2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

SSL-RGB2IR: Semi-supervised RGB-to-IR Image-to-Image Translation for Enhancing Visual Task Training in Semantic Segmentation and Object Detection

Aniruddh Sikdar[1], Qiranul Saadiyean[2], Prahlad Anand[3], Suresh Sundaram[4]

2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 979-8-3503-7770-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/IROS58592.2024.10802815

Abstract— **The scarcity of annotated infrared (IR) image datasets limits deep learning networks from achieving performances comparable to those achieved with RGB data. To address this, we introduce a novel semi-supervised RGB-to-IR Image-to-Image Translation model (SSL-RGB2IR) that generates synthetic IR data from RGB images. Our model effectively preserves the IR characteristics in the generated images from both synthetic and real-world data. Compared to existing image-to-image translation techniques, training models on this generated IR data significantly improves performance in downstream tasks like segmentation and detection. Notably, in sim-to-real transfer, the segmentation model trained on SSL-RGB2IR generated IR images outperforms baselines and other Image-to-Image (I2I) models. Furthermore, for real-world applications utilizing EO/IR fusion images, this approach solves the well-known challenge of co-registering EO and IR images, which often have inherent misalignment's due to differing sensor characteristics. Our code is available at https://github.com/prahlad-anand/ssl-rgb2irhttps://github.com/prahlad-anand/ssl-rgb2ir.**

I. INTRODUCTION

Deep learning models have greatly enhanced the scene parsing capabilities of autonomous vehicles [24], [25], [22], [2]. Recent research in robotics and computer vision has increased interest in thermal infrared (IR) imaging because of its effectiveness in harsh weather conditions and low-light environments [23]. Despite the utility of IR cameras, the images they generate usually contain less semantic information compared to electro-optical (EO) RGB images. This disparity often leads to significant drops in the performance of deep learning models for downstream tasks [3]. The fusion of RGB and IR data helps overcome limitations associated with individual sensor modalities, providing a more comprehensive understanding of the environment. This combined approach addresses the individual limitations of modalities, such as the low resolution of IR and the incapacity of RGB sensors to function effectively in low-light conditions [28], [6].

However, multi-modal techniques encounter three primary challenges: (1) noisy IR images resulting from sensor noise, (2) scarcity of co-registered IR images with EO images, and (3) inadequate IR data for training models based on IR images. Images captured by IR cameras typically exhibit low resolution and blurred object boundaries. Training large deep-learning models requires extensive annotated IR data, which is often unavailable. Manual annotation is both time-consuming and expensive, further complicating the acquisition of sufficient data for training purposes. Computer simulations can provide annotated IR data, but they require prior knowledge of IR objects [7]. Approaches capable of using annotated RGB images to generate high-quality annotated IR images can offer a solution to the challenges mentioned above.

Generative Adversarial Networks (GANs) based translation techniques [4], [15], [1] have been used for RGB data to generate annotated IR images. However, these bi-domain-based models tend to generate artifacts in the resulting IR images when the input images differ significantly from the training data. While these approaches offer a solution, the generated images often lack the distinctive style and characteristics of real IR data. Contrastive learning-based methods [19] also struggle to learn semantic relations when significant discrepancies exist. To overcome the limitations of bi-domain-based methods, Dong *et al.* introduced an edge-guided and style-controlled multidomain RGB to IR translation network [14]. It is based on a multi-domain translation framework that uses disentangled content and style latent vectors for image translation. Their emphasis on edge consistency, particularly for simulated data, leads to a decline in performance when faced with a substantial domain gap, as observed in the transition from simulated to real-world datasets.

Given the challenges associated with the scarcity of co-registered EO-IR pairs, supervised image-to-image translation methods encounter difficulties. In view of these limitations, we introduce Semi-supervised RGB-to-IR Image-to-Image Translation (SSL-RGB2IR) for enhancing vision task training in semantic segmentation and object detection. A supervised training branch is introduced within the existing unsupervised framework. By leveraging generated paired data, this approach aims to alleviate the challenges associated with unsupervised training. The semi-supervised framework is capable of learning from both co-registered and unco-registered images. SSL-RGB2IR has three stages, the first is the pseudo pair generation stage, followed by semi-supervised training. The synthetic co-registered images

[1] Aniruddh Sikdar is with the Robert Bosch Centre for Cyber-Physical Systems, Indian Institute of Science, Bangalore. aniruddhss@iisc.ac.in

[2] Qiranul Saadiyean is with the Department of Aerospace Engineering, Indian Institute of Science, Bangalore. qiranuls@iisc.ac.in

[3] Prahlad Anand is with the School of Computer Science and Engineering, Vellore Institute of Technology, Vellore. prahladanand01@gmail.com

