Synthesizing Anyone, Anywhere, in Any Pose

Håkon Hukkelås Frank Lindseth
Norwegian University of Science and Technology
hakon.hukkelas@ntnu.no

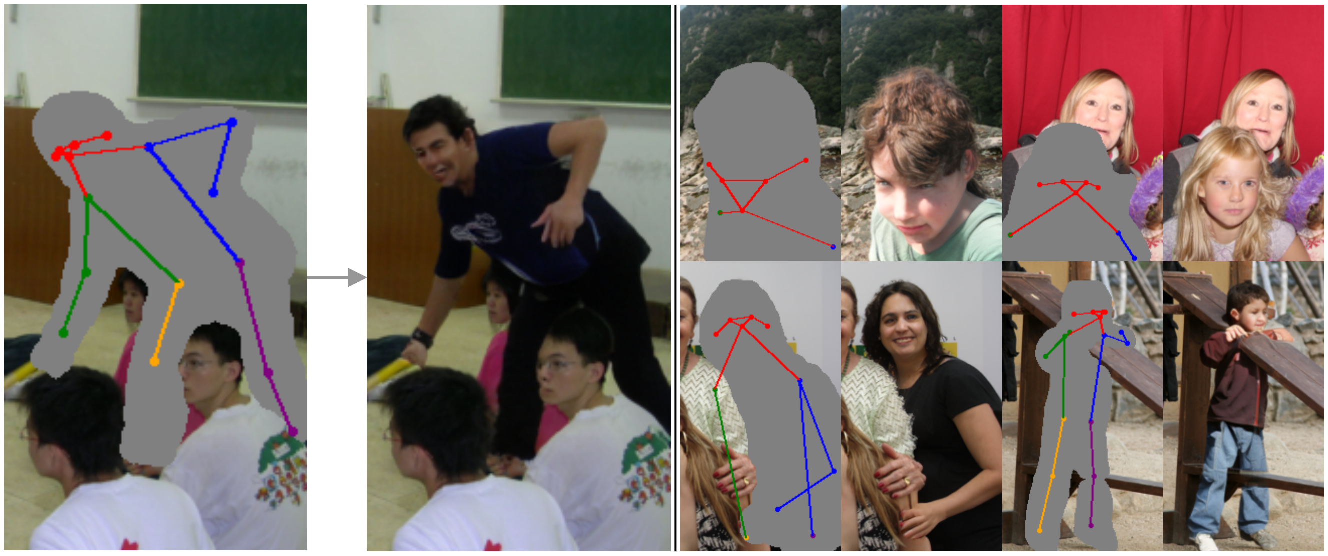


Figure 1. TriA-GAN can synthesize realistic human figures given a masked image and a sparse set of keypoints.

Abstract

We address the task of in-the-wild human figure synthesis, where the primary goal is to synthesize a full body given any region in any image. In-the-wild human figure synthesis has long been a challenging and under-explored task, where current methods struggle to handle extreme poses, occluding objects, and complex backgrounds.

Our main contribution is TriA-GAN, a keypoint-guided GAN that can synthesize Anyone, Anywhere, in Any given pose. Key to our method is projected GANs combined with a well-crafted training strategy, where our simple generator architecture can successfully handle the challenges of in-the-wild full-body synthesis. We show that TriA-GAN significantly improves over previous in-the-wild full-body synthesis methods, all while requiring less conditional information for synthesis (keypoints vs. *DensePose). Finally, we show that the latent space of TriA-GAN is compatible with standard unconditional editing techniques, enabling text-guided editing of generated human figures.*

1. Introduction

Given any image with a missing region, can you imagine a human appearance fitting into it? If there is a football next to the missing region, does your imaginary person change? This is a fascinating and difficult problem because countless possible solutions could fit the context. We refer to this task as in-the-wild human figure synthesis. Addressing this problem requires a complex understanding of human appearances and how they vary based on different environmental conditions, viewpoints, poses, and sizes of the missing region. Such a system would have widespread applications in content creation, fashion [37], or even for anonymization purposes [18].

Human figure synthesis is a well-established research field with many high-level goals. However, *in-the-wild* human figure synthesis is a difficult and under-explored task. Previous methods focus on simpler tasks, such as transferring a known appearance into a given pose [2, 4], transferring garments [14, 52], or full-body synthesis into a plain background [9]. Often they disregard the key difficulties of in-the-wild-synthesis, such as overlapping objects, partial bodies, complex backgrounds, and extreme poses. In fact, recent studies filter out these difficult cases from their dataset to improve synthesis quality [9, 10]. To the best of our knowledge, only a handful of research studies have tackled these challenges, with a focus on full-body synthesis for anonymization [18, 20] [1]. While previous methods [18] generate visually pleasing results, they heavily rely on DensePose estimation and struggle in complex scenarios. In addition, the generated images are hard to edit [18].

A key issue of current methods for in-the-wild human figure synthesis is their reliance on DensePose annotations [18,20]. The available datasets with such annotations are either limited in size [12,20] or automatically annotated [18]. We argue that this reliance constrain these methods, either by overfitting on small datasets [20] or by the numerous annotation errors arising from DensePose [18].

This paper explores full-body synthesis conditioned on sparse 2d-keypoints, eliminating the need for expensive DensePose annotations. However, this increases the modeling complexity considerably, as the generative model must now infer both the body's texture *and* its structure. We find that current GANs [18] struggle to synthesize realistic human figures without DensePose correspondences.

Our contributions address the challenge of scaling up GANs to handle in-the-wild full-body synthesis without DensePose correspondences. Key to our method is replacing the conventional GAN discriminator with Projected GANs [53]. By combining Projected GANs with a thoughtfully designed training strategy, our method can generate coherent bodies with visually pleasing textures.

Our contributions can be summarized as follows. First, we adapt Projected GANs [53] for image inpainting (Sec. 3.1), and propose a novel mask-aware patch discriminator (Sec. 3.2). Secondly, we investigate the representational power of pre-trained feature networks used by the discriminator (Sec. 3.3). Our experiments reflect that the previously used classification networks [53, 54] are poorly suited for discriminating human figures. Instead, we use a combination of self-supervised feature networks for the discriminator, which significantly improves sample quality. Finally, we propose a progressive training technique for U-Net [50] architectures (Sec. 3.4), enabling us to easily scale up to high resolutions and larger model sizes.

Our contributions culminate into a new state-of-the-art for in-the-wild human figure synthesis. As far as we know, our approach is the first to generate nearly photorealistic humans without DensePose annotations while effectively dealing with extreme poses, complex backgrounds, partial bodies, and occlusions. Source code: `http://github.com/hukkelas/deep_privacy2/`.

