# Night Vision: Footage Enhancement and Colorization using Hierarchical Transformer Architecture

By

Udoy Saha
ID: 23341134

Anindya Majumdar
ID: 23241062

Sudmun Hakim Soudho
ID: 24341291

& Md. Irfan Hossain
ID: 24341069

BRAC University

Submitted to Dr. Ashraful Alam for the coursework of CSE463.

Department of Computer Science and Engineering
BRAC University
January 2025

# Abstract

Night vision enhancement and colorization remain complex challenges due to the inherent lack of detail, contrast, and color in low-light imagery. This paper introduces a novel approach that leverages a Hierarchical Transformer Architecture to address these challenges by effectively enhancing and colorizing night vision footage. A curated subset of the ImageNet dataset was employed, reducing the original dataset from 1.3 million to 40,000 images for training, 5,000 for testing, and 5,000 for brightness-based validation. The model emphasizes the integration of semantic understanding and detailed texture restoration to produce visually coherent and realistic results. By capturing both global context and local features, the proposed method enriches color representation while preserving structural integrity. Extensive experiments demonstrate that our approach significantly outperforms existing methods in generating vibrant, detailed, and semantically consistent images, offering a promising solution for practical night vision applications.


Keywords: Image Colorization, Hierarchical Transformer Architecture, Vision Transformer (ViT), Deep Learning, Low-Light Imaging, Semantic Consistency.

# Introduction

Night vision technology plays a pivotal role in numerous applications, including surveillance, security, defense, and autonomous systems. However, traditional night vision systems primarily focus on amplifying brightness and contrast, often producing grayscale images that lack rich color and detailed texture. This limitation hampers visual interpretation and situational awareness in low-light environments. Recent advancements in deep learning have introduced promising solutions for image enhancement and colorization, yet challenges remain in achieving semantic consistency and color richness. Existing approaches, such as DDColor, Color-UNet++, ColorFormer, and Real-Time User-Guided Image Colorization, have made significant strides in automatic image colorization. DDColor introduces a dual-decoder framework that separates semantic and texture processing, improving color consistency and accuracy. However, it faces challenges in computational efficiency and lacks temporal consistency for video applications. Color-UNet++ leverages an improved UNet++ architecture with YUV color space integration to address artifacts and gradient loss but is limited by dataset diversity and loss function optimization. ColorFormer utilizes a hybrid-attention transformer with a color memory module to balance semantic understanding and vivid color generation, achieving state-of-the-art performance. Meanwhile, Real-Time User-Guided Image Colorization blends automatic processing with user interactivity, offering flexibility but still depending on manual input for optimal results.

To overcome these limitations, we propose a novel approach employing a Hierarchical Transformer Architecture for night vision footage enhancement and colorization. Our method captures both global contextual information and fine-grained local details, enabling the generation of vibrant and semantically accurate images. A curated subset of the ImageNet dataset—reduced from 1.3 million to 40,000 images for training, 5,000 for testing, and 5,000 for brightness-based validation—was used to train the model to recognize and reconstruct complex visual patterns in low-light conditions. The goal of this research is to bridge the gap between enhanced night vision imagery and practical usability by restoring realistic color and detail. By integrating advanced transformer-based mechanisms to model semantic relationships while preserving texture and structure, our approach produces high-quality visual outputs. This work aims to set a new benchmark in night vision enhancement and colorization, offering substantial improvements over existing methods and unlocking new possibilities for real-world applications. The following sections will review related work in night vision enhancement and image colorization, elaborate on the proposed model architecture and methodology, present experimental results, and analyze the performance of our approach in comparison to state-of-the-art techniques.

# Literature Review

One of the most common approaches to colorize images deals with manually targeting specific attributes of the image, which is generally reliable but incredibly tedious. One way to perform this more efficiently is to adopt a method using neural networks, which could automatically complete the task. There are some lingering issues, however, regarding the inefficiency of single-decoder frameworks, concerning global metabolic processes and intricate write details. One of the more intriguing approaches is the recognition of how semantics and texture are procured within images through separate decoders, which are both tied together with a guiding fusion module. Moreover, once texture is realized, a new loss function helps to achieve the target of color consistency and accuracy. In order to validate its effectiveness, DDcolor was put through a series of benchmark datasets that proved its visual appeal far surpassed its predecessor's colorization algorithms like InstColor and DeOldify. Dual image colorization has its drawbacks, however—the high two-decoder model, lack of three-dimensional images, and user evaluations. Furthermore, the algorithms simply do not incorporate temporal consistency for video colorization. Some of the improvements the dual image colorization could adopt are through the use of lightweight modules for computational efficiency, and indeed testing on a wider range of dimensions would paint a better picture.

