Computers & Graphics 116 (2023) 73–81

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on CAD/Graphics 2023

DeepPortraitDrawing: Generating human body images from freehand sketches

Xian Wu [a,1], Chen Wang [b,1], Hongbo Fu [c], Ariel Shamir [d], Song-Hai Zhang [b,*]

[a] Bytedance, Beijing, 100028, China
[b] Tsinghua University, Beijing, 100084, China
[c] City University of Hong Kong, 518057, Hong Kong
[d] Reichman University, Herzliya, 4610101, Israel

ARTICLE INFO

Article history:
Received 24 April 2023
Received in revised form 11 July 2023
Accepted 5 August 2023
Available online 9 August 2023

Keywords:
Generative adversarial networks
Image-to-image generation
StyleGAN

ABSTRACT

Various methods for generating realistic images of objects and human faces from freehand sketches have been explored. However, generating realistic human body images from sketches is still a challenging problem. It is, first because of the sensitivity to human shapes, second because of the complexity of human images caused by body shape and pose changes, and third because of the domain gap between realistic images and freehand sketches. In this work, we present DeepPortraitDrawing, a deep generative framework for converting roughly drawn sketches to realistic human body images. To encode complicated body shapes under various poses, we take a local-to-global approach. Locally, we employ semantic part auto-encoders to construct part-level shape spaces, which are useful for refining the geometry of an input pre-segmented hand-drawn sketch. Globally, we employ a cascaded spatial transformer network to refine the structure of body parts by adjusting their spatial locations and relative proportions. Finally, we use a style-based generator as the global synthesis network for the sketch-to-image translation task which is modulated by segmentation maps for semantic preservation. Extensive experiments have shown that given roughly sketched human portraits, our method produces more realistic images than the state-of-the-art sketch-to-image synthesis techniques.

1. Introduction

Creating realistic human images benefits various applications, such as fashion design, movie special effects, and educational training. Generating human images from freehand sketches can be more effective since even non-professional users are familiar with such a pen-and-paper paradigm. Sketches can not only represent the global structure of a human body but also depict the local appearance details of the body as well as garments.

Deep generative models, such as generative adversarial networks (GANs) [1] and variational auto-encoders (VAEs) [2], have recently made a breakthrough for image generation tasks. Based on these generative models, many methods [3–6] have been proposed to generate desired images from input sketches by solving a general image-to-image translation problem. Some other methods have focused on generating specific types of images, including human faces [7,8], human hairs [9] and foreground objects [10]. Such methods can better handle freehand sketches by incorporating the relevant domain knowledge.

Compared to many other types of images, human body images have more complicated intrinsic structures and larger shape and pose variations, making the sketch-based synthesis task difficult for the following reasons. First, existing public human portrait image datasets [11] only cover a small subset of all possible human images under various changing conditions of pose, shape, viewpoint, and garment. Since the existing sketch-to-image translation techniques often use pairs of images and their corresponding edge maps for training, they may fail to generate desired results when a test sketch is under very different conditions. Second, hand-drawn sketches, especially those created by users with little drawing skills, can hardly describe accurate body geometry and structure and look very different from edge maps extracted from the training images (Fig. 1). Simultaneously, style-based generative models [12,13] conditioned on human poses have demonstrated impressive performance for controllable human body image synthesis [14,15], which motivates us to exploit it with sketch condition.

In this work, we present *DeepPortraitDrawing*, a novel deep generative approach for generating realistic human images from coarse, rough freehand sketches (Fig. 2). Instead of trying to

[*] Corresponding author.
E-mail addresses: thuwx95@gmail.com (X. Wu), cw.chenwang@outlook.com (C. Wang), hongbofu@cityu.edu.hk (H. Fu), arik@idc.ac.il (A. Shamir), shz@tsinghua.edu.cn (S.-H. Zhang).

[1] The first two authors contributed equally to this work, the order is random.

https://doi.org/10.1016/j.cag.2023.08.005
0097-8493/© 2023 Elsevier Ltd. All rights reserved.

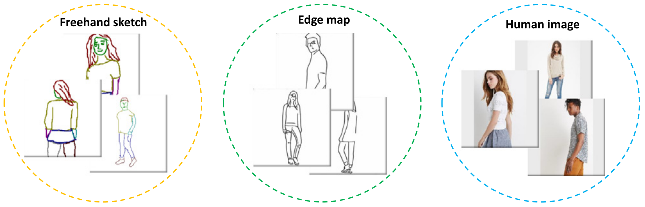

Fig. 1. There are huge gaps between freehand sketches with human images and the extracted edge maps. The freehand sketches, especially by those with few drawing skills, might not describe the local geometry or global structure of a human body accurately.

increase the generalization ability of sketch-to-image algorithms, our key idea is to project an input test sketch to part-level shape spaces constructed based on image-based training data. This can assist to bridge the gap between the training and test data, and also the gap between freehand sketches and realistic images. This idea makes sense for our task since roughly drawn sketches do not provide hard constraints for geometric interpretation. By properly recombining part-level information in different training images we are able to cover a significant portion of all possible human images.

To this end, we take a local-to-global approach to encode complicated body shapes under various poses. For each semantic body part, we employ an auto-encoder to define a part-level latent shape space by training on part-level edge maps extracted from images. Our system takes as input a semantically segmented sketch, whose individual body parts are projected onto the constructed part-level shape spaces. This results in a geometrically refined sketch map and a corresponding parsing map (i.e., labeled regions). Next, we employ a cascaded spatial transformer network to structurally refine the sketch and parsing maps by adjusting the locations and relative proportions of individual body parts. Finally, we use a global sketch conditioned and paring modulated StyleGAN to produce a photo-realistic human image from the transformed maps.

Extensive experiments demonstrate the effectiveness and practicability of our method. We are able to satisfy novice users' need for creating visually pleasing human images from hand-drawn sketches. In our self-collected dataset of freehand sketches, our method produces visually more pleasing results with more realistic local details, compared to the previous sketch-based image generation techniques (Fig. 7). The main contributions of our paper can be summarized as follows:

- We are the first to deal with roughly drawn human body sketches and synthesize realistic images accordingly;
- We present a local-to-global deep generative solution to geometrically and structurally refine an input sketched human before image synthesis.
- We collect a hand-drawn sketch dataset of human images (containing 308 segmented sketches), which can facilitate future research.

2. Related work

2.1. Sketch-to-image generation

Generating desired images from hand-drawn sketches is a difficult task, since sketches often exhibit different levels of abstraction. To address this domain gap, traditional methods take a *retrieval-composition* approach, essentially considering sketches as soft constraints. For example, a pioneering work by Chen et al. [16] first retrieves images from the Internet using input sketches with text descriptions, and fuses the retrieved foreground and background images into desired pictures. A similar idea is used in PhotoSketcher [17]. PoseShop [18] constructs image scenes with human figures but requires users to provide 2D poses for retrieval. Since such retrieval-based approaches directly reuse portions of existing images for re-composition, their performance is highly dependent on the scale of image datasets, as well as the composition quality.

