Supplementary for iColoriT: Towards Propagating Local Hint to the Right Region in Interactive Colorization by Leveraging Vision Transformer

This supplementary material presents quantitative results on unconditional colorization (Section 1), the effect of the number of hints sampled during training (Section 2), quantitative analysis of the hint propagation range (Section 3), and additional qualitative results (Section 4). We also attach a demo video demonstrating the use case of iColoriT.

1. Unconditional Colorization Using iColoriT

Similar to recent point-interactive colorization approaches [6, 9], iColoriT is capable of generating colorized images without any user hints. We compare the Fréchet inception distance (FID) score [1] of the colorization results generated without any user hints with a pre-trained Inception V3 [7] network. We exclude the conventional filter-based approach by Yin *et al*. [8] since the method does not operate without user hints. The FID score is widely used to evaluate the how realistic the generated images are compared to the original color images. Our approach achieves a low FID score and generates realistic colors even without user hints.

Methods	FID
Zhang *et al*. [9]	7.29
Su *et al*. [6]	6.56
Yin *et al*. [8]	-
iColoriT	**4.89**

Table 1. Fréchet inception distance (FID) score of unconditional results of iColoriT and baselines. iColoriT is able to generate realistic colors even without user hints.

2. Number of Hints Sampled During Training

The number of simulated user hints sampled during training may alter the performance since iColoriT directly learns how to propagate the color hints. We train our model provided with different number of hints sampled from various uniform distributions (*e.g.*, $\mathcal{U}(0, 16)$, $\mathcal{U}(0, 32)$, $\mathcal{U}(0, 64)$, $\mathcal{U}(0, 128)$, $\mathcal{U}(0, 256)$). To our surprise, the number of hints provided during training did not have an immense effect on the final performance. As plotted in Fig. 1, the PSNR measured on ImageNet ctest [2] are similar for most models. We empirically find that sampling the number of hints from $\mathcal{U}(0, 128)$ demonstrates a high performance in most regions including PSNR given 10 user hints (PSNR@10) which we choose as a representative indicator. We also observe that the performance of models given a relatively small number of hints (*i.e.*, $\mathcal{U}(0, 16)$) does not necessarily outperform other models when evaluated with a small number of hints (*e.g.*, PSNR@5). We presume that the model learns to appropriately propagate hints to relevant regions when trained with a more diverse number of hints. Future work focusing on how to sample the simulated hints is an interesting subject.

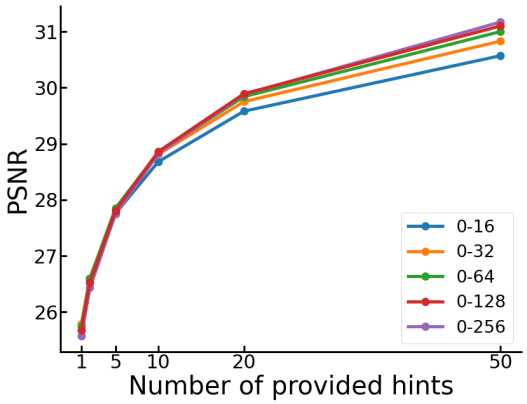

Figure 1. PSNR achieved by models trained by sampling different number of simulated user hints. The numbers on the legend indicate the range of the uniform distribution. The PSNR is measured in the ImageNet ctest [2] validation set.

3. Measuring the Hint Propagation Range

Not only can iColoriT accurately reflect user hints and achieve a higher PSNR, iColoriT is capable of propagating user hints to longer distances if needed. In order to measure how far a hint propagates to further regions, we present the hint propagation range (HPR) measure. Given an image I_{pred}^{t-1} colorized with $t-1$ number of color hints, HPR@ t indicates the average distance from the newly provided t-th hint to the pixels that have been colorized by the t-th hint. We define a pixel to have been *colorized* if the mean squared

