# Improving Thermal to RGB Image Translation Using Enhanced pix2pix Models: A Comparative Study with CycleGAN

Theekshana Aturupane
Department of Computer Science and Software Engineering
Auckland University of Technology
Auckland, New Zealand
wbd6141@autuni.ac.nz

Nirmal Kankanamge
Department of Computer Science and Software Engineering
Auckland University of Technology
Auckland, New Zealand
gjq3318@autuni.ac.nz

Nilanjan Fernando
Department of Electrical and Electronic Engineering
Auckland University of Technology
Auckland, New Zealand
msd7060@autuni.ac.nz

***Abstract-*** **In the domain of computer vision, the process of translating thermal to RGB images in low-light conditions encounters unique challenges impacting its accuracy. The study proposes an innovative method to address these challenges by leveraging an enhanced image-to-image translation framework using conditional adversarial networks, commonly referred to as pix2pix. The proposed method uses the paired dataset comprising visible-infrared imagery for low-light vision and the Teledyne forward-looking infrared advanced driver assist systems dataset, meticulously curated for training low-light object detection to train the model. Following a comprehensive evaluation against the Unpaired Image-to-Image Translation using the Cycle-Consistent Adversarial Network, recognised as the CycleGAN framework, and employing established evaluation metrics, the refined pix2pix model exhibits an improved accuracy, precision, and recall.**

***Keywords- GAN, CGAN, CycleGAN, pix2pix, Thermal Image Translation, Structural-Aware GAN, Thermal Image Colourisation, LLVIP, FLIR ADAS***

## I. INTRODUCTION

In modern imaging techniques, thermal imaging fulfils the limitations of traditional visual sensors to identify objects in low-light conditions. In thermal imaging, the conversion of thermal to visible spectrum data (thermal to RGB image translation) is vital in applications, including image-to-image translation, image fusion, and pedestrian detection. However, thermal imaging in low light conditions poses significant challenges, including temporal misalignment between the capture of thermal and RGB images, impacting the accuracy of the translated data. Therefore, it is important to study how traditionally used Cycle Generative Adversarial Network (CycleGAN) models process thermal to RGB image translation in low-light conditions and develop methods to address identified limitations. This study will focus on addressing the limitations of CycleGAN by using an enhanced image-to-image translation with conditional adversarial networks (pix2pix) model capable of synchronising and processing temporally disparate data effectively.

This study utilises two distinct data sets, namely the Visible-infrared Paired Dataset for Low-light Vision (LLVIP), which provides paired visible and infrared images curated for low-light applications [1] and the Teledyne Forward Looking Infrared Advanced Driver Assist Systems (FLIR ADAS) dataset, which offers asynchronous thermal and visible frames tailored for object detection systems [2], to evaluate the efficacy of the proposed pix2pix model.

This study drew inspiration from the recent research advancements in thermal image colourisation [3], structural aware generative adversarial networks (GAN) [4], methodologies that utilise single-mode lightweight encoders [5], sparse GANs for thermal infrared image generation [6] and image fusion technologies for infrared and visible images [7], in developing the enhanced pix2pix model following industry standards.

The study analyses the efficacy of handling temporally misaligned RGB-thermal image pairs of the enhanced pix2pix model by conducting a comparative analysis against CycleGAN model using evaluation metrics. The study aims to provide insight into advancing thermal image translation techniques in low-light vision.

## II. LITERATURE REVIEW

In recent years, thermal to RGB image conversion has received significant advancements where many research efforts focused on enhancing the visual quality and realism of generated images. However, most studies focus on RGB to thermal image conversion, leaving a significant gap in researching the inverse process.

The introduction of conditional generative adversarial networks (CGAN) [8] enabled the conditional generation of RGB images from thermal images, offering fine-grained control over the image generation process compared to traditional GAN architectures. At present, many studies focus on analysing its efficacy in thermal to RGB image translation.

