Unbalanced Feature Transport for Exemplar-based Image Translation

Fangneng Zhan [1,2], Yingchen Yu [1,2], Kaiwen Cui [1], Gongjie Zhang [1], Shijian Lu [1], Jianxiong Pan [2],
Changgong Zhang [2], Feiying Ma [2], Xuansong Xie [2], Chunyan Miao [1]
[1] Nanyang Technological University [2] DAMO Academy, Alibaba Group

Abstract

Despite the great success of GANs in images translation with different conditioned inputs such as semantic segmentation and edge maps, generating high-fidelity realistic images with reference styles remains a grand challenge in conditional image-to-image translation. This paper presents a general image translation framework that incorporates optimal transport for feature alignment between conditional inputs and style exemplars in image translation. The introduction of optimal transport mitigates the constraint of many-to-one feature matching significantly while building up accurate semantic correspondences between conditional inputs and exemplars. We design a novel unbalanced optimal transport to address the transport between features with deviational distributions which exists widely between conditional inputs and exemplars. In addition, we design a semantic-activation normalization scheme that injects style features of exemplars into the image translation process successfully. Extensive experiments over multiple image translation tasks show that our method achieves superior image translation qualitatively and quantitatively as compared with the state-of-the-art.

1. Introduction

Conditional image-to-image translation aims to generate images from certain given conditional inputs such as semantic segmentation [30, 40], layout [20], and key points [36]. With the advance of Generative Adversarial Networks (GANs), it has made rapid progress and achieved quite promising translation performance in recent years. However, most existing methods have very loose control over the translation process which often affects the translation quality greatly and so the wide application of image translation in various tasks. Optimal style control is still an open challenge in high-fidelity realistic image translation.

Several prior works attempted to tackle the style control challenge by using a latent code that is encoded by either Variational Auto-Encoder (VAE) [30] or style encoder [3]. However, latent codes often impair style control accuracy as they do not have sufficient capacity to capture detailed style information. A different approach is to inject specific style codes for different semantic regions [56], but it is specifically designed for conditional input of semantic segmentation and cannot well generalize to other conditional inputs. Recently, Zhang et al. [51] explore to establish dense semantic correspondence between conditioned input and a given style exemplar so as to offer dense style guidance in translation. However, it constructs the semantic correspondence based on cosine similarity that often leads to many-to-one matching (i.e. multiple conditional input features match to the same exemplar feature as illustrated in Fig. 1) and missing of details in image translation.

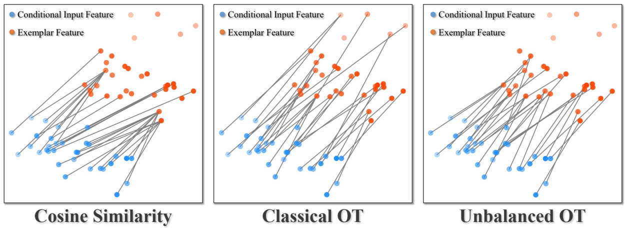

Figure 1. Different feature matching in image translation: *Cosine Similarity* tends to match each feature separately which often leads to many-to-one matching. Classical optimal transport (*Classical OT*) suppresses the many-to-one matching problem but it matches all feature points including undesired outliers (existing between deviational feature distributions). Our designed unbalanced optimal transport (*Unbalanced OT*) mitigates many-to-one matching and avoid outlier matching effectively.

We designed *UNITE*, UNbalanced optImal feature Transport for Exemplar-based image translation that achieves high-fidelity image generation with faithful style to given exemplars. UNITE consists of a *feature transport network* and a *translation network* that are inter-connected and can be jointly optimized in training. The *feature transport network* introduces optimal transport [31] which matches two sets of features as a whole and effectively overcomes many-to-one matching as in the widely adopted cosine similarity [51] that matches individual features separately. To tackle the distribution deviations between conditional inputs and exemplars, we design an unbalanced optimal transport technique that adaptively learns the mass (or weight) of each individual feature for effective transport between distributions of different masses. In the *translation network*, we design a semantic-activation normalization scheme that injects the aligned features into the translation process, where the exemplar features are transported in a multi-stage manner for preserving rich and complicated textural details. Extensive experiments show that UNITE translates images with superior realism and fidelity.

The contributions of this work can be summarized in three aspects. First, we propose a conditional image translation framework that introduces optimal transport for proper feature alignment and faithful style control in image translation. Second, we design an unbalanced optimal transport technique with adaptive mass learning scheme that is capable of aligning features with deviational distributions, and a multi-stage transport strategy that can preserve complex textures at different scales. Third, we design a novel semantic-activation normalization that is capable of injecting the aligned style features into the image translation process effectively.

2. Related Work

2.1. Image-to-Image Translation

GAN-based image-to-image translation has been investigated extensively due to its wide applications in different tasks such as domain adaptation [32, 44], data augmentation [49, 45], image editing [41, 15, 42], image composition [46, 43], etc. Existing works explored different conditional inputs such as semantic segmentation [10, 40, 30], scene layouts [35, 53, 20], key points [27, 29, 48], edge maps [10, 55, 18], etc. for photo-realistic image translation. On the other hand, optimal style control remains a critical yet challenging task that has attracted increasing attention in recent years. For example, [9] and [26] transfer style codes from exemplars to source images via adaptive instance normalization (AdaIN) [8]. [30] uses variational autoencoder (VAE) [13] to encode exemplars for image translation. [3] employs a style encoder for style consistency between exemplars and the translated images.

Different from the aforementioned methods that adopt latent vectors for style control, [51] learns dense semantic correspondences between conditional inputs and exemplars for image translation. Similar ideas have been explored in other translation tasks such as image colorization [6, 47] that also employs exemplars to build up semantic correspondences. On the other hand, most existing works use cosine similarity to build up semantic correspondences which often suffer from many-to-one matching and resultant feature missing. We introduce optimal transport for feature matching that treats the whole feature set as a whole and overcomes the many-to-one matching effectively.