2. Related Work

2.1. Full-body Human Synthesis

Synthesizing human bodies has a range of applications, and previous studies have a large variety of high-level goals. We categorize human synthesis into *transfer-based* and *synthesis-based* models. *Transfer-based* methods transfers a source appearance (or garment [14, 52]) into a new pose [2, 33, 39, 47, 52, 56], motion [4] or scene [57]. While some of these methods are applicable for in-the-wild human figure synthesis [57, 67], they require a source appearance that limits the synthesized identities to a texture bank or an image dataset of appearances. In contrast, our method can directly synthesize novel identities. For the latter goal, *synthesis-based* methods can synthesize the appearance either conditioned on a pose [40, 60, 68], scene [8, 18, 20], or unconditionally [5, 9, 10]. Several of these methods are applicable for in-the-wild human synthesis [18, 40, 60], but they are limited to low-resolution [8, 40], struggle to handle complex backgrounds [9, 60], and only a few handles overlapping objects [18, 20].

Independent of the goal, most methods use a form of pose information to enhance synthesis quality through DensePose annotations [18, 20, 43, 52], semantic segmentations [5, 60, 67], sparse keypoints [2, 4, 8, 14, 33, 39, 40, 47, 56, 57, 67], or a 3d pose of the body [32, 68].

Previous studies primarily focus on GAN-based methods, but recent studies have employed diffusion models [59] for human figure synthesis [22]. Our work focuses on GANs as they offer fast sampling of high-quality images.

2.2. Generative Adversarial Networks

Generative Adversarial Networks [11] (GANs) have long been a leading generative model for a range of full-body synthesis tasks. GANs are notoriously difficult to train, and a notable research focus has been on achieving stable training of the generator, where different techniques such as novel objectives [1], architectures [24, 26–28], training strategies [25], and regularization [13, 41] has been proposed to improve stability and synthesis quality. Recently introduced Projected GANs [53] use pre-trained feature networks for the discriminator to reduce training time and improve image quality, which was later extended for high-resolution image synthesis on the ImageNet [6] dataset [54]. We continue this line of research, where we adapt projected GANs for conditional synthesis.

2.3. Image Inpainting

Image inpainting [3] aims to complete missing regions in natural images. Unlike general image inpainting, we complete missing regions that contain human figures appearing at random regions in natural images. GANs have long been the leading methodology for free-form image in-

[1] Note that other studies address similar tasks [40, 60], but they focus on simpler datasets (*i.e.* Market1501 [75], DeepFasion [37]) with few overlapping/occluding objects.

painting [46, 70], where most prior work focuses on architectural changes to the generator. For example, to handle missing values [19, 35, 70], generate higher resolution [69], utilize auxiliary information [23, 31, 42], or improve the receptive field via attention mechanisms [71] or fourier convolutions [62]. Previous methods adapt a traditional GAN discriminator, often patch discriminators [21, 36, 49, 65, 70], combined with perceptual image similarity losses [36, 49] and pixel-wise l_1 loss [49, 65]. As far as we know, we are the first to adapt Projected GANs [53] for image inpainting, where we exclusively train on the adversarial objective.

3. TriA-GAN - A Keypoint-Guided GAN

In this section, we gradually introduce changes to improve synthesis quality (Tab. 1). **Config A** (Sec. 3.1) starts with a StyleGAN-based [27] U-Net [50] architecture, similar to the architecture used in [18], trained with Projected GANs [54] using EfficientNet-Lite0 [63]. **Config B** introduces our Mask-Aware Discriminator objective (Sec. 3.2), and **Config C** replaces EfficientNet-lite0 with ViT-L16$_{\text{MAE}}$ and RN50$_{\text{CLIP}}$ (Sec. 3.3). **Config D** introduces our progressive training technique (Sec. 3.4) and finally, **Config E** increases the generator model size. To reduce training time, we ablate our method on low-resolution images (72 × 40). Finally, Sec. 3.5 increases the resolution to 288 × 160. Appendix A includes experimental and architecture details.

Problem Formulation We formulate in-the-wild full-body synthesis as an image inpainting task. Our goal is to complete the missing regions of a corrupted image $\bar{I} = I \odot M$, where I is the ground truth image, M is the mask indicating missing regions ($M_i = 1$ for known pixels and 0 for missing), and $\odot$ is element-wise multiplication. To improve synthesis quality, we condition the generator on 17 keypoints following the COCO [34] keypoint format

Dataset We conduct our experiments on the FDH dataset [18]. The FDH dataset is a large unfiltered dataset, where models trained on FDH adapt well to in-the-wild settings [18]. The dataset consists of 1.87M training images and 30K validation images. Each image includes a single human figure as the subject, but the same image can include several individuals. Each image is annotated with a 2d keypoint annotation, a segmentation mask indicating the human to be inpainted, and pixel-to-surface correspondences (*i.e.* surface of a T-shaped 3D body). Note that TriA-GAN does not use pixel-to-surface correspondences.

We find that a large amount of the keypoint annotations in the FDH dataset are incorrect. Thus, we automatically re-annotate all images with ViTPose [66] (see Appendix B).

Table 1. Iterative development of our method. Each addition is added on top of the previous. Config A-C are trained until the discriminator has observed 50M images.

Configuration	FID ↓	FID$_{\text{CLIP}}$ ↓	PPL ↓	OKS ↑
A: Baseline	1.73	1.74	55.8	0.916
B: + Mask-Aware Discriminator	1.65	1.63	52.8	0.912
C: + Improved Feature Nets	1.79	0.47	49.2	0.951
D: + Progressive Growing	1.66	0.40	52.0	**0.954**
E: + Larger G (62M → 110M)	**1.62**	**0.30**	**52.0**	0.948

Pose Representation We represent keypoints as a one-hot encoded spatial map, specifically $P \in \{0,1\}^{K\times H\times W}$ where $K = 17$ and $P_{k,y,x} = 1$ for keypoint k with location (x, y) and P is 0 otherwise. In addition, we include a spatial map (S) drawing the human skeleton. Specifically, the spatial map $S \in \{0,1\}^{6\times H\times W}$ is one-hot encoded into 6 categories, where lines connect closeby joints in the body, separated into 6 classes (left/right arm/leg, torso, head). The one-hot encoded pose and the skeleton map are concatenated with the input image of the generator.

