MISC: Multi-condition Injection and Spatially-adaptive Compositing for Conditional Person Image Synthesis

Shuchen Weng[#1] Wenbo Li[#2] Dawei Li[2] Hongxia Jin[2] Boxin Shi[*1]
[1]NELVT, Dept. of CS, Peking University [2]Samsung Research America AI Center
{shuchenweng,shiboxin}@pku.edu.cn, {wenbo.li1,dawei.l,hongxia.jin}@samsung.com

Abstract

In this paper, we explore synthesizing person images with multiple conditions for various backgrounds. To this end, we propose a framework named "MISC" for conditional image generation and image compositing. For conditional image generation, we improve the existing condition injection mechanisms by leveraging the inter-condition correlations. For the image compositing, we theoretically prove the weaknesses of the cutting-edge methods, and make it more robust by removing the spatially-invariance constraint, and enabling the bounding mechanism and the spatial adaptability. We show the effectiveness of our method on the Video Instance-level Parsing dataset, and demonstrate the robustness through controllability tests.

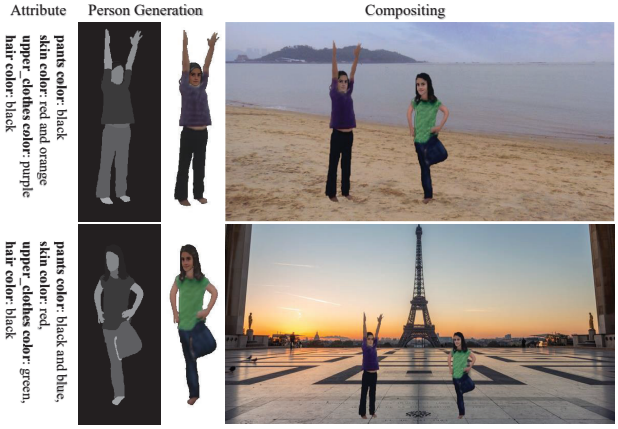


Figure 1. Illustration of the conditional person image synthesis problem, which includes a generation and a compositing phase.

1. Introduction

Conditional person image synthesis is vital in many application scenarios, *e.g.*, augmented reality and data creation. This problem can be formulated as two phases (Figure 1): *(i)* conditional *generation* phase: on the basis of a geometry condition, generating the fine-grained textures for this person following the pattern and color conditions as precisely as possible; *(ii)* adaptive *compositing* phase: adjusting the color tone of the generated person adaptively towards different backgrounds, so that the generated textures not only look realistic by themselves, but also remain realistic after being composited with the background.

For conditional image generation, recent works [10, 12, 14, 20] suggest that a proper injection of conditions into the generation pipeline is crucial to the generation quality. However, existing methods are designed based on either the assumption of a unique condition or the independence among multiple conditions, and few of them attempt to improve the condition injection by leveraging the inter-condition correlations. Specific to the conditional person generation phase, it takes three conditions as input, *i.e.*, geometry, pattern (*viz.*, gray-scale textures) and color, which are instantiated herein as the parsing mask, Gaussian noise, and multi-valued color attributes, respectively. The geometry condition is visually concrete, while the pattern and color are abstract and are *correlated* with the geometry. Thus, the conditional person generation can be regarded as a visual concretization of the abstract conditions which is constrained by the innate concrete conditions.

On the other hand, an image compositing model is necessary for harmonizing the generated person foreground and the target background. Cutting-edge works, *e.g.*, [1, 24], perform the compositing by adjusting the global color tone of the foreground with an inferred *spatially-invariant* affine transformation. However, we discover that such a spatially-invariant transformation at the pixel level could lead to the over-saturated effects due to the unbounded value range of the transformed foreground image. Therefore, a more robust compositing model is desired for handling our task.

In this paper, we present MISC (Multi-condition Injection and Spatially-adaptive Compositing), an end-to-end trainable deep neural network to address the above mentioned problems for conditional person image synthesis. MISC includes a conditional person generation model and

[#] Equal contributions. * Corresponding author.

a spatially-adaptive image composition model.

The generation model is a multistage convolutional neural network where two new modules are introduced for injecting the pattern and color conditions respectively based on their inferred or inherent correlations with the determinate geometry condition. Regarding the pattern condition, we format it as Gaussian noise and propose a conditional normalization layer, namely *geometry-guided adaptive instance normalization* (GAIN), to modulate the activations using the Gaussian noise while constraining the steepness of image gradients through a geometry-guided gate without harming the controllability of the condition. For injecting the color condition, we use a *bipartite network projection* method to project the attribute embeddings (encoded by a multi-valued attribute encoder) to their inherently associated spatial locations of a person geometry, and a pretrained cross-modality similarity model to enhance the semantic meaningfulness of the attribute embeddings.

The compositing model is another convolutional neural network which takes the generated foreground person image and a provided background image to infer the per-pixel color transformation parameters for the foreground image. Compared to previous spatially-invariant compositing model, it adjusts the color tone of the generated foreground with high robustness and training stability.

We conduct extensive experiments, and perform detailed ablation study and controllability study on Video Instance-level Parsing dataset [23]. We show that MISC can achieve more convincing person synthesis results than baselines with other related architectures. MISC binds the color condition with the pattern condition for synthesizing the realistic non-uniform textures, which distinguishes MISC from the naive global adjustment for the hue channel of an image.

2. Related Work

There are four main ways of injecting conditions. *(i)* Many methods [9, 18] input the condition directly through the first layer of a feed-forward network, which suffers from the "condition dilution" problem as indicated in [10, 14]. *(ii)* Some works [4, 21] tile the input condition uniformly, and concatenates the tiled condition to the intermediate feature maps within the generation pipeline. Since not all information encoded in the condition is useful to every spatial location, such a uniform fashion causes extra burdens for the generation model on the information selection. *(iii)* An improved module of the uniform injection is the attentive aggregation, *e.g.*, [11, 20], which estimates the usefulness of the conditional information, and attentively aggregates the useful conditional information for each spatial location. The effectiveness of this module depends on the reliability of the usefulness estimation. *(iv)* The conditional normalization is proposed to alleviate the "condition dilution" problem by performing an affine transformation after each normalization operation. The affine parameters, which are inferred through a network from the input condition, are responsible for modulating the activations either element-by-element [14] or channel-by-channel [10]. The element-wise transformation is tailored for the visually concrete condition with spatial dimensions, *e.g.*, parsing mask, while the channel-wise one is much more general and not limited to spatial-explicit condition, and thus should be more suitable for our abstract pattern condition, *i.e.*, Gaussian noise.