2.2. Optimal Transport

Optimal transport (OT) [38] provides a principal way of comparing distributions and offers optimal plans for matching distributions. As a linear programming problem, classic OT is computationally intensive and [5] presents entropy regularized optimal transport that is differentiable and can be solved by the Sinkhorn-Knopp algorithm [34, 14] efficiently. On the other hand, the classical optimal transport has a typical constraint that they can only handle distributions with equal mass and thus become inapplicable while facing unbalanced distributions with different masses and deviations as widely existed in various tasks. Different approaches have been reported to address this new challenge. For example, [2] presents a unified treatment of unbalanced optimal transport that allows for both static and dynamic formulations. [21] introduce an entropic version of unbalanced optimal transport.

In recent years, optimal transport has been widely explored in various computer vision tasks such as domain adaptation [4], semantic matching [23], style transfer [16], etc. In this work, we adapt unbalanced feature transport for aligning deviational features between conditional inputs and exemplars for high-fidelity image translation.

3. Proposed Method

Our UNITE consists of a feature transport network (in blue and orange) and a translation network (in green) which are inter-connected as shown in Fig. 2. The feature transport network aligns the features of conditional inputs and exemplars and the translation network produces the final synthesis, more details to be described in the following subsections.

3.1. Feature Transport Network

The feature transport network aims to transport the feature of exemplars to be aligned with that of conditional inputs, thus providing accurate style guidance for the image translation. As shown in Fig. 2, both conditional input and exemplar are fed to two feature extractors F_X and F_Z to extract two sets of feature vectors $X = (x_1, \cdots, x_n) \in \mathbb{R}^d$ and $Z = (z_1, \cdots, z_n) \in \mathbb{R}^d$, where n denotes the number of feature vectors and d denotes the feature dimension.

To align feature sets X and Z, most existing methods [51, 6, 47] build a dense correspondence matrix between X and Z by measuring the Cosine similarity between any two feature vectors. As each feature vector x_i is matched to the feature vector z_j with the maximum Cosine similarity separately, multiple feature vectors in X may correspond to the same feature vector in Z (i.e. many-to-one matching), which leads to blurry translation as illustrated in Fig. 3. To avoid many-to-one matching between sets of feature vectors, we introduce optimal transport method to align the

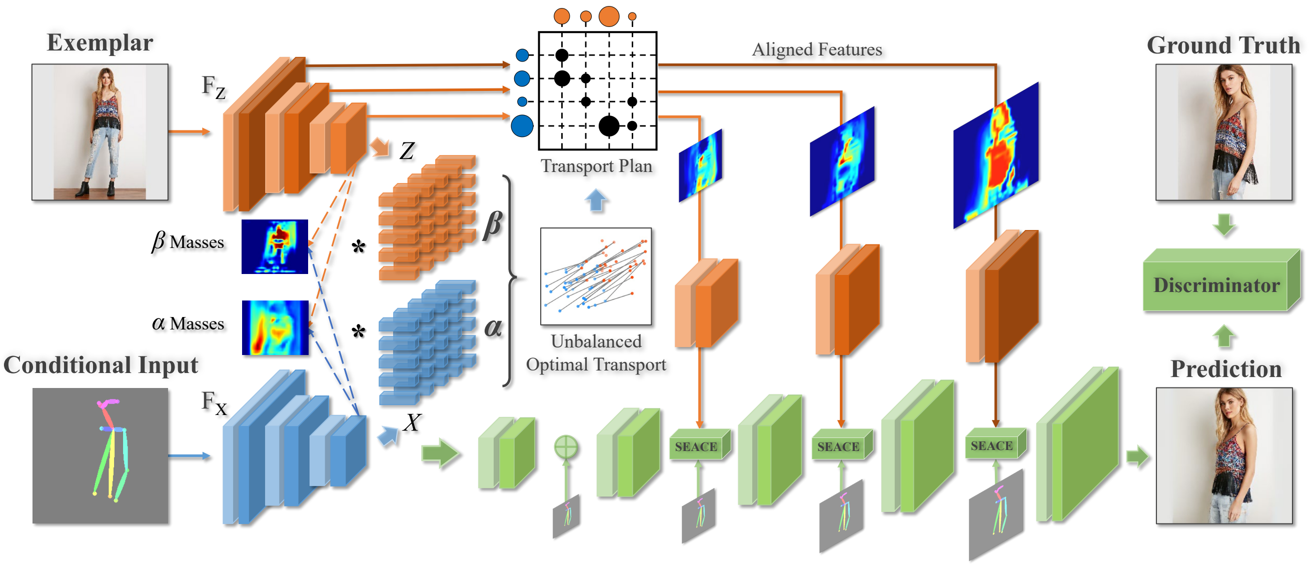

Figure 2. The framework of our proposed network: The *Conditional Input* and *Exemplar* are fed to feature extractors F_X and F_Z to extract feature vectors X and Z. The mass (or weight) of the feature vectors (α and β masses) are then determined collectively by X and Z. The weights and the feature vectors form two sets of Dirac masses $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which are further aligned through *Unbalanced Optimal Transport*. With an obtained *Transport Plan*, the feature of the *Exemplar* is transported in a multi-stage manner to be aligned with that of the *Conditional Input*. The aligned features will be injected into the translation network through a proposed SEmantic-ACtivation (dE)normalization (SEACE) to synthesize the final output image. (Blue and orange parts for feature transport network, green part for translation network)

features of conditional inputs and exemplars.

Classical Optimal Transport. The classical Optimal transport aims to determine the best transport plan (namely the minimum amount of total work required) to transform one measure into another with the same mass. Here the 'work' is evaluated by the product of the cost and the amount of mass to be transported. With constraints on the total masses in transport, optimal transport penalizes the many-to-one matching effectively.

