Generation of Human Images with Clothing using Advanced Conditional Generative Adversarial Networks

Sheela Raju Kurupathi[1,2][a], Pramod Murthy[2][b] and Didier Stricker[1,2]
[1]Department of Computer Science, Technical University of Kaiserslautern, Kaiserslautern, Germany
[2]Augmented Vision, German Research Center for Artificial Intelligence, Kaiserslautern, Germany

Keywords: Conditional GANs, Human Pose, Market-1501, DeepFashion.

Abstract: One of the main challenges of human-image generation is generating a person along with pose and clothing details. However, it is still a difficult task due to challenging backgrounds and appearance variance. Recently, various deep learning models like Stacked Hourglass networks, Variational Auto Encoders (VAE), and Generative Adversarial Networks (GANs) have been used to solve this problem. However, still, they do not generalize well to the real-world human-image generation task qualitatively. The main goal is to use the Spectral Normalization (SN) technique for training GAN to synthesize the human-image along with the perfect pose and appearance details of the person. In this paper, we have investigated how Conditional GANs, along with Spectral Normalization (SN), could synthesize the new image of the target person given the image of the person and the target (novel) pose desired. The model uses 2D keypoints to represent human poses. We also use adversarial hinge loss and present an ablation study. The proposed model variants have generated promising results on both the Market-1501 and DeepFashion Datasets. We supported our claims by benchmarking the proposed model with recent state-of-the-art models. Finally, we show how the Spectral Normalization (SN) technique influences the process of human-image synthesis.

1 INTRODUCTION

The idea of generating realistic human images has been of great value in recent times due to their varied applications in e-commerce for fashion shopping and also in synthesizing training data for person detection, person identification (Chen et al., 2019). Due to advances in Artificial Intelligence (AI), we can see the rapid growth of integrating every aspect into AI. The human-image generation has been one of the most crucial tasks over the past few decades. There exist two problems that need to be dealt with while generating human images, one is the representation of the human pose, and the other is the generation of the appearance details like clothing textures.

We have various ways to represent the poses like 2D skeletons (stick figures), segmentation masks, 3D pose skeletons, dense pose, as shown in Figure 1. For generating the clothing textures, we can use warping and clothing segmentation techniques. A wide range of deep learning models like PG^2 (Ma et al., 2017), Pix2pixHD (Wang et al., 2018), Deformable GANs (Siarohin et al., 2018) has been used for generating the human poses along with the clothing. However,

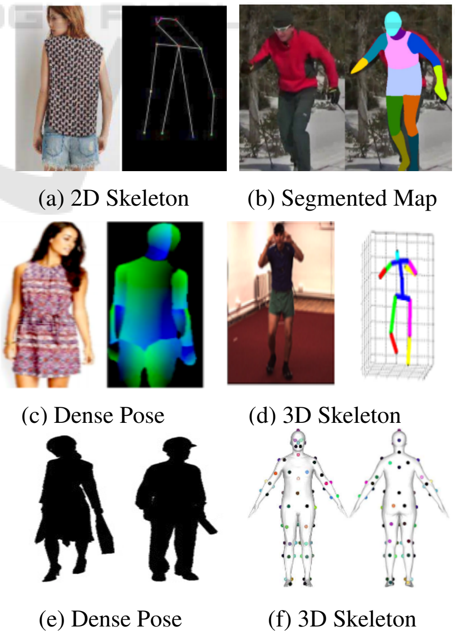
(a) 2D Skeleton
(b) Segmented Map
(c) Dense Pose
(d) 3D Skeleton
(e) Dense Pose
(f) 3D Skeleton

Figure 1: Different pose representations.

[a] https://orcid.org/0000-0003-4530-9717
[b] https://orcid.org/0000-0002-8016-8537

Kurupathi, S., Murthy, P. and Stricker, D.
Generation of Human Images with Clothing using Advanced Conditional Generative Adversarial Networks.
DOI: 10.5220/0009832200300041
In *Proceedings of the 1st International Conference on Deep Learning Theory and Applications (DeLTA 2020)*, pages 30-41
ISBN: 978-989-758-441-1

these models still suffer to generate human images with accurate pose and clothing due to many variations in the textures, appearance, and shape.

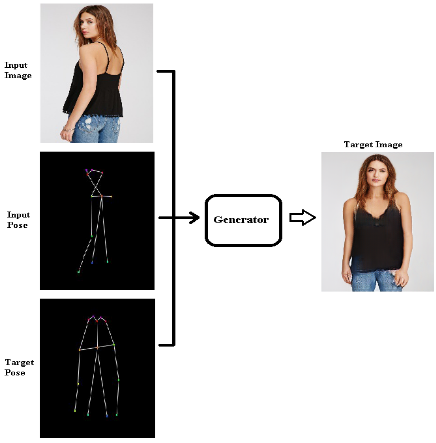

Figure 2: The conditional pose generation task. The input image source: DeepFashion (Liu et al., 2016b).

Data availability to train a deep learning model is very scarce in both 2D and 3D domains. By using the proposed models, we can generate the human-images in rare poses, which can be used as synthetic datasets for humans for further research purposes. For generating the human-image with accurate pose and appearance details, the model needs to know information about the human body poses. To avoid the expensive annotations for poses, we represent the pose related information using the 2D keypoints representing the 2D coordinates for each joint in the image. We use the HumanPoseEstimator (HPE) (Cao et al., 2017) to estimate the 2D coordinates for all the joints, and the number of keypoints is 18. Using these key points, the model learns the positions of joints in the human body. We can also use other pose representations depending on the application.

We have used Conditional GANs along with SN to deal with the problem of generating humans with pose and clothing details. The main aim can be seen from Figure 2, that given an input image of a person, the pose of that person, and a target pose to a Generator to output an image (target image) of the person in the target pose. However, training these networks requires high computational power and a sufficient amount of training data to achieve the desired results (Stewart, 2019). In this paper, we do not make any assumptions about the backgrounds, objects, etc. and we do not use any representations to denote the clothing information like segmentation, which makes the network to learn different clothing textures by itself.

