Masked and Adaptive Transformer for Exemplar Based Image Translation

Chang Jiang[1], Fei Gao[1,2*], Biao Ma[1], Yuhao Lin[1], Nannan Wang[3], Gang Xu[1]
[1] School of Computer Science and Technology, Hangzhou Dianzi University
[2] Hangzhou Institute of Technology, Xidian University [3] ISN State Key Laboratory, Xidian University
{jc233, aiartma, linyh, gxu}@hdu.edu.cn, {fgao, nnwang}@xidian.edu.cn

Abstract

We present a novel framework for exemplar based image translation. Recent advanced methods for this task mainly focus on establishing cross-domain semantic correspondence, which sequentially dominates image generation in the manner of local style control. Unfortunately, cross-domain semantic matching is challenging; and matching errors ultimately degrade the quality of generated images. To overcome this challenge, we improve the accuracy of matching on the one hand, and diminish the role of matching in image generation on the other hand. To achieve the former, we propose a masked and adaptive transformer (MAT) for learning accurate cross-domain correspondence, and executing context-aware feature augmentation. To achieve the latter, we use source features of the input and global style codes of the exemplar, as supplementary information, for decoding an image. Besides, we devise a novel contrastive style learning method, for acquire quality-discriminative style representations, which in turn benefit high-quality image generation. Experimental results show that our method, dubbed MATEBIT, performs considerably better than state-of-the-art methods, in diverse image translation tasks. The codes are available at https://github.com/AiArt-HDU/MATEBIT.

1. Introduction

Image-to-image translation aims at transfer images in a source domain to a target domain [16, 50]. Early studies learn mappings directly by Generating Adversarial Networks (GANs), and have shown great success in various applications [2, 42]. Recently, exemplar based image translation [29, 30, 45], where an exemplar image is used to control the style of translated images, has attracted a lot of attention. Such methods allow high flexibility and controllability, and have a wide range of potential applications in social networks and metaverse. For example, people can transfer a facial sketch to an artistic portrait, in the style of oil paintings or avatars. Despite the remarkable progress, yielding high-fidelity images with consistent semantic and faithful styles remains a grand challenge.

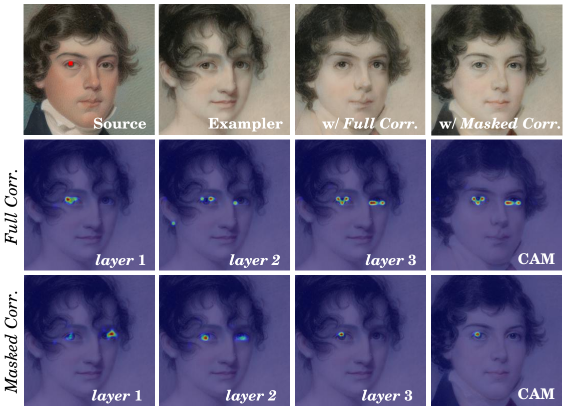


Figure 1. Visualization of correspondence maps. The red point is the query position. *Full Corr.* and *Masked Corr.* denote the full correspondence [45] and masked one in our method, respectively. CAM denotes visualization by *Class Activation Mapping* [48].

Early pioneering works [15, 21, 35] attempt to globally control the style of generated images. However, such methods ignore spatial correlations between an input image and an exemplar, and may fail to produce faithful details. Recently, some advanced methods [25, 44, 45, 49] first establish the cross-domain semantic correspondence between an input image and an exemplar, and then use it to warp the exemplar for controlling local style patterns. In these methods, the quality of generated images relies heavily on the learned correspondence [39]. Unfortunately, cross-domain semantic matching is challenging, since there is no reliable supervision on correspondence learning [45]. As a result, potential matching errors ultimately lead to degraded artifacts in generated images.

To combat this challenge, we propose to boost the matching accuracy on one hand, and to diminish the role of matching in image generation on the other hand. Inspired by the great success of Transformers [6, 10, 26, 41], we first devise a *Masked and Adaptive Transformer* (MAT) for learning ac-

*Corresponding Author

curate cross-domain correspondence and executing context-aware feature augmentation. Previous works [44, 45, 49] have used the vanilla attention mechanism [41] for learning full correspondence. However, the initial attention typically involves ambiguous correspondences (2nd row in Fig. 1). To mitigate these limitations, in MAT, we use a masked attention to distinguish the correspondence as reliable or not, and then reliability-adaptively aggregate representations. Besides, the *Feed-Forward Network* (FFN) [41] in vanilla transformers neglects contextual correlations inside an image. We thus replace FFN by an adaptive convolution block [28], where the coordinate attention [12] and depth-wise separable convolution [5] are used to capture contextual correlations and to improve efficiency. With a joint consideration of matching reliability and contextual correlations, MAT gradually focuses on accurate correspondences and emphasizes on features of interest (3rd row in Fig. 1).

In addition, to boost both the semantic consistency and style faithfulness, we supplementally use semantic features of the input image and global style codes of the exemplar for decoding an image. To this end, we first design our whole network following the U-Net architecture [16]. Besides, we devise a novel contrastive style learning (CSL) framework for acquiring discriminative style representations. Recently, Zhang et al. [47] propose a similar CSL method, where the target exemplar is used as a positive sample, and the other exemplars as negative ones. Differently, we use low-quality images, generated during early training stages, as negative samples. In this way, our style codes are desired to discriminate not only subtle differences in style, but also those in perceptual quality. Ultimately, the learned *global* style codes, cooperating with the *local* style control induced by MAT, in turn benefit high-quality image generation.

With the proposed techniques above, our full model, dubbed MATEBIT, diminishes the impact of position-wise matching on image quality, and integrates both local and global style control for image generation. Experimental results show that MATEBIT generates considerably more plausible images than previous state-of-the-art methods, in diverse image translation tasks. In addition, comprehensive ablation studies demonstrate the effectiveness of our proposed components. Finally, we perform interesting applications of photo-to-painting translation and Chinese ink paintings generation.

