Synthetic Thermal Image Generation from Multi-Cue Input Data

Patricia L. Suárez[1] [a] and Angel D. Sappa[1,2] [b]
[1]*Escuela Superior Politécnica del Litoral, ESPOL, Facultad de Ingeniería en Electricidad y Computación, CIDIS, Campus Gustavo Galindo Km. 30.5 Vía Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador*
[2]*Computer Vision Center, Edifici O, Campus UAB, 08193 Bellaterra, Barcelona, Spain*
{*plsuarez, asappa*}*@espol.edu.ec, asappa@cvc.uab.es*

Keywords: Synthetic Thermal Images, Generative Approach, Multi-Cue Input Data.

Abstract: This paper presents a novel approach for generating synthetic thermal images using depth and edge maps from the given grayscale image. In this way, the network receives the fused image as input to generate a synthetic thermal representation. By training a generative model with the depth map fused with the corresponding edge representation, the model learns to generate realistic synthetic thermal images. A study on the correlation between different types of inputs shows that depth and edge maps are better correlated than grayscale images, or other options generally used in the state-of-the-art approaches. Experimental results demonstrate that the method outperforms state-of-the-art and produces better-quality synthetic thermal images with improved shape and sharpness. Improvements in results are attributed to the combined use of depth and edge maps together with the novel loss function terms proposed in the current work.

1 INTRODUCTION

The use of thermal images has evolved from its origins in military applications to widespread use in industrial inspection, medical diagnostics, and more recently in areas such as precision agriculture and urban infrastructure management. In industrial inspection, thermal imaging can be used to detect hotspots in electrical and mechanical equipment, helping prevent failures and improve operational efficiency (Rippa et al., 2021). In medical diagnostics, thermal imaging allows early detection of diseases by identifying abnormal temperature patterns in the body (Casas-Alvarado et al., 2020), (Qu et al., 2022). In addition, there are approaches that provide a more comprehensive and accurate view of the environment, enhancing real-time decision making (Teutsch et al., 2022). In agriculture, for instance, it is important to mention the efforts to improve the efficiency of the Crop Water Stress Index (CWSI) to monitor plant water stress using thermal imaging (Pradawet et al., 2023). Key challenges in this domain include precise material properties estimation, incorporating atmospheric effects, and validating synthetic images against real-world data. Continuing, advances in sensor technology and computing have enabled the creation of high-resolution thermal images and the use of these images to perform the task of automatic person detection in thermal images using convolutional neural network models originally intended for detection in RGB images (Krišto et al., 2020).

The ability to detect and visualize temperature differences accurately and noninvasively has opened new possibilities and led to significant advances in each of the fields mentioned above. The main limitation to overcome is the lack of large-scale thermal image datasets and public benchmarks needed for deep learning-based solutions. This limitation can be addressed through the use of synthetic thermal images. As technology continues to advance, synthetic thermal image generation is expected to expand its applications and improve its precision and utility.

In recent years, research has focused on employing advanced computational techniques to generate thermal-like images from RGB inputs. Leveraging numerical heat transfer modeling and advanced technologies such as GANs, researchers are pioneering innovative methods to create synthetic thermal images for diverse applications. By learning from datasets of paired RGB and thermal images, these models can capture the underlying patterns and characteristics of

[a] https://orcid.org/0000-0002-3684-0656
[b] https://orcid.org/0000-0003-2468-0031

Suárez, P. L. and Sappa, A. D.
Synthetic Thermal Image Generation from Multi-Cue Input Data.

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 3: VISAPP*, pages 275-282
ISBN: 978-989-758-728-3; ISSN: 2184-4321

thermal imagery, enabling them to produce realistic thermal-like representations from RGB inputs (e.g., (Li et al., 2023), (Blythman et al., 2020)). These efforts have led to the creation of various methodologies, including deep learning models like Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). The field of conditional synthetic thermal image generation has become a valid alternative, offering new solutions to the challenges associated with obtaining real thermal images (Mizginov and Danilov, 2019).

