CoCosNet v2: Full-Resolution Correspondence Learning for Image Translation

Xingran Zhou[1*] Bo Zhang[2] Ting Zhang[2] Pan Zhang[4] Jianmin Bao[2]
Dong Chen[2] Zhongfei Zhang[3] Fang Wen[2]
[1]Zhejiang University [2]Microsoft Research Asia [3]Binghamton University [4]USTC

Abstract

We present the full-resolution correspondence learning for cross-domain images, which aids image translation. We adopt a hierarchical strategy that uses the correspondence from coarse level to guide the fine levels. At each hierarchy, the correspondence can be efficiently computed via PatchMatch that iteratively leverages the matchings from the neighborhood. Within each PatchMatch iteration, the ConvGRU module is employed to refine the current correspondence considering not only the matchings of larger context but also the historic estimates. The proposed CoCosNet v2, a GRU-assisted PatchMatch approach, is fully differentiable and highly efficient. When jointly trained with image translation, full-resolution semantic correspondence can be established in an unsupervised manner, which in turn facilitates the exemplar-based image translation. Experiments on diverse translation tasks show that CoCosNet v2 performs considerably better than state-of-the-art literature on producing high-resolution images.

1. Introduction

Image-to-image translation learns the mapping between image domains and has shown success in a wide range of applications [28, 10, 38, 45, 58]. Particularly, exemplar based image translation allows more flexible user control by conditioning the translation on a specific exemplar with the desired style. However, simultaneously producing high quality while being faithful to the exemplar is non-trivial, whereas it becomes rather challenging for producing high-resolution images.

Early studies [9, 19, 55, 54, 47, 5] directly learn the mapping through generative adversarial networks [14, 35], yet they fail to leverage the information in the exemplar. Later, a series of methods [12, 17, 39] propose to refer to the exemplar image during the translation, by modulating the feature normalization according to the style of the exemplar image. However, as the modulation is applied uniformly, only the global style can be transferred whereas the detailed textures are washed out in the final output.

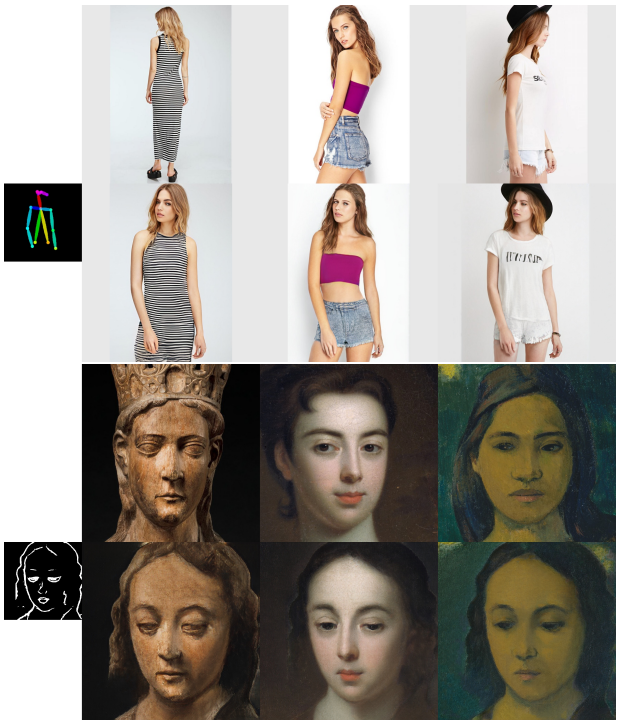

Figure 1: Image translation at resolution 512×512 for pose-to-body (DeepFashion) and at resolution 1024×1024 for edge-to-face (MetFaces). For each task, the 1st row shows the exemplar images, and the 2nd row shows the translation outputs.

Very recently, CoCosNet [56] established the dense semantic correspondence between cross-domain images. In this way the network could make use of the fine textures from the exemplar, which eases the hallucination for the local textures. However, prohibitive memory footprint occurs when estimating high-resolution correspondence, as the matching requires to compute the pairwise similarities among all locations of the input feature maps, while low-resolution correspondences (*e.g.*, 64×64) cannot guide the network to leverage the fine structures from the exemplar.

In this paper, we propose the cross-domain correspondence learning, *in full-resolution* for the first time, which leads to high-resolution translated images in photo-realistic quality, as the network can leverage the meticulous details

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

from the exemplar. To achieve that, we draw inspiration from PatchMatch [3] which is advantageous in computational efficiency and texture coherency as it iteratively propagates the correspondence from the neighborhood rather than searching globally. Nonetheless, directly applying PatchMatch to high-resolution feature maps for training is infeasible and the reasons are threefold. First of all, this algorithm is not efficient enough for high-resolution images when the correspondence is initialized randomly. Second, at the early training phase, the correspondence is chaotic and the backward gradient flows to the incorrectly corresponded patches, making the feature learning difficult. Moreover, PatchMatch fails to consider a larger context when propagating the correspondence estimate and requires a large number of iterations to converge.

To tackle these limitations, we propose the following techniques to learn the full-resolution correspondence. 1) We adopt a hierarchical strategy that makes use of the matchings from the coarse level to guide the subsequent, finer levels so that the searching at the fine levels may start with a good initialization. 2) Enlightened by the recent success of recurrent refinement [41, 7, 44], we employ convolutional gated recurrent unit (ConvGRU) to refine the correspondence within each PatchMatch iteration. The GRU-assisted PatchMatch considers a larger context as well as the historic correspondence estimates, which considerably improves the correspondence quality. Besides, it greatly benefits the feature learning as the gradient can now flow to a larger context than just a few corresponded patches. 3) Last but not least, the proposed hierarchical GRU-assisted PatchMatch is fully differentiable, and learns the cross-domain correspondence in an unsupervised manner, which is very challenging especially in high-resolution.

