Neural3D: Light-weight Neural Portrait Scanning via Context-aware Correspondence Learning

Xin Suo
ShanghaiTech University
Shanghai, China
suoxin@shanghaitech.edu.cn

Minye Wu
ShanghaiTech University
Shanghai, China
wumy@shanghaitech.edu.cn

Yanshun Zhang
Dgene
Shanghai, China
yanshun.zhang@dgene.com

Yingliang Zhang
Dgene
Shanghai, China
yingliang.zhang@dgene.com

Lan Xu
ShanghaiTech University
Shanghai, China
lxuan@connect.ust.hk

Qiang Hu
ShanghaiTech University
Shanghai, China
huqiang@shanghaitech.edu.cn

Jingyi Yu
Shanghai Engineering Research Center of Intelligent Vision and Imaging, School of Information Science and Technology,
ShanghaiTech University
Shanghai, China
yujingyi@shanghaitech.edu.cn

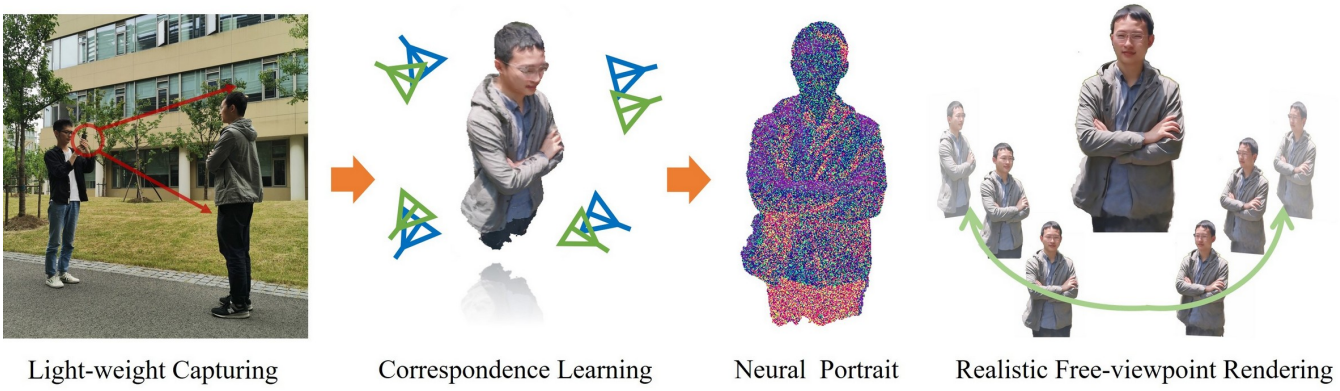

Figure 1: Illustration of our Neural3D system, which achieves convenient and realistic neural reconstruction and free-viewpoint rendering of human portraits from only a single portable RGB camera.

ABSTRACT

Reconstructing a human portrait in a realistic and convenient manner is critical for human modeling and understanding. Aiming at light-weight and realistic human portrait reconstruction, in this paper we propose *Neural3D*: a novel neural human portrait scanning system using only a single RGB camera. In our system, to enable accurate pose estimation, we propose a context-aware correspondence learning approach which jointly models the appearance, spatial and motion information between feature pairs. To enable realistic reconstruction and suppress the geometry error, we further adopt a point-based neural rendering scheme to generate realistic and immersive portrait visualization in arbitrary virtual view-points. By introducing these learning-based technical components into the pure RGB-based human modeling framework, we can achieve both accurate camera pose estimation and realistic free-viewpoint

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413734

rendering of the reconstructed human portrait. Extensive experiments on a variety of challenging capture scenarios demonstrate the robustness and effectiveness of our approach.

CCS CONCEPTS

• **Human-centered computing** → *Virtual reality*;

KEYWORDS

Human scaning; Correspondence estimation; Neural networks; Neural rendering

ACM Reference Format:
Xin Suo, Minye Wu, Yanshun Zhang, Yingliang Zhang, Lan Xu, Qiang Hu, and Jingyi Yu. 2020. Neural3D: Light-weight Neural Portrait Scanning via Context-aware Correspondence Learning . In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3413734

1 INTRODUCTION

Robust perception and understanding with humans present can enable numerous applications, such as human analysis and recognition, computer games, and virtual/augmented reality (VR/AR). How to reconstruct a realistic 3D model of a human target especially under an extremely light-weight capture setup, i.e. using only a single RGB camera, evolves as a cutting-edge yet bottleneck technique, which has recently attracted substantive attention of both the multimedia and computer graphics communities.

Despite tremendous advances in vivid and robust human portrait scanning using the RGBD sensors [12, 15, 29, 43, 49, 53], the utilization of depth sensor in these approaches brings inherent constraint for working under general illumination especially for outdoor capturing scenarios. On the other hand, pure RGB-based human modeling approaches [2, 3, 32, 36] usually adopt per-vertex coloring or atlas texturing schemes to provide vivid rendering results in a novel view, which suffers from geometry reconstruction error, leading to visually unpleasant results. With the significant progress of deep learning techniques, recent neural rendering approaches [1, 39–41] can further generate more realistic 2D results in the novel views. However, reliable correspondence estimation and subsequently camera pose estimation for all the input images remains the key fundamental problem of RGB-based human modeling. Early solution relies on handcrafted features such as SIFT [31] and RANSAC [18] to mitigate the influence of outlier, which is based on local content analysis and fragile to excessive outliers. Recent approaches [17, 27, 33, 34, 51, 54] utilize deep neural networks to generate more reliable feature detectors and descriptors or exploit local context and motion normalization for outlier rejection. However, these techniques emphasize unanimously on the context-encoding aspect while ignoring the underlying geometry or topology such as relative positions between individual feature points, leading to high sensitivity to outliers.

In this paper, we attack the above challenges and propose *Neural3D*, a novel light-weight neural human portrait reconstruction system, which can generate realistic rendering results of the target in the novel viewpoints, only using about 80 RGB images roughly around the target from a single RGB camera. Our novel pipeline brings aspects inherent in RGB-based human modeling to both the neural rendering and data-driven corresponding optimization.