By using deep learning models, (e.g., conditional GANs [19]), recent sketch-based image synthesis works adopt a *reconstruction*-based approach. Some works [3,20,21] aim at general-purpose image-to-image translation and can handle sketches as one of the possible input types. Other works focus on using sketches as the condition for GANs. For example, Scribbler [4] can control textures in generated images by grayscale sketches and colorful strokes. Contextual-GAN [5] updates latent vectors for input sketches through backpropagation and produces images by a pre-trained model. SketchyGAN [6] and iSketchNFill [10] are able to generate multi-class images for diverse sketches by introducing gated conditions. Gao et al. [22] propose an approach to produce scene images from sketches, by generating each foreground object instance and the background individually. Later, Ho et al. [23] propose a coarse-to-fine generation framework and incorporate human poses to synthesize human body images. While impressive results were presented in the above works, these techniques do not generalize well to rough or low-quality sketches, which have very different characteristics compared to image edge maps used for training the generative models.

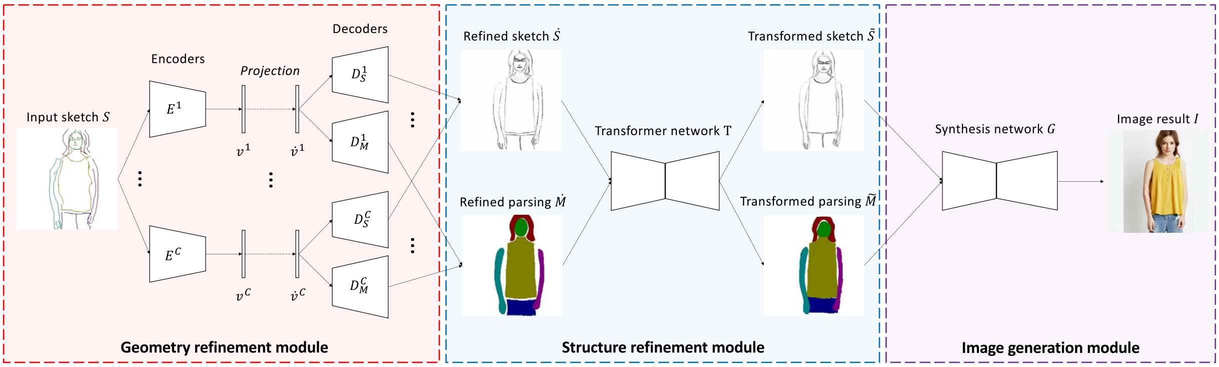

Fig. 2. Pipeline of the proposed *projection-transformation-reconstruction* approach to generate human body images from freehand sketches. Firstly, individual body parts of an input sketch are projected onto the underlying part-level manifolds and decoded into a geometrically refined sketch map and a parsing map, based on an auto-encoder architecture. Secondly, the individual parts of the refined sketch map and the parsing map are transformed via a cascaded spatial transformer network, to refine the global structure of the human body. Thirdly, the transformed maps are fed into the global synthesis network to generate a new human image and then the face refinement network to enhance the facial details.

Additionally, since sketches are largely used as hard constraints in these techniques, the synthesized images would inherit geometric distortions if they exist in the input sketches (Fig. 7). Yang et al. [24] transformed input sketches to synthesize human faces and improve robustness towards free-hand sketches, but only to a limited extent. Yang et al. [25] proposed to utilize the latent space of StyleGAN to generate face images semantically and geometrically consistent with the input sketch. DrawingInStyles [26] also generated high-quality face generation by encoding and mapping the input sketches to a pretrained StyleGAN.

Our approach is inspired by DeepFaceDrawing [8], which takes a *projection-reconstruction* approach for synthesizing realistic human face images from sketches. The key idea of DeepFaceDrawing is to refine the input sketches before synthesizing the final image. This refinement is achieved by projecting the input sketches to component-level spaces spanned by edge maps of realistic faces. DeepFaceDrawing achieves impressive results even for rough or incomplete sketches and can be further extended to face editing [27]. However, it is limited to the synthesis of frontal faces. We extend their approach to synthesizing human body images under various poses and viewpoints. Our extension explicitly uses the semantic information in the whole pipeline, and contributes a spatial transformation module, essentially leading to a *projection-transformation-reconstruction* pipeline.

2.2. Label-to-image generation

There are many semantic synthesis approaches generating images from segmentation label maps. For example, Pix2pix [3] is a general image-to-image translation framework based on a U-Net [28] generator and a conditional discriminator, based on which Pix2pixHD [21] enables high-resolution generation by proposing multi-scale networks and feature matching loss. Chen and Koltun [29] present a cascaded refinement network and use multi-layer perceptual losses to achieve photographic images from segmentation maps. GauGAN [30] introduces the SPADE layer to control image styles directly by semantic segmentation. Zhu et al. [31] present a semantically multi-modal synthesis model to generate images with diverse styles for each semantic label. LGGAN [32] combines local class-specific subgenerators and a global image-level generator for semantic scene generation. DAGAN [33] present two novel attention modules to capture spatial-wise and channel-wise attention individually. CollageGAN [34] refined results of a base generator based on a segmentation map with multiple class-specific models. Diffusion models [35] have also shown great potential for the image-to-image task by conditioning the denoising process on an input image [36,37]. Although segmentation labels can be used to generate plausible images, they are less expressive than sketches in describing local details and geometric textures of user-desired images. (e.g., collars and sleeves in Fig. 7)

2.3. Human body image generation

Human-body image synthesis is challenging in that the human visual system is sensitive to human shapes, so it is necessary to make the global body structure reasonable and produce realistic local textures. Most researchers have focused on the pose-guided person image synthesis [38,39], which transfers the same person's appearance from a source image in target poses. To achieve this, some methods utilize component masks [40,41], human parsing [42–44], or correspondence flows [45–47] to transform local source features into target areas, thus preserving the appearance of the same person in target poses. Several methods [48–51] extract a surface texture map from a source human body image and then synthesize a target human image with it using neural rendering or generative models. Besides pose, other approaches synthesize human images with different controls. For example, FashionGAN [52] encodes the shape, appearance, and text, allowing to edit of garment textures of human images through text descriptions. Ak et al. [53] and Men et al. [44] use attribute vectors to represent appearance information and then control the clothes and textures of human images via such attribute vectors. Dong et al. [54] leverage a parsing map as guidance and introduce an attention normalization layer to edit human images by sketches and colors. Many researchers have also attempted to address the virtual try-on problem [55,56], i.e., dressing a source person with given clothes through proper geometric transformations. However, both pose-guided and controllable methods cannot generate a brand-new human image from scratch in that the former is constrained by the textures from source images and the latter only supports attribute editing, while we focus on generating body textures and garments according to hand-drawn sketches.

More recently, several research proposed synthesizing human body images with a style-based [12,13] generator for its ability to generate high-quality images. InsetGAN [57] produces human portraits by inserting generated body parts onto a global canvas. Other works investigated pose-conditioned StyleGAN for synthesizing human body images [14,15] or virtual on [58]. These methods replace the constant low-dim tensor with pose features and modulate in 1D or 2D with textures to control both shape and appearance in StyleGAN. Inspired by them, we also employ StyleGAN as our generation module, but with sketch features as condition and parsing map modulation.

