

# EgoRenderer: Rendering Human Avatars from Egocentric Camera Images

Tao Hu<sup>1\*</sup>, Kripasindhu Sarkar<sup>2</sup>, Lingjie Liu<sup>2</sup>, Matthias Zwicker<sup>1</sup>, Christian Theobalt<sup>2</sup>

<sup>1</sup>Department of Computer Science, University of Maryland, College Park

<sup>2</sup>Max Plank Institute for Informatics, Saarland Informatics Campus

## Abstract

We present *EgoRenderer*, a system for rendering full-body neural avatars of a person captured by a wearable, egocentric fisheye camera that is mounted on a cap or a VR headset. Our system renders photorealistic novel views of the actor and her motion from arbitrary virtual camera locations. Rendering full-body avatars from such egocentric images come with unique challenges due to the top-down view and large distortions. We tackle these challenges by decomposing the rendering process into several steps, including texture synthesis, pose construction, and neural image translation. For texture synthesis, we propose *EgoDPNet*, a neural network that infers dense correspondences between the input fisheye images and an underlying parametric body model, and to extract textures from egocentric inputs. In addition, to encode dynamic appearances, our approach also learns an implicit texture stack that captures detailed appearance variation across poses and viewpoints. For correct pose generation, we first estimate body pose from the egocentric view using a parametric model. We then synthesize an external free-viewpoint pose image by projecting the parametric model to the user-specified target viewpoint. We next combine the target pose image and the textures into a combined feature image, which is transformed into the output color image using a neural image translation network. Experimental evaluations show that *EgoRenderer* is capable of generating realistic free-viewpoint avatars of a person wearing an egocentric camera. Comparisons to several baselines demonstrate the advantages of our approach.

## 1. Introduction

The goal of this work is to render full-body avatars with realistic appearance and motion of a person wearing an egocentric fisheye camera from arbitrary external camera viewpoints (Figure 1). Such egocentric capture and rendering

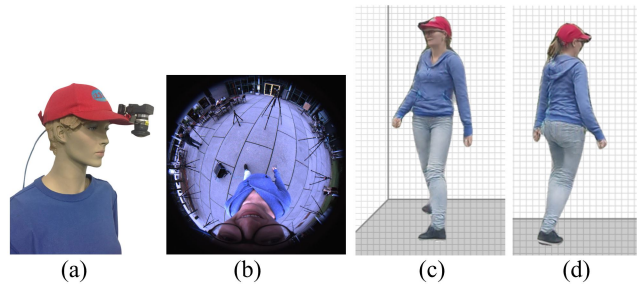


Figure 1: Based on a wearable fisheye camera setup (a), we propose *EgoRenderer*, which is trained for a single person and can produce full-body avatars of the person from new viewpoints and in new poses (c)(d) by taking as input an egocentric image (b) captured by the fisheye camera.

enables new applications in sport performance analysis or health care. Real-time free-viewpoint rendering of self-embodied avatars is also important in virtual reality (VR) and augmented reality (AR) applications, notably telepresence. A key advantage of our approach is that it uses a lightweight and compact sensor that could be mounted to glasses, headsets or caps, and that it is fully mobile. Therefore, actors can freely roam and are not limited to stay in confined spaces visible to external multi-camera setups.

We approach free-viewpoint neural avatar rendering from our egocentric view by a combination of new solutions for egocentric pose estimation, appearance transfer, and free-viewpoint neural rendering; each of these need to be tailored to the challenging egocentric top-down fish-eye perspective with strong distortions and self-occlusions. Most established pose estimation methods employ external outside-in camera views [27, 28, 37, 7, 9] and are not directly applicable to our setting.

Some recent approaches are designed to estimate 3D skeletal pose from head-mounted fisheye cameras [54, 49]. However, our setting requires a denser pixel wise estimation of egocentric pose and shape, as dense correspondences are prerequisite to transfer the texture appearance of a person from egocentric to external views (Figure 3). Similarly, re-

\*Work partly conducted during TH's internship at MPI-INF

cent neural rendering-based pose transfer methods enable creation of highly realistic animation videos of humans under user-specified target motion [4, 1, 19, 51]. However, all of these are tailored to external outside-in views, such that target motions already need to be specified as skeletal pose or template mesh sequences from the extrinsic camera view. We face the additional challenge of transferring appearance and pose from the starkly distorted egocentric view to the external view. To enable highly realistic appearance and pose transfer of the actor wearing the camera to an arbitrary external view, even in more general scene conditions, EgoRenderer decomposes the rendering pipeline into **texture synthesis**, **pose construction**, and **neural image translation**, as shown in Figure 4.

**Texture Synthesis.** In contrast to most aforementioned image-based pose transfer methods for outside-in views [4, 38], EgoRenderer explicitly builds an estimation of surface textures of a person on top of a parametric body model. Specifically, we extract explicit (color) textures from egocentric images and learn implicit textures from a multi-view dataset in a training phase. We then combine them to form our full texture representation of the person. Compared to static color texture maps, the learned implicit textures better capture detailed appearance variation across poses and viewpoints. To extract (partial) textures for visible body parts from egocentric images, we create a large synthetic dataset (see Figure 2) and train an Ego-DPNet network tailored to our setup to infer the dense correspondences between the input egocentric images and an underlying parametric body model, as shown in Figure 3.

**Pose Construction.** Different from earlier neural human rendering methods that expect target poses as input, irrespective of where these targets are from [38], EgoRenderer works end-to-end. We have to exactly reproduce pose and appearance seen in an egocentric image from any external viewpoint. We support neural rendering of the target view by projecting the 3D parametric model from egocentric camera space to the target viewpoint, enabling us to also transfer the partially visible texture appearance.

**Neural Image Translation.** Pose construction enables us to render the 3D model in the desired external view using both implicit textures and color textures. We transform these images into the final color image by means of a neural image translation network. Experiments show that this implicit to explicit rendering approach outperforms direct pose-to-image translation.

Our qualitative and quantitative evaluations demonstrate that our EgoRenderer system generalizes better to novel viewpoints and poses than baseline methods on our test set. To summarize, our contributions are as follows:

1) A large synthetic ground truth training dataset of top-down fisheye images and an Ego-DPNet network tailored to our fisheye camera setup to predict dense correspondence to

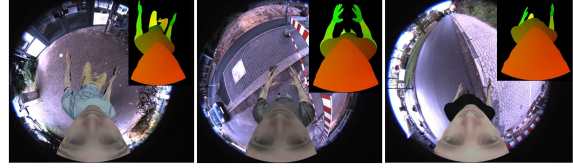


Figure 2: Examples from our synthetically rendered fisheye training dataset (top-right: ground truth DensePose). Our dataset features a large variety of poses, human body appearance, and realistic backgrounds.