The present work introduces a novel approach to generate thermal images, from the given visible spectrum images, by using a CycleGAN architecture. The main novelty lies in using a multi-cue source of information to feed the network—i.e., the depth and edge maps corresponding to the given visible spectrum image are used as input to the network. Input depth maps are obtained by an off-the-shelf architecture that generates depth maps for a given 2D image. Additionally, by incorporating the edges of the real image accentuates image features, improving contour definition on the synthesized thermal image. The use of depth maps is motivated by the correlation between depth information and the thermal energy propagation model in the given scene. The contribution of this work can be highlighted as follows:

- Present a novel CycleGAN-based generative architecture that uses multi-cue input—depth and edge maps—to generate thermal images from visible spectrum images improving the accuracy of synthesized thermal images.
- A multiple loss functions to ensure effective training by addressing different aspects of image quality. This combination of losses enhances spatial coherence, feature preservation, and structural similarity, leading to highly accurate and realistic synthesized thermal images.

The manuscript is organized as follows. Section 2 provides a review of related work in the fields of thermal image synthesis. Section 3 presents the approach proposed for generating synthetic thermal images, based on the usage of both depth and edge maps. Experimental results and comparisons with different approaches are given in Section 4. Finally, conclusions are presented in Section 5.

2 RELATED WORK

This section reviews state-of-the-art approaches proposed in the literature for generating synthetic thermal images using deep learning architectures. Through a detailed examination of various strategies and techniques employed in these approaches, we seek to understand the advances and challenges in this field, thereby establishing the foundation for the proposed methodology. In the field of thermal imaging, high-quality image synthesis using convolutional networks has become a promising approach. Researchers have explored multiple methodologies that leverage conditional GANs and prior information to generate thermal images that are both realistic and visually appealing. Therefore, conditioning generative networks with complementary data such as semantic information, distances, or physical characteristics of the scene have managed to obtain significant improvements in the definition and quality of synthetic thermal images.

One of the approaches is the work of (Zhang et al., 2018) where the authors propose to use an image-to-image translation model to generate synthetic thermal infrared data from the more readily available labeled RGB data. With this synthetic data, they create an extensive labeled dataset of synthetic thermal infrared sequences. These sequences can be used to train end-to-end optimal features specifically for thermal infrared tracking. A similar strategy has been proposed in Pons et al. (Pons et al., 2020); the authors develop a method to generate thermal images from the visible spectrum for facial emotion recognition using a CycleGAN; the authors propose to translate images from the visible spectrum to thermal images, thus optimizing the capabilities of emotional recognition systems. Güzel et al. (Güzel and Yavuz, 2022) investigate the use of CycleGAN training a paired image schema rather than unpaired, based on a better simulation of the normalization of the electromagnetic spectrum to improve the quality of the generated images. In Imtiaz et al. (Imtiaz et al., 2021), a Lightweight Pyramid Network (LPNet) is proposed for image synthesis. Their approach is based on Laplacian-Gaussian pyramidal decomposition and subsequent reconstruction to improve the thermal signature and maintain the contours of objects in the images. Continuing with convolutional architectures, Pavez et al. (Pavez et al., 2023) present a deep learning technique for the generation of high-quality synthetic thermal images. Their study introduces a database with paired visible and thermal facial images, proposing a cross-spectral facial recognition framework that facilitates quick and easy integration into existing facial recognition systems.

Continuing with the review, we can mention the work presented in (Suárez and Sappa, 2023), where a generic framework is proposed to generate thermal images of any environment, from the corresponding RGB images. In this case, an unpaired adversarial

