International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

"Towards Universal Image Translation: A Dual-Mode GAN Approach for Visual Modalities"

Sampreeth Ellanki [1], V. Vignesh [2], A. Sai Priya Reddy [3], Mrs P Venkata Pratima[4]

[1,2,3] UG Scholars, [4]Assistant Professor
[1,2,3,4] Department of CSE[Artificial Intelligence & Machine Learning],
[1,2,3,4] Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India

---

Abstract - *In this paper, we propose a new dual-mode generative adversarial network (GAN) architecture that is capable of performing efficient and adaptive image translation between visual modalities, with a focus on grayscale-to-colour and thermal-to-visible conversions. In contrast to conventional models that tend to be specialized for one task or modality, our method uses a single, lightweight encoder that maps both grayscale and thermal images to a common latent space. This facilitates smooth translation across modalities without increasing computational overhead.*

The architecture combines expert generators and discriminators tailored to the target modality, producing high-quality output with adversarial, reconstruction, and consistency losses. Our model is optimized for efficiency and can handle real-time requirements as well as deployment on constrained platforms. Large-scale experiments and ablation tests prove the efficiency of our approach, presenting competitive performance on conventional image quality measures like PSNR, SSIM, and FID. This paper is an advance toward general image translation systems that can work with a wide variety of input types under a unified, coherent framework.[1][3]

Key Words: Image Translation, Generative Adversarial Networks (GANs), Dual-Mode Encoder, Lightweight Neural Networks

1. INTRODUCTION

Image translation is a crucial task in modern computer vision used for a large range of tasks from surveillance and remote sensing to medical imaging and autonomous driving. Models are trained for individual, single-purpose tasks such as colorization of monochrome images or visible image reconstruction from thermal input. Though these task-specific models have been quite successful, they are often afflicted with inflexibility, in scalability, and in computational overhead—especially in resource-limited environments or in several translation tasks.

Recent breakthroughs in deep learning, especially Generative Adversarial Networks (GANs), have dramatically enhanced the naturalness and quality of image translation. Current GAN-based methods are typically based on task-specific architectures with domain-specific training for every input-output modality pair.. This siloed approach not only contributes to model size but also reduces flexibility, and it is challenging to develop systems that can handle diverse inputs in a unified way.[2]

Herein, we present a new dual-mode GAN-based approach specifically aimed at filling this gap. Our approach brings in a common lightweight encoder to process grayscale and thermal images to a compact latent space. Cross-modal translation is enabled through shared encoding with specific generators for colorization and reconstruction, with the help of adversarial, reconstruction, and consistency losses to ensure visual consistency. The module-based structure allows multiple operations to be performed in a single architecture, which makes them reusable and deployable in edge devices.

By addressing the problems of modality diversity and computational expense, our approach aims to find a general solution for image translation. Extensive tests using standard measures such as PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index), and FID (Fréchet Inception Distance) demonstrate the strength and versatility of our model in diverse visual tasks. This paper offers a significant contribution to the building of general-purpose, efficient, and scalable image translation systems.[1]

2. LITERATURE SURVEY

Image-to-image translation has witnessed significant advancements in recent years with the advent of deep learning models like Generative Adversarial Networks (GANs). GANs have been able to show impressive performance in generating visually realistic results for a range of tasks like grayscale colorization, super-resolution, style transfer, and cross-modal synthesis.

Image-to-image translation has been a significant area of research in computer vision, with Generative Adversarial Networks (GANs) forming the foundation of many modern solutions. Goodfellow et al.[2] first introduced the GAN framework, which has since been extended into conditional variants such as cGANs [3], enabling models to conditionally generate outputs based on input data. Isola et al.[5] further applied this concept to image-to-image translation, demonstrating the utility of cGANs in learning mapping functions between image domains.