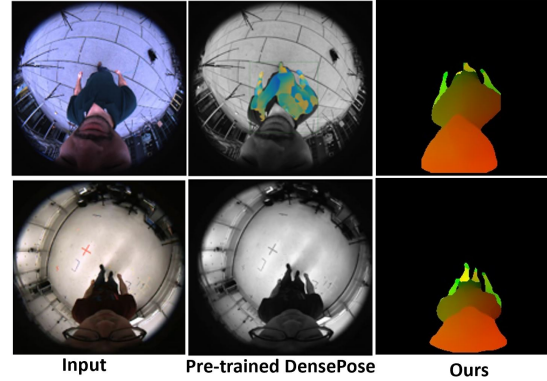


Figure 3: DensePose [9] performs poorly on images captured by our setup, sometimes failing to detect the human (second row). Our DensePose predictions are on the right.

a parametric body model from egocentric images.

2) An end-to-end EgoRenderer system that takes single egocentric images as input and generates full-body avatar renderings from external user-defined viewpoints.

## 2. Related Work

Our approach is closely related to many sub-fields of visual computing, and below we discuss a small subset of these connections.

**Neural Rendering.** Neural rendering is a class of deep image and video generation approaches that combines generative machine learning techniques with physical knowledge from computer graphics to obtain controllable outputs. Many neural rendering methods [43, 44, 29, 16, 47] learn implicit representations of scene instead of explicitly modeling geometry, such as DeepVoxels [43], SRNs [44], NeRF [29] and NSVF [16]. However, only a few neural scene representations handle dynamic scenes [12, 22]. Our neural rendering approach is inspired by the aforementioned scene specific methods. To handle the dynamic nature of the scene, we learn a person specific implicit texture map on a parametric model of humans. This learned implicit texture stack, along with the pose dependent appearance from the egocentric camera, is used by a high fidelity generator to produce realistic renderings of the person.

**Pose Transfer and Human Re-enactment.** Pose trans-

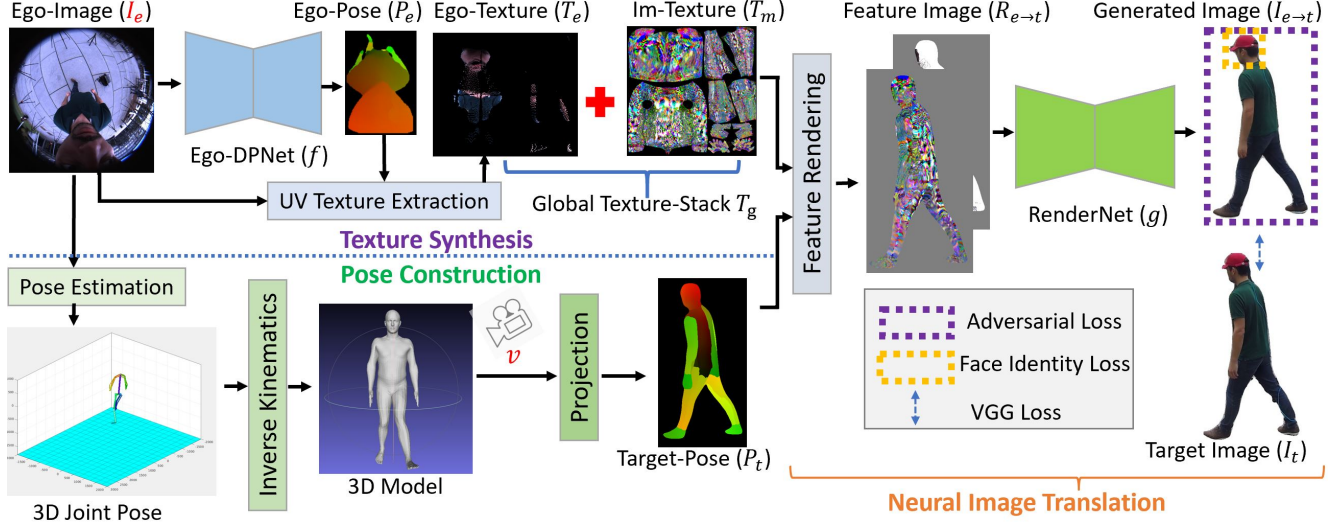


Figure 4: Pipeline overview. Given an egocentric image  $I_e$  of a person and a user-defined viewpoint  $v$ , EgoRenderer synthesizes the full body avatar  $I_{e \rightarrow t}$  from viewpoint  $v$ . EgoRenderer is custom tailored to our cap-mounted fisheye camera setup, and decomposes the rendering process into texture synthesis, pose construction and neural image translation. For texture synthesis, given an input image  $I_e$  of a person in egocentric camera space, we first predict the dense correspondences ( $P_e$ ) between the input image  $I_e$  and an underlying parametric mesh model using Ego-DPNet, which allows a partial UV Texture-map  $T_e$  to be extracted for the body regions visible in the image. We also learn an implicit texture stack ( $T_m$ ) during training, and concatenate  $T_e$  and  $T_m$  as the global texture representation  $T_g$ . For pose construction, given a user-defined viewpoint  $v$ , we synthesize a target pose image  $P_t$  by projecting the parametric model from egocentric space to the target viewpoint. In the *Feature Rendering* step, the parametric body mesh is textured with the global texture stack  $T_g$  to produce intermediate feature images  $R_{e \rightarrow t}$ . An image-translation network RenderNet converts the feature images to the final realistic image  $I_{e \rightarrow t}$ .

fer, first introduced by [24], refers to the problem of re-rendering a person from a different viewpoint and pose from the appearance of a single image. Most approaches formulate this problem as an image-to-image mapping problem, i.e. given a reference image of a target person, mapping the body pose in the form of renderings of a skeleton [4, 39, 33, 14, 58], dense mesh [18, 51, 17, 35, 31, 8] or joint position heatmaps [24, 1, 25], to real images. To better map the appearance of the reference to the generated image, some methods [17, 35] first map the person appearance in screen space to UV space and feed the rendering of the person in the target pose with the UV texture map into an image translation network. However, these methods generally works in external camera setup with regular viewpoints. We propose a new method that extracts appearance from the top-down egocentric fisheye camera using a novel network, and combines it with a learned person specific neural texture for a high fidelity generation.

