STACKED MULTI-SCALE ATTENTION NETWORK FOR IMAGE COLORIZATION

Bin Jiang[*†], *Fangqiang Xu*[†], *Jun Xia*[†], *Chao Yang, Wei Huang, Yun Huang*
College of Computer Science and Electronic Engineering
Hunan University, Changsha, China

ABSTRACT

Deep convolutional networks (CNNs) show their potential in image colorization for producing plausible results. Recently, the attention mechanism further boosts the performances of CNNs by constructing channel and spatial interactions. However, existing attention methods are performed in a single-scale manner, which is hard to capture multi-scale information interactions in a limited computational cost. This can limit the performance of the network to reconstruct color channels. In this paper, we propose a stacked multi-scale attention network (SMSANet) for image colorization. The core idea is to perform the attentions of a feature map in a multi-scale manner so that sufficient interactions are conducted to efficiently and adaptively capture multi-scale and long-range dependencies. By stacking the multi-scale attention layers in different convolutional layers, the SMSANet can focus on more discriminative features to reconstruct color channels. Moreover, a salient loss is designed to further refine the generated image at both pixel-level and object-level. Extensive experiments on the ImageNet dataset have demonstrated that SMSANet outperforms the state-of-the-art automatic colorization methods.

Index Terms— Stacked Mluti-Scale Attention, Automatic Image Colorization, Salient loss

1. INTRODUCTION

Automatic image colorization, whose goal is to transform a gray image into a kind of colorful style, is an active research topic in computer vision recently. At the emergence of deep learning and large-scale datasets, automatic image colorization methods [1, 2, 3, 4] using convolutional networks (CNNs) have achieved significant successes. Since they do not require any user interventions, such as color scribbles [5, 6, 7, 8] and reference images [9, 10, 11, 12], they become a recent trend.

Most of the automatic colorization methods construct autoencoder-type models to reconstruct color images, such as utilizing the pre-trained VGG [13] network as the encoder and building the corresponding decoder to compose the final model. Besides, some recent methods pay attention to the details of the generated images. For example, Xiao *et al.* [14] and Xia *et al.* [15] propose edge loss to constrain the edge of the generated image, while Su *et al.* [16] adopt the results of object detection to extract object-level features. Although significant progress have been made in image colorization via deep learning, it is still extremely crucial and challenging to design an elegant architecture to collect more diverse and discriminative features for specific information reconstruction.

We observe that: 1) image features are treated with equal importance, which makes the training of the network difficult and ultimately leads to poor image coloring. Attention mechanism may alleviate this problem, but most existing methods are performed in a single-scale manner, resulting in insufficient feature interactions. This motivates us to establish a more suitable attention mechanism. 2) usual objective functions treat the errors of all pixels equally, which weaken foreground objects and objects with rich textures. This motivates us to build a loss function more focused on regions with rich textures and foreground objects.

In this paper, we propose a novel stacked multi-scale attention network (SMSANet) for automatic image colorization. Specifically, we propose a stacked multi-scale attention (SMSA) mechanism for finely learning more discriminative features for diverse scenarios. The core idea is to perform attentions of a feature map in a multi-scale manner, so that sufficient interactions are conducted to efficiently and adaptively capture multi-scale and long-range dependencies. For a given feature, we first adopt a pooling operator to aggregate the multi-scale features like CBAM [17] and SENet [18]. Next, we construct different scales by splitting features into some patches with different scaling parameters. Then multi-scale attention map can be easily obtained by convolution or fully connected (FC) operations. Note that a sequence of learnable weights is presented to estimate the importance of each attention maps. Different from existing works, SMSA takes a variety of feature scales into account to obtain more discriminative features for diverse scenarios. Moreover, the coloring results of the image foreground are particularly important, so we design an efficient and effective salient loss to regularize the details of the results from both pixel and object levels.

