A MULTIMODAL INTRINSICS-GUIDED THERMAL-AWARE FRAMEWORK FOR RGB LOW-LIGHT IMAGE ENHANCEMENT

Simone Melcarne[⋆] *Jean-Luc Dugelay*[⋆]

[⋆] Eurecom Research Center, Digital Security Department, Sophia Antipolis, France

ABSTRACT

Low-light image enhancement is crucial in situations where visible sensors might suffer from severe noise and information loss (*e.g.,* nighttime surveillance). Recent approaches investigate auxiliary modalities invariant to illumination to improve the performance, such as thermal infrared imaging. We propose a Multimodal Intrinsics-Guided Framework that integrates RGB and thermal data to reconstruct well-lit images. Our method utilizes a two-stage pipeline: first, we employ an intrinsic decomposition strategy to separate reflectance and shading components through knowledge distillation, where a teacher network guides a student model in reconstructing consistent intrinsic components; then, a refinement stage restores fine structures and visual details. We train the proposed model on synthetic data from HDRT dataset and demonstrate strong generalization to real-world benchmarks such as LLVIP and V-TIEE, outperforming state-of-the-art methods in most evaluation metrics. *Code is available at*: https://github.com/simonemelc/TIRGlow

***Index Terms*—** Low-light Image Enhancement, Multimodal Fusion, Thermal Infrared, Intrinsic Decomposition.

1. INTRODUCTION

Low-light image enhancement (LLIE) plays a key role in computer vision, as images captured at night often suffer from low contrast, heavy noise, and color distortion. These degradations reduce visual quality and strongly affect downstream tasks, especially in nighttime surveillance, where reliable scene interpretation is required for monitoring, tracking, and safety-related applications. Traditional enhancement methods mainly adjust intensity and contrast, but they often amplify noise or introduce visual artifacts, limiting their effectiveness in real-world scenarios [1]. Recent learning-based approaches improve performance but still struggle when operating on severely under-exposed images with low signal-to-noise ratios [2]. To address these limitations, there is a growing interest in using auxiliary sensors that are less sensitive to illumination changes, such as LiDAR [3], event cameras [4], and thermal infrared (TIR) sensors [2, 5]. In particular, TIR imaging captures emitted heat rather than reflected light, providing stable structural and textural cues under poor lighting conditions. Motivated by this trend, our work focuses on integrating thermal information with low-light RGB data to achieve more robust and visually consistent image enhancement. The main contributions from this paper are: 1) We employ a knowledge distillation strategy where a teacher network guides the student to disentangle intrinsic properties, allowing better physical consistency and color fidelity compared to standard methods. 2) We adapt a multi-scale attention gating mechanism to the multimodal domain to selectively integrate thermal features and inject structural details only where the visible signal is degraded, preventing the propagation of thermal noise common in naive fusion approaches. 3) We validate a realistic low-light simulation pipeline on HDRT [6] data that enables robust zero-shot generalization to real-world benchmarks (LLVIP [7] and V-TIEE [8]), outperforming state-of-the-art methods especially in structure and color preservation.

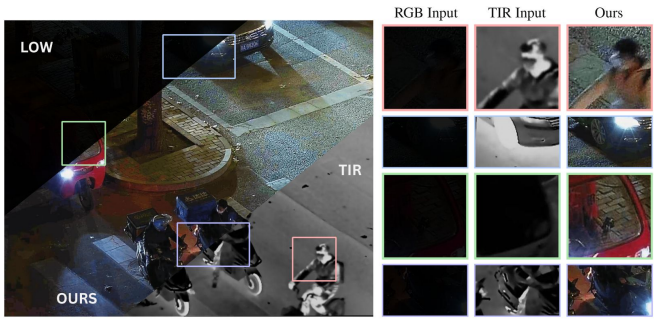

Fig. 1. A visual example on a very low-light image from LLVIP: the proposed method fuses low-light RGB and TIR to recover missing textures and colors.

2. RELATED WORK

Traditional LLIE. Classical techniques adopted to solve LLIE task include gamma-correction and histogram equalization; however these early approaches do not take illumination into account, leading to physically inconsistent results [9, 10].

This work was partly supported by the European Union's Horizon Europe research and innovation program under Grant Agreement No 101094831 for the Converge-Telecommunications and Computer Vision Convergence Tools for Research Infrastructures project.

Retinex-based formulations, instead, successfully treat the problem through the reflectance-illumination decomposition which improve global contrast but often amplify noise or introduce halos and color artifacts under complex illumination [11]. Deep learning solutions have significantly improved LLIE performance by combining Retinex priors with CNN-based decompositions, while more recent transformer-based methods [1] leverage self-attention to capture long-range dependencies and non-uniform illumination patterns. GAN-based [12, 13] and diffusion-inspired [14] enhancement models further improve perceptual quality by learning realistic brightness, texture, and color distributions in low-light conditions. Nevertheless, large-scale evaluations and recent LLIE challenges [15] show that most existing models still struggle on extremely dark images.

Thermal-Assisted LLIE. Several works exploit auxiliary modalities that are less affected by illumination, such as TIR imaging [2, 5]. Thermal-guided enhancement approaches typically use thermal edges or textures as guidance maps to regularize the enhancement of visible images, improving structural consistency without excessively boosting RGB noise [2]. More recent multimodal frameworks explicitly fuse RGB and thermal features through attention mechanisms [8]. Beyond pure enhancement, Zhao et al. [16] introduces visible–infrared information synthesis for severe low-light scenarios, arguing that single-modality enhancement is fundamentally limited when large regions lack valid visible signal.

3. METHODOLOGY

3.1. Intrinsic Image Decomposition

We base our work on the physical principles of image formation. Intrinsic Image Decomposition (IID) theory [17] states that an image I can be modeled as the pixel-wise product of two distinct layers: reflectance R and shading S. Adopting the standard Lambertian-world assumption, this decomposition is formulated as:

$$I = R_{RGB} \odot S_{gray}. \tag{1}$$

$R_{RGB} \in \mathbb{R}^{3\times H\times W}$ represents the illumination invariant material properties and colors of the scene, $S_{gray} \in \mathbb{R}^{1\times H\times W}$, instead, is a gray-scale image that captures the light distribution and the surface geometry. The separation between illumination and reflectance has a long history in LLIE with recent studies further showing that solving reconstruction or enhancement tasks in the intrinsic domain, rather than in luminance, improves color fidelity and simplifies statistical modeling [18, 19].

