Empirical Study on Performance–Perception Discrepancy in RGB–Thermal Monocular Depth Estimation under Varying Illumination

Kang Min Ji
Dongguk University, 30 Pildong-ro 1-gil, Jung-gu, Seoul 04620, South Korea
*Corresponding author: kangminji411@gmail.com

Abstract

With the advancement of multimodal perception technologies, integrating visible (RGB) and thermal infrared (THR) information has become a key approach to enhancing the robustness of visual systems under complex illumination conditions. While existing studies primarily focus on improving quantitative accuracy through multimodal fusion, less attention has been paid to the perceptual differences and consistency between modalities. This study investigates the performance–perception discrepancy in multimodal depth estimation under varying illumination scenarios. Through comparative experiments between RGB and THR modalities, the analysis reveals that THR exhibits superior numerical performance (e.g., lower RMSE and AbsRel) in low-light and nighttime conditions, yet suffers from perceptual degradation such as over-smoothing and structural blurring. Moreover, by referencing findings in multimodal object detection, this phenomenon is shown to be task-general, arising from the distinct spatial frequency responses of different modalities. The presented results provide empirical evidence and theoretical insight for future research on multimodal feature fusion and perceptual consistency optimization.

Index Terms— Depth Estimation, Multimodal, Illumination Robustness, Quantitative Evaluation, Visual Consistency.

1 Introduction

Recent advances in computer vision have enabled machines to perceive and reconstruct 3D structures from visual data with remarkable accuracy, particularly through deep learning–based monocular depth estimation (MDE) methods [1, 25]. These models have achieved significant progress in autonomous driving, robotics, and scene understanding [3, 17]. However, RGB-based depth estimation systems remain sensitive to environmental variations—especially in challenging illumination conditions such as nighttime or adverse weather—where visible light becomes unreliable.

Thermal infrared (THR) imaging provides a promising complementary modality in such environments. By capturing infrared radiation emitted by objects above absolute zero, thermal cameras can perceive structures that are invisible to the RGB spectrum, offering illumination invariance and robustness to occlusion and haze [11, 5]. As illustrated in prior multimodal vision studies, THR sensors have been successfully used for tasks such as salient object detection, pedestrian recognition, and image segmentation, thanks to their strong response to heat-emitting objects even in complete darkness [11, 5]. Nevertheless, when applied to depth estimation, thermal images introduce new challenges: they often exhibit low resolution, limited texture, and reduced semantic richness, making fine-grained 3D reconstruction difficult.

To address these challenges, multimodal fusion between RGB and THR modalities has emerged as a viable strategy. Recent works have explored RGB–Thermal integration through two primary paradigms: (1) feature-level fusion networks that combine spatial and channel-wise cues from both modalities to enhance representation learning [20, 27], and (2) cross-modal distillation frameworks that transfer geometric knowledge from large-scale RGB foundation models to thermal networks using confidence-aware consistency objectives [15, 24, 2]. Despite these advances, the performance–perception inconsistency remains a largely overlooked issue in multimodal depth estimation: quantitative metrics such as RMSE or AbsRel may improve significantly through thermal guidance, while the resulting depth maps exhibit visual artifacts such as texture loss or oversmoothing.

This discrepancy highlights a fundamental property of multimodal depth perception—the asymmetric contribution of RGB and THR features. RGB imagery captures high-frequency textures and detailed semantics but degrades rapidly in dark scenes, whereas THR imagery maintains structural continuity at the cost of visual sharpness. Existing multimodal fusion models [14, 4] often overlook this imbalance by treating both modalities uniformly, resulting in fused representations that may optimize numerical performance but fail to achieve perceptual coherence.

In this study, we systematically analyze the performance–perception discrepancy in multimodal depth estimation under varying illumination conditions. Using a series of controlled experiments comparing RGB, THR, and fused modalities, we observe that thermal-based estimation yields superior quantitative metrics but perceptually degraded visual results. We further discuss the relevance of this phenomenon to other multimodal tasks, such as object detection and salient

object segmentation, where similar modality-dependent trade-offs have been observed [19, 26]. The findings of this study provide empirical evidence and analytical insights for developing future multimodal perception systems that balance accuracy with perceptual fidelity.

