COAT: Correspondence-driven Object Appearance Transfer Supplementary material

Sangryul Jeon[1]
srjeon@berkeley.edu
Zhifei Zhang[2]
zzhang@adobe.com
Zhe Lin[2]
zlin@adobe.com
Scott Cohen[2]
scohen@adobe.com
Zhihong Ding[2]
zhding6@gmail.com
Kwanghoon Sohn[3]
khsohn@yonsei.ac.kr

[1] ICSI
UC Berkeley
Berkeley, USA

[2] Adobe Research
San Jose, USA

[3] Yonsei University
Seoul, South Korea

1 Implementation Details

FPN [7]-based architecture is employed for the encoder E, and StyleGAN2 [5] is adopted as the decoder D. All the input images are resized to 256×256 before fed into the encoder, and we sample the latent maps to multiple scales, *i.e.*, 64×64, 32×32, and 16×16. For the input indices of the modulation layers in D, 16×16 latent maps are provided to the layers indexing from 1 to 3, 32×32 maps to layers from 4 to 7, and 64×64 maps to layers from 8 to the last (*e.g.*, in our case, the 18th layer is the output layer with the scale of 1024×1024). In the image generation, the original style latent codes are always fed to layers indexing from 10 to the last in D. Therefore, we can set n between 1 and 9 to control texture transfer. In our experiments, $n = 8$ which means that the content latent codes are provided to the modulation layers indexing from 1 to 7, and the aligned style latent codes are fed to layers from 8 to 9. The encoder E_{mod} is constructed by a series of convolutions with the stride of 2 and LeakyReLU activation functions. The temperature $\tau = 0.03$. The threshold $\gamma = 256$.

First, we only learn the parameters of the encoder by freezing the pre-trained StyleGAN2 decoder. In the losses, we set the weights $\alpha = 1$ and $\beta = 0$ for 500k iterations, and then we fine-tune the whole network setting $\beta = 10$ for another 500k iterations.

1.1 Datasets

Here, we provide more details on the employed four datasets.

(1) ***Human Face:*** For training data, all 70,000 images from the FFHQ [4] dataset were used. For evalutions, we use the official test split of the CelebA-HQ [8] dataset, 2,824 test images. Our decoder outputs 1024×1024 resolution image for human face category.
(2) ***Animal Face:*** We used the offical train-test split of the AFHQ [3] wild dataset consisting of 4,738 and 500 images, respectively. Our decoder outputs 512×512 resolution image for animal face category.
(3) ***Car:*** 8,144 images from the training split of the Stanford Cars [6] dataset were used to learn our method. For evaluations, we used randomly selected 1,000 images from the test set due to its large test split (8,041 images). Our decoder outputs 512×384 resolution image for car category.
(4) ***Horse:*** We used the LSUN horse [11] dataset for training and testing images. As the train-test split is not identified, we randomly select 20,000 images to be used for training and 2,000 images for testing. Our decoder outputs 256×256 resolution image for horse category.

1.2 Losses

To learn our networks in an unsupervised manner, we utilized the commonly used reconstruction loss $\mathcal{L}_{rec}$ that encourages the networks to keep the consistency between the original image and predicted one. Specifically, the reconstruction loss, that consists of Mean Square Error (MSE) for pixel-wise similarity and LPIPS [12] for perceptual similarity, is applied to both content and style images such that

$$\mathcal{L}_{rec} = \lambda_{mse}\mathcal{L}_{mse} + \lambda_{LPIPS}\mathcal{L}_{LPIPS}, \tag{1}$$

where

$$\mathcal{L}_{mse} = \sum_{k\in\{1,2\}} \sum_i ||I_i^k - \bar{I}_i^k||_2, \tag{2}$$

$$\mathcal{L}_{LPIPS} = \sum_{k\in\{1,2\}} \sum_i ||P(I^k)_i - P(\bar{I}^k)_i||_2, \tag{3}$$

$\bar{I} = D(E_{\text{mod}}(E(I)))$, and P is the perceptual feature extractor.