3. Method

We propose a *projection-transformation-reconstruction* approach for generating realistic human body images from freehand sketches. As illustrated in Fig. 2, it is achieved through three modules operated in sequence: a geometry refinement module, a structure refinement module, and an image generation module. The geometry refinement module takes a semantically segmented sketch as input and refines the geometry of its individual body parts by retrieving and interpolating the exemplar body parts in the latent spaces of the learned part-level auto-encoders. This module results in a refined sketch map and a corresponding parsing map. The structure refinement module spatially transforms the sketch and parsing maps to better connect and shape individual parts, and refine the relative proportions of body parts. Finally, the image generation module translates the transformed maps into a realistic human body image.

3.1. Geometry refinement module

This module aims to refine an input freehand sketch by using human portrait images to train several part-level networks. This has two advantages. First, locally pushing the input sketch towards the training edge maps, and second reducing the geometric errors in the input sketch. This assists the image generation module in generating more realistic images.

Due to the complexity of human images, it is very unlikely to find in our training dataset an image that is globally similar to an input sketch (Fig. 7). On the other hand, it is much easier to retrieve similar body parts and learn a component-level shape space for each body part. We thus follow the idea in DeepFaceDrawing [8] to perform manifold projection at the component level.

DeepFaceDrawing has focused on the synthesis of frontal faces and relies on a shadow interface to guide users to sketch face components that are well aligned with the training examples.

This alignment is critical for synthesizing realistic faces with DeepFaceDrawing. In contrast, we aim to handle portrait images under various poses and viewpoints. Hence, we cannot use a single layout template for body components. Instead, we propose to use the semantic segmentation information through the entire pipeline, since semantic labels provide a natural way to establish corresponding body parts in different images.

Let S denote a test sketch or a training edge map. We assume that S has been semantically segmented into $C = 8$ parts, including hair, face, top-clothes, bottom-clothes, left and right arms, left and right legs. We denote the part sketches as $\{S^c\}_{c=1,\ldots,C}$. Each body part S^c is cropped by a corresponding bounding box (S^c will be a white image if part-c is absent from S). We use an auto-encoder architecture to extract a feature vector for each body part to facilitate the subsequent manifold projection task, as illustrated in Fig. 2.

In the testing stage, given a semantically segmented sketch denoted as $\{S^c\}_{c=1,\ldots,C}$, we project its body parts to the underlying part-level manifolds for geometric refinement. We adopt the Locally Linear Embedding (LLE) algorithm [59] to perform manifold projection without explicitly constructing each part-level manifold. Specifically, each part sketch S^c is first encoded into a latent vector v^c by a corresponding encoder E^c. Based on the local linear assumption, we use a retrieve-and-interpolate approach. In more detail, we first retrieve K nearest neighbors $\{v_k^c\}_{k=1,\ldots,K}$ for v^c in the latent space $\{v_i^c\}$ for part c using the Euclidean distance. $\{v_i^c\}$ collected from a set of training images can be considered as the samples that build the underlying part-level manifold for part c. We then interpolate the retrieved neighbors to approximate v^c by minimizing the mean squared error as follows:

$$\min \|v^c - \sum_{k=1}^{K} w_k^c \cdot v_k^c\|_2^2, \quad s.t. \sum_{k=1}^{K} w_k^c = 1, \tag{1}$$

where $K = 10$ in our experiments and w_k^c is the unknown weight of the kth vector candidate. For each body part, $\{w_k^c\}$ can be found independently by solving a constrained least-squares problem. After the weights $\{w_k^c\}$ are found, we can calculate the projected vector $\dot{v}^c$ by linear interpolation:

$$\dot{v}^c = \sum_{k=1}^{K} w_k^c \cdot v_k^c. \tag{2}$$

Next, the sketch decoder D_S^c and the mask decoder D_M^c for part c process the projected vector $\dot{v}^c$, resulting in a refined part sketch $\dot{S}^c$ and a part mask $\dot{M}^c$, respectively. Finally, all projected part sketches $\{\dot{S}^c\}$ and masks $\{\dot{M}^c\}$ are combined together to recover the global body shape, resulting in a geometry-refined sketch map $\dot{S}$ and a human parsing map $\dot{M}$.

In the training stage, we first train the encoder E^c and the sketch decoder D_S^c to avoid the distraction from the mask branch. Since E^c and D_S^c need to reconstruct the input S^c with consistent shapes and fine details, we employ the L_2 distance as the reconstruction loss to train them. Then, we fix the weights of the parameters in E^c and train the mask decoder D_M^c. We use the cross-entropy loss for this training since it is a binary segmentation task.

3.2. Structure refinement module

The geometry refinement module focuses only on the refinement of the geometry of individual body parts in a sketch. However, relative positions and proportions between body parts in a hand-drawn sketch might not be accurate. We thus employ the structure refinement module to refine the relative positions and proportions of body parts to get a globally more consistent body image.

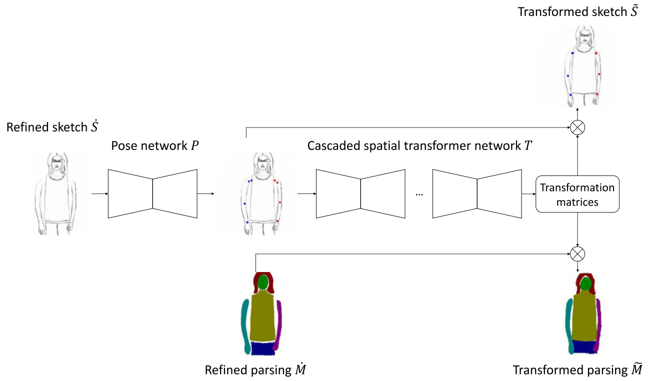
Transformed sketch $\tilde{S}$
Refined sketch $\dot{S}$
Pose network P
Cascaded spatial transformer network T
...
Transformation matrices
Refined parsing $\dot{M}$
Transformed parsing $\tilde{M}$

Fig. 3. Illustration of the structure refinement module. The keypoints of individual body parts (e.g., the arms and shoulders) are better connected and their relative length is globally more consistent after this step.

To refine the body structure, we use the pose keypoints (see Fig. 3), which provide a simple and effective way to represent a human body structure. According to the physiological characteristics of human beings, the positions of pose keypoints should obey two rules. First, a joint of a body part should connect to the same joint of its neighboring body part. Second, the relative length of different body parts should be globally consistent. Therefore, we aim to transform the keypoints of different body parts and make them conform to these rules.

As illustrated in Fig. 3, we first utilize a pose estimation network P to predict heatmaps H^c for the position of each keypoints from each refined part sketch map $\dot{S}^c$. Note that we need to predict the same joint repeatedly for neighboring body parts. Then, we leverage all the part heatmaps $\{H^c\}$ as guidance to recover the global structure of the sketched human body. The different body parts should preserve proper relative lengths, and connect with each other based on the inherent relationships among them. To achieve this, we apply affine transformations to the body parts predicted by a spatial transformer network [60] T, so that the part heatmaps $\{H^c\}$ are transformed to reasonable locations $\{\tilde{H}^c\}$ learned from real human poses. We apply the same predicted affine transformations to the refined part sketch maps $\{\dot{S}^c\}$ and the part mask maps $\{\dot{M}^c\}$, resulting in $\{\tilde{S}^c\}$ and $\{\tilde{M}^c\}$, respectively.

