Computer Vision: Project Proposal

Jiaqi Liu
liu00687@umn.edu

Tony Liu
liux4408@umn.edu

Prithvi Raj Botcha
botch025@umn.edu

Abstract

We tackle the HUMBI Pose-Guided Human Rendering Challenge and try to beat the benchmark scores. As a baseline, we used the VUNet that is proved to work well for synthesizing poses from a single still image. In an attempt to overcome the extreme generality and potential loss of the image appearance, while preserving the shape/pose, that comes with VUNets; we later propose to add another Generative network to the existing U-net as a possible improvement on the image construction.

1. Introduction

Human pose transfer, the generation of an image of a person in an unseen target pose given a source image, has a wide range of applications. A few of these include fashion try-on, animation, and virtual reality. Each of these rely heavily on the synthesized image being believable and realistic while also maintaining high fidelity.

The problem is challenging as the target pose contains regions that are occluded in the source image so data needs to be accurately inferred to unseen regions. Additionally, variance in textures as well as the diversity of people and the wide range of poses a person can take often creates misalignment issues.

2. Related works

In human pose transfer, generative adversarial networks (GANs) or variational autoencoders (VAEs) are typically used [9] [7]. There are generally two approaches to human pose transfer.

The first approach learns to map the provided input to produce the desired output. Methods following the first approach usually begin by extracting the pose of the subject in the given image to compare to the target pose. This is often done using some variant of OpenPose developed by Cao et al [4]. PG2, developed by Ma et al., follows the first approach by first generating a coarse initial image with the correct pose before further refining textures. [16]. Zhu et al. proposed generating images in a progressive fashion by using the target pose as a guide to gradually move patches of the image until it resembles the target pose [26]. Li et al. furthers this by using a two step update where image and pose features are first merged and then a pose-guided image update is used to reduce ambiguities in inferring pixels in occluded regions [11]. Instead of using target poses made up of joint keypoints, Liu et al. used segmentation masks to represent the target pose leading to decreased computation load [14].

The second approach typically breaks down the image into several components and modifies each piece before combining them to produce the final image. Mat et al. first decompose the source image into foreground, background, and textures where each is manipulated and encoded separately. The final image is then synthesized by decoding these features [17]. Balakrishnan et al. further separate the body into parts such as the legs or arms. These body parts are then spatially transformed into their targeted positions to be fused together [1]. Zhao et al. extract latent pose features from a sequence of interpolated poses between the source and target pose [25]. Sun et al. expands on this idea by using bidirectional convolutional LSTM to extract texture features for more realistic synthesized images [21]. Men et al. uses two paths to encode information. The first path embeds pose features to latent space and a human parser [6] is used to extract attributes of body parts where they are encoded using a global texture encoder Style blocks then transfer the textures before a decoder is used to reconstruct the final image. [18]. Tang et al. hone in on local attributes by using a novel architecture called XingGAN composed of two crossing branches made up of concatenated blocks. The first is Shape-Guided Appearance-Based Generation branch and the second is the Appearance-Guided Shape-Based Generation branch. The outputs of blocks in each path are fed to both branches [22]. [12] [8] both use appearance flow between source and target pose to synthesize the output image.

3. Baseline Evaluations

VUNets [5] are used to train and evaluate some set of the HUMBI training data to get some decent similarity scores as can be observed in the Fig 1.

Major limitations that can be seen including appearance

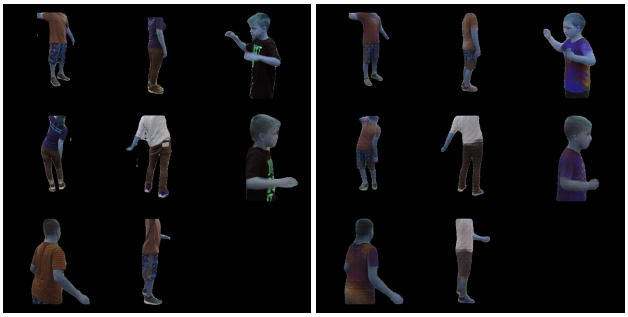

Figure 1. Base truth (left) and the synthesized images (right)

misalignment (w.r.t. shape), low detail especially in facial features and hands as well as clothes with dense patterns. Rare samples can generate highly dissimilar results because of excessive generalizations by the generator.

