

AV2T-Gen: Aerial Visible to Thermal Generation with Environment and Vehicle State Guidance

Kun Yang¹, Yuxiang Liu², Yihan Wang², Shen Yan², Maojun Zhang²,
Yu Liu^{2(✉)}, Xue Wang¹, and Qing Wang¹

¹ Northwestern Polytechnical University, Xi'an, China

² National University of Defense Technology, Changsha, China
jasonyuliu@nudt.edu.cn

Abstract. Accurate visible-to-thermal image generation holds significant application value in fields such as vehicle detection and autonomous driving. However, existing methods often suffer from the problem that the generated results are inconsistent with environmental conditions and vehicle states (e.g., moving/static). Moreover, the scarcity of synchronized and calibrated visible-thermal image pairs has further hindered the development of these fields. We propose AV2T-Gen, an innovative environment and vehicle guided visible-to-thermal generation framework built upon the Instruct-Pix2Pix Diffusion Models for aerial imagery, which distinguishes itself from existing methods by explicitly incorporating environment parameters and vehicle state information for the first time. The framework first maps images to the latent space using a VAE to process image features efficiently, and then generates directional features for vehicles in different environmental conditions using multimodal feature information. Meanwhile, we also proposed a visible-thermal paired dataset with environmental conditions and vehicle states. Experimental results have demonstrated that the generated thermal images outperform existing approaches in terms of physical realism and vehicle detail discrimination significantly, with more accurate reflection of environmental and vehicle characteristics. The code is available at <https://github.com/NPU-CVPG/AV2T>.

Keywords: Visible-to-Thermal Translation · Diffusion Models · Environment Guidance · Multimodal Perception

1 Introduction

Thermal imaging technology has become a crucial means of achieving reliable scene understanding in complex environments. Its ability to present contrasts caused by temperature differences is independent of external lighting conditions. This importance is even more prominent in RGB-T fusion tasks [1].

✉ Corresponding author.

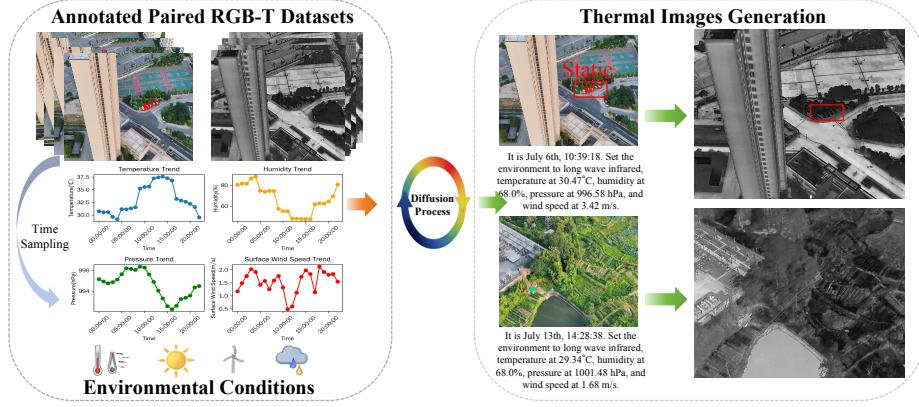


Fig. 1: Our method takes environmental conditions and target motion states as conditions to generate thermal images that both comply with radiometrically plausible and are consistent with the context. Additionally, we propose a novel paired visible-thermal imaging dataset, which is annotated with rich environmental and target metadata.

Recent studies have explored this technology to improve tasks such as pedestrian detection [4, 20], segmentation [27, 31], and vehicle detection [16, 41]. These methods typically rely on paired RGB images and thermal images. However, capturing thermal images requires specialized hardware, including infrared cameras and sensors, making dataset collection both expensive and labor-intensive. Meanwhile, accurately aligned visible-thermal image pairs are even more challenging to obtain. These limitations have prompted the use of generative models to synthesize thermal images from RGB inputs as a more scalable and flexible alternative. This approach ensures perfect alignment of details between authentic RGB images and their synthesized thermal image counterparts, providing high-quality data for tasks that require precise cross-modal correspondences, such as dense feature matching [6, 26] and keypoint matching [17]. Benefiting from these advantages, recent studies [8, 29, 37, 38] have demonstrated significant improvements in model performance on downstream tasks by leveraging synthesized RGB-T datasets.

Early methods relied on Generative Adversarial Networks (GANs) [13, 15, 21], but they often struggled with overfitting in low-data regimes and failed to preserve physical realism. Diffusion models (DMs) [19] later emerged as a more stable alternative, leveraging their ability to model complex distributions. However, many still treat translation as a style transfer task, neglecting critical factors such as vehicle-specific thermal signatures and environmental variability. ThermalGen [24] employs a flow-based generative model to develop an RGB-T conversion model that generates high-fidelity thermal images across diverse perspectives, sensor characteristics, and scenario categories. F-ViT [22] introduced a paradigm shift by integrating foundation models (*e.g.*, SAM [12], Grounded DINO [18], and CLIP [23]) to guide diffusion, outperforming state-of-the-art

methods across near-infrared (NIR), medium-wave infrared (MWIR), and long-wave infrared (LWIR) bands while generalizing well to out-of-distribution scenarios.

Despite this progress, key gaps remain. Existing methods often lack explicit mechanisms to model dynamic environmental parameters (*e.g.*, ambient temperature) or vehicle states (*e.g.*, motion), leading to generated thermal images that do not align with real-world physics. Additionally, while F-ViT enables multi-band translation via text prompts, few approaches combine the strengths of foundation-model guidance with fine-grained control over scene-specific factors. Moreover, the scarcity of datasets annotated with environment and vehicle metadata limits the ability to train models that generalize across complex, unconstrained scenarios.

