Human Pose Transfer with High-Resolution Clothing Outfit

Kundan Sai Prabhu Thota[1] and Pramod Murthy[2]

[1] thota@rhrk.uni-kl.de
[2] pramod.murthy@dfki.de

Abstract. In this paper, we study the displacement of clothes on a 3D human body for different poses. Our goal is to train a Conditional Variational Auto-Encoder GAN to reconstruct a clothed human body for different poses without losing the cloth dynamics. We use Graph Convolutional Neural Networks to learn the feature space of the cloth displacement. To preserve cloth wrinkles, we pass constructed mesh into a patch-wise discriminator to minimize regressive loss. In addition to that, we condition the cloth type and body pose to reconstruct the human body with the cloth as an extra layer on top of the SMPL body model.

Keywords: SMPL, Graph CNN's, GAN's, Conditional Variational Auto-Encoder

1 Introduction

Human generative models [8, 1, 6] successfully estimate the statistics of human pose and shape, but still miss an important component: clothing, which leads to several problems in various applications. For example, when body models like [13, 9, 12] is used to estimate a 3D clothed human from the 2D image, produce results that have a significant domain gap between synthesized and real images of humans. Estimating human body shape under clothing is useful for many applications, including virtual try-on, fitness tracking. It's also a key component in the displacement of a virtual cloth over a minimally clothed human body, reducing the designer's and animator's workload or understanding simulations through deep learning. However, most of the current literature in modeling, recovery, and generating clothes is focused on 2D data [5, 17] is because of two factors. First, deep learning approaches require a lot of data for training. But, there are very few 3D public datasets available for research purposes. Second, 3D garments contain large variability in terms of shape, sizes, fabrics, or textures, among others, increasing the complexity of representation of 3D garment generation. Furthermore, rendering 3D clothes regardless of topologies post-training is difficult.

Goal. The goal of the project is to create a generative model of human bodies with the cloth as a displacement layer over SMPL that is low-dimensional, easy to pose, differentiable. While training, we give cloth displacement(can be achieved by subtracting minimal clothed from clothed in a canonical pose) as input to the network. The additional layer is compatible with the SMPL model, mapping one pose and many clothing offsets. The cloth generation is a probabilistic task and conditioned with a one-hot encoded vector, different poses. The model can capture the stochastic nature of clothes depending on poses which is important for realistic cloth modeling.

Dataset. We used CAPE [7] dataset for the project. The CAPE dataset is a 3D dynamic dataset scanned using a 4D scanner of clothed humans that contains 3D mesh scans captured at 60 FPS in motion. It follows SMPL mesh topology, all frames in correspondence. It precisely captures minimally clothed body shape under clothing. It contains large pose variations (both posed and unposed that is in the canonical pose for each frame). It contains SMPL body pose parameters for each frame. The CAPE dataset contains 10 male and 5 female subjects, 600+ motion sequences, 140K+ frames, 4 different types of outfits: short upper garment short lower garment, short upper garment long lower garment, long upper garment short lower garment, long upper garment long lower garment. All these garments and pose sequences are captured under certain light conditions.

Overview. We represent cloth as an extra displacement graph layer over an SMPL body. Each node in the graph represents three-dimensional point cloud data of the 3D humans. We train a graph convolutional neural network with a Variational Auto-Encoder GAN framework to reconstruct clothed humans in different poses. During the test phase, we sample the mesh from the learned feature space distribution. We model the GAN using a patch-wise discriminator to calculate the regression loss and edge loss of the reconstructed mesh to the actual mesh.

2 Related Work

There are mainly two methods for modeling, capturing, reconstruction of clothing: (1) capturing and reconstruction methods, and (2) modeling with parameters, they are detailed as follows.

