SAMDiff-TTR: SAM2-Augmented Diffusion with Test-time Refinement for Small-Sample Infrared-to-Visible Image Generation

Lingyun Tian[1], Qiang Shen[1,2], Yang Cao[1], Dezhi Sun[1], Zilong Deng[1,2]
[1]College of Mechatronical Engineering, Beijing Institute of Technology, Beijing 100081, China
[2]National Key Laboratory of Proximity Detection and Control, Beijing 100081, China
Corresponding author: Qiang Shen (e-mail: shen2bit@163.com)

ABSTRACT Infrared-to-visible image generation provides an effective solution for augmenting visible-spectrum samples in small-sample scenarios and offers important support for cross-spectral information representation in electromagnetic sensing systems, thereby benefiting downstream visual perception tasks such as object detection in autonomous navigation and unmanned aerial vehicle (UAV) applications. However, existing methods often suffer from insufficient semantic consistency and weak generalization when only limited paired cross-modal data are available. To address this issue, we propose SAMDiff-TTR, a novel infrared-to-visible image generation framework that combines a diffusion-based generative model with semantic guidance from a pre-trained SAM2 model and adaptive test-time refinement (TTR). Specifically, a SAMSeg pre-training stage with a Multi-scale Feature Fusion Module (MFFM) is introduced to extract multi-scale semantic features from infrared images for visible-image generation. A denoising U-Net equipped with ResNet Attention (ResAtt) blocks is then employed to synthesize high-quality visible images. Furthermore, a Boundary- and Appearance-aware Test-Time Refinement (BATTR) strategy is developed to enhance robustness under limited-data conditions by jointly enforcing boundary preservation, semantic alignment, and visible-domain appearance constraints during inference without requiring paired visible references. Experiments on the DroneVehicle dataset demonstrate that SAMDiff-TTR achieves competitive image quality and consistently improves downstream object detection under small-sample settings, providing an effective cross-spectral data augmentation framework for infrared electromagnetic imaging and intelligent visual perception.

INDEX TERMS infrared-to-visible image translation, diffusion models, semantic guidance, electromagnetic imaging, cross-spectral data augmentation

I. INTRODUCTION

THERMAL infrared-to-visible image generation is of great importance for data augmentation in cross-modal visual learning, especially in scenarios where strictly aligned infrared-visible image pairs are difficult to obtain. In practical applications such as autonomous navigation, surveillance, and unmanned aerial vehicle (UAV) operations, thermal infrared (TIR) imaging provides stable target responses under varying illumination conditions and therefore serves as an important sensing modality[1, 2]. However, downstream visual perception models are still predominantly developed for the visible spectrum and rely heavily on texture, color, and fine-grained appearance cues. As a result, the lack of sufficient visible-domain samples, particularly those corresponding to infrared observations, becomes a key bottleneck for cross-modal learning and data-driven perception tasks.

This problem is especially pronounced when strictly registered infrared-visible pairs are required. In real-world acquisition, differences in sensor imaging mechanisms, viewpoints, resolutions, fields of view, temporal synchronization, and platform motion make it difficult to collect large-scale paired data with accurate pixel-level alignment. Even when infrared and visible images are captured simultaneously, residual geometric mismatch and content inconsistency are often unavoidable, which substantially increases the cost of dataset construction and limits the number of usable training samples. Under such conditions, infrared-to-visible image generation provides a practical solution for small-sample data augmentation by synthesizing semantically consistent visible images from infrared inputs, thereby alleviating the shortage of paired training data, enriching visible-domain sample diversity, and improving the generalization capabil-

ity of downstream models.

Recent advances in infrared-to-visible generation can be broadly categorized into supervised and unsupervised paradigms. Supervised methods, often built upon generative adversarial networks (GANs)[3, 4], learn the mapping from infrared images to visible counterparts using paired training data. For example, PearlGAN improves ToDayGAN[5] by introducing guided attention mechanisms to enhance semantic retention during translation. Unsupervised approaches, represented by CycleGAN[6] and MornGAN[7], alleviate the reliance on strictly paired data through cycle-consistency constraints or memory-guided translation strategies. Although these methods have shown effectiveness, GAN-based models often suffer from training instability, mode collapse, and sensitivity to hyperparameter settings, which may lead to inconsistent generation quality, especially when available training samples are limited and scene content is complex.

Diffusion models have recently emerged as a powerful alternative for generative modeling, providing superior sample quality and better training stability than conventional GAN-based frameworks. Starting from the foundational formulation of diffusion probabilistic modeling[8] and its practical development as Denoising Diffusion Probabilistic Models (DDPMs)[9], diffusion models synthesize high-fidelity images through a progressive denoising process. Their potential has been demonstrated in several cross-modal and image restoration tasks, such as SAR image coloring[10] and underwater image enhancement[11], where preserving structural information and fine details is essential. These properties make diffusion models particularly suitable for infrared-to-visible generation under limited-data settings, where stable training and strong generative fidelity are both required. However, their application to small-sample cross-modal data augmentation remains insufficiently explored, especially in scenarios where semantic consistency and downstream utility must be jointly considered.

For small-sample infrared-to-visible generation, semantic guidance is especially important. Since the generated visible images are intended not only to appear realistic but also to support downstream learning, they should preserve object boundaries, structural layouts, and category-related cues from the infrared source images. Pre-trained foundation models, such as Segment Anything Model 2 (SAM2)[12], provide a promising means of extracting robust multi-scale semantic priors to guide the generation process. In addition, models trained with limited paired data are more susceptible to domain discrepancies caused by variations in illumination, background complexity, and sensor characteristics. Under such conditions, Test-Time Refinement (TTR)[13] becomes important for improving the robustness of generated samples. Unlike conventional test-time refinement strategies that rely on paired visible references, the proposed refinement in this work is reference-free and operates only on the infrared input and the generated result, making it more suitable for practical small-sample data augmentation scenarios.

In this work, we propose SAMDiff-TTR, a novel infrared-to-visible image generation framework for small-sample data augmentation. The proposed method integrates diffusion modeling, SAM2-based semantic guidance, and test-time refinement into a unified pipeline. Specifically, a SAMSeg pre-training stage is introduced to extract robust semantic priors from infrared images, where a Multi-scale Feature Fusion Module (MFFM) is designed to enhance the representation of object boundaries and contextual structures. Based on these semantic cues, a denoising U-Net equipped with ResNet Attention (ResAtt) blocks and a Multi-Receptive Field Module (MRFM) is constructed to improve the quality and fidelity of visible-image generation. Furthermore, a Boundary- and Appearance-aware Test-Time Refinement (BATTR) strategy is developed to enhance the robustness of generated samples under limited-data conditions. By jointly imposing boundary consistency, semantic consistency, and visible-domain appearance constraints during inference, the proposed BATTR improves the structural reliability and visible-domain plausibility of generated images without requiring paired visible references. It should be emphasized that the proposed framework is designed for training-stage data augmentation under small-sample conditions, rather than real-time inference. The generated visible images are used to enhance the training dataset for downstream perception tasks.

The main contributions of this work are summarized as follows:

1. We reformulate infrared-to-visible image generation as a small-sample data augmentation problem and propose a unified SAMDiff-TTR framework to improve both generation quality and downstream utility under limited-data conditions.

2. We introduce a SAMSeg pre-training stage with a Multi-scale Feature Fusion Module (MFFM) to extract robust semantic features from infrared images, thereby strengthening structure preservation and semantic coherence in generated visible images.

3. We design a denoising U-Net with ResNet Attention (ResAtt) blocks and a Multi-Receptive Field Module (MRFM) to capture diverse contextual information and enhance the visual fidelity of the generated results.

4. We propose a Boundary- and Appearance-aware Test-Time Refinement (BATTR) strategy that performs reference-free adaptation using boundary consistency, semantic alignment, and visible-domain appearance constraints, improving the robustness and practical utility of generated samples under limited-data conditions. Experiments on the DroneVehicle dataset[14] show that the proposed method achieves strong image generation performance and improves downstream object detection under small-sample settings. These results suggest that SAMDiff-TTR can serve as an effective cross-modal data augmentation solution for robust visual perception in practical applications.

II. METHOD

A. OVERVIEW

To address the problem of infrared-to-visible image generation for small-sample data augmentation, we propose SAMDiff-TTR, a diffusion-based framework that integrates semantic guidance from a pre-trained SAM2 with an enhanced U-Net architecture. The proposed method consists of three key stages: (1) SAMSeg pre-training for semantic feature extraction, (2) diffusion-based generation for visible image synthesis, and (3) BATTR for reference-free adaptive refinement under limited-data conditions. An overview of the framework is presented in Figure 1.

