Diverse Semantic Image Synthesis via Probability Distribution Modeling

Zhentao Tan[1], Menglei Chai[2], Dongdong Chen[3], Jing Liao[4], Qi Chu[1]*,
Bin Liu[1], Gang Hua[5], Nenghai Yu[1]*
[1]University of Science and Technology of China [2]Snap Inc. [3]Microsoft Cloud AI
[4]City University of Hong Kong [5]Wormpex AI Research LLC
{tzt@mail., qchu@, flowice@, ynh@}ustc.edu.cn
mchai@snap.com, {cddlyf,ganghua}@gmail.com, jingliao@cityu.edu.hk

Abstract

Semantic image synthesis, translating semantic layouts to photo-realistic images, is a one-to-many mapping problem. Though impressive progress has been recently made, diverse semantic synthesis that can efficiently produce semantic-level multimodal results, still remains a challenge. In this paper, we propose a novel diverse semantic image synthesis framework from the perspective of semantic class distributions, which naturally supports diverse generation at semantic or even instance level. We achieve this by modeling class-level conditional modulation parameters as continuous probability distributions instead of discrete values, and sampling per-instance modulation parameters through instance-adaptive stochastic sampling that is consistent across the network. Moreover, we propose prior noise remapping, through linear perturbation parameters encoded from paired references, to facilitate supervised training and exemplar-based instance style control at test time. Extensive experiments on multiple datasets show that our method can achieve superior diversity and comparable quality compared to state-of-the-art methods. Code will be available at https://github.com/tzt101/INADE.git

1. Introduction

Image synthesis has recently seen impressive progress, particularly with the help of generative adversarial networks (GANs) [8]. Besides stochastic approaches that generate high-quality images from random latent variables [18, 19], conditional image synthesis is attracting equal or even more attention due to the practical advantages of its controllability. The conditional input, to guide the synthesis, can be of various forms, including RGB images, edge/gradient maps, semantic labels, etc. In this work, semantic image synthesis is one particular task that aims to generate a photo-realistic image from a semantic label mask. In particular, we further explore its diversity and controllability without loss of generation quality. Some samples are shown in Figure 1.

Previous works [15, 41] propose solutions within the general image-to-image translation framework, which directly feeds the semantic mask into the encoder-decoder network. For higher quality, some recent methods [30, 50, 36] adopt spatially-varying conditional normalization to avoid the loss of semantic information due to conventional normalization layers [40]. Although proven successful in synthesizing certain types of content, these methods lack controllability over the generation diversity, which is particularly important for such a one-to-many problem. Some methods [49, 43] attempt to yield multimodal results by incorporating variational auto-encoder (VAE) or introducing noises. However, these methods only support global image-level diversity. To obtain finer-grained controllability, a recent work [51] proposes to use group convolution for different semantics to achieve semantic-level diversity. However, it is computationally expensive and difficult to be extended to support diversity at the instance level.

In this paper, we attempt to achieve controllable diversity in semantic image synthesis from the perspective of semantic probability distributions. The intuition is to treat each semantic class as one distribution, so that each instance of this class could be drawn from this distribution as a discrete sample. Following this idea, we propose a novel semantic image synthesis framework, which is naturally capable of producing diverse results at semantic or even instance level.

Specifically, our method contains three key ingredients. Firstly, we propose *variational modulation models* (§ 3.2) that extend discrete modulation parameters to class-wise continuous probability distributions, which embed diverse styles of each semantic category in a class-adaptive manner. Secondly, based on the variational models built per normalization layer, we further develop an *instance-adaptive sampling* method (§ 3.3) that achieves instance-level diversity

*Corresponding author.

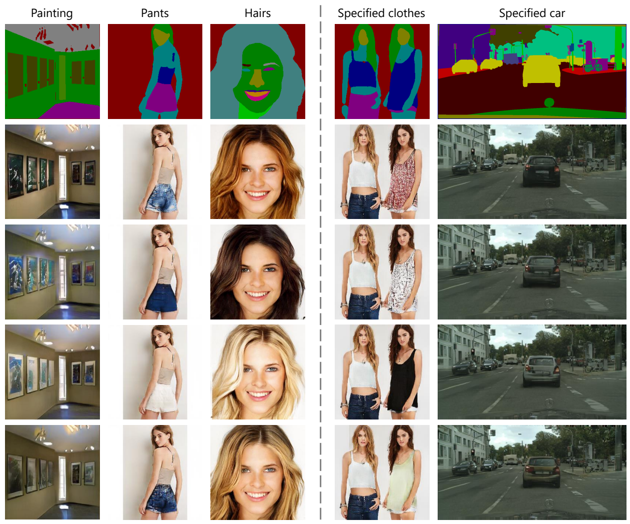

Figure 1. Semantic-level (left three columns) and Instance-level (right two columns) multimodal images generated by the proposed method. Text of each column indicates which semantic class or instance will be changed in the following results.

by stochastically sampling modulation parameters from the variational models. We harmonize the sampling across the network via consistent randomness and a learnable transformation function for each normalization layer. Finally, to more efficiently embed the instance diversity to the modulation models, we propose *prior noise remapping* (§ 3.4) that transforms the noise samples with perturbation parameters encoded from arbitrary references. We adopt this step to facilitate supervised training and enable test-time reference-based style guidance. Inspired by [30, 38, 37], the proposed method is called INADE (**IN**stance-**A**daptive **DE**normalization).

To evaluate the proposed method, we conduct extensive experiments on multiple datasets, including Cityscapes [3], ADE20K [47], CelebAMask-HQ [23, 17, 29], and DeepFashion [28]. Both quantitative and qualitative results show that our method significantly outperforms state-of-the-art methods by achieving much better instance-level diversity while keeping comparable generation quality.

