

CM-Diff: A Single Generative Network for Bidirectional Cross-Modality Translation Diffusion Model Between Infrared and Visible Images

Bin Hu¹ Chenqiang Gao^{1,2*} Shurui Liu² Fang Chen³
Junjie Guo¹ Fangcen Liu¹ Junwei Han^{1,4}

¹Chongqing University of Posts and Telecommunications ²Sun Yat-sen University

³University of California, Merced ⁴Northwestern Polytechnical University

{haydengauthiero, liuufc67, gjj893866738, liushr6688}@gmail.com, gaochq6@mail.sysu.edu.cn, fchen20@ucmerced.edu, jhan@nwpu.edu.cn

Abstract

Image translation is one of the crucial approaches for mitigating information deficiencies in the infrared and visible modalities, while also facilitating the enhancement of modality-specific datasets. However, existing methods for infrared and visible image translation either achieve unidirectional modality translation or rely on cycle consistency for bidirectional modality translation, which may result in suboptimal performance. In this work, we present the bidirectional cross-modality translation diffusion model (CM-Diff) for simultaneously modeling data distributions in both the infrared and visible modalities. We address this challenge by combining translation direction labels for guidance during training with cross-modality feature control. Specifically, we view the establishment of the mapping relationship between the two modalities as the process of learning data distributions and understanding modality differences, achieved through a novel Bidirectional Diffusion Training (BDT). Additionally, we propose a Statistical Constraint Inference (SCI) to ensure the generated image closely adheres to the data distribution of the target modality. Experimental results demonstrate the superiority of our CM-Diff over state-of-the-art methods, highlighting its potential for generating dual-modality datasets.

Introduction

Infrared (IR) and visible (VIS) imaging are two common modalities in visual perception within the computer vision community. Infrared images are captured on the basis of the thermal radiation emitted by objects, which makes them invariant to illumination conditions and robust against environmental disturbances such as low light and smoke (Liu et al. 2024). Nevertheless, they lack color, texture, and detailed information, which limits their ability to analyze fine-grained objects. On the other hand, visible images are formed by the reflection of light from objects, providing rich color, texture, and detailed information, enabling precise object representation and fine-grained analysis. This disparity in the characteristics of infrared and visible imaging creates a complementary relationship between the two modalities, making their integration or alternation crucial for robust real-world applications. Given the complementary strengths and limitations of visible and infrared imaging, dual-modality methods often outperform single-modality methods, particularly in tasks like object detection

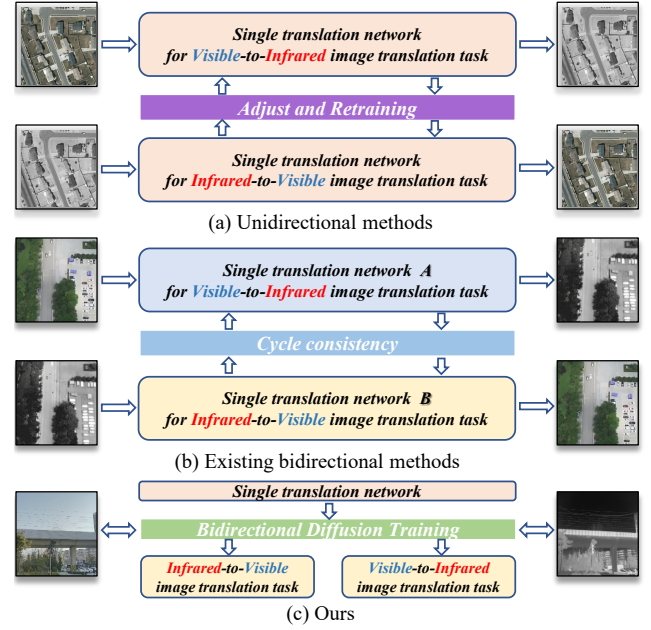


Figure 1: Comparison of infrared-visible image translation methods: (a) Unidirectional methods require network modifications and retraining to enable translation in the opposite direction; (b) Bidirectional methods employ two independent generative networks and utilize cycle consistency for simultaneous training; (c) CM-Diff introduces Bidirectional Diffusion Training (BDT), eliminating the need for separate networks and reducing the computational overhead and re-training time associated with unidirectional methods.

(Guo et al. 2024; Fu et al. 2023; Li et al. 2023) and semantic segmentation (Liu et al. 2023b; Ji et al. 2023; Deng et al. 2021). However, these methods usually rely on a large, well-aligned dual-modality dataset, while in practice, acquiring such registered dual-modality datasets is challenging and costly. Thus, a natural idea is to leverage image translation techniques to generate dual-modality datasets.

Recent infrared-visible image translation methods have achieved promising results using Generative Adversarial Networks (GANs) (Liu et al. 2018; Wang et al. 2024; Han et al. 2024a; Luo et al. 2022; Di et al. 2019; Yadav, Singh,

and Dubey 2023; Han et al. 2024b) and Denoising Diffusion Probabilistic Models (DDPMs) (Ding et al. 2025; Sasaki, Willcocks, and Breckon 2021; Nair and Patel 2023; Lin et al. 2024; Oliveira et al. 2024), which can be broadly classified into two categories: bidirectional image translation methods and unidirectional image translation methods. Bidirectional methods (Zhu et al. 2017; Liu, Breuel, and Kautz 2017; Han et al. 2021; Luo et al. 2022; Sasaki, Willcocks, and Breckon 2021) typically employ two independent generative networks to establish mappings between the infrared and visible modalities, often relying on a cycle consistency method, as shown in Fig. 1 (b). Although the cycle consistency based methods have proven to be effective in ensuring modality correspondence, it not only increases the complexity of the network architecture but also frequently leads to perceptual blurriness and loss of fine details in the generated outputs (Zhu et al. 2017; Maqsood, Abid et al.). Unidirectional methods (Zhang et al. 2018b; Di et al. 2019; Isola et al. 2017; Yadav, Singh, and Dubey 2023; Ma et al. 2024; Boroujeni and Razi 2024; Han et al. 2024b) leverage a single generative network to model the mapping in one direction, either infrared-to-visible image translation or vice versa. Consequently, they require separate training processes for each translation direction to achieve bidirectional modality translation, as shown in Fig. 1 (a).

In this paper, we propose a unified framework called CM-Diff. The core idea is to enable joint learning of infrared and visible image data using a Bidirectional Diffusion Training (BDT), as shown in Fig. 1 (c). This method incorporates implicit prompts through input channel positions, combined with explicit directional label embedding. This method is further refined by a multi-level Cross-modality Feature Control (CFC), which enhances semantic control over the translation process. Consequently, it eliminates reliance on the cycle-consistency method while enabling more precise mapping between the two data modalities. Furthermore, we observe that after implementing bidirectional conversion, the issue of color distortion becomes critically severe when directly applying the original ancestral sampling method (Koller and Friedman 2009; Yang et al. 2023) for infrared-visible image translation. To address this, we propose a Statistical Constraint Inference (SCI), which leverages statistical mean constraints tailored to different datasets and scenarios. By incorporating these statistical properties, we mitigate abnormalities caused by the inherent randomness in the reverse Markov chain of DDPM, ensuring that the generated images maintain high fidelity and align closely with real-world distributions. The main contributions of this paper can be summarized as follows:

- **A unified CM-Diff framework for bidirectional IR→VIS translation.** We propose CM-Diff with a Bidirectional Diffusion Training (BDT) that encodes direction labels and source-domain features into a single U-Net, enabling IR→VIS and VIS→IR mappings without dual networks or additional consistency losses.
- **Statistical Constraint Inference (SCI) for distributional alignment.** By integrating domain-specific statistics (e.g., thermal variance) into the reverse diffusion process,

SCI corrects sampling drift and aligns the generated image distribution with real data, effectively reducing anomalies and color distortion.

- **Extensive experimental validation.** Comprehensive experiments on standard benchmarks demonstrate that our method outperforms existing approaches in both quantitative metrics and qualitative visual consistency, addressing modality absence and misalignment issues.

Related Work

Cross-modality translation between infrared and visible images is a challenging task due to the significant modality gap in texture, color, and semantics. Existing approaches can be broadly categorized into GAN-based, DDPM-based, and more recent bidirectional generative modeling techniques.

GAN-based Cross-Modality Image Translation

Generative Adversarial Networks have been widely used for IR-VIS translation tasks, often leveraging pixel-level supervision or perceptual loss to bridge the modality gap. For IR→VIS, TVA-GAN (Yadav, Singh, and Dubey 2023) incorporates attention and inception-based features to improve local structural detail, while I2V-GAN (Li et al. 2021) adds cyclic and perceptual constraints to achieve temporal coherence in video translation. ROMA (Yu et al. 2022) and EADS (Ren et al. 2023) enhance semantic alignment by integrating edge or similarity priors. In contrast, VIS→IR translation focuses on thermal realism. For example, VTF-GAN (Ordun, Raff, and Purushotham 2023) introduces Fourier losses to recover frequency-aware details, and IC-GAN (Boroujeni and Razi 2024) employs color-space mapping to reduce appearance distortion in remote sensing.

While GANs produce sharp outputs, they are prone to mode collapse and training instability. More importantly, most methods are unidirectional, requiring separate training for each direction. Bidirectional translation using GANs typically relies on two models with cycle consistency (e.g., CycleGAN variants), which adds training overhead and may degrade translation quality.

DDPM-based Cross-Modality Image Translation

Diffusion models offer an alternative that avoids GAN-specific instabilities by gradually denoising from noise via a Markov chain. T2V-DDPM (Nair and Patel 2023) pioneers their use in IR-to-VIS face translation under low-light conditions, while UNIT-DDPM (Sasaki, Willcocks, and Breckon 2021) jointly models both modalities through dual-domain Markov chains. However, these methods remain unidirectional and require retraining to switch translation direction.

Some diffusion-based editing methods support partial reversibility. SDEdit (Meng et al. 2021), for instance, enables image synthesis through partial denoising but assumes conditional masks and is primarily used in unpaired settings. EGSDE (Zhao et al. 2022) introduces energy-guided sampling, yet it remains unpaired and task-specific. These techniques are not directly comparable due to assumptions not holding in our paired IR-VIS domain.

Bidirectional Diffusion and Bridge Models

To eliminate the need for dual generators or cycle loss, several works propose bidirectional learning within a single model. Bidirectional Consistency Models (BCM) (Li and He 2024) introduce a consistency loss to regularize one diffusion model for both directions, but they operate in unpaired domains. In contrast, our setting assumes paired supervision, which allows stronger data-driven training signals.

Similarly, Cycle-Free CycleGAN (Kwon and Ye 2021) avoids cycle loss by using an invertible generator in medical imaging, which still differs architecturally and assumes invertibility—an assumption hard to satisfy in IR-VIS tasks due to non-invertible thermal content. Schrödinger Bridge methods such as PSB (Liu et al. 2023a) and CBD (He et al. 2024) formulate bidirectional generation as stochastic transport, relying on bridging distributions. These methods are theoretically elegant but often require iterative sampling or multi-step alignment and are more suited to unsupervised or unpaired tasks.

Unlike these approaches, our method is the first to enable bidirectional IR $\leftrightarrow$ VIS generation using a single shared diffusion model. The direction of translation is encoded via the timestep embedding, without requiring auxiliary supervision or additional inference passes.

Method

Given a set of paired infrared images $\mathcal{X}^{\text{IR}} = \{\mathbf{x}_1^{\text{IR}}, \dots, \mathbf{x}_N^{\text{IR}} \mid \mathbf{x} \in \mathbb{R}^{H \times W \times C_{\text{IR}}}\}$ and visible images $\mathcal{X}^{\text{VIS}} = \{\mathbf{x}_1^{\text{VIS}}, \dots, \mathbf{x}_N^{\text{VIS}} \mid \mathbf{x} \in \mathbb{R}^{H \times W \times C_{\text{VIS}}}\}$, our goal is to train a model $\mathcal{Z}_N$ capable of translating infrared images $\mathcal{X}^{\text{IR}}$ into their corresponding visible counterparts $\hat{\mathcal{X}}^{\text{VIS}}$, while also being able to perform the reverse translation from visible images $\mathcal{X}^{\text{VIS}}$ to infrared images $\hat{\mathcal{X}}^{\text{IR}}$. To accomplish this task, we propose a bidirectional end-to-end framework, CM-Diff.

Bidirectional Diffusion Training (BDT)

Paired infrared (IR) and visible (VIS) images contain similar semantic information, but their data distributions differ significantly due to the distinct imaging characteristics of each modality. To simultaneously learn these two data distributions and establish a bidirectional mapping relationship, we propose a Bidirectional Diffusion Training (BDT). Our strategy uses a U-Net architecture network, as shown in Fig. 2 (a), to simultaneously learn and distinguish the visible modality data $\mathbf{x}_0^{\text{VIS}} \in \mathcal{X}^{\text{VIS}}$ and infrared modality data $\mathbf{x}_0^{\text{IR}} \in \mathcal{X}^{\text{IR}}$, and institutes an effective bidirectional mapping to ensure that the content of the source modality image corresponds effectively to the translated target modality image, thereby achieving a stable training process.

