End-to-End Image Colorization with Multiscale Pyramid Transformer

Tongtong Zhao[1*], Gehui Li[2*], Shanshan Zhao[3*]
[1] Presight AI, Abu Dhabi, UAE
[2] Dalian University of Technology, Dalian, China
[3] Inception Institute of Artificial Intelligence, Abu Dhabi, UAE
tongtong.zhao@presight.ai, gehuili90@gmail.com, shanshan.zhao@g42.ai

***Abstract*—Image colorization is a challenging task due to its ill-posed and multimodal nature, leading to unsatisfactory results in traditional approaches that rely on reference images or user guides. Although deep learning-based methods have been proposed, they may not be sufficient due to the lack of semantic understanding. To overcome this limitation, we present an innovative end-to-end automatic colorization method that does not require any color reference images and achieves superior quantitative and qualitative results compared to state-of-the-art methods. Our approach incorporates a Multiscale Pyramid Transformer that captures both local and global contextual information and a novel attention module called Dual-Attention, which replaces the traditional Window Attention and Channel Attention with faster and lighter Separable Dilated Attention and Factorized Channel Attention. Additionally, we introduce a new color decoder called Color-Attention, which learns colorization patterns from grayscale images and color images of the current training set, resulting in improved generalizability and eliminating the need for constructing color priors. Experimental results demonstrate the effectiveness of our approach in various benchmark datasets, including high-level computer vision tasks such as classification, segmentation, and detection. Our method offers robustness, generalization ability, and improved colorization quality, making it a valuable contribution to the field of image colorization.**

***Index Terms*—Colorization, Classification, Detection, Segmentation**

I. Introduction

In the realm of computer vision, the pursuit of image colorization has emerged as a vibrant field of inquiry, boasting a plethora of practical applications. However, the challenge of this task lies in its ill-posed and multimodal nature, leading to incorrect or unsaturated results in traditional approaches that rely on reference images or user guides. With the advent of deep learning, researchers have been exploring automatic colorization using convolutional neural networks (CNNs). However, these methods may not be sufficient due to the lack of semantic understanding. To address this issue, some methods incorporate generative adversarial networks (GANs) as generative priors. Nonetheless, these methods have limited representation space, leading to inappropriate coloring results or artifacts. More recently, transformer-based methods have been proposed, but they require additional training of multiple modules or rely on dataset-level empirical distributions, which are not adaptable to different contexts.

*These authors contributed equally to this work.

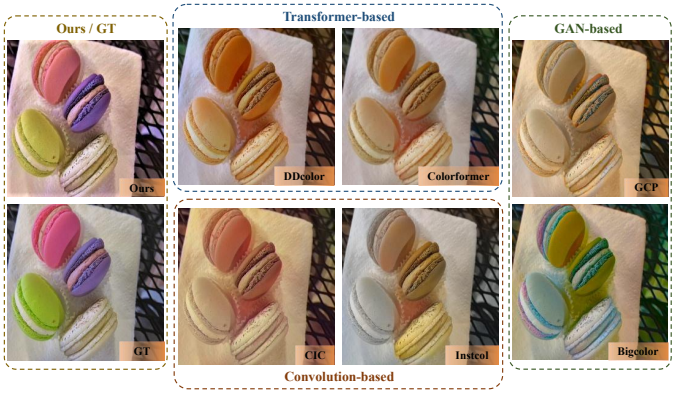


Fig. 1: **Visual comparison with SOTA methods.** A method is proposed to extract features using a transformer encoder and perform end-to-end coloring with learnable color queries. It is able to generate more natural and vivid coloring results compared to other methods.

Within this landscape, we unveil an cutting-edge approach to end-to-end automatic colorization that outperforms prior models, negating the need for color reference imagery and delivering both quantitatively and qualitatively unparalleled results compared to the contemporary methods. We introduce the Multiscale Pyramid Transformer, a design that adeptly harnesses both the micro and macro facets of contextual information, yielding colorizations that are not only precise but resoundingly lifelike. Central to our model is a faster and lighter attention mechanism, named Dual-Attention, which supplants the conventional Window and Channel Attention with an agile and powerful pair: the Separable Dilated Attention and the Factorized Channel Attention. The former scales down the attentional complexity from quadratic to linear without sacrificing spatial acuity, whereas the latter accomplishes a similar reduction in complexity, all without compromising channel depth.

Moreover, we present the Color-Attention decoder, a novel entity adept at deciphering colorization motifs from a juxtaposition of grayscale and color images within the current training ensemble, thereby amplifying its adaptability across different datasets and obviating the construction of color priors. Our empirical analysis, spanning a variety of benchmark datasets and encompassing paramount computer vision undertakings

such as classification, segmentation, and detection, attest to the efficacy of our methodology. Our contributions, therefore, lie in the introduction of a Multiscale Pyramid Transformer that seamlessly captures nuanced contextual clues, a Dual-Attention mechanism that redefines efficiency in computational processing, and the Color-Attention decoder, a beacon of generalizability and precision in color pattern learning. Our findings underscore our method's robustness, its unparalleled capacity for generalization, and its superior colorization finesse, making it a valuable contribution to the field of image colorization. Our contributions include the following:

- A Multiscale Pyramid Transformer that ensnares both micro and macro context for colorization that is both precise and visually engaging.
- A pioneering Dual-Attention module that leverages Separable Dilated Attention and Factorized Channel Attention to streamline the attentional complexity to linear terms without forgoing long-range dependence, thereby enhancing computational efficiency.
- An innovative color decoder, termed Color-Attention, which is proficient in learning colorization patterns directly from the grayscale and color images of the current training set, thereby boosting generalizability across datasets and eliminating the necessity for predicated color priors.
- Our experimental outcomes vividly demonstrate the supremacy of our model across diverse benchmark datasets, including in high-stakes computer vision tasks like classification, segmentation, and detection, thereby evidencing our model's robustness, generalization capacity, and superior colorization quality compared to leading-edge methods

II. Related Work

A. Colorization Methods

Colorization is a popular research area in computer vision that has seen significant improvements in recent years due to the advancements in deep learning methods and large-scale datasets. Existing reference-based colorization methods have struggled to achieve semantic consistency due to the lack of color knowledge. To overcome this issue, data-driven colorization methods such as deep learning have been introduced. Although these methods have shown promising results, they still face challenges such as the preference in color prior building and varying degrees of semantic discontinuity. To address these limitations, the Color-Attention module is introduced in this paper, which is an end-to-end module that learns colorization laws based on the training set. Additionally, the Dual-Attention Block is proposed to address the issue of semantic discontinuity.

In recent years, hierarchical designs in vision transformers have become increasingly popular. However, local window attention can limit their ability to model globally. To compensate for this, some designs slide the window, use horizontal and vertical attention, or introduce multi-axis self-attention modules. While multi-attention transformer structures have been explored in high-level designs, they are not commonly used for low-level tasks. This paper introduces the Multiscale Pyramid transformer, which integrates multi-scale perceptual field design into low-level tasks, providing a new solution for these types of problems.

The field of colorization has seen significant advancements in recent years with the development of deep learning methods and large-scale datasets. Reference-based colorization methods have been used to integrate grayscale input with color knowledge, but their results have been unsatisfactory due to the lack of semantic consistency. To address this, data-driven colorization methods have been implemented, with [1] being the first deep learning method for image colorization. [2] proposed a novel method for automatic image colorization by transforming the colorization problem into a classification problem by quantizing the color space. More recently, [3] used pre-trained GANs to recover vivid colors and Colorformer [4] introduced a transformer into the encoder part to improve the acquisition of semantic information. DDColor's [5] core idea involves establishing two decoders, one focusing on pixel semantic information and the other on coloring.

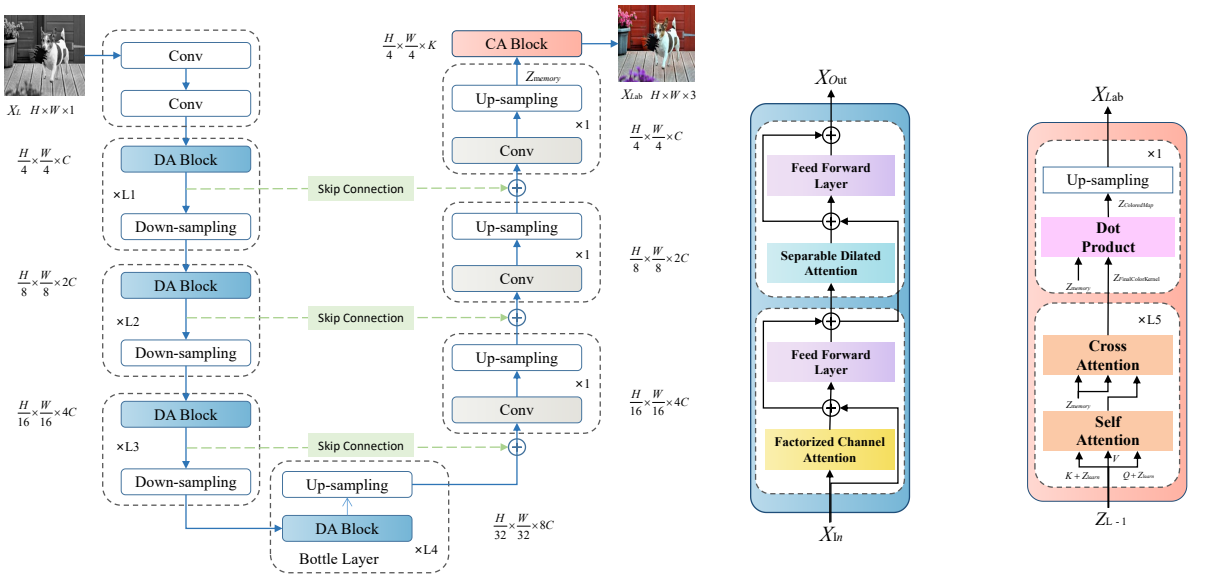
However, there are still issues with existing methods such as specific preferences in color prior building and varying degrees of semantic discontinuity. To overcome these limitations, this paper introduces the Color-Attention module, an end-to-end module that automatically learns colorization laws based on the training set. Additionally, the Dual-Attention Block is proposed to address the issue of semantic discontinuity.

