

Article

Not peer-reviewed version

HDCGAN-Based UAV Remote Sensing Image Enhancement and Evaluation Method with WPID

[Kelly Chen Ke](#), [Min Sun](#)^{*}, [Xinyi Wang](#), [Dong Liu](#), Hanjun Yang

Posted Date: 3 February 2026

doi: 10.20944/preprints202602.0145.v1

Keywords: UAV low-light images; TIR images; deep learning; low-rank image fusion; image quality assessment; weakly paired image dataset(WPID); SAM semantic segmentation



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

HDCGAN-Based UAV Remote Sensing Image Enhancement and Evaluation Method with WPID

Kelly Chen Ke, Min Sun *, Xinyi Wang, Dong Liu and Hanjun Yang

Institute of Remote Sensing and Geographic Information Systems, Peking University, 5 Summer Palace Road, Beijing 100871, China

* Correspondence: sunmin@pku.edu.cn

Highlights

What are the main findings?

- A novel low-light UAV remote sensing image enhancement method is proposed, based on a weakly paired dataset constructed for HDCGAN, improving image quality.
- A new structural consistency evaluation framework is developed, utilizing the Segment Anything Model (SAM) for objective assessment of enhancement performance.

What are the implications of the main findings?

- The proposed method enhances the usability of low-light UAV images in public security and disaster emergency applications.
- The new evaluation system provides a robust tool for assessing and comparing low-light enhancement techniques, advancing the field of remote sensing image quality evaluation.

Abstract

Remote sensing images acquired by UAVs under nighttime or low-light conditions suffer from insufficient illumination, leading to degraded image quality, detail loss, and noise, which restrict their application in public security and disaster emergency scenarios. Although existing machine learning-based enhancement methods can recover part of the missing information, they often cause color distortion and texture inconsistency. This study proposes an improved low-light image enhancement method based on a Weakly Paired Image Dataset (WPID), combining the Hierarchical Deep Convolutional Generative Adversarial Network (HDCGAN) with a low-rank image fusion strategy to enhance the quality of low-light UAV remote sensing images. First, YCbCr color channel separation is applied to preserve color information from visible images. Then, a Low-Rank Representation Fusion Network (LRRNet) is employed to perform structure-aware fusion between thermal infrared (TIR) and visible images, thereby enabling effective preservation of structural details and realistic color appearance. Furthermore, a weakly paired training mechanism is incorporated into HDCGAN to enhance detail restoration and structural fidelity. To achieve objective evaluation, a structural consistency assessment framework is constructed based on semantic segmentation results from the Segment Anything Model (SAM). Experimental results demonstrate that the proposed method outperforms state-of-the-art approaches in both visual quality and application-oriented evaluation metrics.

Keywords: UAV low-light images; TIR images; deep learning; low-rank image fusion; image quality assessment; weakly paired image dataset(WPID); SAM semantic segmentation

1. Introduction

UAV images have significant application value in various fields; however, their quality is highly dependent on the illumination conditions during image acquisition. In generally, Nighttime Color Visible (NCV) images captured under low-light environments are often unsuitable for direct

application. Nevertheless, in scenarios such as disaster response and public security, there is a strong demand for reliable scene information acquired at nighttime or under low-illumination conditions. Therefore, enhancing the quality of low-light images is of great importance for supporting subsequent processing tasks and improving the practical applicability of UAV images. In recent years, low-light (including nighttime) image enhancement has attracted increasing attention, particularly with the rapid development of machine learning-based methods, which have substantially improved both the visual quality of enhanced images and their effectiveness in downstream application tasks.

Nevertheless, under nighttime low-illumination conditions, color visible (CV) cameras suffer from substantial losses of color and texture information during imaging, resulting in information attenuation and detail degradation in the acquired images. Even with advanced machine learning methods, it remains challenging to effectively restore authentic scene information. Therefore, illumination enhancement alone cannot fundamentally resolve the quality degradation of low-light images [1].

To address this issue, many studies have introduced thermal infrared (TIR) images to compensate for the scene information lost by CV cameras. Among various approaches, the fusion of TIR and CV images has proven to be particularly effective, as it integrates the complementary characteristics of both modalities, enhances visual quality, and facilitates downstream tasks such as target detection and environmental perception.

Although TIR and CV image fusion can partially mitigate the information loss in low-light CV images, TIR images themselves lack sufficient color and texture details. Consequently, the fused results still suffer from incomplete perceptual content and constrained color representation, making them inadequate for downstream tasks requiring high semantic representation and visual interpretability. Particularly for tasks relying on abundant visual prior information (e.g., target recognition and image segmentation), the structural and appearance authenticity of images directly affects the accuracy and robustness of model predictions.

Despite the significant quality improvements achieved by TIR and low-light CV image fusion, there remains a substantial gap between the fused images and Daytime Color Visible (DCV) images. To address this issue, existing studies [2] have proposed machine learning algorithms to transform fused images into DCV images, thereby further improving image quality and better supporting downstream image analysis and processing tasks. Leveraging deep learning models for this transformation, particularly by learning the imaging characteristics under natural daytime illumination, effectively enriches the image information in terms of color, brightness, and texture. Moreover, boundary features and environmental context can be better restored, thereby enhancing the semantic interpretability of images in complex scenes.

Existing studies have demonstrated that Generative Adversarial Networks (GANs) [3] outperform other models in image enhancement and transformation [4]. However, similar to most machine learning algorithms, current GAN-based enhancement methods are highly dependent on the quality and availability of training datasets. Existing GAN models can be broadly categorized into supervised and unsupervised learning paradigms according to whether paired or unpaired datasets are used, both of which have inherent limitations:

Supervised methods, represented by LightNet [5] and pix2pix [6], rely heavily on strictly paired image datasets, where each low-quality input is matched with a corresponding high-quality reference. Such methods require precise pixel-level alignment between paired images to achieve optimal training performance. Although such methods can ensure favorable enhancement results through explicit mapping relationships, they are hardly applicable to outdoor UAV scenarios. For instance, UAVs cannot acquire strictly paired images for the same region. Even with post-processing, it is difficult to overcome interference caused by dynamic changes in ground objects (e.g., moving vehicles and pedestrians) or parallax variations during repeated UAV acquisitions of the same scene. Moreover, manual pixel-level annotation and registration of such images are extremely costly and time-consuming, making the construction of large-scale datasets impractical.

In contrast, unsupervised methods, represented by CycleGAN [7] and Retinex-GAN [8], can utilize unpaired datasets, which significantly reduces the difficulty of data acquisition. However, by completely discarding the correspondence information between input and output images during the training process, these methods often introduce severe color distortion and localized degradations, such as speckle-like noise [8].

Some improved GAN models (e.g., PearlGAN [10]) have enhanced the transformation performance from night TIR images to DCV images, however, the generated results still suffer from texture and color distortion. The fundamental reason lies in the lack of semantic consistency constraints between pre- and post-transformation images. Yang et al. proposed an HDCGAN model [2], which integrates Hyperdimensional Computing (HDC) into GANs to address semantic inconsistency in unpaired image translation. Although HDCGAN substantially improves the performance of translating fused NCV and TIR images into DCV images, the use of unpaired datasets still fails to provide adequate constraints on semantic consistency, resulting in information loss and semantic deviation in the transformed results.



Figure 1. Examples of weakly paired UAV images acquired over the same area along an identical flight route during daytime and nighttime. All data were collected in Xianghe County, Hebei Province, China, on June 30 2024 using a DJI-M300 UAV with a XT2 thermal infrared (TIR) camera featuring a built-in visible camera. (a) Daytime images; (b) Nighttime images.

In addition, accurately evaluating the quality of transformed images is a crucial task. Existing studies [11] have shown that images with higher evaluation scores do not necessarily perform better than those with lower scores in downstream tasks. Since the primary goal of this study is to meet the requirements of downstream applications, it should be noted that most existing image evaluation metrics mainly focus on perceptual quality and semantic preservation, but fail to effectively assess structural preservation, especially in terms of performance in downstream tasks such as target detection and image segmentation. To address this issue, we propose a stable and objective structure-based quality evaluation method for enhanced images to more accurately assess their effectiveness in practical applications.

Overall, the main objectives of this study are: (1) to propose the construction of a WPID to improve the transformation performance of GAN models on low-light images by introducing semantic consistency constraints; (2) to establish an evaluation method that integrates existing image quality metrics (FID [12], KID [13], LPIPS [14]) with downstream task accuracy metrics, based on the correlation characteristics of WPI.

The remainder of this paper is organized as follows. Section 2 presents a review of related work. Section 3 introduces the construction of the WPID and details the improved HDCGAN model. Section 4 presents the experimental setup and analyzes the results. Section 5 discusses the experimental results in further detail. Section 6 concludes the study.

2. Related Work

Existing studies on low-light image enhancement are broadly categorized into three types: (1) traditional low-light image enhancement methods, which primarily aim to enhance image brightness and contrast via conventional image processing techniques and physical modeling; (2) deep learning-based low-light image enhancement methods, encompassing supervised, unsupervised, and weakly supervised learning paradigms designed to accommodate diverse data acquisition scenarios; (3) methods integrating visible and TIR image fusion with enhancement. This section first reviews these three categories of methods, then discusses the advantages and limitations of existing image quality evaluation metrics, and finally identifies the key issues that this paper aims to address.

2.1. Traditional Low-Light Image Enhancement Methods

Early-stage low-light image enhancement methods primarily relied on basic linear or nonlinear transformation techniques, such as linear brightness stretching or nonlinear Gamma correction. These techniques feature simple principles and high computation efficiency; however, they lack adaptability to local image characteristics, resulting in overall limited enhancement effects. Subsequently, global and local histogram adjustment techniques emerged as the mainstream. A representative example is the Contrast-Limited Adaptive Histogram Equalization (CLAHE) method proposed by Pisanò et al. (1998) [15], which mitigates noise amplification by constraining local contrast and greatly enhances the practical utility of histogram-based techniques. Subsequent studies further optimized this framework. For instance, Lee et al. (2012) proposed a contrast enhancement method based on hierarchical difference representation [16], which enhances the preservation of local detail.