We show that our method, called CocosNet v2, achieves significantly higher quality images than the state-of-the-art litearture due to the full-resolution cross-domain correspondence. More importantly, our approach is able to generate visually appealing image translation results in high-resolution, *e.g.*, images at 512×512 and 1024×1024 (Figure 1). We summarize our major contributions as follows:

- We propose to learn full-resolution correspondence from different domains in order to capture meticulously realistic details from an exemplar image for image translation.
- To achieve that, we propose CoCosNet v2, a hierarchical GRU-assisted PatchMatch method, for efficient correspondence computation, which is simultaneously learned with image translation.
- We show that the full-resolution correspondence leads to significantly finer textures in the translation output. The translated images demonstrate unprecedented quality at large resolutions.

2. Related works

PatchMatch. Correspondence matching is a fundamental problem in computer vision [6, 27, 50, 31, 11, 13, 49]. The prohibitively high computational challenge has been largely alleviated by the pioneering work, PatchMatch [3]. The key insights stem from two principles: 1) good patch matches can be found via random sampling; 2) images are coherent such that matches can be propagated to nearby areas. Due to its efficiency, PatchMatch has been successfully applied to different tasks [25, 4, 2, 16, 11]. However, traditional PatchMatch can only find matches with image and is unsuitable to deep neural networks. Recently, [11] proposes to make the whole matching process differentiable and enables the feature learning and correspondence learning end-to-end. However, this method is still computational prohibitive to learn high-resolution correspondence during training. In contrast, we apply PatchMatch in hierarchy, and propose a novel GRU-assisted refinement module to consider a larger context, which enables a faster convergence and a more accurate correspondence. It is worth noting that [24, 26] use PatchMatch for style transfer, but they operate on the pre-trained VGG features and require the input to be natural images, whereas we allow the feature learning for arbitrary domain inputs such as pose or edge.

Image-to-image translation. Image translation methods [19, 47, 39, 59, 52, 22, 28, 43] typically resort to a conditional generative adversarial network and optimize the network through either paired data with explicit supervision or unpaired data by enforcing cycle consistency. Recently, exemplar-based image translation [18, 40, 46, 32, 42, 1, 53] have attracted a lot of interest due to its flexibility and improved generation quality. While most methods transfer the global style from the reference image, a recent work, CoCosNet [56] proposes establishing the dense semantic correspondence to the cross-domain inputs, and thus better preserves the fine structures from the exemplar. Our work is closely related to CoCosNet [56] but has a substantial improvement. We aim to compute dense correspondence on full-resolution whereas [56] can only find the correspondence on a small scale. Due to the full-resolution correspondence, our network can leverage finer structures from the exemplar, and thus achieves a superior quality on high-resolution outputs.

3. CoCosNet v2

Given an image x_A in the source domain $\mathcal{A}$ and an image y_B in the target domain $\mathcal{B}$, we propose to learn full-resolution cross-domain correspondences that aim to capture finer details and serve as a better guidance in exemplar-based image translation. Specifically, x_A and y_B are first represented as multi-level features (Section 3.1). Thereafter the correspondences are established starting from low-

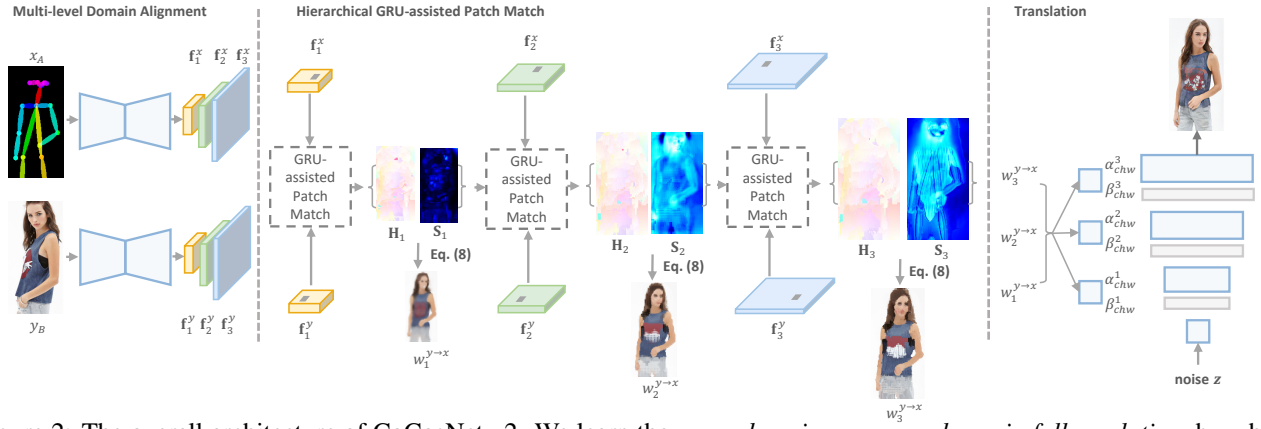

Figure 2: The overall architecture of CoCosNet v2. We learn the *cross-domain correspondence in full resolution*, by which we warp the exemplar images ($w_i^{y\to x}$) and feed them into the translation network for further rendering. The full-resolution correspondence is learned hierarchically, where the low-resolution result serves as the initialization for the next level. In each level, the correspondence can be efficiently computed via differentiable PatchMatch, followed by ConvGRU for recurrent refinement.

resolution to full-resolution, which are further used to warp the exemplar to align with x_A (Section 3.2). Finally, the warped exemplars are passed through a translation network to generate the desired output image (Section 3.3). We illustrate the whole network architecture in Figure 2.

