COAT: Correspondence-driven Object Appearance Transfer

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

Abstract

Semantic correspondence is playing an increasingly important role in photorealistic style transfer, especially on objects with prior structural patterns like faces and cars. Unlike traditional methods that are blind to object/non-object regions and spatial correspondence between objects, we propose a new model called correspondence-driven object appearance transfer (COAT), which leverages correspondence to spatially align texture features to content features at multiple scales. Our model does not require extra supervision like semantic segmentation or body parsing and can be adapted to any given generic object category. More importantly, our multi-scale strategy achieves richer texture transfer, while at the same time preserving the spatial structure of objects in the content image. We further propose the correspondence contrastive loss (CCL) with hard negative mining during the training, boosting appearance transfer with improved disentanglement of structural and textural features. Exhaustive experimental evaluation on various objects demonstrates our superior robustness and visual quality as compared to state-of-the-art works.

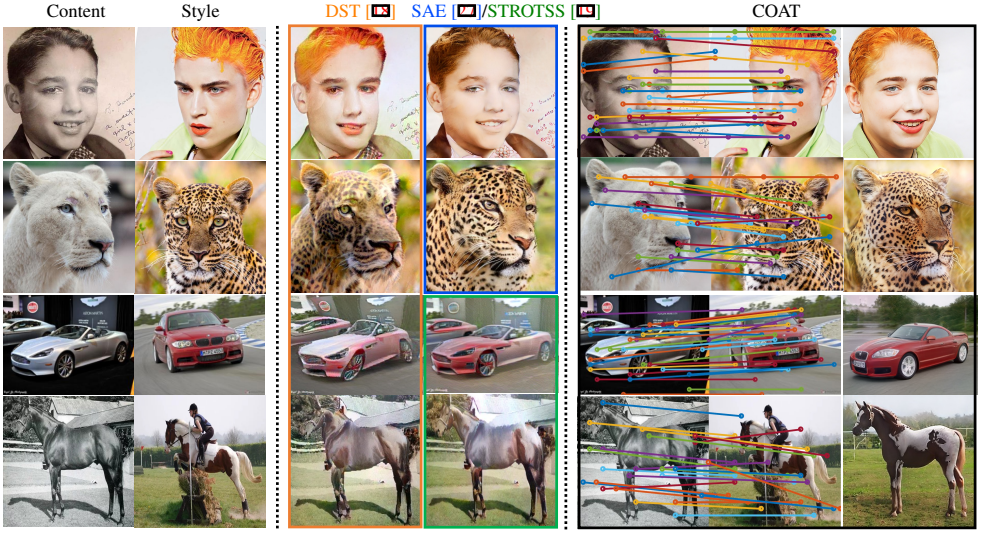


Figure 1: The proposed COAT vs. the state-of-the-art methods. Driven by correspondences estimated between content and style images, we achieve more accurate transfer of fine-grained texture and object appearance, while preserving object structure from the content.

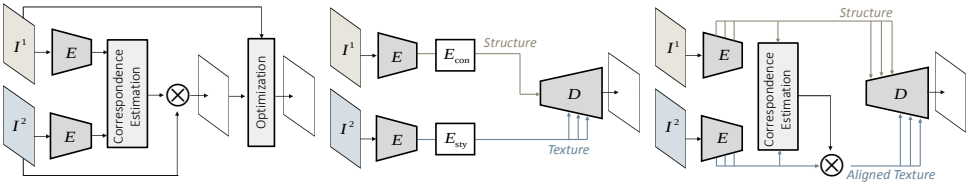
1 Introduction

The great success of recent image-to-image translation techniques [12, 13, 36, 42] enables new classes of image manipulation by injecting the generic visual knowledge drawn from external exemplar images into a target image, *e.g.*, image inpainting [21, 28], photorealistic style transfer [13, 25], object replacement [9, 42], *etc*. These exemplar-driven image manipulation methods traditionally factorize the visual information into two components, *i.e.*, content and style, to re-render the structure of original content image using the style from one or more style exemplars.

