Supplementary Material to "ColorFormer: Image Colorization via Color Memory assisted Hybrid-attention Transformer"

Xiaozhong Ji[1*], Boyuan Jiang[1*], Donghao Luo[1], Guangpin Tao[1], Wenqing Chu[1], Zhifeng Xie[2], Chengjie Wang[1†], Ying Tai[1†]

[1] Youtu Lab, Tencent
{xiaozhongji,byronjiang,michaelluo,
guangpintao,wenqingchu,jasoncjwang,yingtai}@tencent.com
[2] Shanghai University
zhifeng_xie@shu.edu.cn

In the supplementary, we provide the following materials:

- Network architecture of ColorFormer.
- Detail and more analysis of Color Memory.
- More visual results on the three benchmarks.
- Discussion about diverse colorization.

1 Network Architecture

We list the detailed architecture of our ColorFormer in Table 1, where an input image size of 256×256 is assumed. For Stage1 to Stage4, "Concat $n \times n$" indicates a concatenation of n $\times$ n neighboring features in a patch. This operation results in a downsampling of the feature map by a rate of n. "96-d" denotes a linear layer with an output dimension of 96. "win. sz. 7×7" indicates a multi-head self-attention module with window size of 7×7. "[ ]$\times$2" means a GLH-Transformer block consisting of a GL-MSA and a SW-MSA. For Stage5 to Stage7, we merge features from corresponding encoder stage and upscale feature map with PixelShuffle operations. Stage8 is our proposed Color Memory module, which stores color priors to enhance features. Stage9 is used to refine features and generate ab maps.

2 Detail and Analysis of CM

Implementation Detail of Memory Build. The detailed building process is described in Algorithm 1.

Ablation study of CM at different positions. We conduct ablation study on inserting CM after each decoder stage. As shown in Table 2, inserting CM at the last stage achieves better performance than at the early stages.

* Equal contribution.
† Corresponding authors.

	Output Size	ColorFormer
Stage1	$64 \times 64 \times 96$	Concat 4x4, 96-d, LN $[\frac{\text{Win.SZ. } 7\times7}{\text{dim } 96,\text{head } 3}] \times 2$
Stage2	$32 \times 32 \times 192$	Concat 2x2, 192-d, LN $[\frac{\text{Win.SZ. } 7\times7}{\text{dim } 192,\text{head } 6}] \times 2$
Stage3	$16 \times 16 \times 384$	Concat 2x2, 384-d, LN $[\frac{\text{Win.SZ. } 7\times7}{\text{dim } 384,\text{head } 12}] \times 6$
Stage4	$8 \times 8 \times 768$	Concat 2x2, 768-d, LN $[\frac{\text{Win.SZ. } 7\times7}{\text{dim } 768,\text{head } 24}] \times 2$
Stage5	$16 \times 16 \times 512$	PixelShuffle, scale 2 Concat feat. from Stage3
Stage6	$32 \times 32 \times 512$	PixelShuffle, scale 2 Concat feat. from Stage2
Stage7	$64 \times 64 \times 256$	PixelShuffle, scale 2 Concat feat. from Stage1
Stage8	$64 \times 64 \times 256$	Color Memory, group 4
Stage9	$256 \times 256 \times 2$	PixelShuffle, scale 4 Concat input Residual Conv. KS. 3×3 Output Conv. KS. 3×3

Table 1: **Details of ColorFormer architecture.**

Position	$\frac{H}{16} \times \frac{W}{16}$	$\frac{H}{8} \times \frac{W}{8}$	$\frac{H}{4} \times \frac{W}{4}$
FID↓	1.97	2.04	**1.71**
CF↑	37.66	37.75	**39.76**

Table 2: **Ablation study of CM at different positions.**

Qualitative comparison of CM module with different numbers of groups We provide the qualitative comparison of CM with different numbers of groups in Figure 1.

Analysis of Multiple Color Priors. To inspect the effect of different groups of color priors, we adjust the weights of the fusion process using only one group. To further analyse the fusion weights, we display the results and the corresponding weights together. As shown in Figure 2, the multiple groups of color priors help produce diverse colorful images. Furthermore, the corresponding weights reflect the relationship between the single-group results and the fused-groups results.

