Learning Shared-Specific Representations for Cross-Spectral Image Generation

Haining Wang[1,2], Na Li[1,2,†], Huijie Zhao[1,2,†], Yifan Da[1,2], Yan Wen[1], Yi Su[3], and Yuqiang Fang[4]

[1] School of Artificial Intelligence, Beihang University, Beijing, China
{wanghaining,lina_17,hjzhao,sy2342116}@buaa.edu.cn, wenyancas@163.com
[2] Key Laboratory of Aerospace Broad Band Detection and Intelligent Perception, Ministry of Industry and Information Technology, Beijing, China
[3] University of Electronic Science and Technology of China, Chengdu, China
364665843@qq.com
[4] Space Engineering University, Beijing, China
fangyuqiang@nudt.edu.cn
[†] Corresponding author

Abstract. Cross-spectral image synthesis from visible to infrared faces significant challenges due to the severe scarcity of infrared data in critical scenarios. Existing methods primarily achieve style transfer through pixel-level mappings, which often suffer from semantic misalignment and unrealistic spectral characteristics. To address this problem, we rethink cross-spectral generation as a decoupling and reconstruction process of shared-specific features between modalities and propose SHASP, a novel generative network. Specifically, an encoding-decoupling module is designed at the front of the generator, which consists of a structure content encoder and a spectral feature encoder to decouple the shared-specific features. A decoding-reconstruction module is designed at the backend of the generator to perform feature combination and reconstruction. We further impose semantic consistency constraints on shared features to ensure precise cross-modal alignment. Extensive experiments on aerial, driving, and monitoring scenarios demonstrate that our method outperforms state-of-the-art approaches. This work reveals that there are common representations that can be mined at the semantic structure level in cross-spectral data, and modality gaps can be accurately modeled via a decoupling-reconstruction mechanism.

Keywords: Image Generation · Visible and Infrared · Shared-Specific Representation · Cross-Spectral

1 Introduction

Visible (VIS) and infrared (IR) images are two important types of cross-spectral data in the field of visual sensing. They operate within different imaging bands and reflect distinct physical mechanisms [21]. VIS images capture clear texture and detail information, while IR images depict the temperature distribution of

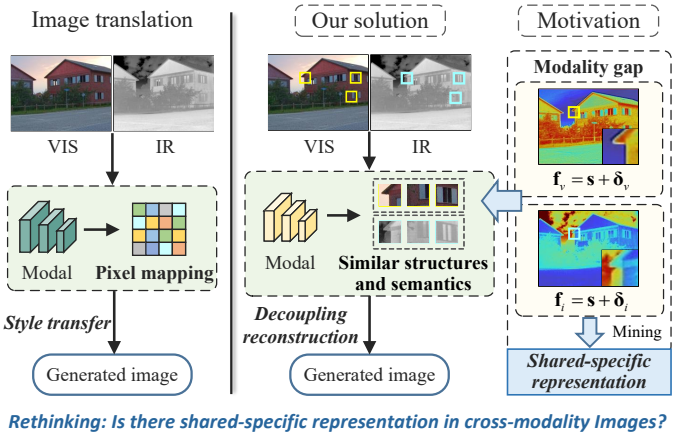


Fig. 1: Illustration of our motivation. Existing methods convert VIS images into a style similar to IR through pixel mapping, whereas our goal is to explore the shared-specific representations between cross-modality, and solve the modality gap in translation.

scenes and targets. Many downstream tasks use the complementary properties of these two cross-modal data for collaborative applications, such as object detection [6, 42], pedestrian re-identification [40, 44], and autonomous driving [41]. However, the scale of IR data collected in many tasks cannot meet the requirements for deep model learning. Therefore, generating equivalent IR data from easily accessible VIS data has emerged as a significant research direction.

Recently, inspired by generative models, increasing efforts have been devoted to cross-spectral generation [14, 31, 39, 46] using GANs [7, 49] and diffusion models [10]. These methods typically embed VIS images into a latent feature space and reconstruct IR images via nonlinear mapping. Although they can capture basic IR imaging styles, they suffer from critical limitations. Most existing methods focus on pixel-level nonlinear mapping between modalities, while failing to explicitly mine shared structural semantics such as scene layout and object contours. Some works introduce contrastive learning [11, 12, 25] to refine details via feature or pixel constraints. However, due to the domain gap between VIS and IR, such constraints are insufficient to ensure semantic structure consistency. In general, current cross-spectral translation methods neglect the explicit modeling of shared and modality-specific features, leading to over-smoothing, lost semantic interpretability, or missing key thermal properties.

Physically, VIS images reflect object surface reflectance, while IR images represent thermal radiation and temperature distribution. Their cross-spectral translation is inherently a cross-domain knowledge transformation task. As illustrated in Fig. 1, despite the huge modality gap caused by different imaging mechanisms, recent visual studies have verified that VIS and IR data exhibit strong statistical regularity in deep semantic structure [26, 33]. Such shared representations can be effectively exploited at the semantic level even under large

domain discrepancies. This contradiction between low-level modality divergence and high-level semantic consistency motivates our work.

Accordingly, we focus on exploring the shared semantic that enables mutual representation between VIS and IR data, as well as how to explicitly model and preserve modality-specific attribute features within the generative translation process. We hypothesize that VIS and IR images can achieve reliable semantic alignment via shared structural features, while modality-specific properties can be accurately disentangled and reconstructed. Based on this hypothesis, we re-formulate cross-spectral image translation as a shared-specific feature representation learning problem, aiming to explore explicit knowledge association and semantic alignment for heterogeneous visual data. This work provides novel insight for solving the semantic gap in heterogeneous data conversion.

Specifically, we propose SHASP, a shared-specific feature learning network for VIS-IR cross-spectral generation. It consists of the generator and discriminator to generate high-fidelity IR images for downstream tasks. Guided by decoupled representation, we introduce an encoding-decoupling module at the front of the generator to disentangle shared structural content and modality-specific spectral features. A decoding-reconstruction module then fuses and reconstructs these features in two stages: cross-modality feature combination and image generation. Finally, we apply dual-branch structural constraints to refine the domain-invariant semantic features extracted by the structure content encoder.

Leveraging the proposed decoupled representation learning framework, our cross-spectral data generation model effectively bridges the modality discrepancy in data translation and provides a robust data augmentation solution for downstream tasks. In summary, our contributions are as follows:

- This work proposes a novel IR generation network, SHASP, which is based on shared-specific feature mining. The generator performs decoupling representation and cross-modality reconstruction in the latent space, which improves the semantic consistency and interpretability of IR generation.
- We design a spectral knowledge decoupling method, using the structure content encoder and the spectral feature encoder to extract shared-specific features, and verify that VIS and IR data have shared structural information that can represent each other in similar scenes.
- We rethink the role of domain knowledge in cross-spectral synthesis, providing novel insights for solving the semantic gap in data conversion, and we offer a robust data augmentation solution for downstream tasks.

