CDL-FusionNet: A Two-Stage Illumination-Invariant and Luminance-Guided Framework for Infrared and Visible Image Fusion

Linyang Wang, Lei Wang*, Hengyang Wang*
School of Computer and Big Data (School of Cyber Security), Heilongjiang University, Harbin, China
2232661@s.hlju.edu.cn, wanglei@hlju.edu.cn, 2025002@hlju.edu.cn
*Corresponding authors: Lei Wang and Hengyang Wang

***Abstract*—Infrared and visible image fusion (IVIF) plays a pivotal role in enhancing scene perception under challenging conditions such as nighttime surveillance and autonomous driving. However, existing deep learning-based approaches often suffer from significant performance degradation in low-light environments. Specifically, direct fusion methods tend to introduce noise and artifacts from visible images, while the prevalent decompose-then-fuse paradigm struggles with imperfect decoupling of illumination and reflectance information. This frequently results in inadequate exploitation of illumination cues, leading to fused images with degraded textures and attenuated target features. To address these limitations, this paper proposes a two-stage fusion framework termed CDL-FusionNet. In the first stage, we develop an illumination-invariant decoupling network incorporating a self-supervised mechanism that enforces output consistency under varying lighting conditions. This enables the extraction of a physically consistent reflectance component and effectively suppresses interference from noise and artifacts. In the second stage, a brightness-guided complementary-enhanced cross-attention module is introduced. This module leverages the illumination map from the first stage as a pixel-level prior to dynamically modulate fusion weights, thereby preserving fine visible details in well-lit regions while enhancing infrared thermal features in dark areas. Extensive qualitative and quantitative evaluations on multiple public datasets demonstrate the superiority of CDL-FusionNet over state-of-the-art methods, particularly in low-light conditions. The proposed framework yields fused images with superior visual quality, higher contrast, and enhanced detail preservation, validating its robustness and effectiveness for dual-modal information integration.**

***Index Terms*—Infrared and visible image fusion, low-light enhancement, Retinex theory, illumination-invariant learning, contrastive learning, cross-attention.**

I. Introduction

Infrared and Visible Image Fusion (IVIF) aims to integrate thermal radiation information from infrared images with the rich texture details of visible images, and plays an important role in autonomous driving, video surveillance, and all-weather perception [1]. Although recent deep learning-based IVIF methods have achieved remarkable progress [2], [3], their performance often degrades severely under extremely low-light conditions [4]. This limitation stems from a fundamental dilemma in existing fusion frameworks.

Direct fusion in low-light environments is challenging because the low signal-to-noise ratio of visible images causes structural textures to be heavily entangled with noise and artifacts, thereby contaminating the fusion results [5]. To alleviate this problem, many studies [6], [7] have adopted a "decompose-then-fuse" strategy [8] inspired by Retinex theory. Early explicit decomposition methods relied on fixed mathematical models or independent subnetworks, but their limited flexibility often introduced decomposition bias [9]. By contrast, implicit decomposition methods perform end-to-end learning for adaptive separation [6], [10], yet they still suffer from two drawbacks: conventional reconstruction and smoothness constraints are insufficient for separating high-frequency textures from artifacts, and the estimated illumination map is usually discarded despite containing valuable physical information for fusion [8].

To address these issues, we propose CDL-FusionNet, a unified two-stage framework. First, we develop an Illumination-Invariant Decoupling Network (IID-Net), which introduces an illumination-invariance self-supervised mechanism into decomposition. Motivated by the observation that structural textures remain relatively stable under illumination changes whereas artifacts are transient, and inspired by contrastive learning [11], this design encourages consistent reflectance representations under varying lighting conditions. By incorporating the InfoNCE loss into a physically constrained decomposition framework [12], [13], IID-Net improves reflectance purity and robustness.

Second, we propose a Brightness-Guided Complementary-Enhanced Cross-Attention (LB-CECA) module to better exploit illumination information. Specifically, the illumination map is used as a pixel-level dynamic prior to guide the fusion process, enhancing visible reflectance details in bright regions while strengthening infrared thermal responses in dark areas.

