D3F-IR: Dual-Domain Deterministic Flow Matching for Visible-to-Infrared Translation

Peiyi Zeng[1*], Diedong Feng[1*], Zhen Liu[1], Zhenming Peng[1], Bing Zeng[1], and Shuaicheng Liu[1†]

University of Electronic Science and Technology of China, Chengdu, China
liushuaicheng@uestc.edu.cn

Abstract. Visible-to-infrared translation provides a scalable pathway to synthesize infrared imagery from abundant visible data, offering a practical complement to costly cross-modal paired data acquisition. However, most existing generative approaches formulate the task as a unified process that implicitly entangles macroscopic thermal semantics with microscopic spatial detail reconstruction. This entanglement often forces a sub-optimal trade-off, leading to blurred target boundaries or physically unnatural thermal artifacts. In this paper, we reformulate the translation process into two explicitly decoupled but tightly coordinated continuous flows, proposing a novel framework named D3F-IR. Leveraging the smooth and deterministic trajectories of flow matching, D3F-IR introduces a semantic-domain flow to model macroscopic thermal distributions within a pretrained latent space, alongside an independent pixel-domain flow to generate fine-grained spatial structures directly in the image space. To seamlessly coordinate the two domains, we design a Pixel-space Velocity Predictor equipped with SEmantic-Aligned Modulation (SEAM) Layers, which establish explicit correspondence between semantic tokens and spatial patches through patch-wise alignment. Experiments on three benchmarks demonstrate that D3F-IR achieves strong performance in both thermal semantic plausibility and pixel-level spatial fidelity. These results highlight that explicit dual-domain decoupling offers a highly effective paradigm for high-quality visible-to-infrared generation. Code is available at https://github.com/WanrenZeng/D3F-IR.

Keywords: Visible-to-Infrared Image Translation · Flow Matching

1 Introduction

Visual-thermal sensor fusion and cross-modality downstream tasks, including low-light perception surveillance and autonomous driving, have become increasingly important for achieving robust perception under adverse environmental conditions [1, 20, 36, 39]. In such environments, visible-light sensors degrade significantly due to their strict dependence on illumination, whereas infrared sensing reliably captures thermal radiance. While deep learning models have shown

* Equal contribution.
† Corresponding author.

tremendous promise in these domains, they typically rely heavily on paired visible and thermal datasets for training. However, acquiring synchronized and precisely calibrated cross-modal image pairs is highly expensive and often restricted by limited viewpoints and specific sensor configurations [12, 35, 38]. Therefore, visible-to-infrared image translation has emerged as a scalable and flexible alternative. Synthesizing thermal images from abundant visible-spectrum data provides a cost-effective pathway to construct paired datasets and enhances the robustness of downstream tasks against various thermal characteristics [45].

Early visible-to-infrared translation methods primarily relied on generative adversarial networks to model the cross-modal mapping [13, 17, 51]. While these methods improve visual realism, they frequently suffer from training instability and tend to generate structural distortions or texture hallucinations due to the limited constraints on global geometric consistency [7, 30]. Subsequently, diffusion models and flow-based generative paradigms have emerged to bring greater stability and controllability to conditional image generation. These advanced frameworks further incorporate vision-language understanding [35], foundation-model priors [31], or style-disentangled mechanisms [45] to effectively enhance generation quality and spectral fidelity.

Despite recent advances in generative modeling, visible-to-infrared translation remains fundamentally challenging because it requires simultaneously inferring macroscopic thermal semantics and preserving microscopic spatial structures from visible inputs that lack explicit temperature cues. Although similar decoupling ideas have been explored in broader vision research, most of them focus on content-style separation, and most advanced V2IR generative frameworks [28, 31, 35] still formulate the task as a unified generation process. This unified formulation implicitly entangles macroscopic thermal semantic assignment and microscopic spatial detail synthesis within a single continuous flow or latent space. Handling these distinct objectives within a unified representation space significantly increases the learning difficulty and often forces a sub-optimal compromise between global thermal plausibility and local texture fidelity.

To overcome the limitations of unified generative processes, we propose D3F-IR, which is a dual-domain deterministic flow matching framework. Our core insight is to explicitly decouple the translation process into two synchronized continuous flows. A semantic-domain flow initialized by a pretrained encoder captures global thermal priors and models macroscopic thermal distributions. Meanwhile, a synchronized pixel-domain flow driven by a carefully designed pixel-space velocity predictor synthesizes fine-grained spatial structures under dynamic guidance from the semantic flow. To ensure coordinated interaction between the two domains, we introduce SEmantic-Aligned Modulation (SEAM) layers based on patch-wise spatial alignment. These layers establish explicit correspondence between semantic tokens and spatial patches within the pixel branch and enable structured semantic guidance during pixel-level synthesis.

Bringing these insights together, we develop a principled generative framework for visible-to-infrared translation. Our main contributions are as follows:

- We propose D3F-IR, a dual-domain deterministic flow matching framework that explicitly decouples thermal semantic modeling and pixel-level spatial detail synthesis, thereby effectively avoiding the sub-optimal trade-offs caused by entangled optimization.
- We design a dedicated Pixel-space Velocity Predictor for structured pixel-domain evolution. It leverages the deterministic nature of flow matching and is equipped with Semantic-Aligned Modulation layers to establish explicit correspondence between semantic guidance and spatial reconstruction.
- Extensive experiments on three standard benchmarks demonstrate that D3F-IR achieves state-of-the-art quantitative performance while producing semantically coherent thermal structures and high-fidelity spatial details.

2 Related Work

Visible-to-Infrared Translation. Generative adversarial networks [6] pioneered early visible-to-infrared translation research. Approaches such as Pix2Pix [13] and CycleGAN [51] established paired and unpaired mapping paradigms. Subsequent architectures further refined spectral fidelity through dedicated loss formulations and structural adjustments [7,17–19,27,30,41]. However, adversarial training remains notoriously unstable and frequently produces unnatural visual artifacts. Other structural paradigms include transformer-based contrastive alignments [43] and saliency-guided spatial refinements [21].

