Semantically Guided Attention-Based SG-U-Net for Thermal Image Colorization

Sargis Hovhannisyan
Institute for Informatics and Automation Problems of NAS RA Yerevan, Armenia
e-mail: sargis.hovhannisyan@ysu.am

Sos Agaian
College of Staten Island (CSI) and Graduate Center, CUNY New York City, USA
e-mail: sos.agaian@csi.cuny.edu

***Abstract*—Thermal infrared cameras are vital for computer vision in challenging conditions where normal cameras fail. For applications like autonomous navigation and surveillance, it's crucial to translate thermal images into full-color representations. This is difficult because thermal cameras capture heat radiation instead of reflected light, leading to hurdles including the complex temperature-to-color relationship, noisy thermal images, and the absence of aligned thermal-RGB datasets.**

This work proposes a new end-to-end deep learning model for this colorization problem. Our pipeline's first step is a unique preprocessing stage that enhances thermal image quality. The main network uses a U-Net-based architecture with attention mechanisms to convert the enhanced thermal data into an RGB image. To ensure contextually correct colors, we employ a multitask learning strategy that combines colorization with semantic segmentation. This forces the model to learn objects in the scene, leading to more logical color output. Quantitative metrics reveal that our model produces images with high perceptual quality. Comparisons show our method surpasses current techniques in generating realistic and semantically coherent color images, achieving state-of-the-art performance.

***Keywords*—Image-to-image translation, multitask learning, thermal image colorization.**

I. Introduction

High-quality imagery is crucial for computer vision, but system performance often degrades in challenging conditions like low illumination or poor visibility [1, 2, 3]. While conventional cameras struggle in such environments, thermal infrared (TIR) imaging provides a robust alternative, offering resilience to illumination changes and the ability to penetrate obscurants like fog [4, 5, 6]. However, raw TIR images lack the color and fine-grained detail necessary for effective human interpretation and automated analysis. This limitation makes the colorization of thermal images—translating them into realistic, full-color visuals—a critical area of research.

Recent thermal colorization methods can be categorized as supervised or unsupervised. Supervised approaches often produce blurry results with poor semantic clarity [7, 8]. Unsupervised image-to-image translation models, such as those based on Generative Adversarial Networks (GANs), can handle unpaired datasets but frequently struggle to generate realistic textures and maintain semantic consistency [9, 10, 11, 12, 13]. These existing methods often fail to produce visually plausible and semantically coherent color images, highlighting a significant gap in the field.

To address these limitations, this paper proposes a novel, end-to-end pipeline for thermal image colorization. Our main contributions are threefold:

1. A dedicated **preprocessing framework** to significantly enhance the quality and detail of input thermal images before colorization.
2. A **multitask learning architecture** that jointly performs colorization and semantic segmentation, using semantic information to guide the generation of contextually accurate colors.
3. A **comprehensive evaluation** demonstrating that our method not only achieves state-of-the-art visual quality but also improves the performance of downstream tasks like object detection and segmentation.

II. Proposed Method

We propose an enhanced thermal infrared colorization network inspired by the PearlGAN [13] architecture, aiming to significantly improve the visual quality and semantic accuracy of translated thermal images, as shown in Figure 1. Our approach introduces critical modifications and additions designed to tackle inherent challenges in thermal imagery, such as noise and limited semantic clarity.

Initially, the attention mechanism in PearlGAN, which utilizes a CycleGAN-based [9] architecture, is completely removed. This decision is motivated by our objective to integrate a more semantically robust attention strategy aligned with the segmentation information. Instead, we embed a dedicated segmentation module and leverage multitask training to explicitly incorporate semantic features. The segmentation network employs a Residual Channel Attention Network [14] architecture, known for its effectiveness in capturing detailed semantic representations.

We introduce a simplified convolution-based attention module that effectively generates channel-wise attention maps from segmentation features to leverage semantic information during colorization. Specifically, the encoder features from our primary thermal translation network are multiplied channel-wise with the generated attention maps before

feeding them into the decoder. This ensures that semantic context directly guides the image translation, enhancing realism and semantic coherence in the resulting colorized images.