The contributions of this work are as follows:

1) **Illumination-Invariant Decoupling Network (IID-Net)**: We introduce an illumination-invariance self-supervised constraint into a Retinex-inspired decomposition framework. This design improves the stability of reflectance representations under varying illumination conditions, leading to more robust and physically consistent decomposition.
2) **Brightness-Guided Cross-Attention (LB-CECA)**: We propose a brightness-guided complementary-enhanced cross-attention module that explicitly incorporates illu-

mination maps as pixel-level priors to guide adaptive fusion.

3) **Unified CDL-FusionNet Framework**: We develop a two-stage "decouple-then-guide" framework that effectively integrates decomposition and fusion processes, achieving improved robustness and task adaptability in low-light scenarios.

II. Related Works

A. Retinex-Based Decomposition Strategies

Retinex theory [8], which models an image as the product of reflectance and illumination, provides an important physical foundation for low-light enhancement and fusion. With the development of deep learning [10], Retinex-based methods have evolved mainly through improved decomposition strategies and loss designs. Early methods typically relied on explicit decomposition using fixed mathematical models or independent subnetworks. For example, Retinex-Net [6] combined reconstruction and illumination smoothness constraints, while KinD++ [14] further improved stability through reflectance consistency and illumination regularization. However, these methods still rely heavily on reconstruction and smoothness terms, which limits their ability to handle complex degradations.

Recent studies have introduced stronger structural constraints and improved generalization. For instance, CMRetinexNet [15] employs consistency supervision under varying illumination conditions. Although Retinex decomposition has also been introduced into fusion tasks, such as L2Fusion [16], DRF [17], and ILLVF [18], these methods still suffer from insufficient coupling between decomposition and fusion or from residual artifacts in reflectance. Therefore, more effective mechanisms are still needed to jointly ensure reflectance–illumination separation and illumination-aware fusion.

B. Locally Adaptive Fusion Strategies

In IVIF, adaptive allocation of modal contributions is critical to fusion quality [19]. Existing fusion strategies have evolved from globally fixed rules to content-adaptive mechanisms. Early deep learning-based methods, such as DenseFuse [20], generally used direct addition or concatenation and thus lacked content awareness. Later methods introduced adaptive fusion, which can be roughly divided into global and local strategies.

Globally adaptive strategies, represented by DDcGAN [5], estimate channel-level importance according to overall scene content. Although more flexible than fixed fusion rules, such strategies remain spatially invariant and cannot effectively handle localized illumination variations. To improve pixel-level refinement, locally adaptive fusion strategies have become increasingly important. One line of work explicitly learns fusion weight maps, as in IFCNN [2], PMGI [7], DDcGAN [5], PIAFusion [21], and TarDAL [22]. Another line exploits attention mechanisms, including Transformer-based models such as DATFuse [23] and CrossFuse [24]. Although these methods achieve strong performance, most of them still rely mainly on implicit feature modeling and lack explicit physical guidance, which limits their interpretability and robustness under unseen low-light conditions.

III. Method

A. Framework Overview

To address the dual challenges of inadequate decoupling and insufficient information utilization in low-light IVIF, we propose CDL-FusionNet, a two-stage fusion framework following a "Decouple-then-Guide" paradigm. As shown in Fig. 1, the framework first disentangles reliable visible information from degraded inputs, and then uses illumination priors to guide adaptive cross-modal fusion.

In the first stage, an Illumination-Invariant Decoupling Network (IID-Net) decomposes the low-light visible image into a reflectance map R_{vis} and an illumination map L_{vis} using a shared encoder and two asymmetric branches. Here, R_{vis} preserves intrinsic visible structures, while L_{vis} serves as a physical prior. A self-supervised illumination-invariance mechanism is introduced to improve the robustness of reflectance–illumination separation. The extracted reflectance R_{vis} and the original infrared image I_{ir} are then encoded into multi-scale features F_{rvis} and F_{ir} for fusion.

