Multi-Pose to Novel-Pose 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 pose guided human image synthesis when multiple source images are available. The core problem solved in this paper is synthesizing target image with single source and progressively improving the prediction with more available sources. We propose a method consisting of three stages: (1)Geometric transformation, which generates coarse predictions directly from source images; (2)Refine network, which makes predictions from stage 1 realistic and valid human while retaining details; (3)Preference score based merging, which merges multiple predictions from stage 2 into a single final result. We follow the related work and carry out experiments on Market-1501. Both qualitative and quantitative results demonstrate that our model can generate realistic human images while maintaining the details of the person. Our method shows competitive performance comparing to state-of-the-art. And it can improve the predictions with more source images. To the best of our knowledge, this is the first attempt to tackle multi pose to novel pose problem.

1. Introduction

Generating realistic images of humans is importance in computer vision and has broad applications, like image editing, person re-identification, in-painting, etc. The recent progress of image generation models such as variational autoencoders(VAE)[12] and generative adversarial networks[8][19] has enlightened great progress in human image synthesis[15][14][1][16]. However, previous researchers mainly consider the situation of using only one source image in each synthesis. But what will happen when multiple source images are used simultaneously? Therefore, we in-depth explore this ”Multi-Pose to Novel-Pose” problem in this paper.

From the perspective of performance, as multiple sources provides more information than a single source, synthesizing with multiple sources can yield better results than single source. For example, when we synthesize a full human image based on two source images from different views, target pose may have unseen parts on each source images. However, using both the sources enables us to make up the unseen parts of one source with the other, while using only one source forces us to guess what the unseen parts look like. Therefore, using both of the images can lead to predictions with higher fidelity. From the perspective of application, the situation of multiple source images exists widely in many scenarios. And it's also wasteful to use only one source image at a time. Moreover, algorithms able to cope with multiple sources are more flexible as they can also work with only one source image. From the above two perspectives, it's reasonable to go beyond the constraints of using only one source image and use multiple source images simultaneously. To the best of our knowledge, we are the first attempt to solve this problem.

The major challenge is how to combine information from different sources. Although this problem has been studied in novel view synthesis, previous researchers mainly focused on the case of rigid objects, and it also hasn't been solved yet. Much effort have been expended in geometry-based methods aiming to directly estimate underlying 3D structures[20][4]. But these fail to work on non-rigid objects such as humans due to the variety and changing of 3D structure and surface. Learning-based methods arise recently[26][21], but they still focus on rigid objects till now. In light of this, the method in this paper utilizes both the unique structure of human and generation ability of learning algorithms, which corresponds to geometric-based models and learning-based models separately. Then we propose self-learned preference score to guide the merging process, which is motivated by self-learned confidence in[21].

The overview of our method is demonstrated in figure 1, containing the following stages:

1. **Geometric Tran formation Stage**: Each human image is parsed into several body parts according to the key points' locations. Then they are transformed onto target image according to target key points' locations using affine transformation. This stage utilizes human structure in form of key points.

