

Exploration and Validation of Specialized Loss Functions for Generative Visual-Thermal Image Domain Transfer

Simon Fischer, Benedikt Kottler^a, Eva Strauß and Dimitri Bulatov^b

Fraunhofer IOSB Ettlingen, Gutleuthausstrasse 1, 76275 Ettlingen, Germany

Keywords: Thermal Infrared, Domain Transfer, Style Transfer, GAN, Loss Function, Image Generation.


Abstract: This paper presents an enhanced approach to visual-to-thermal image translation using an improved InfraGAN model, incorporating additional loss functions to increase realism and fidelity in generated thermal images. Building on the existing InfraGAN architecture, we introduce perceptual, style, and discrete Fourier transform (DFT) losses, aiming to capture intricate image details and enhance texture and frequency consistency. Our model is trained and evaluated on the FLIR Adas dataset, providing paired visual and thermal images across diverse contexts, from urban traffic scenes. To optimize the interplay of loss functions, we employ hyperparameter tuning with the Optuna library, achieving an optimal balance among the components of the loss function. First, experimental results show that these modifications lead to significant improvements in the quality of generated thermal images, underscoring the potential of advanced loss functions for domain transfer tasks. This work contributes a refined framework for generating high-quality thermal imagery, with implications for fields such as surveillance, autonomous driving, and facial recognition in challenging environmental conditions.


1 INTRODUCTION

Image transfer from the visual to the thermal or infrared domains has multiple military and civil applications. For the former, target detection, precision guidance, and training autonomous vehicles in challenging illumination and weather conditions are among the first use cases that come to mind (Xiong et al., 2016; Suárez and Sappa, 2024). One may imagine an additional screen showing the driver the infrared view of the night scene with possible obstacles. Gaming applications are related to this field; in order to achieve immersive simulations, realistic night views are desired. As for the latter, RGB-to-thermal and RGB-to-infrared transfer also support artistic applications, enabling creative photography and design by showcasing scenes in a different spectrum. In forensic science, these techniques assist in reconstructing crime scenes by uncovering hidden details like heat signatures that may help in investigations. Last but not least, from the point of view of environmental monitoring, surface temperature retrieval from remote sensing data is an elegant way to infer potentially risky areas of the scene. One may think about

Urban Heat Islands, where trapping and multiple radiations contribute to the increase of temperatures in metropolitan centers in comparison to their surroundings (Bulatov et al., 2020). These physical processes are difficult to measure and to simulate due to the precise knowledge of material properties estimation and the need to incorporate atmospheric effects and to validate synthetic images against real-world data (Suárez and Sappa, 2024). Furthermore, for direct measurements, multiple temperature boxes (Kottler et al., 2023) or thermal scanning robots (López-Rey et al., 2023) must be employed, which, on the one hand, produce large amounts of data, and, on the other hand, may be stolen and not provide any data at all. Satellite data allow for relatively broad coverage of areas and time; however, they have, in turn, a too coarse resolution so that 3D effects remain unconsidered.

This article is supposed to make use of two important latest trends: the omnipresence of optical data and tremendous progress in generative style transfer. Billions of images are taken worldwide by smartphone cameras every day, land to a large share on social networks, and not seldomly are used for training by large corporations. For thermal and infrared images, such wide training data are unavailable or have

^a  <https://orcid.org/0000-0002-0498-0646>

^b  <https://orcid.org/0000-0002-0560-2591>

been created only recently. Thus, we make use of the recently published Generative Adversarial Network called *InfraGAN* (Özkanoglu and Ozer, 2022). The main contribution of this article is the idea to add new loss functions to improve the performance on the image transfer task from the visual into the thermal domain.

The paper is structured as follows: Section 2 will summarize the main findings in the aforementioned research field. All the reader is supposed to know about *InfraGAN* and its loss functions is reported in Section 3. The methodology is presented in Section 4. The results and conclusions are reported in Sections 5 and 6, respectively.