Algorithm 1 Process of Memory Build

Require: Colorful images set $\mathcal{X}$, pre-trained model $\mathrm{M}(\cdot)$, output feature map size p
Ensure: *keys*: $\mathbf{S} \in \mathbb{R}^{m\times k}$, n groups of *values*: $\mathbf{C}_1$, $\mathbf{C}_2$, ..., $\mathbf{C}_n \in \mathbb{R}^{m\times 2}$
Initialize feature list $\mathcal{S} = []$, color list $\mathcal{C} = []$
for all $\mathbf{I}$ such that $\mathbf{I} \in \mathcal{X}$ **do**
  $\mathbf{s} = \mathrm{M}(\mathbf{I})$, where $\mathbf{s} \in \mathbb{R}^{p\times p\times c}$
  Add $\mathbf{s}_{i,j} \in \mathbb{R}^c$ to $\mathcal{S}$, where $1 \le i, j \le p$
  Resize $\mathbf{I}$ to $p \times p$ and convert it into CIELAB color space
  Extract the ab values, denoted as $\mathbf{c} \in \mathbb{R}^{p\times p\times 2}$
  Add $\mathbf{c}_{i,j}$ to $\mathcal{C}$, where $1 \le i, j \le p$
end for
Perform PCA on $\mathcal{S}$ to reduce the dimension from c to k
Perform K-means clustering on $\mathcal{S}$ to get centers $s_1, s_2, \ldots, s_m$ and label $\mathcal{Y}$
$\mathbf{S} = [s_1, s_2, \ldots, s_m] \in \mathbb{R}^{m\times k}$
Divide $\mathcal{C}$ into $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_m$ based on $\mathcal{Y}$
for all $\mathcal{C}_i$ **do**
  Clustering $\mathcal{C}_i$ and sort the centers $c_{i1}, c_{i2}, \ldots, c_{in}$ by K-means
end for
$\mathbf{C}_j = [c_{1j}, c_{2j}, \ldots, c_{mj}] \in \mathbb{R}^{m\times 2}$, $j \in 1, 2, \ldots, n$
return $\mathbf{S}$, [$\mathbf{C}_1$, $\mathbf{C}_2$, ..., $\mathbf{C}_n$]

Fig. 1: **Qualitative comparison of CM with one/four groups.**

3 Visual Results

Colorfulness Outlier. We notice that ColorFormer does not achieve high Colorfulness (CF) [3] score on CelebA-HQ [4] datasets compared to ChromaGAN [6] and ColTran [5]. The reason is that CF is not the higher the better for human face colorization, therefore the scores that are too higher than Ground Truths should be considered as outliers. We display visual results on CelebA-HQ in Figure 3, as well as the CF scores of each image. Obviously, the results with extremely high CF show poor visual quality and our results look more better.

More Results. Here, we display more visual results of ImageNet validation [2] in Figure 4, and COCO-Stuff [1] in Figure 5. Since Wu *et al.* [7] didn't release the results of COCO-Stuff, we do not include their results.