3.2. Teacher Network

A significant challenge in training intrinsic models is the lack of ground-truth reflectance and shading data in the context of LLIE. We address this limitation through a *knowledge distillation* strategy where we utilize as a teacher network a pretrained model from Careaga et al. [20] which proved to deliver high-quality decomposition in the wild. This teacher processes a normal-exposure image I_{gt} to extract reflectance R_{gt} and shading S_{gt} maps, serving then as targets for our student network. Following the preprocessing solution from [20], the input image is first normalized to the $[0,1]$ range and linearized (I_{lin}) using a standard gamma value of 2.2. The estimated R_{gt} is bounded to $[0,1]$, instead S_{gt} is unbounded to physically account for light intensities that exceed the normalized dynamic range of the input image.

3.3. Student Network

During the training phase, the proposed student network first predicts the intrinsic decomposition from paired low-light RGB and thermal inputs, using the teacher's output as supervision (R_{gt}, S_{gt}). Then, a refinement stage is added to predict the final well lit image $\hat{I}_{out}$. The framework operates as an end-to-end system and it is summarized in Figure 2 (more details can be found in the Supplementary). The network utilizes a dual-encoder CNN to extract features from both low-light RGB I_{low} and thermal T frames. Inspired by Han et al. [21], we implement a multi-scale attention gating mechanism [22] at every resolution level l to selectively filter thermal features. First (FB in Figure 2), we fuse the encoders features (F_I^l, F_T^l) to obtain the raw skip connection $\mathcal{X}^l$ and then (AG in Figure 2) we generate an attention mask $\mathcal{M}^l$ using the upsampled features $\hat{G}^l$ from the decoder as a structural reference; the raw skip connection $\mathcal{X}^l$ is then multiplied with $\mathcal{M}^l$ to obtain the final gated skip connection $\hat{\mathcal{X}}^l$:

$$\hat{\mathcal{X}}^l = \mathcal{X}^l \odot \mathcal{M}^l, \quad \mathcal{M}^l = \sigma\Big(W_\psi \phi\big(W_x \mathcal{X}^l + W_g \hat{G}^l\big)\Big) \tag{2}$$

where σ and ϕ denote the Sigmoid and ReLU activations respectively. W_x and W_g are learnable 1×1 kernels that project inputs to an intermediate feature space, while W_ψ aggregates them into a single-channel attention map. This mask $\mathcal{M}^l \in [0,1]$ acts as a spatial filter, suppressing background noise while highlighting thermal-guided structural details. At this point, the decoder predicts $\hat{R}$ and $\hat{S}$. Given the unbounded nature of the shading, we follow Dille et al. [18] and use an inverse shading map $\hat{D} = (1+\hat{S})^{-1}$ to prevent numerical instability and gradient explosion during loss minimization. We recover then $\hat{S} = (1-\hat{D})/\hat{D}$ and using the intrinsic law we produce the first linear coarse reconstruction $\hat{I}_{coarse} = \hat{R} \odot \hat{S}$. Then, we concatenate $\hat{I}_{coarse}$ with the thermal input T before feeding into the refinement net, consisting of a sequence of Residual Blocks [23] that learn a residual error map $\mathcal{E}$ to be added to $\hat{I}_{coarse}$ in order to correct potential artifacts or color shifts and finally obtain the prediction $\hat{I}_{out}$ after a Sigmoid (σ). Note that since we train the full pipeline end-to-end against the RGB ground truth, this allows the student to compensate for potential failure points that may derive from the teacher's

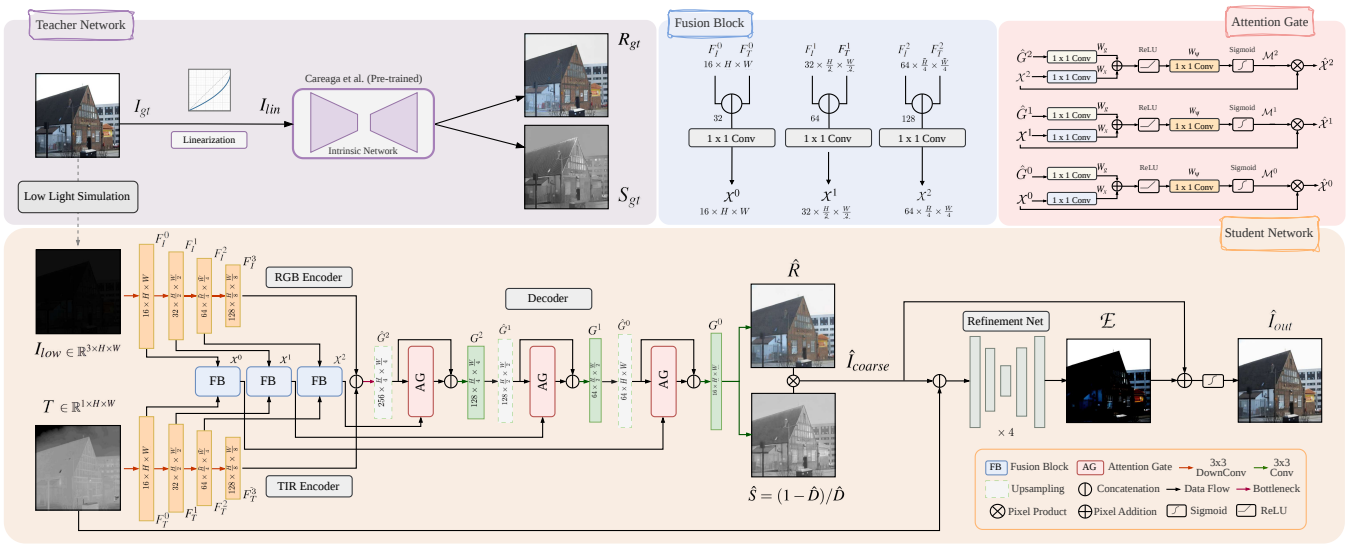

Fig. 2. Overview of the proposed framework: the teacher net provides intrinsic supervision; the student net fuses RGB and TIR features via Fusion Blocks (FB) and refines the resulting skip connections using Attention Gates (AG). The model first predicts $\hat{I}_{\text{coarse}}$, and then applies a Residual Refinement Net guided by the thermal input to produce the final enhanced result $\hat{I}_{\text{out}}$.

priors, ensuring the final results is not strongly impacted by its limitations.