In this work, we introduce AV2T-Gen, a novel framework that bridges these gaps by integrating environment-awareness and vehicle-state modeling into a diffusion-based translation pipeline (Fig. 1 right). Building on the foundation of model guidance, our approach explicitly incorporates dynamic context, such as ambient conditions and vehicle motion states, to generate thermal images that are both physically consistent and contextually aligned. Furthermore, we present a new paired visible-thermal dataset (Fig. 1 left) annotated with rich environmental and vehicle metadata, addressing the data scarcity in context-aware translation tasks.

Thus, our contributions are as follows:

- Environment parameters are systematically integrated into multimodal input. A precise modal alignment mechanism ensures that generated infrared images accurately reflect environmental conditions, boosting their physical realism and scene adaptability.
- Leveraging vehicle states (*e.g.*, static/dynamic), fine-grained generation of infrared features is achieved. This closely aligns the generated results with the vehicle’s actual state, enhancing detail discriminability and state consistency in infrared images.
- A paired aerial-based dataset of visible and infrared images with a continuous time series is constructed. It provides synchronized, calibrated image pairs along with detailed annotations of environmental conditions and vehicle states, offering high-quality data support for cross-modal generation and fusion tasks that better align with real dynamic scenes.

2 Related Work

2.1 Synthetic Thermal Data in Downstream Tasks

Recent studies have shown a growing trend of leveraging synthetic thermal data to enhance downstream tasks. XoFTR [35] applies random cosine transforms to RGB images to improve cross-modal alignment performance. This alignment capability can serve downstream tasks that rely on cross-modal collaboration; AVIID [8] directly utilizes synthetic aerial thermal data for the downstream

task of vehicle detection; STGL [37] and STHN [35, 38], by synthesizing RGB-T datasets, have respectively improved the performance of two types of downstream tasks: satellite-thermal image retrieval and homography estimation. These studies demonstrate that thermal image synthesis is a practical and feasible solution to address data scarcity in downstream tasks. In this study, we propose a general-purpose RGB-T image translation model that generates high-quality synthetic thermal data for widespread application in improving the performance of various downstream tasks.

2.2 GAN-Driven RGB-Thermal Image Translation

In the existing literature, the conversion from visible light to thermal imaging has been extensively studied, with most approaches addressing this problem using Generative Adversarial Networks (GANs). Early studies [8, 10, 14, 30] primarily employed GAN-based methods [7] to synthesize photorealistic thermal images, adopting models like Pix2Pix [9, 32], an image-to-image translation model based on conditional adversarial networks and Cycle-GAN [43], which achieves stable training by introducing additional consistency loss. However, these early GAN models were not explicitly designed for the visible-to-thermal imaging task.

To address this specificity, researchers have proposed task-oriented GAN-based approaches, such as ThermalGAN [13], which leverages thermal imaging features for human recognition, and InfraGAN [21], which enables infrared-domain translation for general scenes via pixel-level losses and structural similarity. In recent years, advances in the field have followed two paths: on one hand, GAN-based research has been refined with methods like EGGAN [15], which utilizes edge map information of natural scenes to guide the conversion process; on the other hand, more recent studies [24, 42] have shifted to exploring diffusion model-based approaches to achieve higher fidelity in thermal image synthesis.

2.3 Diffusion-Model-Powered RGB-Thermal Image Translation

As an alternative to the aforementioned GAN-based methods, the recently proposed PID [19] adopts diffusion models to accomplish the conversion task. Additionally, this study introduces physically meaningful loss functions to preserve the physical properties of the generated infrared images. Another notable diffusion-based method, F-ViT, can produce high-quality conversion results across three scenarios: LWIR, MWIR, and NIR.

Beyond GAN and diffusion model frameworks, recent style transfer methods [3, 5, 28, 33, 40] have also explored text-guided and style-image-guided transfer paradigms. However, RGB-T image conversion faces unique challenges due to fundamental differences between modalities: on the one hand, the RGB and thermal domains cannot directly share weights; on the other hand, effective conversion requires learning complex cross-modal relationships across different RGB-T datasets. These challenges make it difficult to directly apply RGB-style transfer methods to robust RGB-T image conversion models.

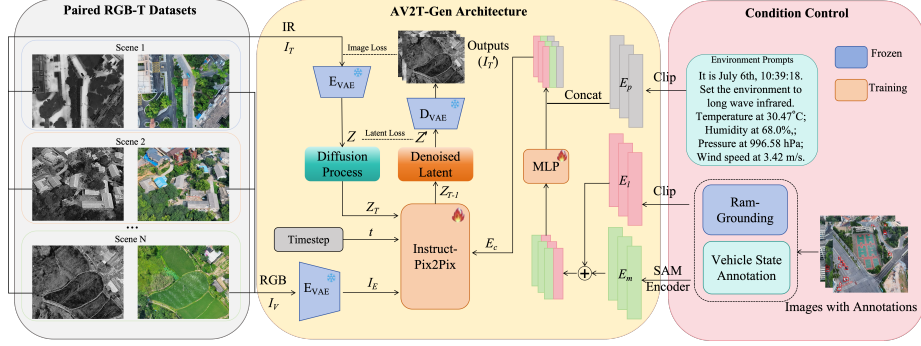


Fig. 2: Method Overview. Environmental condition statements are additionally incorporated as inputs and encoded into Prompt Embeddings via CLIP. Concurrently, vehicle state annotations are fused with universal labels derived from Ram-Grounding; bounding boxes are fed into the SAM to generate Mask Embeddings, while labels are encoded into Label Embeddings using CLIP. Collectively, these text-based embeddings (*i.e.*, Prompt Embeddings and Label Embeddings) and mask-based embeddings (*i.e.*, Mask Embeddings) are simultaneously employed as control conditions for the UNet architecture. In the latent space processing pipeline, the VAE first encodes visible-light and infrared images into latent vectors I_E and Z , respectively. Subsequently, Gaussian noise is injected into the latent vectors to yield Z_T . This vector Z_T is then concatenated with I_E to form the UNet’s input. The UNet performs predictive modeling to output the latent vector Z' , which is ultimately decoded by the VAE to facilitate the supervised training process of the entire framework.