In the second stage, a Brightness-Guided Fusion Network integrates F_{rvis}, F_{ir}, and L_{vis}. A Complementary-Enhanced Cross-Attention (CECA) module uses the illumination prior to guide pixel-wise adaptive fusion, and the fused features are finally reconstructed into the output image. The core intuition of this design is that low-light degradation should be mitigated before cross-modal interaction. By first reducing illumination-dependent corruption and then explicitly reusing the estimated illumination map during fusion, CDL-FusionNet reduces the risk of propagating visible artifacts into the fused result and improves the interpretability of fusion decisions.

B. Illumination-Invariant Decomposition Network (IID-Net)

Conventional implicit decomposition methods mainly rely on "reconstruction + smoothness" constraints, which are often insufficient for separating high-frequency structures from unstructured artifacts. To address this limitation, we introduce an Illumination-Invariance Constraint Mechanism within an Illumination-Invariant Decoupling Network (IID-Net). The key observation is that reflectance remains relatively stable under illumination changes, whereas illumination and noise artifacts are transient. Based on this prior, IID-Net employs self-supervised learning to encourage the network to capture illumination-invariant structural features and thus purify the reflectance component.

In low-light conditions, visible observations are jointly influenced by intrinsic reflectance and illumination variation. Without explicit constraints, illumination fluctuations may leak into the reflectance branch and degrade structural representations. By introducing an illumination-invariant contrastive constraint, reflectance features extracted from the same scene under different illumination conditions are encouraged to remain consistent in the embedding space, thereby suppressing illumination-induced variance and improving reflectance purity.

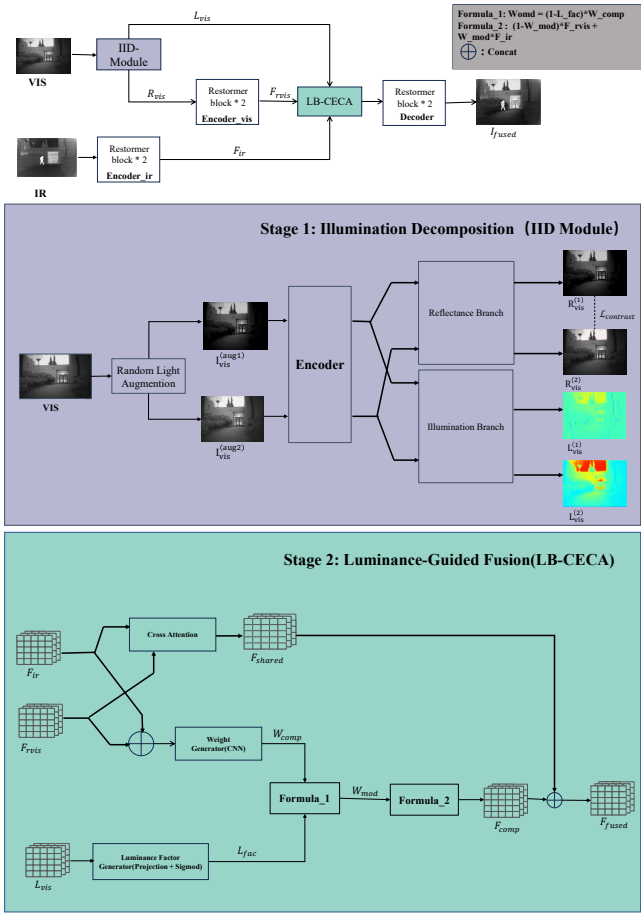

Fig. 1. Overall architecture of the proposed CDL-FusionNet. The network consists of two stages: (1) an Illumination-Invariant Decomposition (IID) Module that separates the visible image into reflectance and illumination maps; (2) a Luminance-Guided Fusion stage (LB-CECA) that integrates modality-specific features from infrared and visible domains.