The main contributions of this paper are as follows:

- We conduct an empirical analysis of RGB–thermal (RGB–THR) depth estimation under varying illumination, revealing a clear performance–perception discrepancy.
- We identify that this discrepancy arises from modality-dependent feature characteristics, where THR improves quantitative accuracy but weakens visual fidelity.
- The findings provide experimental evidence and insight for subsequent research on multimodal feature fusion and perceptual consistency optimization.

2 Related Works

2.1 RGB and Thermal Imaging in Visual Tasks

Multimodal perception has emerged as a key paradigm in computer vision to improve robustness against illumination changes, occlusions, and environmental degradation. Among various modality combinations, visible-light (RGB) and thermal infrared (THR) imaging form a particularly complementary pair. RGB sensors capture reflected visible light, providing rich texture and color cues essential for semantic understanding and fine spatial delineation. In contrast, THR cameras sense long-wave infrared radiation emitted by objects, enabling reliable perception under adverse or low-illumination conditions [23].

The integration of RGB and THR data has been explored across numerous vision tasks, including pedestrian detection, salient object detection, semantic segmentation, and scene understanding [7, 8, 13]. For example, multispectral detectors trained on datasets such as KAIST and LLVIP have shown that thermal cues can significantly enhance nighttime pedestrian recognition [8, 7]. In salient object detection, cross-modality interaction modules and attention-guided fusion methods [13, 28] leverage complementary modality information to achieve robust target localization under dynamic lighting. Similarly, RGB–THR fusion in semantic segmentation improves feature stability at the object boundary level [21], confirming the benefit of multimodal integration in challenging environments.

The advantages of RGB–THR fusion extend beyond conventional image analysis. Recent studies have applied multispectral fusion to domains such as autonomous driving [22], UAV-based surveillance [16], and robotics [9], where the goal is to achieve all-day, all-weather perception. These applications emphasize that while RGB features provide geometric and semantic richness, THR inputs ensure visibility and structural consistency across varying conditions—highlighting the importance of effective cross-modal fusion for real-world deployment.

2.2 Multimodal Fusion Strategies and Emerging Challenges

With the success of deep learning, multimodal fusion has evolved from handcrafted feature concatenation to learned feature-level and attention-based strategies. Early fusion approaches such as pyramid-based blending or weighted averaging [18, 12] mainly focused on pixel-level enhancement without learning task-specific representations. Modern deep fusion frameworks employ dual-stream encoder–decoder architectures, where modality-specific features are extracted independently and later integrated via cross-attention or adaptive weighting [13, 10]. These designs enable networks to selectively exploit complementary signals, improving the robustness and adaptability of multimodal perception systems.

More recently, researchers have introduced transformer-based and frequency-aware models to enhance cross-modal representation learning. Transformer architectures offer global context modeling between RGB and THR modalities [16], while frequency-domain analyses reveal that different modalities contribute unevenly across spatial frequency bands—RGB features dominate high-frequency detail, whereas THR features emphasize low-frequency structure [6]. Such insights have motivated new fusion pipelines that adaptively balance modalities according to scene characteristics.

Despite remarkable progress in multimodal fusion, most existing studies still focus primarily on quantitative performance indicators such as accuracy or mIoU, while the perceptual quality of fused results has received far less attention. In practice, numerical improvements do not necessarily imply perceptually consistent or visually coherent outputs. Our experiments clearly reveal that the inclusion of thermal information can stabilize numerical accuracy yet sometimes lead to degraded visual realism. This observation indicates that numerical metrics alone cannot comprehensively represent the overall quality of multimodal perception. Therefore, our study emphasizes the need to re-examine RGB–THR fusion from a dual perspective—quantitative performance and perceptual fidelity—to achieve a more balanced and interpretable evaluation of multimodal systems.

3 Comparative Evaluation

3.1 Experimental Setup

Dataset.

All experiments in this study are conducted on the **Multi-Spectral Stereo (MS^2) Dataset** [14], a large-scale outdoor benchmark designed for multisensor perception and depth estimation research, as illustrated in Fig. 1.

The dataset provides synchronized recordings from stereo RGB, stereo near-infrared (NIR), and stereo thermal (THR) cameras, together with stereo LiDAR scanners and GPS/IMU navigation units. This comprehensive sensor suite enables precise geometric calibration and temporal synchronization across modalities, supporting detailed investigation of multimodal visual perception under real-world conditions.