Recently, advanced generative models have demonstrated superior cross-modal synthesis capabilities. Diffusion paradigms investigate latent decomposition [28] alongside vision-language pretraining [35] and foundation model integration [2,16,24,31]. UNSB [14] formulates the translation process via stochastic transport bridges. Furthermore ThermalGen [45] introduces a flow-based architecture for thermal image synthesis. Despite these advances, existing single-network paradigms entangle macroscopic thermal semantic assignment and microscopic spatial detail synthesis. This entanglement restricts the model capacity and forces a compromise between global temperature distribution and local texture rendering. Our work diverges from these methods by explicitly decoupling the translation process into a dual-domain framework to simultaneously achieve thermal semantic plausibility and pixel-level spatial fidelity.

Flow Matching Models. Flow matching formulates generative modeling as learning continuous-time probability paths via the estimation of time-dependent vector fields [22]. Notably, recent theoretical analyses reveal its favorable geometric generality and superior guidance flexibility compared to standard diffusion frameworks [4, 8, 50]. Consequently, subsequent algorithmic developments have extensively explored rectified formulations that produce straighter transport trajectories [3] as well as consistency-based techniques to achieve significantly accelerated generation [5, 46].

Deterministic variants of flow matching further replace stochastic noise priors with structured source distributions to reduce sampling variance and improve trajectory stability. RAW-Flow [25] explores this deterministic formulation and

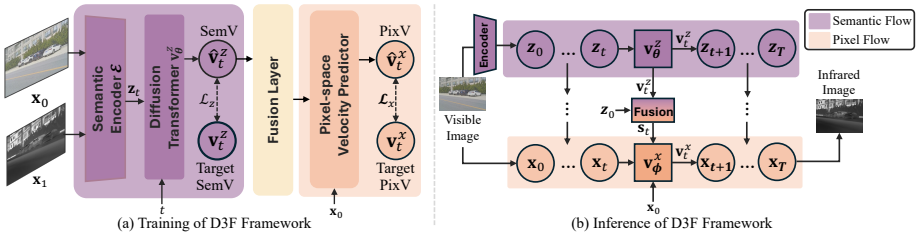

Fig. 1: Overview of the dual-domain deterministic flow matching framework of D3F-IR. (a) During training, a Diffusion Transformer predicts $SemV$, the semantic-domain velocity, which is integrated into a guidance signal to supervise a pixel-space velocity predictor for $PixV$, the pixel-domain velocity. (b) During inference, the semantic and pixel states are updated synchronously over T steps to generate the infrared image, where $\mathbf{x}_0$ and $\mathbf{x}_1$ denote the visible input and the target infrared image, respectively.

demonstrates improved reconstruction quality. However, existing deterministic flow applications typically operate within a single latent or image space. In contrast, we extend this paradigm by orchestrating two synchronized deterministic flows across semantic and pixel domains.

3 Methodology

We introduce D3F-IR, a visible-to-infrared translation framework designed to explicitly decouple thermal semantic modeling from the synthesis of pixel-level spatial details. The proposed approach employs a novel dual-domain deterministic flow matching framework, whose overall training and inference pipelines are illustrated in Figure 1. A semantic encoder based on the DINO architecture [37] first extracts semantic latents from the input RGB image to initialize the semantic-domain flow. Subsequently, a Diffusion Transformer [34] drives the evolution of this semantic flow toward the target infrared semantic latent. At each iteration, the predicted semantic velocity field from the Transformer provides dynamic guidance to the pixel-domain flow through a dedicated pixel-space velocity predictor. By iteratively advancing the pixel-domain flow under this guidance, the framework generates infrared images that achieve both thermal semantic plausibility and high spatial fidelity.

3.1 Dual-Domain Deterministic Flow Matching

The D3F-IR framework jointly models two synchronized continuous flows defined in the semantic domain and the pixel domain, respectively. Given a source visible image $\mathbf{x}_0$ at time $t = 0$ and a corresponding target infrared image $\mathbf{x}_1$ at $t = 1$, the semantic latent variable at time t is denoted by $\mathbf{z}_t$, while the image variable is denoted by $\mathbf{x}_t$. This dual-domain formulation facilitates the evolution of representations at different abstraction levels while maintaining rigorous coordination through dynamic guidance between the two flows.

Deterministic flow matching [25] models data transformation as a continuous-time ordinary differential equation (ODE) that transports a variable from an initial state $\mathbf{y}_0$ to a target state $\mathbf{y}_1$ over the interval $t \in [0, 1]$. In this formulation, the temporal evolution of the variable $\mathbf{y}_t$ is governed by a velocity field $\mathbf{v}_\psi$ that predicts the instantaneous direction of motion:

$$d\mathbf{y}_t = \mathbf{v}_\psi(\mathcal{I}_t)\, dt, \quad t \in [0, 1], \tag{1}$$

where $\mathbf{y}_t$ is either $\mathbf{z}_t$ or $\mathbf{x}_t$, and $\mathcal{I}_t$ denotes the information set available to the velocity predictor at time t. This domain-specific information set is configured to capture the necessary temporal and conditional dependencies for each flow.