4. Method

In this section, we would first give an overview of GAN. Then we introduce our method that implement GAN after the baseline model. One of the appealing properties that GAN provides is that it allows the network to learn similarity between images, so that the discriminator is able to distinguish the true images from the fake images. The proposed method is motivated in a way so that it not only take advantage of having feature-wise similarity metric for the reconstruction error, but also improving the quality of the data sample outputted from the original baseline models.

4.1. Generative adversarial network

A GAN model is made up of a generative model Gz that maps latent variable z to data sample space, and a discriminator model that assigns probability $p = \mathrm{Dis}(\boldsymbol{x}) \in [0, 1]$ where x is a true sample. The total loss function defined for GAN is

$$\mathcal{L}_{\mathrm{GAN}} = \log(\mathrm{Dis}(\boldsymbol{x})) + \log(1 - \mathrm{Dis}(\mathrm{Gen}(\boldsymbol{z})))$$

with x being a training sample and z being the latent variable.

4.2. VUNET

V-Unet models shape and appearance when generating images.The authors built a U-Net-like [19] generator function to learn about images x based on the estimated shape $\hat{y}$,but still conditioning on the latent space p(z). The decoder in U-Net-like network and the encoder F_θ forms a variational autoencoder where latent variable z is mapped from image x to latent space.The loss function defined in vunet is

$$\begin{aligned}\mathcal{L}(x, \theta, \phi) = &-KL\left(q_\phi(z \mid x, \hat{y}) \| p_\theta(z \mid \hat{y})\right) \\ &+ \sum_k \lambda_k \left\|\Phi_k(x) - \Phi_k\left(G_\theta(\hat{y}, z)\right)\right\|_1\end{aligned}$$

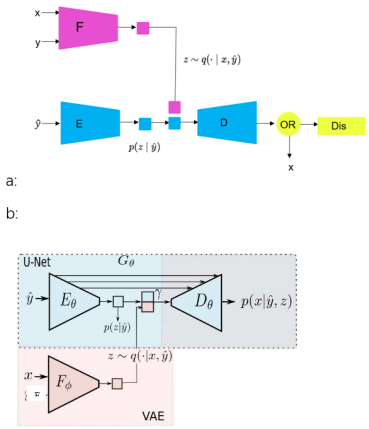

Figure 2. (a) Overview of our work, blue part (E + D) is the U-Net like network that guarantees optimal transfer of spatial information from input to output data samples. The pink part (F) is the encoder which map input data to latent space. The yellow part is the discriminator that is trained to classify decoded samples and samples from the dataset.(b) Previouw work: The baseline model that we used and got inspired on.

where $-KL\left(q_\phi(z \mid x, \hat{y}) \| p_\theta(z \mid \hat{y})\right)$ is the prior regularization term ($\mathcal{L}_{\mathrm{prior}}$) and $\sum_k \lambda_k \left\|\Phi_k(x) - \Phi_k\left(G_\theta(\hat{y}, z)\right)\right\|_1$ which is the reconstruction error term ($\mathcal{L}_{\mathrm{vunet_{rec}}}$). Φ is the VGG19 that mesure the perceptual similarity and λ_k, k are just hyper-parameters that control the contribution of each layers' perceptual loss to the total loss.

4.3. Proposed Method

Given the limitations of the baseline model, we hope to improve the quality of the data sample produced by the generator in the VUnet model. We propose our approach to combine the reconstruction error with the properties of images learned from the discriminator network of GAN. The GAN component of the model is implemented so as to take advantage of its high-quality generative model with encoder that map input data to latent space. [10] [3]

