Thermal Image Synthesis: Bridging the Gap Between Visible and Infrared Spectrum

Patricia L. Suárez[1(✉)] and Angel D. Sappa[1,2]

[1] ESPOL Polytechnic University, FIEC, CIDIS, Guayaquil, Ecuador
{plsuarez,asappa}@espol.edu.ec
[2] Computer Vision Center, 08193 Bellaterra, Barcelona, Spain
asappa@cvc.uab.es

Abstract. This paper proposes a novel approach for obtaining high-quality thermal image-like representations that can be used as inputs in various thermal image compressive sensing applications. We address the challenge of low-resolution/quality thermal images by generating synthetic thermal image representations using a contrastive cycled GAN network from low-cost visible images. These representations can then be used to improve the quality of low-quality thermal images of the same scene. Experimental results demonstrate the effectiveness of the proposed approach on different datasets.

Keywords: Synthesizing Thermal Images · Contrastive Loss · Relativistic Loss · Spectral Normalization

1 Introduction

Thermal sensors are playing an increasingly important role in enhancing the accuracy and reliability of computer vision applications. However, they do come with limitations, particularly the lower resolution of thermal images compared to standard visual imaging techniques. High-resolution thermal cameras exist that can overcome this limitation, but they are cost-prohibitive, making them less accessible for widespread use [24]. Additionally, the nature of thermal imaging requires different manipulation and processing techniques compared to visible spectrum images, further complicating its integration into existing systems.

To address these issues, researchers have been exploring innovative ways to leverage the benefits of using thermal imaging while mitigating its limitations through thermal image super-resolution algorithms (e.g. [16–18]). Another

G. Bebis et al. (Eds.): ISVC 2024, LNCS 15046, pp. 384–396, 2025.
https://doi.org/10.1007/978-3-031-77392-1_29

promising approach to tackle this limitation is by means of data fusion strategies, where information from multiple sensors operating in different spectral bands are merged to obtain a better resolution representation. This multispectral imaging strategy allows for the integration of thermal and visible light data, providing a more complete view that can improve the interpretation and utility of the produced images [19]. Deep learning plays a key role in this research, offering new ways to process and analyze large data sets quickly and accurately. These technologies are essential for applications that require real-time decision-making, such as autonomous vehicle navigation, and predictive maintenance in industries, among others [26].

One promising approach for enhancing thermal images is to incorporate visible images as guidance information. This is known as guidance image processing and has been explored in various studies [22]. However, there is a need to develop a more sophisticated approach to generate high-quality thermal images that take into account the different spectral bands captured by the sensors [13]. Some approaches utilize feature-level guidance rather than image pixel-level guidance, using edges from one image to enhance the other image. This edge-based guiding technique enables the reconstruction of higher-frequency features, as demonstrated in previous works such as [27,29].

The majority of the methods mentioned above utilize deep learning approaches, specifically Convolutional Neural Networks (CNNs), which outperform traditional methods in terms of efficiency. One of the limitations of CNNs is the substantial amount of data required for training, especially when dealing with paired images (thermal and visible spectrum). To address these limitations, in the current work a model capable of generating synthetic thermal images from visible images without the need for paired data is proposed. The ability to generate synthetic thermal images from visible images without the need for paired data has numerous practical applications, such as in surveillance systems, search and rescue operations, and medical imaging. The contribution of this paper can be highlighted in:

- Implement a modified contrastive loss [11] to enhance the features obtained from the generator. The implementation of this loss function enables the model to focus on image regions with high affinity while disregarding those with low affinity in the latent spaces
- Apply spectral normalization [14], to improve the style of the synthetic images and the vanished gradients problem during the training of the model.
- Introduce the relativistic GAN loss function, to enhance the stability, and convergence of the generator to produce high-quality synthetic thermal images that closely resemble real ones.

The manuscript is organized as follows. Section 2 presents works related to the generation of synthetic images. Section 3 presents the proposed cycled GAN-modified architecture. Experimental results and comparisons are given in Sect. 4. Finally, conclusions are presented in Sect. 5.

2 Related Work

Image synthesis has been a popular area of research in computer vision, and deep learning-based techniques have shown remarkable success in this field. Different CNN models have been proposed to generate synthetic images and their further applications; in [28] the authors propose using synthetic thermal images to train a tracking model. By transforming paired and unpaired images, the tracking model with thermal images achieves improved results. Among the various deep learning models, generative models have gained significant attention in recent years for their ability to generate high-quality images from a given distribution. One of the most widely used generative model is the generative adversarial network (GAN), which has proven to be highly effective in generating realistic images, [21,23]. GANs consist of a generator network that learns to generate images and a discriminator network that learns to distinguish between real and generated images. The approach proposed in [6] aims to address the limitations of current data sets for pedestrian detection in thermal imaging scenarios by generating synthetic thermal images from their visible counterparts using domain matching.