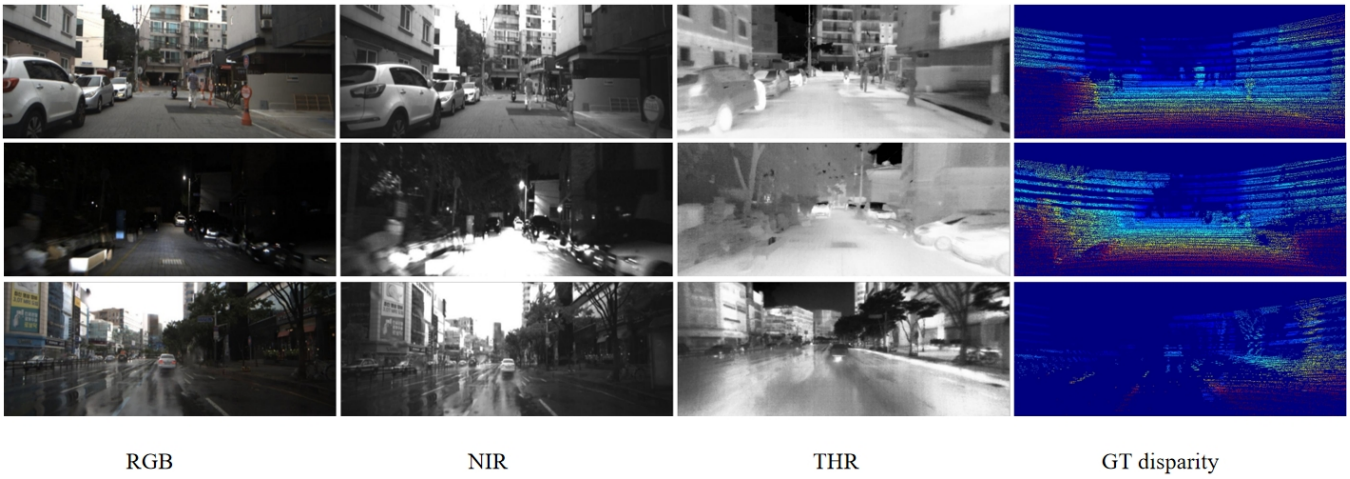

Figure 1: Overview of the Multi-Spectral Stereo (MS^2) dataset. The dataset provides synchronized RGB, NIR, and thermal stereo images captured under diverse environmental conditions (day, night, and rain), along with LiDAR, GPS, and IMU data for geometric consistency.

The MS^2 dataset comprises approximately 184 K rectified and synchronized stereo image pairs captured across diverse environments, including urban streets, residential areas, campus roads, and suburban regions. Each location was recorded multiple times under varying illumination and weather conditions, covering clear, cloudy, and rainy days, as well as morning, daytime, and nighttime scenes. Such diversity provides a valuable basis for studying modality-specific behaviors under challenging visual scenarios. In addition to multi-spectral imagery, the dataset also includes projected LiDAR depth maps, odometry information in both camera and LiDAR coordinate systems, and GPS/IMU trajectories to ensure metric-scale consistency.

In this work, we use theleft RGB andleft THR images, together with their corresponding LiDAR-projected depth maps, to perform a controlled comparison between visible-spectrum and long-wave infrared sensing for monocular depth estimation. Both modalities are spatially aligned and temporally synchronized, ensuring consistent supervision during training and evaluation. The input resolution is fixed at 640×256 pixels, and we follow the official preprocessing protocol of the dataset to maintain alignment and radiometric consistency across all samples.

The MS^2 dataset revisits the same physical locations under different illumination and weather conditions, providing a dense set of multi-condition correspondences for reliable cross-modal analysis. This feature enables systematic evaluation of modality-dependent robustness across structured and unstructured environments, as well as across varying visibility levels such as day, night, and rain. Its synchronized multi-sensor design and environmental diversity make MS^2 a suitable benchmark for analyzing how RGB and thermal modalities contribute to stable and reliable depth perception in complex outdoor scenes.