3 Method

3.1 Overview

The pipeline of the proposed AV2T-Gen is shown in Fig. 2. Our method aims to generate thermographic images with vehicle consistency from visible-light inputs by leveraging vehicle-state annotations and foundation models as additional information inputs, combined with a diffusion model integrated with environmental context. This workflow consists of three interconnected components.

3.2 VAE-Based Latent Space Encoding and Decoding

To capture and reconstruct the structural and thermal properties of images in a compact latent space, we employ a VAE for both visible images and thermal images. Specifically, the VAE architecture adopts the architecture used in Instruct-Pix2Pix [2], achieving stable performance across both visible and thermal domains.

Visible Image Encoding. The pre-trained VAE, which is frozen during the training process, encodes the visible input image into a latent vector I_E . This step compresses redundant information while extracting high-level structural

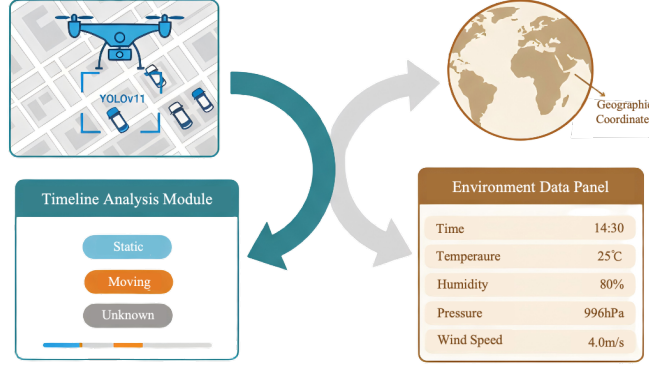


Fig. 3: Condition Acquisition. We used YOLOv11 for vehicle detection on an aerial image sequence to determine whether the vehicle is moving, thereby obtaining the vehicle state labels; meanwhile, we acquired the environmental condition information of the current location through the aerial’s GPS data.

features (e.g., vehicle shapes, scene layouts). The process can be expressed as:

$$I_E = \mathcal{V}_E(I_V), \quad (1)$$

where $\mathcal{V}_E$ denotes the VAE Encoder, and I_V is the input visible light image.

Thermal Image Encoding and Reconstruction. For thermal images I_T , the VAE (frozen) encodes them into latent vectors Z . During the diffusion-based generation stage, the goal is to learn a mapping from visible-light-derived latent feature to thermal-like Z' . Subsequently, the VAE decoder reconstructs the final thermal image from the generated latent vector, expressed as:

$$I'_T = \mathcal{V}_D(Z'), \quad (2)$$

where $\mathcal{V}_D$ denotes the VAE decoder.

3.3 Foundation Model-Driven Context Extraction

To incorporate semantic information, vehicle states, and environmental guidance into thermal image generation, we leverage foundation models (Ram-Grounding [18, 39] for bounding boxes and label association, SAM [12] for segmentation, and CLIP [23] for text-image alignment) to extract context-rich embeddings. This ensures the physical accuracy of the generated infrared images.

Acquisition of Vehicle State Annotations and Environmental Prompts.

As shown in Fig. 3, for vehicle state annotation, we first trained a visible-light vehicle detection model from a drone perspective using the YOLOv11 [11] model. Then, based on temporal sequences, we determined whether the same vehicle is moving. Based on this, we can determine the state of each vehicle: stationary, moving, or unknown. Algorithm details are provided in the supplementary material.

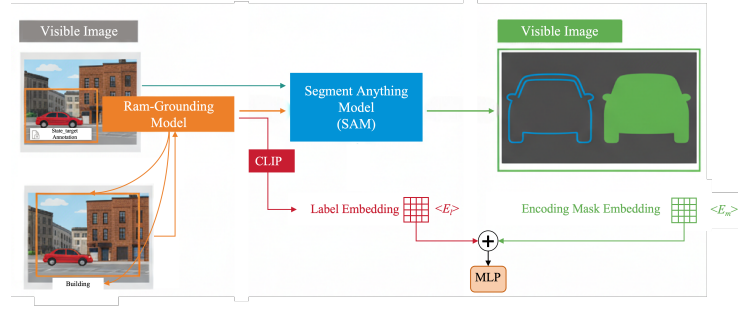


Fig. 4: Acquisition of E_m and E_l . We obtained general labels and bounding boxes via the foundation model Ram-Grounding, which were combined with vehicle state annotations. Finally, E_m and E_l were derived using SAM and CLIP.

Regarding environmental conditions, we determine the shooting location and time based on the GPS data and the capture time of thermal images. We obtain the corresponding ecological conditions-time, temperature, humidity, air pressure, and wind speed using the latitude and longitude of the relevant region and an open-source climate platform. We then use interpolation algorithms to convert discrete data into continuous values, enabling us to obtain environmental conditions at any time point. Finally, we combine these ecological conditions to generate prompts.