3.4. Objective Functions

We optimize the network end-to-end using a composite objective function:

$$\mathcal{L}_{total} = \alpha(\mathcal{L}_R + \mathcal{L}_S) + \beta\mathcal{L}_{phy} + \gamma\mathcal{L}_{ref} \tag{3}$$

Reflectance and Shading. The reflectance ($\mathcal{L}_R$) and shading ($\mathcal{L}_S$) terms are supervised using a smooth ℓ_1 pixel criterion ($\mathcal{L}_{pix}$) combined with a structural edge loss ($\mathcal{L}_{edge}$) to preserve high-frequency details. For reflectance, we additionally employ a CIELAB loss ($\mathcal{L}_{lab}$) to enforce chromatic consistency. As mentioned before, we use the inverse shading representation $D = (1+S)^{-1}$ for computing the loss:

$$\mathcal{L}_R(R,\hat{R}) = \lambda_p\mathcal{L}_{pix} + \lambda_e\mathcal{L}_{edge} + \lambda_c\mathcal{L}_{lab} \tag{4}$$

$$\mathcal{L}_S(D,\hat{D}) = \lambda_p\mathcal{L}_{pix} + \lambda_e\mathcal{L}_{edge} \tag{5}$$

Physical Consistency and Refinement. We enforce the intrinsic constraint via a reconstruction loss against $\hat{I}_{coarse}$ and the ground truth linear image I_{lin}. Also in this case, we apply the loss in the inverse domain for numerical stability:

$$\mathcal{L}_{phy} = \mathcal{L}_{pix}\left((1+I_{lin})^{-1}, (1+\hat{I}_{coarse})^{-1}\right) \tag{6}$$

The refinement module predicts a linear image $\hat{I}_{out}$; to align with human perception, we compute losses in the sRGB domain applying a gamma correction. The total loss compares $\hat{I}_{out}^{1/\gamma} = (\hat{I}_{out})^{\frac{1}{2.2}}$ against the sRGB ground truth I_{gt}:

$$\mathcal{L}_{ref}(\hat{I}_{out}^{1/\gamma}, I_{gt}) = \lambda_p\mathcal{L}_{pix} + \lambda_e\mathcal{L}_{edge} + \lambda_c\mathcal{L}_{lab} + \lambda_\phi\mathcal{L}_{perc} \tag{7}$$

Detailed definitions of individual loss components (*i.e.*, $\mathcal{L}_{pix}$, $\mathcal{L}_{edge}$, $\mathcal{L}_{lab}$, $\mathcal{L}_{perc}$) and the specific hyperparameter values are provided in the Supplementary.

4. EXPERIMENTAL SETUP

4.1. Training Data Synthesis

To train the model, we require a large amount of triplets of low-light RGB I_{low}, thermal T, and well-lit ground truth I_{gt}. We use the HDRT dataset [6] which contains 10,000 HDR-TIR pairs and provides per each scene under-, over-, and well-exposed SDR renderings. Although the dataset includes a real low-light image (the under-exposed one), we observed that its exposure reduction was not sufficiently aggressive to represent severe darkness conditions so we opted to select the well-exposed image as the ground truth and apply to it a low-light simulation pipeline as follows:

$$I_{low} = Q\left[\left(\kappa \cdot I_{gt}^{\gamma} + n_{shot} + n_{read}\right)^{1/\gamma}\right] \tag{8}$$

Equation 8 first converts sRGB images from HDRT in the linear domain through a gamma operation ($\gamma = 2.2$) and applies an exposure gain $\kappa \in [0.0005, 0.005]$. Then, following [2, 25], we add noise (*shot* and *read* components). While true Poisson-Gaussian modeling is physically accurate only for RAW data, approximating it on linearized post-ISP data remains standard practice. Finally, the image is brought back to sRGB ($1/\gamma$) and quantized (Q). See supplementary for additional details.

Table 1. Quantitative comparison across three datasets. *Training Data* indicates the source domain used for training. *Input* reports the input sensor configuration. *Setting* distinguishes between *Supervised* and *Zero-Shot*. Best scores in red, second-best in blue.

Method	Training Data	Input	Setting	PSNR↑	SSIM↑	LPIPS↓	ΔE_{00}↓	#Params
HDRT								
EnGAN [12]	Unpaired (LOL, RAISE, HDR)	RGB	Zero-Shot	13.97	0.627	0.440	16.02	8.64M
LLFlow [24]	LOL	RGB	Zero-Shot	13.71	0.652	0.416	19.02	38.38M
RetinexFormer [1]	LOL-v2	RGB	Zero-Shot	16.12	0.668	0.363	13.06	1.61M
RT-X Net [8]	LLVIP	RGB+TIR	Zero-Shot	15.26	0.539	0.419	16.92	1.64M
RT-X Net [8] *(Retrained)*	HDRT	RGB+TIR	Supervised	26.02	0.827	0.183	5.27	1.64M
Ours	HDRT	RGB+TIR	Supervised	28.99	0.883	0.112	3.51	1.38M
LLVIP								
EnGAN [12]	Unpaired (LOL, RAISE, HDR)	RGB	Zero-Shot	19.83	0.704	0.320	11.02	8.64M
LLFlow [24]	LOL	RGB	Zero-Shot	13.90	0.561	0.449	18.67	38.38M
RetinexFormer [1]	LOL-v2	RGB	Zero-Shot	16.15	0.599	0.363	13.99	1.61M
RT-X Net [8]	LLVIP	RGB+TIR	Supervised	27.03	0.622	0.234	8.89	1.64M
Ours	HDRT	RGB+TIR	Zero-Shot	25.80	0.795	0.220	4.93	1.38M
V-TIEE								
EnGAN [12]	Unpaired (LOL, RAISE, HDR)	RGB	Zero-Shot	18.94	0.574	0.412	14.04	8.64M
LLFlow [24]	LOL	RGB	Zero-Shot	16.91	0.589	0.470	16.57	38.38M
RetinexFormer [1]	LOL-v2	RGB	Zero-Shot	17.13	0.584	0.439	16.36	1.61M
RT-X Net [8]	LLVIP	RGB+TIR	Zero-Shot	18.21	0.347	0.503	17.71	1.64M
Ours	HDRT	RGB+TIR	Zero-Shot	20.56	0.614	0.473	11.65	1.38M

Table 2. Ablation study on HDRT Dataset. Best results in red.