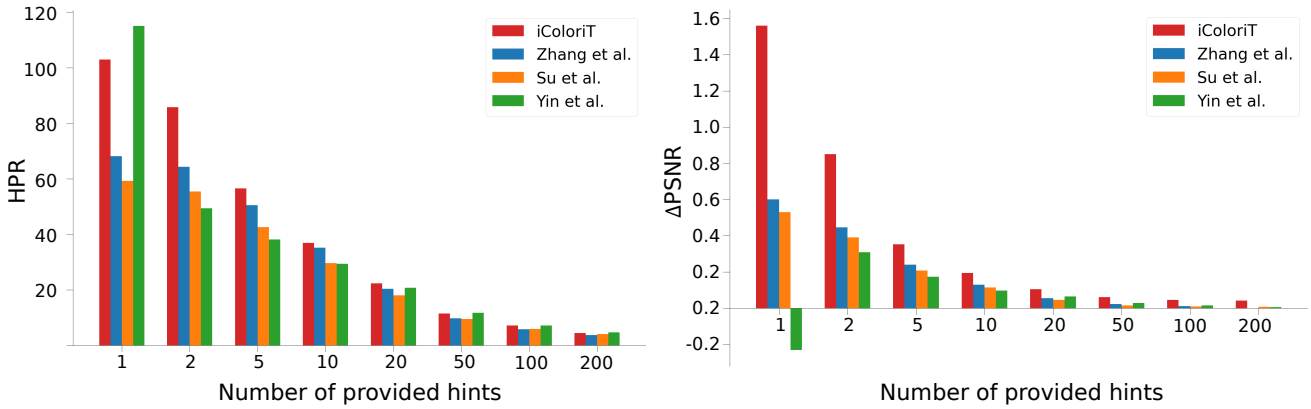

Figure 2. Average hint propagation range (HPR) and PSNR gain (ΔPSNR) when given an additional hint. All scores are measured in the ImageNet ctest [2] dataset. iColoriT shows both high HPR and a high PSNR gain at all stages of the colorization process.

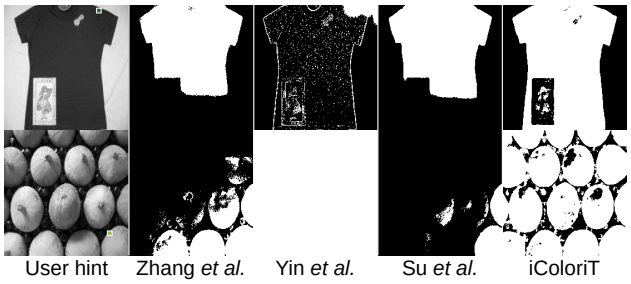

Figure 3. Visualization of the colorized region $\mathcal{C}_1$ for different baselines. iColoriT is able to propagate the hint to further regions if needed, while other approaches do not necessarily colorize all relevant regions.

error (MSE) between the initial value and the altered value is larger than 2.3 in the CIELab color space [5], which is the just-noticeable-difference (JND) perceived by the human eye [4]. Given a set of coordinates $(x_i, y_i) \in \mathcal{C}_t$ of pixels colorized by hint h_t and the coordinate of h_t (x_h, y_h), we calculate HPR with,

$$\mathcal{C}_t = \{(x, y) \mid \text{MSE}(I^t_{xy}, I^{t-1}_{xy}) > \text{JND}\}, \quad (1)$$

$$\text{HPR@}\,t = \frac{1}{|\mathcal{C}_t|} \sum_{(x_i, y_i) \in \mathcal{C}_t} \sqrt{(x_i - x_h)^2 + (y_i - y_h)^2}, \quad (2)$$

which is the average Euclidean distance from h_t to all locations in $\mathcal{C}_t$.

We measure the HPR across ImageNet ctest [2] at different stages of the colorization process and plot the results in Figure 2. Note that the HPR measure itself does not assess whether a model is appropriately reflecting the color hints provided by the user. However, when examined together with PSNR gain (ΔPSNR) in Figure 2 where the PSNR hugely improves at earlier stages, we can conclude that iColoriT reflects the color hints to further regions in a constructive manner. The optimization-based method proposed by Yin *et al*. [8] often overly propagates the initial color

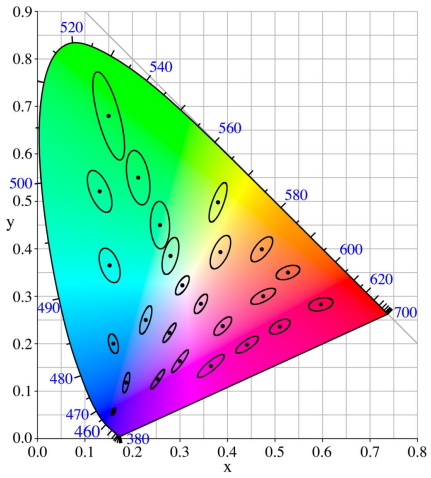

Figure 4. The MacAdam ellipse from MacAdam's paper [3] illustrating the just-noticeable-difference in the *xy* chromaticity diagram.

hint to the entire image, which does not contribute to improving the PSNR or the perceptual quality compared to the original grayscale image. Learning-based baselines [6,9] on the other hand, tend to locally colorize images which hinders further PSNR gain. iColoriT takes advantage of both aspects and propagates color hints to further regions while hugely improving the PSNR.