A recent work [27] has shown impressive results by encoding images into disentangled structure and style latent code. However, their style code, which is a vector without spatial information, cannot effectively encode local texture information. Therefore, such a global style code may cause mismatching of local textures and distortion of global structures, especially for objects with certain structural patterns. More specifically, the global style code is not rich enough to model fine-grained textures, and it would also stylize the object structure to fit its global texture, which is not desired in object appearance transfer.

A more promising approach to transfer local textures better is to spatially align the style image to the content image before style transfer. There have been several works [18, 22] in this direction, which warp the style image to the content one by pixel-wise correspondences. However, their transfer results heavily rely on the quality of the correspondences. For instance, their performance will degrade drastically if encountering larger appearance and/or geometry variations between content and style inputs.

To alleviate the aforementioned limitations, we propose a new model called correspondence-driven object appearance transfer (COAT), which integrates correspondence estimation and multi-scale style transfer in a unified architecture. By aligning the styles to the content based on the correspondence at multiple scales, we obtain style and content features that



(a) Image warping [18, 23] (b) Latent swapping [27] (c) COAT

Figure 2: Comparison of representative image style transfer methods: (a) image warping method [18, 23], (b) latent swapping method [27], and (c) the proposed COAT. Our key idea is to incorporate correspondence estimation into the hierarchy of encoder and decoder, which would boost fine-grained texture transfer and preserve the structure of the objects in the content image at the same time.

encode more spatial and local textural information than the global styling code in existing works. On the other hand, we utilize the layer-wise representation of StyleGAN [16] where high-level structural information of content feature are fed to the early layers of the decoder and low-level textures from aligned style feature to the late ones, thus achieving both accurate object appearance transfer and robust preservation of content structure against large geometry/appearance variation. Beyond the network design, correspondence also benefits the disentangling of structural and textural features during training. We further propose the Correspondence Contrastive Loss (CCL) for mining hard negative samples that are selected based on correspondences. Hard negatives can help the model avoid coupling of content and style. Extensive experiments are conducted to demonstrate the superior performance of the proposed COAT in terms of visual quality and robustness against large geometry/appearance variation as compared to the state-of-the-art methods.

2 Related Works

Classical approaches for style transfer [7, 11, 19] commonly model the image into style and content components, encouraging them to specifically characterize the given image visually. To this end, several works leveraged modern neural networks [2, 14, 35] to measure content using perceptual distance, and style as global texture statistics, e.g., a Gram matrix. SAE [27] extended this framework with an autoencoder architecture where the images are encoded into style and content latent codes and mixed to generate a hybrid image. Despite of the plausible results, their style code only contains global texture distribution and thus has limited capability of transferring fine-grained local textures. Though some methods [18, 22, 32] alleviated this by leveraging dense semantic correspondences to consider local information, obtaining precise correspondences for all pixels is extremely challenged, causing the structure of the result hybrid image not to be consistent to the content image.

In contrast to these methods, our model can achieve both content preservation and localized stylization at the same time, by incorporating the correspondence estimation within the semantic hierarchy of the autoencoder architecture.

3 Approach

Given a content-style image pair $\{I^1, I^2\}$, we aim to transfer the appearance of the style image I^2 to the content image I^1, while at the same time preserving the high-level structure of the

content image. This objective typically involves two stages: 1) embedding these images into the latent space through encoders, and 2) synthesizing the transferred image through an optimization-based algorithm [19, 37] or a decoder [5, 12].

To achieve more precise transfer of local textures, a possible approach [18, 22] (Fig. 2 (a)) is to first align the content and style images by dense correspondence before generating the transferred image. Starting from extracting the latent maps $\{F^1, F^2\} \in \mathbb{R}^{h\times w\times d}$ through the encoder E, where h, w, and d denote the spatial resolution and channel dimension, pixel-wise similarity score is computed as a cosine distance,

$$S_{i,j} = \frac{F_i^1 \cdot F_j^2}{||F_i^1|| \cdot ||F_j^2||}, \; S \in \mathbb{R}^{(h\times w)\times(h\times w)} \tag{1}$$

where i denotes each pixel on the latent maps. A dense correspondence map is then estimated by applying existing techniques [1, 17] to this similarity matrix to finally align style and content images. Although this method could better localize detailed textures than the previous approaches [7, 19], its performance heavily relies on the quality of correspondence field, which often suffers from large variations of appearance and geometry. The results in later experiments (Figs. 6 and 7) demonstrate that this method will show significant degradation if encountering more drastic geometry variations. By contrast, we utilize correspondence to align features and generate transferred results through a powerful decoder, as illustrated in Fig. 2 (c), which more semantically and thus more robustly transfer appearance from style image to content image.

