Journal of Imaging Science and Technology® 69(5): 1–14, 2025.

Automatic Image Colorization with Semantic Segmentation and Multipath Deep Networks

Jie-Sen Wang
Graduate Institute of Applied Science and Technology, National Taiwan University of Science and Technology, 43, Keelung Rd., Sec. 4, Taipei 106335, Taiwan

Hung-Chung Li
Bachelor Program in Intellectual Creativity Engineering, National Chung Hsing University, 145 Xingda Rd., South Dist., Taichung 40227, Taiwan

Pei-Li Sun
Graduate Institute of Color and Illumination Technology, National Taiwan University of Science and Technology, 43, Keelung Rd., Sec. 4, Taipei 106335, Taiwan
E-mail: plsun@mail.ntust.edu.tw

Abstract. *A fully automated colorization model that integrates image segmentation features to enhance both the accuracy and diversity of colorization is proposed. In the model, a multipath architecture is employed, with each path designed to address a specific objective in processing grayscale input images. The context path utilizes a pretrained ResNet50 model to identify object classes while the spatial path determines the locations of these objects. ResNet50 is a 50-layer deep convolutional neural network (CNN) that uses skip connections to address the challenges of training deep models. It is widely applied in image classification and feature extraction. The outputs from both paths are subsequently fused and fed into the colorization network to ensure precise representation of image structures and to prevent color spillover across object boundaries. The colorization network is designed to handle high-resolution inputs, enabling accurate colorization of small objects and enhancing overall color diversity. The proposed model demonstrates robust performance even when training with small datasets. Comparative evaluations with CNN-based and diffusion-based classification approaches show that the proposed model significantly improves colorization quality.*

Keywords: *colorization, multipath networks, convolutional neural network, semantic segmentation*

[DOI: 10.2352/J.ImagingSci.Technol.2025.69.5.050402]

1. INTRODUCTION

Since the advent of photography, the colorization of grayscale images has been a topic of considerable interest. This technology can provide additional semantic information, enhancing the readability and interpretability of image content while also improving visual effects. Traditional grayscale image colorization methods typically require users to manually provide color and image information for the process [1–3]. However, these approaches are labor-intensive and carry the risk of inaccuracies due to user-provided erroneous color information.

Received May 14, 2025; accepted for publication Oct. 7, 2025; published online Oct. 24, 2025. Associate Editor: Samuel Morillas.

1062-3701/2025/69(5)/050402/14/$25.00

Driven by the rapid development and technological breakthroughs of deep learning, automatic colorization of grayscale images has become an important research topic in recent years. Early convolutional neural network (CNN) architectures for colorization used simple and straightforward designs [4–6], primarily consisting of networks with increased depth achieved by stacking multiple convolutional layers. Although these architectures were well designed, they required large datasets for effective learning, limiting their practicality in scenarios with limited data availability. In subsequent studies, some approaches reformulated the colorization problem as a classification task by learning color distributions from large-scale natural image datasets [7]. Other methods employed pixel histogram modeling to capture multimodal color possibilities and avoid single-point estimation [8]. In addition, research combining local and global semantic features has effectively improved color consistency for objects such as buildings and the sky [9]. Exemplar-based methods, on the other hand, transfer colors from reference images to grayscale inputs, thereby enhancing the realism of specific objects in street scenes [10].

In CNN-based colorization methods, user inputs in the form of dots or doodles are often incorporated [11–14]. However, this approach is associated with an increased workload and requires a certain level of expertise from the user. Consequently, the process can be time-consuming. The employment of generative adversarial networks (GANs) or variational autoencoders is a common practice in achieving diverse colorization. The GANs utilize a competitive framework in which the generator endeavors to generate colors that are indistinguishable to the discriminator while the discriminator's objective is to differentiate between genuine and generated colors [15–19]. However, GAN-based methods often exhibit suboptimal performance when dealing with objects that have consistent colors. Additionally, they are characterized by high computational cost and significant resource consumption.

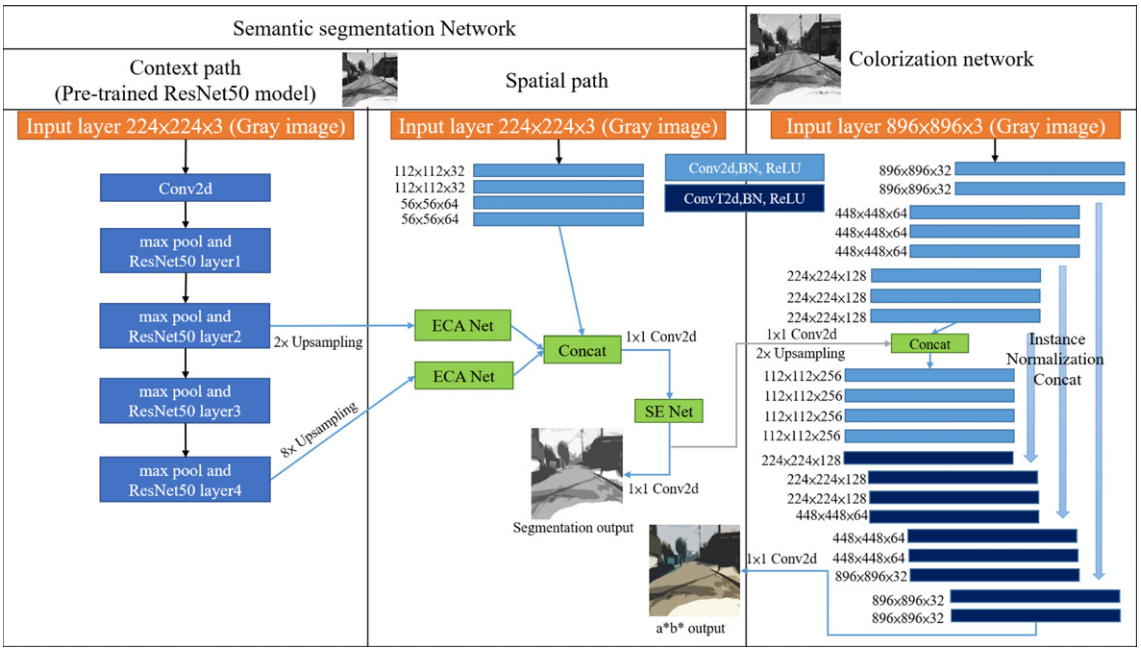

Figure 1. The architecture of the proposed colorization model.

Notwithstanding the elevated memory demands associated with the training process, multipath neural networks demonstrate a remarkable ability to accurately capture semantic information [9, 20–24]. By capitalizing on both local and global image features, these networks achieve enhanced colorization precision in grayscale images, thereby improving the overall quality of the colorization process. Transformer-based models [25–27] have recently garnered considerable attention due to their capacity to extract salient image features through multihead attention mechanisms. Another notable approach is the diffusion model [28, 29], which incorporates incremental noise during training to augment image diversity through denoising. However, both methods are highly data-dependent and necessitate substantial datasets for effective implementation.

With the advancement of generative models, diffusion models and multimodal conditional control methods have been introduced into the field of automatic colorization, further enhancing detail and texture. However, recent studies have also revealed several limitations. First, the one-to-many nature of mapping grayscale to color persists: models in open-domain street scenes tend to generate low-saturation or conservative colors, with limited ability to recover rare hues, as shown in Figure 1 [28, 30, 31]. Second, insufficient semantic understanding often leads to incorrect color predictions when the model encounters uncommon objects or distinctive signs, reducing the realism of the results [9, 28]. Furthermore, although exemplar-based methods can improve colorization accuracy, they are highly dependent on the structural similarity and alignment quality of the reference image. When significant differences exist, these methods can cause color shifts or unnatural transfers [10].

Another challenge lies in the trade-off between controllability and stability. Although conditions such as text descriptions, strokes, and reference images allow users greater control over colorization results, current methods still suffer from issues like color bleeding and unstable condition alignment [32, 33]. On the data side, most existing approaches rely on natural image datasets such as ImageNet and Places, which lack training sources specifically tailored for street scene colorization. As a result, their generalization ability in cross-domain applications, such as low-light environments and historical photographs, remains limited [31]. Although some studies have begun to adopt street scene datasets, such as Cityscapes [34] and Mapillary Vistas [35], these datasets were initially designed for semantic segmentation and autonomous driving, rather than being optimized for colorization tasks.