Implementation details. For network implementation, we employ the ConvNeXt-Tiny backbone as the feature extractor owing to its balance between computational efficiency and representational capacity. For the RGB modality, the model is initialized with ImageNet-pretrained weights to leverage general visual priors and accelerate convergence. Since thermal (THR) images are single-channel inputs lacking color information, two configurations are explored to ensure fair evaluation. In the first configuration, the THR image is replicated across three channels to match the input dimension of the pretrained ConvNeXt-Tiny model, thereby enabling weight transfer from the RGB domain. In the second configuration, the model is trained with a single-channel input using the same backbone structure but without pretrained initialization, allowing the network to learn modality-specific representations from scratch.

To maintain a fair comparison between modalities, we do not employ any data augmentation strategies such as random flipping, color jittering, or cropping. Both models are trained under identical hyperparameter settings, including optimizer configuration, learning rate schedule, and batch size. The implementation is based on PyTorch and executed on an NVIDIA RTX 4090 GPU with 24 GB of memory.

Training configuration. All models are trained for **25 epochs** using the **Adam optimizer** with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and an initial learning rate of 1×10^{-4}. A cosine decay schedule is adopted to gradually reduce the learning rate over time, ensuring stable convergence. The batch size is set to **8**, and all experiments are conducted on an NVIDIA RTX 4090 GPU with 24 GB of memory. We follow the official data preprocessing and normalization procedure of the MS^2 dataset to maintain consistency across modalities. No additional data augmentation or modality-specific tuning is applied during training to ensure a fair comparison between RGB and thermal inputs.

The overall loss function combines a **scale-invariant depth loss** and an **edge-aware smoothness term**, which are widely used in monocular depth estimation [4]. The final objective is

defined as:

$$L = L_{\text{si}} + \lambda L_{\text{sm}}, \quad \lambda = 0.1, \tag{1}$$

where the scale-invariant term L_{si} measures relative depth consistency in logarithmic space:

$$L_{\text{si}} = \frac{1}{n}\sum_i d_i^2 - \frac{1}{n^2}\left(\sum_i d_i\right)^2, \tag{2}$$

with $d_i = \log D_i - \log D_i^*$ representing the difference between predicted and ground-truth depth in log scale. The smoothness regularizer encourages spatial coherence while preserving image edges:

$$L_{\text{sm}} = \sum_{i,j}\left(|\partial_x D_{i,j}|e^{-|\partial_x I_{i,j}|} + |\partial_y D_{i,j}|e^{-|\partial_y I_{i,j}|}\right), \tag{3}$$

where D denotes the predicted depth map and I the corresponding input image. This combination ensures both global depth consistency and local structural smoothness in the predicted maps.

3.2 Evaluation Metrics

To quantitatively evaluate the performance of depth estimation, we adopt three widely used metrics—**Absolute Relative Error (AbsRel)**, **Root Mean Square Error (RMSE)**, and **Threshold Accuracy (δ_i)**—which were originally introduced by Eigen *et al.* [4].

These metrics jointly capture both the numerical deviation from the ground-truth depth and the relative structural consistency across scenes. All evaluations are performed on the official test split of the MS^2 dataset using the unfiltered LiDAR depth maps as reference.

Absolute Relative Error (AbsRel). This metric measures the mean relative deviation between the predicted depth D_i and the ground-truth depth D_i^*:

$$\text{AbsRel} = \frac{1}{n}\sum_{i=1}^{n}\frac{|D_i - D_i^*|}{D_i^*}. \tag{4}$$

A lower AbsRel value indicates a smaller proportional error, implying that the predicted depth magnitudes are closer to their true values. Because it normalizes the difference by D_i^*, AbsRel is particularly sensitive to near-range regions where depth changes rapidly, making it an effective indicator of local depth fidelity.

Root Mean Square Error (RMSE). RMSE evaluates the overall Euclidean distance between prediction and ground truth, reflecting global consistency across the entire image:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(D_i - D_i^*)^2}. \tag{5}$$

Unlike AbsRel, RMSE penalizes large absolute deviations more heavily, and is therefore dominated by outlier pixels or distant regions. A smaller RMSE value corresponds to a globally smoother and numerically stable depth prediction.

