Bi-level Feature Alignment for Versatile Image Translation and Manipulation

Fangneng Zhan[1,2⋆], Yingchen Yu[2⋆], Rongliang Wu[2], Jiahui Zhang[2], Kaiwen Cui[2], Aoran Xiao[2], Shijian Lu[2§], and Chunyan Miao[2]

[1] Nanyang Technological University, Singapore
[2] Max Planck Institute for Informatics, Germany
fzhan@mpi-inf.mpg.de, {shijian.lu,ascymiao}@ntu.edu.sg
{yingchen001,ronglian001,jiahui003,kaiwen001,aoran.xiao}@e.ntu.edu.sg

Abstract. Generative adversarial networks (GANs) have achieved great success in image translation and manipulation. However, high-fidelity image generation with faithful style control remains a grand challenge in computer vision. This paper presents a versatile image translation and manipulation framework that achieves accurate semantic and style guidance in image generation by explicitly building a correspondence. To handle the quadratic complexity incurred by building the dense correspondences, we introduce a bi-level feature alignment strategy that adopts a top-k operation to rank block-wise features followed by dense attention between block features which reduces memory cost substantially. As the top-k operation involves index swapping which precludes the gradient propagation, we approximate the non-differentiable top-k operation with a regularized earth mover's problem so that its gradient can be effectively back-propagated. In addition, we design a novel semantic position encoding mechanism that builds up coordinate for each individual semantic region to preserve texture structures while building correspondences. Further, we design a novel confidence feature injection module which mitigates mismatch problem by fusing features adaptively according to the reliability of built correspondences. Extensive experiments show that our method achieves superior performance qualitatively and quantitatively as compared with the state-of-the-art.

1 Introduction

Image translation and manipulation aim to generate and edit photo-realistic images conditioning on certain inputs such as semantic segmentation [32,43], key points [39,5] and layout [19]. It has been studied intensively in recent years thanks to its wide spectrum of applications in various tasks [35,30,42]. However, achieving high fidelity image translation and manipulation with faithful style control remains a grand challenge due to the high complexity of natural image styles. A typical approach to control image styles is to encode image features into a latent space with certain regularization (e.g., Gaussian distribution) on

⋆ denotes equal contribution, § denotes corresponding author.

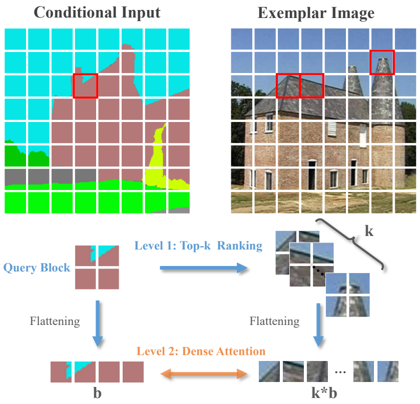


Fig. 1. Bi-level feature alignment via ranking and attention scheme: With a query block from the *Conditional Input*, we first retrieve the top-k most similar blocks from the *Exemplar Image* through a differentiable ranking operation, and then compute dense attention between features in query block and features in retrieved top-k blocks. Such bi-level alignment reduces the computational cost greatly, and allows to build high-resolution correspondences.

the latent feature distribution. For example, Park *et al.*[32] utilize VAE [4] to regularize the distribution of encoded features for faithful style control. However, VAE struggles to encode the complex distribution of natural image styles and often suffers from *posterior collapse* [25] which leads to degraded style control performance. Another strategy is to encode reference images into style codes [3,65] to provide style guidance in image generation, while style codes often capture the global or regional style without an explicit style guidance for generating texture details.

To achieve more accurate style guidance and preserve details from exemplar, Zhang *et al.* [59] explore to build cross-domain correspondences with Cosine similarity to achieve exemplar-based image translation. Zhou *et al.* [63] propose a GRU-assisted Patch-Match [1] method to build high-resolution correspondences efficiently. Since the textures within a semantic region share identical semantic information, the existing methods tend to build correspondences based on the semantic coherence without considering the structure coherence within each semantic region. Warping exemplars with such pure semantic correspondence may cause destroyed texture patterns in the warped exemplars, and consequently result in inaccurate guidance for image generation.

This paper presents **RABIT**, a **R**anking and **A**ttention scheme with **Bi**-level feature alignment for versatile **I**mage **T**ranslation and manipulation. To mitigate the quadratic computational complexity issue of building the dense correspondence between conditional inputs (semantic guidance) and exemplars (style guidance), we design a bi-level alignment strategy with a Ranking and Attention Scheme (RAS) which builds feature correspondences efficiently at two levels: 1) a top-k ranking operation for dynamically generating block-wise ranking matrices; 2) a dense attention module that achieves dense correspondences between features within blocks as illustrated in Fig. 1. RAS enables to build high-resolution correspondences and reduces the memory cost from $\mathcal{O}(L^2)$ to $\mathcal{O}(N^2 + b^2)$ (L is the number of features for alignment, b is block size, and $N = \frac{L}{b}$). However, the top-k operation involves index swapping whose gradient cannot be back-propagated in networks. To address this issue, we approximate

用patch
降低计算复杂度

① 先从大patch找offset
② 细patch 匹配

the top-k ranking operation with a regularized Earth Mover's problem [34] which enables gradient back-propagation effectively.

As in [59,63], building correspondences based on semantic information only often leads to the losing of texture structures and patterns in warped exemplars. Thus, spatial information should also be incorporated to preserve the texture structures and patterns and yield more accurate feature correspondences. A vanilla method to encode the position information is concatenating the semantic features with the corresponding feature coordinates via coordconv [22]. However, the vanilla position encoding builds a single coordinate system for the whole image which ignores the position information within each semantic region. Instead, we design a semantic position encoding (SPE) mechanism that builds a dedicated coordinate system for each semantic region which outperforms the vanilla position encoding significantly.