Introducing a cycle-consistent adversarial network capable of learning bidirectional mappings between two image domains without paired data [9] revolutionised image translation using CycleGAN models, which ensured maintaining the semantic consistency of the translated images and preserving essential features, enabling high-quality image synthesis. However, apart from its immense success in image translation tasks, further research is required to assess the applicability of CycleGAN in thermal to RGB image conversion.

In contrast to CycleGAN, the pix2pix models offered a supervised approach to image translation, leveraging paired datasets to learn direct mappings between input and output images [10] by integrating with CGAN and enabling precise control over the translation process to yield visually compelling results with structural consistency and fine-grained details. Similar to CycleGAN, despite its effectiveness in RGB to thermal image translation, the potential of its inverse process remains relatively unexplored.

Recent advancements in GAN architectures, including the attention-based pix2pix model that integrates attention mechanisms to enhance spatial alignment and feature preservation in the translation process [11], contributed to developing specialised models tailored to thermal to RGB image translation. The proposed model achieves significant performance improvements in generating translated RGB images, especially in complex scenarios with varying dynamics and illumination, by selectively attending to informative regions in the thermal images.

Despite advancements, the process lacks accurate capture of minute details in the thermal to RGB image translation due to differences in sensor technology and environmental issues, including varying spectral characteristics and temperature distribution, which require advanced models that can distinguish complex relationships between thermal and RGB features.

Though existing literature focuses on RGB to thermal image translation, exploring the inverse process would provide greater insight and contribute to unlocking the full potential of image translation.

## III. METHODOLOGY

### A. Selection of CycleGAN and pix2pix Models

The selection of models required careful consideration of the architecture and applicability to address the specific requirements of the study. While generative models, including Deep Convolutional Generative Adversarial Network (DCGAN) and Deep Dream, demonstrate impressive capabilities in image generation, CycleGAN and pix2pix offered distinct advantages in the ability to address the specific requirements of the study.

➢ **The CycleGAN Architecture**

The CycleGAN model was selected because of its ability to learn bidirectional mappings between two domains without paired data, which is an advantage in situations where acquiring paired data is challenging or impractical. Unlike DCGAN, which generates images from random noise and DeepDream, which lacks bidirectional mapping, CycleGAN enables the translation of images between domains using bidirectional mapping, causing it to be well-suited for thermal image translation.

**Generator**

- Structured as a U-Net, characterized by an encoder-decoder framework with skip connections facilitating the preservation of spatial details crucial for accurate thermal image translation.
- Leverage power of residual connections within the generator architecture, promoting training stability and facilitating the propagation of gradients, ensuring effective learning across the thermal and RGB images.
- Incorporates attention mechanisms to enable model focus on salient regions within thermal images, enhancing translation quality and preserving critical details during the conversion process.

**Discriminator**

- Utilised a PatchGAN discriminator architecture, which evaluates image patches rather than the entire image, allowing fine-grained discrimination and generating high-quality details required for thermal image translation.
- Integrates spectral normalisation and feature matching techniques to aid in stability and to enable consistent performance across diverse thermal image datasets

As seen in eq. (1), the mathematical representation of the CycleGAN objective function is as follows;

$$L_{CycleGAN} = L_{GAN}\left(G_{T\_RGB}, D_{RGB}\right) + L_{GAN}\left(G_{RGB\_T}, D_{T}\right) \quad (1)$$

Where:

- $L_{GAN}$ denotes the adversarial loss, which guides the generator to produce more realistic images,
- $G_{T\text{-to-RGB}}$ and $G_{RGB\text{-to-T}}$ represent the generators responsible for translating thermal to RGB images and vice versa, respectively, and;
- $D_{RGB}$ and $D_T$ represent the discriminators for RGB and thermal images, respectively.

➢ **The Enhanced pix2pix Architecture**

The enhanced pix2pix model is a fork of the original model [10] modified to increase its accuracy in thermal to RGB image translation. It uses paired images for training, offering deterministic results.