Training Strategy: For each visible input, we generate two illumination-augmented views, I_{vis}^{aug1} and I_{vis}^{aug2}. These views are processed by a shared encoder and two decoder branches to produce reflectance maps R_{vis}^1 and R_{vis}^2 and illumination maps L_{vis}^1 and L_{vis}^2. This dual-view strategy enables the network to learn reflectance representations that remain stable under varying lighting conditions. Compared with reconstruction-only supervision, such a design explicitly encourages the model to preserve scene-consistent structures under illumination perturbations, which is critical for suppressing illumination leakage into the reflectance branch.

The total loss of IID-Net is

$$\mathcal{L}_{\text{stage1}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{tv}}\mathcal{L}_{\text{tv}} + \lambda_{\text{contrast}}\mathcal{L}_{\text{contrast}}, \tag{1}$$

where the reconstruction loss $\mathcal{L}_{\text{rec}}$ ensures decomposition fidelity [17], the total variation loss $\mathcal{L}_{\text{tv}}$ enforces smooth illumination [25], and the contrastive loss $\mathcal{L}_{\text{contrast}}$ imposes structural invariance. By pulling together reflectance representations of the same image under different lighting conditions while separating those of different images, this loss explicitly suppresses illumination-dependent artifacts.

Formally,

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(\text{sim}(f(R_{vis}^{aug1}), f(R_{vis}^{aug2}))/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{[k\neq i]} \exp(\text{sim}(f(R_{vis}^{i}), f(R_{vis}^{k}))/\tau)}, \tag{2}$$

where $f(\cdot)$ denotes a non-linear projection head, $\text{sim}(\cdot,\cdot)$ represents cosine similarity, and τ is the temperature hyperparameter. For a batch size of $N = 8$, we obtain $2N = 16$ augmented views, where the paired view from the same source image serves as the positive sample and the remaining views act as negatives. The projection head consists of two layers with ReLU activation and maps features into a 128-dimensional latent space.

C. Brightness-Guided Complementary-Enhanced Cross-Attention (LB-CECA) Module

To improve illumination utilization and provide explicit physical guidance for fusion, we propose the LB-CECA module. It takes reflectance features F_{rvis}, infrared features F_{ir}, and the illumination map L_{vis} as input, and outputs adaptively fused features.

1) Shared Feature Extraction Branch: To capture structural information common to both modalities, we employ a cross-attention mechanism [24]. The visible features are used as Query, while the infrared features serve as Key and Value, yielding the shared representation F_{shared}.

2) Dynamic Weight Generation Branch: This branch predicts adaptive fusion weights W_{mod} at the pixel level. First, F_{rvis} and F_{ir} are concatenated and processed by a lightweight CNN-based generator to estimate an initial complementarity weight map. Meanwhile, the illumination map L_{vis} is passed through a projection network and a Sigmoid activation to obtain a normalized brightness factor L_{fac}. The final weight map follows a physically intuitive rule: visible details are emphasized in bright regions, whereas infrared responses dominate in dark regions.

3) Feature Fusion: The complementary feature is first obtained by dynamic weighted fusion:

$$F_{comp} = W_{mod} \cdot F_{ir} + (1 - W_{mod}) \cdot F_{rvis}. \tag{3}$$

The final fused feature is then computed as

$$F_{fused} = F_{comp} + F_{shared}. \tag{4}$$

This design preserves both adaptively selected complementary information and shared cross-modal structures. In this way, the shared branch preserves modality-consistent structural information, while the dynamic-weight branch emphasizes spatially complementary responses under illumination guidance. Their combination enables the fused representation to remain both structurally stable and locally adaptive.

The total loss for the fusion stage is

$$\begin{aligned}\mathcal{L}_{stage2} = \lambda_{int}\mathcal{L}_{int} + \lambda_{grad}\mathcal{L}_{grad} + \lambda_{ssim}\mathcal{L}_{ssim} \\ +\lambda_{prec}\mathcal{L}_{prec} + \lambda_{decomp}\mathcal{L}_{decomp} + \lambda\mathcal{L}_{contrast}^{final}.\end{aligned} \tag{5}$$

This objective combines six constraints from complementary perspectives. Intensity and gradient losses $\mathcal{L}_{int}$ and $\mathcal{L}_{grad}$

preserve content and edge consistency [26]. SSIM loss $\mathcal{L}_{ssim}$ and perceptual loss $\mathcal{L}_{prec}$ improve structural fidelity and naturalness [27], [28].

