Object-Aware NIR-to-Visible Translation

Yunyi Gao[1], Lin Gu[2,3], Qiankun Liu[1], and Ying Fu[1]✉

[1] School of Computer Science & Technology, Beijing Institute of Technology
{yiiclass, liuqk3, fuying}@bit.edu.cn
[2] RIKEN AIP lin.gu@riken.jp
[3] The University of Tokyo

Abstract. While near-infrared (NIR) imaging is essential for assisted driving and safety monitoring systems, its monochromatic nature hinders its broader application, which prompts the development of NIR-to-visible translation tasks. However, the performance of existing translation methods is limited by the neglected disparities between NIR and visible imaging and the lack of paired training data. To address these challenges, we propose a novel object-aware framework for NIR-to-visible translation. Our approach decomposes the visible image recovery into object-independent luminance sources and object-specific reflective components, processing them separately to bridge the gap between NIR and visible imaging under various lighting conditions. Leveraging prior segmentation knowledge enhances our model's ability to identify and understand the separated object reflection. We also collect the Fully Aligned NIR-Visible Image Dataset, a large-scale dataset comprising fully matched pairs of NIR and visible images captured with a multi-sensor coaxial camera. Empirical evaluations demonstrate our method's superiority over existing methods, producing visually compelling results on mainstream datasets. Code is accessible at: https://github.com/Yiiclass/Sherry.

Keywords: NIR-to-Visible · Object-Aware Method · Color Recovery

1 Introduction

Near-infrared (NIR) imaging boasts unique benefits over conventional imaging, including superior atmospheric penetration [11], anti-interference solid features, and undetectability by the human eye. These attributes have driven the adoption of NIR imaging in various fields, including medical diagnostics [4], agriculture [9], transportation [24], assisted driving [22] and surveillance [23], particularly under low-visibility or night-time conditions. However, NIR images lack the luminance and chrominance compared to visible images, which are essential for detailed and intuitive visual interpretation. In contrast, VIS images are more user-friendly, offering rich details and vivid colors that provide a more intuitive visual experience. Furthermore, high-level vision tasks are primarily based on visible images, with a vast collection of VIS images driving advancements in the field. Therefore,

✉ Corresponding Author

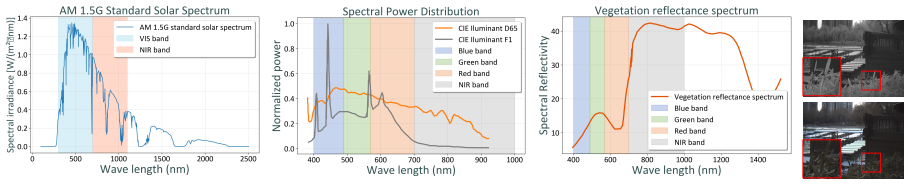

Fig. 1: (a): AM 1.5G standard spectrum of the Earth's surface [36] (b): Spectral power distribution of CIE standard light source. (c): Spectral reflectance curves of green vegetation [13], where chlorophyll absorbs mainly blue-violet and red light, resulting in low reflectance in the visible range. (d): Green vegetation exhibits heightened reflectance in the near-infrared spectrum, yielding pronounced camera response values.

developing effective NIR-to-VIS (NIR2VIS) translation techniques is essential to unlock the full potential of NIR imaging across high-level vision applications.

With the advancement of deep learning, convolutional neural networks have been applied to NIR2VIS translation tasks. Several studies [6,10,12,28–30,33,38] have endeavored to estimate the corresponding VIS images from single NIR images via an end-to-end network. However, these methods are predominantly data-driven and often overlook the fundamental distinctions between NIR and visible (VIS) images, leading to limited interpretability.

In real-world scenarios, illumination spectral variations are significant under different lighting conditions. Figure 1a and Figure 1b depict the spectrum of outdoor and indoor light sources, respectively. While this variation is less obvious for visible images thanks to adjustments like white balance in image signal processors (ISP), its effect on the NIR band could not be easily ignored. Additionally, the limited availability of paired NIR-VIS data limits the effectiveness of current NIR2VIS tasks. Currently, mainstream paired NIR-VIS data collection entails the laborious process of swapping filters at different intervals with merely a scale of hundreds. This process results in a dataset that is not only limited but also misaligned, thereby limiting image capture to static scenes exclusively.

To handle the disparities between NIR and visible imaging, we propose an object-aware framework specifically designed to enhance the NIR2VIS translation, which consists of an image decompose block, a luminance estimate block, and an object-guided reflection block. As depicted in Figure 2, acknowledging the substantial variance in illumination across the visible and NIR spectra, the image decompose block splits the image into object-independent luminance sources and object-specific reflections. This decomposition considers luminance as the light energy impinging upon the scene, unaffected by the scene's objects, and uses a luminance estimate block to ascertain visible luminance. On the other hand, we perceive reflection amount varieties, indicating that objects display distinct reflection spectra across the NIR and visible spectral ranges. To facilitate the understanding of reflective disparities among distinct objects across the NIR and visible light spectra, the object-guided reflection block leverages the prior knowledge of the state-of-the-art segmentation model to estimate reflections of

distinct objects. The visible image is then recovered by combining the visible luminance and reflection components. Furthermore, to address the limitations of existing NIR-VIS datasets, we introduce the Fully Aligned NIR-VIS Image Dataset (FANVID) collected using a multi-sensor coaxial prism camera. Our FANVID contains 5144 pairs of fully aligned high-resolution NIR-VIS images.

Our main contributions can be summarized as follows:

- Observing the large variance for illumination on visible and NIR ranges, we demonstrate that decomposing the illumination and object reflectance to process them separately can effectively enhance NIR2VIS translation task.
- Incorporation of segmentation as an object-aware prior knowledge can facilitate the estimation of object reflectance.
- We collect a Fully Aligned NIR-VIS Image Dataset (FANVID) containing fully paired data in dynamic scenes. The experimental results on mainstream NIR-VIS datasets indicate our method's superiority over leading methods and yield more visually appealing results.

