technologies

MDPI

Article

Pix2Next: Leveraging Vision Foundation Models for RGB to NIR Image Translation

Youngwan Jin [1,2], Incheol Park [1,2], Hanbin Song [1,2], Hyeongjin Ju [1,2], Yagiz Nalcakan [1] and Shiho Kim [1,*]

[1] School of Integrated Technology, Yonsei University, Incheon 21983, Republic of Korea; thatnn@yonsei.ac.kr (Y.J.); incheol97@yonsei.ac.kr (I.P.); thdgksqls369@yonsei.ac.kr (H.S.); wngudwls000@yonsei.ac.kr (H.J.); ynalcakan@yonsei.ac.kr (Y.N.)
[2] BK21 Graduate Program in Intelligent Semiconductor Technology, Incheon 21983, Republic of Korea
* Correspondence: shiho@yonsei.ac.kr

Abstract: This paper proposes Pix2Next, a novel image-to-image translation framework designed to address the challenge of generating high-quality Near-Infrared (NIR) images from RGB inputs. Our method leverages a state-of-the-art Vision Foundation Model (VFM) within an encoder–decoder architecture, incorporating cross-attention mechanisms to enhance feature integration. This design captures detailed global representations and preserves essential spectral characteristics, treating RGB-to-NIR translation as more than a simple domain transfer problem. A multi-scale PatchGAN discriminator ensures realistic image generation at various detail levels, while carefully designed loss functions couple global context understanding with local feature preservation. We performed experiments on the RANUS and IDD-AW datasets to demonstrate Pix2Next's advantages in quantitative metrics and visual quality, highly improving the FID score compared to existing methods. Furthermore, we demonstrate the practical utility of Pix2Next by showing improved performance on a downstream object detection task using generated NIR data to augment limited real NIR datasets. The proposed method enables the scaling up of NIR datasets without additional data acquisition or annotation efforts, potentially accelerating advancements in NIR-based computer vision applications.

Keywords: image-to-image translation; IR; NIR; thermal; image generation; autonomous driving

check for updates

Academic Editor: Juan Gabriel Avina-Cervantes

Received: 18 March 2025
Revised: 7 April 2025
Accepted: 9 April 2025
Published: 11 April 2025

Citation: Jin, Y.; Park, I.; Song, H.; Ju, H.; Nalcakan, Y.; Kim, S. Pix2Next: Leveraging Vision Foundation Models for RGB to NIR Image Translation. *Technologies* **2025**, *13*, 154. https://doi.org/10.3390/technologies13040154

Copyright: © 2025 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 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Visible-range cameras (e.g., RGB cameras), which capture images within the spectrum of light detectable by the human eye, often have limitations in challenging conditions such as low light, adverse weather, or situations where the object of interest lacks sufficient contrast against the background. To address these challenges, one potential solution is utilizing imaging technologies that extend beyond the visible spectrum [1]. In particular, this study focuses on the Near-Infrared (NIR) spectrum. NIR cameras operating beyond the visible range demonstrate significant advantages, such as capturing reflections from materials and surfaces in a manner that enhances detection and contrast. For example, NIR cameras can penetrate fog, smoke, or even certain materials, making them valuable in applications such as surveillance, autonomous vehicles, and medical imaging where visible-range cameras might fail to capture essential details [2] (Figure 1).

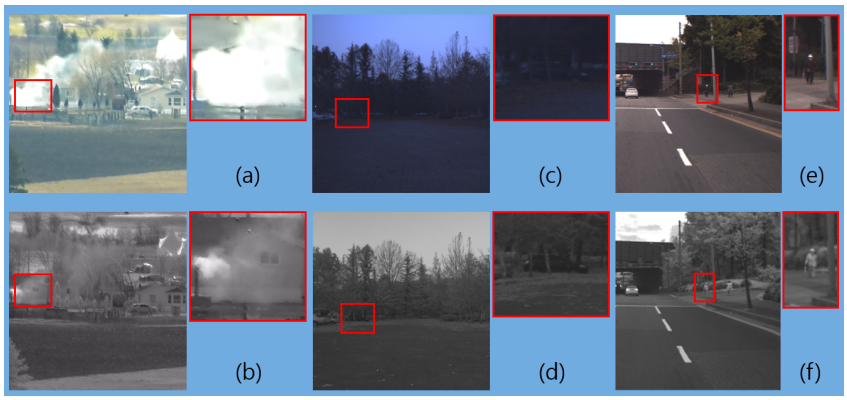


Figure 1. The top row (**a**,**c**,**e**) presents outputs from the RGB camera, while the bottom row (**b**,**d**,**f**) displays the corresponding NIR images. Objects (house in (**b**), pedestrian in (**d**), and car in (**f**)) that are not clearly discernible in the RGB images are distinctly visible in the NIR domain [3].

In the context of autonomous driving tasks, which is the main focus of our work, as shown in Figure 1e,f, some objects that remain undetectable in visible-light images become distinguishable when captured in the NIR range. Thus, incorporating NIR spectral information into imaging systems can substantially improve the performance of computer vision models across a wide range of autonomous driving tasks. However, the primary challenge lies in the lack of sufficient datasets for training perception models utilizing images from non-visible wavelength ranges. Training perception models for autonomous driving requires large datasets, often consisting of millions of annotated images. As illustrated in Figure 2, most publicly available datasets used in autonomous driving, such as KITTI [4], nuScenes [5], Waymo Open [6], Argoverse [7], and BDD100k [8] predominantly consist of visible-wavelength-range (RGB) image data. In contrast, the availability of publicly accessible NIR-based datasets, such as KAIST MS2 [9], IDD-AW [10], RANUS [11], RGB-NIR Scene [12], and TAS-NIR [13], remains limited in terms of data size, making it challenging to train robust models that can take advantage of the NIR spectrum's perception capabilities.

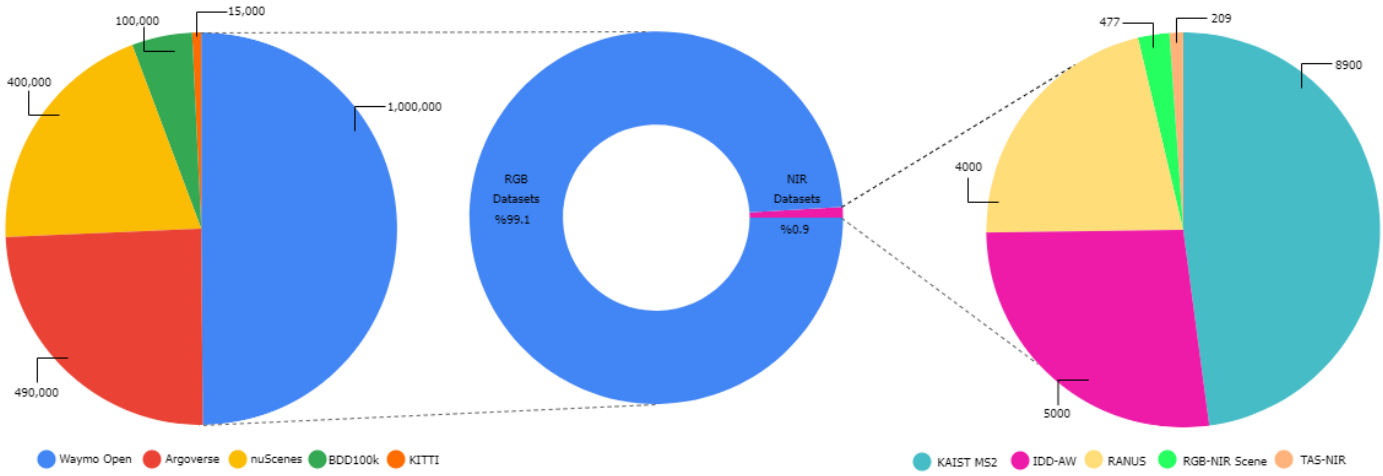


Figure 2. Comparison and distribution of publicly available autonomous driving-based RGB vs. NIR datasets.

Image-to-image (I2I) translation methods offer a promising solution to these challenges. Researchers have proposed various techniques for generating NIR [14–17], Mid-Wave Infrared (MWIR) [18], and Long-Wave Infrared (LWIR) [19–23] images from RGB inputs. These approaches cover a wide range of methods, including unsupervised image translation, supervised translation architectures, and adapted RGB-to-RGB image generation techniques. However, despite these advancements, a fundamental challenge remains

unsolved: accurately reproducing the physical properties unique to different infrared wavelengths. This difficulty arises because infrared signals are influenced by material-specific reflectance and emissivity characteristics, which are not directly captured by RGB-based or conventional translation models. As a result, existing methods often fail to fully preserve or reconstruct the distinct spectral signatures that are critical for downstream tasks. Transferring the physical properties of the visible spectral range to a specific IR range needs more accurate local and global representations of the scene. To address this, we propose Pix2Next (Figure 3), a novel RGB to NIR translation model that incorporates a global feature enhancement strategy based on a vision foundation model (VFM).

