GAN-Based PAN-to-RGB Image Translation for Remote Sensing Data

Xiaowei He[a], Yingzi Xiong[a], Xiao Ling[a], Qinghong Sheng[a], Xiao Xu[b]

[a] College of Astronautics, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, Jiangsu, China
hexiaowei@nuaa.edu.cn(Xiaowei He), xyz1771@nuaa.edu.cn(Yingzi Xiong),
xlingsky@nuaa.edu.cn(Xiao Ling), qhsheng@nuaa.edu.cn(Qinghong Sheng)
[b] Yangtze Delta Region Institute of Intelligent Sensing (Nantong), Nantong 226010, Jiangsu, China
uniais@uniais.cn

Keywords: Image translation, PAN-to-RGB , Multiscale features, Remote sensing imagery.

Abstract

Despite the rapid development of satellite sensors, acquiring high-resolution true-color images remains challenging. The reliance on paired panchromatic and multispectral images, as required by traditional techniques such as pansharpening, presents a significant challenge in scenarios where such data are not available. In this paper, a GAN-based PAN-to-RGB model is proposed to establish a novel framework for high-resolution, high-fidelity true-color images generation from remote sensing data. Symmetric luminance-color decoders are leveraged to learn the mapping between luminance, color, and spatial multi-scale features, effectively overcoming the color desaturation, inaccuracies, and distortion common in existing algorithms. By combining CNNs for local feature modeling and transformers for global feature modeling, high-resolution, high-fidelity RGB images were generated in CIELAB space. We conducted experimental validation on Gaofen-7 satellite data, and the results demonstrated that the proposed algorithm effectively mitigates the issues of desaturated, inaccuracy, and distorted colors, achieving significant improvements in key metrics such as FID, CF, and ΔCF.

1. Introduction

Despite the rapid development in satellite sensor technology, the acquisition of high-resolution true-color images remains a significant challenge. Satellites such as GaoFen-7, WorldView, and QuickBird still rely on panchromatic cameras to capture high spatial-resolution panchromatic images (PAN), while multispectral cameras are utilized to obtain lower spatial-resolution multispectral images (MS) (Chen et al., 2024a, Ozcelik et al., 2021). The final high-resolution RGB images are always generated through pansharpening techniques. These techniques effectively combine spatial information from PAN images with spectral information from MS images(Han et al., 2024, Cui et al., 2025). However, these approaches require dual inputs, which imposes limitations in practical applications. This paper proposes an end-to-end translation model trained on PAN-RGB image pairs to facilitate the mapping of input images from the source domain (PAN) to the target domain (RGB). Within this framework, enhancing the performance of the translation model is of great importance. The rapid advancement of deep learning technology, noted for its robust feature extraction and generalization capabilities, provides a novel perspective for image translation. Deep learning models suitable for the PAN-to-RGB translation task are primarily categorized into two categories:

Dual-channel color generation networks. InstColor performs color generation at both the image level and the object level separately, followed by a fusion process to optimize color prediction results for multi-target images (Su et al., 2020). CT2 reformulates the color generation problem as a classification task through color quantization, which significantly enhances the saturation of the predicted outcomes (Weng et al., 2022). ColorFormer incorporates a color memory module to model contextual semantic information across grayscale images in diverse scenes, thereby yielding semantically correct predictions (Ji et al., 2022). DDColor introduces a pioneering query-based Transformer that leverages its strengths in global feature modeling (Kang et al., 2023). By progressively refining semantically-aware color queries, this approach achieves end-to-end high-fidelity image generation. These dual-channel color generation methods operate by converting the ground-truth color image into the CIELAB color space and subsequently generating only the two chromatic channels. As a result, the generated outputs may suffer from color distortion and desaturation.

Three-channel image generation networks. This class of approaches specifically targets the color distortion issues inherent in dual-channel color generation methods. Pix2Pix establishes a pixel-wise image generation framework, constructing a pixel-level mapping between grayscale and color images, which serves as the foundational architecture for this category of methods (Isola et al., 2017). MUGAN focuses on image texture features by extracting multi-scale high-level semantic features and designing an attention mechanism module (Liao et al., 2023). Consequently, this approach yields prediction results characterized by reasonable color distributions and sharp edges. DDGAN enhances the quality of generated color images in terms of edge structure and visual perception by improving the network's feature extraction capability and focusing on basic visual features (Chen et al., 2023). FRAGAN integrates semantic information by combining spatial and channel interaction features from both encoding and decoding stages, thereby enhancing the details, semantics, and contextual capabilities of the generated images (Chen et al., 2024b). However, three-channel image generation methods predict all RGB channels simultaneously. While various strategies have been devised to enhance semantic correctness, these methods often suffer from blurred boundaries, which can easily lead to low color saturation and inaccuracies in the generated results.

