Secure Systems Design and Technology Development. Veliko Tarnovo, Bulgaria
Proceedings of the International Scientific Conference, 2026. Volume III, 249-255

Comparative Analysis of Deep Learning Architectures for Monochromatic Image Colorization

Ihor Panasenko
Software engineering department
Kharkiv National University of Radioelectronics
Kharkiv, Ukraine
ihor.panasenko1@nure.ua

Kirill Smelyakov
Software engineering department
Kharkiv National University of Radioelectronics
Kharkiv, Ukraine
kyrylo.smelyakov@nure.ua

Anastasiya Chupryna
Software engineering department
Kharkiv National University of Radioelectronics
Kharkiv, Ukraine
anastasiya.chupryna@nure.ua

Paulius Sakalys
Faculty of Electronics and Informatics
Vilniaus kolegija
Higher Education Institution
Vilnius, Lithuania
p.sakalys@eif.viko.lt

Loreta Savulioniene
Faculty of Civil Engineering
Vilniaus kolegija
Higher Education Institution
Vilnius, Lithuania
l.savulioniene@viko.lt

Dainius Savulionis
Faculty of Electronics and Informatics
Vilniaus kolegija
Higher Education Institution
Vilnius, Lithuania
d.savulionis@eif.viko.lt

Abstract – colorizing monochromatic images is crucial for enhancing the informativeness of visual data in modern information technology and computer vision systems. However, automatic colorization poses an inherently ill-posed mathematical challenge due to the multimodal nature of color distributions, where multiple valid color mappings exist for any given grayscale input. This article conducts a comparative analysis of classical algorithms (such as the Welch and Levin methods) against advanced deep learning pipelines. The evaluated architectures range from baseline convolutional neural networks (CNNs) and standalone U-Net models to generative adversarial networks (GANs) and a novel hybrid Fusion GAN that integrates global semantic priors extracted via a ResNet-18 backbone. All models were trained and rigorously evaluated in the CIE Lab* color space, which effectively separates luminance from chrominance. To ensure a robust evaluation, experiments were conducted using a diverse benchmark of 2,000 images from the COCO dataset. Quality assessment combined traditional metrics like Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) with the Learned Perceptual Image Patch Similarity (LPIPS) metric, prioritizing human-like visual fidelity. Experimental results yielded counterintuitive insights. While the proposed Fusion GAN successfully surpassed classical methods, baseline CNNs, and standard GANs across most benchmarks, the standalone U-Net architecture secured the highest overall ranking. Specifically, U-Net achieved the top SSIM score of 0.945 (indicating superior structural preservation), the lowest LPIPS of 0.180 (best perceptual quality), and the fastest inference speed, enabling real-time applications.

Keywords – computer vision, image colorization, deep learning, GANs, U-Net, CNNs, LPIPS, SSIM, PSNR, COCO dataset.

I. Introduction

Restoring color to black-and-white images is much more than just a visual update or cosmetic restoration. In fields such as medicine, historical archive processing, restoration of old photographs, or analysis of satellite data, color plays a key role in understanding what exactly is depicted in the image [1]. The quality and accuracy of such processing directly affect the effectiveness of any subsequent analysis in computer vision systems [2]. However, from a mathematical standpoint, colorization remains a complex task: since the L* luminance channel does not contain information about hues, the same photo can be colored in an infinite number of ways.

Previously, attempts were made to solve this problem using user input. For example, the Welch algorithm transferred colors from a similar reference image based on texture [3], while the Levin method "blended" color between neighboring pixels with similar brightness [4]. Both of these approaches yielded consistent results, provided that the user's instructions were accurate, but at the same time required a lot of manual work, which made automatic processing of large archives impossible.

The transition to data-driven methods began with CNN-based approaches [5], followed by encoder–decoder

Online ISSN 3044-7771
https://doi.org/10.68302/std2026.vol3.217

architectures with skip connections [6], conditional GANs [7][8], and most recently transformer-based models [9]. Each generation reduced a specific failure mode but introduced new trade-offs between accuracy, perceptual realism, and inference speed.

Despite these successes, in practice, it is still difficult to find the ideal balance between image realism, accuracy, and system performance. This work proposes our own solutions to these challenges by developing Fusion GAN. We proposed a model that uses semantic cues from the ResNet-18 network [10] to better understand the overall context of the image.