2 RELATED WORK

Recently deep learning models have shown substantial improvement in the neural image synthesis (Isola et al., 2017), providing numerous applications in the field of virtual reality and gaming. One such an emerging class of models that are being vastly researched and well studied in recent years are Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). GANs are aimed to generate novel data that has similar characteristics to that of real-world data. The main idea of our proposed models is to guide the generation process explicitly by an appropriate pose representation like 2D stick figures to enable direct control over the generation process.

There have been many deep models that have been proposed to deal with the task of human image generation. The most commonly used deep learning architectures are AlexNet (Krizhevsky et al., 2012), ResNet (He et al., 2016), etc. We use Generative Adversarial Network architecture to solve the problem related to human-image generation. GANs (Goodfellow et al., 2014) are a particular class of artificial intelligence algorithms that comprise mainly two neural networks Generator *G* and Discriminator *D*, which play a zero-sum game. The applications of GANs include neural image synthesis, image in-painting, super-resolution, semi-supervised learning, and more. For our problem, we have chosen image-to-image translation in which GANs take the human image as input from one domain and translate it into another domain without any alignment between the domains. Conditional pose generation helps to synthesize a new image of a person, given a reference image of the person and a target pose. Most of the models focus on detection, pose, and shape estimation of people from images. The most critical task would be to transfer the one pose of person to another person or the same person in a different pose. The subjects can have different deformable objects in the foreground and the background, thus making the model difficult to learn. It is also challenging to learn the pose and the clothing details simultaneously. Isola et al. (Isola et al., 2017) proposed a conditional GAN for image-to-image translation, where a given scene representation from one domain is translated into another representation.

Recently Siarohin et al. (Siarohin et al., 2018) proposed a person image generation method conditioned on a given image of the person and the novel (target) pose of the person to synthesize the new image of that same person in the novel pose. Jetchev et al. (Jetchev and Bergmann, 2017) proposed the Conditional Analogy Generative Adversarial Network (CAGAN), which learns to swap the clothing of the person and paint realistically looking images with a target cloth article, given pairs of humans and clothes. In contrast, in our case, we focus on human image generation along with pose and clothing. Neverova et al. (Neverova et al., 2018) adopted DensePose (Alp Güler et al., 2018) as its pose representation for human pose transfer. Ma et al. (Ma et al., 2017) proposed a more general approach to synthesize person images in any arbitrary (random) pose. Similarly, a conditioning image of the person and a target new pose defined by 18 joint locations is the input to our proposed models. The generation process is divided into two different stages as pose generation and texture refinement. Horiuchi et al. (Horiuchi et al., 2019) addressed the problem of human image generation by using deformable skip connections, self-attention, and spectral normalization in GAN. Inpainting modules are used in the recent models to achieve a considerable level of detail in both image resolution and texture of cloth. There are 3D clothing models which automatically captures real clothing to estimate body shape, pose, and to generate new body shapes. Specifically, we have U-Net based architectures that are commonly used for pose-based person-image generation tasks (Ma et al., 2017), (Lassner et al., 2017). Because of local information in input and the output images is not aligned, the skip connections are not well-suited to handle large spatial deformations. Contrary to this, the proposed models use deformable skip connections to deal with this misalignment problem and to share the local information from the encoder to the decoder.

3 PROPOSED MODELS

We address the problem of transferring the person's appearance from a given pose to the desired target pose using a Deformable GAN (Siarohin et al., 2018) architecture with Spectral Normalization (SN) and adversarial hinge loss in the generator and discriminator. We also use warping in the discriminator and skip connections in the generator. We discuss the effect of architectural changes on the performance of human-image generation by comparing different variants.

3.1 Approach

With the inspiration from neural image synthesis (Isola et al., 2017), the proposed human-image generation technique uses Deformable GAN (DGAN) (Siarohin et al., 2018) architecture as the base model and has a similar architecture of DGAN as shown in Figure 3. We discuss in detail the different variants of our proposed architecture. In variant-1, the Spectral Normalization (SN) is integrated into both Generator G and Discriminator D. In the variant-2, the Generator G and Discriminator D losses are modified by adding the hinge adversarial loss along with Spectral Normalization (SN). In variant-3, the Generator G with skip connections, Discriminator D with Warping (W) along with RMSprop optimizer are used. In variant-4, all three variants are combined to observe their combined effect on the overall generation of the target image. The models need to preserve the appearance details like texture from the input image along with the pose information from the target pose. The model first extracts the pose of the person in the 2D skeleton with an HPE (Cao et al., 2017) model. All the four variants of the model are evaluated on the Market-1501 and DeepFashion datasets, and results are reported in the next section to show how the proposed model performs comparatively to existing state-of-the-art approaches PG^2 (Ma et al., 2017), DGAN (Siarohin et al., 2018) in conditional image synthesis.

3.2 Variants

3.2.1 Variant-1: Spectral Normalization (SN)

GANs are well known to be unstable during their training and more sensitive to the choice of hyper-parameters. One of the challenges in the training of GANs is controlling the performance of the discriminator. In higher dimensional spaces, the density ratio estimation by the discriminator is often inaccurate and unstable while training the model. Therefore, the generator networks fail to learn the multi-modal structure of the target data distribution. A novel weight normalization technique known as Spectral Normalization (SN) is used to stabilize the training of the generator and discriminator of GAN. SN has been one of the recent popular normalization techniques (Miyato et al., 2018), which stabilizes the training and avoids unwanted gradients preventing the parameters from exploding. To ensure that the generated image has better resolution without loss of texture details, we need to use SN (Miyato et al., 2018) in adversarial training. Spectral normalization (Miyato et al., 2018) normalizes the spectral norm of the weight matrix W of a convolution layer such that it satisfies the Lipschitz constraint $\sigma(W) = 1$.

