Cross-domain Correspondence Learning for Exemplar-based Image Translation

Pan Zhang[1] *, Bo Zhang[2], Dong Chen[2], Lu Yuan[3], Fang Wen[2]
[1]University of Science and Technology of China [2]Microsoft Research Asia [3] Microsoft Cloud+AI

Abstract

We present a general framework for exemplar-based image translation, which synthesizes a photo-realistic image from the input in a distinct domain (e.g.*, semantic segmentation mask, or edge map, or pose keypoints), given an exemplar image. The output has the style* (e.g.*, color, texture) in consistency with the semantically corresponding objects in the exemplar. We propose to jointly learn the cross-domain correspondence and the image translation, where both tasks facilitate each other and thus can be learned with weak supervision. The images from distinct domains are first aligned to an intermediate domain where dense correspondence is established. Then, the network synthesizes images based on the appearance of semantically corresponding patches in the exemplar. We demonstrate the effectiveness of our approach in several image translation tasks. Our method is superior to state-of-the-art methods in terms of image quality significantly, with the image style faithful to the exemplar with semantic consistency. Moreover, we show the utility of our method for several applications.*

1. Introduction

Conditional image synthesis aims to generate photo-realistic images based on certain input data [18, 45, 52, 6]. We are interested in a specific form of conditional image synthesis, which converts a semantic segmentation mask, an edge map, and pose keypoints to a photo-realistic image, given an exemplar image, as shown in Figure 1. We refer to this form as *exemplar-based image translation*. It allows more flexible control for multi-modal generation according to a user-given exemplar.

Recent methods directly learn the mapping from a semantic segmentation mask to an exemplar image using neural networks [17, 38, 34, 44]. Most of these methods encode the style of the exemplar into a latent style vector, from which the network synthesizes images with the desired style similar to the examplar. However, the style code only characterizes the global style of the exemplar, regardless of spatial relevant information. Thus, it causes some local style "wash away" in the ultimate image.

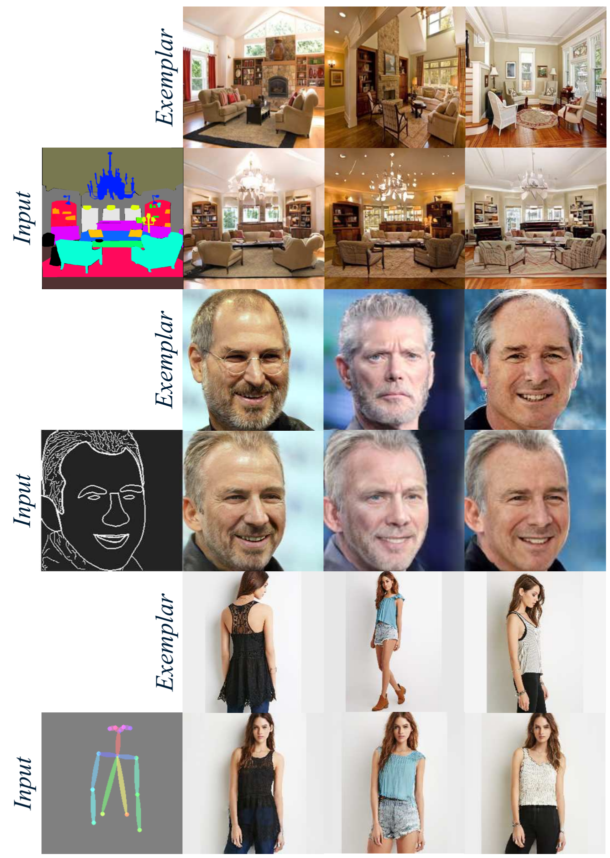

Figure 1: **Exemplar-based image synthesis.** Given the exemplar images (1st row), our network translates the inputs, in the form of segmentation mask, edge and pose, to photo-realistic images (2nd row). Please refer to *supplementary material* for more results.

To address this issue, the *cross-domain correspondence* between the input and the exemplar has to be established before image translation. As an extension of Image Analogies [14], Deep Analogy [27] attempts to find a dense semantically-meaningful correspondence between the image pair. It leverages deep features of VGG pretrained on

*Author did this work during the internship at Microsoft Research Asia.

real image classification tasks for matching. We argue such representation may fail to handle a more challenging mapping from mask (or edge, keypoints) to photo since the pretrained network does not recognize such images. In order to consider the mask (or edge) in the training, some methods [10, 46, 5] explicitly separate the exemplar image into semantic regions and learns to synthesize different parts individually. In this way, it successfully generates high-quality results. However, these approaches are task specific, and are unsuitable for general translation.

How to find a more general solution for *exemplar-based image translation* is non-trivial. We aim to learn the dense semantic correspondence for cross-domain images (*e.g.*, mask-to-image, edge-to-image, keypoints-to-image, etc.), and then use it to guide the image translation. It is weakly supervise learning, since we have neither the correspondence annotations nor the synthesis ground truth given a random exemplar.

In this paper, we propose a *CrOss-domain COrreSpondence network* (*CoCosNet*) that learns cross-domain correspondence and image translation simultaneously. The network architecture comprises two sub-networks: 1) *Cross-domain correspondence Network* transforms the inputs from distinct domains to an intermediate feature domain where reliable dense correspondence can be established; 2) *Translation network*, employs a set of spatially-variant de-normalization blocks [38] to progressively synthesizes the output, using the style details from a warped exemplar which is semantically aligned to the mask (or edge, keypoints map) according to the estimated correspondence. Two sub-networks facilitate each other and are learned end-to-end with novel loss functions. Our method outperforms previous methods in terms of image quality by a large margin, with instance-level appearance being faithful to the exemplar. Moreover, the cross-domain correspondence implicitly learned enables some intriguing applications, such as image editing and makeup transfer. Our contribution can be summarized as follows:

- We address the problem of learning dense cross-domain correspondence with weak supervision—joint learning with image translation.
- With the cross-domain correspondence, we present a general solution to exemplar-based image translation, that for the first time, outputs images resembling the fine structures of the exemplar at instance level.
- Our method outperforms state-of-the-art methods in terms of image quality by a large margin in various application tasks.