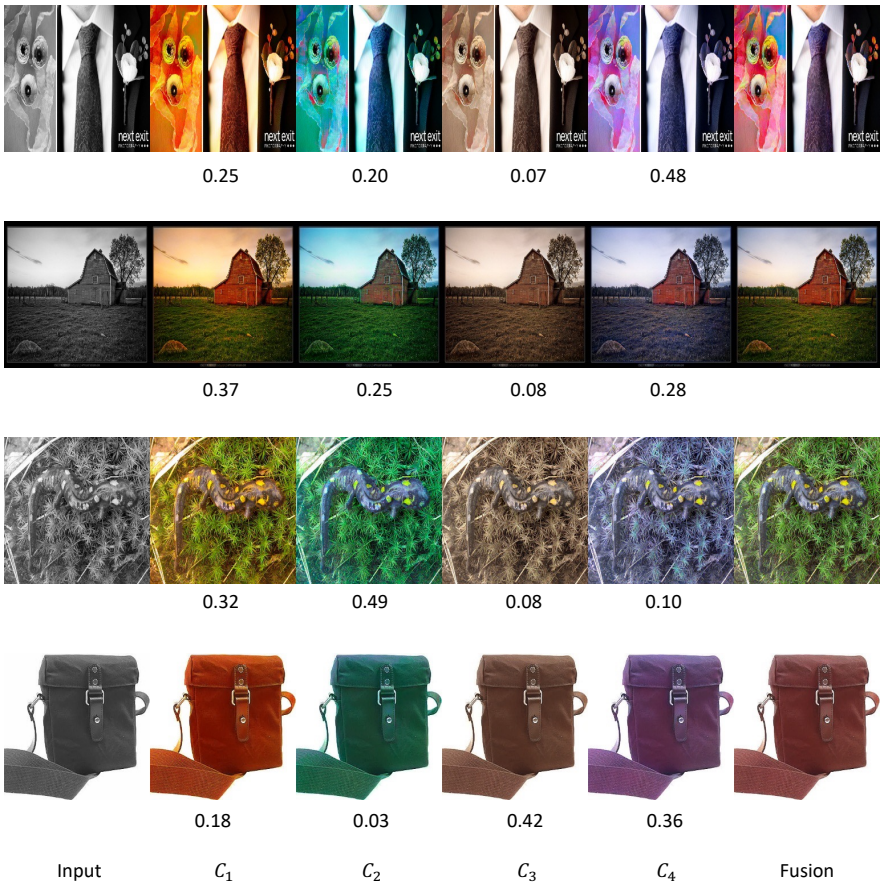

Fig. 2: **Images with different single group of color priors.** The leftmost image is the gray input and the rightmost image is the results of our model. The numbers below the images indicate fusion weights out from the encoder.

4 Discussion

Diverse colorization. Our method can be modified to sample the proportion of different color prior groups (*i.e.*, $\lambda_1, \ldots, \lambda_n$ in Sec. 3.3), which is now determined by the encoder output weights. To demonstrate this potential, we modify the inference model to use random sampling instead of the encoder output to generate images, as shown in Figure 6.

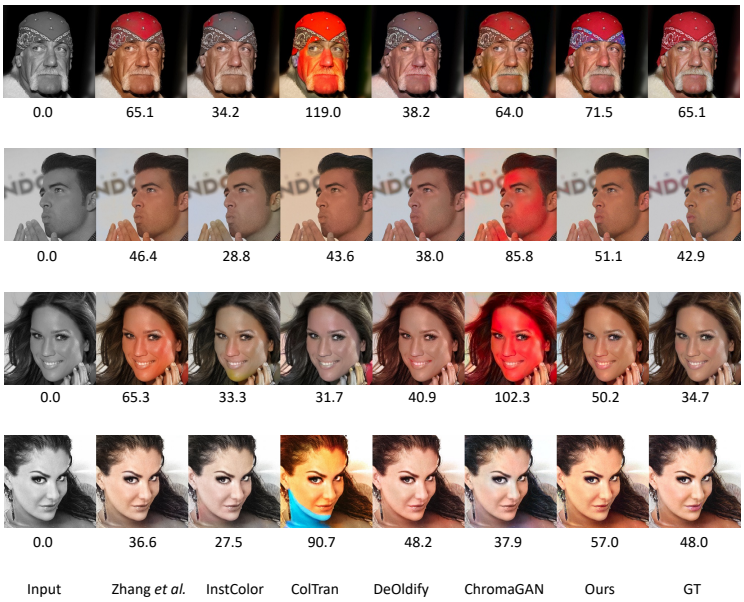

Fig. 3: **More visual comparisons with previous automatic colorization methods on CelebA-HQ.**

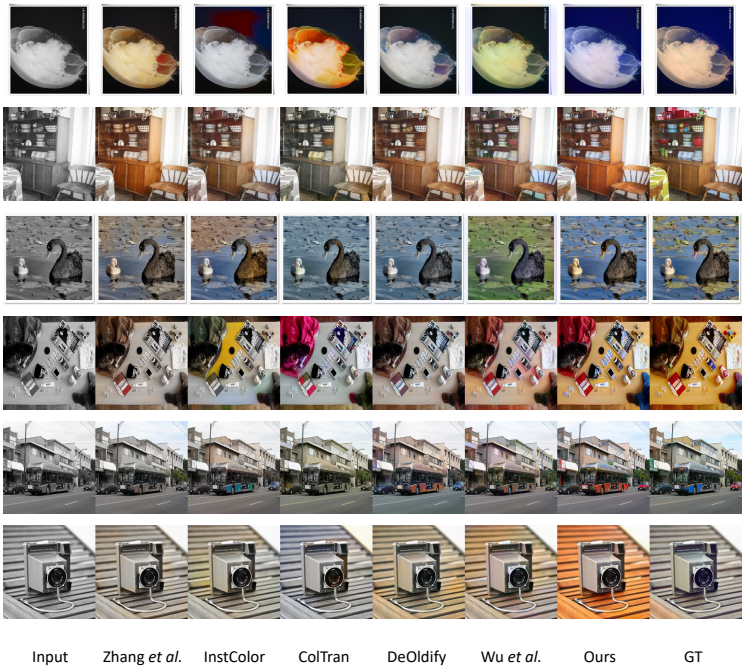

Fig. 4: **More visual comparisons with previous automatic colorization methods on ImageNet.**