2. Related Work

2.1. Conditional Image Synthesis

Conditional image synthesis aims at generating photo-realistic images conditioned on different types of input. We are interested in a special form of it, called semantic image synthesis, which takes segmentation layouts as input. Many impressive works have been proposed for this task. The most representative work, Pix2Pix [15] adopts an encoder-decoder generator for unified image-to-image translation. Pix2pixHD [41] improves Pix2Pix by proposing coarse-to-fine generator and discriminators. Subsequent methods [32, 27, 39, 45, 37, 50] further explore how to synthesize high quality images from semantic masks and achieve significant improvements. Besides using class-level semantic masks, some works also consider instance-level information for image synthesis, since the semantic mask itself does not provide sufficient information to synthesize instances especially in complex environments with multiple of them interacting with each other. Some works [41, 30, 37] extract boundary information from the instance map and concatenate it with the semantic mask. While recent work [6] proposes to use the instance map to guide convolution and upsampling layers for better exploiting both semantic and instance information. Different from these methods, we are interested in taking full advantage of information from instance maps to achieve instance-level diversity control.

2.2. Diversity in Image Synthesis

Diversity is a core target for image synthesis, which aims to generate multiple possible outputs from a single input image. Early conditional image synthesis networks either trained with paired data, like Pix2Pix [15] and Pix2pixHD [41], or with unpaired data, like CycleGAN [48], DiscoGAN [20] and UNIT [26], are single-modal. Later, some multimodal unpaired image synthesis networks [13, 24, 1] are proposed. However, constrained by the reconstruction loss, the semantic image synthesis task trained with paired data is more difficult to support diversity. To tackle this problem, BicycleGAN [49] enforces the bijection mapping between the noise vector and target domain, and DSCGAN [43] proposes a simple regularization method. More recently, a variational autoencoder architecture is used to handle multimodal synthesis by [30, 37, 27]. However, these multimodal image synthesis networks only support diversity at the global level. To further control the diversity at the semantic level, the method proposed by [9] builds several auto-encoders for each face component to extract different component representations. GroupDNet [51] unifies the generation process in only one model, but still requires high computing resources, and the use of group convolution layer makes it difficult to extend to the instance level. In contrast, we propose a novel instance-aware conditional normalization framework that allows diverse instance-level generation with less overhead.

3. Method

We are interested in the task of semantic image synthesis, which is defined as to map a semantic mask $\boldsymbol{m} \in \mathbb{L}_m^{H\times W}$ to a photo-realistic image $\boldsymbol{o} \in \mathbb{R}^{3\times H\times W}$. Here, $\boldsymbol{m}$ is a class-level label map with each pixel representing an integer index to a pre-defined set of semantic categories $\mathbb{L}_m = \{1, 2, \ldots, L^m\}$. Each pair of input $\boldsymbol{m}$ and output $\boldsymbol{o}$ is spatially-aligned and of the same dimension $H \times W$,

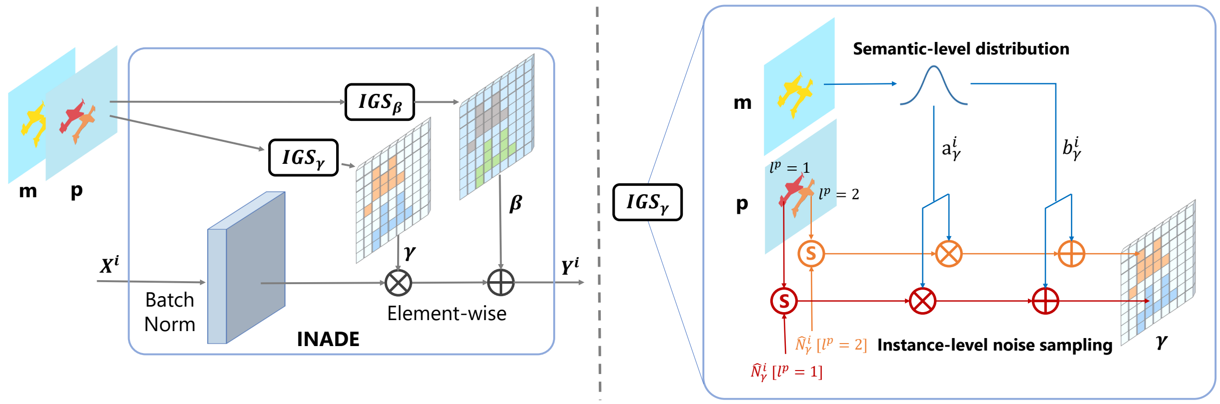

Figure 2. The illustration diagram of the proposed **IN**stance-**A**daptive **DE**normalization (INADE). It combines semantic-level distribution modeling and instance-level noise sampling. IGS denotes the **I**nstance **G**uided **S**ampling which is similar to the guided sampling in [37].

so that the synthesized content in $\boldsymbol{o}$ should comply with the corresponding semantic labels in $\boldsymbol{m}$.

In addition to this basic formulation [41, 30, 37], instance-aware semantic image synthesis [6] adopts the instance map $\boldsymbol{p} \in \mathbb{L}_p^{H \times W}$ as an extra input, which differentiates different object instances sharing a same semantic label by denoting each individual instance with a unique index from the instance label set $\mathbb{L}_p = \{1, 2, \ldots, L^p\}$ in the image. By enforcing an identical semantic label within each instance, pixels belonging to a same instance label l^p in $\boldsymbol{p}$ should always have a same semantic label l^m in $\boldsymbol{m}$. We represent instance to semantic label mapping as a function $l^m = \mathcal{G}(l^p)$.

Overall, image synthesis with instance information can be basically formulated as a function $\mathcal{T}(\boldsymbol{m}, \boldsymbol{p}) : (\mathbb{L}_m, \mathbb{L}_p) \to \mathbb{R}^3$. And feed-forward image translation neural networks, trained in a supervised manner, can be used to model this function. In the following, we introduce the proposed method with both the inputs of semantic and instance maps. When there is no instance label, $\boldsymbol{p}$ degenerates into $\boldsymbol{m}$. And the diversity of the synthesized images changes from the instance level to the semantic level.

3.1. Conditional Normalization

Our solution to semantic image synthesis is based on a novel instance-level conditional normalization method. Before introducing our method, here we give a brief overview of the general framework of conditional normalization first.