**Egocentric Systems.** Because of their mobility and flexibility, egocentric systems have made significant progress in recent years. Applications of egocentric systems can be divided into face, gesture, and full body. [6, 5, 48, 21, 3, 46, 15] study face estimation with a head-mounted camera. [41, 45, 26, 2, 30, 32, 42] perform gesture and activity recognition by using a head-mounted or chest-worn cam-

era. In the case of full body, most of the methods either opt for an inside-out configuration [36, 10, 55, 56], or employ fisheye cameras for a large field of view [54, 34]. Rhodin et al. [34] propose the first full-body capture method with a helmet-mounted stereo fisheye camera pair. Xu et al. [54] and Tome et al. [49] use a much more compact and flexible monocular setting for full body pose estimation. However, our setting learns a denser pixel wise estimation of egocentric pose and shape (see Figure 3).

### 3. Approach

Our EgoRenderer system is built on a single cap-mounted fisheye camera (similar to [54]) that captures egocentric images of persons at run time. Given an egocentric image  $I_e$  of a person, we render a full body avatar of the person from a user-defined viewpoint, which we call external free-viewpoint space in this paper. Our system decomposes the rendering process into texture synthesis, pose construction, and neural image translation, as shown in Figure 4. First, for texture synthesis, we build a global texture stack  $T_g$  that combines explicit textures  $T_e$  from egocentric images and an implicit texture stack  $T_m$ , which is learned in a training phase, as the texture representations of the person. Our system represents body pose and shape as a



parametric mesh [23]. For pose construction, given a user-defined target viewpoint, we synthesize a target pose image by projecting the parametric model from egocentric camera space to the target viewpoint. The third step renders the 3D model with the global texture stack and generates implicit feature images  $R_{e \rightarrow t}$  of the person, which are then transformed to the final rendered color image  $R_{e \rightarrow t}$  by a neural image-translation network *RenderNet* in the fourth step.

### 3.1. Input and Output

Our input is an egocentric image  $I_e$  of a person and a target viewpoint  $v$ , and the output is a photorealistic image  $I_{e \rightarrow t}$  of the person from the target viewpoint. In our training stage, we take pairs of images ( $I_e$ ,  $I_t$ ) of the same person as input. In this paper, we call  $I_e$  images from egocentric space, and  $I_t$  images from external free-viewpoint space.

### 3.2. Texture Synthesis

**Extracting Partial UV Texture Maps from Egocentric Space.** The pixels of the input egocentric images are transformed to UV texture space through dense correspondences between the input images and an underlying SMPL [23] model. DensePose [9] (pre-trained on the COCO-DensePose dataset), can predict 24 body segments and their part specific UV coordinates on images captured with regular cameras and mostly from chest high viewpoints. However, DensePose [9] fails on the egocentric images in our setup, as shown in Figure 3.

Capturing a large amount of annotated DensePose data for egocentric data is a mammoth task that would require large amounts of human labor. To solve this problem, we rendered a large, synthetic egocentric fisheye training dataset that enables training of a deep neural network to predict dense correspondences on our hardware setup. Our dataset contains 178,800 synthetic images with ground truth UV coordinate maps, and features various poses, body appearances and backgrounds, as shown in Figure 2. With this synthetic dataset, we train an Ego-DPNet ( $f$ ) network to predict the DensePose  $P_e$ , that is,  $P_e = f(I_e)$ . More details of the dataset can be found in the supplementary material.

Ego-DPNet was built on the DensePose-RCNN architecture, and we trained it in multi-stage using transfer learning. Different from the original DensePose-RCNN, which takes sparse annotated ground truth points and relies on an inpainting teacher network to interpolate dense ground truth, our Ego-DPNet directly takes synthetic images as dense ground truth without the teacher network. In addition, another difference is that the input image size of Ego-DPNet is fixed in training and testing due to our specific fisheye camera setup. The performance of Ego-DPNet is shown in Figure 3. With the DensePose prediction  $P_e$  and the input image  $I_e$ , we extract the partial UV texture map  $T_e$  by an *UV Texture Extraction* module  $u$ , that is,  $T_e = u(I_e, P_e)$ .

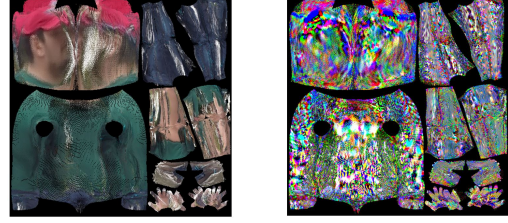


Figure 5: The impact of learning on the implicit texture stack  $T_m$ . Left: initial state; right: final state.

**Learning Implicit Textures in Training.** Besides textures from egocentric space, we also learn a  $d$ -dimensional implicit texture stack ( $T_m$ ) from training images to capture detailed appearance variation across poses and viewpoints. Note that we initialize  $T_m$  with an explicit texture map that is the average texture of the training images. The initial and final state of  $T_m$  during training is shown in Figure 5. We provide our experiments with 3 channels in  $T_m$ .

**Global Texture Stack.** We concatenate explicit textures  $T_e$  and implicit textures  $T_m$  into the *global texture stack*  $T_g = [T_e, T_m]$  to represent dynamic textures of the person. In our paper,  $T_g$  has 6 channels, 3 from  $T_e$  and 3 from  $T_m$ .

### 3.3. Pose Construction

Given an egocentric image  $I_e$  and a target user-defined viewpoint  $v$ , the second step is to synthesize the target pose of the person under the viewpoint, which is represented by a DensePose image  $P_t$  in Figure 4. First, a pose estimation module (Mo2Cap2 [54] in our experiments) is used to extract the 3D joint pose of the person, which is used to drive a 3D SMPL model by *Inverse Kinematics*. The target pose is then rendered by projecting the 3D model to viewpoint  $v$ .

### 3.4. Intermediate Feature Image Rendering

Using the global texture stack  $T_g = [T_e, T_m]$  and the target pose image  $P_t$ , we use a *Feature Rendering* operation  $r$  to produce a 6-dimensional feature image  $R_{e \rightarrow t}$ , that is,  $R_{e \rightarrow t} = r(T_e, T_m, P_t)$ . The operation  $r$  is implemented by differentiable, bilinear sampling in our experiments.