[4] Suresh Sundaram is with the Department of Aerospace Engineering, Indian Institute of Science, Bangalore. vssuresh@iisc.ac.in

generated during the pseudo-pair generation stage facilitate the supervised training branch in effectively learning pixel-wise correspondence, thereby enhancing the generation of fine details in IR images. This process is guided by patch-wise contrastive learning loss, which ensures fidelity in IR representation. Embedded within the unsupervised training branch, our semi-supervised framework aids in attaining the desired IR features while preserving the distinctive attributes of authentic IR imagery. To summarize, the main contributions of this paper are:

- Semi-supervised RGB-to-IR Image-to-Image Translation network is introduced, designed to generate IR images from RGB images.
- It comprises both supervised and unsupervised branches. It employs a Generative Adversarial Network (GAN) framework with contrastive learning-based losses for both branches.
- The performance of models trained on images generated by SSL-RGB2IR is compared with other image-to-image translation (I2I) networks. The former consistently outperform the latter in downstream tasks such as segmentation and detection.
- In the sim-to-real scenario, the performance of SSL-RGB2IR is compared with other image-to-image (I2I) models. IR images are generated for both synthetic and real datasets. The baseline segmentation model trained using IR images from SSL-RGB2IR outperforms both the baseline and other image-to-image translation networks. This proves the superior quality of the generated images produced by SSL-RGB2IR.

II. Related Work

A. Paired EO to IR image translation

Numerous previous studies have tackled the task of image translation from both RGB to IR and vice versa [29].

Isola *et al.* [9] initially introduced image-to-image translation using Conditional Adversarial Networks (Pix2Pix), focusing on paired image to image translation for multiple tasks including day to night and labels to street scene. Kniaz *et al.* [12] proposed ThermalGAN, utilizing thermal histograms and feature descriptors as a thermal signature for IR generation of people for the task of person reidentification, rather than using the entire RGB image. Ozkanöglu *et al.* [18] proposed InfraGAN, introduced structural similarity as an additional loss function to enhance the generator and introduced pixel-level comparison for the discriminator, resulting in a +8% improvement in performance over ThermalGAN for the VEDAI dataset [20]. Although these methods focus on generating IR from RGB and compare metrics including Structural Similarity Index Measure (SSIM) and Peak signal-to-noise ratio (PSNR), no method evaluates the performance of generated IR for downstream tasks such as detection to compare performance against real IR.

B. Unpaired EO to IR image translation

Acquiring large amounts of paired EO and IR data can be a significant hurdle in training models for tasks like RGB to IR conversion using paired image translation methods. Zhu *et al.* [30] addressed this challenge by introducing CycleGAN, a method capable of learning image translation between domains even without paired examples. Their approach relied on a cycle consistency loss to maintain the desired style during translation.

Building on this concept, researchers have explored utilizing both CycleGAN and Pix2Pix for training EO-IR networks. Lichao *et al.* [29] translated RGB videos into TIR using Pix2Pix and CycleGAN, then computed the Euclidean distance between the translations and the TIR ground-truth images. These studies demonstrated the superiority of Pix2Pix for EO to IR translation, emphasizing the value of a paired training signal compared to the unpaired, despite the larger unpaired dataset. These findings underscore the potential limitations of unpaired training for EO-IR tasks. Dong *et al.* [14] proposed Edge-guided Multi-domain RGB-to-TIR Image Translation in view of these limitations. Their unsupervised learning approach incorporated a style vector to generate realistic TIR images with minimal artifacts while preserving key image details. A novel edge-guided loss was introduced to retain the essential dynamic details in the translated TIR image. However, a common shortcoming among these existing methods is the tendency for poor generalization on unseen images, often resulting in a loss of key details in the translated image. This is particularly true for methods relying solely on paired image translation models, leading to unreliable generated IR data. Furthermore, very few of the models have tested the generated IR for object detection tasks.

Our proposed model addresses these limitations by: **Preserving key features of real IR**: Our approach aims to generate translated IR images that closely resemble real IR data.

Object detection and segmentation: We evaluate the suitability of generated IR images for object detection and segmentation tasks, demonstrating effectiveness in replicating real IR functionality.

Enabling broader applications: Through the generation of reliable EO-IR pairs, our model has application in relevant tasks such as image fusion.