To formulate the feature alignment as an optimal transport problem and derive the constraints of total masses, we encode the conditional input feature X and exemplar feature Z as Dirac masses: $\alpha = \sum_{i=1}^{n} \alpha_i \delta_{x_i}$ and $\beta = \sum_{i=1}^{n} \beta_i \delta_{z_j}$, where the masses $\alpha_i, \beta_i \geq 0$ and feature vectors x_i, z_i denote the locations of α_i, β_i. Then we define a distance matrix C, where each entry C_{ij} in C gives the cost of moving mass α_i to mass β_j which can be defined by: $C_{ij} = 1 - \frac{x_i^\top \cdot z_j}{||x_i|| \, ||z_j||}$ A transport plan T can be defined, where each entry T_{ij} is the amount of masses transported between α_i and β_j. Then the classical optimal transport problem can be formed as:

$$OT(\alpha, \beta) = \min_T (\sum_{i,j=1}^{n} C_{ij} T_{ij}) = \min_T \langle C, T \rangle \qquad (1)$$

$$\text{subject to} \quad (T\vec{1}) = \alpha, \quad (T^\top \vec{1}) = \beta$$

The constraints of the total masses $(T\vec{1}) = \alpha$ and $(T^\top \vec{1}) = \beta$ naturally penalize the many-to-one matching in optimal transport as illustrated in Fig. 3.

Unbalanced Optimal Transport. For classical optimal transport, the total masses of the two measures should be the same, namely $\sum_{i=1}^{n} \alpha_i = \sum_{j=1}^{n} \beta_j$. But for conditional inputs and exemplars, their features are usually not perfectly matched so have different total masses. For example, the conditional input (key-point map) in Fig. 2 does not contain feet which exist in the exemplar, so the feature of feet region in the exemplar is treated as outliers in optimal transport and should not be matched to any feature of the conditional input. However, classical optimal transport inevitably matches all features, leading to inaccurate or false matching as illustrated in Fig. 3. We handle it by introducing a relaxed version of classical optimal transport, namely unbalanced optimal transport (UOT or unbalanced OT) [2] that aims to determine an optimal transport plan between measures of different total masses. We formulate unbalanced OT by replacing the 'Hard' conservation of masses in (1) by a 'Soft' penalty with a divergence metric. An unbalanced OT problem can thus be formulated as follows:

$$\min_T \left[\langle C, T \rangle + \tau \mathrm{KL}(T\vec{1} || \alpha) + \tau \mathrm{KL}(T^\top \vec{1} || \beta) \right] \qquad (2)$$

where τ is regularization parameter, KL is the Kullback-Leibler divergence which is defined as $\mathrm{KL}(a||b) = \sum_{i=1}^{n} a_i \log(\frac{a_i}{b_i}) - a_i + b_i$.

We employ cross-inner product to generate the masses $\alpha_i, \beta_j (i,j \in [1,n])$ associated with each feature vector. The masses are highly correlated with specific conditional inputs and exemplars, thus it should be determined collectively by both of them. Intuitively, the feature vector that is more related with another feature set should have higher mass. We therefore determine the mass of a feature vector by computing its relevance with another feature set:

$$\alpha_i = x_i \cdot \frac{\sum_{i=1}^{n}(z_i)}{n}, \; \beta_j = z_j \cdot \frac{\sum_{j=1}^{n}(x_j)}{n} \qquad (3)$$

The mass parameters are adaptively updated in training. They capture the mass of each single feature vector accurately and mitigate the false matching problem effectively.

To implement UOT in a differentiable manner, an entropic regularization term $H(T) = -\sum_{i,j=1}^{n} T_{ij} \log T_{ij}$ is introduced. An entropic UOT problem can be defined by:

$$\min_{T} \left[\langle C, T \rangle + \tau \mathrm{KL}(T\vec{1} || \alpha) + \tau \mathrm{KL}(T^{\top}\vec{1} || \beta) - \eta H(T) \right]$$

where η is the regularization coefficients that denotes the smoothness of the transport plan T. In our network, η is fixed at 0.0001 empirically.

To obtain T, we consider the Fenchel-Legendre dual form of the entropic UOT that is defined by:

$$\max_{u,v} \left[-F^{*}(-u) - G^{*}(-v) - \eta \sum_{i,j} \exp\left(\frac{u_i + v_j - C_{ij}}{\eta}\right) \right] \qquad (4)$$

where F^{*} and G^{*} are the Legendre conjugate of KL divergence which can be computed by:

$$F^{*}(u) = \max_{z} z^{\top} u - \tau \mathrm{KL}(z || \alpha) = \tau \langle e^{u/\tau}, \alpha \rangle - \alpha^{\top}\vec{1}$$

$$G^{*}(v) = \max_{x} x^{\top} v - \tau \mathrm{KL}(x || \beta) = \tau \langle e^{v/\tau}, \beta \rangle - \beta^{\top}\vec{1}$$

Then the Sinkhorn algorithm [5] can be applied to (4) for approximating UOT solution, with a desired transport plan T encoded by optimal dual vectors u and v as below:

$$T_{ij} = \alpha_i \beta_j \, \exp \frac{1}{\eta} \left[u_i + v_j - C_{ij} \right] \qquad (5)$$

Multi-Stage Feature Transport. With the transport plan, the exemplar features can be transported to be aligned with conditional input features for translation. Different from CoCosNet [51] that warps exemplar images directly, we adopt a multi-stage manner to transport exemplar features as shown in Fig. 2. This multi-stage transport helps to preserve detailed exemplar features especially for textures with complicated patterns as illustrated in Fig. 5.

3.2. Translation Network

The translation network aims to synthesize images under the semantic guidance of conditional inputs and style guidance of aligned exemplar features. The overall architecture of the translation network is similar to SPADE [30] as illustrated in Fig. 2 (green part). More details of the network structure are available in the supplementary material.