In addition, conditional inputs and exemplars are seldom perfectly matched, e.g., conditional inputs could contain several semantic classes that do not exist in exemplar images. Under such circumstances, the built correspondences often contain errors which lead to inaccurate exemplar warping and further deteriorated image generation. We tackle this problem by designing a CONfidence Feature Injection (CONFI) module that fuses the features of conditional inputs and warped exemplars according to the reliability of the built correspondences. Although the warped exemplar may not be reliable, the conditional input always provides accurate semantic guidance in image generation. The CONFI module thus assigns higher weights to the conditional input when the built correspondence (or warped exemplar) is unreliable. Experiments show that CONFI helps to generate faithful yet high-fidelity images consistently.

The contributions of this work can be summarized in three aspects. First, we propose a versatile image translation and manipulation framework which introduces a ranking and attention Scheme for bi-level feature alignment that greatly reduces the memory cost while building the correspondence between conditional inputs and exemplars. Second, we introduce a semantic position encoding mechanism that encodes region-level position information to preserve texture structures and patterns. Third, we design a confidence feature injection module that provides reliable feature guidance in image translation and manipulation.

2 Related Work

2.1 Image-to-Image Translation

Image translation has achieved remarkable progress in learning the mapping between images of different domains. It could be applied in different tasks such as style transfer [10,7,20], image super-resolution [16,21,15,58], domain adaptation [36,30,8,41,49], image composition [57,48,55,51,54] etc. To achieve high-fidelity and flexible translation, existing work uses different conditional inputs such as semantic segmentation [12,43,32,53,56], scene layouts [38,60,19,52], key points [27,29,5,50], edge maps [12,6], etc. However, effective style control remains a challenging task in image translation.

Style control has attracted increasing attention in image translation and generation. Earlier works such as [14] regularize the latent feature distribution to control the generation outcome. However, they struggle to capture the complex textures of natural images. Style encoding has been studied to address this issue. For example, [11] and [26] transfer style codes from exemplars to source images via adaptive instance normalization (AdaIN) [10]. [3] employs a style encoder for style consistency between exemplars and translated images. [65] designs semantic region-adaptive normalization (SEAN) to control the style of each semantic region individually. However, encoding style exemplars tends to capture the overall image style and ignores the texture details in local regions. To achieve accurate style guidance for each local region, Zhang *et al.* [59] build dense semantic correspondences between conditional inputs and exemplars with Cosine similarity to capture accurate exemplar details. To mitigate the quadratic complexity issue and enable high-resolution correspondence building, Zhou *et al.* [63] introduce the GRU-assisted Patch-Match to efficiently establish the high-resolution correspondence.

2.2 Semantic Image Editing

The arise of generative adversarial network (GANs) brings revolutionary advances to image editing [64,9,31,2,33,45,44,46]. As one of the most intuitive representation in image editing, semantic information has been extensively investigated in conditional image synthesis. For example, Park *et al.* [32] introduce spatially-adaptive normalization (SPADE) to inject guided features in image generation. MaskGAN [17] exploits a dual-editing consistency as auxiliary supervision for robust face image manipulation. Instead of directly learning a label-to-pixel mapping, Hong *et al.* [9] propose a semantic manipulation framework HIM that generates images guided by a predicted semantic layout. Upon this work, Ntavelis *et al.* [31] propose SESAME which requires only local semantic maps to achieve image manipulation. However, the aforementioned methods either only learn a global feature without local focus (e.g., MaskGAN [17]) or ignore the features in the editing regions of the original image (e.g., HIM [9], SESAME [31]). To better utilize the fine features in the original image, Zheng *et al.* [61] adapt exemplar-based image synthesis framework CoCosNet [59] for semantic image manipulation by building a high-resolution correspondence between the original image and the edited semantic map.

3 Proposed Method

The proposed RABIT consists of an alignment network and a generation network that are inter-connected as shown in Fig. 2. The alignment network learns the correspondence between a conditional input and an exemplar for warping the exemplar to be aligned with the conditional input. The generation network produces the final generation under the guidance of the warped exemplar and the conditional input. RABIT is typically applicable in the task of conditional

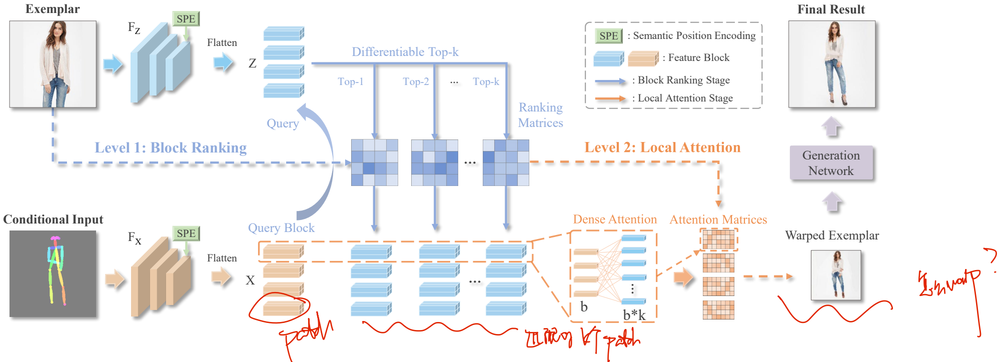


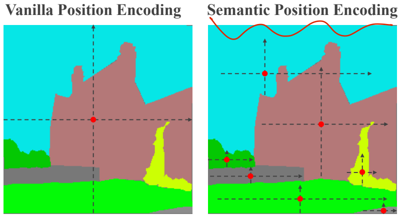
Fig. 2. The framework of the proposed RABIT: *Conditional Input* and *Exemplar* are fed to feature extractors F_X and F_Z to extract feature vectors X and Z where b local features form a feature block. In the first level, each block from the conditional input serves as the query to retrieve top-k similar blocks from the exemplar through a differentiable ranking operation. In the second level, *Dense Attention* is then built between the b features in query block and $b * k$ features in the retrieved blocks. The built *Ranking Matrices* and *Attention Matrices* are combined to warp the exemplar to be aligned with the conditional input as in *Warped Exemplar*, which serves as a style guidance to generate the final result through a generation network.

image translation with extra exemplar as style guidance. It is also applicable to the task of image manipulation by treating the exemplars as the original images for editing and the conditional inputs as the edited semantic. The detailed loss functions can be found in the supplementary materials.