Threshold Accuracy (δ_i). To assess relative correctness independent of absolute scale, we follow the standard accuracy criterion proposed in [4]. For each pixel, the ratio between prediction and ground truth is computed, and the percentage of pixels satisfying the threshold condition is reported as

$$\text{Accuracy}(\delta_i) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left(\max\left(\frac{D_i}{D_i^*}, \frac{D_i^*}{D_i}\right) < \delta_i\right) \tag{6}$$

where $\mathbb{I}(\cdot)$ denotes an indicator function that equals 1 when the condition is satisfied and 0 otherwise. The threshold values are set to $\delta_i \in \{1.25,\ 1.25^2,\ 1.25^3\}$, corresponding to increasing levels of tolerance. Higher δ_i values represent looser error bounds, while δ_1 measures strict accuracy and δ_3 captures broader alignment.

This metric effectively measures how many pixels fall within a fixed multiplicative error bound, providing a complementary view to absolute-error measures.

Together, these three metrics provide a comprehensive assessment of depth estimation quality. AbsRel emphasizes relative precision in nearby regions, RMSE captures global numerical stability, and the threshold-based accuracy highlights structural consistency under scale variations. By jointly analyzing these indicators, we can evaluate not only the quantitative reliability of each modality but also its robustness to illumination and texture variations present in the MS^2 dataset.

3.3 Experimental Results

3.3.1 Quantitative Comparison

Table 1 presents the quantitative evaluation results of monocular depth estimation using RGB and thermal (THR) modalities under three illumination conditions—daytime, nighttime, and rainy—on the MS^2 dataset. Both networks were trained under identical optimization and data processing settings to ensure fair comparison. Across all environments, the THR-based model consistently achieves lower error metrics and higher accuracy rates, indicating that the thermal modality provides more stable geometric cues and greater robustness to illumination changes.

Under the *daytime* condition, the two modalities exhibit comparable performance, with THR showing a modest improvement in most metrics. The AbsRel and RMSE values of THR (0.08 and 2.96, respectively) are slightly lower than those of RGB (0.09 and 3.45). This marginal gap is attributed to the rich texture and color gradients available in RGB images under sufficient illumination, which enable reliable depth inference through photometric cues.

In the *nighttime* condition, the superiority of the thermal modality becomes prominent. The AbsRel decreases from 0.121 to 0.081, and RMSE drops from 4.11 to 2.84, reflecting a substantial reduction in overall depth error. Furthermore, the δ_1 accuracy improves from 0.872 to 0.938, confirming that THR preserves more consistent structural correspondence when visual information is degraded by low-light noise and contrast loss. Notably, the performance of THR remains

Table 1: Quantitative comparison between RGB and THR modalities under different illumination conditions on the MS^2 dataset. Lower is better for AbsRel, SqRel, RMSE, and RMSE(log); higher is better for δ_i.

Condition	Input	AbsRel ↓	SqRel ↓	RMSE ↓	RMSE(log) ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Day	RGB	0.090	0.427	3.454	0.117	0.911	0.979	**0.998**
	THR	**0.080**	**0.342**	**2.955**	**0.112**	**0.941**	**0.981**	0.995
Night	RGB	0.121	0.619	4.105	0.153	0.872	0.981	0.989
	THR	**0.081**	**0.335**	**2.844**	**0.112**	**0.938**	**0.980**	**0.991**
Rainy	RGB	0.139	0.897	4.841	0.182	0.841	0.913	0.981
	THR	**0.115**	**0.549**	**3.785**	**0.159**	**0.875**	**0.952**	**0.987**

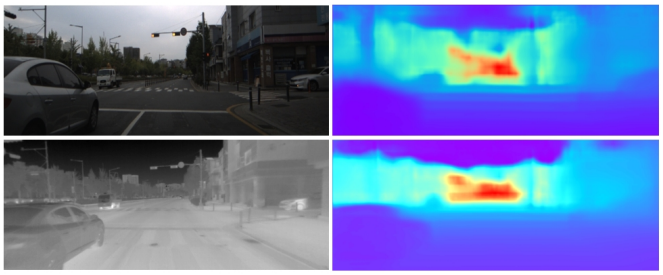

Figure 2: Qualitative comparison of RGB- and THR-based depth predictions under the *daytime* condition. The RGB modality exhibits clearer edges and richer local textures, while THR outputs appear smoother and more homogeneous.