Mask Embeddings via SAM. First, using the Ram-Grounding model, common vehicle labels and bounding boxes are automatically identified from visible light images. Concurrently, by integrating automatically annotated vehicle-state files, complete labels and their corresponding bounding boxes are obtained. Next, the visible-light images and bounding boxes are processed by SAM (Segment Anything Model) to generate object masks (e.g., for vehicles, buildings). These masks are encoded as mask embeddings E_m , which capture the precise boundaries of vehicles.

Embedding Fusion with MLP. As shown in Fig. 4, CLIP encodes the complete labels individually into label embeddings E_l , which capture semantic information. The mask embeddings E_m and label embeddings E_l are concatenated and then fed into an MLP (trained) to generate integrated contextual embeddings. These embeddings are further combined with prompt embeddings E_p to form the final guiding conditional input. This process is expressed as:

$$E_c = (E_p, \text{MLP}(E_m, E_l)), \quad (3)$$

where E_c denotes the final conditional embedding vector.

Physical Interpretation of Contextual Guidance. While AV2T-Gen does not explicitly simulate the full radiative transfer equations, it effectively incorporates environmental contextual priors into the generation process. Specifically, the model captures material-dependent emissivity via semantic-material grounding (facilitated by SAM and Ram-Grounding). Simultaneously, it implic-

Table 1: We conducted training and testing separately for each scene, and also performed combined training and testing using all scenes. In all cases, our method achieves the best performance.

Method	Scene 1				Scene 2				Scene 3				Scene 4				All scenes			
	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	LPIPS $\downarrow$
DiffV2IR [24]	23.32	0.701	108.40	0.36	22.76	0.651	82.01	0.37	21.62	0.632	135.75	0.39	23.23	0.693	85.20	0.37	22.35	0.620	108.54	0.39
PID [19]	23.35	0.660	100.12	0.36	22.15	0.593	84.11	0.37	22.26	0.628	84.94	0.36	21.75	0.623	83.16	0.37	22.33	0.621	106.13	0.36
sRGB-TIR [15]	23.43	0.708	107.49	0.36	22.80	0.652	82.03	0.38	18.58	0.552	80.85	0.40	23.23	0.692	84.67	0.37	22.34	0.619	110.83	0.39
ThermalGen [36]	24.30	0.771	93.73	0.33	23.02	0.710	69.83	0.26	23.91	0.751	71.02	0.35	23.97	0.738	72.73	0.32	23.94	0.734	83.16	0.37
Ours	25.27	0.791	78.11	0.25	24.00	0.710	58.53	0.23	24.42	0.761	60.08	0.28	24.64	0.760	52.50	0.25	24.08	0.750	67.45	0.29

itly approximates atmospheric attenuation by utilizing ambient humidity and air pressure as effective proxy variables.

3.4 Diffusion-Based Thermal Generation with Denoising UNet

The denoising UNet (trained) is the core of the diffusion process, responsible for transforming visible-derived latent I_E , random noise Z_T , and conditional embedding E_c into thermal-like latent vector Z' .

Diffusion Process. Starting from a noisy version of Z , the U-Net iteratively removes noise over multiple timesteps. At each timestep t , the network takes the current latent Z_{T-1} and the integrated context embeddings as input. Given by:

$$Z_{T-1} = \mathcal{U}(Z_T, E_c, I_E), \quad (4)$$

where $\mathcal{U}$ denotes the UNet. Until denoising is performed to obtain the predicted latent vector Z' , the predicted thermal image is obtained using Eq. 2.

QKV Attention Mechanism. Inside the UNet, QKV attention modules enable the network to capture long-range dependencies in the latent space, ensuring that structural details (guided by SAM masks) and semantic consistency (guided by CLIP labels and prompts) are preserved during denoising.

Loss Function. We use the L_1 loss calculated from the latent vector Z' predicted by the UNet and the real latent vector Z as the latent space loss L_{Latent} . Meanwhile, we employ the Sobel operator to extract the edges of the predicted image I'_T and the real image I_T , and calculate the L_1 loss of these edges as the edge loss L_{Edge} to preserve better scene structure. Additionally, we use VGG [25] to compute the perceptual loss $L_{perception}$, which enhances the preservation of detailed information.

$$L = \lambda_1 L_{Latent} + \lambda_2 L_{Edge} + \lambda_3 L_{perception}, \quad (5)$$

where we empirically set $\lambda_1 = 1.0$, $\lambda_2 = 0.5$ and $\lambda_3 = 0.25$.

4 Experiments

4.1 Experiment Setup

Data Division. Our paired dataset of visible-infrared aerial views consists of infrared images captured across different scenarios and months, spanning various

Table 2: Distribution of Paired RGB-T Data. This table shows the number of samples across different scenes as well as the split between the training set and test set. All images are from an aerial perspective.

Scene Category	Resolution	Quantity	Split Ratio
Scene 1 (substation)	640×512	3,175	9/1
Scene 2 (school)		9,428	
Scene 3 (square)		10,505	
Scene 4 (driving school)		6,564	
All scenes	640×512	29,672	9/1

daytime periods. We split each scenario into a training set and a validation set, with detailed dataset information provided in Tab 2. Additionally, we combined all scenarios into a unified set to ensure the generalization performance and accuracy of our method.

Evaluation Metrics. Following the evaluation metrics and temperature assessment indicators used in numerous visible to infrared conversion papers, we use PSNR, SSIM [34], and FID of temperature values as the evaluation metrics.