II. Related Work

Classic algorithms. Semi-automatic colorization methods dominated early research. The Welch algorithm transferred hues from a reference color image by matching texture and luminance in the Lab color space [3], while the Levin method propagated user-applied color strokes to neighboring areas with similar brightness [4]. Both approaches were computationally light but required significant manual input, making fully automated archive processing infeasible.

The era of convolutional neural networks (CNNs). Zhang et al. reframed colorization as per-pixel classification over 313 quantized ab bins, avoiding the dominance of desaturated "safe" tones and substantially increasing output vibrancy [5]. Iizuka et al. extended this with a dual-stream architecture – one branch capturing global scene context, the other capturing local texture – a design principle that directly motivates the Fusion GAN model [11] explored in this work.

Encoder-decoder architectures and skip connections. U-Net [6] introduced long skip connections between encoding and decoding layers, allowing fine spatial detail to bypass heavy compression stages and significantly reducing color bleeding at object boundaries. Stable training of such deep architectures depends critically on proper weight initialization; biologically inspired stochastic initialization strategies accelerate convergence and prevent neuron saturation [12].

Conditional Generative Adversarial Networks (cGANs). The Pix2Pix framework [7][8] radically transformed colorization from a regression task into an adversarial image-translation process; its PatchGAN discriminator recovers vivid textures and color transitions typically lost by L1 regression. DeOldify further refined this approach with Self-Attention mechanisms and a NoGAN pre-training strategy, achieving strong results on historical photographs.

Transformers and diffusion models. DDColor [9] combines a pixel decoder with a query-based transformer color decoder, leveraging cross-attention to reduce inter-object color bleeding artifacts. Diffusion models treat colorization as iterative denoising and often surpass GANs in quality, but require 50–1,000 inference steps – prohibitive for interactive web applications where real-time visual processing is essential.

Evaluation metrics. PSNR and SSIM measure pixel-level and structural fidelity but correlate poorly with human perception. The LPIPS metric [13], calibrated against large-scale human similarity judgments via deep feature distances, is now the standard perceptual quality measure. This work applies all three metrics in combination to capture the trade-offs that no single score can reveal.

III. Methodology

A. Color Space and Problem Statement

All models under study operate in the CIE Lab* color space. This color space was chosen because it allows for a perfect separation of information regarding luminance and color. The L* luminance channel (normalized from 0 to 1) serves as the sole input to the neural network, and the model's task is to predict the two color channels: a* and b* (normalized from -1 to 1 using the Tanh function).

From a mathematical standpoint, we train the function (1), which must reconstruct realistic shades based solely on luminance data.

$$f: L \rightarrow (a, b) \quad (1)$$

The choice of the Lab color space is due to its "perceptual uniformity": equal mathematical distances within it roughly correspond to how the human eye perceives differences between colors. It is also important that the L* channel remains unchanged during processing, which guarantees the complete preservation of the structure and sharpness of the original image [5].

B. Dataset and Preprocessing

For our experiments, we used the validation set of the COCO 2017 dataset. From this, we formed a training subset of 2,000 images (random selection), which we split in a 90/10 ratio for training and validation, respectively. The choice of COCO is explained by the enormous variety of objects – over 80 categories. This allows us to test how well the architectures capture the scene context without being tied to a narrow topic (e.g., only faces or only landscapes).

Data preprocessing included resizing to 256×256 pixels and conversion to the Lab format (D65 white point standard). To improve model robustness during training, augmentation was applied: random cropping and horizontal mirroring of frames.

C. Architectures

Basic CNN. This model is built on the principle of a symmetric encoder-decoder, but without using feed-through connections. The process works as follows: four blocks compress the input data to a minimal size (bottleneck), after which four transposed convolution blocks attempt to reconstruct the color map ab.

We trained this network using the standard loss function MSE (L2). In our work, this model serves as a baseline: it clearly demonstrates the main drawback of simple regression – the desaturation or "sepia" effect. Since the model attempts to minimize the mean error when faced with ambiguity, it selects the "safest" gray color [5].

U-Net. The next stage of our research was the classic U-Net architecture [6]. It has eight downsampling stages and seven upsampling stages, allowing the input data to be compressed into a 1×1×512 bottleneck. The main feature of this model is its long skip connections: they transmit feature maps from each encoder layer directly to the corresponding decoder layers. This ensures that when assigning colors, the neural network clearly focuses on object boundaries regardless of the resolution level. The final layer of the system uses bilinear upsampling in combination with the Tanh function.

