remote sensing

MDPI

Article

HDCGAN+: A Low-Illumination UAV Remote Sensing Image Enhancement and Evaluation Method Based on 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; 2401210166@stu.pku.edu.cn (K.C.K.); wxinyi0616@stu.pku.edu.cn (X.W.); liudong1006@stu.pku.edu.cn (D.L.); hanjuny@stu.pku.edu.cn (H.Y.)

* Correspondence: sunmin@pku.edu.cn

Highlights

What are the main findings?

- A novel low-illumination UAV remote sensing image enhancement method is proposed, based on a Weakly Paired Image Dataset (WPID) 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-illumination UAV images in public security and disaster emergency applications.
- The new evaluation system provides a robust tool for assessing and comparing low-illumination enhancement techniques, advancing the field of remote sensing image quality evaluation.

Abstract

Remote sensing images acquired by UAVs under nighttime or low-illumination 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-illumination 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-illumination 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.

Check for updates

Academic Editor: Zhenwei Shi

Received: 28 January 2026
Revised: 16 March 2026
Accepted: 24 March 2026
Published: 26 March 2026

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

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

1. Introduction

Images acquired by drones in low-illumination environments, such as at night, are often difficult to use directly. This greatly restricts their ability to meet the demand for around-the-clock applications in fields such as disaster response and public safety. If the images captured under these low-illumination conditions could be enhanced to satisfy application requirements, the operational capabilities of drones would be significantly improved.

Research on low-illumination image enhancement can be broadly categorized into three types: (1) traditional enhancement methods, which center on image processing and physical modeling, aiming to improve image brightness and contrast; (2) deep learning-based enhancement methods, which can be roughly divided into supervised and unsupervised approaches, using machine learning to establish a mapping relationship from low-illumination images to normally lit images; and (3) methods for fusing and enhancing Nighttime Color Visible (NCV) images and Thermal infrared (TIR) images, which utilize thermal infrared data to compensate for the information lost by NCV images in low-illumination environments, and further employ machine learning algorithms to construct a mapping relationship between the fused images and normal images.

(1) Traditional Low-Illumination Image Enhancement Methods: In the early stages, traditional low-illumination image enhancement methods were primarily based on fundamental linear or nonlinear transformations, such as linear brightness stretching or nonlinear Gamma correction. These methods are characterized by simple principles and high computational efficiency. However, they lack adaptability to local image features, resulting in limited enhancement effects.

Subsequently, methods based on global and local histogram adjustment became dominant. Representative works include those by Pisano et al. (1998) [1,2], Lee et al. (2013) [3], and RNC et al. (2016) [4]. Since these methods adjust images solely at the level of pixel intensity distribution and do not address the physical essence of image formation, they struggle to balance contrast enhancement with color fidelity.

Another category of traditional methods is based on the Retinex theory, which can effectively address the shortcomings of histogram adjustment methods. The Retinex theory was first proposed by Land et al. (1971) [5] and has been continuously improved by researchers such as Jobson et al. (1997) [6], Rahman et al. (1996) [7], Palma-Amestoy et al. (2009) [8], Fu et al. (2015) [9], and Guo et al. (2017) [10]. These improvements have led to good enhancement results for low-illumination images. Nevertheless, issues such as noise amplification and over-enhancement of illumination still persist.

Overall, traditional methods are inherently limited by hand-crafted priors and distribution optimization. This leads to a trade-off between denoising, edge preservation, and color fidelity, making it difficult to completely eliminate artifacts such as blocky patterns, halos, noise amplification, and color casts. Furthermore, computational complexity increases sharply with image resolution, making it difficult to satisfy the demands for high-quality and high-temporal (real-time) enhancement. To address these limitations, the research focus has gradually shifted toward data-driven deep learning methods. By training on large-scale datasets, these networks can automatically learn the mapping from low-illumination (NCV) images to Daytime Color Visible (DCV) images, thereby overcoming the shortcomings of hand-crafted priors and seeking a new balance between contrast, detail, color, and processing speed.

(2) Deep Learning-Based Low-Illumination Image Enhancement Methods: Deep learning-based methods can be broadly categorized into supervised and unsupervised learning approaches, depending on the training datasets employed. Supervised learning refers to the use of strictly pixel-level paired NCV and DCV images, whereas unsupervised learning refers to the use of unpaired data.

Supervised learning methods, represented by the pix2pix model proposed by Isola et al. (2017) [11] and the LightenNet model proposed by Li et al. (2018) [12], benefit from strict semantic consistency constraints provided by paired images, thus generating transformation results with high geometric accuracy. However, the performance of these models is entirely dependent on the quality and scale of the paired image datasets. Most importantly, the datasets must feature pixel-level precise alignment to achieve ideal training results.

Although these methods can guarantee effective enhancement through explicit mapping relationships, they are extremely difficult to apply to outdoor scenarios. For instance, drones cannot acquire strictly paired data of the same area at different times. Even with post-processing techniques, it is difficult to overcome interference caused by dynamic changes in ground objects (such as moving vehicles and people) and parallax variations caused by different imaging angles when the drone captures the same scene at different times. Manual pixel-level registration and annotation of such image data by professionals would be prohibitively expensive, rendering the construction of large-scale paired datasets infeasible.

Due to the stringent requirements for dataset construction in supervised enhancement methods, researchers have attempted to train neural networks using unpaired datasets. Among these, Generative Adversarial Networks (GANs) are typical unsupervised learning models. They do not rely on explicit labels or strictly paired samples; instead, by designing specific loss functions and structural constraints, these models can achieve high-quality image enhancement and transformation.

The CycleGAN model proposed by Zhu et al. (2017) [13] was first applied to low-illumination image enhancement by Qu et al. (2019) [14]. This model utilizes a cycle-consistency loss function to ensure high similarity between the converted image and the original. However, the presence of cycle-consistency constraints and dual discriminators often leads to issues such as halos, loss of detail, and unstable training in the output images. To address the issues of detail degradation and color distortion in low-illumination images, Mao et al. (2023) proposed the Retinex-GAN model [15], which combines the classic Retinex image decomposition theory with the GAN mechanism. This unsupervised method brightens low-illumination images without relying on large paired datasets while also adapting to complex scenes.

Furthermore, to address the task of low-illumination image enhancement, some studies have proposed more lightweight GAN architectures. For example, UEGAN enhances image brightness and contrast by learning the illumination mapping relationship between low-illumination and normally lit images, achieving good results in natural image enhancement tasks [16].

Despite the good enhancement effects achieved by unsupervised methods such as the original GAN, CycleGAN, Retinex-GAN, and UEGAN using unpaired datasets, the transformed or enhanced images often exhibit abnormal spots, edge artifacts, and other color and semantic mutations. The primary reason is that using unpaired datasets means there is no structural or semantic correlation between the image pairs. Compared to paired images, these methods lack structural and semantic constraints between the image pairs. For example, the PSGAN model proposed by Liu et al. (2020) [17] for remote sensing image enhancement tends to generate artifacts and struggles to maintain consistency in ground object structures when processing remote sensing images with complex ground scenes.

(3) Methods for Fusion and Enhancement of NCV and TIR Images: For low-illumination environments, the imaging process inevitably results in the loss of a large amount of detail. It is difficult to achieve ideal information restoration and supplementation using a single enhancement algorithm or deep learning model. To address this issue, fusing NCV and TIR imagery to enhance image quality has become a mainstream solution. Early representative research includes the adaptive multi-scale TIR and NCV image fusion method proposed by Pei et al. (2024) [18]. Although TIR images can compensate for information lost during NCV imaging to some extent, they exhibit significant shortcomings in color reproduction and detail representation. Consequently, the visual quality of the fused images still struggles to match that of DCV images.

To further improve the structural representation capability of NCV and TIR images, Yang et al. (2024) proposed the HDCGAN model [19], which introduces Hyperspherical Directional Cosine (HDC) space to enhance semantic consistency during the image enhancement process. However, this model adopts unpaired training datasets. Due to the lack of an effective semantic consistency constraint mechanism in unpaired data, the generated enhanced images still suffer from problems such as missing detail information, color block artifacts, and structural contour distortion. Furthermore, although the CHITNet fusion network proposed by Du et al. (2025) [20] innovates in the information transformation mechanism for TIR and CV image fusion, this research primarily focuses on structural information transformation and lacks specific modeling for natural color restoration and visual consistency in the fused images.

In recent years, fusion methods based on deep learning models, such as DenseFuse (2019) [21] and NestFuse (2020) [22] proposed by Li et al., SeAFusion (2022) [23] proposed by Ma et al., and the more recently proposed RFN-Nest [24] and DDcGAN [25], have achieved significant progress in the structural fusion of TIR and CV images, effectively enhancing edge and texture information. However, these methods typically output grayscale fusion results or pseudo-color images. Their research focus is concentrated on structural information fusion rather than natural color reconstruction. Achieving natural and stable color restoration while maintaining fused structural information remains an important problem awaiting a solution in the current field of infrared and visible image fusion and enhancement.

Through the above comprehensive analysis, it can be found that various improved GAN models [26] are currently the most competitive models for enhancing image quality. Their most prominent advantage is that their training does not rely on strict paired datasets. However, precisely because unpaired datasets lack effective constraints on semantic consistency, the training results of GANs inevitably suffer from color and semantic distortions. To address this issue, this paper proposes the use of Weakly Paired Image Datasets (WPIDs) instead of unpaired datasets.

Beyond the development of the enhancement methods discussed above, the accurate evaluation of enhanced image quality and its effectiveness in downstream applications remains another critical concern. Existing image quality evaluation metrics have certain limitations; most metrics focus on perceptual experience (e.g., sharpness, color) and semantic consistency, failing to establish effective evaluation criteria for the critical issue of whether image structures are completely preserved. They also do not fully consider the practical application effects in downstream tasks such as image segmentation.

Currently, mainstream image transformation and enhancement quality evaluation methods primarily utilize three categories of metrics: (1) pixel-level error metrics; (2) perception-based metrics; and (3) no-reference image quality metrics.

(1) Pixel-Level Error Metrics: These primarily include MSE, PSNR, and SSIM. Mean Squared Error (MSE) [27] is the most fundamental metric in image quality assessment. Its

advantages lie in simple calculation and differentiability, making it convenient to use as a loss function during the training phase. However, MSE essentially only measures brightness error and struggles to reflect subjective quality in tasks such as image enhancement. Peak Signal-to-Noise Ratio (PSNR) [28] is the logarithmic form based on MSE, used to measure the ratio between signal and noise. The Structural Similarity Index (SSIM) [29] is one of the most widely used pixel-level structural perception metrics, integrating three dimensions: luminance, contrast, and structure, reflecting the perceptual similarity between images. It effectively supplements the shortcomings of PSNR, but still has certain blind spots regarding fine-grained differences in complex structures.

