Multi-View Human Image Synthesis

Ziqi Pang
1600013029
pangziqi@pku.edu.cn

Yichen Xie
1500012720
1500012720@pku.edu.cn

Abstract

In this paper, we address the task of multi pose to novel pose human image synthesis, where we aim at synthesizing a target human image in arbitrary pose from multiple given source images. The pipeline proposed in this paper makes separate predictions corresponding to each source image, then combines them together for the final result. The components of our method includes: (1)geometric transformations for generating coarse predictions directly from source images; (2) a refine network making predictions from the previous stage into realistic and valid human while retaining the details. (3) define and compute our preference over each intermediate prediction and merge them together. The experiments and evaluations are conducted on Market-1501[25], and the results demonstrate that our model is able to generate realistic human images while maintaining the details of the person. To the best of our knowledge, we are the first to tackle multi pose to novel pose problem.

1. Introduction

Generating realistic images of humans is of great importance in computer vision and has broad applications in different tasks, like image editing, person re-identification, in-painting, etc. The recent progress of image generation models such as variational autoencoders(VAE)[12], generative adversarial networks(GANs)[8][19] has enlightened great progress in this area[15][14][1][16]. In these previous work, researchers have paid great effort in solving the problem of synthesizing with only one source image. But what about more than one source images?

From the perspective of performance, however, synthesizing from multiple source images logically yields better results than from single source image, as more information is provided. To better demonstrate the its advantage, it's helpful to think about a situation in which we try to synthesize a full human image based on two source images each corresponding to a different view. Methods using both of the two source images will make up for the unseen area in one image from the other image, thus lead to predictions with high fidelity and confidence. But the methods using only one source image will always have to purely guess what the other side looks like. From the perspective of application, with multiple source images being available in many cases, it's also wasteful to use single source image. Moreover, the model able to work with multiple source images also have the ability to work with single source image, which means it will be endowed with better flexibility in this way. Therefore, we try to go beyond the constraints of "one source image" and explore the ways to synthesis human images from multiple source images, which is called multi-pose to novel pose synthesis.

The challenge of this task lies in merging the information from different information sources. The problem of novel view synthesis has already been studied, but mostly in the case of rigid objects and not solved till now. A great amount of effort have been expended in geometry-based methods aiming to directly estimate the underlying 3D structures exploiting the knowledge geometry[20][4]. Although this kind of methods are successful with rigid objects, they are unable to work with non-rigid objects like humans due to the large variety and changing of 3D structure and surface. Learning-based methods rise recently[26][21], and they still focus on rigid objects. These work can be vaguely divided into two categories: pixel generation and flow prediction. However, both of the ideas face difficulties in human image generation. Pixel generation directly regresses the pixel value and lacks the awareness of the structure of human body, while flow prediction predicts the location for each target image patch on source image patch and lacks the ability to generate regions that are not present in the source images. In this sense, methods exploiting both the unique structure of human and the power of generative models are needed in novel view human image synthesis.

To achieve this, we propose a pipeline as is demonstrated in figure 1, containing the following stages:

1. **Geometric Tran formation Stage**: Each human image is parsed into several parts according to the key points locations, then geometrically transformed onto target image corresponding to target key points locations. This stage takes human structure into account