3.1 Alignment Network

The alignment network aims to build the correspondence between conditional inputs and exemplars, and accordingly provide accurate style guidance by warping the exemplars to be aligned with the conditional inputs. As shown in Fig. 2, conditional input and exemplar are fed to feature extractors F_X and F_Z to extract two sets of feature vectors $X = [x_1, \cdots, x_L] \in \mathbb{R}^d$ and $Z = [z_1, \cdots, z_L] \in \mathbb{R}^d$, where L and d denote the number and dimension of feature vectors, respectively. Then X and Z can be aligned by building a $L \times L$ dense correspondence matrix where each entry denotes the Cosine similarity between the corresponding feature vectors in X and Z.

Semantic Position Encoding. Existing works [59,63] mainly rely on semantic features to establish the correspondences. However, as textures within a semantic region share the same semantic feature, the pure semantic correspondence fails to preserve the texture structures or patterns within each semantic region. Thus, the position information of features can be facilitated to preserve the texture structures and patterns. A vanilla method to encode the position information is employing a simple coordconv [22] to build a global coordinate for the full image. However, this vanilla position encoding mechanism builds a single coordinate system for the whole image, ignoring region-wise semantic differences. To preserve the fine texture pattern within each semantic region, we design a semantic position encoding (SPE) mechanism that builds a dedicated coordinate for each semantic region as shown in Fig. 3. Specifically, SPE treats the center of each semantic region as the origin of coordinate, and the coordinates within each semantic region are normalized to [-1, 1]. The proposed SPE outperforms the vanilla position encoding significantly as shown in Fig. 6 and to be evaluated in experiments.



基于patch
去卷积

Fig. 3. The comparison of vanilla position encoding and the proposed semantic position encoding (SPE). Red dots denote the coordinate origin.

Bi-level Feature Alignment. On the other hand, building correspondence has quadratic complexity which incurs large memory and computation costs. Most existing studies thus work with low-resolution exemplar images (e.g. 64 × 64 in CoCosNet [59]) which often struggle in generating realistic images with fine texture details. In this work, we propose a bi-level alignment strategy via a novel ranking and attention scheme (RAS) that greatly reduces computational costs and allows to build correspondences with high-resolution images as shown in Fig. 6. Instead of building correspondences between features directly, the bi-level alignment strategy builds correspondences at two levels, including the first level that introduces top-k ranking to generate block-wise ranking matrices dynamically and the second level that achieves dense attention between the features within blocks. As Fig. 2 shows, b local features are grouped into a block, thus the features of conditional input and exemplar are partitioned into N blocks ($N = L/b$) as denoted by $X = [X_1, \cdots, X_N] \in \mathbb{R}^{bd}$ and $Z = [Z_1, \cdots, Z_N] \in \mathbb{R}^{bd}$. In the first level of top-k ranking, each block feature of the conditional input serves as a query to retrieve top-k block features from the exemplar according to the Cosine similarity between blocks. In the second level of local attention, the features in each query block further attends to the features in the top-k retrieved blocks to build up local attention matrices within the block features. The correspondence between the exemplar and conditional input can thus be built much more efficiently by combining such inter-block ranking and inner-block attention.

The ranking and attention scheme employs a top-k operation that ranks the correlative blocks. However, the original top-k operation involves index swapping whose gradient cannot be computed and so cannot be integrated into end-to-end network training. Inspired by Xie *et al.* [47], we tackle this issue by formulating the top-k ranking as a regularized earth mover's problem which allows gradient computation via implicit differentiation. Earth mover's problem aims to find a transport plan that minimizes the total cost to transform one distribution to another. Consider two discrete distributions $U = [\mu_1, \ldots, \mu_N]^\top$ and

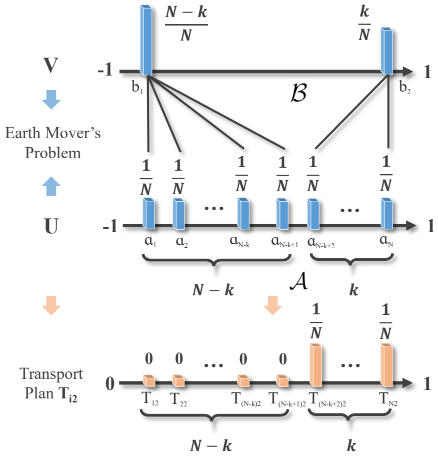


Fig. 4. Illustration of the earth mover's problem in top-k retrieval. Earth mover's problem is conducted between distributions U and V which is defined on supports $\mathcal{A} = [a_1, \cdots, a_N]$ and $\mathcal{B} = [b_1, b_2]$. *Transport Plan* T_{i2} indicates the retrieved top-k elements.

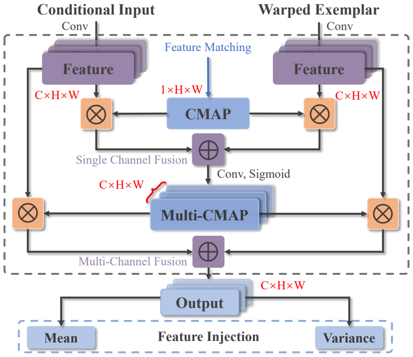


Fig. 5. Illustration of confidence feature injection: Conditional input and warped exemplar are initially fused with a confidence map (CMAP) of size $1 \times H \times W$. A multi-channel confidence map (Multi-CMAP) of size $C \times H \times W$ is then obtained from the initial fusion which further fuses the conditional input and warped exemplar in multiple channels.