Evaluating Sample Quality We evaluate sample quality with Fréchet Inception Distance (FID) [16] and FID$_{\text{CLIP}}$[2]. Additionally, we report latent disentanglement via Perceptual Path Length (PPL) [27], which correlates with consistency and stability of shapes [28].

Furthermore, we introduce a new metric for assessing the sample quality of generated human figures, namely Object Keypoint similarity (*OKS*), that compares the generated pose to the ground truth keypoints. The motivation behind this metric is to obtain a metric that is not influenced by the feature network used by the discriminator. Projected GANs [53] are known to achieve artificially good scores on feature-based metrics [30], which makes it challenging to make quantitative comparisons across different types of feature networks. This is evident from our experiments, where Config B (which uses ImageNet features for the discriminator) generates severely more corrupted images than Config E but still achieves a similar ImageNet FID.

Object Keypoint Similarity (*OKS*) is calculated by predicting keypoints with ViTPose [66], then computing the OKS to the ground truth keypoints following COCO [34]. Compared to direct Euclidean distance, OKS considers that predicted keypoints can deviate slightly from the ground truth keypoints, where the acceptable deviation varies for different keypoints (*e.g.* the shoulder keypoint can deviate more than the eye keypoint).

[2] ImageNet-FID scores images containing ImageNet objects higher and is insensitive to faces [30]. These issues are diminished with FID$_{\text{CLIP}}$, where we use features from a CLIP [48] pre-trained ViT-B/32.

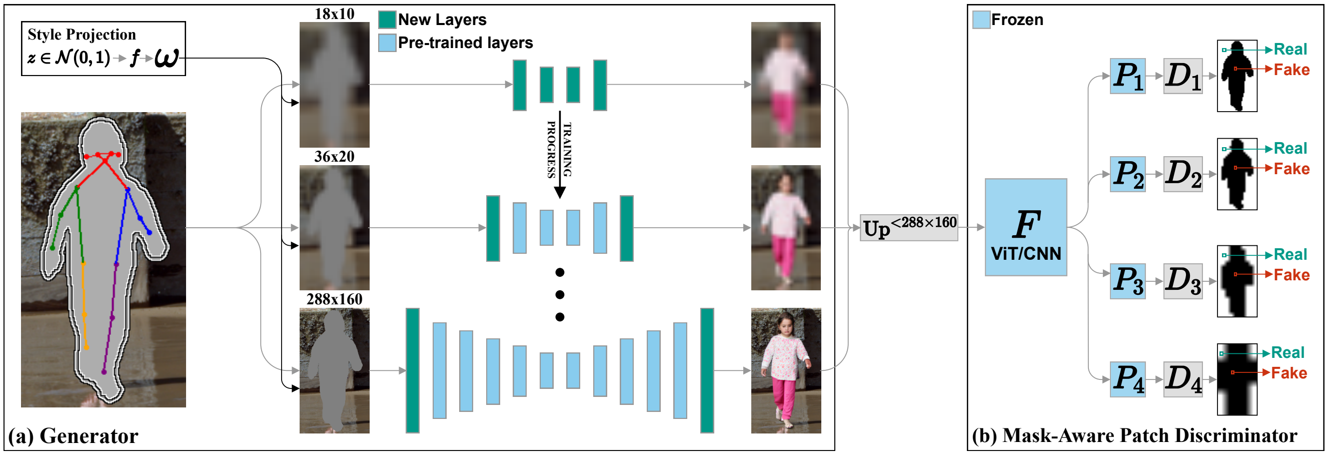


Figure 2. (a) Our generator fills in the missing region given 17 keypoints. The generator layers employ adaptive instance normalization to condition the generator on ω, where ω is the output of the style mapping network. Config D&E is trained progressively starting at 18×10 resolution, then increased by adding layers to the start/end of the encoder/decoder. Note that all layers remain trainable throughout training. (b) For each feature network F, we use four shallow patch discriminators operating its features (with different spatial resolutions), where each feature is projected through random differentiable operations (P_1-P_4). Given the projected features, each discriminator predicts if a given patch corresponds to a real or fake image region.

3.1. Projected GANs for Image Inpainting

Projected GANs [53] employ pre-trained feature networks to discriminate between real and fake images. Given an image I, the adversarial objective is formulated as

$$\min_G \max_{D_\ell} \sum_{\ell \in \mathcal{L}} \mathbb{E}_{I \sim p_{data}} \left[\log\left(D_\ell\left(P_\ell\left(I\right)\right)\right)\right] + \mathbb{E}_{z \sim p_z}\left[\log\left(1 - D_\ell\left(P_\ell\left(G\left(z, \bar{I}\right)\right)\right)\right)\right], \tag{1}$$

where $\{D_\ell\}$ is a set of independent discriminators operating on its feature projector P_ℓ. Each projector is frozen during training and consists of a pre-trained feature network F, where features from F are randomly projected with differentiable operations. For the baseline (**Config A**), we use EfficientNet-Lite0 [63] as F following [53], which we later revisit in Section 3.3. For each discriminator D_ℓ, we adopt a patch discriminator architecture, described in Section 3.2.

Equation (1) does not enforce consistency between the condition ($\bar{I}$) and the generated image, yielding a generator that learns to completely ignore $\bar{I}$ in practice. Thus, we enforce condition consistency by masking the output of the generator. Specifically, we set $G(z, \bar{I}) = \tilde{I} \odot (1 - M) + \bar{I} \odot M$, where $\tilde{I}$ is the output of the last layer in G.