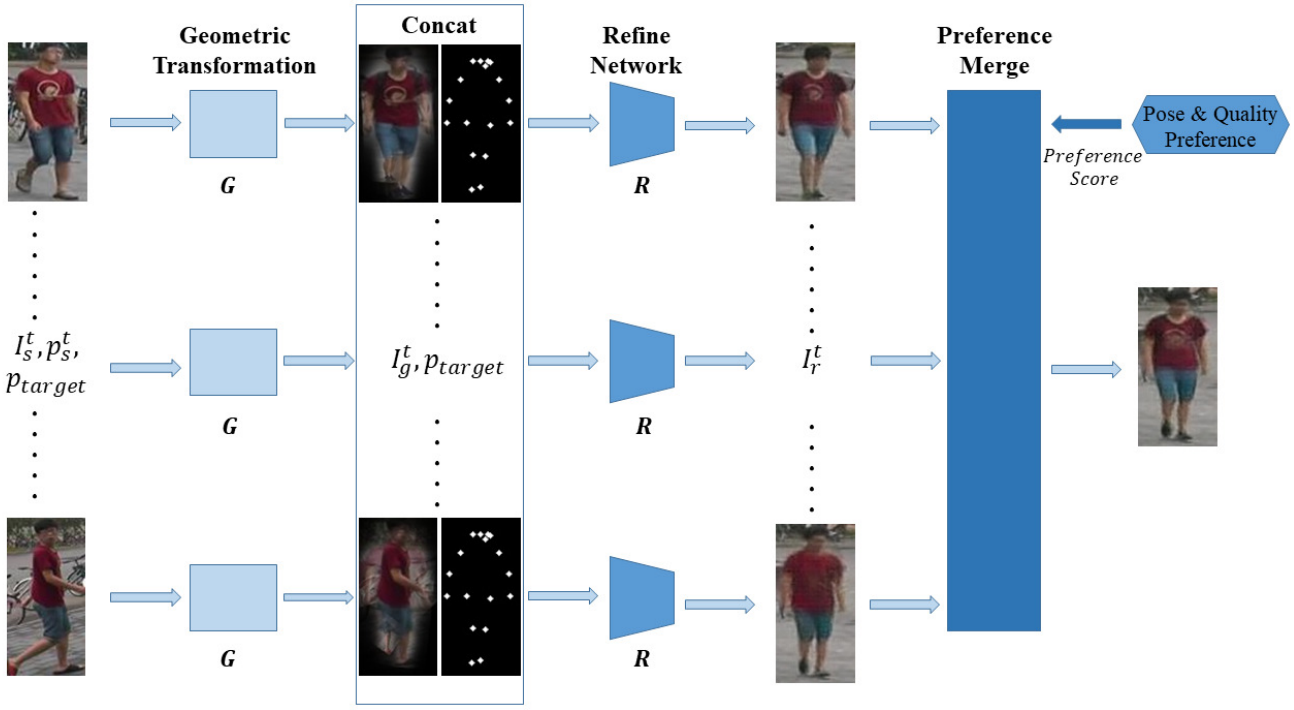

Figure 1. Pipeline Overview

by way of key points and parts.

2. **Refine Stage**: Given an original result from the first stage and target pose information, we exploit the power of generative convolutional neural network to refined the predictions. With deep convolutions in neural network having large scale of receptive fields, this refine network has the ability of propagating information among different parts in the image, thus being able to fix twisted areas and make up missing regions.

3. **Merging Stage**: In this stage we define preference score for further merging intermediate predictions from the previous stage. Generally speaking, we prefer parts (1) originally seen from the source images; (2)from images in high quality. As for(1), we use pose information to determine if one part on target image is visible on source images; as for two, we use a self-defined metric to judge the quality. Finally, we select a best intermediate prediction for each part and merge them together into the final result.

Our contributions can be summarized as follows: (1) Proposing a new task of generating human images according to multiple source images; (2) Pipeline leveraging the inherent structure of human, making up for the unseen parts and combining information from different sources; (3) Experiments showing the high quality of our generated images and the progressive improvement with more source images available; (4) Demonstration of the unique ability of our method to maintain details from source images.

2. Related Work

Novel view synthesis has close relationship with image generation. Image generation has experienced rapid progress in recent years. The most representative group of methods are generative adversarial networks[8][19][9][27][3]. These models use discriminators as supervisor to force the generator to generate realistic and sharp images. Other than that, variational aotoencoder(VAE)[12] also arises as a typical method on generation tasks, especially when conditional information is given. VAE has shown its strong power in conditional generation problems, and has been widely applied as a result[2][13][5]. Generally speaking, VAE encodes source information into latent variables, then recovers the full information in a decoder from latent variables. And it needs mentioning that, these two methods are not contradictory and usually used together, which endows the models with greater ability on synthesis and generation tasks.