Let $\boldsymbol{X}^i \in \mathbb{R}^{C^i \times H^i \times W^i}$ be the activation tensor to the i-th normalization layer, where C^i, H^i, W^i are the channel depth, height, and width, respectively. In the channel-wise normalization framework similar to [14], we can generally formulate the normalization operations as two steps: In the normalization step, $\boldsymbol{X}^i$ is normalized to $\hat{\boldsymbol{X}}^i$ by channel-wise mean and standard deviation $\{\boldsymbol{\mu}^i, \boldsymbol{\sigma}^i\} \in \mathbb{R}^{C^i}$ in the mini-batch containing $\boldsymbol{X}^i$. Then, the modulation step scales and translates $\hat{\boldsymbol{X}}^i$ with learned modulation parameters $\{\boldsymbol{\gamma}^i, \boldsymbol{\beta}^i\} \in \mathbb{R}^{C^i \times H^i \times W^i}$, which are not necessarily channel-wise constant. Let $\boldsymbol{Y}^i$ be the output, for each element $(k \in C^i, x \in H^i, y \in W^i)$ in the tensor, we have:

$$\begin{aligned} \hat{\boldsymbol{X}}^i_{k,x,y} &= (\boldsymbol{X}^i_{k,x,y} - \boldsymbol{\mu}^i_k)/\boldsymbol{\sigma}^i_k, \\ \boldsymbol{Y}^i_{k,x,y} &= \boldsymbol{\gamma}^i_{k,x,y} \hat{\boldsymbol{X}}^i_{k,x,y} + \boldsymbol{\beta}^i_{k,x,y}. \end{aligned} \tag{1}$$

For conditional normalization [5, 12], the modulation parameters $\boldsymbol{\gamma}^i$ and $\boldsymbol{\beta}^i$ are learned with extra conditions. Specifically, for semantic image synthesis, the modulation is usually conditioned on the semantic mask $\boldsymbol{m}$ [30, 37].

3.2. Variational Modulation Model

Conditional normalization (§ 3.1), either spatially-adaptive [30] or class-adaptive [37], has been proven helpful for semantic image synthesis. The semantic-conditioned modulation is able to largely prevent the "wash-away" effect of semantic information caused by repetitive normalizations. However, challenges still exist to achieve promising generation results with semantic-level or even instance-level diversity, given that normalization is solely conditioned on the semantic map and only global randomness is used to diversify the image styles [30]. Semantic-level diversity is realized by [51] through group convolution, but using this convolution cuts off the possibility of its extension to instance-level diversity through instance map. Recent efforts on instance-aware synthesis [41, 6] are majorly focused on better object boundaries, but not the diversity and realism of each individual instance. Due to the lack of proper instance conditioning, existing methods tend to converge instances with the same semantic label into a similar style, which significantly harms the diversity of generation.

The key to instance-level diversity is a proper combination of uniform semantic-level distributions that deterministically decide the general features of a particular semantic label, and instance-level randomness that introduces allowed diversity covered by the semantic distribution models. Therefore, we model the modulation parameters as

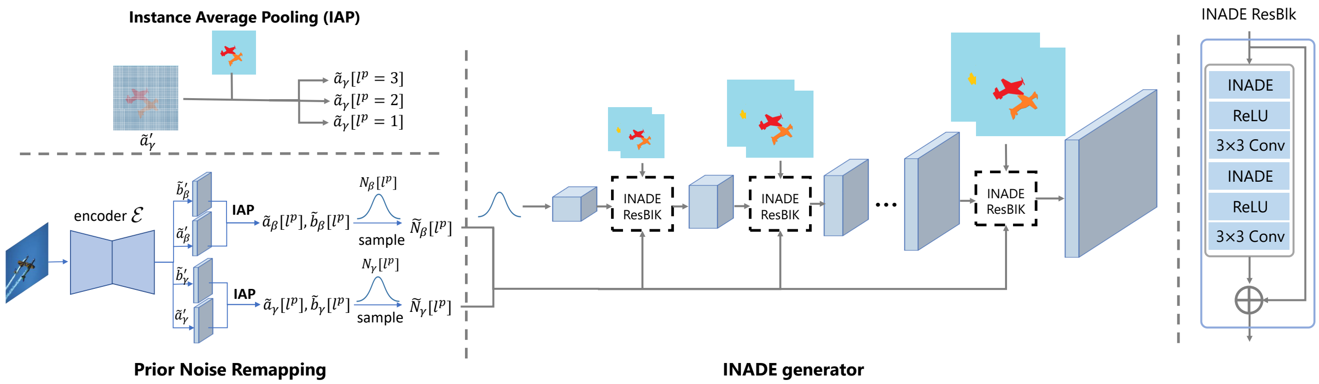

Figure 3. The overall framework of the proposed INADE generator, which consists of a remapping encoder $\mathcal{E}$ and INADE generator. $\mathcal{E}$ is used to transform the noise sample based on arbitrary references (§ 3.4), while the generator consists of several INADE ResBlks.

parametric probability distributions for each semantic label $l^m \in \mathbb{L}^m$, instead of discrete values. With such a, namely, variational modulation model, given an instance $l^p \in \mathbb{L}^p$, instance-level diversity is achievable via sampling modulation parameters from the probability distributions of the corresponding semantic label $\mathcal{G}(l^p)$. For the sake of simplicity and efficiency, following [37], we make the modulation parameters spatially-invariant and only depend on the local instance labels.

Specifically, for each semantic category l^m, its channel-wise modulation parameters are modeled as learnable probability distributions, which are built for each normalization layer in the network respectively. Formally, for the i-th layer with channel depth C^i, we have $\{\boldsymbol{a}^i_\gamma, \boldsymbol{b}^i_\gamma, \boldsymbol{a}^i_\beta, \boldsymbol{b}^i_\beta\} \in \mathbb{R}^{L^m \times C^i}$ as the distribution transformation parameters of γ and β, respectively. All of them are treated as learnable parameters that are jointly trained with the network. Given stochastic noise matrices $\{\boldsymbol{N}^i_\gamma, \boldsymbol{N}^i_\beta\} \in \mathbb{R}^{L^p \times C^i}$ from the same distribution for sampling, the corresponding modulation parameters of one instance label l^p in $\boldsymbol{p}$ are:

$$\begin{aligned} \boldsymbol{\gamma}^i[l^p] &= \boldsymbol{a}^i_\gamma[\mathcal{G}(l^p)] \otimes \boldsymbol{N}^i_\gamma[l^p] + \boldsymbol{b}^i_\gamma[\mathcal{G}(l^p)], \\ \boldsymbol{\beta}^i[l^p] &= \boldsymbol{a}^i_\beta[\mathcal{G}(l^p)] \otimes \boldsymbol{N}^i_\beta[l^p] + \boldsymbol{b}^i_\beta[\mathcal{G}(l^p)], \end{aligned} \tag{2}$$

where $\otimes$ represents element-wise multiplication, and $[\cdot]$ accesses the vector from a matrix in the row-major order.