III. Methodology

A. Problem Formulation

Robotic perception can be significantly enhanced by leveraging infrared (IR) images, especially in low-light conditions. Deep learning models often face a decrease in performance when trained on IR data. This is due to the reduced semantic information carried by IR imagery and the relative scarcity of IR data compared to traditional RGB images. Therefore, our objective is to generate IR images from electro-optical (EO) images using image-to-image translation. Given the need for a semi-supervised image-to-image (I2I) translation framework capable of learning from both co-registered and unco-registered images, we introduce a semi-supervised RGB-to-IR model aimed at generating IR data from RGB images. The objective is to learn mapping

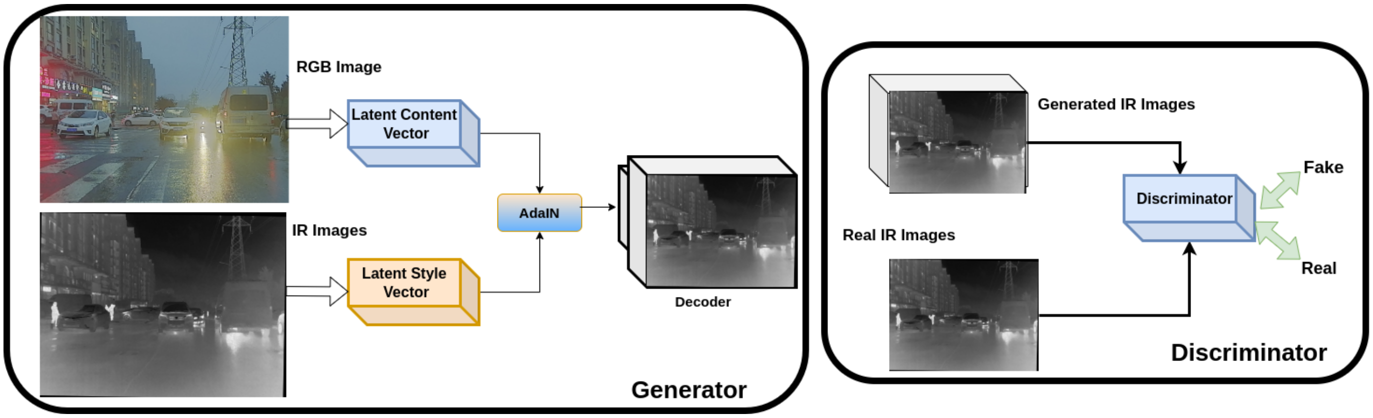

Fig. 1. The pseudo EO-IR pair generation network consists of two encoders within the generator: a content encoder for EO and a style encoder for IR. Each encoder produces a feature vector, which is an input to the adaptive instance normalization. Additionally, the discriminator is responsible for predicting whether the generated images are real or fake.