2 Related Works

2.1 Cross-spectral Image Generation

Cross-spectral image generation is a key technology in visual computing. It aims to build a cross-domain mapping relationship between VIS and IR fields. Existing research faces the core challenge of the cross-modality semantic gap caused by distinct imaging principles. Previous studies [2, 5, 24, 29, 34] have conducted

feature conversion learning based on the VIS and IR images. Among them, methods based on paired data in dual-domain [17, 18] learn the IR thermal feature distribution through registered training samples, but their over-reliance on prior assumptions about data distribution leads to limited cross-domain generalization. In order to break through data constraints, some studies [23, 32, 38] introduced the thermal information to enhance the feature expression of the target area through semantic segmentation and attention masking, but the feature representation gap still exists. Some methods [16, 48] adopt an edge-guided strategy to enhance contour preservation, but the specific feature of IR images still need to be considered. In addition, image translation frameworks [15, 20, 35, 36] achieve style transfer through latent space transformation, and their structure-preserving mechanisms provide important references for cross-spectral data generation, such as patch contrastive loss [13], cycle consistency loss [49]. Recent frequency domain analysis methods provide a new paradigm for cross-modality translation. Band decoupling based on wavelet transform [43] and joint learning of Fourier spectrum [47] achieve hierarchical modeling by separating high and low-frequency components. Laplacian pyramid decomposition [50] refines high-frequency details after completing contour mapping of low-frequency. Furthermore, the latest work combining visual language models has also achieved good results. Although these methods can demonstrate diverse generation capabilities, the fundamental physical differences between VIS and IR still constitute a significant technical barrier to data generation, which further motivates our exploration on explicit disentanglement of shared and specific characteristics.

2.2 Cross-modality Knowledge Representation

According to visual theory, there are similarity, symbolism, and reproducibility between VIS and IR images [28]. In terms of data representation, the physical mechanisms and feature relationships contained in the two modalities can be mined to decouple the multi-level and multi-scale latent generative factors within the data. Currently, some studies learn and decompose cross-modality knowledge through decoupled representation [8, 22]. In the domain separation network [45], the infrared spectra feature z is subdivided into the variable component z_S and the invariant component z_C. According to the implementation method of knowledge representation, the existing cross-modality representation can be divided into joint representation, collaborative representation, and encoding-decoding representation. Joint representation maps different modality information into a shared feature space, and modality interaction is required for prediction during inference [1, 3]. Collaborative representation learns a separate feature extraction model for each modality, which can focus on the differences and similarities between modalities [37]. Encoding-decoding representation converts one modality's feature to another's latent space, and then maps the multi-modal features into a vector space [4]. In summary, knowledge representation plays an important role in the fields of VIS-IR image translation, cross-modality object detection and related tasks. However, few works explicitly unify these representation paradigms for cross-spectral image generation.

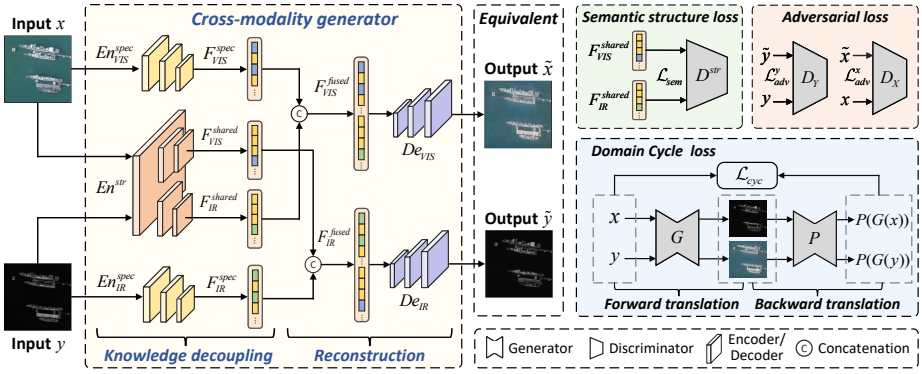


Fig. 2: Overall framework of the proposed SHASP. We designed the cross-modality generator, which consists of knowledge decoupling and reconstruction process. Specifically, an encoding-decoupling module is designed at the front of the generator, and a decoding-reconstruction module is designed at the back of the generator. Then, semantic consistency constraints for shared features are designed to achieve more precise alignment and refinement.

3 Methodology

3.1 Preliminary and Problem Formulation

For cross-spectral image translation, the generation model is required to accurately reconstruct details and global features, while having effective learning capability under unpaired data. Therefore, the architecture based on GAN is adopted to enable the model to have data fitting and adversarial capabilities.

Given image data $\{x \in X\} \subset \mathbb{R}^{H\times W\times 3}$ in VIS domain and image data $\{y \in Y\} \subset \mathbb{R}^{H\times W\times 1}$ in IR domain under similar viewing angles, the task of this work is to mine the shared and modality-specific domain knowledge of cross-spectral data and use such knowledge to guide the IR generation. The entire generation network consists of the generators G and P, and the discriminators D_X and D_Y, forming a bidirectional mapping, which can avoid semantic drift in unidirectional mapping. The generator G learns the decoupling and reconstruction of shared-specific features from the source modality to generated modality. The generator P performs reverse mapping from the generated modality to source modality, and the discriminator is responsible for judging the authenticity of the generated cross-modality image. The generation process is as follows, where $f_{X\to Y}$ represents the mapping and generation of learning the IR modality from the VIS.

$$\widetilde{y} = f_{X\to Y}\left(x,\ G(X;Y),\ D\right) \tag{1}$$

According to mathematical analyze, we found that the difference in deep features between VIS and IR in neural networks is much smaller than the energy of shared structures:

$$\|\mathbf{f}_v - \mathbf{f}_i\|_2 = \|\boldsymbol{\delta}_v - \boldsymbol{\delta}_i\|_2 \le \|\boldsymbol{\delta}_v\|_2 + \|\boldsymbol{\delta}_i\|_2 \ll 2\,\|\mathbf{s}\|_2 \tag{2}$$

where $\mathbf{s}$ is the shared structural feature independent of modality, $\boldsymbol{\delta}_v$ and $\boldsymbol{\delta}_i$ are modality-specific features. $\mathbf{f}_v$ and $\mathbf{f}_i$ are deep features extracted by neural networks for VIS and IR, respectively. Therefore, our goal is to distinguish shared features from specific features of VIS and IR data in the generator and guide the network in knowledge mapping. Intuitively, the generator consists of an encoding-decoupling module and a decoding-reconstruction module. Through adversarial and cycle-consistent training, the entire conversion network can learn the shared-specific features and use them as domain knowledge to guide the generation. The network structure is shown in Fig. 2.