3.3. Instance-Adaptive Modulation Sampling

Our multimodal synthesis method follows the basic form of conditional modulation (§ 3.1), but further extends the conditional inputs to include not just the segmentation mask $\boldsymbol{m}$, but also the instance map $\boldsymbol{p}$ and random noises to initiate sampling, as shown in Figure 2. Utilizing our variational modulation models (§ 3.2), we are able to generate diverse modulation parameters obeying the same set of probability distributions. However, considering that the generation network contains multiple conditional normalization layers, a unified sampling solution is still essential to harmonize all these layers. A straight-forward approach, independent stochastic sampling for each normalization layer, could potentially introduce inconsistency and cause the diversity to be severely neutralized. Therefore, in this paper, we propose an instance-adaptive modulation sampling method that achieves consistent instance sampling across multiple normalization layers with unequal channel depths.

To initialize, for each layer i, we resize and convert each input pair of semantic mask $\boldsymbol{m}$ and instance map $\boldsymbol{p}$ into the one-hot format as $\boldsymbol{M}^i \in \mathbb{B}^{L^m \times H^i \times W^i}$ and $\boldsymbol{P}^i \in \mathbb{B}^{L^p \times H^i \times W^i}$, respectively, which will then be used as the conditional inputs to that layer, as shown in Figure 3. Here, $\mathbb{B}$ represents the Boolean domain, and L^m, L^p are the aforementioned total numbers of semantic/instance labels.

For the sake of simplicity, since scale $\boldsymbol{\gamma}^i$ and shift $\boldsymbol{\beta}^i$ are generated similarly and independently, without loss of generality, here we take scale $\boldsymbol{\gamma}^i$ as the example. The sampling contains the following steps.

First of all, random samples $\boldsymbol{N}_\gamma \in \mathbb{R}^{L^p \times C^0}$ are independently sampled from the standard normal distribution: $\boldsymbol{N}_\gamma \sim \mathcal{N}(0, 1)$. We use the same set of random noise samples $\boldsymbol{N}_\gamma$ for all instances of normalization layers in the network, which helps enforce consistent instance styles throughout the network. Here C^0 is a hyper-parameter that defines the number of the initial sampling channels.

To sample the modulation parameters for each normalization layer i, we translate the initial samples $\boldsymbol{N}_\gamma$ to $\hat{\boldsymbol{N}}^i_\gamma$ with a learnable linear transformation mapping $\mathcal{F}^i_\gamma$: $\mathbb{R}^{L^p \times C^0} \to \mathbb{R}^{L^p \times C^i}$:

$$\hat{\boldsymbol{N}}^i_\gamma = \mathcal{F}^i_\gamma(\boldsymbol{N}_\gamma), \tag{3}$$

where C^i is exactly the channel depth of i-th activations $\boldsymbol{X}^i$, so that the output $\hat{\boldsymbol{N}}^i_\gamma \in \mathbb{R}^{L^p \times C^i}$ assigns a transformed sample for each instance per each channel, in a spatially-invariant manner. Thus, the same source of randomness helps achieve style consistency, while the learnable transformations enforce compatible target dimensions and preserve certain ability to adapt the samples for each layer.

Finally, given the distribution transformation parameters $\boldsymbol{a}_\gamma^i, \boldsymbol{b}_\gamma^i$ and transformed noise samples $\hat{\boldsymbol{N}}_\gamma^i$, the scale parameters $\boldsymbol{\gamma}^i$ are calculated with Equation 2. And similarly for the shift parameters $\boldsymbol{\beta}^i$.

3.4. Prior Noise Remapping

While our variational modulation models (§ 3.2) help achieve instance-level diversity, the noises, sampled regardless of instance styles (§ 3.3), can potentially introduce ambiguities during the supervised training (especially for the popular perceptual and feature matching losses), since similar noise samples can possibly correspond to instances of distinct styles. This will affect the effective diversity of generated instances and prohibit the possibility to control the instance styles with certain references.

In light of this, we propose a prior noise remapping step, during which a set of linear perturbation parameters are encoded from given references, to remap the noise samples while preserving the original distribution, in order to provide guidance to embed more meaningful instance diversity in the modulation models. To achieve this, we adopt a noise remapping encoder $\mathcal{E}(\boldsymbol{r})$ that translates the reference image $\boldsymbol{r} \in \mathbb{R}^{H\times W}$ into four perturbation maps $\{\tilde{\boldsymbol{a}}'_\gamma, \tilde{\boldsymbol{b}}'_\gamma, \tilde{\boldsymbol{a}}'_\beta, \tilde{\boldsymbol{b}}'_\beta\} \in \mathbb{R}^{H\times W}$, which are per-pixel linear transformation parameters, including scale $\tilde{\boldsymbol{a}}$ and shift $\tilde{\boldsymbol{b}}$, for both $\boldsymbol{N}_\gamma$ and $\boldsymbol{N}_\beta$. A instance aware partial convolution [10, 25] is used to avoid information contamination between different instances. Based on these dense perturbation maps, we apply an instance average pooling layer to each of these maps to get the instance-wise perturbation parameters $\{\tilde{\boldsymbol{a}}_\gamma, \tilde{\boldsymbol{b}}_\gamma, \tilde{\boldsymbol{a}}_\beta, \tilde{\boldsymbol{b}}_\beta\} \in \mathbb{R}^{L^p}$. Take $\boldsymbol{N}_\gamma$ as an example, for an instance label $l^p \in \mathbb{L}^p$ and its occupying pixels $\boldsymbol{x}(l^p) = \{x|\boldsymbol{p}[x] = l^p\}$, we have

$$\begin{aligned} \tilde{\boldsymbol{a}}_\gamma[l^p] &= (\textstyle\sum_{x\in\boldsymbol{x}(l^p)}\tilde{\boldsymbol{a}}'_\gamma[x])/|\boldsymbol{x}(l^p)|, \\ \tilde{\boldsymbol{b}}_\gamma[l^p] &= (\textstyle\sum_{x\in\boldsymbol{x}(l^p)}\tilde{\boldsymbol{b}}'_\gamma[x])/|\boldsymbol{x}(l^p)|. \end{aligned} \tag{4}$$