In the SAMSeg pre-training stage, a frozen SAM2 encoder is employed to extract multi-scale semantic features from infrared images. Although SAM2 is pre-trained mainly on visible-domain data, it provides structural semantic priors such as object boundaries, spatial layouts, and region-level organization, which are relatively invariant across infrared and visible modalities. Therefore, the frozen encoder is used to preserve these transferable structural representations for subsequent generation. These features are further processed by the proposed MFFM and a Segmentation Decoder (Seg Decoder) to train an initial semantic segmentation model. The learned semantic representations provide crucial guidance for preserving object boundaries and structural consistency during the subsequent visible-image generation process.

In the diffusion generation stage, a Denoising U-Net enhanced with ResAtt blocks in both the encoder and decoder is used to synthesize visible images from infrared inputs. In particular, deep features derived from the pre-trained SAMSeg model are fused with intermediate features in the U-Net encoder, enabling the generation network to exploit richer semantic context and thus produce results with improved semantic coherence and enhanced visual realism. This design allows the generated images to better serve as augmented visible-domain samples in limited-data scenarios. In the BATTR stage, the model is further refined on test samples through a reference-free adaptation strategy that jointly enforces boundary consistency, semantic consistency, and visible-domain appearance constraints. This design improves the adaptability of the generation framework to domain discrepancies and appearance variations under limited-data conditions, thereby enhancing the robustness and practical utility of the generated images for downstream data augmentation.

B. SAMSEG PRE-TRAINING STAGE

The SAMSeg pre-training stage is designed to learn robust semantic priors from infrared images for subsequent infrared-to-visible image generation under small-sample conditions. Since the target task is cross-modal data augmentation rather than standalone semantic segmentation, the purpose of this stage is not merely to produce segmentation maps, but to distill multi-scale semantic representations that can provide reliable structural and category-level guidance for visible-image generation. Given an infrared image $I_{\text{inf}} \in \mathbb{R}^{C\times H\times W}$, a frozen SAM2 encoder extracts multi-scale feature representations $\{F^k_{\text{SAM}}\}^K_{k=1}$, where K denotes the number of feature scales. By freezing the encoder, the model preserves the transferable structural priors learned by SAM2 while avoiding overfitting caused by fine-tuning a large-scale encoder under limited infrared-visible paired data. This design also reduces the number of trainable parameters and enables more stable semantic feature extraction from infrared inputs.

1) Multi-scale Feature Fusion Module

The Multi-scale Feature Fusion Module (MFFM), illustrated in Figure 2, is introduced to aggregate multi-scale features $\{F^k_{\text{SAM}}\}^K_{k=1}$ extracted by the SAM2 encoder into a fused feature map $F_{\text{fuse}} \in \mathbb{R}^{C_{\text{out}}\times H\times W}$. The goal of this module is to enhance semantic richness and structural sensitivity, thereby providing stronger semantic guidance for downstream visible-image generation. Using a multi-branch convolutional architecture, the MFFM captures contextual information at multiple spatial scales from an input feature map $x \in \mathbb{R}^{C_{\text{in}}\times H\times W}$, which is beneficial for preserving object boundaries, target regions, and scene layout information in the generated visible images.

The design of different kernel sizes and dilation rates in the MFFM is motivated by the multi-scale characteristics of objects in the DroneVehicle dataset. Specifically, targets such as cars typically occupy relatively small spatial regions, while larger objects such as trucks and buses span broader areas and require wider contextual perception. To address this variation, the proposed four-branch structure progressively enlarges the receptive field through increasing dilation rates (3, 5, and 7), enabling the model to capture contextual information at multiple spatial scales. Smaller dilation rates focus on local details and object boundaries, while larger dilation rates provide broader contextual cues that are important for scene layout understanding. This multi-scale design is consistent with commonly used context aggregation strategies in semantic segmentation and helps improve both structural fidelity and robustness to scale variation in infrared-to-visible image generation.

The MFFM consists of four parallel convolutional branches, a fusion layer, and a residual connection, as detailed in Figure 2.

1) Branch 0: A 1×1 convolution is used to adjust the channel dimension from C_{in} to C_{out} while preserving fine-grained local information:

$$x_0 = \text{Conv}^{C_{\text{in}}\to C_{\text{out}}}_{1\times 1}(x). \tag{1}$$

2) Branch 1: Medium-scale contextual information is captured using factorized convolutions (1×3 and 3×1) for parameter efficiency, followed by a 3×3 dilated convolution with dilation rate 3 to enlarge the

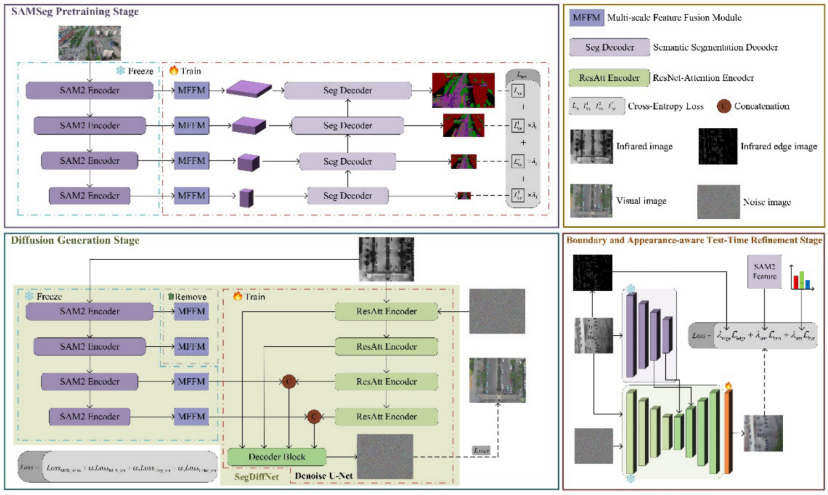

FIGURE 1. Overview of the SAMDiff-TTR framework

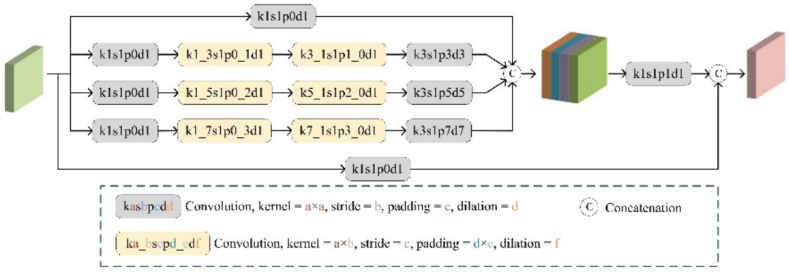

FIGURE 2. Structure of the MFFM

receptive field:

$$x_1 = \mathrm{Conv}_{3\times3,d=3}^{C_{\mathrm{out}}\to C_{\mathrm{out}}}\Big(\mathrm{Conv}_{3\times1}^{C_{\mathrm{out}}\to C_{\mathrm{out}}}\big(\mathrm{Conv}_{1\times3}^{C_{\mathrm{out}}\to C_{\mathrm{out}}}\big(\mathrm{Conv}_{1\times1}^{C_{\mathrm{in}}\to C_{\mathrm{out}}}(x)\big)\big)\Big). \tag{2}$$

3) Branch 2: A larger receptive field is obtained by using 1×5 and 5×1 convolutions, followed by a 3×3 dilated convolution with dilation rate 5:

$$x_2 = \mathrm{Conv}_{3\times3,d=5}^{C_{\mathrm{out}}\to C_{\mathrm{out}}}\Big(\mathrm{Conv}_{5\times1}^{C_{\mathrm{out}}\to C_{\mathrm{out}}}\big(\mathrm{Conv}_{1\times5}^{C_{\mathrm{out}}\to C_{\mathrm{out}}}\big(\mathrm{Conv}_{1\times1}^{C_{\mathrm{in}}\to C_{\mathrm{out}}}(x)\big)\big)\Big). \tag{3}$$

4) Branch 3: The largest-scale contextual patterns are modeled using 1×7 and 7×1 convolutions together with a 3×3 dilated convolution with dilation rate 7:

$$x_3 = \mathrm{Conv}_{3\times3,d=7}^{C_{\mathrm{out}}\to C_{\mathrm{out}}}\Big(\mathrm{Conv}_{7\times1}^{C_{\mathrm{out}}\to C_{\mathrm{out}}}\big(\mathrm{Conv}_{1\times7}^{C_{\mathrm{out}}\to C_{\mathrm{out}}}\big(\mathrm{Conv}_{1\times1}^{C_{\mathrm{in}}\to C_{\mathrm{out}}}(x)\big)\big)\Big). \tag{4}$$