In terms of colorization evaluation, existing assessments still mainly rely on the peak signal-to-noise ratio and the structural similarity index. However, these pixel-level metrics cannot adequately reflect the realism and usability of street scene colorization [36–40], for instance, whether traffic light colors are correctly reproduced or how the results affect autonomous driving tasks. Therefore, future research should not only establish benchmark datasets specifically for street scene colorization but also introduce evaluation metrics grounded in human visual perception. Furthermore, integrating human subjective assessments with task-oriented performance measures will be essential for a more comprehensive evaluation of the practical value of these models in real-world applications.

In image colorization, a thorough understanding of semantic information is crucial to ensure the authenticity of the results. Such understanding enables sensible color assignments, for example, recognizing that cats are unlikely to be blue while leaves are typically green. In the realm of image segmentation, network architectures proposed in [41–45]

exhibit a close correlation with semantic and spatial location information despite adopting divergent design approaches. Based on these concepts, this study proposes a similar framework to support and enhance image colorization.

To meet the application needs of autonomous driving and intelligent transportation, the automatic colorization of grayscale street scene images has gradually attracted increasing attention. However, in the extant literature on grayscale image colorization, user-provided information or reference images are often required, with limited emphasis placed on road-specific colorization. This study introduces an automated method for colorizing grayscale road images. Image segmentation was incorporated into the design to address the challenge of road colorization due to the presence of various artificial objects. By capturing multiple local textures and objects and integrating this information into the colorization network, our model can effectively colorize elements such as buildings, trucks, and the sky without human intervention.

Although diffusion models have gained popularity in colorization, our CNN-based model exhibits superior learning capability, achieving improved performance on both small and large datasets. Our approach leverages the CIELAB color space to predict chromaticity components of an image. The proposed model comprises three key elements: a contextual network, a spatial network, and a colorization network. The contextual network learns semantic information about objects, the spatial network identifies their positions, and the colorization network integrates this information to generate the final colorized output. Experimental results demonstrate that our method outperforms the state-of-the-art diffusion model, achieving superior colorization accuracy and diversity.

The motivation and objective of this study are to enable the practical application of colorization in real-world scenarios while delivering high-quality results. However, due to the diversity of real-world colors, this task presents considerable challenges. Our research primarily employs the proposed model training architecture to first validate its effectiveness on a specific dataset, with subsequent work focusing on conducting a more comprehensive investigation and optimization of the model's generalization capability.

2. METHODS

This section presents a comprehensive account of the proposed CNN architecture illustrated in Fig. 1. The architecture comprises three principal components: the context pathway, the spatial pathway, and the colorization pathway. The context path furnishes data regarding the immediate content of the image, including the sky, building, and tree. In contrast, the spatial path provides information regarding the exact spatial position of these contextual elements within the image. The outputs of the context and spatial paths were integrated after important features were extracted by Efficient Channel Attention (ECA-Net) [46] and Squeeze and Excitation (SE-Net) [47], which enhanced channel-wise attention by adaptively weighting feature maps based on their importance, thereby facilitating effective semantic segmentation before being passed to the coloring path. A detailed explanation of ECA-Net and SE-Net is provided in Section 2.3. The incorporation of semantic information into the colorization network yielded three primary benefits: (1) improved precision in color prediction, (2) mitigation of color overflow problems, and (3) enhanced diversity in color applications.

In the early stages of this study, a single-path architecture was tested, but the recognition and segmentation performance on grayscale images was found to be suboptimal. This was likely due to the absence of color information, which typically provided important cues for semantic discrimination. To address this limitation, a dual-path architecture was adopted. The contextual path, based on an ImageNet pretrained model, enhanced semantic understanding and improved generalization across diverse scenes. The spatial path focused on preserving edge and structural details, thereby helping to prevent color bleeding during the colorization process.

In the proposed model, the process of downsampling (DS) was achieved through the implementation of convolution (Conv2d) [48], a layer that extracts spatial features by sliding filters over the input to generate feature maps. For upsampling (US), transposed convolution (ConvT2d) [49] was employed, which reverses the convolution operation to increase spatial resolution in a learnable manner. Following each convolutional layer, batch normalization (BN) [50] was applied to normalize the activations within each mini-batch, thereby accelerating training and improving model stability. To introduce nonlinearity, the rectified linear unit (ReLU) [51] was used.

The CNN pathways were connected and optimized via an end-to-end training process. The entire framework operated in a CIELAB-type color space, which consists of three channels: the lightness channel (L^*) and two chromatic channels (a^* and b^*). The input of the three pathways was a grayscale image. The input grayscale image was derived from an RGB-to-grayscale conversion based on the ITU-R BT.709 standard ($Y_{709} = 0.2126R + 0.7152G + 0.0722B$), which is closely aligned with the sRGB standard. The outputs from the contextual and spatial paths represented semantic segmentation results while the chromatic path produced the image planes for a^* and b^* color channels, separately. Since the primary objective of image colorization is to generate perceptually realistic chromatic channels from a grayscale image, this study adopted the method proposed by Iizuka [9], where the input grayscale image is treated as the lightness channel in the LAB color space transformation and the calculation of color differences. The final colorized images are represented in the sRGB color space. The conversion from CIELAB to sRGB was carried out in OpenCV-Python using the *cvtColor* function with the COLOR_Lab2RGB flag. This process involved two steps: (1) transforming CIELAB values into CIE 1931 XYZ values with the D65 illuminant as the reference white and (2) converting the XYZ values into sRGB in compliance with the IEC 61966-2-1:1999 standard [52].

To leverage the superior feature extraction capabilities of the pretrained ResNet50 [53] model, this study adopted the recommended input size of 224 × 224 × 3 in the semantic segmentation network. However, using the same resolution in the coloring network may result in the loss of details of small objects (such as brake lights and traffic signals). To address this issue, this study adopted a high-resolution input of 896 × 896 × 3 in the coloring network to preserve the details of small objects and improve coloring accuracy. This study assumed that the semantic information of large objects is sufficient to provide the overall contextual guidance required for coloring.

2.1 *Context Path*

An image is typically composed of both foreground and background elements, with the background often occupying a significant portion of the image area. In CNNs, extracting features at multiresolution enables the capture of both global and local information. This is typically achieved by adjusting the stride, which is a parameter that determines how the convolutional filter moves across the image, thereby affecting the resolution of the extracted feature maps. To this end, a pretrained ResNet50 model was employed to construct the contextual pathways. This approach can effectively capture image features even when the input image is in grayscale.

Specifically, image features were extracted at resolutions of 7 × 7 and 28 × 28 and then upsampled to 56 × 56, thereby ensuring a consistent resolution for subsequent processing. This multiscale approach allowed for the extraction of detailed image features while retaining the broader contextual information essential for subsequent processing as presented in Table I. Images typically consist of foreground and background elements, with the background usually occupying a large portion of the image area. In CNNs, feature extraction at multiple resolutions can capture both global and local information. This architectural configuration aimed to enhance the accuracy of image recognition. To this end, ECA-Net and SE-Net were incorporated to refine and selectively enhance relevant image features, thereby ensuring the optimal representation of both foreground and background information and ultimately improving model performance.

2.2 *Spatial Path*

Images commonly consist of multiple objects, whose spatial arrangements are crucial for accurate interpretation. As the depth of a neural network increases, its capacity to preserve absolute position information decreases. To address this issue, we have devised a wide and shallow architectural configuration comprising just four convolutional layers in the spatial path as detailed in Table II. The initial convolutional layer employed a kernel size of 5 × 5 and a stride of 2, enabling the capture of greater spatial detail at an early stage while minimizing positional loss. This architectural configuration is advantageous because it preserves absolute positional information, which is vital for tasks that necessitate precise localization. Furthermore, our design addresses color overflow issues, guaranteeing accurate color differentiation between adjacent objects or regions within the image.

Table I. Context path network architecture.