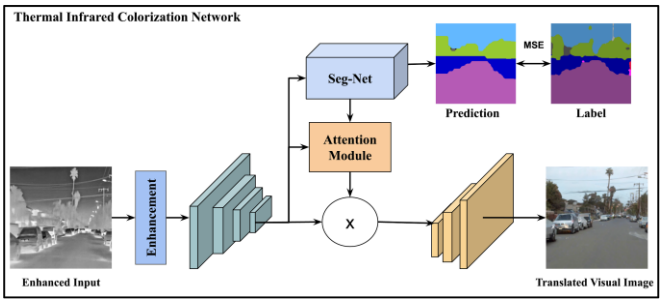

Figure 1. Architecture of the proposed method for thermal image translation

A critical innovation of our methodology is including a pre-colorization thermal image enhancement step using the [15] enhancement method. This enhancement addresses the intrinsic noise and low-quality characteristics of thermal infrared imagery. By improving image visibility and extracting clearer, more informative features, the enhancement substantially aids subsequent processing tasks, particularly segmentation and edge extraction. Enhanced images exhibit fewer artifacts and improved detail retention, which is crucial for accurate semantic labeling and feature extraction.

We adopt an unsupervised semantic segmentation technique described in [16] for segmentation supervision. This method generates high-quality segmentation maps from thermal imagery without manual labeling. Leveraging a memory regularization approach accurately predicts reliable semantic boundaries and classes within thermal images, significantly enhancing the training accuracy of our segmentation module.

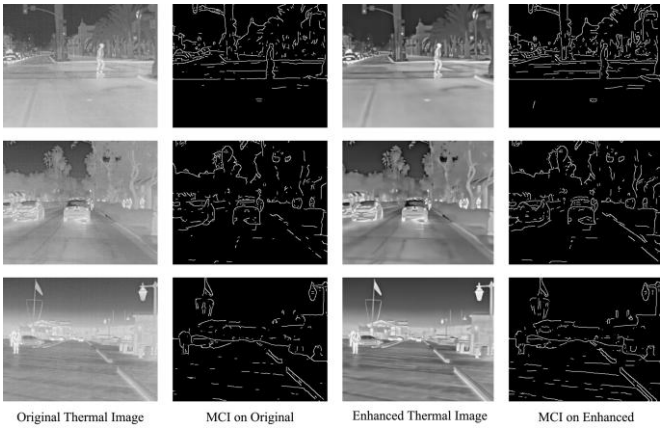

Figure 2. MCI edge detection method results before and after the enhancement method

We preserve the core PearlGAN training paradigm regarding loss functions, excluding losses directly tied to their original attention mechanism. Specifically, we retain edge-aware loss functions, acknowledging their importance in maintaining structural details. Additionally, our enhanced thermal input significantly reduces artifacts previously observable in edge maps generated by the MCI method [17] initially integrated into PearlGAN, as shown in Figure 2. Visual assessments further highlight that our enhancement step markedly improves the fidelity and accuracy of edge delineation.

Our proposed method systematically integrates thermal image enhancement, semantically-informed attention, and segmentation-guided multitask training. This comprehensive approach significantly improves the visual quality and semantic accuracy of colorized thermal infrared images, showing notable advancements compared to baseline methods.

III. Experimental Results

This section presents a comprehensive qualitative and quantitative analysis comparing our proposed method, the Semantically Guided U-Net (SG-U-Net), against prominent GAN-based thermal-to-visible image translation techniques, such as CycleGAN, DRIT++, U-GAT-IT, UNIT, and PearlGAN. We aim to demonstrate our method's significant improvements regarding realism, detail preservation, and overall structural accuracy.