### 3.5. Neural Image Translation

In the final step, the feature image  $R_{e \rightarrow t}$  is translated to a realistic image  $I_{e \rightarrow t}$  using a translation network  $g$ , which we call *RenderNet*,  $I_{e \rightarrow t} = g(R_{e \rightarrow t})$ . *RenderNet* is built on the Pix2PixHD [52] architecture. The discriminator for adversarial training of *RenderNet* also uses the multiscale design of Pix2PixHD, and we use a three scale discriminator network for the adversarial training.

### 3.6. Training Details and Loss Functions

We train the EgoRenderer in two stages. We first train the Ego-DPNet on our synthetic dataset, and then train *RenderNet* on real data. Note that *RenderNet* is person specific.

**Ego-DPNet.** For better generalization on real world imagery, we train Ego-DPNet in multiple stages using transfer learning. It was first pre-trained on the COCO-DensePose dataset [9] to learn good low-level features from real images with usual camera optics. Then we fine tune it on our synthetic dataset to predict DensePose on egocentric images.

**RenderNet.** In training our system takes pairs of egocentric inputs and ground truth images ( $I_e, I_t$ ) of the same person as input. The output of RenderNet can be expressed as

$$I_{e \rightarrow t} = g \circ r(u(I_e, f(I_e)), T_m, P_t),$$

where all operations  $g, r, f, u$  are differentiable. For speed, we pre-compute Ego-Pose  $P_e = f(I_e)$  and  $P_t$ , and directly read them as input in training. We optimize the parameters of RenderNet  $g$  and the implicit texture stack  $T_m$ .

**Loss Functions.** We apply the combination of the following loss functions to train RenderNet:

- **Perceptual Loss.** We use a perceptual loss based on the VGG Network [11], which measures the difference between the activations on different layers of the pretrained VGG network [40] applied on the generated image  $I_{e \rightarrow t}$  and ground truth target image  $I_t$ ,

$$L_p = \sum \frac{1}{N_j} |p^j(I_{e \rightarrow t}) - p^j(I_t)|,$$

where  $p^j$  is the activation and  $N_j$  the number of elements of the  $j$ -th layer in the pretrained VGG network.

- **Adversarial Loss.** We use a multiscale discriminator  $D$  of Pix2PixHD [52] to leverage an adversarial loss  $L_{adv}$  in our system.  $D$  is conditioned on both the generated image and rendered feature image.

- **Face Identity Loss.** We use a pre-trained network to ensure that RenderNet and the implicit texture stack preserve the face identity on the cropped face of the generated and ground truth image,

$$L_{face} = |N_{face}(I_{e \rightarrow t}) - N_{face}(I_t)|,$$

where,  $N_{face}$  is the pretrained SphereFaceNet [20]. The final loss is then

$$L_G = \lambda_p L_p + \lambda_{face} L_{face} + \lambda_{GAN} L_{adv}.$$

The networks are trained using the Adam optimizer [13] with an initial learning rate of  $2 \times 10^{-4}$ ,  $\beta_1 = 0.5$ . The loss weights are set empirically to  $\lambda_{GAN} = 1$ ,  $\lambda_p = 10$ ,  $\lambda_{face} = 5$ . Note that the initial learning rate of  $T_m$  is  $2 \times 10^{-3}$ , 10 times of the learning rate of RenderNet.

## 4. Experiments

**Datasets.** Since there are no public datasets suitable for our project, we captured 4 datasets by ourselves, and the rendered human avatars in our study are shown in Figure 6. We refer to them as H1, H2, H3 and H4. H1 and H2 were captured in outdoor scenes, while H3 and H4 were captured in a studio. We have 11 multi-view cameras for the H1 and H2 datasets, and 8 for H3 and H4, and use the training/test split of 80%/20% to train and evaluate our methods.



Figure 6: Renderings produced by our methods (for all subjects H1, H2, H3, H4 in our study). All renderings are produced from new viewpoints and poses unseen in training.

### 4.1. Baselines

We denote our method as Im-TeX, and compare Im-TeX with 5 other systems, Pix2PixHD [52], Fea-Net [35], and three variants, Ex-TeX, Only-Ego, and Only-MV. All these baseline methods have similar architectures as ours, and we provide them with the same input and loss functions in training, so all these methods are directly comparable.

- 1) **Pix2PixHD.** We used Pix2PixHD’s code with minimal modifications that lead to better performance. Pix2PixHD directly translates a target pose  $P_t$  to a generated image  $I_{e \rightarrow t}$ .
- 2) **Fea-Net.** We refer to the human re-rendering method [35] as Fea-Net (short for FeatureNet), which is the state of the art in human re-enactment and novel view rendering when this project was developed. Different from our method, which maintains a global implicit texture stack, Fea-Net uses a network to extract 16-dimensional feature maps from egocentric images that serve as implicit textures. Though Fea-Net was originally trained on multiple identities to do garment and motion transfer, it can also be directly applied on our person specific task.
- 3) **Ex-TeX.** As a variant of our implicit texture method, Ex-TeX works with an explicit and static texture stack. Compared with Im-TeX, the texture stack ( $T_m$ ) in Figure 4 is not updated during training. Hence we call it an explicit texture stack as shown in Figure 5-left.
- 4) **Only-Ego.** We also consider a variant that does not have the texture stack  $T_m$ , which means  $T_g = T_e$ , and in feature rendering, we only sample textures from egocentric images. We call this method Only-Ego.
- 5) **Only-MV.** In contrast to Only-Ego, another variant is Only-MV, which does not take textures from egocentric images, but only from training images  $I_t$ . In this case,  $T_g = T_m$ . Only-MV can be seen as an extension of Deferred Neural Rendering [47].