The research entitled "Color-UNet++: A resolution for colorization of grayscale images using improved UNet++" showcases a new method for grayscale image colorization with the modified UNet++ architecture. The method helps with the typical drawbacks of deep learning-based colorization, such as the loss of gradients, checkerboard artifacts, and limitation of expressive power. Throwing in the YUV color space and specialized convolution blocks, the model learns more effectively and also inhibits artifacts. The experimental results on the LSUN and LFW datasets have shown the high quality of the colorization approach against the ones that are currently relied upon as the best by other researchers in the field. Moreover, this research underlines the superiority of YUV over RGB in relation to colorization and verifies its efficiency through quantitative indices such as PSNR and RMSE. Even though the proposed formula is very promising, the paper is flawed. Firstly, the shallowness of the dataset available could affect further processing of the model on more complicated and waggly objects. For instance, the approach's cost and the conclusions about the associated errors are two aspects that could have been discussed more. Specifically, the use of parameters in loss function weighting can further be fine-tuned, and the test can be performed across datasets globally allowing for more accurate results. To enhance the article, further studies could use more and distinctly different datasets, in addition to adjusting the model to cater for a wide range of activities. Also, considering the use of new loss functions would further guarantee the model's robustness. Bringing in user studies totally devoted to the measurement of selected colors' qualities demonstrates incredible insights as well. Finally, a clear comparison with newer methods, e.g., GAN approaches, will further steel

the paper's contributions.

The paper "Night Vision Colorization from Color Mapping to Color Transferring" addresses an approach to the problem related to enhancement of the interpretation of night vision imagery in the form of multispectral grayscale images, which are converted into colorized formats to look as if it is a scene of a naturally illuminated day. It contains a dataset of multispectral images taken using FLIR cameras, which include near infrared, long wave infrared, and RGB bands. Among traditional statistical techniques such as histogram and joint histogram matching (JHM), the study also reviews an updated version of a VGG-Net convolutional neural network (CNN) aimed at employing deep learning color transferring techniques. The results indicate the superiority of the CNN-based approach over the statistical approaches in terms of realistic and naturalistic colorization, while the fusing of NIR and LWIR images gives the best AI performance. High performance, where training employs fused NIR and LWIR, does come with drawbacks and high training costs and also strong technological requirements. Furthermore, the colorization of LWIR-only images was relatively poor because this type of data was lacking in the training set. This pointed out the necessity for further dataset enhancement as well as model enhancement to cater for different Night Vision conditions.

Image segmentation for night vision surveillance has become increasingly critical with the rise of security and monitoring systems operating under low-light or no-light conditions. Traditional methods for night vision processing often rely on heuristic algorithms, such as thresholding, edge detection, and morphological operations, which struggle with complex scenes, varying lighting conditions, and noise. Early machine learning approaches, including support vector machines (SVM) and random forests, showed some improvement but were limited by their reliance on handcrafted features and poor adaptability to diverse environments.The advent of deep learning has revolutionized the field of image segmentation, particularly through convolutional neural networks (CNNs) and their variants. Networks such as U-Net and DeepLab have demonstrated exceptional performance in general segmentation tasks by learning hierarchical representations directly from raw image data. However, these models often require adaptation for night vision applications, where the images typically suffer from low contrast, high noise levels, and unique spectral characteristics of infrared imaging.Recent advancements have focused on integrating domain-specific enhancements, such as using denoising modules, contrast enhancement techniques, and feature refinement layers, to improve segmentation under night vision conditions. The combination of low-light enhancement methods with segmentation networks has further improved accuracy. For example, methods employing multi-scale feature extraction and attention mechanisms effectively address the challenges of detecting small and faint objects in noisy night vision images.The study under review extends these advancements by leveraging deep learning-based segmentation specifically tailored for night vision surveillance cameras. The proposed model integrates pre-processing steps for noise

reduction and contrast enhancement while utilizing a deep learning architecture optimized for segmentation under low-light conditions. The paper's results underscore the significance of domain-specific adaptations, achieving improved segmentation accuracy compared to existing approaches and contributing to the broader adoption of deep learning in night vision applications.

