3D Reconstruction of Human Body from 2D Images with Graph Attention Network and Discriminator as Prior

Arda Arslan
aarslan@student.ethz.ch

M. Eren Akbiyik
eakbiyik@student.ethz.ch

Gökberk Özsoy
goezsoy@student.ethz.ch

ABSTRACT

3D reconstruction of humans from 2D images task requires grasping the exact human pose and shape as seen in the image, accounting for the occluded or missing parts, and recreating 3D body mesh using this information. We propose an end-to-end trainable pipeline, which first extracts high level body coefficients directly from image, then refines these with a graph attention network considering human body structure, and finally validates these with a discriminator if they are realistic or not. Together with an augmentation strategy, our method is one of the best for minimizing mean squared error between predicted 3D body mesh coordinates and ground truths.

1 INTRODUCTION

3D human pose estimation using 2D images of humans aims to create 3D human mesh with a specific pose and shape recovered from input RGB image. The parameterization of 3D mesh is done by a generative human body model, SMPL [13] which takes body's joint angles and shape coefficients and outputs 6890 vertices necessary to create 3D human body mesh.

Even though the task is ambitious and has drawn lots of attention, it is inherently challenging because of human body's complex articulation, and inevitable information loss when projecting from 3D to 2D. In addition, the datasets created for solving this task is generated in a controlled environment which hardly captures the variety as in in-the-wild images.

In this paper, we propose an end-to-end recovery pipeline, where we directly estimate SMPL parameters from an RGB image using a deep learning model and optimize the whole system with respect to the objective of predicting 3D mesh vertices as close as possible to ground truths. This way our work can be classified as single-stage estimation of body model parameters as in [7], and differs from multi-stage approaches which include intermediate targets requiring separate optimization procedures [15]. Furthermore, there are some efforts on directly predicting 6890 vertices instead of low dimensional pose and shape parameters [4].

The building blocks of our model are as follows: a robust pre-trained feature extractor to estimate initial pose and shape parameters, a graph attention network whose nodes are connected considering human skeleton to refine pose parameters, and a fully connected network to refine shape parameters. These refined parameters are then fed to SMPL model to obtain 3D mesh. The model defined up to this point is called generator. In addition, a discriminator is trained to differentiate unrealistic pose parameters, and push the generator to produce realistic ones. The performance of the model is evaluated by the standard mean squared error between the predicted mesh vertices and the ground truth mesh vertices, and we used a subset of Human3.6M dataset [1, 6].

With a gradual and aggressive augmentation strategy that will be explained further, the possibility of severe over-fitting is reduced, which allows us to train the model for 200.000 iterations. In addition, the graph attention network captures body joints' dependencies better than plain feature extractor and enables refined representations. Finally, discriminator acts a prior, reducing the chance of the generator outputting unrealistic SMPL coefficients. With these ideas, we managed to get 0.00781 public score on the test set, which is among the best scores.

2 RELATED WORK

Following the development of SMPL [13] model, computer vision community showed increasing interest in the recovery of 3D human mesh using only 2D images. Bogo et al. [15] used a multi-stage approach where they first extracted 2D body joint locations from RGB image using CNN-based method, DeepCut [16], and then used a model to fit predicted 2D joints to SMPL parameters. Kanazawa et al. [7] were inspired by generative adversarial networks where they put a discriminator on predicted SMPL parameters to limit the generator to produce realistic human body model coefficients. Kolotouros et al. [10] presented SPIN, a self-improving approach for training a neural network for 3D body retrieval, with the combination of a regression and an optimization-based method. Recently, estimating human mesh coordinates directly has became another popular approach. For instance, Kolotouros et al. [11] proposes a graph CNN which encodes human skeleton structure and enables convolutional mesh regression of 3D vertex locations using features extracted from RGB image.

3 METHOD

The model consists of four main structures to extract features from the image data, convert those features to SMPL shape and pose parameters separately, and provide regularization with a discriminator. A diagram can be seen in Figure 1.

3.1 Pre-processing

The relevant parts of the input images as labeled in the dataset are cropped, and resized into 224x224 with padding to preserve the aspect ratio. Resizing here can be seen as a form of geometric transformation, namely isotropic scaling, which is also included in the augmentation pipeline below.

