Supplementary Material: CT^2: Colorization Transformer via Color Tokens

Shuchen Weng[#1], Jimeng Sun[#2], Yu Li[3], Si Li[2], and Boxin Shi[*1]

[1] NERCVT, School of Computer Science, Peking University
[2] School of Artificial Intelligence, Beijing University of Posts and Telecommunications
[3] International Digital Economy Academy
{shuchenweng,shiboxin}@pku.edu.cn
{sjm,lisi}@bupt.edu.cn
liyu@idea.edu.cn

6 Appendix

6.1 Model Variants

We build four variants of our model based on ViT [2], as detailed in Tab. 4. The corresponding quantitative results are shown in Tab. 5. As can be seen, while CT^2-Tiny obtains decent results, a larger model size can further improve the performance. The quantitative results reported in the main paper are the results of CT^2-Large.

Table 4. Details of CT^2 model variants.

Model	Layers	Hidden size	MLP size	Heads	Params
CT^2-Tiny	14	192	768	3	11.69M
CT^2-Small	14	384	1536	6	45.75M
CT^2-Base	14	768	3072	12	181.03M
CT^2-Large	26	1024	4096	16	462.98M

Table 5. Quantitative results of model variants. ↑ (↓) means higher (lower) is better.

Model	FID↓	PSNR↑	SSIM↑	LPIPS↓	colorful↑	△colorful↓
CT^2-Tiny	7.17	22.31	0.90	0.22	41.49	5.18
CT^2-Small	6.87	22.44	0.91	0.22	41.26	5.07
CT^2-Base	6.24	22.97	0.91	0.21	38.45	2.27
CT^2-Large	**5.51**	**23.50**	**0.92**	**0.19**	**38.48**	**2.17**

[#] Equal contributions. [*] Corresponding author.

6.2 Additional Qualitative Results

We qualitatively compare our CT^2 with 5 CNN-based colorization methods, *e.g.*, CIC [10], Deoldify [1], ChromaGAN [7], InstColor [6], and GCP [9]. We also compare our method with 4 advanced transformer-based methods, *e.g.*, SwinIR [5], Uformer [8], MAE [3], and ColTran [4]. Detailed descriptions are presented in the Sec. 4 of the main paper. Note that ChromaGAN [7], InstColor [6], and GCP [9] use additional external priors, while CT^2 does not involve any prior. Specifically, ChromaGAN [7] uses image category labels to optimize the class distribution as the side-task; InstColor [6] takes the well-pretrained detection model to extract object-level features; and GCP [9] inputs the image category labels into a well-pretrained GAN to generate a reference image as the colorization guidance. The comparison results can be found in Fig. 9 and Fig. 10, showing the consistent advantage of our method in producing vivid and realistic colorization results.

6.3 Additional Ablation Study

We conduct additional ablation study experiments to test the impact of luminance interval number and decoder layer number. As the results shown in Tab. 6, the performance improves while gains decrease after a certain number is reached (4 for intervals number and 2 for layers number). As a trade-off, we compromise slightly lower performance in exchange for faster running speed and fewer parameters. These ablation study experiments are conducted on the CT^2-Tiny variant. We also show more qualitative ablation study results on the effect of LSM, color attention, and color query, which are listed in Fig. 11.

Table 6. Quantitative ablation study results. ↑ (↓) means higher (lower) is better.

Category	Number	FID↓	PSNR↑	SSIM↑	LPIPS↓	colorful↑	△colorful↓	Params
Interval	1	7.51	20.99	0.820	0.264	41.56	5.24	11.69M
	2	7.23	22.15	0.891	0.231	**41.52**	5.20	11.69M
	4	7.17	**22.31**	0.901	0.224	41.49	5.18	11.69M
	10	**7.15**	22.17	0.903	0.221	41.43	**5.12**	11.69M
	100	**7.15**	22.30	**0.905**	**0.220**	41.50	5.19	11.69M
Decoder Layer	1	7.42	22.20	0.895	0.230	**41.51**	5.20	**11.25M**
	2	7.17	**22.32**	0.901	**0.222**	41.43	**5.12**	11.69M
	3	7.13	22.30	0.902	0.234	41.48	5.17	12.14M
	4	**7.12**	22.31	**0.904**	0.224	41.45	5.14	12.58M

6.4 Application

We show more colorization results on legacy black and white photos in Fig. 12, demonstrating the generalization capability of our CT^2. In addition, more colorization results of CT^2 on test set could be found in Fig. 13, Fig. 14, and Fig. 15. Our method can always produce high quality colorization results.