2. Relate Work

Exemplar Based Image Translation. Recently, exemplar based image translation has attracted increasing attention. For example, Park et al. [35] learn an encoder to map the exemplar image into a global style vector, and use it to guide image generation. Such a global style control strategy enables style consistency in whole, but fails to produce subtle details. Most recently, researchers propose a matching-then-generation framework [39]. Specially, they first establish dense correspondence between an input and an exemplar, and then reshuffle the exemplar for locally control the style of synthesize images. For example, Zhang et al. [45] establish position-wise correspondence based on the Cosine attention mechanism and warp the exemplar correspondingly. Afterwards, the warped image dominates the generation of images in the manner of SPADE [35]. To reduce the cost of matching in high-resolution image generation, Zhou et al. [49] introduce a hierarchical refinement of semantic correspondence from ConvGRU-PatchMatch. Besides, Liu et al. [25] used a dynamic pruning method for learning hierarchical sparse correspondence. They also use reliability-adaptive feature integration to improve the quality of generated images.

Previous methods merely use global or local style control, and the latter relies heavily on the learned correspondence. Besides, they consider little about contextual correlations inside an image. In this paper, we use both global and local style control to boost the style consistency. Besides, we take contextual correlations into consideration and execute reliability-adaptive feature augmentation.

Transformers. Transformers [41] have shown incredible success from the field of natural language processing (NLP) [19] to computer vision (CV) [6, 26]. Multi-head attention (MHA) and FFN are key components in a Transformer, and have been used in exemplar based image translation. However, they induce unreliable matching results and neglect context correlations in feature translation. In our MAT, we combat these limitations by replacing them with a masked attention and a context-aware convolution block, respectively. Recently, researchers use semantic masks to facilitate representation learning [4, 8, 37], where a mask predictor is required. Differently, we use a ReLU function to mask over the attention layer, for distinguishing correspondence as reliable or not (Sec. 3.1). In general, MAT follows a concise and efficient architecture.

Contrastive Learning. Contrastive learning has shown its effectiveness in various computer vision tasks [9, 13, 34]. The basic idea is to learn a representation by pushing positive samples toward an anchor, and moving negative samples away from it. Different sampling strategies and contrastive losses have been extensively explored in various downstream tasks. For example, Chen et al. [3] and He et al. [9] obtain positive samples by augmenting original data. In the field of image translation, Park et al. [34] propose patch-wise contrastive learning by maximizing the mutual information between cross-domain patches. Similarly, Zhang et al. [47] use contrastive learning for acquiring discriminative style representations. In the task of exemplar based image translation, Zhan et al. [44] use contrastive learning to align cross-domain images to a consistent semantic feature space, so as to boost the accuracy of matching. Differently, we

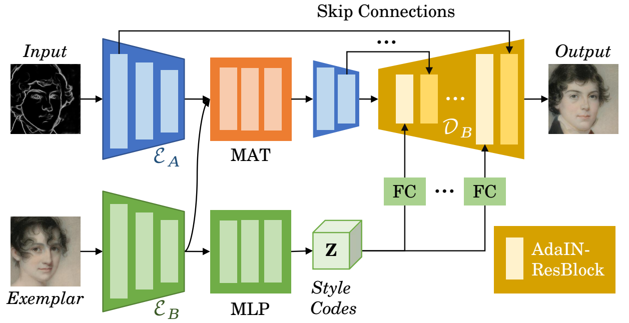


Figure 2. Overview of our image translation network, MATEBIT.

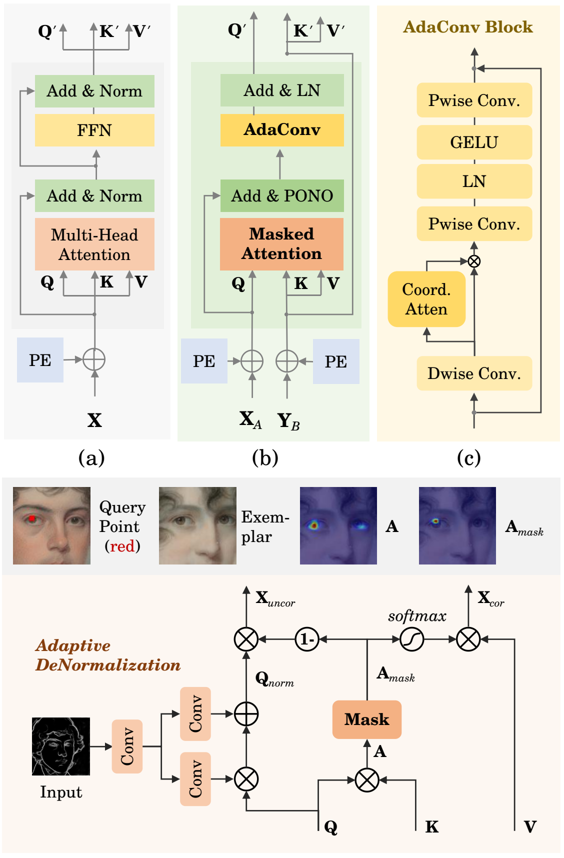
use early generated images as negative samples, so that the learned style representations can discriminate subtle differences in both style and perceptual quality (Sec. 3.2).

3. The Proposed Method

Given an input image x_A in domain $\mathcal{A}$ and an exemplar image y_B in domain $\mathcal{B}$, our goal is to generate a target image x_B which preserves semantic structures in x_A but resembles the style of similar parts in y_B. Fig. 2 shows an overview of our translation network $\mathcal{G}$. Specially, we first align x_A and y_B to an intermediate feature space by encoders $\mathcal{E}_A$ and $\mathcal{E}_B$, respectively. Afterwards, we use a *Masked and Adaptive Transformer* (MAT) for correspondence learning and feature augmentation. Finally, a decoder $\mathcal{D}_B$ produces an output image $\hat{x}_B$ based on the augmented features, as well as the source features and target style codes. Details are described below.

3.1. Masked and Adaptive Transformer (MAT)

In order to establish accurate cross-domain correspondence, we propose a novel and concise Transformer architecture, i.e. MAT. In general, the architecture of MAT (Fig. 3b) follows that of vanilla Transformers (Fig. 3a) [41]. Differently, we use masked attention to distinguish reliable and unreliable correspondence, instead of using multi-head attention. Besides, we use *Positional Normalization* (PONO) [23] and an *Adaptive Convolution* (AdaConv) block [28], instead of LN and MLP-based FFN, respectively. MAT is desired to gradually concentrate on accurate matching, and to reliability-adaptively augment representations with contextual correlations.