We also visualize $\mathcal{C}_1$, the pixel coordinates colorized by the initial hint h_1, in Figure 3 to intuitively understand the regions of which the hint alters the image. We believe that the self-attention mechanism is central for selectively colorizing the relevant regions regardless of the distance from the user-provided color hint. Qualitative results supporting this claim are provided in Section 4.

For a clearer understanding, we provide an illustration of the MacAdam ellipse which visualizes the just-noticeable-difference of colors. The colors within the same ellipse indi-

cate that the colors are indistinguishable to the human eye. Although the areas of the ellipses are not uniform in the *xy* chromaticity diagram, the JND in the CIELab color space is known to be roughly constant.

4. Additional Qualitative Results

In this section and the remaining pages of the supplementary, we provide additional qualitative results. We present colorized images produced by users through a demo user interface in Fig. 5. Notice that iColoriT produces plausible results with minimal user hints. Figs. 6 to 9 compares the results from different baseline models and iColoriT. Furthermore, uncurated results from the ImageNet ctest [2] of iColoriT are provided in Figs. 10 to 13. The figures are sorted according to the number of hints given to the model. Most images do not contain color-bleeding artifacts or partially colorized images. Finally, we attach a demo video demonstrating an interactive colorization scenario and a use case of iColoriT.

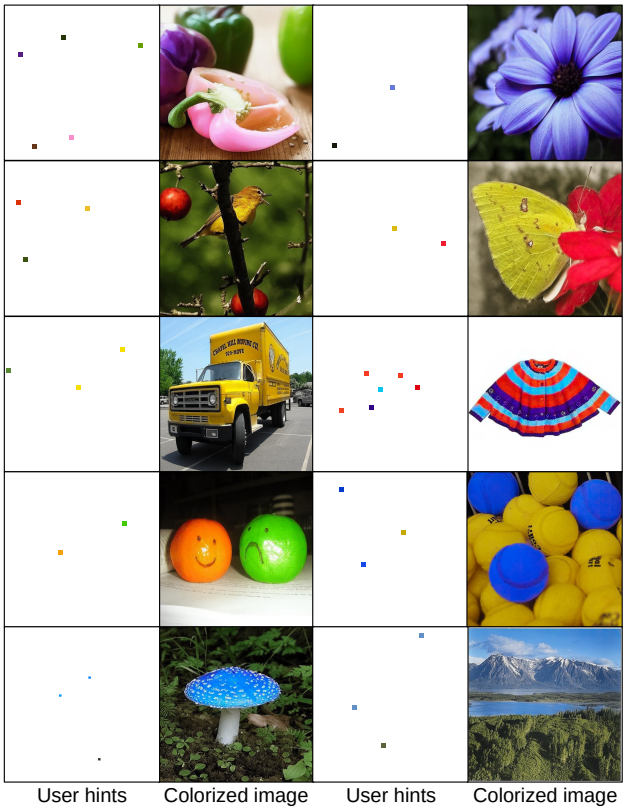

Figure 5. Colorization results produced by user-provided hints. Bottom right is a photo by Ansel Adams of Grand Teton, 1941.

