

DiffV2IR: Visible-to-Infrared Diffusion Model via Vision-Language Understanding

Lingyan Ran
Northwestern Polytechnical University

Lidong Wang
Northwestern Polytechnical University

Guangcong Wang
Great Bay University

Peng Wang
Northwestern Polytechnical University

Yanning Zhang
Northwestern Polytechnical University

Abstract

The task of translating visible-to-infrared images (V2IR) is inherently challenging due to three main obstacles: 1) achieving semantic-aware translation, 2) managing the diverse wavelength spectrum in infrared imagery, and 3) the scarcity of comprehensive infrared datasets. Current leading methods tend to treat V2IR as a conventional image-to-image synthesis challenge, often overlooking these specific issues. To address this, we introduce DiffV2IR, a novel framework for image translation comprising two key elements: a Progressive Learning Module (PLM) and a Vision-Language Understanding Module (VLUM). PLM features an adaptive diffusion model architecture that leverages multi-stage knowledge learning to infrared transition from full-range to target wavelength. To improve V2IR translation, VLUM incorporates unified Vision-Language Understanding. We also collected a large infrared dataset, IR-500K, which includes 500,000 infrared images compiled by various scenes and objects under various environmental conditions. Through the combination of PLM, VLUM, and the extensive IR-500K dataset, DiffV2IR markedly improves the performance of V2IR. Experiments validate DiffV2IR's excellence in producing high-quality translations, establishing its efficacy and broad applicability. The code, dataset, and DiffV2IR model will be available at <https://github.com/LidongWang-26/DiffV2IR>.

1. Introduction

Visible images are the most common form of digital imagery, while infrared images hold significant importance in various practical applications, such as thermal imag-

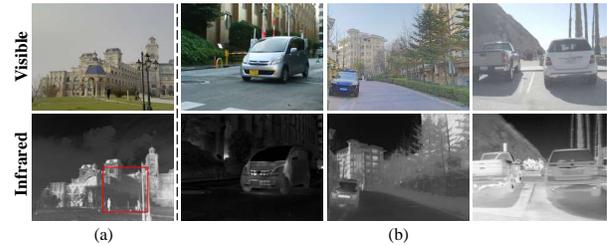


Figure 1. Main challenges of V2IR. (a) Semantic-aware translation, in which the context information of shadow influences the infrared image a lot. (b) Diverse infrared radiations. Even similar visual scenes from different infrared cameras show the diversity of infrared imagery. For the second column to the fourth column, the infrared intensity significantly changes.

ing, night vision monitoring, and environmental sensor data analysis. Currently, lots of foundation visual models are built on annotated large visible datasets. An intuitive way to narrow the gap between visual models and infrared models is to translate visual images into infrared images (V2IR).

However, V2IR presents a difficult task due to three critical challenges. **First**, V2IR is highly dependent on semantic information. Infrared imaging, also known as thermal imaging, captures and visualizes the infrared radiation emitted by objects. Therefore, V2IR relies on scene understanding such as object semantics and context information (e.g., solar radiation, lighting, and shadow). **Second**, infrared imaging primarily relies on infrared radiation, which is typically divided into several different wavelength ranges such as Near-Infrared (NIR), Short-Wave Infrared (SWIR), Mid-Wave Infrared (MWIR), and Long-Wave Infrared (LWIR). Different infrared radiation intensity leads to various pixel values

for the same scene. Images captured in the same wavelength range might still vary due to differences in infrared camera sensors. *Third*, unlike visual images that can be easily captured by anyone with widely-used phones and traditional cameras, infrared images are limited to a few people who have infrared cameras. Therefore, there are only a few public infrared datasets captured by different types of infrared camera. It is unclear how to train a large foundation model of infrared image generation on limited datasets.

Existing methods make initial attempts at V2IR. A common way is to directly formulate the V2IR task as image-to-image translation, with methods such as Variational Autoencoders (VAEs) [21, 30, 53], Generative Adversarial Networks (GANs) [2, 23, 37], and diffusion models [17, 54, 55]. To exploit semantic information (Challenge 1) and reduce the impact of diverse infrared radiation from different infrared cameras (Challenge 2), these methods integrate different low-level semantics into image-to-image translation models, such as edge prior, structural similarity, geometry information, and physical constraint. For example, InfraGAN [48] uses structural similarity as an additional loss function and a pixel-level discriminator. EG-GAN [33, 40] focused on edge preservation. DR-AVIT [14] achieved diverse and realistic aerial visible-to-infrared image translation by integrating a geometry-consistency constraint. TM-GAN [42] incorporates the image-matching process into image-to-image translation. PID [43] incorporated strong physical constraints and used a latent diffusion model. Although these methods significantly improve V2IR, they do not make full use of semantics and do not consider different infrared radiation from different infrared cameras.

To tackle the challenges of V2IR, in this paper, we present DiffV2IR, a novel V2IR diffusion framework that integrates vision-language understanding into a diffusion model with multi-stage knowledge learning. Specifically, to achieve semantic-aware V2IR translation, we extract a detailed scene description by integrating a Vision-Language Understanding Module (VLUM) into the optimization (Challenge 1). To achieve stable V2IR translation trained on the datasets that contain different infrared radiation from different infrared cameras, we propose a Progressive Learning Module (PLM) that features an adaptive diffusion model architecture that leverages multi-stage knowledge learning to transition from full-range to target wavelength (Challenge 2). To train a large-scale V2IR diffusion model, we assembled an extensive infrared dataset named IR-500K, comprising 500,000 infrared images. The IR-500K integrates nearly every substantial publicly accessible infrared dataset. This fusion of scale, diversity, and accessibility establishes the dataset as a pivotal resource for enhancing infrared image generation technologies(Challenge 3). With DiffV2IR and IR-500K, our work significantly improves the performance of V2IR.

Overall, the main contributions are: 1) We propose a novel DiffV2IR framework that integrates a multi-modal vision-language model into a unified optimization and thus achieves semantic-aware V2IR translation. 2) To enable DiffV2IR to perform stable V2IR translation on an infrared dataset with various infrared radiation, we propose a progressive learning module that leverages multi-stage knowledge learning. 3) To train a large DiffV2IR model, we collect a large infrared dataset, IR-500K. Experiments demonstrate the effectiveness of DiffV2IR.

2. Related Works

Diffusion Models. Recently, diffusion models such as those in [47, 56, 69] have made significant strides in image generation. Acting as generative systems, these models emulate physical diffusion processes by gradually introducing noise to learn how to produce clear images. Unlike traditional generative models like GANs [12, 39, 44] and VAEs [30, 53], diffusion models generate superior-quality samples for high-resolution image creation and provide a training process less prone to mode collapse. The concept of diffusion models was first presented in [55]. DDPMs [17] proposed denoising diffusion probabilistic models, which captured significant attention and broadened the application of diffusion models in image generation. Efforts have since been aimed at enhancing their efficiency and production quality. Latent diffusion models (LDMs) [54] executed the diffusion process in a compressed latent space, greatly reducing computational overhead. These models have excelled in image generation and denoising. However, they remain underutilized in multispectral image translation.