Human image synthesis is a sub-area of image generation mentioned above. But it has received great attention these years. Instead of generating images of random humans, a more challenging and useful task was proposed[15]

and thoroughly studied, that is, synthesizing the person in a new pose given an arbitrary image of the same person. The difficulty of this task lies in the large variety of human poses and surfaces. [14] first propose a model generating people in clothing. Then [15] propose a coarse-to-fine two stage model in generative adversarial manner to solve human image synthesis. [24][16] both follow this method, but add modifications to make the process of generation more controllable and at the same time improve the performance. [5] thinks the other way, who considers the source image as information source of pattern and target pose as the shape for result, then they use VAE to encode these information into latent variables. In order to take human structure into consideration, [1] parses human into different parts and warps them before refinement. However, to the best of knowledge, none of the human image synthesis work has considered the setting of given multiple source images. And as is mentioned in introduction, we believe that this new problem is harder and fits into more applications at the same time.

The difficulty for multi-view synthesis lies in merging information from different sources, especially when neural networks haven't developed the ability of progressively understanding an object in multiple shots. Multi-view synthesis is an important task in computer vision, but most of the previous work focuses on rigid objects. [7] focusing on the multi-view stereo problem by regressing directly to output pixel values. [23][22] propose the methods to directly generate pixels of a target view. [17][26] predicts a flow to move the pixels from the source to the target view, followed by an image completion network. Recently, [21]propose a method combining directly generating and flow predicting together with a self-learned confidence, thus enlighten us to define preference score to merge different intermediate predictions.

3. Method

When synthesizing a novel pose from multiple source images, we want our model to (1) Directly use the information from source images so as to maintain as much details as possible; (2) Hallucinate missing information to improve the validity of the image; (3) Progressively improve the predictions when more source images are available. To realize these objectives, we design a pipeline containing three modules, which is shown in figure 1. To put (1) into practice, we use geometric transformations to move the parts into target locations. As for (2), we use a refine network to make images realistic. In the third stage, we aggregate the mediate predictions according to a self-defined preference score based on image quality and poses.

3.1. Overview and Notations

Our goal is to synthesize a target image I_{target} given a target pose p_{target} and N (image, pose) pairs $(I_s^1, p_s^1), (I_s^2, p_s^2), ..., (I_s^N, p_s^N)$. The pose of the images are represented as a vector, consisting the key points' locations in the images. We define the geometric transformation stage as function $G(\cdot)$. Given t-th source image and its corresponding pose (I_s^t, p_s^t), it generates a prediction $I_g^t = G(p_{target}, I_s^t, p_s^t)$, where I_g^t is the predicted image. After this, refine network is defined as function $R(\cdot)$, generating advanced prediction I_r^t in form of $I_r^t = G(p_{target}, I_r^t)$. In the last stage, we use a shallow fully connected network represented as function $P(\cdot)$, which outputs our preference towards each human body part on I_r^t by $preference^t = P(p_s^t, p_{target})$. Finally, the predictions $I_r^1, I_r^2, ..., I_r^N$ are aggregated according to their preference score.

3.2. Stage-1: Geometric Transformation

Geometric transformations are applied to move each body parts from source images into target locations according to target pose. So as to get the original pose information, AlphaPose[6] is used to extract the locations of key points. With the pose information, masks for human body parts are generated in heat map methods. The reason for us using heat map instead of human parsing is based on the following observation: heat map methods usually gives a slightly larger mask than the actual foreground, thus it seldom misses important information; however, current human parsing algorithms try to stick the mask to human silhouettes and lose essential information about clothes sometimes.

After extracting body parts from source images, coarse predictions are made by way of doing affine transformation on each part. The transformation matrixs are computed according to the source and target key points' locations. After these two steps, $I_g^1, I_g^2, ..., I_g^N$ are generated.