Implementation Details. During training, we ran 100 epochs. Our experimental environment consists of 8 NVIDIA RTX 4090 graphics cards (with 24GB memory each), and the resolution of both training and testing images is set to 640×512 . Additional details, including the dataset used, training configurations, and other relevant information, are available in our supplementary materials.

Computational Complexity and Inference. Our complete AV2T-Gen framework contains approximately 15.5B parameters, which are primarily concentrated in the foundation models (RAM, SAM) and the Diffusion Model. The total inference time for a 640×512 image is approximately 10.00 seconds on a single RTX 4090 GPU. We emphasize that AV2T-Gen is intentionally designed as an offline high-fidelity data generation engine addressing the scarcity of paired aerial RGB-T data. Thus, it prioritizes physical consistency and thermal realism over real-time onboard deployment.

4.2 Comparison Experiment

Comparison Methods. We conducted training and testing both on individual scenarios and across all scenarios. To evaluate our infrared image conversion performance, we compared our method with the DiffV2IR [24], PID [19], sRGB-TIR [15], and the ThermalGen [36]. To ensure fairness, we used the same datasets and identical training and testing settings. Our method incorporates vehicle states and environmental conditions, making it more physically realistic and enabling it to yield more accurate results with additional inputs. **Results.** As shown in Tab. 1, our AV2T-Gen outperforms other methods in quantitative metrics. Fig. 5 presents the conversion results of different methods. Our method produces more realistic, detail-rich conversions of both the environment and vehicles than other methods, thanks to additional scene-aware inputs.



Fig. 5: We conducted tests in four scenarios to verify the adaptability and performance of our method. The figure above shows that our method outperforms other methods in both environmental authenticity and vehicle prediction.

Efficiency and Fairness. As shown in Tab. 3, the diffusion process accounts for over 94% , while the large-scale foundation models contribute negligible latency. Crucially, these models remain frozen; only the MLP and UNet are trained. Thus, performance gains stem from our proposed guidance mechanism and physical priors rather than an unfair advantage in trainable capacity.

Cross-Dataset Generalization. To verify the cross-perspective robustness and ensure our model does not merely overfit to specific dataset statistics, we conducted zero-shot and fine-tuned evaluations on the public LLVIP benchmark [10]. As demonstrated in Tab. 4, even under zero-shot settings, AV2T-Gen consistently outperforms state-of-the-art baselines like F-VITA and ThermalGen. This confirms that our explicit MLP-based physical conditioning effectively learns transferable physical radiance mappings rather than memorizing scenario-specific correlations.

Table 3: Inference Latency Analysis. We measure the runtime cost of each module on a single RTX L40 GPU with input resolution 640×512 . All time values are recorded in seconds, and the total latency includes data reading and writing I/O overhead.

Module	RAM	SAM	CLIP	DM	Total (Including I/O)
Time (s)	0.048	0.143	0.034	9.45	10.00

Table 4: Cross-dataset generalization on the LLVIP benchmark. We compare the Zero-shot and Fine-tuned performance of our AV2T-Gen against baseline methods. The results demonstrate that our framework maintains superior physical fidelity and robustness across both settings, indicating effective learning of transferable physical radiance mappings rather than overfitting.

Method	Zero-shot Setting				Fine-tuned Setting			
	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	LPIPS $\downarrow$
DiffV2IR [24]	9.42	0.251	281.30	0.61	20.85	0.612	105.40	0.41
PID [19]	10.15	0.315	255.60	0.56	21.43	0.645	98.20	0.39
sRGB-TIR [15]	10.87	0.332	242.10	0.53	21.76	0.658	92.50	0.37
ThermalGen [36]	12.58	0.388	196.20	0.47	23.12	0.735	74.60	0.30
Ours	13.82	0.435	175.50	0.43	23.68	0.765	60.80	0.27

Table 5: Pedestrian Generation Quality on the LLVIP. We report PSNR, SSIM, FID and LPIPS for evaluating the rendering quality.

Method	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	LPIPS $\downarrow$
DiffV2IR [24]	21.02	0.635	102.40	0.38
PID [19]	21.85	0.662	95.10	0.35
ThermalGen [36]	23.15	0.742	70.42	0.28
Ours	24.32	0.785	52.15	0.21

Focus and Pedestrian Analysis. While the framework is optimized for these aerial scenarios, we also evaluate it on the LLVIP benchmark. We fine-tuned AV2T-Gen on the LLVIP dataset and calculated metrics for localized pedestrian regions using ground-truth masks. As shown in Tab. 5, our method achieves superior performance in fine-grained pedestrian generation compared to existing baselines.

Cross-Scenario Transferability. Furthermore, to evaluate potential scene-specific bias within our dataset, we trained a model variant solely on three scenes and tested it on the unseen scene. As shown in Tab. 6, the minimal performance drop between the in-domain setup and the cross-scenario setup validates our model’s capacity to extract generalized physical priors across distinct environments.

Table 6: Comprehensive cross-scenario generalization analysis. To evaluate the robustness of our physical priors, we perform a leave-one-out cross-validation across all four scenes (e.g., training on Scenes 2, 3, and 4, then testing on Scene 1). Compared with the in-domain baseline performance (see Tab. 1), the highly competitive metrics achieved in these cross-scenario settings demonstrate that AV2T-Gen effectively learns generalized, environment-aware thermal representations.