Meanwhile, a recent work [27] (Fig. 2 (b)) introduced an autoencoder architecture similar to ours. They learn to disentangle the latent code into structure and texture components, thus the structure component will mainly carry spatial information and the texture component will focus on embedding the global texture distribution. The encoders E_{con} and E_{sty} in Fig. 2 (b) aim to extract content and style/texture related features, respectively. However, the high-level texture component mostly embeds the global style rather than capturing rich details from the style image. Therefore, the spatial structure of the final results can be misguided by the global texture component.

To alleviate the aforementioned limitations, we propose COAT (Fig. 2 (c)) which extracts multi-scale features and their correspondences for robust style transfer with aligned texture maps. More specifically, multi-scale features from the content image are considered as the structure features fed to high-level layers of the decoder. On the other hand, multi-scale features from the style image are aligned to obtain aligned texture features, which are skipped to low-level layers of the decoder, providing more fine-grained textures without distorting the spatial structure from the content image. To further enhance the goal of object appearance transfer, *i.e.*, greedy texture transfer without structure distortion of the object, we introduce correspondence contrastive loss (more discussion in Sec. 3.2), where positive and negative samples for a given query patch are determined based on the correspondences rather than random policy widely used in existing works [4, 8, 26].

3.1 Network Design

As shown in Fig. 3, COAT can be divided into three sub-modules: 1) latent extraction that extracts feature maps from input image pair via the encoder E, 2) latent alignment that spatially aligns multi-scale features from the style image to those from the structure input, and 3)

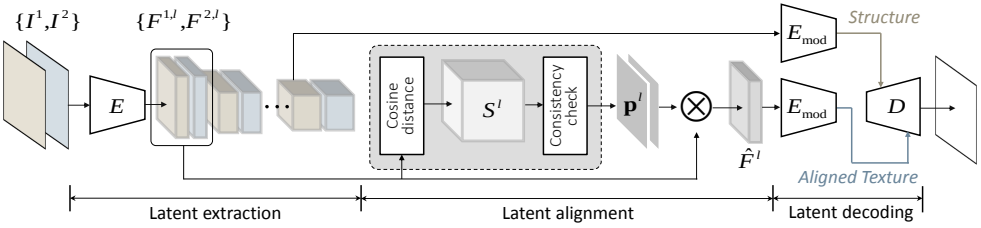
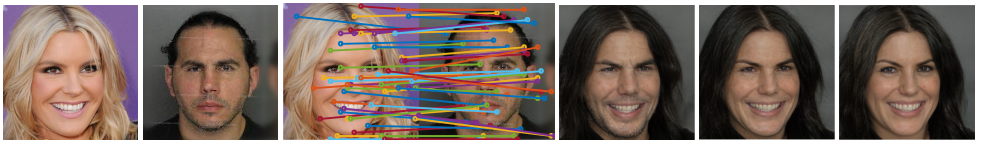


Figure 3: Overview of the proposed COAT, which consists of three sub-modules, *i.e.*, latent extraction, latent alignment, and latent decoding. Image features are extracted by the encoder E, the correspondence estimation (highlighted by gray box) is conducted at multiple scales, and manipulated features are fed to the decode D for generating transferred results.



(a) Content (b) Style (c) Correspondence ($l = 2$) (d) $n = 5$ (e) $n = 7$ (f) $n = 9$

Figure 4: Visualization of texture control by manipulating the layer index n in the decoder D. A smaller n results in more aggressive transfer since richer textural information is passed to the decoder.

latent decoding that consumes the structure features and aligned texture feature to generate appearance transferred image.