Output size	Operator	Stride	Filter
112 × 112	Conv2d	2	7 × 7 × 3 × 64
56 × 56	max pool	2	3 × 3
	Conv2d	1	ResNet50 layer1
28 × 28	max pool	2	3 × 3
	Conv2d	1	ResNet50 layer2
14 × 14	max pool	2	3 × 3
	Conv2d	1	ResNet50 layer3
7 × 7	max pool	2	3 × 3
	Conv2d	1	ResNet50 layer4

Table II. Spatial path network architecture.

Output size	Operator	Stride	Filter size
112 × 112	Conv2d, BN, ReLU	2	5 × 5 × 3 × 32
112 × 112	Conv2d, BN, ReLU	1	3 × 3 × 32 × 32
56 × 56	Conv2d, BN, ReLU	2	3 × 3 × 32 × 64
56 × 56	Conv2d, BN, ReLU	1	3 × 3 × 64 × 64

2.3 *Fusing Context and Spatial Features*

In the context of the color model, the efficient transmission of both spatial and contextual information is paramount. This was accomplished by integrating spatial and contextual data and applying an attention model to selectively extract the most relevant information. At the intermediate resolution level of 56 × 56, two branches were created: one directed towards the color neural network and the other dedicated to image segmentation. This dual-branch structure ensured that the position and content of each object within the image are accurately represented, enhancing both the precision of color application and the clarity of object boundaries.

The ECA-Net has been shown to improve the accuracy of classification results. The attention channel mechanism was employed in the branch path of the context path, and the number of channels in the two branches is equivalent, utilizing 1 × 1 convolution. Following the connection of the spatial path to the context path, a 1 × 1 convolution and SE-Net were employed to determine the significance of different channels and enhance salient features. The attention calculation is shown in Eq. (1). The ECA-Net and SE-Net can be calculated by Eqs. (2) and (3).

$$\tilde{X}_c = \alpha_c \cdot X_c, \quad \forall\, c \in \{1, 2, \ldots, C\}, \tag{1}$$

where X_c refers to the original input feature map of the cth channel and $\tilde{X}_c$ denotes the output feature map of the cth channel after being weighted by the attention coefficient α_c. The attention weight α_c is computed differently depending on the method.

The input features of ECA-Net and SE-Net are $X \in \mathbb{R}^{H\times W\times C}$, which aggregates spatial information into a channel descriptor $z \in R^C$ through global average pooling. The ECA-Net module introduces a lightweight and effective mechanism to capture channel-wise dependencies without dimensionality reduction.

$$\alpha = \sigma(Conv1D_k(z)), \quad z_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{i,j,c},$$
$$k = \left|\frac{\log_2 C}{\gamma} + \frac{b}{\gamma}\right|_{\text{odd}}, \tag{2}$$

where $Conv1D_k(\cdot)$ denotes a 1D convolution with a kernel size of k applied to the channel dimension, where γ and b are hyperparameters typically set at 2 and 1, respectively, and $|\cdot|_{\text{odd}}$ denotes rounding to the nearest odd integer. The parameter $X_{i,j,c}$ represents the value at the spatial position (i, j) in the cth channel of the input feature map $X \in \mathbb{R}^{H\times W\times C}$. Global average pooling is applied across the spatial dimensions. This formulation ensured that the kernel size scales reasonably with increasing channel numbers while preserving efficient local cross-channel interaction.

The SE-Net block enhanced channel-wise feature representations by modeling inter-channel dependencies. The SE block first applied a squeeze operation. This was followed by an excitation operation that captures channel-wise dependencies using two fully connected (FC) layers with a ReLU activation.

$$\alpha = \sigma(W_2 \cdot \delta(W_1 \cdot z)), \quad z_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{i,j,c}, \tag{3}$$

where $W_1 \in R^{(C/r)\times C}$ and $W_2 \in R^{C\times(C/r)}$ are the weight matrices of the FC layers, $\delta(\cdot)$ denotes the ReLU function, and $\sigma(\cdot)$ is the sigmoid [54] function, which maps input values to the range $(0, 1)$ to represent probabilities or attention weights. The parameter $X_{i,j,c}$ represents the value at the spatial position (i, j) in the cth channel of the input feature map $X \in \mathbb{R}^{H\times W\times C}$. Global average pooling is applied across the spatial dimensions. The reduction ratio r is a hyperparameter (typically set at 16) that controls the capacity and complexity of the excitation operation. The ECA-Net and SE-Net attention weights $\alpha \in \mathbb{R}^C$ were obtained after applying the sigmoid activation $\sigma(\cdot)$.

As both models are lightweight, they did not result in a substantial increase in the computational complexity of the model. Presently, the feature size is set at $56 \times 56 \times 256$. Finally, the output was upsampled and output with different 1×1 convolutional layers for image segmentation and integration with the colorization network. The segmentation was performed using the Softmax activation function [55], which converts raw output values into a normalized probability distribution across classes.

2.4 *Colorization Network*

Three design principles are particularly important in the coloring model's design: high-resolution image input, U-Net architecture with instance normalization (IN) [56], and optimal timing for incorporating image segmentation information.

If the same $224 \times 224 \times 3$ input size is employed as in the context and spatial paths, essential details of smaller objects, such as brake lights, will be at risk of being lost due to lower resolution. To address this issue, an input image size of 896×896 was employed within the colorization network, ensuring that even small objects retain sufficient detail for accurate color application.

The U-Net architecture was employed extensively across a range of domains. In CNNs, downsampling is typically accompanied by doubling the number of channels to compensate for the loss of information caused by the reduction in image resolution. However, this approach cannot fully preserve all feature details, often resulting in incomplete feature recovery during upsampling. Instance normalization, followed by concatenation with the upsampling layers, was adopted to address this issue. This strategy helped restore lost information and proved beneficial in high-contrast industrial scenes or when working with semantic masks containing uniform color regions.

In the colorization model, the input image was initially downsampled to a resolution of 112×112 while texture, edges, contrast, and related attributes were extracted at this stage. This information was inadequate for accurate colorization. To address this limitation, semantic features were incorporated, facilitating a more comprehensive reconstruction of colorization results during the upsampling process. The aforementioned three details are presented in Table III.

The colorization network was trained using the Huber loss function in the neural network, with the hyperparameter set at 0.5. However, incorporating semantic information from the images, such as the presence of specific objects like a bus or a building, is necessary for optimal coloring results. Concurrent training on the coloring and semantic segmentation networks was conducted to address this limitation. The training process utilized 11 categories of data, with segmentation labels provided for each category. These labels enabled the division of an image into multiple local regions, which was especially advantageous for accurate local image coloring. The 11 categories included building, bus, car, road, sidewalk, sky, traffic sign, tree, truck, vegetation, and wall.

The final colorized image was generated by combining semantic segmentation and training with the Huber loss function. However, the extent of colorization in an image was contingent upon psychophysical factors. To enhance the perceptual quality of the results, we incorporated perceptual loss into the training process. Unlike the traditional mean square error (MSE), perceptual loss is more aligned with subjective visual perception. In the proposed algorithm, the colorized CIELAB image was first converted to the sRGB color space using the procedure described earlier. Then, a perceptual loss was applied for further refinement, ensuring that the final output aligned closely with human visual expectations. The Huber loss L_{Huber}, cross-entropy loss L_{CE}, and perceptual loss $L_{\text{perceptual}}$ can be calculated by

Table III. Colorization network architecture.