2 Related Work

In this section, we first review work related to NIR2VIS translation. Subsequently, we introduce the predominant NIR-VIS datasets. Next, we discuss the colorization task, which shares the goal of recovering colors. Finally, we detail the image-to-image conversion task, similar to the NIR2VIS task.

NIR2VIS. In recent years, advancements in deep learning have prompted numerous studies exploring neural networks to bridge the spectral gap between NIR and VIS images. Initial endeavors, such as Limmer *et al.* [29], introduced an end-to-end NIR2VIS translation utilizing a multi-scale deep convolutional neural network. This approach, which bypasses the need for prior knowledge, directly infers low-frequency RGB values and integrates NIR's high-frequency attributes to construct the final VIS output. The subsequent adoption of generative adversarial networks (GANs) in studies [6, 12, 28, 33, 38] signifies a key shift, enabling conversation of images from unpaired NIR-VIS datasets. Nevertheless, the inherent spectral overlap between the NIR and VIS camera sensitivities poses a challenge in achieving precise color correspondence, limiting the effectiveness of these methods. After that, Liu *et al.* [30] proposed modifications to the single-band NIR spectrum input, introducing a technique to maximize RGB three-channel variance through deep retrieval and optimal multiplexing of the NIR spectrum. Despite this advancement, the method faces challenges in simultaneously recovering luminance and chrominance from limited NIR-VIS datasets, resulting in artifacts and unsatisfactory visual outcomes.

NIR-VIS Datasets. To gather pairs of NIR-VIS data, various camera systems have been designed, primarily falling into three types. 1) Place an NIR CUT filter in front of the camera to obtain VIS images and an NIR bandpass filter to obtain NIR images, with paired images obtained by manually switching filters. 2) Use specifically designed color filter arrays (CFAs) to automatically switch

filters, where different sections of the CFAs capture NIR and VIS data, respectively. 3) Employ a multi-sensor coaxial camera to capture paired NIR-VIS data. This system uses a prism with multiple sensors to acquire NIR and VIS data concurrently. This method is also adopted for data acquisition in this study.

The mainstream datasets are primarily acquired using the first two methods. For instance, the EPFL [2] dataset comprises 477 pairs of NIR-VIS images with a resolution of 1024×768. Lv *et al*. [32] establish a collection of 714 pairs of NIR-VIS images, with a resolution of 2048×3072, which is currently unavailable. The DVD [20] dataset contains 307 sets of NIR-VIS images with a resolution of 1920×1080. Since NIR-VIS images are obtained by switching filters, this approach complicates data collection and results in imperfectly aligned images, mostly of static scenes. Another dataset, FMSVP [34], features 64 videos with a resolution of 240×320, captured using a coaxial camera system. However, the low resolution and redundancy of information across video frames limit its utility.

Our dataset, FANVID, meticulously compiles 5144 pairs of NIR-VIS images with a resolution of 2048×1536 using a multi-sensor coaxial camera system. The data in this dataset are precisely matched and include dynamic scene information, showcasing its superiority in quality and comprehensiveness. The development and utilization of NIR-VIS datasets have been pivotal in advancing NIR2VIS research. The initial reliance on manual methods to obtain paired NIR and VIS images, as seen in datasets like EPFL, posed significant alignment challenges, particularly in dynamic scenes. The introduction of multi-sensor coaxial camera systems represented a paradigm shift, enabling the simultaneous acquisition of NIR and VIS data with precise alignment. This technological advancement facilitated more accurate and dynamic dataset collections. Our comprehensive FANVID dataset exemplifies these improvements, surpassing existing collections in alignment accuracy, image resolution, and scene dynamics.

Image Colorization. Image colorization fundamentally involves inferring color information from grayscale images. This process is typically categorized into two main approaches: guided-based and automatic. Guided-based colorization leverages reference images or prior knowledge. For instance, an intuitive and interactive method based on color palettes was introduced by [5]. Additionally, various methods employing deep learning frameworks [15, 43] have been developed for colorization using instance-specific images. Conversely, automatic colorization functions without reliance on external references. Cheng *et al*. [8] introduced a pioneering deep learning-based colorization technique. Subsequent studies [17, 47, 48] have refined network architectures to enhance visual output. InstColor [37] was developed to enable object-level colorization using object detection. More recent research, such as that by Ji *et al*. [19] and Weng *et al*. [42], has adopted transformer models to exploit global visual features, facilitating comprehensive information extraction for colorization tasks.

Compared to VIS, NIR images lack both luminance and chrominance details, whereas grayscale images only lack chrominance. Consequently, the NIR2VIS translation presents a greater challenge than traditional colorization.

Image-to-image Translation. Initially, Pix2Pix [18] utilized conditional generative adversarial networks for paired image-to-image translation, marking a foundational step in the field. Subsequently, CycleGAN [52] introduced a cycle consistency loss, enabling translation across unpaired image domains. In parallel, Liu *et al*. introduced the UNIT framework [31], based on a shared latent space concept, while Yi *et al*. developed DualGAN [45], which employs a dual learning mechanism to enhance the quality of translations. This progress was followed by the emergence of various image denoising [7, 26, 49, 50] and restoration methods [27, 39, 41, 44, 51]. More recently, the field has seen the advent of diffusion-based models [16,46], which have demonstrated significant effectiveness.

While image-to-image translation and NIR2VIS share similarities, they possess fundamental differences. Image-to-image translation typically utilizes a three-channel RGB image as input, whereas NIR2VIS processing involves a single-channel NIR image. This difference in channel information inherently affects the translation results' efficiency and accuracy.