$$\overline{W}_{SN}(W) = W/\sigma(W) \tag{1}$$

where $\sigma(W) = u^T W v$. For each layer, the vectors u and v are randomly initialized. So, it replaces every weight W by $W/\sigma(W)$, and the task is to compute the value of $\sigma(W)$ efficiently. By applying singular value decomposition naively at each step to compute $\sigma(W)$ might be computationally expensive. So, the power iteration method is used to estimate the $\sigma(W)$ spectral norm of each layer. The value of Lipschitz constant is the only hyperparameter that needs to be tuned, and the model does not require intensive tuning of the hyperparameter for improving performance. This technique is computationally cheap and easy to incorporate into existing implementations. Spectrally Normalized GANs (SN-GANs) (Miyato et al., 2018) are capable of generating images of better or equal quality in comparison to the previous training stabilization techniques. In the proposed model, for a given input image of a person and a target pose, DGAN (Siarohin et al., 2018) initially extracts the pose of the person in the form of a 2D skeleton with a Human Pose Estimator (HPE) (Cao et al., 2017). The data is then processed with a fully convolutional network encoder to obtain a feature representation of the input image, the extracted pose, and the target pose. Later, using the pose information for each specific body part, an affine transformation is computed and applied to move the feature-map content corresponding to that body part. The spectral normalization is used in both the encoder and decoder of the discriminator and the generator networks, as shown in Figure 3. It improves the results by decreasing the degradation of the error signal during back-propagation. As the weights change slowly, only a single power iteration for each step of learning needs to be performed. Therefore, spectral normalization for GANs is more computationally efficient than other regularization techniques such as weight clipping (Arjovsky et al., 2017) and gradient penalty (Gulrajani et al., 2017).

3.2.2 Variant-2: Spectral Normalization with Hinge Loss (SN + H)

A discriminator is trained with conditional adversarial losses to classify whether the given input is real or fake with negative log-likelihood. Adversarial hinge losses are used to optimize the probability that a given real data is realistic than a randomly sampled generated (fake) data and vice versa. It leads to more stable and robust training. The variant-2 makes use of an additional loss called as adversarial hinge loss. The adversarial hinge losses for the Generator G and Discriminator D can be calculated as follows:

$$\mathcal{L}^{\mathcal{H}}{}_{G}(G,D) = \mathbb{E}_{x\in X_I} \max(0, 1 - D(\hat{x})) + \mathbb{E}_{x^*\in X_T} \max(0, 1 + D(x^*)) \quad (2)$$

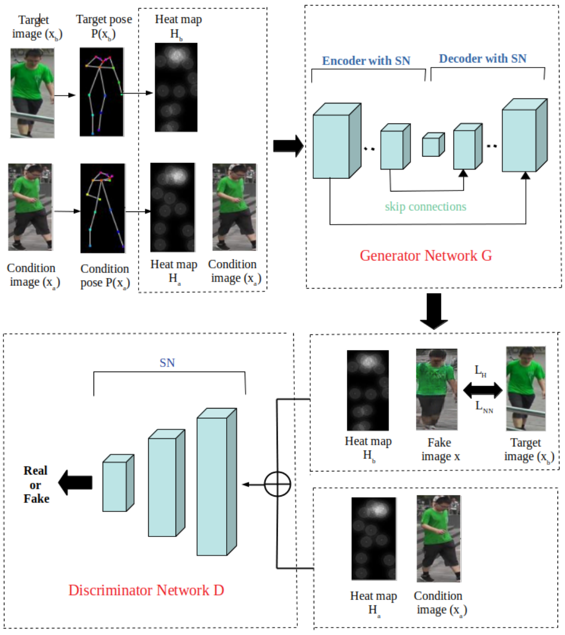

Figure 3: Simple architecture of the proposed model. SN represents Spectral Normalization after each Convolutional layer. L_H : adversarial Hinge loss, L_{NN} : Nearest Neighbour loss.

$$\mathcal{L}^{\mathcal{H}}{}_{D}(G,D) = \mathbb{E}_{x^*\in X_T} \max(0, 1 - D(x^*)) + \mathbb{E}_{x\in X_I} \max(0, 1 + D(\hat{x})) \quad (3)$$

These adversarial hinge loss functions are integrated into the generator and discriminator of the model along with nearest-neighbour loss and minimize the overall objective function. The relativistic hinge loss (Jolicoeur-Martineau, 2018), which modifies the output of the discriminator is not used in our proposed model. Relativistic adversarial losses (Jolicoeur-Martineau, 2018) optimize the probability that a given real data is highly realistic than randomly generated fake data and vice versa. The relativistic average hinge loss used in (Horiuchi et al., 2019) leads to a min-min problem instead of the normal min-max problem of adversarial training. The proposed model is trained with spectral normalization along with adversarial hinge loss on the Market-1501 and DeepFashion datasets.

3.2.3 Variant-3: Discriminator with Warping (W)

In variant-3, the architecture of the Discriminator D is modified. The discriminator uses an affine transformation layer between the first two convolution layers. We performed an ablation study to analyze the various effects of the different components of the proposed model. We used our proposed model for evaluating all the methods by amputating the parts of the full model similar to the base model (Siarohin et al., 2018).

The qualitative and quantitative results are found to be better than the benchmark model (Siarohin et al., 2018) by using this architecture for Discriminator D, which are reported in the next section.

3.2.4 Variant-4: Spectral Normalization with Hinge Loss and Discriminator with Warping (SN + H + W)

In variant-4, the variants SN + H and W are combined in order to know the effectiveness of these variants for the human-image generation. The use of SN with adversarial hinge loss stabilizes the training, whereas the use of discriminator with Warping (W) in the Full model helps to generate the human-images with appearance details close to the target image. The model is trained with both the Nearest-neighbour L_{NN} and adversarial Hinge L_H losses. The SN is used after every convolution and fully connected layers in both the Generator G and Discriminator D. The layers in the Generator G and Discriminator D are the same as in the variant-3 with SN and adversarial hinge loss. The model is trained on both the Market-1501 and DeepFashion datasets. Both the qualitative and quantitative results of the model, in comparison to other variants and benchmark models DGAN (Siarohin et al., 2018), PG^2 (Ma et al., 2017) have been described in detail in the next section.

3.3 Loss Functions

Recently with advances in Generative Adversarial Framework, different loss functions are used to tackle different optimization problems. The Generator G and Discriminator D are trained using the standard conditional adversarial loss $\mathcal{L}_{cGAN}$ along with the nearest-neighbour loss $\mathcal{L}_{NN}$ similar to the base model (Siarohin et al., 2018).