The main contributions of this paper are summarized as follow:

- We propose a novel automatic image colorization framework named stacked multi-scale attention network (SMSANet). To our best knowledge, it is the first attempt to establish a multi-scale attention mechanism in automatic image colorization.
- A Stacked Multi-Scale Attention mechanism is proposed to obtain distinctive and specific features from multiple scales for various images. Since SMSA integrates multiple information from multiple feature scales instead of a single scale, it is more comprehensive and reasonable than existing attention mechanisms.
- We present a salient loss to regularize object-level image details for further performance improvement.
- Extensive experimental results on ImageNet dataset have demonstrated that the proposed SMSA achieves better performance in comparison with state-of-the-art methods.

[*]This work was supported in part by the National Natural Science Foundation of China under grant 62072169 and 62172156, and the National Key Research and Development Program of China under grant 2020YFB1713003.

[†]These authors contributed equally to this work and share first authorship.

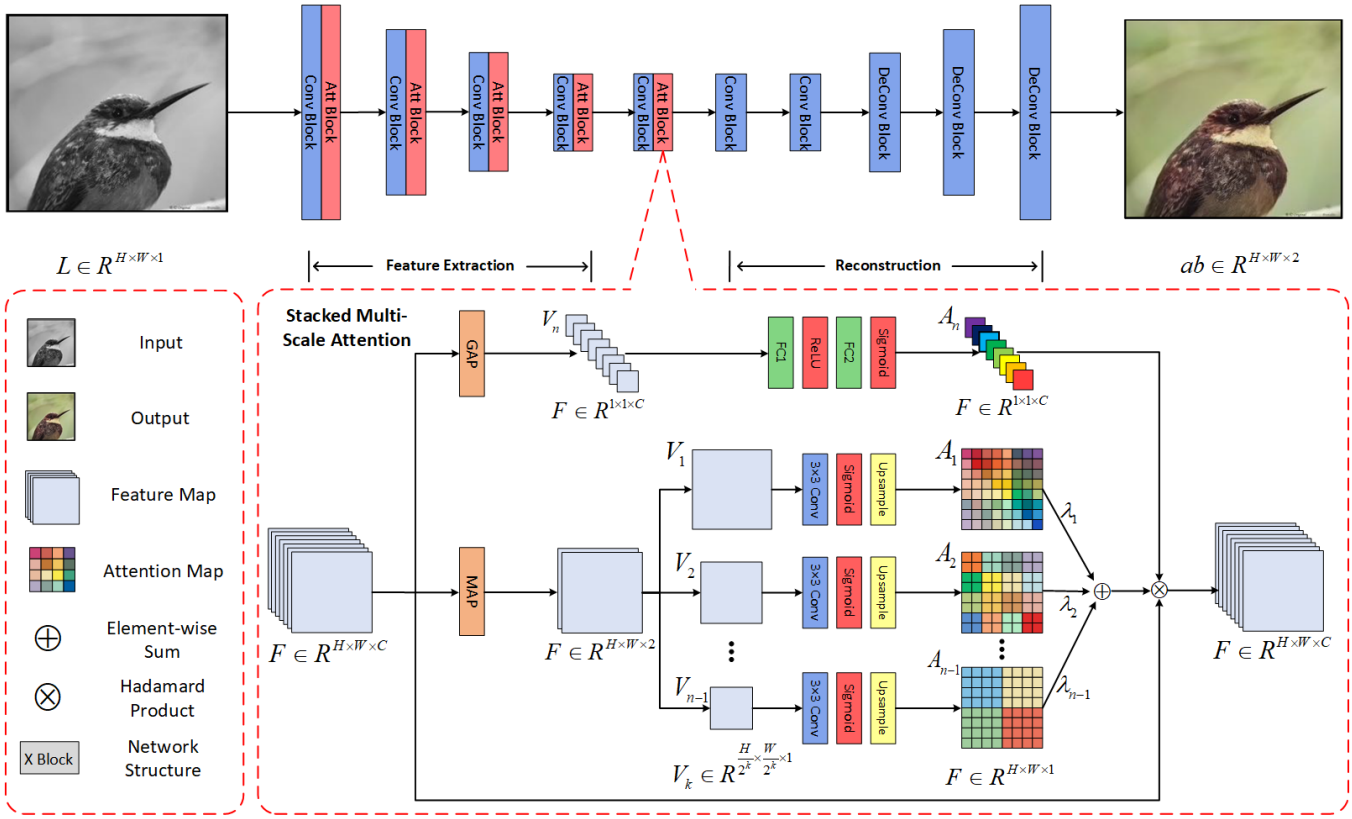

Fig. 1. The architecture of the proposed SMSANet approach, which contains feature extraction module, stacked multi-scale attention module and reconstruction module. Specially, the same color represents the same attention weight in the attention map.