2 RELATED WORK

Ever since the high success of works like (Isola et al., 2016) and (Zhu et al., 2017), generative adversarial networks (GANs), introduced by (Goodfellow et al., 2014) form the standard approach for image-to-image translation. In particular, this is true for subsets of the problem like visual-thermal image domain transfer (see for example (Ma et al., 2024), (Ordun et al., 2023) and (Özkanoglu and Ozer, 2022)). Nevertheless, there are exceptions like (Sun et al., 2023) which rely on the use of transformers in the generation process. (Ordun et al., 2023) further introduce a diffusion model and compares its result to those of the GAN.

The network introduced by (Özkanoglu and Ozer, 2022) stands out by its encoder-decoder structure that is used not only for the generator but also the discriminator, applying a discriminator loss function to the whole image. Further, the authors expand the generator loss by an additional term based on the Structural Similarity Index Measure (SSIM (Wang et al., 2004)) which improves the overall results.

The authors of (Ordun et al., 2023) compare their introduced GAN to a conditional Denoising Diffusion Model. They are able to show that in the case of facial images, the visual-to-thermal transfer of their GAN outperforms the diffusion-based state-of-the-art approach. This confirms GANs forming a state-of-the-art model in image domain transfer.

Further, in their GAN, the authors introduced the use of a new loss function called Fourier Transform Loss. This approach was earlier used in the task of image super-resolution (Fuoli et al., 2021). Their idea is to transfer both the generated and the real thermal image into the frequency domain and to compare their amplitude and phase. “The motivation is to not only map the visible-to-thermal pixel space, but also achieve similarity between high and low frequencies such as

hair, teeth, and glasses.” We use this idea and adapt it for our purposes.

Recent studies have expanded these methodologies. For example, (Suárez and Sappa, 2024) introduce a depth-conditioned approach to generating thermal-like images, further advancing the contextual adaptation of thermal image synthesis techniques. Additionally, (Liu et al., 2021) explore diverse conditional image synthesis through a contrastive GAN approach, showcasing a method to encourage variation in generated outputs. Another recent study by (Yu et al., 2023) addresses the complexities in unpaired infrared-to-visible video translation, focusing on fine-grained, content-rich patch transfers.

While approaches like diffusion models and transformers offer alternatives, GANs remain widely used in visual-to-thermal image translation. In our work we try to further enhance them with the focus on loss functions.

3 PRELIMINARIES

This section provides the *InfraGAN* architecture and its core loss functions, preparing the groundwork for further modifications and optimizations detailed in the methodology section.

3.1 *InfraGAN* Model Architecture

The **generator** in *InfraGAN* is based on a U-Net, which consists of an encoder-decoder structure. The encoder progressively down-samples the input image through a series of convolutional layers, each followed by batch normalization and LeakyReLU activation functions. The decoder mirrors the encoder’s structure, progressively up-sampling the compressed features to the original resolution using transposed convolutions. Skip connections are introduced between corresponding encoder and decoder layers, allowing information from the encoder to flow directly to the decoder, preserving fine-grained image features. The final layer produces the generated infrared image.

The **discriminator** of *InfraGAN* uses a U-Net-based architecture designed for classification at both the image (global) and pixels (local). Similar to the generator, the discriminator’s encoder (D_{enc}) down-samples the input image to extract essential features. However, here the encoder is trained to detect patterns and textures specific to real infrared images, assisting in distinguishing real from generated images. The discriminator’s decoder (D_{dec}) up-samples features extracted by the encoder to classify individual

pixels. This pixel-level classification provides fine-grained authenticity checks across the image, helping the discriminator to enforce more detailed supervision on the generator. This dual-output structure enhances the discriminator’s ability to guide the generator to produce highly realistic infrared images.

The generator and discriminator are trained together in an adversarial setup. The generator aims to create increasingly realistic infrared images to “fool” the discriminator, while the discriminator continually improves at distinguishing real from generated images. Over time, this adversarial process pushes the generator to produce lifelike infrared images with detailed, realistic features.