(2) Perception-Based Metrics: These primarily include FID, KID, and LPIPS. FID [1] is currently one of the most representative deep perception metrics, effectively measuring the realism and statistical consistency of images as a whole, and is widely used in image enhancement tasks. KID [30] uses Maximum Mean Discrepancy to provide an unbiased estimate of the difference in image distributions. Compared to FID, it is more stable, being especially suitable for small samples and scenes with large distribution deviations. In addition, KID avoids the numerical instability caused by matrix square root operations and has gradually been adopted as a supplement or alternative to FID in recent years. LPIPS [31] is based on pre-trained convolutional networks, evaluating visual consistency by comparing the distance between local features of images, capable of capturing fine-grained perceptual differences, making it suitable for evaluating visual improvement effects in tasks such as image enhancement and style transfer.

(3) No-Reference Image Quality Metrics: These primarily include NIQE and BRISQUE. NIQE [32] assesses the degree to which a given image deviates from the distribution of natural image statistics, but has stability issues in certain cases. BRISQUE [33] models the local spatial structure statistics of images, capable of detecting quality defects such as blur and noise, but is highly sensitive to low-quality images.

In summary, although the aforementioned metrics can evaluate the quality of enhanced images to a reasonable extent, they all suffer from certain shortcomings. The primary issues include the difficulty in assessing local distortions and color shifts, as well as their reliance on high-quality prior data. Furthermore, they cannot be used to evaluate the performance of generated images in downstream tasks (such as object detection and classification). To address this, Han et al. [34] proposed the RmAP metric. Its main idea is to compare the mAP (mean Average Precision) of object detection between real images and generated images, thereby reflecting the structural restoration quality of the generated images and their effectiveness in downstream tasks like object detection. However, the RmAP metric requires reliance on real annotation data and lacks the ability to reflect information regarding non-target areas (such as background textures), thus limiting its application scenarios.

To evaluate the performance of enhanced images in downstream tasks as accurately and objectively as possible, in this paper, we introduce an evaluation method based on the segmentation results of the Segment Anything Model (SAM). SAM is a universal vision segmentation foundation model proposed by Meta, possessing zero-shot segmentation capabilities [35]. Although it has not been specifically fine-tuned for UAV data, experiments show that it has relatively stable structural perception capabilities in remote sensing scenarios. We attempt to establish a stable and objective structural perception evaluation mechanism for enhanced image quality to solve the problem of accurately evaluating the enhancement effects of low-illumination images.

The main contributions of this paper are as follows:

(1) This work presents a Weakly Paired Image Dataset (WPID) construction strategy for UAV remote sensing image acquisition under low-illumination conditions, which addresses the challenge of building strictly paired datasets in outdoor scenarios.
(2) A HDCGAN+ image enhancement approach trained on WPID is introduced, which improves semantic consistency and stability throughout the image transformation process.
(3) We further develop a structural consistency evaluation framework built upon SAM segmentation outputs. Integrated with conventional perceptual metrics, this framework enhances the evaluation accuracy for practical performance in downstream remote sensing tasks.

2. Methods

2.1. Construction of WPID

To further improve the translation performance of GAN models, this study proposes the construction of a weakly paired image dataset to strengthen semantic consistency constraints at the training dataset level and enhance the learning performance of GAN models. For outdoor environments involving drone applications, it is extremely difficult to obtain datasets with strict pixel-level alignment; however, it is relatively easy to acquire datasets where the scene semantics are largely consistent and the images are roughly aligned. Suppose we use a UAV to acquire nighttime and daytime images of a specific area; this can be achieved by flying the same route over the area and capturing images once at night and once during the day. Although the UAV's positioning and camera attitude control may involve some error, and factors such as wind and other weather conditions may vary, making it impossible to guarantee that the two images taken at the same waypoint will be perfectly aligned, we can ensure that the two images will have at least 80% overlap in the absence of extreme weather changes. In this paper, we define images with essentially consistent semantic content and spatial overlap of at least 80%, but which do not achieve strict pixel-level alignment, as weakly paired images datasets (WPIDs) to distinguish them from strictly paired image datasets. As shown in Figure 1, it is evident that aside from significant differences in brightness and missing information in the low-illumination image, the two images exhibit excellent semantic consistency.



(**a**) (**b**)

Figure 1. Examples of weakly paired imagery acquired by a drone along the same flight path in the same area during daytime and nighttime; the data was collected on 30 June 2024 in Xianghe County, Hebei Province, using a DJI M300 drone platform (DJI, Shenzhen, China) equipped with a DJI XT2 thermal imaging camera (DJI, Shenzhen, China) (with an integrated visible-light lens); (**a**) shows the daytime drone imagery, and (**b**) shows the nighttime drone imagery.

The WPID proposed in this paper refers to a collection of image pairs captured at the same location but at different times; although these pairs contain some misalignment, they

are sufficient for training machine learning models. Generally, the cost of constructing such datasets is lower than that of precisely paired datasets, and they are suitable for unsupervised learning models that preserve structural information.

We used the DJI M300 drone platform equipped with a DJI XT2 thermal infrared imager (with a built-in visible-light lens) to collect data in Xianghe County, Langfang City, Hebei Province (as shown in Figure 2a), Hancheng Town, Tangshan City. Taking the data collection in Xianghe County as an example, the flight path planned on the remote controller is shown in Figure 2b. The collection area included various features such as roads, vegetation, water bodies, and buildings. Daytime images were captured at 13:00–15:00 in the afternoon, while nighttime images were acquired at 20:00–22:00 in the evening. Owing to the low-illumination conditions during nighttime imaging, the resulting images suffer from severe noise and exhibit certain blurring effects between targets and backgrounds.



(**a**) (**b**)

Figure 2. Construction of a weakly paired image dataset. (**a**) Shows the location where the WPID was collected (Lujiawu Village, Xianghe City, Hebei Province, China), with the red frame highlighting the core study area. (**b**) Shows the DJI drone controller and flight path planning, where the blue color code indicates the planned flight route.

Since the images were collected along the same flight path, an equal number of images were acquired at the same altitude during both daytime and nighttime. Specifically, each set of images captured in Xianghe consisted of 297 images, while each set captured in Hancheng consisted of 220 images. All acquired images are of reliable quality and were used to construct the WPID for the experiment. Given the input image size constraints of the original HDCGAN model, the dataset was uniformly cropped and resized to a resolution of 512 × 512. Figure 3 shows some examples from the dataset, where (a) represents a DCV image, (b) represents an NCV image, and (c) represents a nighttime TIR image. It should be noted that DCV (a) and NCV (b) constitute a typical weakly paired relationship.

During the data preprocessing stage, this study used Pix4Dmapper 2.0 (Pix4D SA, Lausanne, Switzerland) to correct lens distortion and performed global affine registration based on feature matching, thereby achieving basic geometric alignment between the daytime and nighttime images. However, due to differences in acquisition times between day and night, the presence of dynamic ground targets, and variations in UAV attitude and viewpoint, local parallax and pixel misalignment remain unavoidable. Therefore, this study defines image pairs that have undergone global registration but still exhibit local parallax and misalignment as "weakly matched images," to distinguish them from datasets that are strictly pixel-perfect.

Without preprocessing or feature matching using software such as Pix4Dmapper 2.0, images captured over the same scene may exhibit spatial misalignment due to differences in acquisition time (daytime versus nighttime) and slight variations in UAV position and attitude. Under normal weather conditions and taking the Xianghe dataset as an example, where the flight altitude during data acquisition was approximately 120 m, empirical observations suggest that the spatial offset between images captured at the same waypoint

is typically within 10%. This value is used to characterize the maximum offset of the weakly paired images considered in this study, rather than being a strict threshold imposed during training or evaluation.



Figure 3. Example of a WPID: (**a1**–**c1**) are scene images acquired at a flight altitude of 120 m in Xianghe County, Hebei Province; (**a2**–**c2**) are scene images acquired at a flight altitude of 50 m in Hancheng County, Tangshan City, Hebei Province. Specifically, (**a1**,**a2**) are daytime visible images (DCV), (**b1**,**b2**) are nighttime visible images (NCV), and (**c1**,**c2**) are nighttime thermal infrared images (TIR).

In this study, we used an XT2 camera(DJI, Shenzhen, China) with a baseline distance of approximately 4 cm between its visible lens and thermal infrared lens. At a flight altitude of 120 m, this corresponds to a deviation of approximately 0.25 pixels. When the images are downsampled to a resolution of 512 × 512, this deviation can be considered negligible for applications that do not require strict registration. Therefore, the spatial offset between the nighttime thermal infrared images (c1, c2) and the corresponding nighttime visible images (b1, b2) is minimal, as shown in Figure 3.

The final dataset comprises multimodal, roughly registered images in DCV, NCV, and TIR (as shown in Figure 3). With its rich structural information and modal diversity, the dataset provides robust support for subsequent weakly paired enhancement experiments. It should be noted that the weakly paired nature of the dataset is primarily intended to simulate the spatial shifts and parallax errors that may occur under real-world UAV

data collection conditions across different time points, thereby enhancing the model's adaptability in practical application scenarios.

Figure 4 illustrates the contrast between the unpaired image samples used in the original HDCGAN and the weakly paired image samples constructed in this paper. Unlike fully unpaired images, WPID preserves scene-level spatial correspondences while allowing for limited spatial shifts, thereby providing structural constraints for model training.



(**a**) (**b**)

Figure 4. 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. [19]; (**b**) WPI data used in this study.

The experiment utilized a DJI M300 drone equipped with an XT2 thermal infrared camera to capture dual-modal imagery during the day (1:00–3:00 PM) and at night (8:00–10:00 PM). Prior to model training, the acquired data underwent the following preprocessing steps:

(1) Use Pix4D to perform image stitching on the NCV, TIR, and DCV images captured by the drone. If the drone's flight altitude is too low or the camera lens exhibits severe distortion, apply lens distortion correction to all images before proceeding with the stitching process;
(2) Perform global geometric registration between NCV and TIR, and between NCV and DCV, based on feature matching, to ensure that images acquired at different times are essentially aligned in pixel space;
(3) Crop the original image to a size of 512 × 512 pixels;
(4) Apply non-local mean (NLM) filtering to thermal infrared images to remove noise.

It can be seen that, following the aforementioned preprocessing, the WPID has essentially achieved pixel-level alignment for static ground features captured by drones; alignment cannot be considered only for dynamic or otherwise changing targets (such as moving people or vehicles).

2.2. HDCGAN+ Model Based on WPID

To further improve the image enhancement performance of GAN models for low- illumination images, we propose a training strategy using a WPID, which primarily includes the following two improvements:

The improvements in this study are primarily reflected in two aspects, rather than modifications to the network architecture:

(1) Replace the original unpaired dataset with a WPID to enhance semantic consistency constraints at the data level;
(2) By adjusting the input fusion method, enable LRRNet to perform adaptive color reconstruction based on luminance channel fusion, thereby generating input images better suited for processing by HDCGAN.