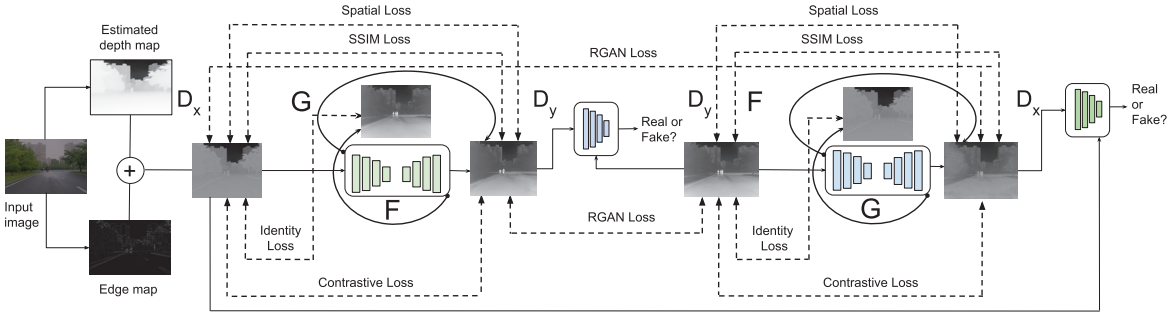

Figure 1: Proposed CycleGAN architecture.

cyclic generative model is proposed. It allows the simulation of the temperature of the objects present in the scene. Another translation model named ToDayGAN is presented in (Anoosheh et al., 2019), which introduces the task of localizing images from the same area captured during the day and night. The proposed model converts nighttime driving images into more useful daytime representations. The problem of domain translation has also been extensively studied in the remote sensing community, focusing on developing models capable of translating coaligned images between modalities such as RGB-IR, Synthetic Aperture Radar (SAR)-Electro-Optical (EO), SAR-IR, and SAR-RGB. This challenge has motivated the research community to organize competitions in various forums to evaluate the performance of different contributions. An example of these competitions is the one held annually at the PBVS-CVPR workshop (e.g., (Low et al., 2024), (Low et al., 2023)).

According to the state-of-the-art, all the reviewed techniques focus on generating representations in a domain other than the given one using generative approaches. However, a crucial aspect that has not been widely explored is the use of additional information to facilitate and improve results. All the reviewed approaches take as an input the given information, which could be the grayscale, RGB, SAR, etc. image. In the present work, we propose an approach that uses multi-cue information as input (i.e., depth and edge maps), instead of the given gray scale image, to improve the generation of thermal images. This change on the representation of the given image, offers a significant improvement in the quality and usefulness of the generated synthetic images, as will be presented in this paper.

3 PROPOSED APPROACH

This section presents the proposed CycleGAN architecture for generating synthetic thermal images based on (Zhu et al., 2017) and inspired by (Suárez and Sappa, 2023). This method takes advantage of the inherent characteristics of the depth maps combined with the corresponding edge map computed from the brightness channel of the HSV color space. This dual-input strategy addresses challenges commonly associated with cross-modal image translation, such as loss of fine details or spatial inconsistency, and contributes to generating thermal images that are both realistic and semantically meaningfulIt is worth mentioning that thermal images capture the thermal energy emitted by the object in the scene, rather than its colors in the visible spectrum. In the proposed model, the depth information is considered since, as will be presented next, it is more correlated with the temperature levels depicted in the thermal images, acting as an indicator of the thermal radiation of the object in the scene. This allows the model to prioritize intensity variations, just like thermal cameras, which focus more on the heat emitted by objects than on their color. As input to the network, the combination of the depth map with the edges of the corresponding 2D images is considered. This resulting fused image provides features that allow contour translation to improve the quality and sharpness of the generated synthetic images. To perform the experiments, we have used the depth map generated by the technique presented in Yin et al. (Yin et al., 2021), which estimates depth with unknown scale and offset and uses 3D point cloud encoders to estimate the missing depth offset and focal length. Figure 1 shows an illustration of the proposed architecture.

Depth maps provide essential information about the three-dimensional structure of the scene, including the surface orientation and the distance between objects. This spatial information is crucial to accurately infer thermal gradients, occlusions, and object boundaries within thermal images. Using depth maps, the model can better understand how thermal energy propagates through the scene, resulting in more accurate and realistic thermal representations. In addition, the inclusion of edge features of the 2D image accentuates the details and contours of objects, im-

proving the definition of shapes and boundaries in the synthesized thermal images. This approach ensures that even the most subtle features are preserved, resulting in a higher-quality thermal image that captures the complexities of the scene.