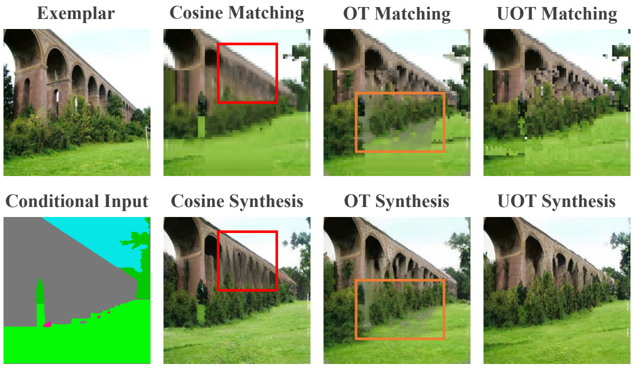

Figure 3. The comparison of different feature alignment methods: For visual comparison, we directly apply the feature alignment result to warp the exemplar. The *Cosine Matching* using cosine similarity often leads to many-to-one matching that introduces blurry feature alignment as highlighted by red box, which further leads to blurry synthesis result as shown in *Cosine Synthesis*. *OT Matching* using classical optimal transport suppresses the many-to-one matching but tends to introduce false matching as highlighted by orange box. Our proposed *UOT Matching* using unbalanced optimal transport mitigates both many-to-one matching and false matching effectively, which achieve the best feature alignment and synthesis fidelity as illustrated in *UOT Synthesis*.

In translation network, the aligned exemplar features are injected into the generation process at multiple stages to control the style of output image. Although style feature injection can be handled by several different approaches such as SPADE [30], all prevalent approaches fail to consider the semantic correlation between style features in feature injection. We designed an innovative semantic-aware injection method to be described in the following subsection.

Semantic-Activation Denormalization. Ideally, the style of a spatial position should be determined by all the style feature with the same semantic instead of only relying on the local feature in the exemplar. In addition, building long-range dependencies between style features is usually beneficial to image generation [50] as it allows to leverage the complementary style features of distant image regions. Based on these observations, we propose a novel SEmantic-ACtivation (dE)normalization (SEACE) to model the long-range dependencies across style features in style injection.

As shown in Fig. 4, two sets of modulation parameters γ_Z and μ_Z are generated from the *Aligned Feature*. To aggregate the style within each semantic region and build their long-range correlation, we introduce a semantic-activation matrix M, which can be obtained from the extracted feature of conditional input $X = (x_1, \cdots, x_n)$ by computing its self-attention $M_{ij} = x_i \cdot x_j$. As there is only semantic feature in conditional input feature, the semantic-

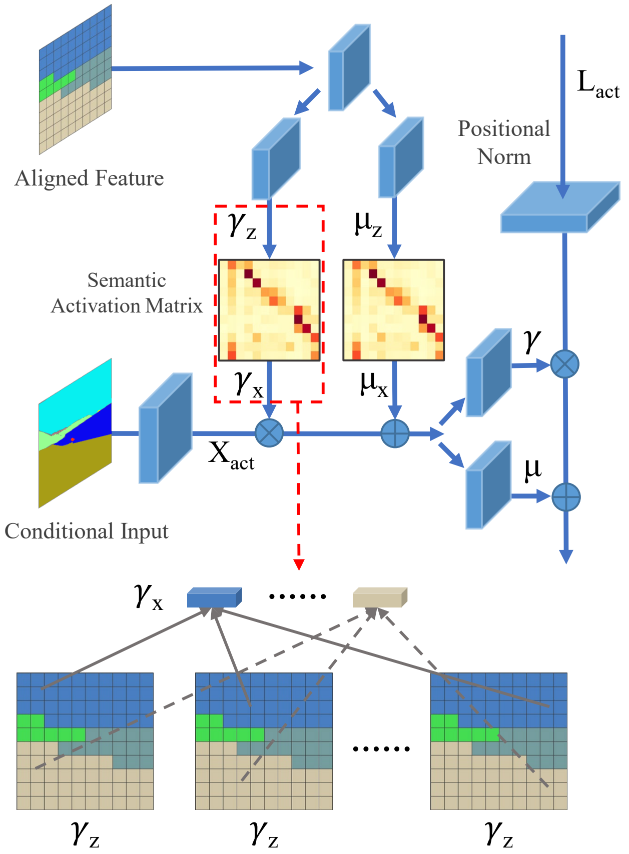

Figure 4. The structure of the proposed SEmantic-ACtivation (dE)normalization (SEACE): To build the long range dependency between style features, a semantic-activation matrix is obtained by computing the self-attention of the condition input features X that are extracted in the feature transport network. With the semantic-activation matrix, γ_X is determined collectively by the entire region in γ_Z with the same semantic as shown at the bottom.

activation matrix accurately measures the self-semantic correlation. Then the semantic-activation matrix is employed to aggregate the modulation parameters by $\gamma_X = M \cdot \gamma_Z$ and $\mu_X = M \cdot \mu_Z$. Thus the feature in each position of γ_X is determined collectively by a region with the same semantic in γ_Z as shown at the bottom of Fig. 4. Meanwhile, the long range correlation between modulation parameters with the same semantic is established.

Specially, instead of modulating the generation network directly with these modulation parameters, we first apply the modulation parameters γ_X and μ_X to modulate the activation X_{act} of the conditional input as follows:

$$X'_{act} = \gamma_X \cdot X_{act} + \mu_X \tag{6}$$

The intuition is that some features cannot be correctly matched if the conditional input contains some parts that do not exist in the exemplar. Thus before injecting the aligned style feature into the generation process, the unmatched feature of conditional input can be effectively corrected according to the accurate semantic information of the conditional input. Then two sets of modulation parameters γ and μ are further generated from the modulated conditional input X'_{act}.