### 4.2. Evaluations on Test Datasets

We consider two regimes: training on single- or multi-camera video sequences. The evaluation is done on the hold-out cameras and hold-out parts of the sequence. That

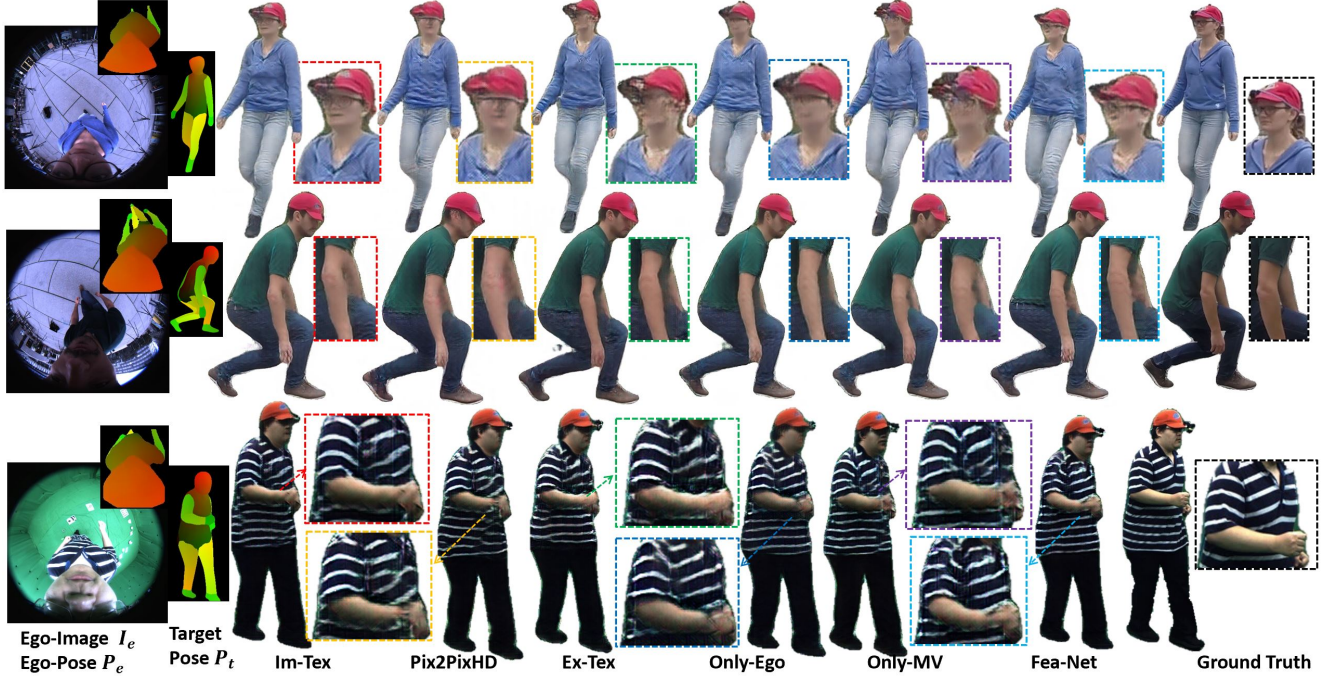


Figure 7: Comparisons of the six methods. All avatars are generated from novel viewpoints and poses unseen in training.

Methods	H1		H2		H3		H4		Mean relative improvements(%)		
	SSIM $\uparrow$	LPIPS $\downarrow$	SSIM	LPIPS	SSIM	LPIPS	SSIM	LPIPS	$RI_{SSIM} \uparrow$	$RI_{LPIPS} \uparrow$	$RI_{PSNR} \uparrow$
Im-Tex	7.529	<b>1.623</b>	<b>6.840</b>	<b>1.617</b>	<b>6.469</b>	<b>1.569</b>	7.535	<b>1.571</b>	<b>1.268</b>	<b>7.562</b>	.748
Pix2PixHD	7.395	1.713	6.798	1.713	6.342	1.640	7.428	1.615	-	2.427	.404
Ex-Tex	7.431	1.691	6.778	1.683	6.326	1.676	7.435	1.629	0.089	3.630	.555
Only-Ego	7.437	1.769	6.782	1.704	6.361	1.660	<b>7.586</b>	1.578	0.240	-	.467
Only-MV	<b>7.543</b>	1.738	6.828	1.626	6.360	1.585	7.505	1.587	1.190	4.263	-
Fea-Net	7.444	1.695	6.823	1.687	6.350	1.630	7.449	1.616	0.487	3.600	<b>.912</b>

Table 1: Quantitative results (multiplied by 10) of single-video training on different datasets, and relative improvements.

is, there is no overlap between the training and test sets in terms of the camera or body poses. For fair comparisons among different methods, we use DensePose results computed on the target frame as input in training and evaluation like other human rendering papers [38, 4].

**Metrics.** The following three metrics are used for comparison, the Structural Similarity Index (SSIM) [53], and the Learned Perceptual Image Patch Similarity (LPIPS) [57], and peak signal-to-noise ratio (PSNR). The recent LPIPS claims to capture human judgment better than existing hand designed metrics, whereas PSNR or SSIM is often inaccurate in capturing the visual perception and would differ starkly with slight changes in texture/wrinkles [57].

**Single video experiments.** We first evaluate our system in a single video case. We use single-camera videos from one of the cameras in our rig, and evaluate the 6 methods on the hold-out cameras and hold-out parts of the sequence. The qualitative results are shown in Figure 7. It can be observed that our results show better realism and preserve more de-

tails, and the quantitative results are provided in Table 1. We also calculate the mean relative improvement ( $RI$ ) of each approach ( $x$ ) over the worst one ( $y$ , indicated by -) on metric  $m$  for all datasets:  $RI_m(x, y) = |m(y) - m(x)|/m(y)$ , where  $m(x)$  is the result of  $x$  on metric  $m$ .  $RI$  is shown in percent. Our Im-Tex outperforms the others on 8 out of 11 metrics, and we achieve the best LPIPS scores on all the datasets. In general, our system is capable of learning avatars from monocular videos, and can be applied in everyday activities for which setting a studio with multi-view cameras to capture training data would be highly challenging.