Item	Output size	Operator	Stride	Filter size
DS1	896 × 896	Conv2d, BN, ReLU	1	3 × 3 × 3 × 32
DS2		Conv2d, BN, ReLU	1	3 × 3 × 32 × 32
DS3	448 × 448	Conv2d, BN, ReLU	2	3 × 3 × 32 × 64
DS4		Conv2d, BN, ReLU	1	3 × 3 × 64 × 64
DS5		Conv2d, BN, ReLU	1	3 × 3 × 64 × 64
DS6	224 × 224	Conv2d, BN, ReLU	2	3 × 3 × 64 × 128
DS7		Conv2d, BN, ReLU	1	3 × 3 × 128 × 128
DS8		Conv2d, BN, ReLU	1	3 × 3 × 128 × 128
Concatenate		Concatenate with Segmentation information		
DS9	112 × 112	Conv2d, BN, ReLU	2	3 × 3 × 256 × 256
DS10		Conv2d, BN, ReLU	1	3 × 3 × 256 × 256
DS11		Conv2d, BN, ReLU	1	3 × 3 × 256 × 256
DS12		Conv2d, BN, ReLU	1	3 × 3 × 256 × 256
US1	224 × 224	ConvT2d, BN, ReLU	2	3 × 3 × 256 × 128
Concatenate		Concatenate with (DS8 + IN)		
US2		ConvT2d, BN, ReLU	1	3 × 3 × 256 × 128
US3		ConvT2d, BN, ReLU	1	3 × 3 × 128 × 128
US4	448 × 448	ConvT2d, BN, ReLU	2	3 × 3 × 128 × 64
Concatenate		Concatenate with (DS5 + IN)		
US5		ConvT2d, BN, ReLU	1	3 × 3 × 128 × 64
US6		ConvT2d, BN, ReLU	1	3 × 3 × 64 × 64
US7	896 × 896	ConvT2d, BN, ReLU	2	3 × 3 × 64 × 32
Concatenate		Concatenate with (DS2 + IN)		
US8		ConvT2d, BN, ReLU	1	3 × 3 × 64 × 32
US9		ConvT2d, BN, ReLU	1	3 × 3 × 32 × 32
US10		Conv2d, BN, sigmoid	1	1 × 1 × 32 × 2

Eqs. (4)–(6). The total loss calculation is shown in Eq. (7).

$$L_{\text{Huber}(x,y)} = \begin{cases} \frac{1}{2}(x-y)^2 & \text{if } |x-y| < \delta \\ \delta \cdot \left(|x-y| - \frac{1}{2}\delta\right), & \text{otherwise,} \end{cases} \tag{4}$$

where x is the colorized image while y represents the ground truth; δ is an adjustable parameter, which is set at 0.5 in this study.

$$L_{\text{CE}} = -\sum_{\text{C}=1}^{\text{C}} y_c \cdot \log(\hat{y}_c), \tag{5}$$

where C is the number of categories, y_c denotes the one-hot encoded indicator for the true class, and $\hat{y}_c$ represents the model's predicted probability for class c.

$$L_{\text{perceptual}(x,y)} = \frac{1}{C_j H_j W_j} \|\varnothing_l(x) - \varnothing_l(y)\|_2^2, \tag{6}$$

where $\varnothing_l(x)$ and $\varnothing_l(y)$ represent the predicted image and the real image from a VGG16 pretrained network. Parameters C_j, H_j, and W_j represent the number of channels, height, and width of feature maps at the jth layer, respectively. The 16th layer was used in this study to ensure that the coloring results in a wide range would be more consistent with the psychophysics.

$$L_{\text{total}} = \lambda_1 \cdot L_{\text{Huber}(x,y)} + \lambda_2 \cdot L_{\text{CE}} + \lambda_3 \cdot L_{\text{perceptual}(x,y)}. \tag{7}$$

The loss weight ratio of λ_1, λ_2, and λ_3 was 10,000:8:15.

The analysis of previous image colorization results revealed that the chroma of generated images is often lower than that of the ground truth. To address this, the a^* and b^* channels were each scaled by a factor of 1.3, effectively increasing the chroma (C^*) of the output images. This factor was determined empirically, as larger values may result in unnatural colorization.

The architecture of our network employed the AdaDelta [57] optimizer, with a learning rate of 0.004 and a batch size of 4. To prevent overfitting, this study employed early stopping, where learning was halted once the test loss did not decrease after 16 consecutive repetitions. The model outputs the image with the lowest loss as the colorization result.

2.5 *GTA5 Dataset*

The GTA5 [43] dataset consists of 24,966 synthetic images with pixel-level semantic annotations, which were rendered using the open-world video game Grand Theft Auto 5. These images depict street scenes in an American virtual city from the perspective of a vehicle. The dataset under consideration encompasses 19 semantic categories. Following a rigorous evaluation of the dataset, 11 categories were selected for further analysis in this study. The selection was based on two criteria: the frequency of segmented objects and the necessity of colorization. The following categories are included: building, bus, car, road, sidewalk, sky, traffic sign, tree, truck, vegetation, and wall. The remaining unused categories are bicycle, person, fence, motorcycle, pole, rider, traffic light, and train. The image resolution of the dataset in question is 1914 × 1052. Prior to importing the model, it underwent a resizing process to align with the study design's dimensions. Subsequently, a comparison was made between the model and the original image, employing the same scale resolution.

The GTA5 dataset offers three notable advantages: (1) a diverse and extensive collection of artifacts, (2) highly complex and varied scenes, and (3) comprehensive label information for complete images. These characteristics make it an optimal choice for training models that require both detailed object recognition and contextual understanding. The training dataset consisted of 20,466 images while the testing dataset comprised 4500 images.

To enhance the model's robustness, a random horizontal flip of the input images was applied with 50% probability. This data augmentation technique helped to reduce overfitting and improve the model's generalization across diverse scenarios.

To validate the performance of the proposed neural network architecture, we conducted a series of experiments under various configurations. Six configurations were evaluated in this study. Method 1 employed the baseline architecture. Method 2 incorporated a pretrained model within

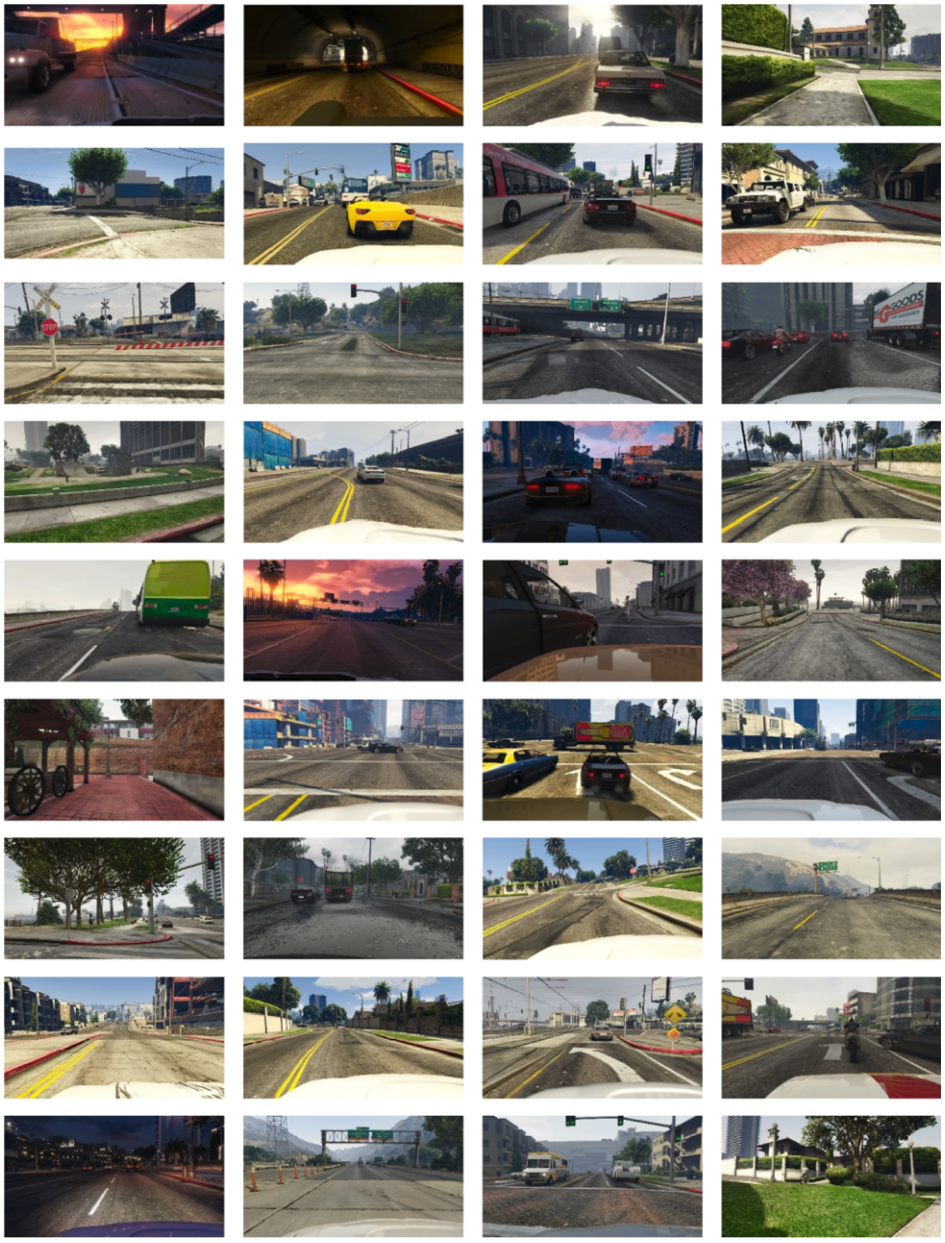

Figure 2. Colorization results on the GTA5 validation set using the proposed model (Method 6).

the context paths. Method 3 modified the input resolution of the colorization network to $896 \times 896 \times 3$. Method 4 added a perceptual loss to the training objective. Method 5 applied instance normalization before the concatenation operation in the colorization network. Method 6 increased the chroma component in the LAB color space by a factor of 1.3.