3.1. Multi-level domain alignment

We first learn a common latent space $\mathcal{S}$ in which the representation contains the semantic contents for both domains and the features can be compared under some similarity metric. Similar to prior work [56], we learn two mapping functions for both domains respectively. We build a pyramid of L latent spaces ranging from low-resolution to high-resolution, instead of creating merely one latent space. For feature extraction, we adopt a U-net architecture to enable rich contextual information being propagated to higher resolution features by means of skip connections.

Formally, let $\mathcal{M}_A$ and $\mathcal{M}_B$ be the corresponding two mapping functions, we have the multi-level latent features,

$$\mathbf{f}_1^x, \cdots, \mathbf{f}_L^x = \mathcal{M}_A(x_A; \theta_{\mathcal{M}_A}), \tag{1}$$

$$\mathbf{f}_1^y, \cdots, \mathbf{f}_L^y = \mathcal{M}_B(y_B; \theta_{\mathcal{M}_B}), \tag{2}$$

where $\mathbf{f}_l^x \in \mathbb{R}^{H_l W_l \times C_l}$ with the height $H_1 < \cdots < H_L$, width $W_1 < \cdots < W_L$, and C_l denotes channel number. Latent features $\{\mathbf{f}_1^x, \cdots, \mathbf{f}_L^x\}$ are enlarged from small resolution to the full resolution. $\{\mathbf{f}_1^y, \cdots, \mathbf{f}_L^y\}$ have similar meanings, whereas, $\theta_{\mathcal{M}_A}$ and $\theta_{\mathcal{M}_B}$ denote the parameters.

3.2. Hierarchical GRU-assisted PatchMatch

It is worth noting the previous works compute dense correspondence field at the low-resolution level because of memory constraints and speed limitations. We propose to exploit the correspondences on the full-resolution feature level, *i.e.*, $\mathbf{f}_L^x$ and $\mathbf{f}_L^y$, and present a novel effective approach that is much less demanding in memory and time.

Coarse-to-fine strategy. Directly establishing the correspondences on full-resolution features not only increases the computational complexity, but also magnifies the noise and ambiguities of small patches. To deal with that, we propose a coarse-to-fine strategy on the pyramid of latent representations. In particular, we start with correspondence matching at the lowest resolution level, and use the matching results as the initial guidance at the subsequent, higher-resolution level. In this way, the correspondence fields of all the levels can be acquired. Formally we have,

$$\mathbf{H}_l = \mathcal{N}_l(\mathbf{H}_{l-1}, \mathbf{f}_l^x, \mathbf{f}_l^y), \tag{3}$$

where $\mathbf{H}_l \in \mathbb{R}^{H_l W_l \times 2K}$ is the nearest neighbor field for $\mathbf{f}_l^x$. Specifically, for a feature point $\mathbf{f}_l^x(\mathbf{p})$, $\mathbf{H}_l(\mathbf{p})$ specifies the locations of its top K nearest neighbors in $\mathbf{f}_l^y$. We have

$$\mathbf{H}_l(\mathbf{p}, 1) = \arg\min_{\mathbf{q}} d(\mathbf{f}_l^x(\mathbf{p}), \mathbf{f}_l^y(\mathbf{q})), \tag{4}$$

as an example. Yet it takes a lot of time to traverse $\mathbf{p}$ and $\mathbf{q}$ exhaustively, especially on the entire full-resolution feature map. Therefore, we propose the GRU-assisted PatchMatch, which attempts an iterative improvement.

GRU-assisted PatchMatch. Essentially, our algorithm can be briefly viewed as performing *propagation* and *GRU-based refinement* iteratively and recurrently until convergence or a fixed number of iterations is reached. The previous level results $\mathbf{H}_{l-1}$ are utilized as the initialization, and are improved gradually by alternating the two steps. We illustrate this matching process in Figure 3.

We denote the correspondence map in the tth step as $\mathbf{H}_{l,t}$, and the initialization correspondence field $\mathbf{H}_{l,0}$ is upsampled from $\mathbf{H}_{l-1}$. The level annotation l is omitted in this subsection without causing confusion. The first step, propagation, stems from the seminal work PatchMatch [3]. It improves the matching of the current patch by examining

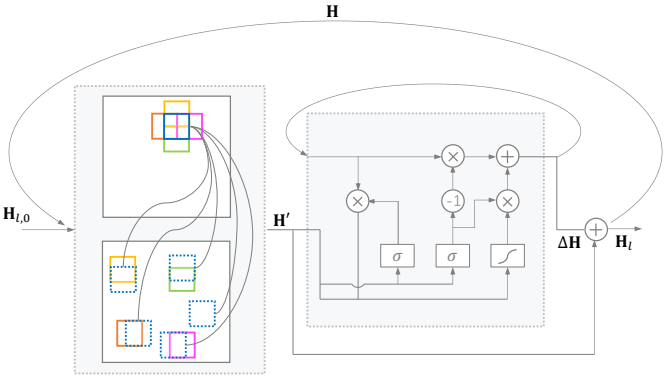

Figure 3: GRU-assisted PatchMatch consisting of (a) propagation and (b) GRU-based refinement. Note that the propagation for all the locations are conducted in parallel.

the already known matching results of its neighborhoods, which we denote as,

$$\mathbf{H}'_t = propagation(\mathbf{H}_t, \mathbf{f}^x, \mathbf{f}^y). \tag{5}$$

where $\mathbf{H}'_t$ are the nearest neighbor field (NNF) propagation results. However, propagation only checks spatially adjacent patches, which makes it heavily relying on the spatial smoothness assumption and tends to be trapped in a local optimum. The random search step in PatchMatch does alleviate this issue to some degree, but it is not enough especially when searching in an extremely large candidate set. Our solution is to look up distant candidates selectively rather than randomly searching, which is guided through a novelly designed refinement module. We expect that, given current offsets, the operator outputs a refinement field that serves as a correction to some incorrectly matched pairs.