It should be noted that this paper does not modify the network architecture, loss functions, or training mechanisms of LRRNet [36] or HDCGAN. Therefore, the method proposed in this paper constitutes an improvement in training strategies and input construction methods, rather than an improvement in the network architecture.

2.2.1. An HDCGAN Training Framework Based on WPID

To improve the quality of NCV, Yang et al. [19] fused it with simultaneously acquired TIR images and then used an HDCGAN model to transform the fused images, aiming to achieve DCV-like results. Yang et al. [19] results demonstrated that HDCGAN achieved better enhancement effects than the then-state-of-the-art (SOTA) models. To improve semantic consistency before and after transformation, the authors introduced the concept of Hyperdimensional Computing (HDC). As mentioned earlier, since the model was trained using unpaired datasets, it still could not overcome the semantic bias in the transformed images caused by semantic inconsistencies within the image pairs in the dataset. Building upon HDCGAN, this paper introduces a WPID training strategy, adjusts the input resolution to 512 × 512, and optimizes data reading and organization methods. These modifications enable the model to train using image datasets with strong semantic correspondence (i.e., WPID), thereby enhancing the structural consistency and stability of the enhanced images. To enable the model to effectively utilize this "semantically consistent but spatially incompletely aligned" supervisory information, this paper adjusts the training data organization and input strategy, allowing the generator to receive weakly paired images in the order of corresponding scenes. This establishes cross-modal semantic associations during training, rather than purely random mapping relationships.

Furthermore, since weakly paired images are not spatially aligned, the model requires greater structural robustness. Therefore, while preserving the original network architecture, this paper modifies the data distribution learned by the model through WPID constraints, causing the generator to be simultaneously influenced by both adversarial loss and weak semantic consistency constraints during the optimization process, thereby mitigating the structural drift issues commonly encountered in unpaired training.

HDCGAN uses unpaired images with a resolution of 256 × 256. In this study, we adjusted the input image size to 512 × 512 by modifying the parameters to preserve the fine-grained structural information in the UAV images. We also adjusted the model's data reading mechanism to enable sequential data reading during training. To distinguish the original HDCGAN from the version described in this paper—which is trained on WPID and operates under high-resolution input conditions—we refer to this training framework as HDCGAN+.

2.2.2. LRR-FusionColor Color Reconstruction

Before inputting the images into the HDCGAN+ model for enhancement, we first fuse the NCV images with the simultaneously acquired TIR images to compensate for the loss of detail in the NCV images during nighttime imaging. To address the limitations of traditional TIR and CV image fusion methods—which typically produce only grayscale images or suffer from severe color restoration distortion—this paper adopts the Low-Rank Representation Fusion Network (LRRNet) [36] proposed by Li et al. as the structural fusion model. Based on the grayscale fusion image output by the low-rank feature extraction network, we propose a color reconstruction method called LRR-FusionColor that utilizes two color spaces. By extracting the color components from the visible image and recombining them with the grayscale fusion results from the NCV and TIR, we achieve a fused image rich in structural details and with natural, consistent color. The generated fused color image serves as the input to the HDCGAN+ model for subsequent low-illumination image

enhancement training and inference, thereby forming a complete "fusion-enhancement" processing workflow.

It should be noted that LRR-FusionColor does not introduce a new structural fusion network, but rather performs a deterministic color reconstruction process based on the fused luminance output by the existing LRRNet. The brightness component (Y) of the visible image is input into the network for structural fusion with the thermal infrared image, and the adaptive color reconstruction process proposed in this paper is then applied to the resulting fused brightness. Although LRRNet has significant advantages in modeling image structure, its fused output is a grayscale image that fails to preserve the color information present in the visible image. This limitation restricts its generalization capabilities in practical tasks such as multimodal perception and object detection.

To this end, this paper proposes the LRR-FusionColor fusion method (as shown in Figure 5), which consists of two collaborative submodules: first, a structural extraction module that, supported by low-rank sparse modeling theory, uses LRRNet to extract redundant primary structures and modal details from TIR and CV; second, the color reconstruction and restoration module (FusionColor), which combines the Cb/Cr color channels of the CV image with the fused grayscale luminance image to reconstruct the RGB color image in the YCbCr color space.

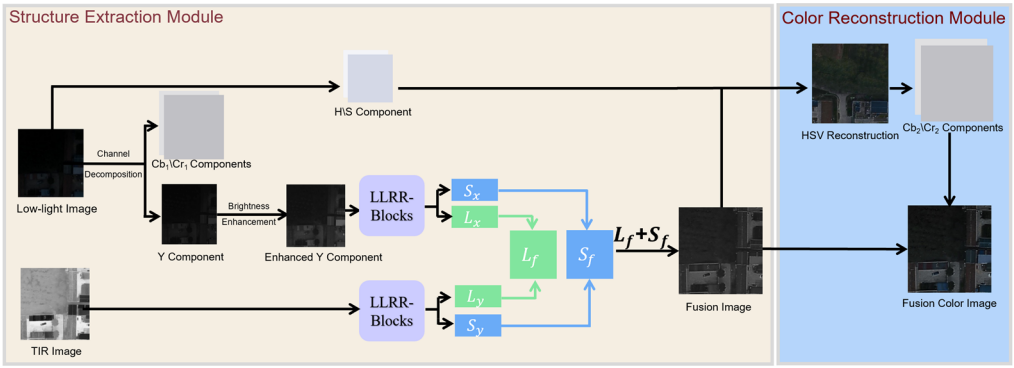


Figure 5. Architecture of the proposed LRR-FusionColor framework.

In the structure extraction module, YCbCr effectively decouples luminance (Y) from chrominance (Cb, Cr), while the HSV color space facilitates independent manipulation of color and luminance. The color reconstruction process is as follows: first, the visible image is converted separately into YCbCr and HSV color spaces to extract the original luminance and chrominance components; second, the V (luminance) channel of the HSV space is replaced with the fused high-frequency grayscale image (Lf + Sf), and the reconstructed image is transformed back to RGB space, then further converted to YCbCr space to obtain structure-guided enhanced chrominance components Cb_{HSV}/Cr_{HSV}.

Although replacing the V channel with a fused grayscale image can introduce clear textural details, when directly converted back to RGB space, the reconstructed image often exhibits spectral distortion or inharmonious color blocks due to the nonlinear coupling between drastic changes in luminance and the original color information. To address this issue, this paper does not use the RGB image reconstructed from HSV directly as the final result, but rather employs it as guidance for chrominance enhancement. We perform a linearly weighted fusion of the enhanced chrominance components (Cb_{HSV}, Cr_{HSV}) obtained from the HSV path with the natural chrominance components of the original visible (Cb_{orig}, Cr_{orig}):

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

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

In this context, $\mathrm{Cb}_{\mathrm{orig}}$ and $\mathrm{Cr}_{\mathrm{orig}}$ represent the chrominance components of the original visible in the YCbCr color space (representing natural colors); $\mathrm{Cb}_{\mathrm{HSV}}$ and $\mathrm{Cr}_{\mathrm{HSV}}$ represent the chrominance components of the HSV color space after replacing the V channel and converting to YCbCr (containing structure-guided color information); $\mathrm{Cb}_{\mathrm{fused}}$ and $\mathrm{Cr}_{\mathrm{fused}}$ are the final chrominance channels; α (0–1) is a weighting coefficient used to balance the original and enhanced colors. α is not a fixed constant but rather a pixel-level adaptive function, whose value is adaptively determined by the degree of deviation of the pixel's chrominance from neutral color. In the YCbCr color space, neutral color corresponds to a chrominance value of 128, and the magnitude of the chrominance deviation from this value reflects the color saturation of the pixel. Therefore, it is defined that:

$$\alpha(\mathrm{x}, \mathrm{y}) = \frac{\left|cb_{orig}(x, y) - 128\right| + \left|cr_{orig}(x, y) - 128\right|}{2 \times 128} \quad (3)$$

where (x, y) represents the pixel position, and α(x, y) $\in$ [0, 1]. This definition has clear physical and perceptual significance: when a pixel approaches grayscale (i.e., its chromaticity value approaches the neutral color 128), the α value approaches 0; at this point, the original visible color information is prioritized to avoid unnatural color shifts in low-saturation regions. When a pixel has high color saturation, the α value approaches 1, thereby allowing more structural information to participate in chromaticity reconstruction to enhance the expression of structural details. This function is monotonically increasing within the [0, 1] interval, enabling adaptive adjustment based on color intensity: it suppresses excessive enhancement in low-saturation regions and allows stronger structural transfer in high-saturation regions, thereby achieving a balance between visual naturalness and structural enhancement. Compared to fixed-weight strategies, this adaptive mechanism avoids the uncertainty associated with manual parameter tuning across different scenes and reduces the risk of color distortion caused by excessive enhancement.

Since the value of α is calculated entirely from pixel chromaticity values using a deterministic mathematical formula and does not rely on empirical settings or machine learning processes, it offers clear reproducibility. This uniform definition was applied across all experiments without separate adjustments for different datasets or scene objects (such as vegetation, roads, buildings, or water bodies), thereby ensuring the method's cross-scene consistency and stability. Consequently, this weighting mechanism does not lead to overfitting for specific object types and ensures good generalization capabilities across different scenes.

In Equations (1) and (2), $\mathrm{Cb}_{\mathrm{orig}}/\mathrm{Cr}_{\mathrm{orig}}$ preserves the color fidelity of natural scenes but lacks detail, whereas $\mathrm{Cb}_{\mathrm{HSV}}/\mathrm{Cr}_{\mathrm{HSV}}$ incorporates the clear structural information introduced by the fused grayscale image but may result in local color shifts. Through the aforementioned adaptive weighted fusion, color blocks and artifacts generated during the HSV color reconstruction process can be effectively suppressed, while structural contrast is enhanced. Finally, by combining the fused luminance channel (Yfused), a fused image is reconstructed that possesses both structural clarity and natural color consistency. The HSV and YCbCr color spaces each have distinct advantages: the HSV color space better preserves human visual consistency and enhances the visual contrast of structural details, while the YCbCr color space effectively separates luminance and chrominance information, thereby maintaining natural color distribution. This paper combines the strengths of both color spaces by using adaptive weights to fuse raw chrominance with enhanced chrominance, thereby preserving structural clarity while avoiding color distortion. This design does not increase the complexity of structural fusion but instead executes an independent color reconstruction process after structural fusion is complete.

Through this adaptive color reconstruction mechanism, the proposed LRR-FusionColor method not only retains the structural representation advantages of LRRNet but also achieves spatially adaptive color restoration, thereby significantly improving the visual quality and practical value of the fused images. The generated fused color images serve as input to the HDCGAN+ model for subsequent low-illumination image enhancement training and inference processes.