2. Related Work

Image-to-image translation The goal of image translation is to learn the mapping between different image domains. Most prominent contemporary approaches solve this problem through conditional generative adversarial network [36] that leverages either paired data [18, 45, 38] or unpaired data [52, 47, 22, 29, 42]. Since the mapping from one image domain to another is inherently multi-modal, following works promote the synthesis diversity by performing stochastic sampling from the latent space [53, 17, 24]. However, none of these methods allow delicate control of the output since the latent representation is rather complex and does not have an explicit correspondence to image style. In contrast, our method supports customization of the result according to a user-given exemplar, which allows more flexible control for multi-modal generation.

Exemplar-based image synthesis Very recently, a few works [39, 44, 34, 40, 2] propose to synthesize photorealistic images from semantic layout under the guidance of exemplars. Non-parametric or semi-parametric approaches [39, 2] synthesize images by compositing the image fragments retrieved from a large database. Mainstream works, however, formulate the problem as image-to-image translation. Huang et al. [17] and Ma et al. [34] propose to employ Adaptive Instance Normalization (AdaIN) [16] to transfer the style code from the exemplar to the source image. Park et al. [38] learn an encoder to map the exemplar image into a vector from which the images are further synthesized. The style consistency discriminator is proposed in [44] to examine whether the image pairs exhibit a similar style. However, this method requires to constitute style consistency image pairs from video clips, which makes it unsuitable for general image translation. Unlike all of the above methods that only transfer the global style, our method transfers the fine style from a semantically corresponding region of the exemplar. Our work is inspired by the recent exemplar-based image colorization [48, 13], but we solve a more general problem: translating images between distinct domains.

Semantic correspondence Early studies [33, 8, 43] on semantic correspondence focus on matching hand-crafted features. With the advent of the convolutional neural network, deep features are proven powerful to represent the high-level semantics. Long et al. [32] first propose to establish semantic correspondence by matching deep features extracted from a pretrained classification model. Following works further improve the correspondence quality by incorporating additional annotations [51, 7, 11, 12, 21, 25], adopting coarse-to-fine strategy [27] or retaining reliable sparse matchings [1]. However, all these methods can only handle the correspondence between natural images instead of cross-domain images, *e.g.*, edge and photorealistic images. We explore this new scenario and implicitly learns the task with weak supervision.

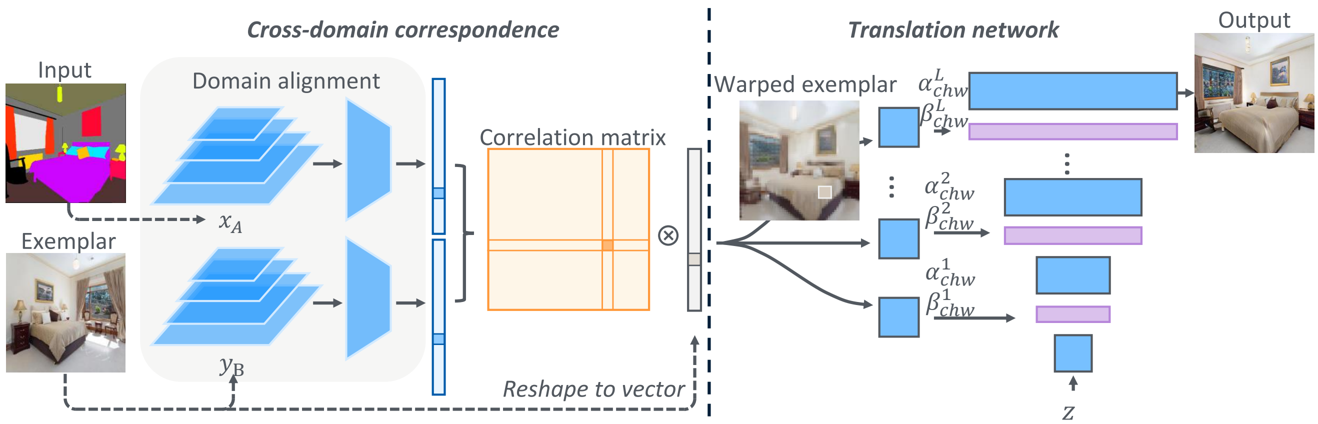

Figure 2: **The illustration of the *CoCosNet* architecture.** Given the input $x_A \in A$ and the exemplar $y_B \in B$, the correspondence submodule adapts them into the same domain S, where dense correspondence can be established. Then, the translation network generates the final output based on the warped exemplar $r_{y \to x}$ according to the correspondence, yielding an exemplar-based translation output.

3. Approach

We aim to learn the translation from the source domain $\mathcal{A}$ to the target domain B given an input image $x_A \in A$ and an exemplar image $y_B \in B$. The generated output is desired to conform to the content as x_A while resembling the style from semantically similar parts in y_B. For this purpose, the correspondence between x_A and y_B, which lie in different domains, is first established, and the exemplar image is warped accordingly so that its semantics is aligned with x_A (Section 3.1). Thereafter, an image is synthesized according to the warped exemplar (Section 3.2). The whole network architecture is illustrated in Figure 2, by the example of mask to image synthesis.