For feature decomposition, given an encoder feature $F \in \mathbb{R}^{C\times H\times W}$, two lightweight projection heads generate base and detail representations:

$$F_{base} = \phi_b(F), \quad F_{detail} = \phi_d(F), \tag{6}$$

where $\phi_b(\cdot)$ and $\phi_d(\cdot)$ are implemented as 1×1 convolutional layers followed by nonlinear activation. The base feature captures low-frequency shared structure, while the detail feature encodes high-frequency modality-specific textures. The feature decomposition loss is defined as

$$\mathcal{L}_{decomp} = (1 - cc(F^{rvis}_{base}, F^{ir}_{base})) + (cc(F^{rvis}_{detail}, F^{ir}_{detail}))^2, \tag{7}$$

where $cc(\cdot,\cdot)$ denotes the Pearson correlation coefficient.

To further enhance image contrast, we introduce

$$\mathcal{L}^{final}_{contrast} = \text{ReLU}\left(\max(\text{SD}(I_{vis}), \text{SD}(I_{ir})) - \text{SD}(F_{fused})\right), \tag{8}$$

where $\text{SD}(\cdot)$ denotes standard deviation.

IV. Experiments

A. Experimental Settings

We evaluate CDL-FusionNet on two representative public datasets: TNO [29], which contains 361 image pairs from diverse nighttime scenarios, and RoadScene [3], which consists of 50 image pairs focusing on traffic targets. Five standard metrics are adopted for quantitative evaluation: Entropy (EN), Mutual Information (MI), Sum of Correlations of Differences (SCD), Visual Information Fidelity (VIFF), and Structural Similarity (SSIM).

We compare our method with seven representative SOTA baselines: FusionGAN [30], IFCNN [2], U2Fusion [3], TarDAL [22], SemLA [31], DATFuse [23], and CrossFuse [24]. All baselines are tested using official codes and released pre-trained weights under a unified environment. Our model is implemented in PyTorch and trained on a single NVIDIA RTX 4090 GPU using Adam with an initial learning rate of 1×10^{-4} and a batch size of 8. The two stages are trained sequentially for 20 epochs each. We observed stable convergence within this schedule, and further training brought only marginal gains. All experiments are conducted with fixed random seeds. Metrics such as VIFF and SSIM are bounded within $[0, 1]$, while EN and MI are unbounded statistical measures.

B. Comparison with State-of-the-Art Methods

1) Quantitative Results: As shown in Table I, CDL-FusionNet achieves the best performance on most metrics across both datasets, indicating strong fusion capability and good generalization. On the TNO dataset, our method ranks first in EN, MI, SCD, and VIFF, suggesting effective information preservation, complementary utilization, and perceptual quality enhancement. In particular, the improvements in MI and SCD indicate that the proposed framework can better retain complementary information from both modalities while preserving structural consistency. Although CrossFuse [24] slightly surpasses our method on SSIM, our model provides a better overall balance between detail preservation and artifact suppression, as reflected by its clear advantage in VIFF and SCD. This result suggests that CDL-FusionNet is more effective in enhancing perceptually important details while maintaining stable cross-modal fusion quality.

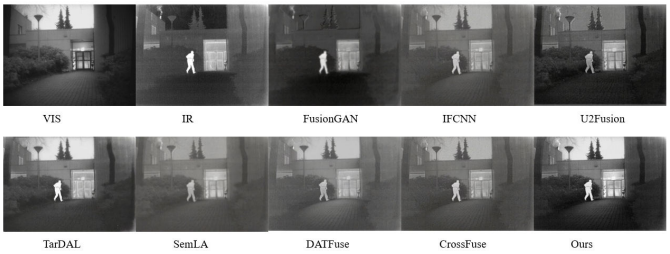

Fig. 2. The fusion results produced by compared methods and the proposed method on TNO.