3.2 Spectral Domain Knowledge Decoupling

This section introduces the designed encoding-decoupling module, which decouples domain knowledge into shared-specific features. The module consists of a structure content encoder and two spectral feature encoders. The composition of the encoders is shown in Fig. 3.

Structure Content Encoder. The purpose of the structure content encoder En^{str} is to extract shared features of the VIS and IR modalities, such as semantic structures and target locations, through partially shared network weights, thereby achieving domain-invariant representation learning for multimodal data. It has dual encoding branches for VIS and IR. Through hierarchical feature extraction, the common semantic structures of cross-modality data are mapped to the same feature space, yielding F_{VIS}^{shared} and F_{IR}^{shared}. Specifically, first, each encoding branch uses a large convolution kernel with a nonlinear activation function to complete the initial feature mapping while retaining spatial details. Then, spatial downsampling is performed through cascaded convolution blocks and channel expansion to retain deep semantic information and obtain semantic features F_{VIS}^{sem} and F_{IR}^{sem}. Finally, the residual structure with shared weights is deployed in the deep layers of the encoder to achieve domain alignment of the feature space through parameter sharing. The process is expressed as follows:

$$\begin{aligned} F_{VIS}^{shared} &= F_{VIS}^{sem} + R(F_{VIS}^{sem},\ \mathrm{W}^{shared}), \\ F_{IR}^{shared} &= F_{IR}^{sem} + R(F_{IR}^{sem},\ \mathrm{W}^{shared}) \end{aligned} \tag{3}$$

where, $R(\cdot,\ \mathrm{W}^{shared})$ represents the residual function with shared weights W^{shared}.

Spectral Feature Encoder. The purpose of the spectral feature encoder En_{VIS}^{spec} and En_{IR}^{spec} is to use modality-specific branches to deeply explore the unique physical properties and style knowledge of different spectral domains. For example, color information and texture composition in the VIS domain, and thermal radiation regions in the IR domain. Specifically, first, a reverse-pyramid feature extraction based on a multi-level convolution unit is performed. Convolution layers of different scales are used to capture global semantics and local details to form a feature extraction network with a reverse pyramid structure.

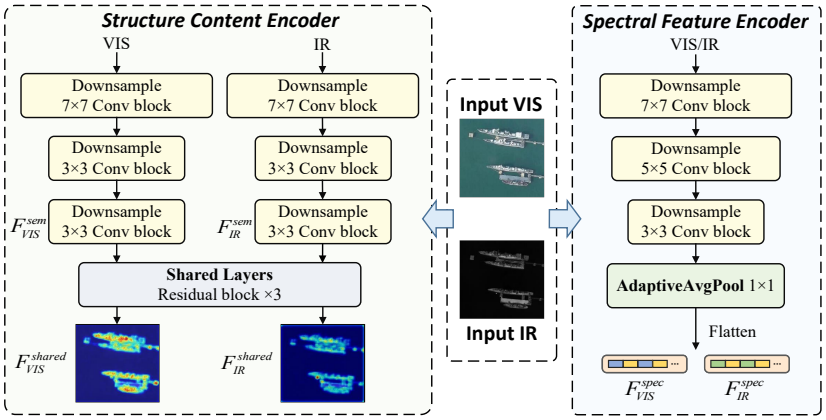


Fig. 3: The composition of the structure content encoder and spectral feature encoder. VIS and IR images share similar modality-invariant structural features.

The shallow convolution block uses a large receptive field convolution kernel to quickly model the semantic features of the global scene, the middle convolution block focuses on the statistical characteristics of the local texture pattern, and the deep convolution block realizes spatial feature compression and nonlinear mapping. Then, adaptive global average pooling and dynamic feature mapping mechanism are added to map the extracted spatial features to a low-dimensional latent space to form a modality-specific feature vector F_{VIS}^{spec} and F_{IR}^{spec}. Through the above design, a nonlinear mapping from the pixel space of the source image to the deep-level semantic attributes can be achieved, providing a modality-specific attribute representations for subsequent decoding and reconstruction. The process can be expressed as follows:

$$\begin{aligned} F_{VIS}^{spec} &= En_{VIS}^{spec}(x),\ x \in X \\ F_{IR}^{spec} &= En_{IR}^{spec}(y),\ y \in Y \end{aligned} \tag{4}$$

3.3 Cross-modality Reconstruction and Refinement

As mentioned above, the encoding-decoupling module focuses on explicitly extracting multi-modal shared-specific features. In order to fully utilize complementary information, this section designs the decoding-reconstruction module that completes feature space combination and image reconstruction. Subsequently, feature refinement constraint is designed to enhance the overall expressive ability.

Combination and Reconstruction. The latter half of the generator is the decoding-reconstruction module, which performs feature fusion and image reconstruction. In the feature combination stage, the module first concatenates the modality-specific feature vectors and shared feature vectors generated by the encoding-decoupling module along the channel dimension. This design combines

the structure semantics and modality attribute information of the image and feeds them as vectors into the subsequent decoder to perform image conversion. The process is as follows:

$$\begin{aligned} F_{IR}^{fused} &= Concat(F_{VIS}^{shared},\ F_{IR}^{spec}), \\ F_{VIS}^{fused} &= Concat(F_{IR}^{shared},\ F_{VIS}^{spec}) \end{aligned} \tag{5}$$

where, F_{IR}^{fused} and F_{VIS}^{fused} are the generated vectors after feature fusion. $Concat(\cdot)$ is the concatenation operation. During VIS-to-IR inference, no real IR image is required. The fused latent feature F_{IR}^{fused} fed to the IR decoder is derived from VIS shared structural features and the learned IR spectral prior, which is implicitly embedded in the IR-specific branch and decoder weights during training. Similarly, F_{VIS}^{fused} combines VIS exclusive features and learned cross-spectral shared parameters captured during training.