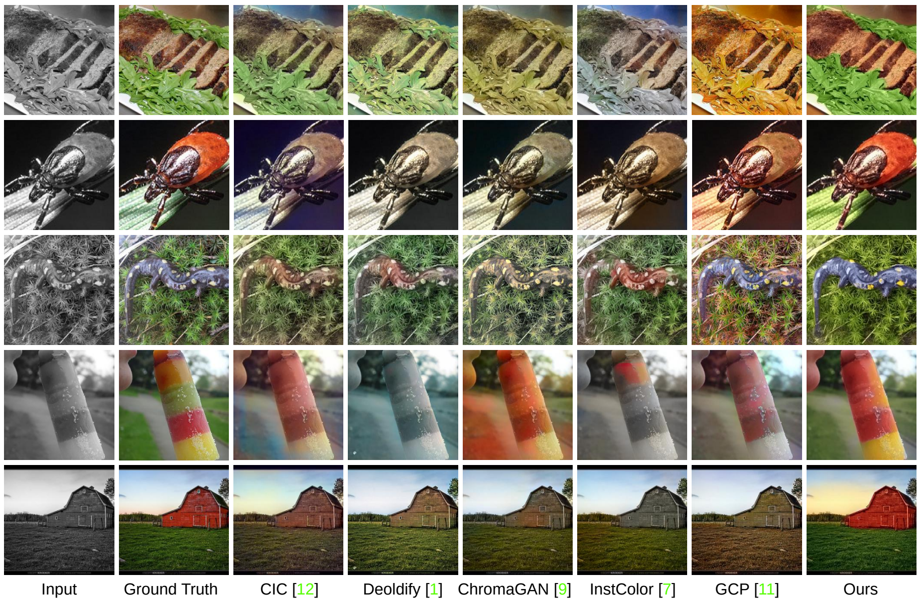

Fig. 9. More comparison results with CNN-based methods.

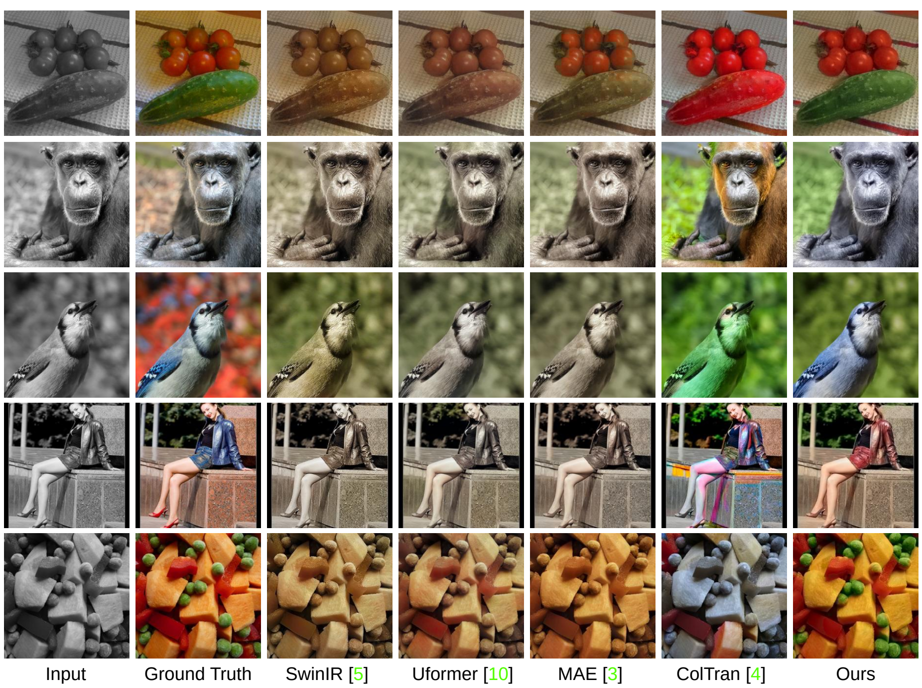

Fig. 10. More comparison results with transformer-based methods.