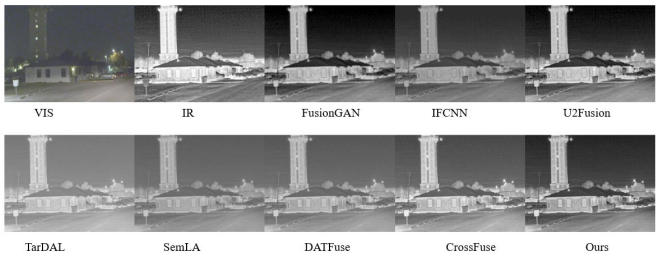

Fig. 3. The fusion results produced by compared methods and the proposed method on RoadScene.

On the RoadScene dataset, our method again achieves the best EN, MI, SCD, and VIFF while remaining competitive on SSIM. The clear advantage in VIFF confirms the robustness of CDL-FusionNet in producing visually faithful results in complex urban scenes. Meanwhile, the superior MI and SCD values indicate that the proposed method can better exploit complementary modal information under challenging lighting conditions. This performance can be attributed to the proposed two-stage design, where IID-Net provides cleaner reflectance representations for subsequent adaptive fusion, and LB-CECA further guides the fusion process with illumination priors. As a result, the model is able to preserve useful visible structures in well-lit regions while enhancing infrared target responses in darker areas.

It should be noted that EN, although widely used in image fusion evaluation, does not always correlate positively with perceptual quality in low-light scenarios, since amplified noise may also increase entropy. Therefore, EN should be interpreted jointly with VIFF, SCD, and qualitative results rather than in isolation. In our experiments, the consistent superiority of CDL-FusionNet on VIFF and SCD, together with its strong performance on MI, provides more convincing evidence that the proposed method improves both visual fidelity and cross-modal information integration, rather than merely increasing statistical complexity.

TABLE I

AVERAGE METRIC VALUES OF EXISTING FUSION METHODS AND THE PROPOSED CDL-FUSIONNET ON THE TNO AND ROADSCENE DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Model	TNO Dataset					RoadScene Dataset				
	EN↑	MI↑	SCD↑	VIFF↑	SSIM↑	EN↑	MI↑	SCD↑	VIFF↑	SSIM↑
FusionGAN [30]	6.35	12.72	1.45	0.49	0.47	6.52	12.98	1.55	0.54	0.55
IFCNN [2]	6.59	13.19	1.71	0.63	0.65	6.84	13.40	1.50	0.62	0.59
U2Fusion [3]	6.75	13.51	1.79	0.68	0.72	6.98	13.79	1.58	0.68	0.62
TarDAL [22]	6.04	12.78	1.31	0.57	0.60	6.54	12.84	1.61	0.59	0.61
SemLA [31]	6.52	13.04	1.53	0.74	0.68	6.78	13.28	1.63	0.72	0.63
DATFuse [23]	6.32	12.64	1.52	0.86	0.69	6.57	12.82	1.68	0.83	0.69
CrossFuse [24]	6.83	13.67	1.76	0.77	**0.73**	6.96	13.78	1.61	0.74	**0.78**
Ours	**7.04**	**13.86**	**1.81**	**0.89**	0.71	**7.12**	**13.96**	**1.75**	**0.86**	0.73

2) Qualitative Comparison: Fig. 2 and Fig. 3 present representative qualitative comparisons. On TNO, several baselines exhibit infrared bias, blurred backgrounds, or visible artifacts. In contrast, our method preserves salient pedestrian targets while maintaining clearer background structures, which is consistent with the improvements in SCD and VIFF. More importantly, the target contours generated by CDL-FusionNet are more distinct, while the surrounding scene details remain relatively natural, indicating that the proposed framework can better balance target saliency and structural continuity.

On RoadScene, some competing methods suffer from edge discontinuity, over-smoothing, or insufficient detail hierarchy. Our method better restores structural details such as local textures and shadow transitions of the lighthouse, while avoiding obvious halo artifacts. In addition, the transitions between bright and dark regions appear more natural in our fused results, showing that the illumination-guided design helps reduce over-enhancement and local distortion. These observations agree well with the quantitative results and further demonstrate the effectiveness of IID-Net and LB-CECA under complex lighting conditions.