3.2 Losses Used in InfraGAN

In InfraGAN, the generator and discriminator are optimized by minimizing a respective loss function composed of multiple terms. Loss functions play a vital role in training neural networks offering a measure of how “similar” the generated output is to the ground truth. While the discriminator loss remains unchanged, we later expand the generator loss by additional terms in Section 4.1. To ensure that the reader can understand the components without having to refer to the original paper (Özkanoglu and Ozer, 2022), we briefly introduce each term here. Let X be the input image in visible domain, and Y be the ground truth thermal image. Then the generated thermal image is denoted by $\hat{Y} = G(X)$, $D(X, Y)$ and $D(X, \hat{Y})$ refers to the binary outputs of the discriminator and $E(\cdot)$ to the expected value.

InfraGAN’s generator loss is composed of the various losses weighted with hyperparameters $\lambda_1, \lambda_2 \in \mathbb{R}$ and is defined as:

$$l_G = l_{\text{cGAN}} + \lambda_1 l_{L1} + \lambda_2 l_{\text{SSIM}}, \quad (1)$$

where the Conditional GAN Loss (l_{cGAN}) encourages the generated images to appear realistic according to the discriminator. It is given by:

$$l_{\text{cGAN}} = \mathbf{E}_X \left[\sum_{i,j} \log \left([D_{\text{dec}}(X, \hat{Y})]_{i,j} \right) \right] + \mathbf{E}_X \left[\log (D_{\text{enc}}(X, \hat{Y})) \right]. \quad (2)$$

The L1 Loss (l_{L1}) measures the pixel-wise differences between the generated and ground truth images:

$$l_{L1} = \frac{1}{N} \sum_{i,j} |\hat{Y}_{i,j} - Y_{i,j}|, \text{ and} \quad (3)$$

the SSIM Loss (l_{SSIM}) is based on the Structural Similarity Index (SSIM) and encourages structural simi-

larity between generated and ground truth images:

$$l_{\text{SSIM}} = \frac{1}{m} \sum_{i=0}^{m-1} (1 - \text{SSIM}(\hat{Y}_i, Y_i)). \quad (4)$$

where the SSIM between two images \hat{Y} and Y is calculated as:

$$\text{SSIM}(\hat{Y}, Y) = \frac{2\mu_{\hat{Y}}\mu_Y + C_1}{\mu_{\hat{Y}}^2 + \mu_Y^2 + C_1} \cdot \frac{\sigma_{\hat{Y},Y} + C_3}{\sigma_{\hat{Y}}^2 + \sigma_Y^2 + C_2}, \quad (5)$$

where the constants C_1 and C_2 are calculated based on the range of pixel values L , where $C_1 = 0,0001 \cdot L^2$, $C_2 = 0,0009 \cdot L^2$.

The discriminator loss combines global and pixelwise discrimination capabilities, defined as:

$$l_D = l_{D_{\text{enc}}} + l_{D_{\text{dec}}}, \quad (6)$$

where the global and pixelwise losses are defined as follows:

$$l_{D_{\text{enc}}} = -\mathbf{E}_{X,Y} [\log D_{\text{enc}}(X, Y)] - \mathbf{E}_X [\log (1 - D_{\text{enc}}(X, \hat{Y}))]. \quad (7)$$

$$l_{D_{\text{dec}}} = -\mathbf{E}_{X,Y} \left[\sum_{i,j} \log (D_{\text{dec}}(X, Y)_{i,j}) \right] - \mathbf{E}_X \left[\sum_{i,j} \log (1 - [D_{\text{dec}}(X, \hat{Y})]_{i,j}) \right]. \quad (8)$$

4 METHODOLOGY

In this section, we outline our approach for refining InfraGAN’s performance. We introduce additional loss functions, perceptual, style, and DFT loss, to capture nuanced image features that enhance realism in thermal image generation. Finally, we then conduct a hyperparameter search, using the Optuna library for Bayesian optimization, to fine-tune the balance among these losses for optimal model outcomes.

4.1 Additional Losses

The additional loss functions, perceptual loss, style loss, and DFT loss, are chosen to address different aspects of image realism and quality. Each of these losses provides unique benefits that collectively guide the network towards generating images that align more closely with human perception and retain realistic textural and frequency characteristics.