B. Hierarchical Structures in Vision Transformers

The evolution of hierarchical structures in vision transformers signifies a pivotal shift in the landscape of this domain. Notable architectures such as PVT [6], CvT [7], Swin Transformer [8], GCVIT [9], Longformer [10], and HaloNet [11] have adopted a strategy of implementing localized window attention to effectively focus on local attributes while simplifying the computational demand to a linear level. This approach, however, introduces a potential limitation in global contextual understanding. To mitigate this, various methods have been employed: Swin Transformer [8] and Longformer [10] utilize a sliding window approach, CSwin Transformer incorporates a unique horizontal and vertical attention mechanism, and Maxvit [12] introduces a distinctive Multi-Axis Self-Attention (MaxSA) module. Additionally, DaViT [13] combines two types of self-attention to enhance its model's global and local information processing. Despite the prevalence of multiple attention structures in transformers for high-level tasks, their integration into low-level tasks remains rare. Addressing this gap, our study presents the Multiscale Pyramid Transformer with Dual-Attention, which amalgamates multi-scale perceptual field design with low-level task execution, proposing an innovative approach to these challenges.

III. Methods

Grayscale image colorization is a fascinating process that aims to restore the missing ab dimension, based solely on the l dimension in the CIELAB color space. The l dimension represents the luminance, while the ab dimension signifies



(a) Multiscale Pyramid Transformer (b) Dual-Attention Block (c) Color-Attention Block

Fig. 2: **The structure of the colorization network.** Multiscale Pyramid Transformer includes Encoder, Bottle Layer, Upsamling Stage, Decoder. (b) Dual-Attention Block includes Factorized Channel Attention, Separable Dilated Attention. (c) Color-Attention block shows the mechanism of colorization.

chromaticity. How to learn the mapping laws between the two is crucial. To accomplish this task, we have developed an innovative encoder-decoder structure, which uses a Dual-Attention Block to extract semantic features from grayscale images. The encoder stage is responsible for identifying the critical features, while the upsampling stage uses shortcut upsampling layers to ensure a smooth transition between layers. At the end of the upsampling stage, a Color-Attention block is introduced to perform the colorization process. This block uses a learnable color query to identify the color distribution from semantic information, which is then multiplied by the feature map to obtain the ab value. Finally, the ab value is combined with the l value to create a beautiful and realistic colored image.

A. Overview of Main Ideas

(1)Dual-Attention. The Dual-Attention framework is designed to optimize spatial analysis through a thoughtful division of the perceptual field, comprising Factorized Channel Attention (FCA) and Separable Dilated Attention (SDA). FCA treats the entire spatial domain as a unified entity, fostering a global understanding of the image. In contrast, SDA differentiates itself by dissecting the perceptual field into multiple distinct regions. Unique to SDA, each of these regions features an expanded receptive field akin to the effect of a dilated convolution, allowing the model to capture a more extensive context within each localized area. This attribute of SDA enables it to glean detailed and expansive insights from various parts of the image, thereby enriching the model's capacity to process and synthesize multi-scale spatial information. Such a dual-attention mechanism empowers the model with a versatile interpretive ability, crucial for both comprehensive high-level analyses and detailed low-level tasks like colorization.

(2)Color-Attention. This new decoder leverages the inherent long-range dependency capabilities of our Dual-Attention framework, requiring only a single level of attention at the end of the process. This simplification not only reduces computational demands but also enhances the adaptability of the colorization process, as it learns directly from the grayscale and color pairs in the training dataset without the need for predefined color priors. This method represents a significant step forward in making colorization more intuitive, efficient, and broadly applicable.

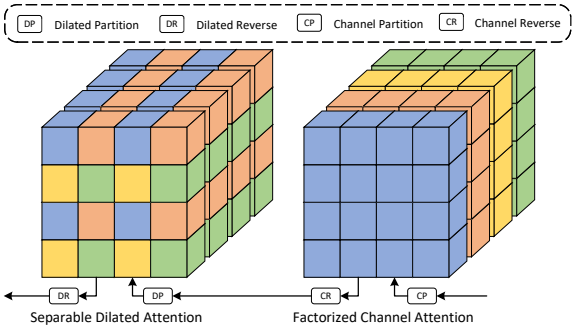


Fig. 3: **A diagram of the window partition of Dual-Attention block.**

B. Semantic Encoder

The network is created by stacking Dual-Attention blocks, featuring channel and dilated attention. These attentions seamlessly work together, gathering multi-scale information to capture even the smallest details of the input image. The network's end goal extends beyond colorization, achieving excellent results in tasks such as classification, segmentation,

and detection, as demonstrated in the transfer experiment. The Decoder is pivotal in the final stage, utilizing the deep semantic features to deliver precise and highly realistic colorization.

$$\mathcal{Z}_{semantic} = Dual - Attention_{\times 4}(Embedding(\mathcal{X}_L)) \quad (1)$$

1) Dual-Attention Block: The Dual-Attention block innovatively combines factorized channel attention (FCA) and separable dilated attention (SDA) to extract global and local information with linear complexity, as illustrated in Fig. 2. This architecture allows the model to efficiently process spatial information at different scales, which is essential for tasks that require nuanced understanding, such as image colorization. The synergy between FCA and SDA within the Dual-Attention block provides a balanced approach to spatial analysis. FCA offers a macroscopic view by aggregating global information, setting the stage for SDA to refine these insights with precise, localized details. This combination ensures that the model benefits from a holistic perspective, enabled by FCA, while also possessing the ability to zoom in on multiple specific regions with enhanced detail, courtesy of SDA. Such a dual-structured attention mechanism empowers the model to perform complex spatial tasks, achieving a deep and nuanced understanding of the image content. The formula is as follows:

$$Dual - Attention(\mathcal{X}) = \begin{cases} \mathcal{X}' = FFN(\mathcal{A}_{Channel}(\mathcal{X})) \\ \mathcal{X}'' = FFN(\mathcal{A}_{Dilated}(\mathcal{X}')) \end{cases} \quad (2)$$

Factorized Channel Attention The FCA mechanism is a streamlined adaptation of channel attention designed to capture global features effectively. It operates by dividing the input tensor into smaller groups along the channel dimension, where attention calculations are performed independently. This group-based approach reduces the computational complexity compared to traditional channel attention mechanisms, enhancing the model's scalability and efficiency. By focusing on global dependencies, FCA allows the model to grasp an overarching understanding of the entire spatial domain, providing a comprehensive context for subsequent localized analysis. Then, the factorized channel attention is computed as follows:

$$\mathcal{A}_{Channel}(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = Softmax\left(\frac{\mathcal{Q}^T\mathcal{K}}{\sqrt{C}}\right)\mathcal{V}^T, \quad (3)$$

where $\mathcal{Q} = \mathcal{X}\mathcal{W}^{QP}\mathcal{W}^{QD}, \mathcal{K} = \mathcal{X}\mathcal{W}^{KP}\mathcal{W}^{KD}, \mathcal{V} = \mathcal{X}\mathcal{W}^{VP}\mathcal{W}^{VD} \in \mathbb{R}^{HW\times H_d}$, $\mathcal{W}^P$ and $\mathcal{W}^D$ are point and depth convolutions, respectively .

$$\mathcal{FCA}(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = \left\{\mathcal{A}_{Channel}(\mathcal{Q}_j, \mathcal{K}_j, \mathcal{V}_j)^T\right\}_{j=0}^{\frac{C}{M}}, \quad (4)$$

where $\mathcal{Q}_i, \mathcal{K}_i, \mathcal{V}_i \in \mathbb{R}^{HW\times M}$ are the query, key, and value projections for the i-th group, obtained by splitting the input tensor $\mathcal{X}$ along the channel dimension into G groups of size C/M. The factorized attention computation is performed within each group, and the results are concatenated to form the final output. The complexity of factorized channel attention is $\Omega(\text{FCA}) = 2MHWC$, which is lower than the original channel attention complexity when M is smaller than C.