3.1. Cross-domain correspondence network

Usually the semantic correspondence is found by matching patches [27, 25] in the feature domain with a pre-trained classification model. However, pre-trained models are typically trained on a specific type of images, *e.g.*, natural images, so the extracted features cannot generalize to depict the semantics for another domain. Hence, prior works cannot establish the correspondence between heterogeneous images, *e.g.*, edge and photo-realistic images. To tackle this, we propose a novel cross-domain correspondence network, mapping the input domains to a shared domain S in which the representation is capable to represent the semantics for both input domains. As a result, reliable semantic correspondence can be found within domain S.

Domain alignment As shown in Figure 2, we first adapt the input image and the exemplar to a shared domain S. To be specific, x_A and y_B are fed into the feature pyramid network that extracts multi-scale deep features by leveraging both local and global image context [41, 28]. The extracted feature maps are further transformed to the representations in S, denoted by $x_S \in \mathbb{R}^{HW \times C}$ and $y_S \in \mathbb{R}^{HW \times C}$ respectively (H,W are feature spatial size; C is the channel-wise dimension). Let $\mathcal{F}_{A \to S}$ and $\mathcal{F}_{B \to S}$ be the domain transformation from the two input domains respectively, so the adapted representation can be formulated as,

$$x_S = \mathcal{F}_{A \to S}(x_A; \theta_{\mathcal{F},A \to S}), \tag{1}$$

$$y_S = \mathcal{F}_{B \to S}(y_B; \theta_{\mathcal{F},B \to S}). \tag{2}$$

where θ denotes the learnable parameter. The representation x_S and y_S comprise discriminative features that characterize the semantics of inputs. Domain alignment is, in practice, essential for correspondence in that only when x_S and y_S reside in the same domain can they be further matched with some similarity measure.

Correspondence within shared domain We propose to match the features of x_S and y_S with the correspondence layer proposed in [48]. Concretely, we compute a correlation matrix $\mathcal{M} \in \mathbb{R}^{HW \times HW}$ of which each element is a pairwise feature correlation,

$$\mathcal{M}(u, v) = \frac{\hat{x}_S(u)^T \hat{y}_S(v)}{\|\hat{x}_S(u)\| \, \|\hat{y}_S(v)\|}, \tag{3}$$

where $\hat{x}_S(u)$ and $\hat{y}_S(v) \in \mathbb{R}^C$ represent the channel-wise centralized feature of x_S and y_S in position u and v, *i.e.*, $\hat{x}_S(u) = x_S(u) - \text{mean}(x_S(u))$ and $\hat{y}_S(v) = y_S(v) - \text{mean}(y_S(v))$. $\mathcal{M}(u, v)$ indicates a higher semantic similarity between $x_S(u)$ and $y_S(v)$.

Now the challenge is how to learn the correspondence without direct supervision. Our idea is to jointly train with image translation. The translation network may find it easier to generate high-quality outputs only by referring to the correct corresponding regions in the exemplar, which implicitly pushes the network to learn the accurate correspondence. In light of this, we warp y_B according to $\mathcal{M}$ and obtain the warped exemplar $r_{y\to x} \in \mathbb{R}^{HW}$. Specifically, we obtain $r_{y\to x}$ by selecting the most correlated pixels in y_B and calculating their weighted average,

$$r_{y\to x}(u) = \sum_v \underset{v}{\mathrm{softmax}}(\alpha \mathcal{M}(u,v)) \cdot y_B(v). \quad (4)$$

Here, α is the coefficient that controls the sharpness of the softmax and we set its default value as 100. In the following, images will be synthesized conditioned on $r_{y\to x}$ and the correspondence network, in this way, learns its assignment with indirect supervision.

3.2. Translation network

Under the guidance of $r_{y\to x}$, the translation network $\mathcal{G}$ transforms the constant code z to the desired output $\hat{x}_B \in B$. In order to preserve the structural information of $r_{y\to x}$, we employ the spatially-adaptive denormalization (SPADE) block [38] to project the spatially variant exemplar style to different activation locations. As shown in Figure 2, the translation network has L layers with the exemplar style progressively injected. As opposed to [38] which computes layer-wise statistics for batch normalization (BN), we empirically find the normalization that computes the statistics at each spatial position, the positional normalization (PN) [26], better preserves the structure information synthesized in prior layers. Hence, we propose to marry positional normalization and spatially-variant denormalization for high-fidelity texture transfer from the exemplar.

Formally, given the activation $F^i \in \mathbb{R}^{C_i \times H_i \times W_i}$ before the i^{th} normalization layer, we inject the exemplar style through,

$$\alpha^i_{h,w}(r_{y\to x}) \times \frac{F^i_{c,h,w} - \mu^i_{h,w}}{\sigma^i_{h,w}} + \beta^i_{h,w}(r_{y\to x}), \quad (5)$$

where the statistic value $\mu^i_{h,w}$ and $\sigma^i_{h,w}$ are calculated exclusively across channel direction compared to BN. The denormalization parameter α^i and β^i characterize the style of the exemplar, which is mapped from $r_{y\to x}$ with the projection $\mathcal{T}$ parameterized by $\theta_\mathcal{T}$, *i.e.*,

$$\alpha^i, \beta^i = \mathcal{T}_i(r_{y\to x}; \theta_\mathcal{T}). \quad (6)$$

We use two plain convolutional layers to implement $\mathcal{T}$ so α and β have the same spatial size as $r_{y\to x}$. With the style modulation for each normalization layer, the overall image translation can be formulated as

$$\hat{x}_B = \mathcal{G}(z, \mathcal{T}_i(r_{y\to x}; \theta_\mathcal{T}); \theta_\mathcal{G}), \quad (7)$$

where $\theta_\mathcal{G}$ denotes the learnable parameter.