2.3. Experimental Setting

2.3.1. Experimental Environment

All experiments in this study were conducted on a DELL PowerEdge T640 tower server. The server is equipped with the following key hardware specifications: four Intel Xeon Silver 4116 processors (Intel Corporation, Santa Clara, CA, USA) with a base clock speed of 2.10 GHz; 128 GB of system memory, consisting of four 32 GB DDR4 memory modules operating at 2666 MHz; and two NVIDIA GeForce RTX 4090 D graphics processing units (GPUs), each equipped with 24 GB of GDDR6X video memory. All training and evaluation of deep learning models were performed on this server using the PyTorch (version 1.10.0) deep learning framework.

To evaluate the computational efficiency of the method, this paper further tested the inference time and GPU memory usage of the model under different input resolutions. Specifically, under the same hardware conditions, we measured the average inference time per image for the HDCGAN+ model with 256 $\times$ 256 and 512 $\times$ 512 input resolutions, and recorded the GPU memory usage. All experiments were conducted on an NVIDIA RTX 4090D GPU.

2.3.2. Evaluation Metrics

This paper introduces Meta AI's Segment Anything Model (SAM) [35] to extract structural information from images using image segmentation algorithms. SAM is a general-purpose segmentation foundation model trained on large-scale visual data that can generate high-quality segmentation masks for any input image without the need for task-specific retraining. Its core architecture consists of an Image Encoder, a Prompt Encoder, and a Mask Decoder, which automatically partitions structural regions of an image by extracting global visual features and predicting pixel-level masks.

SAM is trained on over a billion segmentation masks and demonstrates strong cross-scenario generalization capabilities. In this paper, SAM is not used for semantic classification of objects, but solely for extracting the boundaries of structural regions in images, thereby establishing an evaluation metric for structural consistency. Since SAM's segmentation relies primarily on structural boundaries and textural variations in the image rather than training on specific object categories, it can reliably reflect structural information in the image. Furthermore, all comparison methods perform segmentation using the same SAM model and parameters, ensuring the fairness and comparability of the evaluation results.

This paper directly employs a publicly available pre-trained SAM model for zero-shot segmentation. Since the SAM training dataset includes drone imagery, and to ensure the generality and reproducibility of this evaluation method, we did not perform any additional fine-tuning. Based on the segmentation masks extracted by SAM, we construct a structural consistency evaluation metric system to quantify the performance of enhanced images in terms of structural preservation, thereby more accurately assessing their performance in subsequent tasks.

Based on the SAM segmentation framework, this paper defines the following three metrics:

(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:

$$\Delta N = \mid N_{SAM}^{gen} - N_{SAM}^{ref} \mid \tag{4}$$

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:

$$\Delta A = \left| A_{Gen}^{SAM} - A_{ref}^{SAM} \right| \tag{5}$$

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) Block Intersection-over-Union (BIoU) distribution. Inspired by object detection IoU, the BIoU distribution calculates the average IoU of block-level matching pairs between the enhanced image and the reference image. A higher BIoU value indicates greater consistency between the structure's position and outline and those of the ground-truth image. The calculation formula is:

$$BIoU(A, B) = \frac{\mid A \cap B \mid}{\mid A \cup B \mid} \tag{6}$$

The average IoU is calculated for all block-wise matches between the enhanced image and the reference image, serving as a measure of structural boundary overlap. A higher BIoU indicates that the structural positions and contours of the enhanced image align more closely with those of the ground-truth image.

3. Experimental Results and Analysis

Based on the method described in Section 2, this section conducts experiments to compare the proposed method with representative existing methods in terms of both metrics and visualization, including: (1) conducting LRR-FusionColor image fusion experiments to compare the color fusion results of nighttime thermal infrared images and visible images; (2) enhancing the fused images using HDCGAN+ and comparing the quality of the enhanced images; and (3) processing the enhanced images using the SAM semantic segmentation algorithm, calculating an evaluation metric system based on the SAM segmentation results, and comparing the performance of the proposed algorithm with that of representative existing algorithms in practical applications.

3.1. Comparison of Experimental Results

3.1.1. CV and TIR Image Fusion and Evaluation

It should be noted that the focus of this study is not on proposing a new structural fusion network, but rather on developing an adaptive color reconstruction method (LRR-FusionColor) to address the color restoration of fusion results. Specifically, the brightness component (Y) is first extracted from the visible light image and structurally fused with the nighttime thermal infrared image using LRRNet to obtain the fused brightness component (Y_fused). Subsequently, based on the same fused luminance component, color restoration is performed using typical color space reconstruction methods such as HSV, Lab, and YCbCr, as well as the LRR-FusionColor method proposed in this paper. This ensures that all methods are based on the same structural information, allowing for a direct comparison of the differences in the color restoration strategies themselves.

To objectively quantify the performance of each method, this paper uses three metrics—Peak Signal-to-Noise Ratio (PSNR) [28], Structural Similarity Index (SSIM) [29], and Information Entropy [37]—to evaluate the fusion results. The statistical data are shown in Table 1 (comparison of quality evaluation metrics for color-fused images from different methods; bold text indicates the best result, and underlined text indicates the second-best result).

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**

Taking all the metrics in Table 1 into account, it can be seen that the FusionColor metric proposed in this paper yields the best results. It should be noted that the Fusion method outputs grayscale fusion results, which allows for the direct calculation of PSNR and SSIM metrics using nighttime thermal infrared (TIR) images. In contrast, the Lab, HSV, YCbCr, and the FusionColor method proposed in this paper all perform color reconstruction based on the fusion results, outputting RGB three-channel color images. Since these differ from single-channel TIR images in terms of channel dimension, PSNR_TIR and SSIM_TIR metrics were not calculated for them. The FusionColor method achieved the second-best result on PSNR_NCV and the best result on SSIM_NCV, indicating that it preserves the brightness and structure of the original nighttime visible-light image well. Although the FusionColor method weakens brightness consistency by adding HSV color information to the YCbCr basis, it yields better color and visual effects, as evidenced by its optimal performance on Entropy.

Furthermore, to further validate the performance of the proposed LRR-FusionColor method against various existing deep learning fusion methods, we performed fusion processing on the experimental data using representative infrared–visible light fusion models, including DenseFuse [21], NestFuse [22], SeAFusion [23], RFN-Nest [24], and DDcGAN [25]. Since these methods typically fuse structural information between infrared images and the luminance component of visible light, their output is often a grayscale image. To ensure comparability with our method, we treated the fusion results of these methods as the luminance component during the experimental process. We then reconstructed the image by combining this with the chrominance channel of the visible light image using

the YCbCr color space, thereby obtaining a color fusion result for quantitative evaluation. Table 2 presents the quantitative evaluation results of three representative deep learning fusion methods compared to our method.

Table 2. Quantitative Comparison Results of Deep Learning Fusion Methods (bold values denote the best performance, while underlined values represent the second-best performance).

Method	PSNR_NCV ↑	SSIM_NCV ↑	Entropy ↑
DenseFuse	12.2889	0.4396	7.1569
NestFuse	9.4209	0.3544	7.1604
SeAFusion	7.5832	0.2677	<u>7.1728</u>
RFN-Nest	<u>16.2036</u>	**0.5714**	6.1819
DDcGAN	12.6792	0.4918	**7.4705**
FusionColor	**20.4789**	<u>0.5137</u>	5.9227

As shown in Table 2, the LRR-FusionColor method proposed in this paper achieved optimal or near-optimal results on both the PSNR_NCV and SSIM_NCV metrics, significantly outperforming deep learning fusion methods such as DenseFuse, NestFuse, SeAFusion, RFN-Nest, and DDcGAN. This demonstrates that the method can more effectively restore natural colors while preserving structural information.

In contrast, methods such as DenseFuse, NestFuse, and SeAFusion primarily focus on the fusion of structural information between infrared and visible light images. Their fusion results typically aim to enhance textures and edges, resulting in certain limitations in color reconstruction. RFN-Nest performs well on the SSIM_NCV metric, indicating strong capabilities in structure preservation; however, its PSNR_NCV and overall visual quality still fall short of the method proposed in this paper. DDcGAN achieves the highest value on the Entropy metric, suggesting that its fusion results contain a high information content regarding texture details, but may simultaneously introduce a certain degree of color bias or noise.

Overall, the LRR-FusionColor method proposed in this paper achieves a better balance between structural preservation and color restoration. It effectively enhances the natural color consistency of the fused image while ensuring the expression of structural information, thereby achieving superior overall performance.

3.1.2. HDCGAN+ Image Enhancement Experiment

We trained HDCGAN+ using the WPID dataset. Since the model's loss function failed to converge, we selected the training iteration with the best enhanced images based on visual inspection as the final result.

To quantitatively evaluate the output results of all algorithms, we first employed the commonly used image enhancement evaluation metrics Fréchet Inception Distance (FID) [1] and Kernel Inception Distance (KID) [30] to assess the differences in feature distribution between the enhanced images and the daytime CV images. The specific calculation results are shown in Table 3 (bolded values represent the optimal results, and underlined values represent the next-best results).

Based on the combined performance of the FID and KID metrics, HDCGAN+ significantly outperforms other algorithms, achieving the lowest FID and KID values. This indicates that the generated images most closely resemble the ground-truth images in terms of feature distribution, demonstrating excellent enhancement results. Among traditional algorithms, HDCGAN performs best, with metrics second only to those of HDCGAN+; CycleGAN and EnlightenGAN each excel in different metrics. RetinexFormer performed the worst, with both metrics far exceeding those of other algorithms, indicating the lowest match between generated and real images. In summary, through observation and

comparison, the introduction of the weak pairing strategy in HDCGAN+ not only addresses the high cost and poor robustness associated with acquiring exact-pairing data but also compensates for the weakness in structural preservation during unpaired training. This significantly enhances the model's generalization ability and practical applicability, demonstrating strong potential in the field of remote sensing image enhancement.

Table 3. 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**

3.1.3. Analysis of the Computational Efficiency of HDCGAN+

Given that the use of high-resolution images as input can significantly impact a model's performance in real-world applications, we compared the computational overhead and operational efficiency of HDCGAN+ when using training datasets of different resolutions. Additionally, we benchmarked it against representative existing models (such as UEGAN) to provide fellow researchers with a clearer understanding of its performance. This paper quantifies the computational efficiency metrics of different models on the test set, including average inference time, inference frame rate (FPS), parameter size, and peak GPU memory usage. To ensure a fair comparison, both UEGAN and HDCGAN+ used input images with a resolution of 512 × 512. The experimental results are shown in Table 4.

Table 4. Comparison of Computational Efficiency Among Different Models.