2. METHOD

2.1. Overview

Our goal is to obtain a plausible color image from a grayscale image. Specifically, our model takes in a grayscale image $L \in R^{H\times W\times 1}$ (in CIE Lab color space) as input and generates two corresponding channels $ab \in R^{H\times W\times 2}$ (in CIE Lab color space). As depicted in Fig. 1, the network consists of three main modules: feature extraction module, stacked multi-scale attention module and reconstruction module. First, we leverage the feature extraction module to progressively extract higher-level semantic information from the given grayscale image. Next, the stacked multi-scale attention module is designed to extract discriminative information for multiple scenarios. Finally, the function of reconstruction module is gradually mapping the features into ab channels. Note that the feature extraction module and reconstruction module are based on Zhang *et al.* [19]. Moreover, all frameworks are optimized by two loss objective functions which more details are described in Section 2.3.

2.2. Stacked Multi-Scale Attention

The Stacked Multi-Scale Attention mechanism is the core component in our model. Existing attention mechanisms [18, 17, 20, 21] can be roughly divided into two categories: spatial attention and channel attention. Specifically, spatial attention is designed to explore inter-positions dependencies, which treats each position information as separate. Channel attention is designed to explore inter-channel dependencies, which treats all position information as a whole. However, both spatial and channel attention maps are obtained from a single scale without interactive information from multiple scales, leading to obvious limitations for diverse scenarios. Thus, we propose stacked multi-scale attention mechanism, which aims to obtain more discriminative features from multiple feature scales. The SMSA is illustrated in Fig. 1, given a feature map $F \in R^{H\times W\times C}$, where H, W and C denote the length of height, width and channel, respectively. First, we convert the given features to multiple scales, which is defined as follows:

$$V_k = \begin{cases} Split(MAP(F)), & k = 1, 2, \cdots n-1 \\ GAP(F), & k = n \end{cases} \tag{1}$$

where $V_k (1 \leq k \leq n)$ means the aggregated feature in different scales, $MAP(\cdot)$, $GAP(\cdot)$ and $Split(\cdot)$ denote channel max- and average-pooling, global aaverage-pooling and split feature operations. Note that the above process divides the feature map $F \in R^{H\times W\times C}$ into multiple spatial scales $[V_1, V_2, \cdots V_k] \in R^{\frac{H}{2^k}\times\frac{W}{2^k}\times 2}$ along the spatial dimension, and divides the feature map $F \in R^{H\times W\times C}$ into channel scale $V_n \in R^{1\times 1\times C}$ along the channel dimension.

Next, we obtain multi-scale attention maps $[A_1, A_2, \cdots A_n]$ from the aggregated feature scales. Notably, we define A_n as the channel scale attention, which is designed to obtain the inter-channel relationships in the channel branch. We define $A_1 - A_n - 1$ as the spatial scale attentions, which is presented to obtain the

inter-position relationships in the spatial branch. In summary, the multi-scale attention is computed as:

$$A_k = \begin{cases} Up(\sigma(f^{3\times3}(V_k))), & k = 1, 2, \cdots n-1 \\ \sigma(W_2 \cdot ReLU(W_1 \cdot V_k)), & k = n \end{cases} \tag{2}$$

where A_k means the $k-th$ scale attention map, σ, $ReLU$ and Up refer to Sigmoid [22], ReLU [23] and up-sample operation, respectively. W_1 and W_2 denote the learnable weight matrices of two fully connected layers, while $f^{3\times3}$ denotes a convolution operation with the kernel size of 3×3. In addition, we adaptively select a sequence of learnable weights for estimating the importance of each attention map with the following function:

$$A_s = \lambda_1 \cdot A_1 + \lambda_2 \cdot A_2 + \cdots + \lambda_{n-1} \cdot A_{n-1} \tag{3}$$

where $\lambda_i \in (0, 1)$ is a series of learnable parameters, A_k is a series of attention maps. Finally, we can refine the given feature like the process in CBAM [17], which is calculated as follows:

$$F' = F \oplus (F \otimes A_s \otimes A_n) \tag{4}$$

where A_s and A_n define the aggregated spatial-scale and channel-scale attention map, $F \in R^{H\times W\times C}$ and $F' \in R^{H\times W\times C}$ are the input and refined feature map, respectively. Moreover, $\oplus$ and $\otimes$ denote element-wise sum and Hadamard product operation.

2.3. Objective Function

Huber loss: Similar to Zhang *et al.* [19], we adopt the Huber loss function with δ = 1 to optimize the whole network, which is defined as follows:

$$\mathcal{L}_{huber}(x, y) = \begin{cases} \frac{1}{CHW} \left\| \frac{1}{2}(x-y)^2 \right\|, & |x-y| < \delta \\ \frac{1}{CHW} \|x-y\| - \frac{1}{2}\delta^2, & |x-y| \geq \delta \end{cases} \tag{5}$$

where x and y denote ground truth image and generated color image, respectively. Note that Huber loss is a loss function with stable and fast training speed and insensitive to outliers.

Salient loss: Although Huber loss is a robust regression loss function, it ignore the object-level details of the image. Inspired by Johnson *et al.* [24] and Xiao *et al.* [14], we propose a novel salient loss that considers both pixel-level and object-level information. Note that the salient loss is calculated by an off-the-shelf salient detection network PFANet [25]. In summary, the salient loss is defined as:

$$\mathcal{L}_{sal}(x, y) = \frac{1}{CHW} \|\phi_j(x) \cdot x - \phi_j(y) \cdot y\| \tag{6}$$

where x and y denote ground truth image and generated color image separately, ϕ is the off-the-shelf salient detection network PFANet [25] and $\phi(\cdot)$ denotes the foreground object binary map generated from PFANet. Compared with existing works, the proposed salient loss improve the foreground coloring results of the generated iamge while keeping negligible parameters and complexity increases.

Total loss: The final loss function formula is balanced by the hyper-parameters α_1 = 1 and is calculated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{huber} + \alpha_1 \mathcal{L}_{sal} \tag{7}$$

where $\mathcal{L}_{total}$ indicates the final loss function which is applied to optimize the whole network. $\mathcal{L}_{huber}$ and $\mathcal{L}_{sal}$ are two loss functions mentioned above.

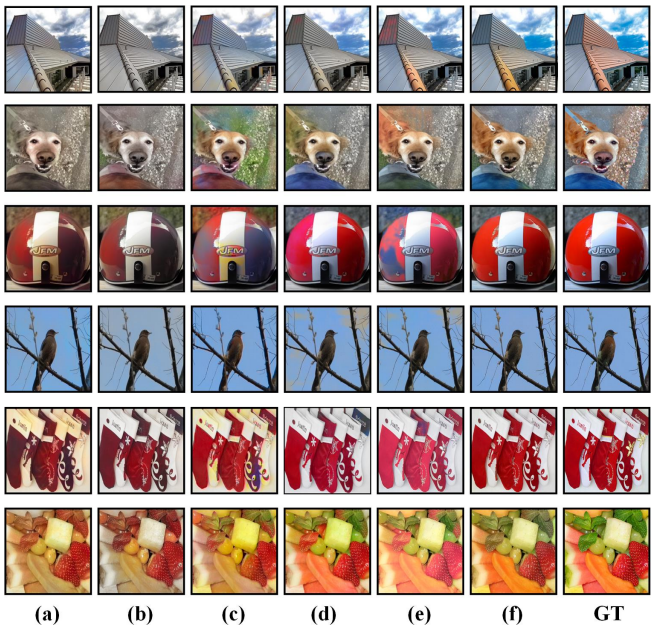

Fig. 2. For the given grayscale image, there are the colorization results of the state-of-the-art methods and the proposed method. (a) Iizuka *et al.* [29] (b) Larsson *et al.* [30] (c) Zhang *et al.* [31] (d) Zhang *et al.* [19] (e) Su *et al.* [16] (f) SMSANet (Ours) and (GT) ground truth image. Note that the three upper rows and three bottom rows denote the results of single-object and multi-object images, respectively.