Image-to-Image Translation. Image translation algorithms are designed to learn either a pixel-wise correspondence or a joint probability distribution to facilitate the translation of images from one domain to another. Pix2Pix [24], a foundational work in the field, utilizes a conditional generative adversarial network (cGAN) [45] to develop a pixel-level map between input and output images. Expanding on this, Pix2PixHD [63] explores techniques for producing high-resolution images of superior quality. These methods require paired images for training. Introducing a different approach, CycleGAN [74], DiscoGAN [29] and DualGAN [70] utilizes unpaired datasets by implementing a cycle consistency loss, which guarantees that the mapping from source to target and back to source retains the original content. Then many models like [3, 6, 7, 65, 67] utilize cycle consistency for unpaired training. [20, 34, 62] assume that the representation can be disentangled into domain-invariant semantic structure features and domain-specific style features. [1, 27, 28, 59] implement attention mechanism in image translation.

With the rapid advancement of diffusion models, a variety of conditional diffusion models that incorporate text

and spatial information have achieved notable success in image translation [5, 8, 11, 35, 61, 66]. InstructPix2Pix [4] employs two large pre-trained models (GPT-3 and Stable Diffusion) to generate an extensive dataset of input-goal-instruction triplet examples and trains a model for image editing based on instructions using this dataset. ControlNet [73] and T2I-Adapter [46] are devoted to making the diffusion process more controllable by introducing various conditions. Although many methods achieve great success in image translation, they do not consider the challenges of V2IR, which heavily depends on scene understanding.

Visible-to-Infrared Image Translation. Several models have attempted to translate visible images to infrared images. Initially, some research focused on generating infrared data tailored for specific tasks like tracking [72] and person re-identification [32], treating it mainly as a pixel generation challenge. Content structure serves as a crucial prior in producing meaningful infrared images. InfraGAN [48] incorporates structural similarity as an auxiliary loss and uses a pixel-level discriminator for V2IR image translation. EG-GAN [33, 40] highlights edge preservation as an effective approach, confirmed by improved outcomes in training deep TIR optical flow and object detection against other benchmarks. VQ-InfraTrans [57] introduces a two-step transfer strategy using a composite encoder and decoder from VQ-GAN [9], alongside a multi-path transformer. DR-AVIT [14] enhances the translation of aerial visible-to-infrared images with disengaged representation learning, separating image representations into a domain-invariant semantic structure space and two domain-specific imaging style spaces. PID [43] further advances this area by integrating significant physical constraints and for the first time employing a latent diffusion model. However, current V2IR methods do not fully exploit key semantic information and struggle to create high-quality infrared images since various infrared imaging.

3. Methodology

In this paper, we propose a DiffV2IR framework based on diffusion models, which can translate visible images into infrared images (V2IR). Different from large-scale visual image datasets that can easily be collected by widely-used phones and RGB cameras, it is difficult to collect diverse visual images in various scenes with limited infrared cameras. To address this problem, we collect a large-scale dataset by combining almost all publicly available infrared datasets (Section 4). However, different infrared cameras might process diverse wavelength spectrums, leading to various infrared images. We design a progressive learning method to learn multi-stage infrared knowledge (Section 3.3). Since the V2IR translation is highly dependent on semantic information (object semantics and context information), we design a semantic-aware V2IR translation module via vision-

language understanding (Section 3.4).

3.1. Preliminary

Diffusion models are a family of probabilistic generative models that progressively destruct data by injecting noise, then learn to reverse this process for sample generation. DDPMs [17] are probabilistic generative models leveraging two Markov chains. The first Markov chain progressively injects noise into the data to transform data distribution into standard Gaussian distribution, while the other stepwise reverses the process of noise injection, generating data samples from Gaussian noise. LDMs [54] significantly reduce resource demand by operating in the latent space, especially dealing with high-resolution images. LDMs mainly consist of an autoencoder with an encoder \mathcal{E} and a decoder \mathcal{D} and a denoising U-Net ϵ_θ . Given an image x , LDMs first encode it into latent space and then add noise to the encoded latent $z = \mathcal{E}(x)$ producing a noisy latent z_t , where t denotes diffusing time step. For conditional diffusion models, condition c is introduced into the denoising process. The denoising U-Net ϵ_θ is trained by minimizing the following objective:

$$L = \mathbb{E}_{\mathcal{E}(x), c, \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2 \right], \quad (1)$$

where LDMs aim to predict the noise added on the encoded latent z at t timesteps under the condition c .

3.2. Overview of DiffV2IR

The pipeline of our DiffV2IR is illustrated in Figure 2. DiffV2IR mainly consists of two components, i.e., Progressive Learning Module (PLM) and Vision-Language Understanding Module (VLUM). Specifically, 1) as for PLM, we first establish foundational knowledge of infrared imaging properties utilizing our collected IR-500K dataset. Then we use visible-infrared image pairs to learn cross-modal transformation and finally conduct the refinement on the specific infrared imaging style. 2) as for VLUM, we incorporate unified vision-language understanding, including detailed language descriptions and segmentation maps, to make DiffV2IR semantic-aware and structure-preserving.

3.3. Visible-to-Infrared Diffusion Model via Progressive Learning

We employ the IR-500K Dataset, as described in Section 4, to train a conditional diffusion model aimed at converting visible images into infrared ones. Typically, fine-tuning diffusion models from a pre-trained checkpoint yields better results than training a diffusion model from scratch. Consequently, our model is constructed upon Stable Diffusion (SD), a pre-trained latent diffusion model with text conditioning [54], to leverage its extensive expertise in the domain of text-to-image translation.