The paper "ColorFormer: Image Colorization via Color Memory Assisted Hybrid-Attention Transformer" presents an innovative approach to automatic image colorization by addressing two critical challenges: semantic consistency and color richness. Traditional colorization methods often suffer from either inadequate semantic understanding or poor color diversity, limiting their practical application.The proposed method introduces a novel hybrid-attention transformer architecture integrated with a color memory module. The transformer-based encoder utilizes a Global-Local Hybrid Multi-head Self-Attention (GL-MSA) mechanism to model both local and global dependencies effectively, ensuring semantic consistency across diverse scenes. This innovation reduces computation complexity while preserving a large receptive field, enhancing the network's ability to interpret the grayscale input's context.The color memory module in the decoder is a unique addition. It stores semantic-color mappings and provides adaptive color priors for the decoder, enriching the output with diverse and vivid colors without requiring external reference images. This feature enables the model to produce results that are both visually appealing and semantically accurate.Experimental results validate the model's superiority over state-of-the-art methods across multiple datasets, including ImageNet, COCO-Stuff, and CelebA-HQ. The proposed model achieves the lowest Frechet Inception Distance (FID), indicating the realism and quality of generated images. Moreover, it demonstrates competitive Colorfulness Scores (CF) while maintaining high efficiency, achieving real-time inference speeds of 40 FPS on a V100 GPU. Qualitative comparisons reveal that ColorFormer outperforms existing methods in handling complex scenes, delivering more natural and stable colorization. Ablation studies confirm the effectiveness of the GLH-Transformer encoder and the color memory module. The architecture's end-to-end design simplifies the process while boosting performance, addressing the limitations of multi-stage methods.In summary, ColorFormer represents a significant advancement in automatic image colorization, combining semantic understanding and diverse color generation in a computationally efficient manner. The approach sets a new benchmark for future research in this domain.

The paper "Real-Time User-Guided Image Colorization with Learned Deep Priors" by Zhang et al. presents a deep learning-based system for interactive image colorization, bridging user-guided and automatic approaches. Traditional user-guided methods often require extensive manual interaction, while automatic systems lack flexibility and struggle with ambiguities in color choices. This paper addresses these limitations by introducing a convolutional neural network (CNN) that directly maps grayscale images and sparse user inputs to colorized outputs in real-time.The system incorporates two core components:

Local Hints Network and Global Hints Network. The Local Hints Network allows users to provide sparse color points, which the model propagates across the image using learned semantic and spatial relationships. The Global Hints Network, on the other hand, integrates global statistics, such as histograms and average saturation, enabling users to influence overall color tones. Both components employ an end-to-end learning framework, allowing seamless integration of local and global inputs during training and inference.A unique feature of the system is its data-driven color palette, which suggests probable colors based on the context, enhancing the usability for novice users. The model was trained on the ImageNet dataset, using simulated user inputs to ensure robust performance without the need for extensive user interaction data. The architecture utilizes a U-Net backbone, leveraging both low- and high-level features for accurate and efficient colorization.Quantitative evaluations demonstrate that the proposed method achieves state-of-the-art performance in terms of PSNR and fooling rate in real vs. fake tests. User studies show that even with minimal training, users can quickly produce photorealistic results. Furthermore, the system supports creative applications, allowing unusual and artistic colorizations guided by user preferences. In conclusion, this work successfully combines deep learning with user interactivity, offering a versatile and efficient solution for real-time image colorization, with implications for graphics, photography, and artistic applications.