3 Object-Aware Framework for NIR2VIS Translation

In this section, we begin with an exposition of the theoretical foundation of our object-aware framework. Next, we present a comprehensive framework formulation, highlighting the roles of individual network modules. Finally, we detail the architectural design of each network module, explicating their specific functions.

3.1 Theoretical Foundation

The human color perception system keeps the perceived color of an object relatively constant under different luminance conditions. For instance, although the spectral distributions of indoor and outdoor light sources significantly differ, human color perception remains stable. Figures 1a and 1b depict indoor and outdoor light source spectra, respectively, showing significant variance in the visible and NIR ranges across different lighting environments. In order to simulate the color stability of the human eye under different lighting conditions, visible light cameras commonly utilize image signal processors (ISP) that perform corrections such as white balance, a procedure often oversimplified in NIR imaging contexts. Consequently, distinguishing the image's luminance source from other components is essential to the task of NIR2VIS.

The classical Retinex theory explicates human color perception and assumes an image can be decomposed into object reflectance ($\boldsymbol{R}$) and environment luminance ($\boldsymbol{L}$) components, where $\odot$ denotes element-wise multiplication:

$$\boldsymbol{I} = \boldsymbol{R} \odot \boldsymbol{L}, \tag{1}$$

Inspired by Retinex theory, our method introduces an innovative framework tailored for the NIR2VIS translation task. Our approach is based on the premise that the luminance component is primarily determined by external lighting conditions and remains invariant across objects, as illustrated in Figures 1a, which

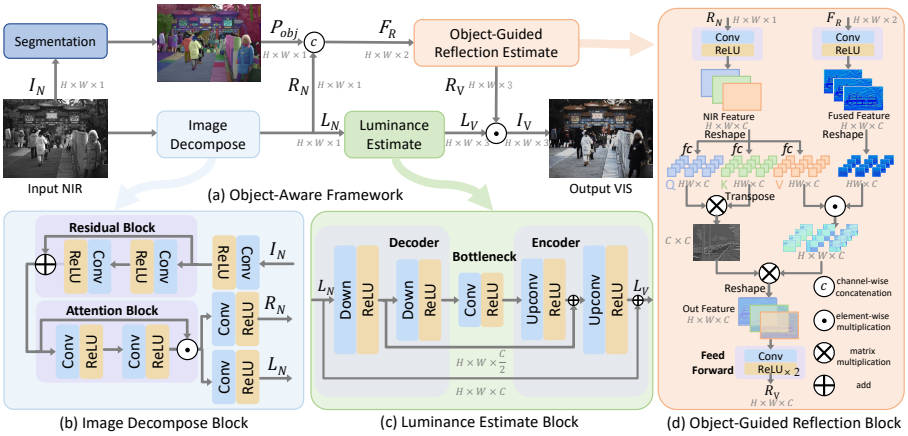

Fig. 2: (a) Overall architecture of the object-aware framework for the NIR2VIS task. (b) The network architecture of the decomposition block incorporates residual and attention modules. (c) Luminance estimate block for encoder and decoder architectures. (d) Object-guided reflection module based on transformer architecture.

depicts the solar spectrum curve. In contrast, the reflection component is intrinsically object-specific and varies between NIR and VIS spectra, exemplified in Figure 1c, which presents the spectral reflectance curve of green plants. Our framework strategically decomposes the image into $\boldsymbol{R}$ and $\boldsymbol{L}$, pertinent to the object's reflection and overall luminance, respectively. This decomposition is critical for understanding NIR2VIS translation, highlighting the necessity to accurately learn the transformation of reflectance properties from NIR to VIS bands, essentially capturing object-specific reflection characteristics.

In our object-aware framework, we focus on isolating the reflectance component to understand the differences in the object's reflection curves between the NIR and visible light bands. This isolation is achieved by mitigating the effects of luminance within the captured scene, ensuring that the reflectance component accurately represents the intrinsic properties of the object, independent of lighting conditions. The luminance component depends entirely on the intensity of the light source and is unrelated to the object. We employ a luminance estimation network to recover this component, thereby delineating the impact of the light source on the overall image composition.

Furthermore, to enhance our model's ability to capture object-specific reflectance nuances, we integrate object priors via a segmentation algorithm [21]. This enhancement bolsters the framework's ability to model the intricate reflectance characteristics unique to each object, thereby elevating the precision of the NIR to VIS image translation.

3.2 Framework Overview

In our proposed framework, we employ a series of networks to decompose and subsequently transfer the NIR into its corresponding visible image. This process begins with the decomposition of the input NIR image $\boldsymbol{I}_N$ into reflection $\boldsymbol{R}_N$ and luminance $\boldsymbol{L}_N$ components using DecomposeNet network:

$$\boldsymbol{R}_N, \boldsymbol{L}_N = \mathit{DecomposeNet}(\boldsymbol{I}_N), \tag{2}$$

Next, LuminanceNet estimates the visible image luminance $\boldsymbol{L}_V$ from $\boldsymbol{L}_N$:

$$\boldsymbol{L}_V = \mathit{LuminanceNet}(\boldsymbol{L}_N). \tag{3}$$

At the same time, we use a segmentation algorithm results as object-aware prior knowledge $\boldsymbol{P}_{obj}$:

$$\boldsymbol{P}_{obj} = \mathit{Segmentation}(\boldsymbol{I}_N). \tag{4}$$

Then, Object-Guided ReflectionNet estimates the VIS image reflection $\boldsymbol{R}_V$ under the guidance of object-aware prior:

$$\boldsymbol{R}_V = \mathit{Object\text{-}Guided\ ReflectionNet}(\boldsymbol{R}_N, \boldsymbol{P}_{obj}). \tag{5}$$

Finally, the synthesized visible light image $\boldsymbol{I}_V$ is obtained by element-wise multiplication of $\boldsymbol{R}_V$ and $\boldsymbol{L}_V$:

$$\boldsymbol{I}_V = \boldsymbol{R}_V \odot \boldsymbol{L}_V. \tag{6}$$

3.3 Network Module Design

Decompose Block. DecomposeNet is the foundational component of our framework, and it is responsible for disentangling input NIR images into reflectance and luminance components. As shown in Figure 2b, this network employs a series of convolutional layers, attention mechanisms, and residual connections to analyze the input image and separate these components effectively. The reflectance output, $\boldsymbol{R}_N$, encapsulates the intrinsic characteristics of the object, while the luminance output $\boldsymbol{L}_N$ represents the scene's lighting conditions. This decomposition is crucial for subsequent stages of the NIR2VIS process, enabling targeted manipulation and enhancement of each component.

