Bridging the Domain Gap towards Generalization in Automatic Colorization

Hyejin Lee[1], Daehee Kim[1,2], Daeun Lee[3], Jinkyu Kim[4*], and Jaekoo Lee[1*]

[1] Department of Computer Science, Kookmin University
[2] Clova AI Research, NAVER Corp.
[3] Department of Statistics, Korea University
[4] Department of Computer Science and Engineering, Korea University
*Corresponding authors: jinkyukim@korea.ac.kr, jaekoo@kookmin.ac.kr

1 Content

This supplementary material provides details on our quantitative evaluation (Section 2) and more diverse qualitative examples (Section 3). We also provide the user study details (Section 4) and examples of our analysis of content-biased features vs. style-biased features in the generator (Section 5). We provide implementation details for transferring content features (Section 6). Lastly, we discuss failure cases of our method (Section 7).

2 Additional Quantitative Analysis

In Table 1 and Table 2, we provide our more detailed scores from our quantitative analysis (in terms of the following four metrics: PSNR, SSIM, IQM, and FID) with PACS [4] and Office-Home [8] datasets, respectively. Note that Office-Home contains a meaningful portion of gray-tone images, which makes our baseline model being placed in the higher rank though it produces poor performance in colorization (compare 2nd vs. 3rd rows in Fig. 3).

3 Additional Qualitative Analysis

We provide more diverse examples to compare colorization performance with conventional colorization approaches (Fig. 1) and with existing domain generalization techniques (Fig. 2 and Fig. 3 on the PACS [4] and Office-Home [8] datasets, respectively).

4 User Study Details

Recall from Section 4.4 in the main paper, we have conducted a user study to quantitatively evaluate the quality of the colorization performance. In this user study, participants are asked to answer the following two questions:

- **Q1** (Naturalness): "Do you think the provided image looks naturally colored?"
- **Q2** (Perceptual Realism): "Which of the following images are the best?"

Images were randomly sampled from each domain on the PACS and Office-Home datasets. Overall 34 human evaluators were recruited offline. We required them to answer 360 questions (each for 180 questions) and 12,240 votes are collected. In Fig. 4, we provide a sample of the questionnaire we used. In Table 3, we also summarize our results from the user study.

5 Content-biased vs. Style-biased Features in Generator

Recall from the Table 2 (model H) in the main paper, we verified our motivation behind transferring content feature statistics by evaluating a variant model of ours where we apply AdaIN to the content of o_l with the style of z (instead of using the content of z with the style of o_l). In Fig. 6, we provide examples of generated images. Examples in the first three columns are produced with our default architecture, while the last three columns are examples of using the content of o_l with the style of z. As consistent with our main paper, constraining the generator towards using content-biased features degrades the overall performance in colorization.

6 Implementation Details for Transferring Content Features

Recall from Section 3.3, we use a style transfer technique using an AdaIN – i.e. given content information from z, we transfer style feature statistics of o_l from the intermediate layer of G. In our experiment in the main paper, we set $l = 5$. As shown in Table 4, we further provide our ablation study on the choice of l. Specifically, we experimented with the following two settings: (i) $l = 5$ and (ii) $l \geq 5$ where we transfer style feature statistics of all o_l for $l \geq 5$. As such style transfer techniques are not generally applied to high-level representations ($l < 5$), we only compare the above two cases. As shown in Table 4, we observe that the best performance is achieved by setting $l = 5$, i.e. adding more style transfer layers slightly degrades the overall performance. This is further confirmed in Fig. 7 where we observe marginal differences. Thus, we set $l = 5$ for better computational efficiency.

7 Failure Cases and Discussion

we use our model to colorize artworks with various textures, drawn by Pollock, and Gogh. In Fig. 5 (a), we observe that our model can colorize in a reasonable way, but sometimes shows sub-optimal results when the domain shift is large, which might be worth exploring as a future work.

Further, in Fig. 5 (b), we observe that dealing with the multi-modal distribution of colors is challenging. If each pixel has a multi-modal distribution of colors, it results in mixed colors. This problem has been repeatedly reported in the community, and it would be worth exploring multi-modal colorization.