The output features from the four branches x_0, x_1, x_2, and x_3, each with dimensions $C_{\mathrm{out}}\times H\times W$, are concatenated along the channel dimension:

$$x' = \mathrm{Concat}(x_0, x_1, x_2, x_3),\quad x' \in \mathbb{R}^{4C_{\mathrm{out}}\times H\times W}. \tag{5}$$

A subsequent 3×3 convolution compresses the channel dimension back to C_{out}, yielding the branch-fused feature x_{branch}:

$$x_{\mathrm{branch}} = \mathrm{Conv}_{3\times3}^{4C_{\mathrm{out}}\to C_{\mathrm{out}}}(x'). \tag{6}$$

Finally, a residual connection is introduced to preserve the original feature response and stabilize optimization. The input x is first projected to the target channel dimension by a 1×1 convolution and then added to the fused branch output. The fused output feature F_{fuse} is obtained through a ReLU activation:

$$x_{\mathrm{res}} = \mathrm{Conv}_{1\times1}^{C_{\mathrm{in}}\to C_{\mathrm{out}}}(x),\quad F_{\mathrm{fuse}} = \mathrm{ReLU}(x_{\mathrm{branch}} + x_{\mathrm{res}}) \tag{7}$$

Through this multi-scale design, the MFFM enhances feature discriminability across local details, object boundaries, and broader contextual regions, providing a stronger semantic foundation for the subsequent generation stage.

2) Semantic Segmentation Decoder

The Seg Decoder takes the fused feature map F_{fuse} produced by the MFFM and generates multi-scale segmentation predictions $\{S_{\mathrm{pred}}^1, S_{\mathrm{pred}}^2, \ldots, S_{\mathrm{pred}}^L\}$, where S_{pred}^L denotes the highest-resolution prediction and L is the number of decoder stages. Although segmentation prediction is used

during pre-training, its more important role in this work is to encourage the encoder and MFFM to learn semantically meaningful and boundary-aware features that can later guide infrared-to-visible generation.

The decoder is mainly composed of Upsampling (Up) modules and Double Convolution (DoubleConv) modules. The Up module increases spatial resolution and integrates features from corresponding encoder stages to recover finer structural details. Given an input feature $x_{\text{up}} \in \mathbb{R}^{C_{\text{up}} \times H \times W}$ and a skip-connection feature $x_{\text{skip}} \in \mathbb{R}^{C_{\text{skip}} \times 2H \times 2W}$, the Up module first applies $2\times$ bilinear upsampling to x_{up}, and then concatenates the upsampled feature with the skip-connection feature along the channel dimension:

$$\begin{aligned} x_{\text{cat}} &= \text{Concat}(x_{\text{up}} \uparrow 2, x_{\text{skip}}), \\ x_{\text{cat}} &\in \mathbb{R}^{(C_{\text{up}}+C_{\text{skip}}) \times 2H \times 2W}. \end{aligned} \tag{8}$$

The concatenated feature x_{cat} is then processed by the DoubleConv module. This module contains two sequential convolutional blocks, each composed of convolution, Batch Normalization, and ReLU activation:

$$\begin{aligned} x_{\text{mid}} &= \text{ReLU}\Big(\text{BatchNorm}\big(\\ &\qquad \text{Conv}_{3\times3}^{(C_{\text{up}}+C_{\text{skip}})\to C_{\text{mid}}}(x_{\text{cat}})\big)\Big), \\ x_{\text{out}} &= \text{ReLU}\Big(\text{BatchNorm}\big(\\ &\qquad \text{Conv}_{1\times1}^{C_{\text{mid}}\to C_{\text{out}}}(x_{\text{mid}})\big)\Big). \end{aligned} \tag{9}$$

where C_{mid} and C_{out} denote the intermediate and output channel dimensions, respectively. The 3×3 convolution captures local spatial patterns, while the 1×1 convolution performs efficient channel transformation.

Starting from the lowest-resolution fused features F_{fuse}, the Seg Decoder progressively applies Up and DoubleConv modules while incorporating skip connections from corresponding SAM2 encoder layers. Multi-scale segmentation predictions S_{pred}^{l} are generated at different resolutions $l = 1, \ldots, L$. To train this stage, a composite segmentation loss L_{seg} is employed to supervise predictions across scales:

$$L_{\text{seg}} = L_{\text{ce}}(S_{\text{pred}}^{L}, S_{\text{gt}}) + \sum_{l=1}^{L-1} \lambda_l L_{\text{ce}}(S_{\text{pred}}^{l}, S_{\text{gt}}^{l}), \tag{10}$$

where L_{ce} denotes the cross-entropy loss, S_{gt} is the ground-truth segmentation map at the original resolution, S_{gt}^{l} is the corresponding downsampled supervision for layer l, and λ_l are weighting coefficients for intermediate predictions.

By jointly optimizing the MFFM and Seg Decoder in this pre-training stage, the model learns multi-scale semantic representations that are sensitive to object boundaries and scene context. These learned semantic priors are then transferred to the diffusion generation stage, where they provide effective guidance for producing structurally consistent and semantically reliable visible images from infrared inputs, especially under limited paired-data conditions.

C. DIFFUSION GENERATION STAGE

This stage employs a Denoising Diffusion Probabilistic Model (DDPM)[9] to synthesize a target visible image $I_{\text{vis}} \in \mathbb{R}^{C \times H \times W}$ conditioned on the input infrared image $I_{\text{inf}} \in \mathbb{R}^{C \times H \times W}$. The generation process is guided by the semantic feature map $F_{\text{fuse}} \in \mathbb{R}^{C' \times H' \times W'}$ learned in the SAMSeg pre-training stage, which is fused with intermediate encoder features to improve semantic consistency and structural reliability. The goal of this stage is to generate visible-spectrum images that preserve semantic structure while maintaining perceptual plausibility that can be used as augmented samples under limited paired-data conditions.

1) Denoising Diffusion Model Formulation

The DDPM framework consists of a forward diffusion process and a reverse denoising process. In the forward process, Gaussian noise is gradually added to the target visible image, while in the reverse process the model learns to recover the clean image from noisy observations under the guidance of the infrared input I_{inf} and the semantic prior F_{fuse}. Specifically, given a clean visible image $I_{\text{vis}}^{(0)} \sim q(I_{\text{vis}})$, the forward diffusion process progressively perturbs it over T timesteps according to a predefined variance schedule $\{\beta_t\}_{t=1}^{T}$:

$$q(I_{\text{vis}}^{(t)} | I_{\text{vis}}^{(t-1)}) = \mathcal{N}\left(I_{\text{vis}}^{(t)}; \sqrt{1-\beta_t} I_{\text{vis}}^{(t-1)}, \beta_t \mathbf{I}\right). \tag{11}$$

This process permits direct sampling of $I_{\text{vis}}^{(t)}$ from $I_{\text{vis}}^{(0)}$:

$$q(I_{\text{vis}}^{(t)} | I_{\text{vis}}^{(0)}) = \mathcal{N}\left(I_{\text{vis}}^{(t)}; \sqrt{\bar{\alpha}_t} I_{\text{vis}}^{(0)}, (1-\bar{\alpha}_t)\mathbf{I}\right), \tag{12}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. As $t \to T$, the sample $I_{\text{vis}}^{(T)}$ approaches an isotropic Gaussian distribution $\mathcal{N}(0, \mathbf{I})$.