In the current work, a relativistic GAN loss, proposed by (Jolicoeur-Martineau, 2018), has been considered, instead of the traditional GAN loss suggested by (Goodfellow et al., 2020). 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 minibatch of randomly generated data, there are more realistic samples than false ones. This leads to better training of the GAN, being more stable training process, and improved image quality. and consequently, more accurate synthetic thermal images. Specifically, the relativistic loss is defined as follows:

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

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

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

In addition, in the current work, a contrastive loss function introduced in (Liu et al., 2021), has been implemented, which helps the model to learn the similarities between the latent spaces it generates. This approach is based on the principles outlined in (Andonian et al., 2021). This loss helps to group similar representations while ensuring that the different ones are distinctly separated. The proposed loss function can be written as:

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

This loss compares 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 (Andonian et al., 2021), 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 bounds of the objective domain 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 $F(y) \approx y$.

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

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,

Additionally, the spatial feature loss is defined. It is a custom function that computes the distances between the input and target tensors by evaluating their spatial features. This loss can be defined as:

$$\mathcal{L}_{\text{spatial}(x,y)} = \sum_{i=1}^{C}\left(\mathcal{L}^{i}_{\text{vertical}} + \mathcal{L}^{i}_{\text{horizontal}} + \mathcal{L}^{i}_{\text{average}}\right), \quad (5)$$

where each spatial feature loss term (vertical, horizontal, average) is computed using the Mean Squared Error (MSE) for each channel, which is defined as:

$$\text{MSE}(x,y) = \frac{1}{n}\sum_{j=1}^{n}(x_j - y_j)^2, \quad (6)$$

where x and y are the input and target tensors, respectively, and n is the number of elements in the tensors x and y. This represents the total number of spatial features (e.g., pixels) over which the error is averaged.

Another index used as a reference is the structural similarity index, proposed in (Wang et al., 2004). This index assesses images by considering the sensitivity of the human visual perception system to alterations in local structure. The underlying concept of this loss function is to help the learning model in generating visually enhanced images. The structural similarity loss is defined as:

$$\mathcal{L}_{\text{SSIM}}(x,y) = 1 - SSIM(x,y), \quad (7)$$

where $SSIM(x,y)$ is the Structural Similarity Index (see (Wang et al., 2004) for details), y represents the output of the neural network that we are trying to optimize, and x is the reference or ground truth image. It represents the target image that the model aims to reproduce or approximate as closely as possible.

Finally, all loss functions presented above are combined in the final loss function ($\mathcal{L}_{\text{final}}$) as follows:

$$\begin{aligned}\mathcal{L}_{\text{final}} = {} & \mathcal{L}_{\text{RGAN}}(G,D,x,y) + \lambda_X \mathcal{L}_{\text{contr}}(G,F,x) \quad (8)\\ & + \lambda_Y \mathcal{L}_{\text{contr}}(F,G,y) + \gamma \mathcal{L}_{\text{identity}}(G,F,x,y) \\ & + \beta \mathcal{L}_{\text{SSIM}}(x,y) + \alpha \mathcal{L}_{\text{spatial}}(x,y), \quad (9)\end{aligned}$$

where λ_X and λ_Y represent the weights attributed to the contrastive loss function for the domains x and