3.2 Image Feature Extraction

In order to extract the necessary predictors from each image, we used a ResNeXt-50 model proposed by Xie et al. [17] pre-trained on the ImageNet [5] dataset. This model performs better on the ImageNet than its more commonly used predecessor ResNet-50 while requiring nearly the same number of parameters [17]. We fine-tuned the last two Bottleneck layers of the pre-trained model.

Gradually increasing augmentation. In order to prevent over-fitting, we have used an augmentation pipeline proposed by Karras et al. [8] prior to feeding the images into the feature extractor. This pipeline is mainly developed for training GANs with limited

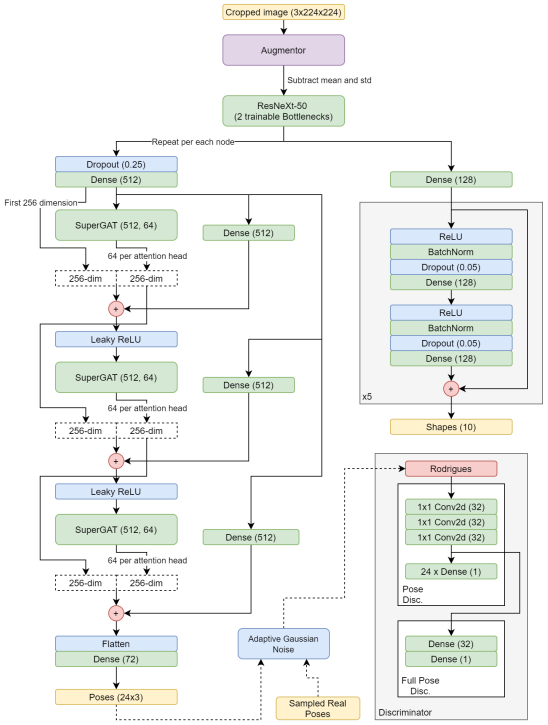

Figure 1: A detailed diagram of the model

data as it provides a framework for differentiable and adaptive augmentation, and the amount of augmentation at any iteration is parametrized with a probability value $p \in [0, 1]$ that stands for the separate probability of applying each augmentation in the pipeline. We have observed that the high initial augmentation impeded the training performance, and chose to gradually increase p during the training until a predetermined maximum. These transformations are directly applied to the cropped image, and includes both the geometric transformations: flipping around y-axis, arbitrary rotations, translations and isotropic scaling; as well as color transformations: brightness, contrast, luma flip, hue rotation, saturation. Cutout is also included, with window side length being 0.25 times the image side lengths.

3.3 Pose Estimation

In order the learn the subject poses from the extracted features, we have utilized a flavor of graph attention networks, SuperGAT, as proposed by Kim et al. [9]. The network is proposed in multiple forms, and we have chosen the scaled dot-product multi-headed attention flavor as recommended for graphs with low average degree and homophily. The initial undirected network is connected to produce a human skeleton as seen in Figure 2, and the symmetric vertices are also linked with additional edges to allow information flow between mirrored limbs [4].

The ResNeXt-50 features are interpolated to each node of the graph attention network with a linear layer, and leaky ReLU activations are used between each layer of the graph network. Each attention head outputs a 64-dimensional vector per node, and these outputs are concatenated before being fed to the next layer. In total, we use 3 layers of SuperGAT with 4 attention heads, using scaled

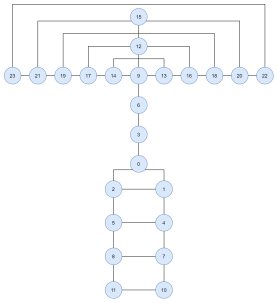

Figure 2: Connectivity of the body template used in the attention network. Numbers correspond to the node indices of SMPL body model.

dot-product attention mechanism (also dubbed as the "SD-type attention" in the original paper).

Skip connections and residual concatenations. In order to ensure healthy gradient flow and prevent over-smoothing of the graph convolutions, we concatenate each output of the graph attention layers with the output from the previous layers. However, this is not enough: as shown by Chen et al., in order to extend graph networks to estimate a K-order polynomial filter with arbitrary coefficients, we require initial residual connections at each layer [3]. We supply this by summing the output of each layer with a learnable linear mapping of the initial node encodings.