In the feature reconstruction stage, the multi-level residual block structure and instance normalization are used to conduct nonlinear feature transformation to achieve preliminary decoding. Subsequently, a transposed convolution structure is used to achieve upsampling to ensure that the spliced features are mapped to high-dimensional image data. Thus, the generator generates the converted IR and VIS images through dual-branch decoding and reconstruction. The process is as follows:

$$\begin{aligned} G(x) &= De_{IR}(F_{IR}^{fused}),\ x \in X \\ G(y) &= De_{VIS}(F_{VIS}^{fused}),\ y \in Y \end{aligned} \tag{6}$$

where, De_{IR} and De_{VIS} are the decoders for generating IR and VIS modalities.

Semantic Refinement. Solely relying on the weight-sharing residual layer within the encoding-decoupling module is insufficient to guarantee maximal structural similarity between the semantic representations extracted from the two domains. To address the limitations of residual learning in capturing structural similarity, we introduce a semantic structure discriminator D^{str}. Specifically, it drives the visible-infrared dual-branch encoder En^{str} to extract domain-invariant structural features through adversarial training, thereby encouraging finer alignment of semantic representations in the latent space. The semantic structure loss $\mathcal{L}_{sem}$ is formulated as a domain-adversarial objective:

$$\mathcal{L}_{sem} = \mathbb{E}_{x\sim p_{VIS}}\left[\log D^{str}\left(En_{VIS}^{str}(x)\right)\right] + \mathbb{E}_{y\sim p_{IR}}\left[\log\left(1 - D^{str}(En_{IR}^{str}(y))\right)\right] \tag{7}$$

3.4 Objective Function

Domain Cycle Loss. We believe it may be difficult to collect strictly paired VIS-IR training data for cross-spectral generation tasks, so it is necessary to incorporate an unsupervised learning mechanism into the network. Therefore, inspired by the cyclic consistency constraint in the CycleGAN model, we have designed a domain cyclic loss to ensure that inter-domain conversion is performed

instead of performing conversion only on paired data. Similar to the mapping process of the generator G, we use the generator P to decouple the features of the equivalent IR and VIS images obtained by conversion and reconstruct them into source domain images. This process can be expressed as:

$$\mathcal{L}_{cyc} = \mathbb{E}_{VIS}[\|P(G(x)) - x\|_1] + \mathbb{E}_{IR}[\|P(G(y)) - y\|_1] \tag{8}$$

where, $G(x)$ represents the generated equivalent IR image $\tilde{y}$, and $G(y)$ represents the generated VIS image $\tilde{x}$.

Adversarial Loss. The adversarial loss $\mathcal{L}_{adv}$ constrains the generated images to follow the data distributions of the corresponding target domains through two discriminators. This loss ensures that the generated image maintains the same basic structure as the source image, such as the location and aspect ratio of the target in the image. Based on the adversarial loss designed in GAN [7], the constraints $\mathcal{L}_{adv}^{y}$ generated by the discriminator D_Y are as follows:

$$\mathcal{L}_{\text{adv}}^{y} = \mathbb{E}_{IR}[\log D_Y(y)] + \mathbb{E}_{VIS}[\log(1 - D_Y(\tilde{y}))] \tag{9}$$

Similarly, the constraints $\mathcal{L}_{adv}^{x}$ generated by the discriminator D_X are as follows:

$$\mathcal{L}_{adv}^{x} = \mathbb{E}_{VIS}[\log D_X(x)] + \mathbb{E}_{IR}[\log(1 - D_X(\tilde{x}))] \tag{10}$$

Therefore, the overall adversarial loss $\mathcal{L}_{adv}$ is the sum of $\mathcal{L}_{adv}^{x}$ and $\mathcal{L}_{adv}^{y}$.

Total Loss. Using the loss terms designed above, the total loss function of the network during training is obtained through weighted combination. The form is as follows:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{sem} + \beta \cdot \mathcal{L}_{cyc} + \delta \cdot \mathcal{L}_{adv} \tag{11}$$

where, α, β, δ are weight coefficients, which adjust the weights of semantic structure loss, domain cycle loss, and adversarial loss, respectively.

4 Experiments

4.1 Experimental Settings

Datasets. We train and evaluate the proposed method on two public datasets and a self-built dataset, including AVIID [9], IRVI [19], and DroneCoast. AVIID is a public road-monitoring dataset covering various objects, such as vehicles, pedestrians, trees and houses. It contains 993 pairs of aligned VIS and IR images of resolution 434×434. IRVI is a public vehicle driving scene dataset, mainly recording vehicle and background from the perspective of the road ahead. The training set contains 17,000 pairs of images, and the test set contains 1,000 pairs of images, each with a resolution of 256×256. DroneCoast is a self-built aerial coastal dataset. We collected VIS-IR video streams of coastal scenes useing binocular mid-wave infrared and color cameras. The dataset contains 939 pairs of training images and 164 pairs of test images of resolution 640×512.

Table 1: Quantitative comparison on two public datasets. Bold and underline denote the best and suboptimal performance. Calculate the mean and standard deviation of three experiments as the sensitivity interval.

Method	AVIID			IRVI		
	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓
CycleGAN [49]	0.713±0.003	24.61±0.17	58.87±1.03	0.559±0.002	18.52±0.24	110.66±1.26
CUT [25]	0.719±0.007	24.45±0.33	69.55±1.45	0.631±0.002	20.77±0.41	89.96±0.97
Swin-UNIT [20]	0.804±0.010	25.96±0.19	86.65±2.16	0.653±0.002	18.92±0.29	<u>89.11±0.82</u>
HnegSRC [12]	0.592±0.008	22.43±0.34	53.64±0.81	0.573±0.004	18.57±0.38	131.08±1.10
RGB-TIR [16]	0.820±0.006	25.59±0.41	**36.80±0.95**	0.619±0.005	19.98±0.25	133.79±0.95
DR-AVIT [8]	0.633±0.012	23.15±0.29	120.58±2.75	0.650±0.002	18.37±0.56	114.06±1.37
DMT [36]	0.753±0.005	26.58±0.22	118.89±2.06	0.669±0.003	21.24±0.37	180.76±3.56
StegoGAN [35]	0.825±0.009	26.05±0.20	43.60±0.67	<u>0.671±0.002</u>	<u>21.68±0.19</u>	97.80±1.03
PatchGCL [13]	0.765±0.008	26.31±0.11	52.58±0.74	0.642±0.001	19.89±0.19	105.75±1.14
DiffV2IR [27]	0.822±0.009	<u>27.04±0.26</u>	45.90±0.83	0.655±0.006	20.70±0.30	99.62±0.80
PID [23]	<u>0.840±0.007</u>	25.77±0.44	46.13±1.14	0.667±0.003	21.46±0.23	91.55±1.24
SHASP (Ours)	**0.852±0.005**	**27.44±0.21**	<u>41.08±0.61</u>	**0.679±0.002**	**21.93±0.21**	**84.30±0.87**