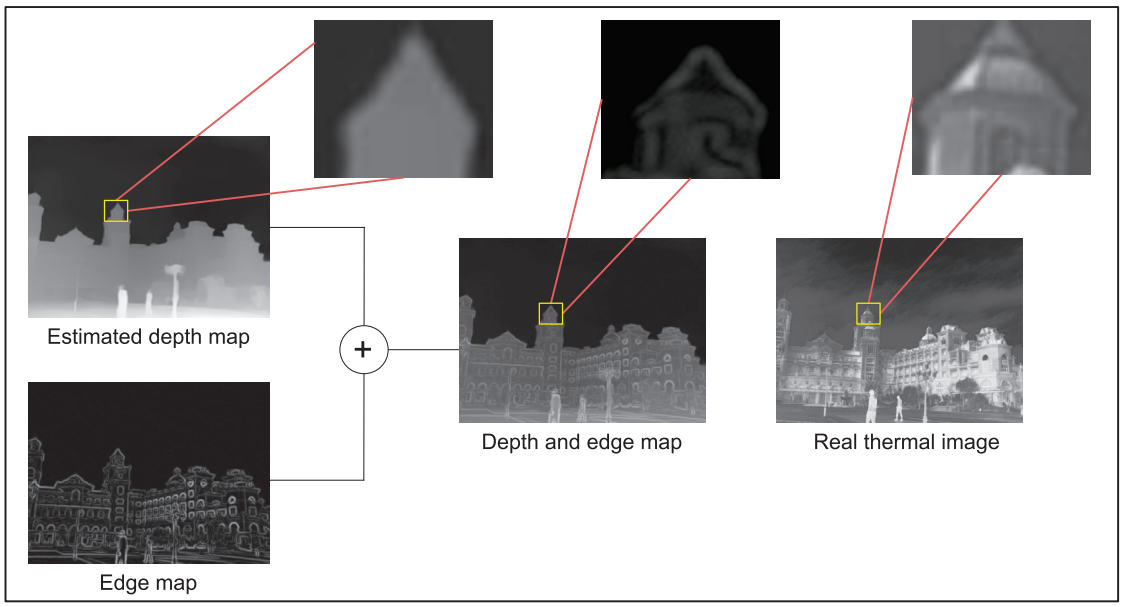

Figure 2: Illustration of the input images (i.e., a combination of estimated depth and edge maps); note how the enhanced estimated depth map, when integrated with the edge map, closely resembles the real thermal image.

y, respectively. These values are empirically determined based on experimental outcomes. γ and β are the weights of the *Identity* and *SSIM* loss functions respectively; α is the weights that control the contribution of *Spatial* feature loss. All of these values have been empirically defined according to the results of the experiments.

4 EXPERIMENTAL RESULTS

This section provides an overview of the quantitative and qualitative results obtained with the proposed architecture. It also, describe the dataset used for training and detailed information about the applied preprocessing techniques used on the images. Furthermore, it performs a comparative analysis using similarity metrics and evaluates the PSNR present on the obtained synthetic images.

4.1 Datasets

The architecture proposed in the current work has been trained with a subset of the M3FD data set (Liu et al., 2022). This data set was created using a binocular optical and infrared sensor and contains 4500 pairs of RGB and thermal images of outdoor scenarios. From this set of images, only 826 pairs were considered for the training process; 30 pairs of images were used for validation and 20 pairs were used to test the trained model. Depth and edge maps have been obtained from this subset of images from the M3FD data set. Depth maps have been obtained using the approach presented in (Yin et al., 2021), while edge maps have been obtained using Sobel edge detector (Gao et al., 2010). Depth and edge maps are combined in a single representation by averaging their values. Depth maps could be also estimated with other state-of-the-art approaches (e.g., (Suárez et al., 2023a), (Suárez et al., 2023b), (Yang et al., 2024), (Bhat et al., 2023)), however in the current work the approach presented in (Yin et al., 2021) has been selected due to good performance in the M3FD dataset.

4.2 Results and Comparisons

The proposed approach is evaluated and compared with the results of two state-of-the-art models for unpaired image translation (i.e., (Zhu et al., 2017) and (Suárez and Sappa, 2023)). As mentioned above, the proposed CycleGAN architecture uses as input the depth and edge map of the given 2D image (Fig. 2 shows an illustration of this multi-cue merging process). The central idea is to take advantage of the spatial characteristics of the images from the depth information to obtain a better representation of the synthetic images. This input data better correlates with the corresponding real thermal image. It can be seen in Table 1 that the merged maps obtain the highest correlation with the given real thermal images. To validate the advantage of using this multi-cue information as input, the proposed CycleGAN architecture

Table 1: Comparisons of different approaches (Corr. with GT: correlation index between the image used as input and the corresponding thermal GT; in (Zhu et al., 2017) and (Suárez and Sappa, 2023) the brightness channels of HSV color space are used as input).