Perceptual Loss for Human-Centric Evaluation

First, we introduce a perceptual loss l_{perc} . This method was introduced by (Johnson et al., 2016). As the name suggests, the loss is supposed to represent the human perception. Therefore, the ground truth and the generated image are evaluated on a layer of a classification network. More precisely, we use the VGG19 network by (Simonyan and Zisserman, 2015).

We set ϕ as the VGG19 network trained on ImageNet (Russakovsky et al., 2015). Further, let $\phi_j(y)$ be the activation of the j -th layer of ϕ . That layer is a convolution layer, and it has the shape $C_j \times H_j \times W_j$. Then the perceptual loss is defined by

$$l_{\text{perc}} = \sum_j \frac{1}{C_j H_j W_j} \|\phi_j(\hat{Y}) - \phi_j(Y)\|_2^2. \quad (9)$$

Style Loss for Textural Consistency

The second loss we want to expand our network with comes from the same paper as the perceptual loss. For the style loss l_{style} we again make use of VGG19 and its layers ϕ_j . As before, we define the activation to have shape $C_j \times H_j \times W_j$. We calculate the Gram matrix $G_j^\phi(y)$ for image y . Its elements are defined by the following formula:

$$G_j^\phi(Y) = \frac{1}{C_j H_j W_j} \cdot \psi \psi^\top, \quad (10)$$

where ψ is $\phi_j(Y)$ reshaped as a matrix. The style loss then is defined by

$$l_{\text{style}} = \sum_{j=1}^J \left\| G_j^\phi(\hat{Y}) - G_j^\phi(Y) \right\|_F^2, \quad (11)$$

where $\|\cdot\|_F$ is the Frobenius norm.

Following the approach in (Kottler et al., 2022), we set $J = 5$ in both the perceptual loss and the style loss.

DFT Loss for Frequency-Based Comparison

Lastly, we introduce the discrete Fourier transform (DFT) loss l_{DFT} . The idea to use the DFT as a loss function was first introduced by (Fuoli et al., 2021) for image super-resolution, and applied to the task of domain transfer by (Ordun et al., 2023). The idea is to transfer the ground truth and the generated image into the frequency domain via DFT and then calculate a difference in this domain. Given the real \mathcal{R} and imaginary I part of the Fourier version of an image, we calculate:

$$l_{\text{DFT}} = \|\mathcal{R}(\hat{Y}) - \mathcal{R}(Y)\|_2^2 + \|I(\hat{Y}) - I(Y)\|_2^2. \quad (12)$$

Unlike (Fuoli et al., 2021) and (Ordun et al., 2023), we do not compare amplitude and phase in the frequency domain but the real and imaginary parts of the image's frequency counterpart. We decided to adjust the approach, because of two observations that lead to some doubts concerning the use of amplitude and phase. Our first observation was that there could be problems when comparing the *periodic* phase values: Imagine two images with values close to 0 and 2π . Ideally, their distance should create a small loss while, in reality, the L1 distance is nearly at maximum. Secondly, we realized that the ranges of amplitude and phase are very different. While the phase is limited to $[0, 2\pi]$, the amplitude can have up to five-digit values. Therefore, simply adding the phase and amplitude differences could create a huge imbalance. Based on these concerns, we decided to calculate the real and imaginary parts of the ground truth and the generated thermal image and then compare these values. This gives us information about the distribution of the image's frequencies without any range-related problems.

4.2 Evaluation Metrics

The evaluation of our model mainly follows the example of (Özkanoğlu and Ozer, 2022). Along with SSIM and L1 metrics that were similarly used as loss functions, it is crucial to use new metrics that were not involved in the training process. Therefore, we add the MSSIM (Mean SSIM), LPIPS (Learned Perceptual Image Patch Similarity) and PSNR (Peak Signal-to-Noise Ratio) metrics. In the following, we explain their structure and why their usage is beneficial. The Mean SSIM builds on SSIM and adds a global perspective by forming the mean over several down-scaled versions \hat{y}_k, y_k of the generated image \hat{Y} and ground truth Y . This also strengthens the noise immunity.

$$\text{MSSIM}(Y, \hat{Y}) = \frac{1}{K} \sum_{k=1}^K \text{SSIM}(Y_k, \hat{Y}_k). \quad (13)$$

The downscaling takes place according to (Wang et al., 2004).