2. **Refine Stage**: We refine the predictions from pre-

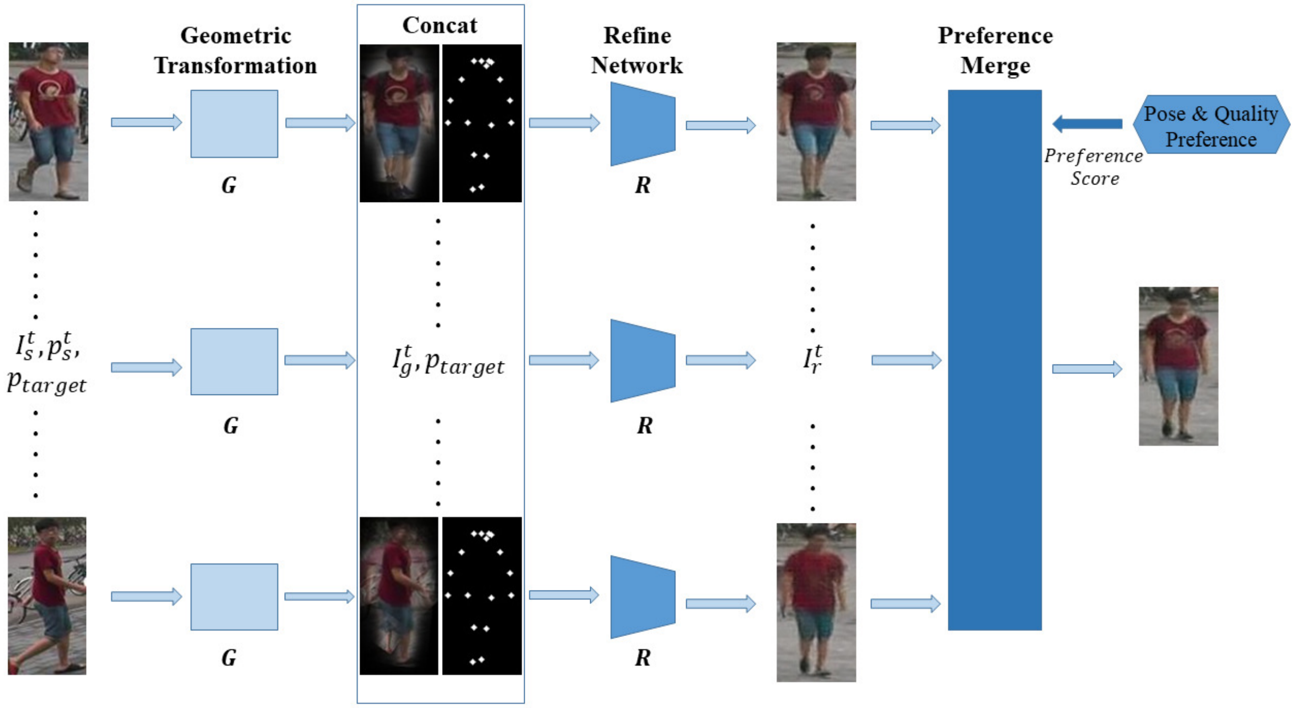

Figure 1. Pipeline Overview

vious stage with a deep convolutional neural network(DCNN). As the receptive field of DCNN is large, this network can propagate information among different parts in the image, thus it's able to fix twisted areas and make up missing regions.

3. **Merging Stage**: With predictions from stage 2, we define preference score to guide the merging of them. The parts preferred are those: (1) originally seen in the source images; (2) in high quality. As for(1), we use a fully connected neural network to process pose information and speculate the visibility of each part; as for (2), we use a trained CNN to evaluate the quality of prediction. Finally, we select the best intermediate prediction for each part and merge them together into a final result.

Our contributions can be summarized as follows: (1) we propose a new task of human image synthesis which is meaningful for applications; (2) we propose a new method for this problem, and out method can also fit in the settings of traditional human image synthesis problem; (3) we carry out experiments and prove the effectiveness of our method both qualitatively and quantitatively; (4) we design experiments to demonstrate of the unique ability of our method to maintain details from source images and progressively improve the predictions with more available source images.

2. Related Work

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. Variational aotoencoder(VAE)[12] also arises as a typical method, especially when conditional information is available. VAE has shown its strong power in conditional generation problems, and has been widely applied into human image synthesis [2][13][5]. In this paper, we follow the scheme of generative adversarial models and deal with the conditional information in other ways.

Human image synthesis 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, pose guided human image synthesis. The difficulty of this task lies in the variety of human poses and deform-ability of surfaces. [14] first propose a model generating people in clothing. Then [15] propose a coarse-to-fine two stage model in generative adversarial manner. [24][16] both share the overall framework this method, with adding modifications to make the process of generation more controllable and performance improved. [5] explores another way, it considers the source image as information source of pattern and target pose as shape. Then

VAE is used to encode these information into latent variables for final synthesis. [1] takes human structure into consideration, it parses human into different parts and warps them before refinement. In this paper, we combine the advantages of [1] and [15]. We leverage the human structure and the generation ability of neural network. Moreover, as none of the human image synthesis work has considered simultaneously using multiple source images, we propose modifications to these methods. enabling them to work with both multiple and single sources.

The difficulty for multi-view synthesis lies in merging information, and it's important because neural networks haven't developed the ability of progressively understanding the structure of an object in few shots. Multi-view synthesis is an important task in computer vision, but most of the previous work focuses on rigid objects. [7] studies multi-view stereo problem by regressing directly to output pixel values. [23][22] propose the methods to generate pixels in a target view. [17][26] predict flows to move pixels from the source onto the target, 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

In order to synthesize human images in high quality, 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. We use geometric transformations to solve (1) by moving parts into target locations. We then use a refine network to make images realistic, so as to satisfy (2). In the third stage, we merge the mediate predictions according to a self-defined preference score based on image quality and poses, thus our model meets the needs of (3).