To address these issues, a three-decoder PAN-to-RGB network

based on multi-scale features (TDP2RNet) is proposed in this paper, which consists of a multi-scale feature encoder and three decoders. The multi-scale feature encoder utilizes a ConvNeXt backbone to effectively capture multi-scale features from PAN imagery. The three decoders are named as the spatial structure decoder, the color decoder, and the luminance decoder. First, the multi-scale feature encoder is employed to extract features from the input PAN images. Subsequently, multi-scale features are reconstructed by the spatial structure decoder. After that, correlations are established among luminance, color, and the multi-scale feature, which are then individually fused with the spatial structure. Finally, the luminance and color channels are concatenated and transformed from the CIELAB space to the RGB space, producing high-resolution, high-fidelity true-color images.

2. Experimental data

As China’s first civilian sub-meter high-resolution optical stereo mapping satellite, the GaoFen-7 (GF-7) is equipped with dual linear array cameras that have both forward-viewing and backward-viewing capabilities. The backward-viewing camera acquires PAN images with a resolution of 0.65 meters and MS images with a resolution of 2.6 meters, while the forward-viewing camera captures PAN images at a resolution of 0.8 meters. The GF7-Pan2RGB dataset specifically designed for PAN-to-RGB tasks in this paper. This dataset is constructed from GF-7 satellite imagery, encompassing geographic coordinates of 31.9°N and 120.9°E, which includes a portion of Nantong City in Jiangsu Province. It features diverse landforms, such as urban buildings, infrastructure, farmland, and water. The Gram Schmidt (Laben and Brower, 1998) image fusion method was employed to merge backward-view PAN images with MS images, resulting in true-color images with a resolution of 0.65 meters, which serve as ground truth. These fused images, together with the backward-viewing PAN images, form the samples utilized for training and testing the TDP2RNet. We selected two regions primarily comprising urban areas, farmland, and water, as shown in Figure 1. These regions were non-overlappingly cropped to obtain 9,995 sample pairs of size 256 × 256, resulting in the formation of the GF7-Pan2RGB dataset. Subsequently, the dataset was divided into training, validation, and test sets in a ratio of 7:1:2.

3. Methods

In the PAN-to-RGB translation task, discrepancies in spatial structure between PAN images and ground-truth images often lead to performance degradation. Specifically, methods that focus on dual color channel generation tend to produce results with color distortion and low saturation, while general three-channel image generation algorithms frequently suffer from issues of color desaturation and inaccuracies. Transformers excel in modeling global feature (Weng et al., 2022). KANG et al. introduced them into image colorization tasks, achieving promising results on natural images (Kang et al., 2023). Taking DDColor as the baseline and accounting for the unique characteristics of the PAN-to-RGB task, we proposed the TDP2RNet. The architecture primarily consists of a multi-scale feature encoder and three decoders: a spatial structure decoder, a symmetric color decoder, and a luminance decoder. The spatial structure decoder upsamples the features extracted by the encoder to obtain multi-scale features for color and luminance prediction. Meanwhile, the symmetric color and luminance

Figure 1. Two regions for making GF7-PAN2RGB dataset.

decoders learn the correlations between color, luminance, and spatial multi-scale semantic representations from multi-scale feature. This process generates high-resolution, high-fidelity true-color images consistent with the distribution of ground-truth images, effectively alleviating issues such as color distortion, desaturation, and inaccuracies prevalent in current PAN-to-RGB translation algorithms. The discriminator adopts the classic PatchGAN architecture (Isola et al., 2017). The overall architecture is illustrated in Figure 2.

3.1 Spatial Structure Decoder

In the proposed spatial structure decoder, the architecture consists of four upsampling layers and four skip connection layers, configured with channel dimensions of 512, 512, 256, and 256, respectively. PixelShuffle (Shi et al., 2016) is leveraged as the upsampling framework. The skip connection layers concatenate the feature maps extracted by the encoder with those obtained by the upsampling layers at the same scale. This structure enables the the decoder to progressively upsample the multi-scale feature maps ($\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$, and $\frac{H}{32} \times \frac{W}{32}$) extracted by the ConvNeXt encoder, facilitating the reconstruction of high-resolution structures. The multi-scale features derived from the upsampling process are subsequently fed into the symmetric color and luminance decoders to learn the correlations between spatial multi-scale semantic representations and color luminance attributes.