The LPIPS metric resembles in its idea the perceptual loss above, as it measures the Euclidean distance between the feature vectors of Y and \hat{Y} . However, for LPIPS, the smaller AlexNet is used instead of VGG19 to obtain the features. Due to the fewer number of weights, LPIPS focus more on low- and mid-level features compared to VGG19.

$$\text{LPIPS}(Y, \hat{Y}) = \sum_{l=1}^L \frac{\omega_l}{H_l W_l} \sum_{h,w} \left[f_l^{h,w}(Y) - f_l^{h,w}(\hat{Y}) \right]^2, \quad (14)$$

where we sum over the last $L = 5$ layers of AlexNet, which are denoted by f . Hereby H_l, W_l is the height and width of the l -th layer, h_l, w_l show the pixel coordinates, $f_l(Y)$ represents the normalized feature from l -th layer, and vector ω_l refers to the trained weights of LPIPS.

Lastly, the PSNR aims to represent the quality of reconstruction of the thermal image from a visual image:

$$\text{PSNR}(Y, \hat{Y}) = -10 \log \text{MSE}(Y, \hat{Y}), \quad (15)$$

where we use the mean squared error

$$\text{MSE}(Y, \hat{Y}) = \frac{1}{HW} \sum_{i=0}^H \sum_{j=0}^W (Y(i, j) - \hat{Y}(i, j))^2 \quad (16)$$

Again, H, W denote height and width of the image.

4.3 Hyperparameter Optimization

The different loss functions that combine to form the generator loss are all weighted by prefactors similar to equation (1). In our case, the additional losses are weighted by the hyperparameters λ_3, λ_4 , and λ_5 as shown below:

$$l_G = l_{\text{cGAN}} + \lambda_1 \cdot l_{\text{L1}} + \lambda_2 \cdot l_{\text{SSIM}} + \lambda_3 \cdot l_{\text{perc}} + \lambda_4 \cdot l_{\text{style}} + \lambda_5 \cdot l_{\text{DFT}}$$

To enhance the interaction of the losses we aim to optimize the hyperparameters λ_1 to λ_5 , which we will summarize in vector Λ . The open-source library `Optuna` provides a framework based on the Bayesian optimization to iteratively find optimal hyperparameters. It is specialized in the optimization of neural network applications. `Optuna` enhances the process's efficiency by pruning trials that are unlikely to yield promising results, thereby saving computational resources. `Optuna` allows for customization of pruner settings. We want it to utilize median performance metrics in order to make better pruning decisions. Additionally, we prohibit pruning before the tenth trial to build up a comprehensive decision pool. We set a minimum threshold of six epochs before pruning can commence, as this has been shown to achieve a good balance between accuracy and efficiency.

Since the framework must test multiple value configurations for Λ , we want the optimization to have a high number of trials. Specifically, we opted for 1000 trials based on extensive research. To maintain consistency with the initial values used in `InfraGAN`'s code, we define the suggested hyperparameters to be integers. We enhance the likelihood of discovering effective hyperparameter combinations by allowing them to range between 1 and 1000. We set them to follow a logarithmic distribution, meaning the values will have logarithmic spacing, thereby

preferring smaller values. This approach allows for nuanced hyperparameters adjustments across a wide integer range, enhancing stability and control during optimization.

To increase the efficiency of the optimization, we use a reduced dataset of 60 varying image pairs from the FLIR dataset. Furthermore, we reduce the number of epochs in the network's training from 200 to 100, facilitating a more time-efficient optimization. Since our goal is not to get a perfectly trained network but to identify an optimal Λ , we prioritize fast optimization iterations. The reduced network parameters are not expected to impair our results.

Our experiments showed that the quality of the network oscillated strongly between subsequent training epochs. Therefore, any metric \mathcal{L} on our network must be smoothed. We first average the value of \mathcal{L} over data batches within the current epoch and, ultimately, return the median over the last five epochs.