Experiment	PSNR ↑	SSIM ↑	LPIPS ↓	ΔE_{00} ↓
(a) w/o Thermal (RGB only)	25.95	0.856	0.148	4.88
(b) Shallow Attention ($\mathcal{M}^0$ only)	27.98	0.875	0.125	3.94
(c) w/o Refinement (Coarse only)	28.22	0.878	0.121	3.73
(d) Full Model	28.99	0.883	0.112	3.51

4.2. Implementation Details

We implemented our framework using PyTorch on a `NVIDIA L40S GPU`. The network was trained for 150 epochs using the `AdamW` optimizer, a weight decay of 1×10^{-4} and a `CosineAnnealingLR` scheduler (learning rate initialized at 1×10^{-4}). During training, we applied consistent geometric data augmentation (flips, rotations) across all modalities and we randomized for each sample the exposure gain and noise parameters from Equation 8, alternating between harder and softer darkness profiles. The input resolution was set to 384×384 pixels with a batch size of 8.

5. RESULTS

5.1. Quantitative Analysis

We evaluate the proposed solution using HDRT, LLVIP and V-TIEE dataset and compare the results against four representative LLIE methods: EnGAN [12], LLFlow [24], RetinexFormer [1], and RT-X Net [8]. Due to unavailability of the code, we could not include other thermal-aware LLIE methods in the comparison [2, 3, 14]. Table 1 reports the quantitative results in terms of standard PSNR, SSIM and LPIPS; we compute also CIE ΔE_{00} scores since a core part of this work is to analyze color preservation properties. Note that methods like [1, 8, 24] use post-processing operations to achieve better results, but they require data augmentation and ground truth information to perform mean alignment, ending up being impractical for fast and target-agnostic scenarios like nighttime surveillance; therefore we disabled this option to run a fair comparison across all models. The significant performance gap on HDRT between our model and the others is justified by the fact that competitors are tested in a Zero-Shot manner using official pre-trained weights, simulating a realistic cross-domain deployment. For this purpose, we retrain RTX-Net on HDRT and demonstrate we still obtain better results. Notably, our framework has robust adaptability across LLVIP and V-TIEE, outperforming the other methods particularly in structural similarity (SSIM) and color fidelity (ΔE_{00}). Finally, we highlight the efficiency of our architecture as it is the most lightweight among the compared methods.

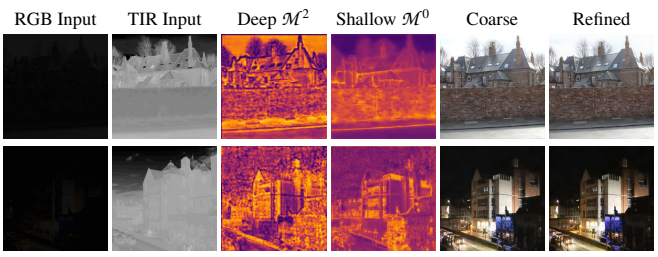

Fig. 3. Visualization of the attention masks (deepest and finest) and of the coarse and refined reconstruction results on HDRT samples.

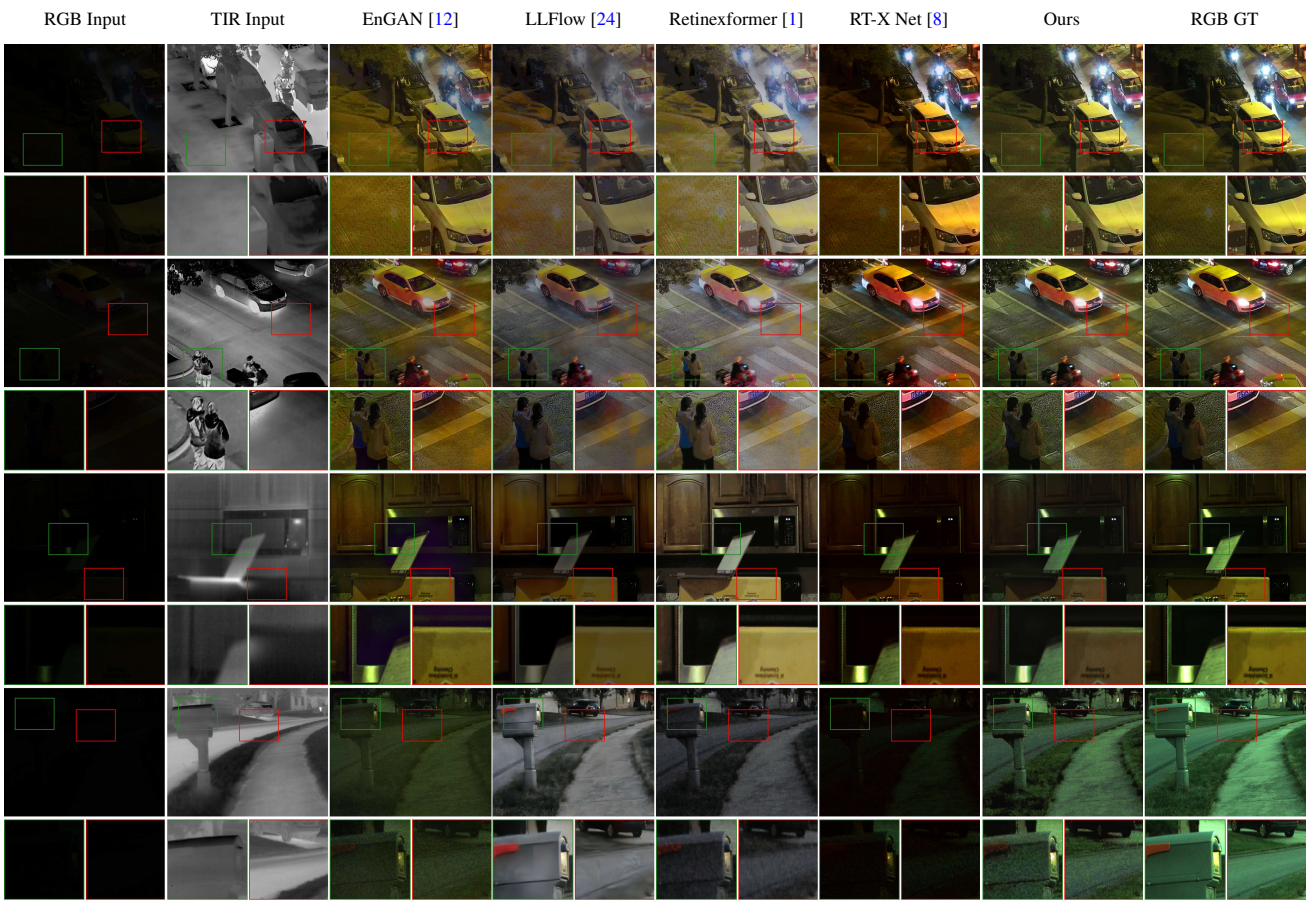

Fig. 4. Visual comparison example. The first two rows show low-light samples from the LLVIP dataset, while the last two rows are from V-TIEE. Each column compares the analyzed methods against the ground truth.