During training, a probability path between the two endpoints is constructed using linear interpolation, providing a deterministic trajectory:

$$\mathbf{y}_t = (1 - t)\mathbf{y}_0 + t\mathbf{y}_1. \tag{2}$$

Under this linear path formulation, the ground-truth velocity field remains constant over time and is equivalent to the displacement between the endpoints, specifically $(\mathbf{y}_1 - \mathbf{y}_0)$. The velocity predictor is thus trained to regress the predicted velocity toward this target displacement:

$$\mathcal{L}_\mathbf{y} = \mathbb{E}_{t\sim\mathcal{U}[0,1]}\Big[\big\|\mathbf{v}_\psi(\mathcal{I}_t) - (\mathbf{y}_1 - \mathbf{y}_0)\big\|_2^2\Big]. \tag{3}$$

The final training objective combines the losses of the semantic and pixel flows using predefined hyperparameters:

$$\mathcal{L}_{\text{flow}} = \lambda_z\mathcal{L}_z + \lambda_x\mathcal{L}_x. \tag{4}$$

Semantic-Domain Flow. The semantic-domain flow operates in a latent space induced by a pretrained semantic encoder. Following the approach in RAE [49], the framework employs a semantic encoder $\mathcal{E}(\cdot)$ built upon a DINO backbone [37], which extracts semantic latents to define the flow endpoints. Specifically, for a visible image $\mathbf{x}_0$ and an infrared image $\mathbf{x}_1$, the semantic latents are:

$$\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0), \qquad \mathbf{z}_1 = \mathcal{E}(\mathbf{x}_1). \tag{5}$$

The encoder provides stable high-level semantic latents. To optimize the encoder for cross-modal visible-to-infrared generation, the backbone parameters are unfrozen and fine-tuned during the training process.

These endpoints capture high-level thermal priors from the visible and infrared domains, respectively. For this domain, the information set is defined as $\mathcal{I}_t = \{\mathbf{z}_t, t, \mathbf{x}_0\}$. The semantic velocity predictor is a Diffusion Transformer [34] conditioned on the source visible image: $\hat{\mathbf{v}}_t^z = \mathbf{v}_\theta^z(\mathbf{z}_t, t \mid \mathbf{x}_0)$. The Transformer outputs the semantic velocity $\hat{\mathbf{v}}_t^z$, which is then combined with the initial latent $\mathbf{z}_0$ through a lightweight fusion layer to produce the guidance signal $\mathbf{s}_t$:

$$\mathbf{s}_t = \text{Fusion}(\hat{\mathbf{v}}_t^z, \mathbf{z}_0), \tag{6}$$

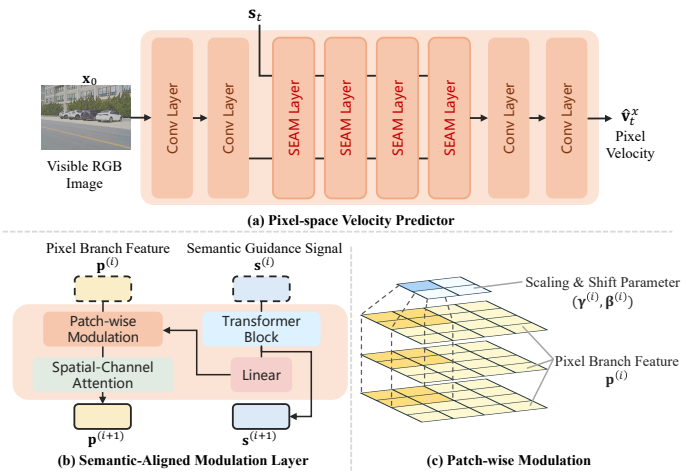

Fig. 2: Illustration of the pixel-space velocity predictor and its core component. (a) Overall architecture of $\mathbf{v}_\phi^x$ consisting of convolutional layers and SEAM Layers. (b) Internal structure of a single Semantic-Aligned Modulation Layer. (c) Details of the patch-wise modulation mechanism. For clarity, the broadcast range in (c) is simplified to 2×2; the actual implementation broadcasts to 4×4 regions.

where the Fusion module comprises a cross-attention block [40] and several linear layers. Specifically, $\mathbf{z}_0$ is linearly projected to form the query, while $\hat{\mathbf{v}}_t^z$ is projected to form the key and value in the cross-attention mechanism. The resulting signal $\mathbf{s}_t$ ensures that pixel-level synthesis remains semantically consistent with the predicted thermal distribution. During inference, at each step, the Transformer outputs the semantic velocity field, which, combined with the current $\mathbf{z}_t$, iteratively computes $\mathbf{z}_{t+1}$ as shown in Figure 1(b).

Pixel-Domain Flow. The pixel-domain flow operates directly in the high-resolution image space, with endpoints $\mathbf{x}_0$ and $\mathbf{x}_1$. For this domain, the information set is defined as $\mathcal{I}_t = \{\mathbf{x}_0, \mathbf{s}_t\}$, where the dependence on time t is effectively maintained through the time-varying guidance signal $\mathbf{s}_t$. Its velocity predictor is a pixel-space module $\mathbf{v}_\phi^x$ that estimates the pixel velocity $\hat{\mathbf{v}}_t^x$ conditioned on the semantic guidance signal $\mathbf{s}_t$ from the semantic-domain flow:

$$\hat{\mathbf{v}}_t^x = \mathbf{v}_\phi^x(\mathbf{x}_0 \mid \mathbf{s}_t). \tag{7}$$

The predictor $\mathbf{v}_\phi^x$ follows a convolutional encoder-decoder structure interleaved with Semantic-Aligned Modulation (SEAM) Layers, as illustrated in Figure 2(a). Given the input visible image $\mathbf{x}_0 \in \mathbb{R}^{B\times 3\times H\times W}$ and the guidance signal $\mathbf{s}_t$, the network first extracts spatial features through convolutional layers that progressively downsample the feature map twice, producing a bottleneck representation denoted as $\mathbf{p}^{(0)}$. The semantic guidance is initialized as $\mathbf{s}^{(0)} = \mathbf{s}_t$, and both representations are jointly processed by the subsequent SEAM Layers.

SEAM Layer. Each Semantic-Aligned Modulation (SEAM) Layer refines the interaction between the semantic guidance signal and the pixel features, as il-

lustrated in Figure 2(b). In the i-th SEAM Layer, the semantic guidance representation $\mathbf{s}^{(i)}$ first passes through a Transformer block to capture dependencies within the guidance signal. The resulting representation is then projected by a linear layer to produce scaling parameters $\boldsymbol{\gamma}^{(i)}$ and shifting parameters $\boldsymbol{\beta}^{(i)}$.