3.2 Symmetric Color and Luminance Decoders

The symmetric color and luminance decoders simultaneously learn the correlations between color, luminance, and spatial multi-scale semantic representations, which improves the alignment between the distributions of predicted images and ground-truth images. The color and luminance decoders share an identical architecture, consisting of three consecutive Transformer-based modules per group, with each group repeated three times. As presented in the Figure 3, these decoders progressively learn the correlations between color luminance

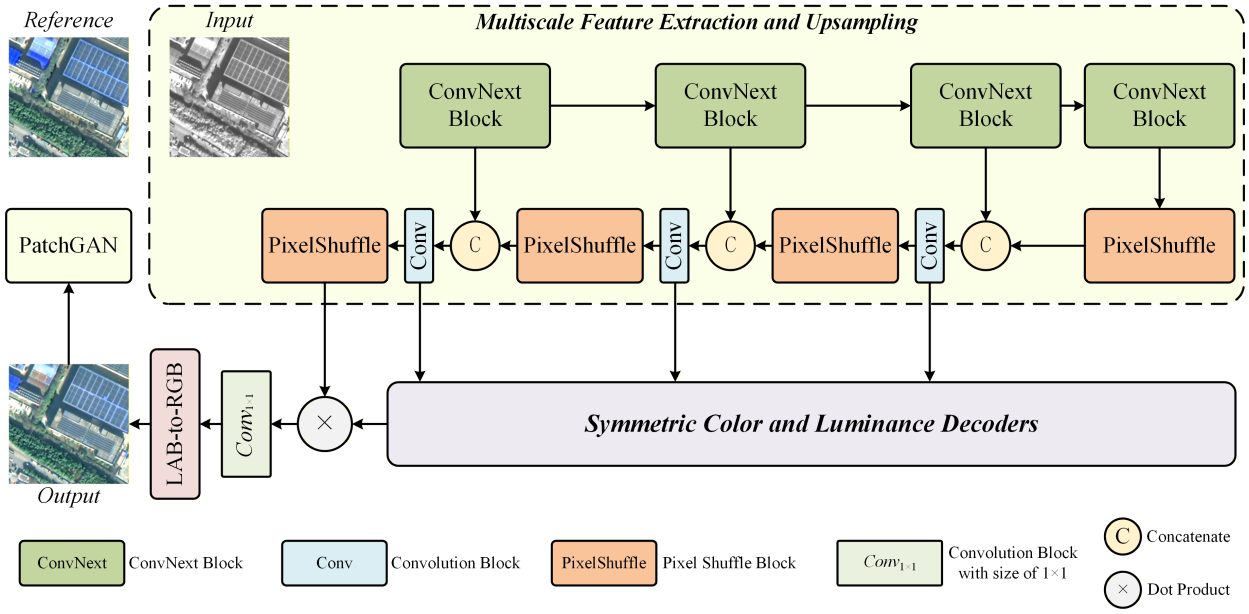

Figure 2. Architecture of proposed TDP2RNet.

and spatial semantic information from the multi-scale feature maps generated by the spatial decoder at resolutions of $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$. The outputs of the two decoders are given by equations 1 and 2.

$$E_c = ColorDecoder(Z_0, F_1, F_2, F_3) \quad (1)$$

$$E_l = lumDecoder(Z_1, F_1, F_2, F_3) \quad (2)$$

where E_c = the outputs of the color decoder
E_l = the outputs of the luminance decoder
Z_0 = the color memory sequences learned via the Transformer-based modules
Z_1 = the luminance memory sequences learned via the Transformer-based modules
F_3 = the feature maps upsampled by the spatial decoder at resolutions of $\frac{H}{4} \times \frac{W}{4}$
F_2 = the feature maps upsampled by the spatial decoder at resolutions of $\frac{H}{8} \times \frac{W}{8}$
F_1 = the feature maps upsampled by the spatial decoder at resolutions of $\frac{H}{16} \times \frac{W}{16}$

3.3 Fusion Modules

We employ two fusion modules. The inputs of the color fusion module are the spatial decoder output $E_i \in R^{C \times H \times W}$ and the color decoder output $E_c \in R^{K_1 \times C}$. C denotes the output dimension. H and W represent the height and width of the feature maps. And K_1 indicates the number of color queries. Similarly, the luminance fusion module receives $E_i \in R^{C \times H \times W}$ and the luminance decoder output $E_l \in R^{K_2 \times C}$, where K_2 indicates the number of luminance queries. The calculations for the two fusion modules are formulated as follows in Equations 3 and 4.