Since neighboring body parts may influence each other, it is very difficult to recover the entire human structure in one step transformation. Therefore, we use a cascaded refinement strategy, employing a multi-step spatial transformer network to update the results iteratively. To leverage the global information, we combine all the part sketch maps as $\dot{S}$ and all the part heatmaps as H, and then feed $\dot{S}$ and H to the spatial transformer network. The transformed sketch map $\tilde{S}$ and heatmaps $\tilde{H}$ in the jth step are the input to the transformer network in the $(j+1)$-th step. In our experiments, we used a three-step refinement, as illustrated in Fig. 4.

To train the pose estimation network P and the cascaded spatial transformer network T, we need to simulate the inconsistencies of the global structure we may find at the test time. We apply random affine transformations to all part edge maps $\{S^c\}$ and part heatmaps $\{H^c\}$ in the training set, except for a selected reference part. We select the top-clothes part (i.e., the upper body) as the reference part and keep it unchanged in our experiments. The pose network P needs to predict all part heatmaps $\{\hat{H}^c\}$ from each randomly transformed edge map $\hat{S}$. We adopt the stacked hourglass architecture [61] for P and use the mean squared error to train it.

The goal of the cascaded spatial transformer network T is to refine the size and location of each body part. Therefore, the

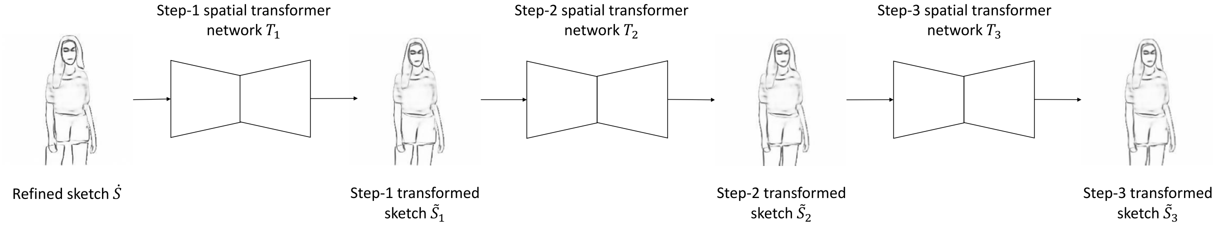
Step-1 spatial transformer network T_1
Step-2 spatial transformer network T_2
Step-3 spatial transformer network T_3
Refined sketch $\dot{S}$
Step-1 transformed sketch $\hat{S}_1$
Step-2 transformed sketch $\hat{S}_2$
Step-3 transformed sketch $\hat{S}_3$

Fig. 4. In our experiments, a geometrically refined sketch map $\dot{S}$ is transformed iteratively for three steps to get a structurally refined sketch map.

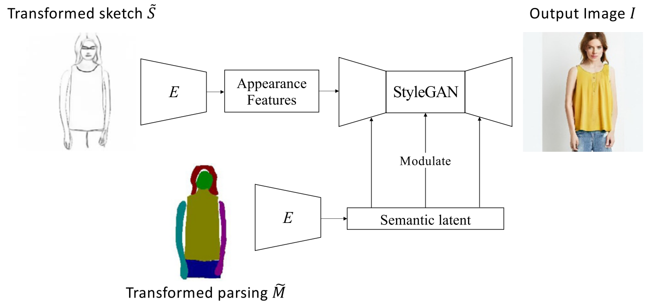

Fig. 5. Illustration of the image generation module. We encode the transformed sketch as appearance features for StyleGAN2 generator and also modulate it with transformer parsing to generate the final human body image.

predicted pose heatmaps $\{\hat{H}^c\}$ should be transformed so that they are as close to the ground truth $\{H^c\}$ as possible. Similarly, we require the randomly transformed part edge maps $\{\hat{S}^c\}$ to be close to the ground-truth part edge maps $\{S^c\}$. We have found that extremely large transformations may lead to training instability. We thus append a regularization term to penalize transformation matrices that are too large. The training struggles to converge without this regularization. The spatial transformer network T_{j+1} in the $(j+1)$-th step is fed with the transformed edge map $\hat{S}_j$ and the combined heatmaps $\hat{H}_j$ in the jth step. Its the initial input is $\hat{S}_0$ and $\hat{H}_0$. The loss function of T can be formulated as:

$$\begin{aligned}\mathcal{L}(T) = \sum_{j=0}^{2}\sum_{c=1}^{C} & \lambda_H \|\mathcal{F}(T^c_{j+1}(\hat{S}_j,\hat{H}_j),\hat{H}^c_j) - H^c\|_2^2 \\ & +\lambda_S \|\mathcal{F}(T^c_{j+1}(\hat{S}_j,\hat{H}_j),\hat{S}^c_j) - S^c\|_2^2 \\ & +\lambda_L \|T^c_{j+1}(\hat{S}_j,\hat{H}_j) - \mathcal{I}\|_2^2,\end{aligned} \tag{3}$$

where $\mathcal{F}$ represents an affine transformation operation and $\mathcal{I}$ denotes the identity matrix. $T^c_{j+1}(\hat{S}_j,\hat{H}_j)$ denotes the predicted transformation matrix for the cth body part in the $(j+1)$-th step. We set $\lambda_H = 100$ and $\lambda_S = \lambda_L = 1$ in our experiment to balance the three terms.

3.3. Image generation module

Finally, we need to generate a desired human image I from the transformed sketch map $\tilde{S}$ and the transformed parsing map $\tilde{M}$ after the structure refinement module, as illustrated in Fig. 5. We devise our generation module based on StyleGAN2 [13] which has achieved superior quality for human image synthesis. Inspired by previous work [15], we extract the sketch map to features of dimension 16 × 16 × 512 with several residual blocks as it contains both pose and texture constraints. With sketch conditions, the global synthesis network G could produce reasonable image patches. However, since the sketch representation inherently lacks semantic information, the generated images sometimes fail to capture human structure, e.g. texture between legs, jackets and pants not separated. Therefore, we encode the parsing map $\tilde{M}$ and use fully connected layers to produce the semantic latent to modulate the generator.