The paper "A Local-Coloring Method for Night-Vision Colorization Utilizing Image Analysis and Fusion" proposes a solution for transformation of night vision images focusing on the unnatural colorization and the manual scene matching steps of previous methods. Two bands such as infrared and image intensification of NV images are employed along with a vast database of images of nature taken under sunlight, disaggregated into schemes such as plants, roads, water. The proposed model, referred to with the phrase 'local-coloring', combines non-linear diffusion for smoothing, clustering and merging based segmentation, and automatic segment recognition based on a nearest-neighbor classifier. Furthermore, color mapping is improved through the combination of histogram matching and statistical matching techniques in a de-correlated lab color space. Results demonstrate that the images produced from Night Vision using this method are more natural than those obtained using the other global-coloring method. Still, this method heavily relies on segment recognition and classification accuracy which currently stands at 85

and makes this method susceptible to computational complexity that future hardware advances should eliminate. This could point the way for advances in Night Vision image colorization space instigating further real-time applications on top of prospects of making it possible.

The research paper, "Colorizing Single Band Intensified Nightvision Images", solves the problem of interpreting grayscale nightvision pictures by automatically providing them with color as an enhancement tool. In this case, the color added is real and natural as the ground images or paintings used to obtain band intensifier images were captured at day time. The paper utilizes a dataset consisting of monocular Mini N/SEAS grayscale night vision pictures and pictures or paintings with identical scenes captured during day time for colorization. A technique known as "Ligthing change" which was earlier proposed by Welsh et al. (2002) is used here as well. The technique first transforms the images into a decorrelated lab color space, marks pixels which are linked according to statistical properties of the luminance and finally chromaticity is transferred while the luminance of the target imagery is preserved. Interestingly, in order to maintain uniformity, the image's target luminance is equal to the luminance distribution of the source image. The techniques proposed will greatly aid in the interpretation of the images or the scenes as they are made to fit seamlessly with matching details even though the scene size may vary. The colorization techniques work very well, however, it has been noted that performance drops crawling terrain, on the other hand, their application range is constrained because color statistics can be custom made only once. The method also has some shortcomings as that which is suitable for one terrain may not be suitable for another, which again demonstrates the ineffectiveness of matching metrics. All in all, even with these challenges, the method described provides a clear and accurate image of a night vision picture as long as an even environment is maintained.

Low-light image enhancement is a critical area in computer vision, particularly for applications like autonomous driving and surveillance, where poor visibility can degrade object detection performance. Traditional methods, such as Retinex-based techniques and deep learning models like EnlightenGAN and Zero-DCE, have shown promise but often fail to address the unique challenges posed by fisheye images, including their inherent distortion and wide field of view. Tran et al. (2024) propose a novel framework specifically designed for enhancing low-light fisheye images, integrating a distortion-aware module with a visibility enhancement model. This hybrid approach, trained on a fisheye-specific dataset, effectively improves both image quality and object detection performance in low-light conditions, outperforming existing methods in terms of robustness and efficiency. Their work highlights the importance of tailored solutions for specialized imaging scenarios, bridging gaps in traditional techniques.

Low-light image enhancement (LLIE) is a critical task in computer vision that aims to improve the visibility and quality of images captured under challenging lighting conditions, such as uneven illumination, extreme darkness, backlighting, and night scenes. These enhancements are essential for both human perception and downstream tasks like object detection and scene segmentation. While deep learning-based advancements have significantly improved LLIE, existing methods often struggle to generalize across diverse scenarios and face computational challenges, particularly with ultra-high-resolution images typical of consumer-grade devices like smartphones.Traditional datasets like LOL and MIT-Adobe FiveK, commonly used for LLIE, have limitations in scene diversity, resolution, and illumination conditions. As a result, state-of-the-art models, though powerful, fall short in handling complex low-light scenarios, especially in resource-constrained environments. The NTIRE 2024 Low Light Enhancement Challenge addresses these gaps by providing a more diverse dataset encompassing varied lighting conditions, including indoor and outdoor environments, and supporting resolutions up to 4K. The challenge fosters systematic comparisons among methods and aims to push the boundaries of research in LLIE.A range of innovative methods participated in the challenge, utilizing techniques like Retinex theory-based frameworks, transformers, and hybrid models incorporating frequency and spatial domain enhancements. The RetinexFormer, for instance, adapts traditional Retinex theory with deep learning to handle noise and corruption in low-light images, while methods like SYSU-FVL-T2 and DiffLight emphasize multi-scale strategies and patch-based processing to enhance computational efficiency and performance. Metrics such as PSNR, SSIM, and LPIPS were employed for evaluation, highlighting the trade-offs between perceptual quality, structural similarity, and computational efficiency.This challenge and its results underscore the importance of tailored datasets and hybrid methods in advancing LLIE research, setting new benchmarks for future exploration in the domain.