Separable Dilated Attention Inspired by the principles of dilated convolution, SDA expands the model's receptive field without increasing computational burden. It achieves this by employing a dilated partition strategy, which allows the attention mechanism to incorporate information from a broader spatial extent, thereby capturing local details across varied scales. To further optimize this process, depthwise separable convolutions are used for QKV (Query, Key, Value) transformations within the attention mechanism, enhancing the model's focus on local context. These modifications not only improve the receptive field but also shifts the computational cost from quadratic to linear, significantly boosting the model's efficiency. Then, the separable dilated attention is computed as follows:

$$\begin{aligned} &\mathcal{A}_{Spatial}(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = Concat\left(head_1, \ldots, head_{\frac{C}{H_d}}\right), \\ &where\ head_i = Softmax\left(\frac{\mathcal{Q}_i\mathcal{K}_i^T}{\sqrt{H_d}}\right)\mathcal{V}_i, \end{aligned} \quad (5)$$

where $\mathcal{Q}_i = \mathcal{X}_i\mathcal{W}_i^{QP}\mathcal{W}_i^{QD}, \mathcal{K}_i = \mathcal{X}_i\mathcal{W}_i^{KP}\mathcal{W}_i^{KD}, \mathcal{V}_i = \mathcal{X}_i\mathcal{W}_i^{VP}\mathcal{W}_i^{VD} \in \mathbb{R}^{HW\times H_d}$, $\mathcal{W}_i^P$ and $\mathcal{W}_i^D$ are point and depth convolutions, respectively .

$$\mathcal{SDA}(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = \{\mathcal{A}_{Spatial}(\mathcal{Q}_j, \mathcal{K}_j, \mathcal{V}_j)\}_{j=0}^{\frac{HW}{G^2}}, \quad (6)$$

where $\mathcal{Q}_j, \mathcal{K}_j, \mathcal{V}_j \in \mathbb{R}^{G^2\times C}$. The dilated partition: $(G\times\frac{H}{G}, G\times\frac{W}{G}, C) \rightarrow (G^2, \frac{HW}{G^2}, C) \rightarrow (\frac{HW}{G^2}, G^2, C)$ and dilated reverse is the inverse process of dilated partition. The computational complexity of separable dilated attention is $\Omega(\text{SDA}) = 2HWCG^2$, which is linear in the input size.

C. Painting Decoder

The colorization model described above is designed to produce highly detailed and accurate colorized images. The upsampling stage is a crucial component of the model, involving the use of three upsampling layers to incrementally enlarge the image. Each stage starts with enlarging the image using convolution and pixel-shuffle techniques, followed by fusing the extracted feature with the corresponding encoder image through the shortcut connection. The final color of the image is obtained through the advanced Color-Attention block, which utilizes embedding for upsampling and information aggregation to predict the ab value.

1) Upsampling Stage: In the upsampling stage, we upsample the semantic features using pixelshuffle. Specifically, we apply pixelshuffle three times with a specified up-sampling multiplier, denoted as f. The output of this stage is denoted as $\mathcal{Z}_{memory}$ and is used as input to the next stage.

$$\mathcal{Z}_{memory} = PixleShuffle_{\times 3}(\mathcal{Z}_{semantic}, f) \quad (7)$$

2) Color-Attention Block: In the color-attention block stage, we first build a learnable color kernel using the upsampled feature maps and a learnable query. The learnable query interacts with the semantic features of the feature graph using cross-attention. The gradient is obtained by back-propagating the target loss back to focus on the color of the object. This allows attention to have learnable parameters that can be continuously adjusted to improve color concern based on

TABLE I: **Quantitative results with SOTA methods on benchmark datasets.** ↑ and ↓ mean higher or lower is desired.

Method	ImageNet				COCO-Stuff				CelebA-HQ				ADE20K			
	FID ↓	CF ↑	Δ CF ↓	PSNR ↑	FID ↓	CF ↑	Δ CF ↓	PSNR ↑	FID ↓	CF ↑	Δ CF ↓	PSNR ↑	FID ↓	CF ↑	Δ CF ↓	PSNR ↑
CIC [2]	19.17	**43.92**	4.83	20.86	27.88	33.84	3.01	22.73	14.97	38.21	4.54	24.54	15.31	31.92	3.12	23.14
Zhang et al. [14]	7.30	27.23	11.86	**24.13**	17.43	25.95	10.90	**24.66**	11.81	36.98	5.77	**26.56**	-	-	-	-
Instcolor [15]	7.36	27.05	12.04	22.91	13.09	27.45	9.40	23.38	13.28	37.08	5.67	24.77	15.44	23.54	11.50	24.27
ChromaGAN [16]	5.16	27.49	11.60	23.12	25.65	27.86	8.99	23.56	14.43	**45.93**	3.18	24.54	-	-	-	-
DeOldify [17]	3.87	22.83	16.26	22.97	13.86	24.99	11.86	24.19	9.48	43.93	1.18	25.20	12.41	17.98	17.06	**24.40**
ColTran [18]	6.14	35.50	3.59	22.30	14.94	36.27	0.58	21.72	10.05	43.62	0.87	22.98	12.03	34.58	0.46	21.86
GCP [3]	3.62	35.13	3.96	21.81	12.10	28.58	8.27	21.58	-	-	-	-	13.27	27.57	7.47	22.03
BigColor [19]	1.24	40.01	0.92	21.24	9.07	**39.84**	2.99	20.46	-	-	-	-	-	-	-	-
Colorformer [4]	1.71	39.76	0.67	23.00	8.68	36.34	0.51	23.91	7.54	42.43	0.32	25.62	8.83	32.27	2.77	23.97
DDColor [5]	1.23	37.72	1.37	23.63	**7.24**	38.48	1.63	23.45	-	-	-	-	10.03	**35.27**	0.23	24.39
Ours	**1.21**	39.33	**0.24**	23.37	7.56	36.57	**0.28**	23.85	**5.68**	42.96	**0.21**	25.94	**9.51**	34.82	**0.22**	24.31

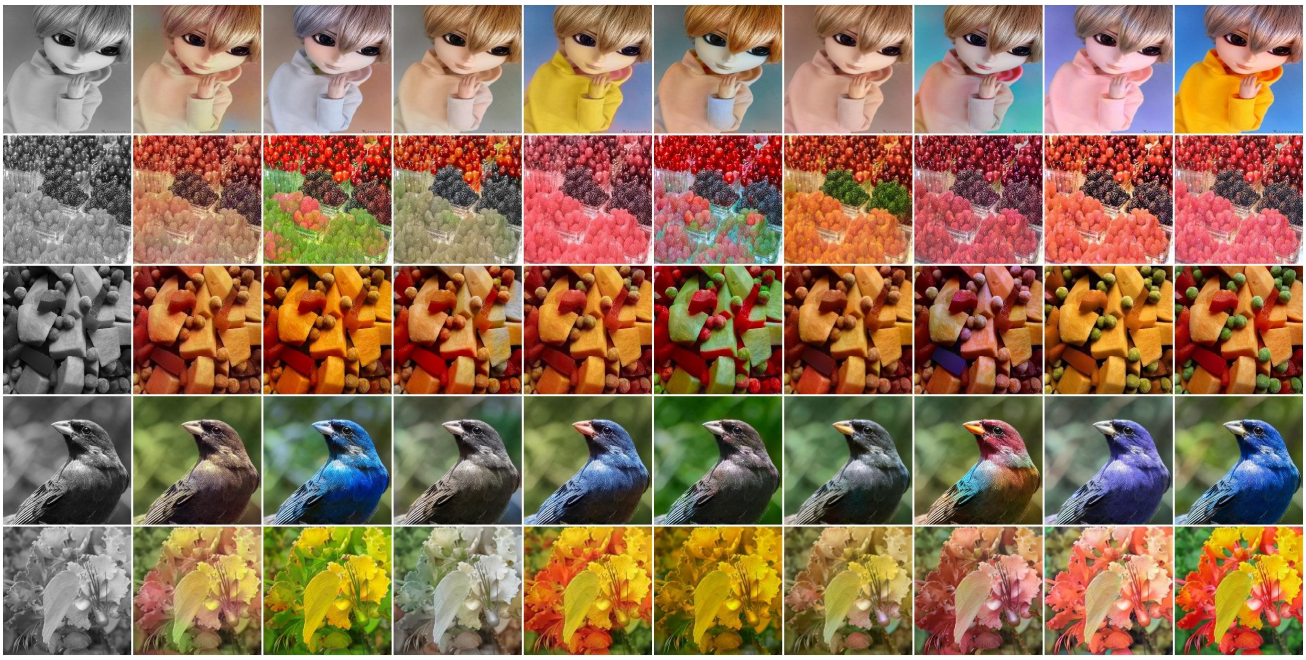


Fig. 4: **Visual comparisons with previous automatic colorization methods.** It should be noted that GT images are for reference, but the evaluation criteria should not be exactly the same color similarity.

the current training samples without the need for color prior, thus achieving end-to-end learning.