Self-supervision loss. SuperGAT comes with a concept of attention loss, by doing negative sampling of the non-linked nodes at each iteration and trying to ensure that the attention scores with the linked nodes are higher, while the scores with the non-linked nodes are lower. We have seen that scarce usage of this loss stabilizes the initial training and acts as a regularizer: the limbs in each side of the body are expected to behave similarly even when one side is occluded, and attention loss forces them to display alike behavior, and allows the network to fill in the blanks in case of missing information.

3.4 Shape Estimation

In parallel with the graph network, the ResNeXt-50 features are also fed to a residual fully-connected network to predict the shape parameters. The network has a depth of 10, and width of 128 at each hidden layer with ReLU activations.

Residual connections. The non-activated output of each two linear layers are summed, similarly with the architecture used in residual convolutional networks.

Usage of Dropout and Batch Normalization. Contrary to the initial opinions that the usage of Dropout and Batch Normalization layers together is counter-productive as Dropout skews the distribution learned by the Batch Normalization layer [12], a recent study suggests that using small amounts of Dropout directly after Batch Normalization layers boost the performance of the deep networks [2]. Accordingly, we have utilized the ordering proposed in this paper, following each ReLU with Batch Normalization and Dropout, with drop probability 0.05.

3.5 Discriminator as a Prior

In order to constrain the training only to sensible outputs for SMPL parameters, rule-based priors to punish straying too far away from the rest state are often used [15]. However, these losses are often tricky to tune and does not adapt according to the training process. Instead, in this work we train a discriminator to ensure that outputs stay in a human-acceptable range towards the end of the training, as similarly used by Kanazawa et al. [7].

For the discriminator, we used publicly available code from [18]. We used separate discriminators for pose and full pose of the subjects. For pose, we have one pose discriminator for each of the 23 joints. For each joint, we first calculate a 3x3 rotation matrix using the Rodrigues formula. Then each rotation matrix is vectorized and fed to a shared 3 hidden layer network with 1x1 convolutions, 32 channels each. This network outputs 32-dimensional embedding for each joint. Each joint's embedding is fed to a network which is only responsible for this particular joint. Each of these 23 networks consists of only an input and a 1-node output layer. We did not use nonlinear activations for pose discriminators' hidden layers. 32-dimensional embeddings for each joint are concatenated and they are also fed to the full pose discriminator. This network has one linear hidden layer with 32 units. It has 1 node at output layer. In total we have 25 discriminators and their output nodes are activated using Sigmoid activations. We used Binary Cross Entropy loss for both the discriminator training and the loss signals which are sent to generator by the discriminators.

Fighting discriminator over-fitting with adaptive additive noise. As the task limits us only to a subset of Human3.6M dataset, the discriminator has the tendency to quickly overfit on the data, providing no useful regularization to the generator network. In order to overcome this, we have extended the adaptive augmentation proposed by Karras et al. [8] for the image generators to regressors, by using additive Gaussian noise, and adjusting its variance across the training depending on the performance of the discriminator. During training, we check each discriminator's prediction on real SMPL parameters which are randomly sampled from the data. If for a given mini-batch, mean of the predictions of a discriminator for real SMPL parameters is above a certain target value, we increase the corresponding discriminator's input noise and if the mean is below this target value, we decrease the input noise. We chose 0.6 as our target value as proposed by Karras et al. [8], considering that a perfect discriminator would assign 1.0 for real SMPL parameters in our design. One intuition behind this approach is the following: As we mentioned before we are using Binary Cross Entropy loss for training the discriminator. If the discriminator assigns a probability of 1.0 to any region in its input space, and if the generator manages to put some nonzero probability mass to this region, then the discriminator's loss approaches to infinity.

3.6 Implementation Details

Both the generator and discriminator networks are trained with AdamW optimizer as proposed by Loshchilov et al. [14], with learning rate 0.0001 and weight decay 0.01. Batch size is set to 16. L_1, L_2, attention loss, discriminator pose loss and discriminator full pose losses are combined with weights 1.0, 1.0, 0.0005, 0.01, 0.01 respectively. The network is trained for 200.000 iterations. Augmentation percentage p is started at 0, and gradually is increased up to 0.75 across 60.000 iterations and then kept as constant.

Architecture	MSE
ResNet-50	0.034036
ResNeXt-50 + SuperGAT	0.008685
ResNeXt-50 + SuperGAT + Discriminator	**0.007817**

Table 1: Public test scores achieved by different architectures. Best score is provided for each architecture.