3.3. Losses for exemplar-based translation

We jointly train the cross-domain correspondence along with image synthesis with following loss functions, hoping the two tasks benefit each other.

Losses for pseudo exemplar pairs We construct exemplar training pairs by utilizing paired data $\{x_A, x_B\}$ that are semantically aligned but differ in domains. Specifically, we apply random geometric distortion to x_B and get the distorted image $x'_B = h(x_B)$, where h denotes the augmentation operation like image warping or random flip. When x'_B is regarded as the exemplar, the translation of x_A is expected to be its counterpart x_B. In this way, we obtain pseudo exemplar pairs. We propose to penalize the difference between the translation output and the ground truth x_B by minimizing the *feature matching loss* [19, 18, 6]

$$\mathcal{L}_{feat} = \sum_l \lambda_l \left\| \phi_l(\mathcal{G}(x_A, x'_B)) - \phi_l(x_B) \right\|_1, \quad (8)$$

where ϕ_l represents the activation of layer l in the pretrained VGG-19 model and λ_l balance the terms.

Domain alignment loss We need to make sure the transformed embedding x_S and y_S lie in the same domain. To achieve this, we once again make use of the image pair $\{x_A, x_B\}$, whose feature embedding should be aligned exactly after domain transformation:

$$\mathcal{L}^{\ell_1}_{domain} = \left\| \mathcal{F}_{A\to S}(x_A) - \mathcal{F}_{B\to S}(x_B) \right\|_1. \quad (9)$$

Note that we perform channel-wise normalization as the last layer of $\mathcal{F}_{A\to S}$ and $\mathcal{F}_{B\to S}$ so minimizing this domain discrepancy will not lead to a trivial solution (*i.e.*, small magnitude of activations).

Exemplar translation losses The learning with pair or pseudo exemplar pair is hard to generalize to general cases where the semantic layout of exemplar differs significantly from the source image. To tackle this, we propose the following losses.

First, the ultimate output should be consistent with the semantics of the input x_A, or its counterpart x_B. We thereby penalize the *perceptual loss* to minimize the semantic discrepancy:

$$\mathcal{L}_{perc} = \left\| \phi_l(\hat{x}_B) - \phi_l(x_B) \right\|_1. \quad (10)$$

Here we choose ϕ_l to be the activation after $relu4\_2$ layer in the VGG-19 network since this layer mainly contains high-level semantics.

On the other hand, we need a loss function that encourages $\hat{x}_B$ to adopt the appearance from the semantically corresponding patches from y_B. To this end, we employ the *contextual loss* proposed in [35] to match the statistics between $\hat{x}_B$ and y_B, which is

$$\mathcal{L}_{context} = \sum_l \omega_l \left[-\log \left(\frac{1}{n_l} \sum_i \max_j A^l(\phi_i^l(\hat{x}_B), \phi_j^l(y_B)) \right) \right], \quad (11)$$

where i and j index the feature map of layer ϕ^l that contains n_l features, and ω_l controls the relative importance of different layers. Still, we rely on pretrained VGG features. As opposed to $\mathcal{L}_{perc}$ which mainly utilizes high-level features, the contextual loss uses $relu2\_2$ up to $relu5\_2$ layers since low-level features capture richer style information (*e.g.*, color or textures) useful for transferring the exemplar appearance.

Correspondence regularization Besides, the learned correspondence should be cycle consistent, *i.e.*, the image should match itself after forward-backward warping, which is

$$\mathcal{L}_{reg} = \left\| r_{y \to x \to y} - y_B \right\|_1, \quad (12)$$

where $r_{y \to x \to y}(v) = \sum_u \mathrm{softmax}_u(\alpha \mathcal{M}(u,v)) \cdot r_{y \to x}(u)$ is the forward-backward warping image. Indeed, this objective function is crucial because the rest loss functions, imposed at the end of the network, are weak supervision and cannot guarantee that the network learns a meaningful correspondence. Figure 9 shows that without $\mathcal{L}_{reg}$ the network fails to learn the cross-domain correspondence correctly although it is still capable to generate plausible translation result. The regularization $\mathcal{L}_{reg}$ enforces the warped image $r_{y \to x}$ remain in domain B by constraining its backward warping, implicitly encouraging the correspondence to be meaningful as desired.

Adversarial loss We train a discriminator [9] that discriminates the translation outputs and the real samples of domain B. Both the discriminator $\mathcal{D}$ and the translation network $\mathcal{G}$ are trained alternatively until synthesized images look indistinguishable to real ones. The adversarial objectives of $\mathcal{D}$ and $\mathcal{G}$ are respectively defined as:

$$\begin{aligned} \mathcal{L}_{adv}^{\mathcal{D}} &= -\mathbb{E}[h(\mathcal{D}(y_B))] - \mathbb{E}[h(-\mathcal{D}(\mathcal{G}(x_A, y_B)))] \\ \mathcal{L}_{adv}^{\mathcal{G}} &= -\mathbb{E}[\mathcal{D}(\mathcal{G}(x_A, y_B))], \end{aligned} \quad (13)$$

where $h(t) = \min(0, -1 + t)$ is a hinge function used to regularize the discriminator [49, 3].

