Supplementary Materials for "DDColor: Towards Photo-Realistic Image Colorization via Dual Decoders"

Xiaoyang Kang   Tao Yang   Wenqi Ouyang   Peiran Ren   Lingzhi Li   Xuansong Xie
DAMO Academy, Alibaba Group

In this supplementary document, we provide the following materials to complement the main manuscript:

- Detailed network architecture of DDColor;
- Additional qualitative results;
- Additional ablation study and visual results;
- Runtime analysis;
- More results on legacy black and white photos.

1. Detailed Network Architecture

We list the detailed architecture of DDColor with a ConvNeXt-T[5] backbone in Table 1. The resolution of the input image is 256×256.

2. Additional Qualitative Results

Here, we show more qualitative comparisons with previous methods on ImageNet[6] validation in Figure 3. As in the main paper, we compare our method with DeOldify [1], Wu *et al*. [8], ColTran [4], CT2 [7], BigColor [3] and ColorFormer [9]. The visual comparisons on COCO-Stuff[2] and ADE20K[10] are also presented in Figure 4 and 5, respectively. It can be seen that our method achieves more natural and vivid results in diverse scenarios, and produces more semantically consistent colors for a variety of objects.

3. Additional Ablation Study and Visual Results

We build four variants of our model with different ConvNeXt[5] backbones, as detailed in Table 2. As can be seen, the backbone plays a key role in image colorization. We choice ConvNeXt-L due to its superior performance.

More visual results on ablations of color decoder, colorfulness loss, and different visual feature scales are shown in Figure 1 and Figure 2.

4. Runtime Analysis

Our method colorizes grayscale images of resolution 256×256 at 25 FPS / 21 FPS using ConvNeXt-T / ConvNeXt-L as the backbone. The inference speed of our end-to-end method is ×96 faster than the previous transformer-based method [4]. All tests are performed on a machine with an NVIDIA Tesla V100 GPU.

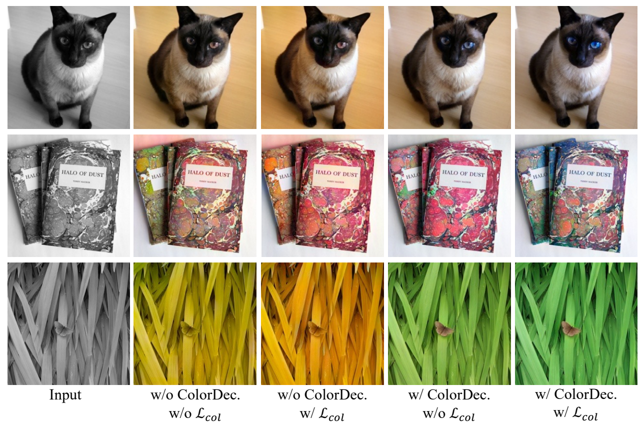


Figure 1. **More visual results of ablation on color decoder and colorfulness loss.**

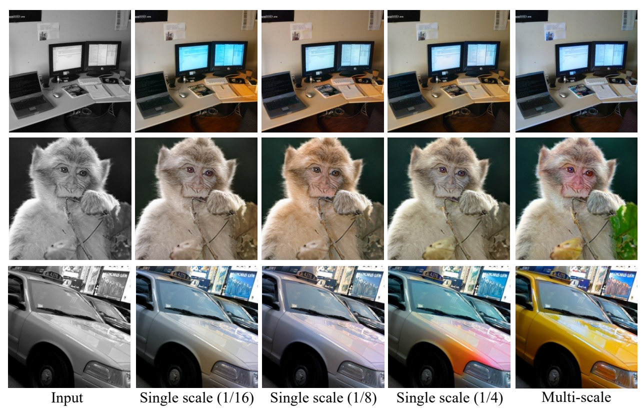


Figure 2. **More visual results of ablation on different feature scales.**

5. More Results on Legacy Black and White Photos

More colorization results on legacy black and white photos are shown in Figure 6, demonstrating the generalization capability of our method.