Arguably, the most important decision in optimizing with `Optuna` is how to define its objective function \mathcal{L} . This function outputs a value representing the network's quality considering the new hyperparameters. Therefore, we need a good measure of how similar the generated thermal image and the ground truth are. We propose employing the LPIPS metric, as seen in equation (14). This metric is not used in the training and therefore does not interfere with the hyperparameters during optimization. Thus, we assign $\mathcal{L}(\Lambda)$ to be the walking average of the LPIPS metric for configuration $\Lambda = (\lambda_1, \dots, \lambda_5)$.

5 RESULTS

In this section, we present the outcomes of our approach, including the dataset used, hyperparameter optimization, and model evaluation. We first describe the dataset that provided paired visual and thermal images, crucial for training and testing `InfraGAN`. We then detail our hyperparameter optimization process to find the optimal balance for the newly integrated loss functions. Finally, we evaluate the model's performance, analyzing the effectiveness of our modifications in producing realistic thermal images.

5.1 Dataset

For our training, we used the `Flir1 Adas` dataset consisting of image pairs of the same motive, one in visual (RGB) and one in thermal (IR) domain. The

¹FLIR dataset, <https://www.flir.com/oem/adas/adas-dataset-form/>.

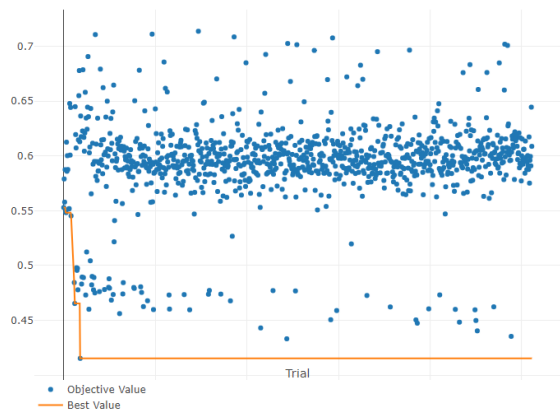


Figure 1: Optimization process of the hyperparameters on the FLIR dataset. Each blue dot represents the activation value $\mathcal{L}(\Lambda_n)$ of the n -th configuration Λ_n . The orange line depicts the best value at time \tilde{n} : $\min_{1 \leq n \leq \tilde{n}} \mathcal{L}(\Lambda_n)$.

FLIR dataset is a publicly available collection of high-resolution images. It includes various traffic scenarios with different light and weather conditions. Multiple scenarios include pedestrians and other road users.

5.2 Hyperparameter Optimization

This section details the optimization process and adjustments made to find the best weight values for each of the generator’s loss functions.

The optimization made on the FLIR dataset included 1000 trials, testing different configurations of vector Λ , and evaluating them using the objective function LPIPS, denoted by \mathcal{L} . A total of 76 out of 1000 trials, approximately 8 percent, were completed, while the rest was pruned.

Figure 1 shows the results of each trial. The orange line represents the best value at the corresponding time \tilde{n} : $\min_{1 \leq n \leq \tilde{n}} \mathcal{L}(\Lambda_n)$. The overall best result was achieved in trial $n = 36$:

$$\min_n \mathcal{L}(\Lambda_n) = \mathcal{L}(\Lambda_{36}) \approx 0.41534, \quad (17)$$

where $\Lambda_{36} = [250, 9, 69, 1, 273]$. The $\mathcal{L}(\Lambda_n)$ values in Figure 1 are divided into two main areas. Except for the first 10 trials, which were not allowed to prune, the upper part of the point cloud, approximately between 0.55 and 0.65, consists entirely of pruned trials. Most of the completed trials yield objective values of 0.5 and lower. A noticeable gap exists between $\mathcal{L}(\Lambda_{36})$ and the next best result, $\mathcal{L}(\Lambda_{487})$. Their ratio of these values is approximately 0.95.

5.3 Evaluation

Here, we assess InfraGAN’s performance with the enhanced loss functions and optimized parameters,

measuring improvements in thermal image generation quality compared to InfraGAN.

Table 1: Quantitative Results: Original InfraGAN vs Our enhanced approach.