Luminance Estimate Block. LuminanceNet is designed to process the isolated luminance component further, refining it to align with the characteristics observed in visible light images. As shown in Figure 2c, through the encoder and decoder structure network, alongside bottleneck processing, the network adjusts the NIR luminance details to match the VIS spectrum. This architecture preserves essential luminance qualities, ensuring accurate translation and realistic integration with the reflectance component during image reconstruction.

Object-Guided Reflection Block. ReflectionNet enhances the quality of the reflectance component. As shown in Figure 2d, it leverages deep learning to enhance image details, focusing on material-specific characteristics. The network

architecture is layered and complex, incorporating an initial embedding stage that transforms the input reflectance into a feature-rich representation. Following this, a series of encoder and decoder stages apply attention mechanisms and convolutional layers to process these features, integrating external segmentation or material-oriented information to guide the refinement process. The attention mechanisms within the network enable focused processing of relevant image areas, enhancing details and reducing artifacts. The final output is a refined reflectance component with enhanced detail and material-specific accuracy, which is crucial for reconstructing a high-fidelity visible light image from NIR data.

4 Fully Aligned NIR-VIS Image Dataset

To mitigate the limitations of existing NIR-VIS datasets and address the challenge of incomplete data pairing, we use a multi-sensor coaxial camera to collect a Fully Aligned NIR-VIS Image Dataset (FANVID).

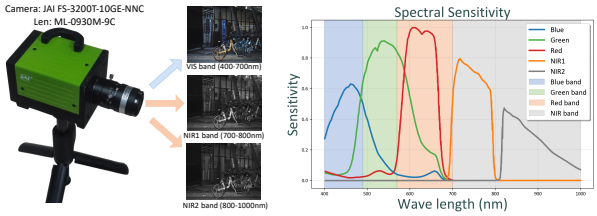

Fig. 3: (a): Camera system to collect FANVID. (b): Spectral sensitivity of camera.

4.1 Camera Settings

For simultaneous NIR and VIS data acquisition, we use the JAI FS-3200T-10GE-NNC, a 3CMOS multispectral coaxial camera with an ML-0930M-9C prism lens. Figure 3 illustrates the camera's configuration and spectral response. It can capture images simultaneously across three spectral bands: 400-700 nm for visible light, 700-800 nm for the NIR spectrum, and 820-1000 nm for the extended NIR range. Precise image alignment requires meticulous synchronization of the camera sensors' exposure times, which is essential for our dataset's accuracy.

4.2 Collected Data

Leveraging the capabilities of the multi-sensor coaxial camera, we collected 5144 image sets from diverse outdoor scenes, including streets, parks, traffic, and campuses. The dataset was gathered over two months, covering various weather conditions such as sunny, cloudy, rainy, and snowy, thereby encompassing a range

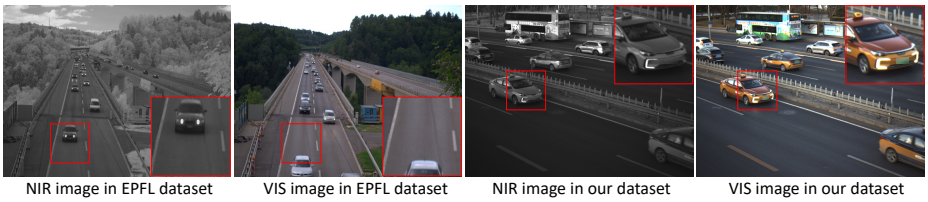

Fig. 4: The prevalent NIR-VIS dataset EPFL [2] exhibits misalignment in dynamic scenes due to manual filter replacement. In contrast, our dataset, FANVID, is acquired using a multi-sensor coaxial camera system, ensuring precise alignment of the images.

of spectral conditions. Each set contains three images that are precisely aligned: one from the visible spectrum (400-700 nm), one from the NIR spectrum (700-800 nm), and one from the extended NIR spectrum (820-1000 nm). Figure 4 delineates the alignment discrepancies in the EPFL dataset [2] due to manual filter adjustments. Our dataset circumvents these issues, enhancing its utility and providing a robust foundation for future research.

4.3 Privacy and Ethics

Our dataset collection was independently reviewed and approved by an Institutional Review Board (IRB). Following IRB guidelines, we prioritize data privacy and implement comprehensive security measures. We adhere to stringent ethical guidelines during data collection, respecting individuals' privacy rights and avoiding recording sensitive areas or activities [14]. To prevent personal identification, we anonymize all identifiable information, such as faces and license plates, following the protocols in [25, 40]. These measures ensure the protection of participants' privacy and uphold the ethical integrity of our research.

5 Experiments

In this section, we first introduce the experimental settings, including datasets, metrics, and compared methods. Next, we conduct quantitative and qualitative comparative experiments to analyze our proposed method and related approaches. Finally, we conduct ablation experiments to examine the effectiveness of our object-aware framework and the object prior.