5.2. Ablation Study

We conducted an ablation study to evaluate the impact of thermal information, multi-scale attention and the refinement stage. Table 2 reports the scores for each experiment: the proposed full model achieves the best performance across all metrics validating the influence of each analyzed component. First, thermal shows to be useful in guiding the restoration; second, limiting the attention mechanism only to the shallowest level proves to be suboptimal, suggesting that filtering features at multiple resolution is more effective; finally, the refinement stage shows its importance in recovering finer details and more faithful results.

5.3. Visual Comparison

Visual comparisons on LLVIP and VTIEE samples are presented in Figure 4. In general, we notice that approaches like EnGAN tend to hallucinate textures or introduce artifacts; LLFlow method often produce overly smooth results; retinex-based transformers like RetinexFormer and RTX-Net instead amplify high-frequency noise in the deepest shadows. In contrast, our framework, leveraging the clean structural priors from thermal, helps maintaining natural structural definition and color saturation. More comparisons, together with the specification of the test sets for LLVIP and V-TIEE, can be found in the Supplementary.

6. CONCLUSION

In this work we present a physics-guided framework for LLIE that explicitly integrates thermal structural cues into the intrinsic decomposition process, targeting extreme scenarios where the RGB signal is severely degraded. While trained on synthetic data from HDRT, experimental results on LLVIP and V-TIEE show that the proposed approach generalizes well and better preserves structure and color than recent state-of-the-art methods, reducing noise, artifacts, and color instability. The conducted ablation studies also confirm the contribution of thermal information and of the attention-based fusion mechanism. Future work will focus on extending the framework to video with real-time constraints and to richer multimodal settings (*e.g.,* event or LiDAR data).

7. REFERENCES

[1] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang, "Retinexformer: One-stage retinex-based transformer for low-light image enhancement," 2023.

[2] Yanpeng Cao, Xi Tong, Fan Wang, Jiangxin Yang, Yanlong Cao, Sabin Tiberius Strat, and Christel-Loic Tisse, "A deep thermal-guided approach for effective low-light visible image enhancement," *Neurocomputing*, vol. 522, pp. 129–141, 2023.

[3] Zhen Wang, Yaozu Wu, Dongyuan Li, Guang Li, Peide Zhu, Ziqing Zhang, and Renhe Jiang, "Lidar-assisted image restoration for extreme low-light conditions," *Knowledge-Based Systems*, vol. 316, pp. 113382, 2025.

[4] Lei Sun, Yuhan Bao, Jiajun Zhai, Jingyun Liang, Yulun Zhang, Kaiwei Wang, Danda Pani Paudel, and Luc Van Gool, "Low-light image enhancement using event-based illumination estimation," 2025.

[5] Zhen Wang, Yaozu Wu, Dongyuan Li, Shiyin Tan, and Zhishuai Yin, "Thermal-aware low-light image enhancement: A real-world benchmark and a new lightweight model," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 8, pp. 8223–8231, Apr. 2025.

[6] Jingchao Peng, Thomas Bashford-Rogers, Francesco Banterle, Haitao Zhao, and Kurt Debattista, "Hdrt: A large-scale dataset for infrared-guided hdr imaging," *Information Fusion*, vol. 120, pp. 103109, 2025.

[7] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, Shengjie Liu, and Wenli Zhou, "Llvip: A visible-infrared paired dataset for low-light vision," 2021.

[8] Raman Jha, Adithya Lenka, Mani Ramanagopal, Aswin Sankaranarayanan, and Kaushik Mitra, "Rt-x net: Rgb-thermal cross attention network for low-light image enhancement," 2025.

[9] Zhan Jingchun, Goh Eg Su, and Mohd Shahrizal Sunar, "Low-light image enhancement: A comprehensive review on methods, datasets and evaluation metrics," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 36, no. 10, Dec. 2024.

[10] Shen Zheng, Yiling Ma, Jinqian Pan, Changjie Lu, and Gaurav Gupta, "Low-light image and video enhancement: A comprehensive survey and beyond," 2024.

[11] Zexin Wang, Letu Qingge, Qingyi Pan, and Pei Yang, "Retinex decomposition based low-light image enhancement by integrating swin transformer and u-net-like architecture," *IET Image Processing*, vol. 18, no. 11, pp. 3028–3041, 2024.

[12] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang, "Enlightengan: Deep light enhancement without paired supervision," 2021.

[13] Lin Wang, Lei Zhao, Tao Zhong, et al., "Low-light image enhancement using generative adversarial networks," *Scientific Reports*, vol. 14, pp. 18489, 2024.

[14] Haiyan Jin, Wenfan Yang, Haonan Su, Yuanlin Zhang, and Bin Wang, "Nar-diff: A noise-adaptive reflectance diffusion model for low-light image enhancement," in *2025 IEEE International Conference on Image Processing (ICIP)*, 2025, pp. 983–988.

[15] Xiaoning Liu, Zongwei Wu, Ao Li, Florin-Alexandru Vasluianu, Yulun Zhang, Shuhang Gu, Le Zhang, Ce Zhu, Radu Timofte, Zhi Jin, et al., "Ntire 2024 challenge on low light image enhancement: Methods and results," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6571–6594.

[16] Chen Zhao, Mengyuan Yu, Fan Yang, and Peiguang Jing, "Viis: Visible and infrared information synthesis for severe low-light image enhancement," 2025.

[17] Elena Garces, Carlos Rodriguez-Pardo, Dan Casas, and Jorge Lopez-Moreno, "A survey on intrinsic images: Delving deep into lambert and beyond," 2021.

[18] Sebastian Dille, Chris Careaga, and Yağız Aksoy, "Intrinsic single-image hdr reconstruction," 2024.

[19] S Melcarne, P Cyriac, J L Dugelay, A Artusi, and F Banterle, "A color preserving tone mapping framework in the intrinsic domain," *Journal of Physics: Conference Series*, vol. 3128, no. 1, pp. 012008, oct 2025.