Implementation Details. Performance is assessed from three complementary perspectives: pixel-level (SSIM, PSNR), feature-level (FID, KID), and decision-level (AP_{50}, $AP_{50:95}$, Precision and Recall) metrics. Model efficiency is gauged by parameter count and inference time. The Adam optimizer is used to update the model parameters during the training. The batch size in training is set to 1, and the initial learning rate is set to 1×10^{-4}. 200 epochs are set for each of the three datasets, and the learning rate decays by 50% every 50 epochs. All experiments are conducted on a workstation with an RTX 4090 GPU. In addition, we train and test the advanced comparative methods including image translation models CycleGAN [49], CUT [25], DMT [36], PatchGCL [13], Swin-UNIT [20], StegoGAN [35], HnegSRC [12], and the IR generation models RGB-TIR [16], DR-AVIT [8], PID [23], DiffV2IR [27].

4.2 Quantitative Evaluation

Tab. 1 and Tab. 2 show the quantitative evaluation on three cross-spectral datasets using the above metrics. Each experiment is conducted three times, with mean and standard deviation as the error fluctuation interval for network training. The proposed method achieves the best results on seven metrics and the second-best results on two metrics. Specifically, on the AVIID and IRVI datasets, SHASP has obvious advantages in maintaining image structural similarity and reducing signal distortion. Meanwhile, SHASP achieves the best KID score, indicating its superior performance in modeling local semantic consistency and fine-grained feature distribution, even though FID is not optimal, which is more sensitive to global intensity statistics. Moreover, we measured the parameters of each model and the inference time on 256×256 images. Although the

Table 2: Quantitative comparison on the collected dataset.

Method	DroneCoast				Params.	Inf. time
	SSIM↑	PSNR↑	FID↓	KID↓	(M)↓	(ms)↓
CycleGAN [49]	0.681±0.001	24.24±0.22	170.90±2.26	0.1660± 0.0031	11.38	46.7±1.5
CUT [25]	0.633±0.002	22.90±0.18	141.04±0.43	0.1252±0.0019	11.38	45.2±1.9
Swin-UNIT [20]	0.731±0.002	**25.16±0.17**	128.19±1.87	0.0953±0.0009	**3.35**	**32.6±0.6**
HnegSRC [12]	0.701±0.003	22.85±0.19	154.47±2.15	0.0987±0.0017	11.38	41.0±0.9
RGB-TIR [16]	0.710±0.003	24.72±0.20	176.66±1.39	0.0809±0.0014	30.05	56.3±1.2
DR-AVIT [8]	0.614±0.002	20.03±0.17	187.45±2.67	0.1152±0.0031	65.04	68.5±1.4
DMT [36]	0.611±0.003	23.32±0.23	280.48±2.51	0.1690±0.0045	398.10	117.1±2.3
StegoGAN [35]	0.603±0.003	20.76±0.26	136.92±1.50	0.0401±0.0010	11.38	44.7±0.8
PatchGCL [13]	0.647±0.002	21.85±0.19	127.62±1.62	0.0545±0.0022	11.38	68.9±0.8
DiffV2IR [27]	0.744±0.001	24.27±0.09	121.62±1.32	0.0492±0.0008	162.20	185.9±5.4
PID [23]	0.711±0.002	23.65±0.15	**117.15±1.47**	0.0569±0.0015	309.18	460.5±15.3
SHASP (Ours)	**0.760±0.002**	24.89±0.12	125.28±1.56	**0.0379±0.0009**	24.81	50.7±1.0

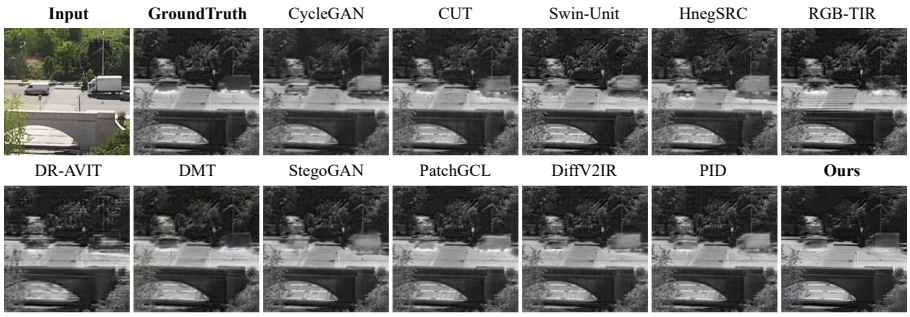


Fig. 4: Qualitative comparisons on the AVIID (Road monitoring scene).

proposed model did not achieve optimal performance in these two indicators, it achieved a balance between computational performance and operational efficiency. Overall, the proposed SHASP demonstrates significant advantages in maintaining structural similarity, reducing signal distortion, and the proximity of feature distribution.

4.3 Visual Qualitative Comparison

The visual results on the three datasets are shown in Fig. 4, Fig. 5, and Fig. 6. It can be seen that the proposed method achieves the best visual quality in maintaining the semantic structure of the VIS modality, which is due to the design of decoupling and mining the shared features of the modalities. In addition, the vehicle regions exhibit clear and distinguishable bright thermal responses in the generated IR images. This indicates that SHASP maintains high fidelity in object edges and structural details from the source domain, while effectively restoring thermal response regions in the IR domain. Compared with

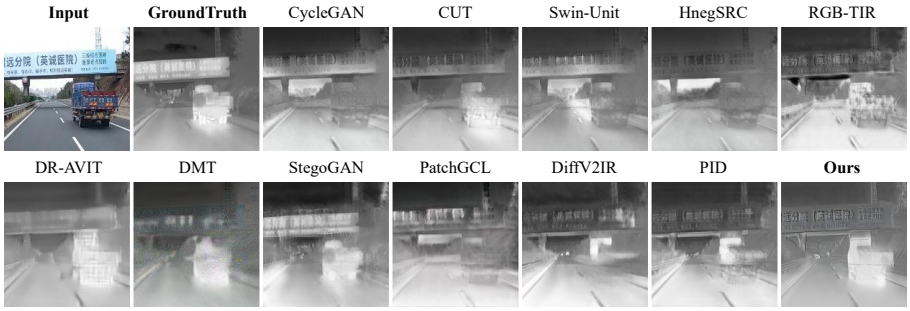


Fig. 5: Qualitative comparisons on the IRVI (Driving scene).