Model	Resolution	Params (M)	GPU Memory (MB)	Time (ms/Image)	FPS
UEGAN	512 × 512	4.16	4468	5.25 ± 1.33	190
HDCGAN+	256 × 256	70.74	3572	42.78 ± 15.51	23.37
HDCGAN+	512 × 512	70.74	3572	153.69 ± 31.04	6.51

The results show that at a resolution of 512 × 512, the average inference time for HDCGAN+ on the test set is 153.69 ± 31.04 ms/image, corresponding to 6.51 FPS. When the input resolution is reduced to 256 × 256, the average inference time decreases to 42.78 ± 15.51 ms/image, and the FPS increases to 23.37. In contrast, UEGAN has an average inference time of 5.25 ± 1.33 ms/image at a resolution of 512 × 512. Since UEGAN primarily focuses on brightness and contrast enhancement for low-illumination images, its network architecture is relatively lightweight, resulting in higher inference speeds. In contrast, HDCGAN+ requires the integration of thermal infrared information for cross-modal structural reconstruction and detail restoration, leading to relatively higher computational complexity; however, it is capable of generating enhanced images with richer structural information and higher semantic consistency.

It should be noted that there are certain differences in the design objectives of UEGAN and HDCGAN+. UEGAN primarily focuses on brightness and contrast enhancement for low-illumination images, with its enhancement process centered on adjusting the overall brightness distribution of the image; consequently, its network architecture is relatively lightweight, resulting in faster inference speeds. In contrast, HDCGAN+ aims not only to enhance image brightness but also to perform cross-modal structural reconstruction by

integrating thermal infrared information, thereby restoring more details and structural information. Consequently, while HDCGAN+ exhibits higher computational complexity than lightweight enhancement models, it is capable of generating enhanced images with richer structural details and stronger semantic consistency.

Overall, although HDCGAN+ is slower in inference speed compared to lightweight enhancement models, its computational efficiency still holds acceptable engineering value for offline processing and subsequent analysis of UAV remote sensing imagery.

3.1.4. Evaluation of Structural Consistency Based on SAM Segmentation

The experiment selected LRR-FusionColor fused images, images enhanced using CycleGAN [13], images enhanced using EnlightenGAN [38], images enhanced using RetinexFormer [39], images enhanced using HDCGAN [19], and images enhanced using HDCGAN+ as the control group, while using the original nighttime images and real daytime images as references. SAM was used to perform full-image segmentation on each image, yielding a set of semantic patches. The segmentation results were then evaluated using two metrics: the difference in the number of patches and the difference in coverage.

To ensure fairness in the results, the following standards were uniformly applied to all images: (1) a resolution of 512 × 512 was used for all images; (2) the SAM algorithm used fixed parameters without separate tuning; (3) all metrics were calculated based on the complete test set.

Using the evaluation metrics proposed in this paper, statistical calculations were performed on the segmentation results of each image, yielding the results shown in Table 5.

Table 5. 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	ΔN ↓	ΔA ↓	BIoU (A, B) ↑
Night	91.6857	0.3588	0.2350
LRR-FusionColor	58.9360	0.3961	0.2855
CycleGAN	61.1111	0.3158	0.2983
EnlightenGAN	**29.7239**	<u>0.2530</u>	0.2474
RetinexFormer	50.6566	0.3567	0.2278
HDCGAN	67.9273	0.2880	<u>0.3048</u>
HDCGAN+ (Ours)	<u>29.7800</u>	**0.2458**	**0.3325**

Table 5 lists the quantitative results of various methods across different structural consistency metrics. All metrics are based on statistical values derived from the images generated by each method; the mean values were calculated after removing outliers using the interquartile range (IQR), with a sample size of approximately 300 images. The results show that HDCGAN+ achieved the best performance in both the difference in block coverage area (ΔA) and the BIoU metric, and also performed close to optimal in the difference in block count (ΔN) metric. This indicates that its generated results are most similar to the daytime reference images in terms of regional coverage consistency, semantic object preservation, and structural boundary representation.

A further comparison of the traditional evaluation metrics (FID, KID) in Table 3 with the structural consistency metrics in Table 5 reveals that the evaluation framework proposed in this paper can more accurately reflect the performance of different methods in practical applications at the structural level. For example, in Table 3, HDCGAN ranks second in terms of FID and KID; however, in Table 5, its performance on the two structural metrics—difference in the number of patches (ΔN) and difference in patch coverage area (ΔA)—lags behind methods such as CycleGAN and EnlightenGAN. This indicates that

traditional metrics primarily measure the similarity in overall feature distribution between generated and real images, but struggle to reflect the image's ability to preserve structural details, such as the retention of semantic objects and the alignment of boundary contours.

In contrast, the structural consistency metrics proposed in this paper provide a multi-dimensional quantitative evaluation of the structural preservation capability of enhanced images through the difference in the number of blocks, the difference in coverage area, and boundary and region matching metrics based on segmentation results. Furthermore, the optimal performance achieved by HDCGAN+ under the new evaluation framework, as shown in Table 5, is consistent with its ability to better preserve semantic objects and structural details in the visual results. Thus, the proposed structural consistency evaluation framework can more objectively and comprehensively assess the structural preservation capability of enhanced images in practical remote sensing applications compared to traditional metrics.

3.2. Visual Comparison of Experimental Results

3.2.1. Visual Comparison of CV and TIR Fusion Results

In our experiments, we employed four representative fusion methods for comparison with our proposed method, including the LRRNet grayscale fusion method (Fusion) [21], the HSV color space fusion method (HSV) [22], the LAB color space fusion method (Lab) [23], and the YCbCr color space fusion method (YCbCr) [40]. The visual output results of the experiments are shown in Figure 6. The numbers 1–5 on the left represent five different experimental samples, while the vertical columns display the processing results of the five methods, including the original images (the left two columns are the NCV and TIR images, respectively) and the method proposed in this paper.

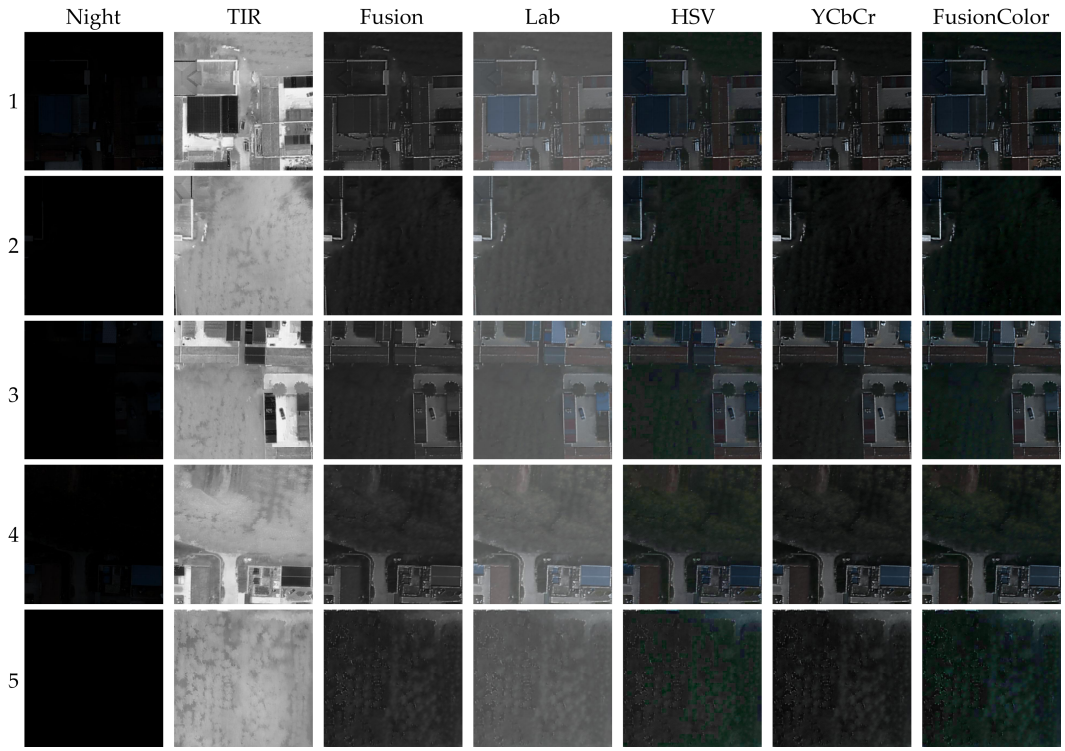


Figure 6. Visual comparison of fusion results obtained using different color restoration methods. The numbers 1–5 on the left correspond to five different experimental regions. From left to right: NCV Images, NTIR Images, Fusion results, Lab fusion results, HSV fusion results, YCbCr fusion results, and the results of the proposed FusionColor method.

As shown in Figure 6, the Fusion grayscale image, which incorporates TIR, exhibits significantly improved contrast and edge sharpness compared to the original NCV, validating the effectiveness of LRRNet in processing multimodal information. Although the HSV and Lab methods achieve color reconstruction, they suffer from color distortion, regional color bias, and loss of detail in low-illumination areas; in particular, the HSV method compromises structural integrity due to missing pixels. The YCbCr method preserves the basic color style but suffers from low brightness and insufficient color information; in contrast, the FusionColor method performs best in areas such as vegetation, roads, and building edges, not only restoring natural colors and ensuring uniform brightness distribution but also effectively reducing color block artifacts and discontinuities, thereby balancing structural preservation and color consistency.

Furthermore, to compare the differences between the method proposed in this paper and representative deep learning methods, we selected DenseFuse, NestFuse, SeAFusion, RFN-Nest, and DDcGAN for comparative experiments, with the results shown in Figure 7. As shown in the figure, the aforementioned methods can enhance the structural information of nighttime images to some extent, making road and building outlines clearer; however, their fusion results primarily focus on enhancing structural details, and there are still certain shortcomings in color restoration. For example, in certain areas, the results generated by these methods appear generally grayish or exhibit low color saturation, making it difficult to restore a natural color distribution consistent with daytime visible-light images; simultaneously, in some low-illumination areas, local details still suffer from insufficient contrast or lack of clear texture representation.



Figure 7. Visual comparison of deep learning fusion methods. The numbers 1–6 on the left correspond to six different experimental regions. From left to right, the images show: nighttime CV images, nighttime TIR images, the DenseFuse fusion result, the NestFuse fusion result, the SeAFusion fusion result, the RFN-Nest fusion result, the DDcGAN fusion result, and the result of the FusionColor method proposed in this paper.

In contrast, the LRR-FusionColor method proposed in this paper is able to more effectively restore the natural colors of a scene while preserving structural details. For example, in the first and fourth sets of examples, the colors of building roofs and road areas are closer to those in daytime visible-light images; in the second and fifth sets of examples, the structural details in dark areas are enhanced while avoiding obvious color distortion. Overall, the FusionColor method achieves a good balance between structural preservation and color restoration, resulting in enhanced images with a more natural visual appearance.