It is widely recognized that deep learning models generally require large-scale datasets to achieve optimal performance. Nevertheless, demonstrating robust performance under small-data conditions is also meaningful, as it highlights the model's ability to generalize in resource-constrained scenarios. Therefore, an additional experiment was conducted under a reduced protocol with 2000 training images and 500 validation images, following the settings adopted by Zabari and Iizuka. Under this protocol, the Zabari and Iizuka models and the proposed method (Method 6) were evaluated under identical conditions. The corresponding comparative results are presented and analyzed in Section 4.

The implementation was developed in Python 3.7.16 with PyTorch 1.12.1 (CUDA 11.3), cuDNN 8.3.2, and OpenCV 3.4.17. Training and inference were conducted on a Windows 10 system equipped with one NVIDIA GeForce RTX 2080 Ti (11 GB), an Intel Core i7-9700F CPU, and 32 GB RAM.

3. RESULTS

Figure 2 illustrates the colorization results on the GTA5 validation set, highlighting the model's ability to colorize small objects with diverse color distributions accurately. Leveraging image segmentation, the model achieved realistic

Table IV. Mean and 95th percentile color differences between the colorized images generated by Methods 1–6 and the ground-truth images (averaged over 4500 test images).

Methods	Basic architecture	Pretrained model	High-resolution input	Perceptual loss	IN	Chroma (C^*) scaled by 1.3	Global color difference (test data)	Traffic light color difference (test data)
Method 1	√						3.8/10.3	20.9/27.6
Method 2	√	√					3.5/8.4	21.2/27.7
Method 3	√	√	√				3.0/8.7	20.6/25.3
Method 4	√	√	√	√			2.8/8.0	0.66/5.4
Method 5	√	√	√	√	√		2.7/7.6	0.61/5.8
Method 6	√	√	√	√	√	√	2.7/7.3	0.52/5.0

(Metrics: mean $\Delta E_{\text{CAM16-UCS}}$/95th percentile of $\Delta E_{\text{CAM16-UCS}}$)

colorization for categories such as pedestrians and vehicles. The entire process was fully automated, requiring no human intervention.

Previous research on the perception of color differences in large printed images [58] demonstrated that statistical measures of extreme color deviations correlate more strongly with perceived image color differences than mean color differences do. The results of image colorization using the proposed methods, in terms of the mean and 95th percentile of CAM16-UCS color differences (denoted as $\Delta E_{\text{CAM16-UCS}}$) [59] between the colorized images and the corresponding ground-truth images, are presented in Table IV. The global color difference refers to the mean and the 95th percentile of $\Delta E_{\text{CAM16-UCS}}$ computed across all pixels in the image. In contrast, the traffic light color difference refers explicitly to the mean and the 95th percentile of $\Delta E_{\text{CAM16-UCS}}$ computed only over pixels corresponding to traffic signal lights. In Methods 1 and 2, a ResNet50 pretrained model was employed to enhance model diversity and to ensure optimal performance even with limited training data. The results are illustrated in Figure 3(a). In Methods 3 and 4, the ability to colorize small objects, such as red traffic signs with traffic horns, was enhanced by increasing the resolution and reducing the perceptual loss as illustrated in Fig. 3(b). Figure 3(c) demonstrates that the ability to colorize artificial objects can be enhanced by instance normalization and by increasing the chroma of the output images by a factor of 1.3 in Methods 4–6.

4. DISCUSSION

To further contextualize these findings, it is necessary to compare the method with representative prior studies. Zabari proposed a text-guided latent diffusion framework for image colorization, which integrates Cold Diffusion with a CLIP (Contrastive Language–Image Pretraining; Radford et al., 2021) [60]-based ranking mechanism to provide flexible and diverse results, albeit at a relatively high computational cost. In contrast, Iizuka designed a CNN-based architecture that performs colorization by fusing global scene priors with local features through image recognition. Their model benefits from implicit semantic guidance via scene classification, enabling natural colorization across a wide variety of images. As shown in Figure 4, the proposed model (Method 6) is compared with Zabari's diffusion model [29] and Iizuka's model [9] for image colorization. The model's performance was evaluated using 2000 training samples and 500 validation samples. Although implementations employed a greater number of training samples, this strategy is not consistently practical because the data collection process is often characterized by its labor-intensive and time-consuming nature, particularly in real-world applications. Notably, the ability to achieve competitive results with a reduced dataset underscores the efficiency of the proposed method and indicates its robust generalization capabilities while substantially reducing training resource demands.

The results indicate that elements such as trees and artifacts are not effectively colorized in the Zabari and Iizuka models. Although the sky exhibits some blue tones, the colorized regions remain imprecise. The proposed model has been demonstrated to exhibit superior performance in identifying both natural and artificial objects. The results of the $\Delta E_{\text{CAM16-UCS}}$ color difference comparison for various semantic categories are presented in Table V.

The results indicate that the proposed model achieves superior color restoration for categories such as sky and trees. This improvement can be attributed to the relatively consistent spatial positions, features, and color distributions of elements like the sky, trees, and vegetation.

In terms of artificial objects, the model employed in this study utilizes a multipath neural network, enabling precise identification of the necessity for distinct colors to denote varied semantic objects.

In the context of image segmentation, the proposed model's categorization of vehicles, including cars, trucks, and buses, enables the discernment of color variations across different categories. This capability is supported by the neural network's learning process, which integrates diverse annotated data. Analysis revealed that the color differences between buses and trucks were within acceptable limits. However, the outcome for cars was less optimal, likely due to the inability to distinguish between different cars solely through image segmentation when multiple cars were present in a single image. However, the Zabari and Iizuka models demonstrated even poorer performance, with the color of the