$$\hat{F_1} = E_c \cdot E_i; \hat{y_{ab}} = Conv_{1\times1}(\hat{F_1}) \quad (3)$$

$$\hat{F_2} = E_l \cdot E_i; \hat{y_l} = Conv_{1\times1}(\hat{F_2}) \quad (4)$$

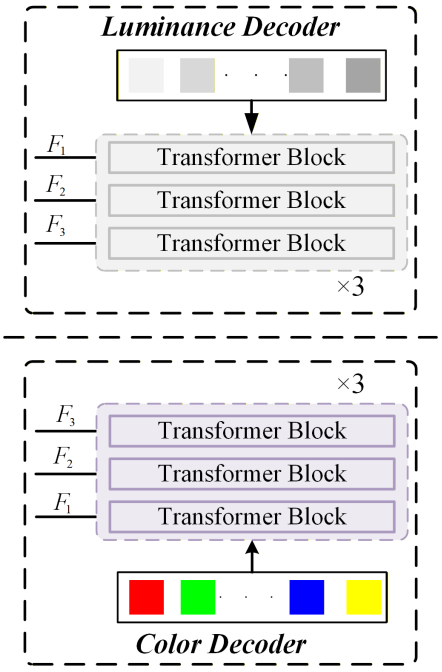

Figure 3. The structure of symmetric color and luminance decoders.

where $\hat{F_1} \in R^{K_1 \times H \times W}$ = the fused color features
$\hat{F_2} \in R^{K_2 \times H \times W}$ = the fused luminance features
$\hat{y_{ab}}$ = the predicted color
$\hat{y_l}$ = the predicted luminance

3.4 Loss Functions

TDP2RNet generates the color, luminance, and spatial structure of true-color images from multi-scale features through three decoders. Consequently, four loss functions are employed to ensure the quality of the generated results, which are color, luminance, semantics, and adversarial loss.

$\mathbf{L_{ab}}$: The L_1 distance between the generated colors and the ground-truth colors is computed to ensure accuracy in color representation in the results.

$\mathbf{L_l}$: The L_1 distance between the generated luminance and the ground-truth luminance is to ensure the accuracy of luminance in the results.

$\mathbf{L_{per}}$: The pretrained VGG16 model is utilized to extract deep features from both the generated results and the ground-truth images. Subsequently, the L_1 distance between these features is computed to enforce semantic fidelity in the output.

$\mathbf{L_{adv}}$: The PatchGAN discriminator is employed to assess the disparity between the generated outputs and the ground-truth images, thereby ensuring reliability of results.

Therefore, the total training loss function as follows equation 5.

$$L_{total} = \lambda_1 L_{ab} + \lambda_2 L_l + \lambda_3 L_{per} + \lambda_4 L_{adv} \tag{5}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ = the weights of the corresponding losses

4. Experiments and Results

4.1 Experimental settings

4.1.1 Implementation Details: All models were trained and tested on a computer equipped with an RTX 4080Ti GPU. The batch size for training the network was set to 4, with an initial learning rate set to 1×10^{-4}. A total of 174900 training iterations were conducted, implementing a 50% learning rate decay every 20000 iterations starting from the 40000th iteration until 140000th iteration. We used the AdamW optimizer and set $\beta_1 = 0.9$, $\beta_2 = 0.99$, $weight\,decay = 0.01$. Additionally, loss weights $\lambda_1 = \lambda_2 = 0.1$, $\lambda_3 = 5$, $\lambda_4 = 1$. For the color decoder and luminance decoder, $K_1 = K_2 = 100$.

4.1.2 Comparison Methods: To evaluate the effectiveness of proposed method, six state-of-the-art algorithms were selected for comparison, including two dual-channel color generation networks (ColorFormer (Ji et al., 2022) and DDColor (Kang et al., 2023)) and four three-channel image generation approaches (Pix2Pix (Isola et al., 2017), MUGAN (Liao et al., 2023), DDGAN (Chen et al., 2023), and FRAGAN (Chen et al., 2024b)).

4.1.3 Evaluation Indicators: The performance of PAN-to-RGB translation models were evaluated using five established metrics: Structural Similarity (SSIM) (Wang et al., 2004), Peak Signal-to-Noise Ratio (PSNR) (Huynh-Thu and Ghanbari, 2008), Fréchet Inception Distance (FID) (Heusel et al., 2017), Colorful (CF), and the difference in CF relative to ground-truth images ΔCF (Hasler and Suesstrunk, 2003). These metrics, which are widely adopted in the relevant literature, ensure a reliable and credible quantitative assessment of the generated outputs (Wang et al., 2025, Xia et al., 2025, Xia et al., 2024).