Latent extraction accepts an image pair $\{I^1, I^2\}$ and extracts their latents leveraging the hierarchy of the CNN-based encoder E, obtaining multi-scale feature maps, *i.e.*, $\{F^{1,l}\}_{l=1}^{L}$ and $\{F^{2,l}\}_{l=1}^{L}$ from the content and style image, respectively. L denotes the number of scales.

Latent alignment aims to better localize fine-grained textures in the style image. We spatially align the style feature maps to the content feature maps at each scale l by estimating correspondence between $F^{2,l}$ and $F^{1,l}$. To this end, we first compute cosine scores following Eq. 1 to obtain the similarity matrix S^l. With the presence of the large intra-class appearance variations, the encoded representation F is not guaranteed to be accurate because of noisy scores in the similarity matrix S^l. To address this, we apply the widely used soft consistency criterion [4, 24, 30] to S^l, such that the correspondences between two pixels are checked forward and backward to determine whether they are consistently correlated. The similarity score incorporating soft consistency can be expressed in the following.

$$Q^l_{i,j} = \frac{\left(S^l_{i,j}\right)^2}{\max_i S^l_{i,j} \cdot \max_j S^l_{i,j}}, \tag{2}$$

where $S^l_{i,j}$ indicates the score on the l-th scale between the i-th pixel from I^1 and j-th pixel from I^2. The $Q^l_{i,j}$ equals 1 if and only if the match between i and j satisfies the forward-backward consistency constraint, and it will be less than one otherwise. To reduce the effect from noisy correspondences, we collect a set of sparse but highly confident correspondences $\mathbf{p}^l = \{(i,j)|Q^l_{i,j} = 1\}$.

Therefore, the alignment from style feature maps to content feature maps can be achieved

(a) Content image I^1 (b) Style image I^2 (c) Transfer result $I^{2\to1}$

Figure 5: Illustration of the collected negative samples by the proposed correspondence contrastive loss (CCL). The green dots indicate positive samples, and red dots are hard negative samples selected based on their correspondence to the query location (blue dot). The gray dots and boxes shows randomly sampled negatives that are much weaker than our selected negatives (red dots).

by swapping the pixels only in $\mathbf{p}^l$,

$$\hat{F}_i^l = \begin{cases} F_j^{2,l}, & \text{if} \quad (i,j) \in \mathbf{p}^l \\ F_i^{1,l}, & \text{otherwise} \end{cases}, \tag{3}$$

where $\hat{F}^l$ denotes the aligned style feature map at the l-th scale. Compared to the latent swapping method [27] that embeds the global texture distribution into a vector without spatial information, our aligned multi-scale feature maps are able to capture and transfer fine-grained details with higher spatial and visual accuracy.

Latent decoding is conducted with the StyleGAN [16]-based decoder leveraging rich hierarchical semantics in its layer-wise representations. To this end, we formulate a modulation encoder E_{mod} that projects feature maps from the previous step (latent alignment) to the vectors to modulate layer in D. Given multi-scale content feature maps and aligned style feature maps $\{F^{1,l}, \hat{F}^l\}_{l=1}^L$, we synthesize a hybrid image through E_{mod} and D. To preserve the structure from the content image, the content feature maps are fed to the first n modulation layers of D. At the same time, to transfer high-fidelity textures from the style image, the rest layers of D are modulated by the aligned style feature maps. The finally transferred image $I^{2\to1}$ is obtained by

$$I^{2\to1} = D(E_{\text{mod}}(\{F^{1,l}, \hat{F}^l\}_{l=1}^L)). \tag{4}$$

By changing the layer index n in the decoder D, we can provide the content and style feature maps to different layers of D, smoothly controlling the amount of transferred texture as exemplified in Fig. 4. Note that the structure of our transferred image is completely determined by the embedded latent of the content image, *i.e.*, no structural distortion caused by wrong correspondence. This is a key advantage over previous approaches [18, 22], which have difficulties in maintaining the original structure due to the noisy correspondences.