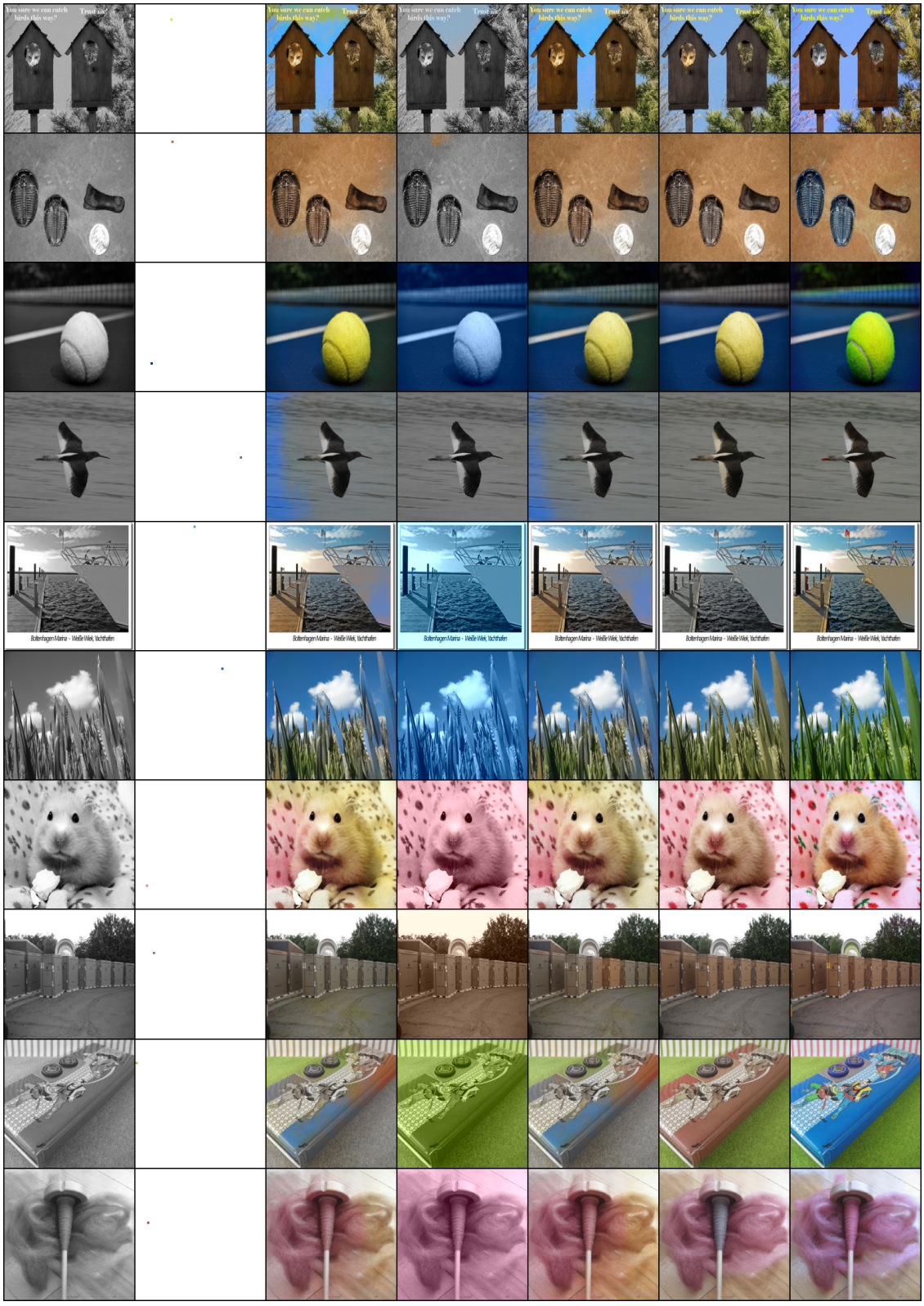

Figure 6. Additional qualitative results compared with baseline approaches. A single hint location is sampled from a uniform distribution.