As recently as 2016, Chandrasekar et al. pointed out that histogram equalization-based methods remain an important technical branch in low-light image contrast enhancement [17]. However, their inherent limitations, including block artifacts, detail loss, and noise amplification, have not been fundamentally addressed. This is mainly because these methods primarily manipulate pixel grayscale distributions without considering the physical mechanisms of image formation, making it difficult to achieve an optimal balance between contrast enhancement and color fidelity.

Methods based on Retinex theory constitute another important category of low-light image enhancement, as they can more effectively mitigate the limitations of histogram adjustment methods. Originally proposed by Land et al. in 1971 [18], Retinex theory assumes that image irradiance consists of illumination and reflectance components, and that low-light image enhancement can be achieved by adjusting the reflectance component. However, as pointed out by Jobson et al. [19], early Retinex-based algorithms exhibited significant limitations due to the difficulty of accurately modeling the reflectance of real scenes, making it challenging to generate images with both high quality and realistic colors.

To address these limitations, Rahman et al. proposed an improved algorithm [20], which achieved good color fidelity but was limited by high computational complexity and prominent halo artifacts. To suppress halo artifacts, Palma-Amestoy et al. proposed a perception-inspired variational framework for color enhancement [21], and Fu et al. proposed a probabilistic approach based on simulated illumination and reflectance estimation for image enhancement [22]. Nevertheless, these algorithms still suffer from issues such as inadequate edge sharpening, color oversaturation, and high computational cost.

To further address these remaining issues, Guo et al. [23] proposed the Low-Light Image Enhancement via Illumination Map Estimation (LIME) method, which refines the illumination map by incorporating structural priors and achieves promising enhancement performance for low-light images. However, it still suffers from problems such as noise amplification and excessive illumination enhancement.

In summary, traditional low-light enhancement methods are inherently constrained by hand-crafted priors and distribution-driven optimization, resulting in an inevitable trade-off among

denoising, edge preservation, and color fidelity. Consequently, it is difficult to fully resolve issues such as block artifacts, halo effects, noise amplification, and color cast. In addition, their computational complexity increases rapidly with image resolution, making it difficult to satisfy the dual requirements of high quality and real-time enhancement. For these reasons, research attention has gradually shifted toward data-driven deep learning methods. By leveraging large-scale datasets for training, these networks can automatically learn the mapping from low-light to normal-light images, thereby overcoming the limitations of hand-crafted priors and achieving a better balance among contrast enhancement, detail preservation, color fidelity, and computational efficiency.

2.2. Deep Learning-Based Low-Light Image Enhancement Methods

Deep learning-based methods can be broadly categorized into supervised and unsupervised learning paradigms. As mentioned earlier, the key distinction lies in the type of training data employed: supervised learning requires strictly pixel-aligned low-light and normal-light image pairs, whereas unsupervised learning is used when such paired data are not available.

As noted in the Introduction, representative supervised learning methods include LightenNet model proposed by Li et al. (2018) for low-light image enhancement [5], which adopts a CNN backbone and is trained on a dataset constructed based on the Retinex model. Experimental results demonstrate that this model is easy to train and computationally efficient, delivering superior enhancement performance compared with traditional methods. Another representative model is Pix2Pix, proposed by Isola et al. [6], which can be applied to the color enhancement of low-light images. Its core advantage lies in its end-to-end training capability with strictly paired images, the U-Net architecture ensures pixel-level alignment, while the PatchGAN discriminator preserves texture fidelity. Consequently, Pix2Pix can generate geometrically accurate transformation results without relying on hand-crafted loss functions. However, Pix2Pix relies heavily on paired image datasets, and its performance is strongly constrained by the quality and scale of the available paired datasets.

Given the strict requirements for dataset construction in supervised enhancement methods, many researchers have sought to train neural networks using unpaired datasets. Among these approaches, Generative Adversarial Networks (GANs) [3] constitute a typical category of unsupervised learning models. They do not rely on explicit labels or strictly paired samples; instead, high-quality image enhancement and transformation can be achieved by designing appropriate loss functions and structural constraints.

Although GAN-based models can generate promising enhancement results, their training process is often unstable, and the termination point of training usually needs to be determined empirically. To improve detail restoration and global consistency during image enhancement and transformation, various strategies have been introduced, such as perceptual loss, structure-preservation loss, and attention mechanisms. Qu et al. [24] first applied CycleGAN [7] to low-light image enhancement. This model employs a cycle-consistency loss to enforce similarity between the transformed and original images. However, it still suffers from issues such as halo artifacts, detail loss, and training instability, mainly due to the cycle-consistency constraint and the use of dual discriminators.

To alleviate detail degradation and color distortion in low-illumination images, Mao et al. proposed the Retinex-GAN model [9], which integrates classical Retinex image decomposition theory with GAN mechanisms. It adopts an unsupervised framework for low-light image enhancement without relying on large-scale paired data and is capable of adapting to complex scenes. Nevertheless, the enhanced results are prone to halo artifacts near light source edges, and sensitive to noise in extremely dark regions, the overall stability of the model still requires further improvement.

Although unsupervised methods typified by GANs, CycleGAN, and Retinex-GAN can achieve satisfactory image enhancement performance with unpaired datasets, the transformed or enhanced images often exhibit undesirable artifacts, such as local noise spots, edge distortions, and abrupt

changes in color and semantics. The fundamental reason is that unpaired datasets lack explicit structural and semantic correspondences between image pairs, leading to insufficient effective constrain compared with paired datasets. For instance, the improved PSGAN model [25] tends to produce artifacts when enhancing remote sensing images with complex ground object scenes, making it difficult to preserve the structural consistency of ground objects.

Given the limitations of both strictly paired and unpaired datasets, this study introduces a WPID to train the GAN model. For UAV remote sensing image enhancement, WPI can be efficiently constructed using multi-temporal UAV images of the same area. This moderate level of correlation enables the model to learn the essential mapping relationships for image enhancement without being constrained by the difficulty of acquiring strictly paired data. Consequently, it achieves a favorable balance between data acquisition cost and enhancement performance, and lays a solid foundation for the application of image enhancement techniques in more complex real-world scenarios.

2.3. Fusion and Transformation Methods of Visible and TIR Images

As noted in the introduction, although deep learning methods can effectively improve the visual quality of low-light images, low-light imaging is inherently prone to significant detail loss. Such loss cannot be fully remedied by a single enhancement algorithm or deep learning model, making it difficult to achieve ideal information recovery and supplementation. To address this issue, fusing NTIR (Nighttime thermal infrared) images to enhance CV image quality has emerged as a mainstream solution, and early representative studies in this field have achieved notable progress. For instance, Pei et al. proposed an adaptive multi-scale fusion method for infrared and CV images [26]. However, TIR images are acquired based on temperature differences and exhibit inherent limitations in color reproduction and fine-detail representation compared with CV images. Consequently, for low-light CV images, even with supplementary information from TIR images, the visual quality of the fused results still fails to match that of DCV images.

To further enhance the quality of fused images, Yang et al. proposed HDCGAN model which was specifically designed to optimize the fusion performance of NTIR and CV images. However, this model is trained on an unpaired dataset, and due to the absence of an effective semantic consistency constraint mechanism, the generated fused images still suffer from problems such as detail loss, color block artifacts, and structural contour distortion. Subsequently, Du et al. proposed the CHITNet fusion network [27], which introduced innovations in the information transfer mechanism for infrared and CV image fusion. Nevertheless, this method still fails to fundamentally address the core issues of detail loss and structural distortion in fused images. Therefore, this study builds upon the HDCGAN model proposed by Yang et al. and further optimizes its existing limitations.

2.4 Evaluation Metrics Related to Image Enhancement and Transformation

Current mainstream quality assessment methods for image transformation and enhancement are broadly categorized into four types: pixel-level error metrics, perceptual metrics, structural and region-based matching metrics, and no-reference image quality assessment (NR-IQA) methods.

Among pixel-level error metrics, Mean Squared Error (MSE) [28] is the most fundamental indicator for image quality assessment. It is defined as the average of the squared differences between corresponding pixels of the predicted and reference images. Its key advantages include simple computation and excellent differentiability, making it suitable for use as a loss function in network training. However, MSE essentially only quantifies pixel-wise intensity differences and fails to effectively capture image structural information and perceptual consistency, making it difficult to accurately reflect subjective visual quality in image enhancement tasks.

Peak Signal-to-Noise Ratio (PSNR) [29] is a logarithmic metric directly derived from MSE and is commonly used to quantify the ratio of signal power to noise power. It is widely applied in image compression and reconstruction tasks. Nevertheless, PSNR still neglects perceptual quality and structural restoration, and therefore exhibits a weak correlation with human visual perception.

The Structural Similarity Index (SSIM) [30] is one of the most widely used structure-aware metrics. It incorporates three components, namely luminance, contrast, and structural information, to evaluate the perceptual similarity between images. Although SSIM effectively addresses the limitations of PSNR, it still exhibits certain shortcomings in capturing fine-grained structural differences in complex scenes.