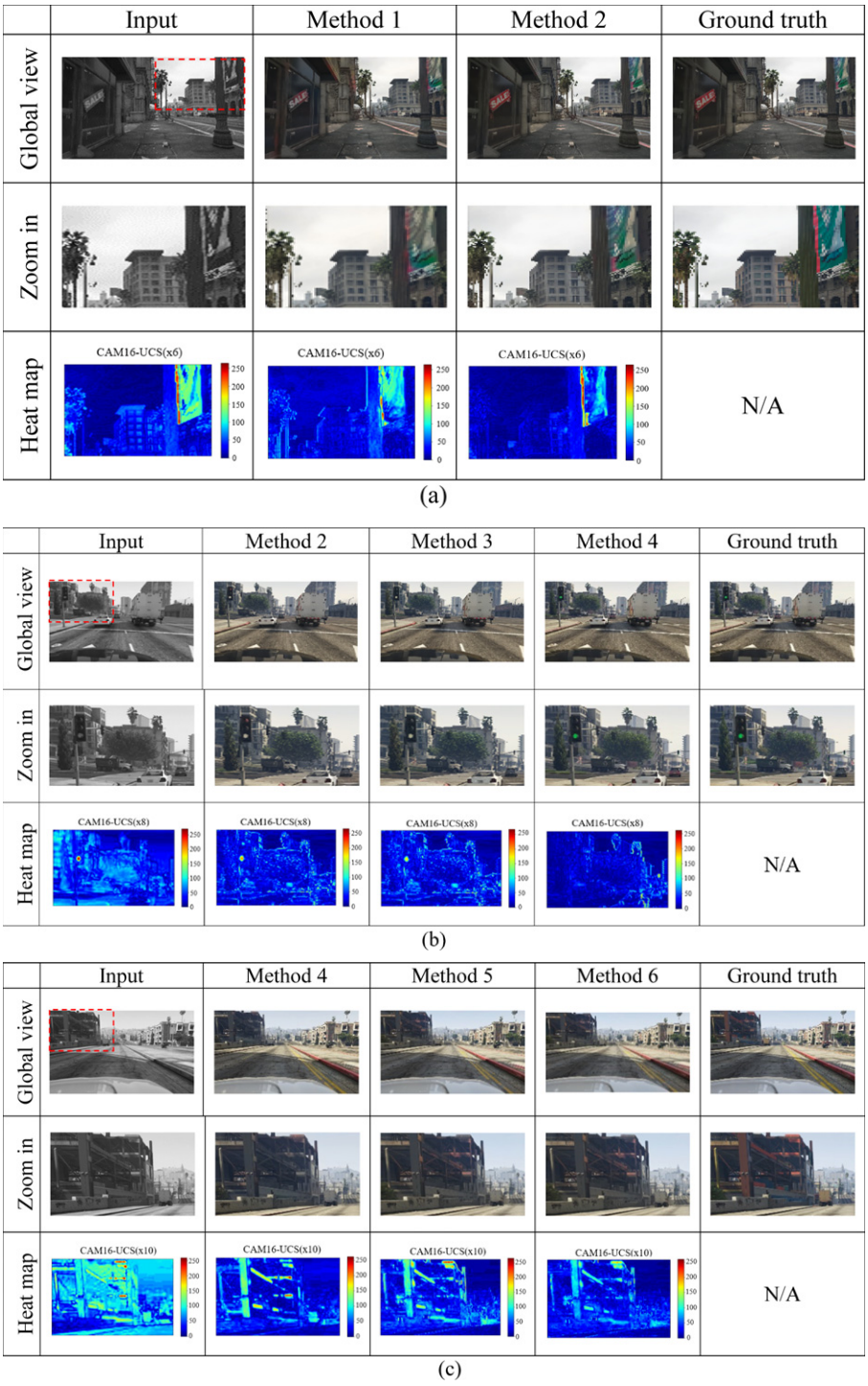

Figure 3. Colorization results of Methods 1–6.

cars often being different from the ground truth and being unable to distinguish the saturated red of the brake lights.

In the case of roads, walls, and sidewalks, these features in the image were very close to each other. Therefore, image segmentation is necessary to determine the location of the road and the presence of people. The proposed model outperformed the competitor's model, as it improved the visibility of roads, walls, and sidewalks.

In the coloring of small objects, such as traffic lights and brake lights, the proposed model significantly outperformed other models due to the use of a high-resolution U-Net architecture in the colorization path, which preserved the loss of image encoding by concatenating the encoder and decoder information.

Coloring buildings presents a unique challenge compared to other artifacts due to the wide range of styles and

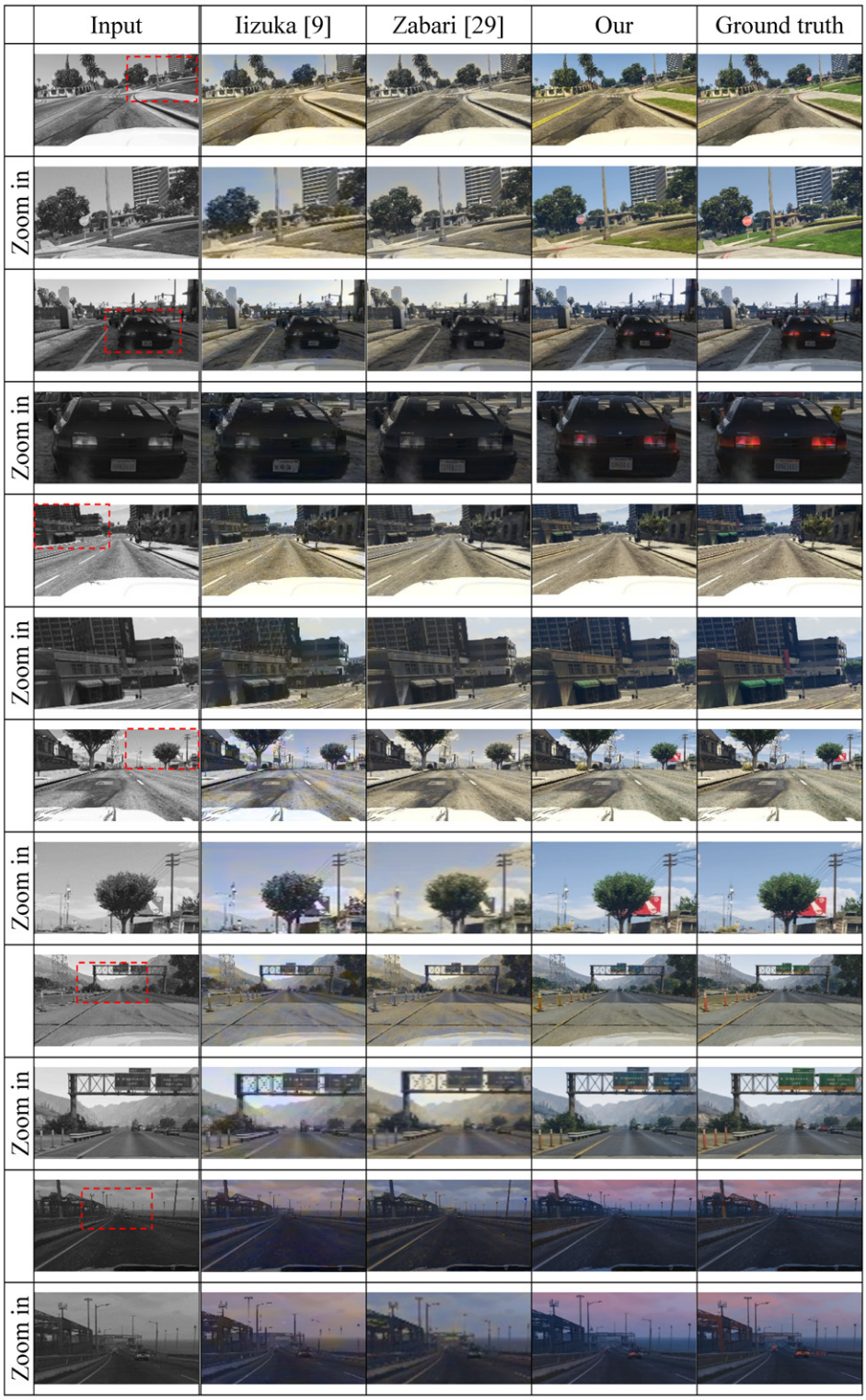

Figure 4. A comparison of Zabari's model, Iizuka's model, and the proposed model (Method 6) in image colorization.

colors found in this category. To address this challenge, we propose a model for segmenting images into various categories, thereby enhancing learning. This approach aimed to enhance the diversity of building colors and minimize the impact on other categories. The results of the proposed model demonstrated the efficacy of this method, surpassing the performance of different models.

The $\Delta E_{\text{CAM16-UCS}}$ calculation method in this study is as follows. First, convert both the colorized image and the ground-truth image from the sRGB space to the

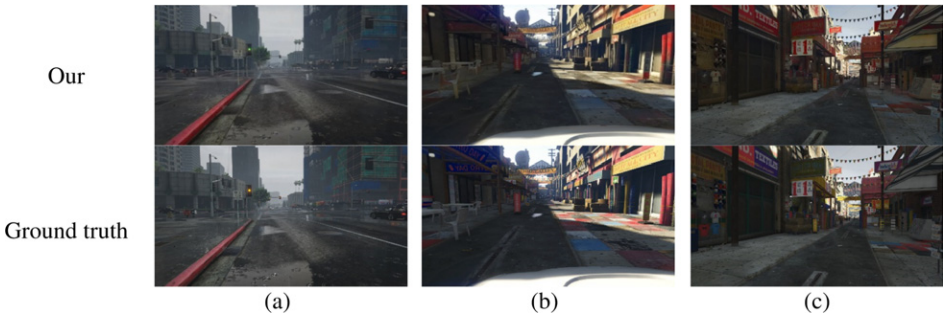

Figure 5. Examples of failed colorization using the proposed model (Method 6).