We apply a *patch-wise modulation* mechanism to inject these semantic signals into the pixel features. Specifically, the modulation parameters are broadcast to the corresponding $p \times p$ spatial regions of the pixel feature map $\mathbf{p}^{(i)}$ (with $p = 4$ in our implementation), as illustrated in Figure 2(c). The pixel features are then modulated through the affine transformation

$$\mathbf{p}'^{(i)} = (1 + \boldsymbol{\gamma}^{(i)}) \odot \mathbf{p}^{(i)} + \boldsymbol{\beta}^{(i)}.$$

Compared with global modulation (e.g., AdaIN [11]) and pixel-wise spatial modulation (e.g., SPADE [32]), the proposed patch-wise modulation provides a balanced trade-off between computational efficiency and local spatial awareness, which is particularly suitable for fine-grained detail reconstruction in the pixel-domain flow.

The modulated features are subsequently refined by a spatial-channel attention module, producing the updated feature map $\mathbf{p}^{(i+1)}$, while the updated semantic guidance representation is propagated to the next layer as $\mathbf{s}^{(i+1)}$.

3.2 Inference

During inference, starting from $\mathbf{x}_0$ and $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$, we iteratively update both states over T discrete steps. At each step t_k, we first compute $\mathbf{v}^z_{t_k}$ from the Diffusion Transformer, derive $\mathbf{s}_{t_k}$ via the fusion layer, and then obtain $\mathbf{v}^x_{t_k} = \mathbf{v}^x_\phi(\mathbf{x}_0 \mid \mathbf{s}_{t_k})$. The states are advanced by simple Euler integration: $\mathbf{z}_{t_{k+1}} = \mathbf{z}_{t_k} + \mathbf{v}^z_{t_k} \cdot \Delta t$ and $\mathbf{x}_{t_{k+1}} = \mathbf{x}_{t_k} + \mathbf{v}^x_{t_k} \cdot \Delta t$. Upon reaching $t = 1$, the final pixel state $\mathbf{x}_1$ is the generated infrared image.

3.3 Training Strategy

We train the proposed D3F-IR framework end-to-end. The dual-domain deterministic flow loss $\mathcal{L}_{\text{flow}}$ has been introduced in Section 3.1, and we utilize it as the primary supervision for the joint evolution across both domains. We further introduce additional losses to enhance training stability and quality.

Pixel Reconstruction Losses. To directly supervise the pixel-space velocity predictor $\mathbf{v}^x_\phi$, image reconstruction losses are applied to two distinct paths. For the predicted path, the guidance signal $\mathbf{s}_t$ is first computed from the predicted semantic velocity $\mathbf{v}^z_\theta(\mathbf{z}_t, t \mid \mathbf{x}_0)$ through the fusion layer. The pixel-space velocity field $\hat{\mathbf{v}}^x_t$ is then obtained as

$$\hat{\mathbf{v}}^x_t = \mathbf{v}^x_\phi(\mathbf{x}_0 \mid \mathbf{s}_t), \qquad \hat{\mathbf{x}}^{\text{pred}}_1 = \mathbf{x}_0 + \hat{\mathbf{v}}^x_t, \tag{8}$$

where $\hat{\mathbf{x}}^{\text{pred}}_1$ denotes the image obtained by applying a single deterministic update from $\mathbf{x}_0$ using the predicted pixel-space velocity.

For the ground-truth (GT) path, the guidance signal $\mathbf{s}^{\mathrm{gt}}$ is computed from the target semantic velocity $(\mathbf{z}_1 - \mathbf{z}_0)$ through the fusion layer. The corresponding pixel-space velocity field $\hat{\mathbf{v}}^x_{\mathrm{gt}}$ is obtained as

$$\hat{\mathbf{v}}^x_{\mathrm{gt}} = \mathbf{v}^x_\phi\big(\mathbf{x}_0 \mid \mathbf{s}^{\mathrm{gt}}\big), \qquad \hat{\mathbf{x}}^{\mathrm{gt}}_1 = \mathbf{x}_0 + \hat{\mathbf{v}}^x_{\mathrm{gt}}, \tag{9}$$

where $\hat{\mathbf{x}}^{\mathrm{gt}}_1$ denotes the image obtained by applying the deterministic update using the pixel-space velocity derived from the target semantic velocity. We compute the reconstruction loss [42, 48] for both $\hat{\mathbf{x}}^{\mathrm{pred}}_1$ and $\hat{\mathbf{x}}^{\mathrm{gt}}_1$,

$$\mathcal{L}_{\mathrm{rec}}(\hat{\mathbf{x}}_1, \mathbf{x}_1) = \|\hat{\mathbf{x}}_1 - \mathbf{x}_1\|_2^2 + \lambda_{\mathrm{ssim}}\big(1 - \mathrm{SSIM}(\hat{\mathbf{x}}_1, \mathbf{x}_1)\big) + \lambda_{\mathrm{lpips}}\mathcal{L}_{\mathrm{LPIPS}}(\hat{\mathbf{x}}_1, \mathbf{x}_1), \tag{10}$$

and utilize the combined term $\mathcal{L}_{\mathrm{rec}} = \mathcal{L}_{\mathrm{rec}}(\hat{\mathbf{x}}^{\mathrm{pred}}_1, \mathbf{x}_1) + \mathcal{L}_{\mathrm{rec}}(\hat{\mathbf{x}}^{\mathrm{gt}}_1, \mathbf{x}_1)$. The GT path provides a stable teacher-forcing signal for the pixel-space velocity predictor.

Overall Objective. The final objective is

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{flow}}\mathcal{L}_{\mathrm{flow}} + \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}}, \tag{11}$$

where λ_{flow} and λ_{rec} control the relative weights of the corresponding terms. All modules are optimized jointly in an end-to-end manner.