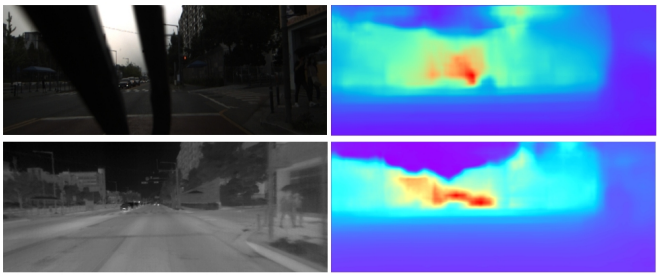

Figure 3: Qualitative comparison under the *rainy* condition. Despite visual occlusions from raindrops and wiper traces, THR maintains structural continuity, whereas RGB preserves distant object details with perceptual contrast.

highly stable between the daytime and nighttime settings, suggesting that thermal imaging is largely invariant to illumination intensity.

For the *rainy* condition, both modalities experience increased errors due to reflection, occlusion, and atmospheric scattering; however, THR still maintains a clear advantage. The AbsRel decreases from 0.139 to 0.115, and RMSE from 4.84 to 3.79, while δ_1 improves from 0.841 to 0.875. These results indicate that the thermal signal provides more coherent depth boundaries under adverse weather, mitigating the degradation effects commonly observed in RGB-based estimation.

In summary, the thermal modality exhibits strong resilience to environmental variations, yielding consistent performance across both well-lit and low-visibility scenarios. The relatively small performance gap between daytime and nighttime conditions further demonstrates the illumination-invariant characteristics of thermal sensing, highlighting its potential as a reliable alternative or complementary input for robust monocular depth estimation.

3.3.2 Qualitative Visualization

To provide a visual understanding of the modality-specific differences, we further present qualitative depth estimation results under three illumination conditions from the MS^2 dataset: daytime, rainy, and nighttime. Each visualization includes the input image and the corresponding predicted depth map for both RGB and THR modalities, highlighting the contrast between numerical stability and perceptual fidelity. Under the *daytime* condition (Fig. 2), the RGB-based prediction displays sharper edges and richer local details. Elements such as traffic lights and nearby vehicles are distinctly represented, demonstrating the benefit of texture and color gradients in depth reconstruction. In contrast, the THR-based map appears smoother and more uniform, suggesting higher numerical consistency but reduced perceptual richness.

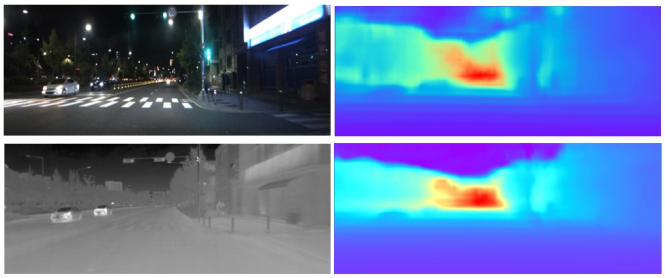

Figure 4: Qualitative comparison under the *nighttime* condition. RGB predictions capture meaningful visual cues such as vehicles and pedestrians, while THR produces smoother yet less visually expressive results.

In the *rainy* scenario (Fig. 3), the scene involves visual occlusion from raindrops and wiper traces. The THR modality maintains structural coherence since it is unaffected by these optical distortions, whereas the RGB prediction exhibits localized degradation. Nevertheless, distant objects such as vehicles remain perceivable in the RGB result, indicating that RGB retains a degree of depth sensitivity even under adverse visual conditions. Under the *nighttime* condition (Fig. 4), RGB predictions capture meaningful spatial cues such as pedestrians and vehicles with higher perceptual contrast, whereas THR

maintains smoother yet less expressive results. Despite the quantitative advantage of THR, the RGB-based outputs deliver more visually coherent depth perception.

In summary, while the THR modality achieves superior numerical accuracy across all conditions, the RGB modality produces perceptually more natural and structurally expressive depth maps. This observation highlights the existence of a performance–perception gap, emphasizing that quantitative superiority does not necessarily correlate with visual plausibility.