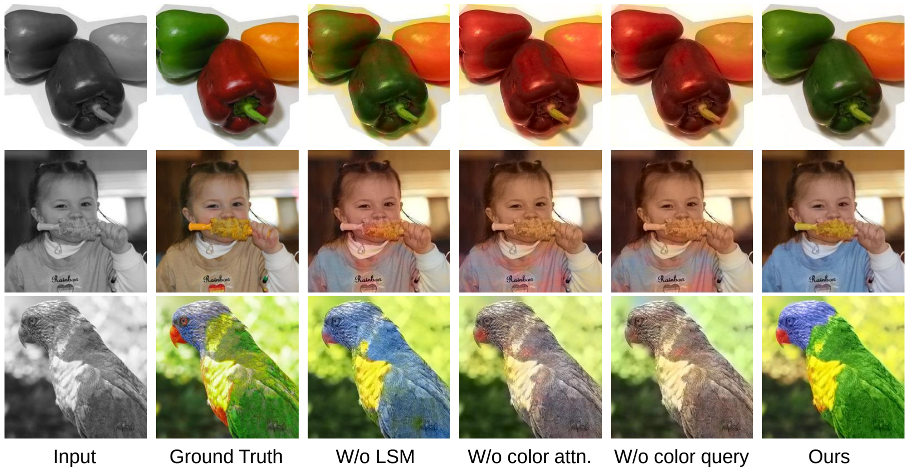

Fig. 11. More ablation study results.

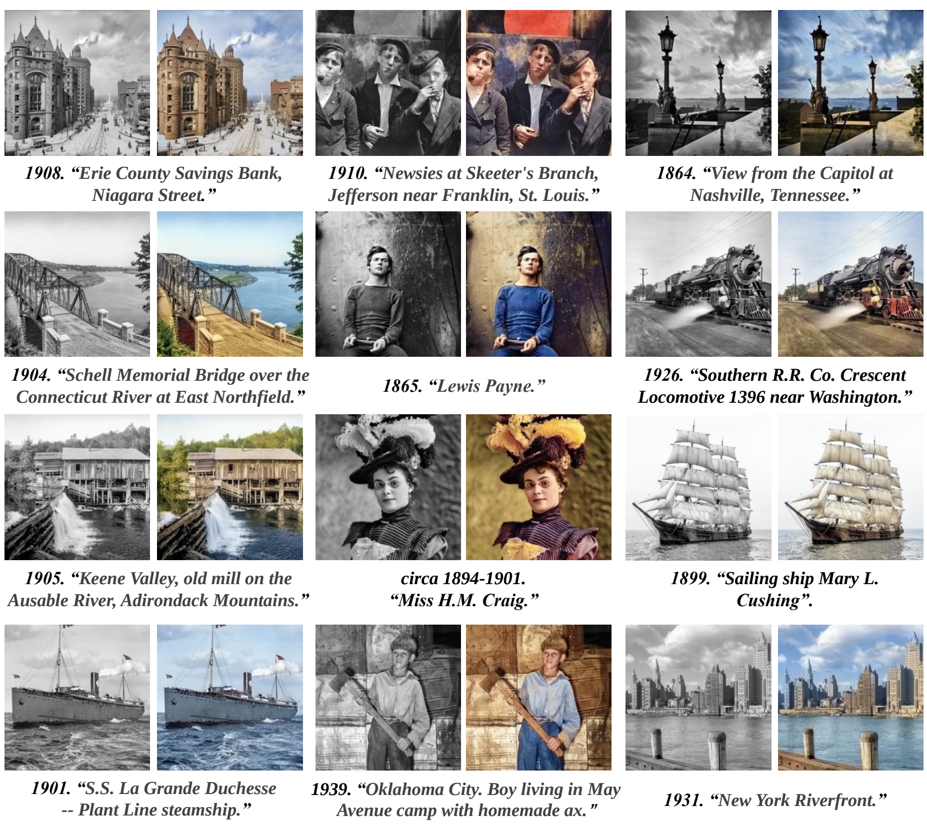

Fig. 12. More application results.