To train the global synthesis network G, we could simply take the edge maps $\{S_i\}$ and the parsing maps $\{M_i\}$ in the training set as input. However, we have found that the synthesis network G trained this way cannot address freehand sketches well. Although the geometry refinement module can refine the geometric shape of an input sketch S, the resulting sketch $\dot{S}$ still differs from edge maps found in the training set. The main reason is that edge maps extracted from natural human images contain many texture details, and these can violate the local linear assumption [59] used in the step of manifold projection. Instead, to simulate the input at the test time, we take the projected version of each edge map in the training set as the input to train G. We retrieve K nearest neighbors in the underlying manifold for each edge map S_i. Then, the edge maps $\{\dot{S}_i\}$ and the parsing maps $\{\dot{M}_i\}$ decoded by the projected vectors are fed into G. Inspired by Albahar et al. [15], the loss function of our generation module is shown in Eq. (4), which includes an adversarial loss L_{adv}, a L_{l_1} loss between the synthesized image and ground truth image, and a perceptual loss L_{vgg}.

$$\mathcal{L}(G) = L_{adv} + L_{l_1} + L_{vgg} \tag{4}$$

4. Experiments

To get the paired data for training, we construct a large-scale sketch dataset of human images from DeepFashion [11], as described in Section 4.1. Section 4.2 introduces the architecture of our proposed networks and the implementation details of model training. We conduct comparison experiments with several sketch-to-image techniques in Section 4.3 to show the superiority of our method for generating human images from hand-drawn sketches. The ablation study in Section 4.4 evaluates the contribution of individual components in our method. Experiments are based on the deep learning frameworks Jittor [62] and Pytorch [63].

4.1. Data preparation

Training the global synthesis network G needs a dataset of paired images and sketches. Similar to previous methods [3,4,8], we extract edge maps from human images in DeepFashion [11] to build our synthetic sketch dataset. At first, we filter the DeepFashion dataset to remove images of the lower body. Then we apply the edge detection method proposed by Im2Pencil [64] to get an edge map for each human image (Fig. 6 from (a) to (b)). By employing the sketch simplification method proposed by Simo-Serra et al. [65], we clean noise curves in the extracted edge maps (Fig. 6(c)) so they resemble hand-drawn sketches more. This results in a new large-scale sketch dataset of human images with paired data. This dataset contains 37,844 pairs in total. We randomly select 2000 pairs as the validation set and the remaining 35,844 pairs as the training set. Both the edge map and human images are of resolution 256 × 256.

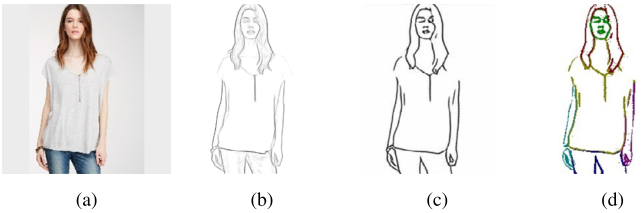

Fig. 6. The process of building our training and validation sets of sketches. (a): Input human image. (b): Edge extraction of (a) by Im2Pencil [64]. (c): Sketch simplification of (b) by the method of Simo-Serra et al. [65]. (d): Part segmentation of (c) by PGN [66].

Our models also require human parsing maps and pose heatmaps for training. We utilize PGN [66] to predict a parsing map for each human image in our dataset. To simplify the problem, we merge several labels in the parsing maps, resulting in $C = 8$ types of body parts altogether. The merged parsing maps are regarded as the ground-truth. These maps also allow us to segment the paired edge maps to obtain semantically segmented edge maps (Fig. 6(d)). To prepare the data for training the transformer network, we first employ OpenPose [67] to predict the 2D pose keypoints from the human images, and then generate pose heatmaps from the keypoints based on the Gaussian distribution to better capture spatial features.

To evaluate the usefulness of our method in practice, we have collected freehand sketches from 12 users (6 males, 6 females). Four of them have good drawing skills, while the others are less proficient. The users were asked to imitate a given human image or just draw an imagined human. They were instructed to draw a segmented sketch part by part, taking around one minute to complete one sketch on average. We have collected 308 hand-drawn sketches of human images in total to construct our test set. We plan to release our dataset of paired human images and synthetic edge maps as well as hand-drawn sketches publicly for future research.

4.2. Implementation details

In the geometry refinement module. We share the left and right arms/legs with the same auto-encoders by leveraging the human body symmetry, so there are in total 6 part auto-encoders. Each part encoder E^c contains five downsampling convolutional layers, with each downsampling layer followed by a residual block. A fully-connected layer is appended in the end to encode the features into the latent vector v^c of 512 dimensions. Similarly, the part decoders D^c_S and D^c_M each contain five upsampling convolutional layers and five residual blocks in total. The final convolutional layers in D^c_S and D^c_M reconstruct the part sketch S^c and the part mask M^c, respectively. To train the structure refinement module, we preprocess the training set by applying random affine transformations, which are composed of translation, rotation, resizing, and shearing transformations. The spatial transformer network T_j in each step consists of five downsampling convolutional layers, five residual blocks, and the last two fully-connected layers to predict the affine transformation matrices for all body parts.

We use the Adam [68] solver to train all the networks. We set the learning rate to 0.0002 initially and linearly decay it to 0 after half iterations. For each part auto-encoder, we first train the encoder E^c and the sketch decoder D^c_S for 100 epochs and then train the mask decoder D^c_M for 50 epochs, costing about 6 h on one Nvidia GTX 1080 Ti. We train the pose estimation network P and the cascaded spatial transformer network T both for 50 epochs, which costs about 50 h on one Nividia GTX 1080 Ti. We set the batch size to 16 for the above networks. We train the synthesis network G for 25k images of batch size 32, taking about 8 days with two NVidia TITAN RTX GPUs. The inference time of the sketch transformation is about 0.1 s per image, and the generation takes about 0.6 s per batch (batch size = 16).

4.3. Comparison with state-of-the-art methods

To demonstrate the effectiveness of our method for synthesizing realistic human images from freehand sketches, we compare our method with four state-of-the-art sketch-based image synthesis methods, including pix2pix [3], pix2pixHD [21], GauGAN [30], DAGAN [33] and Palette [36]. For a fair comparison, we train all the models on our training set for the same epochs as our method. Please note that we employ the first-stage generator of pix2pixHD [21], since the image resolution of our dataset is limited to 256 × 256. We also compare our method with a sketch-based image retrieval approach. To achieve this, we train an auto-encoder for an entire edge map and collect all latent vectors in the training set. Given an input sketch, we encode it into a vector and retrieve the nearest neighbor from the training set. We regard the human image corresponding to the nearest vector as the retrieval result.

Fig. 7 shows several representative results of our method and the other five approaches on our test sketches. Compared to the four state-of-the-art sketch-to-image synthesis techniques, our method performs much better with visually more pleasing results. Specifically, our method produces more realistic texture details and more reasonable body structures, owing to the geometry and structure refinement guided by the semantic parsing maps. Compared to the sketch-based image retrieval approach, our method can produce brand-new human images which respect user inputs more faithfully (the body pose of the last example in Fig. 7).

To further evaluate the results, we have applied FID [69] as a quantitative metric, which measures perceptual distances between generated images and real images. Table 1 shows that our method outperforms the other sketch-to-image synthesis methods [3,21,30], indicating more realistic results by our method. However, as claimed by [8], this perceptual metric might not measure the quality of results correctly, since it does not take the geometry and structure of the human body into consideration. Therefore, we also conducted a user study to compare our method with the three sketch-to-image synthesis techniques [3,21,30]. Since Palette [36] has a really high FID (127.9) and unreal background, we omit it for further comparison. We randomly selected 30 sketches from the test set and showed each sketch along with the four results by the compared methods in a random order to users, who were asked to pick the most realistic results. There were 17 participants in total, resulting in 510 votes. Our method received significantly more votes than the other methods, as shown in Table 1. The participants were also asked to give a score of faithfulness for each result by GauGAN [30] (we select it as the representative one of the sketch-to-image synthesis methods), the sketch-based image retrieval method, and our method. The scores ranged from 1 to 10, the higher the better. Table 1 shows that the results of our method conform with input sketches better than the image retrieval method and are comparable to GauGAN [30].