The reverse process aims to reconstruct the original visible image starting from pure noise $I_{\text{vis}}^{(T)} \sim \mathcal{N}(0, \mathbf{I})$ by iteratively sampling from the conditional distribution

$$p_\theta(I_{\text{vis}}^{(t-1)} | I_{\text{vis}}^{(t)}, I_{\text{inf}}, F_{\text{fuse}}), \tag{13}$$

which is parameterized by a denoising network θ. The network predicts the noise component at timestep t conditioned on the noisy visible image $I_{\text{vis}}^{(t)}$, the timestep t, the infrared input I_{inf}, and the semantic feature map F_{fuse}. Accordingly, the reverse transition is modeled as

$$\begin{aligned} p_\theta(I_{\text{vis}}^{(t-1)} | I_{\text{vis}}^{(t)}, I_{\text{inf}}, F_{\text{fuse}}) = \mathcal{N}\big(&I_{\text{vis}}^{(t-1)}; \\ &\mu_\theta(I_{\text{vis}}^{(t)}, t, I_{\text{inf}}, F_{\text{fuse}}), \sigma_t^2 \mathbf{I}\big), \end{aligned} \tag{14}$$

where the mean is given by

$$\begin{aligned} \mu_\theta(I_{\text{vis}}^{(t)}, t, I_{\text{inf}}, F_{\text{fuse}}) = \frac{1}{\sqrt{\alpha_t}}\Big(&I_{\text{vis}}^{(t)} \\ &- \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \theta(I_{\text{vis}}^{(t)}, t, I_{\text{inf}}, F_{\text{fuse}})\Big), \end{aligned} \tag{15}$$

and σ_t^2 is set to either β_t or $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$. The sampling step is then expressed as

$$I_{\text{vis}}^{(t-1)} = \mu_\theta(I_{\text{vis}}^{(t)}, t, I_{\text{inf}}, F_{\text{fuse}}) + \sigma_t z, \quad z \sim \mathcal{N}(0, \mathbf{I}). \tag{16}$$

By incorporating the semantic feature map F_{fuse} as an additional condition, the denoising network receives strong semantic priors beyond the raw infrared observation. This enables the model to generate visible images that better preserve object categories, structural layout, and fine details, which is especially important when the generated results are intended to serve as augmented samples for downstream learning.

2) Semantic-guided Diffusion Network

The denoising U-Net, termed Semantic-guided Diffusion Network (SegDiffNet), adopts a symmetric encoder-decoder architecture with skip connections and is designed to estimate the noise term θ. The model takes as input the noisy visible image $I_{\text{vis}}^{(t)} \in \mathbb{R}^{C\times H\times W}$, the timestep t, the infrared image $I_{\text{inf}} \in \mathbb{R}^{C\times H\times W}$, and the semantic feature map $F_{\text{fuse}} \in \mathbb{R}^{C'\times H'\times W'}$.

(a) Encoder and Decoder Modules Both the encoder and decoder adopt the ResAtt structure, which is built upon residual learning and attention enhancement to strengthen feature extraction. Compared with standard Transformer-based attention, ResAtt is more suitable for the small-sample infrared-to-visible generation setting considered in this work because its convolutional residual design provides stronger local inductive bias, lower computational complexity, and more stable optimization under limited data. Moreover, the proposed task focuses on preserving local structural details, object boundaries, and region-level semantics from infrared inputs, where convolution-based attention is more aligned with the structural-preservation objective than global dependency modeling. The encoder progressively extracts hierarchical representations through cascaded convolution and downsampling operations, reducing the spatial resolution from $H \times W$ to lower-resolution latent feature maps. The decoder mirrors this process through upsampling and convolutional reconstruction. Skip connections are introduced between corresponding encoder and decoder stages to preserve fine-grained details and stabilize feature propagation.

As shown in Figure 3, the ResAtt block processes an input feature $F \in \mathbb{R}^{B\times C\times H''\times W''}$ through two sequential MR-ResNet blocks, followed by an attention module. Each MR-ResNet block contains a 3×3 convolution, a Multi-Receptive Field Module (MRFM), Batch Normalization, ReLU activation, and a residual connection. This design allows the network to capture diverse spatial contexts while maintaining effective feature reuse. The MRFM replaces the second conventional convolution in a standard residual block and uses parallel convolutional branches with different kernel sizes (3×3, 5×5, and 7×7) and dilation rates (1, 2, and 3) to enlarge the receptive-field diversity:

$$\begin{aligned} F_{\text{conv}} &= \text{ReLU}(\text{BatchNorm}(\text{Conv}_{3\times3}(F))), \\ F_{\text{MRFM}} &= \text{MRFM}(F_{\text{conv}}), \\ F_{\text{res}} &= \text{ReLU}(\text{BatchNorm}(F_{\text{MRFM}})) + F. \end{aligned} \tag{17}$$

The resulting intermediate features are then processed by the attention module to produce the final feature map F_{att}. Through the combination of residual learning, multi-receptive-field aggregation, and attention-based enhancement, the ResAtt structure improves semantic expressiveness and structural sensitivity while avoiding the heavier data and computation requirements of Transformer-based attention, thereby facilitating more stable denoising and higher-fidelity visible-image generation under the small-sample setting.

(b) Timestep Embedding The diffusion timestep t is embedded using sinusoidal positional encoding:

$$\begin{aligned} \text{Emb}(t) = \text{concat}(&\sin(2\pi t \cdot \omega_1), \cos(2\pi t \cdot \omega_1), \ldots, \\ &\sin(2\pi t \cdot \omega_d), \cos(2\pi t \cdot \omega_d)). \end{aligned} \tag{18}$$

The resulting embedding vector $\text{Emb}(t) \in \mathbb{R}^d$ is projected through a fully connected layer and injected into the network to provide timestep-dependent conditioning for denoising.

(c) Semantic Feature Guidance The semantic feature map $F_{\text{fuse}} \in \mathbb{R}^{C'\times H'\times W'}$ is extracted from deep layers of the SAMSeg network, where semantic abstraction is relatively strong. During diffusion generation, F_{fuse} is concatenated with the encoder feature map $F_{\text{enc}} \in \mathbb{R}^{C_{\text{enc}}\times H'\times W'}$ along the channel dimension to obtain

$$F_{\text{concat}} \in \mathbb{R}^{(C'+C_{\text{enc}})\times H'\times W'}. \tag{19}$$

A 1×1 convolution is then used to project the concatenated feature back to the encoder channel dimension:

$$F_{\text{fused}} = \text{Conv}_{1\times1}(F_{\text{concat}}). \tag{20}$$

The fused representation F_{fused} is subsequently fed into the decoder. This mechanism injects semantic priors learned from infrared images into the denoising process, enabling the generation network to better preserve object boundaries, semantic categories, and scene structure in the synthesized visible images.

3) Loss Function

The diffusion generation stage is trained by minimizing the following composite objective:

$$\begin{aligned} L_{\text{total}} = L_{\text{MSE\_noise}} &+ \alpha_1 L_{\text{MSE\_vis}} \\ &+ \alpha_2 L_{\text{Edge\_vis}} + \alpha_3 L_{\text{Color\_vis}}, \end{aligned} \tag{21}$$

where $L_{\text{MSE\_noise}}$ is the mean squared error loss for noise prediction, which supervises the denoising network to estimate the injected Gaussian noise accurately; $L_{\text{MSE\_vis}}$ is the reconstruction loss between the generated visible image and the corresponding target visible image; $L_{\text{Edge\_vis}}$ is an edge preservation loss that penalizes structural discrepancies between generated and target visible images; and $L_{\text{Color\_vis}}$ is a color consistency loss computed in the Lab color

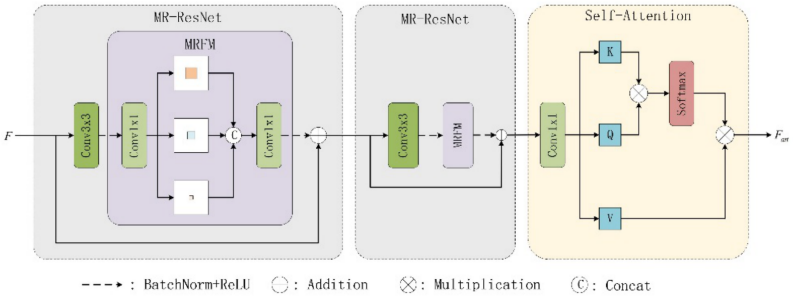

FIGURE 3. Illustration of the ResAtt structure

space to encourage realistic visible-domain appearance. The weighting coefficients are set to $\alpha_1 = 0.3$, $\alpha_2 = 0.3$, and $\alpha_3 = 0.2$. This loss design jointly supervises generation quality from the perspectives of diffusion denoising accuracy, pixel-level reconstruction, structural preservation, and visible-domain appearance consistency. As a result, the model can generate visible images that are not only faithful to the infrared source in semantic structure, but also visually suitable for use as augmented samples in downstream small-sample learning tasks.

D. BOUNDARY AND APPEARANCE-AWARE TEST-TIME REFINEMENT

To further improve the robustness of infrared-to-visible image generation under limited-data conditions, we introduce a reference-free adaptive optimization strategy at test time, termed Boundary- and Appearance-aware Test-Time Refinement (BATTR). Unlike the original setting of brightness-guided refinement with visible references, the proposed BATTR is designed for small-sample data augmentation scenarios where no paired visible image is available during testing. Its objective is to adapt the generation model to the input infrared sample by enforcing cross-modal boundary consistency, semantic consistency, and visible-domain appearance plausibility.