[20] Chris Careaga and Yağız Aksoy, "Intrinsic image decomposition via ordinal shading," *ACM Transactions on Graphics*, vol. 43, no. 1, pp. 1–24, Nov. 2023.

[21] Jin Han, Yixin Yang, Peiqi Duan, Chu Zhou, Lei Ma, Chao Xu, Tiejun Huang, Imari Sato, and Boxin Shi, "Hybrid high dynamic range imaging fusing neuromorphic and conventional images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[22] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert, "Attention u-net: Learning where to look for the pancreas," 2018.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[24] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex C Kot, "Low-light image enhancement with normalizing flow," *arXiv preprint arXiv:2109.05923*, 2021.

[25] Feifan Lv, Yu Li, and Feng Lu, "Attention guided low-light image enhancement with a large scale low-light simulation dataset," 2020.

SUPPLEMENTARY MATERIAL: A MULTIMODAL INTRINSICS-GUIDED THERMAL-AWARE FRAMEWORK FOR RGB LOW-LIGHT IMAGE ENHANCEMENT

Simone Melcarne[⋆] *Jean-Luc Dugelay*[⋆]

[⋆] Eurecom Research Center, Digital Security Department, Sophia Antipolis, France

1. DETAILED NETWORK ARCHITECTURE

The proposed framework employs a dual-encoder U-Net architecture. During our experimentation, we found out that utilizing Group Normalization (GN) instead of Batch Normalization throughout the network was helpful in mitigating the training instability and achieving overall better results.

1.1. Fundamental Building Blocks

In this section we detail the hyperparameters of the core computational units:

- `ConvBlock`: The standard processing unit consists of a 3×3 Convolution (with padding 1, bias=False), followed by Group Normalization ($N_{groups} = 8$) and a LeakyReLU activation (slope $\alpha = 0.2$).
- `ResBlock`: Utilized in the refinement stage, it consists of a `ConvBlock` followed by a second 3×3 Convolution and Group Normalization. A residual connection adds the input feature map to the output before a final LeakyReLU activation ($\alpha = 0.2$).

1.2. Encoder-Decoder Configuration

The backbone is a symmetric 4-level U-Net. To maintain a lightweight architecture, we set the initial network width to $C = 16$ channels for both inputs.

- **Dual Encoders:** We utilize two independent encoders. Downsampling is performed via strided convolutions ($k = 3, s = 2, p = 1$), doubling the channels at each stage ($16 \to 32 \to 64 \to 128$).
- **Bottleneck:** The bottleneck concatenates the deepest features from both modalities ($128 + 128$) and processes them via a `ConvBlock` with 256 channels, preserving the spatial resolution of $H/8 \times W/8$.
- **Decoder & Fusion:** Upsampling is performed via bilinear interpolation followed by a concatenation with the gated skip connections. The skip connections are fused via a 1×1 convolution that reduces the concatenated channels ($C_l + C_l$) back to C_l before entering the Attention Gate.
- **Prediction Heads:** The decoder branches into two heads for Reflectance (R) and Inverse Shading (D). Each head consists of: 3×3 Conv $\to$ LeakyReLU $\to$ 3×3 Conv $\to$ Sigmoid, mapping the 16-channel feature map to 3 channels (for R) or 1 channel (for D).

1.3. Refinement Module

The refinement network is designed to recover details lost during the coarse prediction. It accepts a 4-channel input obtained by concatenating $\hat{I}_{coarse}$ (3 ch) and the original thermal input T (1 ch). First, a `ConvBlock` maps the 4-channel input to a hidden dimension of 32 channels; then a sequence of 4 `ResBlocks` processes the features at full resolution ($H \times W$) to capture fine structural dependencies; finally a 1×1 convolution projects the features to 3 channels, producing the residual map $\mathcal{E}$. The final output is obtained as $\hat{I}_{out} = \sigma(\hat{I}_{coarse} + \mathcal{E})$, where σ is the Sigmoid function.

2. LOSS COMPONENTS DESCRIPTION

In this section, we present more in detail the implementation of the loss functions used to supervise the network. For all formulations, we use Y to denote the ground truth and $\hat{Y}$ the prediction; N is the total number of pixels.

Pixel Reconstruction Loss ($\mathcal{L}_{pix}$). To compute the photometric error, we employ the Smooth ℓ_1 loss, which is less sensitive to outliers than ℓ_2. It is defined as:

$$\begin{aligned} \mathcal{L}_{pix}(\hat{Y}, Y) &= \frac{1}{N} \sum_p \rho_\delta\big(\hat{Y}_p - Y_p\big), \\ \rho_\delta(d) &= \begin{cases} 0.5 \cdot d^2/\delta, & \text{if } |d| < \delta, \\ |d| - 0.5 \cdot \delta, & \text{otherwise,} \end{cases} \end{aligned} \tag{1}$$

where δ is set to 0.01. This loss is always applied with weight $\lambda_p = 1.0$.

Edge Loss ($\mathcal{L}_{edge}$). To preserve structural sharpness, we use a composite loss constraining both gradient magnitude and orientation. Let ∇ denote the spatial gradient operator and $M(\cdot) = \|\nabla \cdot\|_2$ its magnitude. The function is defined as:

Fig. 1. Low light simulation from HDRT ground truth normal-exposed image using different values for the exposure gain. Best if viewed in digital format.

$$\begin{aligned} \mathcal{L}_{edge}(\hat{Y},Y) &= \mathcal{L}_{mag} + 0.5 \cdot \mathcal{L}_{dir} \\ \mathcal{L}_{mag} &= \left\| M(\hat{Y}) - M(Y) \right\|_1 \\ \mathcal{L}_{dir} &= \frac{1}{N} \sum_p \mathbf{1}_{\{M(Y)_p > \tau\}} \cdot \left(1 - \frac{\nabla \hat{Y}_p \cdot \nabla Y_p}{M(\hat{Y})_p \cdot M(Y)_p + \varepsilon} \right) \end{aligned} \tag{2}$$

where $\mathbf{1}_{\{\cdot\}}$ is a function that aims to mask flat regions (threshold $\tau = 0.05$) to ensure numerical stability and p is any pixel. The weight here is always set to $\lambda_e = 0.8$.

Color Loss ($\mathcal{L}_{lab}$). To decouple chromaticity from luminance, we first convert RGB images to the CIELAB space and minimize the error specifically on the chrominance channels (a,b):

$$\mathcal{L}_{lab}(\hat{Y},Y) = \left\| \hat{Y}_a - Y_a \right\|_1 + \left\| \hat{Y}_b - Y_b \right\|_1 \tag{3}$$

We explicitly exclude the luminance (L) channel to enforce color consistency independent of brightness. We set the weight $\lambda_c = 1.0$.