There are five main methodologies for adjusting the color tone of images for the compositing purpose. *(i)* Some methods [17, 19] investigate the matching between the low-level handcrafted features of the foreground and those of the background. Their limitation lies in the *generalization* because of the assumption that the contents or color tones of foreground and background are highly correlated. *(ii)* Tsai *et al*. [16] explore the color tone adjustment together with the semantic segmentation as a dual task, which causes the extra *computational burdens*. *(iii)* Cun and Pan [3] assume the *availability* of the ground-truth realistic composited images, so they can apply the reconstruction loss for training. *(iv)* Chen and Kae [1] aim to infer the spatially-invariant color-tone affine parameters using a feed-forward network. *(v)* Some other methods [1, 6, 15] use a lightweight segmentation based adversarial discriminator which plays against the compositing model by identifying the composited foreground area. Considering the above issues, we design our compositing model by improving the fourth and fifth methodologies to be more robust in our task.

3. Conditional Person Image Synthesis

The conditional person image synthesis is formulated as a generation-compositing setting, under which a person image y is formulated as a composition of a foreground y^{f} and the remaining regions as the provided background y^{b}:

$$y = m \odot y^{\mathrm{f}} + (1 - m) \odot y^{\mathrm{b}}. \tag{1}$$

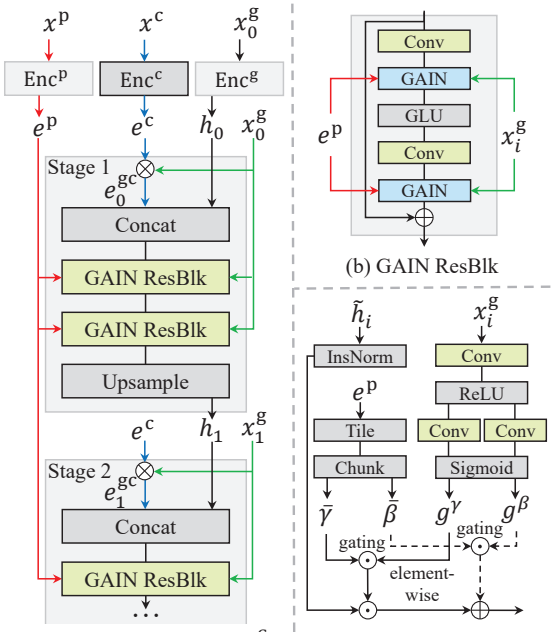
Our generation model F^G aims at mapping a combination of conditions $\mathbf{x}$ to a raw person foreground $\hat{y}^{\mathrm{f}}$ of which the semantic information corresponds to $\mathbf{x}$. The generation phase is formulated as $\hat{y} = F^{\mathrm{G}}(\mathbf{x})$. Following Eq. (1), $\hat{y}$ can be decomposed into $\hat{y}^{\mathrm{f}}$ and $\hat{y}^{\mathrm{b}}$. Our compositing model F^{C} estimates the contrast and brightness parameters, (ρ, τ), based on $\hat{y}^{\mathrm{f}}$ and a desirable background y^{b}:

$$(\rho, \tau) = F^{\mathrm{C}}(\hat{y}^{\mathrm{f}}, y^{\mathrm{b}}). \tag{2}$$

The color tone of $\hat{y}^{\mathrm{f}}$ can thus be adjusted towards y^{b} through an affine transformation with (ρ, τ). Then, we can synthesize a complete person image y^{s} by blending the adjusted y^{f} with y^{b} as in Eq. (1).

3.1. Conditional generation

The input conditions $\mathbf{x}$ to F^{G} include the geometry x^{g}, pattern x^{p} and color x^{c}. The definitions are as follows:



(a) Conditional generator F^{G} (b) GAIN ResBlk (c) GAIN

Figure 2. (a) The multistage conditional generator F^{G} takes as input three conditions, *i.e.*, pattern x^{p}, color x^{c}, and geometry x_i^{g} for the $(i+1)$-th stage. x^{p} and x^{c} are encoded by $\mathrm{Enc}^{\mathrm{p}}$ and $\mathrm{Enc}^{\mathrm{c}}$ to form e^{p} and e^{c} for condition injection. x_0^{g} is encoded by $\mathrm{Enc}^{\mathrm{g}}$ to form the initial feature maps h_0. The red and blue arrows indicate that e^{p} and e^{c} are injected into F^{G} under the guidance (green arrows) of x_i^{g}. e_i^{gc} is the spatially-specific color condition and $e_i^{\mathrm{gc}} = e^{\mathrm{c}} \otimes x_i^{\mathrm{g}}$, where $\otimes$ denotes a bipartite network projection. (b) A ResBlk with GAIN for injecting e^{p}. (c) In GAIN, the injection of e^{p} is performed by a geometry-guided affine transformation.

(i) $x^{\mathrm{g}} \in \mathbb{L}^{N_{\mathrm{g}} \times H \times W}$ is a body part parsing mask, where $\mathbb{L} \in \{0, 1\}$, and N_{g}, H and W represent the number of body parts, image height and width, respectively. Entry $x^{\mathrm{g}}[k, i, j]$ indicates the coverage of the k-th body part for the location (i, j) on the image plain. Let $m \in \mathbb{L}^{H \times W}$ be an overall mask which is formed by a max-pooling along dimension N_{g} of x^{g}. *(ii)* $x^{\mathrm{p}} \sim \mathcal{N}(0, 1)$ is a Gaussian noise vector, which has two unique properties. First, it is free of annotation. Second, latent space of the Gaussian noise is continuous, so it should provide high variety in generating clothing textures and decorative design. *(iii)* $x^{\mathrm{c}} \in \mathbb{L}^{N_{\mathrm{c}} \times N_{\mathrm{v}}}$ denotes the multi-valued attributes, where $\mathbb{L} \in \{0, 1\}$, N_{c} and N_{v} represent the number of attributes, *e.g.*, coat color, and the number of values, *e.g.*, blue, respectively. Entry $x^{\mathrm{c}}[i, j] = 1$ indicates the existence of the j-th value of the i-th attribute. Multiple values are allowed to co-exist for an attribute.

Among the three conditions, x^{g} is visually concrete, while x^{p} and x^{c} are abstract. Therefore, the mapping of these conditions to a person foreground can be regarded as a visual concretization process of the abstract conditions.