(1) Translation Direction Guidance (TDG). We first employ an implicit input channel position encoding to determine whether the data distribution being learned is from the infrared modality or the visible modality. Specifically, for the infrared-to-visible translation part, the input consists of three components: the noisy data $q(\mathbf{x}_t^{\text{VIS}} \mid \mathbf{x}_0^{\text{VIS}}) \in \mathbb{R}^{H \times W \times C_{\text{VIS}}}$ in the visible modality at time step t , the infrared modality image data $\mathbf{x}_0^{\text{IR}} \in \mathbb{R}^{H \times W \times C_{\text{IR}}}$, and the edge

feature maps $E^{\text{IR}} = \text{DexiNed}(\mathbf{x}^{\text{IR}}) \in \mathbb{R}^{H \times W \times E}$ extracted from the infrared image using the edge detection network DexiNed (Soria et al. 2023). For the consistency of subsequent operations, we duplicate the single-channel infrared image across three channels to match the visible format, ensuring $C_{\text{VIS}} = C_{\text{IR}} = 3$. Using edge feature maps extracted from infrared and visible images as part of the input prompt can ensure the consistency of the target image (Ren et al. 2023). To construct the network input, we concatenate these data along the channel dimension, resulting in the following input representation:

$$\mathcal{Z}_{\text{IR-to-VIS}}^t = q(\mathbf{x}_t^{\text{VIS}} \mid \mathbf{x}_0^{\text{VIS}}) \oplus \mathbf{x}_0^{\text{IR}} \oplus E^{\text{IR}}, \quad (1)$$

where $\mathcal{Z}_{\text{IR-to-VIS}}^t \in \mathbb{R}^{H \times W \times (C_{\text{VIS}} + C_{\text{IR}} + E)}$, $\oplus$ denotes the concatenation operation, with the target modality data, source modality data, and edge features corresponding to channel index ranges $\{1, 2, \dots, C_{\text{VIS}}\}$, $\{C_{\text{VIS}} + 1, C_{\text{VIS}} + 2, \dots, C_{\text{IR}} + C_{\text{VIS}}\}$, and $\{C_{\text{IR}} + C_{\text{VIS}} + 1, C_{\text{IR}} + C_{\text{VIS}} + 2, \dots, C_{\text{IR}} + C_{\text{VIS}} + E\}$, respectively. This implicit channel encoding effectively distinguishes the data from different modalities without requiring explicit direction labels indicating the data source. Additionally, we introduce label embedding layers τ_ϕ to process the translation direction labels, which represent: C^{IR} corresponding to the image translation direction from visible to infrared, learning the data distribution of the infrared modality.

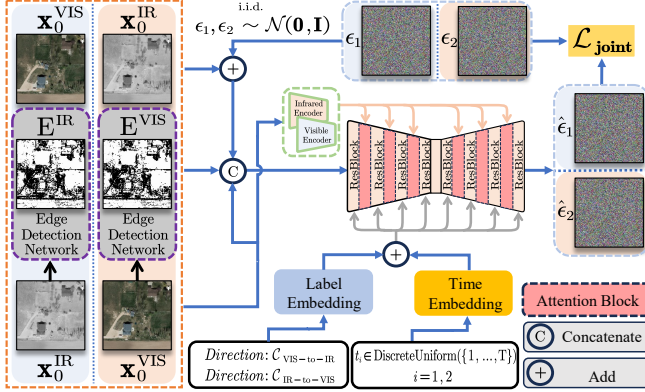
(2) Cross-modality Feature Control (CFC). To achieve effective target-modality image content control based on the source-modality image, in addition to the direct input of edge feature maps, we also train two independent U-Net encoders, one serving as the infrared modality encoder $\mathcal{Z}_E^{\text{IR}}$ and the other as the visible modality encoder $\mathcal{Z}_E^{\text{VIS}}$, to extract modality features F_d from the infrared or visible images. These modality features are then fed into the cross-attention layer of $\mathcal{Z}_N$ to enhance the control of the target modality image content. The query Q_n^i , key K_n^i , and value V_n^i are computed as follows:

$$Q_n^i = \text{Conv}_q^i(F_{g_i}^n), \quad K_n^i, V_n^i = \text{Conv}_{kv}^i(F_{d_i}^n), \quad (2)$$

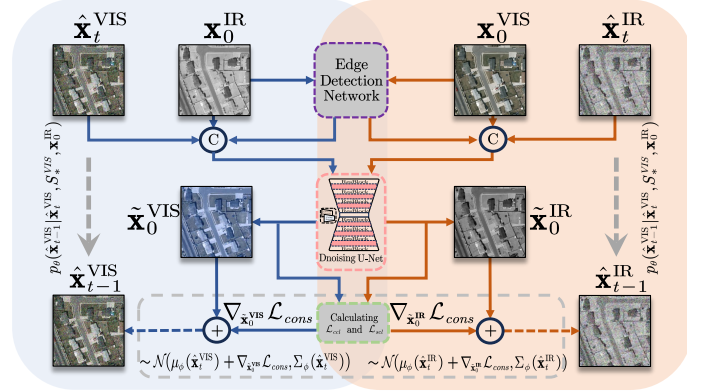
where $F_{g_i}^n$ is the input vector obtained by flattening the noise feature map of the i -th cross-attention part in the n -th layer of network $\mathcal{Z}_N$, and $F_{d_i}^n$ is the feature map extracted from the i -th part in the n -th layer of the modality encoder. Through the cross-modality attention mechanism, we compute the attention weights A^i and obtain the updated feature output $F_{g_{i+1}}^n$:

$$F_{g_{i+1}}^n = F_{g_i}^n + \text{softmax}\left(\frac{Q_n^i \cdot (K_n^i)^T}{\sqrt{d}}\right) \cdot V_n^i. \quad (3)$$

(3) Modality Joint Learning Loss Function. Once the model effectively learns to capture the characteristics of both data distributions, we compute the noise prediction loss in each translation direction. Specifically, when the infrared image provides guidance, the mean squared error (MSE) between the predicted noise in the noisy visible image and the true noise is calculated, yielding $\mathcal{L}_{\text{IR-to-VIS}}$. Conversely, when the visible image provides guidance, the corresponding MSE is computed for the infrared image, resulting



(a) Bidirectional Diffusion Training (BDT)



(b) Statistical Constraint Inference (SCI)

Figure 2: Overview of CM-Diff. (a) Bidirectional Diffusion Training (BDT) : Learning domain-specific data distributions and establishing robust bidirectional mappings. (b) Statistical Constraint Inference (SCI) : Ensuring consistency between the statistical distributions of translated images and those in the training data.

in $\mathcal{L}_{\text{VIS-to-IR}}$. The loss functions are defined as:

$$\begin{aligned}\mathcal{L}_{\text{IR-to-VIS}}(\phi) &= \mathbb{E}_{t_1, \mathbf{x}_0^{\text{IR}}, \epsilon_1} \left[\left\| \epsilon_1 - \epsilon_\phi(\mathcal{Z}_{\text{IR-to-VIS}}^{t_1}, J^{\text{VIS}}, t_1) \right\|_2^2 \right] \\ \mathcal{L}_{\text{VIS-to-IR}}(\phi) &= \mathbb{E}_{t_2, \mathbf{x}_0^{\text{VIS}}, \epsilon_2} \left[\left\| \epsilon_2 - \epsilon_\phi(\mathcal{Z}_{\text{VIS-to-IR}}^{t_2}, J^{\text{IR}}, t_2) \right\|_2^2 \right],\end{aligned}\quad (4)$$

where $t_1, t_2 \stackrel{\text{i.i.d.}}{\sim} \text{DiscreteUniform}(\{1, \dots, T\})$, $\epsilon_1, \epsilon_2 \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I})$, $J = \tau_\phi(\mathcal{C})$. The modality joint learning loss $\mathcal{L}_{\text{joint}}$ is as follows:

$$\mathcal{L}_{\text{joint}}(\phi) = \lambda_{\text{IR-to-VIS}} \cdot \mathcal{L}_{\text{IR-to-VIS}} + \lambda_{\text{VIS-to-IR}} \cdot \mathcal{L}_{\text{VIS-to-IR}}. \quad (5)$$

Algorithm 1: CM-Diff Training

Input: visible (VIS) image $\mathbf{x}^{\text{VIS}}$, infrared (IR) image $\mathbf{x}^{\text{IR}}$, infrared-to-visible direction label $\mathcal{C}^{\text{IR}}$, visible-to-infrared direction label $\mathcal{C}^{\text{VIS}}$.

- 1: **repeat**
- 2: $\mathbf{x}^{\text{VIS}} \in \mathcal{X}^{\text{VIS}}, \mathbf{x}^{\text{IR}} \in \mathcal{X}^{\text{IR}}$
- 3: $t_1, t_2 \stackrel{\text{i.i.d.}}{\sim} \text{DiscreteUniform}(\{1, \dots, T\})$
- 4: $\epsilon_1, \epsilon_2 \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I})$
- 5: $\mathbf{x}_{t_1}^{\text{VIS}} \leftarrow \sqrt{\alpha_{t_1}} \cdot \mathbf{x}_0^{\text{VIS}} + \sqrt{1 - \alpha_{t_1}} \cdot \epsilon_1$
 $\mathbf{x}_{t_2}^{\text{IR}} \leftarrow \sqrt{\alpha_{t_2}} \cdot \mathbf{x}_0^{\text{IR}} + \sqrt{1 - \alpha_{t_2}} \cdot \epsilon_2$
- 6: $J^{\text{VIS}} \leftarrow \tau_\phi(\mathcal{C}^{\text{VIS}}), J^{\text{IR}} \leftarrow \tau_\phi(\mathcal{C}^{\text{IR}})$
- 7: $E^{\text{VIS}} \leftarrow \text{DexiNed}(\mathbf{x}_0^{\text{VIS}}), E^{\text{IR}} \leftarrow \text{DexiNed}(\mathbf{x}_0^{\text{IR}})$
- 8: **Gradient descent step on:**

$$\nabla_\phi \left[\lambda_{\text{IR-to-VIS}} \cdot \left\| \epsilon_1 - \epsilon_\phi(\mathcal{Z}_{\text{IR-to-VIS}}^{t_1}, J^{\text{VIS}}, t_1) \right\|_2^2 + \lambda_{\text{VIS-to-IR}} \cdot \left\| \epsilon_2 - \epsilon_\phi(\mathcal{Z}_{\text{VIS-to-IR}}^{t_2}, J^{\text{IR}}, t_2) \right\|_2^2 \right]$$
- 9: **until** Converged

Statistical Constraint Inference (SCI)

Our statistical constraint inference (SCI) module aims to guide the IR $\rightarrow$ VIS generation toward semantically and thermally consistent outputs. Related ideas include Diffusion

Posterior Sampling (DPS) (Chung et al. 2022), which constrains diffusion inference via approximate posteriors. However, SCI differs by using domain-specific statistics (e.g., thermal variance) instead of learned masks or latent priors, making it more suitable for infrared-visible translation where semantic cues are inherently weak.

In order to enhance bidirectional diffusion inference with cross-modality statistical awareness, we propose Statistical Constraint Inference (SCI), which constrains the denoising trajectory using prior statistical properties derived from paired training data. The method consists of three components: (1) baseline inference, (2) constrained denoising using statistical priors, and (3) loss-guided gradient-based sampling.

(1) Baseline Inference Formulation. To adapt the denoising diffusion process to our Bidirectional Diffusion Training (BDT) framework, we reformulate the reverse transitions in each direction using conditional generative functions. Specifically, the denoising step is defined as:

$$\begin{aligned}\mathbf{x}_{t-1}^{\text{VIS}} &= \mu_\phi(\mathcal{Z}_{\text{IR-to-VIS}}^{t_1}, J^{\text{VIS}}, t_1) + \Sigma_\phi(\mathcal{Z}_{\text{IR-to-VIS}}^{t_1}, J^{\text{VIS}}, t_1) \epsilon_1, \\ \mathbf{x}_{t-1}^{\text{IR}} &= \mu_\phi(\mathcal{Z}_{\text{VIS-to-IR}}^{t_2}, J^{\text{IR}}, t_2) + \Sigma_\phi(\mathcal{Z}_{\text{VIS-to-IR}}^{t_2}, J^{\text{IR}}, t_2) \epsilon_2.\end{aligned}\quad (6)$$

where $\mathcal{Z}^t$ denotes the latent embedding extracted from the conditional encoder at time step t along each direction, and J represents modality-specific guidance such as edge or semantic priors. The functions $\mu_\phi(\cdot)$ and $\Sigma_\phi(\cdot)$ produce the mean and variance of the denoising distribution at step t .

This formulation allows both visible and infrared modalities to be reconstructed from Gaussian noise by iteratively incorporating the complementary information provided by the opposite modality. Unlike unidirectional DDPM, our bidirectional structure ensures mutual enhancement and robust translation under strong structural constraints.

(2) Statistical-Constrained Reverse Sampling. To enhance semantic fidelity and statistical realism in generated

images, we incorporate domain-specific statistical priors into the reverse denoising process by modifying the standard unconditional transition distribution $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$ into a constrained form $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, w)$, where w encodes statistical properties of the target domain, such as edge distributions or pixel-level histograms.

Following the framework of classifier-free guidance (Dhariwal and Nichol 2021), the conditional distribution can be approximated as:

$$\begin{aligned} \log p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, w) &= \log(p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)p(w | \mathbf{x}_t)) + z_1 \\ &\approx \log p(h) + z_2, \quad h \sim \mathcal{N}(\mu_\theta + g\Sigma_\theta, \Sigma_\theta), \end{aligned} \quad (7)$$

where $g = \nabla_{\mathbf{x}_t} \log p(w | \mathbf{x}_t)$, z_1 and z_2 are constants.