4 EVALUATION

The vanilla ResNet-50 predictor without graph network performs considerably worse than its SuperGAT counterpart as seen in Table 1. The addition of the discriminator provides a small boost in performance in qualitative results, yet the main advantage is in the qualitative analysis: we have observed that the poses are much more human-like when the discriminator is added. Although not listed in the above list, experiments with different graph networks such as GCNs and simpler graph attention structures always gave worse validation results, and residual connections consistently improved both the performance and training speed.

5 DISCUSSION

Graph attention network has proven to be quite successful in human body modeling, and it has efficiently trained even under harsh augmentation techniques such as cutout, displaying resilience against occlusion and missing information.

Tuning the weights of the loss signals that are sent from discriminator to the generator turned out to be crucial. When these weights are too large, training of the generator becomes unstable and even during the initial iterations of the experiments it becomes obvious that the generator will not do better in the following iterations. While tuning the discriminator on the validation set, we could not find a setting which yielded a substantial improvement on validation error. However, we observed that the body models generated by the generator became more realistic by the help of the discriminator.

The main bottleneck of the model is the shape regression: we have qualitatively observed that even when the pose is fully captured by the graph attention network, lack of diversity of shapes in the training set causes bad subject height prediction. Although this was expected, and was the reason for diverging from Kanazawa et al.'s [7] approach that included also a shape discriminator, we have not observed an improvement in the results when smaller shape regressors or other regularization techniques are used.

6 CONCLUSION

We combined two state-of-the-art approaches, graph neural networks and discriminator-based regularization, to achieve satisfactory results on the Human3.6M dataset. The lack of shape variation was the main limiting factor in the performance of the model, and we believe that with the inclusion of another set of subjects to improve this shortcoming, the designed model can provide better performance than the state-of-the-art architectures used for 3D human body reconstruction.

REFERENCES

[1] Cristian Sminchisescu Catalin Ionescu, Fuxin Li. 2011. Latent Structured Models for Human Pose Estimation. In *International Conference on Computer Vision*.

[2] Guangyong Chen, Pengfei Chen, Yujun Shi, Chang-Yu Hsieh, Benben Liao, and Shengyu Zhang. 2019. Rethinking the Usage of Batch Normalization and Dropout in the Training of Deep Neural Networks. *CoRR* abs/1905.05928 (2019). arXiv:1905.05928 http://arxiv.org/abs/1905.05928

[3] Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. 2020. Simple and Deep Graph Convolutional Networks. arXiv:cs.LG/2007.02133

[4] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. 2020. Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose. *Lecture Notes in Computer Science* (2020), 769–787. https://doi.org/10.1007/978-3-030-58571-6_45

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[6] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (jul 2014), 1325–1339.

[7] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2017. End-to-end Recovery of Human Shape and Pose. *CoRR* abs/1712.06584 (2017). arXiv:1712.06584 http://arxiv.org/abs/1712.06584

[8] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training Generative Adversarial Networks with Limited Data. arXiv:cs.CV/2006.06676

[9] Dongkwan Kim and Alice Oh. 2021. How to Find Your Friendly Neighborhood: Graph Attention Design with Self-Supervision. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Wi5KUNlqWty

[10] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. 2019. Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop. arXiv:cs.CV/1909.12828

[11] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. 2019. Convolutional Mesh Regression for Single-Image Human Shape Reconstruction. arXiv:cs.CV/1905.03244

[12] Xiang Li, Shuo Chen, Xiaolin Hu, and Jian Yang. 2018. Understanding the Disharmony between Dropout and Batch Normalization by Variance Shift. *CoRR* abs/1801.05134 (2018). arXiv:1801.05134 http://arxiv.org/abs/1801.05134

[13] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

[14] Ilya Loshchilov and Frank Hutter. 2017. Fixing Weight Decay Regularization in Adam. *CoRR* abs/1711.05101 (2017). arXiv:1711.05101 http://arxiv.org/abs/1711.05101

[15] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[16] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. 2016. DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation. arXiv:cs.CV/1511.06645

[17] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2016. Aggregated Residual Transformations for Deep Neural Networks. *CoRR* abs/1611.05431 (2016). arXiv:1611.05431 http://arxiv.org/abs/1611.05431

[18] Xiong Zhang. 2018. PyTorch HMR. https://github.com/MandyMo/pytorch_HMR