Geometric transformation is an attractive method in human image synthesis, even compared to neural networks methods[10]. Geometric transformation methods are not only easy to control and implement, but yields to details with high fidelity. The patches in predictions are directly sampled and moved from source images, and they can still maintain the details of source identity after transformation.

3.3. Stage-2: Refine Network

Due to the unavoidable error in pose estimation and oversimplicity in using affine transformations, images generated in stage1 suffer from the twisted textures, and mismatched body parts. Therefore, a refine network is necessary for such synthesis tasks. The input to the refine network are the predictions from previous stage $I_g^1, I_g^2, ..., I_g^N$ and their corresponding target pose. The output of the refine network are images $I_r^1, I_r^2, ..., I_r^N$.

The network structure is modified from PNGAN[18]. The modifications aim at: (1) maintaining the details from input images; (2) keeping the ability of neural network to

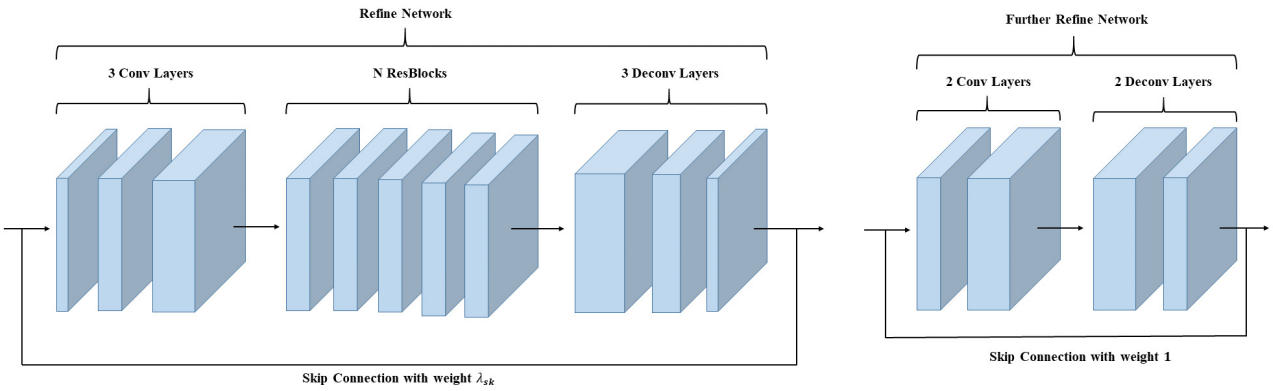

Figure 2. Refine Network Structure

compensate for missing areas. As for (1), we add a skip connection from input to output on the original PNGAN. However, as some of the images from stage 1 is highly invalid, we design the skip-connection to play a "guiding" role instead of "dominant role" to satisfy (2), so we set the weight of this skip connection to be λ_{sk} instead of one, which is 0.5 in the actual experiment. Despite of this, the output images still suffer from minor twisting and blurry, which is shown in figure 2. So we add another refine block for further refinement on top of modified PNGAN and these two make the whole refine network.

To generate sharp images, we use the training methods of generative adversarial networks, that is, leverage a discriminator to improve the performance of refine network. The loss functions for training refine network are three. The first evaluates the difference between predictions and ground truth target images.

$$L_{GT} = \|I_{target} - I_r^t\| \tag{1}$$

The adversarial loss is defined below,

$$L_{adv}^D = L_{bce}(D(I_target), 1) + L_{bce}(D(I_r^t), 1) \tag{2}$$

$$L_{adv}^G = L_{bce}(D(I_r^t), 1) \tag{3}$$

The overall loss function for refine network is

$$L_G = \lambda_w L_{GT} + L_{adv}^G \tag{4}$$

where λ_w is the weight of distance to ground truth.

3.4. Stage-3: Preference Based Aggregation

After the first two stages of processing, the predictions are already valid human beings. However, these synthesized images are often incomplete and many of the areas are predicted purely without any direct visual cue from source images (e.g. synthesize an image of one's back given an image in the front). This occurs often in occlusion cases or when the difference between source pose and target pose large. Therefore, enlightened by self-learned confidence in [21], we tend to propose a module predicting our preference for each part on intermediate predictions.