Figure 3. Qualitative comparison of existing methods

A. Qualitative comparison

Figure 3 illustrates qualitative comparisons of translation outcomes across different methodologies on selected thermal images. The translation results highlight several apparent shortcomings in baseline approaches. CycleGAN tends to generate unrealistic textures and frequently oversimplifies structural details, causing essential objects like vehicles and pedestrians to blend into their surroundings or become indistinct. DRIT++, while generally producing brighter images, often leads to overexposure, merging adjacent objects, and losing finer details such as trees and pedestrians. U-GAT-IT achieves improved realism but struggles significantly with clarity, particularly in preserving detailed object boundaries and avoiding object-background fusion. UNIT performs well for larger structures but typically fails to preserve minor details, resulting in blurred representations of distant objects like pedestrians and finer tree branches. PearlGAN, despite demonstrating effectiveness in retaining prominent objects, introduces excessively dark or indistinct regions, reducing the visibility of smaller objects and often inaccurately merging closely positioned items. In contrast, our proposed method consistently yields superior results, clearly translating thermal images into visually coherent daytime scenarios. Our approach accurately captures fine structural details, maintains natural illumination without unnatural brightness or darkness, and provides more apparent differentiation among closely positioned objects. The

qualitative comparison shows that our method demonstrates enhanced realism, producing translations that accurately reflect scene contents and maintain structural and semantic integrity.

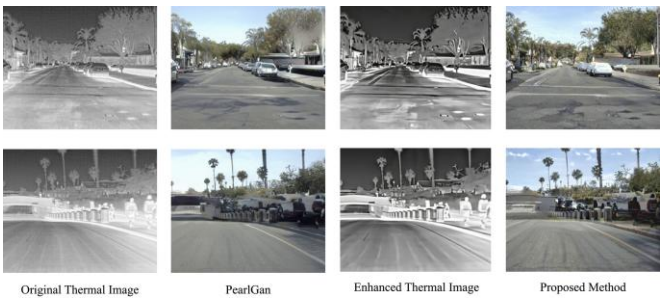

Figure 4. Qualitative comparison with the baseline method

Figure 4 further illustrates a detailed comparative evaluation between our method and the baseline PearlGAN, emphasizing the impact of our thermal image enhancement process. The visual results demonstrate noticeable improvements when employing our pre-colorization enhancement step. Enhanced thermal images result in significantly more accurate translations, characterized by sharper edges, improved visibility of structural details, and fewer translation artifacts than those generated directly from the original thermal images. This comparison demonstrates the efficacy and necessity of incorporating the enhancement stage into the colorization pipeline, substantially improving translation quality and ensuring semantic consistency and visual realism.

Table 1. Quantitative comparison of the proposed pipeline with existing methods

	BRISQUE	NIQE	PIQE
CycleGAN	22.17	8.57	12.70
UNIT	25.50	7.81	9.12
UGATIT	26.14	6.73	7.63
DRIT++	24.23	8.56	9.49
PearlGAN	23.43	6.15	9.25
SG-U-Net	**21.61**	**5.94**	**6.82**

B. Quantitative comparison

To objectively evaluate the performance of our proposed method, we employed three widely recognized non-reference image quality assessment metrics:

- Natural Image Quality Evaluator (NIQE) [18]: This metric assesses image quality based on statistical patterns found in natural images.
- Perception-based Image Quality Evaluator (PIQE) [19]: This metric evaluates the visual perceptual quality of images by quantifying noticeable distortions.
- Blind Image Spatial Quality Evaluator (BRISQUE) [20]: This method measures image quality relying solely on statistical regularities of spatial features without requiring reference images.

A lower score indicates better image quality, improved realism, and greater perceptual authenticity across all these metrics. The comparative performance results for these metrics are detailed in Table 1. Our proposed method consistently achieves the lowest scores compared to existing techniques, highlighting its effectiveness in producing images with enhanced perceptual realism and natural visual characteristics.

Table 2. Evaluation of object detection accuracy on translated images using various methods (IoU: 0.50)

	Person	Bicycle	Car	mAP
CycleGAN	21	3.2	38.4	20.9
UNIT	17.5	11.2	20.5	16.4
UGATIT	11.8	1.1	37.6	16.5
DRIT++	20.6	2.1	48.5	23.8
PearlGAN	59.3	25.1	75.4	53.3
SG-U-Net	**63.8**	**29.8**	**79.1**	**57.5**

Additionally, to validate the practical utility of our generated images, we assessed object detection performance using the detection model [21]. We computed Average Precision (AP), which measures the precision of detection at different recall thresholds, and used mean Average Precision (mAP) across multiple object classes as a comprehensive performance measure. As detailed in Table 2, our method obtained the highest mAP scores, clearly outperforming all methods across all evaluated object categories.