A positional normalization [19] with variance γ_p and mean μ_p is applied to the activation of the translation network L_{act} to preserve the structure information synthesized in prior layers, followed by a denormalization with γ and μ as follows:

$$L'_{act} = \gamma \frac{L_{act} - \mu_p}{\gamma_p} + \mu \tag{7}$$

3.3. Loss Functions

The feature transport network and translation network are trained jointly, and will drive each other to achieve better translation. For clarity purpose, we denote the conditional input and exemplar as X and Z, the ground truth as X', the generated image as Y, the feature extractor network for conditional input and exemplar as F_X and F_Z, the translation network as G, the discriminator as D.

Feature Transport Network. First, the transported features should be cycle consistent, i.e. the original features should be able to be recovered from the transported features. We thus employ a cycle-consistency loss as follows:

$$\mathcal{L}_{cyc} = ||T^\top \cdot T \cdot Z - Z||_1 \tag{8}$$

where T is the transport plan. As the two feature extractor networks F_X and F_Z aim to extract semantic information, the extracted features from the conditional input X and the corresponding ground truth X' should be consistent. A feature consistency loss can thus be defined as follows:

$$\mathcal{L}_{cst} = ||F_X(X) - F_Z(X')||_1 \tag{9}$$

Translation Network. Several losses are employed in the translation network to drive the generation of high-fidelity images. As the semantic of the generated image should be consistent with the conditional input X or the ground truth X', we employ a perceptual loss $\mathcal{L}_{perc}$ [11] to penalize the semantic discrepancy as below:

$$\mathcal{L}_{perc} = ||\phi_l(Y) - \phi(X')||_1 \tag{10}$$

where ϕ_l represent the activation of layer l in pre-trained VGG-19 [33] model. To ensure the consistency of statistics between the generated image Y and the exemplar Z, a contextual loss in [28] is adopted as follows:

$$\mathcal{L}_{cxt} = -\log(\sum_i \max_j CX_{ij}(\phi_l^i(Z), \phi_l^j(Y))) \tag{11}$$

where i and j are the indexes of the feature map in layer ϕ_l. Besides, a pseudo pairs loss $\mathcal{L}_{pse}$ as described in [51] is included in training.

Table 1. Comparing UNITE with state-of-the-art image translation methods: The comparisons were performed over four public datasets with 3 widely used evaluation metrics FID, SWD and LPIPS.

Methods	**ADE20K**			**COCO-Stuff**			**DeepFashion**			**CelebA-HQ**		
	FID ↓	SWD ↓	LPIPS ↑	FID ↓	SWD ↓	LPIPS ↑	FID ↓	SWD ↓	LPIPS ↑	FID ↓	SWD ↓	LPIPS ↑
Pix2pixHD[40]	81.80	35.70	N/A	121.2	44.82	N/A	25.20	16.40	N/A	42.70	33.30	N/A
Pix2pixSC[39]	56.23	24.52	0.378	77.63	26.34	0.307	28.49	21.13	0.172	49.39	33.20	0.193
StarGAN v2[3]	98.72	65.47	0.451	153.2	61.87	0.394	43.29	30.87	**0.296**	48.63	41.96	0.214
SPADE[30]	33.90	19.70	0.344	49.27	19.78	0.254	36.20	27.80	0.231	31.50	26.90	0.187
SelectionGAN[37]	35.10	21.82	0.382	52.41	20.32	0.277	38.31	28.21	0.223	34.67	27.34	0.191
SMIS[57]	42.17	22.67	0.416	58.21	22.65	0.311	22.23	23.73	0.240	23.71	22.23	0.201
SEAN[56]	24.84	10.42	0.499	37.74	16.31	0.355	16.28	17.52	0.251	18.88	19.94	0.203
CoCosNet[51]	26.40	10.50	0.560	35.23	14.54	0.391	14.40	17.20	0.272	14.30	15.30	0.208
UNITE	**25.15**	**10.13**	**0.571**	**33.65**	**12.18**	**0.401**	**13.08**	**16.65**	0.278	**13.15**	**14.91**	**0.213**

Table 2. Comparing UNITE with state-of-the-art image translation methods over evaluation metrics semantic consistency and style consistency (on dataset ADE20k [54]).

Methods	**Semantic Consistency**		**Style Consistency**	
	VGG_{42} ↑	VGG_{52} ↑	VGG_{M} ↑	VGG_{V} ↑
Pix2PixSC [39]	0.840	0.751	0.941	0.932
SPADE [30]	0.861	0.772	0.934	0.884
StarGAN v2 [3]	0.741	0.718	0.919	0.907
SelectionGAN [37]	0.843	0.785	0.951	0.912
SMIS [57]	0.862	0.787	0.951	0.933
SEAN [56]	0.868	0.791	0.962	0.942
CoCosNet [51]	0.878	0.790	0.986	0.965
UNITE	**0.883**	**0.795**	**0.990**	**0.969**

The discriminator adopts the same architecture with Patch-GAN [10]. With the adversarial loss $\mathcal{L}_{adv}$, the model can be optimized with the following objective:

$$\begin{aligned}\mathcal{L} = \min_{F_X, F_Z, G} \max_{D} (&\lambda_1 \mathcal{L}_{cyc} + \lambda_2 \mathcal{L}_{cst} + \lambda_3 \mathcal{L}_{perc} \\ &+ \lambda_4 \mathcal{L}_{cxt} + \lambda_5 \mathcal{L}_{pse} + \lambda_6 \mathcal{L}_{adv})\end{aligned} \tag{12}$$

where the weights λ balance the losses in objective.

4. Experiments

4.1. Experimental Settings

Datasets: We experiment over multiple public datasets that handle different conditional image translation tasks.