For training, we chose the L1 (MAE) loss function. The stability of such a deep network is ensured by a special strategy for stochastic weight initialization, inspired by biological neural systems [12]. This solution helps the model converge better during training and minimizes the risk of activations "getting stuck" in the deep layers of the encoder-decoder.

Conditional GAN (cGAN / Pix2Pix). In this model, we transition from simple regression to adversarial training [8]. The role of the Generator here is performed by the U-Net described above, while the Discriminator of the PatchGAN architecture [7] is responsible for evaluating the results. Its structure includes blocks of progressive resolution reduction and a final layer that outputs a score for individual image patches (fragments). The discriminator analyzes the pair (L, $ab_{predicted}$) and attempts to determine whether the color map is real or fake.

The total loss function of the Generator (2) is as follows:

$$L_G = L_{GAN}(G, D) + \lambda_{L1} \cdot L_{L1}(G) + \lambda_P \cdot L_{per}(G) \quad (2)$$

where $\lambda_{L1} = 100$ (according to [7]), and L_{per} is the perceptual loss based on feature matching from the VGG-16 network at layers {3, 8, 15, 22} [14]. To train the Discriminator, we used BCEWithLogitsLoss with one-sided label smoothing (the "true" value = 0.9). Both networks were optimized using the Adam algorithm with the following parameters: $lr = 2 \times 10^{-4}$, $\beta_1 = 0.5$, $\beta_2 = 0.999$.

Fusion GAN. This architecture is an extension of cGAN, where we added a mechanism for understanding global context. To do this, we integrated the base ResNet-18 network [10], adapting its first layer to work with single-channel brightness (by averaging the weights of pre-trained RGB filters). The U-Net encoder generates a feature vector, which is fused at the "bottleneck" with a 512-dimensional global hint from ResNet-18. The final fused vector is formed through a linear layer with a ReLU function (3):

$$z_{fused} = ReLU\left(W_f \cdot \left[z_{enc}; g_{global}\right]\right) \quad (3)$$

After that, the fused vector is passed to the decoder for the final image reconstruction. The ResNet-18 base network is trained jointly with the generator, while the discriminator and loss functions remain the same as in the cGAN stage. This principle of multi-branch data fusion has proven effective not only in colorization but also in many other critical computer vision tasks [11][15].

D. Evaluation Metrics

To objectively verify the results, we use three complementary metrics:

- PSNR (Peak Signal-to-Noise Ratio). This metric measures the technical accuracy of reconstruction at the pixel level. The higher the PSNR value (in decibels), the fewer digital distortions in the image.
- SSIM (Structural Similarity Index). This metric evaluates how well the model has preserved the brightness, contrast, and overall structure of objects. The value ranges from 0 to 1, where a higher score indicates a better match to the original.
- LPIPS (Learned Perceptual Image Patch Similarity) [13] is a modern measure of perceptual similarity, calibrated to reflect how a human evaluates an image. Unlike the previous metrics, a lower value is better here, as it indicates high naturalness and realism of colors.

We consider it critically important to analyze this data holistically. For example, a "conservative" model may achieve a high PSNR simply because it selects muted, almost gray colors (minimal pixel error), but its LPIPS score will be low due to a lack of realism. At the same time, generative adversarial networks (GANs) often produce vibrant colors, which improves LPIPS, but may suffer from the appearance of artifacts that lower SSIM.

IV. EXPERIMENTAL RESULT

A. Quantitative Comparison

The results that were received after running test colorizing on the set of images with different scenes that were not included in the learning dataset are illustrated in Table I.

TABLE I. QUANTITATIVE EVALUATION ON VALIDATION SUBSET

Architecture	PSNR (dB)	SSIM	LPIPS	Inference (ms)
Baseline CNN	~22.0	0.71	0.38	~22
U-Net	28.53	0.945	0.180	~118
cGAN	~24.5	0.887	0.165	~120
Fusion GAN	~24.1	0.881	0.168	~130

For the PSNR and SSIM metrics, a higher value indicates a better result, whereas for the LPIPS metric and inference time, a lower value is better. As can be seen from the data, the U-Net architecture delivers the best overall performance across all metrics simultaneously within a training budget of 20 epochs.