More specifically, to enable reliable inlier correspondence estimation, we first propose a context-aware end-to-end hybrid scheme to measures the matching score of each candidate feature (SIFT in our implementation) pair from various images, consisting of an appearance module and a motion module. The former module jointly extracts the appearance and the spatial information in the image domain of the candidate pair, while the motion module encodes the correspondence's coordinates into the motion feature space with the self-attention mechanism to extract both the local and global motion information. Second, based on all these inlier feature pairs via the hybrid matching scheme above, we further perform a global bundle adjustment to generate the accurate camera poses and initial geometry of the human target, followed by a shape-from-silhouette refinement so as to generate a dense initial geometry. Finally, to enable realistic rendering of the human portrait in arbitrary viewpoints and suppress the influence of geometry error, a neural portrait scheme is adopted, consisting of a projection and rasterization module for feature projection as well as a U-Net based feature decoder. Our neural scheme enables realistic free-viewpoint rendering of the human portrait.

In summary, the main contributions of Neural3D include:

- We propose a novel neural portrait scanning system for realistic free-viewpoint rendering from only a single portable RGB camera.
- To enable accurate pose estimation, we propose a context-aware correspondence learning approach to jointly model the appearance, spatial and motion information.
- We combine the initial geometry with an effective point-based neural rendering scheme to provide realistic and immersive portrait visualization.

2 RELATED WORK

Human Modeling. Acquiring 3D geometric content and realistic rendering for human modeling from real world is an essential task for many applications in multimedia and computer graphics communities. High-quality human models can be created using 3D scanning devices, such as laser scan [4, 5] or a multi-view studio-level setup [14, 21, 22, 44] with a controlled imaging environment. Those systems are usually costly and the synchronizing and calibrating multi-camera systems is cumbersome, leading to the high restriction of the wide applications for daily usage.

The availability of commodity depth cameras enabled low-cost human modeling without complicated multi-view setup [12, 29, 30, 45, 47]. These methods utilize the TSDF volume [16] for both geometry representing and camera localization. To capture and reconstruct the full body, researchers adopt a sparse multi-view setup [48, 49] with only three or four depth sensors for human modeling. However, the active IR-based depth sensors are unsuitable for outdoor capture, and their high power consumption limits the mobile application. Recently, with the advent of deep neural networks, purely RGB-based monocular methods have been proposed to encode various prior information of human models such as motion [23, 24, 50], geometry [20, 28, 36, 37], garment [8] or appearance [25]. However, such methods still relies on per-vertex

coloring or atlas texturing schemes which suffers from geometry error and visually unpleasant results. Recent neural rendering algorithms [1, 39–41] bring huge potential to enable more realistic 2D rendering results in the novel views. However, researchers pay less attention to strengthen the human scanning process with neural rendering technique, especially for light-weight capture in the real-world scenario. Comparably, our approach not only combines pure RGB-based human portrait scanning with neural rendering but also benefits from a robust camera pose optimization scheme via corresponding learning, which is the key fundamental problem of RGB-based human modeling.

Correspondence Learning. Reliable correspondence estimation and subsequently camera pose estimation serves as the key fundamental problem of a variety of RGB-based applications such as in the wild scene reconstruction and 3D scanning. Early solution relies on handcrafted features such as SIFT [31], SURF [7] and ORB [11], etc. which can tackle the lighting, perspective and scale variations for correspondence matching. Then, the RANSAC [18] algorithm is adopted to reject the substantial outliers of those hand-crafted feature matching, which has been the gold standard for outlier rejection for decades.

With the advent of deep neural networks, recent learning-based approaches have achieved significant progress for both feature estimation and outlier rejection. LIFT [51] proposes an end-to-end feature estimation network consisting of image-based detector, an orientation estimator and a rotation-corrected descriptor, while SuperPoint [17] employs a learning architecture in a self-supervision manner to simultaneously detect the keypoints and compute the descriptors. Besides, global methods such as ContextDesc[26] utilizes a context-aware network to aggregate both the spatial and visual context of the whole image representations. In contrast to these methods that are devoted to effective extracting feature descriptors, in this paper we focus on the post-phase, i.e. distinguishing the false matching from the true inlier correspondences.

As for outlier rejection, Universal RANSAC (USAC) [35] adopts the generalized hypothesize-and-verify framework and incorporates the practical RANSAC variants to improve both the speed and accuracy performance. DSAC [9] adopts a probabilistic selecting process to make it differential so that the complete process can be trained in an end-to-end manner, while Marginalizing Sample Consensus (MAGSAC) [6] proposes to find the optimal model through weighted least-squares fitting without estimation of an inlier-outlier threshold. Recently, Neural-Guided RANSAC(NG-RANSAC) [10] employs deep networks to first estimate the confidence of the putative correspondences being inliers to guide the matching process with improved model hypothesis searching. As the most closely related to our approach, PointCN [27] employs the PointNet-like [33, 34] architecture to classify every pair of correspondences as either inlier or outlier and then uses the weighted eight-point algorithm for essential matrix estimation. Specifically, PointCN [27] utilizes context normalization to learn global context features that encode additional camera motion in terms of correspondences distributions. In a similar vein, Zhang *et al.* [54] utilize novel differential pooling and unpooling operations on correspondences by learning a soft assignment matrix that implicitly clusters correspondences with respect to local context. However, these approaches above only focus on the context of the correspondence instead of the spatial information across the images, and the context normalization operation cannot directly model the relations among the correspondences and update the correspondences feature according to the similarity.

In contrast, our correspondence learning approach utilize a KNN graph to capture the spatial information among all the key points and a self-attention mechanism to obtain the motion similarity among the correspondences. Our approach explicitly exploits the appearance, spatial and motion relationship jointly between putative correspondences for more reliable outlier rejection.