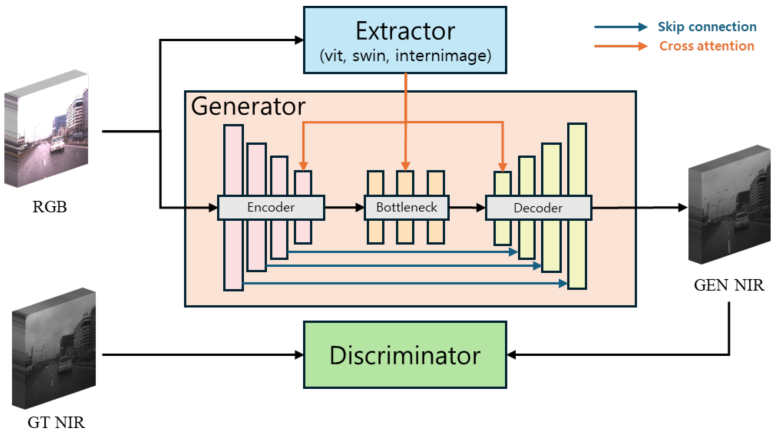


Figure 3. Overall architecture of the Pix2Next method. The Generator and Discriminator architectures are primarily based on the Pix2pixHD [24] framework. However, to achieve fine-grained scene representation, we integrated an Extractor module with cross-attention mechanisms applied to various layers of the Generator.

Pix2Next is specifically designed to accurately reflect the nuances of the NIR spectrum. As illustrated in Figure 4, generated NIR images from RGB images by our proposed method maintain fine details and critical spectral features of the translated domain. When comparing the generated images with ground truth (GT), it can be observed that the model successfully preserves essential information, such as edges and object boundaries, during the translation to the NIR spectrum. With this robust performance, the proposed model sets a new benchmark for RGB to NIR image translation and achieves state-of-the-art (SOTA) results by surpassing existing image-to-image translation (I2I) methods in six different metrics, which we will explore in detail in Section 4.

Furthermore, to assess the impact and utility of the generated NIR images on a classical autonomous driving perception task, we utilized our proposed model to scale up the NIR dataset for a downstream task. By leveraging the BDD100k data, we expanded the existing NIR dataset and observed improved performance in training when using these scaled-up data compared to previous results. This demonstrates the effectiveness of our method in enhancing the dataset for better performance in real-world autonomous driving scenarios.

The main contributions of our study are summarized as follows:

1. Overcoming the challenge of limited NIR data: We address the scarcity of NIR data compared to RGB data by employing I2I translation to generate NIR images from RGB images. This allows us to expand the NIR dataset by transferring annotations from RGB images, circumventing the need for direct NIR data acquisition and annotation efforts.
2. Introducing an enhanced I2I model—Pix2Next—and demonstrating its improved performance: Existing I2I models fail to accurately capture details and spectral characteristics when translating RGB images into other wavelength domains. To overcome

this limitation, we propose a novel model, Pix2Next, inspired by Pix2pixHD. Our model achieves SOTA performance in generating more accurate images in alternative wavelength domains from RGB inputs.

3. Validating the utility of generated NIR data for data augmentation: To evaluate the utility of the translated images, we scaled up the NIR dataset using our proposed model and applied it to an object detection task. The results demonstrate improved performance compared to using limited original NIR data, validating the effectiveness of our translation model for data augmentation in the NIR domain.



Figure 4. Example of RGB to NIR generation using the proposed method.

2. Related Work

2.1. Image-to-Image Translation

Image-to-image translation is a critical task in computer vision that involves converting images from one domain to another while retaining the underlying structure and content. This field has wide-ranging applications, including style transfer [25–27], image super-resolution [28,29], remote sensing [30–32], and domain adaptation [33,34]. The advent of deep learning, particularly Generative Adversarial Networks (GANs) [35], has significantly advanced the capabilities of I2I translation.

One of the earliest and most influential models in I2I translation is Pix2pix [36], which operates using paired datasets to learn the mapping between input and output domains and employs a conditional GAN framework where the generator is trained to produce images that the discriminator classifies as real, thereby learning to generate high-quality and realistic outputs. Building upon Pix2pix, Pix2pixHD [24] was developed to handle the challenges associated with generating high-resolution images. It introduced several improvements over the original Pix2pix, including a multi-scale discriminator and a coarse-to-fine generator architecture, which together enable the production of more detailed and realistic images. While Pix2pix and Pix2pixHD rely on paired datasets, CycleGAN [37] extends I2I translation to unpaired datasets by introducing a cycle consistency loss, which ensures that the translation from source to target and back to source preserves the original content. This innovation significantly broadened the applicability of I2I translation models to domains where paired datasets are unavailable.

More recently, models such as BBDM [38] were proposed using the diffusion process for image-to-image translation, and it has demonstrated competitive performance across various benchmarks. BBDM combines the strengths of GANs and Brownian Bridge diffusion processes to generate high-quality images with better output stability and diversity. BBDM represents a further evolution in the field, addressing some of the limitations of

earlier models, such as mode collapse in GANs and the need for extensive training data. UVCGAN [39] enhances the CycleGAN framework for unpaired image-to-image translation by incorporating a UNet-Vision Transformer (ViT) hybrid generator and advanced training techniques. UVCGAN retains strong cycle consistency while improving translation quality and preserving correlations between input and output domains, which are crucial for tasks like scientific simulations. These advancements illustrate the continuous evolution of I2I translation models, with each iteration improving upon the limitations of previous methods.

2.2. NIR/IR-Range Imaging

Infrared (IR), especially NIR imaging, is crucial in various applications that require capturing information beyond the visible spectrum, such as night-time surveillance, automotive safety, and medical diagnostics [40–42]. NIR imaging, which operates within the 700 to 1000 nanometer (nm) wavelength range (Figure 5), is particularly valuable in challenging conditions and for highlighting features that are not visible in standard RGB images.

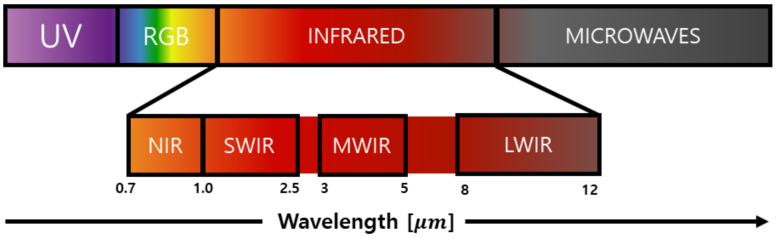


Figure 5. Diagram of the electromagnetic spectrum focusing on the infrared range.

Recent advancements have integrated NIR/IR imaging with deep learning techniques to significantly improve tasks such as human recognition and object detection under challenging conditions ([43–45]). These approaches are crucial for applications in autonomous driving and surveillance, where compromised visibility demands robust detection and recognition capabilities.

A major challenge in this field is the limited availability of annotated NIR/IR datasets, which hampers the effective training of deep learning models. To overcome this obstacle, researchers have explored the generation of synthetic NIR/IR images from RGB inputs. Aslahishahri et al. [16] employed a Pix2pix framework based on conditional GANs to produce NIR aerial images of crops. In another study focusing on person re-identification, Kniaz et al. [19] proposed ThermalGAN, which converts RGB images into LWIR images using a BicycleGAN-inspired [46] framework. Building on the concepts introduced by ThermalGAN, Özkanoğlu et al. [21] developed InfraGAN specifically for generating LWIR images in driving scenes, employing two distinct U-Net-based architectures. Additionally, Mao et al. [47] introduced C2SAL, an effective style transfer framework for generating images in the NIR domain within the driving scene context. C2SAL's approach emphasizes content consistency learning, which is applied to refined content features from a content feature refining module, which enhances the preservation of content information. Furthermore, their style adversarial learning ensures style consistency between the generated images and the target style. Notably, similar to our work, C2SAL was evaluated on the RANUS benchmark, and we have included their approach in our comparative analysis. More recently IRFormer [20] introduces a lightweight Transformer-based approach to enhance visible-to-infrared (VIS-IR) translation. This model addresses limitations like unstable training and suboptimal outputs in earlier methods by integrating a Dynamic Fusion Aggregation Module for robust feature fusion and an Enhanced Perception Attention Module to refine details under low-light or occluded conditions.

These methods have facilitated the scaling up of NIR/IR datasets without requiring extensive manual annotation, thereby enabling the training of more robust models for various NIR/IR imaging applications.

3. Method

The Pix2pixHD model uses coarse-to-fine generator architectures to transfer the global and local details of the input image to the generated image. With Pix2Next, we extended this framework by employing residual blocks within an encoder–decoder architecture instead of using separate global and local generators. Residual blocks are integral to our design, as they allow the network to maintain critical feature details by facilitating identity mappings through shortcut connections. These connections help to address the vanishing gradient problem, ensuring stable training and enabling the network to learn more complex transformations essential for high-quality image generation.

To further improve the preservation of fine details and overall image context, we integrate a vision foundation model (VFM) into our architecture, which serves as a feature extractor. Vision foundation models, trained on diverse large-scale visual datasets, possess deep knowledge of environmental patterns. This integration provides the advantage of capturing global features that work synergistically with the local features learned by the encoder–decoder structure. These features are combined throughout the network using cross-attention mechanisms, which help align and merge the global and local features during the image generation process. This approach is key to accurately capturing the specific characteristics and subtle details of the NIR domain, resulting in translated images of higher quality and reliability. To the best of our knowledge, our method is the first application of a VFM [48] into an RGB-to-NIR translation model. This novel integration idea allows our model to capture complex patterns, resulting in significant improvements in the quality and precision of the translated NIR images.