FLIR dataset	SSIM	MSSIM	LPIPS	L1	PSNR
InfraGAN	0.2401	0.3429	0.2275	0.3039	16.3238
Our approach	0.2683	0.3534	0.2558	0.2979	16.4590

The quantitative evaluation of our enhanced approach compared to the original InfraGAN model is presented in Table 1. The table compares the metrics SSIM, MSSIM, LPIPS, L1, and PSNR. Our approach demonstrates improvements across several metrics, indicating enhanced image quality. For SSIM, our method surpasses the original model, reflecting better preservation of spatial details. Similarly, MSSIM shows an improvement of approximately 0.01, suggesting enhanced structural consistency. While the LPIPS metric shows a slight increase and therefore a minor decrease in perceptual quality, our approach shows a modest improvement in L1 loss, indicating more precise image reconstruction in terms of pixel-level accuracy. Additionally, the PSNR metric improves from 16.3238 to 16.4590, reflecting better overall image fidelity. Despite the slight increase in LPIPS, overall, our enhanced model demonstrates significant improvements in most metrics.

Figure 2 provides a comparison of qualitative results between the original InfraGAN algorithm and our enhanced approach. The rows of the Figure showcase various scenes from the FLIR dataset, differing in scenario and exposure. The Figure shows that InfraGAN often suffers from artifacts in its generated images. These artifacts can obscure important details and diminish the images’ utility. In contrast, our model successfully mitigates these artifacts, resulting in cleaner and more coherent images. However, it is noteworthy that the images exhibit a ”smooth” appearance, similar to the effect of a blurring filter. This characteristic may limit the texture and detail of the images. Nevertheless, the qualitative and quantitative results highlight the potential of the new loss functions in improving the perceptual quality of generated images.

6 CONCLUSION

In this work, we explored enhancements of InfraGAN for visual-to-thermal image translation by introducing additional loss functions—perceptual, style, and DFT losses—that capture finer image details and improve realism. We trained and tested our model with the FLIR dataset consisting of traffic scenes. Through

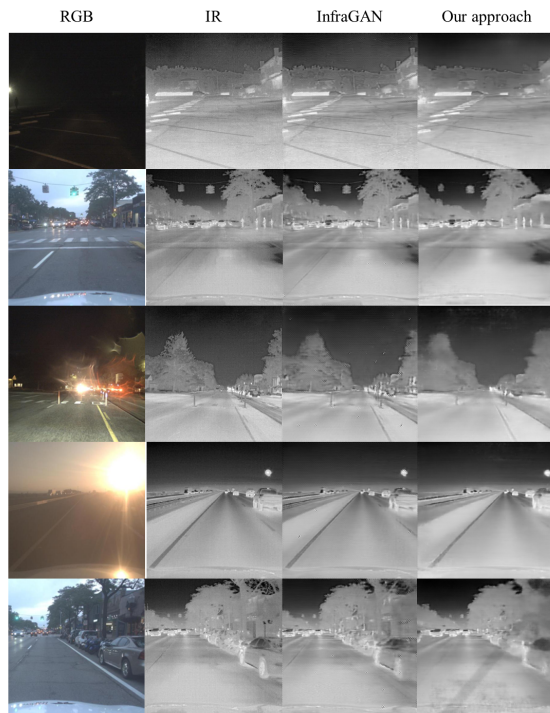


Figure 2: Comparison of Qualitative Results: Original InfraGAN algorithm vs. our enhanced approach. The rows of the Figure showcase various scenes from the FLIR dataset, differing in exposure and content.

hyperparameter optimization with the Optuna framework, we refined the trade-off between the loss components, significantly enhancing InfraGAN’s performance and establishing a framework for further experiments.

Our results indicate that these modifications improve InfraGAN’s ability to generate high-fidelity thermal images with more accurate detail and structural consistency. This approach demonstrates the effectiveness of advanced loss configurations in domain transfer tasks, contributing valuable insights to the field of image synthesis and domain translation. Future work could extend this methodology to other domains and explore additional optimization techniques for further performance gains.