**Multi-video comparisons.** We also conduct multi-video experiments where each method was trained on multiple videos from different viewpoints in Table 2. We train the 6 methods with 9 multi-view cameras for H1, and 4 multi-view cameras for H4. Each method performs better when more video sequences are added in training, and especially our method can reconstruct more details, as shown in Figure 9. More comparisons on multi-video training are provided



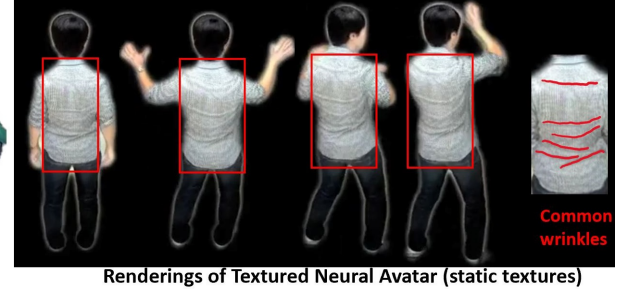
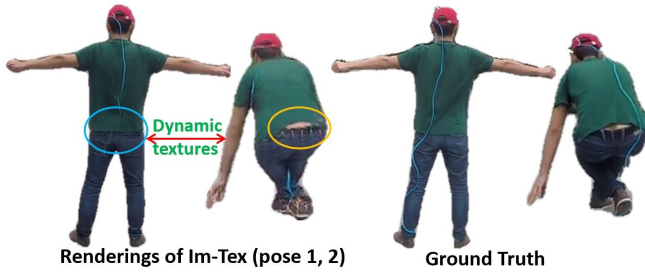


Figure 8: Comparisons to Textured Neural Avatars [38]. Our method (left) produces realistic time- and pose-dependent appearance detail, such as pose-dependent wrinkles in clothing, or the shifting of the shirt. In contrast, Textured Neural Avatars cannot produce the same level of realism, since a static texture with fixed clothing wrinkles is warped into new poses.



Figure 9: Renderings between single- and multi-video training. From left to right: egocentric image and estimated DensePose, single-video result, multi-video result, ground truth. With multi-video training, more details are restored, such as the right hand.

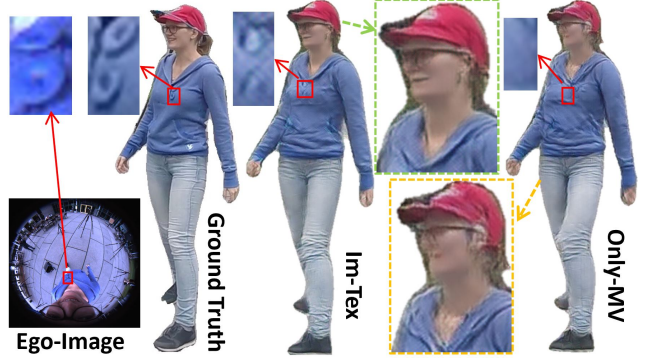


Figure 10: Compared with Only-MV, the use of egocentric texture of our Im-Tex leads to improvements visible in faces, neck and collar. We even restore the tiny clothes buttons visible in the Ego-Image.

	H1			H4		
	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR
ImT	8.134	<b>1.174</b>	2.104	<b>7.893</b>	<b>1.246</b>	<b>1.780</b>
P2P	8.038	1.218	2.077	7.823	1.301	1.753
ExT	7.993	1.258	2.059	7.716	1.374	1.722
Ego	8.038	1.228	2.075	7.825	1.287	1.763
MV	<b>8.174</b>	1.217	<b>2.140</b>	7.770	1.343	1.732
FNet	8.022	1.220	2.095	7.744	1.372	1.736

Table 2: Quantitative comparisons of multi-video training on H1 outdoor and H4 indoor datasets. SSIM and LPIPS are multiplied by 10, PSNR by 0.1. The methods are listed as follows: Im-Tex, Pix2PixHD, Ex-Tex, Only-Ego, Only-MV, Fea-Net.



Figure 11: In applications, given an egocentric camera image (left), our methods can generate plausible avatars from different viewpoints, even with high elevation degree (the last four) unseen in training. We show the same example frame as Figure 4.

in Figure 13 and the supplementary material.

### 4.3. Comparisons to Textured Neural Avatar (TNA)

We compare our method against static texture based method TNA in Figure 8, where ours can produce time- and pose-dependent appearance details, whereas TNA cannot.

### 4.4. Ablation Study

We study the advantages of the learned neural texture over other human rendering methods introduced in Section 4.1, including Pix2PixHD [52], Fea-Net [35], Ex-Tex, Only-Ego, and Only-MV. Note that the main difference

among these methods is the texture synthesis module. We get the following conclusions by comparing different methods according to the quantitative and qualitative results in Table 1, Figure 7-13 and our supplementary material. We use  $a > b$  to denote that method  $a$  outperforms method  $b$ .

1) Texture synthesis methods (e.g. Im-Tex, Only-MV) outperform the direct pose-to-image translation method (Pix2PixHD), implied by **Only-MV** > **Pix2PixHD**, where the only difference is that Only-MV has an implicit texture stack  $T_m$ . 2) Textures from egocentric images help improve

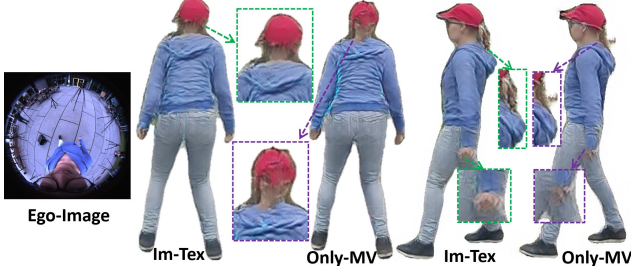


Figure 12: Comparisons with Only-MV on occluded parts.

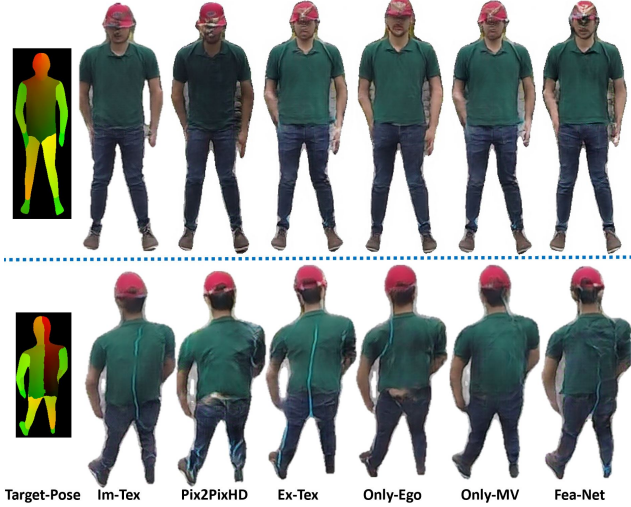


Figure 13: Renderings of different views in applications.

the rendering quality, implied by **Im-Tex** > **Only-MV**. 3) The implicit texture stack  $T_m$  matters, implied by **Im-Tex** > **Only-Ego**. 4) Implicit texture stacks capture detailed appearance variation across poses and viewpoints better than explicit color texture stacks, implied by **Im-Tex** > **Ex-Tex**. 5) For our human specific task, maintaining a globally shared texture stack works better than extracting temporary features from egocentric images for each frame, implied by **Im-Tex** > **Fea-Net** [35]. Different from Im-Tex, Fea-Net has a special FeatureNet to extract high dimensional features from egocentric inputs. Though Fea-Net has almost twice as many parameters as ours, its performance is not as good as ours on this person specific task.