Trying to overcome the limitation of paired images, [12] proposes the usage of a cycled GAN network to translate visible images to the thermal domain while maintaining consistency using a disparity map. Also in [3] the authors propose a method for learning to count the number of pedestrians by utilizing synthetic images instead of real ones. Similarly, using deep learning the approach proposed in [11] aims to enhance the scene context in night vision applications utilizing a GAN network to map context information. Also, another approach that uses synthetic data is presented in [25] where the authors introduce a novel approach to procedural world modeling, offering physically accurate and highly variable image synthesis for real-time applications. In [9] a semantic image segmentation technique is presented; the authors propose to address lighting and environmental limitations by incorporating both real and synthetic thermal infrared camera images to guide contour extraction. Following the synthetic image generator approaches, in [2] a new method is proposed to use synthetic images to generate an almost unlimited dataset for depth training. On the other hand, in the work presented in [8], a cross-modal generative network is introduced to generate synthetic thermal images for training a people re-identification model. This approach utilizes thermal images and introduces object notations to enhance the accuracy of the re-identification results.

A more recent generative model is the diffusion model (DM), which has gained significant popularity due to its ability to generate high-resolution images with fine details. Diffusion models operate by iteratively adding noise to a partially generated image and then removing the noise using a sequence of learned transformations. Another work related to image synthesizing is presented in [20], where the authors discuss the use of diffusion models in image synthesis and their ability to control the image generation process without retraining. They also introduce cross-attention layers to allow for general conditioning inputs such as text or bounding boxes.

3 Proposed Approach

The method proposed in the current work for obtaining synthetic thermal images from visible spectrum images combines several state-of-the-art techniques. The architecture used in this approach is based on a cycled GAN, which is a method for transferring unpaired domains, as described in [30]. The current work builds on the seminal research by refining the generator process to achieve a more accurate translation of pixel intensity in the far-infrared spectrum. A contrastive loss is also proposed, to improve the quality of the obtained images. The contrastive loss, as presented in [11], allows the architecture to establish the correlation between input embeddings that are in proximity to the region being processed. By using cosine similarity to determine the similarity of nearby regions, the loss method can determine differences based on their orientation, rather than just their magnitude, as is the case with the L1 loss.

To obtain synthetic thermal images from visible spectrum images, the choice of color space is particularly crucial, since it directly affects the model's ability to accurately simulate the temperatures of the objects presented in the images. In experiments performed with the proposed approach, it could be determined that changing the RGB color space of the input images to HSV led to significantly improved results. Specifically, in the HSV color space, the brightness channel was found to be particularly useful for accurately transforming visible information into thermal, resulting in a high-fidelity representation of temperatures and maintaining good contours and details in generated thermal images.

The current work also incorporates a relativistic GAN loss, proposed by [7], in place of the traditional GAN loss suggested by [5]. The relativistic GAN loss considers that in each mini-batch, at least 50% of the generated data are false. The learning divergence is then minimized based on this assumption. This approach is beneficial because it enables us to estimate that in a mini-batch of randomly generated data, there are more realistic samples than false ones. This leads to better training of the GAN, being a more stable training process, and improved image quality; and consequently, more accurate synthetic thermal images. Specifically, the relativistic loss is defined as follows:

Specifically, the relativistic loss is defined as follows:

$$\mathcal{L}_G^{RGAN} = \mathbb{E}_{(x,y)\sim(\mathbb{P},\mathbb{Q})}\left[g\left(C\left(y\right) - C\left(x\right)\right)\right], \tag{1}$$

$$\mathcal{L}_D^{RGAN} = \mathbb{E}_{(x,y)\sim(\mathbb{P},\mathbb{Q})}\left[f\left(C\left(x\right) - C\left(y\right)\right)\right], \tag{2}$$

where, $\mathbb{E}_{(x,y)\sim(\mathbb{P},\mathbb{Q})}$ corresponds to the expectation over the real x data sampled from the distribution $\mathbb{P}$ and the fake y data sampled from the distribution $\mathbb{Q}$. The $f\left(C\left(x\right) - C\left(y\right)\right)$ is the function that measures the difference between the scores of the real and fake data for the discriminator and $g\left(C\left(y\right) - C\left(x\right)\right)$ is the function that measures the difference between the scores of the fake and real data for the generator.

The architecture uses a contrastive loss function, which helps the model learn the similarities between the latent spaces it generates. This approach is based on the principles outlined in [1].