However, directly estimating $p(w | \mathbf{x}_t)$ is unreliable because $\mathbf{x}_t$ is corrupted by time-dependent noise. Therefore, inspired by (Ho, Jain, and Abbeel 2020), we introduce $\tilde{\mathbf{x}}_0$ —an intermediate prediction of the clean image:

$$\tilde{\mathbf{x}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} \cdot (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon_\theta(\mathbf{x}_t, t)). \quad (8)$$

This prediction captures more reliable semantic content than $\mathbf{x}_t$, and is used as a proxy for evaluating $p(w | \tilde{\mathbf{x}}_0)$.

To construct a practical approximation of $p(w | \tilde{\mathbf{x}}_0)$, we define:

$$p(w | \tilde{\mathbf{x}}_0) = \frac{1}{\mathcal{M}} \exp \{ -[\lambda_Q \cdot \mathcal{Q}(\tilde{\mathbf{x}}_0, w) + \lambda_K \cdot \mathcal{K}(\tilde{\mathbf{x}}_0)] \}, \quad (9)$$

where $\mathcal{Q}(\cdot)$ measures the distance between the statistics of $\tilde{\mathbf{x}}_0$ and the target domain (e.g., intensity histograms or edge maps), and $\mathcal{K}(\cdot)$ is a regularization term that penalizes visual artifacts such as blur or noise. The scale parameters λ_Q and λ_K control the relative influence of each constraint.

Accordingly, the gradient g becomes:

$$\nabla_{\tilde{\mathbf{x}}_0} \log p(w | \tilde{\mathbf{x}}_0) = -\lambda_Q \cdot \nabla_{\tilde{\mathbf{x}}_0} \mathcal{Q}(\tilde{\mathbf{x}}_0, w) - \lambda_K \cdot \nabla_{\tilde{\mathbf{x}}_0} \mathcal{K}(\tilde{\mathbf{x}}_0). \quad (10)$$

By injecting the guidance term $\Sigma_\theta \cdot \nabla_{\tilde{\mathbf{x}}_0} \log p(w | \tilde{\mathbf{x}}_0)$ into the reverse process, we steer the denoising trajectory toward outputs that respect both structural fidelity and global statistical properties, while maintaining the stochastic integrity of the original diffusion model.

Finally, the updated prediction of $\mathbf{x}_{t-1}$ is derived from the posterior distribution:

$$\begin{aligned} q(\mathbf{x}_{t-1} | \mathbf{x}_t, \tilde{\mathbf{x}}_0) &= \mathcal{N}(\mathbf{x}_{t-1}; \mu_q(\mathbf{x}_t, \tilde{\mathbf{x}}_0), \Sigma_q(t)), \\ \mu_q(\mathbf{x}_t, \tilde{\mathbf{x}}_0) &= \frac{(1 - \alpha_t) \cdot (1 - \bar{\alpha}_{t-1}) \cdot \mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}} \cdot \beta_t \cdot \tilde{\mathbf{x}}_0}{1 - \bar{\alpha}_t}, \\ \Sigma_q(t) &= \frac{(1 - \alpha_t) \cdot (1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \cdot \mathbf{I}. \end{aligned} \quad (11)$$

The overall process of the SCI is presented in Fig. 2 (b).

(3) Constraint Loss Function for Statistical Regulation. To implement the conditional term $p(w | \tilde{\mathbf{x}}_0)$ in our SCI framework, we define two differentiable losses—*channel constraint loss* $\mathcal{L}_{ccl}$ and *statistical constraint loss* $\mathcal{L}_{scl}$ —which correspond to $\mathcal{K}(\cdot)$ and $\mathcal{Q}(\cdot)$, respectively. These are combined into the overall constraint objective:

$$\mathcal{L}_{cons} = \lambda_{ccl} \cdot \mathcal{L}_{ccl} + \lambda_{scl} \cdot \mathcal{L}_{scl}. \quad (12)$$

Algorithm 2: CM-Diff Inference (Visible-to-Infrared)

Input: Visible image $\mathbf{x}_0^{\text{vis}}$, infrared-to-visible direction label $\mathcal{C}^{\text{IR}}$.

Output: Translated IR modality image $\hat{\mathbf{x}}_0^{\text{IR}}$

```

1: Initialize a pre-trained edge detection network DexiNed
2:  $\mathbf{x}_0^{\text{vis}} \in \mathcal{X}^{\text{vis}}$ ,  $\hat{\mathbf{x}}_T^{\text{IR}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
3:  $E^{\text{vis}} \leftarrow \text{DexiNed}(\mathbf{x}_0^{\text{vis}})$ 
4:  $J^{\text{IR}} \leftarrow \tau_\phi(\mathcal{C}^{\text{IR}})$ 
5: for  $t = T, \dots, 1$  do
6:    $\mu, \Sigma = \mu_\phi(\hat{\mathbf{x}}_t^{\text{IR}}), \Sigma_\phi(\hat{\mathbf{x}}_t^{\text{IR}})$ 
7:    $\tilde{\mathbf{x}}_0^{\text{IR}} = \frac{\hat{\mathbf{x}}_t^{\text{IR}}}{\sqrt{\bar{\alpha}_t}} - \frac{\sqrt{1 - \bar{\alpha}_t} \cdot \epsilon_\phi(\hat{\mathbf{x}}_t^{\text{IR}}, \mathbf{x}_0^{\text{vis}}, E^{\text{vis}}, J^{\text{IR}}, t)}{\sqrt{\bar{\alpha}_t}}$ 
8:    $\mathcal{L}_{cons} = \lambda_{scl} \cdot \mathcal{L}_{scl}(\tilde{\mathbf{x}}_0^{\text{IR}}, w) + \lambda_{ccl} \cdot \mathcal{L}_{ccl}(\tilde{\mathbf{x}}_0^{\text{IR}})$ 
9:   Sample  $\hat{\mathbf{x}}_{t-1}^{\text{IR}}$  by  $\mathcal{N}(\mu + \nabla_{\tilde{\mathbf{x}}_0^{\text{IR}}} \mathcal{L}_{cons}, \Sigma)$ 
10: end for
11: return  $\hat{\mathbf{x}}_0^{\text{IR}}$ 
```

Channel Constraint Loss. We penalize discrepancies in per-channel histograms between the noisy estimate $\tilde{\mathbf{x}}_0$ and the empirical prior histograms from the target modality:

$$\mathcal{L}_{ccl} = \sum_{\delta \in \{R, G, B\}} \sum_{i=1}^B \frac{(h_i^{\text{pred}, \delta} - h_i^{\text{prior}, \delta})^2}{h_i^{\text{pred}, \delta} + h_i^{\text{prior}, \delta} + \epsilon}, \quad (13)$$

where $h_i^{\text{pred}, \delta}$ and $h_i^{\text{prior}, \delta}$ are the i -th histogram bin counts of the predicted and prior distributions for channel δ , B is the number of bins, and $\epsilon = 10^{-6}$ avoids division by zero.

Statistical Constraint Loss. To align the generated image’s first- and second-order channel statistics with the target domain, we define:

$$\mathcal{L}_{scl} = \sum_{\delta \in \{R, G, B\}} |\mu_\delta^{\text{pred}} - \mu_\delta^{\text{prior}}| + |\sigma_\delta^{\text{pred}} - \sigma_\delta^{\text{prior}}|, \quad (14)$$

where $\mu_\delta^{\text{pred}}, \sigma_\delta^{\text{pred}}$ are the mean and standard deviation of channel δ in $\tilde{\mathbf{x}}_0$, and $\mu_\delta^{\text{prior}}, \sigma_\delta^{\text{prior}}$ are the corresponding empirical values from the training data.

By backpropagating the gradients of $\mathcal{L}_{cons}$ into the reverse sampling step, we adjust the denoising mean via

$$\mu \leftarrow \mu + \Sigma \cdot \nabla_{\tilde{\mathbf{x}}_0} \mathcal{L}_{cons},$$

thereby enforcing both histogram consistency and statistical alignment without modifying the original diffusion model parameters.

The overall inference process using the SCI is presented in Algorithm 2.

Experiments

We conduct all our experiments on VEDAI (Razakarivony and Jurie 2016), AVIID (Han et al. 2023), and M³FD (Liu et al. 2022) datasets to evaluate the performance of image translation methods.

Comparison with State-of-the-Art Methods

(1) Quantitative Results. The quantitative comparison results are listed in Tab. 1. The evaluation metrics include

Table 1: Quantitative comparison of the CM-Diff with state-of-the-art methods. The best performance for each metric is highlighted in bold, and the second-best scores are marked with underline fonts for reference.

Method	AVIID (infrared $\rightarrow$ visible)				VEDAI (infrared $\rightarrow$ visible)				M ³ FD (infrared $\rightarrow$ visible)				AVIID (visible $\rightarrow$ infrared)				VEDAI (visible $\rightarrow$ infrared)				M ³ FD (visible $\rightarrow$ infrared)			
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$
T2V-DDPM (Nair and Patel 2023)	22.03	0.759	0.074	63.40	14.72	0.645	0.249	137.65	11.35	0.373	0.421	144.55	23.75	0.883	0.050	23.88	14.97	0.835	0.122	54.09	11.33	0.502	0.364	138.90
CycleGAN (Zhu et al. 2017)	19.58	0.521	0.185	45.34	12.29	0.671	0.223	144.46	15.02	0.519	0.382	122.77	18.58	0.452	0.272	87.13	19.03	0.740	0.204	106.65	14.42	0.396	0.378	113.75
UNIT (Liu, Breuel, and Kautz 2017)	23.02	0.702	0.140	59.69	12.51	0.107	0.414	176.83	13.96	0.421	0.396	136.60	<u>25.03</u>	0.744	0.138	50.20	13.86	0.194	0.405	175.24	16.86	0.577	0.349	117.75
MUNIT (Huang et al. 2018)	23.34	0.694	0.141	48.37	14.74	0.402	0.318	163.14	13.06	0.370	0.450	144.65	23.56	0.664	0.173	63.66	19.25	0.528	0.104	51.76	15.86	0.549	0.374	126.22
DCLGAN (Han et al. 2021)	18.08	0.492	0.291	80.46	14.73	0.502	0.297	176.15	13.62	0.364	0.431	119.28	17.31	0.424	0.333	93.57	18.11	0.493	0.282	196.44	12.99	0.440	0.474	191.5
I2V-GAN (Li et al. 2021)	15.23	0.403	0.378	134.97	14.31	0.518	0.321	202.39	12.47	0.391	0.447	179.09	15.98	0.443	0.398	147.49	19.33	0.664	0.230	143.94	11.31	0.382	0.445	143.67
ROMA (Yu et al. 2022)	19.10	0.531	0.173	33.47	12.88	0.128	0.629	250.61	17.37	0.464	0.293	83.46	19.83	0.541	0.193	40.36	<u>20.20</u>	0.613	0.174	83.17	19.00	0.606	0.251	75.96
CUT (Park et al. 2020)	14.14	0.338	0.460	122.71	16.87	0.535	0.272	142.45	15.21	0.401	0.371	104.66	18.78	0.486	0.270	76.74	15.84	0.561	0.359	228.34	9.33	0.231	0.631	314.04
Ours	23.67	0.835	0.054	22.80	19.89	0.774	0.185	94.11	<u>17.03</u>	0.554	0.184	66.19	27.96	<u>0.871</u>	<u>0.054</u>	21.33	24.03	0.883	0.096	46.18	23.54	0.765	0.116	57.26

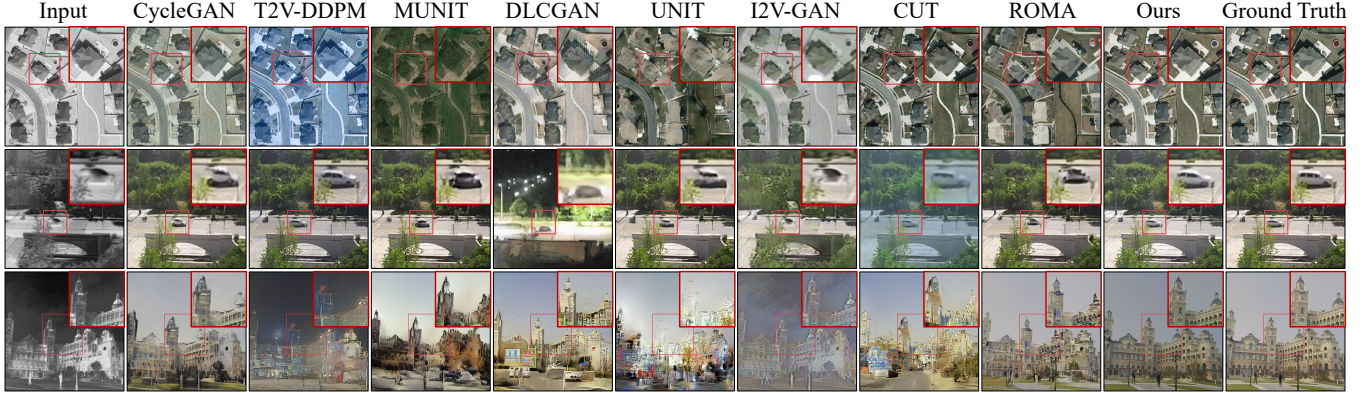


Figure 3: Qualitative results of CM-Diff for the infrared-to-visible image translation task. The first row corresponds to the VEDAI dataset, the second row to the AVIID dataset, and the third row to the M³FD dataset.