2.1 3D human shape reconstruction under the cloth

Reconstructing 3D humans from 2D images is one of the classical computer vision problems that have many solutions. Models like [8, 6] output 3D mesh from the input image but lack an additional cloth layer. For reconstructing clothed human bodies, there are existing methods like volumetric [10] or bi-planar representations [2] to model the body and clothes as a whole. These methods lack to produce parametric results, which means they can't generate results that control the pose, shape, and deformations. There are methods based on the SMPL[8] body model. In this method, clothing is often considered an offset layer over human bodies [14]. Also, the produced results are parametric, leaving us to control pose and shape deformations. Estimation of the human body under the occluded cloth is an under-constrained problem. To overcome this, most of the existing models like SCAPE [1] or its variants exploit the statistical human bodies. Many methods like [4] estimate human shapes from a single 3D scan. As the estimated models are rotation invariant, we cannot extend human poses and shapes to sequences. Also, there are methods like [11], where they proposed a layered model for cloth and estimated the shape by tracing the nearest areas where the garment is close to the body. Stoll et al. at [16] estimate the minimally clothed body under a clothed template, but the method requires manual input. Their focus is to estimate the approximate shape, which is used as a proxy in collision detection.

2.2 DRAPE (DRessing Any PErson)

Simulating human bodies of different shapes in different poses with realistic cloth displacement is possible through PBS(Physically Based Simulation) method. DRAPE [3] (DRessing Any PErson) is

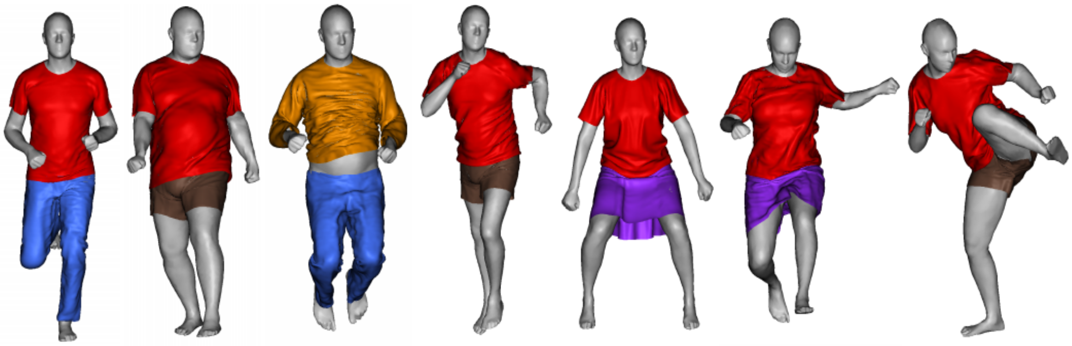

Fig. 1: *DRAPE*. sequence of dressed avatars in different poses. Image is taken from [3]

a model based on the physics simulation behind the cloth on bodies for different shapes and poses. This model has the property of learning cloth deformations in different shapes and poses. Given a body model with the known pose and shape parameters, the model takes input and learns the displacement of clothes iteratively to possess realistic wrinkles. DRAPE is used to dress animated avatars with a learned model of cloth dynamics. As this method is automated, we can use this to dress a large number of virtual characters with known shapes and pose parameters. A sequence of results is shown in Fig.1.

2.3 Dress 3D people in Generative Clothing

CAPE(Cloth Auto Person Encoding) learns displacement of clothes on human bodies. This model uses graph convolutions to simulate clothes, and cloth generation is a probabilistic task over SMPL body models. Ranjan et al. at [15] with up and down mesh sampling layers to capture the expressions in the human face. Although this architecture works well for human faces, the sampling layers make it difficult to capture the dynamics of the cloth, which are key in this aspect. In the CAPE model, they capture the local features with the help of a PatchGAN architecture to 3D meshes.

3 Implementation

To learn a model to predict clothed human bodies, we factorize the problem into two parts: (1) minimally clothed body mesh and (2) a clothing layer as a displacement on the human body. As the minimal clothed human body follows SMPL topology and is mostly used, we can extend the generated clothed human with the same number of vertices and geometry.