4 Experiments

In this section, we comprehensively evaluate the proposed D3F-IR framework. We first detail the experimental setup, followed by quantitative and qualitative comparisons against state-of-the-art methods. Finally, we conduct thorough ablation studies to validate our core architectural designs.

4.1 Datasets and Implementation Details

Datasets. To evaluate the effectiveness of the proposed visible-to-infrared translation framework, three widely used public benchmarks are employed: KAIST [12], FLIR [38], and M^3FD [23]. For the KAIST dataset, the data partition protocol defined in PID [28] is strictly followed, which contains 12,538 training pairs and 2,504 testing pairs. For the FLIR dataset, the preprocessing procedure introduced in DiffV2IR [35] is adopted, where image pairs with black borders in the infrared images are discarded. After filtering, the FLIR dataset contains 3,462 training pairs and 1,013 testing pairs. For the M^3FD dataset, the dataset split provided in DiffV2IR [35] is directly used, resulting in 3,550 training pairs and 640 testing pairs. All image pairs are spatially resized to 640×512 for the KAIST and FLIR benchmarks, and to 512×384 for M^3FD.

Since recent methods based on large-scale pretraining provide strong baseline performance, additional experiments are conducted on the dataset subset introduced in ThermalGen [45]. Both a pretraining-then-fine-tuning setting and a from-scratch setting are evaluated for comparison.

Table 1: Quantitative comparison on KAIST [12], FLIR [38], and M^{3}FD [23]. Best results are **bold**; second-best are underlined.

Method	KAIST [12]				FLIR [38]				M^{3}FD [23]			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
Pix2Pix [13]	21.569	0.746	0.496	130.422	16.774	0.410	0.618	180.746	16.943	0.631	0.413	170.680
CycleGAN [51]	20.248	0.692	0.519	73.565	16.056	0.390	0.602	79.876	14.652	0.590	0.382	111.853
InfraGAN [30]	12.233	0.560	0.587	86.287	17.145	0.450	0.675	91.955	16.382	0.616	0.628	85.433
EGGAN-U [18]	14.141	0.573	0.481	85.916	14.328	0.412	0.579	120.267	14.859	0.564	0.411	173.512
DR-AVIT [7]	19.453	0.714	0.495	99.642	14.036	0.372	0.637	80.667	14.270	0.601	0.390	116.469
StegoGAN [44]	17.329	0.630	0.501	69.916	14.546	0.368	0.692	137.266	13.640	0.502	0.484	209.521
UNSB [14]	19.181	0.676	0.492	67.175	15.941	0.368	0.650	84.749	12.227	0.412	0.495	132.335
ControlNet [47]	14.698	0.582	0.600	89.149	16.559	0.392	0.648	80.329	17.839	0.577	0.658	94.876
T2I-Adapter [29]	14.353	0.627	0.541	66.635	15.334	0.392	0.640	87.037	14.830	0.627	0.643	86.370
F-ViTA [31]	16.210	0.653	0.330	61.487	18.346	0.452	0.338	67.997	19.074	0.685	0.240	69.638
PID [28]	23.306	0.801	0.285	60.490	17.777	0.532	0.461	83.829	19.207	0.690	0.323	159.482
ThermalGen† [45]	23.743	0.801	0.261	**35.958**	17.823	0.490	0.430	78.652	20.573	0.744	0.214	74.326
DiffV2IR† [35]	19.986	0.733	0.270	47.405	19.596	0.472	0.309	**55.399**	20.704	0.721	0.217	**63.154**
D3F-IR (Ours)	**24.938**	0.830	0.246	75.909	**20.094**	0.541	0.360	66.716	21.530	0.768	0.213	103.264
D3F-IR (Ours)†	24.738	**0.834**	**0.219**	40.312	19.581	**0.542**	**0.268**	60.964	**22.045**	**0.780**	**0.191**	65.750

† indicates methods pre-trained on external large-scale infrared datasets.

Implementation Details. The proposed D3F-IR framework is optimized end-to-end with the AdamW [15] optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.95$, no weight decay) and a Cosine Annealing with Warm Restarts scheduler [26] (base LR $= 4 \times 10^{-4}$, $T_0 = 25$, $T_{\text{mult}} = 1$, min LR $= 4 \times 10^{-6}$). It employs the smallest ViT-based DINOv3 model (ViT-S/16). For the pretraining variant, models are first trained on a large-scale augmented infrared dataset for 80 epochs, then fine-tuned on each of the benchmark datasets (KAIST, FLIR, M^{3}FD) for 20 epochs. From-scratch versions are trained directly on these datasets for 80 epochs. All experiments are conducted on a single NVIDIA A100 GPU with a batch size of 1 and 8 steps of gradient accumulation. During inference, the deterministic flow is solved using an Euler ODE solver with 10 steps. Notably, our approach also offers hardware-overhead advantages compared to other generative methods. More details are provided in the **Supplementary Material**.

Evaluation Metrics. We systematically evaluate the generated infrared images using four widely adopted metrics. Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [42] are employed to measure pixel-level and structural fidelity. Learned Perceptual Image Patch Similarity (LPIPS) [48] is utilized to gauge perceptual realism. Additionally, we report the Fréchet Inception Distance (FID) [9]. To ensure a completely fair and unbiased distribution measurement across varying resolutions, all FID scores are computed using the CLEAN-FID [33] implementation.

4.2 Comparison with State-of-the-Art Methods

Quantitative Results. We conduct an extensive comparison of D3F-IR with 13 representative visible-to-infrared translation approaches, including 7 conventional deep learning or GAN-based models (Pix2Pix [13], CycleGAN [51], Infra-

Fig. 3: Qualitative results on benchmarks. D3F-IR achieves better structural fidelity and thermal plausibility, especially in challenging regions with unreliable visible cues.