(a) (b) (c)

(d)

Figure 3. Detailed architectures. (a) Vanilla Transformer block, (b) MAT block, (c) AdaConv block, and (d) *Masked Attention*.

Masked Correspondence Learning. Let $\mathbf{X}_A \in \mathbb{R}^{H\times W\times C}$ and $\mathbf{Y}_B \in \mathbb{R}^{H\times W\times C}$ be the representations of x_A and y_B in the intermediate feature space, with height H, width W, and C channels. We first map $\mathbf{X}_A$ to the query $\mathbf{Q} \in \mathbb{R}^{HW\times C}$, and $\mathbf{Y}_B$ to the key $\mathbf{K} \in \mathbb{R}^{HW\times C}$ and value $\mathbf{V} \in \mathbb{R}^{HW\times C}$, by using 1×1 convolutions, respectively. As shown in Fig. 3d, we add positional encoding (PE) to $\mathbf{X}_A$ and $\mathbf{Y}_B$, for embedding spatial correlations. Afterwards, we learn the initial correspondence $\mathbf{A} \in \mathbb{R}^{HW\times HW}$ following the Cosine attention mechanism [45], i.e.

$$\mathbf{A}(u,v)=\frac{\tilde{\mathbf{Q}}(u)\tilde{\mathbf{K}}(v)^T}{||\tilde{\mathbf{Q}}(u)||\cdot||\tilde{\mathbf{K}}(v)||}, \tag{1}$$

with $\tilde{\mathbf{Q}}(u)=\mathbf{Q}(u)-\bar{\mathbf{Q}}(u)$, $\tilde{\mathbf{K}}(v)=\mathbf{K}(v)-\bar{\mathbf{K}}(v)$, where $u, v \in [1, ..., HW]$ are position indices; $\bar{\mathbf{Q}}(u)$ and $\bar{\mathbf{K}}(v)$ are the means of $\mathbf{Q}(u)$ and $\mathbf{K}(v)$, respectively. $\mathbf{A}(u,v)$ is the matching score between $\mathbf{Q}(u)$ and $\mathbf{K}(v)$.

Previous methods [44, 45] typically use the initial correspondence map $\mathbf{A}$ to reshuffle an exemplar for controlling local patterns in image synthesis. However, induced by the difficulties in cross-domain correspondence learning, $\mathbf{A}$ involves unreliable match scores (Fig. 3d). As a result, the reshuffled image will lead to implausible artifacts in generated images. To combat this limitation, we distinguish initial matching scores as reliable or not, according to their signs [32]. The masked correspondence map becomes:

$$\mathbf{A}_{mask} = \mathrm{ReLU}(\mathbf{A}), \tag{2}$$

In DynaST [25], two networks are used to predict the reliability mask of correspondence. However, it's challenging to effectively train the network, because there is no supervision on matching during training. In contrast, ReLU contains no learnable parameters and ultimately leads to superior performance over DynaST (Sec. 4.1).

Reliability-Adaptive Feature Aggregation. For regions with reliable correspondence in x_A, we use $\mathbf{A}_{mask}$ to warp the value features, $\mathbf{V}$, derived from the exemplar:

$$\mathbf{X}_{cor} = \tilde{\mathbf{A}}_{mask}\mathbf{V}, \text{ with } \tilde{\mathbf{A}}_{mask} = \text{softmax}(\alpha \cdot \mathbf{A}_{mask}), \tag{3}$$

where α is a scaling coefficient to control the sharpness of the softmax function. In default, we set its value as 100.

For regions with unreliable correspondence in x_A, $\mathbf{X}_{cor}$ provides an average style representation of $\mathbf{V}$. We further extract complementary information from the query, $\mathbf{Q}$, derived from the input. Inspired by SPADE [35], we first transfer $\mathbf{Q}$ to the target domain by using pixel-wise modulation parameters (i.e., $\boldsymbol{\gamma}$ for scale and $\boldsymbol{\beta}$ for bias) learned from x_A. The modulation is formulated by:

$$\mathbf{Q}_{norm} = \boldsymbol{\gamma}(x_A)\frac{\mathbf{Q} - \mu(\mathbf{Q})}{\sigma(\mathbf{Q})} + \boldsymbol{\beta}(x_A), \tag{4}$$

where $\mu(\mathbf{Q})$ and $\sigma(\mathbf{Q})$ are the mean value and standard deviance of $\mathbf{Q}$. Afterwards, we select the translated features of unreliably corresponded regions in x_A by:

$$\mathbf{X}_{uncor} = (1 - \sum\nolimits_j \mathbf{A}_{mask}) \odot \mathbf{Q}_{norm}, \tag{5}$$

where the summation is along the second dimension; $\odot$ denotes point-wise production with broadcasting. Since $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are learned from the input image x_A, the modulated features preserve the semantic information of x_A. Besides, constraints on the generated image will push the selected features convey to the style of y_B.

Ideally, $\mathbf{X}_{cor}$ and $\mathbf{X}_{uncor}$ would complement each other and facilitate both semantic consistency and style relevance in image generation. To this end, we integrate $\mathbf{X}_{cor}$, $\mathbf{X}_{uncor}$, and $\mathbf{Q}$ by:

$$\mathbf{X}_{agg} = \text{PONO}(\mathbf{X}_{cor} + \mathbf{X}_{uncor} + \mathbf{Q}). \tag{6}$$

In PONO [23], features at each position are normalized dependently. Compared to LN in vanilla transformers and DynaST [25], PONO boosts the flexibility in reliability-adaptive feature aggregation.

Context-Aware Feature Augmentation. Inspired by ConvNeXT [28], we replace FFN by an AdaConv block to position-adaptively emphasize informative representations. Besides, we use the *coordinate attention* (CoordAtten) module [12] to capture contextual correlations.