C. Ablation Studies

To validate the contribution of each component, we conduct ablation studies on both datasets using the complete model as the baseline.

w/o IID-Net: Removing the decomposition stage leads to clear declines in SCD, VIFF, and SSIM, although EN becomes higher. This suggests that raw low-light visible inputs introduce noise and illumination artifacts into fusion, which may be mistakenly counted as information by the entropy metric. The degraded SCD, VIFF, and SSIM indicate that structural fidelity and perceptual quality are weakened. This confirms that IID-Net plays an important role in purifying visible structural inputs and providing a cleaner basis for subsequent fusion.

w/o LB-CECA: Replacing LB-CECA with naive element-wise summation degrades MI, SCD, VIFF, and SSIM. This demonstrates that static feature aggregation is insufficient for handling spatially varying illumination conditions. In contrast, LB-CECA uses the illumination map as an explicit pixel-level prior to adaptively regulate modal contributions, allowing visible details to dominate in bright regions while strengthening infrared responses in dark regions. The improved results verify the effectiveness of this physically guided adaptive fusion mechanism.

w/o $\mathcal{L}_{contrast}$: Removing the illumination-invariant contrastive loss yields relatively high EN but lower VIFF and SSIM. This mirrors the phenomenon observed in the "w/o IID-Net" setting: without the invariance constraint, residual illumination artifacts remain in the reflectance branch and may be misinterpreted by EN as additional information. However, the drop in VIFF and SSIM reveals that these artifacts reduce perceptual fidelity and structural stability. This confirms that the proposed contrastive constraint helps suppress illumination-induced variance and improves reflectance purity.

Overall, the ablation results confirm that IID-Net, illumination-invariant contrastive learning, and LB-CECA all contribute positively to the final fusion performance.

V. CONCLUSION

In this paper, we present CDL-FusionNet, a novel infrared and visible image fusion framework specifically tailored for low-light scenarios. Adopting a "Decouple-then-Guide" two-stage paradigm, the proposed method systematically addresses the performance degradation observed in existing approaches, which stems from inadequate decoupling and insufficient information utilization under low-light conditions. In the

TABLE II

OBJECTIVE RESULTS OF ABLATION STUDIES ON THE TNO AND ROADSCENE DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Model	TNO Dataset					RoadScene Dataset				
	EN↑	MI↑	SCD↑	VIFF↑	SSIM↑	EN↑	MI↑	SCD↑	VIFF↑	SSIM↑
w/o IID-Net	7.23	11.66	0.79	0.74	0.63	6.95	12.50	0.85	0.69	0.65
w/o LB-CECA	7.01	11.98	1.58	0.79	0.67	6.98	13.15	1.55	0.82	0.68
w/o $\mathcal{L}_{contrast}$	**7.34**	12.06	1.74	0.76	0.68	**7.18**	13.30	1.72	0.81	0.69
Ours	7.04	**13.86**	**1.81**	**0.89**	**0.71**	7.12	**13.96**	**1.75**	**0.86**	**0.73**

first stage, the Illumination-Invariant Decoupling Network (IID-Net) incorporates a self-supervised mechanism based on illumination-invariant contrastive learning. This design effectively extracts a pure and physically consistent reflectance component, thereby significantly suppressing interference from noise and artifacts. In the second stage, the Brightness-Guided Complementary-Enhanced Cross-Attention (LB-CECA) module leverages the illumination map as a pixel-level dynamic weight regulator. This facilitates the spatially adaptive fusion of infrared and visible features across regions with varying illumination intensities. Extensive experimental results on multiple public datasets demonstrate that our method outperforms existing state-of-the-art approaches in key evaluation metrics, including VIFF, SCD, and MI. Furthermore, it exhibits superior visual performance in terms of structural restoration and perceptual quality. Finally, ablation studies validate the effectiveness and necessity of each core module within the framework.