**Generator**

- Embraces a U-Net architecture augmented with symmetric skip connections, facilitating seamless translation from thermal to RGB domains and allowing preservation of spatial details crucial in thermal to RGB image translation.
- Introduced additional convolution layers and adjusted the filter counts in each layer to capture spatial and spectral features in thermal images to enhance the representational power.
- Incorporate residual connections within the generator architecture to enhance training stability by promoting gradient flow to mitigate the vanishing gradient issue.

**Discriminator**

- Introduce enhancements to fortify training stability, model robustness and spectral normalisation with feature-matching techniques to ensure consistent performance across diverse thermal datasets.

- Utilised a patch-based discriminator architecture to allow fine-grained discrimination and offer realistic high-frequency details critical for accurate thermal image translation.

As seen in eq. (2), the mathematical representation of the CycleGAN objective function is as follows;

$$L_{pix2pix} = L_{GAN}(G, D) + \lambda L_{L1}(G) \qquad (2)$$

Where:

- $L_{GAN}$ denotes the adversarial loss,
- $L_{L1}$ represent the $L_1$ loss, and;
- $\lambda$ controls the relative importance of the two components.

As described earlier, the enhanced pix2pix model introduces several modifications, including additional convolutional layers to capture complex spatial details, adjusting filter counts and introducing residual connections between layers to effectively learn complex mappings between thermal and RGB domains, including spectral normalisation, which regulates discriminator weights, preventing mode collapse and enhance the ability to distinguish subtle differences between actual and generated images and feature matching which encourage the generator to produce realistic looking images matching statistics of actual images in feature space to deceive the discriminator.

Furthermore, the enhanced pix2pix utilises a patch-based discriminator which evaluates image patches rather than entire images, allowing the capture of fine-grained details and local structures, resulting in sharper and more coherent image translations. By focusing on smaller image regions, the discriminator can provide informative feedback to the generator, guiding it to produce high-quality image translations with greater fidelity. The modifications in the enhanced pix2pix model allow it to generate accurate translations of thermal images to RGB.

B. Data Pre-processing

As illustrated in Fig. 1, the data was processed before usage to ensure accuracy and compatibility.

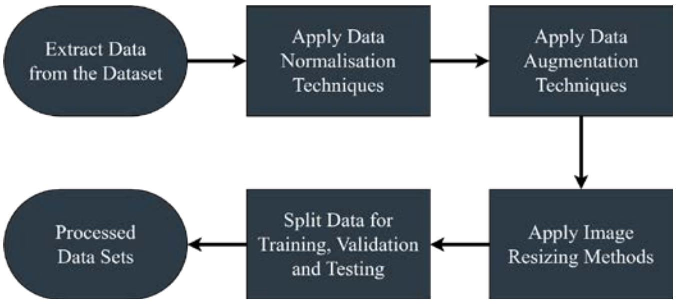

Extract Data from the Dataset
Apply Data Normalisation Techniques
Apply Data Augmentation Techniques
Apply Image Resizing Methods
Split Data for Training, Validation and Testing
Processed Data Sets

Fig. 1. Process of preparation of images for training, validation and testing. Source: Primary

Data pre-processing included four steps;

i. Data Normalisation: Normalise thermal and RGB images to a range of [0, 1] to standardise pixel intensity values to ensure consistency and uniformity in the data distribution across images, facilitating smoother convergence during training.

ii. Data Augmentation: Includes random rotation, flipping, and scaling to include variation in training samples to prevent overfitting and improve the generalisation capability of models.

iii. Image Resizing: Resize all images to a predefined dimension compatible with the input requirements of models to ensure uniformity in image dimensions to enable seamless processing and minimise computational overhead during training.

iv. Split Images: Split images into training, validation, and testing sets using dedicated Python libraries to ensure the availability of independent datasets for model training, validation, and evaluation, thereby enabling robust performance assessment.

A total of 1600 images, including augmented images after randomisation, were used to conduct the analysis. The dataset was split, allocating 1280 images for training and distributing the remainder equally for validation and testing purposes (160 each). This 8:1:1 splitting ratio, widely recognised as the industry standard, was adopted to uphold the integrity of the evaluation process. The application of this ratio-based distribution was consistent across both the LLVIP and FLIR ADAS datasets, ensuring objectivity and minimising potential biases.