3.1. Network Architecture

The Pix2Next architecture is composed of three key modules (Figure 6). The extractor module is responsible for extracting detailed features from input RGB images, which are then fed into the generator module's encoder, bottleneck, and decoder layers via cross-attention. The generator module, designed with an encoder–bottleneck–decoder framework, focuses on generating NIR images and incorporates U-Net-inspired skip connections to facilitate information flow between the encoder and decoder layers. Finally, the discriminator module is implemented as a multi-scale patch-based GAN, featuring three discriminators operating at different resolutions. This multi-resolution approach enables the image generation process to be optimized in a coarse-to-fine manner. Algorithm 1 describes the training steps of the proposed method. Unlike previous approaches, our architecture combines the strengths of a VFM with attention mechanisms. This integration enables Pix2Next to more effectively capture global and local features than traditional methods.

In the following sections, we will delve into the specifics of each module. First, we will examine the Feature Extractor (Section 3.1.1), which leverages state-of-the-art VFMs to capture rich and contextual image representations. We will then explore the structure and innovations of our generator (Section 3.1.2), which synthesizes high-quality images by adopting an encoder–bottleneck–decoder structure with novel mechanisms for feature integration and attention. Lastly, we will discuss the details of the discriminator architecture (Section 3.1.3) and its role in enhancing the generation of high-quality, realistic NIR images.

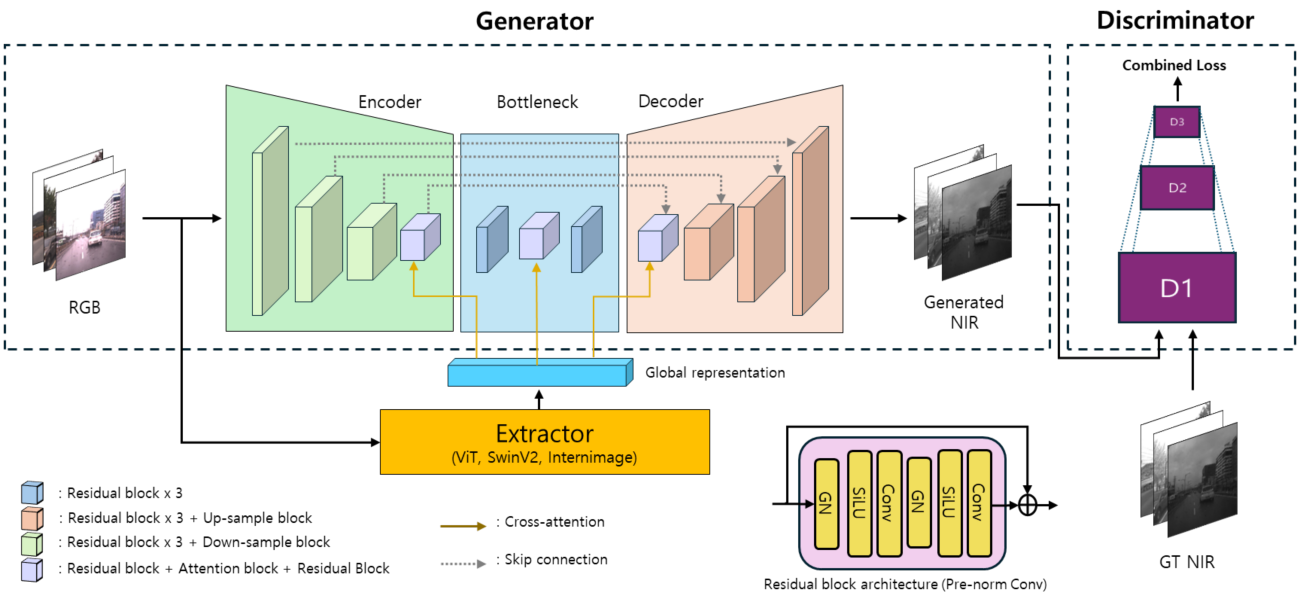


Figure 6. Detailed architecture of Pix2Next. Extractor features are fed into the encoder, bottleneck, and decoder layers, leveraging VFM representations for high-quality NIR image generation.

Algorithm 1 Training for RGB-to-NIR Image Translation with Multi-Scale Discriminators

Require: Paired dataset of RGB images X and NIR images Y
Require: Initialized generator G
Require: Three discriminators $D = \{D_1, D_2, D_3\}$ for multi-scale discrimination
Require: Hyperparameters λ_{FM}, λ_{SSIM}
Require: Learning rates η_G, η_D
Require: Number of iterations N, batch size B
for iteration $= 1$ to N **do**
  Sample mini-batch of B RGB images $x \in X$ and NIR images $y \in Y$
  Feature Extraction with VFM: $f = \text{VFM}(x)$
  Generate NIR images: $\hat{y} = G(x, f) = G(z)$
  Multi-Scale Discriminator Updates
  Create multi-scale real and generated images $\{y_i\}$ and $\{\hat{y}_i\}$ for $i = 1, 2, 3$
  for each discriminator D_i in D **do**
    Compute discriminator loss $\mathcal{L}_{D_i}$ using y_i and $\hat{y}_i$
    Update D_i by minimizing $\mathcal{L}_{D_i}$ with learning rate η_D
  end for
  Generator Update
  Compute GAN loss: $\mathcal{L}_{\text{GAN}} = \sum_{i=1}^{3} \mathcal{L}_{\text{GAN}_i}$
  Compute feature matching loss: $\mathcal{L}_{\text{FM}}$ using intermediate features from $\{D_i\}$
  Compute SSIM loss: $\mathcal{L}_{\text{SSIM}}$ between $\hat{y}$ and y
  Compute total generator loss:
    $\mathcal{L}_G = \mathcal{L}_{\text{GAN}} + \lambda_{\text{FM}}\mathcal{L}_{\text{FM}} + \lambda_{\text{SSIM}}\mathcal{L}_{\text{SSIM}}$
  Update G by minimizing $\mathcal{L}_G$ with learning rate η_G
end for

3.1.1. Feature Extractor

Our proposed model employs a state-of-the-art VFM as our feature extractor to capture detailed global representations from input images. Specifically, we utilize the Internimage [48] architecture due to its exceptional performance in capturing long-range dependencies and adaptive spatial aggregation on various vision tasks. The primary role of the feature extractor in our model architecture is to generate a comprehensive global representation of the input image, which is then used to guide the image translation process in

the generator. This method allows our model to maintain the global context and structural integrity of the RGB image during the NIR translation. We implement the feature extractor as follows:

- Input Processing: The RGB input image ($256 \times 256 \times 3$) is fed into the InternImage model.
- Feature Extraction: The InternImage model processes the input through its hierarchical structure of deformable convolutions and attention mechanisms.
- Global Representation: The output of the final layer of InternImage serves as our global feature representation. This global representation is then used in the cross-attention mechanisms throughout our generator's encoder, bottleneck, and decoder stages.

The selection of InternImage as our feature extractor is motivated by its ability to capture both fine-grained local details and broader contextual information. The deformable convolutions in InternImage allow for adaptive receptive fields, enabling the model to focus on the most relevant parts of the image for our translation task. This global representation serves as a guiding framework for our generator, ensuring that local modifications during the translation process remain coherent with the overall image structure and content. To validate the effectiveness of our chosen feature extractor, we conducted ablation studies comparing InternImage with other architectures such as ResNet [49], ViT [50], and Swin Transformer [51]. Our experiments demonstrated that InternImage outperformed other models in our RGB to NIR translation task, providing a more informative global representation that led to improved translation quality.

3.1.2. Generator

The generator in our proposed model adopts an encoder–bottleneck–decoder architecture (Table 1) designed to process 256×256 RGB images. The key components of our generator are as follows:

- Encoder: Seven blocks progressively increase the channel depth from 128 to 512, utilizing residual and downsample layers with an attention layer in the final block.
- Bottleneck: Three blocks maintain 512 channels, combining residual and attention layers for complex feature interactions.
- Decoder: Seven blocks gradually reduce channel depth from 512 to 128, using upsample layers alongside residual and attention layers.
- Normalization: Group normalization with 32 groups is applied throughout the network.

Our approach significantly diverges from the conventional Pix2pixHD architecture incorporating several key innovations. Unlike Pix2pixHD's separate global and local generators, we implement a single, deeper encoder–bottleneck–decoder structure. This design is enhanced with skip connections inspired by the U-Net architecture [52], which concatenates features from the encoder with those in the decoder. These connections facilitate the fusion of multi-scale feature representations to enhance the accuracy of the generated output and effectively preserve intricate details throughout the image synthesis process. Additionally, we introduce a cross-attention mechanism that utilizes features extracted by the VFM Feature Extractor. This mechanism is applied at each stage of the generator—the encoder, bottleneck, and decoder—allowing for effective integration of global contextual information with local features. The cross-attention operation can be formulated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where $Q \in \mathbb{R}^{n \times d_q}$ is the query matrix derived from the current layer features, $K \in \mathbb{R}^{m \times d_k}$ and $V \in \mathbb{R}^{m \times d_v}$ are the key and value matrices derived from the feature extractor output, n is the number of query elements, m is the number of key/value elements, and d_k is the dimension of the keys.

Table 1. Pix2Next generator architecture. B = block; res = residual; attn = attention; up = upsample; down = downsample. × n denotes n consecutive identical layers. For residual: [in, out channels]. For attention: [hidden dim, heads]. For up/downsample: [channels].