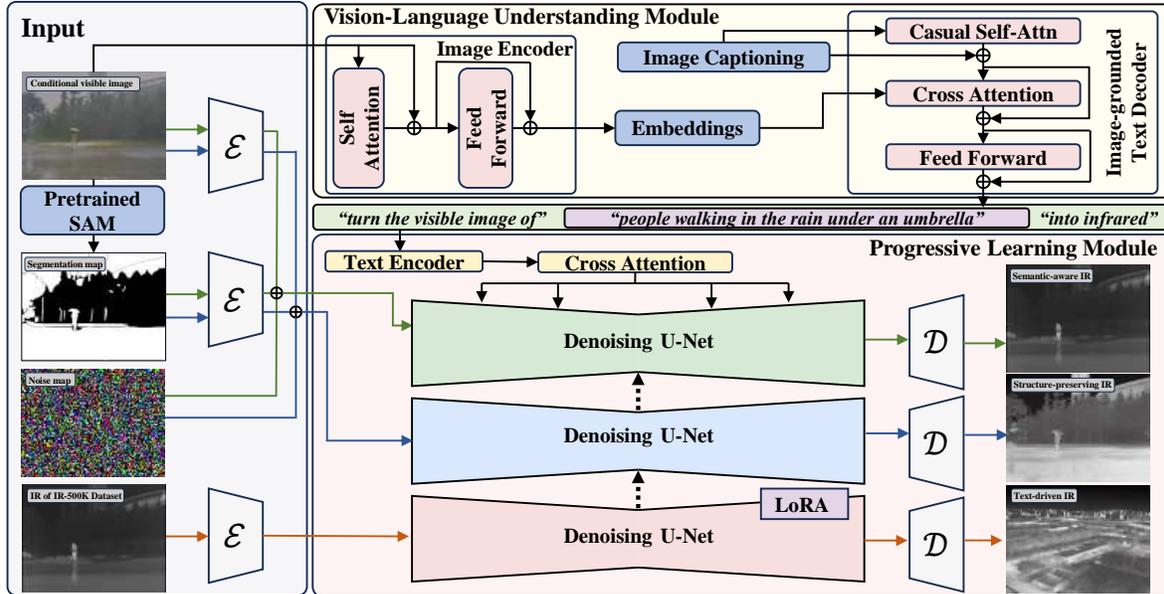


Figure 2. Framework overview of our DiffV2IR. DiffV2IR mainly consists of two components, i.e., Progressive Learning Module (PLM) and Vision-Language Understanding Module (VLUM). We use PLM for multi-stage knowledge learning and VLUM for semantic preserving in the V2IR task. The three U-Nets from bottom to top respectively denote the infrared representation internalization phase, the cross-modal transformation learning phase, and stylization refinement phase of PLM. The VLUM is introduced during PLM to make DiffV2IR semantic-aware.

Although pre-trained diffusion models are capable of generating high-quality visible images based on textual prompts, we observed that their performance degrades when the prompt includes the term “infrared.” This indicates a generally poor comprehension of infrared modality for most pre-trained diffusion models, with even less proficiency in translating visible images to infrared ones. To address this, we introduce PLM, a progressive learning strategy. This approach firstly enables the diffusion model to fill the gap between infrared modality and visual modality, then develops its capability of performing general visible-to-infrared image translation, and finally allows it to generate infrared images in a specified style.

Phase #1: infrared representation internalization, which aims to establish foundational knowledge of infrared imaging properties. This is the initial phase in our progressive learning strategy designed to integrate infrared knowledge. We achieve this by fine-tuning a stable diffusion model using Low-Rank Adaptation (LoRA) [19] on IR-500K dataset, all prompted with the same phrase, “an infrared image”. Throughout this tuning process, the weights of the pre-trained model remain fixed, while smaller trainable rank decomposition matrices are inserted into the model, enhancing training efficiency and minimizing overfitting risks. As a result of this progressive learning phase, the diffusion model associates infrared characteristics with the term “infrared”, enabling the generation of infrared-

style images from textual prompts, without losing other vital information.

Phase #2: cross-modal transformation learning, which aims to map visible-to-infrared (V2IR) modality differences through paired supervision. Then the dataset consisting of about 70,000 visible-infrared image pairs is utilized for diffusion model learning the mapping relationship between visible and infrared images. Subsequent to this stage, the model can generate images well consistent with the characteristics of infrared modality under the guidance of corresponding visible images. As the style of infrared images is relevant to many factors such as wavelength range and infrared camera sensors, the infrared images in our collected dataset have a high diversity. This diversity strengthens the generalization capability of our model when facing all kinds of visible images, which enables the model to serve as a pre-trained model for visible-to-infrared image translation.

Phase #3: stylization refinement, which aims to adapt infrared outputs to spatio-temporal variations and environmental dynamics. Although the pre-trained model is now capable of translating visible image into high-quality infrared one, the diversity of the training dataset makes it hard to generate infrared images in a specific style. To compensate for this shortcoming, we introduce the last training phase of our proposed progressive learning using a small dataset containing image pairs of visible images and in-

frared images in the desired style.

The diffusion model gradually advances through three progressive stages of enhanced learning, ultimately evolving into a model proficient in style-controllable transformation from visible images to infrared images.

3.4. Semantic-aware V2IR Translation via Vision-Language Understanding

V2IR is highly dependent on semantic information. Infrared imaging captures and visualizes the infrared radiation emitted by objects on the basis of their temperature and radiant existence. Therefore, V2IR relies on scene understanding such as object semantics and context information (e.g., solar radiation, lighting, and shadow). We integrate VLUM into a unified optimization framework and thus achieve semantic-aware V2IR translation. What’s more, we also incorporate additional embeddings of the segmentation map for better structure preserving.

To enhance the content awareness of the translation process, we use Blip [36] to create vision-language embeddings derived from visible images. This vision-language model provides a detailed description of key objects that determine the existence of radiants, along with contextual information such as weather, lighting, and other environmental elements that influence temperature. We employ a similar method to SD for text conditioning, utilizing a CLIP-based text encoder [51] that takes text as input and applies a cross-attention mechanism to incorporate the encoded tokens. Thanks to the robust text-image alignment capability of the pre-trained stable diffusion model and the internalization of infrared representation via the progressive learning module, vision-language capabilities enable DiffV2IR to comprehend the correspondences and distinctions in cross-modality images more effectively. In addition, to maintain structural integrity in the translation process, we add embeddings from the segmentation map generated by SAM [31], which have a rich knowledge of layout and structure. We merge the conditioning from visible images and the segmentation map by concatenating them with the noise map after latent encoding and by adding extra input channels to the first convolutional layer of the denoising U-Net. The weights of the newly introduced input channels are initialized using zero initialization.

Moreover, the Classifier-free Guidance mechanism [16] is utilized to enhance the controllability of generated images using conditional inputs. This technique is often used in conditional image generation to achieve a balance between sample quality and diversity. In our approach, the score network incorporates three types of conditioning: a visible image c_V , a segmentation map c_S , and a vision-language c_T . During training, certain conditionings are randomly set to none to allow unconditional training, with 2% of examples varying from fully unconditioned to having only one con-

ditioning. To balance the control strength of the three conditionings, we introduce the following guide scales: s_V for the visible image, s_S for the segmentation map, and s_T for vision-language. The score estimation is formulated as in Eq. 2. Each conditioning is assigned a guidance scale to adjust its intensity, resulting in a score estimate that combines conditional and unconditional outputs with specific weights.

$$\begin{aligned} \tilde{\epsilon}_\theta(z_t, c_V, c_S, c_T) &= \epsilon_\theta(z_t, \emptyset, \emptyset, \emptyset) \\ &+ s_V \cdot (\epsilon_\theta(z_t, c_V, \emptyset, \emptyset) - \epsilon_\theta(z_t, \emptyset, \emptyset, \emptyset)) \\ &+ s_S \cdot (\epsilon_\theta(z_t, c_V, c_S, \emptyset) - \epsilon_\theta(z_t, c_V, \emptyset, \emptyset)) \\ &+ s_T \cdot (\epsilon_\theta(z_t, c_V, c_S, c_T) - \epsilon_\theta(z_t, c_V, c_S, \emptyset)) \end{aligned} \quad (2)$$

4. IR-500K Dataset

To achieve superior quality in visible-to-infrared image translation, this research brings together nearly all large publicly available datasets [10, 13, 22, 25, 38, 41, 50, 52, 58, 60], supplemented by some additional data sourced online. The result is an extensive multi-wavelength database comprising 500,000 infrared images. These images represent a wide range of scene types, diverse object categories, and various camera perspectives, such as natural landscapes, cityscapes, driving environments, aerial scene understanding, and surveillance contexts. Each image is captured at a high resolution, showcasing rich visual details, and is carefully chosen to meet the model’s learning needs. These images serve as essential components in the understanding of infrared imaging, accompanied by the label text description “An infrared image”.