3.2 Losses

Reconstruction and regularization losses are adopted following the literature of latent space learning [29, 34]. The reconstruction loss $\mathcal{L}_{rec}$ includes the commonly used Mean Square Error (MSE) and LPIPS [40], which learn both pixel-wise and perceptual similarities. For the regularization loss $\mathcal{L}_{reg}$, we employ the loss introduced by [34] which encourage the extracted latent vectors to be smoothly distributed within the latent space of StyleGAN [16].

Correspondence Contrastive Loss (CCL) is proposed to further encourage our networks to separate the structure from texture/appearance. CCL introduces a novel contrastive loss that

associates the patches that have a similar structure to each other while disassociating them from other patches although with similar textures. A recent work [26] attempted to use the contrastive setting for unpaired image-to-image translation by collecting negative samples with randomly cropped patches. However, such random negatives cannot efficiently distinguish where the textures we are interested in are located, and they often contain background clutters or occluded regions that would distract the learning process.

Unlike random cropping, the proposed CCL will identify negative samples by ranking the patches based on the similarity scores estimated in the step of latent alignment. More specifically, given a query position i and its positive correspondence $(i, j) \in \mathbf{p}$, negative samples $\mathbf{n}$ are collected with a threshold γ,

$$\mathbf{n}^l(i, j) = \{k | \text{rank}(S^l_{i,j}) > \gamma, k \neq i\}, \tag{5}$$

where $\text{rank}(\cdot)$ returns the rank of values sorting in the descending order. As illustrated in Fig. 5, the collected negative samples consistently capture the relevant textures to the given query position, thereby providing harder negatives than random samples during training.

Finally, we minimize the following objective

$$\mathcal{L}_{CCL} = \sum_{(i,j)\in\mathbf{p}} -\log \frac{C_{i,i}}{\sum_{\mathbf{n}(i,j)} C_{i,\mathbf{n}(i,j)} + C_{i,i}}, \tag{6}$$

where $C_{i,j} = \exp((F^{2\to1}_i \cdot F^1_j)/(\tau \cdot ||F^{2\to1}_i|| \cdot ||F^1_j||))$ and $F^{2\to1} = E(I^{2\to1})$. The superscript l is omitted for readability. Thus, the total loss can be written as a weighted summation of the above three losses

$$\mathcal{L} = \mathcal{L}_{rec} + \alpha\mathcal{L}_{reg} + \beta\mathcal{L}_{CCL}. \tag{7}$$

where α and β are parameters to balance these losses.

4 Experiments

We conduct our evaluations on four datasets, *i.e.*, CelebA-HQ [23] for human face, AFHQ [5] for animal face, Stanford Cars [20] for car, and LSUN [38] for horse. The official train-test splits of the datasets are used in our training and evaluation, except for the human face where the FFHQ [15] dataset was used for training. Note that we attempted to compare COAT to the other correspondence-based transfer methods [39, 41] or StyleGAN2-based image synthesis methods [6, 10, 33], but their settings are different from us, *e.g.* distinct domain between input images, which makes the direct comparison infeasible. Please refer to the supplemental materials for implementation details and more experiments including visual comparisons.

4.1 Results

We evaluate how consistently our method can transfer the texture of the style image while preserving the structure from the content image. However, due to the subjective nature of aesthetic properties defining style and content [18, 19], it is challenging to quantitatively evaluate the performance of appearance transfer. To address this, we follow the protocol in previous works [18, 22, 27], *i.e.*, a human evaluation study using Amazon Mechanical Turk (AMT) designed with Two-alternative Forced Choice (2AFC). More specifically, given our result and that from a certain baseline, the participants are asked to choose which better

(a) Content (b) Style (c)STROT. [19] (d) DST [18] (e) Star. [6] (f) SAE [27] (g)COAT

Figure 6: Visual comparison of appearance transfer on CelebA-HQ [23] and AFHQ [6] datasets. Given input pairs of (a) content and (b) style images, transferred results are obtained from (c) STROTSS [19], (d) DST [18], (e) StarGAN2 [6], (f) SAE [27], and (g) the proposed COAT.

preserves the content and which better transfers the style, respectively. Moreover, we further ask which they like better overall.