Module	Configuration
Encoder	B1: res[128, 128] → res[128, 256] → res[256, 256] B2: down[256] B3: res[256, 256] → res[256, 512] → res[512, 512] B4: down[512] B5: res[512, 512] × 3 B6: down[512] B7: res[512, 512] → attn[128, 4]
Bottleneck	B1: res[512, 512] × 3 B2: res[512, 512] → attn[128, 4] → res[512, 512] B3: res[512, 512] × 3
Decoder	B1: res[512, 512] → attn[128, 4] → res[512, 512] B2: up[512] B3: res[512, 512] → res[512, 256] → res[256, 256] B4: up[256] B5: res[256, 256] × 3 B6: up[256] B7: res[256, 128] → res[128, 128]

This architectural design enables our model to capture and process multi-scale features more effectively, balancing global and local information. The combination of these elements achieves a balance between high-quality image generation, computational efficiency, generalization capability, and preservation of fine details. As a result, our model demonstrates significant improvements over previous approaches in image-to-image translation by producing detailed and contextually coherent translations from RGB to NIR domains. The use of VFM with cross-attention at multiple blocks distinguishes our approach from existing methods and contributes to the preservation of fine details and structural consistency.

3.1.3. Discriminator

We adopt the multi-scale PatchGAN architecture from Pix2pixHD for our study as the discriminator. This design utilizes three discriminators (D1, D2, D3) operating on different image scales: the original resolution and two down-sampled versions (by factors of two and four, respectively). Each discriminator uses a PatchGAN structure, divides the input image into overlapping patches, and classifies each as real or fake. The network consists of four convolutional layers (kernel size 4, stride 2), followed by leaky ReLU activations and instance normalization. The final layer produces a one-dimensional output for each patch. The varying scales result in different receptive fields: D1 focuses on fine details, while D3 captures more global structures.

Utilizing three varying resolution-focused discriminators enables more realistic image generation at various levels of detail, balanced local and global consistency, stable and reliable feedback to the generator, and computational efficiency compared to full-image discriminators. We maintained this discriminator architecture from Pix2pixHD due to its

proven effectiveness in similar image-to-image translation tasks and its compatibility with our enhanced generator.

3.2. Loss Function

We enhanced the model's performance by incorporating additional loss components into the standard loss function of generative adversarial networks [35]. Specifically, we added the Structural Similarity Index Measure (SSIM) [53] loss and the feature matching loss [24] to the traditional GAN loss.

Our key contribution lies in the novel combination of GAN, SSIM, and feature matching losses specifically optimized for NIR image generation. While these individual losses have been used separately in various contexts, their combined application in the NIR domain translation presents unique advantages: (1) the GAN loss ensures overall image quality; (2) the SSIM loss specifically preserves the structural information crucial for NIR imagery; and (3) the feature matching loss maintains domain-specific details across the RGB-NIR translation.

3.2.1. GAN Loss

The standard loss function of GANs is defined through adversarial learning between the generator and the discriminator. The generator aims to produce samples that closely resemble the real data distribution, while the discriminator attempts to distinguish between real and generated samples. This process can be defined by the following equation:

$$\begin{aligned} \min_G \max_D \mathcal{L}_{GAN}(G, D) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] \\ + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \end{aligned} \tag{2}$$

3.2.2. SSIM Loss

The SSIM loss was introduced to optimize the structural similarity between the generated and target images directly. SSIM measures the structural similarity between two images, modeling how the human visual system perceives structural information in images by considering luminance, contrast, and structure [53]. The SSIM loss is defined as follows:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{3}$$

$$\mathcal{L}_{\text{SSIM}} = 1 - \text{SSIM}(x, G(z)) \tag{4}$$

where μ_x and μ_y are the mean luminance values of the images, σ_x and σ_y are the standard deviations, σ_{xy} represents the covariance, and c_1 and c_2 are small constants added for stability. As the SSIM value ranges from -1 to 1, $\mathcal{L}_{\text{SSIM}}$ takes values between 0 and 2, where values closer to 0 indicate greater structural similarity between the two images.

By incorporating SSIM in our loss function, we ensure that our model is optimized to preserve important structural information in the image translation process. This leads generated images to be numerically similar and perceptually close to the target images.

3.2.3. Feature Matching Loss

Since RGB and NIR are different domains, the preservation of the details has higher importance. In order to penalize low-quality representations and stabilize the training of Pix2Next, we employ a feature matching loss. This loss encourages the generator to produce images that match the representations in real images at multiple feature levels of the discriminator. The feature matching loss is defined as follows:

$$\mathcal{L}_{\text{FM}}(G, D) = \mathbb{E}_{x \sim p_{\text{data}}(x)} \sum_{i=1}^{T} \frac{1}{N_i} \left\| D^{(i)}(x) - D^{(i)}(G(z)) \right\|_1 \tag{5}$$

where $D_k^{(i)}$ denotes the i-th layer feature extractor of discriminator D_k, T is the total number of layers, and N_i is the number of elements in each layer.

This loss computes the L1 distance between the feature representations of real and synthesized image pairs. By minimizing this difference across multiple layers of the discriminator, the generator learns to produce images that are statistically similar to real images at various levels of abstraction.

3.2.4. Combined Loss

To optimize the generation process effectively, we combine the previously explained loss functions into a comprehensive total loss ($\mathcal{L}_{\text{total}}$). This combined loss leverages the strengths of each individual component to guide the model toward producing high-quality NIR images. The total loss function is formulated as follows:

$$\begin{aligned} \mathcal{L}_{\text{total}} = \min_G \Big[\big(\max_{D_1, D_2, D_3} \sum_{k=1,2,3} \mathcal{L}_{\text{GAN}}(G, D_k) \big) \\ + \lambda_1 \sum_{k=1,2,3} \mathcal{L}_{\text{FM}}(G, D_k) \Big] + \lambda_2 \mathcal{L}_{\text{SSIM}} \end{aligned} \tag{6}$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GAN}} + \lambda_1 \mathcal{L}_{\text{FM}} + \lambda_2 \mathcal{L}_{\text{SSIM}} \tag{7}$$

where λ_1 and λ_2 are hyperparameters that control the relative importance of the SSIM and feature matching loss terms, respectively. In our final model, we set both λ_1 and λ_2 to 10, based on empirical experiments that showed optimal performance with these values. This combined loss function enables the model to preserve the high-quality image generation capability characteristic of GANs while simultaneously enhancing structural consistency through SSIM and feature matching.

4. Experiments

4.1. Datasets

We conducted our experiments using the RANUS [11] and IDD-AW [10] datasets, which are urban scene datasets that have spatially aligned RGB-NIR images. The RANUS dataset is particularly well suited to our research on domain translation between RGB and NIR images. The RANUS dataset consists of images with a resolution of 512 × 512 pixels and includes a total of paired 4519 RGB-NIR images. The dataset was collected over 50 different sessions and routes, covering a diverse range of scenes and objects. We randomly selected 40 out of the 50 image sequences, representing 80% of the dataset, to train our model, while the remaining 10 image sequences were reserved for testing to evaluate our model's performance on unseen categories and environments. In other words, this split strategy allowed us to assess Pix2Next's ability to generalize to new scenes that were not encountered during the training phase.

To enhance data quality, we conducted additional preprocessing steps, including a manual review to identify and remove mismatched frames that were not correctly aligned in time between the RGB and NIR image pairs. The final dataset utilized in our experiments encompassed a total of 3979 images, precisely 3179 images used for training and 800 images used for testing. Similarly, the IDD-AW dataset was employed to evaluate our model's robustness in unstructured driving environments and adverse weather conditions,

including rain, fog, snow, and low light. This dataset contains paired RGB-NIR images with pixel-level annotations, captured using a multispectral camera to ensure high-quality alignment between modalities. A total of 3430 images were used for training and 475 for testing, following the dataset's predefined split.

4.2. Training Strategy

The experiments in this study were conducted on a system equipped with four NVIDIA GeForce RTX 4090 Ti GPUs. During the training process, all images were resized to 256 $\times$ 256 to ensure efficient use of GPU memory. This choice was made to optimize performance given the hardware constraints. All models were trained around 1000 epochs, ensuring sufficient convergence. Additionally, a cosine scheduler with warmup was applied to adjust the learning rate dynamically. This scheduler gradually increases the learning rate during the warmup phase and then decreases it following a cosine function. The initial learning rate was set to 1×10^{-4} for all model training.

Evaluation Metrics

To evaluate the quality of the translated images, we employ four widely used metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [53], Fréchet Inception Distance (FID) [54], and Root Mean Square Error (RMSE). SSIM evaluates structural similarity, PSNR and RMSE measure pixel-level differences, and FID assesses the statistical similarity between generated and real images. We further enhance our evaluation approach with two additional perceptual evaluation metrics: Learned Perceptual Image Patch Similarity (LPIPS) [55] and Deep Image Structure and Texture Similarity (DISTS) [56]. LPIPS uses features from a pre-trained neural network to measure image similarity in a way that aligns with human visual perception, while DISTS evaluates both structural and textural similarities between images, also designed to mimic human visual perception.

Additionally, we include pixel-wise Standard Deviation (STD) as a supplementary metric. Pixel-wise STD measures the spatial variability of pixel intensities, indicating how consistently the translation method reproduces local image textures and details. By employing this comprehensive set of metrics, we objectively assess our model's performance from multiple perspectives, gaining a clearer understanding of both its strengths and limitations, particularly in terms of the perceptual quality of the generated images.

4.3. Quantitative and Qualitative Evaluations

We evaluate the performance of our proposed method, Pix2Next, against several image-to-image translation models on the RANUS and IDD-AW datasets. As shown in Tables 2 and 3, Pix2Next consistently outperformed the competing methods across all metrics, achieving state-of-the-art results on both the RANUS and IDD-AW datasets.