Table V. Comparison of mean color differences among various semantic categories.

	Building	Bus	Car	Road	Sidewalk	Sky	Traffic sign	Tree	Truck	Vegetation	Wall
Iizuka [9]	9.9	8.9	9.7	8.9	9.8	8.8	11.5	10.0	7.3	8.8	9.4
Zabari [29]	8.6	7.8	7.7	5.4	7.6	8.2	10.4	8.7	6.9	7.2	7.9
Ours	4.5	0.26	4.8	3.1	3.9	3.2	1.4	4.6	3.4	3.6	4.1

(Unit: $\Delta E_{\text{CAM16-UCS}}$)

CAM16-UCS J'a'b' space, using a D65 reference white and an adapting field luminance of 20 cd/m^2 under dim surround conditions. Then, compute the Euclidean distance between them in the J'a'b' space. Statistics are computed only for pixels belonging to the corresponding category in the segmentation mask (i.e., $\Delta E_{\text{CAM16-UCS}}$ is calculated for all pixels of each object), and no low-pass filtering is applied, primarily to preserve image detail. However, such pixel-wise statistics cannot account for perceptual phenomena such as visual masking or color assimilation. To further validate whether the model aligns with human subjective perception and to address these limitations, psychophysical experiments can serve as a valuable extension of this study.

The proposed method has two primary limitations. First, although the coloring accuracy of small objects can be improved by using a high-resolution U-Net structure, certain inaccuracies still persist. For instance, as shown in Figure 5(a), sign lights that are originally yellow may be incorrectly colored as green in the output. Second, referring to Figs. 5(b) and 5(c), in more complex scenes, the model sometimes applies unnatural colors, such as gray or red, which negatively impacts the overall realism of the image.

Although the proposed model performs well on the synthetic image dataset GTA5, it is still necessary to further validate its performance on real-world images. The BDD100K [61] dataset was released by the Berkeley Artificial Intelligence Research laboratory in collaboration with the Berkeley DeepDrive Industrial Consortium. It is one of the largest and most diverse publicly available datasets of driving videos. The dataset consists of 720p-resolution driving videos collected across multiple regions of the United States, including metropolitan areas such as New York City and the San Francisco Bay Area. The model (Method 6) proposed in this study achieves satisfactory results in terms of color filling performance on real-world scene datasets as shown in Figure 6. However, compared to the colorized outputs, the color differences from the ground-truth images are still relatively high. For memory colors (e.g., trees, roads, sky), the colorization results are satisfactory, likely because these categories exhibit relatively consistent canonical colors across scenes. An additional observation is that sky regions exhibit lower chroma, likely due to a dataset-induced bias in GTA5, where skies tend to appear less vividly blue. In the coloring of small objects, the red color of brake lights is consistently present, indicating that the high-resolution input of the proposed model effectively enhances the coloring results, particularly ensuring accuracy in coloring small objects on the road. Although the coloring of nonmemory colors is less accurate, the overall image still demonstrates noticeable color diversity. Additionally, based on the color filling performance of the proposed model, transfer learning can be applied to train the model on different application datasets, thereby improving color filling performance and reducing color differences. Furthermore, incorporating panoptic segmentation may also enhance the coloring of vehicles.

5. CONCLUSIONS

In this study, a novel automatic image colorization method was proposed, which integrates a multipath neural network with semantic segmentation to enhance the accuracy of color prediction. The experimental results on the GTA5 dataset demonstrated that the proposed method significantly improved color fidelity and object edge preservation compared to existing CNN and diffusion models. Using a high-resolution training dataset, the proposed method achieved small global color differences on 4500 test images relative to the ground truth, with mean $\Delta E_{\text{CAM16-UCS}} = 2.7$ and 95th percentile $\Delta E_{\text{CAM16-UCS}} = 7.3$. Moreover, even when trained on a small dataset, the method consistently outperformed the CNN and diffusion models across all categories. However, in highly complex scenes, this method did not produce ideal coloring results.

Figure 6. Colorization results for the BDD100K dataset using the proposed model (Method 6).

Future work will focus on improving image segmentation through two main directions. One direction is the application of panoptic segmentation to achieve more comprehensive scene understanding. Another direction is the incorporation of semantic guidance for objects with characteristic colors, such as taxis or airplanes of specific airlines. These strategies are expected to reduce color misclassification and further enhance color accuracy and diversity in complex scenes. In addition, existing color difference formulas are considered insufficient to fully capture human perception in complex images. To address this limitation, a preliminary experimental framework has been designed to recruit participants for subjective evaluations. Specifically, participants will compare the colorized results generated by the deep learning model with the corresponding ground-truth color images and provide naturalness ratings. This evaluation aims to assess the perceptual performance of colorization models from the perspective of human visual perception. This study can help us gain a deeper understanding of the key aspects of perceived color differences in image colorization.

Availability of Materials

Code and model availability: The source code and trained model can be provided upon reasonable request for research use. Requests can be directed to D10822501@mail.ntust.edu.tw or plsun@mail.ntust.edu.tw. Redistribution and commercial use are prohibited.

REFERENCES

[1] T. Chen, Y. Wang, V. Schillings, and C. Meinel, "Grayscale image matting and colorization," *Proc. Asian Conf. Comput. Vis.* (Springer, Jeju, Korea, 2004), pp. 1164–1169.

[2] L. Yatziv and G. Sapiro, "Fast image and video colorization using chrominance blending," IEEE Trans. Image Process. **15**, 1120 (2006).

[3] A. Y. S. Chia, S. Zhuo, R. K. Gupta, Y. W. Tai, S. Y. Cho, P. Tan, and S. Lin, "Semantic colorization with internet images," ACM Trans. Graph. **30**, 1 (2011).

[4] F. M. Carlucci, P. Russo, and B. Caputo, "$(DE)^2CO$: deep depth colorization," IEEE Robot. Autom. Lett. **3**, 2386 (2018).

[5] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," *Proc. IEEE Int'l. Conf. Comput. Vis.* (IEEE, Piscataway, NJ, 2015), pp. 415–423.

[6] Z. Hu, O. Shkurat, and M. Kasner, "Grayscale image colorization method based on U-net network," Int. J. Image, Graph. Signal Process. **16**, 70 (2024).

[7] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," *Proc. Eur. Conf. Comput. Vis.* (Springer, Amsterdam, Netherlands, 2016), pp. 649–666.

[8] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," *Proc. Eur. Conf. Comput. Vis.* (Springer, Amsterdam, Netherlands, 2016), pp. 577–593.

[9] 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. **35**, 1 (2016).

[10] M. He, D. Chen, J. Liao, P. V. Sander, and L. Yuan, "Deep exemplar-based colorization," ACM Trans. Graph. **37**, 1 (2018).

[11] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays, "Scribbler: controlling deep image synthesis with sketch and color," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (IEEE, Piscataway, NJ, 2017), pp. 5400–5409.

[12] R. Zhang, J. Y. Zhu, P. Isola, X. Geng, A. S. Lin, T. Yu, and A. A. Efros, "Real-time user-guided image colorization with learned deep priors," ACM Trans. Graph. **36**, 1 (2017).

[13] Y. Xiao, P. Zhou, Y. Zheng, and C. S. Leung, "Interactive deep colorization using simultaneous global and local inputs," *Proc. IEEE Int'l. Conf. Acoust., Speech, Signal Process* (IEEE, Piscataway, NJ, 2019), pp. 1887–1891.

[14] Y. Ci, X. Ma, Z. Wang, H. Li, and Z. Luo, "User-guided deep anime line art colorization with conditional adversarial networks," *Proc. ACM Int'l. Conf. Multimedia* (ACM, Seoul, Korea, 2018), pp. 1536–1544.

[15] A. Deshpande, J. Lu, M. C. Yeh, M. J. Chong, and D. Forsyth, "Learning diverse image colorization," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (IEEE, Piscataway, NJ, 2017), pp. 6837–6845.