FID (Heusel et al. 2017), SSIM (Wang et al. 2004), LPIPS (Zhang et al. 2018a), and PSNR for both the infrared-to-visible image translation task and the visible-to-infrared image translation task. Overall, our method demonstrates superior performance compared to unidirectional and bidirectional modality translation methods.

Compared to the DDPM-based method T2V-DDPM (Nair and Patel 2023), the proposed method achieves a 7.44% improvement in PSNR, along with a 64.04% reduction in FID for the infrared-to-visible image translation task on the AVIID dataset. On the M³FD dataset, the enhancements are also pronounced, with a 53.04% increase in PSNR and a 56.29% reduction in LPIPS. This improvement stems from the limitations of diffusion-based unidirectional methods, which focus solely on learning the data distribution in the target modality and establishing mappings within it, thereby missing deeper semantic connections. In contrast, BDT can simultaneously learn the data distributions of both modalities. Additionally, the CFC is utilized to enhance the semantic guidance of the source modality image.

Compared to GAN-based unidirectional translation methods, our proposed method achieves improvements across the three datasets. On the AVIID dataset, for the infrared-to-visible image translation task, FID decreases by 72.20% and PSNR increases by 48.88% compared to CUT (Park et al. 2020), while FID decreases by 31.88% and PSNR increases by 23.93% compared to ROMA (Yu et al. 2022). This improvement can be attributed to the proposed DDPM-based method, which transforms the data distribution into a series of noise distributions to avoid the problem of mode collapse. Additionally, it employs SCI during the sampling phase

to incorporate observed statistical information, thereby improving the consistency between the generated results and the real data distribution.

Compared to GAN-based bidirectional translation methods, our proposed method also demonstrates performance gains. On the AVIID dataset, for the infrared-to-visible image translation task, PSNR increases by 1.41% compared to MUNIT (Huang et al. 2018). The observed improvement can be ascribed to the adoption of a BDT, which effectively encodes directional embeddings and edge information while leveraging multi-scale features extracted by the source modality encoder. This method mitigates modality distribution ambiguity arising from the structural and semantic similarities between infrared and visible images.

(2) Qualitative Results. We compare the infrared-to-visible (Fig. 3) and visible-to-infrared (Fig. 4) translation results of state-of-the-art methods with our approach. In the first row, using the VEDAI dataset, we observe that all methods transfer semantic information well, but our method outperforms DCLGAN (Han et al. 2021) and UNIT (Liu, Breuel, and Kautz 2017) in terms of color and detail. This is due to the limitations of DCLGAN and UNIT in accurately defining object boundaries and utilizing spatial information. In the second row, methods like CycleGAN (Zhu et al. 2017), MUNIT (Huang et al. 2018), and ROMA (Yu et al. 2022) produce blurred or distorted objects. While CUT (Park et al. 2020), T2V-DDPM (Nair and Patel 2023), and our method generate clearer object structures, our method stands out by improving color fidelity through inference guidance with the SCI. The third row shows challenging

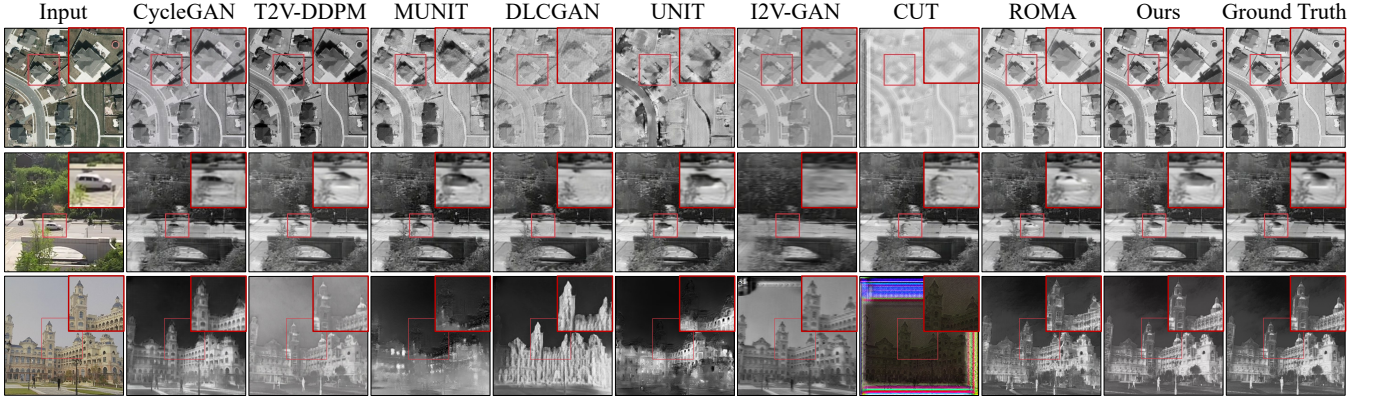


Figure 4: Qualitative results of CM-Diff for the visible-to-infrared image translation task. The first row corresponds to the VEDAI dataset, the second row to the AVIID dataset, and the third row to the M³FD dataset.

samples with complex backgrounds and pedestrians. Methods like T2V-DDPM and CUT tend to mistakenly render background buildings as sky. In contrast, our method accurately reconstructs architectural details and sky backgrounds, thanks to the CFC module that ensures precise feature generation and noise control.

Ablation Study

(1) Effect of CFC. We conducted ablation experiments to demonstrate the effectiveness of the source modality features and edge feature map cues in preserving content consistency in translation results. The quantitative results are presented in Tab. 2. It can be observed that the removal of the modality encoders and the edge detection part leads to a notable drop in the overall performance in both translation directions. We can see that external control using modality and edge features can effectively help the network distinguish boundaries of objects in the image.

Table 2: Effect of CFC.

CFC	infrared $\rightarrow$ visible				visible $\rightarrow$ infrared			
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$
w/o	19.17	0.761	0.201	101.13	24.05	0.880	0.098	47.84
w/	19.89	0.774	0.185	94.11	24.06	0.883	0.096	46.18

(2) Effect of TDG. In Tab. 3, we investigate the impact of TDG on the quality of translation results for the two directions. When direction labels are passed through the label embedding layer and used to guide the network, the PSNR and SSIM metrics for the infrared-to-visible image translation task improve by 7.16% and 6.91%, respectively. Similarly, for the visible-to-infrared translation task, these metrics increase by 5.29% and 14.52%, respectively. These results demonstrate that TDG is indispensable during training, as it helps the model recognize noisy data and provide crucial guidance. Moreover, during inference, the TDG corresponds to assisting the model in executing the correct translation direction, further highlighting its importance.

Table 3: Effect of TDG.

TDG	infrared $\rightarrow$ visible				visible $\rightarrow$ infrared			
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$
w/o	18.56	0.724	0.253	142.79	22.85	0.771	0.181	163.87
w/	19.89	0.774	0.185	94.11	24.06	0.883	0.096	46.18

(3) Effect of SCI. As listed in Tab. 4, in the infrared-to-visible image translation task, the $\mathcal{L}_{scl}$ plays a significant role in enhancing modality translation performance, resulting in an improvement of 5.85% in PSNR and 1.57% in SSIM compared to the case without SCI. In the visible-to-infrared image translation task, it results in an improvement of 52.70% in PSNR, 5.87% in SSIM, 26.15% in LPIPS, and 19.97% in FID. Moreover, in the visible-to-infrared image translation task, incorporating the $\mathcal{L}_{ccl}$ further boosts the model’s performance, leading to an improvement of 0.16% in PSNR and 0.21% in FID.

Table 4: Effect of SCI.

$\mathcal{L}_{ccl}$	$\mathcal{L}_{scl}$	infrared $\rightarrow$ visible				visible $\rightarrow$ infrared			
		PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$
×	×	18.78	0.762	0.205	105.02	15.73	0.834	0.130	57.83
×	✓	19.88	0.774	0.185	93.65	24.02	0.883	0.096	46.28
✓	×	19.54	0.775	0.201	97.39	15.64	0.840	0.129	56.88
✓	✓	19.89	0.774	0.185	94.11	24.06	0.883	0.096	46.18

Conclusion

This paper introduces CM-Diff, a unified framework that utilizes DDPM for bidirectional translation between infrared and visible images. Our approach leverages the diffusion model and incorporates a Bidirectional Diffusion Training (BDT) to distinguish and learn modality-specific data distributions while establishing robust bidirectional mappings. We have also proposed a Statistical Constraint Inference (SCI) for maintaining consistency between the statistical distributions of the translated images and training images. Extensive experiments have shown the effectiveness of our method.

CM-Diff: A Single Generative Network for Bidirectional Cross-Modality Translation Diffusion Model Between Infrared and Visible Images (Supplementary Material)

Section A evaluates the effectiveness of the proposed Statistical Constraint Inference (SCI) strategy through pixel intensity frequency histogram analysis. The results demonstrate that SCI enhances infrared-to-visible translation by effectively suppressing unnatural color artifacts and aligning the translated image distributions with those of the ground truth. Details of the CM-Diff architecture are also provided, including the Translation Direction Guidance (TDG), Cross-Modality Feature Control (CFC), and SCI modules.

Section B presents a comprehensive performance evaluation of cross-modality translation methods on object detection tasks using CFNet on the VEDAI dataset. Experimental results indicate that CM-Diff consistently outperforms state-of-the-art approaches such as ROMA and CUT in both infrared-to-visible and visible-to-infrared directions, achieving superior mAP scores. Moreover, translation experiments conducted on the Rain100H dataset further validate the model’s generalization capability beyond the IR-VIS domain.

Section C presents ablation and comparison studies that validate the effectiveness of key components in CM-Diff. Varying the constraint scale factors λ_{ccl} and λ_{scl} shows that moderate values (e.g., 20) offer the best balance between fidelity and diversity. Among the tested histogram distance metrics, χ^2 achieves the lowest FID and best perceptual quality. For structural guidance, DexiNed outperforms traditional edge detectors by better preserving semantic contours. Finally, comparisons with general I2I methods (UGATIT, UNSB, DDBM) highlight CM-Diff’s superior performance in cross-modality alignment, confirming the importance of domain-specific design.

Section D provides qualitative results on the M³FD, AVIID, and VEDAI datasets, illustrating the robustness of CM-Diff across a variety of challenging scenes. Analysis of training losses reveals that IR-to-VIS translation presents relatively lower optimization difficulty. Additional implementation details are also included, such as the use of a U-Net backbone with multi-scale attention, DexiNed-based edge fusion, and AdamW optimization. Evaluation metrics including FID, SSIM, LPIPS, and PSNR are discussed, along with detailed descriptions of the datasets used in the experiments.

A. Method and Module Analysis

Overall Framework

Figure 7 depicts the complete CM-Diff architecture, which comprises three core components:

- (1) **Translation Direction Guidance (TDG)**: steers the diffusion noise prediction toward the reverse gradient of the target modality;
- (2) **Cross-Modality Feature Control (CFC)**: injects semantic information across modalities at multiple feature scales to ensure content consistency;
- (3) **Statistical Constraint Inference (SCI)**: applies histogram-based statistical constraints ($\mathcal{L}_{ccl}$, $\mathcal{L}_{scl}$) during

Table 5: Quantitative evaluation of visible-to-infrared image translation performance of various translation methods based on CFNet.

Method	P	R	mAP_{50}	mAP_{75}	mAP
ROMA (Yu et al. 2022)	0.133	0.027	0.036	0.022	0.012
CUT (Park et al. 2020)	0.418	0.157	0.187	0.159	0.082
I2V-GAN (Li et al. 2021)	0.379	0.113	0.179	0.144	0.077
Ground Truth	0.471	0.368	0.449	0.306	0.168
Ours	0.313	0.198	0.281	0.249	0.138

sampling to reduce distribution shift.

SCI Module Deep Dive and Analysis

The SCI design draws inspiration from guided diffusion and distribution alignment techniques. By adjusting the mean and covariance of the target modality in the noise steps, it enforces closer adherence to the true data distribution. Histogram comparisons in Figures 5 and 6 confirm SCI’s efficacy for bidirectional infrared–visible translation.

As illustrated in Fig. 5, the proposed Statistical Constraint Inference (SCI) strategy introduces significant improvements in the infrared-to-visible image translation task. Specifically, when employing SCI, the frequency distributions of pixel intensities for each RGB channel shift notably towards lower intensity values. This shift is particularly evident in the green and blue channels, where high-frequency components associated with unnatural intensity peaks are substantially suppressed. As a result, the occurrence of color artifacts—such as oversaturation or unnatural hues—is greatly reduced in the translated visible images.

The effectiveness of SCI is further validated in the reverse translation task, i.e., visible-to-infrared conversion. As shown in Fig. 6, the translated infrared images exhibit histogram distributions that are more consistent with those of the ground truth infrared images when SCI is applied. This consistency reflects a more faithful preservation of modality-specific statistical properties during translation.

Overall, these observations highlight the advantage of integrating SCI into our framework. By enforcing statistical alignment between the translated output and the target domain, SCI not only improves the perceptual realism of the results but also mitigates the color distribution anomalies typically introduced by Bidirectional Diffusion Training (BDT). Consequently, the proposed **CM-Diff** model, when guided by SCI, demonstrates superior capability in performing high-fidelity, artifact-free cross-modality translation within a unified generative framework.