As reported in Table 1, the collected 20,000 user votes over five baselines and four datasets demonstrate that our method outperforms all baseline methods in texture transfer. All numbers in the Texture column are greater than 50, which means that over half of the users prefer our results as compared to the other methods. For structure preservation, WCT^2 gets more votes, *i.e.*, 46.9% users vote us and 53.1% vote WCT^2. The reason is that WCT^2 tends to keep the image structure and only change color. Visual comparisons in Fig. 7 demonstrated the limited texture transfer capacity of WCT^2. For the overall evaluation, the proposed COAT outperforms all baselines.

The qualitative comparisons in Figs. 6 and 7 show our advantages over the baselines in object structure preservation and accurate texture transfer. Only encoding global texture distribution like SAE [27] fails to capture fine-grained local texture. For content preservation, the methods based on image warping, *e.g.*, WST [22] and DST [18], are difficult to maintain the structure from content image due to inaccurate correspondence estimation. While our method slightly lags behind WCT^2 that modifies only minor style changes for a photorealistic stylization, the performance of WCT^2 on style preservation is on the contrary far behind our method.

4.2 Ablation Studies

To examine the effects of our key components, *i.e.*, latent alignment and correspondence contrastive loss (CCL), we conduct a series of ablation studies for appearance transfer task on CelebA-HQ dataset [23]. We adopt Self-similarity Distance [19] and Single-Image FID [31] to measure the distance of two images in content and style, respectively. The self-similarity

(a) Content (b) Style (c)WCT^2 [37] (d) STROT. [19] (e) DST [18] (f) WST [23] (g) COAT

Figure 7: Visual comparison of appearance transfer on Stanford Cars [20] and LSUN Horse [38] datasets. Given input pairs of (a) content and (b) style images, transferred results are obtained from (c) WCT^2 [37], (d) STROTSS [19], (e) DST [18], (f) WST [23], and (g) the proposed COAT.

Methods	Preference of **COAT** in terms of		
	Structure	Texture	Overall
DST [18]	69.4	72.5	83.0
WST [23]	77.8	73.1	91.3
WCT^2 [37]	46.9	64.7	63.8
STROTSS [19]	63.3	68.9	73.2
SAE [21]	63.0	57.8	71.8
COAT w/o align	48.1	70.8	69.7
COAT w/o CCL	72.7	63.3	78.7

Table 1: User study for appearance transfer and ablation studies. Each number indicates the percentage of users that prefer our results as compared to the corresponding method in the left column. A number over 50 means our results are visually better.

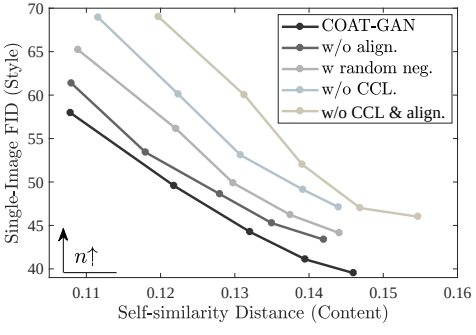


Figure 8: Ablation studies for removing the CCL and/or latent alignment (align). Each curve is created by iterating n, a larger n will result in less texture transfer (large single-image FID) and higher structure fidelity (smaller self-similarity distance). A curve closer to the bottom-left indicate better performance.

Distance computes the self-similarity map of the features extracted from a pretrained network. The single-image FID calculates the Frechet Inception Distance (FID) between two feature distributions of given image pair. We also conducted user studies on the four datasets.

The effects of latent alignment. As reported in the user study as shown in Table 1 (the row of COAT w/o align), the results of our full model is preferred about two times more than the one without latent alignment in texture transfer and overall quality. In Fig. 8, compared to w/o align, the full model achieves lower distances in both single-image FID and self-similarity distance, which demonstrates that using sparse but confident matches can boost the structure preservation and stylization at the same time.