Method	PSNR	SSIM	Corr. with GT
Zhou et al. (Zhu et al., 2017)	*15.299*	*0.663*	*-0.2607*
Suárez et al. (Suárez and Sappa, 2023)	*18.112*	*0.706*	*-0.2607*
Prop. App. with Gray Scale	*17.131*	*0.673*	*-0.2898*
Prop. App. with Pseudothermal	*17.084*	*0.616*	*0.2487*
Prop. App. with Depth+edges	*18.317*	*0.713*	*0.3011*

has been also trained considering a grayscale image as input as well as a pseudothermal image as input (Tuzcuoğlu et al., 2024). Results from these three alternatives are presented below.

Fig. 3 shows some illustrations of the obtained synthetic thermal images and comparisons with state-of-the-art approaches. Table 1 presents the quantitative results of the averages obtained by the models: (Zhu et al., 2017), (Suárez and Sappa, 2023) and the proposed model trained with different inputs: *i*) grayscale image; *ii*) pseudothermal image; and *iii*) fusion of depth and edge maps. SSIM and PSNR metrics demonstrate that the architecture trained by using as an input the depth and edge maps reaches the best results.

5 CONCLUSION

This paper presents a novel approach for generating synthetic thermal images using depth maps and corresponding edge maps as inputs. This method produces a well-generalized model capable of generating clear thermal-like images that closely resemble real thermal images. This combination enhances the realism and accuracy of the synthesized thermal images, as evidenced by improved performance metrics compared to those of the existing methods. Future work will explore the usage of other depth estimation techniques. Furthermore, exploring advanced generative model architectures and improved training techniques could enhance the quality and resolution of the synthetic images. Finally, developing new evaluation metrics and loss functions could further refine its quality and sharpness.

ACKNOWLEDGEMENTS

This work was supported in part by Grant PID2021-128945NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe", in part by the Air Force Office of Scientific Research Under Award FA9550-24-1-0206 and in part by the ESPOL project CIDIS-003-2024. The authors acknowledge 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 2021SGR01499.

REFERENCES

Andonian, A., Park, T., Russell, B., Isola, P., Zhu, J.-Y., and Zhang, R. (2021). Contrastive feature loss for image prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1934–1943.

Anoosheh, A., Sattler, T., Timofte, R., Pollefeys, M., and Van Gool, L. (2019). Night-to-day image translation for retrieval-based localization. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5958–5964. IEEE.

Bhat, S. F., Birkl, R., Wofk, D., Wonka, P., and Müller, M. (2023). Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*.

Blythman, R., Elrasad, A., O'Connell, E., Kielty, P., O'Byrne, M., Moustafa, M., Ryan, C., and Lemley, J. (2020). Synthetic thermal image generation for human-machine interaction in vehicles. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE.

Casas-Alvarado, A., Mota-Rojas, D., Hernández-Ávalos, I., Mora-Medina, P., Olmos-Hernández, A., Verduzco-Mendoza, A., Reyes-Sotelo, B., and Martínez-Burnes, J. (2020). Advances in infrared thermography: Surgical aspects, vascular changes, and pain monitoring in veterinary medicine. *Journal of Thermal Biology*, 92:102664.

Gao, W., Zhang, X., Yang, L., and Liu, H. (2010). An improved sobel edge detection. In *2010 3rd International conference on computer science and information technology*, volume 5, pages 67–71. IEEE.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Güzel, S. and Yavuz, S. (2022). Infrared image generation from rgb images using cyclegan. In *2022 Interna-*