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

In this stage, we use geometric transformations to move body parts on source images onto target images, and get a coarse prediction. In order to extract the body parts, we apply AlphaPose[6] to estimate the locations of key points. Then each body part is masked out using Gaussian heat map. With the pose information extracted, transformation matrix can be computed according to source pose and target pose. After applying affine transformations onto each part, we generate $I_g^1, I_g^2, ..., I_g^N$, which are the predictions of this stage.

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, so 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 inconsistency between source and target, images generated in stage 1 suffer from problems like twisted textures, mismatched borders, false structure, etc. However, the essential information have already been put into the right place. In light of this, we use a network to refine the predictions from stage 1. The input to the 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 compensate for missing areas. As for (1), we add a skip connection from input layer to output layer. 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" in order to satisfy (2). So we set the weight of this skip connection to be λ_{sk} instead of 1, and it is 0.5 in our experiment. Apart from the skip connection, the output images still suffer from minor twisting and blurry. So we add another refine block for further refinement on top of modified PNGAN and these two make the whole refine network. The structure is demonstrated in Figure 2.

To generate sharp images, we use the training methods

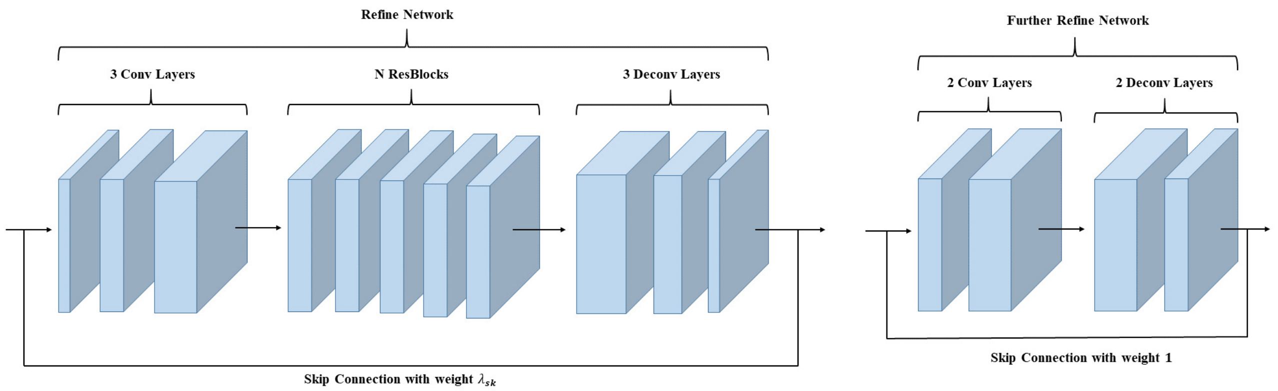

Figure 2. Refine Network Structure

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 images. 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

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]

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 in this project so we haven't bother generating the background. Instead, the background is directly pasted on the predictions. Therefore, instead of comparing raw SSIM with previous methods, we choose to compare the SSIM of our full images towards the mask-SSIM of previous work. In this way, the task become harder and the comparison is supposed to be fair. However, we still list their SSIM for reference. The result is shown in Table 1.

Moreover, our method works well at maintaining the details of person. That is to say, our method not only predicts human images with valid shape and texture, but also find visual cues from source images then put them in the right locations. This is largely because of the skip connection in the refine network, which will also be discussed in ablation study. To demonstrate this, we sample two source images with nearest pose to the target pose so as to provide essential detail information. Then the synthesized images are inspected to see if they select and keep the right and essential details from source onto target. As is shown in the pictures of figure 3, the ability of our model maintaining texture and patterns is up to our expectation.

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

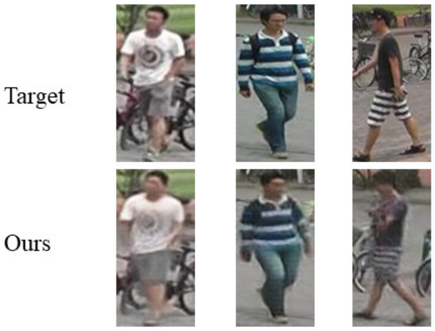

Figure 3. Sample from Predictions. Details are pattern on image1 and texture on image2, 3.

4.3. Ablation Study

In the ablation study section, we focus on 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 improves 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 final predictions are better. It is logically right as the source images may make up for each other's invisible parts and as a result generate better images.

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.