Total loss In all, we optimize the following objective,

$$\begin{aligned} \mathcal{L}_\theta = \min_{\mathcal{F},\mathcal{T},\mathcal{G}} \max_{\mathcal{D}} \; & \psi_1 \mathcal{L}_{feat} + \psi_2 \mathcal{L}_{perc} + \psi_3 \mathcal{L}_{context} \\ & + \psi_4 \mathcal{L}_{adv}^{\mathcal{G}} + \psi_5 \mathcal{L}_{domain}^{\ell_1} + \psi_6 \mathcal{L}_{reg}, \end{aligned} \quad (14)$$

where weights ψ are used to balance the objectives.

Table 1: **Image quality comparison.** Lower FID or SWD score indicates better image quality. The best scores are highlighted.

	ADE20k		ADE20k-*outdoor*		CelebA-HQ		DeepFashion	
	FID	SWD	FID	SWD	FID	SWD	FID	SWD
Pix2pixHD	81.8	35.7	97.8	34.5	62.7	43.3	25.2	**16.4**
SPADE	33.9	19.7	63.3	21.9	31.5	26.9	36.2	27.8
MUNIT	129.3	97.8	168.2	126.3	56.8	40.8	74.0	46.2
SIMS	N/A	N/A	67.7	27.2	N/A	N/A	N/A	N/A
EGSC-IT	168.3	94.4	210.0	104.9	29.5	23.8	29.0	39.1
Ours	**26.4**	**10.5**	**42.4**	**11.5**	**14.3**	**15.2**	**14.4**	17.2

Table 2: **Comparison of semantic consistency.** The best scores are highlighted.

	ADE20k	ADE20k-*outdoor*	CelebA-HQ	DeepFashion
Pix2pixHD	0.833	0.848	0.914	0.943
SPADE	0.856	0.867	0.922	0.936
MUNIT	0.723	0.704	0.848	0.910
SIMS	N/A	0.822	N/A	N/A
EGSC-IT	0.734	0.723	0.915	0.942
Ours	**0.862**	**0.873**	**0.949**	**0.968**

4. Experiments

Implementation We use Adam [23] solver with $\beta_1 = 0, \beta_2 = 0.999$. Following the TTUR [15], we set imbalanced learning rates, $1e{-}4$ and $4e{-}4$ respectively, for the generator and discriminator. Spectral normalization [37] is applied to all the layers for both networks to stabilize the adversarial training. Readers can refer to the supplementary material for detailed network architecture. We conduct experiments using 8 32GB Tesla V100 GPUs, and it takes roughly 4 days to train 100 epochs on the ADE20k dataset [50].

Datasets We conduct experiments on multiple datasets with different sorts of image representation. All the images are resized to 256×256 during training.

- ADE20k [50] consists of ∼20k training images, each image associated with a 150-class segmentation mask. This is a challenging dataset for most existing methods due to its large diversity.
- ADE20k-*outdoor* contains the outdoor images extracted from ADE20k, as the same protocol in SIMS [39].
- CelebA-HQ [30] contains high quality face images. We connect the face landmarks for face region, and use Canny edge detector to detect edges in the background. We perform an edge-to-face translation on this dataset.
- Deepfashion [31] consists of 52,712 person images in fashion clothes. We extract the pose keypoints using the OpenPose [4], and learn the translation to human body.