Table 1. Colorization performance comparison in terms of four image quality evaluation metrics: PSNR, SSIM, IQM, and FID. To observe DG algorithm with baseline performance degradation in the domain generalization setting, we also compare each model with the non-domain generalization (DG) setting, *Abbr.*: P (Photo), A (Art Painting), and C (Cartoon). Data: PACS.

Models	PSNR ↑				SSIM ↑				IQM ↑				FID ↓			
	P	A	C	Avg.	P	A	C	Avg.	P	A	C	Avg.	P	A	C	Avg.
Zhang et al. [9]	29.62	29.66	32.39	30.56	0.68	0.65	0.61	0.65	1.85	1.93	1.90	1.89	37.35	23.69	31.31	30.78
Zhang et al. (*non-DG setting*)	33.77	31.49	35.17	-	0.87	0.76	0.74	-	1.84	1.83	1.85	-	10.01	16.95	11.92	-
Iizuka et al. [2]	30.11	29.57	32.74	30.81	0.72	0.65	0.72	0.80	1.87	1.86	2.00	1.91	26.00	22.35	22.20	23.52
Iizuka et al. (*non-DG setting*)	33.68	31.66	35.13	-	0.86	0.78	0.82	-	1.88	1.84	1.80	-	14.66	15.61	12.26	-
pix2pix [3]	29.68	29.52	31.94	30.38	0.66	0.63	0.60	0.63	1.77	1.83	1.81	1.80	24.19	26.40	25.26	25.28
pix2pix [3] (*non-DG setting*)	31.60	29.99	33.09	-	0.81	0.68	0.63	-	1.86	1.82	2.00	-	21.69	26.36	15.80	-
pix2pix + DANN [1]	29.62	28.41	32.34	30.12	0.65	0.49	0.64	0.59	1.77	1.44	1.83	1.68	28.74	24.88	22.54	25.39
pix2pix + CORAL [7]	29.84	29.56	32.37	30.59	0.67	0.64	0.64	0.65	1.82	1.81	1.9	1.84	26.88	21.43	25.57	24.63
pix2pix + GroupDRO [6]	29.78	29.46	32.30	30.51	0.66	0.64	0.65	0.65	1.86	1.82	1.77	1.82	32.77	21.07	23.94	25.93
pix2pix + SagNet [5]	29.80	29.56	32.64	30.67	0.68	0.63	0.72	0.68	1.78	1.82	1.82	1.81	28.31	31.86	22.20	27.46
Ours	29.89	29.74	32.51	30.71	0.69	0.66	0.64	0.66	1.78	1.84	2.02	1.88	22.98	31.01	20.78	24.92

Table 2. Colorization performance comparison in terms of four image quality evaluation metrics: PSNR, SSIM, IQM, and FID. *Abbr.*: C (ClipArt), A (Art), and R (RealWorld). Data: Office-Home.

Models	PSNR ↑				SSIM ↑				IQM ↑				FID ↓			
	C	A	R	Avg.	C	A	R	Avg.	C	A	R	Avg.	C	A	R	Avg.
Zhang et al. [9]	35.20	30.78	30.77	32.25	0.75	0.74	0.72	0.74	1.55	1.72	1.61	1.63	15.28	21.50	16.54	17.77
Iizuka et al. [2]	35.41	30.6	32.00	32.67	0.82	0.74	0.80	0.79	1.46	1.50	1.47	1.48	15.15	23.01	13.58	17.25
pix2pix [3]	33.98	31.13	32.96	32.69	0.69	0.73	0.81	0.74	1.65	1.71	1.52	1.63	19.92	23.48	14.84	19.41
pix2pix + DANN [1]	30.01	31.38	31.82	31.07	0.48	0.77	0.76	0.67	1.48	1.61	1.50	1.53	33.14	20.72	13.73	22.53
pix2pix + CORAL [7]	32.59	31.75	32.11	32.15	0.68	0.77	0.75	0.73	1.41	1.72	1.57	1.57	15.92	19.17	13.17	16.09
pix2pix + GroupDRO [6]	33.78	29.75	32.41	31.98	0.76	0.69	0.79	0.75	1.36	1.54	1.52	1.47	15.09	25.12	14.84	18.35
pix2pix + SagNet [5]	34.58	30.89	30.29	31.92	0.68	0.72	0.61	0.67	1.58	1.70	1.55	1.61	16.24	17.73	22.12	18.69
Ours	33.75	31.20	31.93	32.29	0.69	0.74	0.75	0.73	1.43	1.62	1.52	1.52	16.01	16.85	11.89	14.92

Table 3. Evaluation of perceptual realism by a user study. Participants were asked to answer two questions for evaluating naturalness and perceptual realism.