Building learnable color kernel The color kernel is built in two steps. Firstly, we use the output of the previous layer, $\mathcal{Z}_{L-1}$, as $\mathcal{QKV} \in \mathbb{R}^{K\times C}$, where $\mathcal{Z}_0$ is initialized to 0, and $L \in [1,\ L5]$. We then sum $\mathcal{QK}$ with $\mathcal{Q}_{learn} \in \mathbb{R}^{K\times C}$, respectively, and perform self-attention. The next query is obtained under the guidance of learnable color queries.

$$\begin{aligned} \mathcal{Q}_{1st} &\leftarrow \mathcal{Z}_{L-1} + \mathcal{Q}_{learn}, \\ \mathcal{K}_{1st} &\leftarrow \mathcal{Z}_{L-1} + \mathcal{Q}_{learn} \\ \mathcal{V}_{1st} &\leftarrow \mathcal{Z}_{L-1} \\ \mathcal{Q}_{2nd} &= \mathrm{Softmax}\left(\mathcal{Q}_{1st}\mathcal{K}_{1st}^{T}\right)\mathcal{V}_{1st} \end{aligned} \tag{8}$$

Secondly, we use the upsampled feature map, $\mathcal{Z}memory$, as $\mathcal{KV} \in \mathbb{R}^{C\times H\times W}$, and perform cross-attention. We repeat this process L5 times to obtain the final color kernel, denoted as $\mathcal{Z}_{L5}$.

$$\begin{aligned} \mathcal{K}_{2nd} \leftarrow \mathcal{Z}_{memory}, \mathcal{V}_{2nd} \leftarrow \mathcal{Z}_{memory} \\ \mathcal{Z}_{L} = \mathrm{Softmax}\left(\mathcal{Q}_{2nd}\mathcal{K}_{2nd}^{T}\right)\mathcal{V}_{2nd} \end{aligned} \tag{9}$$

Reconstructing Missing Color Next, we reconstruct the missing color by multiplying the final color kernel, $\mathcal{Z}_{FinalColorKernel}$, with the upsampled feature map, $\mathcal{Z}_{memory}$. This yields the colored map, denoted as $\mathcal{Z}_{Coloredmap}$. After embedding upsampling, we add it to the ab value to obtain the colored image, denoted as $\mathcal{X}_{Lab}$.

$$\mathcal{Z}_{Coloredmap} = \mathcal{Z}_{FinalColorKernel} \cdot \mathcal{Z}_{memory} \tag{10}$$

After embedding upsampling, it is added to the ab value. Finally, the colored image is obtained.

$$\begin{aligned} \mathcal{X}_{ab} &= Upsampling(\mathcal{Z}_{Coloredmap}) \\ \mathcal{X}_{Lab} &= \mathcal{X}_{L} + \mathcal{X}_{ab} \end{aligned} \tag{11}$$

D. Objectives

Pixel loss. We utilize the L1 norm to calculate the pixel loss, measuring the disparity in color between the generated color image $\hat{y}$ and the true color image y. This loss function aids

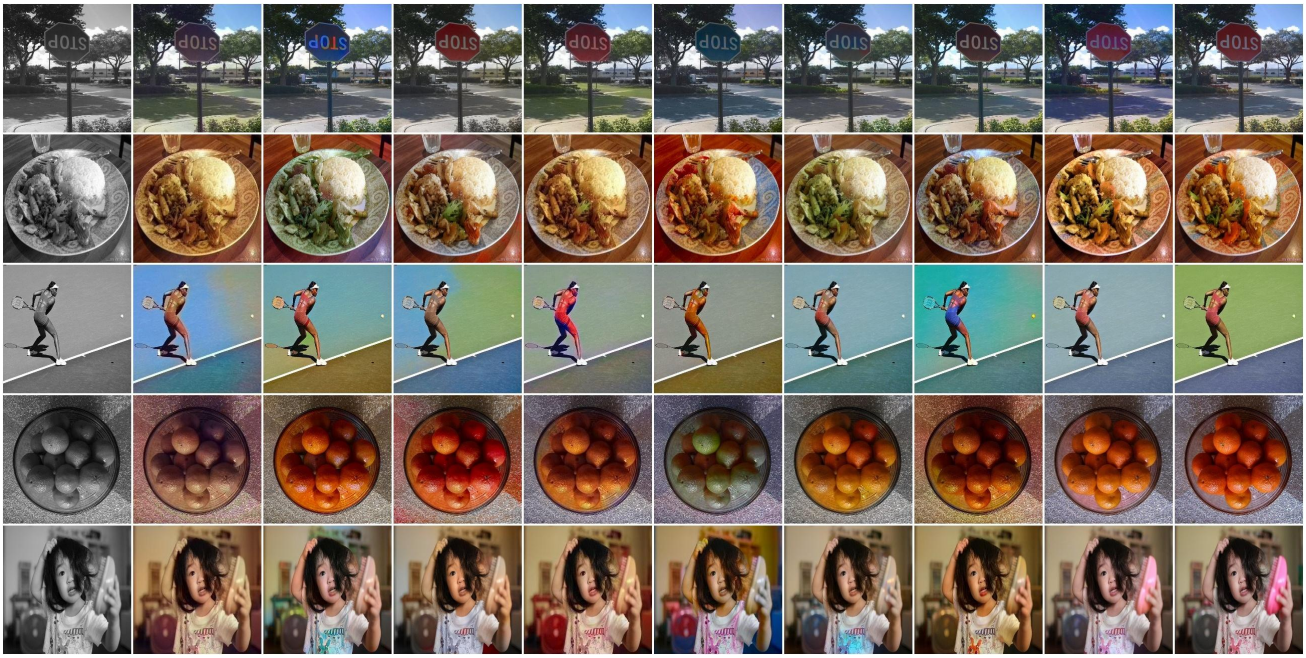


Fig. 5: **More visual comparisons with previous automatic colorization methods on COCO-Stuff.**

the generator in producing colors that are closer to the actual images by providing supervision at the pixel level.

$$\mathcal{L}_{pix} = ||y - \hat{y}||_1 \tag{12}$$

Perceptual loss. To enhance the quality of the generated images, we employ the VGG19 network [20], to extract features from both $\hat{y}$ and y and minimize the discrepancy in their semantic content as proposed by Johnson et al. [21]

$$\mathcal{L}_{per} = \sum_{l=1}^{5} w_l ||\Phi_l(\hat{y}) - \Phi_l(y)||_1 \tag{13}$$

Here, $\Phi_l(\cdot)$ represents the output from the layer conv_l_1 in the VGG19 network, and w_l denotes the layer-specific weight, assigned values of $\frac{1}{16}$, $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$, and 1.0, respectively.

Adversarial Loss. To further refine the image quality, we incorporate a discriminator, following the approach of Goodfellow et al. [22], creating a competitive scenario with the generator. We apply the discriminator from the Wasserstein GAN, as defined by Arjovsky et al. [23] The adversarial loss functions are defined as follows:

$$L_d = \frac{1}{m}\sum_{i=1}^{m} D_w\left(x^{(i)}\right) - \frac{1}{m}\sum_{i=1}^{m} D_w\left(G_\theta\left(z^{(i)}\right)\right) \tag{14}$$

$$L_g = -\frac{1}{m}\sum_{i=1}^{m} D_w\left(G_\theta\left(z^{(i)}\right)\right) \tag{15}$$

Full Objectives. The comprehensive loss for the generator combines the above losses, weighted by their respective coefficients, as follows:

$$\mathcal{L}(y, \hat{y}) = \lambda_{pix}\mathcal{L}_{pix} + \lambda_{per}\mathcal{L}_{per} + \lambda_g\mathcal{L}_g \tag{16}$$

IV. Experiments

A. Datasets and Implementation Details

Dataset. Our experiments are conducted on three datasets: ImageNet [24], COCO-Stuff [25], CelebA-HQ [26] and ADE-20K [27]. The ImageNet training subset serves as our training set, while its validation subset is used for testing. Furthermore, we validate the versatility of our method by evaluating it on COCO-Stuff, CelebA-HQ and ADE-20K datasets without additional fine-tuning.

Evaluation Metrics. Initially, PSNR is considered for evaluation but its effectiveness for this specific task is questionable. Thus, our primary metrics are Frechet inception distance (FID) [28] and Colorfulness Score (CF) [29], where FID measures the similarity between the distributions of generated and real images, and CF quantifies the vibrancy of the generated image's colors.

Implementation Details. The network is optimized using the Adam algorithm [30], with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a learning rate of $1e-4$. The weights for the pixel, perceptual, and adversarial losses are set to $\lambda_{pix} = 0.1$, $\lambda_{per} = 5.0$, and $\lambda_g = 1.0$, respectively. Encoder settings include a dilated window $G = 8$, head dimension and channel group dimension $M = 24$ and channel dimension $C = 96$. The encoder layers are configured as $L1 = 1, L2 = 1, L3 = 3, L4 = 1$, and $L5 = 6$. In the Color-Attention block, K is set to 64, and the upsampling factor is 4. The model trains over $200,000$ iterations with a batch size of 16, undergoing learning rate reductions by 0.5 at $80,000$, $120,000$, and $160,000$ iterations. Both training and evaluation are conducted on 256×256 images using 8 Tesla V100 GPUs.