3. EXPERIMENT

3.1. Dataset

We systemically evaluate the proposed SMSANet on ImageNet [26] dataset, which has been widely used in image colorization. Specifically, ImageNet is a large-scale image dataset with 1.2M images, which contains 1000 different kinds of images. We train the model on the original training split and randomly select 1000 images to be test set from the original test split.

3.2. Experiment settings

The training progress stops when iterations are up to 40K and all image is transformer into CIE Lab color space with 256×256 sizes. Moreover, our framework is optimized by Adam [27] with a learning rate of 1e-5 and momentum parameters β_1 0.9 and β_2 = 0.999. All of the experiments are implemented on Intel Core™ i5-8600K CPU@3.6GHz, 64GB, NVIDIA GeForce GTX 1080Ti 11GB GPU with Python 3.6.0 and Pytorch 1.6.0 [28]. Codes and models will be available at: https://github.com/LionXFQ/SMSANet.

3.3. Comparison with the State-of-the-art Algorithms

In this part, we evaluate the proposed method against the following state-of-the-art automatic image colorization approaches Iizuka *et al.* [29], Larsson *et al.* [30], Zhang *et al.* [31], Zhang *et al.* [19] and Su *et al.* [16].

The visual results of different approaches on ImageNet dataset are shown in Fig.2. From Fig. 2 (a and c), we observe that both of

Table 1. Comparison results of average PSNR and SSIM on ImageNet dataset. The bold values indicate the best results.

Method	PSNR ↑	SSIM ↓
Iizuka *et al.* [29]	23.2727	0.9000
Larsson *et al.* [30]	24.6277	0.9095
Zhang *et al.* [31]	21.5602	0.8757
Zhang *et al.* [19]	25.7688	0.9136
Su *et al.* [16]	26.0771	0.9259
Xia *et al.* [15]	26.3200	0.9090
SMSANet (Ours)	**26.6579**	**0.9301**

them suffer from some color distortion where the results look unrealistic. The results of Larsson *et al.* [30] are illustrated in Fig. 2 (b), which are grayer than other approaches in some cases. In addition, there are still some artifacts in the results of Zhang *et al.* [31], Zhang *et al.* [19] and Su *et al.* [16] as shown in Fig. 2 (c-e). Compared with these methods, the results of our algorithm (Fig. 2 (f)) contain sharper colors and details without obvious artifacts, which are closer to the ground truth. Moreover, we observe that the proposed method obtains not only good performance on single target images (three upper rows), but also achieves convincing results on multi-target images (three bottom rows). This is because SMSANet not only achieves SMSA to extract discriminative information but also adopts a salient loss to focus on image object-level details. Furthermore, we give the quantitative comparison with five approaches in Table 1. Compared with the state-of-the-art approaches, the proposed method achieves competitive results in two widely used evaluation metrics: PSNR and SSIM. Specifically, SMSANet obtains 26.6579 dB and 0.9301 in terms of PSNR and SSIM on the ImageNet dataset, respectively. In summary, the visual and evaluation results on the ImageNet dataset demonstrate the effectiveness of the proposed algorithm for automatic image colorization.

3.4. Ablation study

In order to evaluate the effectiveness of SMSA and salient loss, we carry out a set of ablation studies on ImageNet dataset. Here the BaseNet refers to a united framework proposed in Zhang *et al.* [19]. We firstly compare SMSA with CBAM [17] and SENet [18], which are two popular attention mechanisms. As shown in Fig. 3, we observe that SMSA performs well in multiple scenarios, while CBAM and SENet produce artifacts in some cases, especially in complex scenes images. Table 2 summarizes the numerical comparison with CBAM and SENet, our method represents a relative increase of 0.2991dB, 0.2186dB, 0.0038 and 0.0029 in PSNR and SSIM. The above experimental results demonstrate that the proposed SMSA is able to alleviate the artifacts problem and more suitable for image colorization than the current attention mechanism.