As shown in Figure 2 (a), we design our conditional generation model F^{G} to be multistage, in which every stage shares the architecture of the stacking of two residual blocks as shown in Figure 2 (b). Let F_{i+1}^{G} denote the architecture of the $(i+1)$-th stage of F^{G}, then we have $h_{i+1} = F_{i+1}^{\mathrm{G}}(h_i, x_i^{\mathrm{g}}, e^{\mathrm{p}}, e^{\mathrm{c}})$, where h_i represents the intermediate features from the previous stage. x_i^{g} denotes the geometry condition with the corresponding spatial resolution to the i-th stage, $e^{\mathrm{p}} = \mathrm{Enc}^{\mathrm{p}}(x^{\mathrm{p}})$ represents the pattern embedding output by a pattern encoder $\mathrm{Enc}^{\mathrm{p}}$, and $e^{\mathrm{c}} = \mathrm{Enc}^{\mathrm{c}}(x^{\mathrm{c}})$ represents the color attribute embedding output by a color encoder $\mathrm{Enc}^{\mathrm{p}}$. The output of each stage, h_{i+1}, can be fed into a conv-tanh block (omitted in Figure 2) to generate an image y_{i+1}^{f}. The resolution of y_{i+1}^{f} and x_i^{g} increases with i.

The goal of F^{G} is to facilitate the visual conretization of the abstract conditions x^{p} and x^{c}. To achieve this goal, it is necessary to have proper condition injection mechanisms for x^{p} and x^{c}. To this end, we propose novel network architectures and training mechanisms for injecting x^{p} (§3.1.1) and x^{c} (§3.1.2) via leveraging their inferred or inherent correlations with the visually concrete condition x^{g}.

3.1.1 Pattern injection

AdaIN. Our pattern condition x^{p} is defined as a Gaussian noise vector. The adaptive instance normalization (AdaIN) [10, 7, 8] is well-known for injecting the Gaussian noise through denormalizing the normalized activations with the inferred affine parameters in a channel-wise fashion. Given a pattern embedding $e^{\mathrm{p}} = \mathrm{Enc}^{\mathrm{p}}(x^{\mathrm{p}})$, AdaIN extracts the affine parameters by chunking $e^{\mathrm{p}} \in \mathbb{R}^{2K}$ into two halves: the slope $\gamma \in \mathbb{R}^K$ and the bias $\beta \in \mathbb{R}^K$. Let $\tilde{h} \in \mathbb{R}^{K \times \tilde{H} \times \tilde{W}}$ denote the input feature maps to AdaIN, $\tilde{h}[k]$ denote the k-th feature map of $\tilde{h}$, and $\gamma[k]$ and $\beta[k]$ denote their k-th entry, respectively. The AdaIN operation is formulated as: $\mathrm{AdaIN}(\tilde{h}[k], \gamma[k], \beta[k]) = \gamma[k]\frac{\tilde{h}[k] - \mu(\tilde{h}[k])}{\sigma(\tilde{h}[k])} + \beta[k]$, where $\mu(\cdot)$ and $\sigma(\cdot)$ denote functions for computing mean and std, respectively. AdaIN has been proved effective in varying the appearance of the rigid geometry such as human faces [10]. Unlike the rigid geometry, the structure of the person geometry is much more diverse in poses. We observe that AdaIN frequently fails in handling the person geometry by yielding mistaken non-uniform textures (see Figure 3). We argue that such mistaken non-uniform textures derive from the pose diversity which causes the scattering of activations for a body part onto multiple feature maps. Because of the channel-wise modulation fashion, it is hard for AdaIN to unify the textures for some body parts, *e.g.*, arms. At the same time, the non-uniform textures are sometimes desirable to other body parts, *e.g.*, coat. This poses a problem on how to control the texture uniformity under the guidance of the body part parsing mask x^{g}.

SPADE. Denormalizing the normalized activations with the inferred affine parameters in an element-wise fashion can be achieved by using spatially-adaptive denormalization

(SPADE) [14]. SPADE is tailored for the input conditions with spatial dimensions. Taking our geometry condition $x^{\mathrm{g}} \in \mathbb{L}^{N_{\mathrm{g}} \times H \times W}$ as an example input, SPADE projects x^{g} via a shallow conv block to the affine parameters $\gamma, \beta \in \mathbb{R}^{K \times \tilde{H} \times \tilde{W}}$, which can be applied to the normalized activations for the element-wise transformation. Compared to AdaIN, the advantage of SPADE is its spatial adaptability which can potentially help control the texture uniformity of different body parts, while its disadvantage lies in the incapability of producing the non-uniform textures due to the uniformity within each part of the conditional parsing mask.

GAIN. By combining the advantages of AdaIN and SPADE, we propose the Geometry-guided Adaptive Instance Normalization (GAIN) for injecting the pattern condition x^{p}. The intuition behind GAIN is to adaptively control the texture uniformity for different body parts under the guidance of x^{g}. Assume that the activations of two locations are respectively scattered on the k_1-th and k_2-th feature maps of $\tilde{h} \in \mathbb{R}^{K \times \tilde{H} \times \tilde{W}}$, which are controlled by the two corresponding entries of affine parameters in AdaIN, *i.e.*, $\gamma[k_1]$ vs. $\gamma[k_2]$, and $\beta[k_1]$ vs. $\beta[k_2]$. If these two locations belong to the same body part which desires uniform textures, we would better have a similarity constraint: $(\gamma[k_1] \sim \gamma[k_2]) \wedge (\beta[k_1] \sim \beta[k_2])$. Note that this similarity constraint should not affect the activations of body parts of no interest, so the constraint should be applied on each channel in a spatially-adaptive fashion. Thus, we introduce two spatially-adaptive gates $g^{\gamma} \in \mathbb{M}^{K \times \tilde{H} \times \tilde{W}}$ and $g^{\beta} \in \mathbb{M}^{K \times \tilde{H} \times \tilde{W}}$ for $\gamma \in \mathbb{R}^{K}$ and $\beta \in \mathbb{R}^{K}$, where $\mathbb{M} \in [0, 1]$. We tile γ and β are to be $\bar{\gamma} \in \mathbb{R}^{K \times \tilde{H} \times \tilde{W}}$ and $\bar{\beta} \in \mathbb{R}^{K \times \tilde{H} \times \tilde{W}}$. The similarity constraint for location (k_1, i_1, j_1) and (k_2, i_2, j_2) can be written as:
$g^{\gamma}[k_1, i_1, j_1] \cdot \bar{\gamma}[k_1, i_1, j_1] \sim g^{\gamma}[k_2, i_2, j_2] \cdot \bar{\gamma}[k_2, i_2, j_2]$,
$g^{\beta}[k_1, i_1, j_1] \cdot \bar{\beta}[k_1, i_1, j_1] \sim g^{\beta}[k_2, i_2, j_2] \cdot \bar{\beta}[k_2, i_2, j_2]$.
Obviously, g^{γ} and g^{β} need to be zeroed out to satisfy the constraint irrespective of the values of $\bar{\gamma}$ and $\bar{\beta}$. We adopt SPADE to infer g^{γ} and g^{β} conditioned on x^{g}.