3.1 SMPL to Clothed Model

The SMPL[8] model factorizes the human body surface into shape (β) and pose (θ) parameters. SMPL follows triangulated mesh topology $\widetilde{T}$ in a rest pose, with N = 6890 vertices. The human skeleton is modeled with a chain of bones linked by n = 24 joints with a Degree of Freedom(DoF) of 3. The resulting pose vector has 3*24+3 = 75 parameters, including translation. Given a shape and pose parameters (β, θ), then new SMPL body T can be defined as:

$$T(\beta, \theta) = \widetilde{T} + D_s(\beta) + D_p(\theta) \tag{1}$$

where $D_s(\beta)$ corresponds to shape deformations and $D_s(\beta)$ corresponds to pose deformations. Following this we add clothing as an additional layer to the SMPL model with parameters: body pose θ, cloth type C, and latent variable z which represents cloth shape and structure. Let $L_{clo}(z, \theta, C)$ be the clothing layer. If we displace this clothing layer on top of SMPL body template T then the resulting displacement SMPL Template is defined as.

$$T_{clo}(\beta, \theta, C, z) = T(\beta, \theta) + L_{clo}(z, \theta, C) \tag{2}$$

Then we apply a skinning function to the Clothed SMPL body for different poses of the clothed human.

3.2 Input to network

Let $V_{minimal}$ be the vertices of a minimal clothed human body, and $V_{clothed}$ be the vertices of a clothed human body. The displacement layer achieved should be a graph V_{input} that follows the same SMPL topology. we can obtain V_{input} by subtracting a minimally clothed body mesh from the clothed body mesh in a canonical pose which is shown in Fig.2

$$V_{input} = V_{clothed} - V_{minimal} \tag{3}$$

The subtraction is performed per-vertex along all the feature dimensions. The result will have ideally non-zero values where there is cloth on the human body.

3.3 Architecture

The clothing Layer $L_{clo}(z, \theta, C)$ is a function of z, a learned feature space variable in a low dimensional space that has information about the cloth shape and its structure. This clothing layer is the V_{input} in Eqn3. The feature space is learned using a graph convolutional VAE-GAN network.

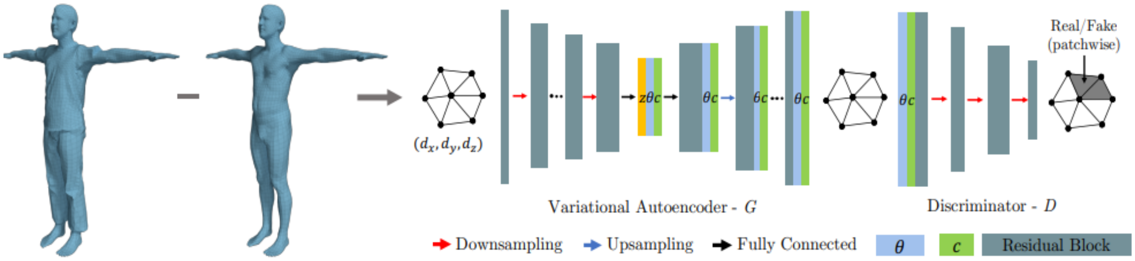

Fig. 2: *Graph VAE-GAN framework*. with cloth displacement as input produces the clothed human in different poses. Image is taken from [7]

As shown in Fig. 2, the model has a generator G with encoder and decoder, where z is the learned feature space of the input cloth displacement m and the output of the generator is reconstructed mesh $\tilde{m}$. The input to the decoder is conditioned on the pose θ and cloth type C, where it maps the low dimensional learned latent space to high dimensional. Leaky ReLU activation function is used in each layer with residual connections in the decoder network. Generator G has upsampling and downsampling layers. Decoder D has downsampling layers to match patch-wise the generated output and the ground truth. The network is differentiable, so back propagates in each epoch and updates the model weights.