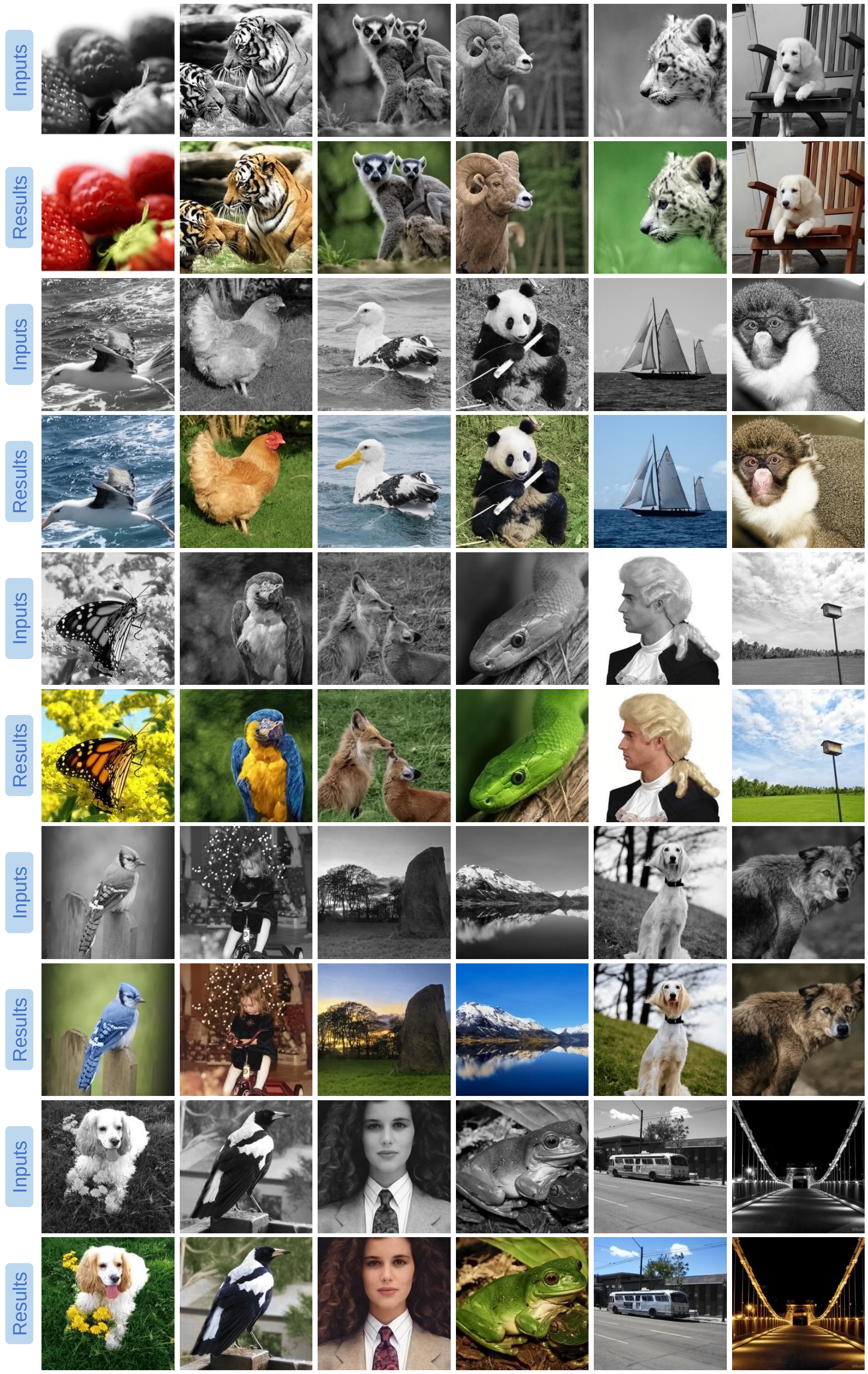

Fig. 13. More results of our CT^2.