Additionally, we selected 70,000 visible and infrared image pairs specifically for precise training purposes, as these pairs are crucial for accurately capturing the differences between the two spectra. These paired datasets undergo strict alignment and segmentation processing to ensure that multi-modal information can be co-learned effectively, enhancing the cross-spectral translation performance. By integrating multiple source datasets, we not only expand the training data volume but also optimize data diversity and relevance for the diffusion model, providing abundant and representative learning materials. With these high-quality large-scale datasets, we believe that diffusion models can effectively capture cross-spectral visual feature relationships, enabling efficient and accurate visible-to-infrared image translation tasks. The processing details of creating the IR-500K dataset will be available at <https://github.com/LidongWang-26/DiffV2IR>.

5. Experiments

5.1. Experimental Settings

Testing Dataset. All of our experimental evaluations are performed using the M³FD dataset [38] and FLIR-aligned dataset [71], both of which offer a rich array of scenes characterized by diverse weather and lighting conditions. The M³FD dataset includes a total of 4,200 precisely aligned infrared and visible image pairs, each with dimensions of 1024 × 768, spread across over 10 different sub-scenarios. It is important to note that images from the same scene tend to be similar, which poses a risk of data leakage if these images are randomly split as training and testing data. To address this, we opt for a manual split of the M³FD dataset, creating a training set composed of 3,550 pairs and a testing set containing 650 pairs, rather than relying on a random division. The differences can be seen as shown in Table 1. The FLIR-aligned dataset, derived from the original FLIR dataset [10], is captured from a driving perspective and has been meticulously aligned [71]. It consists of 5,142 image pairs. However, because of differences in the receptive fields between visible and infrared images, some visible images exhibit noticeable black borders. To improve the quality of the dataset, we refined to select 4,489 image pairs and divided them into training (80%) and testing (20%) subsets. Each image maintains a resolution of 640 × 512.

Table 1. Data leak affects model performance. Experiments conducted on the identical M³FD dataset but with varying segmentations exhibit significant discrepancies, suggesting that randomly dividing the M³FD dataset might lead to data leakage.

Dataset	FID↓	PSNR↑	SSID↑
M ³ FD (randomly split)	45.89	21.70	0.7196
M ³ FD (manually split)	70.29	19.30	0.6620

Visual Quality Assessment. We assess the quality of translated images using established standard metrics, such as Fréchet Inception Distance (FID) [15], Peak Signal-to-Noise Ratio (PSNR) [18], and Structural Similarity Index (SSIM) [64].

Implementation Details. The experiments in this study were carried out on a system equipped with a NVIDIA A800 GPU. Throughout training, images were first resized to 286 x 286 and then randomly cropped to 256 x 256 to enhance training speed and efficiency. For models provided with a recommended configuration, the experiments were executed under those specified settings. Our DiffV2IR and models lacking recommended configurations were trained for approximately 100 epochs to ensure proper convergence. For techniques requiring text input, the same text prompt from phase #3 of PLM training in our DiffV2IR was supplied. For style transfer methods that require both

a style and content image, we randomly chose an infrared image from the training set as the style reference and used a visible image as the content to be translated.

5.2. Comparison with SOTA Methods

We evaluate our DiffV2IR model compared to fifteen cutting-edge methods that have emerged in recent years, many of which necessitate additional training before deployment. These methods include GAN-based methods such as Pix2Pix [24], CycleGAN [74], EGGAN-U [33, 40], DR-AVIT [14], StegoGAN [65], UNSB [26] and methods based on diffusion models like InstructPix2Pix [4], ControlNet [73], FCDiffusion [11], T2I-Adapter [46], Pix2PixTurbo [49], PID [43], among others. In addition, we also examine pre-trained models (CSGO [68]), training-free approaches (StyleID [8]), and few-shot methods (OS-ASIS [5]) for comparative analysis.

Quantitative Comparisons. Table 2 provides information on the performance of V2IR translations.

On the M³FD dataset, Pix2Pix [24] stands out among GAN-based approaches with the highest PSNR and SSIM, signifying excellent pixel-level precision, though it also records one of the worst FID scores. CycleGAN [74], DR-AVIT [14], and UNSB [26] offer a more balanced performance, yet their results remain unsatisfactory. Although EGGAN-U [33, 40] achieves the second-best SSIM among GAN methods, its overall performance along with StegoGAN [65] is not commendable. Regarding diffusion models, ControlNet [73] and T2I-Adapter [46] achieve impressive PSNR and SSIM scores by transforming visible images into feature maps for conditional guidance in the denoising process, with T2I-Adapter [46] reaching a low FID of 114.63. On the other hand, FCDiffusion [11] struggles with input visible image structure preservation, showing the worst SSIM. Methods like OSASIS [5], training-free approaches such as StyleID [8], and the pre-trained CSGO model [68] fail to deliver good metrics due to insufficient training data and limited infrared understanding. PID [43], another diffusion model for infrared generation, presents reasonable PSNR and SSIM but suffers from a high FID. In contrast, InstructPix2Pix [4] and Pix2PixTurbo [49] excel with the second and third best metrics across all methods.

In contrast to the M³FD dataset, the FLIR-aligned dataset presents a less complex scenario. GAN-based methods demonstrate improved outcomes over their performance on the M³FD dataset, with Pix2Pix [24] achieving the highest PSNR among all methods. Nevertheless, the FID score for Pix2Pix [24] remains significantly elevated. Similar to their results on the M³FD dataset, CycleGAN [74] and DR-AVIT [14] maintain balance across three metrics, whereas EGGAN-U [33, 40], StegoGAN [65], and UNSB [26] continue to exhibit subpar performance. As for diffusion-based models, ControlNet [73], T2I-Adapter [46], and FCDiffu-

Table 2. Quantitative comparison with the state-of-the-arts. The best results are highlighted in **bold** and the second best results are underlined. The methods in the first half of the table are GAN-based, while the latter half are based on diffusion models.