Generator. In the training phase, an encoder takes the input V_{input} and maps the high dimensional cloth displacement in the low dimensional space z. During this process, graph convolutions at different layers capture the import features of the cloth displacement. A decoder is trained to map or reconstruct the input structure from the low dimensional latent space z. Both encoder and decoder are feed-forward neural nets with residual connections. Residual connections save the model from losing important cloth details like wrinkles. During the test phase, an encoder is not needed. We can sample the latent variable z from the Gaussian prior distribution, then only a decoder here acts as a generator.

Discriminator. A patchwise discriminator is used for calculating the loss between the ground truth and reconstructed mesh. The patchwise discriminator is used further to enhance the cloth detail capturing. Loss is not calculated on the global structure to classify the generated mesh as real or fake. It calculates the loss over local patches. So, the discriminator will have the ability to capture the fine details of the cloth. The discriminator has four graph convolution layers where the loss on vertices is calculated at each layer output. It allows the discriminator at each level to classify the generated mesh as real or fake.

Conditional model. We condition cloth type C and Pose θ as the conditional input to the decoder so that model learns better for which values the output mesh is varying. Cloth type C is a one-hot encoded vector, and the pose parameter is a rotational matrix constructed using Rodrigues equation. Layers at the end of the encoder and the beginning of the decoder are fully connected so that we can append the cloth type vector C and pose parameter matrix θ as an embedding to the latent variable z.

3.4 Loss functions

We compute reconstruction loss using L1 loss, not L2 because L1 loss encourages less smoothing comparatively. L1 loss of mesh m can be calculated by

$$\mathcal{L}_{reconstruction} = \mathbb{E}_{m \sim p(m), z \sim q(z|m)}[|G(z, \theta, C) - m|_1] \tag{4}$$

Let e be the edge in the edge-set ϵ that is the edge of the input mesh, $\tilde{e}$ be the edge in the edge-set $\tilde{\epsilon}$ that is the edge of the reconstructed mesh. While reconstruction, we minimize the edge loss because upsampling may cause edge linkage mismatch, which will result in wrinkles loss. We apply the L2 loss on edges because we are not aiming for a smooth surface.

$$\mathcal{L}_{edge} = \mathbb{E}_{e \in \epsilon, \tilde{e} \in \tilde{\epsilon}}[|e - \tilde{e}|_2] \tag{5}$$

KL divergence loss to calculate the divergence between the latent space z and the Gaussian prior distribution.

$$\mathcal{L}_{KL} = \mathbb{E}_{m \sim p(m)}[KL(q(z|m)||\mathcal{N}(0, I))] \tag{6}$$

Generator and Discriminator are trained on the GAN loss where generator G tries to minimize the loss against the discriminator D that tries to maximize.

$$\mathcal{L}_{GAN} = \mathbb{E}_{m \sim p(m)}[log(D(m, \theta, C)] + \mathbb{E}_{z \sim q(z|m)}[log(1 - D(G(z, \theta, C)))] \tag{7}$$

So total loss is calculated as the weighted sum of all these individual losses

$$\mathcal{L}_{total} = \mathcal{L}_{reconstruction} + \gamma_{edge}\mathcal{L}_{edge} + \gamma_{KL}\mathcal{L}_{KL} + \gamma_{GAN}\mathcal{L}_{GAN} \tag{8}$$

4 Evaluation Results

4.1 Experiment

We conducted experiments on three different parameters z, θ, C by sampling one parameter while keeping the other two parameters fixed. It will show how sampling affects the generated 3D human. We present quantitative and qualitative results in this section by conducting three different experiments on both male and female subjects. The dataset consists of raw data of each subject with different pose sequences like a flying eagle, basketball goal, topspin. Motion sequences of the subjects are captured at 60 fps. There are 31036 training samples and 5128 test samples for the male dataset and 21096 training samples and 5441 test samples for the female dataset.

4.2 Quantitative results

The experiments are conducted with γ_{KL}: 0.0008, γ_{edge}: 1.0, z: $\mathbb{R}^{18}$, θ: $\mathbb{R}^{24}$, C: $\mathbb{R}^{8}$, pose-type: rotational matrix using Rodriguous equation, learning rate: 0.008, batch size: 6.