- ADE20k [54] consists of 20k training images and each image is associated with a 150-class segmentation mask. This is a challenging dataset to most existing methods due to its rich data diversity. We conduct image generation by using its semantic segmentation as conditional inputs.
- COCO-Stuff [1] augments COCO [22] with pixel-level stuff annotations including 80 thing classes and 91 stuff classes. We use its layout as conditional inputs. Following [19], objects covering less than 2% of the image are ignored and images with 3 to 8 objects are used in experiments.
- CelebA-HQ [25] consists of 30,000 high quality face images. We use its edge maps as conditional inputs. The face landmarks are connected as face edges, and the edges in the background are detected by Canny edge detector.
- Deepfashion [24] contains 52,712 person images with various appearances and poses. 29,000 images are selected as training set and the rest as validation set. We use its key points as conditional inputs in experiments.

Evaluation Metrics: We adopt several evaluation metrics to assess image translation performance. *Fréchet Inception Score (FID)* [7] is adopted to measures the distance between the distribution of generated images and real images. We also adopt *Sliced Wasserstein distance (SWD)* [12] to measure statistical distance of low level patch distributions. Besides, *Learned Perceptual Image Patch Similarity (LPIPS)* [52] is adopted to evaluate the diversity of the translated images with different exemplars, which computes the perceptual distance between image features extracted by AlexNet [17].

We also adopt and extend the metrics in [51] to evaluate semantic consistency and style consistency. Specifically, a pre-trained VGG model [33] is used to extract high-level features ($relu4\_2$ and $relu5\_2$) of the ground truth and generated images that capture semantic features. The semantic consistency (VGG_{42} and VGG_{52}) is defined by the distance between the extracted high-level features as computed by cosine similarity. Similarly, the pre-trained VGG model is applied to extract the low-level feature ($relu1\_2$) of the generated images and exemplars that capture style features. The style consistency (VGG_{M} and VGG_{V}) is defined by the distance of channel-wise mean and standard deviation as computed by cosine similarity.

Besides, we conduct user study (US) to evaluate the images generated under different ablation settings. 100 pairs of generated images were shown to 20 users who select the image with the best visual quality.

Implementation Details: The learning rate for translation network and discriminator is $1e$-4 and 4e-4 (the feature transport network is optimized jointly with the translation network). We use Adam solver with $\beta_1 = 0$ and $\beta_2 = 0.999$. The experiments are conducted on 4 32GB Tesla V100 GPUs with synchronized BatchNorm applied.

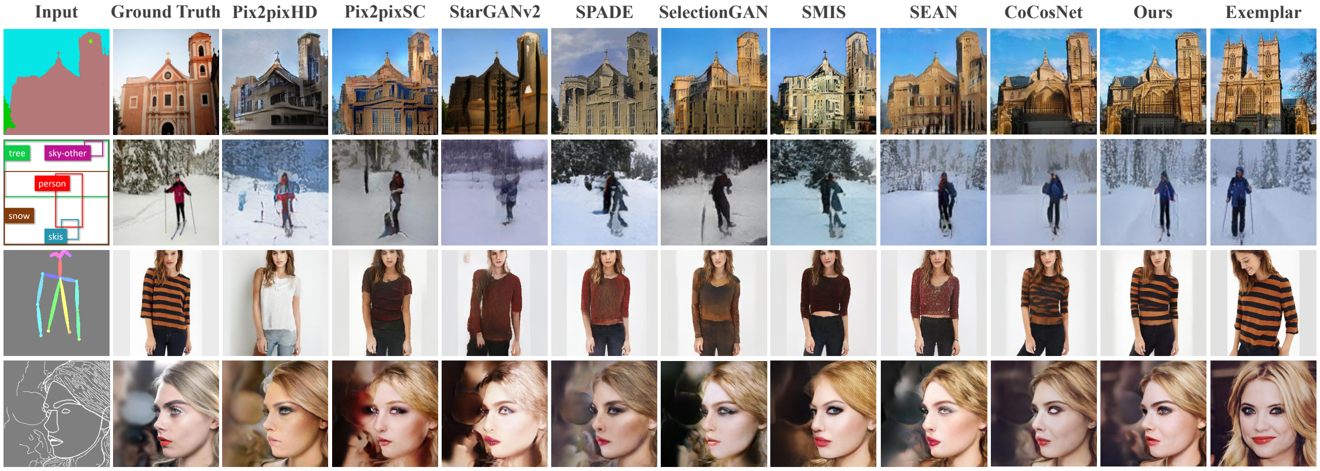

Figure 5. Qualitative illustration of UNITE and state-of-the-art image translation methods over four different types of conditional inputs.

The feature size for optimal transport is 64 × 64 with feature dimension of 128. The image size is set at 256×256 for generation tasks using semantic map, edge map, keypoints, and 128 × 128 for generation task using layout which is consistent with [35].

4.2. Experimental Results

We compare UNITE with several state-of-the-art translation methods including 1) Pix2pixHD [40], a supervised image translation method ; 2) Pix2PixSC [39], an example-guided image synthesis model based on Pix2PixHD [40]; 3) StarGAN v2[3], a model for multi-modal translation with support for style encoding from reference images; 4) SPADE [30], a supervised translation method that supports style injection from an exemplar image; 5) SelectionGAN [37], a guided translation framework with cascaded semantic guidance; 6) SMIS [57], a network for semantically multi-modal synthesis task with all group convolutions; 7) SEAN [56], a conditional translation network that can control the style of each individual semantic region; 8) CoCosNet [51], a leading exemplar-based translation framework that works by building cross-domain correspondences.

Quantitative Results: In quantitative experiments, all methods synthesize images with the same exemplars except Pix2PixHD [40] which synthesizes images directly without exemplar guidance (it doesn't support style injection from exemplars). As shown in Table 1, we compare UNITE with state-of-the-art methods in image quality as measured by FID and SWD and image diversity as measured by LPIPS. We can observe that UNITE outperforms all compared methods over all metrics and tasks consistently. Specifically, UNITE achieves the best FID and SWD which is largely attributed to our designed unbalance optimal transport in accurate feature alignments and semantic-activation normalization in effective style feature injection. Besides generation quality, UNITE achieves the best generation diversity in LPIPS, thanks to the multi-stage feature transport that aligns features in different scales to faithfully preserve rich textures in exemplars.