4 Discussion

The experimental findings presented in this study reveal a distinctive divergence between numerical performance and perceptual quality across RGB and thermal (THR) modalities in monocular depth estimation. Although the THR-based model consistently outperforms its RGB counterpart in quantitative indicators—achieving lower AbsRel and RMSE values under all illumination conditions—the qualitative analysis demonstrates that RGB predictions exhibit greater visual coherence, sharper boundaries, and richer structural expressiveness. This paradoxical outcome reflects the fundamental difference in how the two modalities encode visual information and how numerical optimization interacts with perceptual realism.

From a signal interpretation perspective, the thermal modality captures scene geometry primarily through radiometric emission differences, resulting in spatially smooth but low-frequency representations. Such inputs tend to minimize local gradient variance and yield stable predictions under low-visibility environments such as rain or night, explaining the superior metric values obtained by THR. However, this very stability comes at the cost of attenuated texture sensitivity, leading to over-smoothing in depth transitions and the loss of high-frequency cues that are critical for perceptual depth perception. In contrast, RGB images, rich in color and luminance gradients, provide abundant local features that enhance fine-grained spatial reconstruction. Consequently, although RGB models are more vulnerable to illumination noise, they preserve edge continuity and scene realism—factors that humans intuitively associate with visual quality.

These findings suggest that numerical accuracy and perceptual realism in depth estimation do not necessarily converge, especially across heterogeneous modalities. The observed performance–perception gap highlights a limitation of existing training objectives, which typically optimize for pixel-wise consistency while overlooking perceptual-level coherence. This misalignment underscores the need for evaluation frameworks that jointly assess numerical and perceptual aspects of depth quality, particularly in multimodal contexts where the data distributions are inherently unbalanced.

Future work should extend these insights toward *modality-aware fusion frameworks* that integrate the complementary strengths of RGB and THR sensing. Such methods could employ frequency-domain alignment or cross-modal attention to adaptively emphasize texture fidelity in RGB while leveraging the radiometric stability of THR under adverse conditions. Moreover, perceptually motivated loss functions and human-centered evaluation metrics may bridge the current gap between objective performance and subjective visual realism. Ultimately, achieving both quantitative robustness and perceptual fidelity will be a crucial step toward building reliable, interpretable, and human-aligned depth estimation systems for real-world applications.

5 Conclusion

This work presents an empirical study on the performance–perception relationship in monocular depth estimation using RGB and thermal (THR) modalities on the MS^2 dataset. Through systematic quantitative and qualitative analyses, we observe a consistent divergence between numerical accuracy and visual realism. Specifically, the THR modality achieves lower depth estimation errors and higher stability under adverse illumination conditions such as rain or night, demonstrating its robustness against environmental variations. However, the RGB modality consistently delivers more perceptually coherent depth maps, preserving edges, textures, and fine details that are visually aligned with human depth perception.

These findings underscore that numerical superiority does not necessarily imply perceptual fidelity, revealing an intrinsic imbalance in current objective functions and evaluation metrics. The study highlights the need to jointly consider perceptual and numerical dimensions when assessing and optimizing depth estimation models. In particular, future research should explore perceptually informed loss functions and multimodal fusion strategies that explicitly leverage the complementary strengths of RGB and THR data. By integrating radiometric stability from thermal sensing with the rich semantic and structural priors of RGB imagery, it may be possible to construct depth estimation frameworks that achieve both quantitative robustness and perceptual consistency across diverse environmental conditions.

References

[1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4009–4018, 2021.

[2] Sri Aditya Deevi, Connor Lee, Lu Gan, Sushruth Nagesh, Gaurav Pandey, and Soon-Jo Chung. Rgb-x object detection via scene-specific fusion modules. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7366–7375, 2024.

[3] Pengxiang Ding, Han Zhao, Wenjie Zhang, Wenxuan Song, Min Zhang, Siteng Huang, Ningxi Yang, and Donglin Wang. Quar-vla: Vision-language-action model for quadruped robots. In *European Conference on Computer Vision*, pages 352–367. Springer, 2024.

[4] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

[5] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115. IEEE, 2017.

[6] Jingxue Huang, Xilai Li, Tianshu Tan, Xiaosong Li, and Tao Ye. Mma-unet: A multi-modal asymmetric unet architecture for infrared and visible image fusion. *arXiv preprint arXiv:2404.17747*, 2024.

[7] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015.

[8] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3496–3504, 2021.