3 SYSTEM OVERVIEW

Recall that Neural3D attempts to reconstruct a realistic human portrait using only a single RGB camera. Fig. 1 illustrates the high-level components of our system, which takes a sequential RGB images as input and generates a neural portrait as output, achieving realistic free-viewpoint rendering results at any capture views. To maintain the light-weight setting and the potential of wide applications for daily usage, our Neural3D only relies on a consumer-level mobile phone to capture about 70 RGB images roughly around the target person. Then, we combine a novel data-driven pose estimation scheme with a neural portrait learning scheme to provide realistic reconstruction. A brief introduction of each main component of our pipeline is provided as follows.

Correspondence Estimation. To enable accurate camera pose estimation which is the fundamental problem of image-based human modeling, we propose a data-driven correspondence evaluation scheme for each image pairs, which jointly considers the spatial and appearance-based feature similarity as well as a motion corrections based on epipolar constraint. Our estimation scheme achieves reliable inlier correspondence selection across all the image pairs for further neural portrait reconstruction.

Geometry Generation. Based on the data-driven correspondences above, a global bundle adjustment scheme is adopted to obtain both the accurate camera poses of all the input RGB frames and an initial sparse 3D geometry of the human target. Furthermore, with the optimized camera poses, we utilize the traditional shape-from-silhouette technique to generate a dense initial geometry of the target, where the human silhouette is obtained by applying the human parsing method to each input RGB image.

Neural Portrait Reconstruction. A straight-forward texturing scheme applied to the initial geometry using the input RGB frames leads to inferior visual results due to the reconstruction error. To this end, we utilize a neural portrait reconstruction scheme to generate high quality rendering results in the new virtual views, so as to enable realistic free-view-point rendering of human portrait. The key component in our neural portrait scheme is a differentiable renderer consisting of a projection and rasterization module for feature projection as well as a U-Net based feature decoder.

4 METHOD

4.1 Correspondence Learning

Reliable correspondence evaluation is critical for accurate camera pose estimation and further RGB-based human modeling. To this

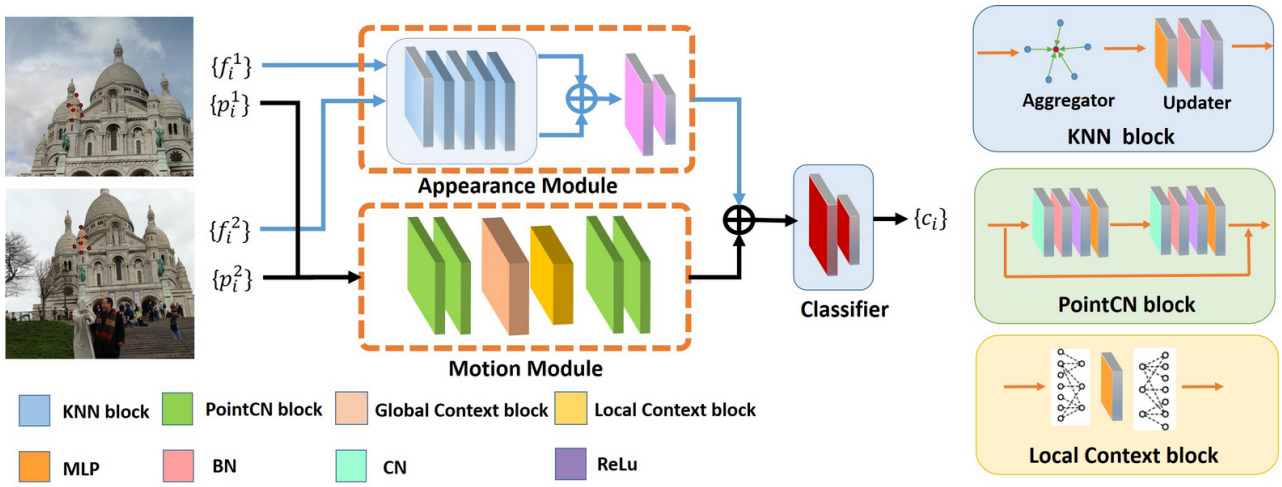

Figure 2: The network architecture of our context-aware correspondence learning scheme, mainly consisting of an appearance module and a motion module, so as to jointly model the appearance, spatial and motion information.

end, we propose a context-aware correspondence learning scheme to distinguish the false matching from the true inlier correspondences, which jointly models the appearance, spatial and motion information between each image pair. As illustrated in Fig. 2, our scheme takes the feature descriptors $\{f_i^1\}, \{f_i^2\}$ and their pixel locations $\{p_i^1\}, \{p_i^2\}$ of the candidate correspondences from each image pair I_1 and I_2 as input, and generates the corresponding matching scores $\{c_i\}$ for robust outlier rejection. Here, $i \in [1, N]$ denotes the index of all the N feature pairs. Specifically, our novel network architecture consists of an appearance module and a motion module. The former module not only encodes the appearance information from the feature descriptions but also utilizes a KNN graph architecture to extract the spatial information across various connected features in the image domain. The motion module encodes the correspondence's coordinates into the motion feature space with a self-attention mechanism and a node-based clustering to extract global and local motion information, respectively. Then, both the matching results from both modules are concatenated to generate the final per-pair matching score for outlier rejection and subsequent global bundle adjustment as well as the initial geometry generation.

Appearance Module. Since the feature descriptions only encodes local textural context from the local image regions, only using feature descriptions cannot handle repeated texture patterns for outlier rejection. To involve more spatial structure information, we adopt a novel KNN graph block in our appearance module, which encodes hybrid embeddings in terms of both and feature descriptions and positions. In each KNN layer of our KNN graph block, we utilize an aggregator and an updater, which are formulated as follows:

$$f_i^{\mathrm{agg}} = \frac{1}{K}\sum_{k=1}^{K} f_{i,k}, \quad f_i^{\mathrm{upd}} = f_i^{\mathrm{agg}} \cdot W. \tag{1}$$

Here, $f_{i,k}$ denotes the k-th nearest detected feature of f_i in the same image and W denotes the parameters of the multi-layer perceptron (MLP) in our updater. Note that for simplification we omit the subscript for image indexing. Furthermore, we employ a MLP-based matching network that concatenates the hybrid keypoint embeddings after the KNN block to estimate their similarities. After such matching, we obtain a hybrid feature map which encodes both the appearance and spatial structural similarities of the correspondence candidates, denoted as $\mathcal{S}_{\mathrm{app}} \in \mathbb{R}^{N\times C}$, where C is the dimension of the hybrid feature space.