GAN [30], EGGAN-U [18], DR-AVIT [7], StegoGAN [44], and UNSB [14]) and 6 recent generative paradigms (ControlNet [47], T2I-Adapter [29], F-ViTA [31], PID [28], ThermalGen [45], and DiffV2IR [35]). Table 1 reports the results.

As shown in Table 1, D3F-IR achieves state-of-the-art performance in pixel-level accuracy and structural fidelity. When trained from scratch on the target benchmarks, D3F-IR attains the best PSNR on all three datasets and yields competitive SSIM and LPIPS, outperforming strong diffusion-based baselines such as PID and F-ViTA.

The results further indicate that additional infrared pre-training can benefit distributional alignment. Externally pre-trained methods (e.g., ThermalGen [45] and DiffV2IR [35]) typically obtain lower FID, whereas the from-scratch setting exhibits a larger FID gap despite higher PSNR and SSIM. To quantify the impact of incorporating an extra pretraining stage before fine-tuning on each target benchmark, an additional variant of the proposed method is evaluated.

Fig. 4: Qualitative results of architecture ablation. Pixel-branch removal causes structural distortions along object boundaries; semantic-branch removal introduces visible-spectrum artifacts and unrealistic thermal patterns; dual-domain framework preserves fine-grained and thermally consistent results.

Table 2: Ablation studies of the dual-domain architecture. Semantic Branch Only and Pixel Branch Only denote configurations where one branch is removed.

Method	FLIR [38]				KAIST [12]			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
Semantic Branch Only	18.938	0.530	0.375	74.542	24.692	0.828	0.253	87.316
Pixel Branch Only	17.542	0.504	0.487	163.088	22.079	0.762	0.338	126.868
Ours	**20.094**	**0.542**	**0.360**	**66.716**	**24.938**	**0.830**	**0.246**	**75.909**

This variant substantially reduces FID across all datasets and improves SSIM and LPIPS. On FLIR, a minor PSNR decrease is observed, which is consistent with a shift in the optimization preference toward perceptual realism and distributional matching after pre-training and fine-tuning, rather than strict pixel-wise regression under the from-scratch regime. Overall, these findings suggest that the FID limitation mainly stems from limited training data and validate the proposed dual-domain framework under both from-scratch and pretraining-then-fine-tuning regimes.

Qualitative Results. Figure 3 presents qualitative comparisons on KAIST [12], FLIR [38], and M^3FD [23]. The visualization includes the visible input, the ground-truth infrared image, and representative results from CycleGAN [51], PID [28], DiffV2IR [35], ThermalGen [45], and the proposed method.

CycleGAN often produces outputs that resemble grayscale conversions of the visible inputs, which fails to reflect plausible thermal distributions and suppresses meaningful infrared contrast. PID frequently introduces noticeable image distortions, leading to warped structures and degraded spatial consistency. Although DiffV2IR and ThermalGen generate visually sharper outputs, the synthesized thermal patterns can be physically implausible, and occasional structural distortions are observed in challenging regions, *e.g.*, human bodies may appear over-smoothed and partially merged. In contrast, the proposed method produces high-fidelity infrared images with clear structures and coherent thermal semantics, exhibiting distribution characteristics that are more consistent with the ground-truth infrared data across all three benchmarks.

Table 3: Ablation studies of the dual-domain flow objectives. Each branch remains in the architecture while the corresponding flow matching loss is disabled.

Method	FLIR [38]				KAIST [12]			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
Semantic Flow Objective Only	18.938	0.530	0.375	74.542	24.692	0.828	0.253	87.316
Pixel Flow Objective Only	19.269	0.530	0.429	103.299	24.514	0.822	**0.245**	83.861
Ours	**20.094**	**0.542**	**0.360**	**66.716**	**24.938**	**0.830**	0.246	**75.909**

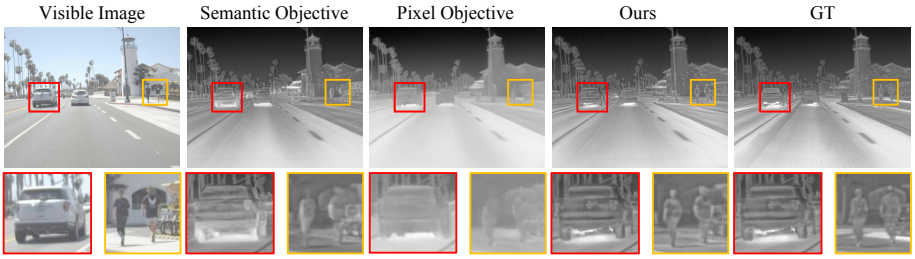

Fig. 5: Qualitative comparison of dual-domain flow objectives. The semantic-only objective introduces structural distortions, while the pixel-only objective produces thermal distributions that deviate from the ground truth.

These improvements are attributed to the joint modeling of a pixel-space flow and a semantic-space flow, which provides a favorable balance between fine-grained detail reconstruction and thermally plausible semantic generation. Moreover, several baselines are noticeably affected by artifacts inherited from the visible inputs, which can lead to incorrect thermal intensities in regions where appearance cues are unreliable. The proposed decoupled design mitigates this issue by separating the treatment of appearance-driven textures and temperature-related semantics. In the first row of Figure 3, a tree trunk covered by white paint under shadow should exhibit low infrared intensity, whereas multiple baselines incorrectly produce an overly high response; the proposed method better preserves the realistic thermal intensity in such cases. In the second-row nighttime scene, D3F-IR also produces sharper human contours with fewer blurry boundaries than competing methods.