In this study, after the decoder modules of SegDiffNet produce the decoded feature map, the resulting feature representation is passed through an MLP layer to generate the final visible image $I_{\text{vis\_pred}} \in \mathbb{R}^{C\times H\times W}$. During test-time refinement, we freeze the parameters of the encoder and decoder modules and update only the MLP layer, thereby maintaining model stability while enabling lightweight sample-wise adaptation. Specifically, given an infrared input image $I_{\text{inf}} \in \mathbb{R}^{C\times H\times W}$, the encoder and decoder first extract the decoded feature map $F_{\text{dec}} \in \mathbb{R}^{C'\times H\times W}$, and the visible prediction $I_{\text{vis\_pred}}$ is obtained as

$$I_{\text{vis\_pred}} = \text{MLP}(F_{\text{dec}}). \tag{22}$$

To refine the generated result without requiring a paired visible reference, we design the following test-time objective:

$$L_{\text{BATTR}} = \lambda_{\text{edge}} L_{\text{edge}} + \lambda_{\text{sem}} L_{\text{sem}} + \lambda_{\text{stat}} L_{\text{stat}}, \tag{23}$$

where λ_{edge}, λ_{sem}, and λ_{stat} are the weighting coefficients for the three loss terms.

The first term, L_{edge}, is a boundary consistency loss that preserves the structural information of the infrared input in the generated visible image. Since thermal infrared images provide stable target contours and salient structural cues, we extract the edge maps of both the infrared image and the generated visible image using the Sobel operator. Let E_{inf} denote the edge map of I_{inf}, and let E_{vis} denote the edge map of the grayscale version of $I_{\text{vis\_pred}}$. The boundary consistency loss is defined as

$$L_{\text{edge}} = \|E_{\text{inf}} - E_{\text{vis}}\|_1. \tag{24}$$

This loss encourages the generated visible image to preserve object boundaries and spatial layouts inherited from the infrared source image.

The second term, L_{sem}, is a semantic consistency loss that constrains the generated visible image to remain semantically aligned with the infrared input. To this end, we reuse the semantic feature extractor obtained from the SAMSeg pre-training stage and keep it frozen during refinement. Let $\phi_l(\cdot)$ denote the semantic feature map extracted from the l-th selected layer. Then, the semantic consistency loss is formulated as

$$L_{\text{sem}} = \sum_{l=1}^{L} \|\phi_l(I_{\text{inf}}) - \phi_l(I_{\text{vis\_pred}})\|_1, \tag{25}$$

where L is the number of selected semantic feature levels. This term helps preserve high-level scene structure and category-related cues during the refinement process.

The third term, L_{stat}, is an appearance statistics loss that regularizes the generated image toward the visible-domain distribution. Because no paired visible target is available at test time, we approximate the visible-domain prior using channel-wise first- and second-order statistics precomputed from the visible training images. For each channel $c \in \{r, g, b\}$, let $\mu(I^c_{\text{vis\_pred}})$ and $\sigma(I^c_{\text{vis\_pred}})$ denote the mean and standard deviation of the generated image, and let μ^c_{train} and σ^c_{train} denote the corresponding statistics

of the visible training set. The appearance statistics loss is defined as

$$L_{\text{stat}} = \sum_{c\in\{r,g,b\}} \left(\|\mu(I^c_{\text{vis\_pred}}) - \mu^c_{\text{train}}\|_1 + \|\sigma(I^c_{\text{vis\_pred}}) - \sigma^c_{\text{train}}\|_1\right). \tag{26}$$

This term suppresses unrealistic color shifts and contrast deviations, and encourages the generated result to exhibit a more plausible visible-spectrum appearance. By optimizing the MLP layer with the above three constraints, BATTR performs lightweight reference-free refinement for each test sample. The boundary consistency term preserves structural details from the infrared input, the semantic consistency term maintains high-level cross-modal alignment, and the appearance statistics term constrains the generated image to remain close to the visible-domain distribution. As a result, the proposed BATTR improves the quality and reliability of generated visible images under small-sample conditions, making them more suitable for downstream data augmentation and visual perception tasks.

III. RESULTS AND DISCUSSION

A. DATASETS AND EVALUATION METRICS

(a) Dataset To validate the effectiveness of SAMDiff-TTR for infrared-to-visible image generation in small-sample data augmentation, we conduct all experiments on the DroneVehicle dataset. DroneVehicle contains 26,969 aligned infrared-visible image pairs captured by drones and provides object annotations for vehicle detection, including car, truck, bus, van, and freight car. Following the official training split, we used the training set to construct a small-sample augmentation setting. Specifically, 30% of the original training pairs are randomly selected as the small-sample subset, while the remaining 70% of the training set is reserved as the augmentation source set. The 30% subset is used to train the proposed infrared-to-visible generation model. After training, the infrared images in the remaining 70% training set are fed into the trained generation model to synthesize their corresponding visible images, which are then used as augmented visible-domain samples. For downstream object detection, three training settings are considered: 1) training with only the 30% real visible data as the small-sample baseline, 2) training with the 30% real visible data together with the generated visible images synthesized from the remaining 70% infrared images, and 3) training with the full original visible training set as the upper-bound reference. All images were resized to 256×256 using bilinear interpolation for image generation experiments.

(b) Evaluation Metrics We evaluate the proposed method from both image generation quality and downstream augmentation effectiveness. For image generation quality, we adopt three widely used metrics: Structural Similarity Index (SSIM)[15], Learned Perceptual Image Patch Similarity (LPIPS)[16], and Fréchet Inception Distance (FID)[17]. SSIM measures the structural similarity between the generated visible image and the corresponding real visible image, where higher values indicate better structural fidelity. LPIPS evaluates perceptual similarity, where lower values indicate that the generated result is perceptually closer to the real visible image. FID measures the distribution discrepancy between generated images and real visible images, with lower values indicating better visible-domain alignment. To assess the utility of generated images for data augmentation, we further evaluate downstream object detection performance using Average Precision (AP) and mean Average Precision (mAP)[18]. Higher mAP values indicate that the generated samples provide more effective support for detector training under limited-data conditions. In addition, qualitative analysis is provided through visualization results such as feature distribution plots and response heatmaps.

B. EXPERIMENTAL SETUP AND IMPLEMENTATION DETAILS

We compare the proposed SAMDiff-TTR with several representative image translation baselines, including GAN-based methods (e.g., Pix2Pix[19], CycleGAN, and PearlGAN) and diffusion-based methods (e.g., T2V-DDPM[20] and WeatherDiffusion[21]). For fair comparison, all competing methods are trained under the same small-sample protocol, namely using the same 30% subset of DroneVehicle training pairs for model training and generating visible images for the infrared inputs in the remaining 70% training set.

SAMDiff-TTR was trained on an NVIDIA RTX A5000 GPU using the Adam optimizer[22], with $\beta_1 = 0.9$, $\beta_2 = 0.999$, a batch size of 4, and a fixed learning rate of 0.0002. The diffusion process is configured with $T = 1000$ timesteps following the standard DDPM setting. The generation model is trained for 200 epochs on the selected 30% training subset, and horizontal flipping was applied as data augmentation during training. After training, the model was used to generate visible images for the infrared samples in the remaining 70% training set, thereby constructing an augmented visible-domain dataset for downstream detection.

For downstream object detection, we adopt MCMF[23] as the detector. To verify the effectiveness of the proposed data augmentation strategy, MCMF is trained under three settings: 1) using only the 30% real visible-infrared training pairs as the small-sample baseline, 2) using the 30% real visible-infrared training pairs together with additional visible-infrared pairs formed by combining the remaining 70% real infrared images with their synthesized visible counterparts, and 3) using the full real visible-infrared training set as the upper-bound reference. All detector settings use the same training configuration to ensure fair comparison. Detection performance is finally evaluated on the original test set of DroneVehicle using AP and mAP.