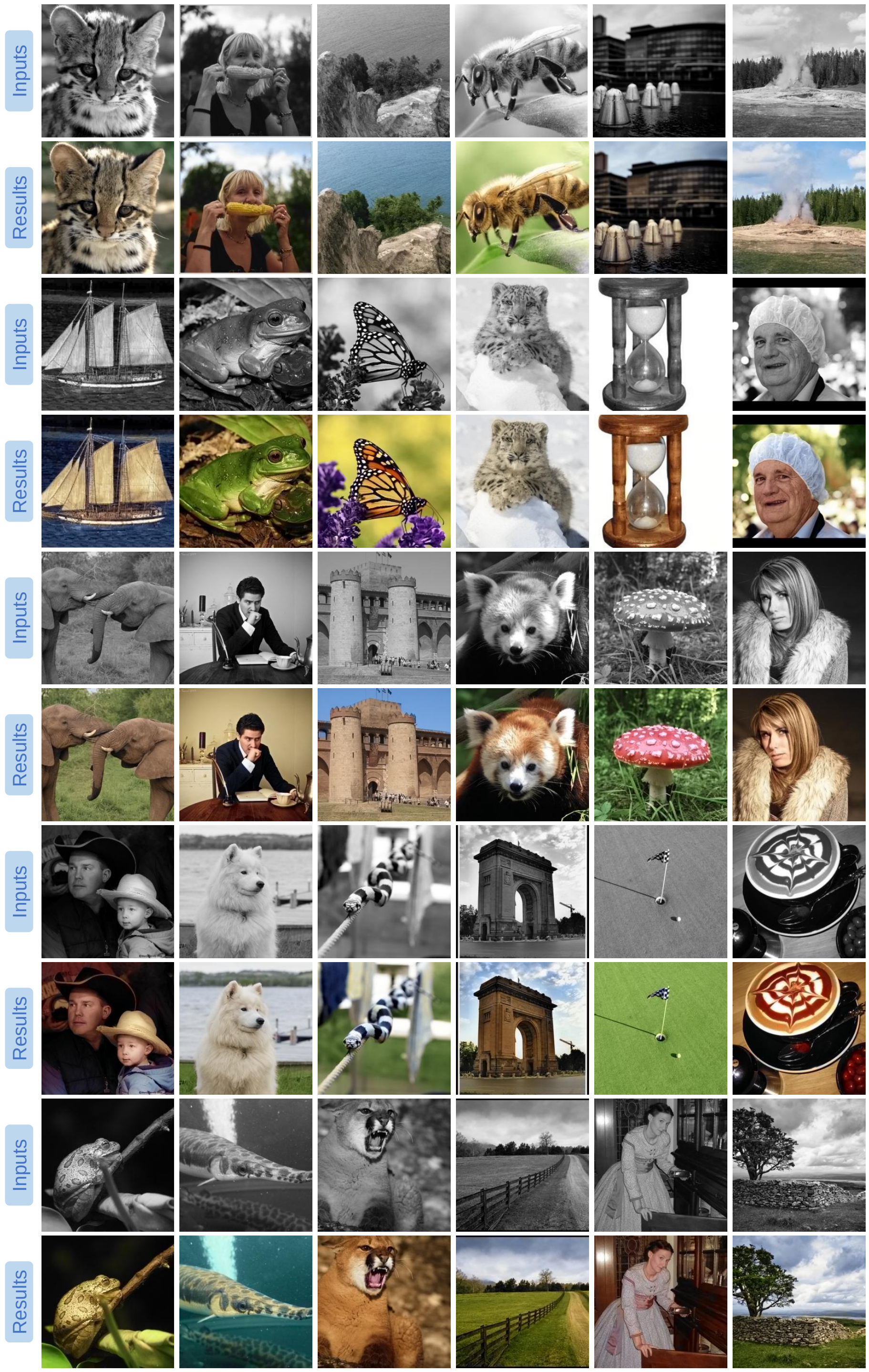

Fig. 14. More results of our CT^2.

