Pixel-Aligned Implicit Functions on Dynamic Meshes

AYUSH AGARWAL and GLEB SHEVCHUK

Pixel-aligned implicit functions (PiFU) are an exiting new method for converting 2D images to 3D human meshes. However, they have only been shown to work on simple, static meshes. To test the limits of this approach, we train it on more dynamic meshes from the SURREAL dataset. We then propose Temporal PiFU, a model that leverages temporal frame neighborhoods to improve 3D mesh prediction. We then compare this approach to the baseline PiFU model on video sequences from SURREAL.

ACM Reference Format:
Ayush Agarwal and Gleb Shevchuk. 2020. Pixel-Aligned Implicit Functions on Dynamic Meshes. *ACM Trans. Graph.* 1, 1 (December 2020), 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Pixel-aligned implicit functions are a new method for converting 2D images into 3D meshes. Unlike keypoint-based methods, PIFu does not require prior information about a mesh. Unlike voxel-based methods, PIFu is resolution-agnostic and can learn accurate reconstructions of a mesh across multiple scales. PIFu's strength lies in its representation – instead of learning a dense representation of a mesh, PIFu learns an implicit function that maps 3D points to whether they exist inside the mesh or not. This allows us to form a direct correspondence between local features aggregated at the pixel level with global 3D shape coordinates. In addition, this allows us to train a model in a sampling-based manner and learn a data-efficient occupancy field. To use PIFu, we simply have to sample points from 3D space, use PIFu's binary outputs to create an occupancy field, then use traditional marching cubes to reconstruct a mesh from that field. This process is shown in Figure 3

However, it is unclear how this approach handles more dynamic, moving meshes like those found in the real world. In the original paper[Saito et al. 2019] and subsequent extensions[Saito et al. 2020], PIFu was only tested on standing meshes exhibiting neutral poses. So, to test the limits of this approach, we first re-train PIFu on the more diverse SURREAL dataset[Varol et al. 2017], which uses motion capture data to animate meshes performing cartwheels, lunges, and other lifelike behaviors. Then, by extending PIFu to include temporal information, we show improved performance on the SURREAL dataset.

Authors' address: Ayush Agarwal, ayush@stanford.edu; Gleb Shevchuk, glebs@stanford.edu.

0730-0301/2020/12-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

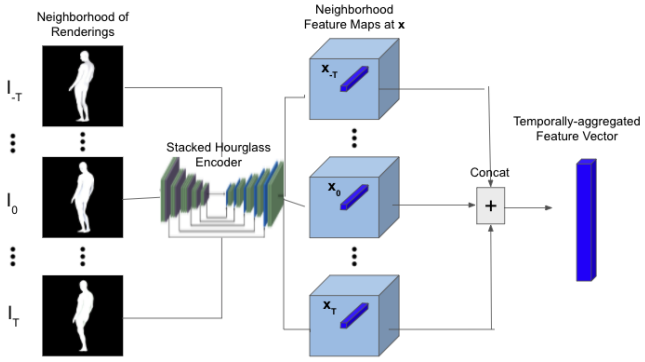

Fig. 1. Our proposed Temporal PIFu extracts features from a neighborhood of renderings around a given central rendering, I_0, for which we are trying to recover the 3D structure for. The feature vectors at the projected position x are aggregated across the neighborhood and passed into the query step.

2 RELATED WORK

Though implicit functions have only recently been adapted to mesh reconstruction, significant prior work has explored 2D image-3D mesh conversion.

One popular approach is to leverage the SMPL parametric human body model[Loper et al. 2015] or other parameteric body representations and regress images to them. The SMPL model, for example, remains popular because it contains fewer than one hundred shape and pose parameters, making it tractable for optimization. Recent work has sought to extend parametric modeling across multiple fronts, including by constructing a differentiable SMPL layer[Kanazawa et al. 2018], performing parametric estimation by training autoregressive models[Zhang et al. 2019], and performing model regression using self-supervision[Tung et al. 2017].

Another approach is to only learn 3D poses from images. Several works estimate body joint positions and skeletons directly in a supervised manner[Pavllo et al. 2019], but they often suffer because they rely on data captured in controlled, indoor settings[Ionescu et al. 2013]).