C. EXPERIMENTS ON THE DRONEVEHICLE DATASET

1) Image Generation Quality

The qualitative generation results on the DroneVehicle dataset are shown in Figure 4. Column (a) shows the input infrared image, and column (b) shows the corresponding

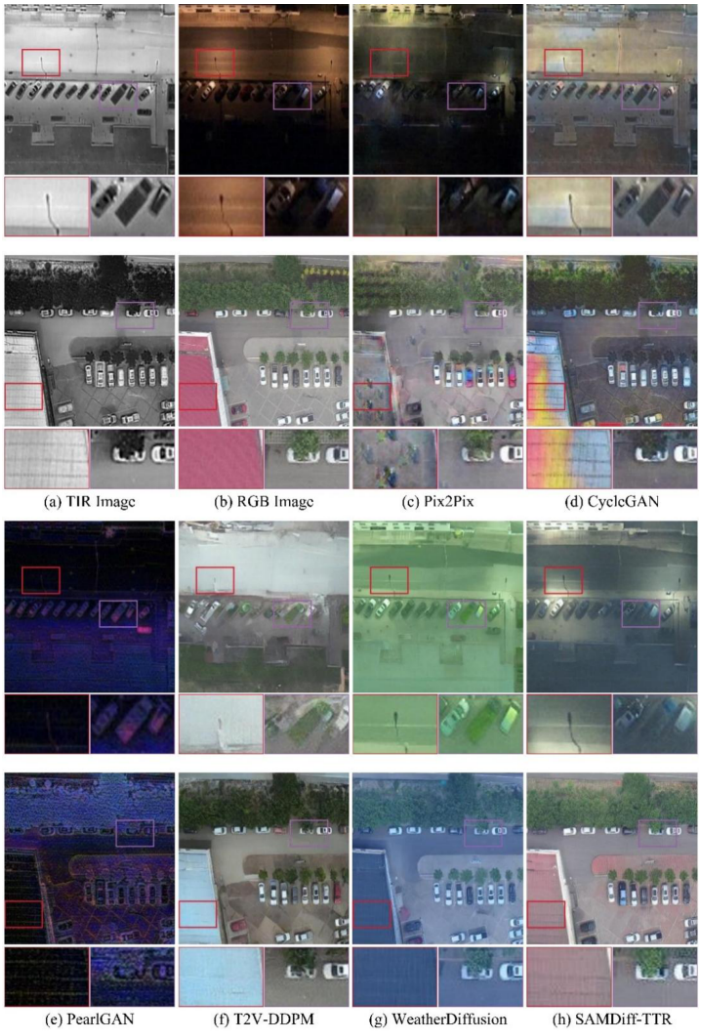

FIGURE 4. Visual comparison of generated results produced by different methods on the DroneVehicle dataset

real visible image. Compared with the competing methods, the images generated by SAMDiff-TTR exhibit clearer object boundaries, more coherent structural layouts, and more realistic visible-spectrum appearance. In particular, in challenging cases involving small objects, partial occlusion, and cluttered backgrounds, SAMDiff-TTR better preserves the semantic structure of vehicles and surrounding scene elements, indicating that the semantic priors introduced by the SAMSeg stage provide effective guidance for infrared-to-visible generation under limited-data conditions.

To quantitatively evaluate image generation quality on the DroneVehicle dataset, we adopt three widely used metrics, namely SSIM, LPIPS, and FID. The corresponding results are reported in Table 1.

As shown in Table 1, SAMDiff-TTR achieves the best overall performance among all compared methods, obtaining the highest SSIM of 0.5201, the lowest LPIPS of 0.3448, and the lowest FID of 35.87. Compared with Pix2Pix, SAMDiff-TTR improves SSIM from 0.4907 to 0.5201, reduces LPIPS from 0.3622 to 0.3448, and significantly lowers FID from 70.18 to 35.87. Compared with the diffusion-based baseline T2V-DDPM, it improves SSIM from 0.3485 to 0.5201, reduces LPIPS from 0.4625 to 0.3448, and decreases FID from 58.39 to 35.87, demonstrating clear advantages in structural fidelity, perceptual quality, and visible-domain distribution alignment.

Compared with WeatherDiffusion, which is the strongest competitor, SAMDiff-TTR further improves SSIM from 0.4295 to 0.5201, reduces LPIPS from 0.4685 to 0.3448, and slightly lowers FID from 37.66 to 35.87. These results indicate that the proposed semantic-guided diffusion framework can achieve a better balance between structural consistency and appearance realism under limited-data conditions. This advantage can be attributed to the semantic priors introduced by the SAMSeg pre-training stage, the multi-scale feature aggregation mechanism, and the reference-free refinement strategy, which jointly enhance the semantic reliability and visual plausibility of the generated images. Although SAMDiff-TTR exhibits a relatively higher inference time compared to lightweight GAN-based models, it is important to note that the proposed method is designed for training-stage data augmentation rather than real-time deployment.

TABLE 1. Image quality assessment and training/inference time on the DroneVehicle dataset

Method	SSIM ↑	LPIPS ↓	FID ↓	Training Time (h)	Inference Time (s)
Pix2Pix	0.4907	0.3622	70.18	5.8	0.06
CycleGAN	0.3165	0.5953	82.44	113.8	0.08
PearlGAN	0.1547	0.6401	114.30	23.4	0.02
T2V-DDPM	0.3485	0.4625	58.39	500	3.31
WeatherDiffusion	0.4295	0.4685	37.66	126.3	1.21
SAMDiff-TTR	0.5201	0.3448	35.87	141.2	1.29

In practical applications, the generated visible images are used to enrich training datasets and improve the performance of downstream perception models. Therefore, the inference latency does not affect the real-time performance of UAV systems. Instead, higher generation quality and semantic consistency are prioritized to achieve more effective training data augmentation.

Overall, the qualitative and quantitative results demonstrate that SAMDiff-TTR can generate high-quality visible images with reliable semantic structure and realistic appearance from infrared inputs under limited-data conditions, providing effective support for subsequent data augmentation and downstream detection.

2) Object Detection

To evaluate the effectiveness of the generated images for downstream data augmentation, we further compare object detection performance on the DroneVehicle dataset. Following the protocol described in the experimental setup, MCMF is trained under three different data conditions: 1) using only the 30% real visible-infrared training pairs as the small-sample baseline, 2) using the 30% real training pairs together with additional generated visible-infrared pairs produced by different generation methods, and 3) using the full real training set as the upper-bound reference. The corresponding AP and mAP results are reported in Table 2.

As shown in Table 2, training the detector with only the 30% real small-sample dataset yields an mAP of 61.10, which is substantially lower than the 71.40 achieved by using the full real dataset. This gap confirms that limited visible-infrared training data significantly constrains downstream detection performance. After introducing generated visible samples for data augmentation, all generation-based methods improve the detector over the small-sample baseline, demonstrating the effectiveness of infrared-to-visible generation for compensating for missing visible-domain data. Among all compared methods, SAMDiff-TTR achieves the best overall detection performance, reaching an mAP of 70.03. This result exceeds the small-sample baseline by 8.93 percentage points and outperforms the strongest comparison method, WeatherDiffusion, by 0.86 percentage points. More importantly, the performance gap between SAMDiff-TTR and the full real dataset is reduced to only 1.37 percentage points, indicating that the generated samples produced by SAMDiff-TTR can effectively narrow the performance loss caused by limited training data.

From the class-wise AP results, SAMDiff-TTR achieved the best performance in all five categories, including Car (87.79), Truck (55.52), Bus (67.19), Van (88.17), and Freight car (51.49). The gains are especially notable in the Truck and Freight car categories, where the small-sample baseline is relatively weak. Compared with the small-sample baseline, SAMDiff-TTR improves AP by 12.59 points for Truck and 14.15 points for Freight car, indicating that the proposed method is particularly effective in enhancing the representation of relatively difficult or underrepresented categories. These results suggest that the generated images preserve not only global scene structure but also category-relevant fine-grained information beneficial for detector training.

The t-SNE visualization in Figure 5 provides further evidence for the effectiveness of the proposed method. Compared with competing methods, the features extracted from samples generated by SAMDiff-TTR exhibit more compact intra-class distributions and clearer inter-class separation. This indicates that the generated visible images are more consistent with the semantic structure of real visible-domain samples, enabling the detector to learn more discriminative feature representations. In particular, the reduced feature dispersion suggests that SAMDiff-TTR better preserves category identity and structural cues during generation, which is consistent with its superior AP and mAP performance in downstream detection.

Overall, the detection results demonstrate that SAMDiff-TTR is not only effective in generating high-quality visible images, but more importantly capable of producing augmented samples that substantially improve downstream detection under small-sample conditions. The proposed method consistently outperforms existing baselines and approaches the performance of training with the full real dataset, which verifies its practical value for infrared-visible data augmentation.

D. ABLATION STUDIES

1) Effect of Semantic Guidance and Composite Objectives

To systematically analyze the contribution of each major component in the proposed SAMDiff-TTR framework to image generation quality and downstream detection performance, we conduct progressive ablation experiments on the DroneVehicle dataset under the small-sample data augmentation setting. The ablation study follows a cumulative design, in which modules are introduced step by step so that the individual and joint effects of different components can be clearly quantified and interpreted.