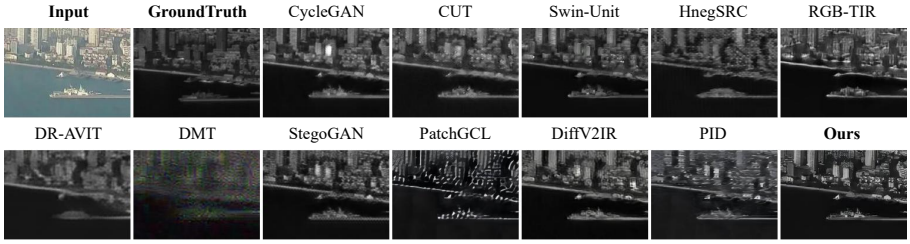
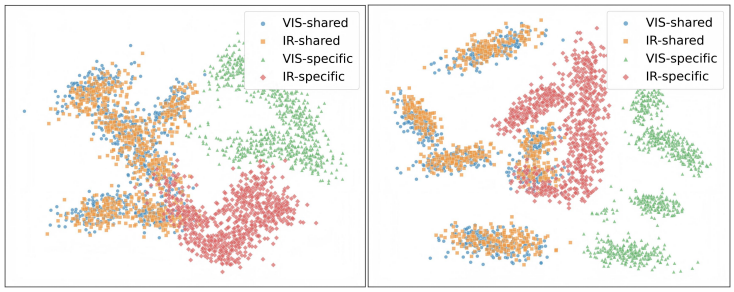


Fig. 6: Qualitative comparisons on the DroneCoast datasets (Aerial coastal scene).

other methods, SHASP also avoids obvious structural distortion and texture artifacts. The results demonstrate that SHASP can effectively extract and generate modality-specific features. Overall, our method generates the most realistic thermal features while preserving the semantic structure in VIS images.

4.4 Ablation Analysis

Effect of Components. Tab. 3 ablates the contribution of each SHASP component on the AVIID and DroneCoast datasets. We first reproduce CycleGAN as the baseline (Model I), then incrementally introduce the structure content encoder and spectral feature encoder into the generator to build Models II, III, and IV. The complete SHASP model is formed by adding semantic refinement constraints to the encoding-decoupling module. The results of Model II and III show that extracting both modality-shared and specific features enhances infrared generation quality. Model IV further demonstrates that joint learning of shared-specific features achieves more significant improvements. To verify scalability, we conduct a data-scalability ablation by enlarging the training data by $4\times$ via label-preserving augmentations. The results demonstrate that our architecture benefits from a larger effective training distribution. We further provide t-SNE plots of globally pooled embeddings. The results reveal tighter overlap between VIS–IR shared embeddings and better separation of modality-specific features. These results are shown in Fig. 7.



t-SNE (AVIID dataset) t-SNE (DroneCoast dataset)

Fig. 7: T-SNE plots of globally pooled embeddings on AVIID and DroneCoast datasets.

Table 3: Ablation study to analyze different components on AVIID and DroneCoast.

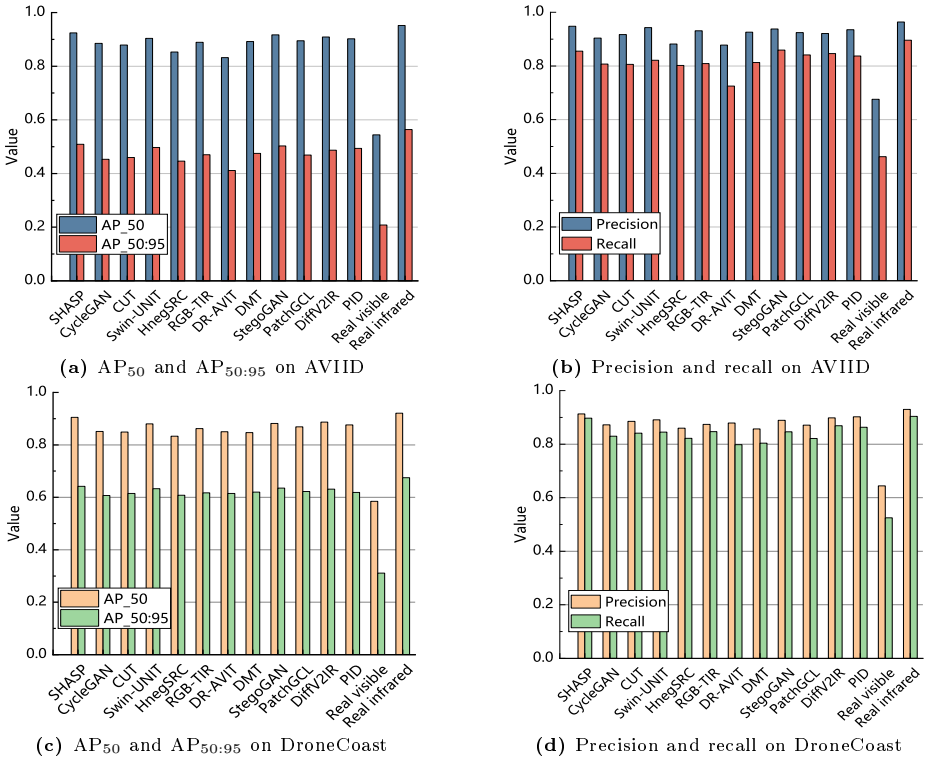
Setup	En^{str}	En^{spec}	Refine	AVIID		DroneCoast	
				SSIM↑	FID↓	SSIM↑	FID↓
I				0.71	58.87	0.68	170.90
II	✓			0.77	50.76	0.72	153.25
III		✓		0.76	54.23	0.72	157.33
IV	✓	✓		0.80	45.39	0.74	136.95
SHASP	✓	✓	✓	0.85	41.08	0.76	125.28
SHASP (4×data)	✓	✓	✓	**0.86**	**40.15**	**0.77**	**124.09**

Table 4: Ablation study on the loss coefficients on the AVIID dataset.

Weight/Metrics	C. group	Experimental group					
α ($\mathcal{L}_{sem}$)	1	2	1	1	2	2	1
β ($\mathcal{L}_{cyc}$)	1	1	2	1	2	1	2
δ ($\mathcal{L}_{adv}$)	1	1	1	2	1	2	2
SSIM↑	0.846	**0.852**	0.844	0.835	0.840	0.847	0.838
FID↓	43.77	**41.08**	42.29	41.62	41.70	42.63	43.45

Furthermore, the ablation analysis verifies the two core issues. First, VIS and IR data have shared information that can represent each other in similar scenes. Second, during the data generation, the feature decoupling-reconstruction mechanism can effectively express the modality-specific attribute.