#### Why does the proposed method outperform the others?

First, we attribute this improvement to the explicit texture synthesis. Compared with Pix2PixHD, which learns implicit correlations between the target poses and RGB avatars, our method explicitly utilizes the semantics of the target poses by Feature Rendering and formulates the rendering as an easier transformation from implicit textures to avatars. Second, ours leveraging the Ego-Texture as input outperforms Only-MV, as shown in Figure 7, 10, where ours shows better realism (e.g. faces and clothes) and even reconstructs the tiny buttons visible in the egocentric images

in Figure 10. We also see that ours outperforms Only-MV even on occluded parts (e.g. the human head in row 1 of Figure 7, the back of human in the second row of Figure 13, and Figure 12), where the RGB textures from the Ego-Texture are not directly used. However, the additional supervision contained in the Ego-Texture makes Render-Net learn better correlation between the occluded part and the visible texture regions, and possibly helps learn implicit shape information useful for more accurate renderings.

**Discussion: egocentric setup.** Why not first transform the fisheye images to normal images, then perform traditional human synthesis? Since we are using a fisheye camera with a  $180^\circ$  field of view (FOV), it is not possible to project (in other words ‘undistort’) our image to a rectilinear image. Although the cropped fisheye image (by reducing the FOV) can be undistorted, we lose the main advantage of using the fisheye image by doing so – full body capturing under a wide range of motions, including fully extended arms, etc. This is also the reason many existing egocentric methods process the fisheye images directly [54, 50].

#### 4.5. Applications

In real applications, as introduced in Figure 4 and Section 3.3, EgoRenderer uses Mo2Cap2 [54] to extract 3D poses and synthesize target poses. The synthesized target poses and comparisons of the six methods are provided in Figure 13 for the example frame in Figure 4. Our method generates higher quality avatars than the other methods. In Figure 11, we show more viewpoints. It can be observed that EgoRenderer synthesizes reasonable avatars from novel viewpoints even with high elevation degrees. In addition, EgoRenderer can work in both local (human-centered) and global coordinate systems with camera tracking. See more details in the supplementary material.

#### 5. Summary and Discussion

We proposed the EgoRenderer to synthesize free-viewpoint avatars from a single egocentric fisheye camera, and our system is an important step towards practical daily full-body reconstruction. We see our approach as the basis for many exciting applications in various areas, such as performance analysis, and human reconstructions in AR and VR. However, our system suffers from certain limitations mainly caused by the inaccuracies of Mo2Cap2 [54]: 1) Mo2Cap2 can estimate 15 joints (e.g., neck, shoulders, elbows, wrists, hips, knees, ankles and toes), whereas 24 joints are required to drive a SMPL model, which may cause inaccuracies in inverse kinematics and unnatural motions. 2) Mo2Cap2 per-frame predictions exhibit some temporal instability at run time, which causes temporal jittering in renderings. We believe these issues can be solved by optimizing the pose estimation module.

**Acknowledgements.** MZ was supported by NSF#1813583.



## References

- [1] Kfir Aberman, M. Shi, Jing Liao, Dani Lischinski, B. Chen, and D. Cohen-Or. Deep video-based performance cloning. *Computer Graphics Forum*, 38, 2019. 2, 3
- [2] Congqi Cao, Yifan Zhang, Yi Wu, Hanqing Lu, and Jian Cheng. Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules. In *IEEE International Conference on Computer Vision*, pages 3783–3791. IEEE Computer Society, 2017. 3
- [3] Y. Cha, T. Price, Z. Wei, X. Lu, N. Rewkowski, R. Chabra, Z. Qin, H. Kim, Z. Su, Y. Liu, A. Ilie, A. State, Z. Xu, J. Frahm, and H. Fuchs. Towards fully mobile 3d face, body, and environment capture using only head-worn cameras. *IEEE Transactions on Visualization and Computer Graphics*, 24(11):2993–3004, 2018. 3
- [4] C. Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. Everybody dance now. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5932–5941, 2019. 2, 3, 6
- [5] Mohamed Elgharib, Mallikarjun BR, Ayush Tewari, Hyeonwoo Kim, Wentao Liu, Hans-Peter Seidel, and Christian Theobalt. Egoface: Egocentric face performance capture and videorealistic reenactment, 2019. 3
- [6] Mohamed Elgharib, Mohit Mendiratta, Justus Thies, Matthias Nießner, Hans-Peter Seidel, Ayush Tewari, Vladislav Golyanik, and Christian Theobalt. Egocentric videoconferencing. *ACM Transactions on Graphics*, 39(6), Dec 2020. 3
- [7] A. Elhayek, E. de Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3810–3818, 2015. 1
- [8] A. K. Grigor’ev, Artem Sevastopolsky, Alexander Vakhitov, and Victor S. Lempitsky. Coordinate-based texture inpainting for pose-guided human image generation. *Computer Vision and Pattern Recognition (CVPR)*, pages 12127–12136, 2019. 3
- [9] R. A. Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 1, 2, 4, 5
- [10] Hao Jiang and Kristen Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3501–3509. IEEE Computer Society, 2017. 3
- [11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. volume 9906, pages 694–711, 10 2016. 5
- [12] Hyeonwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37, 2018. 2
- [13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 5
- [14] Bernhard Kratzwald, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Towards an understanding of our world by GANing videos in the wild. arXiv:1711.11453, 2017. 3
- [15] Hao Li, Laura C. Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma. Facial performance sensing head-mounted display. *ACM Trans. Graph.*, 34(4):47:1–47:9, 2015. 3
- [16] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and C. Theobalt. Neural sparse voxel fields. *ArXiv*, abs/2007.11571, 2020. 2
- [17] Lingjie Liu, Weipeng Xu, Marc Habermann, Michael Zollhöfer, Florian Bernard, Hyeonwoo Kim, Wenping Wang, and Christian Theobalt. Neural human video rendering by learning dynamic textures and rendering-to-video translation. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–1, 05 2020. 3
- [18] Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Hyeonwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. Neural rendering and reenactment of human actor videos. *ACM Transactions on Graphics (TOG)*, 2019. 3
- [19] Lingjie Liu, Weipeng Xu, M. Zollhöfer, H. Kim, F. Bernard, Marc Habermann, W. Wang, and C. Theobalt. Neural animation and reenactment of human actor videos. *ArXiv*, abs/1809.03658, 2018. 2
- [20] Weiyang Liu, Y. Wen, Zhiding Yu, Ming Li, B. Raj, and Le Song. Sphreface: Deep hypersphere embedding for face recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6738–6746, 2017. 5
- [21] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Trans. Graph.*, 37(4), July 2018. 3
- [22] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph. (SIGGRAPH)*, 38(4), 2019. 2
- [23] M. Loper, Naureen Mahmood, J. Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34:248:1–248:16, 2015. 4
- [24] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 405–415, 2017. 3
- [25] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. *Computer Vision and Pattern Recognition (CVPR)*, 2018. 3
- [26] Minghuang Ma, Haoqi Fan, and Kris M. Kitani. Going deeper into first-person activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1894–1903. IEEE Computer Society, 2016. 3
- [27] Dushyant Mehta, H. Rhodin, D. Casas, P. Fua, Oleksandr Sotnychenko, Weipeng Xu, and C. Theobalt. Monocular 3d