Method	Reference	M ³ FD dataset			FLIR-aligned dataset		
		FID↓	PSNR↑	SSIM↑	FID↓	PSNR↑	SSIM↑
Pix2Pix [24]	CVPR ₁₇	182.14	17.19	0.5672	98.81	19.79	0.4327
CycleGAN [74]	ICCV ₁₇	114.71	14.98	0.5271	59.74	16.58	0.4091
EGGAN-U[33, 40]	ICRA ₂₃	149.12	13.87	0.5455	113.51	15.76	0.4253
DR-AVIT[14]	TGRS ₂₄	116.96	14.19	0.5449	65.96	16.30	0.4355
StegoGAN[65]	CVPR ₂₄	183.56	13.19	0.4303	87.57	12.44	0.3752
UNSB[26]	ICLR ₂₄	115.94	14.07	0.4885	85.61	9.95	0.3179
ControlNet[73]	ICCV ₂₃	140.14	15.17	0.5572	119.69	11.98	0.2783
FCDiffusion[11]	AAAI ₂₄	170.14	11.60	0.2854	180.58	10.89	0.2860
T2I-Adapter[46]	AAAI ₂₄	114.63	15.98	0.5976	91.61	12.33	0.3689
OSASIS[5]	CVPR ₂₄	243.21	14.59	0.5642	192.44	14.51	0.3774
StyleID[8]	CVPR ₂₄	135.97	12.67	0.4317	94.28	10.69	0.3086
CSGO[68]	ARXIV ₂₄	185.32	10.33	0.4147	178.04	9.63	0.3288
Pix2PixTurbo[49]	ARXIV ₂₄	98.12	16.80	0.5964	90.72	15.92	<u>0.4590</u>
PID[43]	ARXIV ₂₄	160.91	16.10	0.5579	<u>43.98</u>	<u>18.89</u>	0.4315
InstructPix2Pix[4]	CVPR ₂₃	<u>81.64</u>	<u>17.92</u>	<u>0.6328</u>	46.29	18.41	0.4481
DiffV2IR	Ours	70.29	19.30	0.6620	39.99	18.63	0.4658

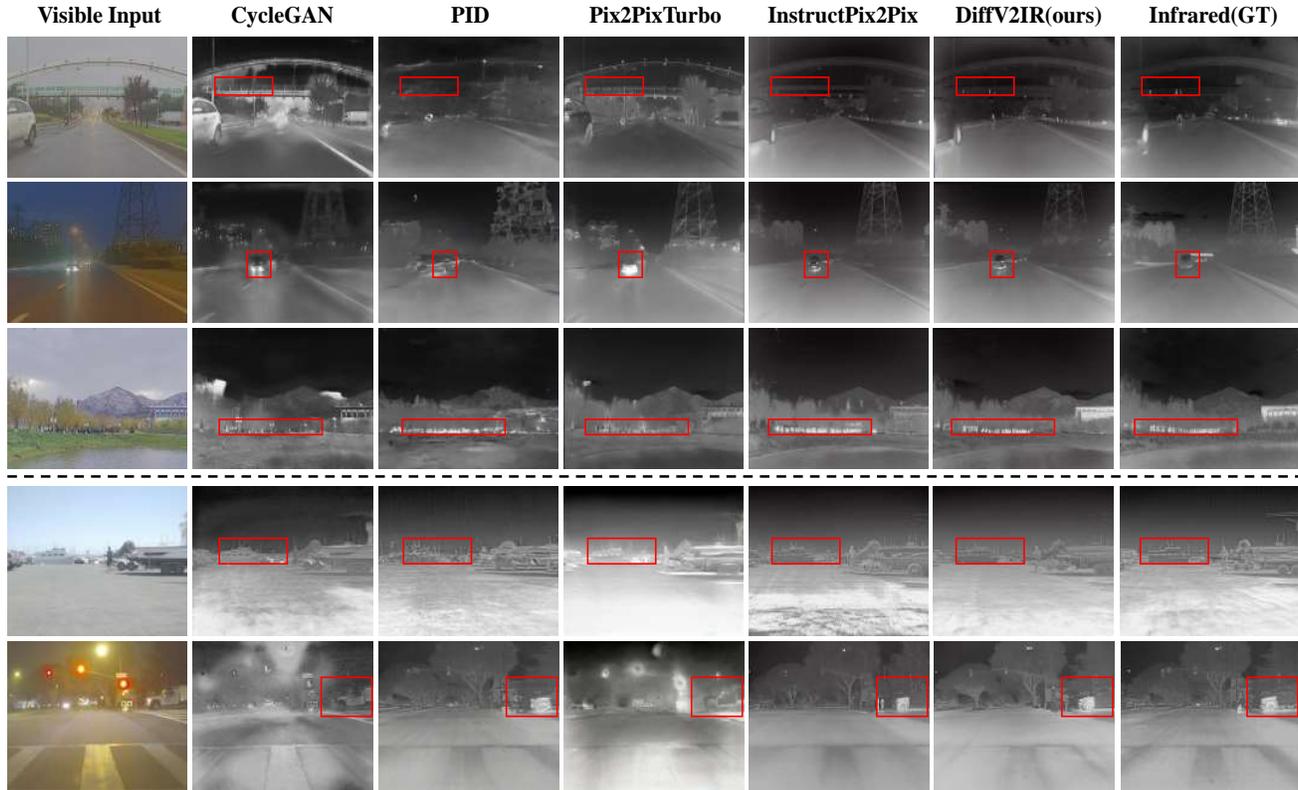


Figure 3. Comparison with SOTA methods. Key differences are highlighted within a red box, such as halos and low-light scenarios. Only the top 5 methods according to assessment metrics are shown. (Top: results from M³FD dataset; Bottom: results from FLIR-aligned dataset.)

sion [11] still struggle to adjust effectively to the V2IR task. Methods lacking additional training also show weak performance. Pix2PixTurbo [49] ranks second best in SSIM, although its FID remains high. PID [43] performs notably well, achieving the second-best FID and PSNR. Instruct-Pix2Pix [4] delivers a well-rounded performance with balanced metrics.

Based on experimental outcomes and previous observations, models utilizing GANs tend to excel with simpler datasets as opposed to more intricate scenes. Furthermore, several techniques are significantly affected by mode collapse, necessitating multiple training rounds to guarantee high-quality generation. InstructPix2Pix [4], Pix2PixTurbo [49], and our proposed DiffV2IR avoid converting the visible image into a noise or feature map, which we believe contributes to their superior performance compared to other methods.

Qualitative Comparisons. As illustrated in Figure 3, the proposed DiffV2IR offers significant improvements in both global image quality and control over local details. Current methods face several key issues, the most critical being incorrect handling of halos around light sources, primarily due to a lack of understanding of the infrared modality. In visible images, halo regions often exhibit high brightness, whereas in infrared images, only objects emitting heat should appear brighter. Another issue arises when processing visible images in low-light conditions, where some methods fail to maintain the image structure and details effectively. In contrast, DiffV2IR leverages enhanced infrared knowledge and vision-language understanding to produce credible images that adhere to physical principles while providing outstanding detail management.

The experiments demonstrate the effectiveness of diffusion models in multi-spectral translation tasks and validate the robustness of our proposed DiffV2IR under different conditions.

5.3. Ablation Study

We perform an ablation study on the M³FD dataset. Table 3 highlights the impact of our progressive learning and vision-language understanding modules.