Effect of Weight Coefficients. We conduct an ablation study to thoroughly investigate the impact of loss weight coefficients on image generation quality, as summarized in Tab. 4. The control group with uniform weights ($\alpha = \beta = \delta = 1$) achieves a balanced performance of 0.846 in SSIM and 43.77 in FID. Among all experimental configurations, adjusting α to 2 while keeping β and δ fixed at 1 yields the most favorable results. This confirms that enhancing the supervision of semantic structure consistency effectively alleviates the cross-



(a) AP_{50} and $AP_{50:95}$ on AVIID

(b) Precision and recall on AVIID

(c) AP_{50} and $AP_{50:95}$ on DroneCoast

(d) Precision and recall on DroneCoast

Fig. 8: Test results of object detection metrics using different training data.

modality domain gap. In contrast, increasing the weight of cyclic consistency or adversarial loss leads to marginal or negative impacts on both perceptual similarity and distribution alignment, indicating that excessive constraints on cycle consistency or adversarial training may introduce unnecessary noise in VIS-IR translation.

4.5 Extension to Object Detection

To further evaluate the effectiveness of knowledge representation in cross-spectral generation, this section applies the IR data generated by the proposed SHASP model to downstream object detection. We perform VIS-to-IR translation on the AVIID and DroneCoast dataset, which contain vehicle and ship targets, respectively. Yolov9 [30] is selected as the benchmark model for object detection. The IR images generated by different translation models are used as the training data, while the remaining real IR images are used as the test data. In addition, we also conduct two control experiments. One group is directly trained using VIS images to evaluate cross-domain detection performance, while the other is trained using real IR images to obtain a non-cross-domain detection reference.

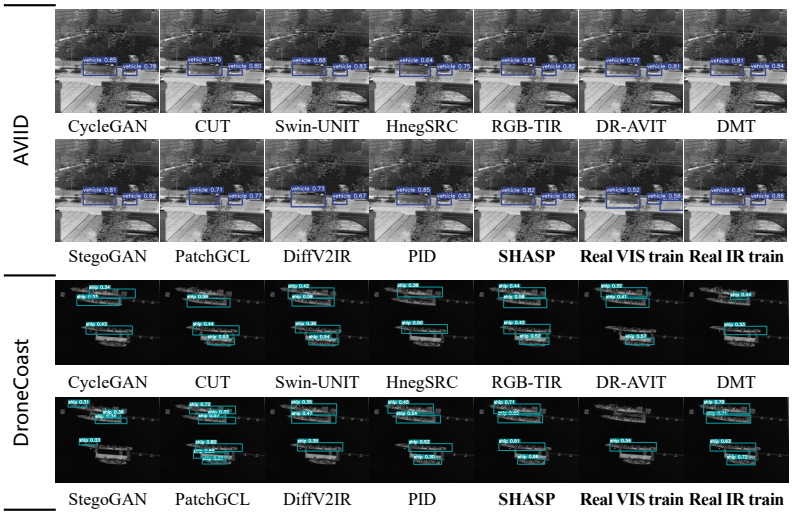


Fig. 9: Applying training data from different generation sources to object detection.

The detection metric results are presented in Fig. 8, and the visualization results are shown in Fig. 9.

The results show that directly using VIS data for cross-domain detection makes it difficult for the detector to learn effective inter-modality generalization. In contrast, the detection performance is significantly improved when using the IR images generated by our SHASP model. This demonstrates that our generated IR images are closer to real IR images in feature distribution and can be effectively applied to downstream detection tasks.

5 Conclusion

This paper proposes SHASP, a novel cross-spectral data generation network for translating VIS images into IR images. Specifically, an encoding-decoupling module is designed in the generator to disentangle shared and modality-specific features. A decoding-reconstruction module is then introduced to combine and reconstruct these features. Finally, semantic features are refined by enforcing consistency constraints on shared feature mapping. Extensive experiments demonstrate that our method achieves state-of-the-art performance in both quantitative and qualitative evaluations, providing an effective data generation solution for downstream tasks. Furthermore, our approach offers new insights into bridging the modality gap in cross-spectral translation and confirms that VIS and IR data share mutually representable structural information. Future work will focus on lightweight model optimization and the extension of the framework to cross-modal temporal signals in videos.

References

1. Chen, J., Yang, L., Yu, W., Gong, W., Cai, Z., Ma, J.: Sdsfusion: A semantic-aware infrared and visible image fusion network for degraded scenes. IEEE Transactions on Image Processing (2025)
2. Dai, W., Lu, L., Li, Z.: Diffusion-based synthetic data generation for visible-infrared person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 11185–11193 (2025)
3. Dong, A., Wang, L., Liu, J., Lv, G., Zhao, G., Cheng, J.: Mfifusion: An infrared and visible image enhanced fusion network based on multi-level feature injection. Pattern Recognition **152**, 110445 (2024)
4. Du, Y., Zhan, J., Li, X., Dong, J., Chen, S., Yang, M.H., He, S.: One-for-all: towards universal domain translation with a single stylegan. IEEE transactions on pattern analysis and machine intelligence (2025)
5. Fu, C., Zhou, X., He, W., He, R.: Towards lightweight pixel-wise hallucination for heterogeneous face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(7), 9135–9148 (2022)
6. Fu, H., Wang, S., Duan, P., Xiao, C., Dian, R., Li, S., Li, Z.: Lraf-net: Long-range attention fusion network for visible–infrared object detection. IEEE Transactions on Neural Networks and Learning Systems **35**(10), 13232–13245 (2023)
7. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
8. Han, Z., Zhang, S., Su, Y., Chen, X., Mei, S.: Dr-avit: Toward diverse and realistic aerial visible-to-infrared image translation. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–13 (2024)
9. 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)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
11. Hu, X., Zhou, X., Huang, Q., Shi, Z., Sun, L., Li, Q.: Qs-attn: Query-selected attention for contrastive learning in i2i translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18291–18300 (2022)
12. Jung, C., Kwon, G., Ye, J.C.: Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18260–18269 (2022)
13. Jung, C., Kwon, G., Ye, J.C.: Patch-wise graph contrastive learning for image translation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 13013–13021 (2024)
14. Kim, J.H., Hwang, Y.: Gan-based synthetic data augmentation for infrared small target detection. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–12 (2022)
15. Kim, S., Jin, C., Diethe, T., Figini, M., Tregidgo, H.F., Mullokandov, A., Teare, P., Alexander, D.C.: Tackling structural hallucination in image translation with local diffusion. In: European Conference on Computer Vision. pp. 87–103. Springer (2024)
16. Lee, D.G., Jeon, M.H., Cho, Y., Kim, A.: Edge-guided multi-domain rgb-to-tir image translation for training vision tasks with challenging labels. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 8291–8298. IEEE (2023)