A comprehensive comparison with the aforementioned methods shows that the proposed LRR-FusionColor method, based on unified structural fusion results, can effectively improve color reconstruction quality. It mitigates the issues of color distortion and structural inconsistency found in traditional fusion methods to a certain extent, thereby further validating the effectiveness of the proposed color reconstruction strategy.

3.2.2. Visual Comparison of Image Enhancement Experiment Results

The experiment compared the original nighttime images prior to conversion, the enhanced nighttime images, and four algorithms: CycleGAN [13], EnlightenGAN [38], RetinexFormer [39], and HDCGAN [19]. Figure 8 displays sample images from seven locations in the experimental area. It is evident that the images generated by HDCGAN+ exhibit the closest color and structural resemblance to daytime visible-light images, clearly reproducing vegetation, buildings, and roads. Furthermore, while EnlightenGAN demonstrates good structural and detail reproduction capabilities, there is a significant discrepancy in color (particularly in vegetation) compared to the original image, resulting in noticeable color differences.

3.2.3. Comparison of Experimental Results for Enhanced Image Segmentation Based on the SAM Model

To validate SAM's ability to represent structural features in UAV remote sensing imagery, this paper presents the segmentation results obtained by applying the SAM model to images generated by various enhancement methods, as shown in Figure 9. It can be observed that SAM effectively extracts major structural regions such as roads, buildings, and vegetation. This indicates that SAM can accurately reflect changes in image structure, thereby providing a reliable foundation for structural consistency evaluation. Figure 9 displays the segmentation results obtained using the SAM algorithm on LRR-FusionColor fused images, images enhanced using five algorithms (CycleGAN, EnlightenGAN, RetinexFormer, HDCGAN, and HDCGAN+), and daytime DCV images. The segmentation results of the original nighttime images exhibit a small number of segments and blurred boundaries, making it difficult to effectively identify target structures; structural regions (such as roads, buildings, and vegetation) in the daytime ground truth images were effectively restored after enhancement of the nighttime images, with good detail continuity.

However, since the SAM model was not specifically trained on the dataset used in this study, a small number of misclassifications and omissions occurred; the segmentation results of the LRR-FusionColor color-fused images showed a significant improvement over the original nighttime images, validating the necessity of fusion and the advantages of the method's details. The segmented regions in the HDCGAN+ enhanced images have clear contours, and their number and distribution are highly similar to those in the daytime reference image, outperforming HDCGAN and other methods. Although EnlightenGAN maintains structure well, it suffers from excessive noise, particularly in vegetation areas. The experimental results indicate that this optimized architecture significantly improves model generation quality and semantic fidelity, demonstrating good adaptability and interpretability in the current experimental scenario.

Overall, HDCGAN+ not only performs well on traditional metrics but also yields segmentation results that are closer to the daytime CV results, further demonstrating that this method possesses superior structural representation capabilities for future remote sensing applications.



Figure 8. Visual comparison of enhancement results between HDCGAN+ and other state-of-the-art image enhancement methods. The numbers 1–7 on the left correspond to seven different experimental regions. 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.



Figure 9. Visualization of segmentation masks produced by different image enhancement methods (colors represent different segmented regions rather than specific land-cover categories). The numbers 1–9 on the left correspond to nine different experimental regions. 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.

4. Discussion and Analysis

This study addresses issues such as the poor quality of UAV remote sensing imagery in low-illumination environments, the data dependency conflicts of existing enhancement methods, and the limitations of evaluation systems. It proposes technical solutions, including the LRR-FusionColor multimodal fusion method, the use of a weak pairing mechanism to improve the enhancement performance of HDCGAN, and a structural consistency evaluation based on SAM segmentation results. The effectiveness of these methods was validated using day–night dual-modal UAV imagery from Xianghe City, Hebei Province. However, this study still has some shortcomings and areas for improvement, which are discussed and analyzed below.

4.1. Sensitivity Analysis of WPID Misalignment

The dataset used in this study was collected from nighttime UAV scenes in Xianghe County and Hancheng, Hebei Province, mainly covering remote sensing images of typical

suburban areas. Although the dataset covers various ground object types and spatial structures and can reflect the actual application environment of UAV nighttime remote sensing to a certain extent, the current data are mainly derived from a relatively single region and do not cover extreme weather conditions (e.g., rain and fog environments) or more complex cross-regional scenarios. Therefore, the experimental results of this paper mainly verify the effectiveness of the method under conventional UAV nighttime imaging conditions. Future research will be further extended to different regions and public datasets to systematically evaluate the cross-scene applicability and generalization ability of the model.

To address the core limitations of difficulty in acquiring strictly paired data and easy semantic distortion of unpaired data in low-illumination UAV remote sensing image enhancement, this study proposes a WPID mechanism. This mechanism only requires images captured day and night in the same area with consistent scene semantics, without strict pixel-level registration.

Since applications such as UAV inspection, security monitoring, and emergency observation usually adopt a repeated route acquisition strategy, daytime and nighttime images of the same area often have high semantic consistency, but inevitably have a certain spatial offset. Based on the characteristics of this practical application, WPID is proposed to reduce the cost of data collection while retaining necessary semantic correspondence.

Experimental results show that under the combined action of the weakly paired training mechanism and the structural improvement of HDCGAN+, the improved model outperforms typical methods such as CycleGAN in FID, KID and other indicators, and can better restore vegetation hue and building texture in subjective visual effects, indicating that the mechanism can maintain good enhancement performance while reducing data acquisition costs.

To further verify the rationality of the weakly paired mechanism and analyze the influence of different weakly pairing degrees on model performance, this paper supplements a WPID offset sensitivity experiment. By controlling the spatial offset degree of daytime color visible (DCV) images relative to nighttime color visible (NCV) images, datasets with different offset ratios including 0%, 10%, 20%, 30%, 50% and 100% were constructed, and HDCGAN+ was trained under the same reference settings. Table 6 shows the results of traditional evaluation indicators under different offset ratios.

Table 6. Quality indicators (FID, KID) and structural consistency indicators (ΔN, ΔA, BIoU) of HDCGAN+ enhanced images under different offset ratios (bold indicates the best; underlined indicates the second best).

Offset	FID ↑	KID ↑	BIoU ↑	ΔN ↓	ΔA ↓
hdcgan_0offset	40.3222	0.0092	**0.1778**	54.7679	<u>0.2718</u>
hdcgan_10offset	40.6786	<u>0.0090</u>	<u>0.1605</u>	51.8978	0.2931
hdcgan_20offset	42.7720	0.0093	0.1524	48.2133	0.2770
hdcgan_30offset	42.6761	0.0100	0.1601	<u>45.8641</u>	**0.2665**
hdcgan_50offset	<u>44.1192</u>	0.0099	0.1459	**45.8472**	0.2816
hdcgan_100offset	**55.6295**	**0.0183**	0.1532	56.1230	0.2936

It can be seen in Table 6 that when the offset ratio is in the range of 0–30%, FID, KID and structural consistency indicators (ΔN, ΔA) change relatively gently, indicating that moderate spatial misalignment will not significantly affect the enhancement performance while maintaining a certain semantic correspondence. With the further increase in the offset ratio, the quality of the generated results gradually decreases. When the offset ratio reaches 100% (completely unpaired), FID and KID increase significantly, indicating that

the difference between the generated image distribution and the real image distribution increases significantly, resulting in obvious deterioration of enhancement performance.

In addition, the BIoU indicator is mainly used to measure the performance of enhanced images in terms of semantic region overlap and boundary structure retention. It can be seen in Table 6 that under the condition of a small offset ratio (0–30%), the BIoU indicator is relatively stable as a whole, indicating that moderate spatial offset will not significantly affect the structural consistency of enhanced images while maintaining a certain semantic correspondence. As the offset ratio further increases to 50% and above, the BIoU indicator gradually decreases, indicating that the semantic region overlap and boundary matching degree begin to weaken. When the offset ratio reaches 100% (completely unpaired), both indicators decrease significantly, indicating that the spatial correspondence between semantic objects is destroyed, resulting in a significant reduction in the structural expression ability of enhanced images.

The increase in FID and KID with the offset ratio indicates that the offset ratio does affect the visual quality of the output images, while the two indicators ΔN and ΔA do not show linear consistency with the change in the offset ratio, which exactly illustrates the inconsistency between the human eye judgment results of enhanced images and the computer processing results in the task link. In addition, during the training process of the GAN model, since its loss function cannot converge automatically, the results relying on manual judgment can only give the optimal results highly dependent on vision and cannot give task-oriented results, which requires further in-depth discussion.

Figure 10 shows the enhancement results under different offset ratios. From left to right in the figure are the daytime reference image and the enhancement results obtained under the offset ratios of 0%, 10%, 20%, 30%, 50% and 100%. Overall, as the spatial overlap between images decreases, the quality of enhanced images gradually decreases. When the offset ratio is in the range of 0–30%, the enhanced results can still maintain the main ground object structures and texture details such as roads and buildings well; however, when the offset ratio continues to increase, the structure retention ability is significantly weakened. Especially under the condition of 100% complete non-overlap, due to the lack of effective semantic correspondence, the fidelity and consistency of some building boundaries and road structures become weaker and weaker compared with the Day Truth, and the quality of enhanced images is the worst at this time.

To further analyze the influence of different offset ratios on enhanced images, the SAM model was used to segment the results of each ratio, and the visualization results of its segmentation masks are shown in Figure 11. From left to right in the figure are the segmentation results of the daytime reference image, and the segmentation results of enhanced images obtained under the offset ratios of 0%, 10%, 20%, 30%, 50% and 100%.

The segmentation results in Figure 11 are basically consistent with the evaluation results of the two indicators ΔN and ΔA in Table 6. The effect of 50% offset is not worse than that of 10% offset. The main reasons for such results are that the training termination condition of HDCGAN+ is the optimal result judged by human eyes, which is not associated with any parameters or information in the task stage. In addition, although WPID greatly enhances the semantic consistency constraint between training data pairs, the structure of the GAN model itself determines that it cannot make good use of this semantic consistency to improve the image enhancement effect. Furthermore, the SAM model has not been specially trained for the data in this study, and its segmentation results have mis-segmentation in some samples, which may also have a certain impact on the structural consistency indicators.



Figure 10. Examples of HDCGAN+ enhancement results under different offset ratios. Numbers 1–6 on the left denote six different experimental regions. From left to right: the original daytime image, enhanced results with 0% offset, 10% offset, 20% offset, 30% offset, 50% offset, and 100% offset, respectively.