Test Scene	Training Scene	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	LPIPS $\downarrow$
Scene 1	S2, S3, S4	23.75 ($\blacktriangledown$ 1.52)	0.742 ($\blacktriangledown$ 0.049)	90.15 ($\blacktriangle$ 12.04)	0.31 ($\blacktriangle$ 0.06)
Scene 2	S1, S3, S4	22.82 ($\blacktriangledown$ 1.18)	0.673 ($\blacktriangledown$ 0.037)	68.45 ($\blacktriangle$ 9.92)	0.28 ($\blacktriangle$ 0.05)
Scene 3	S1, S2, S4	23.15 ($\blacktriangledown$ 1.27)	0.722 ($\blacktriangledown$ 0.039)	71.24 ($\blacktriangle$ 11.16)	0.34 ($\blacktriangle$ 0.06)
Scene 4	S1, S2, S3	23.28 ($\blacktriangledown$ 1.36)	0.714 ($\blacktriangledown$ 0.046)	64.12 ($\blacktriangle$ 11.62)	0.32 ($\blacktriangle$ 0.07)

Table 7: Quantitative ablation study. We report PSNR, SSIM, FID and LPIPS for evaluating the rendering quality.

$L_{perception}$	L_{edge}	Vehicle States	Env. Prompt	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	LPIPS $\downarrow$
	$\checkmark$	$\checkmark$	$\checkmark$	23.14	0.716	72.19	0.30
$\checkmark$		$\checkmark$	$\checkmark$	22.90	0.722	68.00	0.30
$\checkmark$	$\checkmark$		$\checkmark$	22.93	0.732	73.83	0.30
$\checkmark$	$\checkmark$	$\checkmark$		22.29	0.700	77.66	0.31
$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	24.08	0.750	67.45	0.29

4.3 Ablation Study

We conducted ablation experiments from three perspectives to demonstrate the effectiveness of each component in our method: whether to use the vehicle state, whether to use environmental conditions, and whether to use perception and edge losses. For all data, the quantitative results of this ablation study are presented in Tab. 7.

Vehicle State. We conducted experiments without using vehicle states. Since the vehicle’s infrared characteristics vary across different states, as shown in Tab. 7, an accurate conversion cannot always be achieved during infrared image generation. Furthermore, as illustrated in Fig. 6, when vehicle states are injected as conditions, infrared image generation becomes more accurate and more consistent with the real world.

Environmental Conditions. We tested the strategy of not injecting environmental conditions. In this strategy, because the input is only visible-light images and lacks additional information, the network tends to learn the average infrared features of the data, resulting in slightly lower style consistency and realism in the generated infrared images. Table 7 presents the quantitative metrics of this strategy, while Fig. 6 shows that infrared thermal image conversion with environmental condition injection is more accurate. To further clarify the individual contributions of different ambient variables and address potential parameter re-

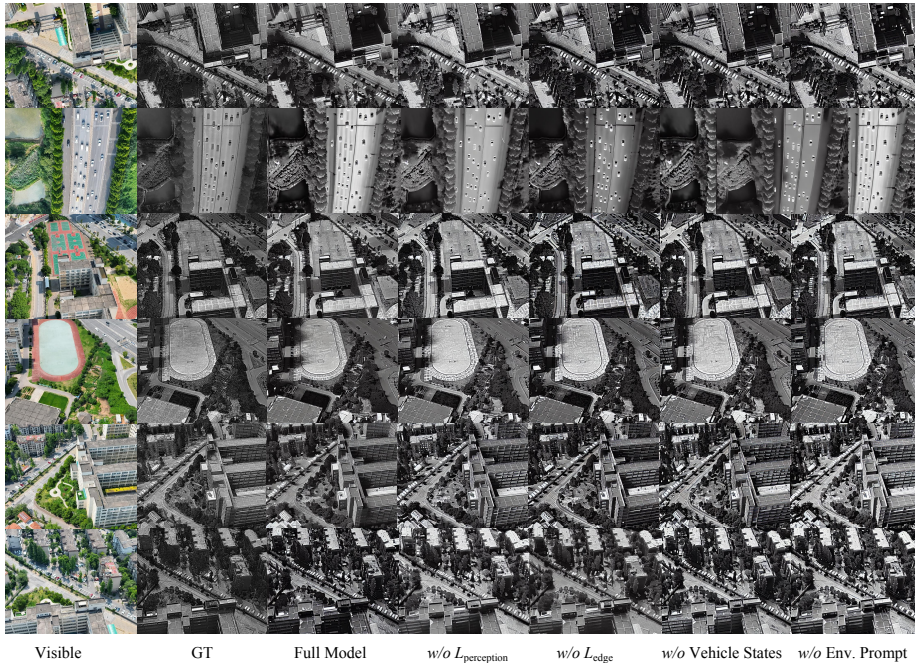


Fig. 6: We conducted visual comparisons under the conditions of using the complete model, not using perceptual loss, not using edge loss, not using vehicle state labels, and not using environmental change prompts, respectively. The experimental results show that our strategy is optimal.

Table 8: Granular ablation of environmental parameters. We evaluate the individual impact of Temperature (T), Humidity (H), Wind Speed (V), and Air Pressure (P) on generation quality. The results indicate an importance ranking of $T > H > V > P$, demonstrating that our model effectively leverages multi-parameter synergy rather than relying on a single dominant variable.

T	H	V	P	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	LPIPS $\downarrow$
	✓	✓	✓	23.41	0.725	71.30	0.30
✓		✓	✓	23.68	0.736	69.85	0.30
✓	✓		✓	23.85	0.742	68.70	0.29
✓	✓	✓		23.97	0.747	67.92	0.29
✓	✓	✓	✓	24.08	0.750	67.45	0.29

dundancy, we conducted a granular sensitivity analysis. As detailed in Tab. 8, the relative significance of factors influencing generation quality follows the descending order of Temperature (T) > Humidity (H) > Wind Speed (V) > Pressure (P). This multi-parameter synergy highlights that our framework performs physically-consistent radiance synthesis rather than generic style transfer.