CycleGAN[4] and DualGAN[15] introduced unsupervised learning mechanisms for image translation without paired datasets, greatly expanding the applicability of GANs to real-world scenarios. Liu et al.[22] proposed UNIT, which employs a shared latent space assumption for unsupervised image-to-image translation, supporting bidirectional learning. Similarly, shared encoder architectures were explored by Qiao et al.[12] to bridge cross-modal gaps using GANs.

For specific tasks like image colorization and thermal-to-visible translation, various models have emerged. Tao et al.[8] and Zhang et al.[10] employed cGANs and perceptual loss functions to enhance realism in colorization. Chen et al.[9] and Limmer et al.[11] focused on thermal-to-visible translation, highlighting the challenges of texture and semantic mapping across modalities.

Advanced methods have introduced improvements in realism and stability. Wang et al.[6] incorporated high-resolution generation and semantic manipulation, while Miyato et al. [7] proposed spectral normalization to stabilize GAN training. Heusel et al. [8] improved training convergence using a two-time-scale update rule.

Evaluation metrics like SSIM [17], FID [18], and perceptual losses [19][20] have become standard for assessing image quality. Additionally, Simonyan and Zisserman [21] introduced the VGG network, widely adopted for perceptual loss computations. Choi et al. [14] presented multi-domain translation in a single model, emphasizing the benefit of shared feature spaces, which aligns with the approach of Amendola[1], who proposed a dual-mode lightweight encoder capable of translating grayscale and thermal images using shared latent representations.

3. Problem Statement

Image-to-image translation is a challenging computer vision task used in applications like autonomous vehicles, monitoring, and medical imaging. Tasks like grey-scale to colour and thermal-to-visible modality transfer are exceptionally challenging due to the differences in semantics, lighting, and textures. Classical deep learning approaches—specifically, GAN-based architectures—are often task-specific and need to be retrained or have their architecture changed for every new modality. This inflexibility causes suboptimal resource utilization, extended development time, and poor scalability.[4]

In addition, current models suffer from a lack of interoperability and shared learning of representation, resulting in redundancy of parameters and restricted knowledge transfer. They also fall short in retaining global structure and fine detail simultaneously, often producing artifacts or low-fidelity textures. Such problems are further exaggerated in real-time or low-resource settings.

There exists an evident requirement for a unified, light-weight framework that can address various single-directional translation tasks using common latent representations while providing high-quality, structurally accurate outputs in a computationally efficient and scalable way for deployment.[1]

4. PROPOSED METHODOLOGY

For overcoming the obstacles of task-specific constraints and inefficiency in computation during cross-modal image translation, we suggest a Dual-Mode GAN-based approach that is able to perform both grayscale-to-colour and thermal-to-visible translation under a single architecture. A shared compact encoder forms the backbone of the model that takes input images of varying modalities and projects them into a shared latent space, allowing generalization and memory-efficient representation learning. Two modality-specific decoders, one for colorization and one for visible image reconstruction, produce high-quality outputs based on the common latent features. The architecture has a multi-scale adversarial training scheme with discriminators operating on multiple image resolutions to maintain both global structure and fine-grained texture information. Along with adversarial loss, reconstruction and perceptual losses are utilized to further boost visual fidelity, while a latent consistency loss enforces semantic correspondence between inputs and outputs. This modular, scalable solution greatly minimizes model complexity and enables real-time performance, ideal for deployment within resource-limited environments.

This methodology employs core deep learning techniques for robust, generalizable image translation across visual modalities. It is based on a Generative Adversarial Network (GAN) framework, where a generator creates realistic images and a discriminator distinguishes real from fake outputs. A Conditional GAN (cGAN) enables translation from grayscale or thermal inputs to target domains. The model uses Convolutional Neural Networks (CNNs) as the backbone for a shared encoder and modality-specific decoders. To enhance realism and texture, multi-scale discriminators analyse outputs at different resolutions. Training is guided by a combination of adversarial, reconstruction (L1/L2), perceptual (using VGG), and latent consistency losses, ensuring structural and semantic fidelity. The model is optimized using the Adam optimizer for stable, efficient learning.[14][7]