Specifically in the second step, we adopt a convolutional gated recurrent unit (ConvGRU),

$$\begin{aligned} z_t &= \sigma(\text{Conv}([h_{t-1}, x_t], \theta_z)) \\ r_t &= \sigma(\text{Conv}([h_{t-1}, x_t], \theta_r)) \\ \hat{h}_t &= \tanh(\text{Conv}([r_t \odot h_{t-1}, x_t], \theta_h)) \\ h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t \end{aligned} \tag{6}$$

where x_t is the input obtained by concatenating features extracted from four variables: $\mathbf{f}^x$, $\mathbf{f}^y$, $\mathbf{O}_t$, $\mathbf{S}_t$. $\mathbf{O}_t$ and $\mathbf{S}_t$ are the current offset and the corresponding matching score,

$$\begin{aligned} \mathbf{O}_t(\mathbf{p}, k) &= \mathbf{H}'_t(\mathbf{p}, k) - \mathbf{p}, \\ \mathbf{S}_t(\mathbf{p}, k) &= \cos(\mathbf{f}^x(\mathbf{p}), \mathbf{f}^y(\mathbf{H}'_t(\mathbf{p}, k)), \end{aligned} \tag{7}$$

where $k = 1, 2, \cdots, K$ considering K nearest neighbors. The initial hidden state is set as $\mathbf{0}$ and the offset update $\Delta\mathbf{H}_t$ is predicted by feeding the output hidden state h_t to two convolutional layers. At last, the offsets are updated by: $\mathbf{H}_{t+1} = \mathbf{H}'_t + \Delta\mathbf{H}_t$ and are passed to the next step.

The benefits of ConvGRU. First, it helps refine the current correspondence estimate making use of a larger context, rather than the local neighborhood. The correspondence can therefore become globally coherent with a faster convergence. Second, the GRU memorizes the history of correspondence estimate, and somehow forecasts the possible corresponding location in the next iteration. Third, the backward gradient can now flow to the pixels in a larger context, rather than at a specific location, which benefits the feature learning and in turn the correspondence.

Differentiable warping function. Unlike conventional applications that directly push the learned correspondences towards ground truth, we do not have the offset ground truth in image-to-image translation. Instead, we leverage the correspondence field in the subsequent translation network to generate high-quality outputs, which pushes the correspondence field to be accurate.

We take the correspondence field to warp the exemplar image y_B and use the warped image $w_l^{y\to x}$ to guide the translation network. Usually, $w_l^{y\to x}$ is obtained by using only the nearest match, *i.e.*, $w_l^{y\to x}(\mathbf{p}) = y_B(\mathbf{H}_l(\mathbf{p}, 1))$. However, the $\arg\min$ operation in Equation 4 is not differentiable. Therefore, we propose to use the following soft warping which is the average of top K possible warping:

$$w_l^{y\to x}(\mathbf{p}) = \sum_{k=1}^{K} softmax(\mathbf{S}_l(\mathbf{p}, k)) y_B(\mathbf{H}_l(\mathbf{p}, k)), \tag{8}$$

where $\mathbf{S}$ is the matching score defined in Equation 7, indicating the semantic similarity.

3.3. Translation network

The translation network $\mathcal{G}$ aims to synthesize an image $\hat{x}_B$ that is desired to respect the spatial semantic structure in x_A while resembling the appearance of similar parts in y_B. Similar to recent conditional generators [37, 54, 34], we employ a simple and natural way that takes a constant code z as input. To preserve the semantic information of the warped exemplar images $w_1^{y\to x}, \cdots, w_L^{y\to x}$, we resort to spatially-adaptive denormalization (SPADE) [39] that learns the modulation parameters adaptively.

Specifically, let the activation before the i^{th} normalization layer be $T^i \in \mathbb{R}^{C_i \times H_i \times W_i}$. we first concatenate the warped images in the channel dimension (upsampling is performed here when necessary). The resulting concatenation is denoted as $\hat{w}^{y\to x} = [w_1^{y\to x}\uparrow, \cdots, w_L^{y\to x}]$ where $\uparrow$ indicates upsampling. Thereafter we project $\hat{w}^{y\to x}$ through two convolutional layers to produce the modulation parameters $\alpha^i_{h,w}$ and $\beta^i_{h,w}$ for style modulation,

$$\alpha^i_{h,w}(\hat{w}^{y\to x}) \times \frac{T^i_{c,h,w} - \mu^i_{h,w}}{\sigma^i_{h,w}} + \beta^i_{h,w}(\hat{w}^{y\to x}), \tag{9}$$

where $\mu^i_{h,w}$ and $\sigma^i_{h,w}$ are calculated mean and standard deviation. Finally, the translation result can be obtained by,

$$\hat{x}_B = \mathcal{G}(z, \hat{w}^{y\to x}; \theta_{\mathcal{G}}), \tag{10}$$

where $\theta_{\mathcal{G}}$ denotes the network parameters.