4.4. Ablation study

We have conducted an ablation study to demonstrate the contributions of the different components of our method. Each time, we remove the parsing map guidance, the projection of latent vectors, the spatial transformation respectively, while keeping the

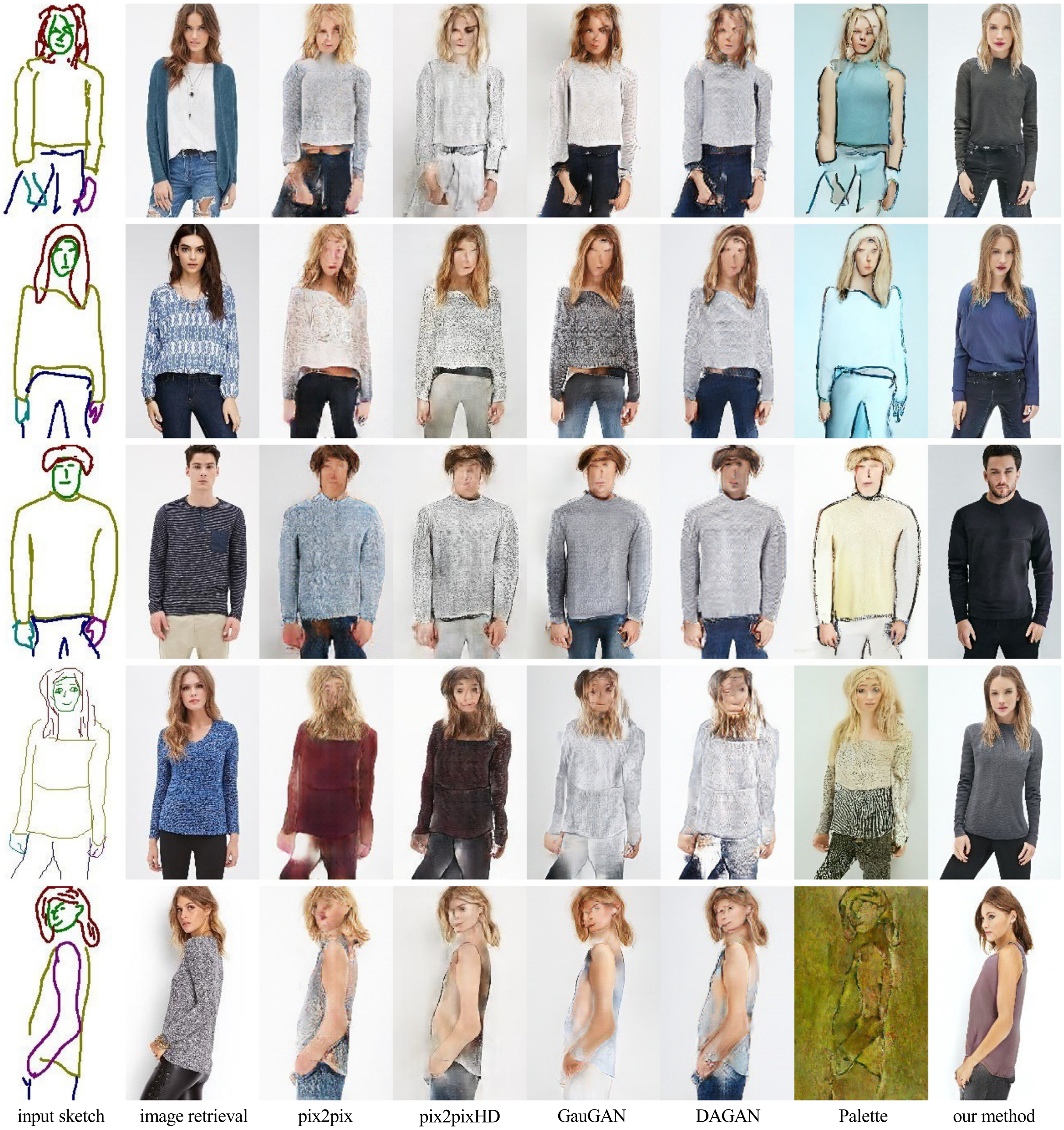

Fig. 7. Comparison results with a sketch-based image retrieval method and four state-of-the-art sketch-based image synthesis methods [3,21,30,33]. Our method can produce visually more pleasing results compared with other baselines.

Table 1
Quantitative evaluation of our method, three sketch-based image synthesis methods [3,21,30], and an image retrieval method. We have used FID [69] as a quantitative metric and conducted a user study to evaluate the realism and faithfulness of the results. The arrow after each metric identifies the improvement direction.

	FID (↓)	Realism (↑)	Faithfulness (↑)
Pix2pix	71.12	7.65%	/
Pix2pixHD	70.87	21.37%	/
GauGAN	51.92	24.71%	**6.21**
Image retrieval	/	/	5.18
Our method	**40.09**	**46.27%**	6.15

other components unchanged. As shown in Fig. 8, without the projection component, our method cannot refine the geometry of local details, resulting in obvious artifacts. Without the spatial transformation component, our method will produce results with incorrect connection relationships of joints (e.g., shoulders in the second and fourth rows) or unreasonable body proportions (e.g., long neck in the third row). Without the guidance of the human parsing map, our method cannot distinguish different body parts, leading to redundant clothes in unnecessary regions (e.g., arms in the first, second and third rows).

5. Conclusion and future work

We have proposed a *projection-transformation-reconstruction* approach for generating realistic human images from hand-drawn sketches. Our method consists of three modules, including a geometry refinement module, a structure refinement module, and an image generation module. The geometry refinement module plays an important role in converting roughly drawn sketches

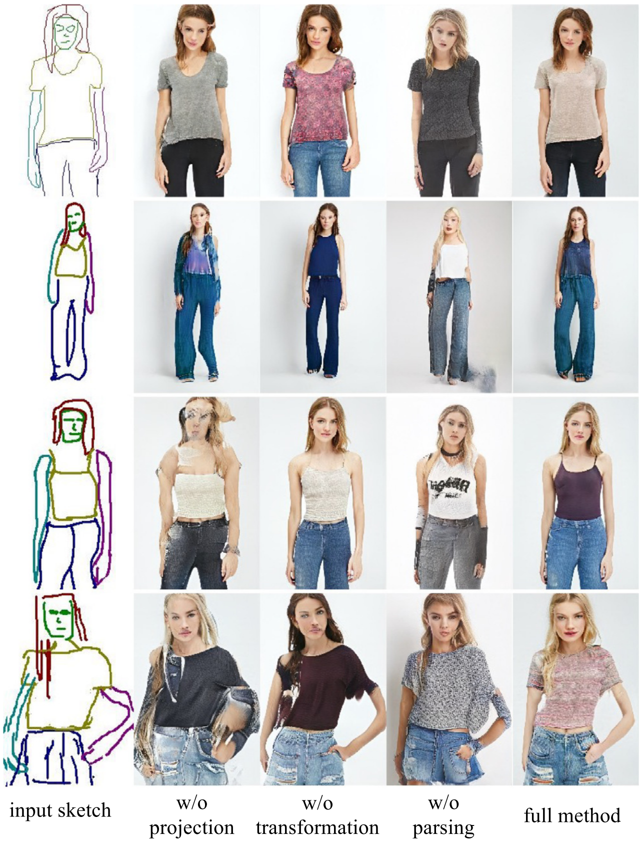

Fig. 8. Comparison results in the ablation study. We remove the projection of latent vectors, the spatial transformation and the parsing map guidance in our method, respectively.