We have integrated the hinge adversarial loss to the variant-2. We train the model using the nearest-neighbour loss $\mathcal{L}_{NN}$ in conjunction with the adversarial hinge loss $\mathcal{L}^H$ by minimizing the final objective function as given below:

$$G^* = arg \min_G \max_D \mathcal{L}_{CGAN}(G,D) + \lambda \mathcal{L}_{NN}(G) + \mathcal{L}^{\mathcal{H}}{}_G(G,D) + \mathcal{L}^{\mathcal{H}}{}_D(G,D) \quad (4)$$

where we use λ as 0.01 in all our experiments. Based on the behavior of λ value leading to the generation of artifacts and blurry results, the value of λ is kept small (Ma et al., 2017). In detailed description of $\mathcal{L}_{NN}$ and other losses can be found in (Siarohin et al., 2018).

4 EXPERIMENTS AND RESULTS

In this Section, we present both quantitative and qualitative evaluation results on publicly available Market-1501 and DeepFashion datasets. We also compare our proposed models with other state-of-the-art methods. We describe in detail about the datasets, evaluation metrics used along with the implementation details for the proposed model.

4.1 Datasets

We used the most commonly available public datasets for humans like Market-1501 (Zheng et al., 2015) and DeepFashion (Liu et al., 2016b) datasets. **Person Re-identification dataset: Market-1501** This dataset contains images of 1,501 persons captured from 6 different surveillance cameras totaling to 32,668 images. It is a very challenging dataset as it contains low-resolution images of size 128 × 64 with the high diversity in illuminations, poses, and backgrounds. The data is divided into train and test sets with 12,936 and 19,732 images, respectively. In the train set, we have 439,420 pairs, each of which is composed of images of the same person with different poses. Then, randomly 12,800 pairs are selected from the test set for testing. **The DeepFashion dataset (In-shop Clothes Retrieval Benchmark)** comprises of 52,712 clothes images, leading to 200,000 pairs of same clothes with two different poses and scales of the persons wearing these clothes. These images have a resolution of 256 × 256 pixels, and the images are less noisy compared to the Market-1501 dataset. The dataset contains images with no background, where the clothing patterns can contain text, graphical logos, etc. Randomly 12,800 pairs are selected from the test set for testing.

4.2 Evaluation Metrics

For benchmarking the performance of the proposed model, widely-used evaluation metrics for human image generation like Structural Similarity (Wang et al., 2004) (SSIM), Inception Score (IS) (Salimans et al., 2016) (based on the entropy of classification neurons), m-SSIM (masked SSIM), m-IS (masked IS), (Detection Scores) DS scores are calculated. Though these measures are widely accepted, we like to show the qualitative results of the generator with adversarial training that are visually appealing compared to present state-of-the-art approaches. The masked scores are obtained by masking out the background of the images and feeding it to the Generator G. Another metric DS (Detection Score) that is based on the detection outcome of the object detector SSD (Single Shot multi-box Detector) (Liu et al., 2016a) is used. It is trained without fine-tuning to the datasets on challenging Pascal VOC 2007 (Everingham et al., 2007). During testing, scores of SSD are computed on each generated image and averaging the SSD score of each generated image; the final DS is obtained. Therefore, DS gives the confidence that a person is present in the image.

4.3 Implementation Details

In all the four variants of the proposed model, both the *G* and *D* are trained with the RMSprop optimizer (learning rate:0.0002 and ρ:0.9). Because of the higher resolution of the DeepFashion dataset, the generator for the DeepFashion dataset of all the four variants has one extra convolution block of 512 kernels and a stride of 2 both in encoder and decoder. The ReLU of the last layer in Discriminator *D* for all the four variants is replaced with the sigmoid activation function. The dropout is used only at the time of training.

4.4 Quantitative Results

We present the results of the ablation study for the variant-3 (W) on Market-1501 and DeepFashion datasets. We also show the comparison of the four variants of our proposed model with other state-of-the-art models DGAN (Siarohin et al., 2018), PG^2 (Ma et al., 2017).

4.4.1 Ablation Study

We have four different methods in variant-3, which are obtained by removing the parts from the Full method described earlier to see the impact of each part. The architecture of the Discriminator *D* includes warping and is the same for all the four methods. The results in Table 1 report the evaluation metrics like SSIM, IS, m-SSIM, m-IS for the Market-1501 dataset and SSIM, IS, DS scores for the DeepFashion dataset. Table 1 gives the results in comparison to the base model (Siarohin et al., 2018). Table 1 gives the respective scores of the base model (Siarohin et al., 2018) and the proposed model for the four different methods. From Table 1, the scores show that there is a significant progressive improvement from the Baseline method to the Full method. We can observe that a combination of all the components gives a large boost in IS and improves SSIM scores. The bold scores represent the highest values obtained for that particular score in comparison with all the models. For example, on the Market-1501 dataset, the SSIM value observed is 0.293, which is highest when compared to all the other SSIM scores of 4 different methods. The highest scores for all the four methods in the case of DeepFashion are obtained by the proposed model when compared with their respective methods, as reported in Table 1. The DS scores for all four different methods on DeepFashion are 0.97, which are close to real data scores that are computed on ground truth images from datasets.

4.4.2 Comparison of the Model with Other State-of-the-Art Models

All the four variants of the proposed model are compared with the other state-of-the-art models like PG^2 (Ma et al., 2017) and Siarohin et al. (Siarohin et al., 2018) in Table 2. The scores for model PG^2 (Ma et al., 2017) were taken from the benchmark paper (Siarohin et al., 2018) that are computed using the code and the network weights released by PG^2 model (Ma et al., 2017). From Table 2, the variant SN has improved the IS scores when compared to Deformable GAN (DGAN) model proposed by Siarohin et al. (Siarohin et al., 2018), whereas the SSIM scores are slightly less. Therefore, the use of SN in the generator and discriminator has increased the IS and m-IS scores, whereas the SSIM scores have slightly decreased. When compared to the PG^2 model (Ma et al., 2017), except for the IS score, all the other scores have been improved. These scores show that the variant SN has outperformed the PG^2 model.