This remapping encoder, together with the main generator, forms a variational autoencoder (VAE) [22]. The remapped noise samples after perturbation are:

$$\tilde{\boldsymbol{N}}_\gamma[l^p] = \tilde{\boldsymbol{a}}_\gamma[l^p]\boldsymbol{N}_\gamma[l^p] + \tilde{\boldsymbol{b}}_\gamma[l^p], \tag{5}$$

where KL-Divergence loss [22] is used to enforce a same normal distribution $\tilde{\boldsymbol{N}}_\gamma \sim \mathcal{N}(0,1)$. These remapped noise samples are used instead of $\boldsymbol{N}_\gamma$ during modulation sampling, as described in § 3.3.

During training, the reference $\boldsymbol{r}$ is exactly the groundtruth paired image. At test time, while initially sampled noises can be used to achieve random style synthesis by default, as described in § 3.3, it is also allowed to provide $\boldsymbol{r}$ as instance references to control the style of the result at the instance level.

4. Experiments

4.1. Implementation Details

Our INADE generator (Figure 3) follows a similar architecture of the SPADE generator [30], but with all the SPADE layers replaced by the INADE layers. Following SPADE [30], the overall loss function consists of four loss terms: conditional adversarial loss, feature matching loss [41], perceptual loss [16] and KL-Divergence loss [22]. Details are provided in the supplementary material.

During training, by default, Adam optimizer [21] ($\beta_1 = 0$, $\beta_2 = 0.9$) is used with fixed epoch number of 200. The learning rates for the generator and the discriminator are set to 0.0001 and 0.0004 respectively, which are gradually decreased to zero after 100 epochs. Noise C^0 has 64 initialized channels, while the input noise has 256 channels, same as [30, 27]. All experiments are implemented in PyTorch [31] and conducted on TITAN XP GPUs.

4.2. Datasets and Metrics

Datasets. Experiments are conducted on five popular datasets: *Cityscapes* [3], *ADE20K* [47], *CelebAMask-HQ* [23, 17, 29], *DeepFashion* [28], and *DeepFashion2*. *DeepFashion2* is built from *DeepFashion* that each image contains two persons, which is only used for testing. More details can be found in the supplementary material.

Quality Metrics. To evaluate the result quality, we adopt two types of metrics following [2, 41, 30, 37]. One is the Fréchet Inception Distance (FID) [11], which measures the distance of distributions between results and real images.

The other one is semantic-segmentation-based, which evaluates the semantic segmentation accuracy on the results by comparing the predicted masks with the groundtruth layouts on both mean Intersection-over-Union (mIoU) and pixel accuracy (accu). State-of-the-art pretrained segmentation models are used for different datasets: DRN [44, 7] for *Cityscapes*, UperNet101 [42, 4] for *ADE20k*, and UNet [34, 35] for *CelebAMask-HQ*. For *DeepFashion*, we use the same UNet-based network but train the model by ourselves. For fair evaluation, we run the model for 10 times and report the average scores when noise input is required.

Diversity Metrics. To evaluate the diversity of the results, we adopt the LPIPS metric as proposed by [46, 33]. Similar to [51], we also adopt two metrics to measure the semantic-level diversity (mCSD and mOCD) and expand them to instance level (mISD and mOID). More details can be found in the supplementary material.

Subjective Metrics. We conduct human evaluations to assess both the quality and the diversity of the methods. For quality, we ask the volunteers to select the most realistic one among the results generated by different methods on the same input. mHE (mean Human Evaluation) denotes the percentage of results being selected for each method.

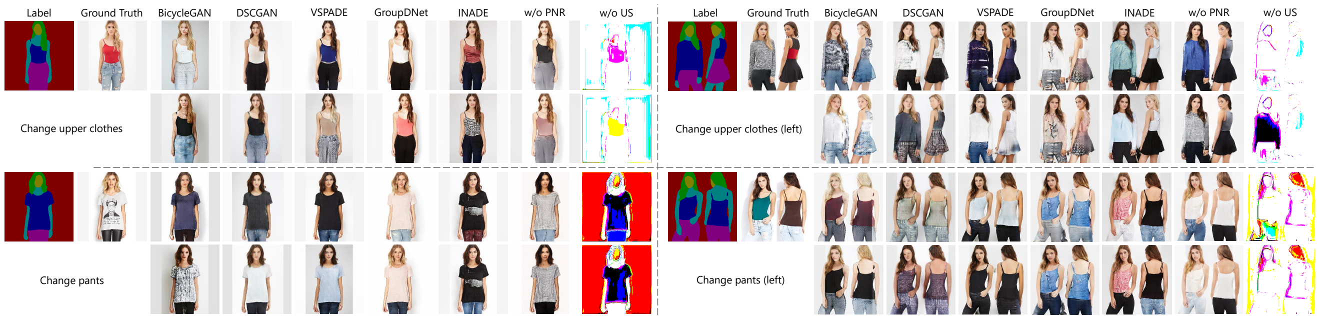

Figure 4. Visual comparison with other multimodal models and two baselines. The results on the left show the performance of class level diversity while the results on the right are for instance level diversity.

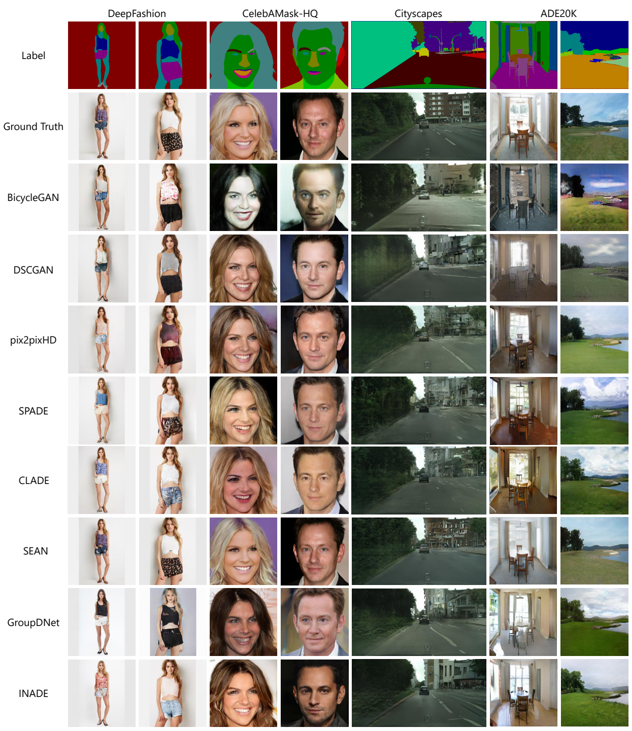

Figure 5. Qualitative comparison with the state-of-the-art semantic image synthesis methods on four datasets: DeepFashion, CelebAMask-HQ, Cityscapes and ADE20K.