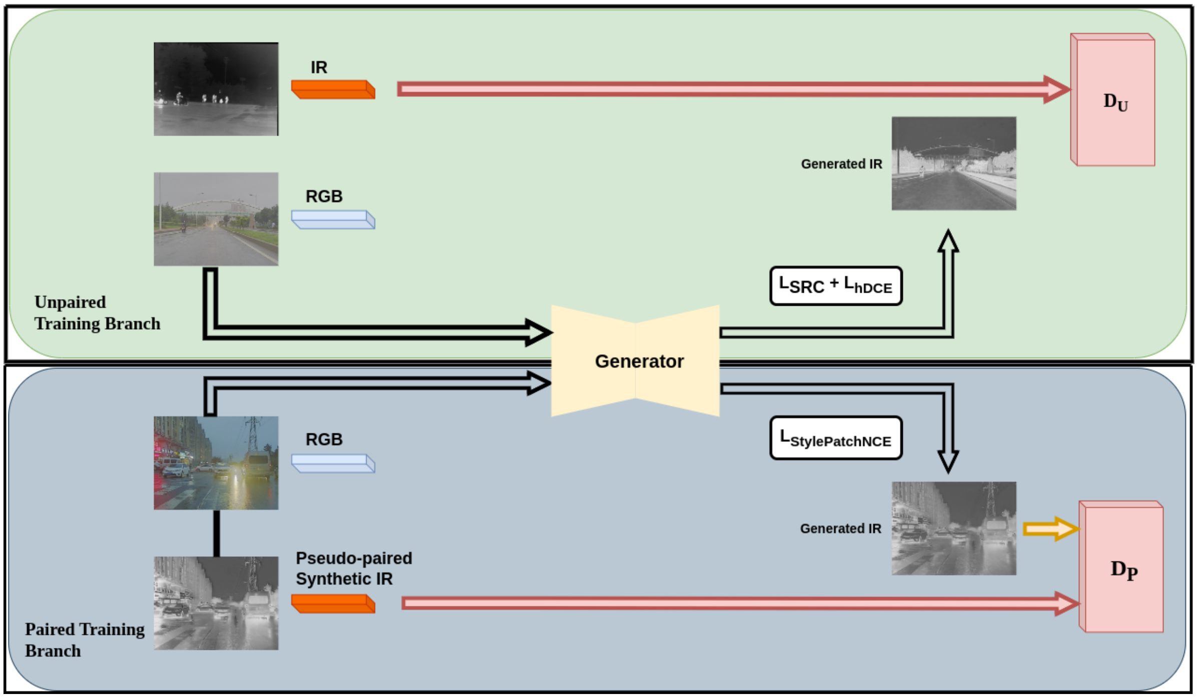

Fig. 2. Framework of SSL-RGB2IR: It includes a single generator shared by both branches, along with discriminators specifically assigned to each branch. It consists of supervised and unsupervised branches. The unsupervised (unpaired) training branch is focused on maintaining the pixel-level details present in the input RGB images when converting them into real IR images. The supervised (paired) training branch acquires the target style from the pseudo-paired images generated during the training process.

$G : X \rightarrow Y$, where X denotes the set of RGB images and Y denotes unpaired IR images, with the help of the pseudo paired dataset $P = \{x_i^p, y_i^p\}_{i=1}^N$.

IV. SSL-RGB2IR: Semi-supervised RGB-to-IR Image-to-Image Translation network

A. Pseudo-pair generation

Generating pseudo-pairs is important for minimizing the domain gap between EO and IR images. For this, the GAN-based unpaired I2I translation model [8] is used. As shown in Fig. 1, the generator has two encoders, one for EO and one for IR, followed by a decoder. The style and content encoders are responsible for generating style and content vectors. Specifically, the content encoder focuses on capturing details such as the edges and textures of the EO image, while the style encoder captures the IR characteristics. The model undergoes training using three types of reconstruction losses [14]. To reconstruct the original RGB image, L^{RGB} is defined as follows:

$$L^{RGB} = E[|G_{RGB}(E^c_{RGB}, E^s_{RGB}) - x_{RGB}|] \qquad (1)$$

where G_{RGB} is the decoder, E^c, E^s and x_{RGB} denote the content, style encoders for RGB, and input RGB image. The content reconstruction loss calculates the disparity between

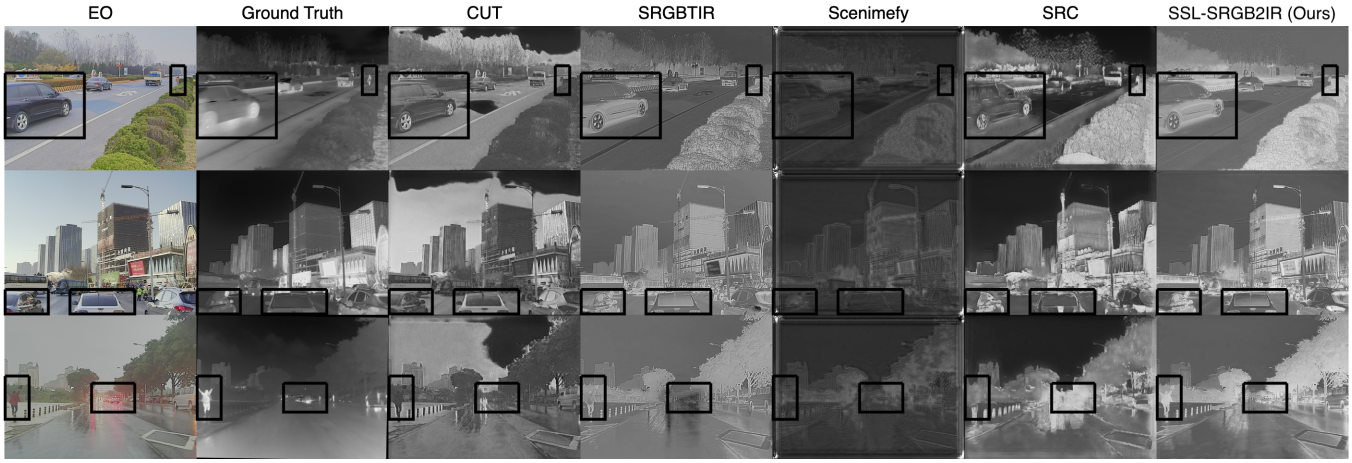

Fig. 3. Comparison of models on RGB to IR image translation on real images from the M3FD dataset. Features of IR images are most accurately portrayed in the proposed model, particularly outperforming baselines in the representation of persons and other key foreground objects crucial to downstream task performance. Objects and regions of particular interest are highlighted across images, demonstrating the superior performance of the proposed model.