$V = [\nu_1, \ldots, \nu_M]^\top$ defined on supports $\mathcal{A} = [a_1, \cdots, a_N]$ and $\mathcal{B} = [b_1, \cdots, b_M]$, with probability (or amount of earth) $\mathbb{P}(a_i) = \mu_i$ and $\mathbb{P}(b_j) = \nu_j$. We define $C \in \mathbb{R}^{N \times M}$ as the cost matrix where C_{ij} denotes the cost of transportation between a_i and b_j, and T as a transport plan where T_{ij} denotes the amount of earth transported between μ_i and ν_j. An earth mover's (EM) problem can be formulated by: $\text{EM} = \min_T \langle C, T \rangle, \quad \text{s.t. } T\vec{1}_M = U,\ T^\top \vec{1}_N = V$ where $\vec{1}$ denotes a vector of ones, $\langle\rangle$ denotes inner product. By treating a correlation scores between a query block and N key blocks as $\mathcal{A} = [a_1, \cdots, a_N], a_i \in [-1, 1]$ and defining $\mathcal{B} = \{-1, 1\}$, $U = [\mu_1, \cdots, \mu_N]$ and $V = [\nu_1, \nu_2]$, it can be proved that solving the Earth Mover's problem is equivalent to select the largest K elements from $\mathcal{A} = [a_1, \cdots, a_N]$. The detailed proof and optimization of the earth mover's problem is provided in supplementary material. Fig. 4 illustrates the earth mover's problem and transport plan T which indicates the top-k elements.

Complexity Analysis. The vanilla dense correspondence has a self-attention memory complexity $\mathcal{O}(L^2)$ where L is the input sequence length. For our bi-level alignment strategy, the memory complexity of building block ranking matrices and local attention matrices are $\mathcal{O}(N^2)$ and $\mathcal{O}(b * (kb))$, where b, N $(N = L/b)$ and k are block size, block number and the number of top-k selection. Thus, the overall memory complexity is $\mathcal{O}(N^2 + b * (kb))$.

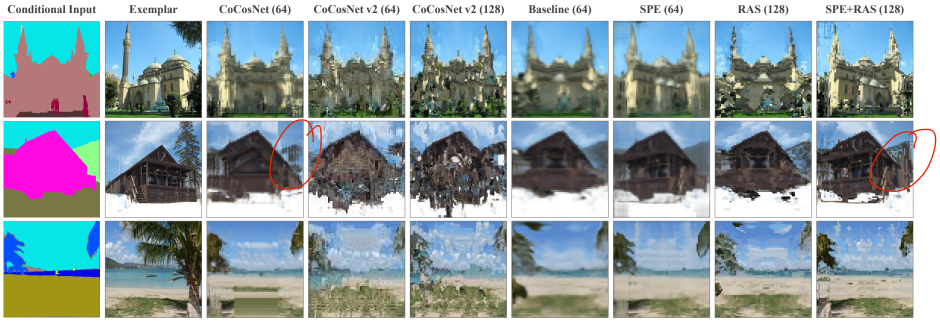


Fig. 6. Warped exemplars with different methods: '64' and '128' mean to build correspondences at resolutions 64 × 64 and 128 × 128. CoCosNet [59] tends to lose texture details and structures, while CoCosNet v2 [63] tends to generate messy warping. The *Baseline* denotes building correspondences with Cosine similarity, which tends to lose textures details and structures. The proposed ranking and attention scheme (RAS) allows efficient image warping at high resolutions, the proposed semantic position encoding (SPE) can better preserve texture structures. The combination of the two as denoted by SPE+RAS achieves the best warping performance with high resolution and preserved texture structures.

3.2 Generation Network

The generation network aims to synthesize images under the semantic guidance of conditional inputs and style guidance of exemplars. The overall architecture of the generation network is similar to SPADE [32]. Please refer to supplementary material for details of the network structure.

State-of-the-art approach [59] simply concatenates the warped exemplar and conditional input to guide the image generation process. However, the warped input image and edited semantic map could be structurally aligned but semantically different especially when they have severe semantic discrepancy. Such unreliably warped exemplars could serve as false guidance and heavily deteriorate the generation performance. Therefore, a mechanism is required to identify the semantic reliability of warped exemplar to provide reliable guidance for the generation network. To this end, we propose a CONfidence Feature Injection (CONFI) module that adaptively weights the features of conditional input and warped exemplar according to the reliability of feature matching.

Confidence Feature Injection. Intuitively, in the case of lower reliability of the feature correspondence, we should assign a relatively lower weight to the warped exemplar which provides unreliable style guidance and a higher weight to the conditional input which consistently provides accurate semantic guidance.

As illustrated in Fig. 5, the proposed CONFI fuses the features of the conditional input and warped exemplar based on a confidence map (CMAP) that captures the reliability of the feature correspondence. To derive the confidence map, we first obtain a block-wise correlation map of size $N \times N$ by computing element-wise Cosine distance between $X = [X_i, \cdots, X_N]$ and $Z = [Z_i, \cdots, Z_N]$.

For a block X_i, the correlation score with Z is denoted by $\mathcal{A} = [a_1, \cdots, a_N]$. As higher correlation scores indicate more reliable feature matching, we treat the peak value of $\mathcal{A}$ as the confidence score of X_i. Similar for other blocks, we can obtain the confidence map (CMAP) of size $1 \times H \times W$ $(N = H * W)$ which captures the semantic reliability of all blocks. The features of the conditional input and exemplar (both of size $C \times H \times W$ after passing through convolution layers) can thus be fused via weighted sum based on the confidence map CMAP: $F = X * (1 - \mathrm{CMAP}) + (T \cdot Z) * \mathrm{CMAP}$ where T is the built correspondence matrix. As the confidence map contains only one channel $(1 \times H \times W)$, the above feature fusion is conducted in $H \times W$ but ignores that in C channel. To achieve thorough feature fusion in all channels, we feed the initial fusion F to convolution layers to generate a multi-channel confidence map (Multi-CMAP) of size $C \times H \times W$. The conditional input and warped exemplar are then thoroughly fused via a full channel-weighted summation according to the Multi-CMAP. The final fused feature is further injected to the generation process via spatial de-normalization [32] to provide accurate semantic and style guidance.