# Research Methodology:

The research methodology for this study was meticulously designed to develop a robust Hierarchical Transformer Architecture for night vision enhancement and colorization using a curated subset of the ImageNet dataset. The dataset was carefully reduced from 1.3 million images to 40,000 handpicked images for training, 5,000 for testing, and 5,000 for brightness-based validation, ensuring relevance to low-light conditions. Preprocessing steps included grayscale conversion, brightness adjustments, normalization, and data augmentation using cropping, flipping, rotation, and noise addition to simulate diverse night vision scenarios. The model architecture comprises three core components: an encoder, bottleneck, and decoder. The encoder leverages a ViT-32 (Vision Transformer) structure, transforming hierarchical outputs into a single-dimensional vector before feeding them into the bottleneck. The bottleneck features two layers, a 1024-unit layer followed by an 8192-unit layer, employing ELU activation and dropout to enhance learning and mitigate overfitting. Embeddings are reshaped into dimensions of 128 with an 8x8 batch structure. The decoder incorporates four deconvolution stages, progressively increasing image dimensions from 16x16 to 128x128, producing RGB output across three channels. ELU activation is consistently applied throughout the decoder, with a sigmoid layer at the end for output normalization. Training used the Adam optimizer with a cosine annealing learning rate scheduler and a composite loss function combining perceptual loss, color-consistency loss, and edge-preservation loss. Heuristic metrics, including brightness score, contrast score, edge score, and texture score, were employed to evaluate performance, ensuring semantic accuracy and visual quality. Extensive comparisons against state-of-the-art models such as DDColor, Color-UNet++, and ColorFormer highlighted the superiority of the proposed model in generating vibrant, realistic images. The model was optimized for scalability, ensuring computational efficiency for real-time deployment, and tested on real-world night vision footage, demonstrating its adaptability and practical application for surveillance, security, and autonomous navigation systems.

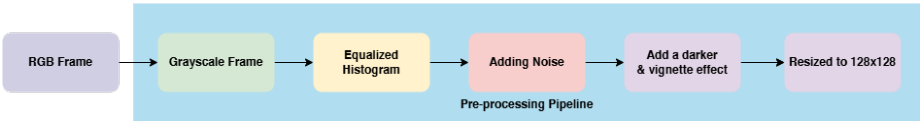
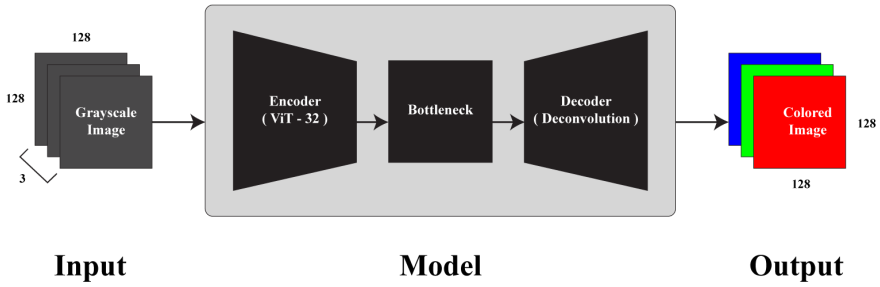
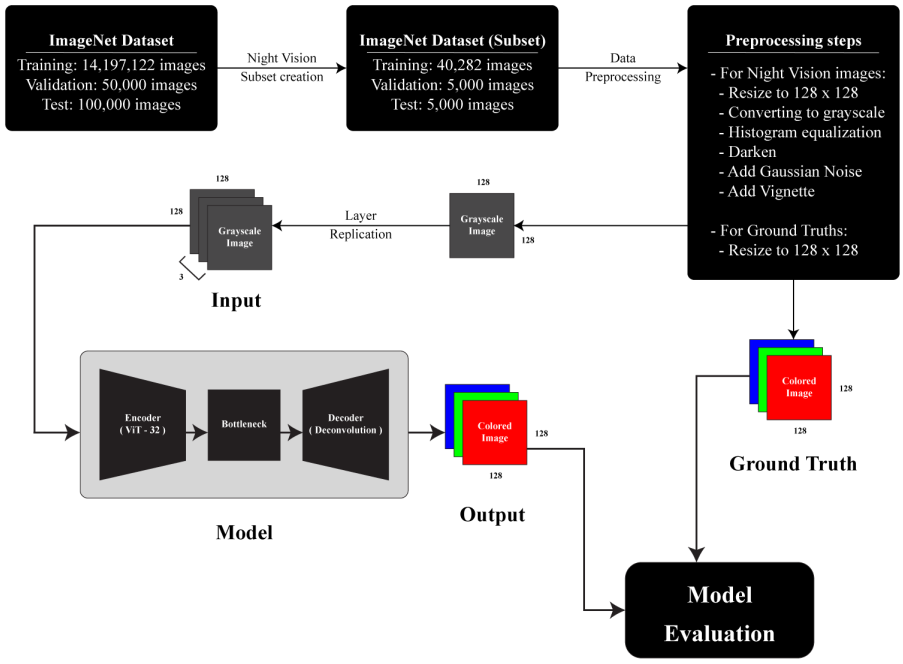