The schematic of our architecture is presented in Figure 2. In addition to the U-Net like network E_θ, D_θ and encoder from VAE F_θ, we introduce a discriminator network Dis(x) that maps either true sample or fake sample to its corresponding probability. Since we want to make use of the similarity metric used from discriminator network, we propose to add another reconstruction error

$$\mathcal{L}_{\mathrm{Dis}_k} = -\mathbb{E}_q\left[\log p\left(\mathrm{Dis}_k(x) \mid z, \hat{y}\right)\right]$$

to vunet loss function where k indicates the kth layer of Discriminator network, $p\left(\mathrm{Dis}_l(x) \mid z, \hat{y}\right) = \mathcal{N}\left(\mathrm{Dis}_l(x) \mid \mathrm{Dis}_l(\tilde{\boldsymbol{x}}), \mathbf{I}\right)$ with $\tilde{\boldsymbol{x}} \sim G_\theta(\hat{y}, z)$. In this case, the total loss function would be

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{prior}} + \alpha\mathcal{L}_{\mathrm{vunet_{rec}}} + (1 - \alpha)\mathcal{L}_{\mathrm{Dis}_k} + \mathcal{L}_{\mathrm{GAN}}$$

Each loss term is defined above and $\mathcal{L}_{\text{GAN}} = \log(\text{Dis}(\boldsymbol{x})) + \log(1 - \text{Dis}(G_\theta(\hat{y}, z))$.

The reason why we want to keep two reconstruction error is to make the decoder learn the spatial information directly from shape and improve the quality at the same time. Notably, as D_θ receives error from two reconstruction error terms and $\mathcal{L}_{\text{GAN}}$, we want to balance the ability to reconstruct and the ability to fool the discriminator. In this case, we use a weight parameter α to control the relative contribution of reconstruction error from generator network and the one from discriminator.

5. Data

The data used is, in its entirety, picked from the HUMBI dataset[24], although the actual data that went to training our model is only a fraction of it. The dataset contains over 360k images, each of resolution 256×256. This large scale of visual data is ideal for learning the latent shape and appearances of the various human poses as the quality of the synthesized the image would highly depend on the trained the data.

The dataset also provided the foreground masks for all the presented images. The mask is generated using YOLACT real time segmentation[2] for people images. Additionally the keypoints are also provided using single-person keypoint generation by OpenPose. We translated the order to match with the *joint order* used in the conditional U-Nets. The mapping is done from gaze/face/hand/body/shoes to the eyes/shoulders/wrists/hips/knees/ankles joints, and those without shoulders and wrists are ignored.

Data for validation and testing are also separately presented.

6. Results

The accuracy scores of our baseline and the proposed methods can be seen in the table 1 and the Codalabs HUMBI challenge submission results are found under the username *botch025*. Both the original VU-nets network and our modified network were trained using only about 10% of the HUMBI dataset. Further, our network was only trained to around 8% completion due to a few time and computational issues like runtimes disconnecting in Colab. The testing is done over the saved checkpoints on this subset of the trained models. As a result there are regions in the synthesized image that contain incorrectly colored pixels and contain lots of noise. Analyzing the output images further, we can see that most of them have fairly accurate poses. This suggests that the model learns pose information more quickly than texture information.

The same can be seen with the results where, while the similarity scores are only modest, the root mean square errors scored pretty good, further boosting the idea that the pose training is quick. Additionally, there are also noisy pixels in blank regions not associated with the subject. Further improvements could be made to boost texture accuracy and reduce noisy pixels by adding a mask so that the model can focus solely on learning texture details about the subject instead of the entire image.

Table 1. Accuracy scores on the test data with the baseline and proposed methods

	MSSIM	MRMSE	SSIM	RMSE
VUNet (Baseline)	0.630372	0.825861	0.630372	0.825861
VUNet + Discriminator (Proposed)	0.179161	0.802072	0.179161	0.802072

The proposed models, with a changed loss function, unfortunately generated only further deteriorated outputs as opposed to the baseline with started. This could be observed visually and with the similarity scores as well. Fewer time allocated to train with the new model might be a reason.

We also attempted to run the testing images with certain pretained models for VUNets that are already made available online, in particular with the COCO[13] dataset and the DeepFashion dataset[15]. The generated images did not look impressive enough visually to make further pursuits in that direction, the most likely reason being that training is done largely on many non-human images.