Among perceptual metrics, Fréchet Inception Distance (FID) [12] is one of the most representative deep perceptual indicators. It effectively quantifies the overall realism and statistical consistency between generated and reference image distributions and has been widely adopted in image enhancement tasks. Kernel Inception Distance (KID) [13] utilizes Maximum Mean Discrepancy to provide an unbiased estimate of the distributional difference between images. Compared with FID, KID exhibits greater stability, particularly in small-sample scenarios and under large distribution discrepancies. Additionally, KID circumvents the numerical instability associated with matrix square-root operations and has gradually been employed as a complement to or replacement for FID in recent years. Learned Perceptual Image Patch Similarity (LPIPS) [14] is based on a pre-trained convolutional neural network and evaluates perceptual similarity by measuring the distance between deep feature representations of image patches. It excels at capturing fine-grained perceptual differences, making it particularly well-suited for assessing visual improvements in tasks such as image enhancement and style transfer.

However, the aforementioned metrics cannot effectively assess the performance of enhanced images in downstream tasks, such as target detection and classification. To address this limitation, Han et al. [4] proposed the RmAP metric. Its core principle is to employ a unified target detection model to quantify the difference in mean Average Precision (mAP) between real and generated images, thereby characterizing both the structural restoration quality of enhanced images and their effectiveness in target detection tasks. Nevertheless, the application of RmAP is constrained by its dependence on real annotated data, which limits its applicability in many practical scenarios. Moreover, RmAP fails to capture information in non-target regions, such as background textures and contextual details.

In terms of NR-IQA, the Natural Image Quality Evaluator (NIQE) [31] constructs a statistical feature distribution model using high-quality natural images and assesses the quality of a given image by quantifying its deviation from this distribution. This method eliminates the reliance on reference images, making it suitable for practical deployment scenarios. However, it may exhibit stability issues under certain conditions. The Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [32] models the local spatial statistical characteristics of images and is effective at detecting quality degradations such as blur and noise. Nevertheless, it is overly sensitive to severely degraded images, which may compromise its robustness in complex practical scenarios.

The aforementioned metrics and NR-IQA methods all have inherent limitations, which primarily manifest in their inability to accurately assess local distortions and color shifts, their heavy reliance on high-quality prior data, and their limited interpretability. Therefore, establishing a comprehensive evaluation system that integrates structure-awareness, local sensitivity, and visual consistency remains a core challenge in current research on image generation quality assessment.

To assess the quality of enhanced images in an accurate and objective manner, this study proposes an evaluation method based on the segmentation results of the Segment Anything Model (SAM) [33]. The proposed method aims to establish a stable and objective structure-aware quality assessment framework, thereby addressing the core challenge of accurately evaluating the effectiveness of low-light image enhancement.

3. Materials and Methods

3.1. LRR-FusionColor Fusion Method

Prior to further enhancement, NCV images are first fused with synchronously acquired TIR images to supplement the detailed information lost during NCV imaging. To overcome the limitation

of traditional TIR-CV fusion methods, which either produce only grayscale images or suffer from severe color distortion, this study proposes a color reconstruction method termed LRR-FusionColor, based on dual color-space collaborative processing. By leveraging the LRRNet model [34], the proposed method combines the grayscale fusion map generated by the low-rank feature extraction network with the color components extracted from CV images to reconstruct color images. Consequently, the fused results preserve rich structural details and exhibit visually natural and consistent color appearance.

The LRRNet model, as illustrated in Figure 2, comprises two symmetric backbone branches that process TIR and gray color visible images independently. Each branch consists of multiple feature decomposition layers (LLRR blocks), which are constructed based on Low-Rank and Redundant Representation (LLRR) modules. These modules decompose the input images from both modalities into low-rank components (representing dominant structural contours) and sparse components (corresponding to fine texture details). Subsequently, the structural components extracted from the two modalities are fused via convolutional operations, followed by the reconstruction of the final fused images through element-wise addition.

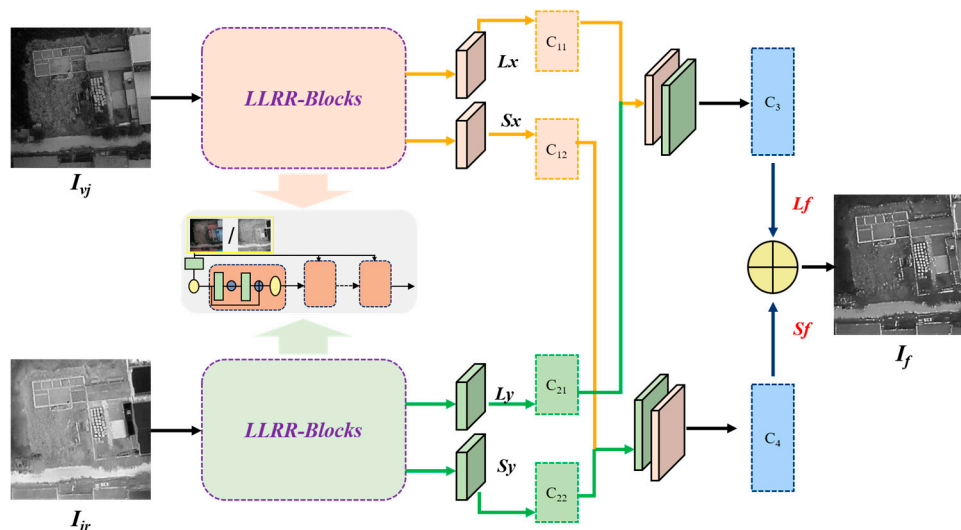


Figure 2. Schematic illustration of the LRRNet network architecture.

Despite the notable advantages of LRRNet in modeling image structural information, its fused output is confined to grayscale images and therefore fails to preserve the color information inherent in CV images. This limitation restricts its applicability and generalization capability in real-world tasks such as multimodal perception and target detection.

To address this limitation, this study proposes the LRR-FusionColor fusion method (as shown in Figure 3), which consists of two collaborative sub-modules: (1) a structure extraction module, based on low-rank and sparse modeling theory, which extracts the dominant structural components and salient modal details from TIR and CV images via LRRNet; (2) a color reconstruction module (FusionColor), which reconstructs RGB color images in the YCbCr color space by combining the CbCr chrominance channels of CV images with the fused grayscale luminance image.

In the Structure Extraction Module, the YCbCr color space is employed to effectively decouple brightness (Y) from color information (Cb, Cr), while the HSV color space enables independent manipulation of intensity and chromaticity. The color reconstruction process is described as follows. First, CV image is converted into the YCbCr and HSV color spaces to extract the original luminance and chrominance components. Second, the fused high-frequency grayscale image is used to replace the V (Value) channel in the HSV space. The reconstructed image is then inversely transformed back to the RGB space and subsequently converted to the YCbCr space to obtain structure-guided enhanced chrominance components, denoted as Cb_{HSV} and Cr_{HSV} .

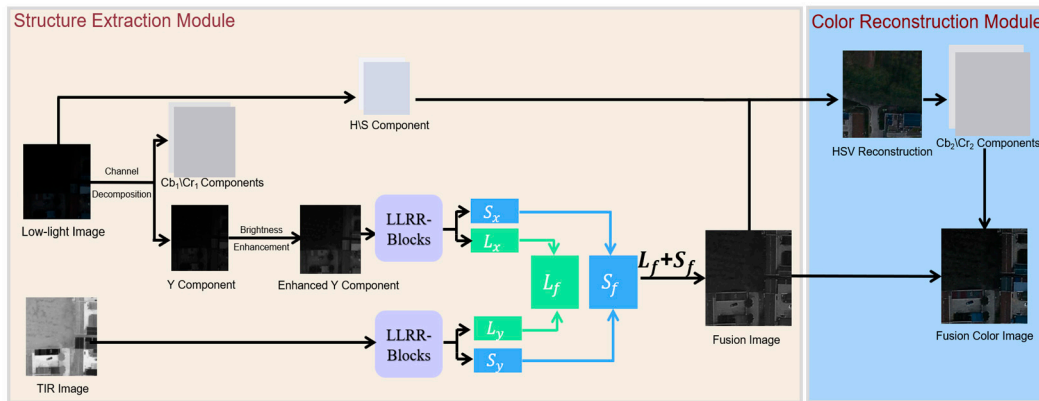


Figure 3. Architecture of the proposed LRR-FusionColor framework. framework.

Although replacing the V channel in the HSV color space with the fused grayscale image can incorporate clearer texture details, a direct inverse transformation back to the RGB space tends to introduce spectral distortion or disordered color blocks. The primary cause lies in the nonlinear coupling between abrupt brightness variations and the original color information. To address this issue, the RGB image reconstructed via the HSV space is not directly adopted as the final output. Instead, it is used as guidance for chrominance enhancement. Specifically, the enhanced chrominance components (Cb_{HSV}, Cr_{HSV}) obtained from the HSV path are linearly blended with the original chrominance components (Cb_{orig}, Cr_{orig}) extracted from the CV image:

$$Cb_{fused} = \alpha \cdot Cb_{orig} + (1 - \alpha) \cdot Cb_{HSV} \quad (1)$$

$$Cr_{fused} = \alpha \cdot Cr_{orig} + (1 - \alpha) \cdot Cr_{HSV} \quad (2)$$

where Cb_{orig} and Cr_{orig} denote the chrominance components of the original CV image in the YCbCr color space, representing the natural color information; Cb_{HSV} and Cr_{HSV} denote the chrominance components converted from the HSV space (after V-channel replacement) to the YCbCr space, which contain structure-guided color information; Cb_{fused} and Cr_{fused} are the final fused chrominance channels; and $\alpha(0-1)$ is a weighting coefficient that controls the balance between the original color information and the enhanced chrominance colors.

Cb_{orig} and Cr_{orig} preserve the color fidelity of natural scenes but lack fine structural details, whereas Cb_{HSV} and Cr_{HSV} contain clear structural information introduced by the fused grayscale image but may suffer from color deviations. Through linear weighted fusion, the proposed method not only effectively suppresses color blocks and artifacts caused by HSV-based transformation but also enhances visual contrast by exploiting the enriched chrominance information. Finally, in combination with the fused luminance channel (Y_{fused}), the method reconstructs fused images that simultaneously exhibit high structural clarity and natural color appearance.