A primary direction for extending this work is to test the methods on additional datasets, such as the Vis-TH dataset for facial expressions (introduced in (Mallat and Dugelay, 2018)). Evaluating the approach on a broader range of data will enhance its generalizability and robustness. Another critical avenue is the modification of the DFT loss. In its current state, the DFT loss behaves similarly to the L^2 norm. Introducing a filter in the DFT loss into a more distinct and potentially effective metric, warranting further exploration. Hyperparameter optimization presents oppor-

tunities for deeper investigation. A key question is whether the optimal hyperparameters differ significantly between datasets or exhibit consistent patterns. Additionally, iterative refinement of hyperparameters should be performed by re-optimizing for each hyperparameter. Furthermore, adopting an analytical approach could further constrain the search space by leveraging inherent relationships, such as the connection between the style loss and perceptual loss. A detailed analysis of the importance of each hyperparameter is also recommended. Understanding parameter importance will inform more targeted and efficient optimization strategies in the future. Finally, alternative accuracy functions beyond LPIPS should be tested to evaluate the model comprehensively. This could provide additional insights into its strengths and areas for improvement. Addressing these recommendations will further refine the methodology and broaden its applicability, leading to more robust and versatile outcomes.

REFERENCES

- Bulatov, D., Burkard, E., Ilehag, R., Kottler, B., and Helmholz, P. (2020). From multi-sensor aerial data to thermal and infrared simulation of semantic 3d models: Towards identification of urban heat islands. *Infrared Physics & Technology*, 105:103233.
- Fuoli, D., Gool, L. V., and Timofte, R. (2021). Fourier space losses for efficient perceptual image super-resolution.
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks.
- Isola, P., Zhu, J., Zhou, T., and Efros, A. A. (2016). Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004.
- Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution.
- Kottler, B., Fischer, S., Strauss, E., Bulatov, D., and Helmholz, P. (2023). Parameter optimization for a thermal simulation of an urban area. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10:271–278.
- Kottler, B., List, L., Bulatov, D., and Weinmann, M. (2022). 3gan: A three-gan-based approach for image inpainting applied to the reconstruction of occluded parts of building walls. pages 427–435.
- 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*, pages 16377–16386.
- López-Rey, A., Ramón, A., and Adán, A. (2023). Hardware/software solutions for an efficient thermal scanning mobile robot. In *ISARC. Proceedings of the International Symposium on Automation and Robotics*

- in Construction*, volume 40, pages 675–682. IAARC Publications.
- Ma, D., Xian, Y., and Li, B. e. a. (2024). Visible-to-infrared image translation based on an improved cgan. *Vis Comput* 40, pages 1289–1298.
- Mallat, K. and Dugelay, J.-L. (2018). A benchmark database of visible and thermal paired face images across multiple variations. In *BIOSIG 2018 - Proceedings of the 17th International Conference of the Biometrics Special Interest Group*. Köllen Druck+Verlag GmbH, Bonn.
- Ordun, C., Raff, E., and Purushotham, S. (2023). When visible-to-thermal facial gan beats conditional diffusion. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 181–185. IEEE.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision*, 115(3):211–252.
- Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition.
- Suárez, P. L. and Sappa, A. (2024). Depth-conditioned thermal-like image generation. In *2024 14th International Conference on Pattern Recognition Systems (ICPRS)*, pages 1–8. IEEE.
- Sun, Q., Wang, X., Yan, C., and Zhang, X. (2023). Vq-infratrans: A unified framework for rgb-ir translation with hybrid transformer. *Remote Sensing*, 15(24).
- Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.
- Xiong, X., Zhou, F., Bai, X., Xue, B., and Sun, C. (2016). Semi-automated infrared simulation on real urban scenes based on multi-view images. *Optics express*, 24(11):11345–11375.
- Yu, Z., Li, S., Shen, Y., Liu, C. H., and Wang, S. (2023). On the difficulty of unpaired infrared-to-visible video translation: Fine-grained content-rich patches transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1631–1640.
- Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251.
- Özkanoglu, M. A. and Ozer, S. (2022). Infragan: A gan architecture to transfer visible images to infrared domain. *Pattern Recognition Letters*, 155:69–76.