Two experiments are conducted on male and female datasets with the number of epochs: 30. Later, to confirm we trained with the number of epochs: 60 and the results are almost the same. We can see different errors in the following table Table:1 trained on different datasets with a different setting.

Table 1: Error table

Subject	Error Mean	Error Median	Vertex Error	Edge error
female	0.00604	0.00461	0.29464	0.26473
male	0.00615	0.00466	0.29675	0.26750

4.3 Qualitative results

Sampling The results in Fig.3 are produced by sampling the latent space z, which means the outfit is displaced on the unseen sampled bodies. Here in this setting, we kept θ, and C fixed and sampled z every time. It has produced various outputs of cloth displacements. Also, captured the wrinkles and rendered them on unseen bodies.

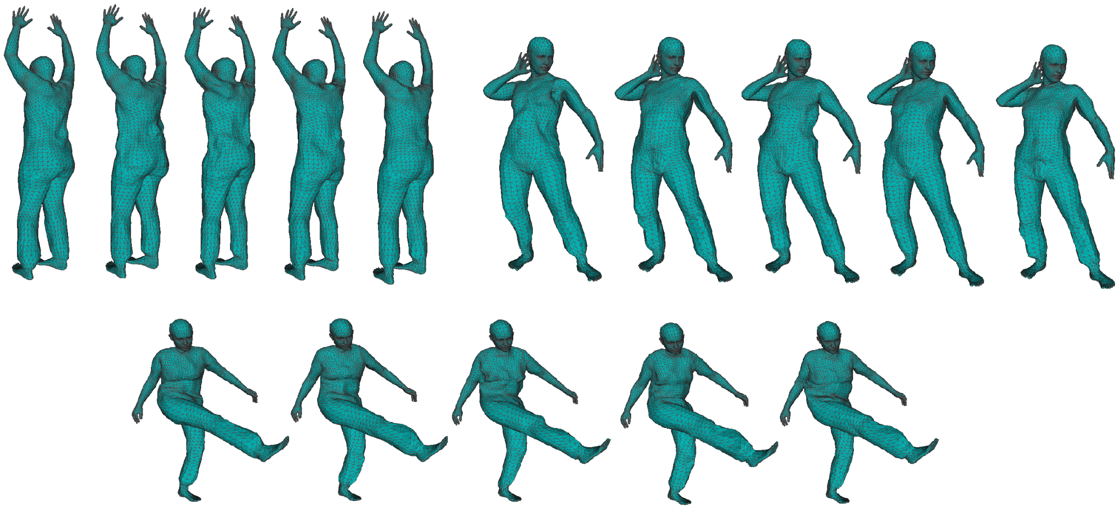

Fig. 3: Clothing sampled from the trained feature space z applied to the three unseen bodies by keeping the cloth parameter C and the pose parameter θ constant. z is sampled five times and produced different rendered displacements.

Pose-dependent deformation The results produced in Fig. 4 are produced by keeping z and C fixed, and θ sampled every time. We can see that cloth deformation changes for different poses.

Ground Truth vs Reconstructed The results produced in Fig. 5 are produced by merging reconstructed mesh and ground truth mesh. We did this to show how details have differed after reconstruction. The wrinkle details are not the same as the ground truth resulting in the color mismatch.