Models	Naturalness ↑	Perceptual Realism ↑
Ours	**41.68%**	**48.35%**
pix2pix + DANN [1]	22.60%	6.66%
pix2pix + CORAL [7]	34.92%	18.01%
pix2pix + GroupDRO [6]	32.08%	13.97%
pix2pix + SagNet [5]	24.07%	13.00%

Fig. 1. Additional colorization performance comparison with conventional colorization approaches. Data: PACS

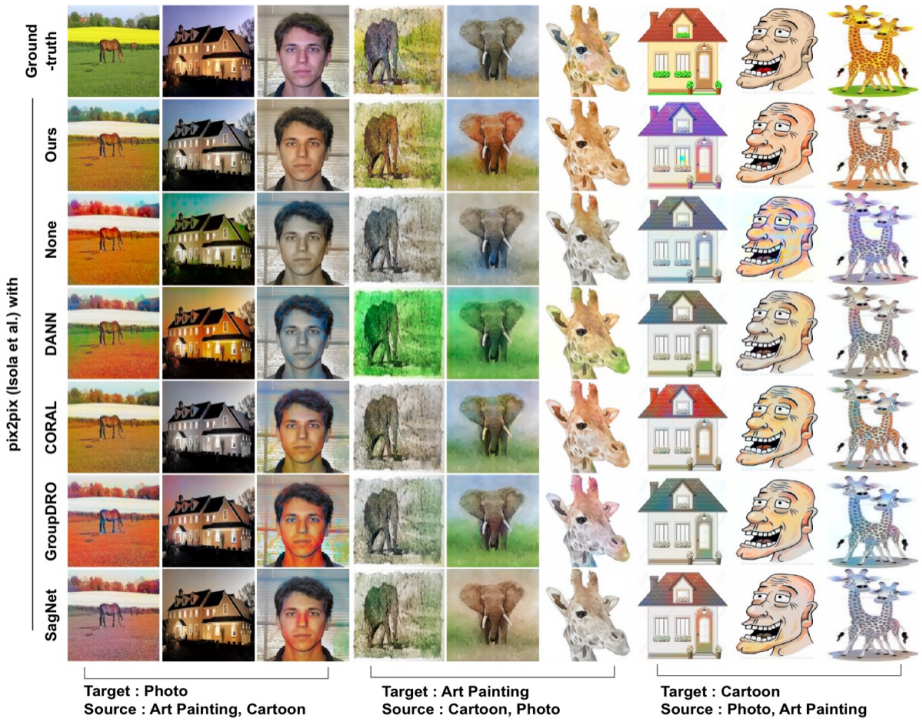

Fig. 2. Additional qualitative colorization performance comparison with four alternative domain generalization techniques. All models are built upon our baseline pix2pix [3] architecture and we add regularization losses to improve the model's generalization power. Data: PACS.

Table 4. Ablation study on the choice of the hyperparameter l. *Abbr.*: P (Photo), A (Art Painting) and C (Cartoon). Data: PACS

Models	PSNR ↑				SSIM ↑				IQM ↑				FID ↓			
	P	A	C	Avg.	P	A	C	Avg.	P	A	C	Avg.	P	A	C	Avg.
$l = 5$	29.89	29.74	32.51	30.71	0.69	0.66	0.64	0.66	1.78	1.84	2.02	1.88	22.98	31.01	20.78	24.92
$l \geq 5$	30.01	29.71	32.70	30.81	0.71	0.66	0.68	0.68	1.77	1.74	2.03	1.84	26.47	31.98	24.38	27.61