4 Loss Functions

The alignment network and generation network are jointly optimized. For clarity, we still denote the conditional input and exemplar as X and Z, the ground truth as X', the generated image as Y, the feature extractors for conditional input and exemplar as E_X and E_Z, the generator and discriminator in the generation network as G and D.

Alignment Network. First, the warping should be cycle consistent, i.e. the exemplar should be recoverable from the warped warped. We thus employ a cycle-consistency loss as follows:

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

where T is the correspondence matrix. The feature extractors F_X and F_Z aim to extract invariant semantic information across domains, i.e. the extracted features from X and X' should be consistent. A feature consistency loss can thus be formulated as follows:

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

Generation Network. The generation network employs several losses for high-fidelity synthesis with consistent style with the exemplar and consistent semantic with the conditional input. As the generated image Y should be semantically consistent with the ground truth X', we employ a perceptual loss $\mathcal{L}_{perc}$ [13] to penalize their semantic discrepancy as below:

$$\mathcal{L}_{perc} = ||\phi_l(Y) - \phi(X')||_1 \tag{1}$$

where ϕ_l is the activation of layer l in pre-trained VGG-19 [37] model. To ensure the statistical consistency between the generated image Y and the exemplar Z, a contextual loss [28] is adopted:

$$\mathcal{L}_{cxt} = -\log(\sum_{i} \max_{j} CX_{ij}(\phi_l^i(Z), \phi_l^j(Y))) \tag{2}$$

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 [59] is included in training.

The discriminator D is employed to drive adversarial generation with an adversarial loss $\mathcal{L}_{adv}$ [12]. The full network is thus optimized with the following objective:

$$\begin{aligned}\mathcal{L} = \min_{F_X, F_Z, G} \max_{D} (&\lambda_1 \mathcal{L}_{cyc} + \lambda_2 \mathcal{L}_{cst} + \lambda_3 \mathcal{L}_{perc} \\ &+ \lambda_4 \mathcal{L}_{cxt} + \lambda_5 \mathcal{L}_{pse} + \lambda_6 \mathcal{L}_{adv})\end{aligned} \tag{3}$$

where the weights λ balance the losses in the objective.

5 Experiments

5.1 Experimental Settings

Datasets. We evaluate and benchmark our method over multiple datasets for image translation & manipulation tasks.

• ADE20K [62] is adopted for image translation conditioned on semantic segmentation. For image manipulation, we apply object-level affine transformations on the test set to acquire paired data (150 images) for evaluations as in [61].

• CelebA-HQ [24] is used for two translation tasks by using face semantics and face edges as conditional inputs. We use 2993 face images for translation evaluations as in [59], and manually edit 100 semantic maps which is randomly selected for image manipulation evaluations.

• DeepFashion [23] is used for image translation conditioned key points.

Implementation Details: The default size for our correspondence computation is 128×128 with a block size of 2×2. The number k in top-k ranking is set at 3 by default in our experiments. The default size of generated images is 256×256.

5.2 Image Translation Experiments

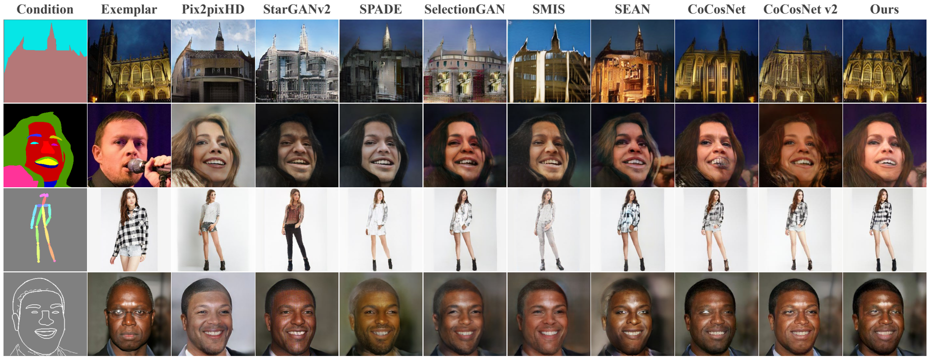
We compare RABIT with several state-of-the-art image translation methods.

Quantitative Results. In quantitative experiments, all methods translate images with the same exemplars except Pix2PixHD [43] which doesn't support style injection from exemplars. LPIPS is calculated by comparing the generated images with randomly selected exemplars. All compared methods adopt three exemplars for each conditional input and the final LPIPS is obtained by averaging the LPIPS between any two generated images.

Table 1 shows experimental results. It can be seen that RABIT outperforms all compared methods over most metrics and tasks consistently. By building explicit yet accurate correspondences between conditional inputs and exemplars,

Table 1. Comparing RABIT with state-of-the-art image translation methods over four translation tasks with FID, SWD and LPIPS as the evaluation metrics.

Methods	ADE20K			CelebA-HQ (Semantic)			DeepFashion			CelebA-HQ (Edge)		
	FID ↓	SWD ↓	LPIPS ↑	FID ↓	SWD ↓	LPIPS ↑	FID ↓	SWD ↓	LPIPS ↑	FID ↓	SWD ↓	LPIPS ↑
Pix2pixHD[43]	81.80	35.70	N/A	43.69	34.82	N/A	25.20	16.40	N/A	42.70	33.30	N/A
StarGAN v2[3]	98.72	65.47	0.551	53.20	41.87	**0.324**	43.29	30.87	**0.296**	48.63	41.96	**0.214**
SPADE[32]	33.90	19.70	0.344	39.17	29.78	0.254	36.20	27.80	0.231	31.50	26.90	0.207
SelectionGAN[40]	35.10	21.82	0.382	42.41	30.32	0.277	38.31	28.21	0.223	34.67	27.34	0.191
SMIS[66]	42.17	22.67	0.476	28.21	24.65	0.301	22.23	23.73	0.240	23.71	22.23	0.201
SEAN[65]	24.84	10.42	0.499	**17.66**	14.13	0.285	16.28	17.52	0.251	16.84	14.94	0.203
CoCosNet[59]	26.40	10.50	**0.580**	21.83	12.13	0.292	14.40	17.20	0.272	14.30	15.30	0.208
RABIT	**24.35**	**9.893**	**0.571**	20.44	**11.18**	0.307	**12.58**	**16.03**	0.284	**11.67**	14.22	0.209



略好一些

Fig. 7. Qualitative comparison of the proposed RABIT and state-of-the-art methods over four types of conditional image translation tasks.