By integrating the color reconstruction module, the proposed LRR-FusionColor method not only retains the structural representation advantages of LRRNet but also achieves color restoration at the visual-semantic perception level, thereby significantly enhancing the practical application value and visual interpretability of the fused images.

3.2. Construction of WPID

The WPID proposed in this study is composed of image pairs captured at the same location but at different times. These pairs may exhibit moderate spatial deviations while retaining sufficient semantic and structural consistency. The construction cost of such a dataset is substantially lower than that of strictly paired datasets, rendering it particularly suitable for learning paradigms that prioritize structural preservation, especially in unsupervised or weakly supervised settings.

The dataset was collected in Xianghe County, Langfang City, Hebei Province, China, as illustrated in Figure 4. A DJI-M300 UAV equipped with a DJI XT2 TIR camera and a built-in visible camera was used for data acquisition. The flight route pre-planned via the remote controller is illustrated in Figure 5. The collected scene cover a variety of land-cover types, including roads, vegetation, water bodies, buildings, and other typical ground features.



Figure 4. Location of the WPID data collection site (Lujiawu Village, Xianghe County, Hebei Province, China).



Figure 5. DJI UAV remote controller and flight route planning.

Daytime images were captured between 13:00 and 15:00 local time, while nighttime images were captured between 20:00 and 22:00 local time. Owing to the low-illumination conditions during nighttime acquisition, the nighttime images exhibit substantial noise and noticeable blurring in both target and background regions. Representative sample images are shown in Figure 6.

By adopting the same flight route, the same number of images (297 images per flight) were collected during the two missions (daytime and nighttime). All the collected images are of reliable quality and were used to construct the WPID for subsequent experiments. Considering the input size limitation of the original HDCGAN model, all images were resized to 512×512 .

Figure 6 presents representative samples from the dataset: (a) DCV images; (b) NCV images; (c) NTIR images. It should be noted that (a) and (b) form a typical weakly paired relationship, as they correspond to the same scene but exhibit an approximate 10% pixel offset due to differences in acquisition time (daytime vs. nighttime) and slight variations in UAV position and attitude.

In contrast, the spatial offset between the NTIR images in (c) and the corresponding NCV images in (b) is minimal. On the DJI XT2 camera, the distance between the visible-light lens and the TIR lens is approximately 4 cm. Given a flight altitude of 120 m, this results in a parallax-induced displacement of only 0.25 pixels. After resizing the image to 512×512 , this offset becomes negligible for applications that do not require strict pixel-level alignment.

The final constructed dataset therefore consists of multimodal image pairs (daytime-nighttime and visible-infrared) that are approximately registered, containing rich structural information and significant modal differences. This dataset provides a solid foundation for subsequent experimental validations, including weakly paired image enhancement and structure preservation evaluation.



Figure 6. Weakly Paired Image Dataset: (a) DCV images; (b) NCV images; (c) Nighttime TIR images.

3.3. HDCGAN+ Model Based on WPID

To improve the quality of NCV images, Yang et al. [2] fused NCV images with synchronously acquired TIR images and then employed the HDCGAN model to refine the fused results, aiming to achieve DCV-like visual effects. Their experimental results demonstrated that HDCGAN outperformed the prevailing state-of-the-art (SOTA) methods at that time. As illustrated in Figure 7, the HDCGAN model is built upon the CycleGAN architecture. To enhance semantic consistency between the input and output images, the authors introduced the concept of Hyperdimensional Computing (HDC) into the GAN framework.

However, as discussed earlier, since HDCGAN is trained on unpaired datasets, it still cannot fully eliminate semantic deviations in the generated images. These deviations mainly arise from the inherent semantic inconsistency between image pairs in the training data.

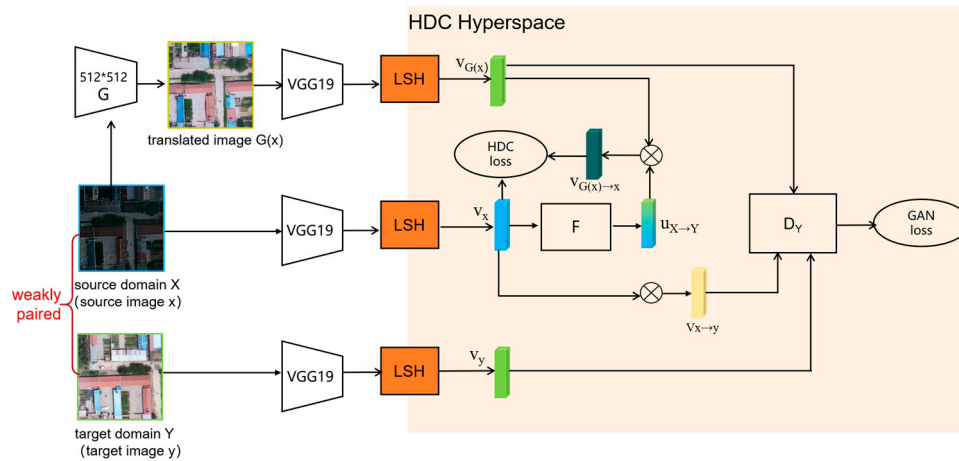


Figure 7. Schematic diagram of the HDCGAN network architecture [2].

The total loss function is composed of an adversarial loss term and a structure preservation loss term, which can be formulated as:

$$\mathcal{L}(G, D_Y, F, X, Y) = \mathcal{L}_{GAN}(G, D_Y, F, X, Y) + \lambda \mathcal{L}_{HDC}(G, X) \quad (3)$$

where the coefficient λ governs the trade-off between the visual authenticity of the generated images, enforced by the adversarial loss term, and the structural consistency, constrained by the HDC-based loss term, thereby facilitating the collaborative optimization of style transfer and structure preservation.

The original HDCGAN is trained on unpaired images with an input resolution of 256×256 pixels. In this study, we adjusted its parameter configuration to support an adaptive input resolution of 512×512 pixels and refined the data loading mechanism to enable sequential data reading during training. For clarity and to distinguish it from the original implementation, the improved model is referred to as HDCGAN+ in the remainder of this paper.

Figure 8 illustrates the differences between the unpaired image samples used in the original HDCGAN model and the WPI samples adopted in this study.



Figure 8. Comparison between unpaired images and WPI. The left side shows unpaired image samples, while the right side shows WPI, covering ground scenes acquired under different time periods and illumination conditions: (a) Image data used by Yang et al.; (b) WPI data used in this study.

3.4. Image Enhancement Evaluation Metrics Based on SAM Segmentation Results

Existing image quality metrics exhibit inherent limitations. Most of them primarily focus on perceptual attributes, such as sharpness and color fidelity, as well as semantic consistency, yet fail to effectively assess the extent to which the structural information of images is preserved. Furthermore, they often overlook the practical performance of enhanced images in downstream tasks, such as image segmentation.

This study incorporates Meta AI's Segment Anything Model (SAM) [33] to extract image structural information through image segmentation. The key advantage of SAM lies in its ability to perform full-image segmentation on arbitrary images without requiring manual annotations, while demonstrating strong structure awareness and excellent cross-scene generalization capability. Based on the segmentation results generated by SAM, a structural consistency-based evaluation framework is established to assess the effectiveness of enhanced images in downstream application scenarios.

Based on the segmentation structure provided by SAM, this study defines the following structural consistency metric:

1) Delta Mask Count (ΔN): This metric reflects the degree of structural preservation before and after image enhancement. A smaller ΔN indicates higher structural integrity of the enhanced image. It is calculated as:

$$\Delta N = |N_{SAM}^{gen} - N_{SAM}^{ref}| \quad (4)$$

where N_{SAM}^{gen} and N_{SAM}^{ref} denote the numbers of segmentation masks extracted by the SAM from the enhanced image and the reference image, respectively. A smaller difference between these two values indicates that the enhanced image better preserves the original structural integrity, maintains the completeness of semantic regions, and avoids significant information loss or region fragmentation.

2) Delta Mask Area (ΔA): This metric quantifies the degree of preservation of the overall semantic structure area before and after image enhancement. A smaller ΔA indicates higher structural consistency between the enhanced image and the reference image. It is defined as follows:

$$\Delta A = |A_{Gen}^{SAM} - A_{ref}^{SAM}| \quad (5)$$

where A_{Gen}^{SAM} and A_{ref}^{SAM} denote the total coverage areas (in pixels) of all segmentation masks extracted by SAM from the enhanced image and the reference image, respectively. ΔA represents the absolute difference between these two coverage areas. A smaller ΔA indicates that the enhanced image better preserves the completeness of semantic regions at the spatial structural level, thereby exhibiting higher consistency with the reference image.

3) Intersection-over-Union (IoU) Distribution: Inspired by the IoU metric widely adopted in object detection, this metric quantifies the spatial overlap between the structural regions of the enhanced image and those of the reference image. A higher IoU value indicates stronger consistency in both the spatial location and contour of structures between the enhanced image and the reference scene. It is defined as follows:

$$IoU(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (6)$$

The average IoU value is computed as the mean over all matched segmentation mask pairs between the enhanced image and the reference image, thereby serving as a quantitative metric for evaluating structural boundary overlap. A higher average IoU indicates that both the structural positions and contours in the enhanced image exhibit stronger consistency with those in the reference image.