the content vector of the RGB image and that of the generated IR image, and is given as,

$$L_c^{RGB} = E[|E_{IR}^C(G_{IR}(C_{RGB}, V_{IR})) - C_{RGB}|] \tag{2}$$

where E_{IR} and G_{IR} denote the content encoder and IR decoder. V_{IR} and C_{RGB} denote the content and style vectors respectively. The style reconstruction loss measures the distinction between the style vector of the original IR image and the generated IR image.

$$L_S^{IR} = E[|E_{IR}^S(G_{IR}(C_{RGB}, V_{IR})) - V_{RGB}|] \tag{3}$$

where E_{IR}^S denotes the style encoder. V_{IR} and C_{RGB} denote the content vector of RGB and the style vector of IR, respectively. The network's overall training objectives are outlined as follows:

$$L_G = L_c^{RGB} + L_S^{IR} + L^{RGB} + L_{GAN} + L_{cyc} \tag{4}$$

where L_{GAN} and L_{cyc} represent the adversarial loss [5] and cyclic loss [8], respectively. Through the generation of pseudo-paired data, we acquire synthetic EO-IR paired images. After the paired samples are generated, they are fed into the SSL-RGB2IR model, which comprises both supervised and unsupervised branches.

B. Supervised Training Branch

In the supervised training branch, the pseudo-paired EO-IR images (X^P, Y^P) are used for semantic mapping across diverse scenes. The reconstruction losses rely on Y^P, rather than real-IR images, thus, resultant generated images resemble the pseudo pairs more strongly. This branch is trained using a conditional GAN framework along with StylePatchNCE loss [10]. The supervised branch operates within a conditional Generative Adversarial Network (GAN) framework [17], employing the conditional adversarial loss:

$$L_{cGAN}(G, D_P) = E_{x,y}[logD_P(x,y)] + E_x[log(1-D_P(x, G(x))] \tag{5}$$

where D_P aims to distinguish between (x, y) and $(G(x), x)$. The StylePatchNCE loss [10] is a patch-level contrastive learning loss used to assist the model in capturing local style information and finer details. This loss is useful in generating IR images and preserving the actual IR characteristics. The training loss for the supervised branch is given by:

$$L_{sup} = L_{cGAN}(G, D_P) + L_{StylePatchNCE}(G, F, Y^P) \tag{6}$$

where G represents the generator, Y denotes the pseudo-label pair, and F denotes the projection head used to train the contrastive learning loss in StylePatchNCE.

C. Unsupervised Training Branch

This branch uses the original high-quality EO-IR (real) dataset to learn the actual target distribution. As seen from Fig. 2 the patches from EO and IR have diverse semantic information, which can be used for unsupervised stylization. The semantic relation consistency loss L_{SRC} [11] and the hard negative contrastive loss L_{hDCE} [11] are used. L_{hDCE} uses a patch-wise contrastive loss mechanism, which progressively elevates the discriminative challenge posed by negative samples, thereby strengthening the model's discriminative capabilities. The training loss for the unsupervised branch is given by:

$$L_{unsup} = L_{GAN}(G, D_U) + L_{SRC} + L_{hDCE} \tag{7}$$

where D_U is the discriminator.

D. Overall Training of SSL-RGB2IR

To train SSL-RGB2IR in a semi-supervised setting, the loss function is defined as:

$$L_{SSL-RGB2IR} = L_{unsup} + L_{sup} \tag{8}$$

In both supervised and unsupervised training scenarios, the Generator G remains the same, while the discriminators are different.

V. Experimental Results

A. Experimental Setup

The effectiveness of the proposed SSL-RGB2IR model is evaluated against other state-of-the-art techniques in image-to-image translation, including edge-guided multi-domain image translation (SRGBTIR) [14], CUT [19], SRC [11] and Scenimefy [10]. To evaluate the performance of deep learning models on downstream tasks of object detection and semantic segmentation, the models are trained using the infrared (IR) images generated using image-to-image translation techniques. All models are evaluated on the official test split. Object detection performance is evaluated on the M3FD dataset [16] using the YOLOv5s and Mask-RCNN models. Semantic segmentation performance is evaluated on the MSRS [26] dataset using the DeepLabV3+. In Sim-2-Real scenarios [27], models undergo training using generated infrared (IR) images from the GTAV dataset [21] and some generated IR samples from the MSRS dataset. Following training, the models are evaluated on the official (real) IR data from the MSRS dataset. For segmentation, the evaluation metric used is the Intersection over Union (IoU), while for object detection, the evaluation metric used is the average Mean Average Precision (mAP).