Remark 3.1. *With g^{γ} and g^{β}, the visually concrete x^{g} guides the injection of the abstract x^{p}. The values of g^{γ} and g^{β}, in turn, reflect the correlations between x^{p} and x^{g}.*

We show the architecture and operation of GAIN in Figure 2 (c). The GAIN operation is formulated as: $\mathrm{GAIN}(\tilde{h}, \bar{\gamma}, \bar{\beta}, g^{\gamma}, g^{\beta}) = (g^{\gamma} \odot \bar{\gamma}) \odot \mathrm{IN}(\tilde{h}) \oplus (g^{\beta} \odot \bar{\beta})$, where $\mathrm{IN}(\cdot)$ represents the instance normalization.

3.1.2 Color injection

We use a series of binary attributes, *e.g.*, blue coat, as our color condition. The binary attributes are widely used in the face editing task, *e.g.*, [2, 12, 20]. These methods inject the color condition by concatenating a multi-hot binary attribute vector uniformly onto the image plain, and expect the generation model to automatically select the useful color information for each location. In these works, the supervision for color comes from an auxiliary attribute classifier which is trained together with the image discriminator. Compared to the common facial attributes (fewer than 20 used in [2, 20, 12]), there are significantly more attributes for the whole body (over 100 in this work). Such a distinction highlights the limitations of the conventional color injection mechanism from two aspects. *(i) Learning burdens:* it brings extra burdens for the generation model in identifying useful information for each region. *(ii) Class number curse:* the auxiliary classifier is known for suffering from handling a large number of classes.

To resolve the "learning burdens" issue caused by the uniform injection, we need to prepare and inject a spatially-specific attribute condition $e^{\mathrm{gc}} \in \mathbb{R}^{K \times H \times W}$, which requires two things: *(i)* a dedicated attribute embedding for each body part, *i.e.*, $e^{\mathrm{c}} \in \mathbb{R}^{K \times N_{\mathrm{g}}}$, where K and N_{g} represent the embedding dimension and the number of body parts, respectively; *(ii)* our geometry condition $x^{\mathrm{g}} \in \mathbb{L}^{N_{\mathrm{g}} \times H \times W}$ which indicates the association between body parts and spatial locations. Then, e^{gc} can be obtained through a bipartie network projection $e^{\mathrm{gc}} = e^{\mathrm{c}} \otimes x^{\mathrm{g}}$. As shown in Figure 2 (a), the color injection can be achieved by concatenating e^{gc} to the input feature maps of each stage.

Next, we introduce how to prepare e^{c}, which is a knob of the color injection: *(i)* Organizing the binary color attributes to be the multi-valued attributes $x^{\mathrm{c}} \in \mathbb{L}^{N_{\mathrm{c}} \times N_{\mathrm{v}}}$ (see §3.1 for definition) according to the inherent associations between binary attributes, *e.g.*, blue coat, and body parts, *e.g.*, coat. *(ii)* Using an attribute encoder $\mathrm{Enc}^{\mathrm{c}}$ to encode each multi-valued attribute into a continuous embedding vector, which as a whole are denoted as $\hat{e}^{\mathrm{c}} \in \mathbb{R}^{K \times N_{\mathrm{c}}}$. *(iii)* Specifying the association matrix between the body parts and attributes as $A \in \mathbb{L}^{N_{\mathrm{g}} \times N_{\mathrm{c}}}$ s.t. $\forall\, i, j\ A[i, j] \geq 0, \forall\, i\ \sum_j A[i, j] = 1$, we have $e^{\mathrm{c}} = \hat{e}^{\mathrm{c}} A$.

To guarantee the meaningfulness of the attribute embeddings, we adopt a cross-modality similarity model (CMSM) to pretrain $\mathrm{Enc}^{\mathrm{c}}$, in which an image encoder and $\mathrm{Enc}^{\mathrm{c}}$ are trained to project the image regions of a body part and its associative attribute(s) to a joint embedding.

The CMSM can also be used to resolve the "class number curse" by replacing the auxilary attribute classifier. Specifically, let $\hat{e}^{\mathrm{c}} \in \mathbb{R}^{K \times N_{\mathrm{c}}}$ and $e^{\mathrm{m}} \in \mathbb{R}^{K \times H \cdot W}$ denote the color attribute embedding and image embedding produced by CMSM, respectively. We can compute a cross-modality ranking loss similarly to computing the DAMSM loss in [20]. Such a CMSM loss can provide the supervision signals for whether the generated images are aligned with the input color conditions.

3.2. Spatially-adaptive compositing

For compositing, we follow the recently proposed pixel transformation method [1] which uses a neural network to

estimate the contrast and brightness transformation parameters given both the foreground and the background images. In [1], they made the spatially-invariant assumption so that the output of the neural network is a set of shared transformation parameters for all pixels in the foreground (i.e., the same transformation is applied for each pixel). Specifically, the compositing model F^{C} is a convolutional neural network which takes a composited image (Eq. (1)) of the foreground image $\hat{y}^{\mathrm{f}}$ and the background image y^{b} as input, and output the contrast and brightness transformation parameters defined w.r.t. the RGB color space as $\rho, \tau \in \mathbb{R}^3$. By tiling ρ and τ spatially to $\bar{\rho} \in \mathbb{R}^{3\times H\times W}$ and $\bar{\tau} \in \mathbb{R}^{3\times H\times W}$, the color tone adjustment is achieved via a spatially-invariant transformation at the pixel level as:

$$y^{\mathrm{f}} = \bar{\rho} \odot \hat{y}^{\mathrm{f}} \oplus \bar{\tau}, \tag{3}$$

where $\odot$ and $\oplus$ indicate the element-wise operations. The spatially-invariance constraint is posed as:

$$\forall k, (i,j),\ \bar{\rho}[k,i,j] = \rho[k],\ \bar{\tau}[k,i,j] = \tau[k], \tag{4}$$