4) Edge $F1$ Score: This metric quantifies the consistency of edge contours between the enhanced image and the reference image by leveraging the $F1$ score to evaluate the effectiveness of edge restoration. A higher Edge $F1$ score indicates that the enhanced image exhibits clearer and more complete edges, as well as a higher degree of structural restoration. It is defined as follows:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (7)$$

where Precision and Recall respectively quantify the degree of edge matching between the enhanced image and the reference image. A higher Edge $F1$ score indicates that the structural edges of the enhanced image are clear and more accurately restored.

5) Average Structural Complexity ($\bar{C}$). This metric is defined as the average ratio of contour length to area and is used to characterize the structural complexity of the enhanced image. A larger

$\bar{C}$ indicates that the enhanced image exhibits richer boundary details and achieves a more refined level of structural restoration. It is defined as follows:

$$C_i = \frac{\text{Perimeter}(M_i)}{\text{Area}(M_i)}, \bar{C} = \frac{1}{N} \sum_{i=1}^N C_i \quad (8)$$

where M_i is the i -th segment. A larger $\bar{C}$ indicates richer boundary details and a more refined level of structural restoration.

4. Experiments and Results

Based on the methods described in Section 3, the experiments were designed to sequentially verify three aspects: (1) conducting LRR-FusionColor image fusion to perform color fusion of NTIR and NCV images; (2) enhancing the fused images using the HDCGAN+ model; (3) applying the SAM semantic segmentation algorithm to the enhanced images, constructing the evaluation index system based on the SAM segmentation results, and comparing the proposed method with representative existing algorithms.

4.1. Experimental Platform and Data Preprocessing

All experiments in this study were conducted on a DELL PowerEdge T640 tower server. The server featured the following key hardware configurations: four Intel Xeon Silver 4116 processors (CPUs) with a base frequency of 2.10 GHz; 128 GB of system memory, composed of four 32 GB DDR4 memory modules running at 2666 MHz; and two NVIDIA GeForce GTX 4090 D graphics processing units (GPUs), each with 24 GB of GDDR6X video memory. All training and evaluation of the deep learning models were performed on this server using the PyTorch deep learning framework.

Consistent with the details in Section 3.2, the experiments utilized a DJI-M300 UAV equipped with an XT2 TIR camera to capture dual-modal images at 12:00 (daytime) and 20:00 (nighttime). Prior to model training, the acquired data underwent the following preprocessing steps:

- (1) correcting lens distortion using Pix4D software;
- (2) performing affine transformation to achieve co-registration of dual-modal and day-night images;
- (3) cropping the original images to 512×512 pixels;
- (4) applying non-local means (NLM) filtering to denoise the TIR images.

4.2. Fusion and Evaluation of CV and TIR Images

We selected four representative existing fusion methods, namely the LRRNet grayscale fusion method (Fusion) [34], the HSV color space fusion method (HSV) [35], the LAB color space fusion method (Lab) [36], and the YCbCr color space fusion method (YCbCr) [37], and compared their performance with that of the proposed LRR-FusionColor method. The experimental results are presented in Figure 9. The vertical axis displays seven image categories, including the original image group, in which the first two columns represent the NCV and TIR images. The horizontal axis depicts five sample images selected from the experimental region.

As illustrated in Figure 9, the grayscale fusion image generated by the Fusion method (fusing with TIR data) exhibits markedly enhanced contrast and edge sharpness compared with the original NCV image, demonstrating the effectiveness of LRRNet in fusing multimodal information. The HSV and Lab methods are capable of achieving color reconstruction, but they suffer from color distortion, regional color deviation, and loss of details in bright areas. In particular, the HSV method degrades structural integrity due to pixel insufficiency.

The YCbCr method preserves the grayscale characteristics and overall color style of the original images, yet it fails to provide sufficient brightness information. In contrast, the proposed FusionColor method achieves the best performance in key ground-object regions, including vegetation, roads, and building edges. It not only restores more natural color tones and a more uniform brightness

distribution, but also effectively mitigates color blocking and halo artifacts, thereby achieving a better balance between structural preservation and color consistency.

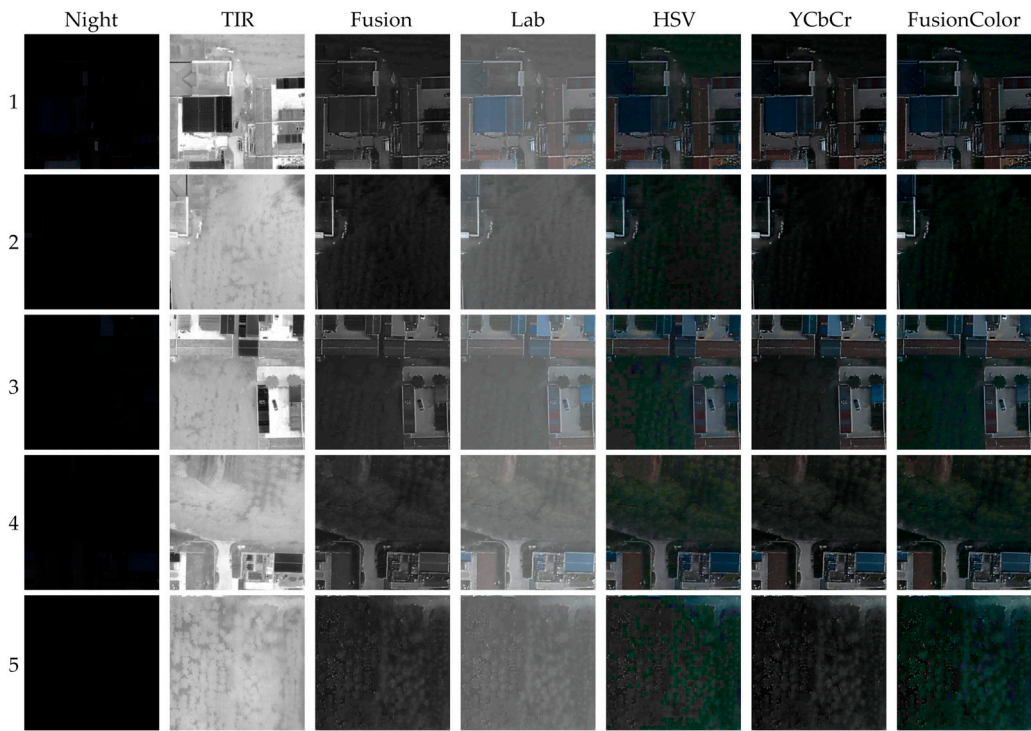


Figure 9. Visual comparison of fusion results obtained using different color restoration methods. From left to right: Nighttime CV Images, NTIR Images, Fusion results, Lab fusion results, HSV fusion results, YCbCr fusion results, and the results of the proposed FusionColor method.

To conduct a quantitative evaluation of the fusion methods, three metrics were adopted: Peak Signal-to-Noise Ratio (PSNR) [29], Structural Similarity Index (SSIM) [30], and Information Entropy (Entropy) [38]. The corresponding statistical results are summarized in Table 1.

Table 1. Quantitative evaluation results of different fusion methods (bold values denote the best performance, while underlined values represent the second-best performance).

Method	PSNR_TIR ↑	SSIM_TIR ↑	PSNR_NCV ↑	SSIM_NCV ↑	Entropy ↑
Fusion	7.855	0.4280	18.4627	0.4087	5.4206
Lab	\	\	9.9169	0.2474	5.5791
HSV	\	\	19.5732	0.4593	5.6654
YCbCr	\	\	20.5635	<u>0.5013</u>	<u>5.9157</u>
<u>FusionColor</u>	\	\	<u>20.4789</u>	0.5137	5.9227

As shown in Table 1, the proposed FusionColor method delivers the best overall performance among the compared methods. Since color reconstruction is performed on the fused images, their dimensionality differs from that of the NTIR images; therefore, SSIM_TIR is computed only between the grayscale fused images and the TIR images. FusionColor ranks second in PSNR_NCV and first in SSIM_NCV, indicating that it preserves the brightness and structural information of the original NCV well. Moreover, by incorporating HSV color information on the top of the YCbCr-based scheme, FusionColor slightly weakens luminance consistency while yielding improved color appearance and visual quality, which is reflected by its best Entropy score.

Integrating the subjective and objective evaluation findings, the proposed LRR-FusionColor framework effectively fuses multimodal structural information from infrared and visible images, thereby addressing the issues of inadequate structural representation and unnatural color

reconstruction in infrared-visible image fusion. It achieves a good balance between color restoration and structural preservation, improves the information content and interpretability of nighttime fused images, and provides better support for subsequent tasks such as image enhancement and target detection.

4.3. HDCGAN+ Image Enhancement

Prior to training the HDCGAN+ model, given that the original dataset (described in Section 3.2) comprises only 297 image pairs, data augmentation techniques were applied to expand the dataset to 1782 samples, including horizontal flipping and rotation. Of these, 300 image pairs were designated as the test set, while the remaining samples were used as the training set.

The experiments incorporated original NCV images (pre-transformation), enhanced nighttime images, and four comparative algorithms, namely CycleGAN [7], EnlightenGAN [39], RetinexFormer [40], and HDCGAN [2], for performance benchmarking. Figure 10 shows the NCV images, the enhancement results generated by the fire compared methods, and the corresponding reference DCV images, providing a comprehensive visual comparison. The enhanced images produced by HDCGAN+ exhibit the closest match to the reference DCV images in terms of both color fidelity and structural consistency, with well-restored vegetation, roads networks, and building structures. In addition, while EnlightenGAN show robust performance in structural and fine-detail restoration, it suffers from notable color distortion, particularly in vegetation-dominated regions.

For the quantitative evaluation of the output results from all algorithms, two widely used image enhancement metrics, FID [12] and KID [13], were adopted to quantify the feature distribution discrepancies between the enhanced images and the reference DCV images. The specific calculation results are summarized in Table 2.