In addition, although the WPID mechanism has shown significant advantages, it still has some shortcomings: first, it is difficult to apply it to extreme low-illumination environments. The experiment only verifies the effect under conventional nighttime low-illumination conditions, and does not involve extreme low-illumination environments such as moonless nights and heavy haze. In such scenes, the noise of visible images increases significantly and information loss is serious, and the model is prone to generate texture artifacts; second, the model has high computational cost. After the input resolution is extended to 512×512, the feature calculation amount of a single image increases significantly, and training and inference still need to rely on high-performance server hardware. At present, it cannot be adapted to the low-power hardware onboard UAVs, making it difficult to meet the real-time application requirements of shooting while enhancing (e.g., live monitoring of nighttime disaster scenes). In addition, when visible images have severe overexposure or structural loss, the structural prior relied on by the LRR-FusionColor module may degenerate, thus limiting the final color restoration effect.

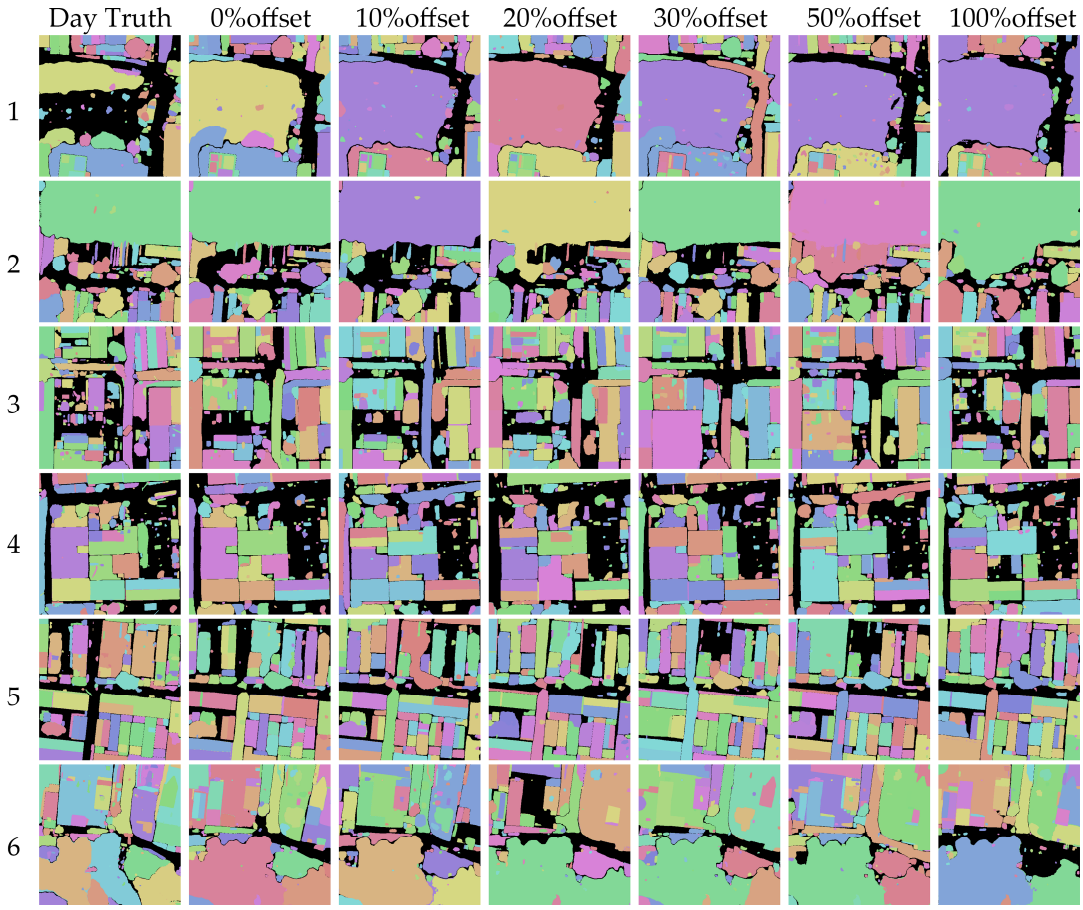


Figure 11. Visualization of SAM segmentation masks for enhanced images under different offset ratios (colors only indicate different ground objects, not object categories). Numbers 1–6 on the left denote six different experimental regions. From left to right: segmentation result of the original daytime image and enhanced results with 0% offset, 10% offset, 20% offset, 30% offset, 50% offset, and 100% offset, respectively.

The experimental data in this paper are mainly from UAV images collected in Xianghe County, Hebei Province. Although the scenes include typical ground object types such as roads, buildings, vegetation and water bodies, the overall regional scope is still relatively limited and does not cover more regional differences and complex surface types. To further verify the applicability of the method under different scene conditions, a comparative experiment is supplemented in the experimental part using UAV image data collected in the suburban area of Hancheng Town, Tangshan. In addition, among the currently publicly available UAV visible–thermal infrared multimodal datasets, data resources with cross-temporal acquisition, natural spatial offset and weakly paired characteristics are still relatively limited, which limits the systematic comparative verification across regions or datasets to a certain extent. Future research will further expand the data collection regions and scene types, and verify the generalization ability of the method under more data conditions.

To solve the above problems, future research will further optimize the WPID mechanism and supporting models from two aspects. First, it will improve the adaptability to extreme low-illumination scenes at the data level by constructing WPID including severe weather conditions, combined with multi-scale noise suppression strategies to improve the robustness of the model under complex imaging conditions; second, it will improve the GAN model, explore its structure adapted to WPID data and lightweight methods, improve its image enhancement effect, and try to improve the natural convergence of the training

process and reduce computational complexity as much as possible, so as to enhance its deployment feasibility in UAV real-time remote sensing applications.

4.2. Verification of Generalization Ability on Cross-Regional Data

To further verify the applicability of the proposed method in different regional scenes, supplementary experiments are carried out on UAV nighttime image data collected in Hancheng Town, Tangshan City. The dataset contains a total of 220 pairs of image samples, including nighttime color visible (NCV) images, thermal infrared (TIR) images and daytime color visible (DCV) images. Although the data have been registered, there may still be a certain spatial offset between images due to factors such as cross-temporal acquisition and UAV attitude changes. Therefore, this experiment mainly analyzes the enhancement effects of different methods through visual comparison without strict quantitative evaluation.

In this experiment, TIR and NCV images are first structurally fused by LRRNet, and the color information of nighttime images is restored by the FusionColor method; then NCV and DCV images are input into the HDCGAN+ model for enhancement and compared with the original HDCGAN method. It should be noted that the original HDCGAN method uses completely unpaired training data, so its generated results may be somewhat different from the method in this paper in local areas. To ensure the consistency of comparison, this paper shows the enhancement results of different methods on the same input samples.

Figure 12 shows the enhancement results of different methods on the Tangshan dataset. From left to right are the enhancement results of FusionColor, CycleGAN, EnlightenGAN, RetinexFormer, HDCGAN, HDCGAN+ and the daytime reference image (Day Truth). From the perspective of human visual judgment, it can be observed that compared with all methods, HDCGAN+ performs more stably in brightness restoration, texture details and structure retention, and its results are visually closer to the daytime reference image, indicating that the proposed method has good transfer ability under cross-regional data conditions.

4.3. Structural Consistency Evaluation Index Based on SAM Segmentation Results

In the evaluation of low-illumination UAV remote sensing image enhancement effects, the block-based structural consistency evaluation system based on SAM constructed in this study makes up for the limitations of traditional evaluation indicators that focus on perceptual quality but neglect structural integrity and downstream task adaptability. Existing mainstream evaluation indicators (e.g., PSNR, SSIM, FID) can only quantify pixel errors or global perceptual similarity, and cannot effectively evaluate whether the image structure is completely retained. However, downstream tasks of UAV remote sensing (e.g., target detection, ground object classification) have high requirements for structural consistency. For example, the integrity of building outlines in disaster emergency scenarios directly affects the judgment of damage degree. The structural consistency evaluation index based on SAM segmentation results proposed in this paper can solve the image quality evaluation for downstream tasks.

In addition, this study uses the zero-shot segmentation ability of SAM to realize full-image structural blocking without additional annotation, and designs two indicators: block number difference and block coverage area difference, focusing on quantifying the enhancement effect from the perspective of global structure retention. The similarities and differences between traditional distribution indicators (FID/KID) and structural consistency indicators (ΔA/ΔN/BIoU) reflect the natural trade-off between perceptual quality and structural retention. The indicators in this paper pay more attention to the structural stability in subsequent remote sensing applications, so SAM indicators are more interpretable at the task level.



Figure 12. Qualitative comparison of enhancement results on cross-regional data. Numbers 1–6 on the left denote six different experimental regions. From left to right: enhancement results of the Tangshan dataset using FusionColor, CycleGAN, EnlightenGAN, RetinexFormer, HDCGAN, and HDCGAN+, respectively, as well as the corresponding daytime reference image (Day Truth).

Experimental results show that the enhanced images of the HDCGAN+ model perform optimally in all structural consistency indicators, and their segmentation masks are highly similar to daytime real images in road continuity and building outline integrity, effectively identifying the pseudo-enhancement problem with acceptable performance of traditional indicators but missing structures, and providing a quantitative basis for the correlation between enhancement effect and practical application value.

However, this evaluation method still has three shortcomings: first, this paper directly adopts the SAM general image segmentation model, which itself has certain segmentation errors. At the same time, it has not been specially optimized for small ground objects (e.g., rural paths, farmland ridges) or low-contrast ground objects (e.g., green belts under low light) common in UAV remote sensing in this study, which may further strengthen the errors and easily lead to missing or mis-segmentation, thereby affecting the accuracy of various structural evaluation indicators; second, the indicator weights proposed in this paper lack scene adaptability. Different application scenarios have different requirements for evaluation dimensions. For example, public security scenarios pay more attention to the structural integrity of people and vehicles, and disaster emergency scenarios pay more attention to the clarity of building damage boundaries. The existing system adopts equal weight calculation, and its application effect in different scenarios needs to be tested and improved. Although theoretically, the quality of images will affect the segmentation of

SAM, there is currently no quantitative method to evaluate this impact. Therefore, SAM evaluation indicators are mainly used for relative comparison of structural consistency rather than absolute semantic accuracy evaluation. Future research can further verify the reliability of the evaluation system by combining UAV remote sensing dedicated segmentation models or manual annotation data.

Subsequent optimization of this evaluation method can be carried out from three aspects: first, improve the segmentation adaptability of SAM to remote sensing ground objects. Use UAV remote sensing ground object annotation datasets containing buildings, roads, vegetation and small ground object categories to fine-tune the SAM decoder, strengthen the model's segmentation ability for low-contrast and small ground objects, and introduce ground object type weight coefficients to assign higher penalty weights to segmentation errors of core ground objects in the scene, so as to improve the accuracy of indicator calculation; second, test the effect of the indicator system in other downstream tasks such as target detection and ground object classification, and incorporate the relevant evaluations of these tasks into the current evaluation system when necessary; third, design a scene-adaptive weight mechanism, construct a weight prediction model based on scene features (e.g., ground object density, lighting conditions, task types), and dynamically adjust the weight of each evaluation indicator by inputting scene information to improve the scene adaptability of the evaluation system.