Fig. 6: **More visual comparisons with previous automatic colorization methods on CelebA-HQ.**

TABLE II: **Effect of Dual-Attention and Color-Attention.** Fix the Encoder and ablate the Color-Attention module in the above table. Instead, Encoder is ablated in the table below. The"CM" and "GLH-Trans" modules originate from ColorFormer. The GLH-Transformer configuration is based on ColorFormer. For ResNet50, Swin Transformer-Tiny, and Twins-Small, we use their ImageNet-1k settings.

Encoder	Decoder	FID ↓	CF ↑	Δ CF ↓	Decoder	FID ↓	CF ↑	Δ CF ↓
GLH-Trans.	w Color-Attention	1.42	39.51	0.42	w/o Color-Attention	1.85	36.70	2.39
ResNet50	w Color-Attention	3.23	36.23	2.86	w/o Color-Attention	4.10	34.27	4.82
Swin-Trans.	w Color-Attention	2.44	38.15	0.94	w/o Color-Attention	2.68	36.03	3.06
Twins	w Color-Attention	2.21	38.21	0.88	w/o Color-Attention	2.48	36.46	2.63
Dual-Attention	w Color-Attention	**1.21**	39.33	**0.24**	w/o Color-Attention	**1.77**	37.42	**1.67**
Decoder	Encoder	FID ↓	CF ↑	Δ CF ↓	Encoder	FID ↓	CF ↑	Δ CF ↓
CM	GLH-Trans.	1.71	39.76	0.67	Dual-Attention	**1.55**	39.44	**0.35**

B. Comparison with State-of-the-Art Methods

Quantitative Comparison. We conducted an extensive evaluation of our method against previously established techniques across three distinct datasets, with the comparative results documented in Table I. We ensured a fair assessment by employing the code and checkpoints provided by the authors of the respective methods. Our approach surpasses all the other methods, exhibiting the lowest FID [28] on all datasets, signifying our model's capability to produce authentic and natural images. Although some methods outrank ours on the color fidelity (CF) metric [29], they have significantly higher FID scores, indicative of their unrealistic and oversaturated color palettes. To circumvent this issue, we adopted the Δ CF measure to evaluate the CF gap with ground truth images. Our Δ CF values remarkably outperform all the other methods on both datasets, indicating that our images have a more realistic color distribution.

Qualitative Comparison. We present the results of colorizing grayscale images in Fig. 4, where we utilize various scenes from the ImageNet validation set and compare our approach with other established schemes. Our approach produces images with more accurate and coherent color distribution. Notably, the colors generated by CIC [2] and GCP [3] are not reflective of reality, especially in scenes dominated by multicolor such as

items and people. Instcolor [15], CT2 [31], and Bigcolor [19] exhibit varying degrees of incoherence in their color output. While Colorformer [4] has a richer color palette, it suffers from serious color errors and spills in several scenarios. Our method generated color images that are both semantically consistent and realistic, demonstrating its superior performance compared to the other techniques.

In addition, our method also perform well on COCO-stuff without fine-tuning, thanks to its pure end-to-end learning approach that can adjust the coloring law according to all samples in the training set. In Fig. 5, our results show that our method can produce both continuous semantics and diverse colors without being overly bright.

Furthermore, we select CelebA-HQ to verify whether our network can normally color a large number of pictures of the same type. In Fig. 6, our network effectively colors the skin of each face, clothes and the background, demonstrating the robustness of our method.

Overall, our method is a promising solution for automatic colorization tasks, especially in scenarios where limited supervision is available. Our approach is robust and can produce high-quality colorization results, making it suitable for a wide range of applications.

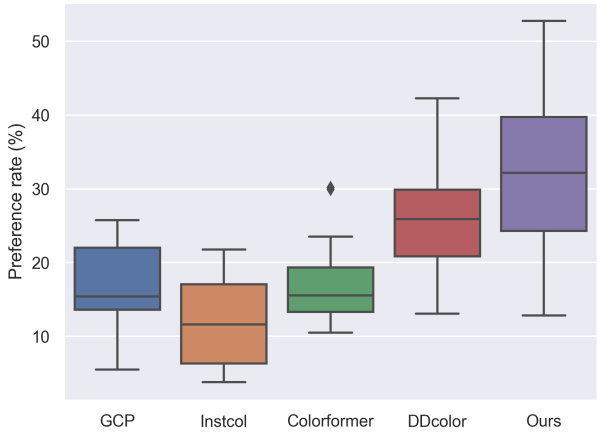


Fig. 7: **Boxplot of user study.** The solid gray line in the boxplot is the percentage preference of the median.

User study We conducted a user study to investigate people's subjective preference for the results of colorization methods, and selected four methods to compare with our method. Our survey had 95 volunteers participating. Specifically, we randomly picked 50 images from the ImageNet validation set and obtained the colorization results of each method to present to the subjects. The subjects chose the best colored images from them. As shown in the figure 7, our method is more preferred by the users.

Runtime and Model Parameters Our method colorizes gray image of 256×256 at 45 FPS with model parameters of 55.2M. All tests are performed using an NVIDIA Tesla V100 32G GPU.

C. Ablation Studies

1) Effect of Dual-Attention and Color-Attention: Table II presents our ablation experiments, where we evaluated the performance of Dual-Attention Block and Color-Attention block by replacing them with pure convolutional structures and transformer-based structures. Specifically, for the transformer block, we utilize the method with multiple attention structures. Furthermore, our proposed method exhibits a smoother perceptual field transition, making it a superior choice for constructing long-range dependencies. As a result, our approach provides better information at multiple scales and achieves a lower FID score. On the other hand, removing the Color-Attention module and replacing it with the same number of Encoder Blocks led to a significant decrease in CF, indicating that the extracted Color-Attention module plays a crucial role in color reconstruction. We observe a similar CF drop when replacing Color-Attention with CM module, highlighting that end-to-end learning helps achieve the goal of unity, leading to improved learning performance.

Model	FID ↓	CF ↑	Δ CF ↓
FCA-SDA	**1.21**	39.33	**0.24**
FCA-DA	1.60	39.43	0.34
CA-SDA	2.52	39.95	0.86
SDA-FCA	2.32	39.62	0.53
DA-FCA	2.58	40.17	1.08
SDA-CA	2.88	40.44	1.35
FCA-FCA	2.66	40.63	1.54
SDA-SDA	2.75	40.89	1.80
FCA	3.68	42.13	3.04
SDA	3.95	42.26	3.17

TABLE III: Attention Arrangement. FCA, SDA denote Factorized Channel Attention and Separable Dilated Attention respectively. CA, DA denote traditional Channel Attention and Dilated Attention

2) Attention Arrangement: Table III elucidates the findings from our comprehensive ablation studies concerning the strategic configuration and amalgamation of attention mechanisms within the Dual-Attention Block framework. This investigation validates the hypothesis that the most effective colorization outcomes are achieved through a meticulous progression of attention mechanisms, specifically transitioning from Factorized Channel Attention (FCA) to Separable Dilated Attention (SDA). This sequence underscores the importance of engaging both global and local receptive fields sequentially to adeptly capture and integrate multi-scale spatial information. The empirical data distinctly highlight the FCA-SDA configuration's superiority, as evidenced by its unparalleled Frechet Inception Distance (FID) score of 1.21 and a minimal Δ CF of 0.24, underscoring the critical synergy between FCA and SDA. Such an arrangement ensures that the model not only grasps broad thematic elements but also hones in on the finer nuances, thus ensuring an unmatched level of colorization precision and detail.

3) Queries number selection: Table IV displays the results of our ablation experiments examining the impact of the number of Learnable Color Queries on Color-Attention. Our findings reveal that the model's performance steadily improves as the number of queries increases, reaching its optimal value

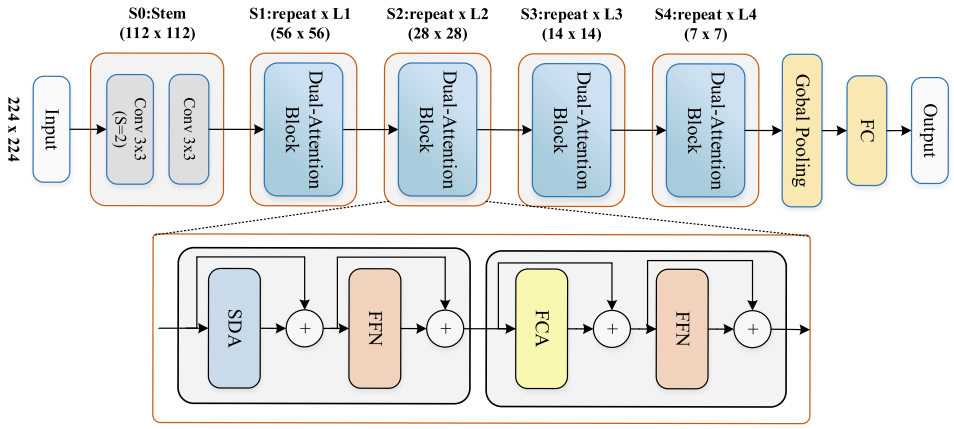


Fig. 8: **The structure used for the classification task.**

TABLE IV: **Queries num.** The queries are gradually increased to find the best number.