B. Baseline CNN

The typical errors of a basic convolutional network have a clear mathematical explanation: they are a direct consequence of using the L2 loss function. When faced with ambiguity in color selection, the network simply averages all possible options. As a result, instead of vivid hues, we get dull gray-brown patches – the so-called "sepia effect" [5]. Furthermore, since this model lacks skip connections, it loses information about fine details, causing colors to often bleed beyond object boundaries. This is confirmed by the numbers: a gap in the PSNR metric of over 6 dB compared to U-Net indicates both low accuracy at the level of individual pixels and serious overall structural distortions.

C. U-Net

The advantage of the U-Net architecture is based on two key mechanisms. First, skip connections transmit accurate information about object boundaries at all stages of processing, ensuring that the generated color perfectly matches the contours [6]. Second, the use of the L1 loss function helps partially mitigate the problem of washed-out colors characteristic of L2. As a result, we obtain structurally accurate images with high generation speed (approximately 118 ms per frame when using a CPU). This speed makes the model an excellent choice for interactive systems where response time is critical [2]. However, the tendency toward loss of saturation still remains, especially in large homogeneous areas (e.g., sky or water), where the network again tends toward averaged "safe" shades due to a lack of chromatic cues.

D. Conditional GAN

Unlike U-Net, the cGAN competitive approach forces the model to abandon "safe" averaging. By analyzing the image in separate fragments (patches), the network successfully restores bright and saturated textures where U-Net produces muted tones – for example, green grass, illuminated surfaces, or patterned fabrics [7]. But this approach has its drawbacks. Since the PatchGAN discriminator evaluates only small local areas, it often loses the overall context of the scene. A fragment on its own may look realistic, but noticeable color artifacts appear when viewed across the entire photo (for example, unnatural spots on skin or clothing).

It is precisely because of such errors that the SSIM structural similarity metric in cGAN is lower than in U-Net (0.887 vs. 0.945). However, the better result on the LPIPS metric (0.165 vs. 0.180 in U-Net) convincingly demonstrates: although GAN loses out in terms of mathematical pixel accuracy, from the perspective of human perception, it generates a significantly more vivid and natural image.

E. Fusion GAN

Using global hints from the ResNet-18 network in the Fusion GAN model [10] allowed the system to recognize the general content of an image much better. This is particularly noticeable in scenes with clear semantics – for example, in landscapes, images of food, or animals with distinctive coloring. In such cases, the results are globally consistent and appear more natural than those from the standard cGAN architecture.

In terms of numerical metrics, COCO Fusion GAN demonstrates results close to those of cGAN. This can be explained by the limited training budget (20 epochs) and small sample size (1,800 images), which proved insufficient to fully leverage all the benefits of the additional fusion parameters. However, the very principle of multi-branch feature fusion has already proven its high effectiveness in other complex tasks, such as traffic accident detection systems, where accurate visual perception is critical for safety [15]. This confirms that the architectural approach we have chosen is promising and goes far beyond the scope of the colorization task alone.

F. Qualitative Observations

One image was selected and colorized for a qualitative (visual) analysis. The selected image contains complex multi-subject scene, featuring people against a backdrop of architectural structures and greenery. Fig. 1 shows the original color image, which serves as a reference for comparison.

Fig. 1. Original color image

The numerical results of the colorization of this image are shown in Fig. 2

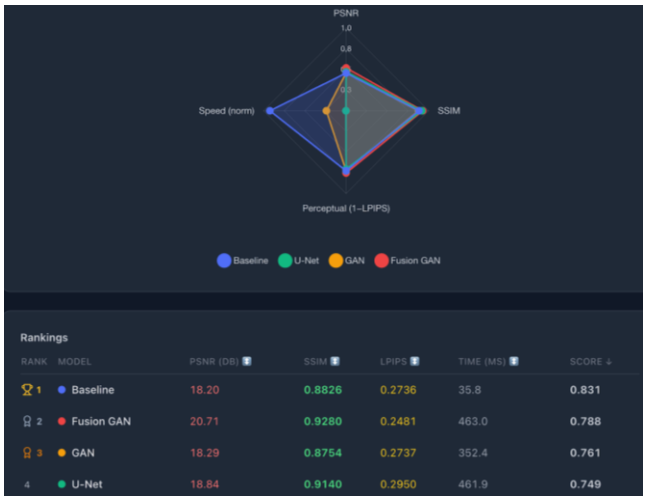

Fig. 2. Numerical results of the colorization of single image

Visual analysis revealed characteristic patterns in the performance of each architecture type.