As in Eq. (2), (ρ, τ) are inferred by F^{C} conditioned on both $\hat{y}^{\mathrm{f}}$ and y^{b}. The supervision comes from two losses: $\mathcal{L}(F^{\mathrm{C}}) = \mathcal{L}_{\text{GAN}} + \mathcal{L}_{\mathrm{R}}$. $\mathcal{L}_{\text{GAN}}$ is a GAN loss driving F^{C} to learn inferring (ρ, τ) which makes $\hat{y}^{\mathrm{f}}$ indistinguishable from y^{b}, and $\mathcal{L}_{\mathrm{R}}$ is defined as a L1 loss for the regularization purpose to discourage dramatic changes from the input foreground. $\mathcal{L}_{\mathrm{R}}$ can be simply written as:

$$\mathcal{L}_{\mathrm{R}}(F^{\mathrm{C}}) = |y^{\mathrm{f}} - \hat{y}^{\mathrm{f}}|. \tag{5}$$

Over-saturated problem. We observe that it sometimes brings the over-saturated effects by directly applying the pixel transformation as in Eq. (3). We argue that this is because such a transformation for the foreground image is unbounded, while the values of the background image are normalized in the range of $[-1, 1]$.

To address the over-saturation problem, we can use a non-linear activation function $\tanh$ to confine the value range of transformed foreground image:

$$y^{\mathrm{f}} = \tanh(\bar{\rho} \odot \hat{y}^{\mathrm{f}} \oplus \bar{\tau}). \tag{6}$$

Pitfall of gradient vanishing. By adding $\tanh$, however, we observe that such a spatially-invariant transformation leads the optimizer to the pitfall of gradient vanishing, in which D_{I} and D_{S} become much more powerful than F^{C}, and the regularization loss $\mathcal{L}_{\mathrm{R}}$ totally fails to assist the optimizer in escaping the pitfall. See Infer-SpInv-B in Figure 5 for the failure cases.

Remark 3.2. *If the regularization loss $\mathcal{L}_R$ is effective, it should be able to assist the optimizer in escaping the pitfall. This, in turn, demonstrates that the migration to the introduction of* $\tanh$ *weakens the effectiveness of $\mathcal{L}_R$.*

Proof. We study the effectiveness of $\mathcal{L}_{\mathrm{R}}$ by checking the gradients deriving from it. By substituting y^{f} in Eq. (5) with Eq. (6), we have $\mathcal{L}_{\mathrm{R}}(F^{\mathrm{C}}) = |\tanh(\bar{\rho} \odot \hat{y}^{\mathrm{f}} \oplus \bar{\tau}) - \hat{y}^{\mathrm{f}}|$. Then, the partial derivative of $\mathcal{L}_{\mathrm{R}}$ w.r.t. $\bar{\rho}$ can be written as: $\frac{\partial \mathcal{L}_{\mathrm{R}}}{\partial \rho} = \ell(y^{\mathrm{f}} - \hat{y}^{\mathrm{f}}) \cdot (1 - \tanh^2(\bar{\rho} \odot \hat{y}^{\mathrm{f}} \oplus \bar{\tau})) \cdot \hat{y}^{\mathrm{f}}$, where $\ell(\cdot)$ is an indicator function that $\ell(\cdot) = 1$ if $y^{\mathrm{f}} > \hat{y}^{\mathrm{f}}$, and otherwise $\ell(\cdot) = -1$. When $|\bar{\rho}| \to \infty$ or $|\bar{\tau}| \to \infty$, $(1 - \tanh^2(\bar{\rho} \odot \hat{y}^{\mathrm{f}} \oplus \bar{\tau})) \to 0$ and $\frac{\partial \mathcal{L}_{\mathrm{R}}}{\partial \rho} \to 0$. The absolute values of $\bar{\rho}$ and $\bar{\tau}$ can be very large under the circumstance of lacking a good initialization for F^{C}. In this case, the gradients from $\mathcal{L}_{\mathrm{R}}$ can be unstable and tiny, making it significantly less effective.

Corollary 3.3. *The bounded version of spatially-invariant transformation requires a good initialization for F^C. Otherwise, the adversarial discriminators become too strong and can easily identify the synthesized fake images. This causes the gradient vanishing for the GAN losses. At this moment when $\mathcal{L}_R$ should eagerly push F^C to improve, the effectiveness of $\mathcal{L}_R$ will be significantly restricted by its unstable and tiny gradients due to the lack of a good initialization for F^C.*

Spatially-adaptive compositing. Inspired by a Verse in Bible, "A threefold cord is not quickly broken", we solve the gradient vanishing problem in Eq. (6) by removing the spatially-invariant constraint in Eq. (4). To be specific, instead of predicting a single set of shared transformation parameters, our spatially-adaptive model outputs a separate set of transformation parameters for each foreground pixel, i.e., $\rho, \tau \in \mathbb{R}^{3\times H\times W}$. The insight here is: the probability that the majority of the learned transformation parameters are too large is small, and thus it greatly reduces the risk of the gradient vanishing pitfall. Our experimental results verify the effectiveness of this adaptive composting method.

3.3. Learning

We train the proposed MISC framework by solving a minimax optimization problem given by

$$\begin{aligned}\min_{\mathbf{D}^{\mathrm{I}}, D^{\mathrm{C}}} \max_{F^{\mathrm{G}}, F^{\mathrm{C}}} \mathcal{L}_{\text{GAN-I}}(\mathbf{D}^{\mathrm{I}}, F^{\mathrm{G}}, F^{\mathrm{C}}) + \lambda_{\text{CM}}\mathcal{L}_{\text{CM}}(F^{\mathrm{G}}) + \\ \lambda_{\text{GAN-C}}\mathcal{L}_{\text{GAN-C}}(D^{\mathrm{C}}, F^{\mathrm{C}}) + \lambda_{\mathrm{R}}\mathcal{L}_{\mathrm{R}}(F^{\mathrm{C}}).\end{aligned} \tag{7}$$

where $\mathcal{L}_{\text{GAN-I}}$, $\mathcal{L}_{\text{CM}}$, $\mathcal{L}_{\text{GAN-C}}$ and $\mathcal{L}_{\mathrm{R}}$ are the GAN loss for the overall image quality, CMSM loss for the color condition, segmentation-based GAN loss and a regularization loss for the compositing performance, respectively.