Query num	FID ↓	CF ↑	Δ CF ↓
50	1.33	38.57	0.52
100	1.27	38.72	0.37
150	**1.21**	39.33	**0.24**
200	1.27	38.43	0.66
500	1.24	38.50	0.49

at around 150. However, as the number of queries continues to increase beyond this point, we observe a diminishing return on performance, with some instances even exhibiting a small decrease. While some previous colorization methods have treated this problem as a classification task for class 313, we posit that the current number of queries is sufficient to represent the embedding of the color space, given that each query can embody multiple colors.

D. Visualizing Final Color Kernel

We aim to illustrate the functionality of our learned color kernel by visually representing it. As depicted in Fig. 9, we employ the color kernel on the feature map output from the prior stage and generate a heatmap. Shades of red represent high activation values while shades of blue represent low activation. We proceeded to visualize the first four color kernels and discovered that each kernel emphasizes different regions of high activation in the image feature map. For instance, in the top row of the image, the first four kernels concentrate on the outside of the car, inside the car, and the road, respectively. This specificity is attributable to the precise multi-scale information offered by the Encoder.

E. Details of ImageNet Classification

The Dual-Attention model is a powerful tool for image classification tasks, and its structure is carefully designed for optimal performance. The building block for the classification task

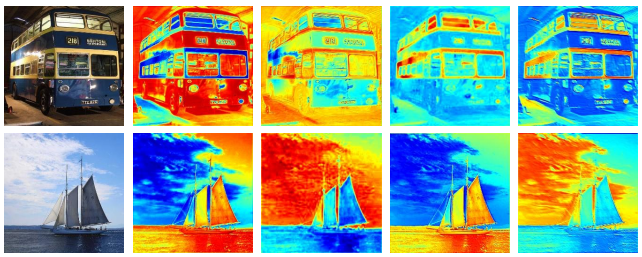


Fig. 9: **Visualization of learned color kernel.**

is the Dual-Attention block, which uses attention mechanisms to capture long-range dependencies between different parts of the input image. The permutation of attention is inverted to facilitate high-level related tasks, which require attention in a detail-to-outline manner. Fig. 8 shows its structure

1) Model Instantiation: After the first convolution building block, the resolution of the feature map is reduced progressively while its dimension is increased correspondingly. Specifically, for an input image with size $H \times W$, the feature map is first reduced to $H/2 \times W/2$, then to $H/4 \times W/4$ with dimension C, $H/8 \times W/8$ with dimension 2C, $H/16 \times W/16$ with dimension 4C and finally to $H/32 \times W/32$ with dimensions 8C. The head dimension for all attention layers is set to 32, and two convolutional layers are used in the S0 stage. For an input image of size 224×224, a square window of size 7×7 is selected.

2) Performance comparisons: Experimental settings for both pretraining and fine-tuning of the Dual-Attention models on ImageNet-1K are provided in Table VI. Table VIII presents three different network configurations that are considered in the experiments. Results for the classification task can be found in Table V and IX, demonstrating the effectiveness of the Dual-Attention model in achieving state-of-the-art performance on the ImageNet-1K dataset. With a similar number of parameters and FLOPs, our model outperforms other models and achieves the state-of-the-art. Dual-Attention-T is 2% better than PVTv2-B2 and Dual-Attention-S is 1% better than COATNET-1. Furthermore, Dual-Attention-B has an increase of 0.5% and 0.35% compared with MAXVIT and

TABLE V: **Performance comparison under ImageNet-1K setting.**

Model	Params (M)	FLOPs (G)	T-1 Acc. (%)	Model	Params (M)	FLOPs (G)	T-1 Acc. (%)
ResNet-50 [32]	25.0	4.1	76.2	ResNet-152 [32]	60.0	11.0	78.3
DeiT-Small/16 [33]	22.1	4.5	79.8	PVT-Large [6]	61.4	9.8	81.7
PVT-Small [6]	24.5	3.8	79.8	DeiT-Base/16 [33]	86.7	17.4	81.8
Swin-Tiny [8]	28.3	4.5	81.2	CrossViT-Base [34]	104.7	21.2	82.2
CvT-13 [7]	20.0	4.5	81.6	T2T-ViT-24 [35]	64.1	14.1	82.3
CoAtNet-0 [36]	25.0	4.2	81.6	CPVT-Base [37]	88.0	17.6	82.3
CaiT-XS-24 [38]	26.6	5.4	81.8	PoolFormer-M48 [39]	73.0	11.9	82.5
ViL-Small [40]	24.6	5.1	82.0	TNT-Base [41]	65.6	14.1	82.8
PVTv2-B2 [42]	25.4	4.0	82.0	ViL-Base [40]	55.7	13.4	83.2
Focal-Tiny [43]	29.1	4.9	82.2	UFO-ViT-B [44]	64.0	11.9	83.3
DaViT-Tiny [13]	28.3	4.5	82.8	Swin-Base [8]	87.8	15.4	83.4
MaxViT-Tiny [12]	31.0	5.6	83.6	CaiT-M24 [38]	185.9	36.0	83.4
Dual-Attention-Tiny(Ours)	28.7	5.3	**83.7**	PyramidTNT-M [45]	85.0	8.2	83.5
ResNet-101 [32]	45.0	7.9	77.4	NFNet-F0 [46]	71.5	12.4	83.6
PVT-Medium [6]	44.2	6.7	81.2	QnA-Base [47]	56.0	9.7	83.7
CvT-21 [7]	32.0	7.1	82.5	PVTv2-B5 [42]	82.0	11.8	83.8
Swin-Small [8]	49.6	8.7	83.1	Focal-Base [43]	89.8	16.0	83.8
ViL-Medium [40]	39.7	9.1	83.3	ConvNeXt-B [48]	89.0	15.4	83.8
CaiT-S36 [38]	68.0	13.9	83.3	Shuffle-B [49]	88.0	15.6	84.0
CoAtNet-1 [36]	42.0	8.4	83.3	CoAtNet-2 [36]	75.0	15.7	84.1
Focal-Small [43]	51.1	9.1	83.5	CSwin-B [50]	78.0	15.0	84.2
CSwin-S [50]	35.0	6.9	83.6	MPViT-B [51]	74.8	16.4	84.3
VAN-Large [52]	44.8	9.0	83.9	GC ViT-B [9]	90.0	14.8	84.4
UniFormer-B [53]	50.0	8.3	83.9	MaxViT-Small [12]	69.8	11.7	84.5
DaViT-Small [13]	49.7	8.8	84.2	DaViT-Base [13]	87.9	15.5	84.6
Dual-Attention-Small(Ours)	51.2	9.1	**84.4**	Dual-Attention-Base(Ours)	77.8	16.2	**84.9**

TABLE VI: **Detailed hyperparameters used in ImageNet-1K experiments .**

	Pretraining	finetuning
Center crop	True	False
Loss type	softmax	softmax
Train epochs	300	150
Train Batchsize	2048	512
Optimizer type	AdamW	AdamW
Learning Rate	5e-4	5e-5
Weight Decay	0.05	1e-8
LR decay schedule	cosine	None
Gradient clip	1.0	1.0
EMA decay rate	None	0.9999

DAVIT.

F. Details of 2D Object Detection and Instance Segmentation

1) Setting in COCO: We evaluate the performance of our Dual-Attention models on the COCO2017 [54] dataset using two-stage object detection and instance segmentation frameworks, specifically the Mask R-CNN [55] and Cascade Mask R-CNN pipelines [55]. The COCO2017 dataset contains 118K training and 5K validation samples. We train our models for 36 epochs using a 3x schedule and initialize them with weights pre-trained on the ImageNet-1k dataset [24]. During training, we use the AdamW optimizer [30] with an initial learning rate of 1e-4. We also employ random depth drops of 0.1, 0.2, and 0.3 for the tiny, small, and base models, respectively.

2) Performance comparisons: In Tables X, the results of experiments based on Mask R-CNN is shown. On the tasks of object detection, our networks with three setups outperform state-of-the-art networks in all metrics, including DAVIT, SWIN. And there is a small reduction in the number of parameters. It is worth mentioning that Dual-Attention-S outperforms DaViT-B and reduces the number of parameters by 39.5%. To further evaluate our models, we use the Cascade Mask R-CNN pipeline and report the performance comparison in Table VII. Our results demonstrate the effectiveness of Dual-Attention models in achieving state-of-the-art performance on the COCO2017 dataset.

G. Details of 3D object detection

1) Setting in nuScenes.: We conduct experiments to evaluate the performance of CenterPoint [58] on the nuScenes [59] dataset, which is a challenging benchmark for 3D detection tasks. In particular, we focus on the two 3D encoders, PointPillars and VoxelNet, to implement the CenterPoint backbone, and replace the Second Block with the Dual-Attention Block in the 2D backbone. To address the long-tail class distribution in the nuScenes dataset, we use ground-truth sampling, which copy-pastes the points in the annotation frame from one frame to another.