The architecture of the AdaConv block is as shown in Fig. 3c. We fist use the depthwise convolution (Dwise) to update representations in each channel separately; and then use two pointwise convolutions (Pwise) to automatically emphasize representations of interest, at every position. The *Gaussian Error Linear Unit* (GELU) activation function and *Layer Norm* (LN) are used after the first Pwise layer [28]. Notably, CoordAtten is used after the Dwise layer for modeling long-range dependencies in an image. Specially, CoordAtten produces cross-channel and position-sensitive attention maps, which helps our model to more accurately locate the representations of interest [12].

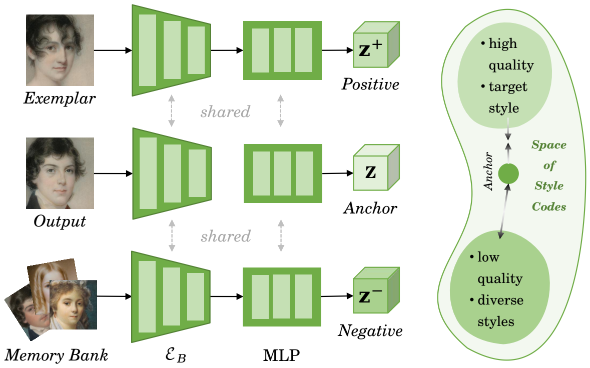


Figure 4. Contrastive style learning. The memory bank consists of divergent low-quality images generated in early training stages.

Finally, the output of a MAT block is obtained with a residual connection, i.e. $\mathbf{X}_{\text{MAT}} = \text{AdaConv}(\mathbf{X}_{agg}) + \mathbf{X}_{agg}$. In the implementation, we stack three MAT blocks in default to gradually refine the correspondence and to augment informative representations (Fig. 1). Empirical verifications will be given in Sec. 4.2.

Benefits of MAT. Fig. 3d illustrates the impact of MAT. The query point locates over the left eye of the source image. Here we show the magnitudes of its correspondence over the exemplar, in the third layer of MAT. Obviously, the original correspondence $\mathbf{A}$ covers both eyes of the exemplar. In contrast, the masked correspondence $\mathbf{A}_{mask}$ accurately concentrates over the left eye. Such superiority significantly boost the quality of ultimate images.

3.2. Contrastive Style Learning (CSL)

In MATEBIT, we use the encoder $\mathcal{E}_B$ to extract local style information $\mathbf{X}_B$, and then a MLP to extract global style codes $\mathbf{z}$. $\mathbf{X}_B$ and $\mathbf{z}$ perform local and global style control on generated images, respectively (Sec. 3.3). To boost the discriminative capacity of style representations, as well as the quality of generated images, we propose a novel *contrastive style learning* (CSL) method (as shown in Fig 4).

In our settings, the exemplars are drawn by human artists and thus considered as high-quality. In contrast, the images generated in early training stages are typically low-quality. Inspired by the idea of contrastive learning [9], we use the exemplar y_B as the positive sample, while a collection of early generated images as negative. Let $\mathbf{z}$ denotes style codes of the generated image $\hat{x}_B$, $\mathbf{z}^+$ that of exemplar y_B, and $\{\mathbf{z}_1^-, \mathbf{z}_2^-, ..., \mathbf{z}_m^-\}$ the style codes of m negative samples. CSL learns style representations by maximizing the mutual information between anchors and positive samples, while minimizing that between anchors and negative samples. Our contrastive style loss is computed by:

$$\mathcal{L}_{style} = -\log \frac{\exp(\frac{\mathbf{z}^T\mathbf{z}^+}{\tau})}{\exp(\frac{\mathbf{z}^T\mathbf{z}^+}{\tau}) + \sum_{j=1}^{m}\exp(\frac{\mathbf{z}^T\mathbf{z}_j^-}{\tau})}, \quad (7)$$

where $\tau = 0.07$ and $m = 1024$. In the implementation, we use a queue to cache negative style vectors.

3.3. Translation network

To boost both the semantic consistency and style faithfulness, we additionally use source semantic features and global style codes for decoding an image. Specially, we design our whole translation network following U-Net (Fig. 2), where the multi-level features in $\mathcal{E}_A$ are skip-connected to the decoder $\mathcal{D}_B$, for supplementing informative sematic structures of the input image x_A. Besides, we use the style codes $\mathbf{z}$ to globally control the style of generated images, in the manner of AdaIN [14]. Specially, $\mathbf{z}$ is mapped to channel-wise modulating factors by fully-connected (FC) layers. In this way, we diminish the impact of correspondence learning on image generation, and provide reliable style control for even unmatched regions.

In summary, our translation network allows both local and global style control, and reuses the semantic features of input images. As a result, the generated image is desired to present consistent semantic to the input x_A and faithful style to the exemplar y_B. More details of our network are available in the supplementary material.

3.4. Loss functions

Our whole network is end-to-end optimized to jointly achieve high-fidelity image generation and accurate correspondence. Following [45], we obtain training triplets $\{x_A, y_B, x_B\}$ from the ready-made data pair $\{x_A, x_B\}$, where y_B is a geometrically warped version of x_B. The generated image is denoted by $\hat{x}_B = \mathcal{G}(x_A, y_B)$. Our loss functions are similar to [45], except for the previous contrastive style loss $\mathcal{L}_{style}$ and the structural loss $\mathcal{L}_{str}$ below.

Semantic alignment Loss. For accurate cross-domain correspondence learning, the encoders $\mathcal{E}_A$ and $\mathcal{E}_B$ should align x_A and x_B to consistent representations. The corresponding semantic alignment loss is:

$$\mathcal{L}_{align} = \|\mathcal{E}_A(x_A) - \mathcal{E}_B(x_B)\|_1 . \quad (8)$$

Correspondence Loss. Ideally, if we warp y_B in the same way as Eq.3, the resulting image should be exactly x_B. We thus constrain the learned correspondence by:

$$\mathcal{L}_{corr} = \left\| \tilde{\mathbf{A}}^T_{mask} y_B \downarrow - x_B \downarrow \right\|_1 , \quad (9)$$

where $\downarrow$ indicates down-sampling y_B and x_B to the size (i.e. width and height) of $\mathbf{X}_A$.