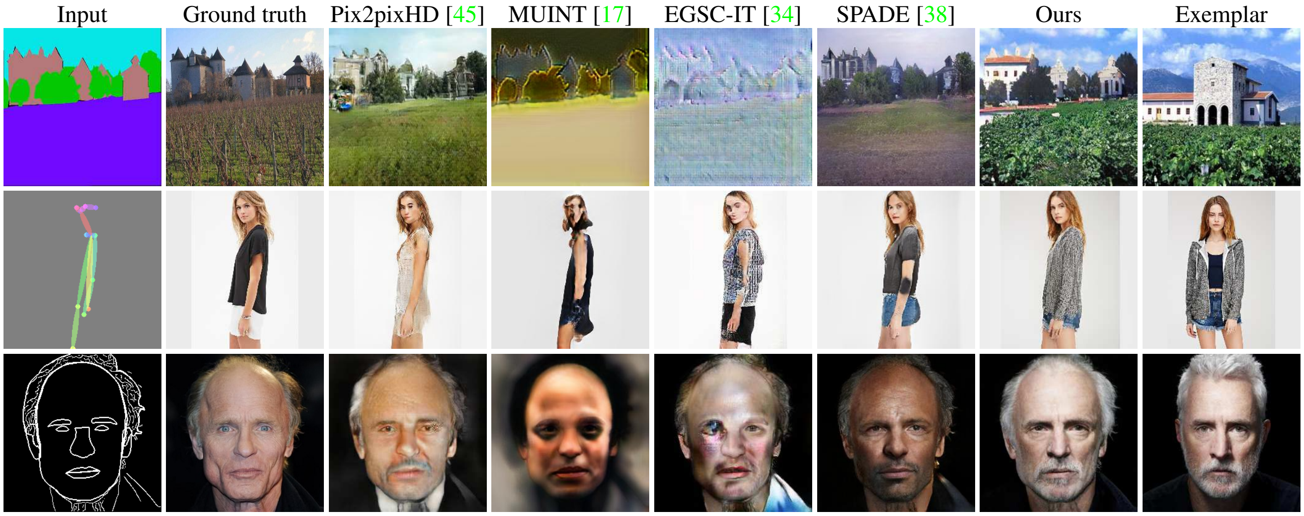

Figure 3: **Qualitative comparison of different methods.**

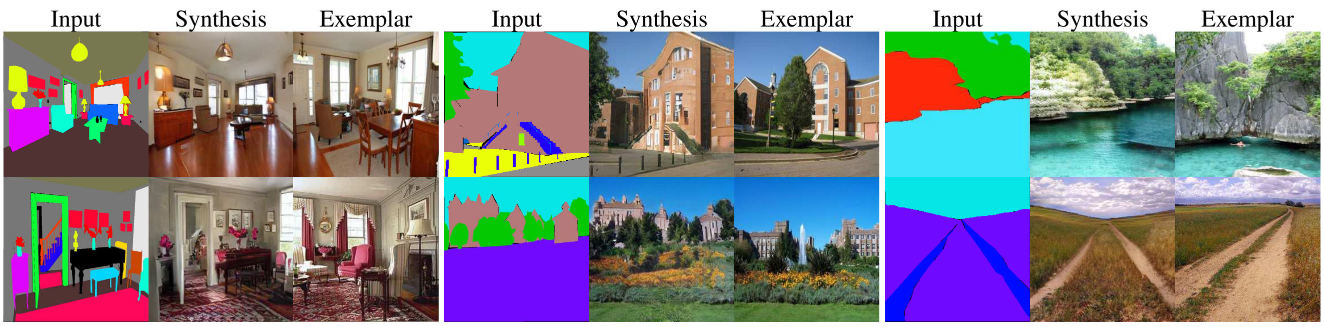

Figure 4: **Our results of segmentation mask to image synthesis (ADE20k dataset).**

Table 3: **Comparison of style relevance.** A higher score indicates a higher appearance similarity relative to the exemplar. The best scores are highlighted.

	ADE20k		CelebA-HQ		DeepFashion	
	Color	Texture	Color	Texture	Color	Texture
SPADE	0.874	0.892	0.955	0.927	0.943	0.904
MUNIT	0.745	0.782	0.939	0.884	0.893	0.861
EGSC-IT	0.781	0.839	0.965	0.942	0.945	0.916
Ours	**0.962**	**0.941**	**0.977**	**0.958**	**0.982**	**0.958**

Baselines We compare our method with state-of-the-art image translation methods: 1) Pix2pixHD [45], a leading supervised approach; 2) SPADE [38], a recently proposed supervised translation method which also supports the style injection from an exemplar image; 3) MUNIT [17], an unsupervised method that produces multi-modal results; 4) SIMS [39], which synthesizes images by compositing image segments from a memory bank; 5) EGSC-IT [34], an exemplar-based method that also considers the semantic consistency but can only mimic the global style. These methods except Pix2pixHD can generate exemplar-based results, and we use their released codes in this mode to train on several datasets. Since it is computationally prohibitive to prepare a database using SIMS, we directly use their reported figures. As we aim to propose a general translation framework, we do not include other task-specific methods. To provide the exemplar for our method, we first train a plain translation network to generate natural images and use them to retrieve the exemplars from the dataset.

Quantitative evaluation We evaluate different methods from three aspects.

- We use two metrics to measure image quality. First, we use the Fréchet Inception Score (FID) [15] to measure the distance between the distributions of synthesized images and real images. While FID measures the semantic realism, we also adopt sliced Wasserstein distance (SWD) [20] to measure their statistical distance of low-level patch distributions. Measured by these two metrics, Table 1 shows that our method significantly outperforms prior methods in almost all the comparisons. Our method improves the FID score by 7.5 compared to previous leading methods on the challenging ADE20k dataset.

Figure 5: **Our results of edge to face synthesis (CelebA-HQ dataset).** First row: exemplars. Second row: our results.

Figure 6: **Our results of pose to body synthesis (DeepFashion).** First row: exemplars. Second row: our results.

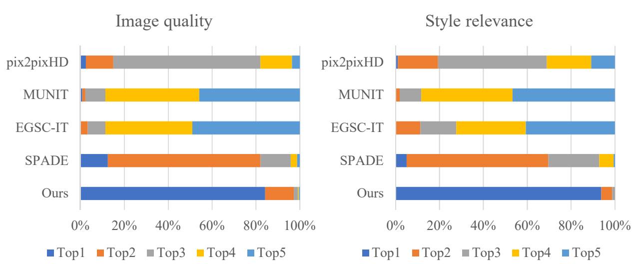

Figure 7: **User study results.**

- The ultimate output should not alter the input semantics. To evaluate the semantic consistency, we adopt an ImageNet pretrained VGG model [3], and use its high-level features maps, $relu3\_2$, $relu4\_2$ and $relu5\_2$, to represent high-level semantics. We calculate the cosine similarity for these layers and take the average to yield the final score. Table 2 shows that our method best maintains the semantics during translation.
- Style relevance. We use low level features $relu1\_2$ and $relu2\_2$ respectively to measure the color and texture distance between the semantically corresponding patches in the output and the exemplar. We do not include Pix2pixHD as it does not produce an exemplar-based translation. Still, our method achieves considerably better instance-level style relevance as shown in Table 3.

Qualitative comparison Figure 3 provides a qualitative comparison of different methods. It shows that our *CocosNet* demonstrates the most visually appealing quality

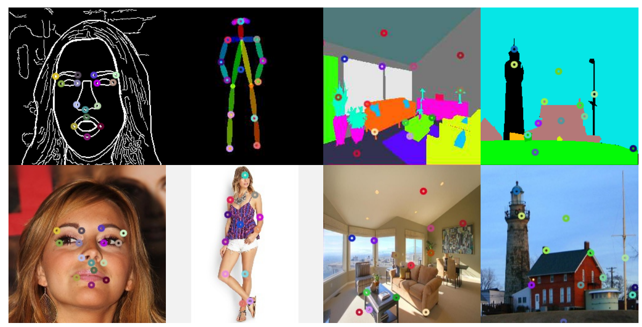

Figure 8: **Sparse correspondence of different domains.** Given the manual annotation points in domain A (first row), our method finds their corresponding points in domain B (second row).