Based on the FID and KID metrics, HDCGAN+ significantly outperforms the other comparative algorithms, achieving the lowest FID and KID values, which indicate the smallest discrepancy between the generated enhanced images and the reference DCV images. Among the comparative algorithms, HDCGAN delivers the best performance, ranking second only to HDCGAN+. CycleGAN and EnlightenGAN demonstrate distinct strengths across the two metrics. RetinexFormer exhibits the poorest performance, with both metrics being substantially higher than those of the other algorithms, resulting in the lowest similarity between the generated images and the reference images.

In summary, the introduction of the weakly paired training strategy in HDCGAN+ effectively addresses the issues of high data acquisition costs and insufficient robustness associated with strictly paired datasets, while compensating for the inadequate structural preservation capability of unpaired training. This strategy significantly enhances the model’s generalization capability and practical applicability, demonstrating strong potential in remote sensing image enhancement.

Table 2. Quantitative evaluation results of enhanced images produced by different algorithms (bold values denote the best performance, while underlined values represent the second-best performance).

group	FID ↓	KID ↓
CycleGAN	58.5054	0.0357
EnlightenGAN	60.1970	0.0291
RetinexFormer	166.6155	0.1544
HDCGAN	<u>53.3045</u>	<u>0.0232</u>
HDCGAN+ (Ours)	34.7739	0.0085

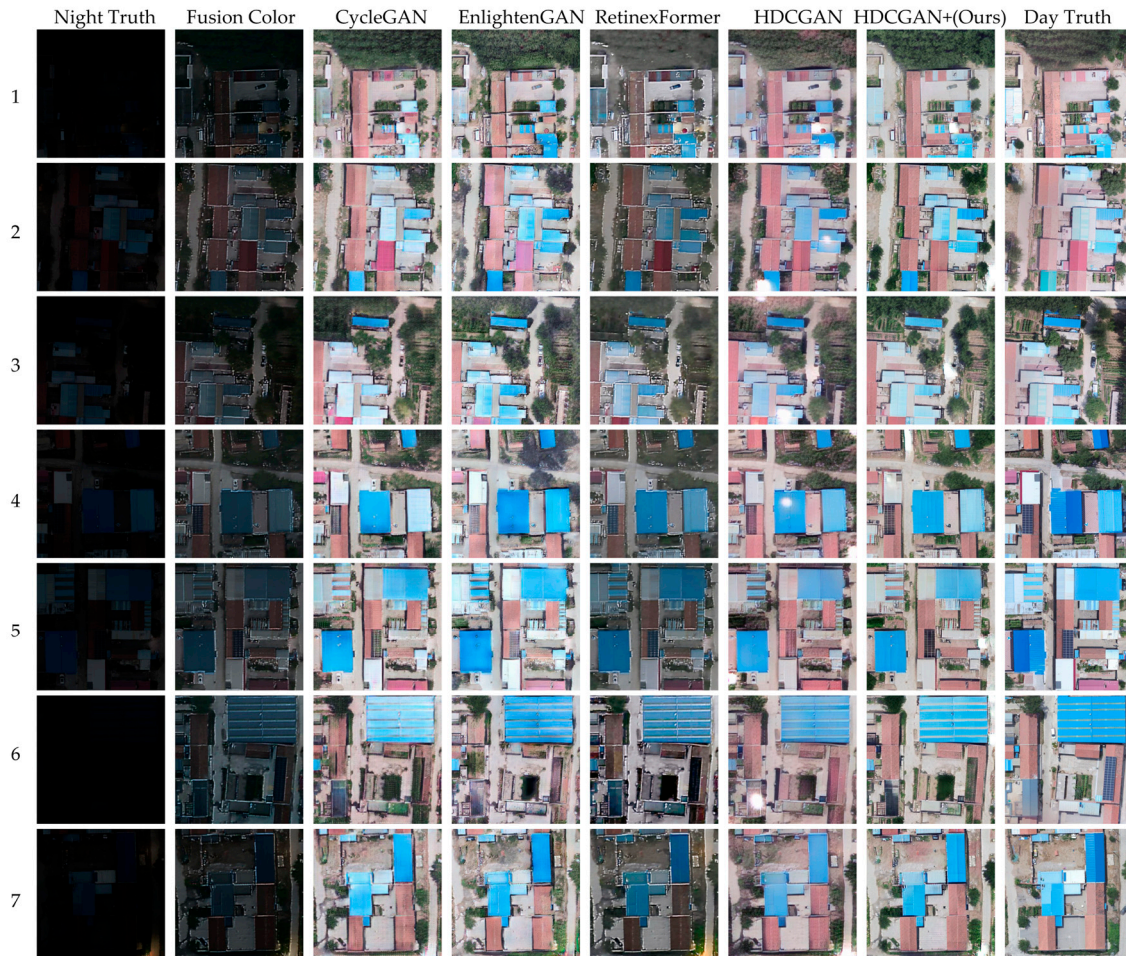


Figure 10. Visual comparison of enhancement results between HDCGAN+ and other state-of-the-art image enhancement methods. From left to right: NCV images, nighttime fusion results, CycleGAN results, EnlightenGAN results, RetinexFormer results, HDCGAN results, HDCGAN+ results, and the reference DCV images developed in this study based on WPID.

4.4. Structural Consistency Evaluation Based on SAM Segmentation

The experiments designated the LRR-FusionColor fused images and the enhanced images generated by CycleGAN, EnlightenGAN, RetinexFormer, HDCGAN, and HDCGAN+ as the comparative set, with the NCV images and the reference DCV images serving as baselines. The SAM model was employed to perform full-image semantic segmentation on each image, yielding the corresponding semantic segmentation masks. Subsequently, five metrics, namely Delta Mask Count (ΔN), Delta Mask Area (ΔA), Intersection-over-Union Distribution (IoU), Edge F1 Score and Average Structural Complexity ($\bar{C}$), were computed to assess the quality of the segmentation results.

To ensure the fairness of the experimental results, the following uniform settings were applied to all images:

- (1) all images used a 512×512 resolution;
- (2) the SAM algorithm was used with fixed parameters, without individual tuning;
- (3) all metrics were computed based on the complete test set;

Figure 11 presents the segmentation results of the original NCV images, the nighttime Fusion images, and the enhanced images produced by CycleGAN, EnlightenGAN, RetinexFormer, HDCGAN, and HDCGAN+, as well as the segmentation results of the original DCV images and the real daytime images, providing a comprehensive visual comparison based on the SAM model. The original nighttime images contain fewer segments and exhibit blurred boundaries, which hinder effective target segmentation. In contrast, the structural regions, such as roads, buildings, and

vegetation, in the DCV images are well preserved in the enhanced nighttime images, with good details continuity.



Figure 11. Visualization of segmentation masks produced by different image enhancement methods (colors represent different segmented regions rather than specific land-cover categories). From left to right: the original NCV masks, nighttime Fusion masks, CycleGAN masks, EnlightenGAN masks, RetinexFormer masks, HDCGAN masks, the proposed weakly paired HDCGAN+ masks, the DCV masks, and the original DCV image.

However, since the SAM model was not fine-tuned on the dataset used in this study, a small number of mis-segmentation and over-segmentation cases still exist. Compared with the original nighttime images, the segmentation results of the LRR-FusionColor fused images are significantly improved, verifying both the necessity of image fusion and the effectiveness of the proposed method. The segments of the HDCGAN+ enhanced images exhibit clear contours, and their quantity and spatial distribution are highly consistent with those of the reference DCV images, outperforming HDCGAN and the other comparative methods.

While EnlightenGAN preserves structural information effectively, it introduces excessive noise, particularly in vegetation-dominated regions. Overall, the experimental results demonstrate that the proposed optimized framework significantly enhances the generation quality and semantic fidelity of the enhanced images, and exhibits favorable generalization capability and interpretability.

Using the proposed evaluation metrics, statistical calculations were performed on the segmentation results of each image. The results are shown in Table 3.

Table 3 presents the quantitative results of structural metrics for different image categories. All metrics are computed based on 297 samples, and the reported values are obtained by averaging the results after removing outliers using the IQR (Interquartile Range) method. HDCGAN+ yields the smallest ΔN and ΔA with respect to real daytime images, indicating superior preservation of semantic objects. It also achieves the highest Block IoU and Edge F1 scores, reflecting its strong structural fidelity. Meanwhile, its moderate structural complexity demonstrates a favorable trade-off between detail richness and contour completeness, without introducing over-smoothing, pseudo-structures, or excessive noises.

Table 3. Structural consistency evaluation results of enhanced images produced by different methods (bold values indicate the best performance, and underlined values indicate the second-best performance).

Group	$\Delta N \downarrow$	$\Delta A \downarrow$	F1 $\uparrow$	IoU(A,B) $\uparrow$	$\bar{C}$
night	91.6857	0.3588	0.2000	0.2350	0.4498
LRR-FusionColor	58.9360	0.3961	0.5770	0.2855	0.4554
CycleGAN	61.1111	0.3158	0.6145	0.2983	0.3489
EnlightenGAN	29.7239	<u>0.2530</u>	<u>0.6248</u>	0.2474	0.3689
RetinexFormer	50.6566	0.3567	0.6089	0.2278	0.4193
HDCGAN	67.9273	0.2880	0.6148	<u>0.3048</u>	0.3590
HDCGAN+(Ours)	<u>29.7800</u>	0.2458	0.6578	0.3325	0.3985

By comparing the traditional metrics (FID and KID) in Table 2 with the structural consistency evaluation metrics in Table 3, it can be observed that the evaluation system proposed in this study provides a more accurate reflection of the performance of different methods in subsequent applications. For example, in Table 2, HDCGAN achieves the second-best performance in terms of FID and KID. However, in Table 3, it underperforms algorithms such as CycleGAN and EnlightenGAN in structural consistency indicators, including ΔN and IoU of segment regions. This discrepancy directly reveals the limitations of traditional metrics: they primarily measure the global distribution similarity between images but fail to capture structural consistency in practical applications, such as semantic object preservation and contour alignment. As a result, relying solely on traditional metrics may lead to evaluation outcomes that deviate from real application requirements.