Other methods attempt to learn volumetric, graph-based, or implicit representations of 3D meshes. In volumetric learning, several works have shown that 3D meshes can be inferred directly from images[Alp Güler et al. 2018][Zhu et al. 2019]. However, some volumetric approaches have been found to be too memory intensive [Varol et al. 2018]. In graph-based learning, groups have shown that graph CNNs can effectively regress SMPL parameters[Kolotouros et al. 2019] and that coarse-to-fine graph CNNs are capable of direct 3D mesh reconstruction [Choi et al. 2020].

Finally, in implicit-function learning, Occupancy Networks [Mescheder et al. 2019], DeepSDFs [Park et al. 2019], and DIST [Liu et al. 2020] have all been proposed to perform human mesh reconstruction. However, they all leverage a single global representation of images to generate the 3D reconstructions and cannot articulate arms,legs,

Fig. 2. While the original PIFu paper was trained on the RenderPeople dataset, which consists of static meshes, we utilized the more dynamic SURREAL dataset.

Parameter	Value
Z normalization factor	2.0
Normalization	group
Num hourglass stacks	4
Num stacked layers	2
Skip connection in hourglass	true
Hourglass dimensions	256
Single-Frame PIFu MLP dimensions	257, 1024, 512, 256, 129, 1
Temporal PIFu MLP dimensions	769, 1024, 512, 256, 129, 1
Last activation	tanh
Temporal PIFu Neighborhood Size	3
Learning Rate	1e-4
Optimizer	Adam
Batch size	2
Num samples	1000
Training meshes	892
Test meshes	100

Table 1. PIFu parameters

or other fine details well. However, PIFu [Saito et al. 2019] and DISN [Xu et al. 2019] overcome this issue by leveraging pixel-aligned 2D local features by projecting 3D query points into 2D feature space. However, PIFu is only trained on static standing poses using expensive ground truth meshes. Also, when applied to videos, PIFu is temporally inconsistent and often leads to large jumps in movement and holes in a mesh.

3 METHODS

Our work aims to aggregate temporal information when learning an occupancy field based on PIFu from video and to train PIFu on dynamic poses. To accomplish this, we train a baseline and temporal version of PIFu on the SURREAL dataset.

3.1 Dataset

Unlike the original PIFu paper, which uses static meshes from RenderPeople, we utilize meshes taken from the SURREAL dataset[Varol et al. 2017] which includes diverse meshes moving according to motion capture data across a wide variety of synthetic environments. Though this dataset also includes a wide variety of information about a scene, we only utilize ground truth meshes and images of each scene. To train PIFu on the SURREAL data, we first had to pre-process it to match the original PIFu requirements. Though we initially tried running PIFu directly on segmented images from SURREAL, we found that the difference in camera and rendering parameters, in addition to the lack of multi-view data, rendered the approach unsuccessful. Therefore, we had to re-process the SURREAL meshes so that they were properly scaled, used the same virtual camera settings as in the original paper, and included images from multiple viewing angles. To make this process simpler, we decided to render these without the original mesh textures. Finally, because this data processing approach required us to create a new mesh for each frame, we only sampled a selection of frames from each video in the SURREAL dataset, significantly decreasing its size by a factor of 5.

3.2 Network Architecture

The original PIFu network consists of two distinct steps: filtering and querying. First, we pass a white-texture rendered image through an hourglass network and we accumulate a set of features for each pixel in the image. Next during the query phase, we pass our training 3D point in, project it into image space, then use the features extracted from the hourglass network at the point to decide whether the point is inside or outside of the mesh.

More formally, we train both an hourglass feature extractor, F, and a PIFu classifier, P. For an input rendering I_0 and a query point X with 2D projection x, the classifier is trained to predict

$$P(F(I_0)_x, z(X)) = \begin{cases} 1 & X \text{ is inside mesh surface} \\ 0 & \text{otherwise} \end{cases}$$

where $F(I_0)_x$ is the feature vector at the projected position x from the stacked hourglass feature map and $z(X)$ is the depth coordinate of X in camera space. This allows the classifier to exploit local 2D details when generating 3D shape.