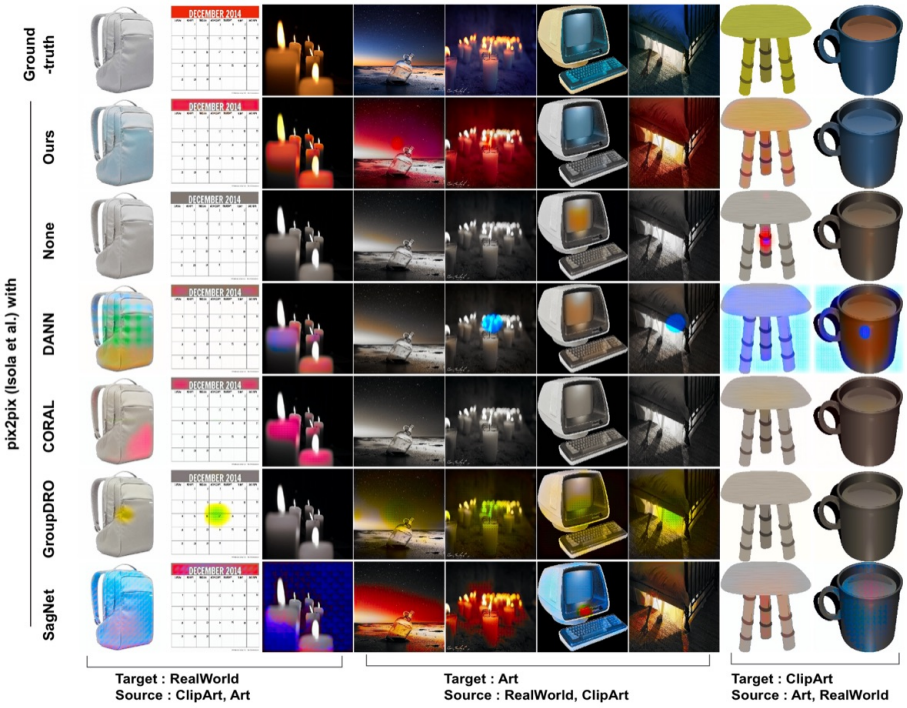

Fig. 3. Additional qualitative colorization performance comparison with four alternative domain generalization techniques. All models are built upon our baseline pix2pix [3] architecture and we add regularization losses to improve the model's generalization power. Data: Office-Home.

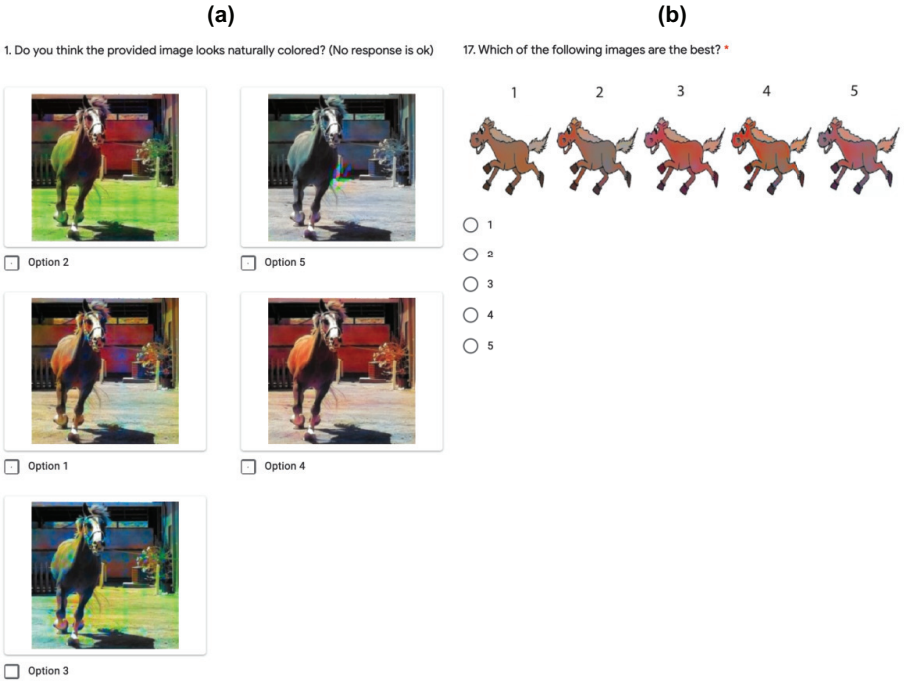

Fig. 4. A sample of the questionnaire for our user study. Two questions are asked for participants to answer: (a) **Q1**(Naturalness: "Do you think the provided image looks naturally colored?") and (b) **Q2**(Perceptual Realism: "Which of the following images are the best?").