5.1 Experimental Settings

Datasets. The datasets used for evaluation are our dataset FANVID, EPFL [2], and ICVL [1]. Our dataset, **FANVID** comprises 5144 sets of data, each containing a visible light image and NIR images in the 700-800 nm and 820-1000 nm bands. We convert the NIR images of the two bands into visible light images. The

dataset is partitioned into 4502 images for training and 642 images for testing. The **EPFL** dataset contains 477 paired images captured with Nikon D90 and Canon T1i cameras equipped with VIS and NIR filters to isolate wavelengths below and above 750 nm, respectively. However, 50 pairs exhibit alignment discrepancies, as shown in Figure 4. This dataset includes 420 images for training and 56 images for testing. The **ICVL** dataset includes 201 sets of hyperspectral images acquired by a Specim PS Kappa DX4 hyperspectral camera and a rotary stage for spatial scanning. This dataset spans 519 spectral bands from 400 nm to 1000 nm, with a spectral resolution close to 1.25 nm. The dataset is divided into 162 images for training and 39 images for testing.

Table 1: Quantitative comparison on the FANVID dataset. FANVID NIR1/NIR2, respectively, indicate using the 700-800nm band NIR1 or 820-1100nm band NIR2 images as input. All the methods have been retrained on both the NIR and RGB domains of our FANVID dataset, ensuring consistency in inputs and uniformity in settings.

Method	FANVID NIR1				FANVID NIR2			
	PSNR ↑	SSIM↑	Delta-E↓	FID↓	PSNR ↑	SSIM↑	Delta-E↓	FID↓
Retinexformer [3]	24.61	0.86	6.55	39.72	22.46	0.79	8.64	51.01
CT2 [42]	17.41	0.68	20.82	52.75	14.80	0.51	29.45	61.43
FastCUT [35]	18.65	0.71	15.94	44.29	16.74	0.63	20.03	58.96
pix2pix [18]	20.10	0.70	12.42	54.34	18.00	0.60	15.90	66.01
CycleGAN [52]	18.63	0.71	16.33	45.72	16.35	0.61	21.58	52.02
NIRcolor [6]	15.71	0.56	27.38	47.70	14.19	0.46	32.39	60.60
TLM [30]	20.65	0.75	11.23	49.79	18.76	0.66	14.47	63.25
Ours	**25.57**	**0.87**	**5.78**	**37.15**	**23.37**	**0.80**	**7.61**	**48.98**

Evaluation Metrics. To assess the performance of our NIR2VIS translation method, we utilize four established metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Fréchet Inception Distance (FID), and Delta-E. Each of these metrics provides a distinct perspective on the quality and accuracy of the image translation. Furthermore, to assess the quality of the NIR2VIS, we annotate 40 images in the FANVID test set and evaluate the translation quality by calculating the recognition rates of people, cars, bicycles, and motorcycles.

Compared Methods. To evaluate the effectiveness of our approach, we conduct three groups of comparative experiments shown in Table 1 and Table 2. All methods are retrained on the NIR and VIS domains of the corresponding datasets. The first group investigates the latest transformer-based image restoration methods, including Retinexformer [3] for low-light enhancement work and CT2 [42] for grayscale image colorization tasks. The second category investigates several classic image-to-image translation algorithms, including pix2pix [18], CycleGAN [52], and FastCUT [35]. Lastly, we compare our proposed method with the recent NIR2VIS methodologies, specifically NIRcolor [6] and TLM [30].

Table 2: Quantitative comparison on the EPFL [2] and ICVL [1] dataset. All the methods have been retrained on both the NIR and RGB domains of the EPFL/ICVL dataset, ensuring consistency in inputs and uniformity in settings.

Method	EPFL [2]				ICVL [1]			
	PSNR ↑	SSIM↑	Delta-E↓	FID↓	PSNR ↑	SSIM↑	Delta-E↓	FID↓
Retinexformer [3]	17.93	0.64	14.89	130.67	27.12	0.89	7.78	88.31
CT2 [42]	12.68	0.29	27.03	116.73	17.96	0.70	20.91	134.53
FastCUT [35]	10.30	0.10	33.77	255.39	18.98	0.65	18.70	169.97
pix2pix [18]	16.90	0.55	16.17	121.02	24.84	0.81	9.70	124.05
CycleGAN [52]	15.13	0.55	21.87	119.64	19.58	0.63	18.25	169.91
NIRcolor [6]	13.99	0.53	29.37	150.60	16.36	0.69	25.52	142.85
TLM [30]	15.63	0.49	19.08	193.17	24.53	0.82	9.61	130.51
Ours	**18.41**	**0.65**	**13.85**	**113.90**	**27.47**	**0.90**	**7.43**	**82.95**

5.2 Main Results

The quantitative results in Tables 1 and 2 demonstrate the superior performance of our approach across various datasets: FANVID, ICVL, and EPFL. The analysis indicates that the image quality restored using the NIR1 band is significantly better than that achieved with the NIR2 band. This disparity is attributed to two primary factors. Firstly, the NIR1 band is spectrally closer to the visible light spectrum, making its properties more similar to visible light compared to the NIR2 band. Secondly, the NIR2 band typically has fewer associated natural light sources, resulting in inherently lower intensity relative to the NIR1 band.

Our methodology not only secures higher PSNR and SSIM values but also exhibits lower FID scores compared to existing methods, underscoring its effectiveness in the detailed recovery of VIS images from NIR sources. Additionally, the reduction in Delta-E values suggests enhanced color accuracy and vibrancy of our result images. This advancement is primarily due to our novel strategy of isolating the reflected light component from ambient brightness, combined with applying object segmentation priors, which facilitates a more refined learning of reflectance relationships from the near-infrared to the visible spectrum.

To complement our quantitative results, Figures 5 and 6 provide qualitative comparisons from the FANVID, ICVL, and EPFL datasets. These visual assessments reveal that conventional methods often produce images marred by artifacts, blur, and desaturation. In contrast, our approach consistently yields images that are clearer, more detailed, and exhibit richer, more accurate colors.