17. Li, B., Ma, D., He, F., Zhang, Z., Zhang, D., Li, S.: Infrared image generation based on visual state space and contrastive learning. Remote Sensing **16**(20), 3817 (2024)
18. Li, J., Fei, K., Sun, Y., Wang, J., Liu, B., Zhou, Z., Zheng, Y., Sun, Z.: Efficient-pip: Large-scale pixel-level aligned image pair generation for cross-time infrared-rgb translation. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8974–8981. IEEE (2024)
19. Li, S., Han, B., Yu, Z., Liu, C.H., Chen, K., Wang, S.: I2v-gan: Unpaired infrared-to-visible video translation. In: Proceedings of the 29th ACM international conference on multimedia. pp. 3061–3069 (2021)
20. Li, Y., Li, Y., Tang, W., Zhu, Z., Yang, J., Liu, Y.: Swin-unit: Transformer-based gan for high-resolution unpaired image translation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 4657–4665 (2023)
21. Liu, J., Wu, G., Liu, Z., Wang, D., Jiang, Z., Ma, L., Zhong, W., Fan, X.: Infrared and visible image fusion: From data compatibility to task adaption. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
22. Luo, B., Wang, H., Wang, J., Zhu, J., Zhao, X., Gao, Y.: Hypergraph-guided disentangled spectrum transformer networks for near-infrared facial expression recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 10101–10109 (2024)
23. 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)
24. Özkanoğlu, M.A., Ozer, S.: Infragan: A gan architecture to transfer visible images to infrared domain. Pattern Recognition Letters **155**, 69–76 (2022)
25. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: European conference on computer vision. pp. 319–345. Springer (2020)
26. Quan, D., Wang, S., Huyan, N., Chanussot, J., Wang, R., Liang, X., Hou, B., Jiao, L.: Element-wise feature relation learning network for cross-spectral image patch matching. IEEE Transactions on Neural Networks and Learning Systems **33**(8), 3372–3386 (2021)
27. 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)
28. Sun, X., Tian, Y., Lu, W., Wang, P., Niu, R., Yu, H., Fu, K.: From single-to multi-modal remote sensing imagery interpretation: A survey and taxonomy. Science China Information Sciences **66**(4), 140301 (2023)
29. Tang, S., Ye, X., Xue, F., Xu, R.: Cross-modality depth estimation via unsupervised stereo rgb-to-infrared translation. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
30. Wang, C.Y., Yeh, I.H., Mark Liao, H.Y.: Yolov9: Learning what you want to learn using programmable gradient information. In: European conference on computer vision. pp. 1–21. Springer (2024)
31. Wang, H., Li, N., Zhao, H., Wen, Y., Su, Y., Fang, Y.: Mappingformer: Learning cross-modal feature mapping for visible-to-infrared image translation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 10745–10754 (2024)
32. Wang, L., Cheng, J., Song, J., Pan, X., Zhang, C.: Learning to measure infrared properties of street views from visible images. Measurement **207**, 112320 (2023)

33. Wang, X., Guan, Z., Qian, W., Cao, J., Liang, S., Yan, J.: Cs2fusion: Contrastive learning for self-supervised infrared and visible image fusion by estimating feature compensation map. Information Fusion **102**, 102039 (2024)
34. Wang, Z., Colonnier, F., Zheng, J., Acharya, J., Jiang, W., Huang, K.: Tirdet: Mono-modality thermal infrared object detection based on prior thermal-to-visible translation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 2663–2672 (2023)
35. Wu, S., Chen, Y., Mermet, S., Hurni, L., Schindler, K., Gonthier, N., Landrieu, L.: Stegogan: Leveraging steganography for non-bijective image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7922–7931 (2024)
36. Xia, M., Zhou, Y., Yi, R., Liu, Y.J., Wang, W.: A diffusion model translator for efficient image-to-image translation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
37. Yang, B., Chen, J., Ye, M.: Shallow-deep collaborative learning for unsupervised visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16870–16879 (2024)
38. Yang, S., Sun, M., Lou, X., Yang, H., Zhou, H.: An unpaired thermal infrared image translation method using gma-cyclegan. Remote Sensing **15**(3), 663 (2023)
39. Yang, X., Chen, J., Yang, Z.: Cooperative colorization: Exploring latent cross-domain priors for nir image spectrum translation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 2409–2417 (2023)
40. Ye, M., Wu, Z., Du, B.: Dual-level matching with outlier filtering for unsupervised visible-infrared person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
41. Yu, Z., Li, S., Shen, Y., Liu, C.H., Wang, S.: On the difficulty of unpaired infrared-to-visible video translation: Fine-grained content-rich patches transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1631–1640 (2023)
42. Yuan, M., Wang, Y., Wei, X.: Translation, scale and rotation: cross-modal alignment meets rgb-infrared vehicle detection. In: European Conference on Computer Vision. pp. 509–525. Springer (2022)
43. Zhang, L., Chen, X., Tu, X., Wan, P., Xu, N., Ma, K.: Wavelet knowledge distillation: Towards efficient image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12464–12474 (2022)
44. Zhang, Y., Yan, Y., Lu, Y., Wang, H.: Adaptive middle modality alignment learning for visible-infrared person re-identification. International Journal of Computer Vision pp. 1–21 (2024)
45. Zhang, Z., Avramidis, S., Li, Y., Liu, X., Peng, R., Chen, Y., Wang, Z.: A bidirectional domain separation adversarial network based transfer learning method for near-infrared spectra. Engineering Applications of Artificial Intelligence **137**, 109140 (2024)
46. Zhao, Z., Wang, C., Li, C., Zhang, Y., Tang, J.: Modality conversion meets superresolution: A collaborative framework for high-resolution thermal uav image generation. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–14 (2024)
47. Zhong, Y., Li, B., Tang, L., Kuang, S., Wu, S., Ding, S.: Detecting camouflaged object in frequency domain. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4504–4513 (2022)

48. 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)
49. 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)
50. Zhu, Z., Li, Y., Li, Y., Yang, J., Chen, P., Liu, Y.: Seit: Structural enhancement for unsupervised image translation in frequency domain. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7820–7827 (2024)