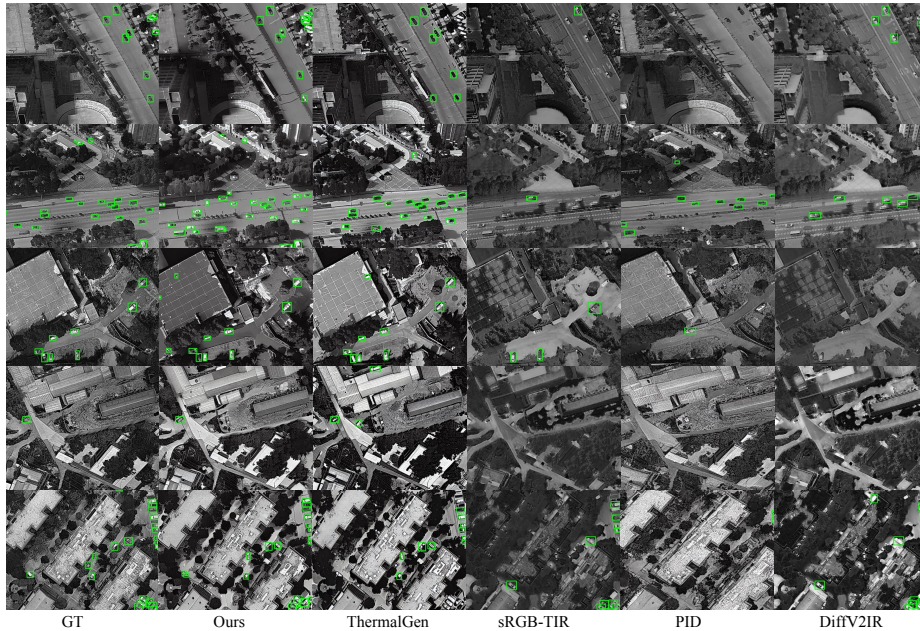


Fig. 7: Green boxes represent the detection results of each method. In the visualized results, the thermal images generated by our method achieve more accurate detection.

Table 9: Detection Comparison. We used a comparative method to convert the test sets of all images into infrared images, and then employed the same detector to perform object detection on these infrared images. The results (unit: %) are as follows.

Method	mAP $\uparrow$	mAP_50 $\uparrow$	mAP_75 $\uparrow$	mAP_s $\uparrow$
DiffV2IR [24]	6.1	17.3	2.6	5.1
PID [19]	0.3	0.7	0.02	0.1
sRGB-TIR [15]	7.33	16.3	5.2	6.0
ThermalGen [36]	11.7	29.4	6.9	10.7
Ours	15.6	35.9	11.1	14.2

Perception and Edge Losses. We performed ablation on the loss function and found that considering only the latent loss leads to the loss of details. Incorporating more image-level losses improves the accuracy of the prediction details. As shown in Tab. 7, adding the pixel-level loss further improves the results. As illustrated in Fig. 6, our loss function significantly enhances the accuracy of thermal image prediction.

4.4 Downstream Task - Vehicle Detection

We used a fine-tuned YOLOv11 to detect vehicles in authentic infrared images, using ground-truth annotations. Then, we converted our method and the

Table 10: Quantitative Evaluation. (*mv/st*: moving/static). We report mAP_{50} (%) metrics for two test sets A and B under moving and static scenarios.

Method	A (mAP_{50})%		B (mAP_{50})%	
	<i>moving</i> ↑	<i>static</i> ↑	<i>moving</i> ↑	<i>static</i> ↑
PID [19]	12.4	15.8	10.5	13.2
DiffV2IR [24]	35.6	38.2	32.4	35.8
ThermalGen [36]	52.4	55.6	48.5	51.2
Ours	80.2	75.6	78.4	73.1

comparative methods to infrared images and used the same detector to detect vehicles. As shown in Tab. 9, our method can provide more accurate detection metrics in the vehicle detection task. As shown in Fig 7, thermal images obtained using our method can achieve more accurate detection results.

We also conducted separate cross-domain experiments in our datasets. As shown in Tab. 10, Setup A assesses feature alignment by testing a detector trained on authentic IR data. Conversely, Setup B validates practical utility by training the detector on synthetic data and testing it on authentic IR. Our strategy effectively generates fine-grained infrared features.

5 Conclusion

In this work, we address the crucial challenge of generating physically consistent thermal images from visible-light data, a task vital for robust scene understanding in scenarios where visible-light cameras fail. By introducing AV2T-Gen, we integrate environmental awareness and vehicle-state modeling into a diffusion-based translation pipeline and explicitly incorporate dynamic context, including ambient conditions, vehicle motion, and sensor properties, thereby ensuring the generated thermal images align with real-world physics. Moreover, we present a novel paired visible-thermal dataset enriched with environmental conditions and vehicle metadata, thereby alleviating the data-scarcity bottleneck for context-aware translation tasks.

Experiments demonstrate that AV2T-Gen outperforms state-of-the-art methods in terms of physical consistency and contextual alignment. It effectively supports downstream tasks such as dense feature matching and object detection, which require precise cross-modal correspondences. While our work advances the field, future research can further integrate real-time sensor data and explore granular vehicle prompts derived from temporal tracking history to refine the physical fidelity of dynamic targets in unconstrained real-world scenarios.