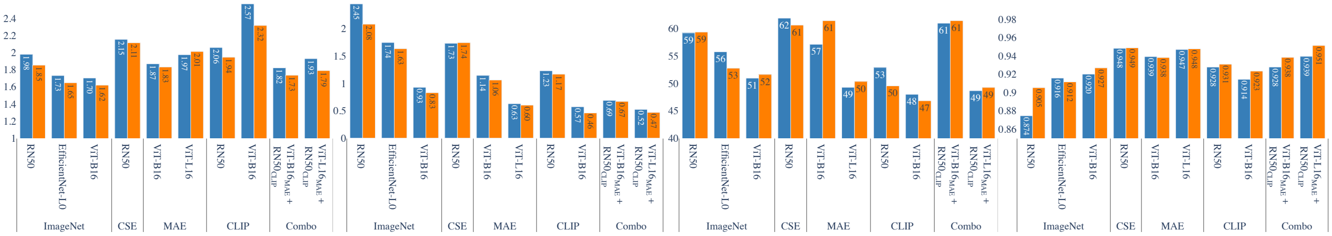
3.1.1 Stabilizing the Generator

Naively adopting projected GANs for image inpainting is unstable to train and prone to mode collapse early in training. This originates from the generator struggling to keep up with the pre-trained discriminator, where the discriminator overpowers the generator early in training. To improve stability, we introduce several modifications to the adversarial setup. First, we blur images inputted to the discriminator at the start of training, where the blur is linearly faded over 4M images. The long blur prevents the discriminator from focusing on the high-frequency edges caused by the masking of the generator output. Previous methods apply discriminator blurring over the first 200k images [26, 54], whereas we find it beneficial to significantly increase this period. Furthermore, the U-net architecture injects the latent code (z) via a mapping network and style modulation following StyleGAN2 [28]. We set the mapping network to 2 layers and reduce the dimensionality of z to 64, following [54]. Furthermore, we scale residual skip connections by $1/\sqrt{2}$ (similar to [28]), and $1/\sqrt{3}$ for skip connections where residual U-net connections are present. Finally, we use instance normalization instead of weight demodulation [28], as we find it more stable to train.

3.2. Mask-Aware Patch Discriminator

Projected GANs [53, 54] adapt four shallow discriminators operating on different feature projections (P_ℓ) with different spatial resolutions. Each discriminator output logits at the same resolution (4×4). In contrast, we find patch discriminators to work better for the image inpainting task, where each discriminator tries to classify local patches instead of the global image. Specifically, each D_ℓ (inputting features from the projection P_ℓ) consists of three convolutions, where the output of D_ℓ is half the spatial resolution of P_ℓ. We find that replacing the discriminator from [53] with a patch discriminator substantially improves performance.

Patch discriminators are widely adapted for image in-



(a) FID ↓

(b) FID$_{\text{CLIP}}$ ↓

(c) PPL ↓

(d) OKS ↑

Figure 3. Comparison of different feature networks with the standard projected GAN objective (Eq. (1)) and mask-aware discriminator objective (Eq. (2)). All models are trained until the discriminator has observed 50M images.

painting [62, 70, 73, 74]. Typically, each patch is classified as belonging to the class of the original image, such that all patches corresponding to a real image are classified as real. However, this introduces ambiguity for the image inpainting task, as certain features (*e.g.* shallow features from CNNs) might exclusively depend on real pixels even though the image is fake due to a limited receptive field. Thus, we propose a mask-aware discriminator objective, where the discriminator's patches are categorized as belonging to the real or fake class based on whether they correspond to a real or fake region in the image. The new objective is given by

$$\begin{aligned}\min_{G}\max_{D_\ell}\sum_{\ell\in\mathcal{L}}\mathbb{E}_{I\sim p_{data}}\left[\log\left(D_\ell\left(P_\ell\left(I\right)\right)\right)\right]+\\ \mathbb{E}_{z\sim p_z}\Bigg[\sum_{y}^{H_\ell}\sum_{x}^{W_\ell}M_\ell^{y,x}\cdot\log\left(D_\ell^{y,x}\left(P_\ell\left(G\left(z,\bar{I}\right)\right)\right)\right)+\\ (1-M_\ell^{y,x})\cdot\log(1-D_\ell^{y,x}(P_\ell(G(z,\bar{I}))))\Bigg],\end{aligned} \tag{2}$$

where $D_\ell \in \mathbf{R}^{H_\ell \times W_\ell}$, and M_ℓ is downsampled from M to $H_\ell \times W_\ell$ via min-pooling.

Equation (2) removes the ambiguous classification of patches due to global class allocation, which provides more detailed and spatial coherent responses to the generator. Furthermore, it introduces an auxiliary task to the discriminator, which is known to improve synthesis quality [45]. In our case, the auxiliary task is to spatially segment the region that corresponds to the generated area.

Figure 3 confirms that Equation (2) improves image quality (FID/FID$_{\text{CLIP}}$) and OKS across a range of feature networks. This includes feature networks with different pre-training tasks and architectures (CNNs and ViTs). Similar segmentation discriminators have been explored before for other tasks [55, 61, 68]. Our work further validate that this concept generalizes to extremely shallow discriminator architectures leveraging pre-trained feature networks, independent on the feature network used as F.

3.3. Discriminative Feature Networks for Human Synthesis

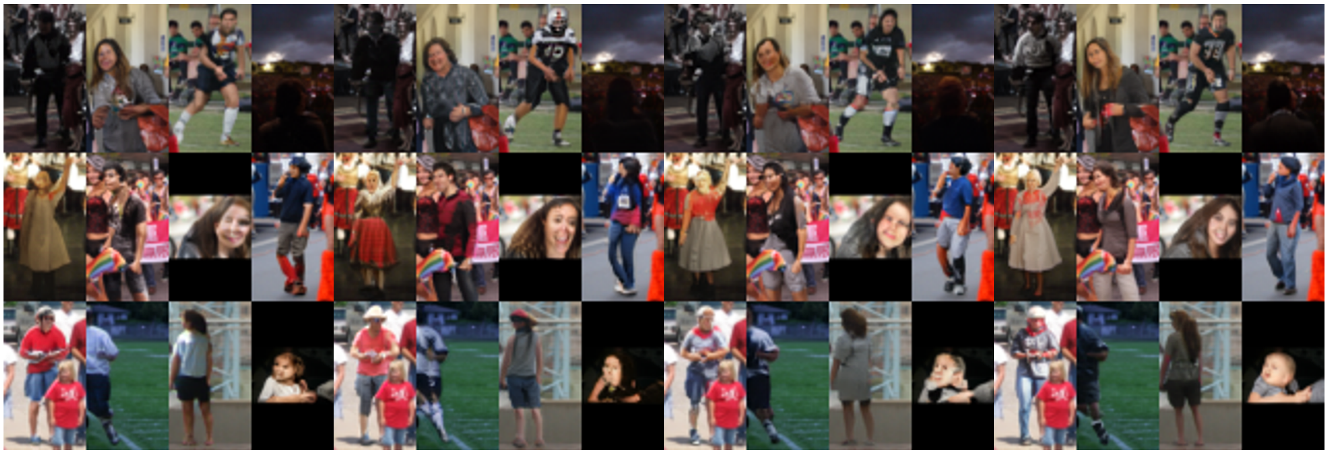
GANs have historically generated impressive results for aligned human synthesis, especially on the FFHQ [27] and CelebA-HQ [25, 38] datasets. However, projected GANs are known to generate artifacts for face synthesis on FFHQ [53] and struggle to generate realistic images of unaligned humans [54] [3]. We find that the poor human synthesis quality originates from an invariance in the pre-trained feature space used by the discriminator. Earlier work [53, 54] has utilized pre-trained ImageNet [6] classification networks. These feature networks learn feature representations for the sole goal of classification; mapping an image to the top-1 class. Hence, they learn to ignore features that are irrelevant to the goal of classification. While this invariance benefits image classification, we find it to hurt discriminative representation for human synthesis.