The second variant, SN + H, has outperformed the DGAN model (Siarohin et al., 2018), whereas only the IS score was less than the PG^2 model (Ma et al., 2017). It can be noticed that the use of SN has significantly increased the DS scores in the first and second variants. The inclusion of adversarial hinge loss along with SN in the GAN framework has improved the results, and this proves the importance of adversarial hinge loss for the human-image generation. The third variant (W), which represents the Full model of the ablation study, has the highest scores than the DGAN model (Siarohin et al., 2018) except for the DS score. In comparison to the PG^2 model (Ma et al., 2017), this variant reports the highest performance than the benchmark PG^2 model with all metrics except the IS metric. Conversely, on the DeepFashion dataset, the full model significantly improves the IS and DS values than the other two benchmark models PG^2 (Ma et al., 2017), DGAN (Siarohin et al., 2018) but returns a slightly lower SSIM value. The fourth variant, which is SN + H + W has obtained better quantitative results than PG^2 model (Ma et al., 2017) except for the IS metric on the Market-1501 dataset. In the case of the DeepFashion dataset, it has significantly

Table 1: Quantitative results: Ablation study results on the Market-1501 and the DeepFashion datasets for the variant-3 of the proposed model. The best results are highlighted in bold. For all the measures, higher is better.

Models	**Market-1501**					**DeepFashion**		
	SSIM	IS	m-SSIM	m-IS	DS	SSIM	IS	DS
Baseline (Siarohin et al., 2018)	0.256	3.188	0.784	3.580	0.595	0.754	3.351	0.96
Ours-Baseline	0.265	3.232	0.787	3.631	0.625	0.759	3.452	**0.97**
DSC (Siarohin et al., 2018)	0.272	**3.442**	0.796	3.666	0.629	0.754	3.352	0.96
Ours-DSC	0.277	3.433	0.797	**3.684**	0.599	**0.764**	3.385	**0.97**
PercLoss (Siarohin et al., 2018)	0.276	3.342	0.788	3.519	0.603	0.744	3.271	0.96
Ours-PercLoss	0.279	3.318	0.802	3.537	0.671	0.762	3.400	**0.97**
Full (Siarohin et al., 2018)	0.290	3.185	**0.805**	3.502	**0.720**	0.756	3.439	0.96
Ours-Full	**0.293**	3.354	**0.805**	3.540	0.571	0.759	**3.584**	**0.97**
Real-Data	1.00	3.86	1.00	3.36	0.74	1.000	3.898	0.98

Table 2: Quantitative results: Comparison of four variants of the proposed model with other state-of-the-art models on Market-1501 and the DeepFashion datasets. SN represents Spectral Normalization, and H represents Adversarial Hinge loss. For all the measures, higher is better.

Models	**Market-1501**					**DeepFashion**		
	SSIM	IS	m-SSIM	m-IS	DS	SSIM	IS	DS
(Ma et al., 2017)	0.253	**3.460**	0.792	3.435	0.39	0.762	3.090	0.95
(Siarohin et al., 2018)	0.290	3.185	**0.805**	3.502	**0.72**	0.756	3.439	0.96
Ours SN	0.280	3.300	0.797	3.528	0.64	**0.764**	3.461	**0.97**
Ours SN + H	0.291	3.239	0.804	**3.592**	0.69	0.763	3.444	**0.97**
Ours W	**0.293**	3.354	**0.805**	3.540	0.57	0.759	**3.584**	**0.97**
Ours SN + H + W	0.291	3.192	**0.805**	3.551	**0.72**	0.756	3.473	**0.97**
Real-Data	1.00	3.86	1.00	3.36	0.74	1.000	3.898	0.98

higher values for IS and DS than the PG^2 model (Ma et al., 2017). When compared to the DGAN model (Siarohin et al., 2018) on the Market-1501 dataset, the values for SSIM, m-SSIM are the same for both the models. Whereas all the other scores are higher for Siarohin et al. (Siarohin et al., 2018) on the Market-1501 dataset. It can be seen that the DS score has increased rapidly by using SN + H along with the Full model of variant-3 (W) when compared to all the other variants. On the DeepFashion dataset, the values for IS and DS are higher for SN + H + W when compared to the benchmark models PG^2 (Ma et al., 2017), DGAN (Siarohin et al., 2018). The fourth variant SN + H + W obtained a higher score for IS than all the other models except for the Full model. Apart from this, the proposed model has reported better results than the two benchmark models PG^2 (Ma et al., 2017), DGAN (Siarohin et al., 2018).

From Table 3, the comparison of the two variants with the state-of-the-art model proposed by (Horiuchi et al., 2019) can be seen. The scores SSIM and L_1 are computed based on the generated and ground truth images. It shows that the SSIM values are slightly lower than the SSIM scores of the variants SN and SN + RH proposed by (Horiuchi et al., 2019). The IS scores are comparatively higher than the proposed model (Horiuchi et al., 2019). It can be observed that the reported IS is 3.300 for SN, and that of the benchmark model (Horiuchi et al., 2019) is 3.066. In the case of SN with adversarial hinge loss, the value of IS has been 3.239, which is higher than the SN + RH variant of Horiuchi et al. (Horiuchi et al., 2019).

Table 3: Quantitative results: Comparison of variants SN and SN + H with the state-of-the-art model (Horiuchi et al., 2019) on the Market-1501 dataset. SN: Spectral Normalization, H: Adversarial Hinge loss, RH : Relativistic Hinge Loss. For all the measures, higher is better.

Models	**Market-1501**		
	SSIM	IS	L_1
Horiuchi et al.-SN	0.289	3.066	0.289
Ours SN	0.280	**3.300**	0.289
Horiuchi et al.-SN + RH	**0.296**	2.973	**0.288**
Ours SN + H	0.291	3.239	**0.288**
Real-Data	1.00	3.86	0.00

4.5 Qualitative Results

Not only the quantitative measures are enough to show how good is the model, but the qualitative results are equally important to see how realistic they are from a human point of view.