Perceptual Loss ($\mathcal{L}_{perc}$). To align results with human perception and preserve semantic consistency, we compute the distance between feature maps extracted from a pre-trained VGG-16 network [1]. We extract maps ϕ_j from four distinct depths: shallow layers (`relu1_2`, `relu2_2`) capturing fine textures, and deeper layers (`relu3_3`, `relu4_3`) capturing structural semantics. The loss is defined as:

$$\mathcal{L}_{perc}(\hat{Y},Y) = \sum_{j=1}^{4} w_j \cdot \left\| \phi_j(N(\hat{Y})) - \phi_j(N(Y)) \right\|_1 \tag{4}$$

where $N(\cdot)$ represents the standard ImageNet mean-variance normalization required by the pre-trained backbone. The weights are set to $w = \{1/32, 1/16, 1/8, 1/4\}$ respectively, assigning higher importance to higher-level semantic features. This component is weighted in the total loss with $\lambda_\phi = 0.2$.

Total Objective. The final objective function combines these components for the reflectance (R), shading (S), physical consistency (phy), and refinement (ref) stages:

$$\mathcal{L}_{total} = \alpha(\mathcal{L}_R + \mathcal{L}_S) + \beta \mathcal{L}_{phy} + \gamma \mathcal{L}_{ref}, \tag{5}$$

where we set $\alpha = 1, \beta = 0.5, \gamma = 1$.

3. PHYSICS-BASED LOW-LIGHT SIMULATION

In this section we extend the description of the adaptive exposure scaling and noise injection part in the low-light simulation.

Adaptive Scaling. During training, instead of a fixed gain, we determine the exposure factor κ dynamically to target specific darkness levels regardless of the initial scene brightness of the image $I_{lin} = I_{gt}^{\gamma}$. We define a target mean intensity t_{mean} sampled uniformly from $[0.0005, 0.005]$. The gain is computed as:

$$\kappa = t_{mean} / (\mu(I_{lin}) + \varepsilon) \tag{6}$$

where $\mu(\cdot)$ represents the mean operator. Figure 1 shows several low-light simulated version of the same image with different values of κ.

Noise Model. Following the signal-dependent nature of photon arrival, we model the degradation process in the linear domain as:

$$I_{noisy} = \underbrace{\kappa \cdot I_{lin}}_{Signal} + \underbrace{\mathcal{N}_s(0, \sigma_s^2) \cdot \sqrt{\kappa \cdot I_{lin}}}_{\text{Shot Noise } n_{shot}} + \underbrace{\mathcal{N}_r(0, \sigma_r^2)}_{\text{Read Noise } n_{read}} \tag{7}$$

where $\mathcal{N}$ denotes the standard Normal distribution. To ensure robustness across different noise levels, we sample $\sigma_s \in [0.001, 0.01]$ and $\sigma_r \in [0.00005, 0.0002]$.

4. ADDITIONAL VISUAL COMPARISON

In Figures 3 and 4 are shown additional results and comparisons between the proposed solution and state-of-the-art methods on LLVIP and V-TIEE dataset. Note that for LLVIP images we used the test set of 70 samples that Jha et al. [2]

made available from their official GitHub page [1]. Regarding V-TIEE images, due to the lack of official documentation detailing the pairing of low-light and well-lit images (ground truths), we performed a manual selection and pairing between the different low-exposure and high-exposure versions within the dataset. This naturally led to the formation of a custom test set. Furthermore, since some well-lit images intended to serve as ground truth presented clearly perceptible artifacts (e.g., under-exposure, noise, incorrect white balance), we also selected images from the dataset that had balanced and clean ground truths, reducing the full set to a subset of 28. We report in Figure 2 the 28 V-TIEE normal light images used in the test as ground truths. While this filtering was strictly intended to avoid evaluating on corrupted targets, we acknowledge that manually selecting a subset may induce bias, and we explicitly highlight these artifact issues as a dataset limitation.

5. REFERENCES

[1] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[2] Raman Jha, Adithya Lenka, Mani Ramanagopal, Aswin Sankaranarayanan, and Kaushik Mitra, "Rt-x net: Rgb-thermal cross attention network for low-light image enhancement," 2025.

[3] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang, "Enlightengan: Deep light enhancement without paired supervision," 2021.

[4] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex C Kot, "Low-light image enhancement with normalizing flow," *arXiv preprint arXiv:2109.05923*, 2021.

[5] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang, "Retinexformer: One-stage retinex-based transformer for low-light image enhancement," 2023.

[1] https://github.com/jhakrraman/rt-xnet

Fig. 2. The 28 V-TIEE images manually selected and used for testing.