Perceptual Loss. The generated image $\hat{x}_B$ should be semantic-consistent with the ground truth x_B in term of semantic. We thus use the perceptual loss:

$$\mathcal{L}_{perc} = \|\varphi_l(\hat{x}_B) - \varphi_l(x_B)\|_1 , \quad (10)$$

where φ_l denotes the activations after layer $relu4\_2$ in pretrained VGG19 [40] , which represent high-level semantics.

Contextual Loss. In addition, the generated image should be in the same style as the exemplar. In addition to the previous contrastive style loss (Eq.7), we additionally use the contextual loss (CX) [31] to constrain on local style consistency. The contextual loss is computed by:

$$\mathcal{L}_{ctx} = -\log\left(\sum_l w_l \mathrm{CX}(\varphi_l(\hat{x}_B), \varphi_l(y_B))\right) \quad (11)$$

where w_l balances the terms of different VGG19 layers.

Structural Loss. The generated image should preserve semantic structures in the input image. Correspondingly, we use the *Learned Perceptual Image Patch Similarity* (LPIPS) [46] between their boundaries as the structural loss:

$$\mathcal{L}_{str} = \mathrm{LPIPS}(\mathcal{H}(\hat{x}_B), \mathcal{H}(x_B)), \quad (12)$$

where $\mathcal{H}$ is the HED algorithm [43], which has been widely used for extracting semantic boundaries in an image.

Adversarial loss. Finally, we add a discriminator $\mathcal{D}$ to distinguish real images in domain $\mathcal{B}$ and the generated images [7]. The adversarial loss is:

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

where $h(t) = \min(0, -1 + t)$ is the hinge loss function [1].

Total loss. In summary, our overall objective function is,

$$\begin{aligned} \min_{\mathcal{G}} \max_{\mathcal{D}} \; & \lambda_1\mathcal{L}_{style} + \lambda_2\mathcal{L}_{align} + \lambda_3\mathcal{L}_{corr} + \lambda_4\mathcal{L}_{str} \\ & + \lambda_5(\mathcal{L}_{perc} + \mathcal{L}_{ctx}) + \lambda_6(\mathcal{L}_{adv}^{\mathcal{G}} + \mathcal{L}_{adv}^{\mathcal{D}}) \end{aligned} \quad (14)$$

where λ denotes the weight parameters.

4. Experiment

Implementation details. We apply spectral normalization [33] to all the layers in the translation network and discriminator. We use the Adam [20] solver with $\beta_1 = 0$ and $\beta_2 = 0.999$. The learning rates for the generator and discriminator are set as $1e{-}4$ and $4e{-}4$ respectively, following TTUR [11]. The experiments are conducted using 4 24GB RTX3090 GPUs. Limited by the computation load, we restrict the resolution of generated images to 256 × 256 in all translation tasks.

Datasets. We mainly conduct experiments on the following datasets. (1) **CelebA-HQ** [22] contains 30,000 facial photos. We chose 24,000 samples as the training set

Table 1. Comparison on the Metfaces [18], CelebA-HQ [22], Ukiyo-e [36],Cartoon [36], AAHQ [24], and DeepFashion [27] datasets.

	CelebA-HQ					Metfaces		Cartoon		Ukiyo-e		AAHQ		DeepFashion		Time ↓
	FID ↓	SWD ↓	Texture ↑	Color ↑	Semantic↑	FID ↓	SWD ↓	FID ↓	SWD ↓	FID ↓	SWD ↓	FID ↓	SWD ↓	FID ↓	SWD ↓	(*s*)
SPADE [35]	31.5	26.9	0.927	0.955	0.922	45.6	26.9	97.5	30.5	45.6	26.9	79.4	32.1	36.2	27.8	0.196
CoCosNet [45]	14.3	15.2	0.958	0.977	0.949	25.6	24.3	66.8	27.1	38.3	13.9	62.6	21.9	14.4	17.2	0.321
CoCosNet-v2 [49]	13.2	14.0	0.954	0.975	0.948	**23.3**	22.4	66.4	27.0	32.1	**11.0**	62.4	22.8	13.0	16.7	1.573
MCL-Net [44]	12.8	14.2	0.951	0.976	**0.953**	23.8	24.5	67.9	27.9	32.4	12.4	64.4	22.2	12.9	16.2	0.309
DynaST [25]	12.0	**12.4**	0.959	0.978	0.952	29.2	28.6	**62.8**	**26.5**	38.9	14.2	67.2	24.0	8.4	11.8	0.214
MATEBIT (ours)	**11.5**	13.2	**0.966**	**0.986**	0.949	26.0	**19.1**	64.4	27.6	**30.3**	11.5	**56.0**	**19.5**	**8.2**	**10.0**	**0.185**