3.4. Loss functions

Our approach is end-to-end differentiable and can be optimized through backpropagation to simultaneously learn the cross-domain correspondence and the desired output. Generally, it is easy to access the semantically aligned data pair $\{x_A, x_B\}$ in different domains, but does not necessarily have the access to the training triplets $\{x_A, y_B, x_B\}$ where x_B shares a similar appearance with y_B while resembling the semantics of x_A. Hence we construct the *pseudo exemplar* $\tilde{y}_B = \mathcal{T}(x_B)$ from x_B by applying geometric distortion, where $\mathcal{T}$ denotes the geometric augmentation.

Domain alignment loss. For successful correspondence, the multi-level representation for x_A and its corresponding counterpart x_B must lie in the same space, therefore we enforce,

$$\mathcal{L}_{align} = \|\mathcal{M}_A(x_A; \theta_{\mathcal{M}_A}) - \mathcal{M}_B(x_B; \theta_{\mathcal{M}_B})\|_1. \quad (11)$$

Correspondence loss. Still, with the pseudo pairs, the warping $w^{\tilde{y}\to x}$ should exactly be x_B. Thus we enforce the correspondence with,

$$\mathcal{L}_{corr} = \sum_l \|w_l^{\tilde{y}_B \to x_A} - x_B\downarrow\|_1, \quad (12)$$

where $\downarrow$ indicates down-sampling to match the size of x_B to the warped image.

Mapping loss. We expect that the cross-domain inputs can be mapped from the latent representation to their corresponding counterparts in the target domain, which helps the semantics-preserving in the latent space,

$$\mathcal{L}_{map} = \|\mathcal{R}(\mathcal{M}_A(x_A; \theta_{\mathcal{M}_A})) - x_B\|_1 \quad (13)$$
$$+ \|\mathcal{R}(\mathcal{M}_B(y_B; \theta_{\mathcal{M}_B})) - y_B\|_1, \quad (14)$$

where $\mathcal{R}$ maps the features to images in the target domain.

Translation loss. The translated output is desired to be semantically similar to the input with the appearance close to that of the exemplar. We propose two losses focusing on the two objectives respectively. One is the perceptual loss to minimize the semantic discrepancy against x_B:

$$\mathcal{L}_{sem} = \|\phi_m(\hat{x}_B) - \phi_m(x_B)\|_1, \quad (15)$$

where we adopt features ϕ_m from high-level layers of pretrained VGG network. The other one is the appearance loss that comprises of a contextual loss (CX) [33] when applying an arbitrary exemplar y_B and a feature matching loss when using a pseudo exemplar $\tilde{y}_B$. The appearance loss encourages the appearance resemblance by leveraging low-level features ϕ_m of VGG. Concretely, the appearance loss is,

$$\mathcal{L}_{app} = \sum_m u_m[-\log(CX(\phi_m(\hat{x}_B), \phi_m(y_B)))$$
$$+ \sum_m \eta_m \|\phi_m(\hat{x}_B) - \phi_m(\tilde{y}_B)\|_1, \quad (16)$$

where u_m controls the relative importance of different VGG layers and η_m is the balancing coefficient.

Adversarial loss. We add a discriminator to distinguish outputs from the real images in the target domain, competing with the generator which tries to synthesize images that are indistinguishable. The adversarial loss is,

$$\mathcal{L}_{adv}^{\mathcal{D}} = -\mathbb{E}[h(\mathcal{D}(y_B))] - \mathbb{E}[h(-\mathcal{D}(\mathcal{G}(x_A, y_B)))], \quad (17)$$
$$\mathcal{L}_{adv}^{\mathcal{G}} = -\mathbb{E}[\mathcal{D}(\mathcal{G}(x_A, y_B))], \quad (18)$$

where $h(t) = \min(0, -1 + t)$ is the hinge loss [54, 5] to regularize the discriminator.

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

$$\mathcal{L} = \min_{\mathcal{M},\mathcal{N},\mathcal{G},\mathcal{R}} \max_{\mathcal{D}} \lambda_1 \mathcal{L}_{align} + \lambda_2 \mathcal{L}_{corr} + \lambda_3 \mathcal{L}_{map}$$
$$+ \lambda_4(\mathcal{L}_{sem} + \mathcal{L}_{app}) + \lambda_5(\mathcal{L}_{adv}^{\mathcal{D}} + \mathcal{L}_{adv}^{\mathcal{G}}), \quad (19)$$

where λ denotes the weighting parameters, $\mathcal{M}$ contains $\mathcal{M}_A$ and $\mathcal{M}_B$, and $\mathcal{N}$ includes $\mathcal{N}_1, \cdots, \mathcal{N}_L$.

4. Experiment

Implementation details. We apply spectral normalization [36] to all the layers for the translation network and discriminator. We use the Adam solver [23] with $\beta_1 = 0$ and $\beta_2 = 0.999$. The learning rates for the generator and the discriminator are set as $1e-4$ and $4e-4$ respectively, following the TTUR [15]. For detailed implementation including network architectures, please see our appendix. The experiments are conducted using 8 32GB Tesla V100 GPUs.