Datasets: Three public datasets are used for benchmarking, i.e., M3FD datasets [16], MSRS [26], and GTAV [21]. The M3FD dataset comprises approximately 4,200 EO-IR image pairs, each with a resolution of 1024 x 768 pixels with 6 categories. The MSRS dataset consists of 1,444 EO-IR image pairs, each with an image size of 480 × 640 and containing 9 classes. The dataset is divided randomly, with 1,083 samples for training and 361 for testing. GTAV consists of 24,966 driving-scene images generated directly from the Grand Theft Auto V game engine. It contains 19 distinct object categories. For the Sim–Real scenario, 840 images have been randomly selected from the training set, which originally consists of 12,403 images. These selected images have a resolution of 1914×1052 pixels.

Implementation Details: For the object detection task, YOLOV5s, and Mask-RCNN models are used, with backbones pre-trained on ImageNet. The YOLOv5 model is trained for 100 epochs with a batch size of 16. Augmentation techniques, such as random rotation by 10 degrees, translation, and random scaling, are applied with a probability of 0.5. The Mask-RCNN model is trained for 24 epochs with a batch size of 4, with a learning rate of 10^{-3}, with a polynomial scheduler. No augmentations are used while training the Mask-RCNN models. Both models are trained using an SDG optimizer. For the semantic segmentation task, the DeepLabV3+ model with an EfficientNet-B3 backbone pre-trained on ImageNet is employed. It is trained utilizing an SGD optimizer with a learning rate of 5×10^{-3}. All models undergo training for 200 epochs, with a batch size of 8. The experiments are performed using an NVIDIA Quadro RTX 5000 GPU.

TABLE I
Performance comparison of object detection models of mAP (%) of SSL-RGB2IR with other state-of-the-art models on M3FD dataset.

I2I Translation models	YOLOV5s (% mAP)	Mask-RCNN
CUT	15.01	6.066
SRGBTIR	18.5	10.734
Scenimefy	11.1	6.47
SRC	17.7	11.66
SSL-RGB2IR (Ours)	**19.4**	**14.69**

TABLE II
Performance comparison of semantic segmentation models of mIoU (%) of SSL-RGB2IR with other state-of-the-art models on MSRS dataset.

Model	Publication	Deeplab V3+ (% IoU)
CUT	ECCV 2020	29.68
SRGBTIR	ICRA 2023	31.88
Scenimefy	ICCV 2023	26.25
SRC	CVPR 2022	30.78
SSL-RGB2IR (Ours)		**32.36**

B. Quantitative Evaluation

Table I presents the object detection performance of SSL-RGB2IR and other image-to-image translation techniques on standard detection models YOLOv5s and Mask-RCNN. Both models demonstrate superior performance when trained on the generated IR data from SSL-RGB2IR. This results in an improvement of 8.3% and 8.22% compared to the contemporary semi-supervised model Scenimefy for YOLOv5s and Mask-RCNN, respectively. Additionally, there is an improvement of 0.9% and 3.95% compared to SRGBTIR on both models. Table II displays the segmentation performance of the DeepLabV3+ model on generated IR images from SSL-RGB2IR and other image-to-image translation models. The results demonstrate that the model trained with SSL-RGB2IR images surpasses all other models in performance. Specifically, it outperforms the Scenimefy model by 6.11% and the SRGBTIR model by 0.48%. Table III presents the segmentation performance of the DeepLabV3+ model when trained on generated IR data from both the GTAV and MSRS datasets. The updated training dataset comprises 840 images from GTAV and 360 images from the MSRS dataset. The trained model's performance is evaluated on the real test set of the MSRS dataset. The baseline model refers to the DeepLabV3+ model trained exclusively on the 360 images from the MSRS dataset. As observed, using simulated images enhances the performance compared to using only real images. The model trained with SSL-RGB2IR generated IR data outperforms all the other models. It outperforms Scenimefy by 3.5% and SRGBTIR by 2.88%. This illustrates the superior IR generation capability of SSL-RGB2IR for both real and simulated datasets.

C. Qualitative Evaluation

For qualitative evaluation, the generated IR images produced by SSL-RGB2IR are compared with those generated by other models. Figures 3 and 4 depict the generated images

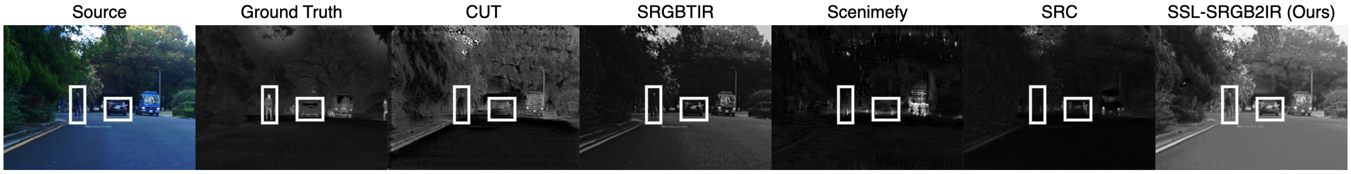

Fig. 4. Comparison of the baseline and the proposed method on images from the MSRS dataset. Baseline methods fail to preserve details from RGB images and result in ambiguous boundaries in generated IR images.