Motion Module. Our motion module essentially encodes the correspondence's coordinates into the motion feature space with a self-attention mechanism and a node-based clustering to extract global and local motion information, respectively. To this end, we apply the PointCN block [27, 54] to encode the motion features of the correspondence's coordinates, which consists of shared-parameter MLP, batch normalization and ReLU layers. Comparably, we introduce both global and local context blocks for the context normalization, since the original normalization in [27, 54] only models the local context, leading to inferior results when the false matches have comparable influence in terms of the mean and variances as the true matches. To fetch the global motion information, we utilize the self-attention mechanism in our global context block to calculate a global affinity between the correspondences. Given the feature map $\mathbf{F} \in \mathbb{R}^{N\times C}$ after the PointCN block, we apply an attention map to it to update the global information and obtain the normalized feature map $\mathbf{F}^*$:

$$\mathbf{F}^* = Attention(\mathbf{F}) \cdot \mathbf{F} = \mathbf{A} \cdot \mathbf{F}. \tag{2}$$

Here, $\mathbf{A}$ is the self-attention map, which is produced by applying the softmax function among the correspondences as follows:

$$\mathbf{A} = Attention(\mathbf{F}) = softmax(\mathbf{F}\mathbf{F}^{\mathbf{T}}/C), \tag{3}$$

where C is the dimension of the hybrid feature space for scale normalization.

As for modeling the local motion, similar to previous work [27, 54], we adopt a node-based clustering mechanism in the motion

Figure 3: The network architecture of the utilized neural portrait generation module.

feature space. To this end, we utilize the differential pooling and unpooling method to learn the assignment matrices $\mathbf{S}_{\mathrm{p}}$ and $\mathbf{S}_{\mathrm{up}}$ between the output feature map $\mathbf{F}^*$ after global normalization and the latent nodes, which are formulated as:

$$\begin{aligned} \mathbf{S}_{\mathrm{p}} &= softmax(\mathbf{M}_{\mathrm{p}}(\mathbf{F}^*)), \\ \mathbf{S}_{\mathrm{up}} &= softmax(\mathbf{M}_{\mathrm{up}}(\mathbf{F}^*)), \end{aligned} \tag{4}$$

where $\mathbf{M}_{\mathrm{p}}$ and $\mathbf{M}_{\mathrm{up}}$ denote the corresponding weight-sharing MLP layers, respectively. Then, based on these assignment matrices, we follow [54] to obtain the hybrid motion feature map $\mathcal{S}_{\mathrm{mot}} \in \mathbb{R}^{N \times C}$ which encodes both the global and local motion similarities of the correspondence candidates. Finally, both $\mathcal{S}_{\mathrm{app}}$ and $\mathcal{S}_{\mathrm{mot}}$ are concatenated into a MLP-based classifier to generates the final matching scores $\{c_i\}$ for robust outlier rejection.

Loss Function. We train our context-aware correspondence learning network with a binary classification loss $\boldsymbol{L}_{\mathrm{c}}$, and an epipolar distance loss $\boldsymbol{L}_{\mathrm{e}}$:

$$\boldsymbol{L} = \boldsymbol{L}_{\mathrm{c}} + \lambda_{\mathrm{e}} \boldsymbol{L}_{\mathrm{e}}. \tag{5}$$

Here, $\boldsymbol{L}_{\mathrm{c}}$ is a typical binary cross entropy loss between estimated inliers set $\mathbf{M}$ and the ground truth inliers set $\mathbf{M}_{\mathrm{gt}}$. The epipolar distance loss $\boldsymbol{L}_{\mathrm{e}}$ is derived from the estimated essential matrix to model the epipolar geometry constraint. Specifically, we adopt the weighted eight-point algorithm to estimate the essential matrix $\mathbf{E}$, which is differentiable with respect to the predicted inliers $\mathbf{M}$ and makes it possible to regress $\mathbf{E}$ in an end-to-end manner. Then, the epipolar distance loss $\boldsymbol{L}_{\mathrm{e}}$ is formulated as the sum of epipolar distance of all inliers as follows:

$$\boldsymbol{L}_{\mathrm{e}} = \sum_{(p_i^1, p_i^2) \in \mathbf{M}} \|\mathbf{dist}(p_i^1, p_i^2, \mathbf{E})\|_2^2, \tag{6}$$

where p_i^1 and p_i^2 are the feature coordinates of the i-th matched correspondence inlier from $\mathbf{M}$. And the epipolar distance is formulated as follows:

$$\mathbf{dist}(p_i^1, p_i^2, \mathbf{E}) = \frac{{p_i^2}^{\mathbf{T}} \mathbf{E} p_i^1}{\sqrt{\|\mathbf{E}p_i^1\|_{[1]}^2 + \|\mathbf{E}p_i^1\|_{[2]}^2 + \mathbf{E}p_i^2\|_{[1]}^2 + \|\mathbf{E}p_i^2\|_{[2]}^2}}, \tag{7}$$

where $\mathbf{t}_{[j]}$ denotes the j-th element of a vector $\mathbf{t}$.

Geometry Generation. Based on the above final per-pair matching score from our corresponding learning scheme, we can reject those outliers to enable robust correspondence matching. To this end, after outlier rejection, we perform a global bundle adjustment to optimize both the camera poses of all the input RGB images and an initial sparse 3D keypoints, which is based on the incremental Structure-from-Motion system [38]. Furthermore, with the optimized camera poses, we utilize the traditional shape-from-silhouette technique [13] to generate a dense initial geometry of the target, where the human silhouette is obtained by applying the human parsing method [19]. Such final dense geometry of the human portrait and the corresponding accurate camera poses of the input RGB images are utilized to generate the neural portrait model in the next section.