We explore different feature networks (including variants of CNNS/ViTs) with widely different pre-training tasks for the discriminator. Specifically, Figure 3 ablate the following feature nets with the following pre-training tasks:

- **IN**: ImageNet Classification: ResNet50 (**RN50**), **ViT-B16** (DeIT variant), EfficientNet-Lite0 (**EN-L0**).
- **CLIP**: Contrastive Language Image Pre-training [48]: **RN50**, **ViT-B16**.
- **MAE**: Masked Autoencoders [15]: **ViT-B16**, **ViT-L16**.
- **CSE**: DensePose estimation [44]: ResNet50 (**RN50**).

We refer to each model as *architecture*$_{task}$, *e.g.* RN50$_{\text{CLIP}}$ refers to ResNet-50 with CLIP pre-trained weights. Directly selecting the best feature network from standard generative metrics (FID/FID$_{\text{CLIP}}$) is ambiguous, as projected GANs are known to achieve unnatural high scores on feature-based metrics [30]. We find that ImageNet models achieve unnatural high FID due to matching pre-training

[3] See the appendix in [54].



(a) ImageNet (ViT-B/16) (b) CLIP (ViT-B/16) (c) MAE (ViT-B/16) (d) ViT-L/16$_{\text{MAE}}$ + RN50$_{\text{CLIP}}$

Figure 4. Qualitative comparison of various feature networks used for the discriminator. It is worth noting that these examples are not curated but selected from the first 12 images from the validation set.

tasks, and ViT scores better on FID$_{\text{CLIP}}$ due to matching architecture [4].

Independent of the architecture, we observe that all ImageNet [6] models generate highly corrupted faces, illustrated in Figure 4. This is most likely due to the invariance of facial descriptors in these feature networks, a phenomenon that has also been observed in [30]. Note that Appendix C includes comparison for all networks in Figure 3.

From the results in Figure 3, **Config C** replaces EfficientNet-Lite0 with ViT-L16$_{\text{MAE}}$ and RN50$_{\text{CLIP}}$. The motivation for pairing these networks is to exploit features with completely different architectures and pre-training tasks. In addition, these networks scores among the best w.r.t. OKS, FID$_{\text{CLIP}}$, and PPL. Finally, RN50$_{\text{CLIP}}$ supplements ViT well, as RN50 operates on the original aspect ratio (288 × 160), whereas ViT is fixed to 224 × 224 [5].

3.4. Progressive Growing

Progressive training [25] is known to improve training stability of GANs and was recently re-introduced for unconditional synthesis with projected GANs [54]. StyleGAN-XL [54] first trains at 16 × 16 resolution, then increases the resolution by adding new layers to the end of the decoder. Note that StyleGAN-XL freezes already trained layers and the style network when training the next stage.

We adopt a straightforward extension to the image-to-image translation case, where we progressively train the U-net architecture by adding layers to the start/end of the encoder and decoder, respectively (see Fig. 2). We observe that adding new blocks to the start of the encoder leads to training instability as it results in significant changes to the input of already-trained layers. To mitigate this, we introduce LayerScale [64] for each residual block with an initial value of 10^{-5} to lessen the contribution of new blocks. Furthermore, we include output skip connections following [27]. Unlike StyleGAN-XL, we avoid freezing any blocks during training as the computational benefit is minimal, given that we need to calculate gradients for layers at the beginning of the encoder. Introducing these changes substantially improves the final image quality (**Config D**)

We note that we experimented with more advanced techniques for progressive training, such as cascaded U-nets [17], or assymetric training of the encoder/decoder (*i.e.* start with a full-resolution encoder and a low-resolution decoder). However, we found that the straightforward progressive training technique was superior in terms of training time and final image quality.

3.5. Scaling Up the Generator

Config E double the number of residual blocks for each resolution in the encoder/decoder, resulting in 110.4M parameters in the generator compared to the previous 62.2M. This model trains stable up to 288 × 160 resolution, which is the maximum resolution of the FDH dataset.

4. Comparison to Surface-Guided GANs

Table 2 compares TriA-GAN to Surface Guided GANs (SG-GAN) [20] trained following DeepPrivacy2 [18], the current state-of-the-art for in-the-wild full-body synthesis. Figure 1 shows synthesis results with TriA-GAN, and Figure 5 compares TriA-GAN to SG-GAN. Appendix D include randomly selected samples.

The main difference between TriA-GAN and SG-GAN [18] is the improved training strategy of TriA-GAN, and the

[4] FID$_{\text{CLIP}}$ is calculated from features of ViT-B/32 following [30].

[5] ViT input resolution is set to 224 × 224 for all models, as ViT features are less robust to changes in resolution from the training resolution.

Table 2. Quantitative comparison of SG-GAN [18] *vs*. ours.

Method	FID ↓	FID_{CLIP} ↓	PPL ↓	OKS ↑
SG-GAN [18]	1.97	1.25	70.2	0.950
TriA-GAN (ours)	**1.68**	**0.43**	**47.8**	**0.972**

sparser conditional information (keypoints *vs*. dense surface correspondences). TriA-GAN improves at handling overlapping objects, partial bodies (*e.g*. intersection with image edges), and synthesis of texture (*e.g*. hair, clothing). Furthermore, TriA-GAN improves at context handling, *e.g*. inferring that an elderly lady is likely to sit at the table (top row, Fig. 5), or that there is a motorcyclist on the bike (3rd row, Fig. 5).

Finally, TriA-GAN is easier to use for downstream tasks, as our method does not rely on DensePose detections. For example, keypoints are easier to edit for interactive editing applications. Furthermore, detecting DensePose is challenging and unreliable for long-range detection, restricting its use in many scenarios (*e.g*. anonymizing pedestrians on the street). See Appendix E for examples of failure cases.

5. Editability of TriA-GAN

StyleGAN [27] is known for its disentangled latent space, and it is widely used for user-guided image editing, such as modifying images through text prompts [29]. However, most methods for editing images focus on unconditional GANs (or class-conditional GANs), and their application to image inpainting is less explored. StyleMC [29] is effective for editing faces with inpainting methods [18], but the same study finds editing human figures in-the-wild much harder [18]. We believe this limitation originates from the DensePose condition, where descriptive conditions can be correlated with specific attributes. This narrows the sampling probability, which makes it harder to find meaningful directions for randomly sampled images.