	Output size	DDColor
Stage 1	64×64×96	Conv. 4×4, 96, stide 4 [Depthwise Conv. 7×7, 96 Conv. 1×1, 384 Conv. 1×1, 96] × 3
Stage 2	32×32×192	[Depthwise Conv. 7×7, 192 Conv. 1×1, 768 Conv. 1×1, 192] × 3
Stage 3	16×16×384	[Depthwise Conv. 7×7, 384 Conv. 1×1, 1536 Conv. 1×1, 384] × 9
Stage 4	8×8×768	[Depthwise Conv. 7×7, 768 Conv. 1×1, 3072 Conv. 1×1, 768] × 3
Stage 5	16×16×512	PixelShuffle, scale 2 Concat feat. from Stage 3 Conv. 3×3, 512
Stage 6	32×32×512	PixelShuffle, scale 2 Concat feat. from Stage 2 Conv. 3×3, 512
Stage 7	64×64×256	PixelShuffle, scale 2 Concat feat. from Stage 1 Conv. 3×3, 256
Stage 8	256×256×256	PixelShuffle, scale 4
Color Dec.	256×100	Conv. 1×1, 256 feat. from Stage 5 Conv. 1×1, 256 feat. from Stage 6 Conv. 1×1, 256 feat. from Stage 7 [Conv. 1×1, 256×3 Cross-attn. Self-attn. Conv. 1×1, 2048 Conv. 1×1, 256] × 9
Stage 9	256×256×100	Dot Product feat. from Stage 8 & feat. from Color Dec.
Stage 10	256×256×2	Concat input Conv. 1×1, 2

Table 1. **Detailed architecture of DDColor**.

Model Name	Backbone	FID↓	CF↑	ΔCF↓	Params
DDColor-T	ConvNeXt-T	4.38	37.66	0.55	55.0M
DDColor-S	ConvNeXt-S	4.25	38.10	0.11	76.6M
DDColor-B	ConvNeXt-B	4.06	38.15	0.06	116.2M
DDColor-L	ConvNeXt-L	**3.92**	**38.26**	**0.05**	227.9M

Table 2. **Backbone variants.** We build four variants of our DDColor based on backbones of different sizes. The overall performance improves with the increase of the scale of the backbone network.