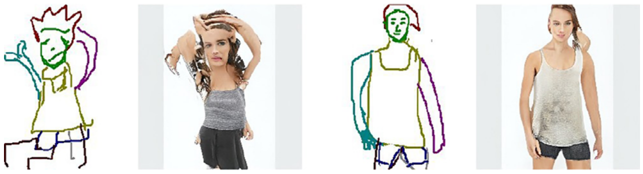

Fig. 9. Less successful cases of our method. Left: our method trained on adult images cannot handle a sketched child well. Right: our method trained on images with mixed genders might fail to respect the gender of an input sketch.

into semantic sketch maps, which are locally similar to the edge maps of real human images. This successfully bridges the gap between realistic images and freehand sketches. The structure refinement module locally adjusts spatial connections between body parts and their relative proportions to get a globally more consistent structure. The image generation module produces visually pleasing human images with fine facial details. Comparison experiments have shown that our approach outperforms three state-of-the-art sketch-to-image synthesis methods, which cannot address freehand sketches well.

Still, the geometry and structure refinement modules are restricted to the data distribution in the training set. Therefore, our method cannot produce human images very different from the images in DeepFashion [11]. For example, as shown in Fig. 9 (Left), our method generates an unsatisfying result for a hand-drawn sketch of a child. We also cannot synthesize images of more complicated body poses such as running, jumping because of the dataset issue. The structure refinement module is also limited to recovering the human body structure of an adult only since there are only adult models in DeepFashion [11]. As we do not divide the latent vectors of different genders for retrieval, our method is sometimes confused with the gender, as shown in Fig. 9 (Right). We will collect more types of human images to improve the generalization ability of our method in future work. It will also be interesting to introduce colorful strokes to control the texture styles more exactly.

CRediT authorship contribution statement

Xian Wu: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Chen Wang:** Conceptualization, Methodology, Software, Writing – review & editing. **Hongbo Fu:** Writing – review & editing. **Ariel Shamir:** Writing – review & editing. **Song-Hai Zhang:** Resources, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Advances in neural information processing systems. 2014, p. 2672–80.

[2] Kingma DP, Welling M. Auto-encoding variational Bayes. 2013, arXiv preprint arXiv:1312.6114.

[3] Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: IEEE conference on computer vision and pattern recognition. 2017, p. 1125–34.

[4] Sangkloy P, Lu J, Fang C, Yu F, Hays J. Scribbler: Controlling deep image synthesis with sketch and color. In: IEEE conference on computer vision and pattern recognition. 2017, p. 5400–9.

[5] Lu Y, Wu S, Tai YW, Tang CK. Image generation from sketch constraint using contextual gan. In: European conference on computer vision. 2018, p. 205–20.

[6] Chen W, Hays J. Sketchygan: Towards diverse and realistic sketch to image synthesis. In: IEEE conference on computer vision and pattern recognition. 2018, p. 9416–25.

[7] Li Y, Chen X, Yang B, Chen Z, Cheng Z, Zha ZJ. DeepFacePencil: Creating face images from freehand sketches. In: ACM international conference on multimedia. 2020, p. 991–9.

[8] Chen S, Su W, Gao L, Xia S, Fu H. DeepFaceDrawing: Deep generation of face images from sketches. ACM Trans Graph 2020;39(4):72:1–72:16.

[9] Xiao C, Yu D, Han X, Zheng Y, Fu H. Sketchhairsalon: Deep sketch-based hair image synthesis. 2021, arXiv preprint arXiv:2109.07874.

[10] Ghosh A, Zhang R, Dokania PK, Wang O, Efros AA, Torr PH, et al. Interactive sketch & fill: Multiclass sketch-to-image translation. In: IEEE international conference on computer vision. 2019, p. 1171–80.

[11] Liu Z, Luo P, Qiu S, Wang X, Tang X. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: IEEE conference on computer vision and pattern recognition. 2016, p. 1096–104.

[12] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 4401–10.

[13] Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of stylegan. In: IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 8110–9.

[14] Sarkar K, Golyanik V, Liu L, Theobalt C. Style and pose control for image synthesis of humans from a single monocular view. 2021, arXiv preprint arXiv:2102.11263.

[15] Albahar B, Lu J, Yang J, Shu Z, Shechtman E, Huang J-B. Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. ACM Trans Graph 2021;40(6):1–11.

[16] Chen T, Cheng MM, Tan P, Shamir A, Hu SM. Sketch2photo: Internet image montage. ACM Trans Graph 2009;28(5):1–10.

[17] Eitz M, Richter R, Hildebrand K, Boubekeur T, Alexa M. Photosketcher: Interactive sketch-based image synthesis. IEEE Comput Graph Appl 2011;31(6):56–66.

[18] Chen T, Tan P, Ma LQ, Cheng MM, Shamir A, Hu SM. Poseshop: Human image database construction and personalized content synthesis. IEEE Trans Vis Comput Graphics 2013;19(5):824–37.

[19] Mirza M, Osindero S. Conditional generative adversarial nets. 2014, arXiv preprint arXiv:1411.1784.

[20] Zhu J-Y, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE conference on computer vision and pattern recognition. 2017, p. 2223–32.

[21] Wang TC, Liu MY, Zhu JY, Tao A, Kautz J, Catanzaro B. High-resolution image synthesis and semantic manipulation with conditional gans. In: IEEE conference on computer vision and pattern recognition. 2018, p. 8798–807.

[22] Gao C, Liu Q, Xu Q, Wang L, Liu J, Zou C. SketchyCOCO: Image generation from freehand scene sketches. In: IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 5174–83.

[23] Ho TT, Virtusio JJ, Chen YY, Hsu CM, Hua KL. Sketch-guided deep portrait generation. ACM Trans. Multimedia Comput., Commun. Appl. (TOMM) 2020;16(3):1–18.

[24] Yang S, Wang Z, Liu J, Guo Z. Controllable sketch-to-image translation for robust face synthesis. IEEE Trans Image Process 2021;30:8797–810.

[25] Yang B, Chen X, Wang C, Zhang C, Chen Z, Sun X. Semantics-preserving sketch embedding for face generation. 2022, arXiv preprint arXiv:2211.13015.