For diversity, we expand the metric proposed by [51] to the instance level. A pair of results, with one random semantic class or instance manipulated, are given to volunteers. The percentage of pairs that are judged to be different in only one area represents the human evaluation, namely SHE (Semantic Human Evaluation) and IHE (Instance Human Evaluation). We invite 20 volunteers for evaluation and the evaluated number of images is 20.

4.3. Quantitative and Qualitative Comparisons

We compare INADE with several SOTA works, including quality-oriented (pix2pixHD [41], SPADE [30], CLADE [38, 37] and SEAN [50]) and diversity-oriented (BicycleGAN [49], DSCGAN [43], and GroupDNet [51]) methods. For a fair comparison, we directly use the pre-trained models provided by the authors when available, otherwise train the models by ourselves using the codes and settings provided by the authors. For SPADE, which has the strategy for multimodal synthesis, we train the model with that extra encoder but ignore it when testing its diversity performance (namely VSPADE). Reference images are not allowed for any of the methods during testing except for SEAN which requires the reference input.

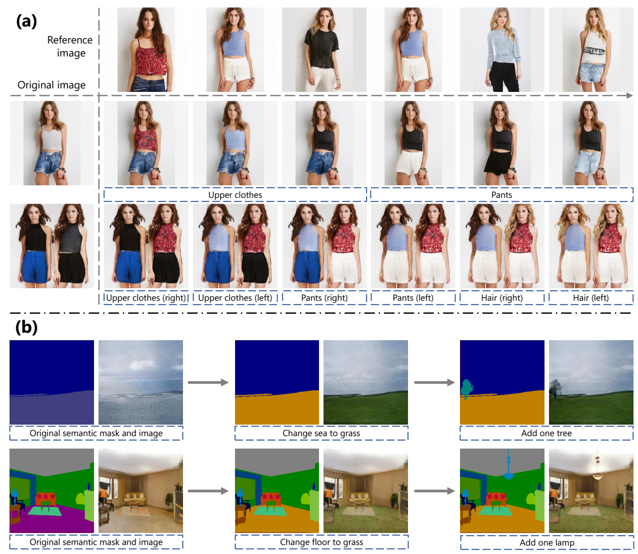

Figure 6. Exemplar applications of the proposed INADE. (a) Results of our method for reference appearance editing. (b) Application of our method for semantic manipulation. The text in the blue dotted box indicates what is changed or edited each time.

4.3.1 Multimodal Image Synthesis

Several methods that support multimodal image synthesis are compared to demonstrate the superior diversity performance of INADE. In additional, we also compare with

Table 1. Comparison with other multimodal methods on diversity. mHE, SHE and IHE are aforementioned metrics. ↑ and ↓ represent the higher the better and the lower the better. **Bold** and <u>underlined</u> numbers are the best and the second best of each metric, respectively.

Methods	DeepFashion						DeepFashion2					
	FID ↓	LPIPS ↑	mCSD ↑	mOCD ↓	mHE ↑	SHE ↑	FID ↓	LPIPS ↑	mISD ↑	mOID ↓	mHE ↑	IHE ↑
BicycleGAN	31.10	**0.225**	<u>0.0465</u>	0.2014	0.0	4.5	33.46	<u>0.286</u>	<u>0.0500</u>	0.2456	0.0	2,5
DSCGAN	29.79	0.146	0.0404	0.1218	0.0	9.3	48.64	0.199	0.0433	0.1633	0.0	4.8
VSPADE	11.11	0.197	0.0450	0.1665	7.5	6.5	<u>22.29</u>	0.222	0.0390	0.1780	<u>43.7</u>	3.3
GroupDNet	**9.72**	0.222	0.0453	**0.0077**	<u>40.0</u>	<u>86.0</u>	22.81	0.281	0.0434	<u>0.0303</u>	8.8	<u>9.3</u>
INADE	<u>9.97</u>	**0.225**	**0.0511**	<u>0.0161</u>	**52.5**	**88.3**	**18.18**	**0.319**	**0.0580**	**0.0187**	**47.5**	**82.8**
w/o PNR	12.09	0.184	0.0289	0.0138	-	-	20.76	0.243	0.0291	0.0189	-	-
w/o US	248.33	0.624	0.0730	0.0370	-	-	265.63	0.633	0.0748	0.0296	-	-

Table 2. Comparison with SOTA methods on result quality. All the numbers are collected by running the evaluation on our machine. Here **M**, **A**, **F**, and **L** represent mIoU, accu, FID, and LPIPS, respectively. Note that the **L** score of SEAN is almost zero even with noise input.