Progressive Learning Module. We conduct experiments using all possible combinations of the three phases, relying exclusively on visible-infrared image pairs without any additional embeddings. Since the original Stable Diffusion functions as a text-to-image framework and models fine-tuned only in phase #1 are unable to perform image translation, we excluded that case. Each training phase contributes positively to the results.

Vision-Language Understanding Module. After establishing the progressive learning approach, we initially incorporated two distinct embeddings for vision-language and segmentation maps independently and subsequently com-

bined them. Although each additional individual embedding enhances the quality of generation, the unified strategy, known as the DiffV2IR method, achieves the highest level of performance.

Hyper-parameters. A limited number of hyperparameters can influence the ultimate performance. Primarily, we experiment with denoising steps and classifier-free guidance scales across three conditional inputs. Table 4 presents the results. The ultimate outcome is a balance among several factors, such as inference consumption and translation quality according to these three evaluation metrics.

Table 3. The ablation study is split into two distinct stages. The initial rows focus on evaluating the training phases of progressive learning excluding VLUM, whereas the later rows assess the types of information in VLUM when combined with PLM.

PLM			VLUM		FID↓	PSNR↑	SSIM↑
#1	#2	#3	Seg-Map	Vision-Language			
×	✓	×	-	-	113.98	12.71	0.4013
✓	✓	×	-	-	112.45	13.50	0.4055
×	×	✓	-	-	81.15	18.47	0.6439
✓	×	✓	-	-	78.10	18.71	0.6481
×	✓	✓	-	-	75.48	19.01	0.6533
✓	✓	✓	-	-	74.79	19.13	0.6557
✓	✓	✓	×	✓	73.92	19.11	0.6563
✓	✓	✓	✓	×	71.63	19.17	0.6585
✓	✓	✓	✓	✓	70.29	19.30	0.6620

Table 4. Ablation study of denoising steps and classifier-free guidance scales. s_T , s_V , s_S denote guidance scale for vision-language, visible image and segmentation mask, respectively. The optimal settings we choose are highlighted in bold.

Hyper-parameters		FID↓	PSNR↑	SSIM↑
steps	50	70.92	19.37	0.6640
	100	70.29	19.30	0.6620
	150	71.41	19.38	0.6616
	200	70.69	19.33	0.6614
s_T	5.0	71.34	19.43	0.6656
	7.5	70.29	19.30	0.6620
	10.0	72.04	19.34	0.6592
s_V	1.0	71.65	19.34	0.6663
	1.5	70.29	19.30	0.6620
	2.0	73.02	19.24	0.6580
s_S	1.0	72.09	19.32	0.6630
	1.5	70.29	19.30	0.6620
	2.0	72.45	19.37	0.6631

6. Conclusion

Converting visible images into infrared images is highly demanded and is not adequately addressed. The primary challenges are generating content with semantic awareness, differences in spectrum appearances, and the scarcity of effective infrared datasets. This study introduces DiffV2IR, an innovative framework for translating visible images into infrared images. By integrating PLM, VLUM and the comprehensive IR-500K dataset, we significantly enhance the V2IR translation performance. Experimental findings confirm the effectiveness of diffusion models in generating superior translations, demonstrating their efficacy and wide-ranging applicability, thereby offering a fresh approach for multi-spectral image generation.

Limitation. DiffV2IR is specialized in translating broad scenes, which may restrict its effectiveness in specific applications such as face image translation.

Potential Negative Impact. DiffV2IR focuses on infrared image synthesis. It might be misused to create misleading content.

References

- [1] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image-to-image translation. *NeurIPS*, 31, 2018. 2
- [2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223, 2017. 2
- [3] Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Hyunjung Shim. Rethinking the truly unsupervised image-to-image translation. In *ICCV*, pages 14154–14163, 2021. 2
- [4] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 3, 6, 7, 8
- [5] Hansam Cho, Jonghyun Lee, Seunggyu Chang, and Yonghyun Jeong. One-shot structure-aware stylized image synthesis. In *CVPR*, 2024. 3, 6, 7
- [6] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, pages 8789–8797, 2018. 2
- [7] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, pages 8188–8197, 2020. 2
- [8] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *CVPR*, pages 8795–8805, 2024. 3, 6, 7, 1
- [9] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 3
- [10] TELEDYNE FLIR. Free teledyne flir thermal dataset for algorithm training. [Online]. <https://www.flir.com/oem/adas/adas-dataset-form/>. 5, 6, 2
- [11] Xiang Gao, Zhengbo Xu, Junhan Zhao, and Jiaying Liu. Frequency-controlled diffusion model for versatile text-guided image-to-image translation. In *AAAI*, pages 1824–1832, 2024. 3, 6, 7, 8
- [12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2
- [13] Zonghao Han, Ziyue Zhang, Shun Zhang, Ge Zhang, and Shaohui Mei. Aerial visible-to-infrared image translation: Dataset, evaluation, and baseline. *Journal of remote sensing*, 3:0096, 2023. 5, 2
- [14] Zonghao Han, Shun Zhang, Yuru Su, Xiaoning Chen, and Shaohui Mei. DR-AVIT: Towards diverse and realistic aerial visible-to-infrared image translation. *IEEE TGRS*, 2024. 2, 3, 6, 7
- [15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 6
- [16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5
- [17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2, 3
- [18] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *ICPR*, pages 2366–2369. IEEE, 2010. 6
- [19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 4
- [20] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, pages 172–189, 2018. 2
- [21] HyeonJoo Hwang, Geon-Hyeong Kim, Seunghoon Hong, and Kee-Eung Kim. Variational interaction information maximization for cross-domain disentanglement. *NeurIPS*, 33: 22479–22491, 2020. 2
- [22] Soonmin Hwang, Jaesik Park, Namil Kim, Yookyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *CVPR*, pages 1037–1045, 2015. 5, 2
- [23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 2
- [24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2, 6, 7, 1
- [25] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. LLVIP: A visible-infrared paired dataset for low-light vision. In *ICCV*, pages 3496–3504, 2021. 5, 2
- [26] Beomsu Kim, Gihyun Kwon, Kwanyoung Kim, and Jong Chul Ye. Unpaired image-to-image translation via neural schrödinger bridge. In *ICLR*, 2024. 6, 7
- [27] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwang Hee Lee. U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *ICLR*, 2020. 2