For the RANUS dataset, Pix2Next achieved a PSNR of 20.83, surpassing the best-performing baseline, Pix2pixHD, by 1.74%. In terms of SSIM, Pix2Next recorded a value of 0.8031, representing a 2.19% improvement over the next best model. Notably, the FID score was significantly reduced to 28.01, achieving a remarkable 42.96% improvement over the strongest GAN-based baseline, CycleGAN. Moreover, Pix2Next achieved lower RMSE (8.24), LPIPS (0.107), and DISTS (0.1252) values, indicating superior accuracy and perceptual quality in the generated images. The improvements in LPIPS and DISTS were particularly significant, with Pix2Next outperforming previous best results by 22.41% and 27.13%, respectively. Additionally, Pix2Next achieved a pixel-wise standard deviation (STD) of 20.37, marking a 13.67% improvement over the closest competitor and highlighting its ability to consistently reproduce local image textures and details.

Table 2. Quantitative comparison of Pix2Next with previous I2I methods on RANUS test set.

Method	Type	PSNR ↑	SSIM ↑	FID ↓	RMSE ↓	LPIPS ↓	DISTS ↓	STD ↓
Pix2pix [1] [36]	G	15.67	0.5406	87.69	9.27	0.2942	0.2141	34.18
Pix2pixHD [1] [24]	G	20.47	0.7409	53.38	8.53	0.1385	0.1742	23.60
CycleGAN [1] [37]	G	17.05	0.6679	42.97	8.98	0.1643	0.1678	33.02
BBDM [1] [38]	D	18.76	0.6614	49.29	8.74	0.1792	0.1637	26.84
C^2SAL [2] [47]	G	16.46	0.63	83.45	-	-	-	-
IRFomer [1] [20]	G	18.96	0.7857	90.89	8.76	0.2132	0.1964	26.15
UVCGAN [1] [39]	G	18.21	0.6711	46.50	8.91	0.1733	0.1656	27.30
Pix2Next (Ours)	**G**	**20.83** (+%1.74)	**0.8031** (+%2.19)	**28.01** (+%42.96)	**8.24** (+%3.45)	**0.107** (+%22.41)	**0.1252** (+%27.13)	**20.37** (+%13.67)

[1]: Models trained from scratch. [2]: Results from the paper. G: GAN-based. D: Diffusion-based. ↑: Higher is better. ↓: Lower is better. Green colored values indicate the difference between best and second best model.

Table 3. Quantitative comparison of Pix2Next with previous I2I methods on IDD-AW test set.

Method	Type	PSNR ↑	SSIM ↑	FID ↓	RMSE ↓	LPIPS ↓	DISTS ↓	STD ↓
Pix2pix [36]	G	29.14	0.8735	42.97	5.66	0.0951	0.1317	11.32
Pix2pixHD [24]	G	28.53	0.8716	63.23	6.04	0.0935	0.1803	11.61
CycleGAN [37]	G	21.17	0.7665	60.26	8.36	0.1664	0.2046	21.16
BBDM [38]	D	19.11	0.6316	122.1	8.72	0.2932	0.3044	27.53
IRFomer [20]	G	27.07	0.9041	88.16	5.76	0.1152	0.1596	12.99
UVCGAN [39]	G	27.63	0.8690	40.09	6.215	0.1077	0.1289	13.12
Pix2Next (Ours)	**G**	**30.41** (+%4.26)	**0.9228** (+%1.95)	**32.81** (+%20.17)	**5.06** (+%11.86)	**0.0663** (+%32.78)	**0.1040** (+%22.44)	**10.55** (+%6.8)

G: GAN-based. D: Diffusion-based. ↑: Higher is better. ↓: Lower is better. All models trained from scratch with same training split, evaluated on same test split. Green colored values indicate the difference between best and second best model.

On the IDD-AW dataset, Pix2Next further demonstrated its robustness under diverse and adverse conditions. It achieved a PSNR of 30.41, reflecting a 4.26% improvement over Pix2pix, the next best-performing model. The SSIM reached 0.9228, representing a 1.95% increase compared to the previous best. The FID score was reduced to 32.71, showing a 20.17% improvement over all baseline methods. Pix2Next also outperformed competing models in terms of RMSE (5.06), LPIPS (0.0664), and DISTS (0.1040), achieving relative improvements of 11.86%, 32.78%, and 22.44%, respectively. The pixel-wise STD reached 10.55, representing a 6.8% improvement over the closest model, further demonstrating its consistency in reproducing fine textures. These substantial improvements across both datasets clearly demonstrate the effectiveness of our proposed model in generating high-quality NIR images under various conditions. For a fair comparison, all experiments were conducted using the default parameters provided by the original implementations of Pix2pix, Pix2pixHD, CycleGAN, BBDM, IRFormer, and UVCGAN.

Figures 7 and 8 showcase the qualitative performance of Pix2Next compared to other image translation methods, including Pix2pix, Pix2pixHD, CycleGAN, BBDM, IRFormer, and UVCGAN, alongside the ground truth (GT). The results clearly demonstrate Pix2Next's superior ability to preserve image details and produce realistic outputs.

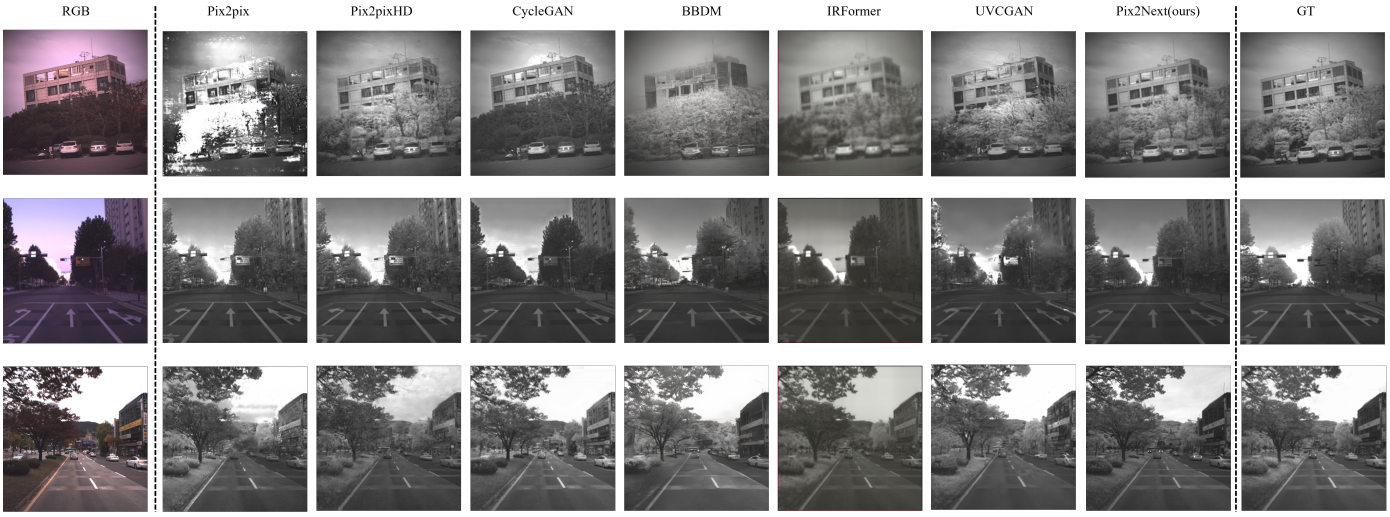


Figure 7. Qualitative evaluation on the RANUS dataset. The results demonstrate consistency with the quantitative comparisons, highlighting that our method produces outputs closest to the ground truth NIR data.

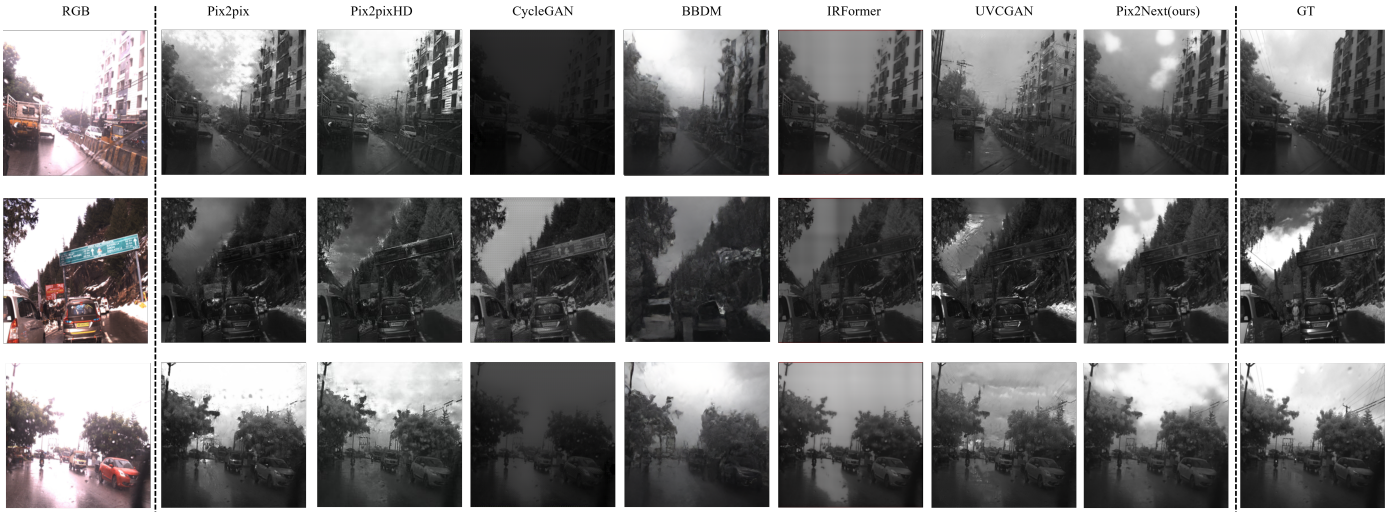


Figure 8. Qualitative evaluation on the IDD-AW dataset. The results demonstrate consistency with the quantitative comparisons.