This loss promotes the grouping of similar representations while ensuring that dissimilar ones are distinctly separated. Contrastive learning requires defining two distributions: A positive input distribution, $\mathbf{x}^+ \sim p^+(\cdot \mid \mathbf{x})$, which samples inputs that are similar to a given input image $\mathbf{x}$ and a negative input distribution, $\mathbf{x}^- \sim p^-(\cdot \mid \mathbf{x})$, which samples inputs that are different from the given input image $\mathbf{x}$. The proposed loss function can be written as:

$$\mathcal{L}_{\text{contr}}(\hat{Y}, Y) = \sum_{l=1}^{L} \sum_{s=1}^{S_l} \ell_{\text{contr}} \ (\hat{v}_l^s, v_l^s, \bar{v}_l^s) \,. \tag{3}$$

This function computes the contrastive loss by comparing the predicted feature vectors $\hat{v}_l^s$ with the true feature vectors v_l^s and their corresponding sets of other feature vectors $\bar{v}_l^s$. According to the authors in [1], the shape of the tensor $V_l \in \mathbb{R}^{S_l \times D_l}$ is determined by the network architecture, where S_l is the number of spatial locations in the tensor. $v_l^s \in \mathbb{R}^{D_l}$ represents the feature vector at the s^{th} spatial location and $\bar{v}_l^s \in \mathbb{R}^{(S_l-1)\times D_l}$ represents the collection of feature vectors at all other spatial locations except s.

To prevent the intensity levels of the pixels from exceeding the objective domain's bounds during the data transformation process, the model also employs the identity loss function. This means that the generative network must retain the most important characteristics, such as the thermal intensity level and object shape, while maintaining the formation model's stability. Specifically, the generative network must ensure that $G(x) \approx x$ and $F(y) \approx y$.

$$\begin{aligned} \mathcal{L}_{\text{identity}}\,(G, F) = {} & E_{x \sim p_{\text{data}}(x)}[\|G(x) - x\|] \\ & + E_{y \sim p_{\text{data}}(y)}[\|F(y) - y\|], \end{aligned} \tag{4}$$

where, G and F are the generative networks, x and y are samples from the data distributions $p_{\text{data}}(x)$ and $p_{\text{data}}(y)$ respectively,

The final loss function (L_{final}) employed in our model is a combination of previous loss components and it can be expressed as follows:

$$\begin{aligned} L_{\text{final}} = {} & \mathcal{L}_{\text{RGAN}}(G, D) + \lambda_X \mathcal{L}_{\text{contr}}(G, F) \\ & + \lambda_Y \mathcal{L}_{\text{contr}}(F, G) + \gamma \mathcal{L}_{\text{identity}}(G, F), \end{aligned} \tag{5}$$

here, λ_X and λ_Y are the weights assigned to the contrastive loss function for the two domains X and Y, respectively. And γ is the weight assigned to the identity loss function. These values have been empirically determined based on the outcomes of the experiments. The contrastive loss component $\mathcal{L}_{\text{contr}}$ measures the similarity of the latent spaces produced by the generator networks G and F on the embedding network for corresponding input images. The identity loss component $\mathcal{L}_{\text{identity}}$ ensures that the intensity levels of the pixels do not deviate significantly from the original domain during the transformation process. The

weight of the identity loss function, γ, is also defined empirically. Finally, the overall objective function of the model is denoted by $\mathcal{L}_{\text{RGAN}}$ and is optimized jointly with the adversarial loss function between the generator G and the discriminator D.

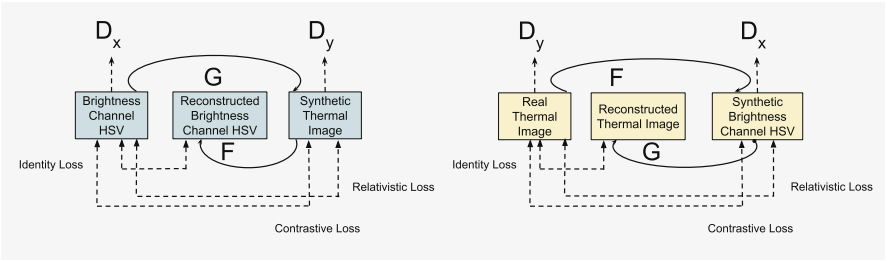

Fig. 1. Cycled GAN architecture proposed.

The architecture (depicted in Fig. 1) incorporates spectral normalization to enhance the quality of the synthetic thermal images generated by the model.