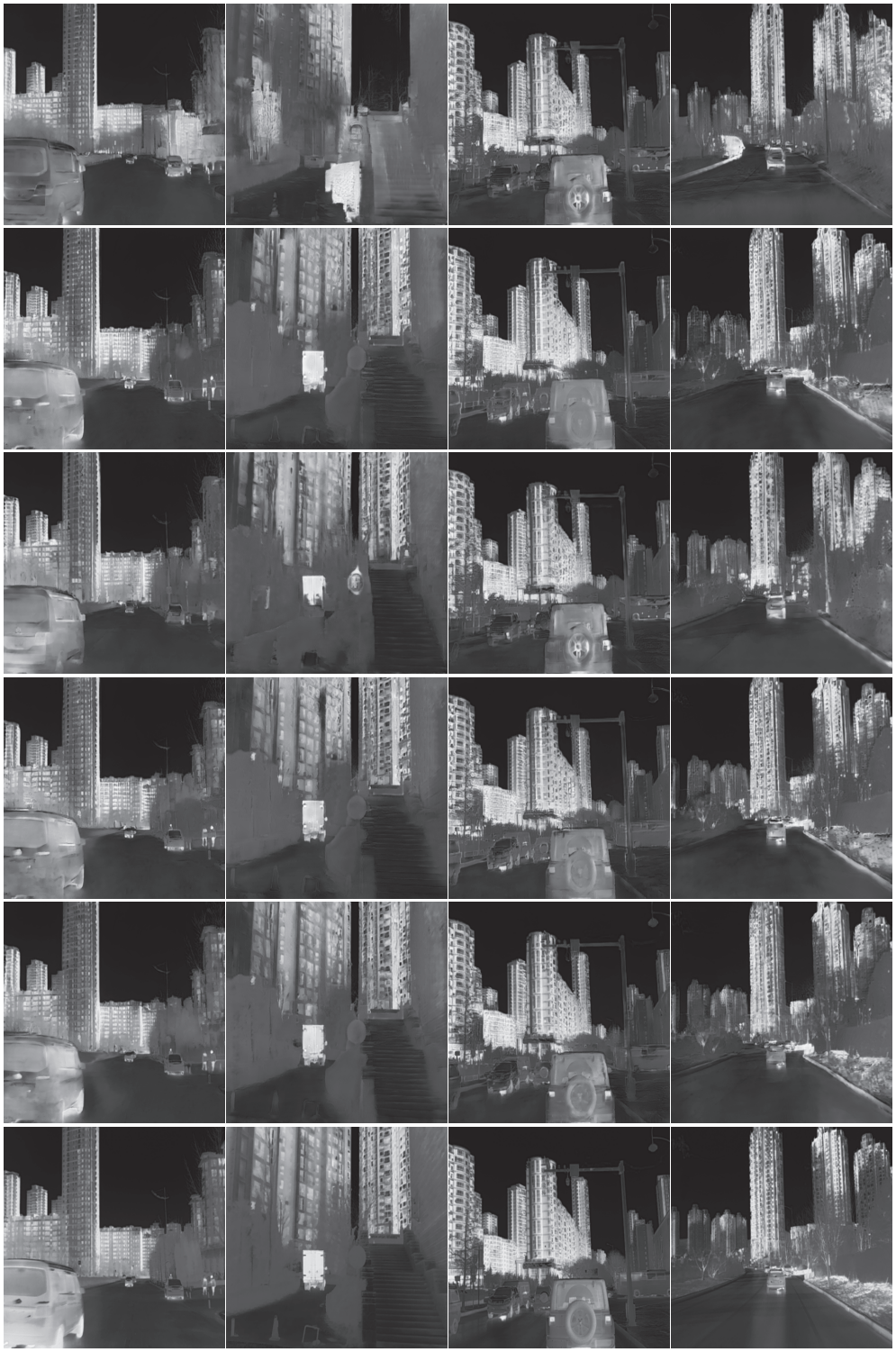

Figure 3: Experimental results from M3FD dataset: (*1st. row*) synthetic thermal images using (Zhu et al., 2017); (*2nd. row*) synthetic thermal images using (Suárez and Sappa, 2023); (*3rd. row*) synthetic thermal images using gray scale as input; (*4th. row*) synthetic thermal images using pseudothermal images as input; (*5th. row*) synthetic thermal images using depth + edge maps as input; (*6th. row*) ground truth images.

tional Conference on INnovations in Intelligent SysTems and Applications (INISTA), pages 1–6. IEEE.