- human pose estimation in the wild using improved cnn supervision. *2017 International Conference on 3D Vision (3DV)*, pages 506–516, 2017. [1](#)
- [28] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, H. Rhodin, Mohammad Shafiei, H. Seidel, Weipeng Xu, D. Casas, and C. Theobalt. Vnect: real-time 3d human pose estimation with a single rgb camera. *ACM Trans. Graph.*, 36:44:1–44:14, 2017. [1](#)
- [29] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. [2](#)
- [30] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proceedings of International Conference on Computer Vision*, October 2017. [3](#)
- [31] Natalia Neverova, Riza Alp Güler, and Iasonas Kokkinos. Dense pose transfer. *European Conference on Computer Vision (ECCV)*, 2018. [3](#)
- [32] Katsunori Ohnishi, Atsushi Kanehira, Asako Kanezaki, and Tatsuya Harada. Recognizing activities of daily living with a wrist-mounted camera. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3103–3111. IEEE Computer Society, 2016. [3](#)
- [33] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. [3](#)
- [34] Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. Egocap: egocentric marker-less motion capture with two fisheye cameras. *ACM Trans. Graph.*, 35(6):162:1–162:11, 2016. [3](#)
- [35] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural re-rendering of humans from a single image. In *European Conference on Computer Vision (ECCV)*, 2020. [3](#), [5](#), [7](#), [8](#)
- [36] Takaaki Shiratori, Hyun Soo Park, Leonid Sigal, Yaser Sheikh, and Jessica K. Hodgins. Motion capture from body-mounted cameras. *ACM Trans. Graph.*, 30(4):31, 2011. [3](#)
- [37] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, Mat Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *CVPR 2011*, pages 1297–1304, 2011. [1](#)
- [38] Aliaksandra Shysheya, E. Zakharov, Kara-Ali Aliev, R. Bashirov, Egor Burkov, K. Isakov, Aleksei Ivakhnenko, Yury Malkov, I. Pasechnik, D. Ulyanov, Alexander Vakhitov, and V. Lempitsky. Textured neural avatars. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2382–2392, 2019. [2](#), [6](#), [7](#)
- [39] Aliaksandr Siarohin, Enver Sangineto, Stephane Lathuiliere, and Nicu Sebe. Deformable GANs for pose-based human image generation. In *CVPR 2018*, 2018. [3](#)
- [40] K. Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015. [5](#)
- [41] S. Singh, C. Arora, and C. V. Jawahar. First person action recognition using deep learned descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2620–2628, 2016. [3](#)
- [42] Suriya Singh, Chetan Arora, and C. V. Jawahar. Trajectory aligned features for first person action recognition. *Pattern Recognit.*, 62:45–55, 2017. [3](#)
- [43] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. Deepvoxels: Learning persistent 3d feature embeddings. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. [2](#)
- [44] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. [2](#)
- [45] Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt. Fast and robust hand tracking using detection-guided optimization. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, June 2015. [3](#)
- [46] Yusuke Sugano and Andreas Bulling. Self-calibrating head-mounted eye trackers using egocentric visual saliency. In Celine Latulipe, Bjoern Hartmann, and Tovi Grossman, editors, *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, pages 363–372. ACM, 2015. [3](#)
- [47] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38, 2019. [2](#), [5](#)
- [48] Justus Thies, Michael Zollöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. FaceVR: Real-Time Facial Reenactment and Eye Gaze Control in Virtual Reality. *arXiv preprint arXiv:1610.03151*, 2016. [3](#)
- [49] Denis Tomè, Patrick Peluse, L. Agapito, and H. Badino. xr-egopose: Egocentric 3d human pose from an hmd camera. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7727–7737, 2019. [1](#), [3](#)
- [50] Denis Tomè, Patrick Peluse, Lourdes Agapito, and Hernán Badino. xr-egopose: Egocentric 3d human pose from an HMD camera. In *IEEE International Conference on Computer Vision*, pages 7727–7737. IEEE, 2019. [8](#)
- [51] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. [2](#), [3](#)
- [52] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. pages 8798–8807, 06 2018. [4](#), [5](#), [7](#)
- [53] Zhou Wang, A. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004. [6](#)
- [54] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo2cap2: Real-time mobile 3d motion capture

- with a cap-mounted fisheye camera. *IEEE transactions on visualization and computer graphics*, PP, 03 2018. 1, 3, 4, 8
- [55] Ye Yuan and Kris Kitani. 3d ego-pose estimation via imitation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 735–750, 2018. 3
  - [56] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time PD control. In *IEEE International Conference on Computer Vision*, pages 10081–10091. IEEE, 2019. 3
  - [57] Richard Zhang, Phillip Isola, Alexei A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6
  - [58] 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 Conference on Computer Vision and Pattern Recognition*, pages 2347–2356, 2019. 3