Except for high quality and rich diversity, the generated image should preserve consistent semantics with conditional inputs and present consistent styles with exemplars. Table 2 shows the semantic consistency and style consistency evaluated by the metrics described in *Evaluation Metrics*. With our UOT for accurate semantic feature matching and SEACE for effective style injection, UNITE achieves the best semantic consistency and style consistency.

Qualitative Evaluation: We compare images as generated by different translation methods as shown in Fig. 5. It can be seen that UNITE achieves faithful styles to the exemplars. SPADE [30], SMIS [57] and StarGAN v2 [3] adopt single latent code to encode image styles, which tend to capture global exemplar styles but miss local details. Although SEAN [56] employs multiple latent codes for feature injection, it still struggles to preserve faithful and detailed exemplar style. CoCosNet [51] can preserve certain details, but it adopts cosine similarity to align features which often lead to many-to-one matching and missing details as demonstrated by blurry textures in CoCosNet synthesized images. Our UNITE instead adopts UOT to achieve accurate feature alignment and a multi-stage transport to preserve the detailed texture. Besides, most existing methods tend to produce various artefacts as they do not build long-range dependency between style features. Our UNITE designs SEACE to explicitly build long-range dependency between style features which leads to superior synthesis fidelity as illustrated.

The proposed UNITE also demonstrates superior diversity in image translation as illustrated in Fig. 6. We can observe that UNITE is capable of synthesizing various real-

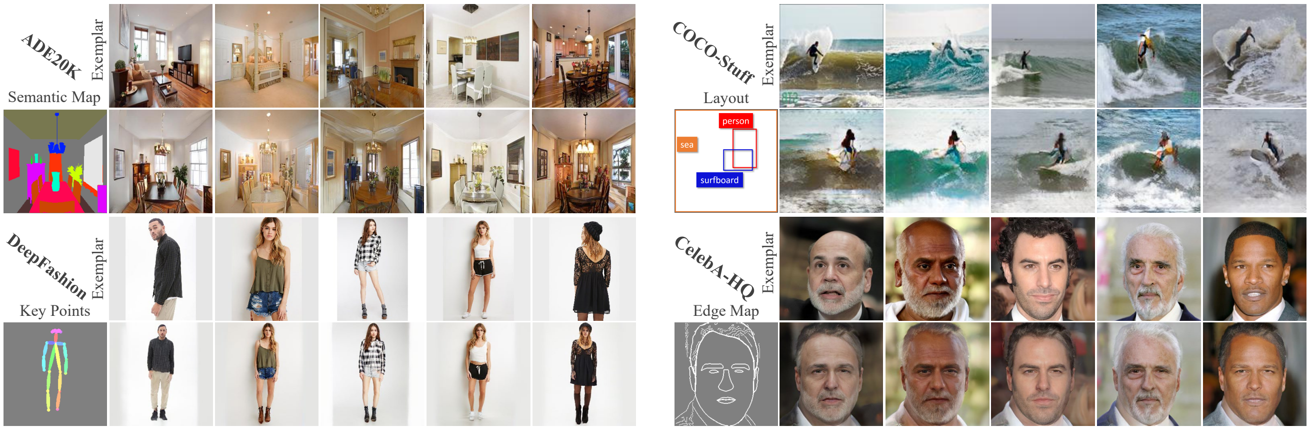

Figure 6. Qualitative illustration of our proposed UNITE with different types of conditional inputs and exemplars.

Table 3. Ablation studies of our UNITE designs over CelebA-HQ [25]: The baseline is SPADE that uses spatial denormalization [30]. COS, OT and UOT mean to include cosine similarity, classical optimal transport and unbalanced optimal transport in feature alignment. SEACE means to use the proposed semantic-activation denormalization to inject style features. MS denotes the multi-stage feature transportation. Model in the last row is the standard UNITE. US denotes the user study metric.

Models	FID ↓	SWD ↓	LPIPS ↑	US ↑
SPADE	31.50	26.90	0.187	0%
SPADE+COS	16.32	16.10	0.201	13%
SPADE+OT	17.87	17.24	0.202	10%
SPADE+UOT	14.02	15.41	0.206	22 %
SEACE+UOT	13.46	15.12	0.208	25 %
SEACE+UOT+MS	**13.15**	**14.91**	**0.213**	**30 %**

istic images with faithful style to the given exemplars.

4.3. Ablation Study

We conduct extensive ablation studies over CelebA-HQ [25] to validate the effectiveness of our designs. As Table 3 shows, SPADE [30] is the baseline which achieves image translation directly without feature alignment. When cosine similarity is included to align features, the translation is improved significantly. While replacing cosine similarity with classical optimal transport, the performance does is clearly aggravated as classical optimal transport introduce many false matchings. However, the translation performance improves clearly when our UOT is included, largely attributed to that UOT adaptively learns the feature masses and suppresses false and many-to-one matching effectively. When replacing SPADE with our proposed SEACE, the FID score is improved clearly by 0.73. Additionally, the SWD and LPIP scores are improved clearly when our proposed multi-stage feature transport is included. We also performed qualitative ablation studies on DeepFashion [24] by removing each of our designs from the complete UNITE model. As Fig. 7 shows, our designed UOT, MS and SEACE all contribute to the high-fidelity realistic image translation clearly.