ACKNOWLEDGMENTS

This work was supported by the Research Funds for Universities of Heilongjiang Province under Grant No. 2025-KYYWF-ZR0426 and the Guidance Project for Key Research and Development Plan of Heilongjiang Province under Grant No. GZ20230015.

REFERENCES

[1] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, 2019.

[2] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," *Inf. Fusion*, vol. 54, pp. 99–118, 2020.

[3] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, 2022.

[4] E. Wang, J. Li, J. Lei, and S. Zhou, "Deep learning-based infrared and visible image fusion: A survey," *J. Softw.*, vol. 18, no. 4, pp. 899–915, 2023.

[5] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.

[6] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep Retinex decomposition for low-light enhancement," *arXiv preprint arXiv:1808.04560*, 2018.

[7] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 12797–12804, 2020.

[8] Z. Rahman, D. Jobson, and G. Woodell, "Retinex processing for automatic image enhancement," *J. Electron. Imaging*, vol. 13, no. 1, pp. 100–110, 2004.

[9] X. Ren, W. Yang, W.-H. Cheng, and J. Liu, "LR3M: Robust low-light enhancement via low-rank regularized Retinex model," *IEEE Trans. Image Process.*, vol. 29, pp. 5862–5876, 2020.

[10] R. Liu, L. Ma, J. Zhang, X. Fan, and Z. Luo, "Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 10556–10565, 2021.

[11] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 1597–1607, 2020.

[13] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 15750–15758, 2021.

[14] Y. Zhang, X. Guo, J. Ma, W. Liu, and J. Zhang, "Beyond brightening low-light images," *Int. J. Comput. Vis.*, vol. 129, pp. 1013–1037, 2021.

[15] T. Zhang, P. Liu, M. Cai, X. Wang, and Q. Zhou, "Cross-modal guided and refinement-enhanced Retinex network for robust low-light image enhancement," *Inf. Fusion*, vol. 124, p. 103380, 2025.

[16] X. Gao, G. Lv, A. Dong, Z. Wei, and J. Cheng, "L2fusion: Low-light oriented infrared and visible image fusion," in *Proc. IEEE Int. Conf. Image Process.*, pp. 2405–2409, 2023.

[17] Y. Gu *et al.*, "Physics driven deep Retinex fusion for adaptive infrared and visible image fusion," *arXiv preprint arXiv:2112.02869*, 2023.

[18] Y. Xie, X. Fan, C. Lin, Z. Xue, and B. Wang, "ILLVFusion: Infrared and low-light visible image fusion based on CNN and transformer," *Opt. Lasers Eng.*, vol. 195, p. 109267, 2025.

[19] E. Wang, J. Li, J. Lei, and S. Zhou, "Deep learning-based infrared and visible image fusion: A survey," *J. Comput. Sci. Explor.*, vol. 18, no. 4, pp. 899–915, 2024.

[20] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, 2019.

[21] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "PIAFusion: A progressive infrared and visible image fusion network based on illumination aware," *Inf. Fusion*, vol. 83–84, pp. 79–92, 2022.

[22] J. Liu *et al.*, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 5792–5801, 2022.

[23] W. Tang, F. He, Y. Liu, Y. Duan, and T. Si, "DATFuse: Infrared and visible image fusion via dual attention transformer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 7, pp. 3159–3172, 2023.

[24] H. Li and X.-J. Wu, "CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach," *Inf. Fusion*, vol. 103, p. 102147, 2024.

[25] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, vol. 60, no. 1–4, pp. 259–268, 1992.

[26] J. Ma *et al.*, "Infrared and visible image fusion via detail preserving adversarial learning," *Inf. Fusion*, vol. 54, pp. 85–98, 2020.

[27] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[28] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, pp. 694–711, 2016.

[29] A. Toet, "TNO image fusion dataset," Figshare, 2014.

[30] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, 2019.

[31] H. Xie *et al.*, "Semantics lead all: Towards unified image registration and fusion from a semantic perspective," *Inf. Fusion*, vol. 98, p. 101835, 2023.