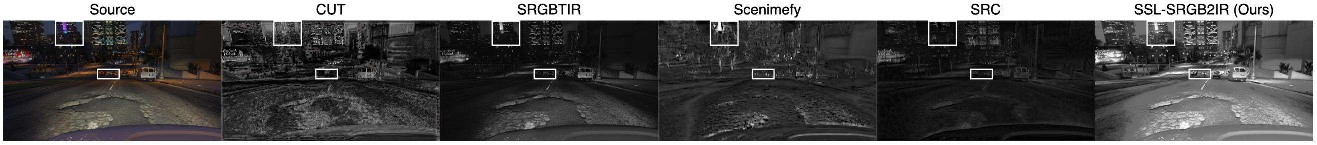

Fig. 5. Comparison of model performance on generalization to synthetic RGB data from the GTAV dataset. The proposed model is much better at preserving structural information and texture details from the input RGB images, while maintaining a style consistent with IR imagery. Further, IR images generated by the proposed model provides the largest boost to accuracy on the downstream task of semantic segmentation.

TABLE III

PERFORMANCE COMPARISON OF SEMANTIC SEGMENTATION MODELS OF MIOU (%) OF SSL-RGB2IR WITH OTHER STATE-OF-THE-ART MODELS FOR SIM-2-REAL SCENARIO.

I2I Translation models	Publication	Deeplab V3+ (% IoU)
Baseline	-	46.85
CUT	ECCV 2020	49.79
SRGBTIR	ICRA 2023	48.17
Scenimefy	ICCV 2023	47.55
SRC	CVPR 2022	48.13
SSL-RGB2IR (Ours)		**51.05**

TABLE IV

PERFORMANCE COMPARISON OF MIOU (%) OF SSL-RGB2IR WITH AND WITHOUT SUPER-RESOLUTION.

Method	DeepLabV3+ (On MSRS)	YOLOV5s (On M3FD)
SSL-RGB2IR + Super-resolution	29.88	12.10
SSL-RGB2IR	**32.36**	**19.4**

of M3FD and MSRS datasets, respectively. SSL-RGB2IR demonstrates its ability to preserve semantic information from RGB images, maintaining edge consistency and demonstrating consistency with IR characteristics, particularly for persons as can be seen in Figure 3. Figure 5 illustrates the generated IR images sourced from the GTAV dataset. The generated images retain pixel-level details present in the original images (not preserved by other models), as shown by features such as cars and buildings, while translating to a style approximating that of infrared images.

D. Ablation Studies

We conduct an ablation study to investigate the impact of super-resolution in SSL-RGB2IR to potentially remove the sensor noise. The EO-IR images are first passed through SR-GAN [13] and then subsequently fed into SSL-RGB2IR. Table IV presents the performance comparison between SSL-RGB2IR and SSL-RGB2IR with super-resolution (SSL-RGB2IR + Super-resolution). Using super-resolution results in a performance decrease for both the downstream tasks of segmentation on the MSRS dataset and object detection on the M3FD dataset.

VI. CONCLUSIONS

This paper presents a Semi-supervised RGB-to-IR Image-to-Image Translation model (SSL-RGB2IR) designed to generate IR images from EO images. The SSL-RGB2IR model consists of two main components: pseudo-pair generation and a semi-supervised learning framework. The pseudo-pair generation module is responsible for generating synthetic EO-IR pairs. Subsequently, the semi-supervised learning framework comprises a supervised learning branch, which uses these synthetic EO-IR pairs, followed by an unsupervised learning branch, using real EO and IR images. The generated infrared (IR) images are of high quality and effectively retain the characteristics typical of IR imagery. Deep learning models consistently achieve superior performance when trained on IR images generated by SSL-RGB2IR, compared to other I2I translation networks. The DeepLabV3+ model, when trained on IR data synthesized by SSL-RGB2IR, demonstrates a 6.11% improvement in segmentation accuracy compared to Scenimefy. In the detection task, both YOLOV5s and Mask-RCNN models, trained on images generated by SSL-RGB2IR, outperform models trained on Scenimefy-generated images by 8.3% and 8.22%, respectively. In the sim-2-real setting, the DeepLabV3+ model trained on IR data generated by SSL-RGB2IR surpasses the baseline model by 4.2% and outperforms Scenimefy-generated images by 3.5%.