In contrast, the proposed structural consistency metrics characterize the overall structural integrity of images from multiple dimensions, including ΔN , ΔA , IoU , and $F1$. Furthermore, the superior performance of HDCGAN+ under these new metrics (Table 3) is consistent with its strong capability in preserving semantic ground objects and its excellent structural representation ability, which agrees with the qualitative and quantitative results discussed earlier. Thus, compared with traditional evaluation metrics, the proposed evaluation system can more objectively and comprehensively assess the practical application quality of enhanced images in remote sensing scenarios.

In summary, HDCGAN+ not only achieves excellent performance on traditional evaluation metrics but also yields SAM segmentation results that are highly consistent with those of the reference DCV images, demonstrating its strong effectiveness and reliability for subsequent remote sensing application tasks.

5. Discussion and Analysis

To address the insufficient quality of UAV remote sensing images under low-light conditions, the data dependency dilemma of existing enhancement methods, and the limitations of current evaluation systems, this study proposes a set of technical solutions, including the LRR-FusionColor multimodal fusion method, an improved HDCGAN with a weakly paired training mechanism, and a structural consistency evaluation framework based on SAM segmentation results. The effectiveness of the proposed approaches is validated using day-night UAV images collected in Xianghe City,

Hebei Province. Nevertheless, this study still has certain limitations and aspects that can be further improved, which are discussed and analyzed as follows:

5.1. Improving HDCGAN with WPID

To address the core limitations of low-light UAV remote sensing image enhancement, namely the difficulty of acquiring strictly paired data and the susceptibility of unpaired data to semantic distortion, this study proposes a weakly paired data mechanism. It requires only day-night image acquisition of the same area with consistent scene semantics, enabling low-cost dataset construction by conducting UAV flights along identical routes at different times. Experimental results demonstrate that the improved HDCGAN+ outperforms methods such as CycleGAN in terms of FID and KID metrics. Subjectively, it more accurately restores vegetation tones and building textures, effectively alleviating the semantic distortion problem commonly observed in unsupervised methods. These results confirm that the proposed mechanism achieves a favorable balance between data acquisition cost reduction and enhancement accuracy preservation.

Despite the notable merits of the weakly paired data mechanism, several limitations persist. First, its adaptability to extreme low-light scenarios remains constrained. The experiments in this study only validate the effectiveness of the proposed method under conventional nighttime low-light conditions and do not encompass extreme scenarios such as moonless nights or heavy haze. In such cases, visible images exhibit substantially increased noise levels and severe information degradation, which may cause the model to produce texture artifacts and result in unstable enhancement performance.

Second, the computational overhead of the model remains relatively high. When the input resolution is scaled to 512×512 , the computational burden associated with feature extraction and representation learning increases considerably. Both the training and inference stages still rely on high-performance server-grade hardware, making it challenging to deploy the model on low-power UAV onboard computing platforms. Consequently, the proposed method has not yet met the real-time processing requirements for on-the-fly image enhancement, such as real-time monitoring in nighttime disaster-response operations.

In future work, we intend to further optimize both the weakly paired dataset and the associated model from two primary aspects.

First, we will improve the framework's adaptability to extreme low-light scenarios. A multi-scale noise suppression module will be incorporated into the data preprocessing stage, which integrates non-local means filtering with a spatial attention mechanism to effectively suppress severe noise while preserving critical ground-object details. In addition, a synthetic extreme low-light dataset simulating conditions such as moonless nights and heavy haze will be developed based on the existing dataset and utilized for model pre-training, thereby enhancing the robustness and generalization capability of the model.

Second, we will adopt a model optimization strategy that integrates lightweight architectural design with resolution adaptability. While retaining the capability to process images at a resolution of 512×512 , conventional convolutions in the generator will be replaced with depthwise separable convolutions, such as those employed in the bottleneck structures of MobileNetV3. Furthermore, a lightweight feature pyramid structure will be introduced to reduce the computational overhead of intermediate feature maps, and floating-point quantization will be applied to further decrease memory consumption and computational complexity. The overarching goal is to substantially improve inference speed while preserving the model's fine-detail representation capability, thereby enabling real-time deployment on UAV onboard computing platforms.

5.2. Structural Consistency Evaluation Metrics Based on SAM Segmentation Results

In evaluating the enhancement performance of low-light UAV remote sensing images, the block-based structural consistency evaluation system proposed in this study effectively overcomes the limitations of traditional metrics, which mainly emphasize perceptual quality while neglecting

structural integrity and adaptability to downstream tasks. Existing mainstream metrics, such as PSNR, SSIM, and FID, are primarily designed to quantify pixel-wise errors or global perceptual similarity, but they fail to accurately reflect whether the structural information of an image is faithfully preserved. In contrast, downstream UAV remote sensing applications, including target detection and ground object classification, impose stringent requirements on image structural consistency. For instance, in disaster emergency scenarios, the integrity of building contours directly affects the accurate assessment of damage severity, highlighting the necessity of structure-aware evaluation criteria.

By leveraging SAM's zero-shot segmentation capability, this study achieves full-image structural partitioning without requiring additional annotations, and designs five metrics, Delta Mask Count (ΔN), Delta Mask Area (ΔA), Intersection-over-Union Distribution (IoU), Edge F1 Score, and Average Structural Complexity ($\bar{C}$), to quantitatively evaluate enhancement performance from three dimensions: global structure preservation, local detail restoration, and boundary clarity. Experimental results demonstrate that the enhanced images generated by the HDCGAN+ model achieve the best performance across all structural consistency metrics. In particular, their segmentation masks exhibit high similarity to those of the reference images in terms of road continuity and building contour integrity. This effectively reveals the pseudo-enhancement phenomenon, in which traditional metrics indicate good performance while critical structural information is actually lost, and provides a solid quantitative basis for establishing the relationship between enhancement quality and practical application value.

Nevertheless, the proposed evaluation method still has inherent limitations in two aspects. First, this study directly employs the general-purpose image segmentation model SAM without targeted optimization for UAV remote sensing characteristics, such as small-scale ground objects (e.g., rural paths and farm ridges) or low-contrast targets (e.g., green belts under low-light conditions). As a result, segmentation omissions or misclassifications may arise, thereby compromising the accuracy of the derived structural evaluation metrics.

Second, the weighting strategy for the proposed metrics lacks scene adaptability. Different application scenarios impose distinct demands on evaluation dimensions. For instance, public security applications prioritize the structural integrity of pedestrians and vehicles, whereas disaster emergency scenarios place greater emphasis on the clarity and continuity of damaged building boundaries. The current evaluation system employs an equal-weighting scheme for all metrics, and its effectiveness across diverse application scenarios requires further validation and optimization.

Future optimization of this evaluation method can be conducted from three aspects:

First, efforts will be made to enhance SAM's segmentation adaptability to remote sensing ground objects. Specifically, the SAM decoder will be fine-tuned using a labeled UAV remote sensing ground object dataset, including buildings, roads, vegetation, and small ground objects, to improve the model's segmentation capability for low-contrast and small targets. Meanwhile, ground object type weight coefficients will be incorporated to impose higher penalty weights on segmentation errors of core ground objects in the scene, thereby enhancing the accuracy of metric calculation.

Second, the efficacy of this metric system in other downstream tasks (e.g., target detection and ground object classification) will be assessed, and relevant evaluation indicators from these tasks will be integrated into the current system when necessary to establish a more comprehensive evaluation framework.

Third, a scene-adaptive weighting mechanism will be developed. A weight prediction model will be built based on scene features, such as ground object density, lighting conditions, and task type, and the weights of each evaluation metric will be dynamically adjusted by inputting scene-specific information, thereby improving the scene adaptability of the evaluation system.

6. Conclusions

This study addresses two challenges in the field of image enhancement, namely appropriate train dataset construction and performance evaluation. Its key contributions are outlined as follows:

First, to address the predicament faced by reliance on strictly paired datasets and the limited accuracy resulting from dependence on unpaired datasets, this study proposes the construction of a WPID and validates its effectiveness. Unlike strictly paired datasets, WPID does not require pixel-level precise registration; instead, it preserves only the essential correlation information between image pairs, such as approximate spatial alignment of the same region and semantic consistency of key ground objects. This design not only significantly reduces the cost of data collection and manual annotation, but also provides necessary quality constraints for enhancement models, thereby effectively alleviating issue such as detail distortion and artifact generation that are prevalent in unsupervised methods. Meanwhile, it relaxes the strict dependence of supervised methods on accurately paired data, demonstrating superior adaptability in challenging practical scenarios, including low-light imaging and remote sensing scenes with complex ground object distributions.

Second, to address the limitations of traditional evaluation systems, which either rely solely on objective metrics such as PSNR and SSIM or focus only on subjective visual perception, and thus fail to comprehensively assess enhancement performance in terms of detail restoration, semantic consistency, and practical applicability, this study establishes a new evaluation framework that integrates objective metrics, subjective visual scoring, and scenario-oriented application metrics, such as ground feature recognition accuracy of remote sensing images and target detection success rates for low-light images. This framework not only enables a more comprehensive and accurate evaluation of enhancement performance, but also provides more targeted guidance for the selection and practical optimization of enhancement methods under different application scenarios.