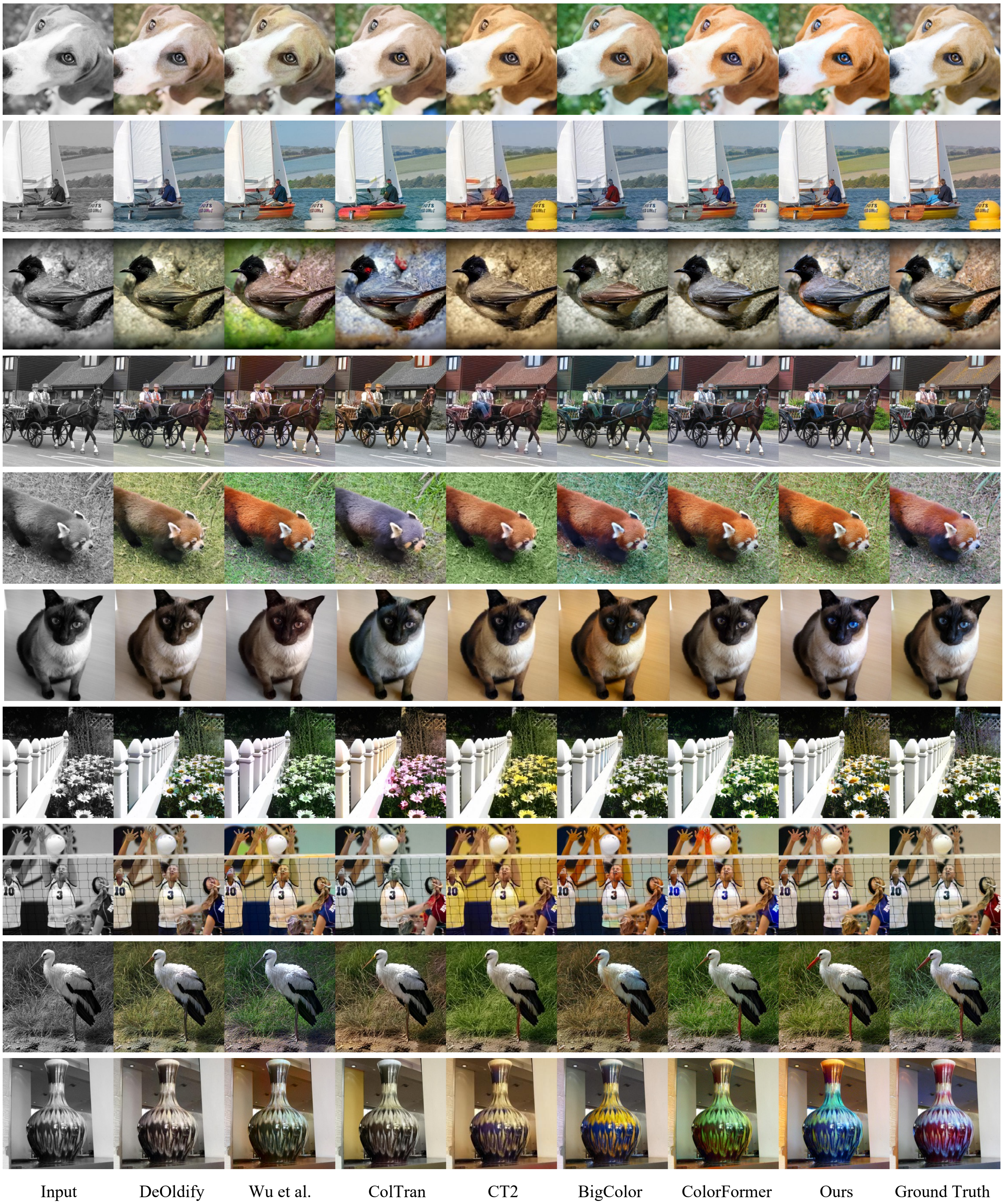
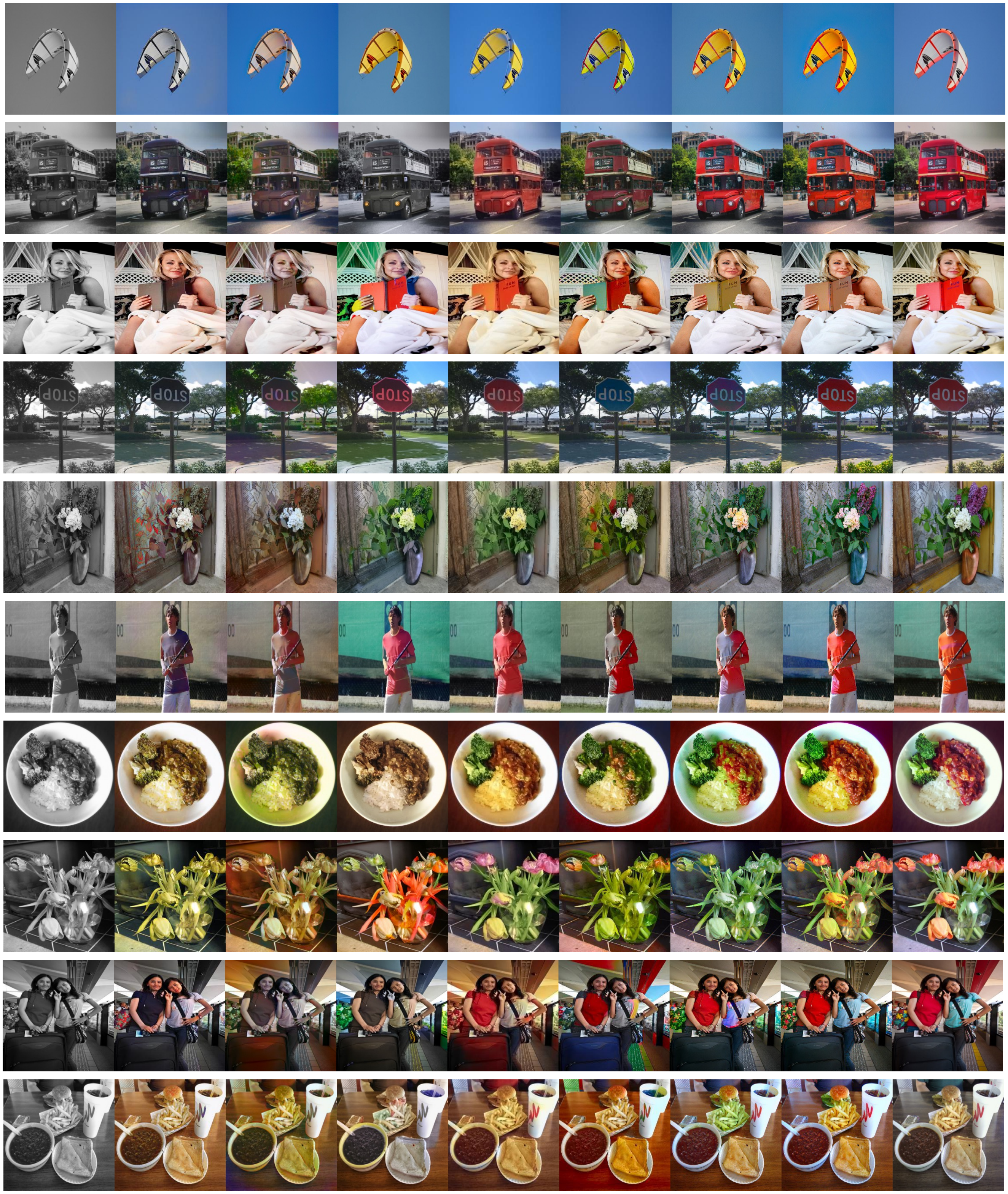