4.2 Results Evaluation

4.2.1 Visual Qualitative Comparison: As demonstrated in the Figure 4 and Figure 5, four representative areas are chosen for a detailed comparison, which are building, shadow, vegetation, and playground. Among all the compared methods, TDP2RNet achieves the superior visual performance. It exhibits the highest color fidelity to the ground truth, with sharp edges, reasonable colors, and high saturation. Furthermore, the image possesses rich textural details and correct semantics. As presented in Figure 4 (a), the blue color of the roof predicted by TDP2RNet most closely approximates the ground truth, despite the presence of desaturated patches. Among the other six methods, ColorFormer and FRAGAN demonstrate relatively better performance with higher blue saturation. In contrast, MUGAN produces a lighter blue hue for the roof. However, it achieves a uniformity in coverage that is comparable to the ground truth. Pix2Pix, DDColor, and DDGAN exhibit obvious abnormal color blocks in their prediction results. In the shadow region shown in Figure 4 (b), both TDP2RNet and the other six algorithms exhibit relatively clear edge details. However, the results of TDP2RNet exhibit colors closest to the ground truth. Similarly, for the vegetation area in Figure 5 (a), our method produces a more vivid green compared to the other methods. Regarding the prediction of the playground runway in Figure 5 (b), ColorFormer and DDColor suffer from obvious color desaturation, resulting in a lighter shade of red. Pix2Pix, DDGAN, and FRAGAN exhibit severe color inaccuracies, such as green and white artifacts. MUGAN achieves the best performance among the comparative methods. However, it still lags behind TDP2RNet in terms of color saturation.

4.2.2 Quantitative Evaluation: Table 1 presents the evaluation metrics, with the best results highlighted in bold. The proposed network in this paper achieves the lowest FID score, demonstrating superior feature distribution consistency with ground-truth images. This indicates that TDP2RNet can generate high-resolution, high-fidelity true-color RGB results. Additionally, TDP2RNet attains the maximum value of the CF indicator, suggesting that it can produce more vivid color results. Since a high CF value does not always indicate superior visual quality, we utilized ΔCF to quantify the difference in the CF indicator between the predicted results and the ground-truth images. TDP2RNet achieves the smallest ΔCF value, indicating the highest color fidelity among the compared methods. The SSIM of TDP2RNet is second only to DDGAN and MUGAN, reaching 0.905, which demonstrates its effective preservation of the high-resolution advantage inherent in PAN images. Although its PSNR performance is lower than that of MUGAN and DDGAN, this likely stems from the joint operation of its luminance and color decoders, which prioritize optimal visual quality over some noise reduction. Therefore, the analysis presented above indicates that while TDP2RNet focus on distribution consistency (measured by FID) and color vibrancy (measured by CF) in relation to ground-truth images, it successfully achieves an optimal balance across all evaluation metrics, including FID, CF, PSNR, SSIM, and ΔCF.

4.3 Ablation Analysis

4.3.1 Pixel loss vs. separate loss: In Table 2, the comparison between model 3 (which calculates the pixel loss of RGB images) and model 4 (the complete TDP2RNet structure and parameters) indicates that calculating the L_1 loss separately for luminance and color is more effective than directly computing the pixel loss of the RGB image. This can likely be attributed to the relative independence of the luminance and color decoders, which allows each component to optimize its specific functionality more effectively.