VII. ACKNOWLEDGEMENT

This work was supported by SwaYaan, Ministry of Electronics and Information Technology (MeitY) and the Centre for Airborne Systems (CABS). We also appreciate the ongoing support from the Kotak IISc AI-ML Centre (KIAC).

REFERENCES

[1] Prahlad Anand, Qiranul Saadiyean, Aniruddh Sikdar, Suresh Sundaram, et al. Supervised image translation from visible to infrared domain for object detection. *arXiv preprint arXiv:2408.01843*, 2024.

[2] Jayabrata Chowdhury, Venkataramanan Shivaraman, Suresh Sundaram, and PB Sujit. Graph-based prediction and planning policy network (gp3net) for scalable self-driving in dynamic environments using deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11606–11614, 2024.

[3] Manash Pratim Das, Larry Matthies, and Shreyansh Daftry. Online photometric calibration of automatic gain thermal infrared cameras. *IEEE Robotics and Automation Letters*, 6(2):2453–2460, 2021.

[4] Chaitanya Devaguptapu, Ninad Akolekar, Manuj M Sharma, and Vineeth N Balasubramanian. Borrow from anywhere: Pseudo multimodal object detection in thermal imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[6] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115. IEEE, 2017.

[7] Yu Hou, Rebekka Volk, and Lucio Soibelman. A novel building temperature simulation approach driven by expanding semantic segmentation training datasets with synthetic aerial thermal images. *Energies*, 14(2):353, 2021.

[8] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[10] Yuxin Jiang, Liming Jiang, Shuai Yang, and Chen Change Loy. Scenimefy: learning to craft anime scene via semi-supervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7357–7367, 2023.

[11] Chanyong Jung, Gihyun Kwon, and Jong Chul Ye. Exploring patchwise semantic relation for contrastive learning in image-to-image translation tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18260–18269, 2022.

[12] Vladimir V Kniaz, Vladimir A Knyaz, Jiri Hladuvka, Walter G Kropatsch, and Vladimir Mizginov. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[13] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[14] Dong-Guw Lee, Myung-Hwan Jeon, Younggun Cho, and Ayoung Kim. 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)*, pages 8291–8298. IEEE, 2023.

[15] Chenglong Li, Wei Xia, Yan Yan, Bin Luo, and Jin Tang. Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(7):3069–3082, 2020.

[16] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5811, 2022.

[17] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[18] Mehmet Akif Özkanoğlu and Sedat Ozer. Infragan: A gan architecture to transfer visible images to infrared domain. *Pattern Recognition Letters*, 155:69–76, 2022.

[19] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020.

[20] Sebastien Razakarivony and Frederic Jurie. Vehicle detection in aerial imagery: A small target detection benchmark. *Journal of Visual Communication and Image Representation*, 34:187–203, 2016.

[21] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016.

[22] Qiranul Saadiyean, SP Samprithi, and Suresh Sundaram. Learning multi-scale context mask-rcnn network for slant angled aerial imagery in instance segmentation in a sim2real setup. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13573–13580. IEEE, 2024.

[23] Aniruddh Sikdar, Jayant Teotia, and Suresh Sundaram. Skd-net: Spectral-based knowledge distillation in low-light thermal imagery for robotic perception. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9041–9047. IEEE, 2024.

[24] Aniruddh Sikdar, Sumanth Udupa, Prajwal Gurunath, and Suresh Sundaram. Deepmao: Deep multi-scale aware overcomplete network for building segmentation in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 487–496, 2023.

[25] Aniruddh Sikdar, Sumanth Udupa, Suresh Sundaram, and Narasimhan Sundararajan. Fully complex-valued fully convolutional multi-feature fusion network (fc 2 mfn) for building segmentation of insar images. In *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 581–587. IEEE, 2022.

[26] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83:79–92, 2022.

[27] Sumanth Udupa, Prajwal Gurunath, Aniruddh Sikdar, and Suresh Sundaram. Mrfp: Learning generalizable semantic segmentation from sim-2-real with multi-resolution feature perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5904–5914, 2024.

[28] Abhinav Valada, Rohit Mohan, and Wolfram Burgard. Self-supervised model adaptation for multimodal semantic segmentation. *International Journal of Computer Vision*, 128(5):1239–1285, 2020.

[29] Lichao Zhang, Abel Gonzalez-Garcia, Joost Van De Weijer, Martin Danelljan, and Fahad Shahbaz Khan. Synthetic data generation for end-to-end thermal infrared tracking. *IEEE Transactions on Image Processing*, 28(4):1837–1850, 2018.

[30] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.