4 Experimental Results

This section presents results of the proposed approach, both quantitatively and qualitatively. Additionally, the data set used for training and the pre-processing techniques applied to the images are described. Finally, a comparative analysis is conducted using the structural similarity index (SSIM) and the peak signal-to-noise ratio (PSNR) to evaluate the performance of the model in generating synthetic images.

4.1 Training Settings

The model training process involves several steps to achieve the desired outcome. Initially, the visible images are converted to the HSV color space, and only the brightness channel is considered for model training. Furthermore, the images are resized to 256 $\times$ 256 pixels during the training process. The model is trained with a learning rate of 0.000273 using the Adam optimizer. Another parameter configured is the value of 0.73 for the β_1 parameter to accelerate the model convergence and enhance the image quality results. To quantitatively evaluate the performance of the model, PSNR and SSIM are selected as metrics. The training process utilizes a TITAN V GPU, and it takes approximately 96 h to complete.

4.2 Datasets

The M3FD data set [10] has been utilized to train the proposed model. The data set was captured using a binocular optical and infrared sensor and consists of 4,500 image pairs of outdoor scenes. For training, 3,000 image pairs were used while 890 pairs were used for testing and the remaining images for validation of the trained model. The images were pre-processed to generate realistic synthetic far-infrared images by transferring them to the HSV color space and selecting the V channel for training. To test the model's robustness, the FLIR ADAS V2 dataset [4] with 300 pairs has been used. Also, an additional proprietary data set called Thermal Stereo, which includes 200 pairs of registered visible-thermal images, was used. The model trained with the M3FD data set was evaluated and compared against the results obtained from other experiments.

4.3 Results and Comparisons

The proposed approach is evaluated by comparing it with unpaired image translation models, as described in papers Zhu et al. [30] and Park et al. [15]. These models are well known for their ability to generate synthetic images from the visible spectrum to another unpaired domain.

Table 1 presents the average results obtained from the model in [30], the approaches presented in [15], and our model. The evaluation process uses samples from the M3FD, FLIR ADAS V2 datasets as well as our dataset Thermal Stereo consisting of outdoor scenes. The SSIM obtained with each dataset, together with their corresponding PSNR values are depicted. Visual representations of the synthetic thermal images generated from these validation sets are depicted in Figs. 2, 3, and 4 for each testing dataset. The evaluation results demonstrate the effectiveness of the approach in producing high-quality synthetic thermal images, as evidenced by the quantitative metrics and visual comparisons.

Table 1. Average results from the validation sets (M3FD, Thermal Stereo and FLIR ADAS V2 datasets). Best results in **bold**.

Approaches	M3FD		Thermal Stereo		FLIR ADAS V2	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Zhu et al. [30]	12.589	0.501	11.939	0.419	12.031	0.462
CUT [15]	13.391	0.672	12.348	0.537	11.986	0.529
FastCUT [15]	13.901	0.703	13.163	0.673	12.432	0.611
Proposed Approach	**14.734**	**0.772**	**17.098**	**0.733**	**13.133**	**0.691**

Fig. 2. M3FD dataset: (*1st. row*) results from [30]; (*2nd. row*) results from [15]; (*3rd. row*) results from [15]; (*4th. row*) results from the proposed approach; (*5th. row*) ground truth images.

Fig. 3. Thermal Stereo dataset: (*1st. row*) results from [30]; (*2nd. row*) results from [15]; (*3rd. row*) results from [15]; (*4th. row*) results from the proposed approach; (*5th. row*) ground truth images.

Fig. 4. FLIR ADAS V2 dataset: (*1st. row*) results from [30]; (*2nd. row*) results from [15]; (*3rd. row*) results from [15]; (*4th. row*) results from the proposed approach; (*5th. row*) ground truth images.

5 Conclusions

The main contribution of the work is the development of a novel approach for transforming images from the visible spectrum to the far infrared (thermal) spectrum using an unpaired cycled GAN network. Also, modifications have been made to the loss and normalization functions to enable the network to simulate not only the shape but also the temperature and texture of the thermal images, to improve the image transformations. The results show better image quality and fidelity compared to the state-of-the-art methods. In future research, further exploration will be conducted to investigate the utilization of other state-of-the-art techniques for thermal image syntheses, such as StyleGAN or VQ-VAE.

Acknowledgements. This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-24-1-0206; and partially supported by the Grant PID2021-128945NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe"; and by the ESPOL project CIDIS-003-2024. The second author acknowledges the support of the Generalitat de Catalunya CERCA Program to CVC's general activities, and the Departament de Recerca i Universitats from Generalitat de Catalunya with reference 2021SGR 01499.