Methods	Cityscapes				ADE20K				CelebAMask-HQ				DeepFashion			
	M ↑	**A** ↑	**F** ↓	**L** ↑	**M** ↑	**A** ↑	**F** ↓	**L** ↑	**M** ↑	**A** ↑	**F** ↓	**L** ↑	**M** ↑	**A** ↑	**F** ↓	**L** ↑
SPADE	**61.38**	<u>93.26</u>	51.98	0	**36.28**	<u>78.13</u>	29.79	0	75.22	94.76	31.40	0	**76.76**	**97.65**	11.22	0
pix2pixHD	60.50	93.06	66.04	0	27.27	72.61	45.87	0	<u>76.11</u>	**95.67**	36.95	0	73.99	97.02	15.27	0
CLADE	60.44	**93.42**	50.62	0	<u>35.43</u>	77.37	30.48	0	75.37	95.05	33.54	0	75.63	97.33	12.76	0
SEAN	56.22	92.28	50.43	0	32.65	76.58	**28.11**	0	75.94	95.03	<u>24.30</u>	0	<u>76.28</u>	97.46	**7.37**	0
BicycleGAN	30.47	78.26	59.87	0.122	5.33	42.68	77.49	**0.443**	65.98	89.77	35.73	<u>0.362</u>	73.09	96.75	31.10	**0.225**
DSCGAN	43.70	87.80	50.84	<u>0.216</u>	8.07	58.10	82.30	0.324	75.98	95.08	52.83	0.198	75.92	96.97	29.79	0.146
GroupDNet	59.20	92.78	<u>41.12</u>	0.073	26.09	73.07	39.11	0.177	**76.13**	<u>95.21</u>	29.39	0.309	76.19	<u>97.48</u>	<u>9.72</u>	0.222
INADE	<u>61.02</u>	93.16	**38.04**	**0.248**	34.96	**78.51**	<u>29.60</u>	<u>0.400</u>	74.08	94.31	**22.55**	**0.365**	76.27	97.44	9.96	**0.225**

two ablation baselines: w/o **US** (without **U**nified **S**ampling, §3.3) and w/o **PNR** (without **P**rior **N**oise **R**emapping, §3.4). w/o **US** means that the noise for each INADE layers are sampled independently, while w/o **PNR** means that the PNR is not used during training. The quantitative results on the *DeepFashion* dataset are summarized in Table 1.

In general, INADE achieves superior performance regarding both quality and diversity compared to previous methods. For single-subject images in the *DeepFashion* dataset, our method exhibits better performance than BicycleGAN, DSCGAN, and VSPADE in terms of FID, and is comparable to GroupDNet. While for multiple-subject images (*DeepFashion2*), INADE shows the lowest FID.

In terms of diversity, our method achieves the best scores on metrics including overall measurement (LPIPS) and semantic-level/instance-level metrics (mCSD/mISD). As for mOCD, methods only support image-level diversity, such as BicycleGAN, DSCAN, and VSPADE, produce much more unwanted changes outside the instance area. Although both being relatively low, INADE is higher than GroupDNet, which is because GroupDNet uses group convolution to prevent premature fusion between features of different classes, while INADE uses conventional convolution for more consistent combinations. For mOID, as the only method that supports instance-wise control, INADE easily gets the best score. More analysis of mOID and mOCD can be found in the supplementary material.

As for subjective evaluations, our method also outperforms others on both semantic- and instance-level cases.

Compared to the two ablation baselines, we find that both prior noise remapping and unified sampling play indispensable roles in our method. Removing prior noise remapping (w/o **PNR**) leads to ambiguities during the supervised training, which seriously affects the quality and diversity of synthesized results. Independently sampling (w/o **US**) for each normalization destroys the consistency of information and significantly degenerates the generation result.

The qualitative comparisons are shown in Figure 4. Although all methods support multimodal synthesis, the quality by BicycleGAN and DSCGAN is not satisfactory. VSPADE achieves good visual quality, but does not support semantic- or instance-level control. GroupDNet is capable of changing the appearance of a specific semantic class (results on the left), but tends to generate identical style for different cloth instances (results on the right). On the contrary, INADE supports both fine-grained multimodal controls with high visual fidelity. As for the ablation baselines, we notice that removing **PNR** significantly decreases the quality, while the whole task fails without **US**.

4.3.2 Semantic Image Synthesis

The quantitative comparisons against semantic image synthesis methods are summarized in Table 2.

Compared to methods that don't support multimodal (i.e. 0 LPIPS), especially SEAN which has additional reference image input, INADE has an advantage or near the best on almost all metrics. It seems that SPADE has slight advantage in segmentation metrics, but INADE still shows its overall superiority when considering FID score and visual results.

Table 3. Comparison with other semantic image synthesis methods on model complexity and efficiency. All the numbers are collected by running the evaluation on Titan XP. Here $\mathbf{P}$ and $\mathbf{T}$ denote the number of generator parameters and inference run time, respectively.

Methods	Cityscapes			ADE20K			CelebAMask-HQ			DeepFashion		
	$\mathbf{P}$ (M)	FLOPs (G)	$\mathbf{T}$ (s)	$\mathbf{P}$ (M)	FLOPs (G)	$\mathbf{T}$ (s)	$\mathbf{P}$ (M)	FLOPs (G)	$\mathbf{T}$ (s)	$\mathbf{P}$ (M)	FLOPs (G)	$\mathbf{T}$ (s)
SPADE	93.05	281.54	0.065	96.50	181.30	0.042	92.54	141.32	0.035	92.21	137.99	0.032
pix2pixHD	182.53	151.32	0.038	182.90	99.30	0.041	182.47	72.17	0.023	182.44	69.91	0.020
CLADE	67.90	75.54	0.035	71.40	42.20	0.024	67.32	42.15	0.022	66.98	42.15	0.019
SEAN	330.41	681.75	0.507	-	-	-	266.90	346.27	0.165	223.23	342.89	0.135
BicycleGAN	54.80	18.40	0.011	54.80	18.40	0.006	54.80	18.40	0.006	54.80	18.40	0.006
DSCGAN	54.00	18.14	0.018	54.00	18.14	0.010	54.00	18.14	0.010	54.00	18.14	0.010
GroupDNet	76.50	463.61	0.224	68.33	383.00	0.088	145.29	225.53	0.090	96.32	291.61	0.062
INADE	77.39	75.25	0.048	90.89	42.19	0.035	85.12	42.18	0.030	84.63	42.92	0.026

Compared to existing multimodal methods, INADE leads all metrics. In terms of quality (e.g. mIoU, acc, and FID), BicycleGAN and DSCGAN are much lower than ours on all datasets. GroupDNet achieves similar or slightly better performance on *CelebAMask-HQ* and *DeepFashion* datasets, but has a significant gap on more complicated scenes such as *Cityscapes* and *ADE20K*. This demonstrates the superiority of the proposed method on synthesizing complex scenes. The LPIPS score shows that all these methods are able to generate multimodal images to some extent. BicycleGAN gets higher scores than ours on some datasets, but is not able to do high-quality synthesis. GroupDNet shows good performance on person-related tasks, but falls into strong bias when dealing with complex scenes which greatly restricts its performance. Therefore, considering both quality and diversity, our method achieves the best overall performance.