B. Performance Evaluation and Expandability

Generated Data Efficacy in Object Detection

We employ the recent small object detection method, CFNet (Yuan et al. 2023), and train it separately on infrared and visible data using the VEDAI dataset. Various translation methods are utilized to perform infrared-to-visible and

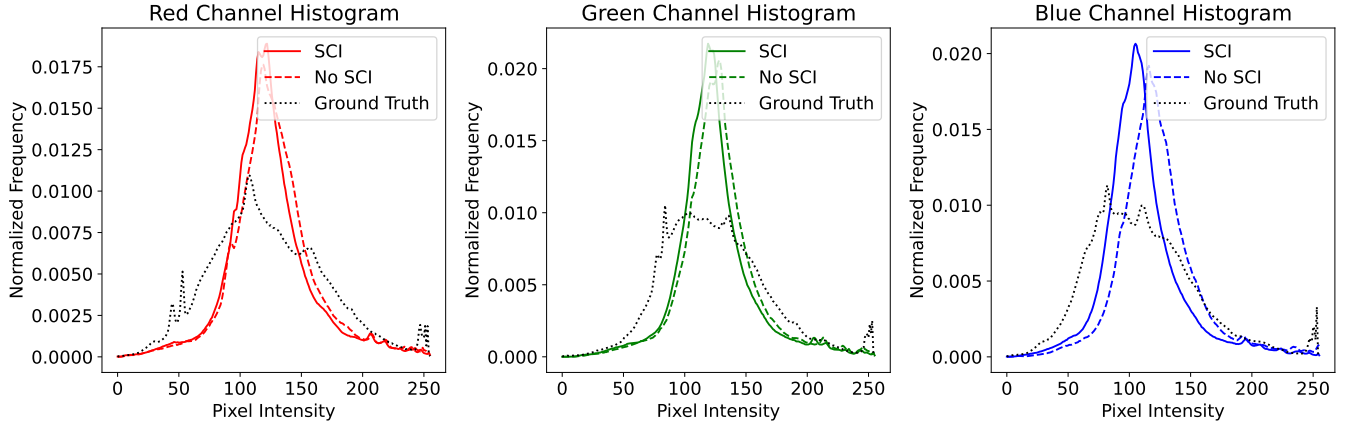


Figure 5: Comparison of histograms for each channel of visible images. The distribution of the images obtained using SCI, without SCI, and the ground truth visible images are presented.

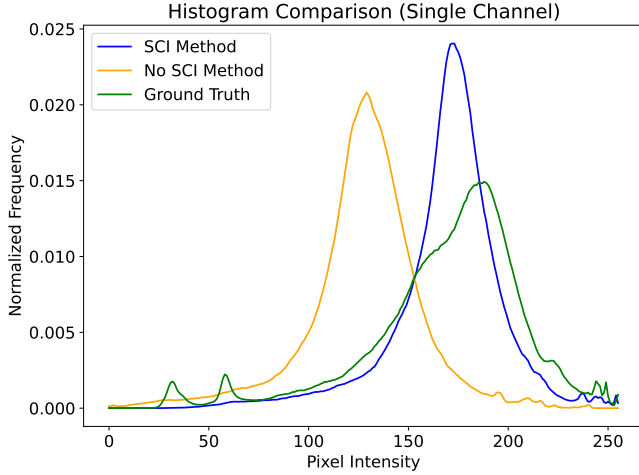


Figure 6: Comparison of histograms for infrared images. The distribution of the images obtained using SCI, without SCI, and the ground truth visible images are presented.

visible-to-infrared modality translation. The translation results are then evaluated using CFINet. For comparison, we selected several high-performing modality translation methods, including ROMA (Yu et al. 2022), CUT (Park et al. 2020), and I2V-GAN (Li et al. 2021). The quantitative results for the infrared and visible modalities are shown in Tables 5 and 6, respectively.

In the comparison of infrared modality translation results, significant differences can be observed in terms of precision (P), recall (R), and mean average precision (mAP). As shown in the table, the Ground Truth achieves the best performance, with the highest precision (0.471) and recall (0.368). It also attains the highest mAP values, including $mAP_{50} = 0.449$, $mAP_{75} = 0.306$, and an overall mAP of 0.168. Among the modality translation methods, our approach demonstrates the best performance, achieving mAP_{50} , mAP_{75} , and overall mAP values of 0.281, 0.249,

Table 6: Quantitative evaluation of infrared-to-visible image translation performance of various translation methods based on CFINet.

Method	P	R	mAP_{50}	mAP_{75}	mAP
ROMA (Yu et al. 2022)	0.240	0.106	0.175	0.136	0.072
CUT (Park et al. 2020)	0.185	0.025	0.068	0.059	0.032
I2V-GAN (Li et al. 2021)	0.168	0.079	0.129	0.090	0.048
Ground Truth	0.237	0.179	0.386	0.217	0.121
Ours	0.241	0.165	0.265	0.206	0.105

and 0.138, respectively. These results are significantly superior to those of ROMA, CUT, and I2V-GAN. Additionally, our method exhibits a recall of 0.198, which is substantially higher than that of other approaches, indicating its effectiveness in the object detection task. Regarding the other modality translation methods, CUT and I2V-GAN perform similarly in terms of precision and mAP, though CUT consistently achieves slightly better results. In contrast, ROMA exhibits the weakest performance across all metrics, suggesting that its translated infrared images contribute minimally to the detection task. In summary, our method generates higher-quality infrared images and substantially improves the performance of object detection tasks.

In the comparison of visible modality translation results, significant differences can be observed in terms of precision (P), recall (R), and mean average precision (mAP). As shown in the table, the Ground Truth achieves the best performance, with relatively high precision (0.237) and recall (0.179). It also attains the highest mAP values, including $mAP_{50} = 0.386$, $mAP_{75} = 0.217$, and an overall mAP of 0.121. Among the modality translation methods, our approach demonstrates superior performance compared to others. Specifically, it achieves mAP_{50} , mAP_{75} , and overall mAP values of 0.265, 0.206, and 0.105, respectively, significantly outperforming ROMA, CUT, and I2V-GAN. Additionally, our method exhibits a recall of 0.165, second only to the Ground Truth, indicating its effectiveness in the object detection task under the visible modality. Regarding

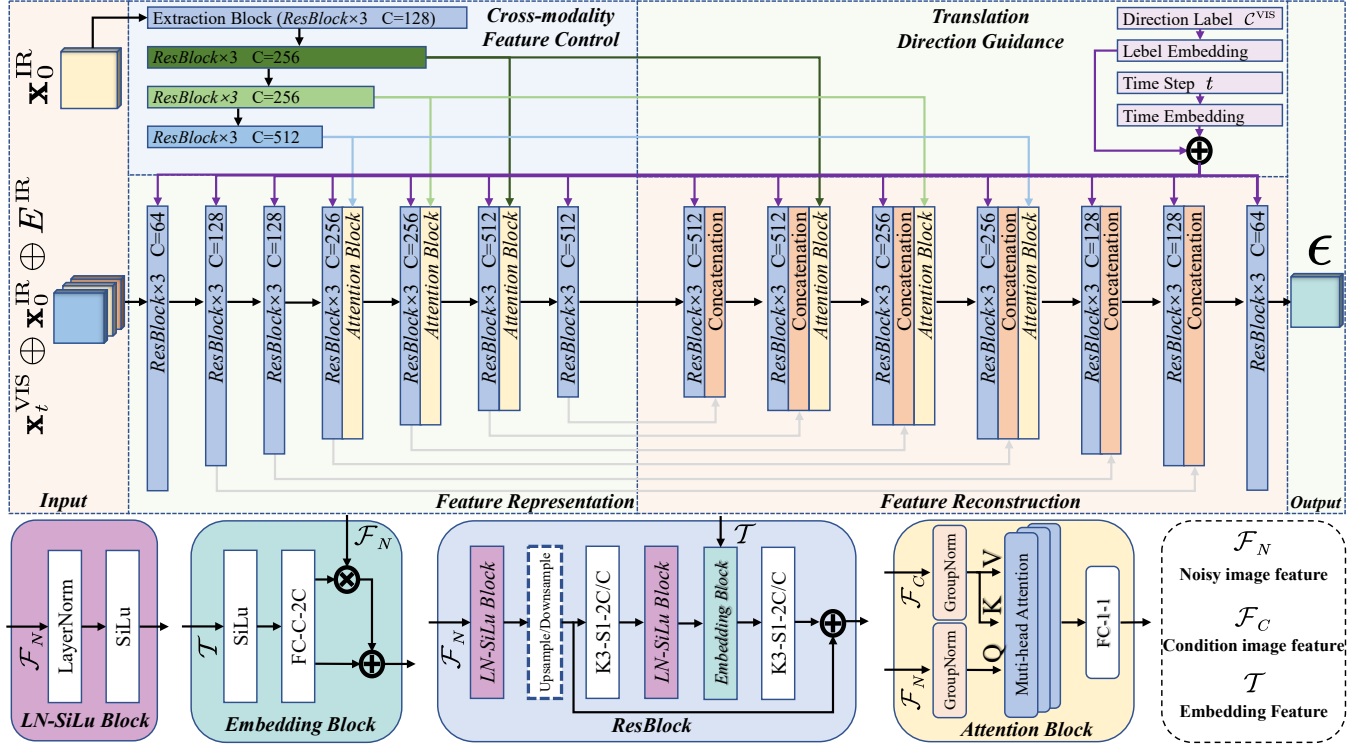


Figure 7: CM-Diff Architecture: Comprising Cross-Modality Feature Control (CFC), Translation Direction Guidance (TDG), and the Feature Representation and Reconstruction Network. CM-Diff processing a noisy input image by integrating edge feature maps and conditional modality images. CFC aligning noise translation with the reverse direction of semantic content in the conditional modality. TDG guiding the predicted noise toward the expected modality’s reverse direction. Finally, the network predicting the output noise.

other translation methods, ROMA performs relatively well in terms of precision and mAP but has a lower recall (0.106) compared to our model. CUT and I2V-GAN demonstrate weaker overall performance, with CUT exhibiting the lowest scores across all metrics, suggesting that its translated visible images are of lower quality and less beneficial for object detection tasks. In summary, our method generates higher-quality visible images and achieves superior performance in object detection tasks.

Finally, we observe that the quality of our infrared translation results is superior to that of visible translation results, both in terms of translation quality and object detection task performance. This observation is also supported by the training loss depicted in Fig. 12, where the loss for visible-to-infrared translation is consistently lower than that for infrared-to-visible translation. This phenomenon can be attributed to the inherent characteristics of infrared and visible images. Specifically, visible images consist of three channels representing the intensity values of red, green, and blue, whereas infrared images contain only a single channel reflecting infrared radiation intensity.

Applicability to Weather Translation task

As shown in Tab. 7. We directly applied our approach to a weather translation task without any architectural modifica-

Table 7: Comparison of different methods on the Rain100H weather translation task.

Method	Rain→ Sunny		Sunny→ Rain	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
CM-Diff (ours)	19.72	0.771	19.81	0.803
SEMI (Wei et al. 2019)	16.56	0.486	/	/
DA-CLIP(Luo et al. 2023)	33.91	0.926	/	/

tion and achieved competitive performance on the Rain100H dataset (Yang et al. 2017), which consists of 100 pairs of synthetic rainy and clean images commonly used for image deraining. This result demonstrates the generalization capability of our method beyond the IR–VIS domain.

C. Ablation and Comparison Studies

Analysis of the Constraint Scale Factor

To evaluate the effectiveness of constraint-based guidance in our generative framework, we conducted a series of ablation experiments focusing on the constraint scale factors λ_{ccl} and λ_{sc1} . As presented in Table 8, we systematically varied these parameters and assessed the corresponding impact on cross-modality translation performance using standard evaluation metrics including PSNR, SSIM, LPIPS, and FID.

The results demonstrate a clear performance trend: as the constraint scale factors increased, the network exhibited consistent improvements across all metrics. This indicates that stronger enforcement of content and structure consistency enhances the model’s capacity to preserve semantic information and spatial alignment during translation. Specifically, setting $\lambda_{ccl} = \lambda_{scl} = 20$ yielded the best overall performance, achieving the highest PSNR score of 19.89 and LPIPS score of 0.185 in the infrared-to-visible translation task, as well as the highest SSIM score of 0.883 in the visible-to-infrared task.

However, further increasing the constraint values beyond this optimal point (e.g., $\lambda = 40$ and $\lambda = 60$) resulted in a slight degradation of performance. For instance, when increasing the scale from 40 to 60, the visible-to-infrared translation task experienced a marginal drop of 0.12% in PSNR and 0.11% in LPIPS. This decline is attributed to the over-constraint of the mean estimation in the diffusion sampling process, which likely suppresses the model’s flexibility in modeling local uncertainty and fine-grained variations. These findings suggest that while constraint-based guidance is beneficial, excessive enforcement may hinder the generative diversity required for high-fidelity modality translation.

Table 8: Effect of the Constraint Scale Factors. The best performance for each metric is highlighted in bold.