The areas on intermediate predictions we prefer are those (1) directly originate from seen parts of the source images (2) have high quality. To achieve the first objective, the model has recognize if an area was originally invisible on one source image and whether it could be replaced by prediction from another source image. As occlusions and invisible views can both be estimated from the pose of human, the module we design only has to take key points locations as input, and the output is our preference score for each body part. With pose being very sparse and low-dimensional, this network can be very shallow. The loss for part i is:

$$L_{pose}^i = \frac{\|I_{target}^t - T_r^t\| \odot mask_t^i}{\|I_{target}^t - T_r^t\|} \tag{5}$$

where $mask$ is the mask for part i on target pose, $\odot$ is the element-wise multiplication and the confidence for each part is normalized by the total difference.

However, this is not the whole story. Apart from the case of unseen parts and occlusions, some intermediate predictions may have lower quality than the others and it is less-preferred as a result. Up till now, although huge effort have been put into evaluation metrics of the quality of image, there hasn't been a common scheme for image synthesis tasks. Enlightened by the idea of perceptual loss and inception score, we use our discriminator in the previous stage as the judge of image quality, which takes both validity and image quality into account.

During the final merging, we simply multiply the part preference score with the discriminator score and choose the part with highest score.

4. Experiment

We evaluate our method on Market-1501[25], which has images of 1501 different persons and each person has several images from different poses. Therefore, this is a perfect dataset for human image synthesis tasks and many of the previous research in this field also conducted experiments on this dataset. We carry out experiments from four perspectives in this section: (1) The performance of our method on human image synthesis tasks; (2) Ablation study on the design of network structure; (3) The ability for our model to progressively improve the quality with more source images; (4) The performance of our model working under only one source image.

4.1. Implementation Details

In the first stage, we use AlphaPose[6] for pose estimation and parse each human into 10 parts according to the locations of 17 key points. Specifically, the parts are: head, torso, left and right forearm, left and right hindarm, left and right thigh, left and right calf.

In the second stage, the λ_{sk} in the refine network is 0.5, and the number of residual block is 6. During training, we set $\lambda_w = 10$. The training process can be completed within 40 epochs. The optimizer we use is Adam[11], with learning rate for refine network and discriminator being 5×10^{-4} and 5×10^{-4} in the first 20 epochs and half for the last 20 epochs. After training the base refine network model, we fix it and train the refine block with Adam and learning rate of 4×10^{-4}. The training process is completed in 20 epochs.

In the third stage, the pose net was trained with a learning rate of 5×10^{-4}. As the representation of poses are very sparse and the pose net is shallow, the whole training process is completed in two epochs. In the quality branch, we use the discriminator for judging. Before apply it into final merging, it was fine tuned using fake and ground truth image pairs for two epochs.

4.2. Performance

Our method shows competitive performance compared to previous work[16][15]. As we focused on human part on this project and haven't bother generating the background, we directly used the background information from target images. Therefore, instead of comparing raw SSIM with previous methods, We compare the SSIM of our full images towards the mask-SSIM of them. In this way, the comparison is supposed to be fair. However, we still list their SSIM for reference. The comparison is shown in Table 1.

However, our methods works well at maintain the details of person. That is to say, our method not only predicts a human with valid shape and texture, but also find the right place to put the visual details. This is largely owed to the skip connection in the refine network, which will be discussed in ablation study. In figure 3, we show the images sampled from our generation. It is demonstrated that, when visual cues are visible, our method is able to synthesize images maintaining detailed information by propagating from other images onto the target one. To test this, we sample two source images with nearest pose to the target so as to provide essential detail information. Then the predictions are inspected to see if the details are kept.