Basic CNN demonstrates (see Fig. 3) the most uniform but washed-out results. In complex areas, a pronounced brown or gray tint often appears, making the image look unnatural.

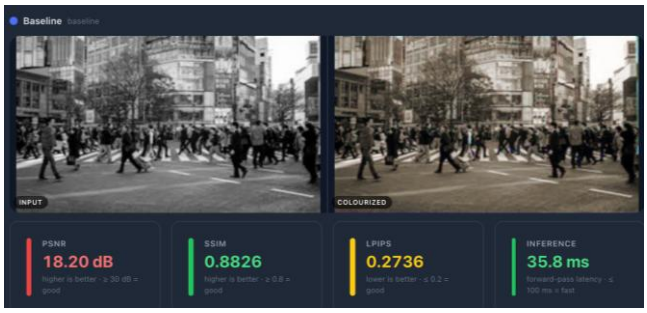

Fig. 3. Colorization result with Baseline CNN model

U-Net demonstrates (see Fig. 4) high structural accuracy. The model handled basic colors well – blue skies, green grass, and natural skin tones. However, on large monochromatic areas, it remains too conservative, avoiding saturated colors.

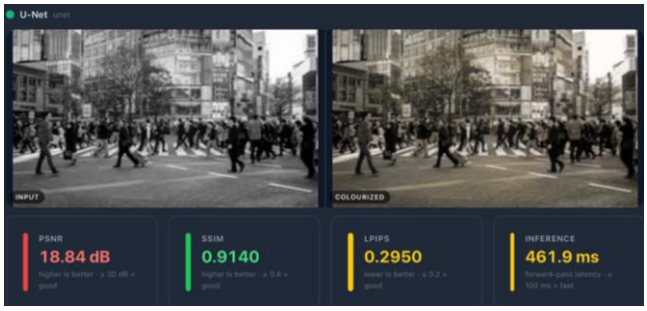

Fig. 4. Colorization result with U-Net model

cGAN provides (see Fig. 5) the highest color diversity at the local level. The network excellently reproduces vivid details – flowers, patterns on clothing, and the play of light. However, global artifacts sometimes appear in background areas where color may be distributed unevenly.

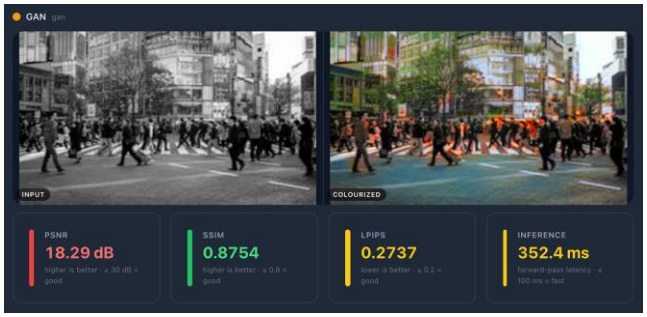

Fig. 5. Colorization result with cGAN model

Fusion GAN (see Fig. 6) appears to be the most balanced in scenes with a clear narrative. It retains the vibrancy of generative adversarial networks but adds logical coherence. At the same time, in very complex compositions with many objects, minor color inaccuracies characteristic of GAN models may still remain.

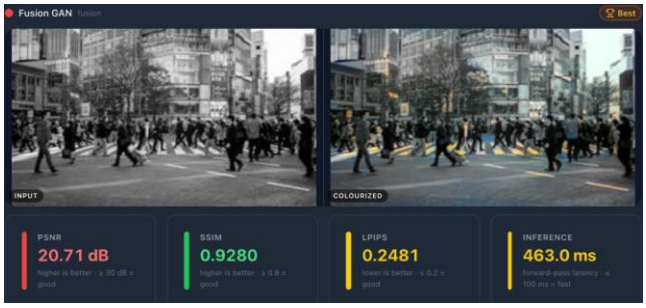

Fig. 6. Colorization result with Fusion GAN model

V. Conclusions

A. Why Does U-Net Lead in the Numbers?

The high performance of the U-Net architecture requires careful interpretation. Traditional metrics, such as PSNR and SSIM, penalize the model for any deviation from the original, even if the result looks natural to the human eye [13]. Since U-Net makes "conservative" predictions close to the dataset's mean values, it automatically incurs fewer penalties for distortion.