To evaluate the efficacy of our translation methodology, we annotate 40 images from the FANVID test set and assess the translation quality by determining the recognition rates of people and vehicles. Specifically, our assessment of conversion performance utilized the YOLOv8n model, pre-trained on the COCO dataset. As indicated in Table 3, our method exhibits superior detection accuracy compared to existing methods. This result demonstrates that our object-aware

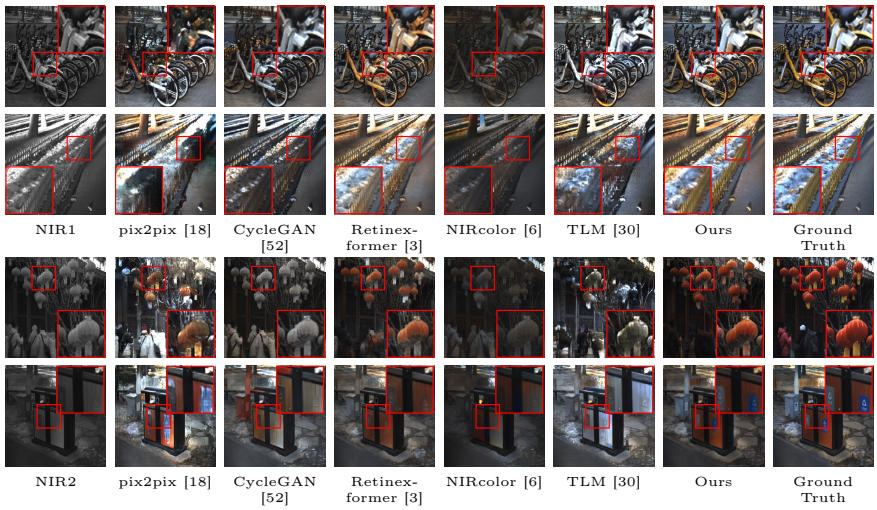

Fig. 5: Visual quality comparison results on our dataset. The first two rows display conversion results using NIR1 band images as input, while the last two rows present results using NIR2 band images as input. Please zoom in for better visualization.

approach significantly enhances the accuracy of object contour detection, providing substantial benefits for subsequent computer vision tasks. Additionally, we evaluated the visible light translation effect using mainstream segmentation methods, as depicted in Figure 7. When the NIR image is utilized as the input, grass areas are erroneously classified as snow. However, the visible light images converted through our method are classified accurately.

Quantitative and qualitative analyses prove our object-aware method's efficacy for the NIR2VIS translation problem. Our method starts from the perspective of physical imaging and decouples the visible light image into object-independent light source components and object-specific reflection components. By introducing segmentation results as object priors, the corresponding visible light image is effectively restored. Our method not only surpasses existing performance techniques but also produces visually appealing results.

5.3 Ablation Study

In this section, we conduct ablation studies to assess the efficacy of our method. Specifically, we omit the decomposition process that enables direct learning from NIR to VIS mapping and exclude the object prior to determine their respective impacts. The results of these ablation experiments are presented in Table 4.

Impact of Image DecomposeNet. To ascertain the effectiveness of the image decomposeNet in our framework, we conducted an experiment where the decomposition process is omitted, thereby forcing the model to learn a direct NIR to VIS mapping (denoted as w/o **Decom** in Table 4). The elimination of

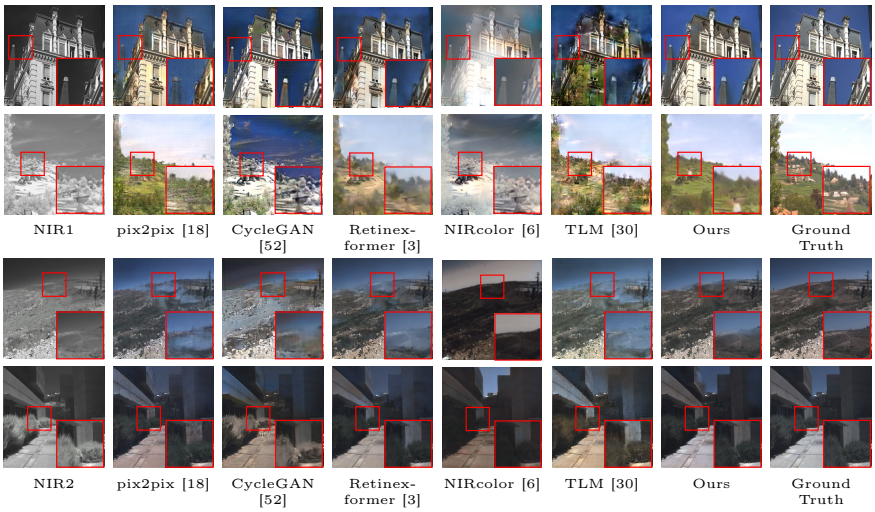

Fig. 6: Visual quality comparison results on the EPFL [2] and ICVL [1] dataset. The first and second rows display the results on EPFL, while the third and fourth rows present the conversion results on ICVL. Please zoom in for better visualization.

Table 3: Detection results on FANVID dataset.

Method	Person	Car	Bicycle	Motorcycle
Retinexformer [3]	0.825	0.855	0.733	0.649
CT2 [42]	0.862	0.903	0.695	0.762
pix2pix [18]	0.706	0.828	0.554	0.575
CycleGAN [52]	0.843	0.915	0.590	0.746
TLM [30]	0.834	0.906	0.492	0.727
Ours	**0.878**	**0.949**	**0.754**	**0.813**

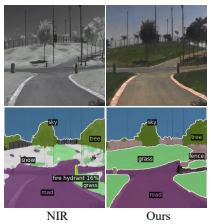

Fig. 7: Segmentation results.