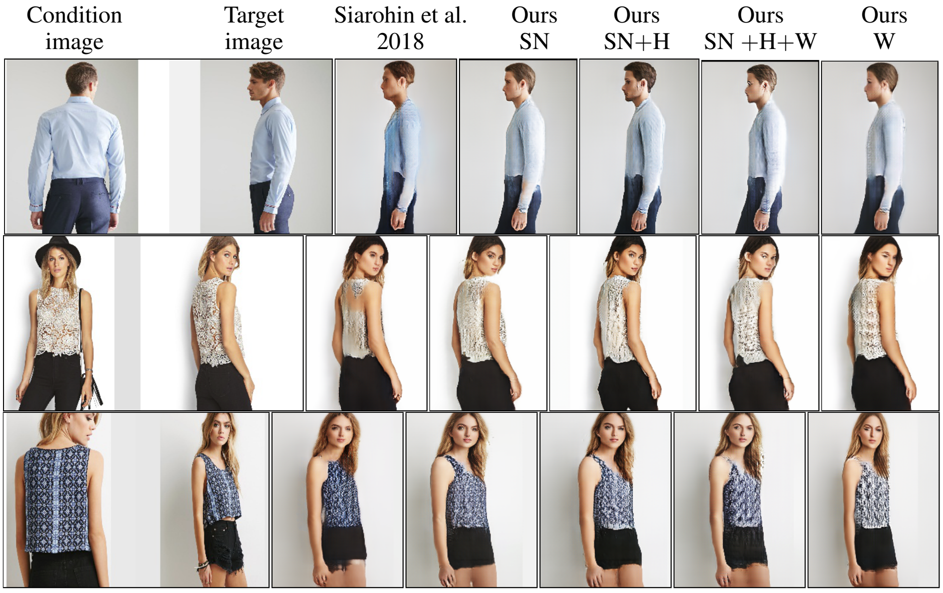

Figure 4: Qualitative results: Comparison of all the four variants SN, SN + H, SN + H + W, W and the benchmark model (Siarohin et al., 2018) on the DeepFashion dataset.

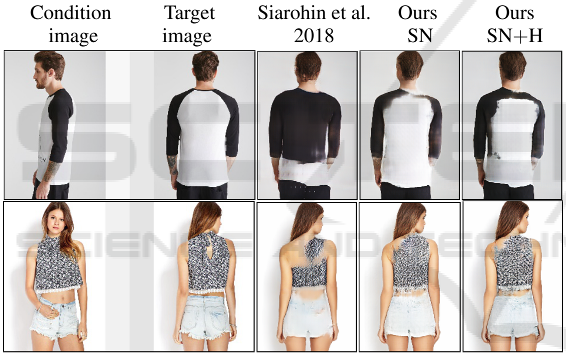

Figure 5: Qualitative results: Comparison of some of the challenging results for SN, SN + H and the benchmark model (Siarohin et al., 2018) on the DeepFashion dataset.

4.5.1 Comparison of Four Variants with Benchmark Model (Siarohin et al., 2018): DeepFashion

From Figure 4, the comparison of the qualitative results for all the four variants SN, SN + H, SN + H + W, W with the benchmark model (Siarohin et al., 2018) on the DeepFashion dataset can be found. The results from the four variants are qualitatively better and sharper than the benchmark model (Siarohin et al., 2018).

In the first row, the result from Siarohin et al. (Siarohin et al., 2018) lacks the texture details, and the generated image is blurry, whereas the result from the four variants looks sharper in texture and appearance details. Even the color of the shirt is similar to that of the target image for all the four variants. In the third and fourth rows, the generated results from the benchmark model (Siarohin et al., 2018) miss the texture details and output the empty spaces when compared to our proposed four variants, which produce crisp appearance details. This shows the inclusion of SN, which has generated realistic images with sharper texture and appearance details.

4.5.2 Comparison of SN and SN+H with Benchmark Model (Siarohin et al., 2018): DeepFashion

Figure 5 shows the qualitative comparison of the results for the benchmark model (Siarohin et al., 2018) and the first two variants SN, SN + H on the DeepFashion dataset. The images in Figure 5 represent the challenging images where (Siarohin et al., 2018) failed to generate the images which look similar to the target image. In the first row, the benchmark model (Siarohin et al., 2018) generates the image with a black patch for the shirt instead of generating it only for the hands, as seen in the images from the SN and SN + H. In the second row, the texture of the cloth is generated with gaps. In contrast, the variants SN and SN + H generate the texture similar to the target image. These results show how our variants are qualitatively better than the benchmark model (Siarohin et al., 2018).

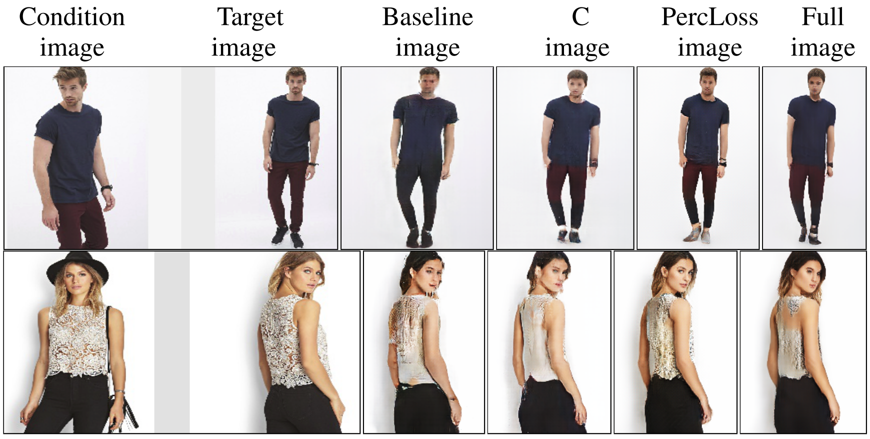

Figure 6: Qualitative results: Ablation study results of Baseline, DSC, PercLoss, Full methods of the benchmark model (Siarohin et al., 2018) on the DeepFashion dataset.

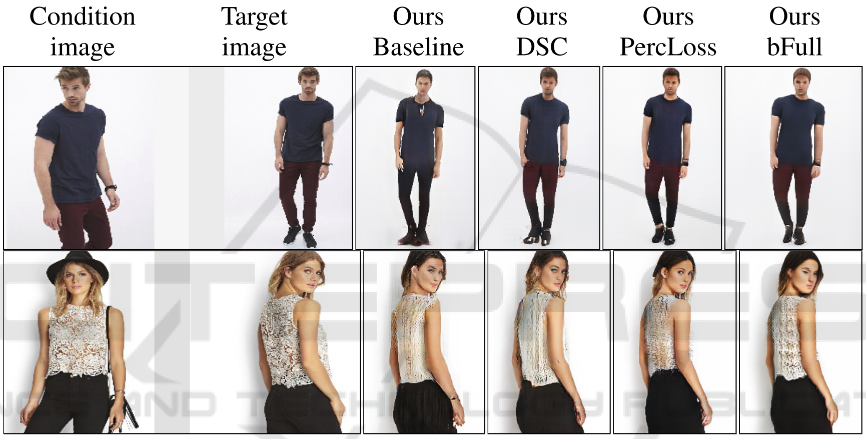

Figure 7: Qualitative results: Ablation study results of Baseline, DSC, PercLoss, Full methods of variant-3 of the proposed model on the DeepFashion dataset.