However, the fact that even the LPIPS metric favors this architecture indicates the great importance of structural accuracy. On diverse datasets (such as the COCO dataset), where there is no single dominant color theme, strict adherence to object boundaries proves to be a more significant factor for perception than high saturation. Our findings align with the results of other COCO-based studies: under limited training time, U-Net-style architectures consistently outperform simpler models across all metrics.

B. The Trade-off Between Mathematical Accuracy and Realism

We found a direct correlation between loss functions and the quality of the final image. Using L2 regression inevitably leads to desaturation – the image becomes dull and grayish [5]. The L1 function somewhat corrects this effect, but true color diversity is restored only by a combination of adversarial loss and VGG-16 perceptual features [14].

At the same time, this approach introduces an element of unpredictability, which degrades technical accuracy metrics at the level of individual pixels. This fundamental trade-off– between pixel-perfect alignment and visual realism –confirms that no single metric can provide the full picture. This is precisely why the combined PSNR/SSIM/LPIPS system is indispensable for comparative studies in the field of colorization.

C. Study Limitations

Our findings have three important limitations. First, a training budget of 20 epochs and a sample of 2,000 images are rather restrictive conditions for GAN models. Such architectures typically require significantly more iterations to fully realize their potential. Second, we did not involve real people in the evaluation (e.g., through "deception level" tests); no automatic algorithm is yet capable of 100% replacing the human eye, especially when it comes to the

results of competitive training [5]. Third, all experiments were conducted at a fixed resolution of 256×256. The performance of models on large images, where understanding the overall context of the scene becomes even more critical, requires separate investigation.

D. Directions for Further Research

Our results open up several promising avenues for the future development of this topic:

- Scaling up training. Increasing the training duration (beyond 100 epochs) on the full COCO 2017 dataset. This will allow GAN-based models to fully realize their qualitative potential, which is currently hindered by time constraints.
- Global analysis based on Transformers (ViT). If we replace ResNet-18 in our Fusion GAN model with the Vision Transformer architecture, the system will be able to understand the complex context of a scene much better. This will solve the "myopia" problem of the local PatchGAN, which has already been demonstrated using the DDColor model [9].
- Improved Initialization. Extending the biologically motivated weight initialization strategy [12] to all network components. This will make training even deeper architectures more stable and reduce the risk of getting stuck during training.
- Human evaluation. Introduction of "authenticity vs. forgery" (AMT) tests. Since no automatic formulas can fully replace human perception, this approach will help evaluate the realism of generated GAN images more objectively [13].
- Application of diffusion models. Using our global cue vector together with latent diffusion models (LDM). This could be the key to achieving truly photorealistic quality while maintaining acceptable processing speed.

E. General Conclusions

In this paper, we conducted a comprehensive and systematic comparison of four neural network architectures for colorizing monochrome images based on the COCO 2017 dataset. The results fully confirmed theoretical expectations. The baseline CNN clearly demonstrated that the use of L2 regression inevitably leads to color averaging and causes an undesirable "sepia effect" [5].

In contrast, U-Net [6] proved to be the most effective tool under conditions of limited training time. Thanks to its cross-connections and L1 loss function, this architecture successfully combines high structural accuracy with an acceptable level of realism and excellent speed (approximately 118 ms per image). The stability of such deep networks is further supported by proper biologically motivated weight initialization [12].

Generative adversarial networks (cGANs) [7] have demonstrated the ability to generate vivid and rich textures; however, due to the local focus of the PatchGAN discriminator, they often made semantic errors across the entire scene. Our Fusion GAN modification successfully mitigated this issue: integrating global cues from ResNet-18 [10] significantly improved contextual integrity in scenes with a clear narrative. It is worth noting that this multi-branch data fusion approach has also proven effective in other critical computer vision systems, such as road hazard detection [15].