Acknowledgements. This work was supported by NSFC under Grant 62571449.

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. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 18392–18402 (2023)
3. Chung, J., Hyun, S., Heo, J.P.: Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8795–8805 (2024)
4. Deng, Q., Tian, W., Huang, Y., Xiong, L., Bi, X.: Pedestrian detection by fusion of rgb and infrared images in low-light environment. In: *2021 IEEE 24th International Conference on Information Fusion (FUSION)*. pp. 1–8. IEEE (2021)
5. Deng, Y., He, X., Tang, F., Dong, W.: Z*: Zero-shot style transfer via attention reweighting. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6934–6944 (2024)
6. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: Roma: Robust dense feature matching. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 19790–19800 (2024)
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. *Communications of the ACM* **63**(11), 139–144 (2020)
8. Han, Z., Zhang, Z., Zhang, S., Zhang, G., Mei, S.: Aerial visible-to-infrared image translation: Dataset, evaluation, and baseline. *Journal of Remote Sensing* **3**, 0096 (2023)
9. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1125–1134 (2017)
10. Jia, X., Zhu, C., Li, M., Tang, W., Zhou, W.: Llvip: A visible-infrared paired dataset for low-light vision. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 3496–3504 (2021)
11. Khanam, R., Hussain, M.: Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725* (2024)
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 4015–4026 (2023)
13. Kniaz, V.V., Knyaz, V.A., Hladuvka, J., Kropatsch, W.G., Mizginov, V.: Thermal-gan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In: *Proceedings of the European Conference on Computer Vision (ECCV) workshops*. pp. 0–0 (2018)
14. Lee, C., Anderson, M., Ranganathan, N., Zuo, X., Do, K., Gkioxari, G., Chung, S.J.: Caltech aerial rgb-thermal dataset in the wild. In: *European Conference on Computer Vision*. pp. 236–256. Springer (2024)
15. 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. *arXiv preprint arXiv:2301.12689* (2023)
16. Liao, G., Gao, W., Li, G., Wang, J., Kwong, S.: Cross-collaborative fusion-encoder network for robust rgb-thermal salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology* **32**(11), 7646–7661 (2022)

17. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17627–17638 (2023)
18. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European Conference on Computer Vision. pp. 38–55. Springer (2024)
19. 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 p. 111816 (2025)
20. Ozcan, A., Cetin, O.: A novel fusion method with thermal and rgb-d sensor data for human detection. IEEE Access **10**, 66831–66843 (2022)
21. Özkanoglu, M.A., Ozer, S.: Infragan: A gan architecture to transfer visible images to infrared domain. Pattern Recognition Letters **155**, 69–76 (2022)
22. Paranjape, J.N., de Melo, C., Patel, V.M.: F-vita: Foundation model guided visible to thermal translation. arXiv preprint arXiv:2504.02801 (2025)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PmLR (2021)
24. 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)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
26. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8922–8931 (2021)
27. Sun, Y., Zuo, W., Yun, P., Wang, H., Liu, M.: Fuseseg: Semantic segmentation of urban scenes based on rgb and thermal data fusion. IEEE Transactions on Automation Science and Engineering **18**(3), 1000–1011 (2020)
28. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1921–1930 (2023)
29. Tuzcuoglu, Ö., Köksal, A., Sofu, B., Kalkan, S., Alatan, A.A.: Xoftr: Cross-modal feature matching transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4275–4286 (2024)
30. Wadsworth, E., Mahajan, A., Prasad, R., Menon, R.: Deep learning for thermal-rgb image-to-image translation. Infrared Physics & Technology **141**, 105442 (2024)
31. Wan, Z., Wang, Y., Yong, S., Zhang, P., Stepputtis, S., Sycara, K., Xie, Y.: Sigma: Siamese mamba network for multi-modal semantic segmentation. arxiv 2024. arXiv preprint arXiv:2404.04256
32. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8798–8807 (2018)
33. Wang, Z., Zhao, L., Xing, W.: Stylediffusion: Controllable disentangled style transfer via diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7677–7689 (2023)
34. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)

35. Xiao, J., Loianno, G.: Uasthn: Uncertainty-aware deep homography estimation for uav satellite-thermal geo-localization. arXiv preprint arXiv:2502.01035 (2025)
36. Xiao, J., Nayak, R., Zhang, N., Tortei, D., Loianno, G.: Thermalgen: Style-disentangled flow-based generative models for rgb-to-thermal image translation. arXiv preprint arXiv:2509.24878 (2025)
37. Xiao, J., Tortei, D., Roura, E., Loianno, G.: Long-range uav thermal geo-localization with satellite imagery. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5820–5827. IEEE (2023)
38. Xiao, J., Zhang, N., Tortei, D., Loianno, G.: Sthn: Deep homography estimation for uav thermal geo-localization with satellite imagery. IEEE Robotics and Automation Letters (2024)
39. Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., Qin, Y., Luo, T., Li, Y., Liu, S., et al.: Recognize anything: A strong image tagging model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1724–1732 (2024)
40. Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., Xu, C.: Inversion-based style transfer with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10146–10156 (2023)
41. Zhou, W., Guo, Q., Lei, J., Yu, L., Hwang, J.N.: Ecffnet: Effective and consistent feature fusion network for rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **32**(3), 1224–1235 (2021)
42. Zhu, G., Pan, H., Wang, Q., Tian, C., Yang, C., He, Z.: Data generation scheme for thermal modality with edge-guided adversarial conditional diffusion model. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 10544–10553 (2024)
43. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2223–2232 (2017)