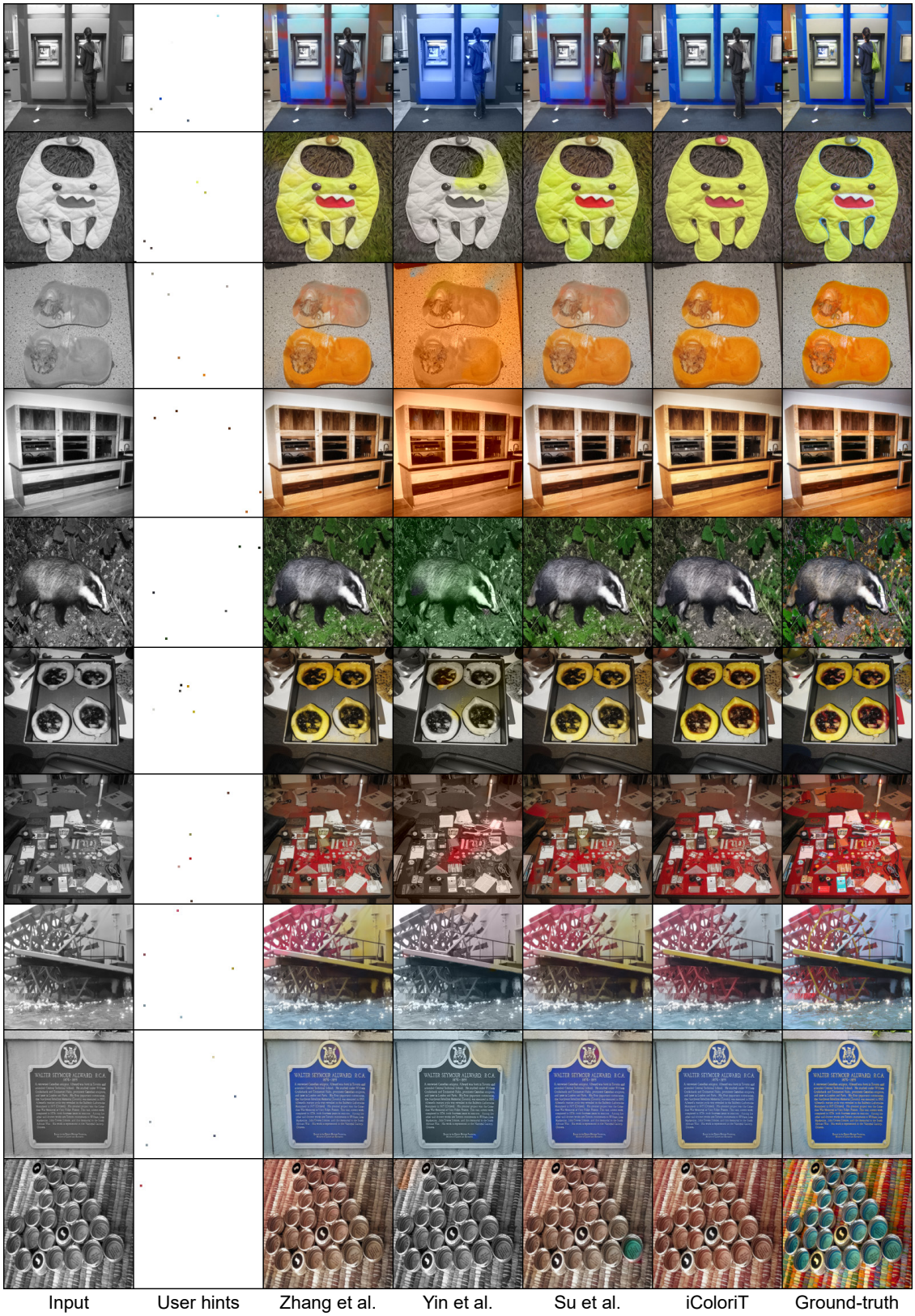

Figure 7. Additional qualitative results compared with baseline approaches. 5 hint locations are sampled from a uniform distribution.

Figure 8. Additional qualitative results compared with baseline approaches. 10 hint locations are sampled from a uniform distribution.