The study also confirmed that a comprehensive system of PSNR/SSIM/LPIPS metrics [13] is indispensable for an objective evaluation of colorization. No single metric alone is capable of revealing all the trade-offs that models make during generation.

Ultimately, our work demonstrates that a relatively simple architecture with properly configured feed-through connections and loss functions remains an extremely powerful solution for practical tasks. At the same time, we have established a clear direction for future research that will help finally bridge the gap between mathematical accuracy and visual realism. And our specially developed web platform [2], which ensures reliable data input and processing, makes this process convenient and fully reproducible for any future experiments.

REFERENCES

[1] S. Anwar, M. Tahir, C. Li, A. Mian, F. S. Khan, and A. W. Muzaffar, "Deep Learning for Image Colorization: A Survey," arXiv preprint arXiv:2008.10774, 2024. [Online]. Available: https://arxiv.org/abs/2008.10774. DOI: https://doi.org/10.48550/arXiv.2008.10774

[2] S. Bielievtsov, I. Ruban, K. Smelyakov, and D. Sumtsov, "Network technology for transmission of visual information," in Selected Papers of the XVIII Int. Sci.-Practical Conf. "Information Technologies and Security" (ITS 2018), Kyiv, Ukraine, Nov. 27, 2018, CEUR Workshop Proceedings, vol. 2318, 2018, pp. 160–175. [Online]. Available: https://ceur-ws.org/Vol-2318/

[3] T. Welsh, M. Ashikhmin, and K. Mueller, "Transferring Color to Greyscale Images," ACM Trans. Graph. (Proc. SIGGRAPH), vol. 21, no. 3, pp. 277–280, 2002. DOI: https://doi.org/10.1145/566654.566576

[4] A. Levin, D. Lischinski, and Y. Weiss, "Colorization Using Optimization," ACM Trans. Graph. (Proc. SIGGRAPH), vol. 23, no. 3, pp. 689–694, 2004. DOI: https://doi.org/10.1145/1186562.1015780

[5] R. Zhang, P. Isola, and A. A. Efros, "Colorful Image Colorization," in Proc. European Conf. Computer Vision (ECCV), Amsterdam, Netherlands, 2016, pp. 649–666. DOI: https://doi.org/10.48550/arXiv.1603.08511

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 2015, pp. 234–241. DOI: https://doi.org/10.48550/arXiv.1505.04597

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 1125–1134. DOI: https://doi.org/10.48550/arXiv.1611.07004

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in Advances in Neural Information Processing Systems (NeurIPS), vol. 27, 2014. DOI: https://doi.org/10.48550/arXiv.1406.2661

[9] X. Kang, T. Yang, W. Ouyang, P. Ren, L. Li, and X. Xie, "DDColor: Towards Photo-Realistic Image Colorization via Dual Decoders," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), Paris, France, 2023, pp. 328–338. DOI: https://doi.org/10.48550/arXiv.2212.11613

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778. DOI: https://doi.org/10.1109/CVPR.2016.90

[11] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification," ACM Trans. Graph. (Proc. SIGGRAPH), vol. 35, no. 4, pp. 110:1–110:11, 2016. DOI: https://doi.org/10.1145/2897824.2925974

[12] O. Zolotukhin, Y. Bodyanskiy, V. Filatov, M. Kudryavtseva, and Y. Yeriemieiev, "Stochastic Initialization for Neural Networks Based on the Analysis of Biological Systems," in Proc. 5th Int. Workshop of IT-Professionals on Artificial Intelligence (ProfIT AI 2025), Liverpool, United Kingdom, Oct. 15–17, 2025, pp. 415–426. [Online]. Available: https://ceur-ws.org/Vol-4164/

[13] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018, pp. 586–595. DOI: https://doi.org/10.1109/CVPR.2018.00068

[14] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," in Proc. European Conf. Computer Vision (ECCV), Amsterdam, Netherlands, 2016, pp. 694–711, DOI: https://doi.org/10.1007/978-3-319-46475-6_43

[15] O. Byzkrovnyi, L. Savulioniene, K. Smelyakov, P. Sakalys, and A. Chupryna, "Comparison of Potential Road Accident Detection Algorithms for Modern Machine Vision System," Vide. Tehnologija. Resursi – Environment, Technology, Resources, vol. 3, pp. 50–55, 2023. DOI: https://doi.org/10.17770/etr2023vol3.7299