4.5.3 Ablation Study Results: DeepFashion Dataset

Figures 6 and 7 depict the qualitative results of the ablation study for the benchmark model (Siarohin et al., 2018) and the variant-3 of the proposed model, respectively, on the DeepFashion dataset. In both Figures, there is a progressive improvement of the results from Baseline to the Full model. When the Baseline results are compared from both Figures, the baseline model (Siarohin et al., 2018) does not capture all the appearance details of the target. The clothing texture is blurred for the baseline model (Siarohin et al., 2018). In contrast, the images generated by the Ours-Baseline model in Figure 7 are better in texture and appearance details than the Baseline model (Siarohin et al., 2018). The results of DSC are better when compared to Baseline as the images are sharper, and texture details are improved in both Figures 6 and 7. When the DSC results are compared in both Figures 6 and 7, the facial features of the Ours-DSC model appears to be sharper. In PercLoss, the results are more refined than the DSC model. The results from the Ours-DSC model show the generation of shoes which are missing in DSC (Siarohin et al., 2018). In most of the cases, the results obtained by the Full model are better than the PercLoss model in both Figures 6 and 7. The generated results for the Ours-Full model are better than the Full model (Siarohin et al., 2018), which can be seen from the texture generated for the images in the last row.

4.5.4 Comparison of Four Variants with Benchmark Model (Siarohin et al., 2018): Market-1501

Figure 8 shows the comparison of the results between the four variants SN, SN + H, SN + H + W, W, and the Siarohin et al. (Siarohin et al., 2018) model.

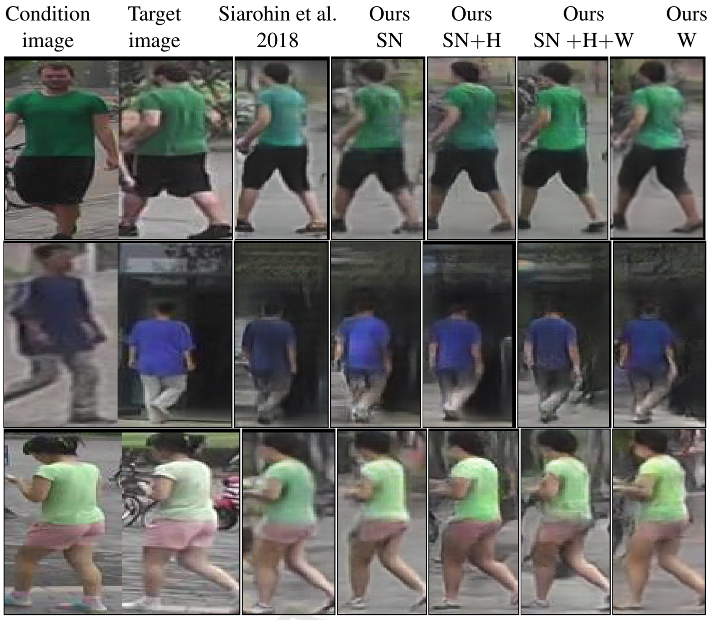

Figure 8: Qualitative results: Comparison of all the four variants SN, SN + H, SN + H + W, W and the benchmark model (Siarohin et al., 2018) on the Market-1501 dataset.

As already known that the images in the Market-1501 dataset have challenging backgrounds and are of low-resolution, the generated images are still blurry. From the images, the results of Siarohin et al. (Siarohin et al., 2018) differ in the color of the shirt from the target image. In contrast, the variants with SN generate the color appropriate to the target. The background of the generated images for the SN variants seems to be better than the base model (Siarohin et al., 2018). However, the results generated are still not sharper in facial features and the texture details in comparison to the target image when the Baseline results are compared from both the Figures.

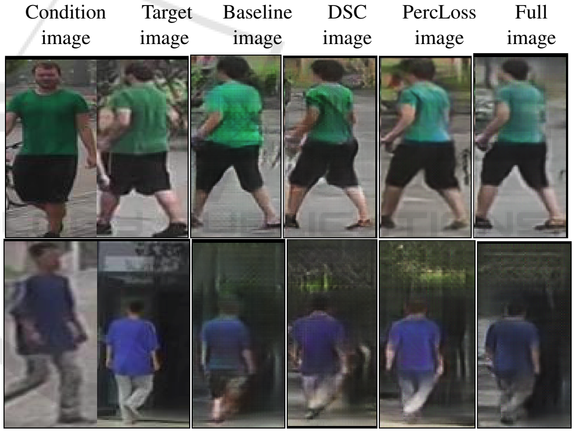

Figure 9: Qualitative results: Ablation study of Baseline, DSC, PercLoss, Full methods of the benchmark model (Siarohin et al., 2018) on the Market-1501 dataset.

4.5.5 Ablation Study Results: Market-1501 Dataset

Figures 9 and 10 depict the qualitative results of the ablation study for the benchmark model (Siarohin et al., 2018) and the variant-3 of the proposed model respectively on the Market-1501 dataset. In both Figures, there is a progressive improvement of the results from Baseline to the Full model. The baseline model (Siarohin et al., 2018) does not capture all the texture detail of the target. This can be observed from the first and second rows of Figure 9 as the color of the shirt has not been generated correctly. The results of DSC are better when compared to Baseline as the appearance details are improved in both Figures 9 and 10. When the DSC results are compared in both Figures 9 and 10, the appearance details, including the color of the clothing of the Ours-DSC model, appear to be better as seen in the images. In PercLoss, the results are more refined than the DSC model, but since the improvements are minor, it is very challenging to differentiate the images.