Figure 6 demonstrate that StyleMC [29] is effective with TriA-GAN to find semantically meaningful directions in the GAN latent space. StyleMC finds global directions by manipulating random images towards a text prompt using a CLIP encoder [48], where the directions are found over 1280 images. We find that StyleMC combined with TriA-GAN can edit a wide range of attributes, even quite specific attributes such as the size of the ears. However, we do note that editing some attributes results in changes to other correlated attributes. For example, the edit ”blond hair” induces slight changes to the skin color. Furthermore, some attributes are more challenging to edit. For instance, introducing ”red lips” to a body inferred as a male can result in significant semantic changes (top row, Fig. 6). It is unclear whether this limitation is a result from the editing technique

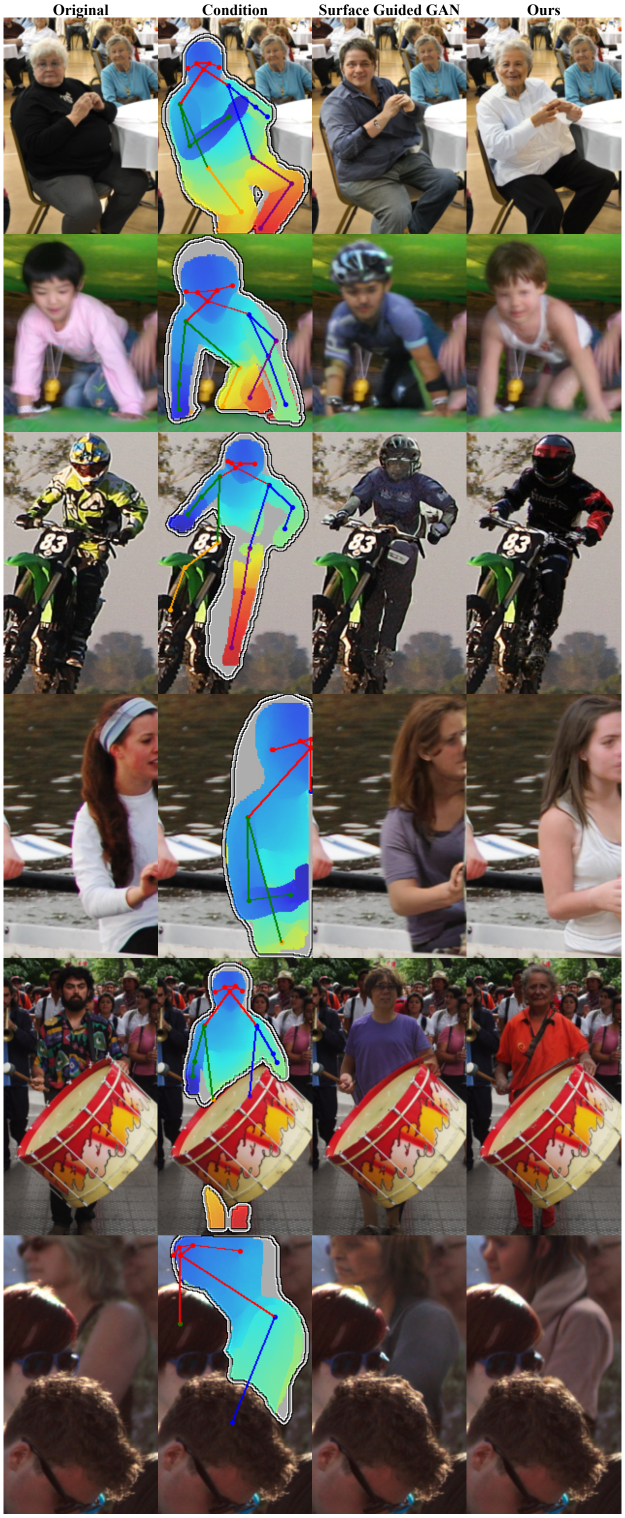


Figure 5. Curated examples comparing Surface Guided GAN [18] to TriA-GAN. Note that surface information is not used for TriA-GAN (shown in blue-yellow tint).

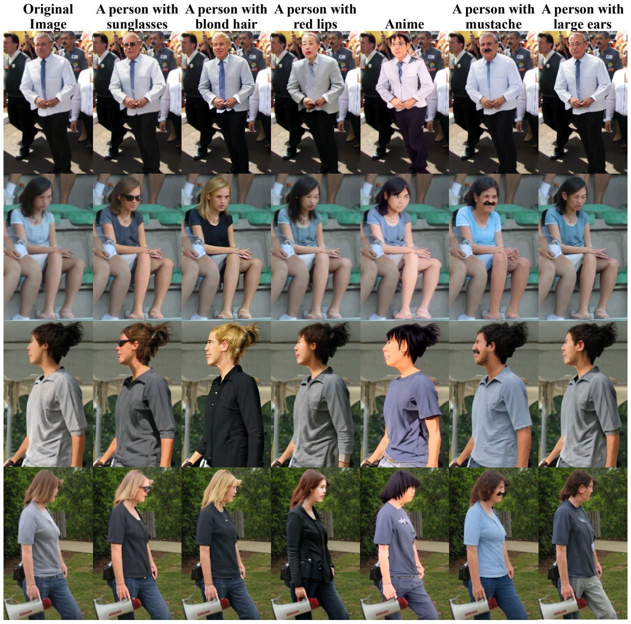


Figure 6. StyleMC [29] edits with TriA-GAN, where a global direction (from text prompt above each column) is added to the style code of the original (leftmost) image.

or TriA-GAN itself. We believe these correlations are inherent in the training datasets of CLIP or TriA-GAN.

6. Conclusion

TriA-GAN has enabled the generation of human figures in any desirable pose and location given a sparse set of keypoints, resulting in a new state-of-the-art for person synthesis on the FDH dataset. Key to our method is leveraging pre-trained feature networks for the discriminator. We demonstrate that a carefully designed training strategy combined with feature networks suited to discriminate human figures substantially improves synthesis quality. TriA-GAN is the first to demonstrate reliable attribute editing of human figures via text prompts, which we believe will be highly practical for many applications.