Additionally, semantic coherence was further examined through segmentation analysis, utilizing the segmentation method [22] for color images based on the Intersection over Union (IoU) metric, as shown in Table 3. Our proposed pipeline achieved higher mean IoU values across the semantic classes, demonstrating a strong retention of semantic information within the transformed imagery.

These quantitative assessments demonstrate that our approach significantly enhances image quality and boosts downstream task performance, confirming its suitability and effectiveness for real-world applications.

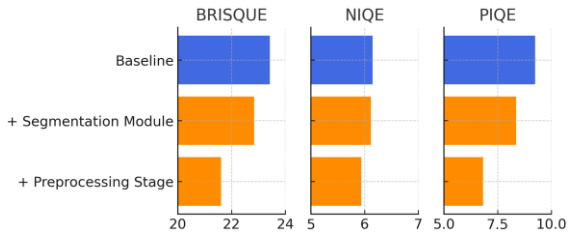

Figure 5. Effectiveness of each module in the proposed pipeline

IV. Ablation Study

We conducted ablation experiments to thoroughly investigate the individual contributions of the image enhancement preprocessing step and the segmentation-guided multitask learning strategy within our proposed pipeline. Specifically, we systematically modified our network by independently removing each component, allowing for a clear assessment of their specific impacts on overall translation performance. The quantitative results of these ablation tests are summarized in Figure 5. These findings clearly illustrate each component's unique and significant contributions toward enhancing the pipeline's effectiveness. The enhancement preprocessing step notably improves the translated images' clarity and structural detail, while the segmentation multitask

Table 3. Semantic segmentation results using various image translation methods (i.e., pixel-wise classification into predefined object categories)

	Road	Building	Sky	Person	Car	Truck	Bus	Traffic Sign	Motorcycle	mIoU
CycleGAN	95.6	39.1	90.8	60.8	78.0	0	0	0	0	40.5
UNIT	96.2	60.3	92.1	64.5	71.5	0	0	0.2	14.6	44.4
UGATIT	94.3	18.4	89.0	23.7	59.0	0	0	0	0	31.6
DRIT++	97.3	29.4	78.4	28.0	78.9	0	0	0	0	34.7
PearlGAN	97.7	73.1	93.4	73.2	82.7	0.1	0	0	0	46.7
SG-U-Net	**98.3**	**79.5**	**96.8**	**89.3**	**92.7**	**1.5**	11.8	0	**18.1**	**54.2**

learning mechanism significantly enriches semantic accuracy and realism. These results confirm that both elements are essential for achieving the highest quality in thermal-to-visible image translation.

V. Conclusion

This paper presents a novel thermal infrared image colorization method integrating a crucial thermal enhancement preprocessing step and a segmentation-guided multitask learning strategy. This preprocessing step significantly reduces inherent noise and enhances structural clarity, critical for accurate colorization. Our method effectively captures detailed semantic representations by employing an RCAN-based semantic segmentation module combined with semantically informed attention mechanisms, further improving the quality and realism of the translated images. Comprehensive experiments demonstrate the clear advantages of the proposed pipeline over existing methods, as indicated by superior performance across established quality metrics and downstream computer vision tasks, including object detection and semantic segmentation. The ablation studies confirm the essential contributions of preprocessing enhancement and multitask learning components. Overall, our approach significantly advances the field of thermal image colorization, offering robust, realistic, and highly applicable visual outputs suitable for practical use in diverse vision-based applications.

Acknowledgment

This work was supported by the Higher Education and Science Committee (project No. 25FAST-1B001) and partly by the Advance Research Grants from the Foundation for Armenian Science and Technology, funded by Sarkis and Nune Sepetjians.

References

[1] H. Gasparyan, S. Hovhannisyan, S. Babayan and S. Agaian, "Iterative Retinex-based decomposition framework for low light visibility restoration," *IEEE Access,* vol. 11, pp. 40298--40313, 2023.