In Eq. (7), $\mathbf{D}^{\mathrm{I}} = \{D_1^{\mathrm{I}}, \ldots, D_i^{\mathrm{I}}, \ldots, D_n^{\mathrm{I}}\}$ is a set of joint-conditional-unconditional patch discriminators [11] for each stage of F^{G}. Given a pair of image and spatially-specific color condition, *i.e.*, (y, e^{gc}), D_i^{I} can be written as: $\mathbf{p}^{\mathsf{u}}[y]_i = D_i^{\mathrm{I}}(y)$, $\mathbf{p}^{\mathsf{c}}[y, e^{\mathsf{c}}]_i = D_i^{\mathrm{I}}(y, e^{\mathrm{gc}})$, where the superscript u and c indicate the "unconditional" and "conditional". $\mathbf{p} = \{p_1, \ldots, p_j, \ldots, p_{N^{\mathrm{pat}}}\}$ is a set of probabilities with each indicating the realness of a patch. The input to D_i^{I} is indicated within the square brackets.

$\mathcal{L}_{\text{GAN-I}}$ helps both F^{G} and F^{C} to produce realistic images, which is defined as follows:

$$\mathcal{L}_{\text{GAN-I}}(F^{\text{G}}, F^{\text{C}}, \mathbf{D}^{\text{I}}) = \\ -\sum_{i=1}^{n} \frac{1}{2N_i^{\text{pat}}} \sum_{j=1}^{N_i^{\text{pat}}} \left(\lambda^{\text{u}} \log p_j^{\text{u}}[\hat{y}_i]_i + \log p_j^{\text{c}}[\hat{y}_i, e_i^{\text{gc}}]_i\right) \\ -\frac{1}{2N_n^{\text{pat}}} \sum_{j=1}^{N_n^{\text{pat}}} \left(\lambda^{\text{u}} \log p_j^{\text{u}}[y^{\text{s}}]_n + \log p_j^{\text{c}}[y^{\text{s}}, e_n^{\text{gc}}]_n\right), \tag{8}$$

where $\hat{y}_i$ is the generated image at the i-th stage of F^{G}, and y^{s} is an image composited by $\hat{y}_n^{\text{f}}$ and a background image. e_i^{gc} (defined in §3.1.2) is a spatial-specific color condition with both the geometry and color information encoded. λ^{u} is a balancing hyperparameter.

As mentioned in §3.1.2, $\mathcal{L}_{\text{CM}}$ is a cross-modality ranking loss for the image modality and the color one, which further enforces the obedience to the color condition with the help of negative examples.

In Eq. (7), D^{C} is a discriminator that learns to separate the composited foreground, which is proposed in [1]. Given a composited image y^{s}, D^{C} can be written as $\mathbf{p}^{\text{s}}[y^{\text{s}}] = D^{\text{C}}(y^{\text{s}})$, where $\mathbf{p}^{\text{s}} = \{p_1^{\text{s}}, \ldots, p_j^{\text{s}}, \ldots, p_{N^{\text{pix}}}^{\text{s}}\}$ is a set of probabilities with each indicating how likely a pixel belongs to the composited foreground regions. Thus, $\mathcal{L}_{\text{GAN-C}}$ is defined as a segmentation loss, which drives F^{C} to learn inferring reasonable contrast and brightness parameters: $\mathcal{L}_{\text{GAN-C}}(F^{\text{C}}, D^{\text{C}}) = -\frac{1}{N^{\text{pix}}} \sum_{j=1}^{N^{\text{pix}}} \log(1 - p_j^{\text{s}}[y^{\text{s}}])$.

$\mathcal{L}_{\text{R}}$ is a pixel-wise L1 loss to regularize the training so as to anchor the transformed foreground regions y^{f} to the original $\hat{y}^{\text{f}}$ which is generated by F^{G}. $\mathcal{L}_{\text{R}}$ is defined as follows: $\mathcal{L}_{\text{R}}(F^{\text{C}}) = \frac{1}{N^{\text{fpix}}} \sum_{j=1}^{N^{\text{fpix}}} \left|y^{\text{f}}[j] - \hat{y}^{\text{f}}[j]\right|$, where N^{fpix} denotes the number of foreground pixels. $y^{\text{f}}[j]$ and $\hat{y}^{\text{f}}[j]$ are the j-th pixel of an image.

Based on the experiments on a held-out validation set, we set the hyperparameters in this section as: $\lambda_{\text{CM}} = 20$, $\lambda_{\text{GAN-C}} = 0.03$, $\lambda_{\text{R}} = 1.0$ and $\lambda^{\text{u}} = 4.0$. Note that our discriminators and losses are not new to image generation. Our contribution is in extending their use to a challenging and novel generation-compositing setting.

4. Experiments

Dataset. We process the VIP person parsing dataset [23] for evaluation. We annotate persons in VIP with 120 attribute classes, and crop the images in VIP to keep one major person in each image. We create the training and test splits, with 42K and 6K images, respectively.

Quantitative Evaluation metrics. Three evaluation metrics are used: *(i)* We use the *Fréchet inception distance* (FID) [5] score to evaluate the general image quality. *(ii)* Inspired by [20], we use R-precision, a common evaluation metric for ranking retrieval results, to evaluate whether the generated image is well conditioned on the given color attribute set. More specifically, given generated image y conditioned on the attribute set x^{c} and 5 randomly sampled attribute sets, we rank these 6 attribute sets by the pre-trained image-to-attribute retrieval model (CMSM). If the ground truth attribute set x^{c} is ranked the highest, we count this a success retrieval. We perform this retrieval task on all generated images and calculate the percentage of success retrievals as the R-precision score. *(iii)* For measuring the compositing quality, we follow [15] to use the manipulation score (M-score) which is the output by a manipulation detection model [22]. The higher M-score, the higher possibility that an image has been manipulated. For each compared method, we randomly pick 100 generated images as inputs to the detection model and calculate the average M-score.

Table 1. The quantitative experiments. ↑ (↓) means the higher (lower), the better. The best performances are highlighted in **bold**. The compared baselines are divided into three categories: pattern, color and compositing.

Category	Methods	FID ↓	R-prcn (%) ↑	M-score ↓
Pattern	AdaIN	18.03	90.78	12.54
	UniPat	16.72	92.73	7.87
Color	w/o $\mathcal{L}_{\text{CM}}$	**15.03**	83.89	5.55
	UniColor	17.24	76.71	12.38
	UniColor w/ AC	41.59	64.23	119.29
Compositing	No-Comp	22.09	93.22	151.39
	varBg	16.86	93.11	7.87
	Infer-SpInv-UB	24.11	87.95	161.07
	Infer-SpInv-B	52.23	36.72	169.93
	Infer-SpAda-UB	17.27	89.30	33.64
Ours	MISC	16.09	**93.59**	**3.86**

The ablation study is performed to evaluate the three major components in MISC, *i.e.*, the pattern injection, color injection, and compositing. Since our problem is new, so no off-the-shelf methods can be directly used for comparison. Thus, we implement 10 baseline methods by disabling modules in MISC or replacing modules with other well-known architectures or training mechanisms. The quantitative comparisons are reported in Table 1, and the qualitative ones are presented in Figure 3 and 5, in which green boxes highlight some noticeable differences. Note that the face generation is not our focus, so we directly copy and paste external faces onto the generated images.