$\lambda_{ccl}, \lambda_{scl}$	infrared $\rightarrow$ visible				visible $\rightarrow$ infrared			
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$
0	18.78	0.762	0.205	105.02	15.73	0.834	0.130	57.83
10	19.88	0.775	0.186	93.66	22.78	0.879	0.099	47.18
20	19.89	0.774	0.185	94.11	24.06	0.883	0.096	46.18
40	19.85	0.774	0.185	93.74	24.13	0.883	0.095	46.01
60	19.84	0.774	0.185	93.68	24.10	0.883	0.096	45.98

Histogram Distance Selection

To enforce statistical alignment between generated and target domains, we evaluate three histogram distance measures—Euclidean, Bhattacharyya, and χ^2 —within the Statistical Constraint Inference framework (Table 9). All metrics achieve comparable PSNR and LPIPS for infrared $\rightarrow$ visible translation and similar SSIM for visible $\rightarrow$ infrared translation, indicating their baseline effectiveness in distribution matching. Notably:

- **Euclidean distance** attains the highest PSNR (24.08 dB) and lowest LPIPS (0.095) for IR $\rightarrow$ VIS, but exhibits a higher FID (46.75), suggesting marginally inferior deep-feature congruence.
- **Bhattacharyya distance** yields PSNR and LPIPS on par with Euclidean for IR $\rightarrow$ VIS and slightly improves FID to 46.71, yet offers no significant benefit for VIS $\rightarrow$ IR.
- **χ^2 distance (proposed)** incurs only a minor PSNR (−0.05 dB) and LPIPS (+0.001) penalty relative to Euclidean in IR $\rightarrow$ VIS, while achieving the lowest FID scores in both directions (46.18 for IR $\rightarrow$ VIS and 94.11 for VIS $\rightarrow$ IR). This reflects its enhanced sensitivity to sparse but extreme discrepancies in pixel-intensity his-

Table 9: Performance comparison of different histogram distance metrics in SCI for bidirectional IR-VIS translation.

Histogram Distance	infrared $\rightarrow$ visible				visible $\rightarrow$ infrared			
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$
Euclidean	24.08	0.882	0.095	46.75	19.88	0.774	0.185	94.43
Bhattacharyya	24.08	0.882	0.096	46.71	19.88	0.774	0.185	94.45
Chi2 (ours)	24.03	0.883	0.096	46.18	19.89	0.774	0.185	94.11

Table 10: Comparison of edge detectors for bidirectional IR-VIS translation.

Edge Detector	infrared $\rightarrow$ visible				visible $\rightarrow$ infrared			
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$
Sobel	23.64	0.774	0.109	54.88	16.90	0.563	0.180	63.14
Canny	21.70	0.729	0.142	64.11	16.51	0.537	0.199	67.77
DexiNed (ours)	24.03	0.883	0.096	46.18	19.89	0.774	0.185	94.11

tograms, yielding crisper and more artifact-free translations.

Consequently, the χ^2 metric provides the optimal compromise between pixel-level fidelity and perceptual realism and is therefore adopted as the default histogram distance.

Edge Detector Selection and Justification

Effective cross-modality feature integration hinges on robust structural priors. Conventional detectors such as Sobel and Canny employ fixed convolutional kernels to infer gradients, but their performance deteriorates in low-contrast or noisy infrared imagery. To overcome these limitations, we integrate DexiNed, a deep learning–based edge extractor optimized for semantic contour delineation and fine-grained detail preservation across heterogeneous visual conditions.

As shown in Table 10, DexiNed consistently outperforms both Sobel and Canny for bidirectional translation tasks. In the infrared $\rightarrow$ visible direction, DexiNed attains a superior PSNR of 24.03 dB and SSIM of 0.883, alongside the lowest LPIPS (0.096) and FID (46.18). Similarly, for visible $\rightarrow$ infrared translation, DexiNed achieves the highest PSNR (19.89 dB) and SSIM (0.774) while delivering competitive LPIPS and FID metrics. These quantitative improvements corroborate that DexiNed’s learned edge representations furnish more reliable and semantically meaningful guidance, thereby enhancing both pixel-level accuracy and perceptual fidelity without incurring substantial computational overhead.

Positioning Against General Image-to-Image Translation Methods

General-purpose I2I models, optimized for RGB-to-RGB mappings, are ill-equipped to reconcile the substantial spectral and statistical disparities inherent in IR-VIS translation. As reported in Table 11, UGATIT (Kim et al. 2019) and UNSB (Kim et al. 2023) achieve only PSNR = 17.86 dB / 11.99 dB and SSIM = 0.602 / 0.415 for IR $\rightarrow$ VIS, with LPIPS = 0.358 / 0.433 and FID = 149.46 / 128.56, respectively—results that reflect both poor structural fidelity and perceptual realism. Even DDBM (Zhou et al. 2023), despite delivering a high PSNR of 26.26 dB, suffers from an excessively large FID of 227.34, indicating severe deep-feature

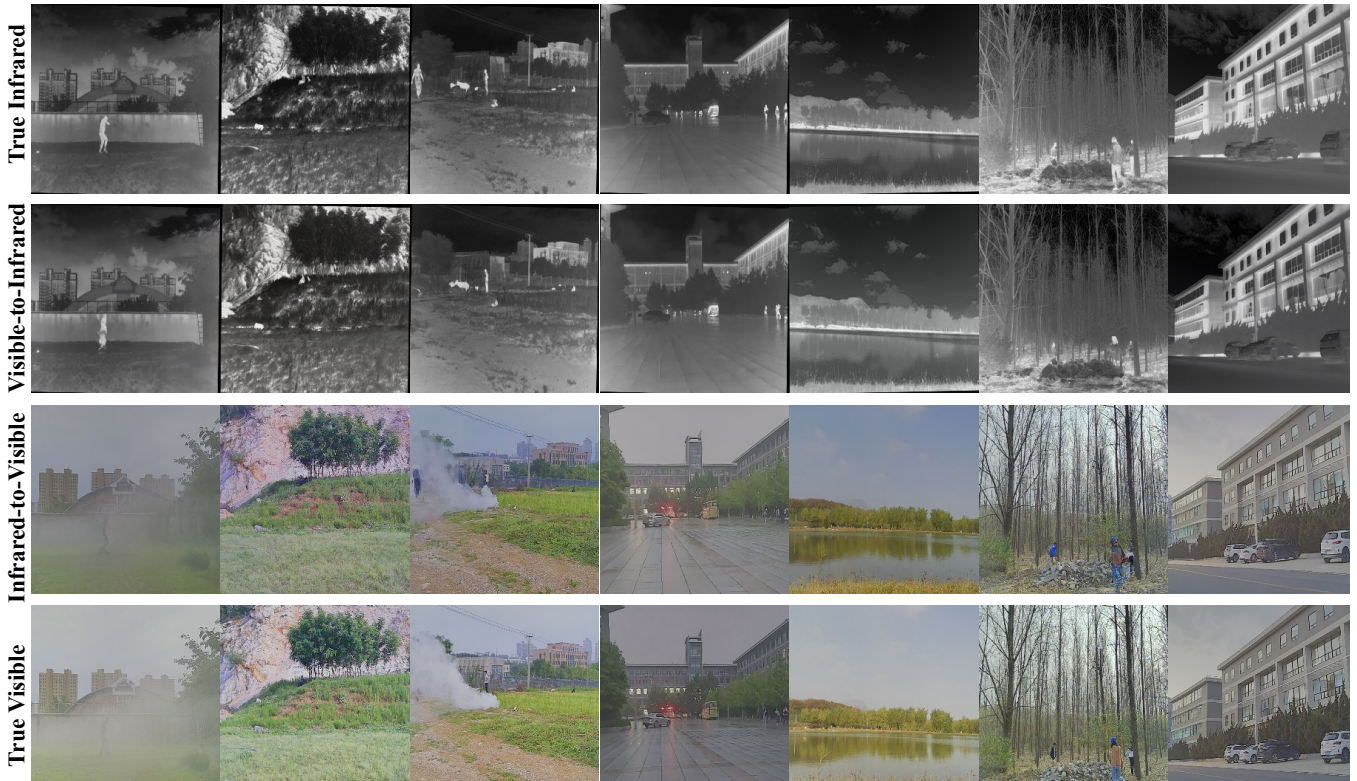


Figure 8: Modality translation results on the M³FD dataset. The first row represents real infrared images, the second and third rows show generated infrared and visible images, respectively, and the fourth row represents real visible images.

Table 11: Performances of Different Methods on VEDAI.

Histogram Distance	infrared $\rightarrow$ visible				visible $\rightarrow$ infrared			
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$
UGATT (ICLR'20)	17.86	0.602	0.358	149.46	16.18	0.461	0.379	133.98
UNSB (ICLR'24)	11.99	0.415	0.433	128.56	14.71	0.408	0.475	148.98
DDBM (ICLR'24)	26.26	0.748	0.405	227.34	22.26	0.567	0.526	241.35
CM-Diff (ours)	23.54	0.765	0.116	57.26	17.03	0.554	0.184	66.19

mismatches. By contrast, CM-Diff attains a well-balanced performance (PSNR = 23.54 dB, SSIM = 0.765, LPIPS = 0.116, FID = 57.26 for IR $\rightarrow$ VIS), and maintains competitive metrics for VIS $\rightarrow$ IR (PSNR = 17.03 dB, SSIM = 0.554, LPIPS = 0.184, FID = 66.19). These quantitative contrasts confirm that our modality-specific bidirectional diffusion framework and statistical constraint inference are critical to bridging the IR-VIS domain gap. Consequently, we exclude general I2I baselines from the main text to preserve focus on specialized IR-VIS comparisons, relegating their evaluation to this supplementary section.

D. Qualitative Results and Experimental Setup

Additional Translation Examples

As illustrated in Figs. 8, 9, and 10, we provide additional qualitative results on the M³FD, AVIID, and VEDAI datasets to further validate the effectiveness and generalizability of our proposed method. Each figure showcases representative samples drawn from diverse scenarios.

In these figures, the first row displays the original infrared images, while the fourth row presents the corresponding visible images. The second and third rows illustrate the translated results obtained using the visible-to-infrared and infrared-to-visible translation pathways, respectively. The input conditions and their corresponding outputs are aligned vertically to facilitate clear comparison.

These samples cover a wide range of complex outdoor environments, including urban streets, residential areas, forested regions, wilderness scenes, and water bodies such as lakes and rivers. Despite significant variability in lighting conditions, object scales, textures, and background complexity, our method—CM-Diff—consistently generates visually coherent and semantically aligned translations across both modalities.

The results confirm that CM-Diff exhibits robust cross-modality translation capabilities, maintaining structural consistency and minimizing color distortion even in challenging real-world scenarios. This highlights the model’s strong adaptability and potential for deployment in practical aerial sensing and remote vision applications.

Ablation Visualizations

To further investigate the contribution of each novel component in our proposed framework, we present qualitative results from a series of ablation experiments conducted on the VEDAI (Razakarivony and Jurie 2016) dataset. These exper-

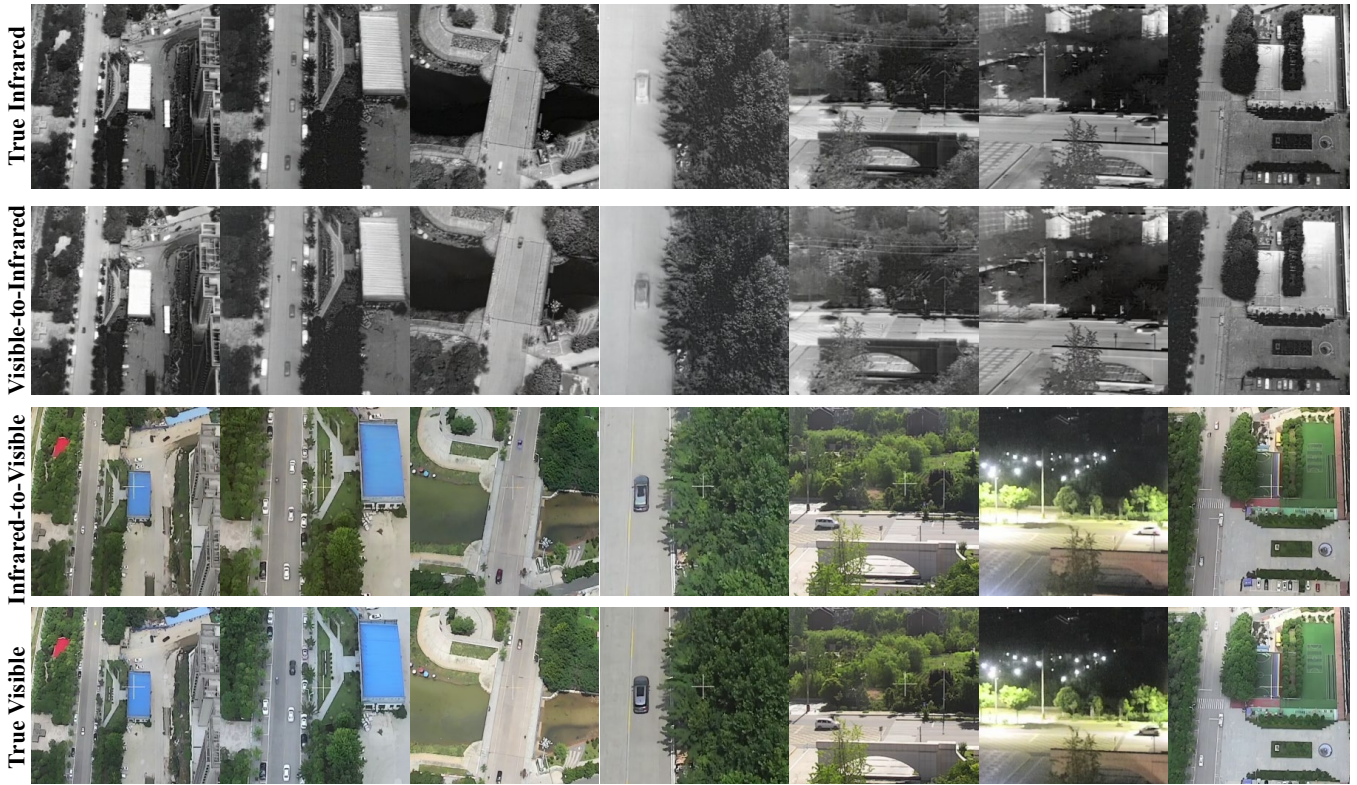


Figure 9: Modality translation results on the AVIID dataset. The first row represents real infrared images, the second and third rows show generated infrared and visible images, respectively, and the fourth row represents real visible images.

iments cover both infrared-to-visible and visible-to-infrared image translation tasks. As illustrated in Fig. 11, each column corresponds to a different model variant, allowing for direct visual comparison of the impact of specific design choices.

