Deep Volumetric Video From Very Sparse Multi-view Performance Capture

Zeng Huang[1,2(✉)], Tianye Li[1,2], Weikai Chen[2], Yajie Zhao[2], Jun Xing[2], Chloe LeGendre[1,2], Linjie Luo[3], Chongyang Ma[3], and Hao Li[1,2,4]

[1] University of Southern California, Los Angeles, USA
Zenghuan@usc.edu
[2] USC Institute for Creative Technologies, Los Angeles, USA
[3] Snap Inc., Los Angeles, USA
[4] Pinscreen, Santa Monica, USA

Abstract. We present a deep learning based volumetric approach for performance capture using a passive and highly sparse multi-view capture system. State-of-the-art performance capture systems require either pre-scanned actors, large number of cameras or active sensors. In this work, we focus on the task of template-free, per-frame 3D surface reconstruction from as few as three RGB sensors, for which conventional visual hull or multi-view stereo methods fail to generate plausible results. We introduce a novel multi-view Convolutional Neural Network (CNN) that maps 2D images to a 3D volumetric field and we use this field to encode the probabilistic distribution of surface points of the captured subject. By querying the resulting field, we can instantiate the clothed human body at arbitrary resolutions. Our approach scales to different numbers of input images, which yield increased reconstruction quality when more views are used. Although only trained on synthetic data, our network can generalize to handle real footage from body performance capture. Our method is suitable for high-quality low-cost full body volumetric capture solutions, which are gaining popularity for VR and AR content creation. Experimental results demonstrate that our method is significantly more robust and accurate than existing techniques when only very sparse views are available.

Keywords: Human performance capture
Neural networks for multi-view stereo · Wide-baseline reconstruction

1 Introduction

Performance capture is essential for a variety of applications ranging from gaming, visual effects to free-viewpoint videos. The increasing popularity of VR/AR

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-01270-0_21) contains supplementary material, which is available to authorized users.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11220, pp. 351–369, 2018.
https://doi.org/10.1007/978-3-030-01270-0_21

technologies has further triggered the needs for volumetric capture systems, which enables an end-to-end solution for capturing dynamic clothed digital humans.

High-end capture solutions use a large number of cameras and active projections [1–3] or controlled lighting conditions [4,5], and are restricted to professional studio settings. More lightweight systems often use a pre-scanned subject-specific template [6–8], but may produce unnatural baked-in details, e.g. clothing folds. Classic shape-from-silhouette approaches [9] and stereo variants [10] use a visual hull to approximate the target geometry and do not rely on a template mesh. However, surfaces with concavities are difficult to model, and the resulting geometries are often rough when a very sparse number of cameras are used. In most cases, a minimum of eight cameras are required to ensure reasonable results.

To make high-quality performance capture more accessible to end users, we propose a passive motion capture technique without requirements for preprocessing or specialized capture hardware. In particular, our approach is able to faithfully capture detailed human shapes from highly sparse, e.g. three or four, camera views without the need of manual image processing, marker tracking, texture cues, or a pre-scanned mesh template.

Reconstruction from highly sparse views is challenging as large regions of the body are often occluded or not observed by multiple cameras. We tackle this challenge by using a novel multi-view convolutional neural network. Inspired by the shape-from-silhouette method, which reconstructs the target surface by fusing multi-view ray projections from 2D silhouettes, we propose to learn a similar 3D probability field that depicts the surface boundary of a human body using multi-view projection constraints. However, instead of calculating the silhouettes directly, which is either tedious to extract manually or error-prone if computed via automatic segmentation, we use a 2D deep neural network to learn discriminative features that could tell whether a 3D sample point is inside or outside the silhouette. In particular, we associate each 3D point in the space where the object occupies with the features extracted from its projections on multi-view image planes using our convolutional neural network. The per-point features are then fed into a classification network to infer its possibilities of lying inside and outside of the human body. By densely sampling the near-surface region, we obtain a high-resolution volumetric probability field that can be used for reconstructing the body geometry at arbitrary resolutions.

As our proposed network implicitly learns the relations between 3D volume and 2D projections, our approach is capable of reconstructing texture-less surfaces and unseen regions, which is not possible with existing muti-view stereo techniques. For varying input views, e.g. different viewing distances and numbers of captured images, we propose a novel scale-invariant symmetric pooling layer to aggregate features from different views. As a result, our approach scales well to different numbers of input views and produces better reconstruction when more views are available. We evaluate the performance of our network using different numbers of views. Our network is only trained on synthetic data generated using a standard 3D rendering software with animated CG characters. Our method can faithfully capture fast and complex motions with a wide range

of occlusion, backgrounds, and clothing. In addition, we compare our technique with state-of-the-art performance capture methods and demonstrate that our approach is significantly more robust and accurate, when only very sparse views are available.

Our main contributions are:

- A novel performance capture technique that is able to robustly reconstruct clothed human bodies from highly sparse camera views, which was not possible using existing techniques.
- A lightweight performance capture framework that does not require background segmentation, marker tracking, texture cues, or a pre-scanned template model.
- A novel multi-view 2D CNN network that maps multi-view images to a dense 3D probability field, which enables high-resolution reconstruction and robust motion capture from texture-less surfaces.
- A large synthetic dataset of clothed human body animations rendered on multiple views, containing 50 characters and 13 animation sequences for each subject.

2 Related Work

Silhouette-Based Multi-view Reconstruction. Visual hulls created from multi-view silhouette images are widely used for multi-view reconstruction, [6,10–15], since they are fast and easy to compute and well approximate the underlying 3D geometry. Further progresses have been made to the visual-hull-based viewing experience [9], smoothing the geometry with fewer cameras [16], and real-time performance [17]. Approaches have also emerged to recover geometric details using multi-view constraints [18–20] and photometric stereo [4,21]. Recently, Collet et al. [1] introduced a system for high-quality free-viewpoint video by fusing multi-view RGB, IR and silhouette inputs.