[26] Su W, Ye H, Chen SY, Gao L, Fu H. Drawinginstyles: Portrait image generation and editing with spatially conditioned stylegan. IEEE Trans Vis Comput Graphics 2022.

[27] Chen SY, Liu F-L, Lai YK, Rosin PL, Li C, Fu H, et al. Deepfaceediting: Deep face generation and editing with disentangled geometry and appearance control. 2021, arXiv preprint arXiv:2105.08935.

[28] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. 2015, p. 234–41.

[29] Chen Q, Koltun V. Photographic image synthesis with cascaded refinement networks. In: Proceedings of the IEEE international conference on computer vision. 2017, p. 1511–20.

[30] Park T, Liu M-Y, Wang TC, Zhu JY. Semantic image synthesis with spatially-adaptive normalization. In: IEEE conference on computer vision and pattern recognition. 2019, p. 2337–46.

[31] Zhu Z, Xu Z, You A, Bai X. Semantically multi-modal image synthesis. In: IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 5467–76.

[32] Tang H, Xu D, Yan Y, Torr PH, Sebe N. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In: IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 7870–9.

[33] Tang H, Bai S, Sebe N. Dual attention GANs for semantic image synthesis. In: ACM international conference on multimedia. 2020, p. 1994–2002.

[34] Li Y, Li Y, Lu J, Shechtman E, Lee YJ, Singh KK. Collaging class-specific gans for semantic image synthesis. In: IEEE international conference on computer vision. 2021, p. 14418–27.

[35] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Adv Neural Inf Process Syst 2020;33:6840–51.

[36] Saharia C, Chan W, Chang H, Lee C, Ho J, Salimans T, et al. Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 conference proceedings. 2022, p. 1–10.

[37] Wang W, Bao J, Zhou W, Chen D, Chen D, Yuan L, et al. Semantic image synthesis via diffusion models. 2022, arXiv preprint arXiv:2207.00050.

[38] Ma L, Jia X, Sun Q, Schiele B, Tuytelaars T, Van Gool L. Pose guided person image generation. In: Advances in neural information processing systems. 2017, p. 405–15.

[39] Ma L, Sun Q, Georgoulis S, Van Gool L, Schiele B, Fritz M. Disentangled person image generation. In: IEEE conference on computer vision and pattern recognition. 2018, p. 99–108.

[40] Balakrishnan G, Zhao A, Dalca AV, Durand F, Guttag J. Synthesizing images of humans in unseen poses. In: IEEE conference on computer vision and pattern recognition. 2018, p. 8340–8.

[41] Siarohin A, Sangineto E, Lathuilière S, Sebe N. Deformable gans for pose-based human image generation. In: IEEE conference on computer vision and pattern recognition. 2018, p. 3408–16.

[42] Dong H, Liang X, Gong K, Lai H, Zhu J, Yin J. Soft-gated warping-gan for pose-guided person image synthesis. In: Advances in neural information processing systems. 2018, p. 474–84.

[43] Han X, Hu X, Huang W, Scott MR. Clothflow: A flow-based model for clothed person generation. In: IEEE international conference on computer vision. 2019, p. 10471–80.

[44] Men Y, Mao Y, Jiang Y, Ma WY, Lian Z. Controllable person image synthesis with attribute-decomposed gan. In: IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 5084–93.

[45] Li Y, Huang C, Loy CC. Dense intrinsic appearance flow for human pose transfer. In: IEEE conference on computer vision and pattern recognition. 2019, p. 3693–702.

[46] Liu W, Piao Z, Min J, Luo W, Ma L, Gao S. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In: IEEE international conference on computer vision. 2019, p. 5904–13.

[47] Ren Y, Yu X, Chen J, Li TH, Li G. Deep image spatial transformation for person image generation. In: IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 7690–9.

[48] Neverova N, Alp Guler R, Kokkinos I. Dense pose transfer. In: European conference on computer vision. 2018, p. 123–38.

[49] Sarkar K, Mehta D, Xu W, Golyanik V, Theobalt C. Neural re-rendering of humans from a single image. In: European conference on computer vision. Springer; 2020, p. 596–613.

[50] Liu L, Xu W, Zollhoefer M, Kim H, Bernard F, Habermann M, et al. Neural rendering and reenactment of human actor videos. ACM Trans Graph 2019;38(5):1–14.

[51] Sarkar K, Liu L, Golyanik V, Theobalt C. HumanGAN: A generative model of humans images. 2021, arXiv preprint arXiv:2103.06902.

[52] Zhu S, Urtasun R, Fidler S, Lin D, Change Loy C. Be your own prada: Fashion synthesis with structural coherence. In: IEEE international conference on computer vision. 2017.

[53] Ak KE, Lim JH, Tham JY, Kassim AA. Attribute manipulation generative adversarial networks for fashion images. In: IEEE international conference on computer vision. 2019, p. 10541–50.

[54] Dong H, Liang X, Zhang Y, Zhang X, Shen X, Xie Z, et al. Fashion editing with adversarial parsing learning. In: IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 8120–8.

[55] Han X, Wu Z, Wu Z, Yu R, Davis LS. Viton: An image-based virtual try-on network. In: IEEE conference on computer vision and pattern recognition. 2018, p. 7543–52.

[56] Wang B, Zheng H, Liang X, Chen Y, Lin L, Yang M. Toward characteristic-preserving image-based virtual try-on network. In: European conference on computer vision. 2018, p. 589–604.

[57] Frühstück A, Singh KK, Shechtman E, Mitra NJ, Wonka P, Lu J. InsetGAN for full-body image generation. In: IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 7723–32.

[58] Lewis KM, Varadharajan S, Kemelmacher-Shlizerman I. TryOnGAN: Body-aware try-on via layered interpolation. ACM Trans Graph 2021;40(4):115:1–115:10.

[59] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science 2000;290(5500):2323–6.

[60] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks. In: Advances in neural information processing systems. 2015, p. 2017–25.

[61] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation. In: European conference on computer vision. 2016, p. 483–99.

[62] Hu SM, Liang D, Yang GY, Yang GW, Zhou WY. Jittor: A novel deep learning framework with meta-operators and unified graph execution. Sci China Inf Sci 2020;63:1–21.

[63] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: An imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 2019;32:8024–35.

[64] Li Y, Fang C, Hertzmann A, Shechtman E, Yang MH. Im2pencil: Controllable pencil illustration from photographs. In: IEEE conference on computer vision and pattern recognition. 2019, p. 1525–34.

[65] Simo-Serra E, Iizuka S, Ishikawa H. Mastering sketching: Adversarial augmentation for structured prediction. ACM Trans Graph 2018;37(1):1–13.

[66] Gong K, Liang X, Li Y, Chen Y, Yang M, Lin L. Instance-level human parsing via part grouping network. In: European conference on computer vision. 2018, p. 770–85.

[67] Cao Z, Hidalgo G, Simon T, Wei S, Sheikh Y. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans Pattern Anal Mach Intell 2021;43(1):172–86.

[68] Kingma DP, Ba J. Adam: A method for stochastic optimization. 2014, arXiv preprint arXiv:1412.6980.

[69] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems. 2017, p. 6626–37.