TABLE 2. Object detection performance comparison of different methods on the DroneVehicle dataset

Method	Car	Truck	Bus	Van	Freight car	mAP
Small-sample dataset	85.22	42.93	54.62	85.39	37.34	61.10
Pix2Pix	86.87	53.04	66.94	87.87	48.82	68.71
CycleGAN	86.94	52.40	65.93	87.92	49.38	68.51
PearlGAN	86.76	52.90	67.14	88.13	49.86	68.95
T2V-DDPM	87.44	51.72	65.17	87.81	48.10	68.07
WeatherDiffusion	87.70	52.99	67.10	87.86	50.18	69.17
SAMDiff-TTR	87.79	55.52	67.19	88.17	51.49	70.03
Full real dataset	88.08	57.14	69.75	88.41	53.62	71.40

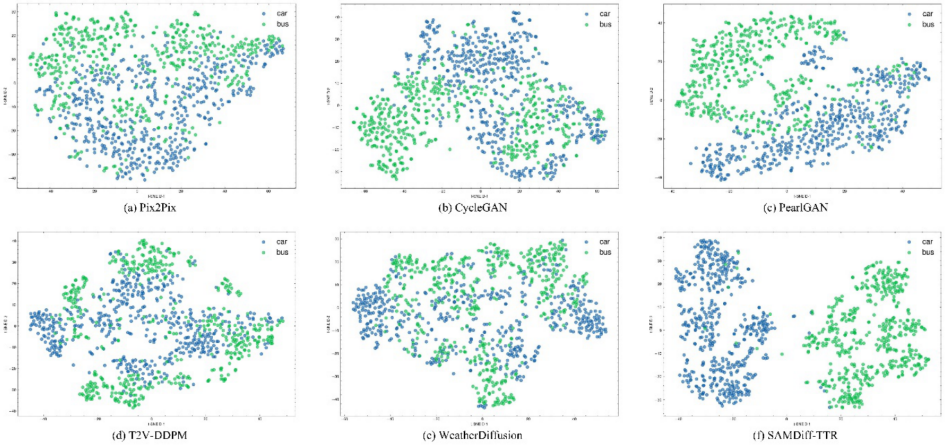

FIGURE 5. The t-SNE visualization of feature distributions

In this study, WeatherDiffusion is adopted as the baseline generation framework, and the model is first trained using only the noise prediction loss $L_{\text{MSE\_noise}}$. On this basis, we progressively introduce: 1) semantic feature guidance extracted from SAM2 (denoted as +SAM); 2) the proposed composite training objective L_{total} (denoted as +Loss); and 3) the proposed Boundary- and Appearance-aware Test-Time Refinement module (denoted as +BATTR). It should be noted that the MFFM serves as the necessary bridge between SAM2 semantic features and the diffusion backbone. Removing it would break the semantic injection pathway. Therefore, MFFM is retained as a fixed component throughout this group of ablation experiments.

TABLE 3. Ablation results on the DroneVehicle dataset.

Baseline	SAM	Loss	BATTR	mAP	SSIM ↑	FID ↓
✓				69.17	0.4295	37.66
✓	✓			69.77	0.4813	36.54
✓	✓	✓		69.91	0.5049	36.12
✓	✓	✓	✓	70.03	0.5201	35.87

Table 3 presents the progressive ablation results on the DroneVehicle dataset. The baseline model, trained with only the noise prediction objective, achieves an mAP of 69.17, an SSIM of 0.4295, and an FID of 37.66, indicating limited structural fidelity and visible-domain alignment under the small-sample setting. After introducing SAM2 semantic guidance (+SAM), mAP increases to 69.77, SSIM improves to 0.4813, and FID decreases to 36.54, showing that semantic priors effectively improve both generation quality and downstream utility. With the further introduction of the composite objective (+Loss), the model continues to improve, reaching an mAP of 69.91, an SSIM of 0.5049, and an FID of 36.12. After additionally incorporating BATTR, the full model achieves the best performance on all metrics, with an mAP of 70.03, an SSIM of 0.5201, and an FID of 35.87. These results show that SAM guidance, the composite loss, and BATTR provide complementary benefits and jointly improve the quality and downstream usefulness of the generated samples.

2) Analysis of the Multi-scale Feature Fusion Module

To verify the practical contribution of MFFM to semantic consistency and structural preservation, we further conduct comparative ablation experiments on different feature fusion structures using the DroneVehicle dataset. Three representative existing fusion modules are selected for comparison, including CAFM[24], GAU[25], and HCF[26].

On this basis, three variants of the proposed MFFM are further designed: 1) MFFM-T, a lightweight version that retains only Branch 0 and Branch 1 to reduce computational complexity; 2) MFFM-S, the standard version with the four-branch multi-scale structure proposed in this work; and 3) MFFM-M, an extended version that introduces additional branches and larger 9×9 convolution kernels to enhance contextual modeling capacity. All methods are evaluated under the same training and testing settings. We use mAP to measure the effectiveness of generated samples for down-

stream detection, and SSIM and FID to evaluate structural fidelity and visible-domain distribution consistency.

As shown in Table 4, the proposed MFFM variants consistently outperform the compared fusion modules in terms of overall generation quality and downstream augmentation effectiveness. Among the baseline fusion modules, HCF achieves the highest SSIM of 0.5198, while GAU obtains the lowest FID of 39.34. However, all three compared methods remain inferior to the proposed MFFM-based designs in terms of either detection performance or overall metric balance. Among the MFFM variants, the standard MFFM-S achieves an mAP of 70.03, an SSIM of 0.5201, and an FID of 35.87, outperforming CAFM, GAU, and HCF on all three metrics. This indicates that the proposed multi-branch multi-scale fusion design is more effective in preserving semantic structure and improving visible-domain alignment for infrared-to-visible generation under the small-sample setting.

TABLE 4. Quantitative comparison of different feature fusion modules on the DroneVehicle dataset

Model	mAP	SSIM ↑	FID ↓
CAFM	69.72	0.5017	42.21
GAU	69.75	0.5043	39.34
HCF	69.54	0.5198	39.76
MFFM-T	69.99	0.5190	37.33
MFFM-S	70.03	0.5201	35.87
MFFM-M	70.07	0.5319	35.79

Comparing different MFFM variants further reveals the trade-off between model complexity and performance. The lightweight version MFFM-T achieves slightly lower performance than MFFM-S, with an mAP of 69.99, an SSIM of 0.5190, and an FID of 37.33, showing that reducing branch capacity weakens multi-scale semantic modeling. The extended version MFFM-M achieves the best numerical results, reaching an mAP of 70.07, an SSIM of 0.5319, and an FID of 35.79. However, its gains over MFFM-S are marginal, suggesting that the added complexity does not bring proportionally significant improvement. Overall, these results show that the proposed MFFM design is more suitable for semantic injection in diffusion-based cross-modal generation than conventional fusion modules. Considering both performance and efficiency, MFFM-S is selected as the default fusion structure in the proposed framework.

3) Analysis of Freezing and Fine-tuning Strategies for the SAM2 Encoder

To empirically justify the use of a frozen SAM2 encoder in the proposed framework, we compare different SAM2 optimization strategies on the DroneVehicle dataset under the same small-sample protocol. The goal is to compare how the trainable range of SAM2 parameters affects generation quality, downstream detection performance, and training stability.

In this set of experiments, three different training strategies are considered for the SAM2 encoder: 1) full freezing, where all SAM2 encoder parameters remain fixed during training, which is also the default configuration in this work; 2) partial fine-tuning, where only the last two Transformer blocks of the SAM2 encoder are updated; and 3) full fine-tuning, where all SAM2 encoder parameters participate in backpropagation. Except for the trainable scope of the SAM2 encoder, all other components, including the MFFM, diffusion generation backbone, and BATTR module, remain unchanged to ensure fair comparison.

For the partial and full fine-tuning settings, a relatively small learning rate is adopted for SAM2 to avoid unstable optimization caused by overly large parameter updates. In addition to standard generation and detection metrics, we further introduce the standard deviation of the late-stage training loss as an indicator of training stability. A smaller value indicates smoother convergence and more stable optimization behavior.

TABLE 5. Ablation results of different SAM2 optimization strategies

SAM2 strategy	FID ↓	SSIM ↑	mAP ↑	Std. ↓
Full freezing	35.87	0.5201	70.03	0.012
Partial fine-tuning	35.20	0.5081	70.00	0.031
Full fine-tuning	35.96	0.5004	69.91	0.084

As shown in Table 5, the full-freezing strategy achieves the best overall balance between generation quality, downstream detection performance, and training stability. It obtains the highest SSIM of 0.5201 and the best mAP of 70.03, while also yielding the lowest training-loss standard deviation of 0.012. These results indicate that freezing the SAM2 encoder provides the most stable optimization behavior and the strongest overall effectiveness under the small-sample setting.