Datasets. We conduct experiments on four datasets:

- DeepFashion [29] consists of $52,712$ high-quality fashionable person images. We adopt the high-resolution version, and conduct pose-to-body synthesis at 512×512 resolution. OpenPose [8] is used for pose extraction.
- MetFaces [21] consists of $1,336$ high-quality human face images at 1024×1024 resolution collected from works of art in the Metropolitan Museum. The images in the dataset exhibit a wide variety in artistic style. We use the HED [51] to obtain the background edges and connect the face landmarks for the face region. On this dataset, we learn the translation from edges to faces.
- ADE20K [57] consists of $20,210$ training and $2,000$ validation images. Each image is paired with a 150-class segmentation mask. Because of its large diversity, it is challenging for most existing methods to perform mask-to-scene translation. As most of the images have short

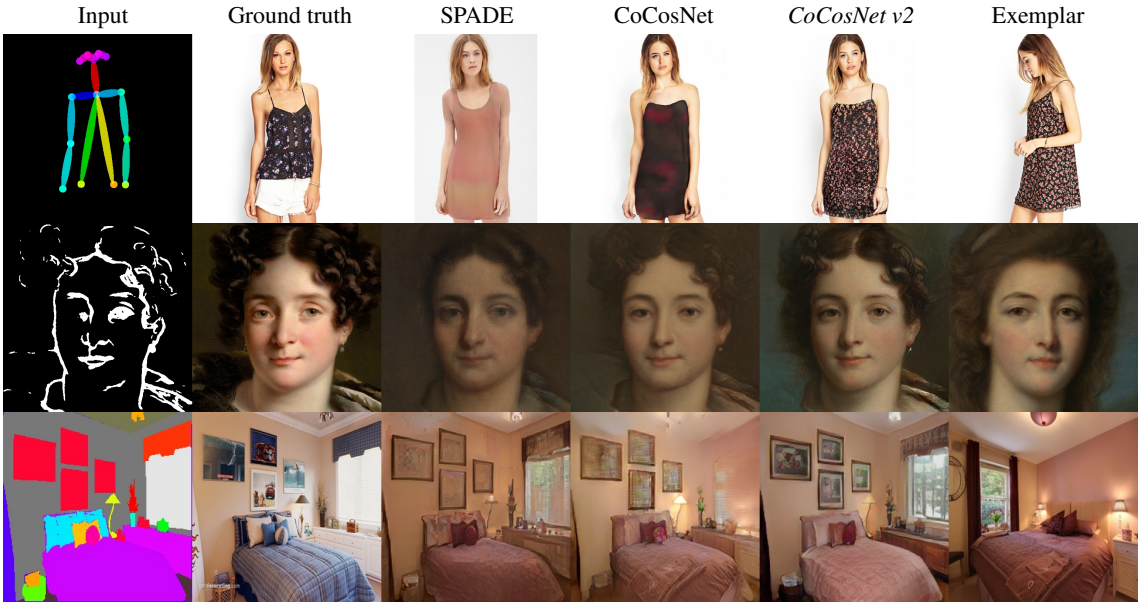

Figure 4: Qualitative comparison on the Deepfashion dataset, the MetFaces dataset, and the ADE20K dataset respectively.

	DeepFashion		MetFaces		ADE20k		ADE20k-*outdoor*	
	FID	SWD	FID	SWD	FID	SWD	FID	SWD
SPADE	34.4	38.0	39.8	30.4	33.9	19.7	63.3	21.9
CocosNet	26.9	29.0	25.6	24.3	26.4	10.5	42.4	11.5
CoCosNet v2	**22.5**	**24.6**	**23.3**	**22.4**	**25.2**	**9.9**	**38.9**	**10.2**

Table 1: Quantitative evaluation of image quality. For both metrics, the lower is better, with the best scores highlighted.

	DeepFashion	MetFaces	ADE20k	ADE20k-*outdoor*
SPADE	0.883	0.915	0.856	0.867
CoCosNet	0.924	0.941	0.862	0.873
CoCosNet v2	**0.959**	**0.963**	**0.877**	**0.895**

Table 2: Quantitative evaluation of semantic consistency. The higher is better with the best scores highlighted.

	DeepFashion		MetFaces		ADE20k	
	Color	Texture	Color	Texture	Color	Texture
SPADE	0.932	0.893	0.949	0.920	0.874	0.892
CoCosNet	0.975	0.944	0.956	0.932	0.962	0.941
CoCosNet v2	**0.987**	**0.961**	**0.972**	**0.956**	**0.970**	**0.948**

Table 3: Quantitative evaluation of style relevance. The higher is better with the best scores highlighted.

side $<$512, we synthesize images at resolution 256$\times$256 on this dataset.

- ADE20K-outdoor is the subset of ADE20K. We follow the same protocol in SIMS [40].

4.1. Comparison with the state-of-the-Art

There are many excellent works that have been proposed for general image translation. We do not compare with those methods that directly learn the translation through networks and fail to utilize the style of exemplars, such as Pix2pixHD [47] and SIMS [40]. We compare with two strong baselines. One is the SPADE [39], a leading approach among the methods [32, 17, 18] that leverage the exemplar style in a global way. We also compare our method with the closest competitor CoCosNet [56] that also leverages cross-domain correspondence but learns at low-resolution. The two works are initially proposed for generating images at resolution 256 $\times$ 256. For a fair comparison, we retrain their models on Deepfashion and MetFaces at resolution 512 $\times$ 512 and make appropriate modifications in order to generate high-quality translation results.

Quantitative evaluation. We first present quantitative evaluation from three directions following [56]. (1) Image quality is evaluated with two widely adopted metrics. One is Fréchet Inception Distance score (FID) [15] that aims to calculate the distance between Gaussian fitted feature distributions of real and generated images. The other one

	L1↓	PSNR↑	SSIM↑
64	82.25	28.03	0.75
64+128	79.56	28.09	0.76
64+128+256	79.10	29.50	0.79
Full 64+128+256+512	**77.84**	**30.03**	**0.82**

Table 4: Ablation study of full-resolution correspondence.