Author Contributions: Conceptualization, K.C.K. and M.S.; Methodology, K.C.K., H.Y. and M.S.; Validation, X.W.; Resources, X.W.; Data curation, D.L.; Writing—review & editing, K.C.K., H.Y., M.S., D.L., and X.W.; Visualization, X.W.; Supervision, M.S.; Funding acquisition, M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 42171327, and supported by the High-Performance Computing Platform of Peking University.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

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

References

1. Jobson, D.J.; Rahman, Z.-U.; Woodell, G.A. Properties and performance of a center/surround Retinex. *IEEE Trans. Image Process.* 1997, 6(3), 451–462. <https://doi.org/10.1109/83.557356>
2. Yang, S.; Sun, M.; Lou, X.; et al. Nighttime Thermal Infrared Image Translation Integrating Visible Images. *Remote Sens.* 2024, 16(4), 666. <https://doi.org/10.3390/rs16040666>.
3. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; et al. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* 2014, 27, 2672–2680.
4. Han, Z.; Zhang, Z.; Zhang, S.; et al. Aerial Visible-to-Infrared Image Translation: Dataset, Evaluation, and Baseline. *J. Remote Sens.* 2023, 3, 1–20. <https://doi.org/10.34133/remotesensing.0096>
5. Li, C.; Guo, J.; Porikli, F.; et al. LightenNet: A Convolutional Neural Network for weakly illuminated image enhancement. *Pattern Recognit. Lett.* 2018, 104, 15–22. <https://doi.org/10.1016/j.patrec.2018.01.010>
6. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 2017, 5967–5976. <https://doi.org/10.1109/CVPR.2017.632>.
7. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proc. IEEE Int. Conf. Comput. Vis.* 2017, 2242–2251. <https://doi.org/10.1109/ICCV.2017.244>.
8. Ma, T.; Liu, J.; Zhang, H.; et al. RetinexGAN: Unsupervised Low-Light Enhancement With Two-Layer Convolutional Decomposition Networks. *IEEE Access* 2021, 9, 56539–56550. <https://doi.org/10.1109/ACCESS.2021.3072331>.

9. Mao, R.; Cui, R. RetinexGAN Enables More Robust Low-Light Image Enhancement Via Retinex Decomposition Based Unsupervised Illumination Brightening. Open Review 2023. <https://openreview.net/forum?id=3SqNZXg24T>.
10. Luo, F.; Li, Y.; Zeng, G.; Peng, P.; Wang, G.; Li, Y. Thermal Infrared Image Colorization for Nighttime Driving Scenes With Top-Down Guided Attention. IEEE Trans. Intell. Transp. Syst. 2022, 23(9), 15808–15823. <https://doi.org/10.1109/TITS.2022.3151276>.
11. Ye, J.; Qiu, C.; Zhang, Z. A survey on learning-based low-light image and video enhancement. Displays 2024, 81, 1–21. <https://doi.org/10.1016/j.displa.2023.102614>.
12. Heusel, M.; Ramsauer, H.; Unterthiner, T.; et al. GANs trained by a two-time-scale update rule converge to a local Nash equilibrium. Adv. Neural Inf. Process. Syst. 2017, 30, 6626–6637.
13. Bińkowski, M.; Sutherland, D.J.; Arbel, M.; et al. Demystifying MMD GANs. arXiv Preprint. <https://doi.org/10.48550/arXiv.1801.01401>.
14. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 2018, 586–595. <https://doi.org/10.1109/CVPR.2018.00068>.
15. Pisano, E.D.; Zong, S.; Hemminger, B.M. et al. Contrast Limited Adaptive Histogram Equalization image processing to improve the detection of simulated spiculations in dense mammograms. J. Imaging Inform. Med. 1998, 11, 193–200. <https://doi.org/10.1007/BF03178082>.
16. Lee, C.; Lee, C.; Kim, C.S. Contrast enhancement based on layered difference representation of 2D histograms. IEEE Trans. Image Process. 2013, 22(12), 5372–5384. <https://doi.org/10.1109/TIP.2013.2284059>.
17. Chandrasekar, R.N.; Anand, C.R.; Preethi, S. Survey on histogram equalization method-based image enhancement techniques. Proc. 2016 Int. Conf. Data Min. Adv. Comput. (SAPIENCE) 2016, 150–158. <https://doi.org/10.1109/SAPIENCE.2016.768415656>.
18. Land, E.H.; McCann, J.J. Lightness and Retinex theory. JOSA 1971, 61(1), 1–11. <https://doi.org/10.1364/JOSA.61.000001>.
19. Jobson, D.J.; Rahman, Z.; Woodell, G.A. Properties and performance of a center/surround Retinex. IEEE Trans. Image Process. 1997, 6(3), 451–462. <https://doi.org/10.1109/83.557356>.
20. Rahman, Z.; Jobson, D.J.; Woodell, G.A. Multi-scale Retinex for color image enhancement. Proc. 3rd IEEE Int. Conf. Image Process. 1996, 3, 1003–1006. <https://doi.org/10.1109/ICIP.1996.560995>.
21. Palma-Amestoy, R.; Provenzi, E.; Bertalmío, M.; et al. A perceptually inspired variational framework for color enhancement. IEEE Trans. Pattern Anal. Mach. Intell. 2009, 31(3), 458–474. <https://doi.org/10.1109/TPAMI.2008.86>.
22. Fu, X.; Liao, Y.; Zeng, D.; et al. A probabilistic method for image enhancement with simultaneous I and reflectance estimation. IEEE Trans. Image Process. 2015, 24(12), 4965–4977. <https://doi.org/10.1109/TIP.2015.2474701>.
23. Guo, X.; Li, Y.; Ling, H. LIME: Low-light image enhancement via illumination map estimation. IEEE Trans. Image Process. 2018, 26(2), 982–993. <https://doi.org/10.1109/TIP.2016.2639450>.
24. Qu, Y.; Ou, Y.; Xiong, R. Low illumination enhancement for object detection in self-driving. Proc. 2019 IEEE Int. Conf. Robot. Biomimetics (ROBIO) 2019, 1738–1743. <https://doi.org/10.1109/ROBIO49542.2019.8961471>.
25. Liu, Q.; Zhou, H.; Xu, Q.; et al. PSGAN: A generative adversarial network for remote sensing image pan-sharpening. IEEE Trans. Geosci. Remote Sens. 2020, 59(12), 10227–10242. <https://doi.org/10.1109/TGRS.2020.3042974>.
26. Pei, S.; Lin, J.; Liu, W.; et al. Beyond night visibility: Adaptive multi-scale fusion of infrared and visible images. arXiv Preprint. <https://doi.org/10.48550/arXiv.2403.01083>.
27. Du, K.; Li, H.; Zhang, Y.; Yu, Z. CHITNet: A complementary to harmonious information transfer network for infrared and visible image fusion. IEEE Trans. Instrum. Meas. 2025, 74, 1–17. <https://doi.org/10.1109/TIM.2025.3527523>.
28. Das, K.; Jiang, J.; Rao, J.N.K. Mean squared error of empirical predictor. Ann. Stat. 2004, 32(2), 818–840. <https://doi.org/10.1214/009053604000000201>.
29. Johnson, D.H. Signal-to-noise ratio. Scholarpedia 2006, 1(12), 2088. <https://doi.org/10.4249/scholarpedia.2088>.

30. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Structural similarity index (SSIM) revisited: A data-driven approach. *Expert Syst. Appl.* 2022, 189, 116087. <https://doi.org/10.1016/j.eswa.2021.116087>.
31. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.* 2013, 20(3), 209–212. <https://doi.org/10.1109/LSP.2012.2227726>.
32. Mittal, A.; Moorthy, A.K.; Bovik, A.C. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* 2012, 21(12), 4695–4708. <https://doi.org/10.1109/TIP.2012.2214050>.
33. Kirillov, A.; Mintun, E.; Ravi, N.; et al. Segment anything. *Proc. IEEE/CVF Int. Conf. Comput. Vis.* 2023, 4015–4026. <https://doi.org/10.1109/ICCV51070.2023.00371>.
34. Li, H.; Xu, T.; Wu, X.-J.; Lu, J.; Kittler, J. LRRNet: A novel representation learning-guided fusion network for infrared and visible images. *IEEE Trans. Pattern Anal. Mach. Intell.* 2023, 45(9), 11040–11052. <https://doi.org/10.1109/TPAMI.2023.3268209>.
35. Smith, A.R. Color gamut transform pairs. *ACM Siggraph Computer Graphics.* 1978, 12(3), 12–19. <https://doi.org/10.1145/965139.807361>.
36. International Commission on Illumination (CIE). Colorimetry — Part 4: CIE 1976 Lab Colour space (ISO/CIE 11664-4:2019). CIE Cent. Bur. 2019. <https://cie.co.at/publications/colorimetry-part-4-cie-1976-lab-colour-space-0>.
37. Xiang, F.M.; Zhu, Z.Y.; Xu, J.; et al. Research on algorithms of color space conversion from YUV to RGB. *Mod. Electron. Tech.* 2012, 35(22), 65–68.
38. Luo, F.; Li, Y.; Zeng, G.; Peng, P.; Wang, G.; Li, Y. Thermal infrared image colorization for nighttime driving scenes with top-down guided attention. *IEEE Trans. Intell. Transp. Syst.* 2022, 23(9), 15808–15823. <https://doi.org/10.1109/TITS.2022.3145476>.
39. Jiang, Y.; Gong, X.; Liu, D.; et al. EnlightenGAN: Deep light enhancement without paired supervision. *IEEE Trans. Image Process.* 2021, 30, 2340–2349. <https://doi.org/10.1109/TIP.2020.3051462>.
40. Cai, Y.; Bian, H.; Lin, J.; et al. Retinexformer: One-stage Retinex-based transformer for low-light image enhancement. *Proc. IEEE/CVF Int. Conf. Comput. Vis.* 2023, 12504–12513. <https://doi.org/10.1109/ICCV51070.2023.01149>.

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