Table 4: **Ablation study.**

	FID ↓	Semantic consistency ↑	Style (color/texture) ↑
w/o $\mathcal{L}_{feat}$	14.4	0.948	0.975 / 0.955
w/o $\mathcal{L}_{domain}$	21.1	0.933	**0.983** / 0.957
w/o $\mathcal{L}_{perc}$	59.3	0.852	0.971 / 0.852
w/o $\mathcal{L}_{context}$	28.4	0.931	0.954 / 0.948
w/o $\mathcal{L}_{reg}$	19.3	0.929	0.981 / 0.951
Full	**14.3**	**0.949**	0.977 / **0.958**

with much fewer artifacts. Meanwhile, compared to prior exemplar-based methods, our method demonstrates the best style fidelity, with the fine structures matching the semantically corresponding regions of the exemplar. This also correlates with the quantitative results, showing the obvious advantage of our approach. We show diverse results by changing the exemplar image in Figure 4-6. Please refer to the supplementary material for more results.

Subjective evaluation We also conduct user study to compare the subjective quality. We randomly select 10 images for each task, yielding 30 images in total for comparison. We design two tasks, and let users sort all the methods in terms of the image quality and the style relevance. Figure 7 shows the results, where our method demonstrates a clear advantage. Our method ranks the first in 84.2% cases in

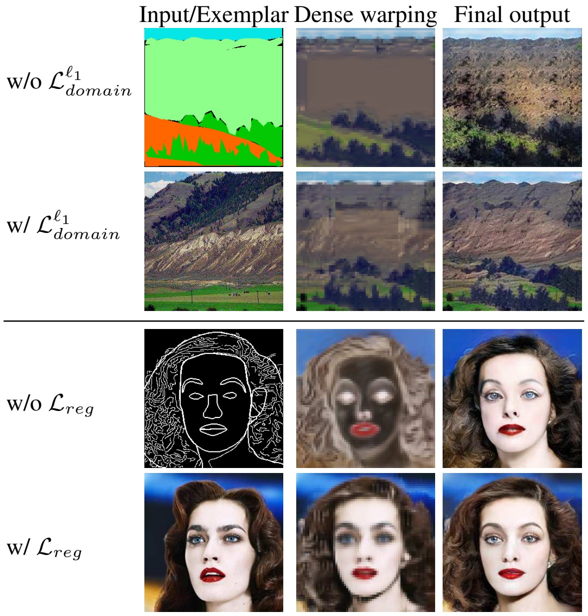

Figure 9: **Ablation study of loss functions.**

evaluating the image quality, with 93.8% chance to be the best in the style relevance comparison.

Cross-domain correspondence Figure 8 shows the cross-domain correspondence. For better visualization, we just annotate the sparse points. As the first approach in doing this, our *CoCosNet* successfully establishes meaningful semantic correspondence which is even difficult for manual labeling. The network is still capable to find the correspondence for sparse representation such as edge map, which captures little explicit semantic information.

Ablation study In order to validate the effectiveness of each component, we conduct comprehensive ablation studies. Here we want to emphasize two key elements (Figure 9). First, the domain alignment loss $\mathcal{L}_{domain}^{\ell_1}$ with data pairs x_A and x_B is crucial. Without it, the correspondence will fail in unaligned domains, leading to oversmooth dense warping. We also ablate the correspondence regularization loss $\mathcal{L}_{reg}$, which leads to incorrect dense correspondence, *e.g.*, face to hair in Figure 9, though the network still yields plausible final output. With $\mathcal{L}_{reg}$, the correspondence becomes meaningful, which facilitates the image synthesis as well. We also quantitatively measures the role of different losses in Table 4 where the full model demonstrates the best performance in terms of all the metrics.

5. Applications

Our method can enable a few intriguing applications. Here we give two examples.

Image editing Given a natural image, we can manipulate its content by modifying the segmentation layout and synthesizing the image by using the original image as the self-exemplar. Since this is similar to the pseudo exemplar pairs we constitute for training, our *CocosNet* could perfectly handle it and produce the output with high quality. Figure 10 illustrates the image editing, where one can move, add and delete instances.

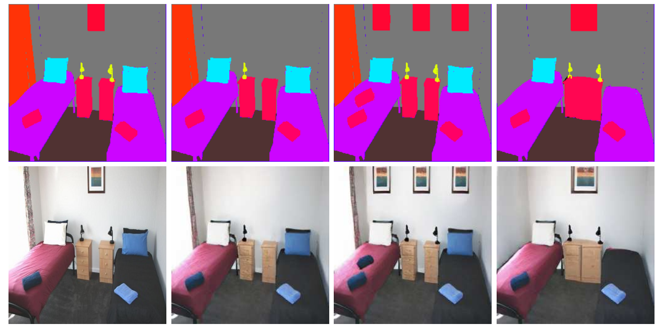

Figure 10: **Image editing.** Given the input image and its mask (1st column), we can semantically edit the image content through the manipulation on the mask (column 2-4).

Figure 11: **Makeup transfer.** Given a portrait and makeup strokes (1st column), we can transfer these makeup edits to other portraits by matching the semantic correspondence. We show more examples in the supplementary material.

Makeup transfer Artists usually manually adds digital makeup on portraits. Because of the dense semantic correspondence we find, we can transfer the artistic stokes to other portraits. In this way, one can manually add makeup edits on one portrait, and use our network to process a large batch of portraits automatically based on the semantic correspondence, which illustrated in Figure 11.