RABIT enables direct and accurate guidance from the exemplar and achieves better translation quality (in FID and SWD) and diversity (in LPIPS) as compared with the regularization-based methods such as SPADE [32] and SMIS [66], and style-encoding methods such as StarGAN v2 [3] and SEAN [65]. Compared with correspondence-based method CoCosNet [59], the proposed bi-level alignment allows RABIT to build correspondences and warp exemplars at higher resolutions (e.g. 128 × 128) which offers more detailed guidance in the generation process and helps to achieve better FID and SWD. While compared with CoCosNet v2 [63], the proposed semantic position encoding enables to preserve the texture structures and patterns, thus yielding more accurate warped exemplars as guidance. Besides generation quality, RABIT achieves the best generation diversity in LPIPS except StarGAN v2 [3] which sacrifices the generation quality with much lower FID and SWD.

Qualitative Evaluations. Fig. 7 shows qualitative comparisons on various conditional image translation tasks. It can be seen that RABIT achieves the best visual quality with faithful styles as exemplars. RABIT also demonstrates superior diversity in image translation as illustrated in Fig. 8.

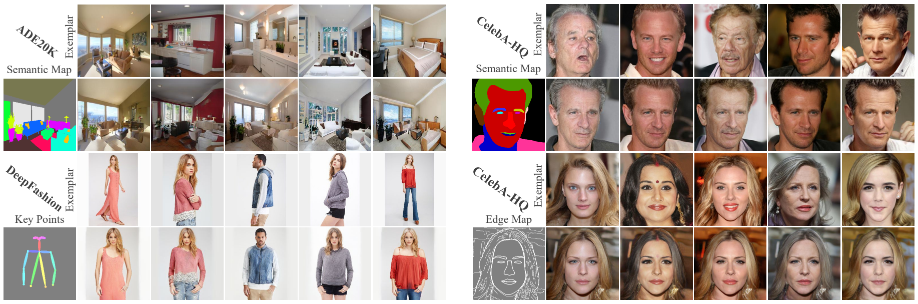


Fig. 8. Illustration of generation diversity of RABIT: With the same conditional input, RABIT can generate a variety of images that have consistent styles with the provided exemplars. It works for different types of conditional inputs consistently.

Table 2. Comparing RABIT with state-of-the art image manipulation methods on ADE20K [62] and CelebA-HQ [24].

ADE20K [62]				CelebA-HQ [24]			
Models	FID ↓	PSNR ↑	SSIM ↑	Models	FID ↓	SWD ↓	LPIPS ↓
SPADE [32]	120.2	13.11	0.334	**SPADE** [32]	105.1	41.90	0.376
HIM [9]	59.89	18.23	0.667	**SEAN** [65]	96.31	35.90	0.351
SESAME [31]	52.51	18.67	0.691	**MaskGAN** [17]	80.89	23.86	0.271
CoCosNet [59]	41.03	20.30	0.744	**CoCosNet** [59]	68.70	22.90	0.224
RABIT	**26.61**	**23.08**	**0.823**	**RABIT**	**60.87**	**21.07**	**0.176**

5.3 Image Manipulation Experiment

RABIT manipulates images by treating input images as exemplars and edited semantic guidance as conditional inputs. We compare RABIT with several state-of-the-art image manipulation methods including 1) SPADE [32], 2) SEAN [65], 3) MaskGAN [18], 4) Hierarchical Image Manipulation (HIM) [9], 5) SESAME [31], 6) CoCosNet [59].

Quantitative Results: In quantitative experiments, all compared methods manipulate images with the same input image and edited semantic label map. Left side of Table 2 shows experimental results over the synthesized test set of ADE20K [62]. It can be observed that RABIT outperforms state-of-the-art methods over all evaluation metrics consistently. Right side of Table 2 shows experimental results over the CelebA-HQ dataset with manually edited semantic maps. It can be observed that RABIT outperforms the state-of-the-art methods by large margins in all perceptual quality metrics.

Qualitative Evaluation: Fig. 9 shows visual comparisons with state-of-art manipulation methods on ADE20K. Fig. 10 shows the editing capacity of RABIT with various types of manipulation on semantic labels. We also compare RABIT with MaskGAN [17] on CelebA-HQ [18] in Fig. 12.

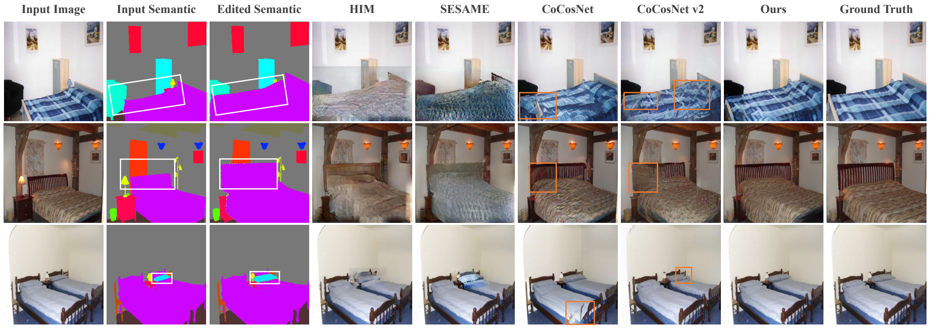


Fig. 9. Qualitative illustration of RABIT and SOTA image manipulation methods on the augmented test set of ADE20K with ground truth as described in [61].