Furthermore, we conduct several experiments to measure the performance of the salient loss. Note that the baseline BaseNet refers to a united framework proposed in Zhang *et al.* [19] and SMSANet is made up of SMSA embedded into BaseNet. As reported in Table 3, salient loss can effectively boost the performances of both BaseNet and SMSANet, which verifies that the proposed salient loss is effective in colorization task.

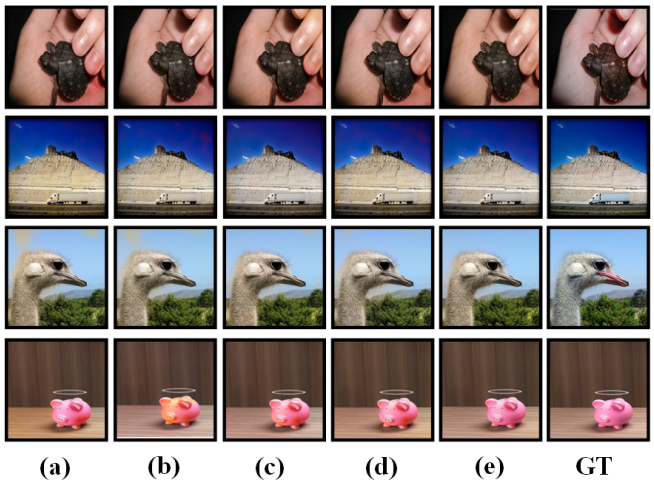

Fig. 3. Ablation study results of different attention mechanism. (a) BaseNet+CBAM, (b) BaseNet+SENet, (c) BaseNet+SMSA (Ours), (d) BaseNet+Salient loss (Ours), (e) Full SMSANet (Ours), (GT) ground truth image.

Table 2. Ablation study of existing attention mechanism and the proposed SMSA. Notably, BaseNet is based on Zhang *et al.* [19].

Method	Attention	PSNR ↑	SSIM ↓
BaseNet [19]	CBAM [17]	26.2487	0.9147
BaseNet [19]	SENet [18]	26.3292	0.9156
BaseNet [19]	SMSA	**26.5478**	**0.9185**

Table 3. Ablation study of huber loss and salient loss.

Method	Loss	PSNR ↑	SSIM ↓
BaseNet [19]	Huber	25.7688	0.9136
SMSANet	Huber	26.5478	0.9185
BaseNet [19]	Huber+Salient	26.5411	0.9294
SMSANet	Huber+Salient	**26.6579**	**0.9301**

4. CONCLUSION

In this paper, we propose a novel stacked multi-scale attention network (SMSANet) for automatic image colorization. The core component SMSA obtains comprehensive and discriminative information via a stacked multi-scale attention mechanism, which effectively provides precise network training guidance and significantly alleviate the artifacts problem. In addition, the proposed salient loss is able to assist in improving the performance of the proposed network in object-level details. Extensive experiments on ImageNet dataset have demonstrated that the proposed method outperforms the state-of-the-art deep colorization methods. In the future, we plan to apply our method to other colorization fields such as: video colorization [32], user-guide [33] or exemplar-based colorization [34].

5. REFERENCES

[1] Zezhou Cheng, Qingxiong Yang, and Bin Sheng, "Deep colorization," in *ICCV*, 2015, pp. 415–423.

[2] Aditya Deshpande, Jason Rock, and David A. Forsyth, "Learning large-scale automatic image colorization," in *ICCV*, 2015, pp. 567–575.

[3] Jiaojiao Zhao, Jungong Han, Ling Shao, and Cees G. M. Snoek, "Pixelated semantic colorization," *Int. J. Comput. Vis.*, vol. 128, no. 4, pp. 818–834, 2020.