References

1. Andonian, A., Park, T., Russell, B., Isola, P., Zhu, J.Y., Zhang, R.: Contrastive feature loss for image prediction. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1934–1943 (2021)
2. Carlucci, F.M., Russo, P., Caputo, B.: A deep representation for depth images from synthetic data. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 1362–1369 (2017)
3. Ekbatani, H.K., Pujol, O., Segui, S.: Synthetic data generation for deep learning in counting pedestrians. In: Proceedings of the International Conference on Pattern Recognition Applications and Methods, pp. 318–323 (2017)
4. FLIR, Thermal, D.: FREE Teledyne FLIR thermal dataset for algorithm training. https://www.flir.com. Accessed 07 Nov 2023
5. Goodfellow, I., et al.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
6. Guo, T., Huynh, C.P., Solh, M.: Domain-adaptive pedestrian detection in thermal images. In: Proceedings of the IEEE International Conference on Image Processing, pp. 1660–1664 (2019)
7. Jolicoeur-Martineau, A.: The relativistic discriminator: a key element missing from standard GAN. arXiv preprint arXiv:1807.00734 (2018)
8. Kniaz, V.V., Knyaz, V.A., Hladůvka, J., Kropatsch, W.G., Mizginov, V.: ThermalGAN: multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11134, pp. 606–624. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11024-6_46
9. Li, C., Xia, W., Yan, Y., Luo, B., Tang, J.: Segmenting objects in day and night: edge-conditioned CNN for thermal image semantic segmentation. IEEE Trans. Neural Netw. Learn. Syst. **32**(7), 3069–3082 (2020)

10. Liu, J., et al.: 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 Conference on Computer Vision and Pattern Recognition, pp. 5802–5811 (2022)
11. Liu, R., Ge, Y., Choi, C.L., Wang, X., Li, H.: DivCo: diverse conditional image synthesis via contrastive generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 16377–16386, June 2021
12. Lu, Y., Lu, G.: An alternative of lidar in nighttime: unsupervised depth estimation based on single thermal image. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, pp. 3833–3843 (2021)
13. Mehri, A., Behjati, P., Sappa, A.D.: TnTViT-G: transformer in transformer network for guidance super resolution. IEEE Access **11**, 11529–11540 (2023)
14. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957 (2018)
15. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for conditional image synthesis. In: Proceedings of the European Conference on Computer Vision (2020)
16. Rivadeneira, R.E., Sappa, A.D., Vintimilla, B.X., Hammoud, R.: A novel domain transfer-based approach for unsupervised thermal image super-resolution. Sensors **22**(6), 2254 (2022)
17. Rivadeneira, R.E., et al.: Thermal image super-resolution challenge in PBVs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4359–4367 (2021)
18. Rivadeneira, R.E., et al.: Thermal image super-resolution challenge results in PBVs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3113–3122 (2024)
19. Rivadeneira, R.E., Velesaca, H.O., Sappa, A.: Object detection in very low-resolution thermal images through a guided-based super-resolution approach. In: 17th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pp. 311–318. IEEE (2023)
20. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10684–10695, June 2022
21. Suárez, P.L., Carpio, D., Sappa, A.D.: Depth map estimation from a single 2D image. In: 17th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pp. 347–353. IEEE (2023)
22. Suárez, P.L., Carpio, D., Sappa, A.D.: Enhancement of guided thermal image super-resolution approaches. Neurocomputing **573**, 127197 (2024)
23. Suárez, P.L., Sappa, A.D.: Toward a thermal image-like representation. In: Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP (2023)
24. Teutsch, M., Sappa, A.D., Hammoud, R.I.: Computer vision in the infrared spectrum: challenges and approaches (2021)
25. Tsirikoglou, A., Kronander, J., Wrenninge, M., Unger, J.: Procedural modeling and physically based rendering for synthetic data generation in automotive applications. arXiv preprint arXiv:1710.06270 (2017)
26. Villa, E., Arteaga-Marrero, N., Ruiz-Alzola, J.: Performance assessment of low-cost thermal cameras for medical applications. Sensors **20**(5), 1321 (2020)
27. Xie, J., Feris, R.S., Sun, M.T.: Edge-guided single depth image super resolution. IEEE Trans. Image Process. **25**(1), 428–438 (2015)

28. Zhang, L., Gonzalez-Garcia, A., Van De Weijer, J., Danelljan, M., Khan, F.S.: Synthetic data generation for end-to-end thermal infrared tracking. Trans. Image Process. **28**(4), 1837–1850 (2018)
29. Zhou, D., Wang, R., Lu, J., Zhang, Q.: Depth image super resolution based on edge-guided method. Appl. Sci. **8**(2), 298 (2018)
30. 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)