions.

Nevertheless, the proposed method does not outperform existing thermal-specific monocular depth estimation models across all evaluation metrics, and its contribution should not be interpreted as achieving state-of-the-art thermal depth estimation performance. Rather, this study focuses on analyzing the behavior of an RGB-pretrained metric depth model in the thermal domain and demonstrating that its outputs can be improved through explicit global scale alignment and pixel-wise refinement. Therefore, the proposed framework should be understood as a refinement-based adaptation approach built on top of DepthPro rather than as a replacement for fully supervised thermal-specific MDE models. In addition, the experiments were conducted only on the MS2 dataset, and other pretrained depth models were not evaluated in this study. Future work will therefore evaluate the proposed framework on additional thermal depth datasets and compare it with other pretrained depth models such as DepthAnythingV2, Metric3D, and UniDepth.

Author Contributions: Conceptualization, H.Y. (HeeJeong Yoo) and H.Y. (Hoon Yoo); methodology, H.Y. (HeeJeong Yoo); investigation, H.Y. (HeeJeong Yoo); writing—original draft preparation, H.Y. (HeeJeong Yoo); writing—review and editing, H.Y. (Hoon Yoo); visualization, H.Y. (Hoon Yoo); supervision, H.Y. (Hoon Yoo); project administration, H.Y. (Hoon Yoo). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a 2026 research grant from Sangmyung University under grant number 2026-A000-0182.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: MS2 dataset is available upon completion of the dataset access survey; download links are provided via an automated email response. https://sites.google.com/view/multi-spectral-stereo-dataset/home (accessed on 15 March 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Shin, U.; Park, J.; Kweon, I.S. Deep depth estimation from thermal image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023), Vancouver, BC, Canada, 18–22 June 2023; pp. 1043–1053. [CrossRef]
2. Shin, U.; Lee, K.; Lee, B.-U.; Kweon, I.S. Maximizing self-supervision from thermal image for effective self-supervised learning of depth and ego-motion. *IEEE Robot. Autom. Lett.* **2022**, *7*, 7771–7778. [CrossRef]
3. Shin, U.; Park, K.; Lee, B.-U.; Lee, K.; Kweon, I.S. 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 (WACV 2023), Waikoloa, HI, USA, 2–7 January 2023; pp. 5787–5796. [CrossRef]
4. Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; Zhao, H. Depth Anything V2. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 21875–21911. [CrossRef]

5. Bochkovskiy, A.; Delaunoy, A.; Germain, H.; Santos, M.; Zhou, Y.; Richter, S.R.; Koltun, V. Depth Pro: Sharp monocular metric depth in less than a second. In Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025), Singapore, 24–28 April 2025.
6. Shin, U.; Park, J. Deep depth estimation from thermal image: Dataset, benchmark, and challenges. *arXiv* **2025**, arXiv:2503.22060. [CrossRef]
7. Lu, Y.; Lu, G. An alternative of LiDAR in nighttime: Unsupervised depth estimation based on single thermal image. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV 2021), Waikoloa, HI, USA, 3–8 January 2021; pp. 3832–3842. [CrossRef]
8. Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; Koltun, V. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 1623–1637. [CrossRef] [PubMed]
9. Yin, W.; Zhang, C.; Chen, H.; Cai, Z.; Yu, G.; Wang, K.; Chen, X.; Shen, C. Metric3D: Towards zero-shot metric 3D prediction from a single image. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2023), Paris, France, 1–6 October 2023; pp. 9009–9019. [CrossRef]
10. Piccinelli, L.; Yang, Y.-H.; Sakaridis, C.; Segu, M.; Li, S.; Van Gool, L.; Yu, F. UniDepth: Universal monocular metric depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024), Seattle, WA, USA, 17–21 June 2024; pp. 10106–10116. [CrossRef]
11. Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; Zhao, H. Depth Anything: Unleashing the power of large-scale unlabeled data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024), Seattle, WA, USA, 17–21 June 2024; pp. 10371–10381. [CrossRef]
12. Atapour-Abarghouei, A.; Breckon, T.P. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 2800–2810. [CrossRef]
13. Zhao, S.; Fu, H.; Gong, M.; Tao, D. Geometry-aware symmetric domain adaptation for monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019), Long Beach, CA, USA, 16–20 June 2019; pp. 9780–9790. [CrossRef]
14. Nguyen, M.H.V.; Tran, T.T.; Nguyen, T.H.; Nguyen, D.D. MonoVINI: Seeing in the dark with virtual-world supervision. In Proceedings of the RIVF International Conference on Computing and Communication Technologies (RIVF 2023), Hanoi, Vietnam, 7–9 December 2023; pp. 272–277. [CrossRef]
15. Liu, L.; Song, X.; Wang, M.; Liu, Y.; Zhang, L. Self-supervised monocular depth estimation for all day images using domain separation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021), Montreal, QC, Canada, 11–17 October 2021; pp. 12717–12726. [CrossRef]
16. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I.S.; Xie, S. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023), Vancouver, BC, Canada, 18–22 June 2023; pp. 16133–16142. [CrossRef]
17. Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; Tao, D. Deep ordinal regression network for monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 2002–2011. [CrossRef]
18. Bhat, S.F.; Alhashim, I.; Wonka, P. AdaBins: Depth estimation using adaptive bins. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021), Nashville, TN, USA, 20–25 June 2021; pp. 4008–4017. [CrossRef]
19. Lee, J.H.; Han, M.-K.; Ko, D.W.; Suh, I.H. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv* **2019**, arXiv:1907.10326. [CrossRef]
20. Yuan, W.; Gu, X.; Dai, Z.; Zhu, S.; Tan, P. Neural window fully-connected CRFs for monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022), New Orleans, LA, USA, 18–24 June 2022; pp. 3906–3915. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.