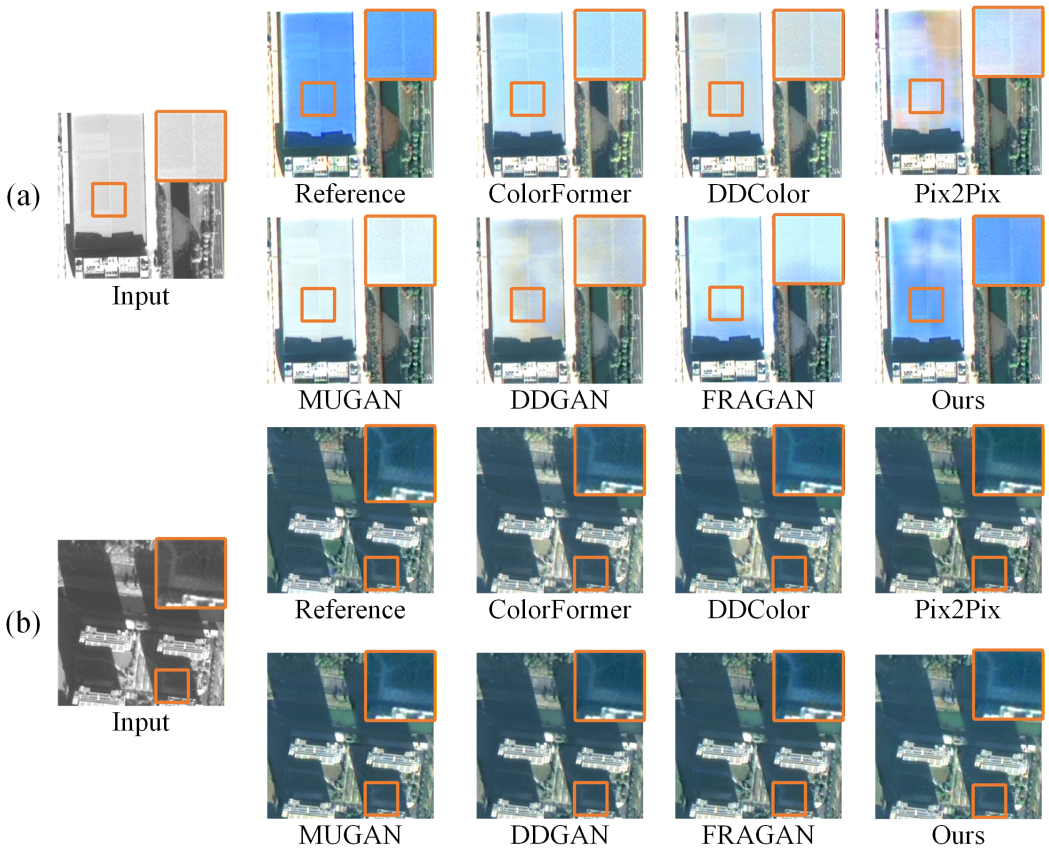

Figure 4. The results of our TDP2RNet and the six comparison methods on (a) Buildings and (b) Shadows.

Methods	Evaluation Indicators				
	FID (↓)	CF (↑)	ΔCF (↓)	PSNR (↑)	SSIM (↑)
ColorFormer	25.101	18.383	3.690	18.532	0.872
DDColor	28.277	20.309	1.764	18.507	0.857
Pix2Pix	28.912	19.239	2.834	28.473	0.883
MUGAN	14.531	19.834	2.239	**29.859**	0.907
DDGAN	20.242	17.920	4.153	29.782	**0.922**
FRAGAN	14.773	20.290	1.783	29.316	0.903
TDP2RNet	**14.109**	**22.699**	**0.626**	29.495	0.905

Table 1. Quantitative comparison of different methods.

4.3.2 Number of decoders: We designed the single decoder model (Model 1) to predict both color and luminance in this section, enabling it to directly output in CIELAB space from the PixelShuffle decoder. These CIELAB images were subsequently converted to RGB images. Conversely, the two-decoder structure (Model 2) simultaneously predicts luminance and color information using a single Transformer-based decoder. Table 2 presents that TDP2RNet demonstrates superior performance in terms of quantitative indicators. The values of FID, PSNR, and SSIM are at their peak, while ΔCF is only slightly lower than that of model 3.

4.3.3 Number of luminance queries: In this paper, the query number for the color decoder is set at 100 (Kang et al., 2023). However, the optimal query number for the luminance encoder requires further verification. Consequently, the luminance query numbers are set at 50, 75, 100, 125, and 150 for the ablation experiments. The results are presented in Table 3. It is evident that the FID, PSNR, and SSIM metrics achieve their optimal values when the number of queries is 100.

5. Conclusion

High-resolution true-color imagery holds significant potential for applications such as true-color 3D visualization and semantic segmentation. However, constrained by current satellite imaging mechanisms, PAN and MS images are often acquired by separate sensors, rendering the acquisition of high-resolution high-fidelity true-color imagery a complex process. In this paper, a GAN-based multiscale feature-based PAN-to-RGB model was proposed to generate high-resolution, high-fidelity true-color images generation from PAN images. We conducted experimental validation on Gaofen-7 satellite data. The spatial structure, texture, and color of the results are consistent with the ground-truth images, and the colors are naturally realistic and vibrant. Furthermore, TDP2RNet demonstrated the best performance among all methods across FID, CF, and ΔCF indicators, with improvements over the comparative methods ranging from 0.422 to 1.049 for FID, 2.39 to 4.779 for CF, and 1.138 to 3.527 for ΔCF.