4.3 Ablation Studies and Analysis

We conduct comprehensive ablation studies to validate the effectiveness of our core methodological designs, and then analyze the dual-domain architecture and flow matching objectives alongside the cross-domain modulation mechanism.
Ablation studies of Dual-domain Architecture. Table 2 evaluates the contribution of the dual-domain architecture by isolating the two branches. Removing the pixel branch and retaining only the semantic flow leads to consistent degradation across all metrics, indicating that semantic latent representations

Fig. 6: Visual comparisons of different generative paradigms. Standard single-domain generative models often struggle to balance global semantic transitions with local structure preservation, while our dual-domain deterministic flow matching framework mitigates these inherent limitations.

Table 4: Comparison of different generative modeling paradigms. Standard Diffusion, Flow Matching, and Deterministic Flow models operate in a single domain.

Paradigm	FLIR [38]				KAIST [12]			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
Diffusion [10]	18.871	0.480	0.484	158.517	23.622	0.800	0.265	77.282
Flow Matching [22]	18.896	0.516	0.399	152.909	22.993	0.796	0.292	104.149
Deterministic Flow [25]	18.938	0.530	0.375	74.542	24.692	0.828	0.253	87.316
Ours	**20.094**	**0.542**	**0.360**	**66.716**	**24.938**	**0.830**	**0.246**	**75.909**

alone are insufficient to reconstruct deterministic microscopic spatial textures. This deficiency also appears in Figure 4, where noticeable structural distortions emerge along object boundaries.

Conversely, removing the semantic branch and relying solely on the pixel flow results in an even larger decline in generative quality, particularly reflected by the substantial increase in FID. Without the global thermal priors provided by the semantic flow, the generated infrared images exhibit strong influence from the visible input, producing visible-spectrum artifacts and unrealistic thermal patterns. In contrast, the complete dual-domain framework preserves fine-grained spatial structures while generating thermally plausible intensity distributions that are consistent with the ground-truth infrared observations.

Ablation studies of Dual-domain Flow Objectives. Table 3 investigates the contribution of the dual-domain flow objectives while keeping the architecture unchanged. Removing the pixel-domain flow objective reduces pixel fidelity and distribution consistency, reflected by lower PSNR and higher FID scores. Because the velocity predictor loses its training signal and collapses into a conventional decoder without explicit flow supervision, the resulting configuration is structurally identical to the Semantic Branch Only setting in the architecture ablation, explaining the identical quantitative behavior. Disabling the semantic flow objective removes the deterministic transport constraint on the semantic branch, making it behave as a standard feature extractor without explicit flow supervision. Although this configuration does not introduce visible-spectrum artifacts observed in the architecture ablation, likely because the pretrained back-

Table 5: Ablation study on the semantic injection mechanisms within SEAM layers.

Injection Method	FLIR [38]				KAIST [12]			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
Concatenation	19.790	0.528	0.361	79.882	24.897	0.829	0.257	85.622
Cross-Attention [40]	19.279	0.523	0.395	114.071	23.302	0.799	0.294	112.592
Patch-wise Modulation	**20.094**	**0.542**	**0.360**	**66.716**	**24.938**	**0.830**	**0.246**	**75.909**

bone [37] provides certain high-level semantic cues, the generated thermal distributions still exhibit noticeable deviations from the ground-truth patterns. Figure 5 supports that the semantic-only objective leads to structural distortions, while the pixel-only objective produces thermal distributions inconsistent with the ground-truth infrared images. Our dual-domain objective yields the most consistent structures and thermally plausible results.

Comparison of Generative Paradigms. Table 4 compares our dual-domain approach with standard single-domain generative baselines. Standard diffusion [10] and standard flow matching [22] paradigms operate exclusively in a single domain. These models struggle to balance global semantic transitions and local structure preservation, leading to sub-optimal scores, especially in PSNR and FID. The standard deterministic flow paradigm improves pixel-level metrics, perceptual realism, and distribution consistency. However, it still falls short compared to our method, as indicated by the inferior PSNR and FID scores. Figure 6 provides qualitative evidence of these differences: diffusion produces visible-spectrum artifacts, flow matching preserves overall thermal appearance but introduces structural distortions, and deterministic flow generates rich textures that deviate from the ground-truth thermal patterns. Our dual-domain deterministic flow matching framework mitigates these limitations by jointly maintaining thermal consistency and structural fidelity. As a result, it outperforms standard single-domain paradigms in pixel fidelity, perceptual realism, and distribution consistency.

Analysis of Semantic Injection in SEAM Layers. Table 5 compares different semantic injection strategies within SEAM layers. Simple concatenation achieves competitive distortion metrics but exhibits inferior distribution alignment, reflected by higher FID scores. Standard cross-attention further degrades both perceptual and distribution consistency, likely due to spatial misalignment between semantic tokens and pixel features. In contrast, our patch-wise modulation consistently achieves the best FID and competitive LPIPS across both datasets. By broadcasting semantic guidance to strictly aligned local patches, it stabilizes global distribution statistics while preserving perceptual fidelity.

5 Conclusion

In this paper, we proposed D3F-IR, a novel visible-to-infrared translation framework that explicitly decouples thermal semantic modeling and pixel-level spatial

detail synthesis. We reformulated the translation process into two coordinated continuous flows: a semantic-domain flow to capture macroscopic thermal distributions, and a synchronized pixel-domain flow to maintain fine-grained spatial structures. To seamlessly integrate these flows, we designed a Pixel-space Velocity Predictor with SEmantic-Aligned Modulation (SEAM) Layers, establishing explicit patch-wise correspondence between the dual domains. Extensive experiments demonstrated that our approach achieved state-of-the-art pixel-level spatial fidelity with competitive thermal semantic plausibility, requiring a much smaller model size than existing generative counterparts.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China under Grant No. 62372091, in part by Natural Science Foundation of Sichuan Province of China under Grant No. 2025ZNSFSC0522, and in part by Hainan Province Key R&D Program under Grant No. ZDYF2024(LALH)001.