We also propose an extension to PIFu for human shape reconstruction from video called Temporal PIFu that aggregates features using F temporally across a neighborhood of frames as depicted in Figure 1. In Temporal PIFu, this temporally aggregated feature vector is passed to the classifier P instead of the single feature vector in the original PIFu.

For our implementation of the single-frame PIFu and temporal PIFu, we used roughly the same settings as in [Saito et al. 2019], which are detailed in Table

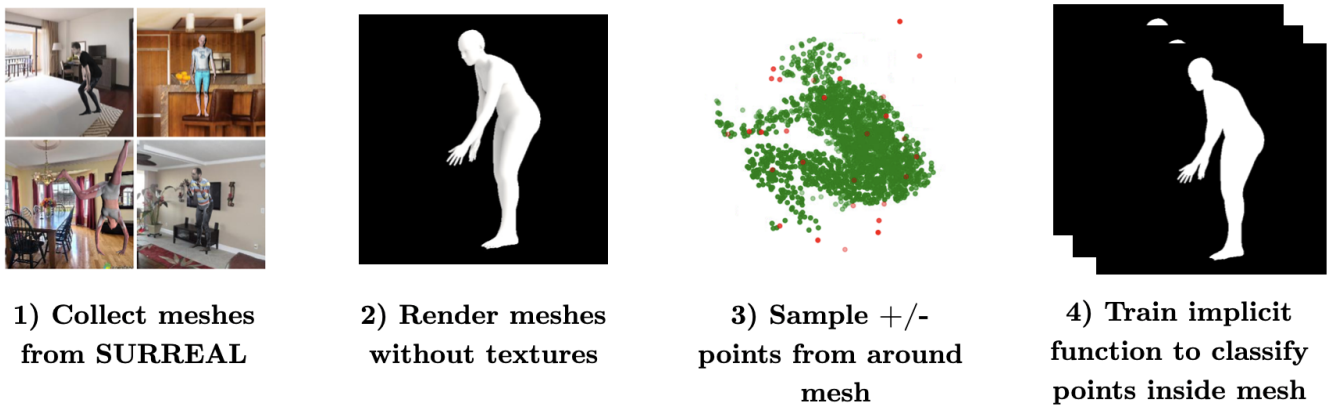

Fig. 3. This pipeline summarizes how we construct training data for our baseline and temporal PIFu models.

Test metric	Baseline PIFu	Temporal PIFu
IOU	0.8118	0.8599
MSE	0.0649	0.0462
Precision	0.8565	0.8980
Accuracy	0.9381	0.9515
Test meshes	100	100

Table 2. Test results

3.3 Training and Evaluation

In order to train the network, we use a standard Mean Squared Error (MSE) loss between the in-mesh/out-of-mesh label predicted by the network and the true labels for all 3D points sampled from a mesh. We mimic the sampling scheme of the original PIFu paper when sampling training points in 3D by both sampling near the mesh surface and sampling random points within a fixed rectangular box.

To further propagate this loss through the hourglass network, we store predictions from all intermediate layers of the network and augment the standard loss, which is between the true labels and the predictions of the final hourglass layer, with the loss between the true labels and predictions at each intermediate layer. Further training details are included in Table 1. We then evaluate the performance of this network on four metrics: mean squared error (MSE), Intersection over Union (IOU), precision, and recall.

4 RESULTS

4.1 Single Frame PIFu

To understand the performance of the original PIFu network, we first benchmarked it by training on the SURREAL dataset. As seen in the quantitative results in Table 2, we achieve surprisingly good performance across all metrics on the set of test meshes, achieving fairly high accuracy of 94 percent. This indicates that, out of all predictions made by the model, 94 percent were correct. These results, however, paint a false picture for the performance of baseline PIFu. This becomes apparent when we look at the single-frame qualitative results in Figure **??**. Here, we see that the model is able to correctly reconstruct most images from a single viewpoint but fails to learn the correct 3D structure of the model when viewed from other angles. In the original PIFu paper[Saito et al. 2019], the authors fixed this issue by forming models from multiple viewpoints at a time. However, this does not fit real world use cases.