We also employed two regularization losses for our encoder that has been shown effective in recent latent space learning literature [2, 10], such that

$$\mathcal{L}_{reg} = \lambda_{avg}\mathcal{L}_{avg} + \lambda_{adv}\mathcal{L}_{adv}. \tag{4}$$

Denoting $\bar{F}$ as the average latent vector of the pretrained StyleGAN2 generator [5], the first loss encourages the extracted latent vectors to be closer to the average one $\bar{F}$

$$\mathcal{L}_{avg} = \sum_{k\in\{1,2,2\to1\}} \sum_l ||E_{\text{mod}}(F^{k,l}) - \bar{F}||_2. \tag{5}$$

Another loss further encourages the individual latent vectors $E_{\text{mod}}(F^l)$ to lie within the distribution of the StyleGAN2 latent space based on the adversarial formulation, such that

$$\mathcal{L}_{adv} = \sum_{k\in\{1,2,2\to1\}} \sum_l \mathcal{L}_E^{k,l} + \mathcal{L}_M^{k,l}, \tag{6}$$

where $\mathcal{L}_E$ and $\mathcal{L}_M$ are the adversarial losses for the encoder E and the discriminator M, respectively as

$$\mathcal{L}_E^{k,l} = -\log M(E_{\text{mod}}(F^{k,l})), \tag{7}$$

$$\mathcal{L}_M^{k,l} = -\log M(\bar{F}) - \log(1 - M(E_{\text{mod}}(F^{k,l}))). \tag{8}$$

The values of balancing parameters are set to

$$\{\lambda_{mse}, \lambda_{LPIPS}, \lambda_{avg}, \lambda_{adv}\} = \{1, 1, 0.0001, 0.1\}$$

.

1.3 Network architecture

Our encoder for modulation E_{mod} consists of a series of 2-strided convolutions with LeakyReLU activations. The discriminator for the adversarial loss in $\mathcal{L}_{adv}$ consists of 4 layer MLP network using LeakyReLU activations.

1.4 User study

As described in Section 4.3 of the main paper, we conducted a user study using Amazon Mechanical Turk. We show the evaluation interfaces in Fig. 1.

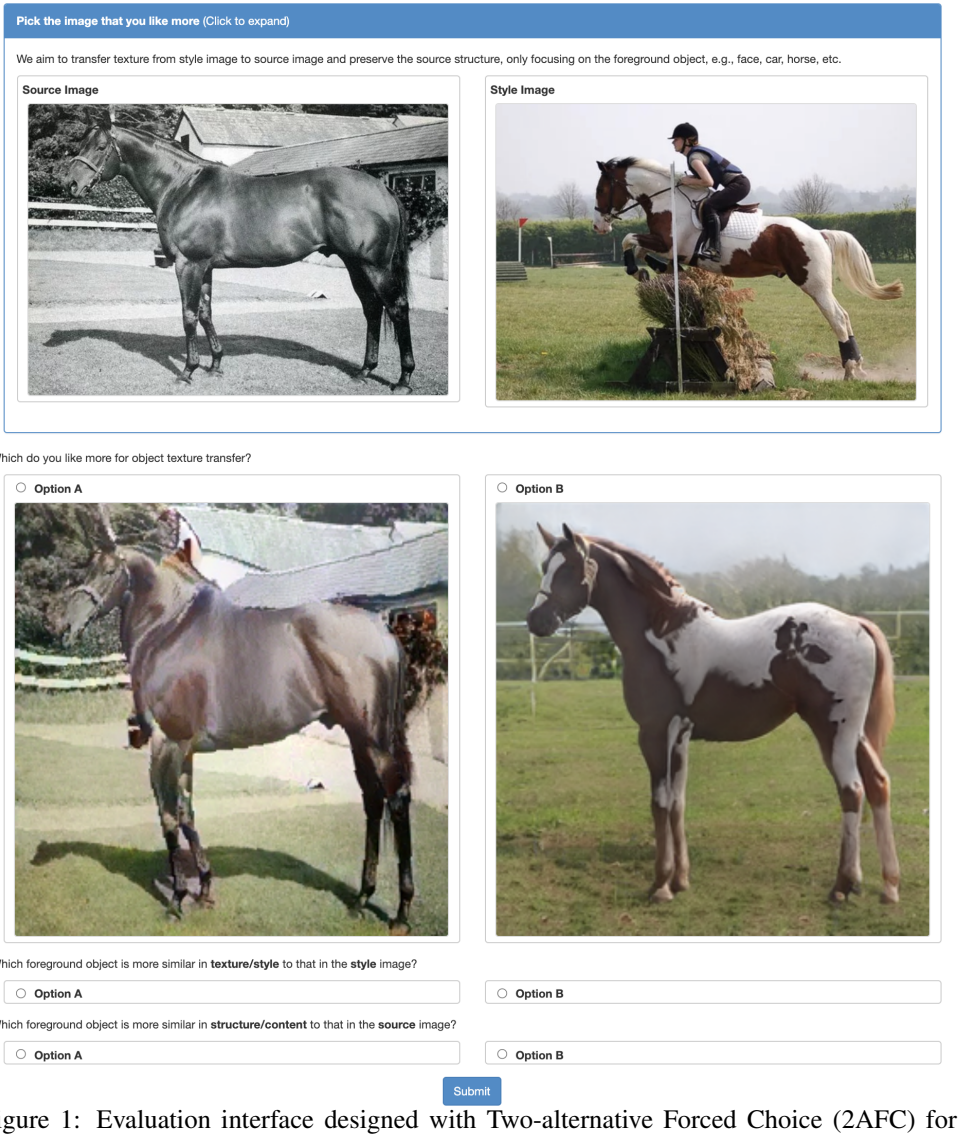

Figure 1: Evaluation interface designed with Two-alternative Forced Choice (2AFC) for measuring the preference on style and content preservation. The participants are also asked to choose which they like better overall.