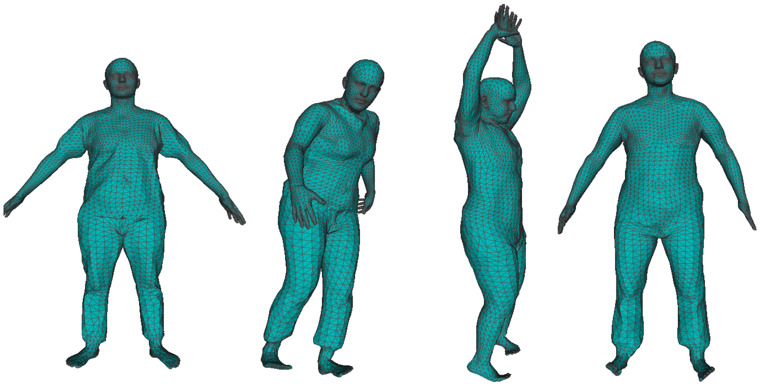

Fig. 4: Example of reconstruction by Z and C fixed, and varying θ

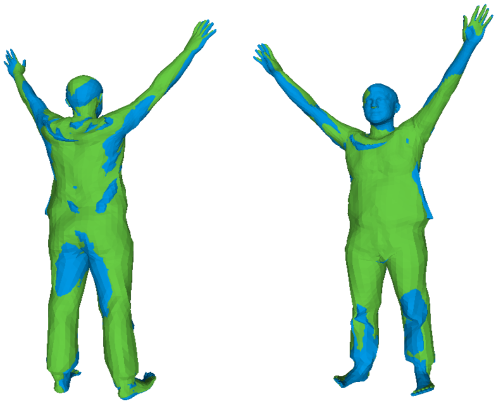

Fig. 5: Merged ground truth and reconstructed mesh to show the difference in details, pose and shape

5 Conclusion

In this paper, we discussed how we shape a model that can generate 3D humans by sampling different parameters like cloth, cloth displacement, and pose. This type of modeling a real-world problem is often used in many applications like fitness tracking, virtual cloth try-on, human pose estimation with clothing, etc.

In this approach, we get the displacement of cloth by taking the difference between the clothed human and the minimal clothed human. There are limitations in this approach where we cannot model clothes like skirts, single garments, etc. This often requires a different approach to get the garment offsets from the clothed human. Also, we cannot model shoes by this method because their geometry is significantly different from the finger and toes.

References

1. Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005.
2. Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2232–2241, 2019.
3. Peng Guan, Loretta Reiss, David A Hirshberg, Alexander Weiss, and Michael J Black. Drape: Dressing any person. *ACM Transactions on Graphics (TOG)*, 31(4):1–10, 2012.
4. Nils Hasler, Carsten Stoll, Bodo Rosenhahn, Thorsten Thormählen, and Hans-Peter Seidel. Estimating body shape of dressed humans. *Computers & Graphics*, 33(3):211–216, 2009.
5. Nikolay Jetchev and Urs Bergmann. The conditional analogy gan: Swapping fashion articles on people images. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2287–2292, 2017.
6. Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8320–8329, 2018.
7. Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3D people in generative clothing. In *Computer Vision and Pattern Recognition (CVPR)*, 2020.
8. Meysam Madadi, Hugo Bertiche, and Sergio Escalera. Smplr: Deep learning based smpl reverse for 3d human pose and shape recovery. *Pattern Recognition*, 106:107472, 2020.
9. Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017.
10. Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4480–4490, 2019.
11. Alexandros Neophytou and Adrian Hilton. A layered model of human body and garment deformation. In *2014 2nd International Conference on 3D Vision*, volume 1, pages 171–178. IEEE, 2014.
12. Sungheon Park, Jihye Hwang, and Nojun Kwak. 3d human pose estimation using convolutional neural networks with 2d pose information. In *European Conference on Computer Vision*, pages 156–169. Springer, 2016.
13. Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018.
14. Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 36(4), 2017. Two first authors contributed equally.
15. Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3d faces using convolutional mesh autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 704–720, 2018.
16. Carsten Stoll, Juergen Gall, Edilson De Aguiar, Sebastian Thrun, and Christian Theobalt. Video-based reconstruction of animatable human characters. *ACM Transactions on Graphics (TOG)*, 29(6):1–10, 2010.
17. Gökhan Yildirim, Calvin Seward, and Urs Bergmann. Disentangling multiple conditional inputs in gans. *arXiv preprint arXiv:1806.07819*, 2018.