7. Conclusion

Conditional U-Nets and generative adversarial networks (GANs) are usually considered the state-of-the-art models that particularly perform very well, often seen to outperform others, in generating diverse and high quality samples [20][23]. We started with the Variational U-Net [5] generative model as the baseline because they are proved to be good for shape-guided image synthesis of any general shape. Training on the training data and validating gave very impressive results to begin with. This model generates the output directly instead of modelling the intricate within-details of the shape and appearance. This makes it difficult to learn new data, which could be seen even with our testing, where rare/underrepresented (e.g. new faces) are often not learnt and the results often tended to be over-generalized versions of training data w.r.t. to the appearance or patterns.

As an attempt to overcome these limitations, we proposed to introduce some kind of supervised learning, and so added another reconstruction error to make use of the similarity metric from the discriminator network. This error is

made a part of the vunet loss function. This new model is then used to train our HUMBI data. We obtained modest results using this approach, and the new outputs with the test dataset is also submitted at the HUMBI competition page.

To conclude, for this project, we attempted the HUMBI human rendering challenge of synthesizing a realistic image of a human given a target image and target pose. We chose our baseline model to be VU-nets and modified this network by adding a discriminator to the output of the VU-nets model. Our submission results show that our network is able to correctly synthesize images with the correct pose. However due to insufficient training, the texture accuracy is not ideal. We managed to reach a decent score with the generation but couldn't beat the benchmark set by HUMBI. Future work would include using masks to narrow the region in the image that the model should focus on learning which would also ameliorate the noisiness seen in current output images.

8. Contributions

- Jiaqi Liu: Training models on VUNets, improvements on method with GAN, wrote section 4 of the report.
- Tony Liu: State-of-the-art research, selecting baseline model, training models with the proposed solution, setting up data, wrote section 1,2, and parts of 6 and 7.
- Prithvi Raj Botcha: Evaluations on the baseline, limitations, setting upon data, testing and submissions, wrote sections 3, 5, 6, 7, 8.

Our submissions of the testing results to the HUMBI Codalab competition page could be seen under the username **botch025**. Currently, we stand at the 4^{th} position on the leaderboard with a similarity score of 63% .

All the code and the instructions to run it is available at https://github.com/cs12b006/cv-pose.

References

[1] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8340–8348, 2018. 1

[2] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation, 2019. 3

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 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, 2017. 1

[5] Patrick Esser, Ekaterina Sutter, and Björn 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. 1, 3

[6] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 932–940, 2017. 1

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[8] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10471–10480, 2019. 1

[9] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[10] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pages 1558–1566. PMLR, 2016. 2

[11] Kun Li, Jinsong Zhang, Yebin Liu, Yu-Kun Lai, and Qionghai Dai. Pona: Pose-guided non-local attention for human pose transfer. *IEEE Transactions on Image Processing*, 29:9584–9599, 2020. 1

[12] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3693–3702, 2019. 1

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 3

[14] Meichen Liu, Kejun Wang, Juihang Ji, and Shuzhi Sam Ge. Person image generation with semantic attention network for person re-identification. *arXiv preprint arXiv:2008.07884*, 2020. 1

[15] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104, 2016. 3

[16] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *arXiv preprint arXiv:1705.09368*, 2017. 1

[17] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2018. 1

[18] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5084–5093, 2020. 1

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 3

[21] Wei Sun, Jawadul H Bappy, Shanglin Yang, Yi Xu, Tianfu Wu, and Hui Zhou. Pose guided fashion image synthesis using deep generative model. *arXiv preprint arXiv:1906.07251*, 2019. 1

[22] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. Xinggan for person image generation. In *European Conference on Computer Vision*, pages 717–734. Springer, 2020. 1

[23] Jiaheng Wei, Minghao Liu, Jiahao Luo, Qiutong Li, James Davis, and Yang Liu. Peergan: Generative adversarial networks with a competing peer discriminator, 2021. 3

[24] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. Humbi: A large multiview dataset of human body expressions, 2020. 3

[25] Wenbin Zhao, Qing Xie, Yanchun Ma, Yongjian Liu, and Shengwu Xiong. Pose guided person image generation based on pose skeleton sequence and 3d convolution. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1561–1565. IEEE, 2020. 1

[26] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2347–2356, 2019. 1