In a qualitative assessment against other methods, Pix2Next delivers images with sharper details and fewer artifacts such as spatial distortion and under-styling. For example, in the first row of Figure 7, Pix2Next effectively maintains the structural integrity of the building and surrounding vegetation, whereas Pix2pix and Pix2pixHD suffer from significant distortions and loss of detail. Similarly, CycleGAN and BBDM generate outputs with visible artifacts and less accurate texture representation, particularly in the foliage and architectural elements. In contrast, Pix2Next closely matches the ground truth images, which highlights its superior capability to maintain both global consistency and fine details.

In the second and third rows, which depict street scenes, Pix2Next again provides the most visually coherent results, with well-preserved road markings, traffic lights, and natural-looking foliage. Other methods, especially Pix2pix and CycleGAN, exhibit significant artifacts and unnatural textures, further underscoring the robustness of Pix2Next in complex scenes. Although BBDM performs relatively well, it still fails to achieve the sharpness and clarity observed in Pix2Next's results.

Overall, Pix2Next consistently delivers the highest quality images across all scenes, closely matching the ground truth and demonstrating superior performance in preserving both global structures and fine-grained details, while significantly reducing visual artifacts compared to existing methods. Additionally, some details are not kept when a scene is captured with NIR cameras such as colors, the effect of light sources, etc. Therefore, models need to learn to preserve some features while also losing others when converting an RGB image to an NIR image. A more detailed analysis of these qualitative differences, including pixel-level comparisons and additional visual examples, is provided in Figure 9.

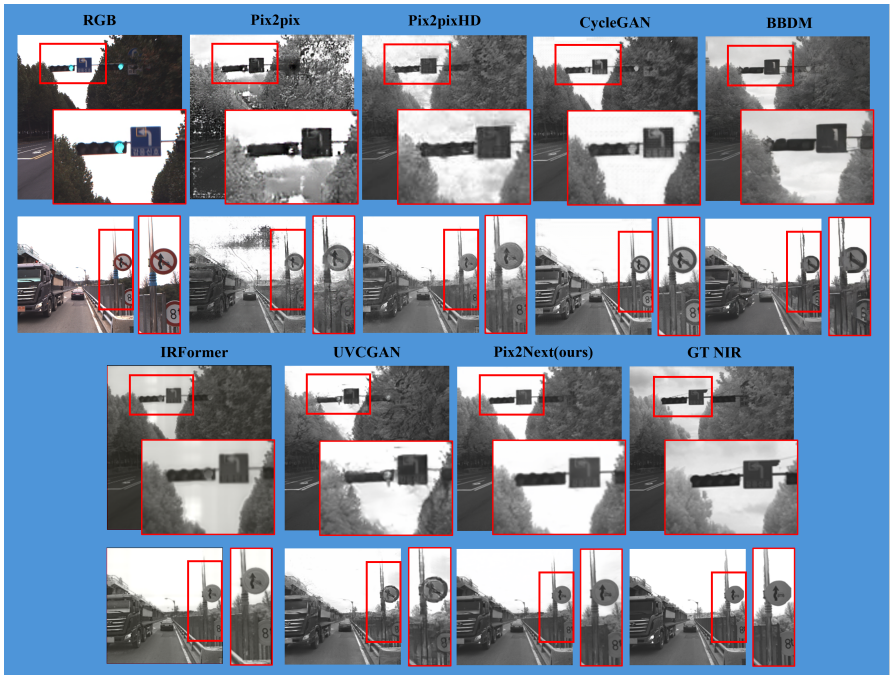


Figure 9. Comparative evaluation of generated images across compared models. Zoomed-in areas show the capability of models to preserve details.

4.4. Ablation Study

4.4.1. Effectiveness of Extractor

To evaluate the effectiveness of the feature extractor in our proposed method, we conducted an ablation study by comparing the performance of the model without a feature extractor (W/O Extractor) to versions using different vision foundation models as feature extractors. As shown in Table 4, the model without a feature extractor yields an FID of 31.26, LPIPS of 0.1116, and DISTS of 0.132. These results indicate that the absence of a feature extractor leads to suboptimal performance. On the other hand, using advanced models like the Vision Transformer (ViT) and SwinV2 shows clear improvements over the absence of an extractor. The ViT-based extractor achieves an FID of 29.05, LPIPS of 0.1185, and DISTS of 0.1338, while using the SwinV2-based extractor results in an FID of 30.24, LPIPS of 0.1117, and DISTS of 0.1299, both outperforming the model without an extractor.

The best results are achieved with the Internimage-based feature extractor, which significantly enhances the model's performance, achieving the lowest FID of 28.01, LPIPS of 0.107, and DISTS of 0.1252. This indicates that the choice of feature extractor is crucial for optimizing model performance, with the Internimage model providing the most significant improvements in image quality and perceptual metrics. A qualitative comparison of the effectiveness of employing a feature extractor is given in Figure 10. As revealed in the figure, the generator can eliminate spatial distortion and under-stylization problems thanks to the inclusion of features obtained from the extractor through cross-attention.



Figure 10. NIR generation comparison with and without the feature extractor module.

Table 4. Effectiveness of extractor.

Model	FID ↓	LPIPS ↓	DISTS↓
W/O Extractor	31.26	0.1116	0.1320
ResNet	35.92	0.1269	0.1524
ViT	29.05	0.1185	0.1338
SwinV2	30.24	0.1117	0.1299
Internimage	**28.01**	**0.1070**	**0.1252**

↓: lower is better. Bold values indicate the best model.

4.4.2. Effectiveness of Attention Position

To determine the optimal position for applying attention mechanisms within our network, we conducted an ablation study comparing two configurations on Pix2Next(SwinV2): applying attention solely at the "B" ottleneck layer (B-attention) versus applying attention across all key stages of the network, meaning the "E"ncoder, "B"ottleneck, and "D"ecoder (EBD-attention) layers. The results of this study are presented in Table 5. When attention is distributed across the encoder, bottleneck, and decoder stages, the model shows notable improvements across all metrics. Specifically, the SSIM increases to 0.8063, and the FID decreases significantly to 30.24, indicating better alignment with the ground truth images. Additionally, LPIPS is reduced to 0.1117 and DISTS to 0.1299, suggesting that applying attention throughout the network leads to better feature representation and more accurate image translation. These findings suggest that distributing attention across multiple stages of the network—rather than concentrating it solely on the bottleneck—leads to superior performance in image translation tasks. The application of attention throughout the encoder, bottleneck, and decoder allows the model to effectively capture and refine features at various levels of abstraction.

Table 5. Effectiveness of attention position.

Model	SSIM ↑	FID ↓	LPIPS ↓	DISTS ↓
B-attention	0.7903	37.02	0.1131	0.1353
EBD-attention	**0.8063**	**30.24**	**0.1117**	**0.1299**

↓: lower is better. ↑: higher is better. Bold values indicate the best model.

4.4.3. Effectiveness of Generator

To assess the effectiveness of the generator design in our proposed method, we conducted an ablation study comparing the performance of the baseline Pix2pixHD model, a modified version of Pix2pixHD where residual blocks are replaced with our extractor (Internimage-based) blocks, and our full model integrating both the Internimage-based feature extractor and our encoder–decoder based generator. The results are summarized in Table 6. The baseline Pix2pixHD model, which uses traditional residual blocks, achieves a PSNR of 20.474, SSIM of 0.7409, FID of 53.38, and RMSE of 8.53. These metrics serve as the foundation for evaluating the enhancements brought by the modifications. By replacing the residual blocks with Internimage blocks, the Pix2pixHD+Internimage model shows improvements in most of the metrics. Specifically, there is a slight increase in PSNR to 20.87 and a reduction in FID to 45.14, indicating better image quality and closer alignment with the ground truth distribution. However, the SSIM decreases to 0.7327. These results suggest that while the integration of Internimage blocks improves certain aspects of image quality, it may not universally enhance all performance metrics. Our full model, which incorporates both the Internimage-based feature extractor and encoder–decoder-based generator, delivers the best performance across all metrics. The substantial improvement in SSIM and FID highlights the effectiveness of our encoder–decoder-based generator architecture.

Table 6. Effectiveness of generator.

Model	PSNR ↑	SSIM ↑	FID ↓	RMSE ↓
Pix2pixHD (Baseline)	20.474	0.7409	53.38	8.53
Pix2pixHD+Internimage	**20.87**	0.7327	45.14	8.35
Ours	20.83	**0.8031**	**28.01**	**8.21**

↓: lower is better. ↑: higher is better. Bold values indicate the best model.