4.2 Neural Portrait

Recall that the dense initial geometry of the human target obtained by the shape-from-silhouette technique usually suffers from incompleteness and reconstruction noise. Thus, a straight-forward texturing scheme applied to the initial geometry using the input RGB frames leads to inferior visual results due to the inherently coarse geometry. To suppress such geometry-related artifact, we utilize a novel neural rendering scheme to synthesize photo-realistic free-viewpoint rendering results of the portrait, based on only a low-fidelity 3D point cloud instead of traditional mesh representation. As illustrate in Fig. 3, our neural portrait scheme assigns a learnable feature vector $\mathbf{f}_i$ for each input 3D point $\mathbf{p}_i$ which encodes both the appearance and contextual information of the human portrait.

Projection and Rasterization. When we render a novel target viewpoint, we generate a view-dependent feature map $\mathbf{M}$. Specifically, we project all points and thus their features onto the target view and splat points into pixel coordinates on image plane linearly. Then we apply the Z-buffer method to maintain correct depth ordering and hence occlusions. For each background pixel, we assign a learnable default feature vector θ_d. The resulting feature map $\mathbf{M}$ is formulated as:

$$\mathbf{M}_q[u, v] = \begin{cases} [\mathbf{f}_i; \vec{d}_i] & (u, v) \in \mathcal{S}_i \\ [\mathbf{f}_0; \mathbf{0}] & otherwise \end{cases} \tag{8}$$

where $\mathcal{S}_i$ is the set of pixels in the viewpoint that maps to the same 3D point $\mathbf{p}_i$ under splatting. We record all point indexes in this feature map, so as to back propagate gradients from feature map to feature vectors.

Rendering. The U-Net architecture has shown great success in many applications, such as image denoising, deblurring, and style transfer. In our setting, we use U-Net to generate a RGB image at target viewpoint from the view-dependent feature map. Similar to [1], we adopt the gated convolution layer [52] in network architecture to handle incomplete and noisy geometry. Projected depth maps and view directions of pixel rays are also fed into the rendering network to improve the rendering results and support view-dependent effects.

To enable light-weight capturing, our method only utilizes a single RGB camera and do not need a large amount of input RGB images to supervise the network training like [1]. Therefore, we utilize generative adversarial network (GAN) for training. We take U-Net as the generator $\mathcal{G}(\cdot)$ to render RGB results and a convolution network with a binary classifier as the role of the discriminator $\mathcal{D}(\cdot)$.

Network Training. To train our neural portrait network, we utilize the initial dense 3D point cloud and the RGB images with optimizied camera poses from previous stage. For each training sample, we set one of the input cameras as the target camera and utilize the corresponding RGB image as groud truth to conduct supervised training for the generator. Meanwhile, we want the discriminator can distinguish whether the input image is from the generator. We utilize the corresponding input RGB image and generated image pairs for training the discriminator.

Specifically, our loss function consists of a GAN loss $\boldsymbol{L}_{gan}$ and an image loss $\boldsymbol{L}_{img}$. The former one is formulated as follows:

$$\boldsymbol{L}_{\text{gan}} = E_x[\log(\mathcal{D}(x))] + E_M[\log(1 - \mathcal{D}(\mathcal{G}(\mathbf{M})))], \tag{9}$$

which intends to force the output distribution of U-net generator closed to the real distribution, so as to provide photo-realistic rendering results. Besides, we also utilize the following image loss:

$$\boldsymbol{L}_{\text{img}} = \sum_{q=1} \|\mathbf{I}_q - \widetilde{\mathbf{I}}_q\|_2^2, \tag{10}$$

where $\mathbf{I}_q$ is the rendering result image in view q; $\widetilde{\mathbf{I}}_q$ is the corresponding ground truth image. The overall loss function is $\boldsymbol{L} = \lambda \cdot \boldsymbol{L}_{gan} + \boldsymbol{L}_{img}$, where $\lambda = 0.2$ in our experiments. Recall that both the projection and rasterization of the 3D feature points are differentiable. Thus, the gradients from loss function can be back-propagated to the entire network including feature vectors of the 3D points. We utilize the gradient-based optimizer Adam in all of our experiments to update both the network parameters and the feature vectors.

5 EXPERIMENTAL RESULTS

In this section, we evaluate our Neural3D system on a variety of challenging scenarios. We first evaluate our correspondence learning scheme, followed by the evaluation of our neural portrait reconstruction, both qualitatively and quantitatively. The limitation and discussion regarding our Neural3D system are provided in the last subsection. Several representative neural portraits reconstructed by our Neural3D system are illustrated in Fig. 4, where the challenging appearance details of the portraits, such as the textures in those face, hand and garment wrinkle regions, are faithfully reconstructed only using a single consumer-lever RGB camera.

Method	Outdoor		Indoor	
	Known	Unknowm	Known	Unknown
RANSAC	7.9	8.8	3.7	2.4
NG-RANSAC	43.1	50.5	23.7	15.9
PointCN	42.6	48.8	21.4	15.8
OANet	43.7	52.6	24.6	17.9
Ours	**43.6**	**54.6**	**28.4**	**17.8**

Table 1: Comparison with other methods on the Known and Unknown test sets about YFCC100M and SUN3D. The results with RANSAC are provided in terms of mAP(%).

model	Known scene	Unknown scene
PCN	42.6/5.7	48.8/11.2
PCN + GM	41.1/10.8	51.1 /21.4
PCN + LM	42.9/16.5	52.6/26.6
PCN + LM + GM	43.1/18.6	53.5/28.4
PCN + LM + GM + AM	43.6/19.3	54.6/30.9

Table 2: Ablation study of our correspondence learning scheme on the known scenes and unknown scenes about YFCC100M datasets. The results with and without RANSAC are provided in terms of mAP(%).