(a) Content (b) Style (c) w/o CCL & align. (d) w/o CCL (e) w/o align. (f) Random neg. (g) COAT

Figure 9: Visual comparison of ablation studies. Given (a) content and (b) style images, the results are generated from models removing (c) both CCL and latent alignment, (d) only CCL, and (e) only latent alignment. In (f), negative samples are randomly selected rather than based on correspondence. (g) is from the full model of the proposed COAT.

The effects of correspondence contrastive loss. We also compare our full model to that trained with randomly collected negatives or trained without CCL. As shown in Table 1 (the row of COAT w/o CCL), Fig. 8, and Fig. 9, the negative samples collected with the guidance of correspondences improves disentanglement of structural and textural components.

The effects of control index n. As reported in Fig. 8, we observe that when the index n increases the single-image FID becomes larger while the self-similarity distance gets smaller. For each curve, both distances change gradually in accordance with the variations of index n. Such gradual visual change can be clearly observed in Fig. 4, and the metrics in Fig. 8 confirm this phenomenon.

5 Conclusion

We proposed the correspondence-driven object appearance transfer GAN (COAT), as well as the correspondence contrastive loss (CCL), to achieve more accurate, robust, and fine-grained object appearance transfer, while at the same time preserving the object structure from the content image. The results of user study and ablation studies demonstrate the effectiveness of our network design and novel training loss. In addition, visual comparison to the state-of-the-art methods on four common objects show the superior performance of COAT in terms of visual quality, texture fidelity, and structure consistency.

6 Acknowledgement

This research was supported by the Yonsei University Researh Fund of 2022 (2022-22-0002).

References

[1] Kfir Aberman, Jing Liao, Mingyi Shi, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Neural best-buddies: Sparse cross-domain correspondence. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018.

[2] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. StyleBank: An explicit representation for neural image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1897–1906, 2017.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[4] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Semantic correspondence with transformers. *arXiv preprint arXiv:2106.02520*, 2021.

[5] 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.

[6] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021.

[7] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[9] Seunghoon Hong, Xinchen Yan, Thomas Huang, and Honglak Lee. Learning hierarchical semantic image manipulation through structured representations. *arXiv preprint arXiv:1808.07535*, 2018.

[10] Jialu Huang, Jing Liao, and Sam Kwong. Semantic example guided image-to-image translation. *IEEE Transactions on Multimedia*, 23:1654–1665, 2020.

[11] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[12] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[15] 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.

[16] 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.

[17] Seungryong Kim, Stephen Lin, Sangryul Jeon, Dongbo Min, and Kwanghoon Sohn. Recurrent transformer networks for semantic correspondence. *arXiv preprint arXiv:1810.12155*, 2018.

[18] Sunnie SY Kim, Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Deformable style transfer. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 246–261. Springer, 2020.

[19] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10051–10060, 2019.

[20] 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.

[21] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.

[22] Xiao-Chang Liu, Yong-Liang Yang, and Peter Hall. Learning to warp for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3702–3711, 2021.

[23] 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.

[24] Juhong Min and Minsu Cho. Convolutional hough matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2940–2950, 2021.

[25] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[26] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020.

[27] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei A Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. *arXiv preprint arXiv:2007.00653*, 2020.

[28] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[29] 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.

[30] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, and J. Sivic. Neighbourhood consensus networks. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, 2018.

[31] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. SinGAN: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4570–4580, 2019.

[32] Yichun Shi, Debayan Deb, and Anil K Jain. WarpGAN: Automatic caricature generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10762–10771, 2019.

[33] Guoxian Song, Linjie Luo, Jing Liu, Wan-Chun Ma, Chunpong Lai, Chuanxia Zheng, and Tat-Jen Cham. Agilegan: stylizing portraits by inversion-consistent transfer learning. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021.

[34] 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.

[35] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, volume 1, page 4, 2016.

[36] Wei Xiong, Yutong He, Yixuan Zhang, Wenhan Luo, Lin Ma, and Jiebo Luo. Fine-grained image-to-image transformation towards visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5840–5849, 2020.

[37] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9036–9045, 2019.

[38] 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.

[39] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020.

[40] 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.

[41] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. Cocosnet v2: Full-resolution correspondence learning for image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11465–11475, 2021.

[42] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.