4.5. Effectiveness of Generated NIR Data

To assess the effectiveness of the NIR data generated by our model, we performed an ablation study on a downstream object detection task. To achieve this, we employed the Co-DETR model [57], which is currently the state-of-the-art object detection model. We followed two different methods while finetuning the Co-DETR model:

In the first method, we used the object annotations in the RANUS dataset and finetuned the model using the training split of the RANUS dataset (Finetune w/ Ranus). In the second method, in order to evaluate the generalizability of our proposed translation model to unseen data, we generated 10,000 NIR images from RGB images of the BDD100k dataset [8] (results are given in Figure 11). These images were used to scale up the RANUS training set (Figure 12), and the newly scaled-up dataset was employed to finetune the Co-DETR model (Finetune w/ Ranus + Gen NIR). Additionally, to establish a baseline for comparison, we also reported the object detection performance of the Co-DETR model on the same test set without any finetuning (RGB-pretrain).

As for the details of the experiment, we merged the “truck”, “bus”, and “car” labeled images into a single “car” class and “bicycle” and “motorcycle” labeled images into a single “bicycle” class while ignoring the remaining classes.



Figure 11. Zero-shot RGB to NIR translation results on BDD100k [8] dataset.

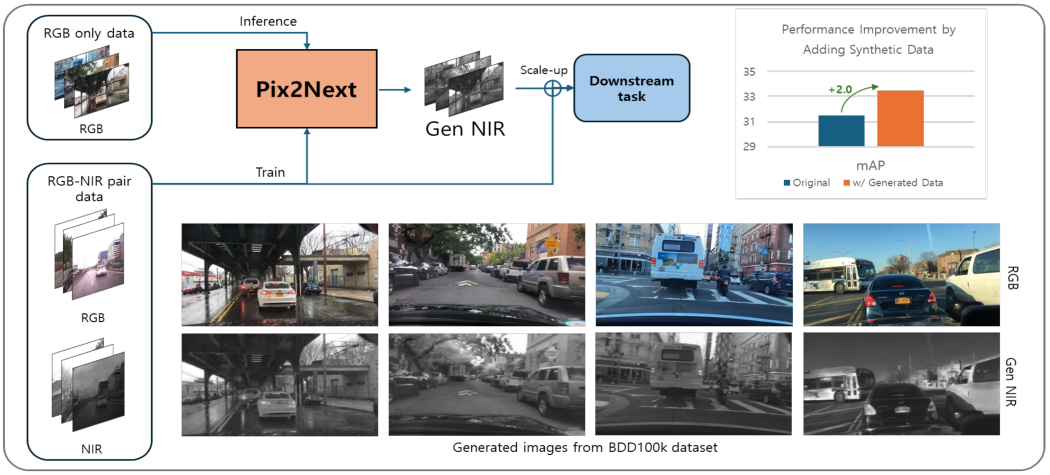


Figure 12. Overview of object detection downstream task pipeline.

As shown in Table 7, the model trained on both the RANUS NIR data and the generated NIR data achieved the highest performance, with a mean Average Precision (mAP) of 0.3347, compared to 0.3149 when trained only on the RANUS data, and 0.2724 when using the RGB-pretrained model without additional NIR training. Notably, the class-specific Average Precision (AP) for bicycles improved significantly from 0.2143 to 0.2829 with the addition of the generated NIR data.

Table 7. Effectiveness of generation data.

Method	mAP	AP_{person}	$AP_{bicycle}$	AP_{car}
RGB_pretrain	0.2724	0.1551	0.1745	0.4874
Finetune w/ranus	0.3149	0.1682	0.2143	**0.5622**
Finetune w/ranus + generated NIR	**0.3347**	**0.1704**	**0.2829**	0.5507

Bold values indicate the best model.

These results demonstrate the effectiveness of using large-scale RGB images and annotations to translate NIR data to scale up the available NIR training dataset without the need for additional NIR data acquisition and annotation. By leveraging our translated NIR data, we significantly enhanced the performance of object detection in the NIR domain, which confirms the value of our method in scenarios where NIR data are limited.

4.6. LWIR Translation

To explore the translation capabilities of our model at different wavelengths, we conducted further experiments on LWIR translation using the aligned FLIR dataset [58]. This dataset comprises 4113 aligned RGB-LWIR image pairs for training and 1029 pairs for testing. Specifically, we trained our Pix2Next (SwinV2) model on the dataset's training set and reported the evaluation results on the same test set, comparing them with other methods from the literature (Table 8).

Our model achieved state-of-the-art performance compared to existing methods as reported in the literature [20]. These results validate the effectiveness of Pix2Next in the LWIR domain and also suggest promising avenues for expanding the translation capabilities to other wavelength images in future work.

Table 8. Quantitative comparison on the LWIR dataset, with results taken from IRFormer [20]. Negative values indicate poorer reconstruction performance.

Method	PSNR ↑	SSIM ↑
CycleGAN [37]	3.45	0.01
Pix2pix [36]	4.19	0.05
UNIT [59]	3.11	0.01
MUNIT [60]	3.65	0.02
BCI [61]	11.14	0.21
IRFormer [20]	17.74	0.48
Ours	**23.45**	**0.66**

↑: higher is better. Bold values indicate the best model.

5. Discussion and Failure Cases

Unlike traditional methods, our model leverages a vision foundation model to extract global features and employs cross-attention mechanisms to effectively integrate these features into the generator. This method enables our model to preserve both the overall structure and fine details of the RGB domain, resulting in generated images that are closer to the ground truth compared to existing methods. As a result, it achieves state-of-the-art image generation performance on the RANUS and IDD-AW datasets.

While the proposed translation model demonstrates robust performance in generating NIR images from RGB inputs, there is still room for improvement, especially in instances where it fails to accurately reproduce certain material properties, as illustrated in Figure 13. Specifically, the model encounters challenges in replicating the unique reflectance characteristics of particular materials, notably cloth, and vehicle lights. This shortcoming may be attributed to an underrepresentation of paired images exhibiting these specific characteristics within our training datasets.

To overcome these challenges, we plan to continuously refine the model architecture. A promising direction is the integration of diffusion-based models, which have demonstrated potential in capturing fine-grained details and enhancing the robustness of image generation across diverse scenarios.



Figure 13. Fail case example: The top row displays the NIR GT images, and the bottom row shows our generated NIR images. The red boxes highlight a failure in representing the material properties of some objects.

6. Conclusions and Future Work

In this paper, we proposed a novel image translation model, Pix2Next, designed to address the challenges of generating NIR images from RGB inputs. Our model leverages the strengths of state-of-the-art vision foundation models, combined with an encoder–decoder architecture that incorporates cross-attention mechanisms, to produce high-quality NIR images from RGB images.

Our extensive experiments, including quantitative and qualitative evaluations as well as ablation studies, demonstrated that Pix2Next outperforms existing image translation models across various metrics. The model showed significant improvements in image quality, structural consistency, and perceptual realism, as evidenced by superior performance in PSNR, SSIM, FID, and other evaluation metrics. Furthermore, our zero-shot experiment on the BDD100k dataset confirmed the model's robust generalization capabilities to unseen data. We validated the utility of Pix2Next by demonstrating performance improvements in an object detection downstream task, achieved by scaling up limited NIR data using our generated images.

In future work, we aim to extend the application of this architecture to other multispectral domains, such as RGB to extended infrared (XIR) translation, to broaden the scope of our model's applicability. Additionally, the label-driven contrastive learning techniques proposed in [62] could be incorporated into our framework if additional annotations (e.g., object labels or bounding boxes) become available. Incorporating category-level or bounding-box information would introduce stronger semantic constraints when mapping to the NIR domain, potentially leading to more accurate object representations and improved performance in downstream tasks like object detection and recognition. By integrating these label-driven constraints into our existing cross-attention mechanism, future work could further enhance both the visual fidelity of the generated images and their practical utility.

Author Contributions: Conceptualization, Y.J. and I.P.; methodology, Y.J. and I.P.; software, Y.J. and I.P.; validation, Y.J. and I.P.; formal analysis, Y.J., I.P. and Y.N.; investigation, H.S., H.J. and Y.N.; resources, S.K.; data curation, H.J.; writing—original draft preparation, Y.J. and I.P.; writing—review and editing, S.K. and Y.N.; visualization, H.S.; supervision, S.K.; funding acquisition, S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a Korea Evaluation Institute of Industrial Technology (KEIT) grant funded by the Korea government (MOTIE) (RS-2022-00154651—3D semantic camera module development capable of material and property recognition).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available upon request. Our code is available at https://github.com/Yonsei-STL/pix2next (accessed on 30 August 2024).