5.1 Evaluation of Correspondence

In this section, we evaluate our context-aware correspondence learning scheme, both qualitatively and quantitatively. To this end, we utilize both the standard outdoor datasets Yahoo's YFCC100M [42] and the indoor datasets SUN3D [46] to validate our scheme quantitatively. We train our network on the subsets of the two datasets and evaluate the correspondence estimation performance on the other subsets of the same scenes as the *Known* test set and the datasets from the other scenes as the *Unknown* test set. For thorough evaluation, we compare against both the traditional baseline technique *RANSAC* and the state-of-the-art approaches including *PointCN* [27], *OANet* [54] and *NG – RANSAC* [10]. We adopt the standard angular difference between the estimation and the ground truth and measure the mean average precision (mAP) under accuracy a threshold (5°) for both rotation and translation. As shown in Tab.1, our approach consistently outperforms the others on both the Known and Unknown test sets for both the indoor and outdoor data sets, which illustrates the effectiveness and generalization ability of our approach to achieve more accurate pose estimation.

We further evaluate the influence of various components in our context-aware correspondence learning scheme, using the above Known and Unknown test sets in the YFCC100M dataset. For the motion module, let *PCN* denote the PointCN block, while *LM* and *GM* denote the local and global motion context blocks, respectively. Besides, let *AM* denote our appearance module. Tab. 2 shows that our full pipeline consistently outperforms the other baseline variations, yielding the highest mAP. This not only highlights the contribution of each algorithmic component but also illustrates that our correspondence learning approach can robustly recover accurate camera poses.

Figure 4: Several examples of our neural portrait scanning results using the proposed Neural3D system.

Then, we further evaluate the influence of our pose estimation scheme in terms of the final neural portrait rendering. To this end, we compare our Neural3D against the variation with the pose estimation results from the 3D reconstruction software [32], denoted as *w/o_pose*. Note that for fair comparison, both w/o_pose and our Neural3D share the same dense initial geometry and neural portrait scheme. For further quantitative analysis, we render their neural portraits into the capturing camera views by taking the input RGB input as reference only in the visible regions. Note that the residuals are calculated as the per-pixel Euclidean distances of the RGB values between the textured results and the color image inputs, where each color channel is normalized to [0,1]. As shown in Fig. 5, the results of w/o_pose suffer from pose localization error and blur rendering results, while our Neural3D with the pose estimation scheme achieves more realistic and immersive rendering, especially for the detailed texture in the face, hair and garment wrinkle regions. This evaluation further illustrates the effectiveness of our context-aware correspondence learning scheme for the neural rendering.

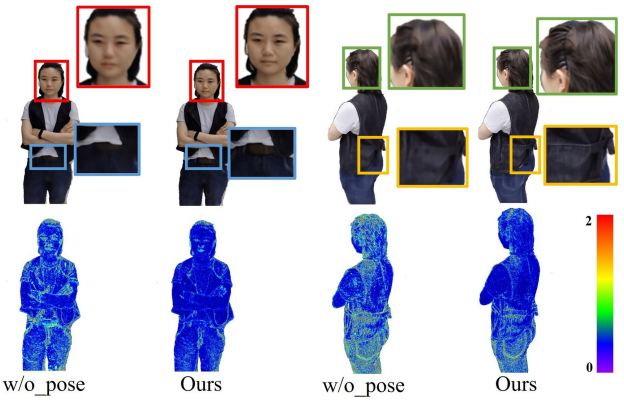

Figure 5: Ablation study of our pose refinement scheme in terms of the final neural portrait rendering results. The blue map indicates the normalized color-coded residual compared with the input color image.

5.2 Evaluation of Neural Portrait

In this subsection, we demonstrate the performance of our neural portrait scheme by comparing it against other state-of-the-art human modeling methods, both qualitatively and quantitatively.

For thorough evaluation, we compare our Neural3D with the traditional shape-from-silhouette technique [13] and the popular 3D reconstruction software [32], denoted as *SFS* and *PhotoScan*, respectively. Note that we apply the refined camera poses with our correspondence learning scheme to SFS; thus both Neural3D and SFS share the same initial dense geometry. Besides, both SFS and PhotoScan utilizes the average texturing scheme provided in the software [32] to achieve textured results. We also apply our camera poses instead of the original poses in PhotoScan for further evaluation, which is denoted as *PhotoScan_ours*. We further compare against the state-of-the-art image-based 3D human reconstruction method [36], denoted as *PIFU*.

As shown in Fig. 6, PhotoScan, PhotoScan_ours and PIFU suffer from severe geometry reconstruction error, while SFS still fails to reconstruct realistic texture details even with a water-tight geometry. In contrast, our Neural3D system achieves significantly better reconstruction results in a realistic and immersive manner, especially for the details in the face and hand regions highlighted by the colored circles in Fig. 6. For quantitative comparison, we render all the reconstruction texturing results into the capturing camera views by taking the input RGB input as reference only in the visible regions and adopt the popular criteria *PSNR* and *SSIM*. As shown in Tab. 3, our Neural3D consistently outperforms the other approaches in terns of both PSNR and SSIM. All these qualitative and quantitative results above illustrate the robustness and effectiveness of our approach to generate not only accurate camera