Notably, the translated results shown in the second, fourth, and fifth columns exhibit clear color distortions and unnatural artifacts. These models lack the statistical guidance provided by our proposed constraint mechanism, and consequently, fail to properly regulate the sampling process during generation. This often leads to aberrant color distribution or poor semantic alignment with the target domain. Such visual artifacts are especially pronounced in high-frequency regions, indicating the importance of statistical consistency in maintaining photorealistic quality.

In contrast, the results shown in the seventh column of Fig. 11 correspond to our full model, which integrates both the cross-modality feature interaction mechanism and the statistical constraint inference (SCI) strategy. These images display enhanced color realism, better-preserved object boundaries, and overall superior visual fidelity—closely resembling the ground-truth visible and infrared images. Moreover, the reduced blurriness and clearer contours in these results also validate the effectiveness of incorporating direction label guidance, which facilitates more targeted and stable translation between the two modalities.

Together, these observations underscore the neces-

sity of combining feature-level cross-modality control, distribution-level statistical constraint, and direction-aware supervision to achieve high-quality and consistent image translation in both directions.

Implementation Details

To effectively integrate feature information across infrared and visible modalities, we incorporate a cross-modality attention mechanism at multiple resolutions within both the encoder and decoder pathways of a U-Net backbone. Specifically, attention modules with 64 channels are applied at feature map resolutions of 8×8 , 16×16 , and 32×32 , enabling rich feature interaction across modalities at various semantic levels. To further enhance edge-awareness, we utilize a pre-trained DexiNed model to extract fine-grained edge information from both infrared and visible inputs. These edge maps are fused into the main network to improve structural detail preservation in the final output.

All input images are uniformly resized to a resolution of 256×256 for both training and evaluation. The U-Net backbone is configured with a base channel width of 128 across all datasets. Training is conducted using the AdamW optimizer with a batch size of 6. The learning rate is initialized to 1×10^{-4} and decayed multiplicatively by a factor of 0.9 every 2,000 iterations. The total training procedure spans 100,000 iterations.

To improve the fidelity of the generation process, we mod-



Figure 10: Modality translation results on the VEDAI dataset. The first row represents real infrared images, the second and third rows show generated infrared and visible images, respectively, and the fourth row represents real visible images.

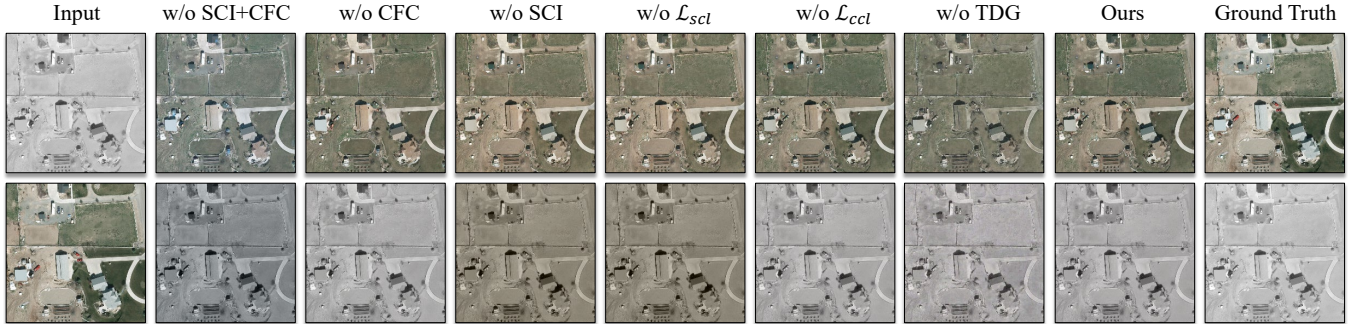


Figure 11: Comparisons of different ablation studies for the visible-to-infrared image translation task. We evaluate the impact of the Statistical Constraint Inference (SCI), which consists of $\mathcal{L}_{ccl}$ and $\mathcal{L}_{scl}$, the Cross-modality Feature Control (CFC) component, and the translation direction guidance (TDG).

ify the original noise schedule by reducing the final noise variance β_T from 0.02 to 0.01, effectively lowering the amount of noise injected during diffusion and thereby reducing degradation artifacts. The revised noise schedule linearly increases β_t from $\beta_1 = 0.0001$ to $\beta_T = 0.01$.

Given the symmetric nature of modality translation between infrared and visible domains, we set both translation loss weights equally as $\lambda_{IR \rightarrow VIS} = \lambda_{VIS \rightarrow IR} = 1.0$, even though the infrared modality typically presents a lower learning difficulty. During the generative phase, we adopt two constraint-based guidance terms: a content-consistency

loss ($\lambda_{ccl} = 20.0$) and a structure-consistency loss ($\lambda_{scl} = 20.0$), both of which help to maintain semantic alignment and spatial coherence across modalities.

All experiments are conducted on a computing platform equipped with an Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz and four NVIDIA V100 GPUs.

Metrics

To objectively assess the performance of our proposed method on cross-modality image translation tasks, we adopt four widely used image quality assessment metrics: Frechet

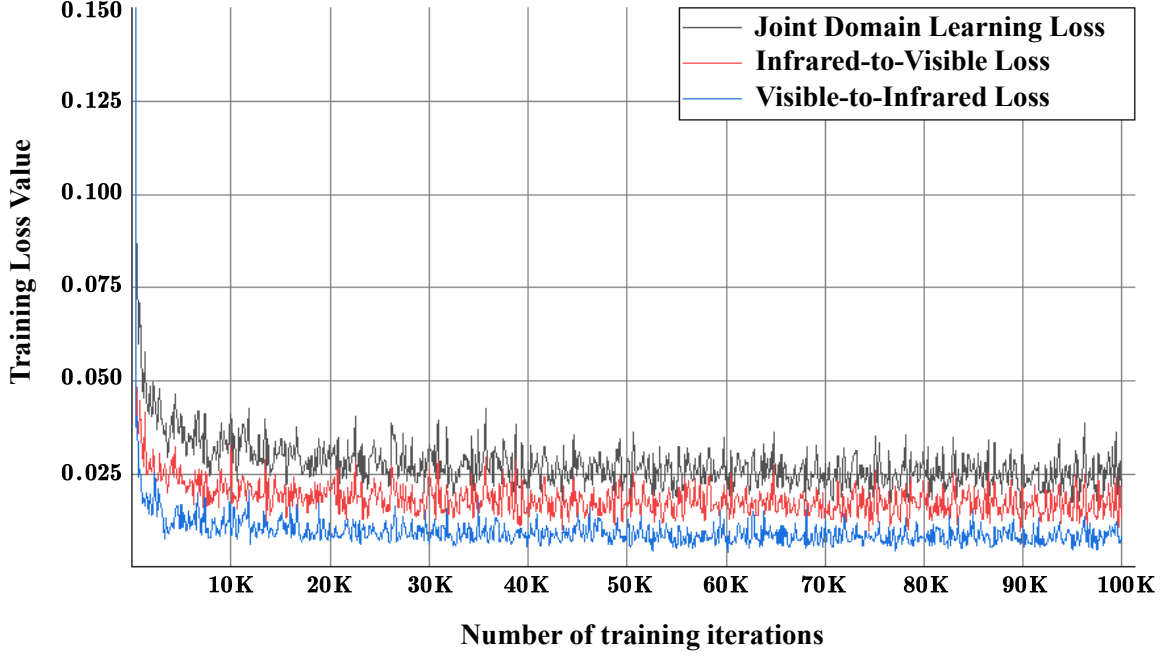


Figure 12: The training loss evolves with the number of iterations. Due to the significant fluctuations in the loss values of individual iterations, which make it challenging to discern overall trends, we instead utilize the average loss value over an entire training epoch and use the iteration count as the horizontal axis.

Inception Distance (FID), Structural Similarity Measure (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), and Peak Signal-to-Noise Ratio (PSNR). These metrics provide a comprehensive evaluation from both perceptual and pixel-level perspectives.

Frechet Inception Distance (FID) (Heusel et al. 2017) quantifies the distance between the feature distributions of real and generated images. Lower FID scores indicate higher similarity between the generated and real image domains in terms of deep features. It is defined as:

$$\text{FID}(v', v) = \|\mu_v - \mu_{v'}\|^2 + \text{tr}(\Sigma_v + \Sigma_{v'} - 2(\Sigma_v \cdot \Sigma_{v'})^{1/2}), \quad (15)$$

where μ_v and $\mu_{v'}$ are the mean feature vectors of real and generated images, respectively, and Σ_v and $\Sigma_{v'}$ are the corresponding covariance matrices. These features are extracted using a pre-trained Inception-V3 network.

Structural Similarity Index Measure (SSIM) (Wang et al. 2004) evaluates the similarity between the translated and ground-truth images based on luminance, contrast, and structural information. SSIM values range from 0 to 1, with higher scores indicating better perceptual quality. The metric is computed as:

$$\text{SSIM}(v', v) = \frac{(2\mu_v \cdot \mu_{v'} + c_1)(2\sigma_{v,v'} + c_2)}{(\mu_v^2 + \mu_{v'}^2 + c_1)(\sigma_v^2 + \sigma_{v'}^2 + c_2)}, \quad (16)$$

where μ_v and $\mu_{v'}$ denote the mean intensities, σ_v^2 and $\sigma_{v'}^2$ represent the variances, and $\sigma_{v,v'}$ is the covariance between v and v' . c_1 and c_2 are stability constants.

Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018a) measures perceptual simi-

larity by comparing deep feature representations of image patches. It is sensitive to semantic differences and has been shown to correlate well with human perception. LPIPS is computed as:

$$\text{LPIPS}(v', v) = \sum_{i=1}^I \sum_{h_i, w_i} \frac{\omega_i \cdot [f_i(v')_{h_i, w_i} - f_i(v)_{h_i, w_i}]^2}{\mathcal{H}_i \cdot \mathcal{W}_i}, \quad (17)$$

where $f_i(\cdot)$ denotes the features extracted from the i -th layer of a pre-trained AlexNet (Krizhevsky, Sutskever, and Hinton 2012), and ω_i is the learned weight for the i -th layer. $\mathcal{H}_i$ and $\mathcal{W}_i$ denote the spatial dimensions of the i -th layer's feature map.

Peak Signal-to-Noise Ratio (PSNR) evaluates image fidelity at the pixel level by measuring the logarithmic ratio between the maximum possible pixel value and the mean squared error (MSE) between the real and translated images. It is defined as:

$$\text{PSNR}(v', v) = 10 \cdot \log_{10} \left(\frac{\mathcal{M} \cdot \mathcal{N} \cdot \text{INTENSITY}_{\max}^2}{\sum_{i=1}^{\mathcal{M}} \sum_{j=1}^{\mathcal{N}} [v_{ij} - v'_{ij}]^2} \right), \quad (18)$$

where $\mathcal{M}$ and $\mathcal{N}$ are the image height and width, respectively, and $\text{INTENSITY}_{\max}$ is the maximum pixel value, which is 255 for 8-bit images.

Datasets

The *VEDAI dataset* (Razakarivony and Jurie 2016) is designed for dual-modality small vehicle detection in aerial imagery. It contains a total of 1,209 paired infrared-visible

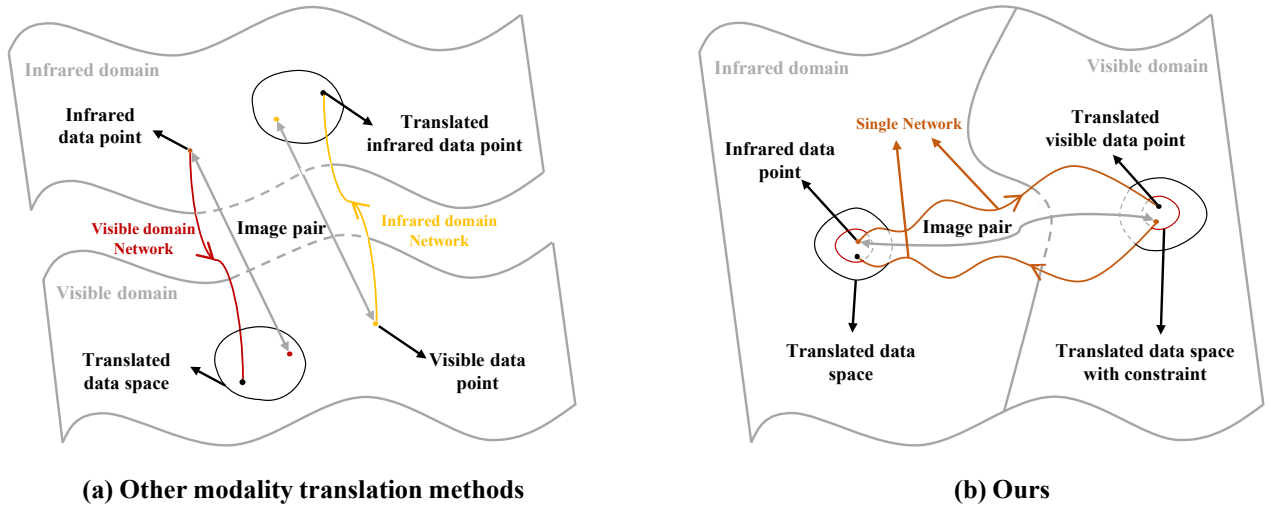


Figure 13: Comparison of bidirectional cross-modality translation frameworks. (a) Other modality translation methods: Independent networks for unidirectional translation (red / yellow arrows) operate in decoupled domains, relying on paired-image constraints but lacking joint optimization. (b) CM-Diff (Ours): A unified network performs simultaneous bidirectional mapping via direction-guided Bidirectional Diffusion Training (BDT), with Statistical Constraint Inference (SCI) enforcing distributional consistency across modalities ("translated data space with constraint").

image samples with resolutions of 1024×1024 and 512×512 , captured under varied scene conditions. The infrared modality is based on near-infrared (NIR) imaging, making the dataset particularly suitable for cross-modality image translation and detection tasks involving small objects with complex backgrounds. The dataset also provides comprehensive pixel-level annotations, which facilitate both supervised training and objective performance benchmarking. In our experiments, we allocate 1,089 image pairs for training and the remaining 120 for testing to ensure a fair and balanced evaluation.