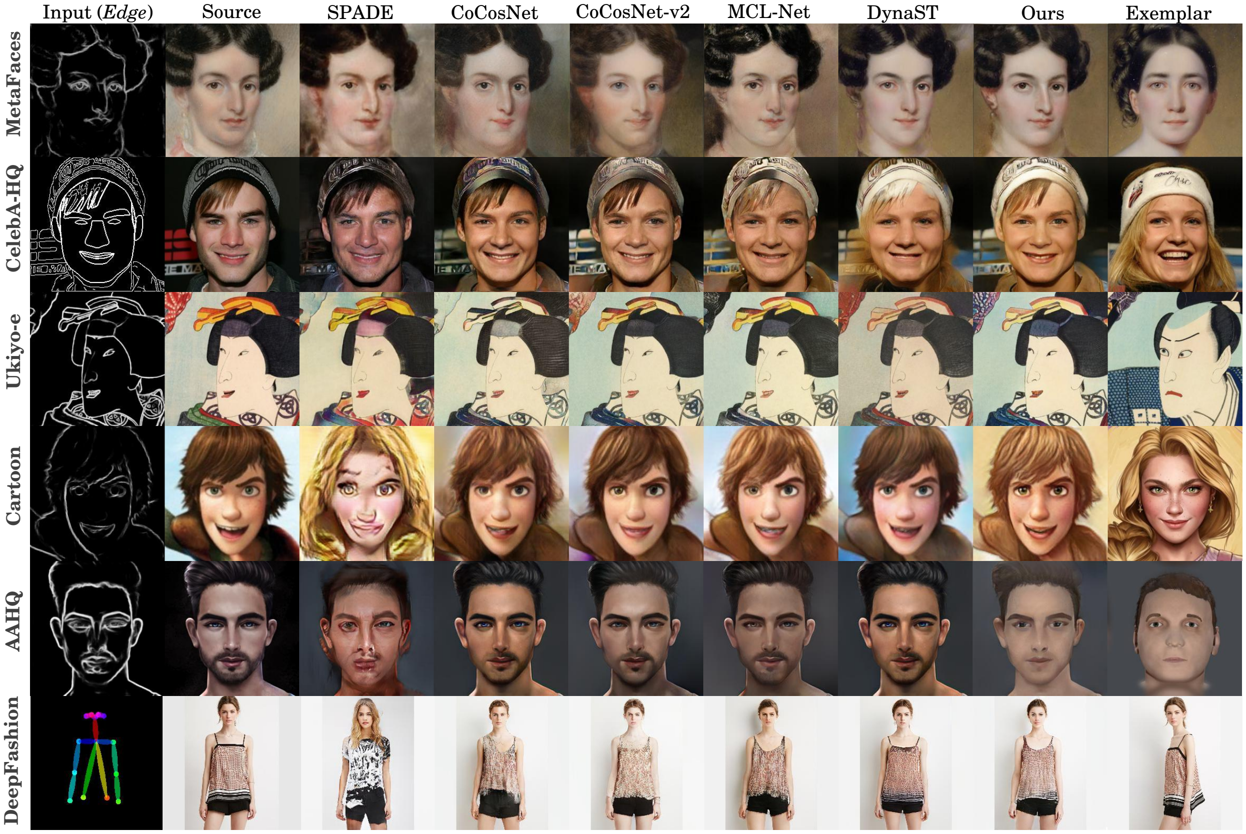


Figure 5. Results on the Metfaces [18], CelebA-HQ [22], Ukiyo-e [36], Cartoon [36], AAHQ [24], and DeepFashion [27]datasets.

and 3000 as the test set. (2) **Metfaces** [18] consists of 1336 high-quality artistic facial portraits. (3) **AAHQ** [24] consists of high-quality facial avatars. We randomly select 1500 samples for training and 1000 samples for testing. (4) **Ukiyo-e** [36] consists of high-quality Ukiyo-e faces. We randomly select 3000 and 1000 samples for training and testing, respectively. (5) **Cartoon** [36] consists of 317 cartoon faces. (6) **DeepFashion** [27] consists of 800,00 fashion images. On CelebA-HQ, we connect the face landmarks for face region, and use Canny edge detector to detect edges in the background. On DeepFashion, we use the officially provided landmarks as input. On the other datasets, we use HED [43] to obtain semantic edges.

4.1. Comparison with state-of-the-art

We select several advanced models, including SPADE [35], CoCosNet [45], CoCosNet-v2 [49], MCL-Net [44], and DynaST [25], for comparison. For a fair comparison, we retrain their models at resolution 256 × 256 under the same settings as ours.

Quantitative evaluation. We adopt several criteria to fully evaluate the generation results. (1) *Fréchet Inception Score* (FID) [38] and *Sliced Wasserstein distance* (SWD) [17] are used to evaluate the image perceptual quality. (2) To assess style relevance and semantic consistency of translated images [45], we compute the *color*, *texture*, and *semantic* metrics based on VGG19 [40]. Specifically, the cosine similarities between low-level features (i.e. $relu1\_2$ and $relu2\_2$) are used to measure *color* and *texture* relevance, respectively; the average cosine similarity between high-level features (i.e. $relu3\_2$, $relu4\_2$, and $relu5\_2$) measures the *semantic* consistency.

The quantitative comparison results are shown in Table 1. Compared to existing methods, our model consistently achieves superior or highly competitive performance across

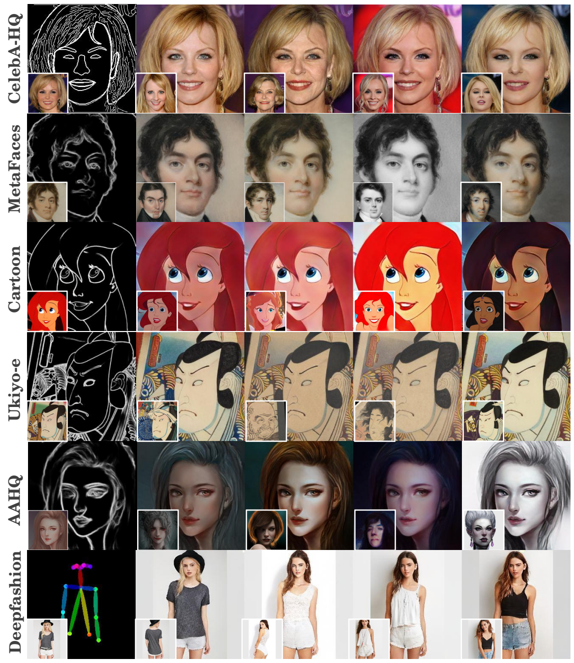


Figure 6. More results generated by MATEBIT. Images shown in the left-bottom corner are source images or exemplars.

all the datasets. Especially, MATEBIT significantly improves the style relevance in both texture and color. On the complicated AAHQ dataset, which contains diverse styles of avatars, MATEBIT dramatically decreases both FID and SWD. Such superiority indicates that our generated images are of better perceptual quality; and present consistent appearance to similar parts in exemplars. We additionally report the average time each method costs for generating an image. Our method shows the best efficiency and is significantly faster than previous methods.

Qualitative comparison. Fig 5 illustrates images generated by different methods. Obviously, previous methods present geometric distortions, blurring artifacts, inconsistent colors, or identity inconsistency. In contrast, MATEBIT consistently produces appealing results, including more results shown in Fig. 6. Specially, our results preserve the semantic structure of input images, and present consistent appearance with semantically similar regions in exemplars. Previous methods suffer serious degradations mainly due to the matching errors in full correspondence learning. In our method, we distinct reliable and unreliable correspondence, and release the role of matching in image generation. As a result, our method stably transfers a source image to the target style of an exemplar.