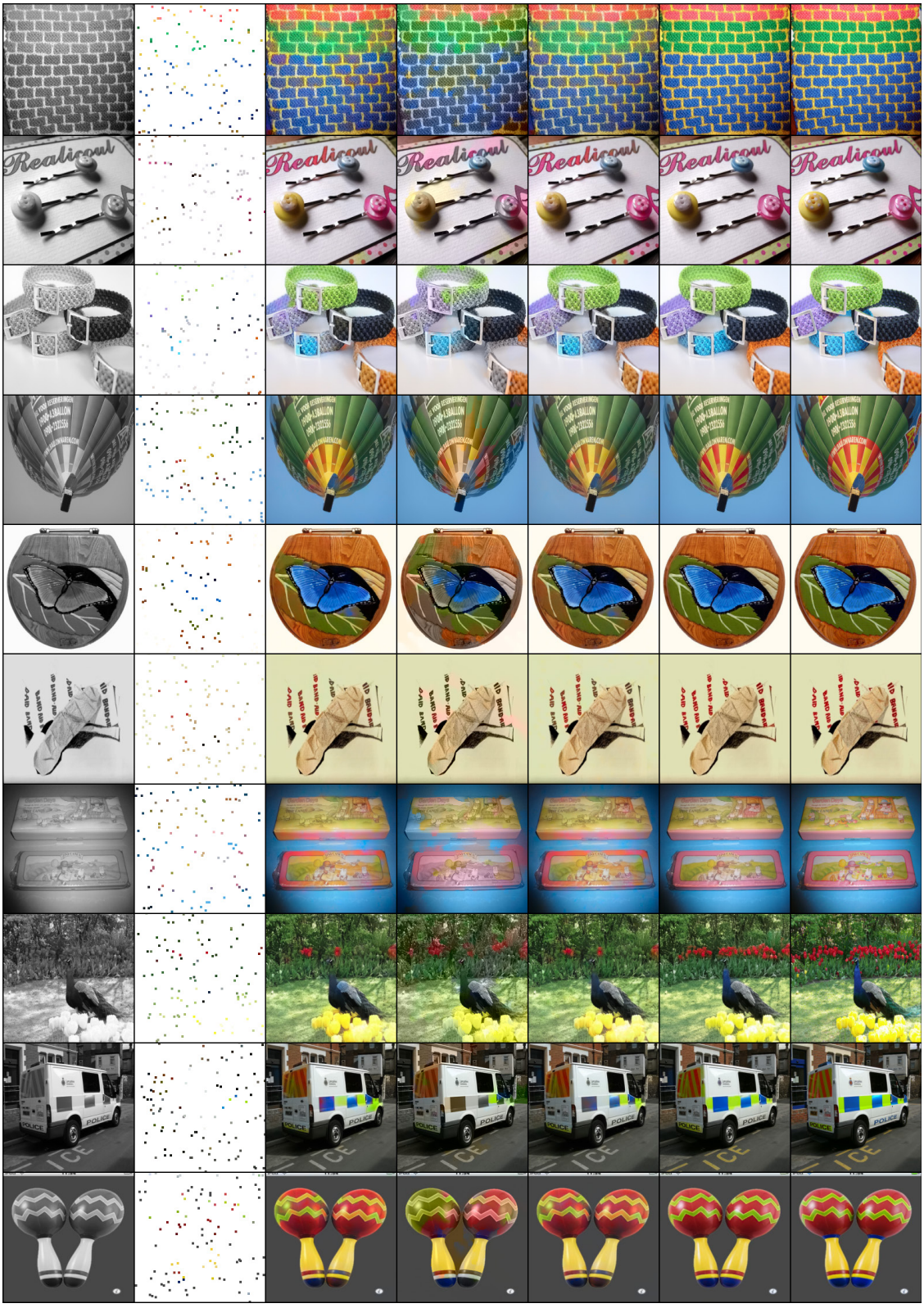

Figure 9. Additional qualitative results compared with baseline approaches. 100 hint locations are sampled from a uniform distribution.

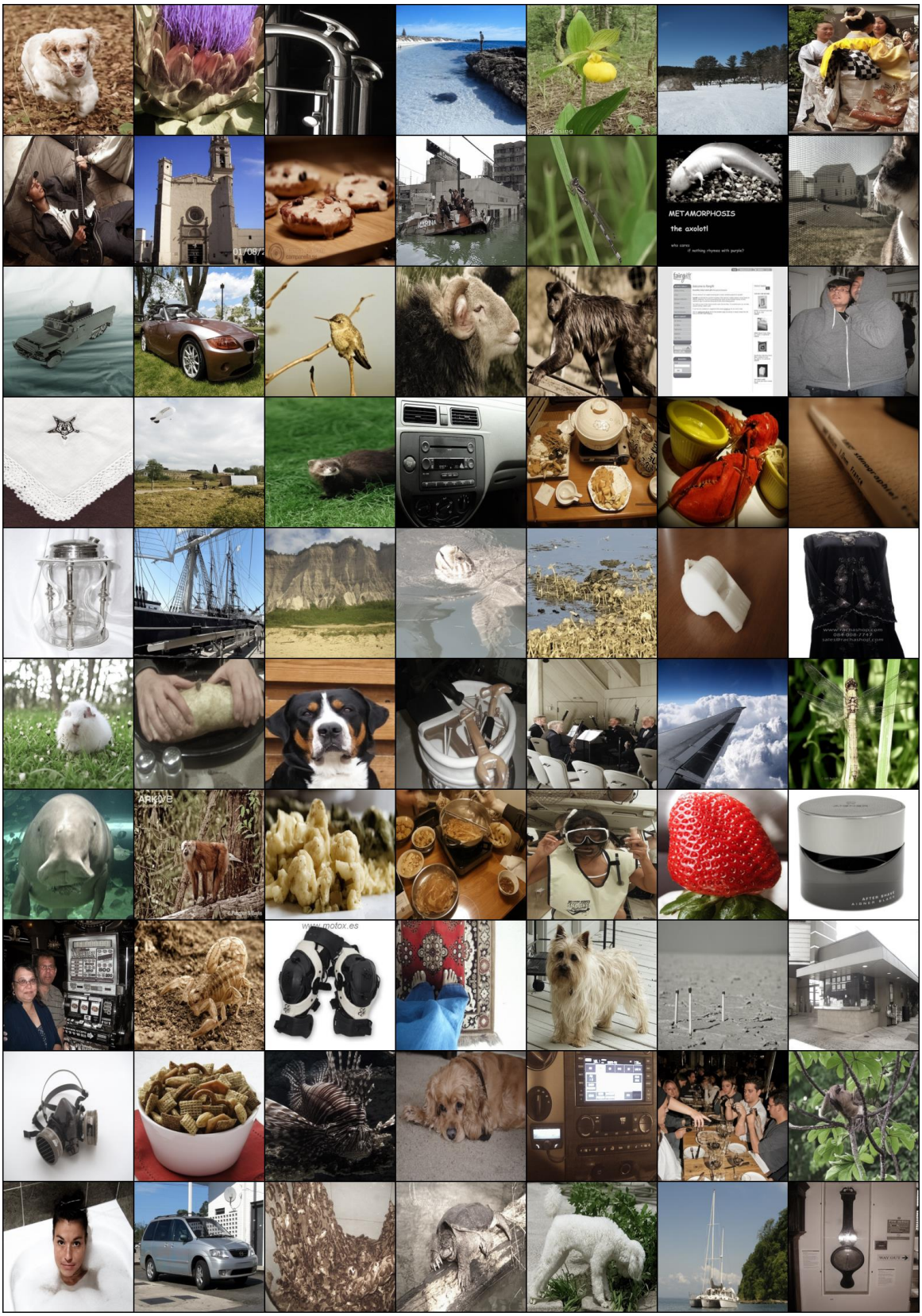

Figure 10. Uncurated images produced with a single hint where the hint location is randomly sampled from a uniform distribution.