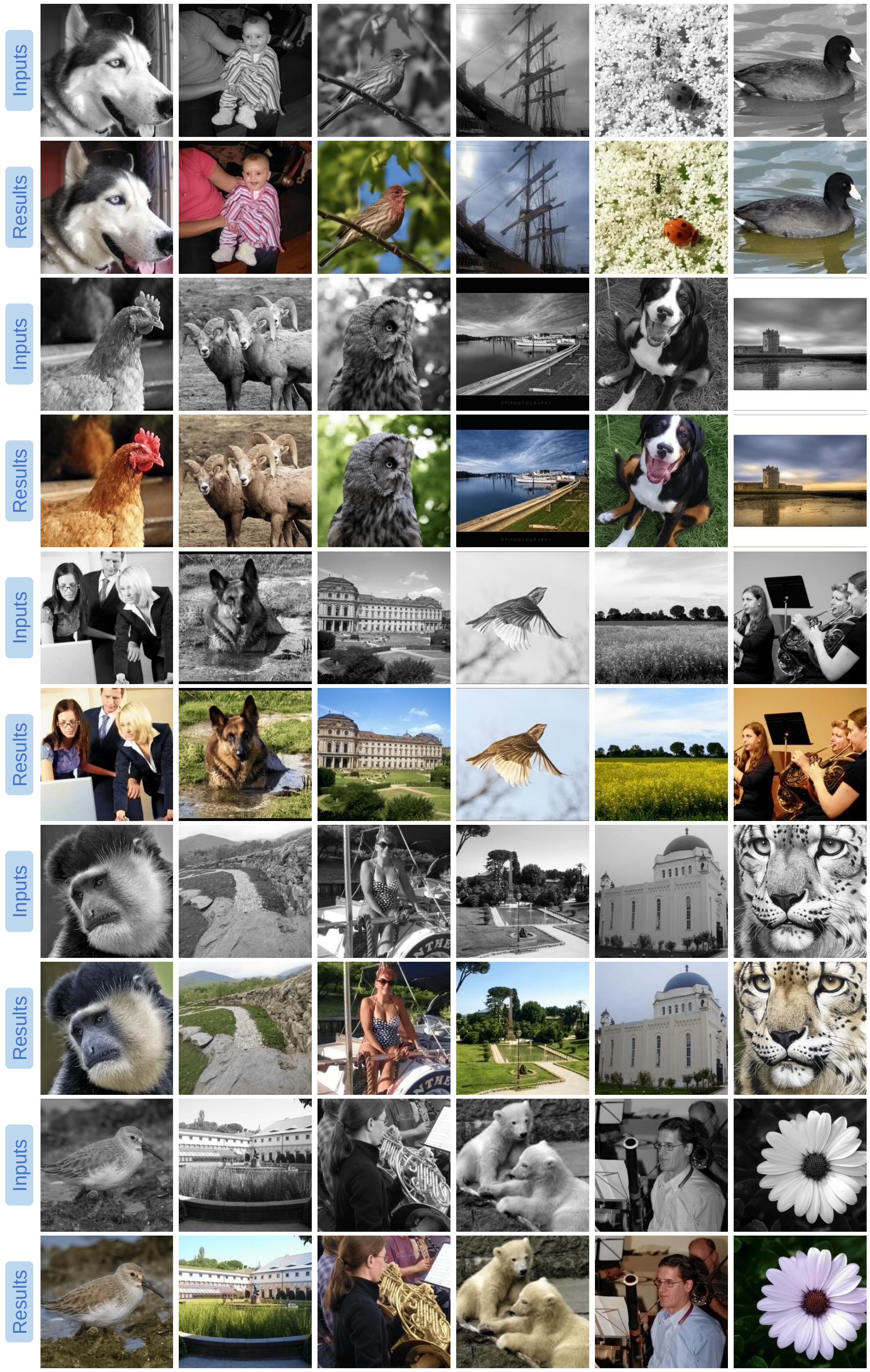

Fig. 15. More results of our CT^2.

References

1. Antic, J.: A deep learning based project for colorizing and restoring old images (and video!), https://github.com/jantic/DeOldify 2
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 1
3. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. arXiv (2021) 2
4. Kumar, M., Weissenborn, D., Kalchbrenner, N.: Colorization transformer. In: ICLR (2021) 2
5. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: ICCV (2021) 2
6. Su, J.W., Chu, H.K., Huang, J.B.: Instance-aware image colorization. In: CVPR (2020) 2
7. Vitoria, P., Raad, L., Ballester, C.: Chromagan: Adversarial picture colorization with semantic class distribution. In: WACV (2020) 2
8. Wang, Z., Cun, X., Bao, J., Liu, J.: Uformer: A general u-shaped transformer for image restoration. arXiv (2021) 2
9. Wu, Y., Wang, X., Li, Y., Zhang, H., Zhao, X., Shan, Y.: Towards vivid and diverse image colorization with generative color prior. In: ICCV (2021) 2
10. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV (2016) 2