4.1. Pattern injection

Compared methods. For pattern injection, we compare MISC with two baselines implemented with AdaIN [10] and SPADE [14], respectively: *(i)* We create *AdaIN* by disabling gates (g^{γ}, g^{β}) for the affine parameters in GAIN. *(ii)* We create *UniPat* by replacing GAIN with SPADE, and tiling the pattern conditions uniformly onto the parsing mask as the input to SPADE.

Texture uniformity. In §3.1.1, we introduce two spatially-adaptive gates g^{γ} and g^{β} conditioning on the body part parsing mask. These two gates can be applied onto the affine parameters of AdaIN, so as to control the texture uniformity under the guidance of the body part parsing mask. By comparing AdaIN and MISC in Figure 3, we can see that with-

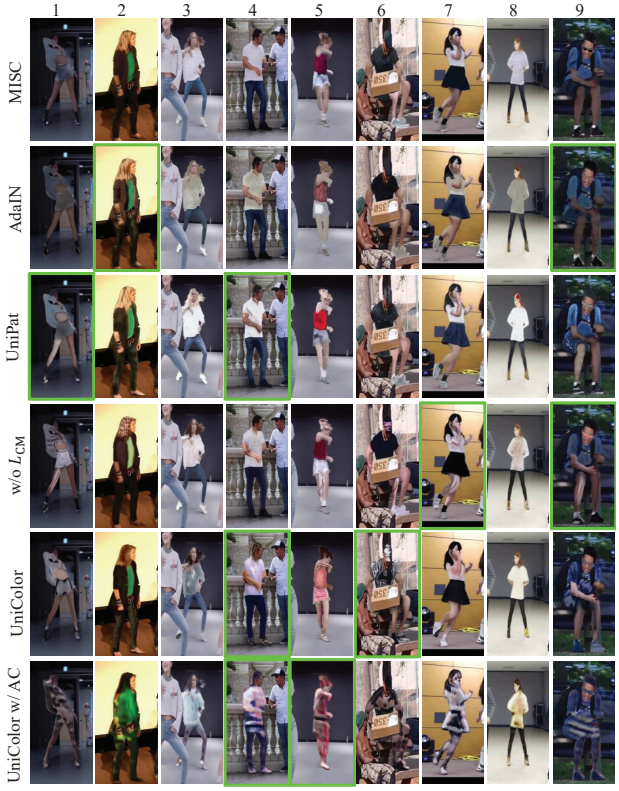


Figure 3. Qualitative comparison for conditional generation. Green boxes highlight some noticeable differences between others and MISC. Please zoom in for details.

out the control of the proposed gating mechanism, AdaIN frequently yields the mistaken non-uniform textures, *e.g.*, arms in Column 2 and legs in Column 9. The comparison between MISC and AdaIN in Table 1 also confirms the necessity and effectiveness of the proposed gating mechanism.

Uniform injection. To compare the proposed GAIN with SPADE, we adapt SPADE to our task as described above. The adaptation makes SPADE a uniform injection mechanism (named *UniPat*) for the pattern condition. As shown in Figure 3, such a uniform injection causes obvious artifacts in multiple occasions, *e.g.*, legs in Column 1 and arms in Column 4. The negative influences of the uniform injection are also reflected in Table 1. Both quantitative and qualitative results echo the justification in §3.1.2 regarding the uniform injection: "it brings extra burdens for the generation model in identifying useful information for each region".

4.2. Color injection

Compared methods. The design of our color injection module is centered on an attribute encoder which enables both the spatially-specific color condition and the cross-modality ranking loss ($\mathcal{L}_{\text{CM}}$ in Eq. (7)). Thus, we create three baselines to study the effectiveness of our color injection module. *(i)* We first create a baseline by removing $\mathcal{L}_{\text{CM}}$.

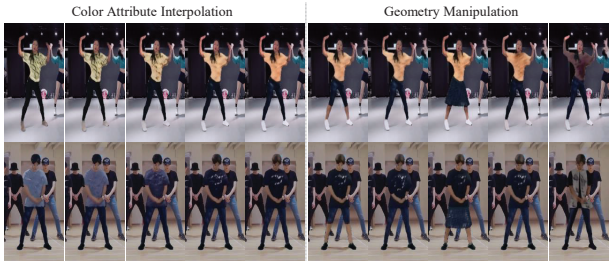


Figure 4. Controllability of color and geometry conditions. Please zoom in for details.

(ii) Then, based on *(i)*, we create *UniColor* by replacing the injection of spatially-specific color condition with the injection of multihot binary attribute vector as in [2, 4, 12]. *(iii)* We create another baseline by incorporating *UniColor* with an auxiliary classifier [13] for attributes.

Cross-modality ranking loss $\mathcal{L}_{\text{CM}}$ is designed to supervise the color condition injection module. By removing $\mathcal{L}_{\text{CM}}$, the color condition injection will only be implicitly supervised by the joint-conditional-unconditional discriminators $\mathbf{D}^{\text{I}}$ defined in §3.3. In Figure 3, we show that the removal of $\mathcal{L}_{\text{CM}}$ leads to wrong colors, *e.g.*, leg color in Column 7 and 9. An interesting phenomenon is shown in Table 1, compared to MISC, the removal of $\mathcal{L}_{\text{CM}}$ leads to significant downgrade on the color metric (R-precision) and slight downgrade on the compositing metric (M-score), while it leads to a saturation of the the image quality metric (FID). This demonstrates the necessity of a comprehensive evaluation with multiple metrics.

Spatially-specific color condition. We evaluate the importance of spatially-specific color condition by comparing *UniColor* against *MISC w/o* $\mathcal{L}_{CM}$. Figure 3 shows that the uniform injection (*UniColor*) frequently leads to drastic failures in manifesting the color conditions, *e.g.*, wrong color of the up-clothes in Column 4 and 6. The quantitative comparison in Table 1 shows the obvious downgrade from *MISC w/o* $\mathcal{L}_{CM}$ to *UniColor*. These all demonstrate the importance of the spatially-specific color condition.