TABLE VII: **COCO object detection and segmentation results with Cascade Mask R-CNN.**

Backbone	Resolution	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}	FLOPs	Pars.
ResNet-50	1280×800	46.3	64.3	50.5	40.1	61.7	43.4	739G	82M
X101-32	1280×800	48.1	66.5	52.4	41.6	63.9	45.2	819G	101M
X101-64	1280×800	48.3	66.4	52.3	41.7	64.0	45.1	972G	140M
ConvNeXt-T	1280×800	50.4	69.1	54.8	43.7	66.5	47.3	741G	-
ConvNeXt-S	1280×800	51.9	70.8	56.5	45.0	68.4	49.1	827G	-
ConvNeXt-B	1280×800	52.7	71.3	57.2	45.6	68.9	49.5	964G	-
Swin-T	1280×800	50.4	69.2	54.7	43.7	66.6	47.3	745G	86M
Swin-S	1280×800	51.9	70.7	56.3	45.0	68.2	48.8	838G	107M
Swin-B	1280×800	51.9	70.5	56.4	45.0	68.1	48.9	982G	145M
As-ViT-L	1024×1024	52.7	72.3	57.9	45.2	69.7	49.8	1094G	139M
UViT-T	896×896	51.1	70.4	56.2	43.6	67.7	47.2	613G	47M
UViT-S	896×896	51.4	70.8	56.2	44.1	68.2	48.0	744G	54M
UViT-B	896×896	52.5	72.0	57.6	44.3	68.7	48.3	975G	74M
MaxViT-T	896×896	52.1	71.9	56.8	44.6	69.1	48.4	475G	69M
MaxViT-S	896×896	53.1	72.5	58.1	45.4	69.8	49.5	595G	107M
Dual-Attention-T	896×896	52.8	72.9	57.4	45.5	70.1	49.3	417G	87M
Dual-Attention-S	896×896	**53.3**	**73.3**	**58.3**	**45.7**	**70.6**	**49.8**	498G	123M

TABLE VIII: **The network configuration.** L denotes number of blocks. D denotes number of channels for each stage.

Stages	Size	Dual-Attention-T	Dual-Attention-S	Dual-Attention-B
S0-Conv-stem	1/2	L=2 D=64	L=2 D=64	L=2 D=128
S1-Dual-Attention	1/4	L=1 D=96	L=1 D=96	L=1 D=128
S2-Dual-Attention	1/8	L=1 D=192	L=1 D=192	L=1 D=256
S3-Dual-Attention	1/16	L=3 D=384	L=8 D=384	L=8 D=512
S4-Dual-Attention	1/32	L=1 D=768	L=1 D=768	L=1 D=1024

TABLE IX: **Performance comparison under ImageNet-1K.** ↑ denote the model is evaluated with resolution of 384×384

Model	Params	FLOPs	Top-1 Acc.
CvT-21 ↑ [7]	32M	24.9G	83.30
Swin-B ↑ [8]	88M	47.0G	84.50
CSwin-B ↑ [50]	78M	47.0G	85.40
CoAtNet-3 ↑ [36]	168M	107.4G	85.80
MaxViT-T ↑ [12]	31M	17.7G	85.24
MaxViT-S ↑ [12]	69M	36.1G	85.74
Dual-Attention-T ↑	**27.6M**	**15.1G**	**85.67**
Dual-Attention-S ↑	**51.9M**	**27.8G**	**85.85**
Dual-Attention-B ↑	**79.3M**	**53.2G**	**86.21**

TABLE X: **COCO object detection and segmentation with Mask R-CNN**

Backbone	Params	FLOPs	AP^b	AP^m
ResNet50 [32]	44.2M	260G	41.0	37.1
PVT-Small [6]	44.1M	245G	43.0	39.9
ViL-Small [40]	45M	174G	43.4	39.6
Swin-Tiny [8]	47.8M	264G	46.0	41.6
Focal-Tiny [43]	48.8M	291G	47.2	42.7
DaViT-Tiny [13]	47.8M	263G	47.4	42.9
Dual-Attention-T	47.5M	281G	**48.3**	**43.2**
ResNeXt101-32x4d [32]	62.8M	340G	44.0	39.2
PVT-Medium [6]	63.9M	302G	44.2	40.5
ViL-Medium [40]	60.1M	261G	44.6	40.7
Swin-Small [8]	69.1M	354G	48.5	43.3
Focal-Small [43]	71.2M	401G	48.8	43.8
DaViT-Small [13]	69.2M	351G	49.5	44.3
Dual-Attention-S	64.9M	369G	**50.1**	**44.8**
ResNeXt101-64x4d [32]	102M	493G	44.4	39.7
PVT-Large [6]	81M	364G	44.5	40.7
ViL-Base [40]	76.1M	365G	45.7	41.3
Swin-Base [8]	107M	496G	48.5	43.4
Focal-Base [43]	110M	533G	49.0	43.7
DaViT-Base [13]	107.3M	491G	49.9	44.6
Dual-Attention-B	99.3M	546G	**50.5**	**45.0**

For data augmentation, we use random flips along the X and Y axes and global scaling with a random factor of [0.95, 1.05], and a random global rotation between $[-pi/8, pi/8]$ for nuScenes. We optimize the model using the AdamW [30] optimizer with a single-cycle learning rate strategy with a maximum learning rate of 1e-3, weight decay of 0.01, and momentum of 0.85 to 0.95. We train the model on 8 V100 GPUs with 20 epochs of batch size 16.

The main evaluation metrics for 3D detection on nuScenes are the mean average precision (mAP) and the nuScenes detection score (NDS). The mAP uses bird's eye view center distances <0.5m, 1m, 2m, and 4m instead of the standard box overlap, while the NDS is a weighted average of mAP and other attribute metrics, including translation, scale, orientation, velocity, and other box attributes. It can be seen that the results in table XI and XII.

2) Setting in Kitti.: For the Kitti 3D detection task, our team focus on detecting three categories: Cars, Pedestrians, and Cyclists. To evaluate the performance of our model, we use several metrics including bird's eye view (BEV), 3D, 2D, and average orientation similarity (AOS).

To improve the 2D detection, we replace the 2D backbone in the widely used SECOND framework [60] with a Dual-

TABLE XI: **Comparison with Centerpoint for 3D detection on nuScenes validation.**

Encoder	Backbone	Car	Truck	Bus	Trailer	Vehicle	Pedestrian	Motorcycle	Bicycle	Traffic_cone	Barrier
VoxelNet [56]	Second	0.851	0.526	0.656	0.364	**0.191**	0.822	0.591	0.409	0.642	0.608
	Dual-Attention	**0.875**	**0.544**	**0.669**	**0.374**	0.190	**0.828**	**0.598**	**0.426**	**0.652**	**0.622**
PointPillars [57]	Second	0.840	0.500	0.628	0.338	0.125	0.776	0.426	0.174	0.527	0.594
	Dual-Attention	**0.852**	**0.511**	**0.643**	**0.362**	**0.173**	**0.791**	**0.597**	**0.291**	**0.595**	**0.576**

TABLE XII: **We also show mean average precision (mAP) and nuScenes detection score (NDS)**

Encoder	BackBone	mAP	NDS	FPS
VoxelNet [56]	Second	56.59	65.15	14
	Dual-Attention	**57.78**	**66.35**	**21**
PointPillar [57]	Second	49.27	59.58	11
	Dual-Attention	**53.91**	**63.51**	**17**

TABLE XIII: **Comparisons between Second and Dual-Attention backbone with 3 classes on the kitti validation set with SECOND .**

Metrics	BackBone	easy	moderate	hard
bbox AP11	Second	82.59	77.82	75.67
	Dual-Attention	**82.85**	**78.59**	**75.76**
bev AP11	Second	79.87	72.53	68.60
	Dual-Attention	**80.95**	**73.55**	**69.00**
3d AP11	Second	75.59	65.19	62.03
	Dual-Attention	**77.16**	**67.33**	**63.15**
aos AP11	Second	80.68	75.53	73.33
	Dual-Attention	**80.83**	**76.34**	**73.41**

Attention Block. The 2D detection is performed on the image plane, while the average orientation similarity measures the average orientation similarity (measured in BEV) of the 2D detection.

To assess the difficulty of the detection task, we evaluate the detection results according to three difficulty levels: easy, medium, and hard. These levels are determined based on the object size, occlusion status, and truncation level of the objects.

During training, we use 8 V100 GPUs and train the model for 40 epochs with a batch size of 16. We follow the official KITTI evaluation protocol, where the IoU threshold is set to 0.7 for the car class and 0.5 for pedestrians and cyclists. The IoU threshold is the same for both the bird's eye view and full 3D evaluation.

To compare our approach with other methods, we use the average precision (AP) metric. The results of our experiments can be found in XIII, which demonstrates the effectiveness of our model in detecting the three categories of objects in the Kitti dataset.

3) Quantitative analysis.: Our research has demonstrated significant improvements in performance on both the KITTI and nusence datasets by using different encoding methods and attention block arrangements.