4.2. Ablation study

Impacts of MAT. We present a comprehensive analysis to justify the important component in our architecture, i.e. MAT. We here modify our full model by (1) removing the MAT module (i.e. w/o MAT), (2) removing ReLU in MAT (i.e. w/o ReLU), (3) replacing MAT with three-layer full correspondence learning modules [45] (i.e. *Full Corr.*), and (4) replacing the AdaConv with FFN [45] (i.e. *w/ AdaConv*). The results in Table 2 show that removing MAT or ReLU dramatically hurts the performance. Besides, using the full correspondence learning in [45] or using FFN also significantly decreases the texture relevance and semantic consistency. Correspondingly, these model variants leads to inferior results in terms of textures or colors, compared to our full model (Fig. 8). Recall the visualized correspondence in Fig. 1, our method learns remarkably accurate correspondence, which ultimately benefits the quality of generated images. In addition, Fig. 7 shows that both the semantic consistency and style realism broadly improve with the number of MAT blocks and peak at three. All these observations demonstrate our motivation that MAT gradually refines cross-domain correspondence and augments informative representations for generating high-quality images.

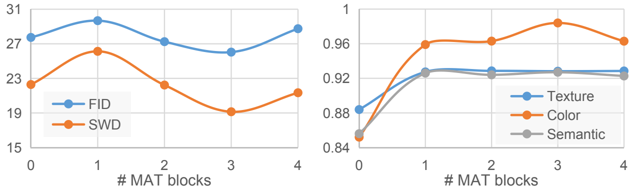


Figure 7. Impact of the number of MAT blocks on performance.

Table 2. Results of ablation studies on MetaFaces.

	FID ↓	SWD ↓	Texture ↑	Color ↑	Semantic↑
w/o MAT	27.7	22.3	0.883	0.852	0.856
w/o ReLU	30.1	20.0	0.916	0.956	**0.928**
w/o MAT (→ Full Corr. [45])	34.1	19.7	0.872	0.969	0.902
w/o AdaConv (→ FFN [41])	34.3	20.1	0.841	0.971	0.896
w/o $\mathcal{L}_{style}$	30.3	20.8	0.874	0.848	0.908
w/o $\mathcal{L}_{style}$ (→ $\mathcal{L}_{\text{CAST}}$ [47])	31.0	19.8	0.904	0.983	0.921
w/o $\mathcal{L}_{str}$	28.4	21.8	0.915	0.983	0.911
w/o skip connection	47.0	27.7	0.925	0.958	0.902
w/o global style **z**	27.6	21.2	0.927	0.961	0.925
Full Model	**26.0**	**19.1**	**0.938**	**0.984**	0.927

Contrastive Style Loss. To verify the effectiveness of the proposed contrastive style learning methodology, we train our model by (1) removing the style loss (i.e. w/o $\mathcal{L}_{style}$) and (2) replacing $\mathcal{L}_{style}$ with the loss used in CAST [47] (i.e. w/ $\mathcal{L}_{\text{CAST}}$). In CAST, only high-quality exemplars in different styles are used as negative samples. Differently, we use low-quality generated images in diverse styles as negative samples. From both Table 2 and Fig. 8, we observe that: (1) without $\mathcal{L}_{style}$, although the generated images show high semantic consistency, they present low style relevance; (2) $\mathcal{L}_{\text{CAST}}$ benefits the style relevance, but leads to inferior performance to $\mathcal{L}_{style}$. These comparison results meet our expectation that: our CSL methodology enables the learned style codes to discriminate subtle divergences between images with different perceptual qualities. Such discriminability facilitates pushes the network to generate high-quality images.

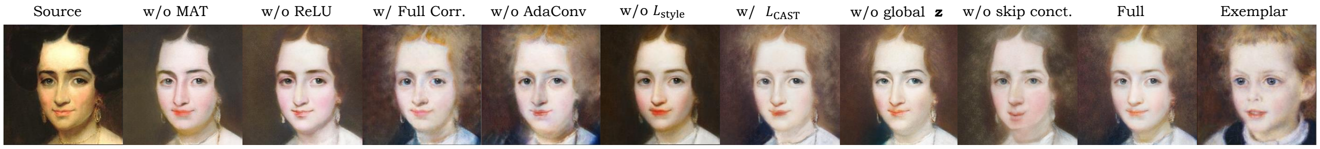


Figure 8. Comparison of generated images by different variants of our model, on Metfaces [18].

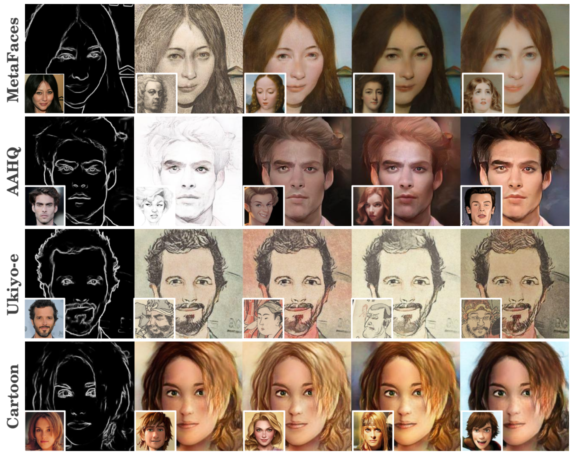


Figure 9. Our method can transfer facial photos to artistic portraits in the style of exemplars.