C. Experimental Setup

Deep learning frameworks, including TensorFlow and PyTorch, were used to conduct the study. The model training and evaluation were accelerated using Graphics Processing Unit (GPU) accelerators.

D. Model Architecture and Training

**CycleGAN Model Training**

- The CycleGAN model comprises a generator and a discriminator network.
- During training, the generator network learns to translate thermal images to RGB images, while the discriminator network distinguishes between translated and actual RGB images.
- Adversarial training was conducted iteratively, with the generator and discriminator networks updated alternately to minimize their respective adversarial and reconstruction losses
- Hyper-parameters, including learning rate, batch size, and number of training epochs, were specified in the Python code and fine-tuned to achieve optimal performance.

**Enhanced pix2pix Model Training**

- The pix2pix model comprises an encoder-decoder generator and a discriminator network.
- The encoder-decoder generator network learns to generate realistic RGB images from thermal images, guided by an adversarial loss and a reconstruction loss.
- Hyper-parameters, including lambda, learning rate, batch size, and training epochs, were defined in the code and adjusted to balance the competing objectives of image realism and structural preservation.

### E. Evaluation Metrics

The evaluation of model performance required computation of metrics, including peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and mean absolute error (MAE) using the provided Python functions. These metrics provide insights into the perceptual and structural quality of the generated outputs by quantitatively assessing the similarity and fidelity of the translated images to ground truth RGB images.

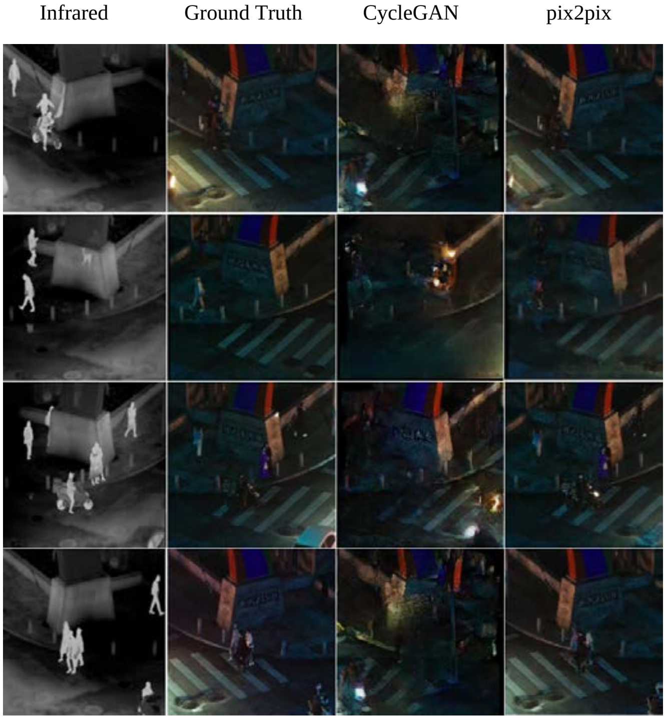
### F. Validation and Testing

The model was tested after the training and validation using the test dataset and was evaluated qualitatively through visual inspection of generated images and quantitatively using the computed evaluation metrics to ensure the robustness and generalization capability of the trained models across diverse datasets.

## IV. ANALYSIS OF DATA

### A. Model Outcomes

Fig. 2 illustrates the results generated by the CycleGAN and enhanced pix2pix models on the LLVIP test dataset.



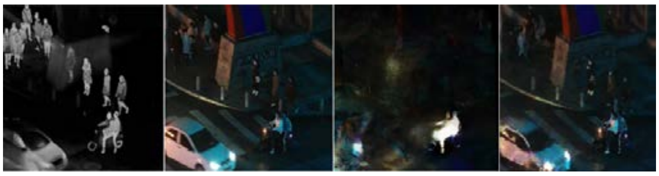


Fig. 2. Results generated by CycleGAN and enhanced pix2pix on LLVIP. Source: Adapted from [1]

Fig. 3 illustrates the results generated by the CycleGAN and enhanced pix2pix models on the FLIR ADAS test dataset.