These results also uncover several significant issues with the single-view PIFu formulation. First, because most of the loss signal comes from heavily sampled parts of a mesh, the PIFu network is incenvitized to correctly classify points near the main body of the mesh and less so with classifying points on mesh appendages. Second, because the points are projected into 3D image space from the image perspective, there are no additional mechanisms for a model to correctly classify points from other viewing angles.

4.2 Temporal

When we benchmark our proposed Temporal PIFu on the same test set as the single-frame PIFu, we see a significant increase in performance with a 4.7% gain in IOU and 30% decrease in MSE loss. This is a promising sign that aggregating temporal information is beneficial to reconstruction, probably because we can utilize information like occlusions that are absent in the single-frame setting. Qualitatively, we compare the output mesh after marching cubes of the temporal PIFu versus the single-frame PIFu in Figure 5. We can see that Temporal PIFu appears significantly closer to the ground truth with far fewer holes in the reconstructed mesh and shows more consistent movement across frames (e.g. the right arm is consistent in temporal PIFu but not in single-frame PIFu). However, there is still room for improvement since arms and feet are still poorly predicted.

5 FUTURE WORK

This exploration points to a number of future directions. First, it is worth further investigating the impact that temporal information makes on mesh reconstruction and whether we can utilize pre-scripted motions to improve it. Next, it would be interesting to combine multiple representations to learn more dynamic, lifelike meshes. For example, if two separate models are learned for body and face reconstruction, this could allow more flexibility for 3D modellers and other graphics artists when virtualizing human characters.

Finally, performing style and mesh transfer on learned meshes can decrease the burden of reconstruction and lead to exciting real-time use cases.

REFERENCES

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7297–7306.

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. *arXiv preprint arXiv:2008.09047* (2020).

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2013. 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 (2013), 1325–1339.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7122–7131.

Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. 2019. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4501–4510.

Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. 2020. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019–2028.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)* 34, 6 (2015), 1–16.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4460–4470.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 165–174.

Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7753–7762.

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. 2019. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE International Conference on Computer Vision*. 2304–2314.

Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 2020. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 84–93.

Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. 2017. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems*. 5236–5246.

Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. 2018. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 20–36.

Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. 2017. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 109–117.

Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. 2019. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In *Advances in Neural Information Processing Systems*. 492–502.

Jason Y Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. 2019. Predicting 3d human dynamics from video. In *Proceedings of the IEEE International Conference on Computer Vision*. 7114–7123.

Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. 2019. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4491–4500.

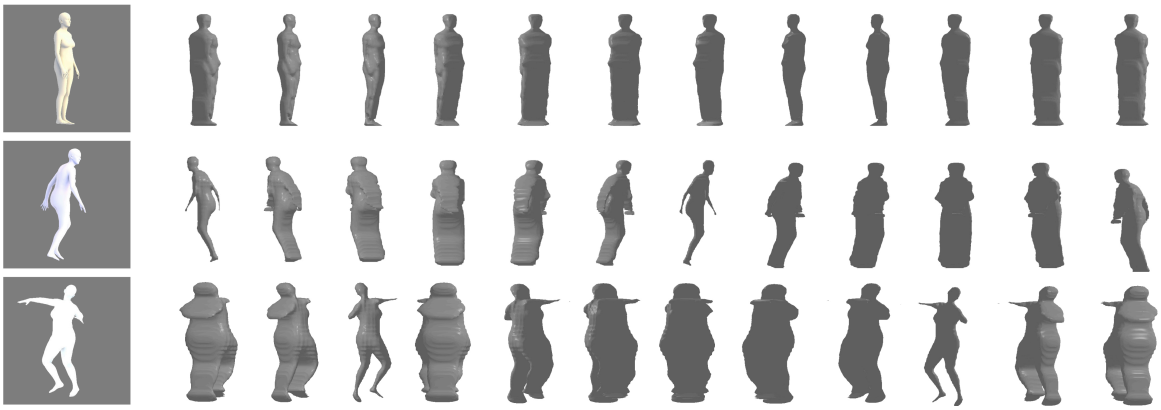

Fig. 4. We visualize predictions from Single-Frame PIFu on three random test images and show different viewing angles of the resulting mesh.

Fig. 5. We visualize predictions from the Single-Frame and Temporal PIFu on a randomly sampled video sequence and compare to the ground truth SURREAL meshes.