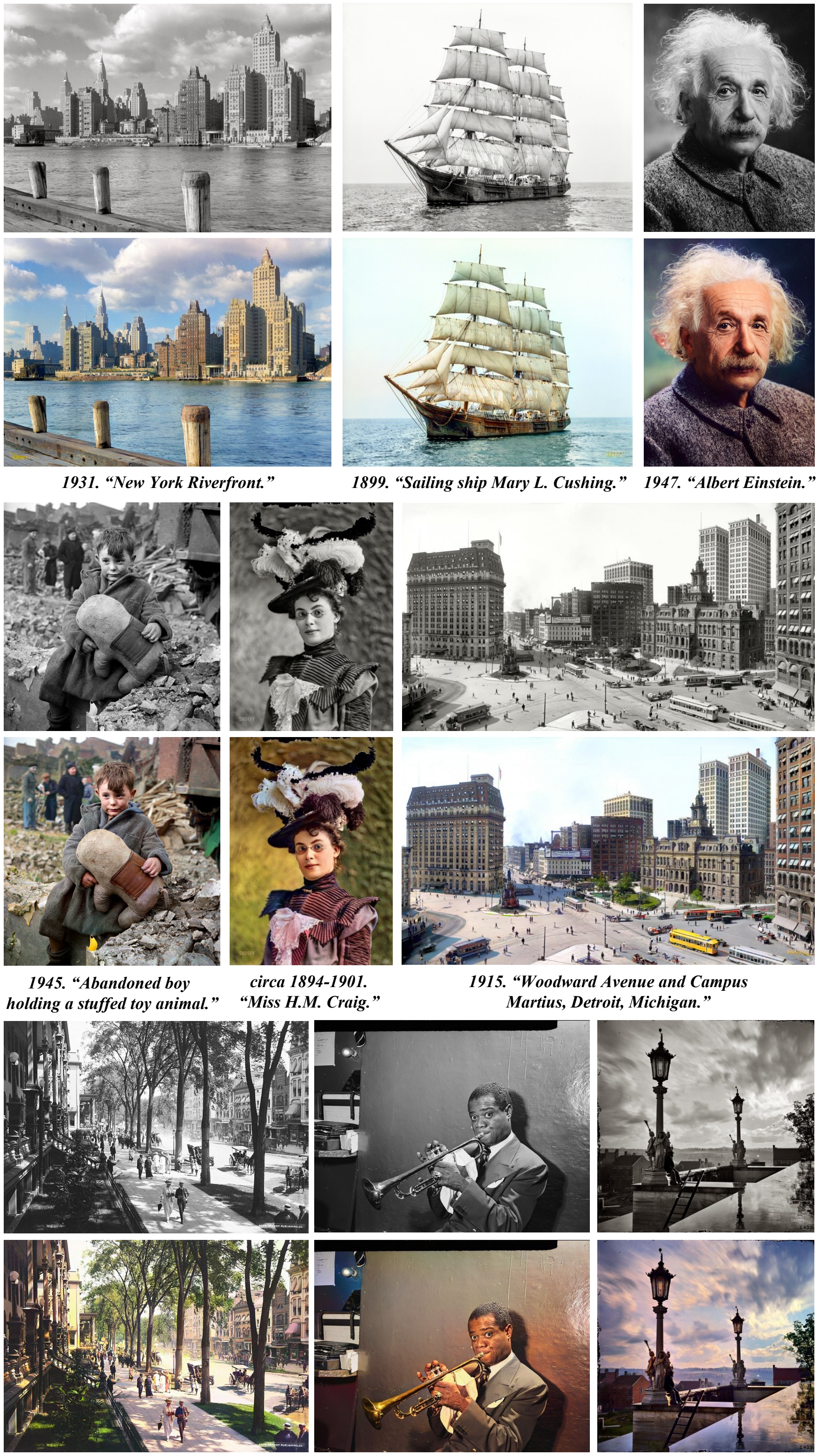
Figure 3. **More qualitative comparisons with previous colorization methods on ImageNet.**