[9] Shehryar Khattak, Christos Papachristos, and Kostas Alexis. Visual-thermal landmarks and inertial fusion for navigation in degraded visual environments. In *2019 IEEE Aerospace Conference*, pages 1–9. IEEE, 2019.

[10] Lina Liu, Xibin Song, Jiadai Sun, Xiaoyang Lyu, Lin Li, Yong Liu, and Liangjun Zhang. Mff-net: Towards efficient monocular depth completion with multi-modal feature fusion. *IEEE Robotics and Automation Letters*, 8(2):920–927, 2023.

[11] Yunpeng Ma, Dengdi Sun, Qianqian Meng, Zhuanlian Ding, and Chenglong Li. Learning multiscale deep features and svm regressors for adaptive rgb-t saliency detection. In *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, volume 1, pages 389–392. IEEE, 2017.

[12] Gemma Piella. A general framework for multiresolution image fusion: from pixels to regions. *Information fusion*, 4(4):259–280, 2003.

[13] Jifeng Shen, Yifei Chen, Yue Liu, Xin Zuo, Heng Fan, and Wankou Yang. Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145:109913, 2024.

[14] Ukcheol Shin, Jinsun Park, and In So Kweon. Deep depth estimation from thermal image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1043–1053, 2023.

[15] Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *2020 IEEE international conference on robotics and automation (ICRA)*, pages 9441–9447. IEEE, 2020.

[16] Shivpal Singh, K Sathya Babu, Vasit Sagan, Chandrakanth Vipparla, K Palaniappan, and Hadi AliAkbarpour. Uav detect, track and follow (dtf) of non-stationary targets in aerial thermal videos. In *2025 17th International Conference on Computer and Automation Engineering (ICCAE)*, pages 165–170. IEEE, 2025.

[17] Wenxuan Song, Han Zhao, Pengxiang Ding, Can Cui, Shangke Lyu, Yaning Fan, and Donglin Wang. Germ: A generalist robotic model with mixture-of-experts for quadruped robot. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11879–11886. IEEE, 2024.

[18] Alexander Toet. Image fusion by a ratio of low-pass pyramid. *Pattern recognition letters*, 9(4):245–253, 1989.

[19] Zhengzheng Tu, Zhun Li, Chenglong Li, Yang Lang, and Jin Tang. Multi-interactive dual-decoder for rgb-thermal salient object detection. *IEEE Transactions on Image Processing*, 30:5678–5691, 2021.

[20] Zhengzheng Tu, Yan Ma, Zhun Li, Chenglong Li, Jieming Xu, and Yongtao Liu. Rgbt salient object detection: A large-scale dataset and benchmark. *IEEE Transactions on Multimedia*, 25:4163–4176, 2022.

[21] Yike Wang, Gongyang Li, and Zhi Liu. Sgfnet: Semantic-guided fusion network for rgb-thermal semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7737–7748, 2023.

[22] Yingjie Wang, Qiuyu Mao, Hanqi Zhu, Jiajun Deng, Yu Zhang, Jianmin Ji, Houqiang Li, and Yanyong Zhang. Multi-modal 3d object detection in autonomous driving: a survey. *International Journal of Computer Vision*, 131(8):2122–2152, 2023.

[23] AN Wilson, Khushi Anil Gupta, Balu Harshavardan Koduru, Abhinav Kumar, Ajit Jha, and Linga Reddy Cenkeramaddi. Recent advances in thermal imaging and its applications using machine learning: A review. *IEEE Sensors Journal*, 23(4):3395–3407, 2023.

[24] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):502–518, 2020.

[25] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. New crfs: Neural window fully-connected crfs for monocular depth estimation. *arXiv preprint arXiv:2203.01502*, 2022.

[26] Qiang Zhang, Nianchang Huang, Lin Yao, Dingwen Zhang, Caifeng Shan, and Jungong Han. Rgb-t salient object detection via fusing multi-level cnn features. *IEEE Transactions on Image Processing*, 29:3321–3335, 2019.

[27] Qiang Zhang, Tonglin Xiao, Nianchang Huang, Dingwen Zhang, and Jungong Han. Revisiting feature fusion for rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5):1804–1818, 2020.

[28] Wujie Zhou, Qinling Guo, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Ecffnet: Effective and consistent feature fusion network for rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1224–1235, 2021.