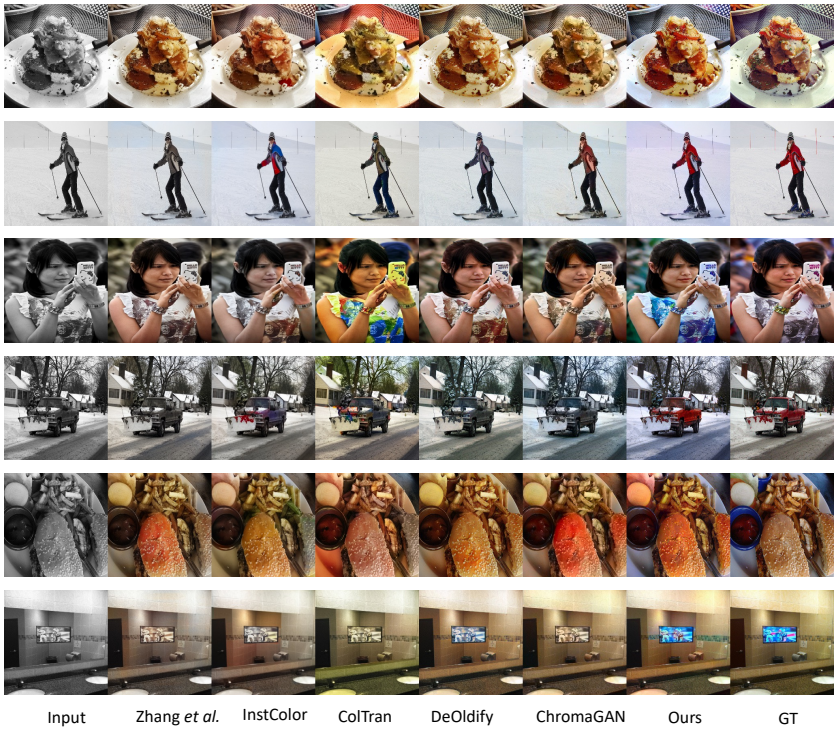

Fig. 5: **More visual comparisons with previous automatic colorization methods on COCO-Stuff.**

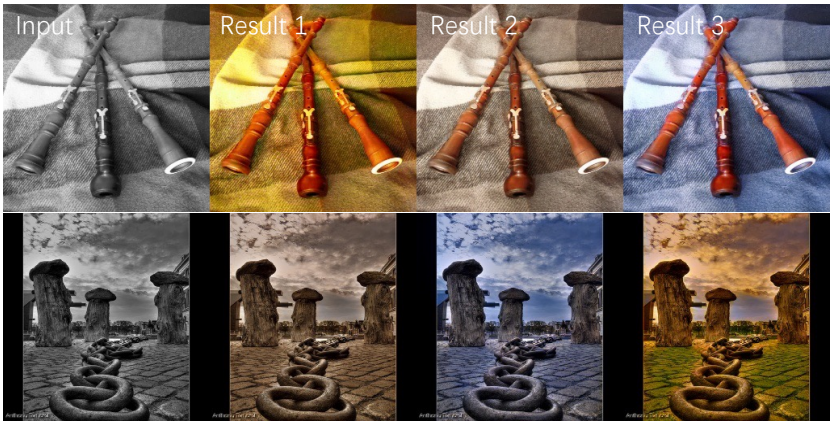

Fig. 6: **Diverse colorization results for a single input.**

References

1. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018) 3
2. Deng, J.: A large-scale hierarchical image database. Proc. of IEEE Computer Vision and Pattern Recognition, 2009 (2009) 3
3. Hasler, D., Suesstrunk, S.E.: Measuring colorfulness in natural images. In: Human vision and electronic imaging VIII. vol. 5007, pp. 87–95. International Society for Optics and Photonics (2003) 3
4. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: International Conference on Learning Representations (2018) 3
5. Kumar, M., Weissenborn, D., Kalchbrenner, N.: Colorization transformer. In: International Conference on Learning Representations (2021) 3
6. Vitoria, P., Raad, L., Ballester, C.: Chromagan: Adversarial picture colorization with semantic class distribution. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2445–2454 (2020) 3
7. Wu, Y., Wang, X., Li, Y., Zhang, H., Zhao, X., Shan, Y.: Towards vivid and diverse image colorization with generative color prior. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14377–14386 (2021) 3