Societal Impact Synthesizing human figures has a range of useful applications everywhere, from content creation to anonymization purposes. However, similar to all learning-based generative models, the synthesized human figures adhere to the sampling probability of the dataset. In our case, the dataset originates from Flickr, which means that our generator follows its biases and is less likely to synthesize people from underrepresented groups on the website. Furthermore, our work focuses on generating lifelike humans, which carries the potential for abuse (*e.g.* DeepFakes). We note that the community has made a concerted effort to address this issue, through initiatives like the DeepFake Detection Challenge [7], or embedding watermarks into images from generative models [72].

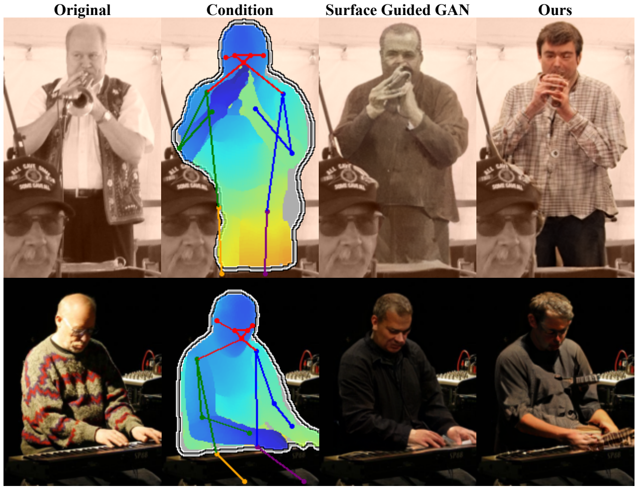


Figure 7. Failure cases of TriA-GAN.

6.1. Limitations

TriA-GAN sets a new state-of-the-art for human figure synthesis in-the-wild. Exploring methods for disentangling the latent space from the pose, body shape, and environment are exciting future avenues. Currently, the sampling space of TriA-GAN is highly dependent on the conditional information, where it can collapse into a single synthesized identity given certain conditions. Disentangled person image generation can mitigate this, by disentangle pose, appearance, and context. However, current methods require datasets with paired images [40, 51], which are less diverse and small.

The key limitation of TriA-GAN is handling more complex interactions with objects (Fig. 7). This is particularly true for generating realistic hands/fingers, *e.g.* when playing the piano. SG-GAN [18] often improve on TriA-GAN in such scenarios if the DensePose information explicitly describes the interaction. But, it still struggles in cases where it is not clear (*e.g.* playing the masked-out trumpet).

TriA-GAN is hard to edit for attributes that are less frequent in the FDH dataset. For example, many images do not contain the lower body and attempting to find editing directions for ”a person wearing red pants” results in editing other attributes as well. Whether this is a limitation to the editing method, or TriA-GAN is an open question.

Acknowledgement The computations were performed on the NTNU IDUN/EPIC computing cluster [58] and the Tensor-GPU project led by Prof. Anne C. Elster. Furthermore, we thank Rudolf Mester for his general support and thoughtful discussion.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017. 2

[2] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing Images of Humans in Unseen Poses. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8340–8348. IEEE, jun 2018. 1, 2

[3] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics*, pages 417–424, 2000. 2

[4] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei Efros. Everybody Dance Now. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, volume 49, pages 5932–5941. IEEE, oct 2019. 1, 2

[5] Bindita Chaudhuri, Nikolaos Sarafianos, Linda Shapiro, and Tony Tung. Semi-supervised Synthesis of High-Resolution Editable Textures for 3D Humans. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 7987–7996, 2021. 2

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, jun 2009. 2, 5, 6

[7] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The DeepFake Detection Challenge (DFDC) Dataset. *arXiv preprint arXiv:2006.07397*, 2020. 8

[8] Patrick Esser and Ekaterina Sutter. A Variational U-Net for Conditional Appearance and Shape Generation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8857–8866. IEEE, jun 2018. 2

[9] Anna Fruhstuck, Krishna Kumar Singh, Eli Shechtman, Niloy J Mitra, Peter Wonka, and Jingwan Lu. InsetGAN for Full-Body Image Generation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7713–7722. IEEE, jun 2022. 1, 2

[10] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. StyleGAN-Human: A Data-Centric Odyssey of Human Generation. In *Computer Vision - ECCV 2022*, volume 13676, pages 1–19. Springer, Cham, 2022. 2

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[12] Riza Alp Guler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense Human Pose Estimation in the Wild. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7297–7306. IEEE, jun 2018. 2

[13] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017. 2

[14] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. VITON: An Image-Based Virtual Try-on Network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7543–7552. IEEE, jun 2018. 1, 2

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2022-June, pages 15979–15988. IEEE, jun 2022. 5

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 3

[17] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded Diffusion Models for High Fidelity Image Generation. *Journal of Machine Learning Research*, 23:1–33, 2022. 6

[18] Håkon Hukkelås and Frank Lindseth. DeepPrivacy2: Towards Realistic Full-Body Anonymization. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1329–1338. IEEE, jan 2023. 1, 2, 3, 6, 7, 8

[19] Håkon Hukkelås, Frank Lindseth, and Rudolf Mester. Image Inpainting with Learnable Feature Imputation. In *DAGM German Conference on Pattern Recognition*, pages 388–403. Springer-Verlag, 2021. 3

[20] Håkon Hukkelås, Morten Smebye, Rudolf Mester, and Frank Lindseth. Realistic Full-Body Anonymization with Surface-Guided GANs. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1430–1440. IEEE, jan 2023. 2, 6

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976. IEEE, jul 2017. 3

[22] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics*, 41(4):1–11, jul 2022. 2

[23] Youngjoo Jo and Jongyoul Park. SC-FEGAN: Face Editing Generative Adversarial Network With User's Sketch and Color. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1745–1753. IEEE, oct 2019. 3

[24] Animesh Karnewar and Oliver Wang. MSG-GAN: Multi-Scale Gradients for Generative Adversarial Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7796–7805. IEEE, jun 2020. 2

[25] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 2, 5, 6

[26] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-Free Generative Adversarial Networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 852–863, 2021. 2, 4

[27] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405. IEEE, jun 2019. 2, 3, 5, 6, 7

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116. IEEE, jun 2020. 2, 3, 4

[29] Umut Kocasari, Alara Dirik, Mert Tiftikci, and Pinar Yanardag. StyleMC: Multi-Channel Based Fast Text-Guided Image Generation and Manipulation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3441–3450. IEEE, jan 2022. 7, 8

[30] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The Role of ImageNet Classes in Frechet Inception Distance. *arXiv preprint arXiv:2203.06026*, 2022. 3, 5, 6