Partial fine-tuning achieves the best FID of 35.20, slightly better than full freezing. However, its SSIM and mAP decrease to 0.5081 and 70.00, respectively, and the training-loss fluctuation increases to 0.031. This suggests that updating a small portion of the SAM2 encoder may slightly improve visible-domain distribution alignment, but does not consistently benefit structural fidelity, downstream utility, or optimization stability.

Full fine-tuning shows a further decline in performance. Its SSIM drops to 0.5004, mAP decreases to 69.91, and the training-loss standard deviation rises sharply to 0.084, indicating substantially worse training stability. This result suggests that, under limited-data conditions, full-parameter fine-tuning of a large pre-trained model such as SAM2 is prone to overfitting and may weaken its transferable semantic priors.

Overall, the results indicate that the full-freezing strategy achieves the best balance between generation quality, downstream performance, and training stability. Although partial fine-tuning yields a slightly better FID, this mainly reflects

improved global distribution alignment rather than structural fidelity, which is more critical for downstream tasks as captured by SSIM and mAP. The inferior performance and higher instability of full fine-tuning further suggest that updating all SAM2 parameters under limited data leads to overfitting and degradation of transferable priors. This can be attributed to the fact that SAM2 encodes modality-invariant structural information, which is better preserved by freezing under the small-sample setting. Therefore, the full-freezing strategy is adopted in all experiments.

IV. CONCLUSION

In this paper, we proposed SAMDiff-TTR, a novel framework for infrared-to-visible image generation in small-sample data augmentation scenarios. By integrating SAM2-based semantic guidance, diffusion-based generation, and reference-free test-time refinement, the proposed method improves both the quality of generated visible images and their effectiveness for downstream detection under limited paired-data conditions. Experiments on the DroneVehicle dataset demonstrate that SAMDiff-TTR achieves strong performance in image generation quality and consistently improves downstream object detection in the small-sample setting. The results show that the proposed method can effectively narrow the performance gap between limited real-data training and full-data training. Ablation studies further verify the effectiveness of the major components, including SAM-based semantic guidance, the composite loss, BATTR, and the MFFM design. In addition, freezing the SAM2 encoder is shown to provide the best balance between generation quality, downstream performance, and training stability. Overall, SAMDiff-TTR provides an effective solution for using infrared-to-visible generation to augment visible-domain samples when strictly aligned paired data are limited. Future work may explore accelerated diffusion strategies to enable real-time deployment.

AUTHOR CONTRIBUTIONS

Conceptualization, Lingyun Tian, Qiang Shen and Yang Cao; methodology, Lingyun Tian and Qiang Shen; formal analysis, Lingyun Tian and Qiang Shen; writing—original draft preparation, Lingyun Tian; writing—review and editing, Lingyun Tian, Qiang Shen and Zilong Deng; visualization, Lingyun Tian and Dezhi Sun; supervision, Qiang Shen. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research was funded by the National Natural Science Foundation of China, grant number 62403055.

REFERENCES

[1] Y. Xi, Y. Zhang, Y. Jiang, et al., "Infrared Small Target Detection Based on Entropy Variation Weighted Local Contrast Measure [J]," Remote Sensing, vol. 17, no. 8, 2025.

[2] C. Zhao, W. Wang, Y. Yan, et al., "A Re-Identification Framework for Visible and Thermal-Infrared Aerial Remote Sensing Images with Large Differences of Elevation Angles [J]," Remote Sensing, vol. 17, no. 11, 2025.

[3] N. Bhat, N. Saggu, Pragati, et al., "Generating Visible Spectrum Images from Thermal Infrared using Conditional Generative Adversarial Networks; proceedings of the 2020 5th International Conference on Communication and Electronics Systems (ICCES), F, 2020 [C]."

[4] F. Luo, Y. Li, G. Zeng, et al., "Thermal Infrared Image Colorization for Nighttime Driving Scenes With Top-Down Guided Attention [J]," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 9, pp. 15808–15823, 2022.

[5] A. Anoosheh, T. Sattler, R. Timofte, et al., "Night-to-day image translation for retrieval-based localization; proceedings of the 2019 International conference on robotics and automation (ICRA), F, 2019 [C]."

[6] J.-Y. Zhu, T. Park, P. Isola, et al., "Unpaired image-to-image translation using cycle-consistent adversarial networks; proceedings of the Proceedings of the IEEE international conference on computer vision, F, 2017 [C]."

[7] F.-Y. Luo, Y.-J. Cao, K.-F. Yang, et al., "Memory-Guided Collaborative Attention for Nighttime Thermal Infrared Image Colorization of Traffic Scenes [J]," IEEE Transactions on Intelligent Transportation Systems, 2024.

[8] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, et al., "Deep Unsupervised Learning using Nonequilibrium Thermodynamics; proceedings of the Proceedings of the 32nd International Conference on Machine Learning, Lille, France, F 07–09 Jul, 2015 [C]," PMLR.

[9] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models; proceedings of the Proceedings of the 34th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, F, 2020 [C]," Curran Associates Inc.

[10] X. Zhang, Y. Li, F. Li, et al., "Ship-Go: SAR Ship Images Inpainting via instance-to-image Generative Diffusion Models [J]," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 207, pp. 203–217, 2024.

[11] J. Song, H. Xu, G. Jiang, et al., "Frequency domain-based latent diffusion model for underwater image enhancement [J]," Pattern Recognition, vol. 160, p. 111198, 2025.

[12] N. Ravi, V. Gabeur, Y.-T. Hu, et al., "SAM 2: Segment Anything in Images and Videos, F, 2024 [C]."

[13] S. Cui, Y. Li, J. Li, et al., "Continual Test-Time Adaptation for Single Image Defocus Deblurring via Causal Siamese Networks [J]," Int J Comput Vision, vol. 133, no. 7, pp. 4134–4157, 2025.

[14] Y. Sun, B. Cao, P. Zhu, et al., "Drone-Based RGB-Infrared Cross-Modality Vehicle Detection Via Uncertainty-Aware Learning [J]," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 10, pp. 6700–6713, 2022.

[15] Z. Wang, A. C. Bovik, H. R. Sheikh, et al., "Image quality assessment: from error visibility to structural similarity [J]," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, 2004.

[16] R. Zhang, P. Isola, A. A. Efros, et al., "The unreasonable effectiveness of deep features as a perceptual metric; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2018 [C]."

[17] M. Heusel, H. Ramsauer, T. Unterthiner, et al., "Gans trained by a two time-scale update rule converge to a local nash equilibrium [J]," Advances in neural information processing systems, vol. 30, 2017.

[18] M. Everingham, S. M. A. Eslami, L. Van Gool, et al., "The pascal visual object classes challenge: A retrospective [J]," International journal of computer vision, vol. 111, pp. 98–136, 2015.

[19] P. Isola, J.-Y. Zhu, T. Zhou, et al., "Image-to-Image Translation with Conditional Adversarial Networks; proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), F, 2017 [C]."

[20] N. G. Nair and V. M. Patel, "T2V-DDPM: Thermal to Visible Face Translation using Denoising Diffusion Probabilistic Models; proceedings of the 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), F, 2023 [C]."

[21] O. Özdenizci and R. Legenstein, "Restoring Vision in Adverse Weather Conditions With Patch-Based Denoising Diffusion Models [J]," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 8, pp. 10346–10357, 2023.

[22] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization [Z]," 2017.

[23] L. Tian, Q. Shen, Z. Deng, et al., "Mask-Guided Cross-Modality Fusion Network for Visible-Infrared Vehicle Detection [J]," IEEE Signal Processing Letters, vol. 32, pp. 1815–1819, 2025.
[24] S. Hu, F. Gao, X. Zhou, et al., "Hybrid Convolutional and Attention Network for Hyperspectral Image Denoising [J]," IEEE Geoscience and Remote Sensing Letters, vol. 21, pp. 1–5, 2024.
[25] X. Qiu, R.-J. Zhu, Y. Chou, et al., "Gated attention coding for training high-performance and efficient spiking neural networks; proceedings of the Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, F, 2024 [C]," AAAI Press.
[26] S. Xu, S. Zheng, W. Xu, et al., "HCF-Net: Hierarchical Context Fusion Network for Infrared Small Object Detection; proceedings of the 2024 IEEE International Conference on Multimedia and Expo (ICME), F, 2024 [C]."