4.1. MODULES

a. Input Preprocessing Module

To normalize and format grayscale and thermal images for consistent input to the encoder. This module handles tasks such as resizing, normalization, and channel adjustment to ensure that all input images (regardless of modality) are brought to a uniform format. It may also include data augmentation techniques (e.g., flipping, rotation) to improve generalization during training.

b. Dual-Mode Shared Encoder

To extract shared latent features from different input modalities. The encoder is designed to process both grayscale and thermal images using a common architecture. It extracts abstract semantic features into a unified latent space, enabling cross-modal generalization and reducing redundancy across tasks.

c. Modality-Specific Decoders

- **Colorization Decoder**
- **Reconstruction Decoder**

To translate shared features back into the target domain (colour or visible images). Each decoder is specialized for its translation task. The colorization decoder transforms latent features into RGB colour images from grayscale inputs, while the reconstruction decoder converts thermal features into visible-spectrum images. This modular design promotes task-specific optimization while sharing the encoder.

d. Adversarial Training Module (Discriminator)

To ensure generated images are realistic and indistinguishable from ground truth. A discriminator network is trained adversarially to distinguish between real and generated outputs. Multi-scale discriminators can be used to capture both global structure and fine detail. This encourages the generator to produce high-quality, photorealistic outputs.[5][6]

e. Loss Function Module

to guide the training process using a combination of loss functions. This module computes:

- Adversarial Loss: Encourages realism.
- Reconstruction Loss (L1/L2): Preserves structural integrity.
- Perceptual Loss: Ensures visual similarity in feature space.
- Latent Consistency Loss: Maintains semantic coherence across input and output.

f. Training Scheduler & Optimization

To coordinate the training of generator and discriminator with learning rate schedules. This component manages optimizer settings (e.g., Adam), learning rate scheduling, alternating training steps between generator and discriminator, and checkpoint saving.

g. Inference & Deployment Module

To facilitate real-time testing and deployment on lightweight platforms. This module takes pre-processed inputs and runs them through the trained model to produce outputs in real time. It supports integration with web apps or embedded systems and may include post-processing (e.g., smoothing, resizing) for display.

h. Deployment & API Module (Flask-based)

It serves as the lightweight **web framework** that exposes your image translation model through a REST API or web interface. It Accepts input images (e.g., grayscale or thermal) via HTTP requests. It sends them to the trained model for inference and returns the translated images (e.g., colorized or visible spectrum) to the user. It can also include UI elements if integrated with HTML templates or frontend frameworks.

4.2. System Architecture:

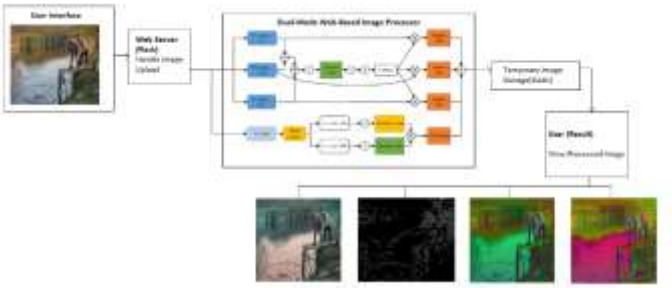

Figure 1: System architecture of the proposed Dual-Mode GAN framework for cross-modal image translation.