Nevertheless, it is acknowledged that the approach presented

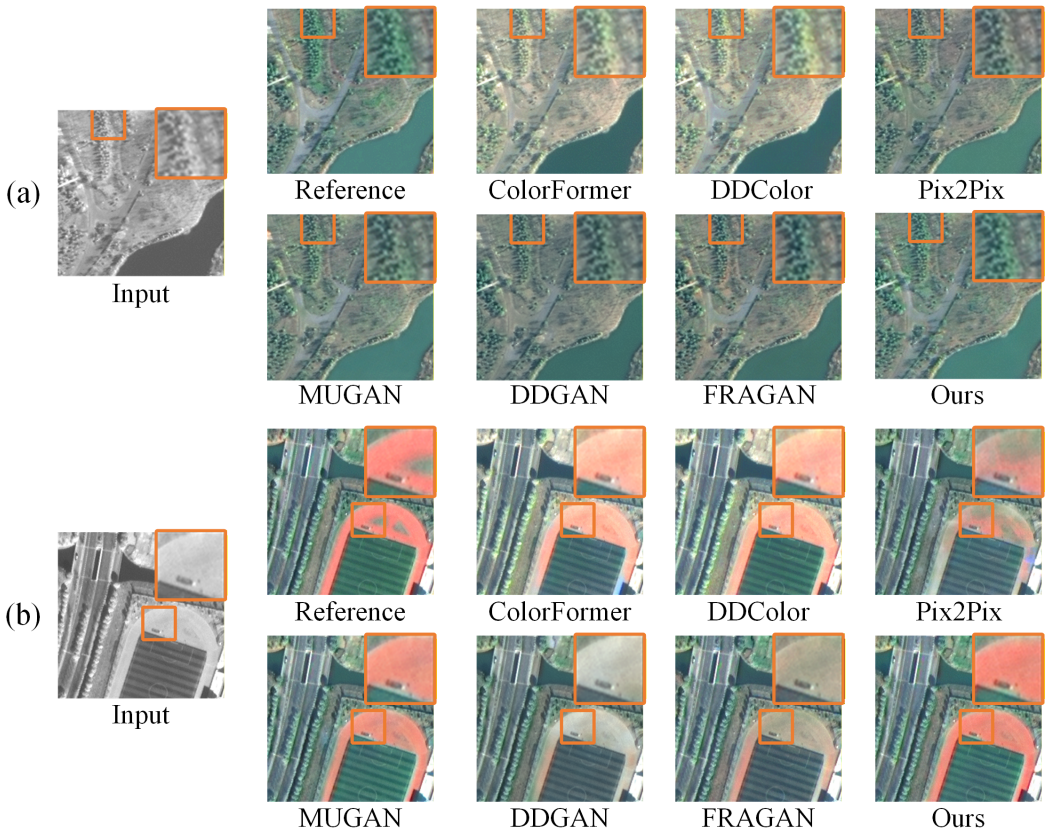

Figure 5. The results of our TDP2RNet and the six comparison methods on (a) Vegetation and (b) Playgrounds.

Methods	Evaluation Indicators				
	FID (↓)	CF (↑)	ΔCF (↓)	PSNR (↑)	SSIM (↑)
model 1	18.240	**23.210**	1.137	27.449	0.885
model 2	14.686	22.804	0.731	29.000	0.893
model 3	17.040	21.550	**0.523**	29.281	0.898
model 4	**14.109**	22.699	0.626	**29.495**	**0.905**

Table 2. Ablation on loss type and number of decoders.

herein still possesses limitations and opportunities for enhancement. Three principal directions for future research are identified:

1. Conducting an in-depth exploration of the correlations between spatial semantic features and color luminance attributes to refine the quality of true-color imagery.
2. Incorporating additional prior knowledge, such as spatial relationships and object classes, to constrain the predictions and correct color inaccuracies.
3. Enhancing the model’s interpretability to ensure the reliability of its predictions from a mechanistic perspective, thereby making it more suitable for downstream tasks such as object detection and semantic segmentation.

Acknowledgements (optional)

This work was supported by in part by the National Natural Science Foundation of China under Grant 42401423; in part by the Natural Science Foundation of Jiangsu Province under Grant BK20241403.