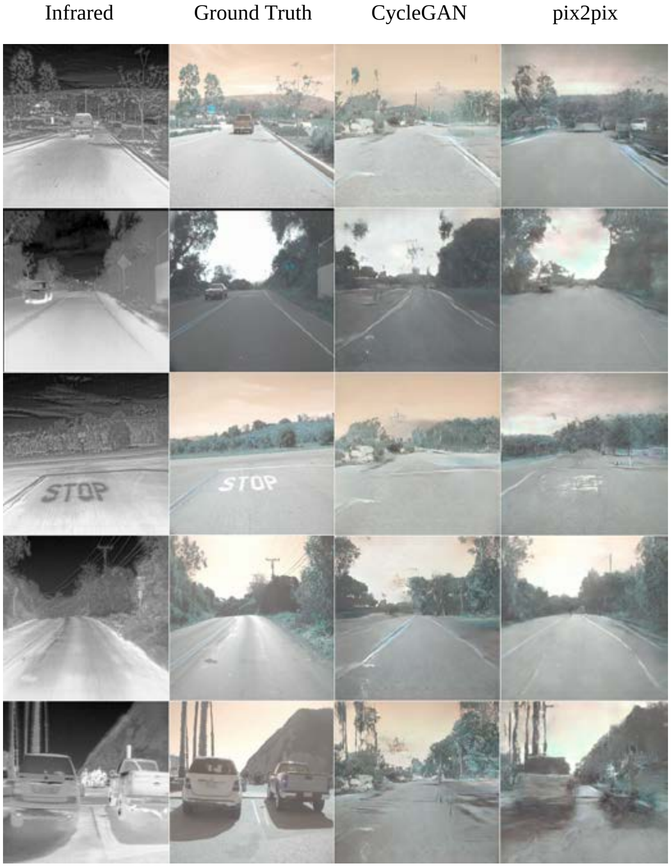


Fig. 3. Results generated by CycleGAN and enhanced pix2pix on FLIR ADAS. Source: Adapted from [2]

### B. Model Evaluation

A detailed analysis assessing the quantitative metrics and the qualitative assessments was conducted on the trained models to demonstrate the performance in thermal to RGB image translation.

**Quantitative Evaluation Metrics**

- Peak Signal-to-Noise Ratio (PSNR): Measures the peak signal-to-noise ratio between the translated RGB images and the ground truth RGB images, where higher values indicate better reconstruction fidelity.
- Structural Similarity Index (SSIM): Computes the similarity between the translated RGB images and the ground truth RGB images based on luminance, contrast, and structure, where values closer to 1 indicate better structural preservation.
- Mean Absolute Error (MAE): Calculates the average absolute pixel-wise difference between the translated

RGB images and the ground truth RGB images, where lower values indicate better image fidelity.

**Qualitative Assessment**

- Visual Inspection: Visual inspection of the generated RGB images to assess the perceptual quality and realism. Key aspects evaluated include colour accuracy, texture preservation, and overall image coherence.
- Comparison with Ground Truth: Side-by-side evaluation of translated RGB images with their respective ground truth RGB counterparts to detect any disparities or artefacts introduced during translation. For evaluation metrics, the method involves computing the pixel-wise differences between the ground truth and predicted images, forming the confusion matrices.

## V. EVALUATION OF DATA

The evaluation results provide valuable insights into the performance of the trained models across different datasets and highlight their effectiveness in thermal-to-RGB image translation. As illustrated in Fig. 4, by analysis of metrics, including accuracy, precision, recall, and F1 score, the strengths, limitations and suitability for specific use cases of each model variant could be determined.