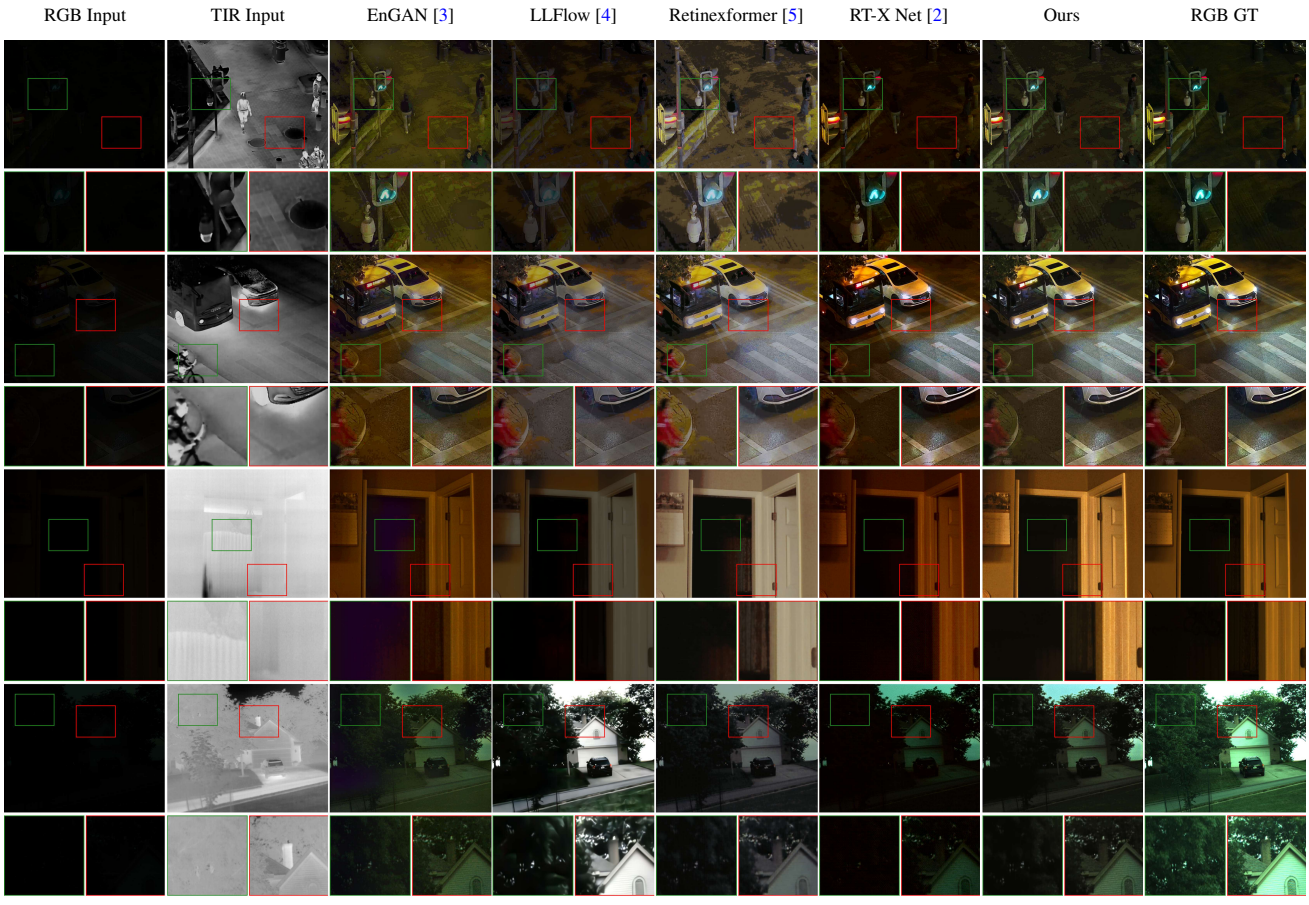

Fig. 3. Visual comparison example. The first two rows show low-light samples from the LLVIP dataset, while the last two rows are from V-TIEE. Each column compares the analyzed method against the ground truth.

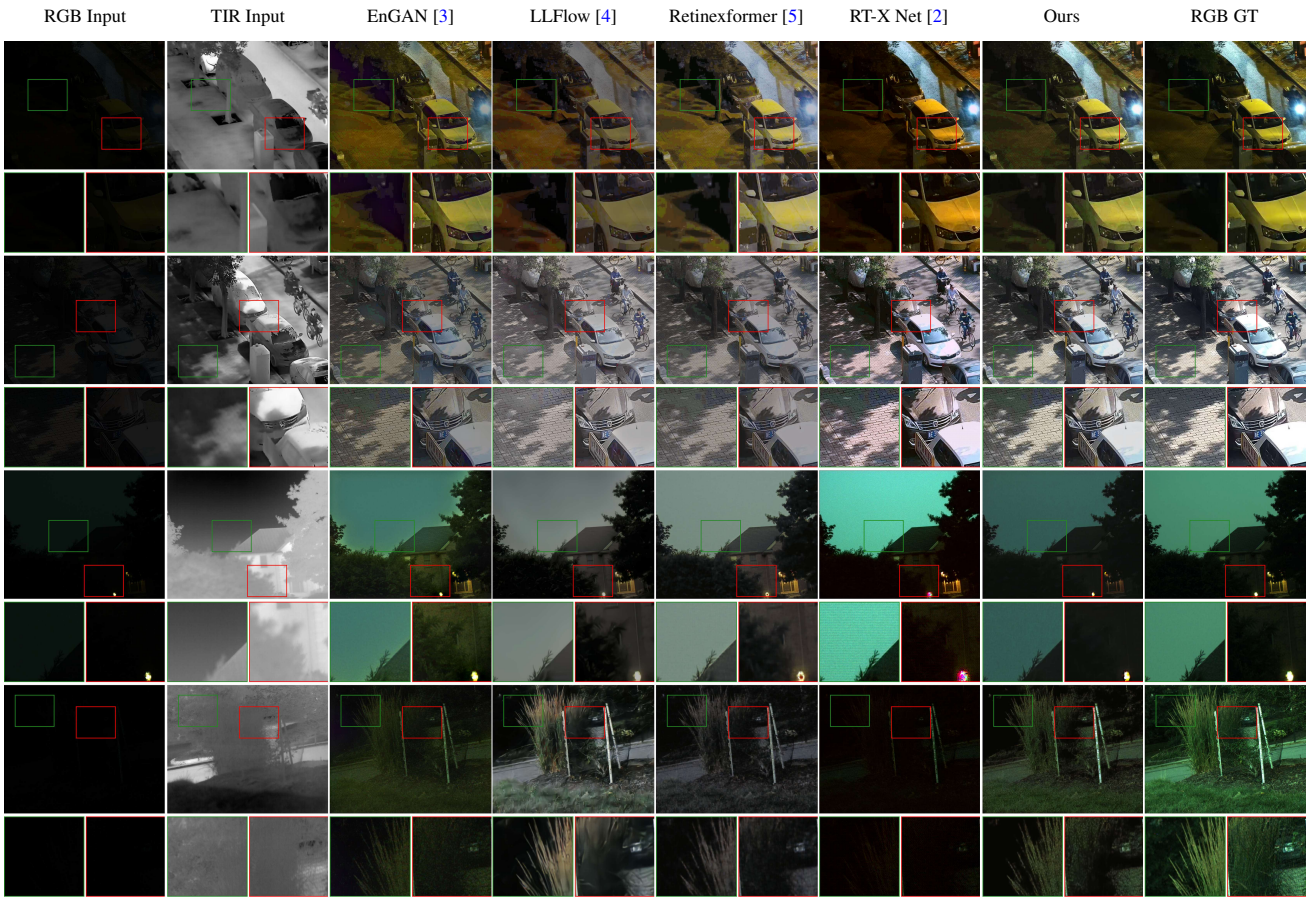

Fig. 4. Visual comparison example. The first two rows show low-light samples from the LLVIP dataset, while the last two rows are from V-TIEE. Each column compares the analyzed method against the ground truth.