[31] Avisek Lahiri, Arnav Kumar Jain, Sanskar Agrawal, Pabitra Mitra, and Prabir Kumar Biswas. Prior Guided GAN Based Semantic Inpainting. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 3

[32] Christoph Lassner, Gerard Pons-Moll, and Peter V. Gehler. A Generative Model of People in Clothing. In *2017 IEEE International Conference on Computer Vision (ICCV)*, volume 2017-Octob, pages 853–862. IEEE, oct 2017. 2

[33] Yining Li, Chen Huang, and Chen Change Loy. Dense Intrinsic Appearance Flow for Human Pose Transfer. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3688–3697. IEEE, jun 2019. 2

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European conference on computer vision*, volume 8693 LNCS, pages 740–755. Springer, Cham, 2014. 3

[35] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018. 3

[36] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent Semantic Attention for Image Inpainting. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4169–4178. IEEE, oct 2019. 3

[37] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104. IEEE, jun 2016. 1, 2

[38] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738. IEEE, dec 2015. 5

[39] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, volume 2017-Decem, pages 406–416, 2017. 2

[40] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled Person Image Generation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 99–108. IEEE, jun 2018. 2, 8

[41] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *International Conference on Learning Representations*, 2018. 2

[42] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. EdgeConnect: Structure Guided Image Inpainting using Edge Prediction. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3265–3274. IEEE, oct 2019. 3

[43] Natalia Neverova, Riza Alp Güler, and Iasonas Kokkinos. Dense Pose Transfer. In *European conference on computer vision*, volume 11207 LNCS, pages 128–143, 2018. 2

[44] Natalia Neverova, David Novotny, Vasil Khalidov, Marc Szafraniec, Patrick Labatut, and Andrea Vedaldi. Continuous Surface Embeddings. In *Advances in Neural Information Processing Systems*, volume 33, pages 17258–17270. Curran Associates, Inc., nov 2020. 5

[45] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *34th International Conference on Machine Learning (ICML)*, volume 6, pages 4043–4055, 2017. 5

[46] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context Encoders: Feature Learning by Inpainting. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544. IEEE, jun 2016. 3

[47] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised Person Image Synthesis in Arbitrary Poses. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8620–8628. IEEE, jun 2018. 2

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 3, 5, 7

[49] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. StructureFlow: Image Inpainting via Structure-Aware Appearance Flow. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 181–190. IEEE, oct 2019. 3

[50] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical image*

computing and computer-assisted intervention, pages 234–241. Springer, 2015. 2, 3

[51] Kripasindhu Sarkar, Vladislav Golyanik, Lingjie Liu, and Christian Theobalt. Style and Pose Control for Image Synthesis of Humans from a Single Monocular View. *arXiv preprint arXiv:2102.11263*, 2021. 8

[52] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural Re-rendering of Humans from a Single Image. In *European conference on computer vision*, volume 12356 LNCS, pages 596–613. Springer Science and Business Media Deutschland GmbH, 2020. 1, 2

[53] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected GANs Converge Faster. In *Advances in Neural Information Processing Systems*, pages 17480–17492, 2021. 2, 3, 4, 5

[54] Axel Sauer, Katja Schwarz, and Andreas Geiger. StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, Vancouver, BC, Canada, aug 2022. Association for Computing Machinery. 2, 3, 4, 5, 6

[55] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. A U-Net Based Discriminator for Generative Adversarial Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8204–8213. IEEE, jun 2020. 5

[56] Chenyang Si, Wei Wang, Liang Wang, and Tieniu Tan. Multistage Adversarial Losses for Pose-Based Human Image Synthesis. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 118–126. IEEE, jun 2018. 2

[57] Aliaksandr Siarohin, Enver Sangineto, Stephane Lathuiliere, and Nicu Sebe. Deformable GANs for Pose-Based Human Image Generation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3408–3416. IEEE, jun 2018. 2

[58] Magnus Själander, Magnus Jahre, Gunnar Tufte, and Nico Reissmann. EPIC: An energy-efficient, high-performance GPGPU computing research infrastructure, 2019. 8

[59] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[60] Sijie Song, Wei Zhang, Jiaying Liu, Zongming Guo, and Tao Mei. Unpaired Person Image Generation With Semantic Parsing Transformation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4161–4176, nov 2021. 2

[61] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. OASIS: Only Adversarial Supervision for Semantic Image Synthesis. *International Journal of Computer Vision*, 130(12):2903–2923, dec 2022. 5

[62] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust Large Mask Inpainting with Fourier Convolutions. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3172–3182. IEEE, jan 2022. 3, 5

[63] Mingxing Tan and Quoc V Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International conference on machine learning*, pages 6105–6114, 2019. 3, 4

[64] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Herve Jegou. Going deeper with Image Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 32–42. IEEE, oct 2021. 6

[65] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-Aware Image Inpainting. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5833–5841. IEEE, jun 2019. 3

[66] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. *arXiv preprint arXiv:2204.12484*, 2022. 3

[67] Chaojie Yang, Hanhui Li, Shengjie Wu, Shengkai Zhang, Haonan Yan, Nianhong Jiao, Jie Tang, Runnan Zhou, Xiaodan Liang, and Tianxiang Zheng. BodyGAN: General-purpose Controllable Neural Human Body Generation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7723–7732. IEEE, jun 2022. 2

[68] Zhuoqian Yang, Shikai Li, Wayne Wu, and Bo Dai. 3DHumanGAN: Towards Photo-Realistic 3D-Aware Human Image Generation. *arXiv preprint arXiv:2212.07378*, 2022. 2, 5

[69] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual Residual Aggregation for Ultra High-Resolution Image Inpainting. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7505–7514. IEEE, jun 2020. 3

[70] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas Huang. Free-Form Image Inpainting With Gated Convolution. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4470–4479. IEEE, oct 2019. 3, 5

[71] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative Image Inpainting with Contextual Attention. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5505–5514. IEEE, jun 2018. 3

[72] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14428–14437. IEEE, oct 2021. 8

[73] Tao Yu, Zongyu Guo, Xin Jin, Shilin Wu, Zhibo Chen, Weiping Li, Zhizheng Zhang, and Sen Liu. Region Normalization for Image Inpainting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12733–12740, apr 2020. 5

[74] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-Resolution Image Inpainting

with Iterative Confidence Feedback and Guided Upsampling. In *European conference on computer vision*, pages 1–17. Springer-Verlag, 2020. 5

[75] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable Person Re-identification: A Benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124. IEEE, dec 2015. 2