References

1. Brenner, M., Reyes, N.H., Susnjak, T., Barczak, A.L.: Rgb-d and thermal sensor fusion: A systematic literature review. IEEE Access **11**, 82410–82442 (2023)
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
3. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. arXiv preprint arXiv:2403.03206 (2024)
4. Feng, R., Yu, C., Deng, W., Hu, P., Wu, T.: On the guidance of flow matching. In: ICML (2025)
5. Geng, Z., Deng, M., Bai, X., Kolter, J.Z., He, K.: Mean flows for one-step generative modeling. In: NeurIPS (2025)
6. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
7. Han, Z., Zhang, S., Su, Y., Chen, X., Mei, S.: Dr-avit: Toward diverse and realistic aerial visible-to-infrared image translation. IEEE Trans. on Geoscience and Remote Sensing **62**, 1–13 (2024)
8. Haviv, D., Pooladian, A.A., Pe'er, D., Amos, B.: Wasserstein flow matching: Generative modeling over families of distributions. In: ICML (2025)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
11. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV (2017)
12. Hwang, S., Park, J., Kim, N., Choi, Y., Kweon, I.S.: Multispectral pedestrian detection: Benchmark dataset and baselines. In: CVPR (2015)

13. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
14. Kim, B., Kwon, G., Kim, K., Ye, J.C.: Unpaired image-to-image translation via neural schrödinger bridge. In: ICLR (2024)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: ICCV (2023)
17. Kniaz, V.V., Knyaz, V.A., Hladuvka, J., Kropatsch, W.G., Mizginov, V.: Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In: ECCV workshops (2018)
18. Lee, D.G., Jeon, M.H., Cho, Y., Kim, A.: Edge-guided multi-domain rgb-to-tir image translation for training vision tasks with challenging labels. In: ICRA (2023)
19. Li, B., Ma, D., He, F., Zhang, Z., Zhang, D., Li, S.: Infrared image generation based on visual state space and contrastive learning. Remote Sensing **16**(20), 3817 (2024)
20. Li, C., Liang, X., Lu, Y., Zhao, N., Tang, J.: Rgb-t object tracking: Benchmark and baseline. Pattern Recognition **96**, 106977 (2019)
21. Li, N., Wang, H., Zhao, H., Ou, W.: Cross-modal visible-to-infrared image translation in remote sensing guided by thermal features. IEEE Trans. on Geoscience and Remote Sensing **63**, 1–16 (2025)
22. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
23. Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., Luo, Z.: Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In: CVPR (2022)
24. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: ECCV (2024)
25. Liu, Z., Feng, D., Jiang, H., Zeng, L., Wang, H., Feng, C., Lei, L., Zeng, B., Liu, S.: Raw-flow: Advancing rgb-to-raw image reconstruction with deterministic latent flow matching. In: AAAI (2026)
26. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. In: ICLR (2017)
27. Ma, D., Xian, Y., Li, B., Li, S., Zhang, D.: Visible-to-infrared image translation based on an improved cgan. The Visual Computer **40**(2), 1289–1298 (2024)
28. Mao, F., Mei, J., Lu, S., Liu, F., Chen, L., Zhao, F., Hu, Y.: Pid: Physics-informed diffusion model for infrared image generation. Pattern Recognition **169**, 111816 (2026)
29. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: AAAI (2024)
30. Özkanoğlu, M.A., Ozer, S.: Infragan: A gan architecture to transfer visible images to infrared domain. Pattern Recognition Letters **155**, 69–76 (2022)
31. Paranjape, J.N., de Melo, C., Patel, V.M.: F-vita: Foundation model guided visible to thermal translation. arXiv preprint arXiv:2504.02801 (2025)
32. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR (2019)
33. Parmar, G., Zhang, R., Zhu, J.Y.: On buggy resizing libraries and surprising subtleties in fid calculation. arXiv preprint arXiv:2104.11222 (2021)

34. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
35. Ran, L., Wang, L., Wang, G., Wang, P., Zhang, Y.: Diffv2ir: visible-to-infrared diffusion model via vision-language understanding. arXiv preprint arXiv:2503.19012 (2025)
36. Shin, U., Lee, K., Kweon, I.S., Oh, J.: Complementary random masking for rgb-thermal semantic segmentation. In: ICRA (2024)
37. Siméoni, O., Vo, H.V., Seitzer, M., aldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
38. Teledyne FLIR: Teledyne flir adas thermal dataset. `https://oem.flir.com/solutions/automotive/adas-dataset-form/`, accessed 2025-11-03
39. Tuzcuoğlu, Ö., Köksal, A., Sofu, B., Kalkan, S., Alatan, A.A.: Xoftr: Cross-modal feature matching transformer. In: CVPR (2024)
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
41. Wang, S., Sun, G., Dong, L., Zheng, B.: Pas-gan: A gan based on the pyramid across-scale module for visible-infrared image transformation. Infrared Physics & Technology **139**, 105314 (2024)
42. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Trans. on Image Processing **13**(4), 600–612 (2004)
43. Wu, H., Li, X., Yang, S., Tan, H., Tang, Y., Liu, T.: Uniivft: Towards a unified framework for infrared-visible fusion and translation. In: ICASSP (2025)
44. Wu, S., Chen, Y., Mermet, S., Hurni, L., Schindler, K., Gonthier, N., Landrieu, L.: Stegogan: Leveraging steganography for non-bijective image-to-image translation. In: CVPR (2024)
45. Xiao, J., Nayak, R., Zhang, N., Tortei, D., Loianno, G.: Thermalgen: Style-disentangled flow-based generative models for rgb-to-thermal image translation. In: NeurIPS (2025)
46. Yang, L., Zhang, Z., Zhang, Z., Liu, X., Xu, M., Zhang, W., Meng, C., Ermon, S., Cui, B.: Consistency flow matching: Defining straight flows with velocity consistency. arXiv preprint arXiv:2407.02398 (2024)
47. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
48. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
49. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. In: ICLR (2026)
50. Zhou, Z., Liu, W.: An error analysis of flow matching for deep generative modeling. In: ICML (2025)
51. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)