6. Conclusion

In this paper, we present the *CocosNet*, which translates the image by relying on the cross-domain correspondence. Our method achieves preferable performance than leading approaches both quantitatively and qualitatively. Besides, our method learns the dense correspondence for cross-domain images, paving a way for several intriguing applications. Our method is computationally intensive and we leave high-resolution synthesis to future work.

References

[1] K. Aberman, J. Liao, M. Shi, D. Lischinski, B. Chen, and D. Cohen-Or, "Neural best-buddies: Sparse cross-domain correspondence," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, p. 69, 2018. 2

[2] A. Bansal, Y. Sheikh, and D. Ramanan, "Shapes and context: In-the-wild image synthesis & manipulation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2317–2326. 2

[3] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018. 5, 7

[4] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1812.08008*, 2018. 5

[5] H. Chang, J. Lu, F. Yu, and A. Finkelstein, "Pairedcyclegan: Asymmetric style transfer for applying and removing makeup," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 40–48. 2

[6] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1511–1520. 1, 4

[7] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker, "Universal correspondence network," in *Advances in Neural Information Processing Systems*, 2016, pp. 2414–2422. 2

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005. 2

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680. 5

[10] S. Gu, J. Bao, H. Yang, D. Chen, F. Wen, and L. Yuan, "Mask-guided portrait editing with conditional gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3436–3445. 2

[11] B. Ham, M. Cho, C. Schmid, and J. Ponce, "Proposal flow: Semantic correspondences from object proposals," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 7, pp. 1711–1725, 2017. 2

[12] K. Han, R. S. Rezende, B. Ham, K.-Y. K. Wong, M. Cho, C. Schmid, and J. Ponce, "Scnet: Learning semantic correspondence," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1831–1840. 2

[13] M. He, D. Chen, J. Liao, P. V. Sander, and L. Yuan, "Deep exemplar-based colorization," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, p. 47, 2018. 2

[14] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin, "Image analogies," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, 2001, pp. 327–340. 1

[15] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637. 5, 6

[16] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510. 2

[17] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 172–189. 1, 2, 6

[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134. 1, 2, 4

[19] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711. 4

[20] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017. 6

[21] S. Kim, D. Min, B. Ham, S. Jeon, S. Lin, and K. Sohn, "Fcss: Fully convolutional self-similarity for dense semantic correspondence," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6560–6569. 2

[22] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1857–1865. 2

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 5

[24] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 35–51. 2

[25] J. Lee, D. Kim, J. Ponce, and B. Ham, "Sfnet: Learning object-aware semantic correspondence," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2278–2287. 2, 3

[26] B. Li, F. Wu, K. Q. Weinberger, and S. Belongie, "Positional Normalization," *arXiv e-prints*, p. arXiv:1907.04312, Jul. 2019. 4

[27] J. Liao, Y. Yao, L. Yuan, G. Hua, and S. B. Kang, "Visual attribute transfer through deep image analogy," *arXiv preprint arXiv:1705.01088*, 2017. 1, 2, 3

[28] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125. 3

[29] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in neural information processing systems*, 2017, pp. 700–708. 2

[30] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, Dec. 2015. 5

[31] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016. 5

[32] J. L. Long, N. Zhang, and T. Darrell, "Do convnets learn correspondence?" in *Advances in Neural Information Processing Systems*, 2014, pp. 1601–1609. 2

[33] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004. 2

[34] L. Ma, X. Jia, S. Georgoulis, T. Tuytelaars, and L. Van Gool, "Exemplar guided unsupervised image-to-image translation with semantic consistency," *ICLR*, 2019. 1, 2, 6

[35] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 768–783. 4

[36] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014. 2

[37] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018. 5

[38] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2337–2346. 1, 2, 4, 6

[39] X. Qi, Q. Chen, J. Jia, and V. Koltun, "Semi-parametric image synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8808–8816. 2, 5, 6

[40] M. Riviere, O. Teytaud, J. Rapin, Y. LeCun, and C. Couprie, "Inspirational adversarial image generation," *arXiv preprint arXiv:1906.11661*, 2019. 2

[41] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241. 3

[42] A. Royer, K. Bousmalis, S. Gouws, F. Bertsch, I. Mosseri, F. Cole, and K. Murphy, "Xgan: Unsupervised image-to-image translation for many-to-many mappings," *arXiv preprint arXiv:1711.05139*, 2017. 2

[43] E. Tola, V. Lepetit, and P. Fua, "Daisy: An efficient dense descriptor applied to wide-baseline stereo," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 5, pp. 815–830, 2009. 2

[44] M. Wang, G.-Y. Yang, R. Li, R.-Z. Liang, S.-H. Zhang, P. Hall, S.-M. Hu *et al.*, "Example-guided style consistent image synthesis from semantic labeling," *arXiv preprint arXiv:1906.01314*, 2019. 1, 2

[45] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807. 1, 2, 6

[46] R. Yi, Y.-J. Liu, Y.-K. Lai, and P. L. Rosin, "Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 743–10 752. 2

[47] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2849–2857. 2

[48] B. Zhang, M. He, J. Liao, P. V. Sander, L. Yuan, A. Bermak, and D. Chen, "Deep exemplar-based video colorization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8052–8061. 2, 3

[49] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," *arXiv preprint arXiv:1805.08318*, 2018. 5

[50] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641. 5

[51] T. Zhou, P. Krahenbuhl, M. Aubry, Q. Huang, and A. A. Efros, "Learning dense correspondence via 3d-guided cycle consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 117–126. 2

[52] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232. 1, 2

[53] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *Advances in Neural Information Processing Systems*, 2017, pp. 465–476. 2