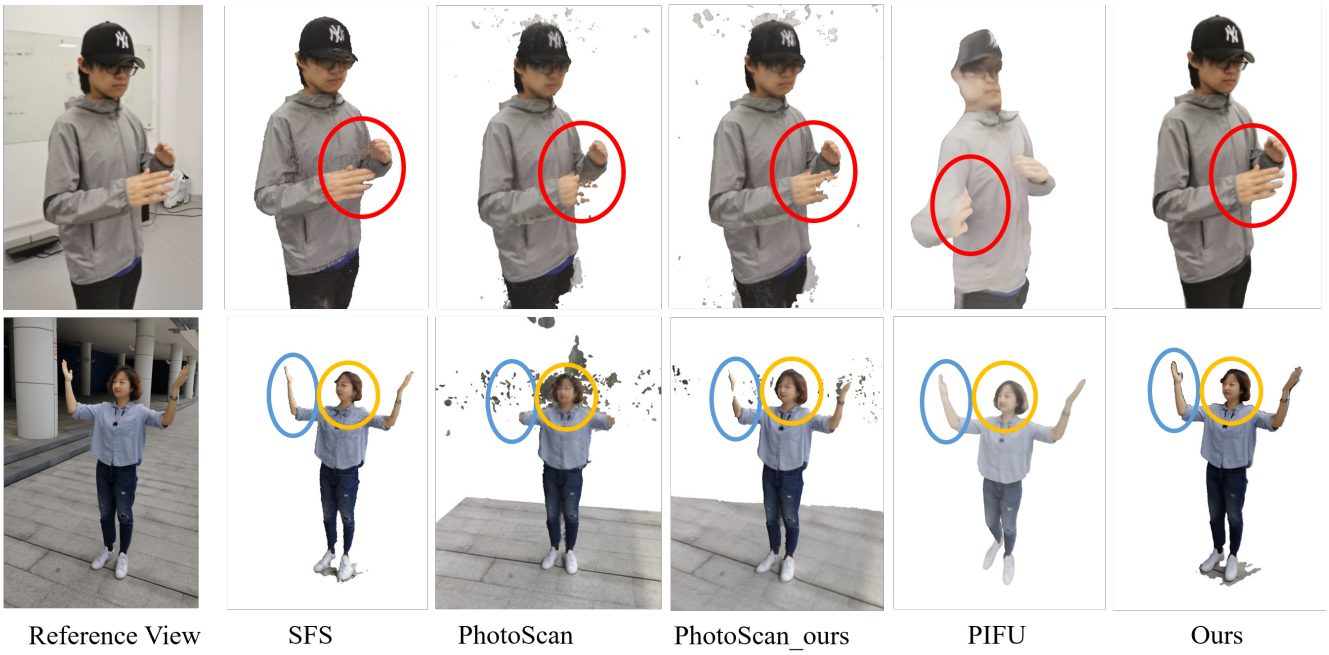

Figure 6: Qualitative comparison. Our neural portrait achieves much more realistic rendering results in the new virtual views.

Method	PSNR	SSIM
PhotoScan	22.5	0.92
PhotoScan_ours	25.4	0.94
SFS	35.2	0.94
Ours	**41.9**	**0.99**

Table 3: Quantitative comparison of our neural portrait against PhotoScan and SFS in terms of PSNR and SSIM.

poses but also realistic rendering results with fine details, from only a single RGB camera.

5.3 Limitations and Discussion

We have demonstrated compelling neural free-viewpoint rendering results of the human portraits in a variety of scenarios. Nevertheless, as the first trial to combine pure RGB-based human modeling with both data-driven correspondence learning and neural rendering, the proposed Neural3D system is subject to some limitations. First, even though our system enables convenient portrait capture in real-time with a portable device, the training process of our neural portrait to provide a visually pleasant rendering results takes about 6 to 8 hours, which is not suitable for some real-time human-computer-interaction (HCI) applications. However, our Neural3D system still brings new possibility for convenient and realistic rendering of a human portrait. Even if the goal is to upload the model for using in a VR/AR game, an overnight process remains valuable. Secondly, our current pipeline focuses on human portrait reconstruction, without modeling the background. Thus, severe blur occurs in the background scenes due to the lack of constraints of our neural rendering in those regions. This could be alleviated in the future by modeling the background of the captured scene explicitly as a static panoramic image in our current framework. Besides, it is an promising direction to further modify the initial geometry using the neural portrait rendering.

6 CONCLUSION

We have presented Neural3D, a novel neural human portrait scanning system using only a single RGB camera, which combines RGB-based human modeling with both data-driven correspondence learning and neural rendering. Our context-aware correspondence learning scheme enables accurate camera pose estimation, while our neural portrait scheme further suppresses the geometry error and generates visually pleasant rendering results in novel views. Our experimental results demonstrate the effectiveness of Neural3D for providing realistic and immersive free-viewpoint rendering results of a human portrait, which compares favorably to the other methods. We believe that it is a significant step to enable convenient and realistic human modeling, with many potential applications in VR and AR, gaming, entertainment and human analysis.

ACKNOWLEDGMENTS

This work was supported by NSFC programs (61976138, 61977047), STCSM (2015F0203-000-06), the National Key Research and Development Program (2018YFB2100500) and SHMEC (2019-01-07-00-01-E00003).

REFERENCES

[1] Kara-Ali Aliev, Dmitry Ulyanov, and Victor Lempitsky. 2019. Neural point-based graphics. *arXiv preprint arXiv:1906.08240* (2019).