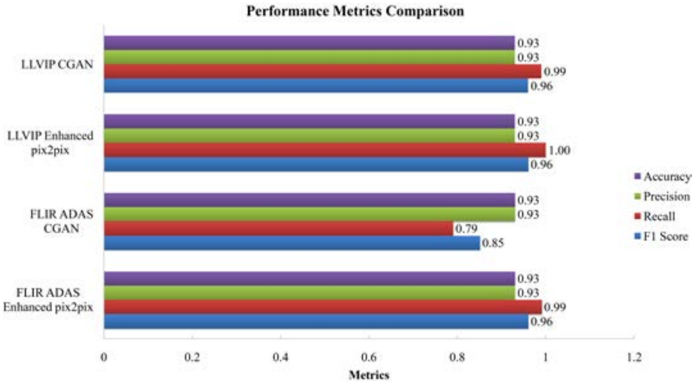


Fig. 4. The performance metrics comparison chart.
Source: Primary

### ➢ LLVIP Dataset

**CycleGAN**

Demonstrate the ability to convert thermal to RGB images accurately and precisely while mitigating erroneous identifications. Higher recall values indicate a higher ability to capture significant thermal attributes. Suitable for scenarios that require intricate preservation of features, including object detection and recognition in surveillance systems. However, ineligible for asynchronous thermal-to-RGB image translation.

**Enhanced pix2pix**

Demonstrated comparable accuracy and precision to the CycleGAN, reporting high recall values indicating high performance in preserving important thermal features. Reported a slightly lower F1 score compared to CycleGAN, indicating a good balance between precision and recall. Suitable for applications that require a balance between feature preservation and spatial accuracy.

### ➢ FLIR ADAS Dataset

**CycleGAN**

Demonstrate the ability to convert thermal to RGB images accurately and precisely while mitigating erroneous identifications. Lower recall and F1 values indicate challenges in capturing and preserving thermal image features during translation. Lower results could be attributed to variations in thermal signatures within the dataset, leading to discrepancies between the generated RGB images and ground truth data. Remains a viable option for applications where spatial alignment is vital.

**Enhanced pix2pix**

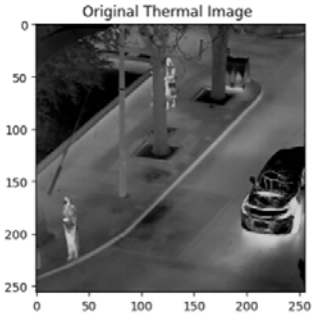
Demonstrated competitive accuracy and precision, with a slight decrease in recall and F1 scores. However, the ability to maintain consistency in thermal image translation across various thermal signatures highlights its robustness and adaptability to complex real-world scenarios.

The evaluation provided valuable insight into the strengths and limitations of different models when performing thermal to RGB image translation. While CycleGAN offers superior feature preservation and spatial accuracy, enhanced pix2pix provides greater flexibility and robustness to dataset variability. Therefore, understanding the strengths and limitations of models allows stakeholders to make informed decisions in using appropriate models for specific use cases.

## VI. CONCLUSION

Evaluation of CycleGAN and enhanced pix2pix models in thermal to RGB image translation provided valuable insight into the model performance characteristics across diverse datasets. While quantitative metrics, including accuracy, precision, recall, and F1 score offer a standardized means of assessing model performance, it is essential to complement these measures with qualitative evaluations to capture the differences in output quality.

Despite comparable quantitative metrics, qualitative analysis revealed distinct advantages of the enhanced pix2pix model particularly in the context of the LLVIP dataset. As illustrated in Fig. 5, the enhanced pix2pix model demonstrates exceptional capabilities in object replication and spatial fidelity, indicating its understanding of scene semantics and finer details, where it accurately replicates vehicle lights and lighted areas with greater intensity, enhancing the realism and visual appeal of the generated RGB images