[16] K. Frans, "Outline colorization through tandem adversarial networks," Preprint, arXiv:1704.08834 (2017).

[17] K. Nazeri, E. Ng, and M. Ebrahimi, "Image colorization using generative adversarial networks," *Proc. Int'l. Conf. Articulated Motion and Deformable Objects* (Springer, Palma de Mallorca, Spain, 2018), pp. 85–94.

[18] P. Vitoria, L. Raad, and C. Ballester, "ChromaGAN: adversarial picture colorization with semantic class distribution," *Proc. IEEE Winter Conf. Appl. Comput. Vis.* (IEEE, Piscataway, NJ, 2020), pp. 2445–2454.

[19] B. Li, Y. Lu, W. Pang, and H. Xu, "Image colorization using CycleGAN with semantic and spatial rationality," Multimed. Tools Appl. **82**, 1 (2023).

[20] S. Guadarrama, R. Dahl, D. Bieber, M. Norouzi, J. Shlens, and K. Murphy, "PixColor: pixel recursive colorization," *Proc. Brit. Mach. Vis. Conf.* (BMVA, London, UK, 2017), pp. 1–12.

[21] J. Zhao, J. Han, L. Shao, and C. G. Snoek, "Pixelated semantic colorization," Int. J. Comput. Vis. **128**, 818 (2020).

[22] J. Zhao, L. Liu, C. Snoek, J. Han, and L. Shao, "Pixel-level semantics guided image colorization," *Proc. Brit. Mach. Vis. Conf.* (BMVA, Newcastle, UK, 2018), p. 156.

[23] M. He, D. Chen, J. Liao, P. V. Sander, and L. Yuan, "Deep exemplar-based colorization," ACM Trans. Graph. **37**, 47 (2018).

[24] J. W. Su, H. K. Chu, and J. B. Huang, "Instance-aware image colorization," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (IEEE, Piscataway, NJ, 2020), pp. 7968–7977.

[25] M. Kumar, D. Weissenborn, and N. Kalchbrenner, "Colorization transformer," *Proc. Int'l. Conf. Learn. Represent.* (ICLR, Virtual, Vienna, Austria, 2021).

[26] S. Weng, J. Sun, Y. Li, S. Li, and B. Shi, "CT^2: colorization transformer via color tokens," *Proc. Eur. Conf. Comput. Vis.* (Springer, Tel Aviv, Israel, 2022), pp. 1–10.

[27] H. Shafiq and B. Lee, "Transforming color: a novel image colorization method," Electronics **13**, 2511 (2024).

[28] H. Wang, X. Chai, Y. Wang, Y. Zhang, R. Xie, and L. Song, "Multimodal semantic-aware automatic colorization with diffusion prior," *Proc. IEEE Int'l. Conf. Multimedia Expo Workshops* (IEEE, Piscataway, NJ, 2024), pp. 1–6.

[29] N. Zabari, A. Azulay, A. Gorkor, T. Halperin, and O. Fried, "Diffusing colors: image colorization with text guided diffusion," *Proc. SIGGRAPH Asia Conf. Papers* (ACM, Sydney, Australia, 2023), pp. 1–11.

[30] J. W. Su, H. K. Chu, and J. B. Huang, "Instance-aware image colorization," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (IEEE, Piscataway, NJ, 2020), pp. 7968–7977.

[31] S. Anwar, M. Tahir, C. Li, A. Mian, F. S. Khan, and A. W. Muzaffar, "Image colorization: a survey and dataset," Inf. Fusion **114**, 102720 (2025).

[32] Z. Liang, Z. Li, S. Zhou, C. Li, and C. C. Loy, "Control color: multimodal diffusion-based interactive image colorization," Int. J. Comput. Vis. (2025).

[33] R. García, G. Randall, and L. Raad, "A short analysis of BigColor for image colorization," Image Process. Online **14**, 144 (2024).

[34] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, . . . , and B. Schiele, "The Cityscapes dataset for semantic urban scene understanding," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (IEEE, Piscataway, NJ, 2016), pp. 3213–3223.

[35] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder, "The Mapillary Vistas dataset for semantic understanding of street scenes," *Proc. IEEE Int'l. Conf. Comput. Vis.* (IEEE, Piscataway, NJ, 2017), pp. 4990–4999.

[36] K. Bajbaa, M. Usman, S. Anwar, I. Radwan, and A. Bais, "Bird's-eye view to street-view: a survey," Preprint, arXiv:2405.08961 (2024).

[37] Y. Li, S. Yang, and J. Liu, "Language-based image colorization: a benchmark and beyond," Preprint, arXiv:2503.14974 (2025).

[38] C. Ma, Z. Shi, Z. Lu, S. Xie, F. Chao, and Y. Sui, "A survey on image quality assessment: insights, analysis, and future outlook," Preprint, arXiv:2502.08540 (2025).

[39] M. Xu, "Image colorization based on transformer," Sci. Rep. **15**, 21311 (2025).

[40] Z. Xu and C. Geng, "Color restoration of mural images based on a reversible neural network," Heritage Sci. **12**, 351 (2024).

[41] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "BiSeNet V2: bilateral network with guided aggregation for real-time semantic segmentation," Int. J. Comput. Vis. **129**, 3051 (2021).

[42] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (IEEE, Piscataway, NJ, 2017), pp. 2881–2890.

[43] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: multi-path refinement networks for high-resolution semantic segmentation," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (IEEE, Piscataway, NJ, 2017), pp. 1925–1934.

[44] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *Proc. Eur. Conf. Comput. Vis.* (Springer, Munich, Germany, 2018), pp. 801–818.

[45] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," *Proc. Med. Image Comput. Comput.-Assist. Interv. (MICCAI)* (Springer, Munich, Germany, 2015), pp. 234–241.

[46] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: efficient channel attention for deep convolutional neural networks," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (IEEE, Piscataway, NJ, 2020), pp. 11534–11542.

[47] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (IEEE, Piscataway, NJ, 2018), pp. 7132–7141.

[48] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE **86**, 2278 (1998).

[49] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high-level feature learning," *Proc. IEEE Int'l. Conf. Comput. Vis.* (IEEE, Piscataway, NJ, 2011), pp. 2018–2025.

[50] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," *Proc. Int'l. Conf. Mach. Learn.* (PMLR, Lille, France, 2015), pp. 448–456.
[51] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," *Proc. Int'l. Conf. Artif. Intell. Stat.* (JMLR, Fort Lauderdale, FL, USA, 2011), pp. 770–778.
[52] International Electrotechnical Commission, "IEC 61966-2-1: Multimedia systems and equipment - Colour measurement and management - Part 2-1: Colour management-Default RGB colour space-sRGB," Standard IEC 61966-2-1, International Electrotechnical Commission, (1999).
[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (IEEE, Piscataway, NJ, 2016), pp. 770–778.
[54] J. Han and C. Moraga, "The influence of the sigmoid function parameters on the speed of backpropagation learning," *Proc. Int'l. Workshop Artif. Neural Netw.* (Springer, Sitges, Spain, 1995), pp. 195–201.
[55] J. S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," Neurocomputing **68**, 227 (1990).
[56] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: the missing ingredient for fast stylization," Preprint, arXiv:1607.08022 (2016).
[57] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," Preprint, arXiv:1212.5701 (2012).
[58] J. Uroz, R. Luo, and J. Morovic, "Perception of colour differences in large printed images," in *Colour Image Science: Exploiting Digital Media*, edited by L. MacDonald and M. R. Luo (John Wiley & Sons, Chichester, UK, 2002), pp. 49–73.
[59] C. J. Li, Z. Q. Li, Z. F. Wang, Y. Xu, M. R. Luo, G. H. Cui, M. Melgosa, H. Brill, and M. Pointer, "Comprehensive color solutions: CAM16, CAT16, and CAM16-UCS," Color Res. Appl. **42**, 703 (2017).
[60] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *Proc. Int'l. Conf. Mach. Learn.* (PMLR, Vienna, Austria, 2021), pp. 8748–8763.
[61] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Mu, V. Koltun, and T. Darrell, "BDD100K: a diverse driving video database with scalable annotation tooling," Preprint, arXiv:1805.04687 (2018).