Figure 11. Uncurated images produced given five hints where the hint locations are randomly sampled from a uniform distribution.

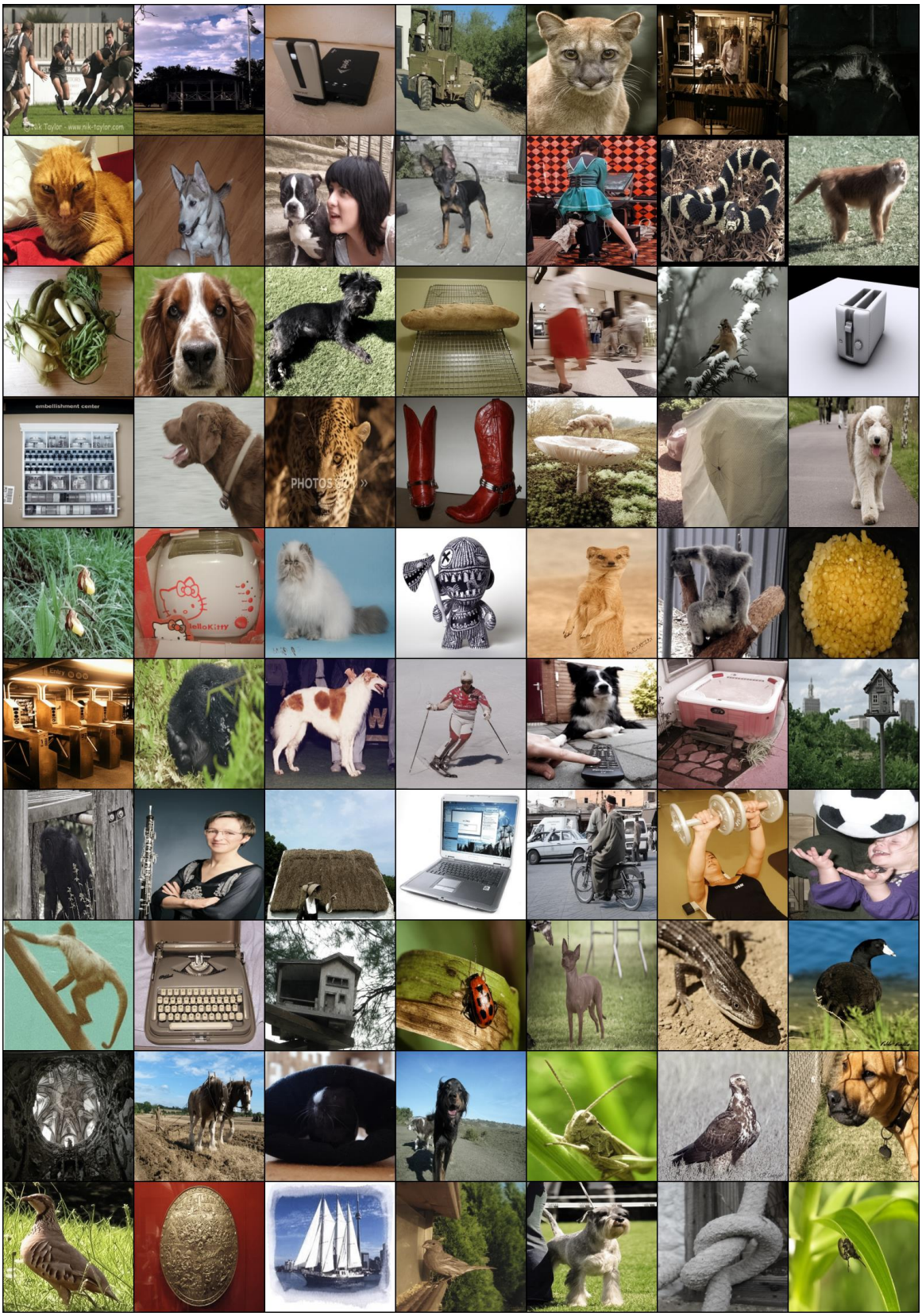

Figure 12. Uncurated images produced given ten hints where the hint locations are randomly sampled from a uniform distribution.