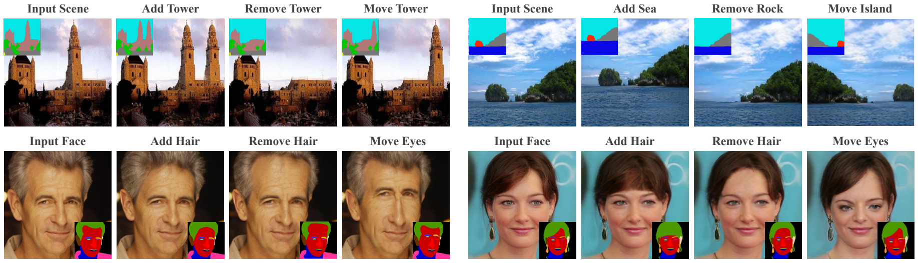


Fig. 10. Various image editing by the proposed RABIT: With input images as the exemplars and edited semantic maps as the conditional input, RABIT generates new images with faithful semantics and high-fidelity textures with little artifacts.

6 User Study

We conduct crowdsourcing user studies through Amazon Mechanical Turk (AMT) to evaluate the image translation & manipulation in terms of generation quality and style consistency. Specifically, each compared method generates 100 images with the same conditional inputs and exemplars. Then the generated images together with the conditional inputs and exemplars were presented to 10 users for assessment. For the evaluation of image quality, the users were instructed to pick the best-quality images. For the evaluation of style consistency, the users were instructed to select the images with best style relevance to the exemplar. The final AMT score is the averaged number of the methods to be selected as the best quality and the best style relevance.

Fig. 11 shows AMT results on multiple datasets. It can be observed that RABIT outperforms state-of-the-art methods consistently in image quality and style consistency on both image translation & image manipulation tasks.

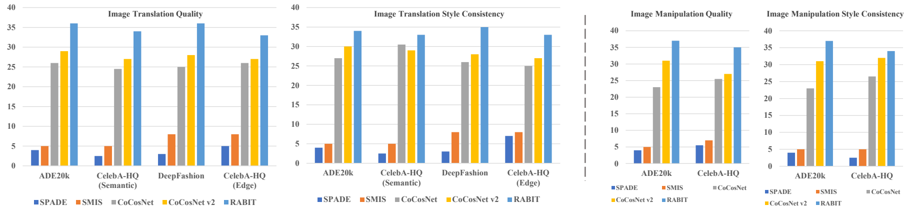


Fig. 11. AMT (Amazon Mechanical Turk) user studies of different image translation and image manipulation methods in terms of the visual quality and style consistency of the generated images.

7 Conclusions

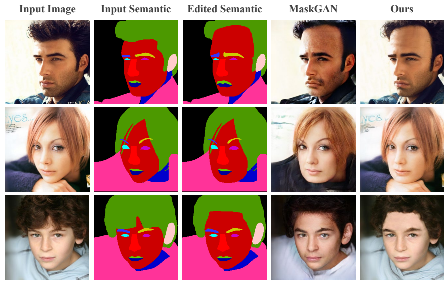


Fig. 12. The comparison of image manipulation by MaskGAN [17] and the proposed RABIT over dataset CelebA-HQ [24].

This paper presents RABIT, a versatile conditional image translation & manipulation framework that adopts a novel bi-level alignment strategy with a ranking and attention scheme (RAS) to align the features between conditional inputs and exemplars efficiently. A semantic position encoding mechanism is designed to facilitate semantic-level position information and preserve the texture patterns in the exemplars. To handle the semantic mismatching between the conditional inputs and warped exemplars, a novel confidence feature injection module is proposed to achieve multi-channel feature fusion based on the matching reliability of warped exemplars. Quantitative and qualitative experiments over multiple datasets show that RABIT is capable of achieving high-fidelity image translation and manipulation while preserving consistent semantics with the conditional input and faithful styles with the exemplar.

Acknowledgement. This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

1. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28**(3), 24 (2009)
2. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8789–8797 (2018)
3. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8188–8197 (2020)
4. Doersch, C.: Tutorial on variational autoencoders. arXiv preprint arXiv:1606.05908 (2016)
5. Dong, H., Liang, X., Gong, K., Lai, H., Zhu, J., Yin, J.: Soft-gated warping-gan for pose-guided person image synthesis. arXiv preprint arXiv:1810.11610 (2018)
6. Fu, Y., Ma, J., Ma, L., Guo, X.: Edit: Exemplar-domain aware image-to-image translation. arXiv preprint arXiv:1911.10520 (2019)
7. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016)
8. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: International conference on machine learning. pp. 1989–1998. PMLR (2018)
9. Hong, S., Yan, X., Huang, T., Lee, H.: Learning hierarchical semantic image manipulation through structured representations. In: NIPS (2018)
10. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1501–1510 (2017)
11. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 172–189 (2018)
12. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
13. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)
14. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
15. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 624–632 (2017)
16. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)
17. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: CVPR. pp. 5549–5558 (2020)

18. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
19. Li, Y., Cheng, Y., Gan, Z., Yu, L., Wang, L., Liu, J.: Bachgan: High-resolution image synthesis from salient object layout. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)
20. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms. arXiv preprint arXiv:1705.08086 (2017)
21. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017)
22. Liu, R., Lehman, J., Molino, P., Petroski Such, F., Frank, E., Sergeev, A., Yosinski, J.: An intriguing failing of convolutional neural networks and the coordconv solution. In: Advances in Neural Information Processing Systems (2018)
23. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1096–1104 (2016)
24. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015)
25. Lucas, J., Tucker, G., Grosse, R., Norouzi, M.: Don't blame the elbo! a linear vae perspective on posterior collapse. arXiv preprint arXiv:1911.02469 (2019)
26. Ma, L., Jia, X., Georgoulis, S., Tuytelaars, T., Van Gool, L.: Exemplar guided unsupervised image-to-image translation with semantic consistency. In: International Conference on Learning Representations (2018)
27. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose guided person image generation. In: Advances in neural information processing systems. pp. 406–416 (2017)
28. Mechrez, R., Talmi, I., Zelnik-Manor, L.: The contextual loss for image transformation with non-aligned data. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 768–783 (2018)
29. Men, Y., Mao, Y., Jiang, Y., Ma, W.Y., Lian, Z.: Controllable person image synthesis with attribute-decomposed gan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5084–5093 (2020)
30. Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K.: Image to image translation for domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4500–4509 (2018)
31. Ntavelis, E., Romero, A., Kastanis, I., Van Gool, L., Timofte, R.: Sesame: Semantic editing of scenes by adding, manipulating or erasing objects. In: ECCV. pp. 394–411. Springer (2020)
32. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2337–2346 (2019)
33. Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F.: Ganimation: Anatomically-aware facial animation from a single image. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 818–833 (2018)
34. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. IJCV (2000)
35. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: Pro-

ceedings of the IEEE conference on computer vision and pattern recognition. pp. 2107–2116 (2017)

36. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2107–2116 (2017)
37. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
38. Sun, W., Wu, T.: Image synthesis from reconfigurable layout and style. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10531–10540 (2019)
39. Tang, H., Xu, D., Liu, G., Wang, W., Sebe, N., Yan, Y.: Cycle in cycle generative adversarial networks for keypoint-guided image generation. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 2052–2060 (2019)
40. Tang, H., Xu, D., Sebe, N., Wang, Y., Corso, J.J., Yan, Y.: 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. pp. 2417–2426 (2019)
41. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7472–7481 (2018)
42. Wan, Z., Zhang, B., Chen, D., Zhang, P., Chen, D., Liao, J., Wen, F.: Bringing old photos back to life. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2747–2757 (2020)
43. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
44. Wu, R., Lu, S.: Leed: Label-free expression editing via disentanglement. In: European Conference on Computer Vision. pp. 781–798. Springer (2020)
45. Wu, R., Zhang, G., Lu, S., Chen, T.: Cascade ef-gan: Progressive facial expression editing with local focuses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5021–5030 (2020)
46. Xia, W., Yang, Y., Xue, J.H., Wu, B.: Tedigan: Text-guided diverse face image generation and manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2256–2265 (2021)
47. Xie, Y., Dai, H., Chen, M., Dai, B., Zhao, T., Zha, H., Wei, W., Pfister, T.: Differentiable top-k with optimal transport. Advances in Neural Information Processing Systems **33** (2020)
48. Zhan, F., Lu, S., Zhang, C., Ma, F., Xie, X.: Adversarial image composition with auxiliary illumination. In: Proceedings of the Asian Conference on Computer Vision (2020)
49. Zhan, F., Xue, C., Lu, S.: Ga-dan: Geometry-aware domain adaptation network for scene text detection and recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9105–9115 (2019)
50. Zhan, F., Yu, Y., Cui, K., Zhang, G., Lu, S., Pan, J., Zhang, C., Ma, F., Xie, X., Miao, C.: Unbalanced feature transport for exemplar-based image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2021)

51. Zhan, F., Yu, Y., Wu, R., Zhang, C., Lu, S., Shao, L., Ma, F., Xie, X.: Gmlight: Lighting estimation via geometric distribution approximation. arXiv preprint arXiv:2102.10244 (2021)
52. Zhan, F., Yu, Y., Wu, R., Zhang, J., Lu, S.: Multimodal image synthesis and editing: A survey. arXiv preprint arXiv:2112.13592 (2021)
53. Zhan, F., Yu, Y., Wu, R., Zhang, J., Lu, S., Zhang, C.: Marginal contrastive correspondence for guided image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10663–10672 (2022)
54. Zhan, F., Zhang, C., Hu, W., Lu, S., Ma, F., Xie, X., Shao, L.: Sparse needlets for lighting estimation with spherical transport loss. arXiv preprint arXiv:2106.13090 (2021)
55. Zhan, F., Zhang, C., Yu, Y., Chang, Y., Lu, S., Ma, F., Xie, X.: Emlight: Lighting estimation via spherical distribution approximation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 3287–3295 (2021)
56. Zhan, F., Zhang, J., Yu, Y., Wu, R., Lu, S.: Modulated contrast for versatile image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18280–18290 (2022)
57. Zhan, F., Zhu, H., Lu, S.: Spatial fusion gan for image synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3653–3662 (2019)
58. Zhang, J., Lu, S., Zhan, F., Yu, Y.: Blind image super-resolution via contrastive representation learning. arXiv preprint arXiv:2107.00708 (2021)
59. Zhang, P., Zhang, B., Chen, D., Yuan, L., Wen, F.: Cross-domain correspondence learning for exemplar-based image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5143–5153 (2020)
60. Zhao, B., Meng, L., Yin, W., Sigal, L.: Image generation from layout. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8584–8593 (2019)
61. Zheng, H., Lin, Z., Lu, J., Cohen, S., Zhang, J., Xu, N., Luo, J.: Semantic layout manipulation with high-resolution sparse attention. arXiv preprint arXiv:2012.07288 (2020)
62. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)
63. Zhou, X., Zhang, B., Zhang, T., Zhang, P., Bao, J., Chen, D., Zhang, Z., Wen, F.: Cocosnet v2: Full-resolution correspondence learning for image translation. In: CVPR (2021)
64. Zhu, J.Y., Krähenbühl, P., Shechtman, E., Efros, A.A.: Generative visual manipulation on the natural image manifold. In: ECCV. pp. 597–613. Springer (2016)
65. Zhu, P., Abdal, R., Qin, Y., Wonka, P.: Sean: Image synthesis with semantic region-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5104–5113 (2020)
66. Zhu, Z., Xu, Z., You, A., Bai, X.: Semantically multi-modal image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5467–5476 (2020)