The AVIID dataset (Han et al. 2023) consists of paired visible-infrared images captured by a dual-spectrum camera mounted on an unmanned aerial vehicle (UAV), across varying illumination and environmental conditions. It is subdivided into three subsets: AVIID-1 includes 993 image pairs of resolution 434×434 featuring road scenes and common vehicles; AVIID-2 contains 1,090 low-light image pairs with visible noise and motion blur; and AVIID-3 offers 1,280 image pairs at 512×512 resolution, covering diverse scenarios such as roads, bridges, and residential streets. While the dataset was originally constructed for the visible-to-infrared image translation task, our framework supports bidirectional translation. Therefore, we adopt the full dataset for both $IR \rightarrow VIS$ and $VIS \rightarrow IR$ tasks. We divide the dataset into 2,674 training pairs and 669 testing pairs.

The M³FD dataset (Liu et al. 2022) provides 4,500 spatially registered visible and infrared image pairs with a resolution of 1024×768 . Although initially designed for multi-sensor fusion and object detection tasks, the high-resolution and scene diversity of the dataset make it well-suited for image-to-image translation applications. The data is collected from multiple urban and suburban locations

in Dalian, China, including the Dalian University of Technology, Golden Stone Beach, and Jinzhou District. Despite slight misalignments between the modalities, the dataset remains valuable for training robust translation models. For our experiments, we utilize 3,780 image pairs for training and reserve 210 pairs for testing.

Table 12: Code and Dataset Sources for Evaluated Methods

Method	Code address
T2V-DDPM	https://github.com/Nithin-GK/T2V-DDPM
CycleGAN	https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix
UNIT	https://github.com/mingyuliuw/UNIT
MUNIT	https://github.com/NVlabs/MUNIT
DCLGAN	https://github.com/JunlinHan/DCLGAN
I2V-GAN	https://github.com/BIT-DA/I2V-GAN
ROMA	https://github.com/BIT-DA/ROMA
CUT	https://github.com/taesungp/contrastive-unpaired-translation
Dataset	Data address
AVIID	Aerial-visible-to-infrared-image-translation
VEDAI	https://downloads.greyc.fr/vedai/
M ³ FD	https://github.com/dlut-dimt/TarDAL

References

- Boroujeni, S. P. H.; and Razi, A. 2024. Ic-gan: An improved conditional generative adversarial network for rgb-to-ir image translation with applications to forest fire monitoring. *Expert Systems with Applications*, 238: 121962.
- Chung, H.; Kim, J.; Mccann, M. T.; Klasky, M. L.; and Ye, J. C. 2022. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*.
- Deng, F.; Feng, H.; Liang, M.; Wang, H.; Yang, Y.; Gao, Y.; Chen, J.; Hu, J.; Guo, X.; and Lam, T. L. 2021. FEANet:

- Feature-enhanced attention network for RGB-thermal real-time semantic segmentation. In *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, 4467–4473. IEEE.
- Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.
- Di, X.; Riggan, B. S.; Hu, S.; Short, N. J.; and Patel, V. M. 2019. Polarimetric thermal to visible face verification via self-attention guided synthesis. In *2019 International Conference on Biometrics (ICB)*, 1–8. IEEE.
- Ding, J.; Du, Y.; Li, W.; Pei, L.; and Cui, N. 2025. LG-Diff: Learning to follow local class-regional guidance for nearshore image cross-modality high-quality translation. *Information Fusion*, 117: 102870.
- Fu, H.; Wang, S.; Duan, P.; Xiao, C.; Dian, R.; Li, S.; and Li, Z. 2023. Lraf-net: Long-range attention fusion network for visible–infrared object detection. *IEEE Transactions on Neural Networks and Learning Systems*.
- Guo, J.; Gao, C.; Liu, F.; Meng, D.; and Gao, X. 2024. DAMSDet: Dynamic Adaptive Multispectral Detection Transformer with Competitive Query Selection and Adaptive Feature Fusion. *ECCV*.
- Han, J.; Shoeiby, M.; Petersson, L.; and Armin, M. A. 2021. Dual contrastive learning for unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 746–755.
- Han, Z.; Zhang, S.; Su, Y.; Chen, X.; and Mei, S. 2024a. Dravit: Towards diverse and realistic aerial visible-to-infrared image translation. *IEEE Transactions on Geoscience and Remote Sensing*.
- Han, Z.; Zhang, S.; Su, Y.; Chen, X.; and Mei, S. 2024b. DR-AVIT: Towards Diverse and Realistic Aerial Visible-to-Infrared Image Translation. *IEEE Transactions on Geoscience and Remote Sensing*.
- Han, Z.; Zhang, Z.; Zhang, S.; Zhang, G.; and Mei, S. 2023. Aerial visible-to-infrared image translation: dataset, evaluation, and baseline. *Journal of Remote Sensing*, 3: 0096.
- He, G.; Zheng, K.; Chen, J.; Bao, F.; and Zhu, J. 2024. Consistency diffusion bridge models. *Advances in Neural Information Processing Systems*, 37: 23516–23548.
- Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.
- Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, 172–189.
- Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.
- Ji, W.; Li, J.; Bian, C.; Zhou, Z.; Zhao, J.; Yuille, A. L.; and Cheng, L. 2023. Multispectral video semantic segmentation: A benchmark dataset and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1094–1104.
- Kim, B.; Kwon, G.; Kim, K.; and Ye, J. C. 2023. Unpaired image-to-image translation via neural schrödinger bridge. *arXiv preprint arXiv:2305.15086*.
- Kim, J.; Kim, M.; Kang, H.; and Lee, K. 2019. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*.
- Koller, D.; and Friedman, N. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.
- Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.
- Kwon, T.; and Ye, J. C. 2021. Cycle-free cyclegan using invertible generator for unsupervised low-dose ct denoising. *IEEE Transactions on Computational Imaging*, 7: 1354–1368.
- Li, L.; and He, J. 2024. Bidirectional consistency models. *arXiv preprint arXiv:2403.18035*.
- Li, Q.; Zhang, C.; Hu, Q.; Zhu, P.; Fu, H.; and Chen, L. 2023. Stabilizing multispectral pedestrian detection with evidential hybrid fusion. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Li, S.; Han, B.; Yu, Z.; Liu, C. H.; Chen, K.; and Wang, S. 2021. I2v-gan: Unpaired infrared-to-visible video translation. In *Proceedings of the 29th ACM international conference on multimedia*, 3061–3069.
- Lin, J.; Zhao, G.; Xu, J.; Wang, G.; Wang, Z.; Dantcheva, A.; Du, L.; and Chen, C. 2024. DiffTV: Identity-preserved thermal-to-visible face translation via feature alignment and dual-stage conditions. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 10930–10938.
- Liu, F.; Gao, C.; Zhang, Y.; Guo, J.; Wang, J.; and Meng, D. 2024. Infmae: A foundation model in infrared modality. *ECCV*.
- Liu, G.-H.; Vahdat, A.; Huang, D.-A.; Theodorou, E. A.; Nie, W.; and Anandkumar, A. 2023a. I²SB: Image-to-Image Schrödinger Bridge. *arXiv preprint arXiv:2302.05872*.
- Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; and Luo, Z. 2022. Target-aware Dual Adversarial Learning and a Multi-scenario Multi-Modality Benchmark to Fuse Infrared and Visible for Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5802–5811.
- Liu, J.; Liu, Z.; Wu, G.; Ma, L.; Liu, R.; Zhong, W.; Luo, Z.; and Fan, X. 2023b. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8115–8124.
- Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30.

- Liu, S.; John, V.; Blasch, E.; Liu, Z.; and Huang, Y. 2018. Ir2vi: Enhanced night environmental perception by unsupervised thermal image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1153–1160.
- Luo, F.; Li, Y.; Zeng, G.; Peng, P.; Wang, G.; and Li, Y. 2022. Thermal infrared image colorization for nighttime driving scenes with top-down guided attention. *IEEE Transactions on Intelligent Transportation Systems*, 23(9): 15808–15823.
- Luo, Z.; Gustafsson, F. K.; Zhao, Z.; Sjölund, J.; and Schön, T. B. 2023. Controlling vision-language models for universal image restoration. *arXiv preprint arXiv:2310.01018*, 3(8).
- Ma, D.; Li, S.; Su, J.; Xian, Y.; and Zhang, T. 2024. Visible-to-Infrared Image Translation for Matching Tasks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.
- Maqsood, R.; Abid, F.; et al. 2023. Cycle Consistency and Fine-Grained Image to Image Translation in Augmentation: An Overview.
- Meng, C.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.
- Nair, N. G.; and Patel, V. M. 2023. T2v-ddpm: Thermal to visible face translation using denoising diffusion probabilistic models. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, 1–7. IEEE.
- Oliveira, G. C.; Ngo, Q. C.; Papa, J. P.; and Kumar, D. 2024. A Stable Diffusion Approach for RGB to Thermal Image Conversion for Leg Ulcer Assessment. In *2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS)*, 158–163. IEEE.
- Ordun, C.; Raff, E.; and Purushotham, S. 2023. When visible-to-thermal facial GAN beats conditional diffusion. In *2023 IEEE International Conference on Image Processing (ICIP)*, 181–185. IEEE.
- Park, T.; Efros, A. A.; Zhang, R.; and Zhu, J.-Y. 2020. Contrastive learning for unpaired image-to-image translation. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, 319–345. Springer.
- Razakarivony, S.; and Jurie, F. 2016. Vehicle detection in aerial imagery: A small target detection benchmark. *Journal of Visual Communication and Image Representation*, 34: 187–203.
- Ren, K.; Zhao, W.; Gu, G.; and Chen, Q. 2023. EADS: Edge-assisted and dual similarity loss for unpaired infrared-to-visible video translation. *Infrared Physics & Technology*, 134: 104936.
- Sasaki, H.; Willcocks, C. G.; and Breckon, T. P. 2021. Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2104.05358*.
- Soria, X.; Sappa, A.; Humanante, P.; and Akbarinia, A. 2023. Dense extreme inception network for edge detection. *Pattern Recognition*, 139: 109461.
- Wang, H.; Li, N.; Zhao, H.; Wen, Y.; Su, Y.; and Fang, Y. 2024. MappingFormer: Learning cross-modal feature mapping for visible-to-infrared image translation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 10745–10754.
- Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.
- Wei, W.; Meng, D.; Zhao, Q.; Xu, Z.; and Wu, Y. 2019. Semi-Supervised Transfer Learning for Image Rain Removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yadav, N. K.; Singh, S. K.; and Dubey, S. R. 2023. TVA-GAN: attention guided generative adversarial network for thermal to visible image transformations. *Neural Computing and Applications*, 35(27): 19729–19749.
- Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; and Yang, M.-H. 2023. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4): 1–39.
- Yang, W.; Tan, R. T.; Feng, J.; Liu, J.; Guo, Z.; and Yan, S. 2017. Deep joint rain detection and removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1357–1366.
- Yu, Z.; Chen, K.; Li, S.; Han, B.; Liu, C. H.; and Wang, S. 2022. ROMA: cross-domain region similarity matching for unpaired nighttime infrared to daytime visible video translation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5294–5302.
- Yuan, X.; Cheng, G.; Yan, K.; Zeng, Q.; and Han, J. 2023. Small object detection via coarse-to-fine proposal generation and imitation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6317–6327.
- Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018a. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.
- Zhang, T.; Wiliem, A.; Yang, S.; and Lovell, B. 2018b. Tv-gan: Generative adversarial network based thermal to visible face recognition. In *2018 international conference on biometrics (ICB)*, 174–181. IEEE.
- Zhao, M.; Bao, F.; Li, C.; and Zhu, J. 2022. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *Advances in Neural Information Processing Systems*, 35: 3609–3623.
- Zhou, L.; Lou, A.; Khanna, S.; and Ermon, S. 2023. Denoising diffusion bridge models. *arXiv preprint arXiv:2309.16948*.
- Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.