	L1↓	PSNR↑	SSIM↑
Only PatchMatch propagation	108.75	20.40	0.67
Only ConvGRU	94.21	22.99	0.74
PatchMatch propagation + conv	87.83	23.54	0.76
PatchMatch propagation + ConvGRU (*ours*)	**81.97**	**28.99**	**0.83**

Table 5: Ablation study of GRU-assisted refinement.

Figure 5: More results at resolution of 512×512 by CoCosNet v2. For each group, 1st row: exemplars; 2nd row: our results.

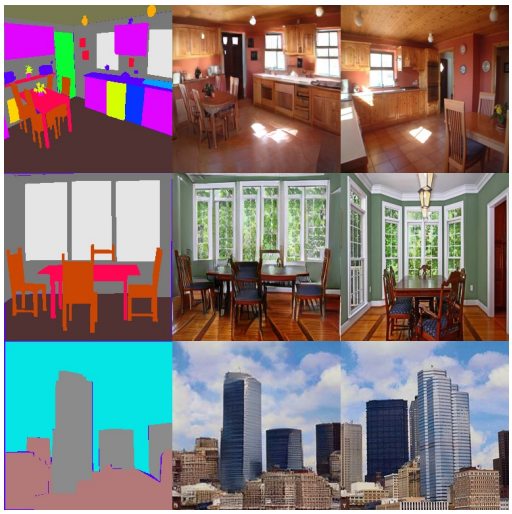

Figure 6: Our results on the ADE20k dataset. Left to right: input, our results, the exemplar.

is sliced Wasserstein distance (SWD) [20] that attempts to measure the Wasserstein distance between the distributions of real images and synthesized ones. Both metrics have been shown that a lower score indicates higher quality images; (2) Semantic consistency is evaluated between the output and the input by calculating the cosine similarity between high-level features representing semantics, *i.e.*, $relu3\_2$, $relu4\_2$ and $relu5\_2$ of an ImageNet pretrained VGG model [5]; (3) Style relevance is evaluated between the output and the exemplar with low-level features, $relu1\_2$, and $relu2\_2$ that mostly encode the color and texture information. The comparison results are shown in Table 1, Table 2, and Table 3 respectively. We can see that CoCosNet v2 significantly outperforms prior competitive methods in the three aspects, suggesting that CoCosNet v2 synthesizes images of higher quality, better preserved semantics and more relevant style.

Qualitative comparison. We show qualitative comparison with the competitors in Figure 4. It can be clearly seen that CoCosNet v2 produces the most visually appealing results and the least visible artifacts. We find that the distinctive patterns in the exemplar have been remarkably well preserved in the semantically corresponding regions of the output, *e.g.*, the texture patterns of the dress in pose-to-body translation, which has been washed out in SPADE and CoCosNet. On the other hand, our output depicts subtle details that are of particular importance to a high-resolution image, demonstrating the advantage of CoCosNet v2. Figure 5-6 shows more diverse results under different exemplars. We also demonstrate 1024×1024 results in Figure 1.

4.2. Ablation study

Full-resolution correspondence. We validate the effectiveness of full-resolution correspondence, which benefits CoCosNet v2 in producing fine textures in the ultimate output. We explore the translation results when correspondence is established at certain level of limited resolution.

Figure 7: Comparison of warped images at different resolution levels. From left to right: edge, warped images at 64^2, 128^2, 256^2, 512^2, output, exemplar. The warped image at 512^2 exhibits more details.

Figure 8: Comparison of warped images for different variants of GRU-assisted refinement. From left to right: exemplar, pose, warped images for using only PatchMatch propagation, only ConvGRU, PatchMatch propagation with convolution, CoCosNet v2 using PatchMatch propagation with convGRU, and ground truth. CoCosNet v2 produces the most faithful warping image.

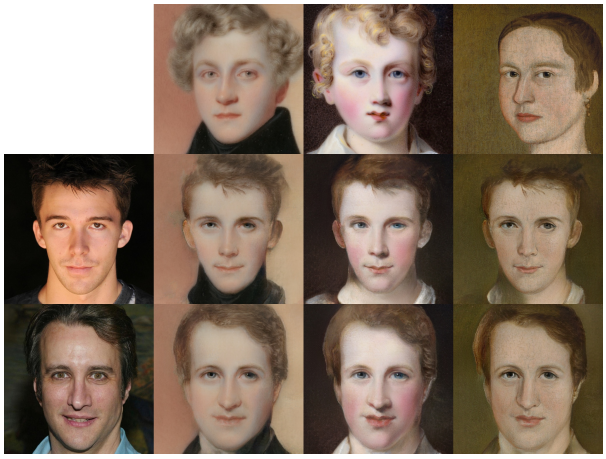

Figure 9: Oil portrait. Given a portrait, CoCosNet v2 can transfer it to a customized oil painting with style from a given exemplar.

Specifically, we vary the largest resolution, *i.e*. the dimension of the latent space, from 64^2 to 512^2 and see how the performance changes. We evaluate the warping on Deepfashion dataset as we consider the person image under a different pose as the exemplar as well as the ground truth. Hence, we can measure the warping with L1, PSNR and SSIM [48]. The numerical results in Table 4 show that hierarchical PatchMatch offers a more accurate correspondence in high-resolution. The qualitative study in Figure 7 shows that full-resolution correspondence captures more details, which further benefits the high-quality synthesis.