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

References

1. Bijelic, M.; Gruber, T.; Ritter, W. Benchmarking image sensors under adverse weather conditions for autonomous driving. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 26–30 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1773–1779.
2. Wu, J.; Wei, P.; Huang, F. Color-preserving visible and near-infrared image fusion for removing fog. *Infrared Phys. Technol.* **2024**, *138*, 105252. [CrossRef]
3. Infiniti Electro-Optics. NIR (Near-Infrared) Imaging (Fog/Haze Filter). 2024. Available online: https://www.infinitioptics.com/technology/nir-near-infrared (accessed on 30 August 2024).
4. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 3354–3361.
5. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11621–11631.
6. Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. Scalability in perception for autonomous driving: Waymo open dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2446–2454.
7. Chang, M.F.; Lambert, J.; Sangkloy, P.; Singh, J.; Bak, S.; Hartnett, A.; Wang, D.; Carr, P.; Lucey, S.; Ramanan, D.; et al. Argoverse: 3d tracking and forecasting with rich maps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8748–8757.
8. Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; Darrell, T. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2636–2645.
9. Hwang, S.; Park, J.; Kim, N.; Choi, Y.; So Kweon, I. Multispectral pedestrian detection: Benchmark dataset and baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1037–1045.
10. Shaik, F.A.; Reddy, A.; Billa, N.R.; Chaudhary, K.; Manchanda, S.; Varma, G. Idd-aw: A benchmark for safe and robust segmentation of drive scenes in unstructured traffic and adverse weather. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 4614–4623.
11. Choe, G.; Kim, S.H.; Im, S.; Lee, J.Y.; Narasimhan, S.G.; Kweon, I.S. RANUS: RGB and NIR urban scene dataset for deep scene parsing. *IEEE Robot. Autom. Lett.* **2018**, *3*, 1808–1815. [CrossRef]
12. Brown, M.; Süsstrunk, S. Multi-spectral SIFT for scene category recognition. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 177–184.
13. Mortimer, P.; Wuensche, H.J. TAS-NIR: A VIS+ NIR Dataset for Fine-grained Semantic Segmentation in Unstructured Outdoor Environments. *arXiv* **2022**, arXiv:2212.09368. [CrossRef]

14. An, L.; Zhao, J.; Di, H. Generating infrared image from visible image using Generative Adversarial Networks. In Proceedings of the 2019 IEEE International Conference on Unmanned Systems (ICUS), Beijing, China, 17–19 October 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 157–161.
15. Yuan, X.; Tian, J.; Reinartz, P. Generating artificial near infrared spectral band from rgb image using conditional generative adversarial network. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2020**, *3*, 279–285. [CrossRef]
16. Aslahishahri, M.; Stanley, K.G.; Duddu, H.; Shirtliffe, S.; Vail, S.; Bett, K.; Pozniak, C.; Stavness, I. From RGB to NIR: Predicting of near infrared reflectance from visible spectrum aerial images of crops. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 1312–1322.
17. Zhao, G.; He, Y.; Wang, Z.; Wu, H.; Cheng, L. Generation of NIR Spectral Band from RGB Image with Wavelet Domain Spectral Extrapolation Generative Adversarial Network. *Comput. Electron. Agric.* **2024**, *227*, 109461. [CrossRef]
18. Uddin, M.S.; Kwan, C.; Li, J. MWIRGAN: Unsupervised Visible-to-MWIR Image Translation with Generative Adversarial Network. *Electronics* **2023**, *12*, 1039. [CrossRef]
19. Kniaz, V.V.; Knyaz, V.A.; Hladuvka, J.; Kropatsch, W.G.; Mizginov, V. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018; pp. 1–20.
20. Chen, Y.; Chen, P.; Zhou, X.; Lei, Y.; Zhou, Z.; Li, M. Implicit Multi-Spectral Transformer: An Lightweight and Effective Visible to Infrared Image Translation Model. *arXiv* **2024**, arXiv:2404.07072. [CrossRef]
21. Özkanoğlu, M.A.; Ozer, S. InfraGAN: A GAN architecture to transfer visible images to infrared domain. *Pattern Recognit. Lett.* **2022**, *155*, 69–76. [CrossRef]
22. Luo, Y.; Pi, D.; Pan, Y.; Xie, L.; Yu, W.; Liu, Y. ClawGAN: Claw connection-based generative adversarial networks for facial image translation in thermal to RGB visible light. *Expert Syst. Appl.* **2022**, *191*, 116269. [CrossRef]
23. Berg, A.; Ahlberg, J.; Felsberg, M. Generating visible spectrum images from thermal infrared. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1143–1152.
24. Wang, T.C.; Liu, M.Y.; Zhu, J.Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8798–8807.
25. Shen, Z.; Huang, M.; Shi, J.; Xue, X.; Huang, T.S. Towards instance-level image-to-image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3683–3692.
26. Huang, J.; Liao, J.; Kwong, S. Unsupervised image-to-image translation via pre-trained stylegan2 network. *IEEE Trans. Multimed.* **2021**, *24*, 1435–1448. [CrossRef]
27. Liang, J.; Zeng, H.; Zhang, L. High-resolution photorealistic image translation in real-time: A laplacian pyramid translation network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9392–9400.
28. Shen, Y.; Kang, J.; Li, S.; Yu, Z.; Wang, S. Style Transfer Meets Super-Resolution: Advancing Unpaired Infrared-to-Visible Image Translation with Detail Enhancement. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 4340–4348.
29. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
30. Li, H.; Gu, C.; Wu, D.; Cheng, G.; Guo, L.; Liu, H. Multiscale generative adversarial network based on wavelet feature learning for SAR-to-optical image translation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
31. Ismael, S.F.; Kayabol, K.; Aptoula, E. Unsupervised domain adaptation for the semantic segmentation of remote sensing images via one-shot image-to-image translation. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
32. Du, W.L.; Zhou, Y.; Zhu, H.; Zhao, J.; Shao, Z.; Tian, X. A semi-supervised image-to-image translation framework for SAR–optical image matching. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
33. Roy, S.; Siarohin, A.; Sangineto, E.; Sebe, N.; Ricci, E. Trigan: Image-to-image translation for multi-source domain adaptation. *Mach. Vis. Appl.* **2021**, *32*, 1–12. [CrossRef]
34. Pizzati, F.; Charette, R.D.; Zaccaria, M.; Cerri, P. Domain bridge for unpaired image-to-image translation and unsupervised domain adaptation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 2990–2998.
35. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial net. *Adv. Neural Inf. Process. Syst.* **2014**, *27*.

36. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
37. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
38. Li, B.; Xue, K.; Liu, B.; Lai, Y.K. Bbdm: Image-to-image translation with brownian bridge diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1952–1961.
39. Torbunov, D.; Huang, Y.; Yu, H.; Huang, J.; Yoo, S.; Lin, M.; Viren, B.; Ren, Y. Uvcgan: Unet vision transformer cycle-consistent gan for unpaired image-to-image translation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 702–712.
40. Kumar, W.K.; Singh, N.J.; Singh, A.D.; Nongmeikapam, K. Enhanced machine perception by a scalable fusion of RGB–NIR image pairs in diverse exposure environments. *Mach. Vis. Appl.* **2021**, *32*, 88. [CrossRef]
41. Luo, Y.; Remillard, J.; Hoetzer, D. Pedestrian detection in near-infrared night vision system. In Proceedings of the 2010 IEEE Intelligent Vehicles Symposium, San Diego, CA, USA, 21–24 June 2021; IEEE: Piscataway, NJ, USA, 2010; pp. 51–58.
42. Liu, S.; Gao, M.; John, V.; Liu, Z.; Blasch, E. Deep learning thermal image translation for night vision perception. *ACM Trans. Intell. Syst. Technol. (TIST)* **2020**, *12*, 1–18. [CrossRef]
43. Govardhan, P.; Pati, U.C. NIR image based pedestrian detection in night vision with cascade classification and validation. In Proceedings of the 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, Ramanathapuram, India, 8–10 October 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 1435–1438.
44. Bhowmick, S.; Kuiry, S.; Das, A.; Das, N.; Nasipuri, M. Deep learning-based outdoor object detection using visible and near-infrared spectrum. *Multimed. Tools Appl.* **2022**, *81*, 9385–9402. [CrossRef]
45. Ippalapally, R.; Mudumba, S.H.; Adkay, M.; HR, N.V. Object detection using thermal imaging. In Proceedings of the 2020 IEEE 17th India Council International Conference (INDICON), New Delhi, India, 10–13 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
46. Zhu, J.Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A.A.; Wang, O.; Shechtman, E. Toward multimodal image-to-image translation. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
47. Mao, K.; Yang, M.; Wang, H. Infrared and Near-Infrared Image Generation via Content Consistency and Style Adversarial Learning. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Beijing, China, 4–7 November 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 618–630.
48. Wang, W.; Dai, J.; Chen, Z.; Huang, Z.; Li, Z.; Zhu, X.; Hu, X.; Lu, T.; Lu, L.; Li, H.; et al. InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 17804–17815.
49. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
50. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
51. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
52. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
53. 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]
54. 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*.
55. 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 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.
56. Ding, K.; Ma, K.; Wang, S.; Simoncelli, E.P. Image quality assessment: Unifying structure and texture similarity. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 2567–2581. [CrossRef] [PubMed]
57. Zong, Z.; Song, G.; Liu, Y. Detrs with collaborative hybrid assignments training. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 6748–6758.
58. FLIR Dataset FREE Teledyne FLIR Thermal Dataset for Algorithm Training. 2024. Available online: https://www.flir.com/oem/adas/adas-dataset-form/?srsltid=AfmBOoqQxC-X5fpQ7xRxOq895ks2E3reFxCPv0l8aMJPX4UWz_4kiAEU (accessed on 30 August 2024).

59. Liu, M.Y.; Breuel, T.; Kautz, J. Unsupervised image-to-image translation networks. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
60. Huang, X.; Liu, M.Y.; Belongie, S.; Kautz, J. Multimodal unsupervised image-to-image translation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 172–189.
61. Liu, S.; Zhu, C.; Xu, F.; Jia, X.; Shi, Z.; Jin, M. Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1815–1824.
62. Liu, C.; Wen, J.; Xu, Y.; Zhang, B.; Nie, L.; Zhang, M. Reliable Representation Learning for Incomplete Multi-View Missing Multi-Label Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, 1–17. [CrossRef]

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.