References

Chen, P., Huang, H., Ye, F., Liu, J., Li, W., Wang, J., Wang, Z., Liu, C., Zhang, N., 2024a. A Benchmark GaoFen-7 Dataset for Building Extraction from Satellite Images. *Scientific Data*, 11(1), 187.

Chen, Y., Zhan, W., Jiang, Y., Zhu, D., Guo, R., Xu, X., 2023. DDGAN: Dense Residual Module and Dual-stream Attention-Guided Generative Adversarial Network for Colorizing near-Infrared Images. *Infrared Physics & Technology*, 133, 104822.

Chen, Y., Zhan, W., Jiang, Y., Zhu, D., Xu, X., Hao, Z., Li, J., Guo, J., 2024b. A Feature Refinement and Adaptive Generative Adversarial Network for Thermal Infrared Image Colorization. *Neural Networks*, 173, 106184.

Cui, Y., Liu, P., Ma, Y., Chen, L., Xu, M., Guo, X., 2025. Pan-sharpening via Predictive Filtering with Element-Wise Feature Mixing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 219, 22–37.

Han, Y., Chi, H., Huang, J., Gao, X., Zhang, Z., Ling, F., 2024. TemPanSharpening: A Multi-Temporal Pansharpening Solution Based on Deep Learning and Edge Extraction. *ISPRS*

Numbers	Evaluation Indicators				
	FID (↓)	CF (↑)	ΔCF (↓)	PSNR (↑)	SSIM (↑)
50	14.540	22.765	0.692	29.276	0.900
75	14.769	22.673	0.600	29.178	0.900
100	**14.109**	22.699	0.626	**29.495**	**0.905**
125	14.160	**23.032**	0.959	29.261	0.902
150	14.466	22.553	**0.480**	29.138	**0.905**

Table 3. Ablation on the number of luminance decoder.

Journal of Photogrammetry and Remote Sensing, 211, 406–424.

Hasler, D., Suesstrunk, S. E., 2003. Measuring colorfulness in natural images. *Human Vision and Electronic Imaging VIII*, 5007, 87–95.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, Red Hook, NY, USA, 6629–6640.

Huynh-Thu, Q., Ghanbari, M., 2008. Scope of Validity of PSNR in Image/Video Quality Assessment. *Electronics Letters*, 44(13), 800.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A. A., 2017. Image-to-Image Translation with Conditional Adversarial Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5967–5976.

Ji, X., Jiang, B., Luo, D., Tao, G., Chu, W., Xie, Z., Wang, C., Tai, Y., 2022. ColorFormer: Image Colorization via Color Memory Assisted Hybrid-Attention Transformer. S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, T. Hassner (eds), *Computer Vision – ECCV 2022*, 13676, Cham, 20–36.

Kang, X., Yang, T., Ouyang, W., Ren, P., Li, L., Xie, X., 2023. DDColor: Towards Photo-Realistic Image Colorization via Dual Decoders. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 328–338.

Laben, C. A., Brower, B. V., 1998. Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening.

Liao, H., Jiang, Q., Jin, X., Liu, L., Liu, L., Lee, S.-J., Zhou, W., 2023. MUGAN: Thermal Infrared Image Colorization Using Mixed-Skipping UNet and Generative Adversarial Network. *IEEE Transactions on Intelligent Vehicles*, 8(4), 2954–2969.

Ozcelik, F., Alganci, U., Sertel, E., Unal, G., 2021. Rethinking CNN-Based Pansharpening: Guided Colorization of Panchromatic Images via GANs. *IEEE Transactions on Geoscience and Remote Sensing*, 59(4), 3486–3501.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., Wang, Z., 2016. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1874–1883.

Su, J.-W., Chu, H.-K., Huang, J.-B., 2020. Instance-Aware Image Colorization. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7965–7974.

Wang, P., Chen, Y., Huang, B., Zhu, D., Lu, T., Mura, M. D., Chanussot, J., 2025. MT_GAN: A SAR-to-optical Image Translation Method for Cloud Removal. *ISPRS Journal of Photogrammetry and Remote Sensing*, 225, 180–195.

Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E., 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4), 600–612.

Weng, S., Sun, J., Li, Y., Li, S., Shi, B., 2022. CT2: Colorization Transformer via Color Tokens. *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, Berlin, Heidelberg, 1–16.

Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., Timotfe, R., Van Gool, L., 2025. DiffI2I: Efficient Diffusion Model for Image-to-Image Translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3), 1578–1593.

Xia, M., Zhou, Y., Yi, R., Liu, Y.-J., Wang, W., 2024. A Diffusion Model Translator for Efficient Image-to-Image Translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), 10272–10283.