- [28] Soohyun Kim, Jongbeom Baek, Jihye Park, Gyeongnyeon Kim, and Seungryong Kim. Instaformer: Instance-aware image-to-image translation with transformer. In *CVPR*, pages 18321–18331, 2022. 2
- [29] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, pages 1857–1865. Pmlr, 2017. 2
- [30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 2
- [31] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *ICCV*, pages 4015–4026, 2023. 5
- [32] Vladimir V Kniaz, Vladimir A Knyaz, Jiri Hladuvka, Walter G Kropatsch, and Vladimir Mizginov. 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*, pages 0–0, 2018. 3
- [33] Dong-Guw Lee, Myung-Hwan Jeon, Younggun Cho, and Ayoung Kim. Edge-guided multi-domain rgb-to-thermal image translation for training vision tasks with challenging labels. In *ICRA*, 2023. 2, 3, 6, 7
- [34] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, pages 35–51, 2018. 2
- [35] Junsung Lee, Minsoo Kang, and Bohyung Han. Diffusion-based image-to-image translation by noise correction via prompt interpolation. In *ECCV*, 2024. 3
- [36] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 5
- [37] Jianxin Lin, Yingce Xia, Tao Qin, Zhibo Chen, and Tie-Yan Liu. Conditional image-to-image translation. In *CVPR*, 2018. 2
- [38] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *CVPR*, 2022. 5, 6, 1, 2
- [39] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. *NeurIPS*, 29, 2016. 2
- [40] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *NeurIPS*, 30, 2017. 2, 3, 6, 7
- [41] Qiao Liu, Xin Li, Zhenyu He, Chenglong Li, Jun Li, Zikun Zhou, Di Yuan, Jing Li, Kai Yang, Nana Fan, et al. Lsotb-thermal: A large-scale high-diversity thermal infrared object tracking benchmark. In *ACM MM*, pages 3847–3856, 2020. 5
- [42] Decao Ma, Shaopeng Li, Juan Su, Yong Xian, and Tao Zhang. Visible-to-infrared image translation for matching tasks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024. 2
- [43] Fangyuan Mao, Jilin Mei, Shun Lu, Fuyang Liu, Liang Chen, Fangzhou Zhao, and Yu Hu. Pid: Physics-informed diffusion model for infrared image generation. *arXiv preprint arXiv:2407.09299*, 2024. 2, 3, 6, 7, 8, 1
- [44] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, pages 2794–2802, 2017. 2
- [45] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2
- [46] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, pages 4296–4304, 2024. 3, 6, 7
- [47] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pages 8162–8171. PMLR, 2021. 2
- [48] Mehmet Akif Özkanoglu and Sedat Ozer. Infragan: A gan architecture to transfer visible images to infrared domain. *Pattern Recognition Letters*, 155:69–76, 2022. 2, 3
- [49] Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*, 2024. 6, 7, 8
- [50] Zhang Pengyu, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *CVPR*, 2022. 5, 2
- [51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. Pmlr, 2021. 5
- [52] RAYTRON. Raytron infrared open source platform. [Online]. <http://openai.raytrontek.com/>. 5
- [53] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, pages 1278–1286. PMLR, 2014. 2
- [54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3
- [55] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265. pmlr, 2015. 2
- [56] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2
- [57] Qiyang Sun, Xia Wang, Changda Yan, and Xin Zhang. Vq-infratrans: A unified framework for rgb-ir translation with hybrid transformer. *Remote Sensing*, 15(24):5661, 2023. 3
- [58] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2022. 5, 2
- [59] Hao Tang, Hong Liu, Dan Xu, Philip HS Torr, and Nicu Sebe. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks.

IEEE transactions on neural networks and learning systems, 34(4):1972–1987, 2021. 2

- [60] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 2022. 5, 2
- [61] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, pages 1921–1930, 2023. 3
- [62] Ben Usman, Dina Bashkirova, and Kate Saenko. Rift: Disentangled unsupervised image translation via restricted information flow. In *WACV*, pages 2420–2429, 2023. 2
- [63] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 2
- [64] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 6
- [65] Sidi Wu, Yizi Chen, Samuel Mermet, Lorenz Hurni, Konrad Schindler, Nicolas Gonthier, and Loic Landrieu. Stegogan: Leveraging steganography for non-bijective image-to-image translation. In *CVPR*, 2024. 2, 6, 7
- [66] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, Radu Timotfe, and Luc Van Gool. Diffi2i: efficient diffusion model for image-to-image translation. *IEEE TPAMI*, 2024. 3
- [67] Shaoan Xie, Mingming Gong, Yanwu Xu, and Kun Zhang. Unaligned image-to-image translation by learning to reweight. In *ICCV*, pages 14174–14184, 2021. 2
- [68] Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. CSGO: Content-style composition in text-to-image generation. *arXiv preprint arXiv:2408.16766*, 2024. 6, 7, 1
- [69] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4): 1–39, 2023. 2
- [70] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, pages 2849–2857, 2017. 2
- [71] Heng Zhang, Elisa Fromont, Sébastien Lefevre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International conference on image processing (ICIP)*, pages 276–280. IEEE, 2020. 6, 1
- [72] Lichao Zhang, Abel Gonzalez-Garcia, Joost Van De Weijer, Martin Danelljan, and Fahad Shahbaz Khan. Synthetic data generation for end-to-end thermal infrared tracking. *IEEE TIP*, 28(4):1837–1850, 2018. 3
- [73] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 3, 6, 7
- [74] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2, 6, 7

DiffV2IR: Visible-to-Infrared Diffusion Model via Vision-Language Understanding

Supplementary Material

To provide more details of our proposed DiffV2IR, this supplementary material includes the following content:

- Section A: Common infrared image datasets that have been widely used in recent years.
- Section B: Intermediate results for a vivid understanding of our proposed modules.
- Section C: Infrared images generated by all the SOTA methods mentioned in our paper.
- Section D: A demo video to better illustrate our motivation and solving approach.

A. Comparison of Our Proposed IR-500K and Other Commonly Used Infrared Datasets

Table I provides a comparison of our proposed IR-500K and other commonly used infrared datasets. Though the scale is not the biggest, IR-500K has the best diversity with various scenarios, multiple camera angles, and all kinds of object. The scenarios of IR-500K include urban scenes, such as campus, parks, roads, infrastructures, and natural scenes such as rivers, lakes, beaches, seas, mountains. The camera angles consist of aerial view by drones, surveillance view by monitors, driving view by in-vehicle cameras, horizontal view by handheld cameras, and so on. The main objects comprise human beings, vehicles, wild animals, natural landscapes, and buildings.

B. Intermediate Results

Figure I presents intermediate outcomes of PLM and VLUM, highlighting their effectiveness. Although original stable diffusion (v 1.5) excels at generating images from text owing to extensive training data, it struggles with the infrared modality due to limited infrared knowledge. Hence, the generated output is far from a normal infrared image. In Phase #1 of PLM, the model internalizes the infrared representation, allowing it to produce high-quality infrared images, although without precise control. From Figure I column 2, the output of PLM Phase #1 shows good appearance as an infrared image. During Phase #2, the model starts to convert visible images into infrared ones with a guidance visible image and a segmentation map, maintaining structural integrity, though stylization remains challenging due to the tight link between semantics and infrared imagery. Finally, with the refinement of stylization in phase #3 and the incorporation of VLUM, the model achieves both structure preservation and semantic awareness. Our final outputs of DiffV2IR can have both good texture matching and struc-

tural preservation.

C. More Visualization Results

We present the visual results of infrared images generated by 15 state-of-the-art (SOTA) methods involved in the experimental comparison section.