Method	SSIM	mask-SSIM
PG^2[15]	0.253	0.792
Disentangled[16]	*	0.614
Ours(one source)	0.682	0.682 (full SSIM)
Ours(four sources)	0.713	0.713 (full SSIM)

Table 1. Performance Comparison on Market-1501[25]

Method	SSIM
PNGAN[18]	0.605
PNGAN + sk	0.636
PNGAN + sk + fr	0.682
PNGAN + sk + fr(2 sources)	0.705
PNGAN + sk + fr(4 sources)	0.713

Table 2. Ablation Study. *sk* for skip connection, *fr* for refinement

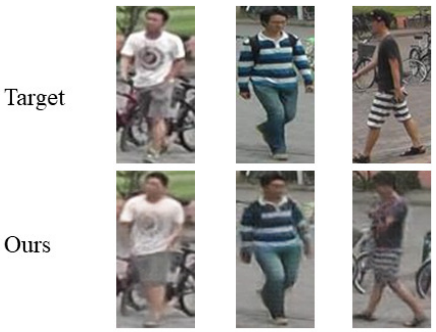

Figure 3. Sample from Predictions. Details are pattern on image1 and texture on image2, 3.

4.3. Ablation Study

In the ablation study section, we focus the following topics: (1) the improvement of skip connection; (2) the progressive improvement with more available source images. By default, the experiments are carried with only one source image.

It's demonstrated in the table that skip connection im-

proves the performance by a large margin. It's because the skip connection directly propagates the information from predictions in geometric transformation stage into the result. Moreover, this also contributes to the final details. To the best of our observations, the results without skip connection are more smooth but lose almost all the details.

The progressive ability of our model is also shown in Table2. It is observed that with more source images available, the better are the final predictions.

5. Conclusion

In this report, we explore the methods for multi pose to novel pose human image synthesis and propose a new pipeline for this. The stage one leverages geometric transformations to put the information from source images into right place. The stage two uses refine network to make the predictions from stage 1 valid. In stage three, preference score is defined and used to guide the aggregation of intermediate predictions from different sources. Experiments on Market-1501 shows that our method is able to generate realistic human images with details.

References

[1] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag. Synthesizing images of humans in unseen poses. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[2] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua. Cvae-gan: fine-grained image generation through asymmetric training. *CoRR, abs/1703.10155*, 5, 2017.

[3] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[4] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 11–20. ACM, 1996.

[5] P. Esser, E. Sutter, and B. Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018.

[6] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.

[7] J. Flynn, I. Neulander, J. Philbin, and N. Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2016.

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

[9] P. Isola, J.-Y. Zhu, T. Zhou, and A. 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, 2017.

[10] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[11] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[13] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.

[14] C. Lassner, G. Pons-Moll, and P. V. Gehler. A generative model of people in clothing. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 6, 2017.

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

[16] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2018.

[17] E. Park, J. Yang, E. Yumer, D. Ceylan, and A. C. Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 702–711. IEEE, 2017.

[18] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue. Pose-normalized image generation for person re-identification. In *European Conference on Computer Vision*, pages 661–678. Springer, 2018.

[19] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[20] P. Sturm and B. Triggs. A factorization based algorithm for multi-image projective structure and motion. In *European conference on computer vision*, pages 709–720. Springer, 1996.

[21] S.-H. Sun, M. Huh, Y.-H. Liao, N. Zhang, and J. J. Lim. Multi-view to novel view: Synthesizing novel views with self-learned confidence. In *European Conference on Computer Vision*, pages 162–178. Springer, 2018.

[22] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Single-view to multi-view: Reconstructing unseen views with a convolutional network. *CoRR abs/1511.06702*, 1(2):2, 2015.

[23] J. Yang, S. E. Reed, M.-H. Yang, and H. Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *Advances in Neural Information Processing Systems*, pages 1099–1107, 2015.

[24] B. Zhao, X. Wu, Z.-Q. Cheng, H. Liu, Z. Jie, and J. Feng. Multi-view image generation from a single-view. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 383–391. ACM, 2018.

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

[26] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In *European conference on computer vision*, pages 286–301. Springer, 2016.

[27] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017.