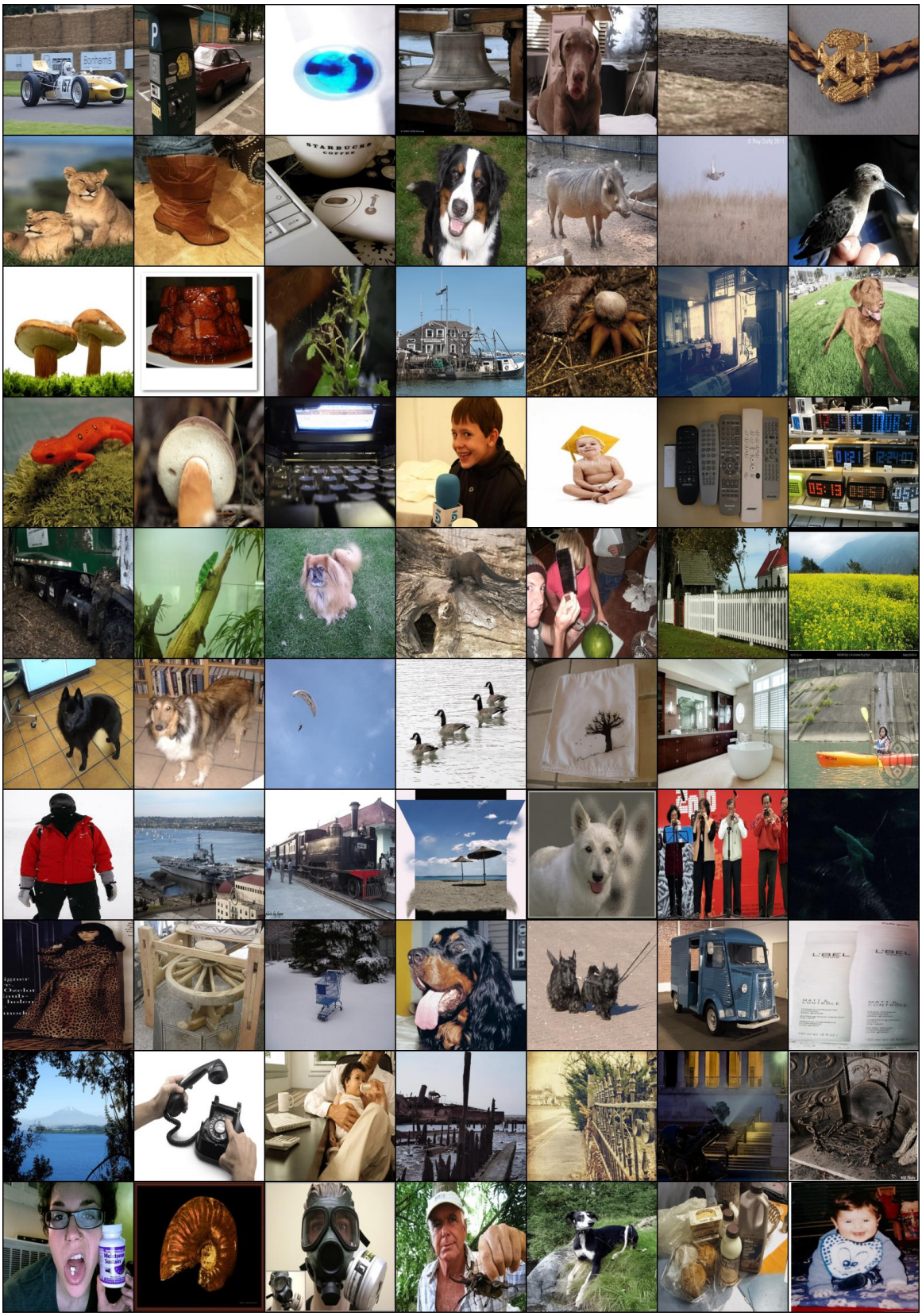

Figure 13. Uncurated images produced given hundred hints where the hint locations are randomly sampled from a uniform distribution.

References

[1] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

[2] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European Conference on Computer Vision (ECCV)*, 2016.

[3] David L. MacAdam. Visual sensitivities to color differences in daylight∗. *J. Opt. Soc. Am.*, 32(5):247–274, May 1942.

[4] Gaurav Sharma. Color fundamentals for digital imaging. *Digital color imaging handbook*, 20:1–114, 2003.

[5] Thomas Smith and John Guild. The cie colorimetric standards and their use. *Transactions of the optical society*, 33(3):73, 1931.

[6] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7968–7977, 2020.

[7] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. arxiv. *arXiv preprint arXiv:1512.00567*, 2015.

[8] Hui Yin, Yuanhao Gong, and Guoping Qiu. Side window filtering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8758–8766, 2019.

[9] 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 TOG*, 2017.