Figure II and Figure III show the visualization results of all the methods mentioned in our experimental part, on M³FD [38] and FLIR-aligned [71] dataset. There are several main problems. The first one is that some models like Pix2Pix [24] and PID [43] exhibit ambiguity, which causes Image quality degradation. The second one is the failure of preserving structure of visible inputs, such as StyleID [8] and CSGO [68]. The third as well as the most common problem is the overlook of semantics and context information.

D. Demo Video

To effectively convey the research’s objectives, challenges, and innovative aspects, we have created a video that outlines three principal challenges in infrared image translation, describes the technical methodology for converting visible to infrared images employed in this study, and highlights preliminary implementation results.

Please refer to the demo video attached in “supplementary.zip”.

Table I. Comparison of our proposed IR-500K and other commonly used visible-infrared datasets.

Dataset	Camera Angle	Scenario	Amount
MSRS [60]	Driving	Road	1,444 pairs
AVIID [13]	Aerial	Road	3,363 pairs
M ³ FD [38]	Horizontal	Campus, Road, Natural scenes	4,200 pairs
FLIR [10]	Driving	Road	9,711 IR/9,233 visible (not aligned)
LLVIP [25]	Surveillance	Street	15,488 pairs
DroneVehicle [58]	Aerial	Road, Urban area	28,439 pairs
Kaist [22]	Driving	Road	95,000 pairs
VTUAV [50]	Aerial	Urban scenes	1.7M pairs from 500 sequences
IR-500K(ours)	Multiple	Multiple	500K IR, 70,000 pairs

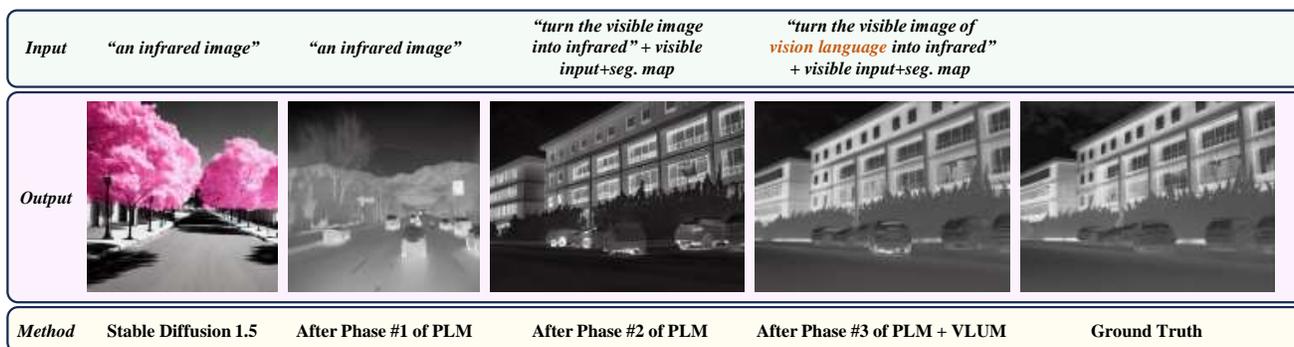


Figure I. Intermediate results of our proposed DiffV2IR.

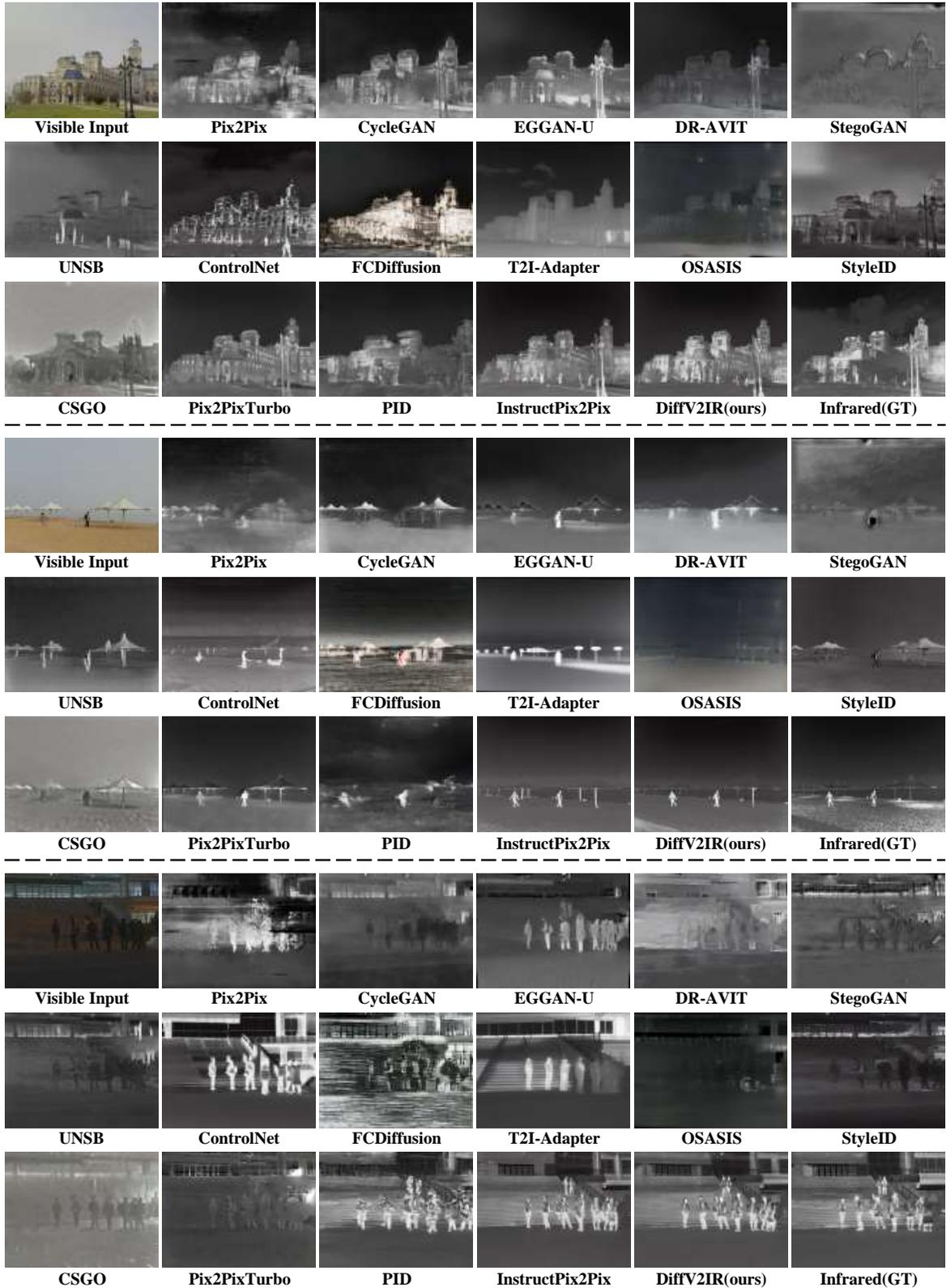


Figure II. Comparison with SOTA methods on M³FD dataset.



Figure III. Comparison with SOTA methods on FLIR-aligned dataset.