DecomposeNet resulted in a reduction of PSNR by 1.24 and 1.16 on the FANVID NIR1 and NIR2 input, respectively. Analogous tendencies were discerned within the EPFL and ICVL datasets. This decline is mirrored across other evaluation metrics as well, underscoring the decomposeNet's vital role in enhancing the NIR2VIS translation accuracy. Typically, our method segregates the luminance and reflection components, enriching the model's ability to adapt and learn the nuanced relationships between NIR reflections and VIS counterparts. The absence of this decomposition compromises the model's efficacy, resulting in less detailed and vibrant image translations.

Influence of Object Prior Integration. Moreover, we explored the significance of integrating object priors into our framework. For this purpose, we replaced the object prior utilization with a self-attention mechanism in computing the decomposed reflection component (denoted as w/o **Prior** in Table 4). The re-

Table 4: Ablation Study of DecomposeNet and object prior.

Dataset	Method	PSNR ↑	SSIM↑	Delta-E↓	FID↓
FANVID NIR1	w/o **Decom**	24.33	0.85	6.78	44.54
	w/o **Prior**	24.71	0.85	6.40	43.03
	Ours	**25.57**	**0.87**	**5.78**	**37.15**
FANVID NIR2	w/o **Decom**	22.21	0.78	8.92	56.41
	w/o **Prior**	22.54	0.79	8.49	56.36
	Ours	**23.37**	**0.80**	**7.61**	**48.98**
EPFL [2]	w/o **Decom**	17.70	0.64	14.95	118.48
	w/o **Prior**	17.92	0.49	14.57	261.48
	Ours	**18.41**	**0.65**	**13.85**	**113.90**
IVCL [1]	w/o **Decom**	26.83	0.89	8.19	85.22
	w/o **Prior**	26.65	0.82	8.08	166.12
	Ours	**27.47**	**0.90**	**7.43**	**82.95**

moval of object priors led to a notable decrease in PSNR values on the FANVID NIR1 and NIR2 input, dropping by 0.86 and 0.83, respectively. A similar trend was observed for the EPFL and ICVL datasets. This decline extended across all employed metrics, emphasizing the object priors' role in enhancing translation accuracy. Object priors equip the network with contextual cues, enabling more precise differentiation and processing of distinct image regions. Consequently, this facilitates more accurate learning of the reflective properties across various objects and scenes, manifesting in improved NIR to VIS image translations.

6 Conclusion

In this paper, we introduce an object-aware framework designed to enhance the NIR-to-visible image translation process. Specifically, by recognizing the disparities between NIR and visible imaging under diverse lighting conditions, we decompose the visible light image reconstruction into two distinct components: object-independent luminance and object-specific reflection elements. This separation allows the framework to effectively discern the differential reflectance properties of objects across NIR and visible light spectra. Furthermore, we incorporate prior knowledge from a state-of-the-art segmentation model, which improves the network's capability to delineate and interpret the reflective dynamics of various objects between the NIR and visible ranges. To facilitate this framework, we collect a well-aligned large-scale NIR-VIS dataset, the Fully Aligned NIR-VIS Image Dataset (FANVID). Quantitative and qualitative experiments validate the effectiveness of our approach, demonstrate its superiority over existing methods, and produce more visually appealing results.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62331006, 62171038, and 62088101), and the Fundamental Research Funds for the Central Universities.

References

1. Arad, B., Ben-Shahar, O.: Sparse recovery of hyperspectral signal from natural rgb images. In: Proceedings of European Conference on Computer Vision. pp. 19–34 (2016)
2. Brown, M., Süsstrunk, S.: Multi-spectral sift for scene category recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 177–184 (2011)
3. 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 International Conference on Computer Vision. pp. 12504–12513 (2023)
4. Cao, L., Huang, D., Zhang, Y., Jiang, X., Chen, Y.: Brain decoding using fnirs. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 12602–12611 (2021)
5. Chang, H., Fried, O., Liu, Y., DiVerdi, S., Finkelstein, A.: Palette-based photo recoloring. ACM Transactions on Graphics **34**(4), 139–1 (2015)
6. Chen, L., Liu, Y., He, Y., Xie, Z., Sui, X.: Colorization of infrared images based on feature fusion and contrastive learning. Optics and Lasers in Engineering **162**, 107395 (2023)
7. Chen, L., Fu, Y., You, S., Liu, H.: Hybrid supervised instance segmentation by learning label noise suppression. Neurocomputing **496**, 131–146 (2022)
8. Cheng, Z., Yang, Q., Sheng, B.: Deep colorization. In: Proceedings of International Conference on Computer Vision. pp. 415–423 (2015)
9. Chiu, M.T., Xu, X., Wei, Y., Huang, Z., Schwing, A.G., Brunner, R., Khachatrian, H., Karapetyan, H., Dozier, I., Rose, G., et al.: Agriculture-vision: A large aerial image database for agricultural pattern analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2828–2838 (2020)
10. Dong, Z., Kamata, S.i., Breckon, T.P.: Infrared image colorization using a s-shape network. In: IEEE International Conference on Image Processing. pp. 2242–2246 (2018)
11. Enomoto, K., Sakurada, K., Wang, W., Fukui, H., Matsuoka, M., Nakamura, R., Kawaguchi, N.: Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 48–56 (2017)
12. Gao, Y., Liu, Q., Gu, L., Ying, F.: Grayscale-assisted rgb image conversion from near-infrared images. Tsinghua Science and Technology (2024), doi: 10.26599/TST.2024.9010115
13. Gates, D.M., Keegan, H.J., Schleter, J.C., Weidner, V.R.: Spectral properties of plants. Applied optics **4**(1), 11–20 (1965)
14. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 18995–19012 (2022)