In most of the cases, the results obtained from the Full model are better than the PercLoss model in both Figures 9 and 10, and the pose information is generated well by all the methods. The generated results of the Ours-Full model are better than the Full model (Siarohin et al., 2018), which can be observed from the generated images in Figure 10. From the condition image in the second row, the image is too blurry, which gives vague information for training the model.

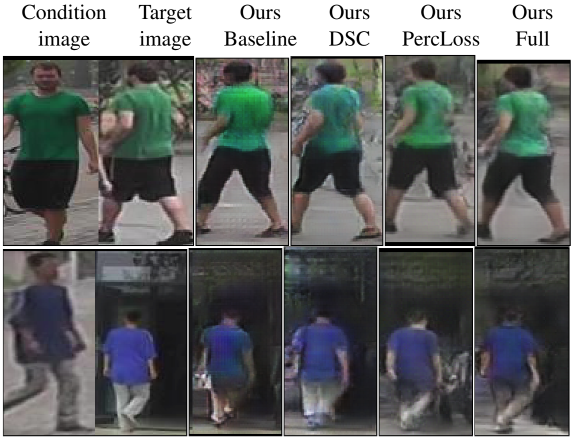

Figure 10: Qualitative results: Ablation study of Baseline, DSC, PercLoss, Full methods of variant-3 of the proposed model on the Market-1501 dataset.

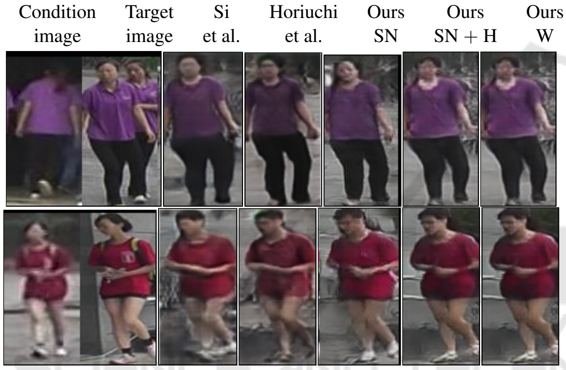

Figure 11: Qualitative results: Results of SN, SN+H, W variants of the proposed model in comparison with the benchmark models (Siarohin et al., 2018), (Horiuchi et al., 2019) on the Market-1501 dataset. Here Si et al. represents (Siarohin et al., 2018).

4.5.6 Comparison of SN, SN+H and W with Benchmark Models (Siarohin et al., 2018), (Horiuchi et al., 2019): Market-1501

Figure 11 shows the qualitative results of Siarohin et al. (Siarohin et al., 2018), Horiuchi et al. (Horiuchi et al., 2019), SN, SN + H, and W models. The clothing color and the shoes of the person in the second row seems to be not generated correctly by Siarohin et al. (Siarohin et al., 2018) and Horiuchi et al. (Horiuchi et al., 2019) when compared to the target image.

5 CONCLUSION

We presented the conditional generative models which exploit the power of deep neural networks for transferring various poses from one person to the same person along with appearance and clothing texture details. The model uses the novelties like Spectral Normalization (SN) and adversarial hinge loss. The use of SN would stabilize the training of GANs and helps to generate images with high quality. The use of adversarial hinge loss has shown to improve the generation of images with sharper details. This model has been divided into four variants as a part of the ablation study to know the importance of each component of the model for the human-image generation. We showed how the four variants differ from each other both qualitatively and quantitatively on the Market-1501 and DeepFashion datasets. Without any data augmentation, the proposed models converge faster and also generalize well to never seen test data. We showed how the proposed model outperforms recent state-of-the-art models for person image generation with pose and clothing details by providing benchmark results on publicly available Market-1501 and DeepFashion datasets. It can be concluded that the previous state-of-the-art models generated the results with loss of appearance and clothing details based on our experiments. We have also reported results that depict that the proposed models have a qualitative improvement over other methods. We presented that generative modeling with SN for person image generation conditioned on pose and appearance in a supervised setting provides good generalization capabilities.

5.1 Future Work

There are many models for GANs which are designed to tackle the problem of human image generation along with pose and clothing. Different models focus on different pose representations, but the combination of a good pose representation along with objective function would improve the person-image generation. In future work, we would like to focus on using other pose representations like 3D keypoints along with existing Spectral Normalization (SN) and hinge loss, which we believe would make better improvements in this area of research. Also, self-attention (Zhang et al., 2018) modules could be integrated into the GAN architectures along with different pose representations for the person-image generation. These attention modules would attend to all the important features of the person to generate accurate appearance details with high quality. We can use a multi-stage model with SN and attention mechanisms to further improve the human-image generation.

REFERENCES

Alp Güler, R., Neverova, N., and Kokkinos, I. (2018). Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299.

Chen, X., Song, J., and Hilliges, O. (2019). Unpaired pose guided human image generation. *CoRR*, abs/1901.02284.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2007). The pascal visual object classes challenge 2007 (voc2007) results.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Horiuchi, Y., Iizuka, S., Simo-Serra, E., and Ishikawa, H. (2019). Spectral normalization and relativistic adversarial training for conditional pose generation with self-attention. In *2019 16th International Conference on Machine Vision Applications (MVA)*, pages 1–5. IEEE.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Jetchev, N. and Bergmann, U. (2017). The conditional analogy gan: Swapping fashion articles on people images.

Jolicoeur-Martineau, A. (2018). The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Lassner, C., Pons-Moll, G., and Gehler, P. V. (2017). A generative model of people in clothing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 853–862.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016a). Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer.

Liu, Z., Luo, P., Qiu, S., Wang, X., and Tang, X. (2016b). Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104.

Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., and Van Gool, L. (2017). Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 406–416.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*.

Neverova, N., Alp Guler, R., and Kokkinos, I. (2018). Dense pose transfer. In *The European Conference on Computer Vision (ECCV)*.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242.

Siarohin, A., Sangineto, E., Lathuilière, S., and Sebe, N. (2018). Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3408–3416.

Stewart, M. (2019 (accessed May 8, 2019)). *Advanced Topics in Generative Adversarial Networks (GANs)*. https://towardsdatascience.com/comprehensive-introduction-to-turing-learning-and-gans-part-2-fd8e4a70775.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807.

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., et al. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. (2018). Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124.