GRU-assisted refinement. We present a comprehensive analysis to justify the important component in our architecture, *i.e*. GRU-assisted refinement. We study three variants that are different in each iteration: 1) using only PatchMatch propagation; 2) using only ConvGRU refinement; 3) using PatchMatch propagation assisted with convolution. The comparison with our full model (PatchMatch propagation assisted with ConvGRU) are presented in Table 5 numerically and Figure 8 visually. We can see that only adopting PatchMatch propagation or ConvGRU produces inferior results. We conjecture that the reason may be 1) only PatchMatch propagation cannot backward the gradient to the correctly matched patches, and hence get trapped in the local optimum; 2) only ConvGRU does not consider neighborhood coherence and thus fails to preserve local textures. Moreover, we find that CoCosNet v2 is better than the third variant, which demonstrates that ConvGRU plays an important role in utilizing the history information.

4.3. Application of oil portrait

We present an intriguing application of oil portrait that transfers a portrait to a custom oil painting with different styles specified by the exemplar. This is achieved by extracting the edges from real faces, *e.g*., images from CelebA [30], and applying the model trained from MetFaces. We show several examples in Figure 9.

5. Conclusion

We propose to learn the semantic correspondence in full-resolution. To achieve that, we introduce an effective algorithm CoCosNet v2 that efficiently establishes the correspondence through iterative refinement in a coarse-to-fine hierarchy. At each level, the propagation and GRU-based propagation are alternatively performed. CoCosNet v2 leads to photo-realistic outputs with fine textures as well as visually appealing images at large resolutions, 512^2 and 1024^2.

Acknowledgments. This work is supported in part by Science and Technology Innovation 2030 – "New Generation Artificial Intelligence" Major Project No.(2018AAA0100904), NSFC (No. U19B2043), Artificial Intelligence Research Foundation of Baidu Inc., the funding from HIKVision and Horizon Robotics.

References

[1] Aayush Bansal, Yaser Sheikh, and Deva Ramanan. Shapes and context: In-the-wild image synthesis & manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2317–2326, 2019. 2

[2] Linchao Bao, Qingxiong Yang, and Hailin Jin. Fast edge-preserving patchmatch for large displacement optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3534–3541, 2014. 2

[3] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009. 2, 3

[4] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *Bmvc*, volume 11, pages 1–11, 2011. 2

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1, 5, 7

[6] Thomas Brox and Jitendra Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE transactions on pattern analysis and machine intelligence*, 33(3):500–513, 2010. 2

[7] Shaofan Cai, Xiaoshuai Zhang, Haoqiang Fan, Haibin Huang, Jiangyu Liu, Jiaming Liu, Jiaying Liu, Jue Wang, and Jian Sun. Disentangled image matting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8819–8828, 2019. 2

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 5

[9] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017. 1

[10] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 1

[11] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4384–4393, 2019. 2

[12] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016. 1

[13] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1033–1038. IEEE, 1999. 2

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017. 5, 6

[16] Yinlin Hu, Rui Song, and Yunsong Li. Efficient coarse-to-fine patchmatch for large displacement optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5704–5712, 2016. 2

[17] 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. 1, 6

[18] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 2, 6

[19] 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

[20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 7

[21] 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, 2020. 5

[22] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017. 2

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

[24] Chuan Li and Michael Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2479–2486, 2016. 2

[25] Yu Li, Dongbo Min, Michael S Brown, Minh N Do, and Jiangbo Lu. Spm-bp: Sped-up patchmatch belief propagation for continuous mrfs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4006–4014, 2015. 2

[26] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *arXiv preprint arXiv:1705.01088*, 2017. 2

[27] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):978–994, 2010. 2

[28] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017. 1, 2

[29] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 5

[30] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 8

[31] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981. 2

[32] Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, and Luc Van Gool. Exemplar guided unsupervised image-to-image translation with semantic consistency. *arXiv preprint arXiv:1805.11145*, 2018. 2, 6

[33] 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

[34] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? *arXiv preprint arXiv:1801.04406*, 2018. 4

[35] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1

[36] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 5

[37] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018. 4

[38] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018. 1

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

[40] Xiaojuan Qi, Qifeng Chen, Jiaya Jia, and Vladlen Koltun. Semi-parametric image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8808–8816, 2018. 2, 6

[41] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3937–3946, 2019. 2

[42] Morgane Riviere, Olivier Teytaud, Jérémy Rapin, Yann LeCun, and Camille Couprie. Inspirational adversarial image generation. *arXiv preprint arXiv:1906.11661*, 2019. 2

[43] Amélie Royer, Konstantinos Bousmalis, Stephan Gouws, Fred Bertsch, Inbar Mosseri, Forrester Cole, and Kevin Murphy. Xgan: Unsupervised image-to-image translation for many-to-many mappings. In *Domain Adaptation for Visual Understanding*, pages 33–49. Springer, 2020. 2

[44] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. *arXiv preprint arXiv:2003.12039*, 2020. 2

[45] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2747–2757, 2020. 1

[46] Miao Wang, Guo-Ye Yang, Ruilong Li, Run-Ze Liang, Song-Hai Zhang, Peter M Hall, and Shi-Min Hu. Example-guided style-consistent image synthesis from semantic labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1495–1504, 2019. 2

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

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 8

[49] Li-Yi Wei and Marc Levoy. Fast texture synthesis using tree-structured vector quantization. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 479–488, 2000. 2

[50] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. Deepflow: Large displacement optical flow with deep matching. In *Proceedings of the IEEE international conference on computer vision*, pages 1385–1392, 2013. 2

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

[52] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017. 2

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

[54] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363. PMLR, 2019. 1, 4, 5

[55] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 1

[56] 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, 6

[57] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 5

[58] Xingran Zhou, Siyu Huang, Bin Li, Yingming Li, Jiachen Li, and Zhongfei Zhang. Text guided person image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3663–3672, 2019. 1

[59] 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. 2