15. He, M., Chen, D., Liao, J., Sander, P.V., Yuan, L.: Deep exemplar-based colorization. ACM Transactions on Graphics **37**(4), 1–16 (2018)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Proceedings of Conference on Neural Information Processing Systems **33**, 6840–6851 (2020)
17. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. ACM Transactions on Graphics **35**(4), 1–11 (2016)
18. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1125–1134 (2017)
19. Ji, X., Jiang, B., Luo, D., Tao, G., Chu, W., Xie, Z., Wang, C., Tai, Y.: Colorformer: Image colorization via color memory assisted hybrid-attention transformer. In: Proceedings of European Conference on Computer Vision. pp. 20–36 (2022)
20. Jin, S., Yu, B., Jing, M., Zhou, Y., Liang, J., Ji, R.: Darkvisionnet: Low-light imaging via rgb-nir fusion with deep inconsistency prior. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1104–1112 (2022)
21. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
22. Kumar, A., Gupta, A., Santra, B., Lalitha, K., Kolla, M., Gupta, M., Singh, R.: Vpds: an ai-based automated vehicle occupancy and violation detection system. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9498–9503 (2019)
23. Li, C., Zhu, T., Liu, L., Si, X., Fan, Z., Zhai, S.: Cross-modal object tracking: Modality-aware representations and a unified benchmark. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1289–1296 (2022)
24. Li, H., Li, C., Zhu, X., Zheng, A., Luo, B.: Multi-spectral vehicle re-identification: A challenge. In: In Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11345–11353 (2020)
25. Li, J., Han, L., Chen, R., Zhang, H., Han, B., Wang, L., Cao, X.: Identity-preserving face anonymization via adaptively facial attributes obfuscation. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3891–3899 (2021)
26. Li, M., Fu, Y., Zhang, Y.: Spatial-spectral transformer for hyperspectral image denoising. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1368–1376 (2023)
27. Li, Z., Wang, P., Wang, Z., Zhan, D.c.: Flowgananomaly: Flow-based anomaly network intrusion detection with adversarial learning. Chinese Journal of Electronics **33**(1), 58–71 (2024)
28. Liang, W., Ding, D., Wei, G.: An improved dualgan for near-infrared image colorization. Infrared Physics & Technology **116**, 103764 (2021)
29. Limmer, M., Lensch, H.P.: Infrared colorization using deep convolutional neural networks. In: IEEE International Conference on Machine Learning and Applications. pp. 61–68 (2016)
30. Liu, L., Chen, Y., Yan, J., Zheng, Y.: Optimal led spectral multiplexing for nir2rgb translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12652–12660 (2022)
31. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. Proceedings of Conference on Neural Information Processing Systems pp. 700–708 (2017)

32. Lv, F., Zheng, Y., Li, Y., Lu, F.: An integrated enhancement solution for 24-hour colorful imaging. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11725–11732 (2020)
33. Mehri, A., Sappa, A.D.: Colorizing near infrared images through a cyclic adversarial approach of unpaired samples. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
34. Niu, M., Zhong, Z., Zheng, Y.: Nir-assisted video enhancement via unpaired 24-hour data. In: Proceedings of International Conference on Computer Vision. pp. 10778–10788 (2023)
35. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: Proceedings of European Conference on Computer Vision. pp. 319–345 (2020)
36. Rühle, S.: Tabulated values of the shockley–queisser limit for single junction solar cells. Solar energy **130**, 139–147 (2016)
37. Su, J.W., Chu, H.K., Huang, J.B.: Instance-aware image colorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7968–7977 (2020)
38. Suárez, P.L., Sappa, A.D., Vintimilla, B.X.: Infrared image colorization based on a triplet dcgan architecture. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 18–23 (2017)
39. Tian, Y., Fu, Y., Zhang, J.: Transformer-based under-sampled single-pixel imaging. Chinese Journal of Electronics **32**(5), 1151–1159 (2023)
40. Trinh, L., Pham, P., Trinh, H., Bach, N., Nguyen, D., Nguyen, G., Nguyen, H.: Pp4av: A benchmarking dataset for privacy-preserving autonomous driving. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision. pp. 1206–1215 (2023)
41. Wei, K., Aviles-Rivero, A., Liang, J., Fu, Y., Huang, H., Schönlieb, C.B.: Tfpnp: Tuning-free plug-and-play proximal algorithms with applications to inverse imaging problems. Journal of Machine Learning Research **23**(16), 1–48 (2022)
42. Weng, S., Sun, J., Li, Y., Li, S., Shi, B.: Ct 2: Colorization transformer via color tokens. In: Proceedings of European Conference on Computer Vision. pp. 1–16 (2022)
43. Xu, Z., Wang, T., Fang, F., Sheng, Y., Zhang, G.: Stylization-based architecture for fast deep exemplar colorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9363–9372 (2020)
44. Ye, Z., He, X., Peng, Y.: Unsupervised cross-media hashing learning via knowledge graph. Chinese Journal of Electronics **31**(6), 1081–1091 (2022)
45. Yi, Z., Zhang, H., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2849–2857 (2017)
46. Zhang, F., You, S., Li, Y., Fu, Y.: Atlantis: Enabling underwater depth estimation with stable diffusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11852–11861 (2024)
47. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: Proceedings of European Conference on Computer Vision. pp. 649–666 (2016)
48. Zhang, R., Zhu, J.Y., Isola, P., Geng, X., Lin, A.S., Yu, T., Efros, A.A.: Real-time user-guided image colorization with learned deep priors. ACM Transactions on Graphics **9**(4) (2017)
49. Zhang, T., Fu, Y., Li, C.: Hyperspectral image denoising with realistic data. In: Proceedings of International Conference on Computer Vision. pp. 2248–2257 (2021)

50. Zhang, T., Fu, Y., Zhang, J.: Deep guided attention network for joint denoising and demosaicing in real image. Chinese Journal of Electronics **33**(1), 303–312 (2024)
51. Zhang, T., Liang, Z., Fu, Y.: Joint spatial-spectral pattern optimization and hyperspectral image reconstruction. IEEE Journal of Selected Topics in Signal Processing **16**(4), 636–648 (2022)
52. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of International Conference on Computer Vision. pp. 2223–2232 (2017)