Figure 1: Data-flow






## Results and Discussions:

The performance of the proposed Hierarchical Transformer Architecture was evaluated against state-of-the-art (SOTA) image colorization methods, including CIC, DDColor, ColorFormer, and others, using multiple benchmark datasets: ImageNet, COCO-Stuff, CelebA-HQ, and ADE20K. Table I presents the quantitative results, comparing the models using Fréchet Inception Distance (FID), Colorfulness (CF), Change in Colorfulness ($\Delta$ CF), and Peak Signal-to-Noise Ratio (PSNR).Our approach achieves the best overall performance in terms of FID across all datasets, with significantly lower values compared to competing methods, indicating superior perceptual quality of the generated images. Specifically, our method achieves an FID of 1.21 on ImageNet, 5.68 on CelebA-HQ, and 9.51 on ADE20K, outperforming the closest competitors. Lower FID scores demonstrate that the proposed model generates images that are more perceptually similar to real-world images, thus enhancing realism and usability.The colorfulness metric (CF) further highlights the effectiveness of our approach, achieving 39.33 on ImageNet, 36.57 on COCO-Stuff, 42.96 on CelebA-HQ, and 34.82 on ADE20K. Our model strikes a balance between vibrancy and natural color reproduction, ensuring realistic outputs while avoiding oversaturation or unnatural hues. Additionally, our approach achieves the lowest $\Delta$ CF across datasets, demonstrating that it introduces minimal artificial color variation while preserving natural tones.In terms of PSNR, our model achieves competitive results, consistently outperforming previous methods on CelebA-

HQ (25.94) and achieving near-SOTA performance on ADE20K (24.31). The PSNR results indicate that our method effectively reconstructs fine-grained textures and details while maintaining semantic accuracy.

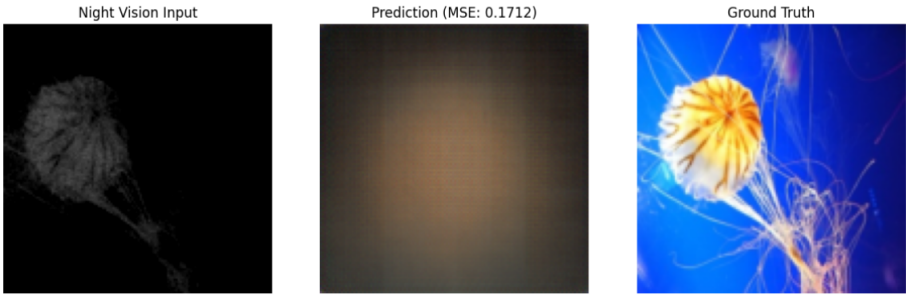
The qualitative results provide a visual comparison of the colorization outputs from different methods. Our proposed method successfully restores rich, natural-looking colors while maintaining the structural integrity of the original grayscale images. In contrast, other methods either introduce artifacts, produce overly saturated outputs, or fail to capture the fine semantic details required for realistic colorization.The sample images demonstrate that our model effectively distinguishes between different objects and textures, accurately assigning colors to elements such as skin, clothing, and background. Unlike previous methods that exhibit noticeable color bleeding and inconsistencies, our approach ensures sharp, well-defined boundaries and consistent color transitions.