4.4. Analysis of Reasons for Differences Between Different Evaluation Indicators

It can be observed in the experimental results that different methods show certain differences between traditional perceptual quality indicators (e.g., FID and KID) and SAM-based structural consistency indicators (e.g., HDCGAN performs well in FID/KID indicators but relatively low in structural consistency indicators). This difference stems from the different goals of evaluation indicators.

FID and KID mainly measure the overall distribution similarity between generated images and real images, focusing on reflecting the overall visual perceptual quality and statistical feature consistency of images. However, such indicators do not directly evaluate the spatial structure accuracy or ground object boundary consistency of images. Therefore, in some cases, even if the image has a good overall visual effect, its local structure may still have offset or distortion.

In contrast, SAM-based structural consistency indicators directly calculate information closely related to boundary consistency such as region number and area based on image segmentation results, and pay more attention to the integrity of ground object structure and the consistency of spatial layout. Such indicators are closer to the actual needs of remote sensing images in downstream tasks such as target recognition, change detection and semantic analysis.

Therefore, the difference between FID/KID and SAM-based indicators does not represent inconsistent evaluation, but reflects the trade-off relationship between perceptual quality and structural consistency of different methods. The HDCGAN+ proposed in this paper performs better in evaluating the performance of enhanced images in specific downstream tasks, indicating that it has better performance in maintaining the integrity of ground object structures, which is of greater significance for the subsequent application and processing of remote sensing images.

5. Conclusions

This study explores two key issues in the field of image enhancement: data acquisition and performance evaluation. Its main innovations lie in two aspects: First, addressing the dilemma where current methods face high acquisition costs due to reliance on strictly

paired datasets yet struggle to ensure accuracy when using unpaired datasets, we propose and validate the effectiveness of constructing weakly paired image datasets. These datasets do not require pixel-level precise registration because they retain only key associative information between images (such as approximate registration of the same region and semantic associations of core features). This approach not only significantly reduces data collection and annotation costs but also provides necessary quality constraints for the enhancement model, effectively addressing issues such as detail distortion and artifacts in unsupervised methods. At the same time, it overcomes the reliance of supervised methods on strictly paired data, demonstrating superior adaptability in scenarios such as low-illumination imagery and remote sensing imagery of complex terrain. Second, addressing the limitations of traditional evaluation systems—which focus solely on objective metrics (such as PSNR and SSIM) or single subjective visual perceptions, making it difficult to comprehensively assess the performance of augmentation results in terms of detail restoration, semantic consistency, and practical application suitability—we have established a new evaluation system that integrates objective metrics, subjective visual scores, and scenario-specific application metrics (such as feature recognition accuracy in remote sensing imagery and object detection success rates in low-illumination imagery).

Experimental results indicate that different evaluation metrics reflect distinct aspects of image quality, with SAM-based structural metrics more directly reflecting the structural preservation capability of remote sensing images in practical applications, thereby holding significant reference value. This framework not only enables a more comprehensive and precise evaluation of enhancement effects but also provides more targeted guidance for selecting and optimizing enhancement methods across different scenarios. Ultimately, the synergy between the WPID and the new evaluation framework provides critical support for the transition of image enhancement technology from the laboratory to real-world complex scenarios while also offering a referenceable paradigm for data construction and performance evaluation for subsequent related research.

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 and 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 Science and Technology Project of Xinjiang Production and Construction Corps, grant number 2017DB005, 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 conflicts of interest.

References

1. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs trained by a two-time-scale update rule converge to a local Nash equilibrium. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6626–6637.
2. Pisano, E.D.; Zong, S.; Hemminger, B.M.; DeLuca, M.; Johnston, R.E.; Muller, K.; Braeuning, M.P.; Pizer, S.M. Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. *J. Digit. Imaging* **1998**, *11*, 193–200. [CrossRef] [PubMed]
3. Lee, C.; Lee, C.; Kim, C.S. Contrast enhancement based on layered difference representation of 2D histograms. *IEEE Trans. Image Process.* **2013**, *22*, 5372–5384. [CrossRef] [PubMed]
4. Chandrasekar, R.N.; Anand, C.R.; Preethi, S. Survey on histogram equalization method-based image enhancement techniques. In *Proceedings of the 2016 International Conference on Data Mining and Advanced Computing*; IEEE: New York, NY, USA, 2016; pp. 150–158. [CrossRef]
5. Land, E.H.; McCann, J.J. Lightness and Retinex theory. *J. Opt. Soc. Am.* **1971**, *61*, 1–11. [CrossRef] [PubMed]

6. Jobson, D.J.; Rahman, Z.-U.; Woodell, G.A. Properties and performance of a center/surround Retinex. *IEEE Trans. Image Process.* **1997**, *6*, 451–462. [CrossRef]
7. Rahman, Z.; Jobson, D.J.; Woodell, G.A. Multi-scale Retinex for color image enhancement. In *Proceedings of the 3rd IEEE International Conference on Image Processing*; IEEE: New York, NY, USA, 1996; Volume 3, pp. 1003–1006. [CrossRef]
8. Palma-Amestoy, R.; Provenzi, E.; Bertalmío, M.; Caselles, V. A perceptually inspired variational framework for color enhancement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 458–474. [CrossRef]
9. Fu, X.; Liao, Y.; Zeng, D.; Huang, Y.; Zhang, X.-P.; Ding, X. A probabilistic method for image enhancement with simultaneous illumination and reflectance estimation. *IEEE Trans. Image Process.* **2015**, *24*, 4965–4977. [CrossRef]
10. Guo, X.; Li, Y.; Ling, H. LIME: Low-light image enhancement via illumination map estimation. *IEEE Trans. Image Process.* **2017**, *26*, 982–993. [CrossRef]
11. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2017; pp. 5967–5976. [CrossRef]
12. Li, C.; Guo, J.; Porikli, F.; Pang, Y. LightenNet: A convolutional neural network for weakly illuminated image enhancement. *Pattern Recognit. Lett.* **2018**, *104*, 15–22. [CrossRef]
13. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the 2017 IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 2017; pp. 2242–2251. [CrossRef]
14. Qu, Y.; Ou, Y.; Xiong, R. Low illumination enhancement for object detection in self-driving. In *Proceedings of the 2019 IEEE International Conference on Robotics and Biomimetics*; IEEE: New York, NY, USA, 2019; pp. 1738–1743. [CrossRef]
15. Mao, R.; Cui, R. RetinexGAN Enables More Robust Low-Light Image Enhancement via Retinex Decomposition Based Unsupervised Illumination Brightening. OpenReview. 2023. Available online: https://openreview.net/forum?id=3SqnZXg24T (accessed on 5 December 2025).
16. Ni, Z.; Yang, W.; Wang, S.; Ma, L.; Kwong, S. Towards unsupervised deep image enhancement with generative adversarial network. *IEEE Trans. Image Process.* **2020**, *29*, 9140–9151. [CrossRef]
17. Liu, Q.; Zhou, H.; Xu, Q.; Liu, X.; Wang, Y. PSGAN: A generative adversarial network for remote sensing image pan-sharpening. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 10227–10242. [CrossRef]
18. Pei, S.; Lin, J.; Liu, W.; Zhao, T.; Lin, C.-W. Beyond night visibility: Adaptive multi-scale fusion of infrared and visible images. *arXiv* **2024**, arXiv:2403.01083. [CrossRef]
19. Yang, S.; Sun, M.; Lou, X.; Yang, H.; Liu, D. Nighttime thermal infrared image translation integrating visible images. *Remote Sens.* **2024**, *16*, 666. [CrossRef]
20. 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*, 5005917. [CrossRef]
21. Li, H.; Wu, X.-J. DenseFuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **2019**, *28*, 2614–2623. [CrossRef]
22. Li, H.; Wu, X.J.; Durrani, T. NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 9645–9656. [CrossRef]
23. Tang, L.; Yuan, J.; Ma, J. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Inf. Fusion* **2022**, *82*, 28–42. [CrossRef]
24. Li, H.; Wu, X.-J.; Kittler, J. RFN-Nest: An End-to-End Residual Fusion Network for Infrared and Visible Images. *Inf. Fusion* **2021**, *73*, 72–86. [CrossRef]
25. Ma, J.; Xu, H.; Jiang, J.; Mei, X.; Ma, J. DDcGAN: A Dual-Discriminator Conditional Generative Adversarial Network for Multi-Resolution Image Fusion. *IEEE Trans. Image Process.* **2020**, *29*, 4980–4995. [CrossRef]
26. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
27. Das, K.; Jiang, J.; Rao, J.N.K. Mean squared error of empirical predictor. *Ann. Stat.* **2004**, *32*, 818–840. [CrossRef]
28. Johnson, D.H. Signal-to-noise ratio. *Scholarpedia* **2006**, *1*, 2088. [CrossRef]
29. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
30. Bińkowski, M.; Sutherland, D.J.; Arbel, M.; Gretton, A. Demystifying MMD GANs. *arXiv* **2018**, arXiv:1801.01401. [CrossRef]
31. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2018; pp. 586–595. [CrossRef]
32. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a completely blind image quality analyzer. *IEEE Signal Process. Lett.* **2013**, *20*, 209–212. [CrossRef]
33. Mittal, A.; Moorthy, A.K.; Bovik, A.C. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* **2012**, *21*, 4695–4708. [CrossRef]

34. Han, Z.; Zhang, Z.; Zhang, S.; Zhang, G.; Mei, S. Aerial visible-to-infrared image translation: Dataset, evaluation, and baseline. *J. Remote Sens.* **2023**, *3*, 96. [CrossRef]
35. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment anything. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2023; pp. 4015–4026. [CrossRef]
36. 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*, 11040–11052. [CrossRef]
37. 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*, 15808–15823. [CrossRef]
38. Jiang, Y.; Gong, X.; Liu, D.; Cheng, Y.; Fang, C.; Shen, X.; Yang, J.; Zhou, P.; Wang, Z. EnlightenGAN: Deep light enhancement without paired supervision. *IEEE Trans. Image Process.* **2021**, *30*, 2340–2349. [CrossRef]
39. Cai, Y.; Bian, H.; Lin, J.; Wang, H.; Timofte, R.; Zhang, Y. Retinexformer: One-stage Retinex-based transformer for low-light image enhancement. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2023; pp. 12504–12513. [CrossRef]
40. Xiang, F.M.; Zhu, Z.Y.; Xu, J.; Cui, Y.B. Research on algorithms of color space conversion from YUV to RGB. *Mod. Electron. Tech.* **2012**, *35*, 65–68.

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.