Specifically, for the KITTI dataset, our model achieve an increase of 0.118%, 0.583%, 1.805%, and 0.109% in bbox, bev, 3d, and aos metrics respectively under the hard condition of KITTI AP11. These improvements are achieved by using the Dual-Attention Block as the 2D backbone in the SECOND framework.

In the nusence dataset, we evaluate the performance of our model using two different encoding methods: voxel and pointpillar. Using the voxel encoder, our model achieve improvements of 2.102% and 1.841% in mAP and NDS respectively. By using the pointpillar encoder, our model achieve even greater improvements of 9.417% and 6.596% in mAP and NDS respectively. Furthermore, the execution speed is greatly improved, with an increase in FPS of about 50%.

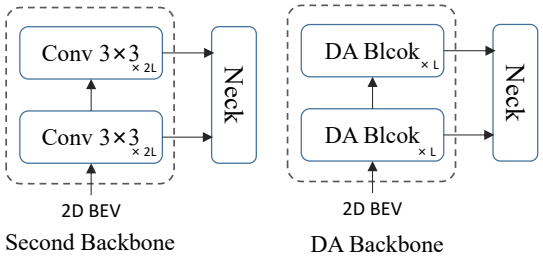


Fig. 10: **Backbone comparison: Second VS Dual-Attention.**

The Fig. 10 shows the structure of the two different backbones used in our experiments, and the reverse ordering of attention blocks is used to achieve better results in both classification and segmentation tasks.

Overall, our results demonstrate the effectiveness of our approach in improving the performance of 3D object detection in challenging real-world scenarios.

V. Conclusion

In conclusion, this manuscript introduces a pioneering approach to automatic image colorization, encapsulated in the novel Multiscale Pyramid Transformer. Central to our method are two key elements: a semantic encoder that discerns the intricate details within images, and a color decoder that adeptly applies vibrant hues. Through rigorous experimentation, we have established that our approach not only meets but surpasses the benchmarks set by existing state-of-the-art techniques, both in terms of quantitative metrics and qualitative assessments. Additionally, the incorporation of a Dual-Attention block signifies our method's adaptability and efficacy across a spectrum of high-level computer vision tasks, including classification, segmentation, and detection, where it delivers exceptional outcomes. We are optimistic about the Multiscale Pyramid Transformer's versatility across a diverse array of low-level tasks and are eager to delve into its further capabilities in upcoming studies.

REFERENCES

[1] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 415–423.

[2] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European conference on computer vision*. Springer, 2016, pp. 649–666.

[3] Y. Wu, X. Wang, Y. Li, H. Zhang, X. Zhao, and Y. Shan, "Towards vivid and diverse image colorization with generative color prior," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 377–14 386.

[4] X. Ji, B. Jiang, D. Luo, G. Tao, W. Chu, Z. Xie, C. Wang, and Y. Tai, "Colorformer: Image colorization via color memory assisted hybrid-attention transformer," in *European Conference on Computer Vision*. Springer, 2022, pp. 20–36.

[5] X. Kang, T. Yang, W. Ouyang, P. Ren, L. Li, and X. Xie, "Ddcolor: Towards photo-realistic image colorization via dual decoders," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 328–338.

[6] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 568–578.

[7] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 22–31.

[8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[9] A. Hatamizadeh, H. Yin, J. Kautz, and P. Molchanov, "Global context vision transformers," *arXiv preprint arXiv:2206.09959*, 2022.

[10] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[11] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, and J. Shlens, "Scaling local self-attention for parameter efficient visual backbones," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 894–12 904.

[12] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxvit: Multi-axis vision transformer," *arXiv preprint arXiv:2204.01697*, 2022.

[13] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, and L. Yuan, "Davit: Dual attention vision transformers," *arXiv preprint arXiv:2204.03645*, 2022.

[14] 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," *arXiv preprint arXiv:1705.02999*, 2017.

[15] J.-W. Su, H.-K. Chu, and J.-B. Huang, "Instance-aware image colorization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7968–7977.

[16] P. Vitoria, L. Raad, and C. Ballester, "Chromagan: Adversarial picture colorization with semantic class distribution," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2445–2454.

[17] J. Antic, "A deep learning based project for colorizing and restoring old images (and video!)," 2019.

[18] M. Kumar, D. Weissenborn, and N. Kalchbrenner, "Colorization transformer," *arXiv preprint arXiv:2102.04432*, 2021.

[19] G. Kim, K. Kang, S. Kim, H. Lee, S. Kim, J. Kim, S.-H. Baek, and S. Cho, "Bigcolor: Colorization using a generative color prior for natural images," *arXiv preprint arXiv:2207.09685*, 2022.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[21] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[23] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 214–223.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[25] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1209–1218.

[26] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[27] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641.

[28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[29] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," in *Human vision and electronic imaging VIII*, vol. 5007. SPIE, 2003, pp. 87–95.

[30] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[31] S. Weng, J. Sun, Y. Li, S. Li, and B. Shi, "Ct2: Colorization transformer via color tokens."

[32] S. Targ, D. Almeida, and K. Lyman, "Resnet in resnet: Generalizing residual architectures," *arXiv preprint arXiv:1603.08029*, 2016.

[33] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357.

[34] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 357–366.

[35] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 558–567.

[36] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3965–3977, 2021.

[37] X. Chu, Z. Tian, B. Zhang, X. Wang, X. Wei, H. Xia, and C. Shen, "Conditional positional encodings for vision transformers," *arXiv preprint arXiv:2102.10882*, 2021.

[38] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 32–42.

[39] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "Metaformer is actually what you need for vision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 819–10 829.

[40] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, and J. Gao, "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2998–3008.

[41] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 908–15 919, 2021.

[42] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022.

[43] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, and J. Gao, "Focal self-attention for local-global interactions in vision transformers," *arXiv preprint arXiv:2107.00641*, 2021.

[44] J.-g. Song, "Ufo-vit: High performance linear vision transformer without softmax," *arXiv preprint arXiv:2109.14382*, 2021.

[45] K. Han, J. Guo, Y. Tang, and Y. Wang, "Pyramidtnt: Improved transformer-in-transformer baselines with pyramid architecture," *arXiv preprint arXiv:2201.00978*, 2022.

[46] A. Brock, S. De, and S. L. Smith, "Characterizing signal propagation to close the performance gap in unnormalized resnets," in *9th International Conference on Learning Representations, ICLR*, 2021.

[47] M. Arar, A. Shamir, and A. H. Bermano, "Learned queries for efficient local attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 841–10 852.

[48] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

[49] Z. Huang, Y. Ben, G. Luo, P. Cheng, G. Yu, and B. Fu, "Shuffle transformer: Rethinking spatial shuffle for vision transformer," *arXiv preprint arXiv:2106.03650*, 2021.
[50] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 124–12 134.
[51] Y. Lee, J. Kim, J. Willette, and S. J. Hwang, "Mpvit: Multi-path vision transformer for dense prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7287–7296.
[52] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," *arXiv preprint arXiv:2202.09741*, 2022.
[53] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "Uniformer: Unifying convolution and self-attention for visual recognition," *arXiv preprint arXiv:2201.09450*, 2022.
[54] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.
[55] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
[56] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4490–4499.
[57] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.
[58] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 784–11 793.
[59] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.
[60] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

VI. Biography Section

Tongtong Zhao Tongtong Zhao, is a Senior data scientist at presight.ai, where she leads a research group specializing in computer vision and machine learning. Her scholarly pursuits are anchored in the design of both efficient and potent deep learning frameworks tailored for real-life deployments, alongside venturing into innovative methodologies aimed at enhancing the clarity and resilience of neural networks. As the primary author of this manuscript, Zhao introduces an avant-garde color decoder, dubbed Color-Attention. This decoder adeptly assimilates colorization motifs from both grayscale and colored images within the prevailing training ensemble, thereby broadening its applicability and obviating the necessity for pre-established color priors.

Gehui Li GeHui Li is currently a student in Dalian University of Technology, specializing in computer vision and deep learning. His research interests include developing novel deep learning architectures for visual recognition tasks and exploring new techniques to enhance the efficiency and interpretability of neural networks. He is the primary author of this paper, which presents an innovative end-to-end automatic colorization method that does not require any color reference images and achieves superior quantitative and qualitative results compared to state-of-the-art methods.. This work highlights the potential of the Dual-Attention block as a general-purpose vision module with strong modeling capabilities, providing a promising avenue for future developments in the field of computer vision.

Shanshan Zhao Shanshan Zhao serving as a distinguished data engineer at the Inception Institute of Artificial Intelligence, spearheads a pioneering research group dedicated to the realms of computer vision and machine learning. Her research interests include developing efficient and effective deep learning algorithms for real-world applications and exploring new techniques to improve the interpretability and robustness of neural networks. She is a primary author of this paper, which presents a new color decoder called Color-Attention, which learns colorization patterns from grayscale images and color images of the current training set, resulting in improved generalizability and eliminating the need for constructing color priors. This work showcases the potential of attention mechanisms as a solution to the challenge of global information acquisition and highlights the effectiveness of the proposed model in addressing computational efficiency concerns while maintaining high accuracy.