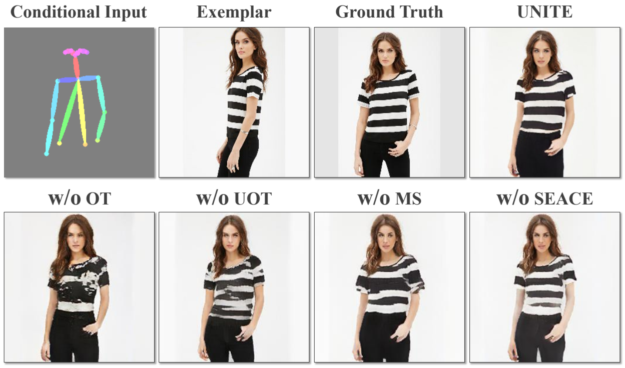

Figure 7. The ablation study of each different design in UNITE as evaluated over a sample from dataset DeepFashion [24]. Specially, 'w/o OT' denotes image translation without feature alignment, 'w/o UOT' denotes using classical OT (without learnt unbalanced weights) to align features.

5. Conclusions

This paper presents UNITE, an exemplar-based image translation framework that adopts unbalanced optimal transport to align the feature between conditional input and exemplar, which effectively transport the style of the exemplar to the conditional input. A multi-stage feature transport manner is applied to preserved more detailed deep features. To inject aligned the style feature into the generation process, we propose a novel semantic-activation normalization which builds the semantic coherence between style features with the same semantic in style injection. Quantitative and qualitative experiments show that UNITE is capable of generating high-fidelity images with consistent semantic with the conditional input and faithful style to the exemplar.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018. 6

[2] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced transport problems. In *arXiv:1607.05816*, 2016. 2, 3

[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. 1, 2, 6, 7

[4] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016. 2

[5] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013. 2, 4

[6] Mingming He, Dongdong Chen, Jing Liao, Pedro V Sander, and Lu Yuan. Deep exemplar-based colorization. *ACM Transactions on Graphics (TOG)*, 37(4):1–16, 2018. 2

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017. 6

[8] 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. 2

[9] 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. 2

[10] 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. 2, 6

[11] 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. 5

[12] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 6

[13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[14] Philip A Knight. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008. 2

[15] Ali Koksal and Shijian Lu. Rf-gan: A light and reconfigurable network for unpaired image-to-image translation. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

[16] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10051–10060, 2019. 2

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 6

[18] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 2

[19] Boyi Li, Felix Wu, Kilian Q Weinberger, and Serge Belongie. Positional normalization. In *Advances in Neural Information Processing Systems*, pages 1622–1634, 2019. 5, 6

[20] Yandong Li, Yu Cheng, Zhe Gan, Licheng Yu, Liqiang Wang, and Jingjing Liu. Bachgan: High-resolution image synthesis from salient object layout. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2

[21] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger-kantorovich distance between positive measures. *arXiv:1508.07941*, 2015. 2

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[23] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4463–4472, 2020. 2

[24] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 6, 8

[25] 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. 6, 8

[26] Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, and Luc Van Gool. Exemplar guided unsupervised image-to-image translation with semantic consistency. In *International Conference on Learning Representations*, 2018. 2

[27] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in neural information processing systems*, pages 406–416, 2017. 2

[28] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 768–783, 2018. 5

[29] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with

attribute-decomposed gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5084–5093, 2020. 2

[30] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 1, 2, 4, 6, 7, 8

[31] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 1

[32] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017. 2

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5, 6

[34] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967. 2

[35] Wei Sun and Tianfu Wu. Image synthesis from reconfigurable layout and style. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10531–10540, 2019. 2, 7

[36] Hao Tang, Dan Xu, Gaowen Liu, Wei Wang, Nicu Sebe, and Yan Yan. Cycle in cycle generative adversarial networks for keypoint-guided image generation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2052–2060, 2019. 1

[37] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2417–2426, 2019. 6, 7

[38] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008. 2

[39] Miao Wang, Guo-Ye Yang, Ruilong Li, Run-Ze Liang, Song-Hai Zhang, Peter M Hall, and Shi-Min Hu. Example-guided style-consistent image synthesis from semantic labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1495–1504, 2019. 6, 7

[40] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 1, 2, 6, 7

[41] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018. 2

[42] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jianxiong Pan, Kaiwen Cui, Shijian Lu, Feiying Ma, Xuansong Xie, and Chunyan Miao. Diverse image inpainting with bidirectional and autoregressive transformers. *arXiv preprint arXiv:2104.12335*, 2021. 2

[43] Fangneng Zhan, Shijian Lu, Changgong Zhang, Feiying Ma, and Xuansong Xie. Adversarial image composition with auxiliary illumination. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

[44] Fangneng Zhan, Chuhui Xue, and Shijian Lu. Ga-dan: Geometry-aware domain adaptation network for scene text detection and recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9105–9115, 2019. 2

[45] Fangneng Zhan, Hongyuan Zhu, and Shijian Lu. Scene text synthesis for efficient and effective deep network training. *arXiv preprint arXiv:1901.09193*, 2019. 2

[46] Fangneng Zhan, Hongyuan Zhu, and Shijian Lu. Spatial fusion gan for image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3653–3662, 2019. 2

[47] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8061, 2019. 2

[48] Changgong Zhang, Fangneng Zhan, and Yuan Chang. Deep monocular 3d human pose estimation via cascaded dimension-lifting. *arXiv preprint arXiv:2104.03520*, 2021. 2

[49] Gongjie Zhang, Kaiwen Cui, Tzu-Yi Hung, and Shijian Lu. Defect-gan: High-fidelity defect synthesis for automated defect inspection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2524–2534, 2021. 2

[50] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363. PMLR, 2019. 4

[51] 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. 1, 2, 4, 5, 6, 7

[52] 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. 6

[53] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8584–8593, 2019. 2

[54] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 6

[55] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances*

in neural information processing systems, pages 465–476, 2017. 2

[56] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020. 1, 6, 7

[57] Zhen Zhu, Zhiliang Xu, Ansheng You, and Xiang Bai. Semantically multi-modal image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5467–5476, 2020. 6, 7