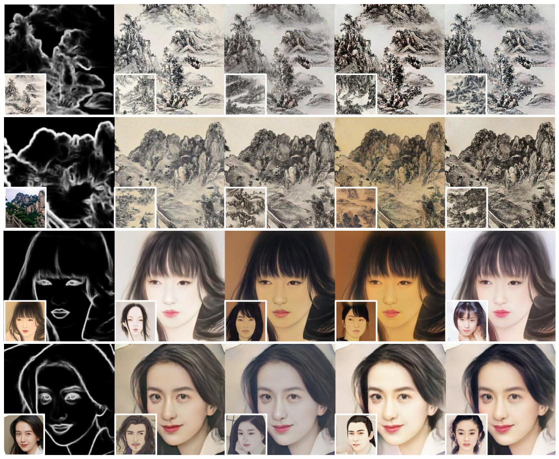


Figure 10. Chinese ink paintings generation (1st & 3rd rows), as well as photo-to-painting translation (2nd & 4th rows).

Skip connections & global style control. In MATEBIT, we use skip connections to supplement input semantic information. Removing skip connections dramatically hurts the semantic inconsistency and the quantitative results. Besides, using global style vector $\mathbf{z}$ increases subtle details, e.g. the colors over the mouth, rings, and hairs.

In summary, MAT learns accurate correspondence and enables context-aware feature augmentation; the contrastive style learning benefits the style control and high-quality image generation; and the U-Net architecture helps the preservation of semantic information. Ultimately, all these benefits make our model significantly outperform previous state-of-the-art methods in generating plausible images.

4.3. Applications

Artistic Portrait Generation. An potential application of our method is transferring a facial photo to an artistic portrait, in the style of an exemplar. We here apply the previously learned models to randomly selected facial photos from CelebA-HQ [22]. As illustrated in Fig. 9, our method can generate appealing portraits with consistent identity and faithful style appearance.

Chinese Ink Painting Generation. To verify the capacity of our model in generating complex images, we additionally apply it to generate Chinese Ink paintings. Specially, we collect paintings of landscapes and facial portraits from the web, and then train and test our model on each subset respectively. Fig. 10 illustrates the results of painting generation and photo-to-painting translation. Obviously, all the generated images show remarkably high quality. Besides, our model successfully captures subtle differences between different exemplars, demonstrating its remarkable capacity in style control.

5. Conclusions

This paper presents a novel exemplar-guided image translation method, dubbed MATEBIT. Both quantitative and qualitative experiments show that MATEBIT is capable of generating high-fidelity images in a number of tasks. Besides, ablation studies demonstrate the effectiveness of MAT and contrastive style learning. Despite such achievements, the artistic portraits transferred from facial photos (Fig. 9) are inferior to those shown in Fig. 6. This may be due to the subtle differences in edge maps between photos and artistic paintings. In the near future, we will explore to solve this issue via semi-supervised learning or domain transfer technologies.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0103202; in part by the National Natural Science Foundation of China under Grants 61971172, 61971339, 62176230, U22A2033, U22A2096; in part by the Zhejiang Provincial Science and Technology Program in China under Grant 2021C01108; in part by the NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Informatization under Grant U1909210; in part by the Technology Innovation Leading Program of Shaanxi under Grant 2022QFY01-15; and in part by Open Research Projects of Zhejiang Lab under Grant 2021KG0AB01.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. 2018. 5

[2] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 40–48, 2018. 1

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[4] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 2

[5] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 2

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 1, 2

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 5

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2, 4

[10] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11936–11945, 2021. 1

[11] 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. *Advances in neural information processing systems*, 30, 2017. 5

[12] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13713–13722, 2021. 2, 4

[13] Xueqi Hu, Xinyue Zhou, Qiusheng Huang, Zhengyi Shi, Li Sun, and Qingli Li. Qs-attn: Query-selected attention for contrastive learning in i2i translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18291–18300, 2022. 2

[14] 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. 5

[15] 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. 1

[16] 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. 1, 2

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 6

[18] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020. 6, 8

[19] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 2

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[21] Thomas N Kipf and Max Welling. Variational graph autoencoders. *arXiv preprint arXiv:1611.07308*, 2016. 1

[22] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020. 5, 6, 8

[23] Boyi Li, Felix Wu, Kilian Q Weinberger, and Serge Belongie. Positional normalization. In *Advances in Neural Information Processing Systems*, pages 1620–1632, 2019. 3, 4

[24] Mingcong Liu, Qiang Li, Zekui Qin, Guoxin Zhang, Pengfei Wan, and Wen Zheng. Blendgan: implicitly gan blending for arbitrary stylized face generation. *Advances in Neural Information Processing Systems*, 34:29710–29722, 2021. 6

[25] Songhua Liu, Jingwen Ye, Sucheng Ren, and Xinchao Wang. Dynast: Dynamic sparse transformer for exemplar-guided image generation. In *European Conference on Computer Vision*, 2022. 1, 2, 3, 4, 6

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1, 2

[27] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 6

[28] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 2, 3, 4

[29] Mandi Luo, Haoxue Wu, Huaibo Huang, Weizan He, and Ran He. Memory-modulated transformer network for heterogeneous face recognition. *IEEE Transactions on Information Forensics and Security*, 2022. 1

[30] Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, and Luc Van Gool. Exemplar guided unsupervised image-to-image translation with semantic consistency. *Proceedings ICLR 2019*, 2019. 1

[31] 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

[32] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3395–3404, 2019. 3

[33] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. 5

[34] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*, pages 319–345. Springer, 2020. 2

[35] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 1, 2, 4, 6

[36] Justin N. M. Pinkney. Aligned ukiyo-e faces dataset, 2020. 6

[37] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021. 2

[38] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. version 0.2.1. 6

[39] Junyoung Seo, Gyuseong Lee, Seokju Cho, Jiyoung Lee, and Seungryong Kim. Midms: Matching interleaved diffusion models for exemplar-based image translation. *arXiv preprint arXiv:2209.11047*, 2022. 1, 2

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR*, 2015. 5, 6

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 2, 3, 7

[42] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *European Conference on Computer Vision*, pages 63–79. Springer, 2018. 1

[43] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 5, 6

[44] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu, and Changgong Zhang. Marginal contrastive correspondence for guided image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10663–10672, 2022. 1, 2, 3, 6

[45] 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, 3, 5, 6, 7

[46] 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. 5

[47] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH*, 2022. 2, 7

[48] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 1

[49] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. Cocosnet v2: Full-resolution correspondence learning for image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11465–11475, 2021. 1, 2, 6

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1