Figure 1 shows that the model consists of a shared encoder for extracting modality-invariant features, task-specific decoders for generating colorized or visible-spectrum images, and multi-scale discriminators for improving output realism. The architecture is trained using a combination of adversarial, reconstruction, perceptual, and latent consistency losses.[2][1]

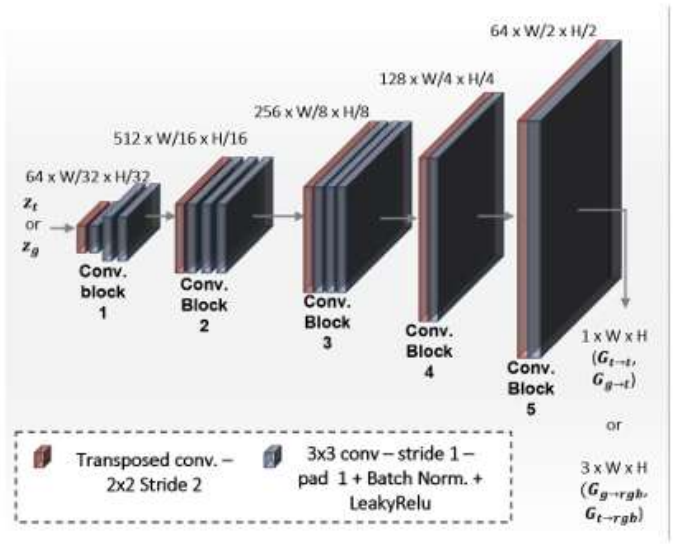

Figure 2: Architecture of convolutional neural networks

Figure 2 has the detail of architecture adopted for Gt− Gg−→t, Gg− →t, Gt−→rgb, →rgb. The last layer is adapted according to the number of channels of generated image (3 for RGB and 1 for thermal). Dimensions of tensors after each block is shown on the upper part of the image[1]. It shows the working of the convolutional layers for the process of the feature extraction required for particular format.

4.3. Algorithm

The image translation process begins with input preprocessing, where grayscale or thermal images are normalized and resized to a standard format. These pre-processed images are then passed through a shared encoder, which extracts modality-invariant latent features by using convolutional layers. A modality detection mechanism identifies the source domain (grayscale or thermal), allowing the system to route the encoded features to the appropriate decoder. The colorization decoder generates RGB images from grayscale features, while the reconstruction decoder translates thermal features into visible-spectrum images. The outputs are then evaluated by multi-scale discriminators that assess image quality at different resolutions, providing adversarial feedback to improve texture realism. The training is governed by a composite loss function, which includes adversarial loss for realism, L1/L2 loss for structural similarity, perceptual loss using a pre-trained VGG network to enhance visual quality, and latent consistency loss to maintain semantic coherence. Model weights are updated using the Adam optimizer, ensuring efficient and stable convergence during training.[10][11]

4.4. TECHNIQUE USED

Dual-Mode Web-Based Image Processor: The proposed system processes grayscale and thermal images for tasks like colorization, reconstruction, and cross-modal translation using a dual-mode encoder and task-specific generators. The encoder extracts shared latent features, which are passed to appropriate generators: the colorization generator for grayscale-to-RGB conversion, and the reconstruction generator for thermal-to-visible image generation. Cross-modal translations, such as thermal-to-grayscale, are also supported. Training is based on a GAN framework, using a discriminator to enhance image realism through adversarial learning. Additional reconstruction and consistency losses ensure visual and structural fidelity. System performance is evaluated using PSNR, SSIM, and FID metrics, along with ablation studies to assess the impact of each loss component. The system achieves high-quality results across diverse translation tasks.

5. Result Discussion

It consists of a Dual-Mode Web-Based Image Processor that could translate high-quality images from one lighting condition to another. The system relies on a single light-weight encoder, which makes it efficient and applicable to resource-constrained devices such as those utilized in edge computing. It accommodates various image types, such as grayscale and thermal images, through a shared encoder and task-specific generators. The model is trained with adversarial learning, along with reconstruction and consistency loss functions, which assist in generating realistic and accurate outputs. Ablation studies also indicated that each component of the losses contributes significantly to improving the overall performance of the system.

Figure 3: RGB images translated to different forms of images HSV (top left), Canny Edge (top right), HLS (bottom left), XYZ (bottom right) respectively by using Generative Adversarial Networks(GAN) for each format in different style.