Imtiaz, S., Taj, I. A., and Nawaz, R. (2021). Visible to thermal image synthesis using light weight pyramid network. In *2021 16th International Conference on Emerging Technologies (ICET)*, pages 1–5. IEEE.

Jolicoeur-Martineau, A. (2018). The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*.

Krišto, M., Ivasic-Kos, M., and Pobar, M. (2020). Thermal object detection in difficult weather conditions using yolo. *IEEE access*, 8:125459–125476.

Li, X., Li, J., Li, Y., Ozcan, A., and Jarrahi, M. (2023). High-throughput terahertz imaging: progress and challenges. *Light: Science & Applications*, 12(1):233.

Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., and Luo, Z. (2022). 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.

Liu, R., Ge, Y., Choi, C. L., Wang, X., and Li, H. (2021). Divco: Diverse conditional image synthesis via contrastive generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16377–16386.

Low, S., Nina, O., Bowald, D., Sappa, A. D., Inkawhich, N., and Bruns, P. (2024). Multi-modal aerial view image challenge: Sensor domain translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3096–3104.

Low, S., Nina, O., Sappa, A. D., Blasch, E., and Inkawhich, N. (2023). Multi-modal aerial view image challenge: Translation from synthetic aperture radar to electro-optical domain results - PBVS 2023. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 515–523.

Mizginov, V. and Danilov, S. Y. (2019). Synthetic thermal background and object texture generation using geometric information and GAN. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:149–154.

Pavez, V., Hermosilla, G., Silva, M., and Farias, G. (2023). Advanced deep learning techniques for high-quality synthetic thermal image generation. *Mathematics*, 11(21):4446.

Pons, G., El Ali, A., and Cesar, P. (2020). ET-CycleGAN: Generating thermal images from images in the visible spectrum for facial emotion recognition. In *Companion publication of the international conference on multimodal interaction*, pages 87–91.

Pradawet, C., Khongdee, N., Pansak, W., Spreer, W., Hilger, T., and Cadisch, G. (2023). Thermal imaging for assessment of maize water stress and yield prediction under drought conditions. *Journal of Agronomy and Crop Science*, 209(1):56–70.

Qu, Y., Meng, Y., Fan, H., and Xu, R. X. (2022). Low-cost thermal imaging with machine learning for non-invasive diagnosis and therapeutic monitoring of pneumonia. *Infrared Physics & Technology*, 123:104201.

Rippa, M., Pagliarulo, V., Lanzillo, A., Grilli, M., Fatigati, G., Rossi, P., Cennamo, P., Trojsi, G., Ferraro, P., and Mormile, P. (2021). Active thermography for non-invasive inspection of an artwork on poplar panel: novel approach using principal component thermography and absolute thermal contrast. *Journal of Nondestructive Evaluation*, 40(1):21.

Suárez, P. L., Carpio, D., and Sappa, A. (2023a). A deep learning based approach for synthesizing realistic depth maps. In *International Conference on Image Analysis and Processing*, pages 369–380. Springer.

Suárez, P. L., Carpio, D., and Sappa, A. (2023b). Depth map estimation from a single 2d image. In *2023 17th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 347–353. IEEE.

Suárez, P. L. and Sappa, A. D. (2023). Toward a thermal image-like representation. In *VISIGRAPP (4: VISAPP)*, pages 133–140.

Teutsch, M., Sappa, A. D., and Hammoud, R. I. (2022). *Computer vision in the infrared spectrum: challenges and approaches*. Springer.

Tuzcuoğlu, Ö., Köksal, A., Sofu, B., Kalkan, S., and Alatan, A. A. (2024). Xoftr: Cross-modal feature matching transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4275–4286.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., and Zhao, H. (2024). Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381.

Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., and Shen, C. (2021). Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 204–213.

Zhang, L., Gonzalez-Garcia, A., Van De Weijer, J., Danelljan, M., and Khan, F. S. (2018). Synthetic data generation for end-to-end thermal infrared tracking. *Transactions on Image Processing*, 28(4):1837–1850.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.