Input DeOldify Wu et al. ColTran CT2 BigColor ColorFormer Ours Ground Truth

Figure 4. **Qualitative comparisons with previous colorization methods on COCO-Stuff.**

Figure 5. **Qualitative comparisons with previous colorization methods on ADE20K.**



1931. "New York Riverfront." *1899. "Sailing ship Mary L. Cushing."* *1947. "Albert Einstein."*

1945. "Abandoned boy holding a stuffed toy animal." *circa 1894-1901. "Miss H.M. Craig."* *1915. "Woodward Avenue and Campus Martius, Detroit, Michigan."*

circa 1900-1915. "Broadway at the United States Hotel Saratoga Springs." *1946. "Louis Armstrong practicing in his dressing room."* *1864. "View from the Capitol at Nashville, Tennessee."*

Figure 6. **More results on legacy black and white photos.**

References

[1] Jason Antic. jantic/deoldify: A deep learning based project for colorizing and restoring old images (and video!). https://github.com/jantic/DeOldify, 2019. 1

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 1

[3] Kim Geonung, Kang Kyoungkook, Kim Seongtae, Lee Hwayoon, Kim Sehoon, Kim Jonghyun, Baek Seung-Hwan, and Cho Sunghyun. Bigcolor: Colorization using a generative color prior for natural images. In *European Conference on Computer Vision (ECCV)*, 2022. 1

[4] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. Colorization transformer. In *International Conference on Learning Representations*, 2021. 1

[5] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 1

[6] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 1

[7] Shuchen Weng, Jimeng Sun, Yu Li, Si Li, and Boxin Shi. Ct2: Colorization transformer via color tokens. In *European Conference on Computer Vision (ECCV)*, 2022. 1

[8] Yanze Wu, Xintao Wang, Yu Li, Honglun Zhang, Xun Zhao, and Ying Shan. Towards vivid and diverse image colorization with generative color prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 1

[9] Ji Xiaozhong, Boyuan Jiang, Luo Donghao, Tao Guangpin, Chu Wenqing, Xie Zhifeng, Wang Chengjie, and Tai Ying. Colorformer: Image colorization via color memory assisted hybrid-attention transformer. In *European Conference on Computer Vision (ECCV)*, 2022. 1

[10] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 1