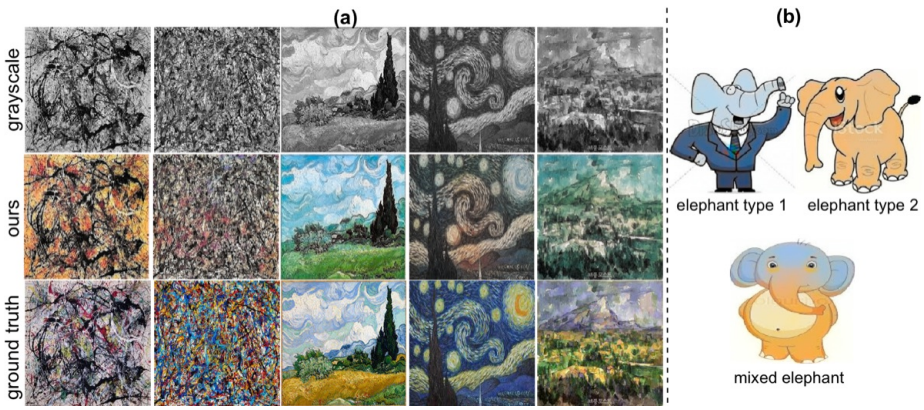

Fig. 5. Failure Cases. In (a), we observe that our model fails to colorize image of large out-of-distribution. In (b), we observe that dealing with the multi-modal distribution of colors is challenging.

Fig. 6. Qualitative Colorization Performance Comparison with z_{AdaIN} (our default setting using the content of z with the style of o_l) and o_{AdaIN} (using the content of o_l with the style of z).

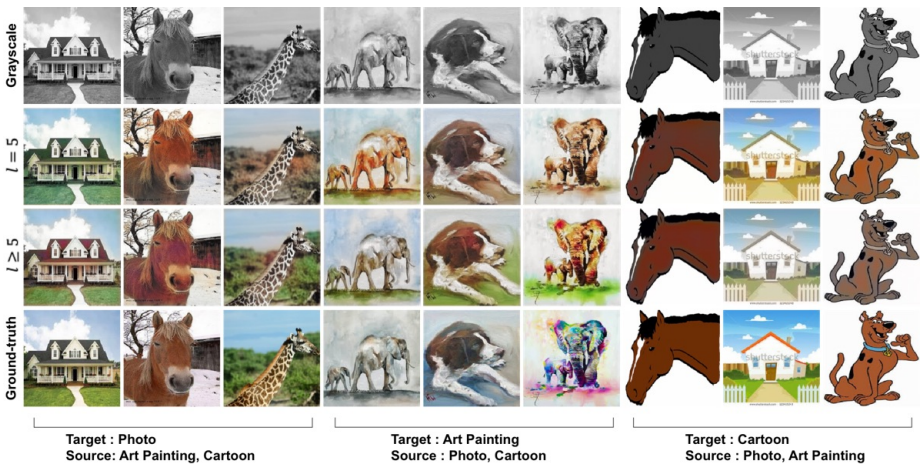

Fig. 7. Ablation study results on the choice of the hyperparameter l. Data: PACS [4]

References

1. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The journal of machine learning research **17**(1), 2096–2030 (2016)
2. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. ACM Transactions on Graphics (ToG) **35**(4), 1–11 (2016)
3. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
4. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
5. Nam, H., Lee, H., Park, J., Yoon, W., Yoo, D.: Reducing domain gap by reducing style bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8690–8699 (2021)
6. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731 (2019)
7. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: European conference on computer vision. pp. 443–450. Springer (2016)
8. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5018–5027 (2017)
9. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: European conference on computer vision. pp. 649–666. Springer (2016)