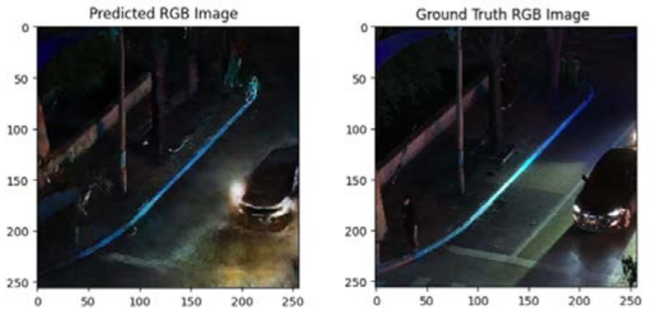


Fig. 5. Replication of special fidelity and scene semantics in thermal to RGB image translation.
Source: Primary

Future research could focus on improving the capabilities of enhanced pix2pix models through advanced techniques, including the progressive growing GAN (PGGAN) and attention mechanisms to improve spatial alignment and feature preservation. Further exploring loss functions tailored to specific applications may aid in addressing challenges related to dataset variability and domain adaptation.

Furthermore, it should be noted quantitative metrics have limitations in fully capturing the perceptual quality and realism of generated images. Therefore, future evaluations should prioritize qualitative assessments leveraging human perceptual studies or adversarial evaluation methods to distinguish minute differences in output quality.

Adopting a holistic approach that integrates quantitative and qualitative evaluations will allow further studies to develop more robust and semantically meaningful models for thermal to RGB image translation applicable in enhancing computer vision in autonomous transportation, surveillance and augmented reality.

## REFERENCES

[1] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, Jan. 2021, "LLVIP: A Visible-infrared Paired Dataset for Low-light Vision," arXiv, doi: https://doi.org/10.48550/arxiv.2108.10831.

[2] Thermal Imaging, Mar. 2024, "FLIR Data Set," Roboflow. [Online]. Available: https://universe.roboflow.com/thermal-imaging-0hwfw/flir-data-set

[3] F. Luo, Y. Li, G. Zeng, P. Peng, G. Wang and Y. Li, "Thermal Infrared Image Colorization for Nighttime Driving Scenes With Top-Down Guided Attention," in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 9, pp. 15808-15823, Sept. 2022, doi: 10.1109/TITS.2022.3145476.

[4] L. Sigillo, E. Grassucci and D. Comminiello, "StawGAN: Structural-Aware Generative Adversarial Networks for Infrared Image Translation," 2023 IEEE International Symposium on Circuits and Systems (ISCAS), Monterey, CA, USA, 2023, pp. 1-5, doi: 10.1109/ISCAS46773.2023. 10181838.

[5] J. Amendola, L. R. Cenkeramaddi and A. Jha, "Image Translation and Reconstruction Using a Single Dual Mode Lightweight Encoder," in IEEE Access, vol. 12, pp. 26787-26799, 2024, doi: 10.1109/ACCESS.2024.3365831.

[6] X. Qian, M. Zhang and F. Zhang, "Sparse GANs for Thermal Infrared Image Generation From Optical Image," in IEEE Access, vol. 8, pp. 180124-180132, 2020, doi: 10.1109/ACCESS.2020.3024576.

[7] H. Lv, B. Deng and X. Li, "Research on Image Fusion Technology of Infrared and Visible Image Based on MST and CNN," 2022 IEEE 4th International Conference on Civil Aviation Safety and Information Technology (ICCASIT), Dali, China, 2022, pp. 1395-1399, doi: 10.1109/ICCASIT55263.2022.9986707.

[8] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," arXiv, Nov. 2014, doi: https://doi.org/10.48550/arXiv.1411.1784.

[9] J. -Y. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2242-2251, doi: 10.1109/ICCV.2017.244.

[10] P. Isola, J. -Y. Zhu, T. Zhou and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 5967-5976, doi: 10.1109/CVPR.2017.632.

[11] Y. Qu, Y. Chen, J. Huang and Y. Xie, "Enhanced Pix2pix Dehazing Network," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 8152-8160, doi: 10.1109/CVPR.2019.00835.