[4] Mohammad Mahdi Johari and Hamid Behroozi, "Gray-scale image colorization using cycle-consistent generative adversarial networks with residual structure enhancer," in *ICASSP*, 2020, pp. 2223–2227.

[5] Anat Levin, Dani Lischinski, and Yair Weiss, "Colorization using optimization," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 689–694, 2004.

[6] Yi-Chin Huang, Yi-Shin Tung, Jun-Cheng Chen, Sung-Wen Wang, and Ja-Ling Wu, "An adaptive edge detection based colorization algorithm and its applications," in *ACM*, 2005, pp. 351–354.

[7] Yingge Qu, Tien-Tsin Wong, and Pheng-Ann Heng, "Manga colorization," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 1214–1220, 2006.

[8] Liron Yatziv and Guillermo Sapiro, "Fast image and video colorization using chrominance blending," *IEEE Trans. Image Process.*, vol. 15, no. 5, pp. 1120–1129, 2006.

[9] Tomihisa Welsh, Michael Ashikhmin, and Klaus Mueller, "Transferring color to greyscale images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 277–280, 2002.

[10] Revital Ironi, Daniel Cohen-Or, and Dani Lischinski, "Colorization by example," in *Rendering Techniques*, 2005, pp. 201–210.

[11] Yu-Wing Tai, Jiaya Jia, and Chi-Keung Tang, "Local color transfer via probabilistic segmentation by expectation-maximization," in *CVPR*, 2005, pp. 747–754.

[12] Guillaume Charpiat, Matthias Hofmann, and Bernhard Schölkopf, "Automatic image colorization via multimodal predictions," in *ECCV*, 2008, pp. 126–139.

[13] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[14] Yi Xiao, Peiyao Zhou, Yan Zheng, and Chi-Sing Leung, "Interactive deep colorization using simultaneous global and local inputs," in *ICASSP*, 2019, pp. 1887–1891.

[15] Jun Xia, Guanghua Tan, Yi Xiao, Fangqiang Xu, and Chi-Sing Leung, "Edge-aware multi-scale progressive colorization," in *ICASSP*, 2021, pp. 1655–1659.

[16] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang, "Instance-aware image colorization," in *CVPR*, 2020, pp. 7965–7974.

[17] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "CBAM: convolutional block attention module," in *ECCV*, 2018, pp. 3–19.

[18] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.

[19] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S. Lin, Tianhe Yu, and Alexei A. Efros, "Real-time user-guided image colorization with learned deep priors," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 119:1–119:11, 2017.

[20] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua, "SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning," in *CVPR*, 2017, pp. 6298–6306.

[21] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *CVPR*, 2020, pp. 11531–11539.

[22] George Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of control, signals and systems*, vol. 2, no. 4, pp. 303–314, 1989.

[23] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[24] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016, pp. 694–711.

[25] Ting Zhao and Xiangqian Wu, "Pyramid feature attention network for saliency detection," in *CVPR*, 2019, pp. 3085–3094.

[26] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[27] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[28] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in pytorch," 2017.

[29] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi 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.*, vol. 35, no. 4, pp. 110:1–110:11, 2016.

[30] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich, "Learning representations for automatic colorization," in *ECCV*, 2016, pp. 577–593.

[31] Richard Zhang, Phillip Isola, and Alexei A. Efros, "Colorful image colorization," in *ECCV*, 2016, vol. 9907, pp. 649–666.

[32] Chenyang Lei and Qifeng Chen, "Fully automatic video colorization with self-regularization and diversity," in *CVPR*, 2019, pp. 3753–3761.

[33] Lvmin Zhang, Chengze Li, Tien-Tsin Wong, Yi Ji, and Chunping Liu, "Two-stage sketch colorization," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 261:1–261:14, 2018.

[34] Zhongyou Xu, Tingting Wang, Faming Fang, Yun Sheng, and Guixu Zhang, "Stylization-based architecture for fast deep exemplar colorization," in *CVPR*, 2020, pp. 9360–9369.