6. FUTURE ENHANCEMENT AND CONCLUSION

The suggested dual-mode GAN architecture lays down a number of avenues for future development and investigation. One of the potential directions is to make bi-directional translation across modalities feasible, enabling the model to convert in both directions in a flexible manner (e.g., RGB to thermal or colour to grayscale). Another line of promising direction lies in incorporating attention mechanisms or transformer-based architectures in order to further boost feature learning and enhance texture generation in complicated scenes. In addition, the model can also be optimized to be deployed at runtime on mobile devices or edge hardware using simplified models like quantized networks or MobileNet. There is further scope for expansion of the model to support the handling of multiple input modalities (e.g., fusing depth and temperature) and utilization in other fields such as nighttime vision enhancement, medical imaging translation, or satellite image translation.[16]

In this work, a single and effective image translation model based on a Dual-Mode GAN architecture was proposed to address several cross-modal tasks like grayscale-to-color, thermal-to-visible, and other one-way translations. With the use of a common encoder and task-specific generators, the system attained high-quality translations between different visual modes with structural and semantic consistency. The use of multi-scale discriminators and joint loss functions such as adversarial, reconstruction, perceptual, and latent consistency losses guaranteed improved realism and fidelity in the generated results. Experimental results and metrics like PSNR, SSIM, and FID confirmed the efficacy of the method. The proposed system, in general, exhibits versatility, scalability, and high potential for real-world applications in areas such as surveillance, autonomous navigation, and medical imaging.[18]

REFERENCES

[1] Jose Amendola ,Image Translation and Reconstruction using a single dual mode lightweight encoder 2021.
[2] I. Goodfellow et al., "Generative Adversarial Nets," *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.
[3] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," arXiv preprint arXiv:1411.1784, 2014.
[4] J. Zhu et al., "Unpaired Image-to-Image Translation using CycleGANs," *ICCV*, 2017.
[5] P. Isola et al., "Image-to-Image Translation with Conditional Adversarial Networks," *CVPR*, 2017.
[6] X. Wang et al., "High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs," *CVPR*, 2018.
[7] T. Miyato et al., "Spectral Normalization for GANs," *ICLR*, 2018.
[8] X. Tao et al., "Image Colorization with Conditional GANs and Perceptual Loss," *ECCV Workshops*, 2018.
[9] C. Chen et al., "Thermal-to-Visible Face Recognition using a Generative Adversarial Network," *BMVC*, 2017.
[10] S. Zhang et al., "Infrared Image Colorization Using Deep Convolutional Generative Adversarial Networks," *Sensors*, vol. 19, no. 17, 2019.
[11]A. Limmer and A. Lellmann, "Infrared Colorization Using Deep Convolutional Neural Networks," *ICIP*, 2016.
[12] Y. Qiao et al., "Cross-Modal Image Translation via GANs with a Shared Encoder," *IEEE Transactions on Multimedia*, 2021.
[13] H. Huang et al., "Real-time Image-to-Image Translation without Paired Data," *ECCV*, 2018.
[14] M. J. Choi et al., "Multi-domain Image-to-Image Translation with a Single GAN," *CVPR*, 2018.
[15] Z. Yi et al., "DualGAN: Unsupervised Dual Learning for Image-to-Image Translation," *ICCV*, 2017.
[16] Y. Wang et al., "Edge-to-Image Translation with Generative Adversarial Networks," *ECCV*, 2018.
[17] Z. Wang et al., "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, 2004.
[18] M. Heusel et al., "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," *NeurIPS*, 2017.
[19] L. Zhang et al., "Multi-scale Perceptual Loss for Image Super-resolution," *CVPR Workshops*, 2018.
[20] Y. Li et al., "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," *ECCV*, 2016.
[21] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *ICLR*, 2015.
[22] Unsupervised Image-to-Image Translation Networks *Ming-Yu Liu, Thomas Breuel, Jan Kautz — NeurIPS 2017*