Auxiliary attribute classifier. As mentioned in §3.1.2, the conventional supervision for colors usually comes from an auxiliary attribute classifier [2, 12, 20] which is trained together with the image discriminator. Therefore, we introduce such a classifier to the *UniColor* baseline to see what influences it will bring. As shown in Figure 3 and Table 1, such a classifier causes the further performance downgrade (compared to *UniColor* and *MISC*) both qualitatively and quantitatively. As mentioned in §3.1.2, the auxiliary classifier suffers from handling a large number of classes, which may explain such a downgrade.

Qualitative controllability study. We demonstrate the robustness of MISC in synthesizing interpolated color attributes and manipulated geometries. The synthesis results are visualized in In Figure 4.

4.3. Compositing

As introduced in §3.2, our compositing model achieves the image compositing by adjusting the color tone of the foreground within the framework of alpha blending. We introduce the bounding mechanism using $\tanh$ to address the over-saturation problem, and we enable the spatial adaptability to help model avoid the pitfall of gradient vanishing. Thus, we study the effectiveness of our compositing model from three aspects: *(i)* the necessity of compositing through alpha blending, *(ii)* the bounding mechanism with $\tanh$, and *(iii)* the spatial adaptability.

Compared methods. We study the necessity of compositing or the necessity compositing within the framework of alpha blending. We create a baseline named *No-Comp* by directly removing the compositing model, and create an another baseline named *varBg* which is beyond the scope of alpha blending, and adjusts the color tone of the foreground together with the background in a black-box image-to-image translation fashion. We also study the impacts of the bounding mechanism (using $\tanh$) and the spatial adaptability. To this end, we implement three more baselines with or without the bounding mechanism or the spatial adaptability (abbreviated as "SpAda"), including *Infer-SpInv-B*, *Infer-SpInv-UB* and *Infer-SpAda-UB*, where "Infer" means that the affine parameters for adjusting the color tone are inferred by our compositing model. "SpInv" indicates the spatial-invariance constraint defined in Eq. 4, which means no spatial adaptability. "B" and "UB" represent "with" and "without" the bounding mechanism, respectively. In this context, our MISC can be represented as *Infer-SpAda-B*. Note that by *Infer-SpInv-UB*, we attempt to implement the compositing model proposed in [1].

Necessity of compositing through alpha blending. By comparing the quantitative and qualitative results of *No-comp* and *MISC*, it shows that the compositing model of MISC indeed significantly improves the image quality. Specifically, without the compositing model, the color tone of the generated textures look artificial, *e.g.*, persons in Column 1 and 5 of Figure 5. *varBg* as a naive compositing approach can achieve generally good results (slightly worse than MISC in all three metrics as shown in Table 1). However, as shown in Figure 5, it is hard for varBg to achieve fine-grained pleasing textures, especially within a small part, *e.g.*, arms in Column 4 and 9. We argue that this is because multiple Conv layers pollute the intermediate feature maps with much contextual information.

Bounding with $\tanh$ & spatial adaptability. As shown in Figure 5, *Infer-SpInv-UB* frequently generates over-saturated effects, *e.g.*, Column 2 and 9. This demonstrates the necessity of the bounding mechanism. However, as justified in §3.2 simply adding the bounding mechanism to *Infer-SpInv-UB* (*i.e.*, *Infer-SpInv-B*) does not overcome the over-saturated problem but introduces the pitfall of gradient vanishing, which are reflected in the total failure results in Figure 5. In §3.2, we propose that by enabling the spatial adaptability of *Infer-SpInv-B* (*i.e.*, *Infer-SpAda-B*, the compositing model of MISC) can resolve the pitfall of gradient vanishing, which can be proved both quantitatively and qualitatively in Table 1 and Figure 3. For readers' interests, we also compare *Infer-SpAda-UB* with our compositing model. Table 1 shows that the removal of the bounding mechanism impacts the compositing metric (M-score) significantly, and Figure 5 also shows the frequent and slight over-saturated effects which are highlighted using small red boxes in Column 2 and 4.

Figure 5. Qualitative comparison for compositing. Green boxes highlight some noticeable differences between others and MISC.

5. Conclusions

We present the MISC framework for conditional person image synthesis. Our contributions include injecting multiple correlated conditions in the person image generation, the spatially-adaptive image compositing method, as well as the complete pipeline for generating photo-realistic images. In experiments, we show the superior performance with both the qualitative and quantitative results.

Acknowledgements

PKU affiliated authors are supported by National Natural Science Foundation of China under Grant No. 61872012, National Key R&D Program of China (2019YFF0302902), and Beijing Academy of Artificial Intelligence (BAAI).

References

[1] Bor-Chun Chen and Andrew Kae. Toward realistic image compositing with adversarial learning. In *CVPR*, 2019. 1, 2, 4, 5, 6, 8

[2] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 4, 7

[3] Xiaodong Cun and Chi-Man Pun. Improving the harmony of the composite image by spatial-separated attention module. *CoRR*, abs/1907.06406, 2019. 2

[4] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *TIP*, 2019. 2, 7

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a nash equilibrium. *NIPS*, 2017. 6

[6] Hao-Zhi Huang, Sen-Zhe Xu, Junxiong Cai, Wei Liu, and Shi-Min Hu. Temporally coherent video harmonization using adversarial networks. *TIP*, 2020. 2

[7] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 3

[8] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 3

[9] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 2

[10] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 2, 3, 6

[11] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *CVPR*, 2019. 2, 5

[12] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. STGAN: A unified selective transfer network for arbitrary image attribute editing. In *CVPR*, 2019. 1, 4, 7

[13] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, 2017. 7

[14] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 1, 2, 4, 6

[15] Shashank Tripathi, Siddhartha Chandra, Amit Agrawal, Ambrish Tyagi, James M. Rehg, and Visesh Chari. Learning to generate synthetic data via compositing. In *CVPR*, 2019. 2, 6

[16] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR*, 2017. 2

[17] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, and Ming-Hsuan Yang. Sky is not the limit: semantic-aware sky replacement. *TOG*, 2016. 2

[18] 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 *CVPR*, 2018. 2

[19] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. GP-GAN: towards realistic high-resolution image blending. In *ACM MM*, 2019. 2

[20] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018. 1, 2, 4, 6, 7

[21] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *TPAMI*, 2019. 2

[22] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Learning rich features for image manipulation detection. In *CVPR*, 2018. 6

[23] Qixian Zhou, Xiaodan Liang, Ke Gong, and Liang Lin. Adaptive temporal encoding network for video instance-level human parsing. In *ACM MM*, 2018. 2, 6

[24] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Learning a discriminative model for the perception of realism in composite images. In *ICCV*, 2015. 1