(a) CelebA-HQ [8] (b) AFHQ [3] (c) Stanford Cars [6] (d) LSUN Horse [11]

Figure 2: Reconstruction results from the proposed COAT on different datasets. In each example pair, the left is the input image, and the right is reconstructed by COAT, which directly feeds encoded latent to the decoder.

Methods	sec ↓	LPIPS ↓			
	Time	CelebA	Car	AFHQ	LSUN
StyleGAN2 [5]	79	0.255	0.385	0.305	0.357
Im2StyleGAN [1]	525	0.161	0.295	0.211	**0.178**
pSp [9]	0.065	0.171	0.287	0.350	0.355
e4e [10]	0.061	0.203	0.315	0.352	0.449
ReStyle [2]	0.132	0.126	0.250	0.211	0.310
COAT	**0.171**	**0.081**	**0.229**	**0.148**	0.202

Table 1: Quantitative comparison of image reconstruction, which is measured by LPIPS [12]. The time indicates average running time per image in second.

2 Additional results

2.1 Image reconstruction

For precise transfer of local textures, it is essential to encode images into the latent space with high fidelity. To this purpose, we validate our method on image reconstruction as compared to the state-of-the-art image generation works, *i.e.*, StyleGAN2 [5] and Im2StyleGAN [1]. Visual examples from our results are shown in Fig. 2, and quantitative comparison measured by LPIPS [12] is reported in Table 1. We compare our approach to the GAN inversion methods based on optimization techniques [1, 5], and the StyleGAN encoders [2, 9, 10].

Our method outperforms the baselines in reconstruction score except for Im2StyleGAN on the Horse dataset. The main reason is the challenging scenarios contained in the LSUN Horse dataset [11], where the objects are mostly unaligned across the instances with non-rigid transformations. Comparing the running time, our COAT runs fast with a single feed-forward pass through the model, while Im2StyleGAN is much more time-consuming since it conducts iterative optimization. Therefore, the better score of Im2StyleGAN is obtained by consuming approximately 2,500 times more computational resources than ours.

2.2 Ablation study on the number of negative samples

We also conduct another ablation study by varying the number of negative samples at level 2. As visualized in Fig. 3, reducing negatives still perform strongly, but utilizing random negatives hurts performance.

2.3 Qualitative results

We uploaded the additional qualitative results on the following link, including the estimated correspondences at each level and the hybrid results in accordance to these correspondences.

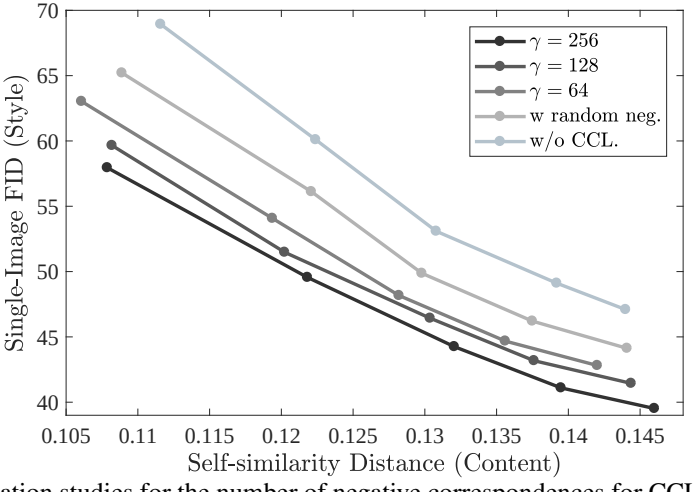

Figure 3: Ablation studies for the number of negative correspondences for CCL. Each curve is created by iterating n, a larger n will result in less texture transfer (large single-image FID) and higher structure fidelity (smaller self-similarity distance). A curve closer to the bottom-left indicate better performance.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN: How to embed images into the StyleGAN latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019.

[2] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021.

[3] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[4] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[5] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[6] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

[7] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[8] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[9] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[10] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *arXiv preprint arXiv:2102.02766*, 2021.

[11] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[12] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.