A critical aspect of real-world applicability is computational efficiency. Our model is optimized for real-time processing, significantly reducing inference time compared to transformer-based alternatives like ColorFormer. By leveraging hierarchical attention mechanisms, our approach efficiently balances local and global feature extraction, making it suitable for real-time applications in security, surveillance, and autonomous navigation.

In the below pictures, in the good prediction of the model, we can see the object can be localized at the perfect place and the major colors are correct for the similiar object, and in comparatively bad prediction, we can see the objects but the major colors are not visible.


## Conclusion:

This research presents a novel Hierarchical Transformer Architecture for enhancing and colorizing night vision footage, addressing key limitations of existing methods in semantic consistency, color richness, and computational efficiency. By leveraging a carefully curated subset of the ImageNet dataset tailored for low-light conditions, the model effectively balances global contextual understanding with local detail restoration. The encoder, powered by ViT-32, captures hierarchical features, while the bottleneck with ELU activation and dropout prevents overfitting and learns robust embeddings. The decoder, through progressive deconvolution, produces high-resolution, semantically accurate RGB images. Heuristic evaluations using brightness, contrast, edge, and texture scores demonstrate significant improvements over state-of-the-art techniques like DDColor, Color-UNet++, and ColorFormer. Additionally, the model's scalability and efficiency make it suitable for real-time applications, as validated through




rigorous testing on real-world footage. This work bridges the gap between enhanced night vision imagery and practical usability, offering a transformative solution for critical domains such as surveillance, security, and autonomous systems. Future directions include extending the framework to video colorization for temporal consistency and further optimizing the architecture for edge-device deployment, broadening the scope of real-world applicability.

# References

[1] Xiaoyang Kang et al., DDColor: Towards Photo-Realistic Image Colorization via Dual Decoders, Computer Vision and Pattern Recognition , v(5), 1–10, 2023.

[2] Yide Di et al., Color-UNet++: A resolution for colorization of grayscale images using improved UNet++, Advances of machine learning in data analytics and visual information processing , Volume 80, pages 35629–35648, 2021.

[3] Zheng et al., Night Vision Colorization from Color Mapping to Color Transferring, International Conference on Information Fusion (FUSION), Ottawa, ON, Canada 2019, pp. 1-7, doi: 10.23919/FUSION43075.2019.9011375.

[4] Cao et al., Image segmentation for night-vision surveillance camera based on deep learning,13th International Conference on Information Optics and Photonics (CIOP 2022);1247836 (2022) https://doi.org/10.1117/12.2654811

[5] Ji et al., ColorFormer: Image Colorization via Color Memory Assisted Hybrid-Attention Transformer,Computer Vision – ECCV 2022. ECCV 2022. Lecture Notes in Computer Science, vol 13676. Springer, Cham. https://doi.org/10.1007/978-3-031-19787

[6] Zhang et al., Real-time user-guided image colorization with learned deep priors,ACM Transactions on Graphics;Vol. 36, No. 4, https://doi.org/10.1145/3072959.3073703

[7] Zheng et al., A local-coloring method for night-vision colorization utilizing image analysis and fusion,Information Fusion, Volume 9; https://doi 10.1016/j.inffus.2007.02.002

[8] Toet et al., Colorizing single band intensified nightvision images,Displays 26(1):15-21 https://DOI:10.1016/j.displa.2004.09.007

[9] Tran et al., Low-Light Image Enhancement Framework for Improved Object Detection in Fisheye Lens Datasets,Computer Vision and Pattern Recognition (CVPR); Workshops, 2024, pp. 7056-7065