[2] S. Hovhannisyan, H. Gasparyan, S. Agaian and A. Ghazaryan, "AED-Net: A single image dehazing," *IEEE access,* vol. 10, pp. 12465--12474, 2022.

[3] S. Hovhannisyan, H. Gasparyan and S. Agaian, "EOD-Net: enhancing object detection in challenging weather conditions using an innovative end-to-end dehazing network," *2023 Twelfth International Conf. on Image Processing Theory, Tools and Applications (IPTA)*, 2023.

[4] E. Yoshida, S. Kobayashi, H. Mukai, N. Uthumphirat, Y. Nakamura, N. Ogino and T. Yokotani, "Proposal and prototyping on wildlife tracking system using infrared sensors," *2022 International Conference on Information Networking (ICOIN)*, 2022.

[5] S. Kim, W. Kim, J. Park and K. Yeo, "Human detection in infrared image using daytime model-based transfer learning for military surveillance system," *2023 14th International Conference on Information and Communication Tech. Convergence (ICTC)*, 2023.

[6] Z. Wang, S. Li and K. Huang, "Cross-Modal Adaptation for Object Detection in Thermal InfraRed Remote Sensing Imagery," *IEEE Geoscience and Remote Sensing Letters,* 2025.

[7] U. Qayynm, Q. Ahsan, Z. Mahmood and M. A. Chcmdary, "Thermal colorization using deep neural network," *2018 15th International Bhurban Conf. on Applied Sciences and Technology (IBCAST)*, 2018.

[8] T. Wang, T. Zhang and B. C. Lovell, "EBIT: Weakly-supervised image translation with edge and boundary enhancement," *Pattern Recognition Letters,* vol. 138, pp. 534--539, 2020.

[9] J.-Y. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *Proceedings of the IEEE international conf. on computer vision*, 2017.

[10] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," Proceedings *of the European conf. on comp. vision (ECCV)*, 2018.

[11] M.-Y. Liu, T. Breuel and J. Kautz, "Unsupervised image-to-image translation networks," *Advances in neural information processing systems,* p. 30, 2017.

[12] H.-Y. Lee, Y.-H. Li, T.-H. Lee and M. S. Aslam, "Progressively unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," *Sensors,* vol. 23, no. 15, p. 6858, 2023.

[13] 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," *IEEE Transactions on Intelligent Transportation Systems,* vol. 23, no. 9, pp. 15808--15823, 2022.

[14] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong and Y. Fu, "Image super-resolution using very deep residual channel attention networks," *Proceedings of the European conf. on computer vision (ECCV)*, 2018.

[15] S. Hovhannisyan, S. Agaian, K. Panetta and A. Grigoryan, "Thermal Video Enhancement Mamba: A Novel Approach to Thermal Video Enhancement for Real-World Applications," *Information,* vol. 16, no. 2, p. 125, 2025.

[16] Z. Zheng and Y. Yang, "Unsupervised scene adaptation with memory regularization in vivo," *arXiv preprint arXiv:1912.11164,* 2019.

[17] K.-F. Yang, C.-Y. Li and Y.-J. Li, "Multifeature-based surround inhibition improves contour detection in natural images," *IEEE Transactions on Image Proc.,* vol. 23, no. 12, pp. 5020--5032, 2014.

[18] A. Mittal, R. Soundararajan and A. C. Bovik, "Making a “completely blind” image quality analyzer," *IEEE Signal processing letters,* vol. 20, no. 3, pp. 209--212, 2012.

[19] N. a. P. D. Venkatanath, M. C. Bh, S. S. Channappayya and S. S. Medasani, "Blind image quality evaluation using perception based features," *2015 twenty first nat. conf. on communic. (NCC)*, 2015.

[20] A. Mittal, A. K. Moorthy and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing,* vol. 21, no. 12, pp. 4695--4708, 2012.

[21] C.-Y. Wang, A. Bochkovskiy and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.

[22] A. Tao, K. Sapra and B. Catanzaro, "Hierarchical multi-scale attention for semantic segmentation," *arXiv preprint arXiv:2005.10821,* 2020.