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018. Detailed Human Avatars from Monocular Video. In *2018 International Conference on 3D Vision (3DV)*. 98–109.
[3] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018. Video Based Reconstruction of 3D People Models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[4] Brett Allen, Brian Curless, and Zoran Popović. 2003. The Space of Human Body Shapes: Reconstruction and Parameterization from Range Scans. *ACM Trans. Graph.* 22, 3 (July 2003), 587–594. https://doi.org/10.1145/882262.882311
[5] artec3d [n.d.]. artec3d. https://www.artec3d.com/. Accessed: 2020-05-24.
[6] Daniel Barath, Jiri Matas, and Jana Noskova. 2019. MAGSAC: marginalizing sample consensus. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10197–10205.
[7] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. SURF: Speeded Up Robust Features. In *Computer Vision – ECCV 2006*, Aleš Leonardis, Horst Bischof, and Axel Pinz (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 404–417.
[8] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. 2019. Multi-Garment Net: Learning to Dress 3D People from Images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE.
[9] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. 2017. DSAC - Differentiable RANSAC for Camera Localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[10] Eric Brachmann and Carsten Rother. 2019. Neural-Guided RANSAC: Learning Where to Sample Model Hypotheses. *arXiv preprint arXiv:1905.04132* (2019).
[11] Michael Calonder, Vincent Lepetit, Mustafa Ozuysal, Tomasz Trzcinski, Christoph Strecha, and Pascal Fua. 2011. BRIEF: Computing a local binary descriptor very fast. *IEEE transactions on pattern analysis and machine intelligence* 34, 7 (2011), 1281–1298.
[12] Wei Cheng, Lan Xu, Lei Han, Yuanfang Guo, and Lu Fang. 2018. IHuman3D: Intelligent Human Body 3D Reconstruction Using a Single Flying Camera. In *Proceedings of the 26th ACM International Conference on Multimedia* (Seoul, Republic of Korea) *(MM '18)*. Association for Computing Machinery, New York, NY, USA, 1733–1741. https://doi.org/10.1145/3240508.3240600
[13] Kong Man Cheung, Simon Baker, and Takeo Kanade. 2003. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, Vol. 1. I–I.
[14] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. 2015. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 69.
[15] Yan Cui and Didier Stricker. 2011. 3D Shape Scanning with a Kinect. In *ACM SIGGRAPH 2011 Posters* (Vancouver, British Columbia, Canada) *(SIGGRAPH '11)*. Association for Computing Machinery, New York, NY, USA, Article 57, 1 pages. https://doi.org/10.1145/2037715.2037780
[16] Brian Curless and Marc Levoy. 1996. A Volumetric Method for Building Complex Models from Range Images. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '96)*. ACM, New York, NY, USA, 303–312. https://doi.org/10.1145/237170.237269
[17] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2018. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 224–236.
[18] Martin A Fischler and Robert C Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24, 6 (1981), 381–395.
[19] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. 2017. Look Into Person: Self-Supervised Structure-Sensitive Learning and a New Benchmark for Human Parsing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[20] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. 2020. ARCH: Animatable Reconstruction of Clothed Humans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
[21] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. 2015. Panoptic Studio: A Massively Multiview System for Social Motion Capture. In *Proceedings of the IEEE International Conference on Computer Vision*. 3334–3342.
[22] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. 2018. Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[23] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2018. End-to-end Recovery of Human Shape and Pose. In *Computer Vision and Pattern Regognition (CVPR)*.
[24] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
[25] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 2019. 360-Degree Textures of People in Clothing from a Single Image. In *International Conference on 3D Vision (3DV)*.
[26] Zixin Luo, Tianwei Shen, Lei Zhou, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. 2019. Contextdesc: Local descriptor augmentation with cross-modality context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2527–2536.
[27] Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. 2018. Learning to find good correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2666–2674.
[28] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. 2019. SiCloPe: Silhouette-Based Clothed People. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
[29] Richard A. Newcombe, Dieter Fox, and Steven M. Seitz. 2015. DynamicFusion: Reconstruction and Tracking of Non-Rigid Scenes in Real-Time. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[30] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. 2011. KinectFusion: Real-Time Dense Surface Mapping and Tracking. In *Proc. of ISMAR*. 127–136.
[31] David Nistér. 2004. An efficient solution to the five-point relative pose problem. *IEEE transactions on pattern analysis and machine intelligence* 26, 6 (2004), 0756–777.
[32] PhotoScan [n.d.]. PhotoScan. http://www.agisoft.com/. Accessed: 2020-05-23.
[33] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 652–660.
[34] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*. 5099–5108.
[35] Rahul Raguram, Ondrej Chum, Marc Pollefeys, Jiri Matas, and Jan-Michael Frahm. 2012. USAC: a universal framework for random sample consensus. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2012), 2022–2038.
[36] 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 *The IEEE International Conference on Computer Vision (ICCV)*.
[37] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 2020. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
[38] Johannes L Schonberger and Jan-Michael Frahm. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4104–4113.
[39] Vincent Sitzmann, Michael Zollhoefer, and Gordon Wetzstein. 2019. Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 1121–1132.
[40] Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred Neural Rendering: Image Synthesis Using Neural Textures. *ACM Trans. Graph.* 38, 4, Article 66 (July 2019), 12 pages. https://doi.org/10.1145/3306346.3323035
[41] Justus Thies, Michael Zollhöfer, Christian Theobalt, Marc Stamminger, and Matthias Nießner. 2018. IGNOR: Image-guided neural object rendering. *arXiv preprint arXiv:1811.10720* (2018).
[42] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.
[43] Jing Tong, Jin Zhou, Ligang Liu, Zhigeng Pan, and Hao Yan. 2012. Scanning 3D Full Human Bodies Using Kinects. *IEEE Transactions on Visualization and Computer Graphics* 18, 4 (2012), 643–650.
[44] Treedy's [n.d.]. Treedy's. https://www.treedys.com/. Accessed: 2019-07-25.
[45] Kangkan Wang, Guofeng Zhang, and Shihong Xia. 2017. Templateless Non-Rigid Reconstruction and Motion Tracking With a Single RGB-D Camera. *IEEE Transactions on Image Processing* 26, 12 (Dec 2017), 5966–5979. https://doi.org/10.1109/TIP.2017.2740624
[46] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. 2013. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE International Conference on Computer Vision*. 1625–1632.
[47] Lan Xu, Wei Cheng, Kaiwen Guo, Lei Han, Yebin Liu, and Lu Fang. 2019. FlyFusion: Realtime Dynamic Scene Reconstruction Using a Flying Depth Camera. *IEEE Transactions on Visualization and Computer Graphics* (2019), 1–1.
[48] Lan Xu, Yebin Liu, Wei Cheng, Kaiwen Guo, Guyue Zhou, Qionghai Dai, and Lu Fang. 2017. FlyCap: Markerless Motion Capture Using Multiple Autonomous Flying Cameras. *IEEE Transactions on Visualization and Computer Graphics* PP, 99 (2017), 1–1. https://doi.org/10.1109/TVCG.2017.2728660
[49] Lan Xu, Zhuo Su, Lei Han, Tao Yu, Yebin Liu, and Lu Fang. 2019. UnstructuredFusion: Realtime 4D Geometry and Texture Reconstruction using CommercialRGBD

Cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), 1–1.

[50] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. 2019. DenseRaC: Joint 3D Pose and Shape Estimation by Dense Render-and-Compare. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[51] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. 2016. Lift: Learned invariant feature transform. In *European Conference on Computer Vision*. Springer, 467–483.

[52] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2019. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*. 4471–4480.

[53] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. 2018. DoubleFusion: Real-Time Capture of Human Performances With Inner Body Shapes From a Single Depth Sensor. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[54] Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. 2019. Learning Two-View Correspondences and Geometry Using Order-Aware Network. *arXiv preprint arXiv:1908.04964* (2019).