Qualitative comparisons on these four datasets are shown in Figure 5. In general, the images generated by INADE are more realistic than others on various datasets, which is consistent with the quantitative results.

4.3.3 Computational and Model Complexity

In this section, we analyze the computational and model overhead of different methods. The quantitative results (generator networks only) are summarized in Table 3.

BicycleGAN and DSCGAN share a similar small network architecture with the least parameters, FLOPs (floating-point operations per second), and run-time cost. However, the quality of the synthesized images is far from satisfactory. CLADE seems to get a good trade-off between performance and efficiency, but still falls short of INADE in terms of overall performance and functionality. Compared to all other methods, INADE achieves the smallest network (parameters and FLOPs), as well as one of the fastest run-time performance. Specifically, compared with GroupDNet, the only method that can achieve semantic-level diversity, our method provides control over both semantic and instance levels with much less overhead, introducing 82% $\sim$ 89% fewer FLOPs and 59% $\sim$ 79% less inference time compared with GroupDNet.

4.4. Applications

Thanks to its superior capability of controllable diverse synthesis, INADE can be used in many image editing applications. Here we show two examples as follows.

Reference appearance editing. With the noise remapping mechanism described in § 3.4, we can extract the instance-wise style from an arbitrary reference image. This makes it possible for INADE to perform reference-based editing to different parts of an image at the instance level. As shown in Figure 6 (a), we can change the appearance of hairstyles, tops, and pants to match the reference.

Semantic manipulation. Similar to most existing semantic image synthesis methods, INADE also supports semantic manipulation. We show some examples in Figure 6 (b), such as changing the semantic class to an object, or insert a new semantic object into the image. And more creative editing results can be achieved by modifying the semantic mask and the instance map.

5. Conclusion

In this paper, we focus on multimodal image synthesis and propose a novel diverse semantic image synthesis method based on instance-aware conditional normalization. Different from previous works, we learn the class-wise probability distributions and perform instance-wise stochastic sampling to generate the per-instance modulation parameters. Our method improves the network's ability to model semantic categories and make it easier to synthesize diverse images at semantic- or instance-level without scarifying the visual fidelity.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (No. U20B2047). Jing Liao is partially supported by an ECS grant from the Research Grants Council of the Hong Kong (Project No. CityU 21209119) and an APRC grant from CityU, Hong Kong (Project No. 9610488). Gang Hua is partially supported by National Key R&D Program of China Grant 2018AAA0101400 and NSFC Grant 61629301.

References

[1] Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *arXiv preprint arXiv:1802.10151*, 2018. 2

[2] 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. 5

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 5

[4] CSAILVision. Pytorch implementation for semantic segmentation/scene parsing on mit ade20k dataset. https://github.com/CSAILVision/semantic-segmentation-pytorch.git, 2019. 5

[5] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016. 3

[6] Aysegul Dundar, Karan Sapra, Guilin Liu, Andrew Tao, and Bryan Catanzaro. Panoptic-based image synthesis. *arXiv preprint arXiv:2004.10289*, 2020. 2, 3

[7] fyu. Dilated residual networks. https://github.com/fyu/drn.git, 2019. 5

[8] 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

[9] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. Mask-guided portrait editing with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3436–3445, 2019. 2

[10] Adam W Harley, Konstantinos G Derpanis, and Iasonas Kokkinos. Segmentation-aware convolutional networks using local attention masks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5038–5047, 2017. 5

[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. In *Advances in neural information processing systems*, pages 6626–6637, 2017. 5

[12] 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. 3

[13] 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

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015. 3

[15] 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

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5

[17] 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. 2, 5

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pages 4401–4410, 2019. 1

[19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 1

[20] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1857–1865. JMLR. org, 2017. 2

[21] D Kingma and J Ba. Adam: A method for stochastic optimization in: Proceedings of international conference on learning representations. 2015. 5

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 5

[23] 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. 2, 5

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

[25] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018. 5

[26] 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. 2

[27] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, et al. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *Advances in Neural Information Processing Systems*, pages 568–578, 2019. 2, 5

[28] 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. 2, 5

[29] 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. 2, 5

[30] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 1, 2, 3, 5, 6

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019. 5

[32] 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

[33] richzhang. Perceptualsimilarity. https://github.com/richzhang/PerceptualSimilarity.git, 2020. 5

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5

[35] switchablenorms. Celebamask-hq. https://github.com/switchablenorms/CelebAMask-HQ.git, 2020. 5

[36] Zhentao Tan, Menglei Chai, Dongdong Chen, Jing Liao, Qi Chu, Lu Yuan, Sergey Tulyakov, and Nenghai Yu. Michigan: multi-input-conditioned hair image generation for portrait editing. *ACM Transactions on Graphics (TOG)*, 39(4):95–1, 2020. 1

[37] Zhentao Tan, Dongdong Chen, Qi Chu, Menglei Chai, Jing Liao, Mingming He, Lu Yuan, Gang Hua, and Nenghai Yu. Semantic image synthesis via efficient class-adaptive normalization. *arXiv preprint arXiv:2012.04644*, 2020. 2, 3, 4, 5, 6

[38] Zhentao Tan, Dongdong Chen, Qi Chu, Menglei Chai, Jing Liao, Mingming He, Lu Yuan, and Nenghai Yu. Rethinking spatially-adaptive normalization. *arXiv preprint arXiv:2004.02867*, 2020. 2, 6

[39] Hao Tang, Song Bai, and Nicu Sebe. Dual attention gans for semantic image synthesis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1994–2002, 2020. 2

[40] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 1

[41] 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, 3, 5, 6

[42] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018. 5

[43] Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. Diversity-sensitive conditional generative adversarial networks. In *International Conference on Learning Representations*, 2018. 1, 2, 6

[44] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017. 5

[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. 2

[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] 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. 2, 5

[48] 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

[49] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*, pages 465–476, 2017. 1, 2, 6

[50] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020. 1, 2, 6

[51] Zhen Zhu, Zhiliang Xu, Ansheng You, and Xiang Bai. Semantically multi-modal image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5467–5476, 2020. 1, 2, 3, 5, 6