Despite the speed and robustness of silhouette-based reconstruction methods, their reliance on visual hulls implies bias against surface concavities as well as susceptibility to artifacts in invisible space.

Human Body Performance Capture. Actor-specific shape priors can be incorporated to improve the reconstruction quality for human body performance capture [7,22–25]. Additional improvements have been proposed using kinematic skeletons [26,27], segmentation of moving subjects [27–32] and human parametric models [33–39], even enabling single-view reconstruction [40–44]. To obtain even higher accuracy and robustness, multi-view depth based approaches are actively explored [45–48]. Orts-Escolano et al. [2] employ active stereo cameras and highly specialized acquisition devices for real-time high-quality capture. Wang et al. [49] use sparse depth sensors and RGB inputs to capture moving subject with textures. In comparison, our method does not require any active sensors and is more accessible to common users.

Further efforts have focused on capturing dynamic details such as clothing folds using shape-from-shading [50], photometric stereo [51,52] or implicit modeling of deformable meshes [53]. To reduce the computational cost for the inverse

rendering problem in many of these approaches, Pons-Moll et al. [54] propose a multi-cloth 3D model to reconstruct both body and clothes from 4D scan sequences, estimating an unclothed body shape using [55,56] and tracking the clothing over time. More recently, Xu et al. [8] reconstruct a human body wearing general clothing. However, this approach requires each actor to be scanned in advance for a template mesh and skeleton. In contrast, our method reconstructs the mesh in a fully automatic way without needing any template model.

Multi-view 3D Deep Learning. Multi-view convolutional neural networks (CNN) have been introduced to learn deep features for various 3D tasks including shape segmentation [57], object recognition and classification [58–60], correspondence matching [61], and novel view synthesis [62–64]. More closely related, a number of previous works apply multi-view CNNs to 3D reconstruction problems in both unsupervised [65] and supervised approaches to obtain the final geometry directly [66,67], or indirectly via normal maps [68], silhouettes [69], or color images [70]. Inspired by the multi-view stereo constraint, others [71,72] have formulated ray consistency and feature projection in a differentiable manner, incorporating this formulation into an end-to-end network to predict a volumetric representation of a 3D object.

Hartmann et al. [73] propose a deep learning based approach to predict the similarity between the image patches across multiple views, which enables 3D reconstruction using stereopsis. In contrast, our approach aims for a different and more challenging task of predicting per-point possibility of lying on the reconstructed surface, and directly connects 3D volume and its 2D projections on the image planes. Closer to our work, Ji et al. [74] propose a learned metric to infer the per-voxel possibility of being on the reconstructed surface in a volumetric shape representation. However, due to the reliance on multi-view stereopsis, these methods [73,74] fail to faithfully reconstruct textureless surfaces and generate dense reconstruction from sparse views. In addition, as both the input images and the output surface need to be converted into volumetric representations, it remains difficult for prior methods to generate high-resolution results. Our approach, on the other hand, can work on textureless surfaces and produce results with much higher resolution by leveraging an implicit representation.

Additionally, Dibra et al. [75] propose a cross-modal neural network that captures parametric body shape from a single silhouette image. However, this method can only predict naked body shapes in neutral poses, while our approach generalizes well to dynamic clothed bodies in extreme poses.

3 Overview

Given multiple views and their corresponding camera calibration parameters as input, our method aims to predict a dense 3D field that encodes the probabilistic distribution of the reconstructed surface. We formulate the probability prediction as a classification problem. At a high level, our approach resembles the spirit of the shape-from-silhouette method: reconstructing the surface according to the consensus from multi-view images on any 3D point staying inside

the reconstructed object. However, instead of directly using silhouettes, which only contain limited information, we leverage the deep features learned from a multi-view convolution neural network.

As demonstrated in Fig. 1, for each query point in the 3D space, we project it onto the multi-view image planes using the input camera parameters. We then collect the multi-scale CNN features learned at each projected location and aggregate them through a pooling layer to obtain the final global feature for the query point. The per-point feature is later fed into a classification network to infer its possibilities of lying inside and outside the reconstructed object respectively. As our method outputs a dense probability field, the surface geometry can be faithfully reconstructed from the field using marching cube reconstruction.

We introduce the multi-view based probability inference network and training details in Sect. 4. In Sect. 5, we will detail the surface reconstruction.

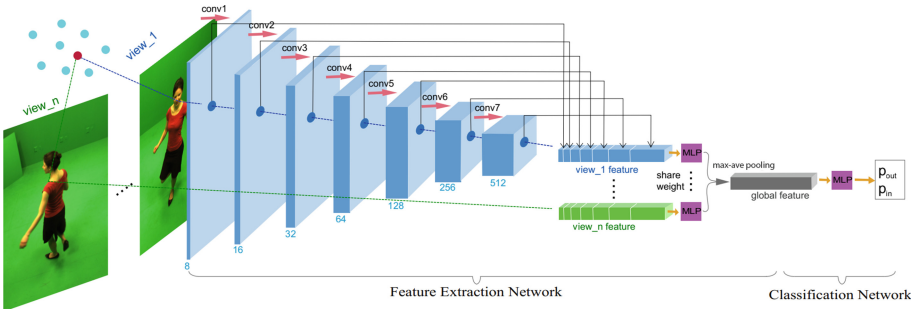

Fig. 1. Network architecture.

4 Multi-view Based Probability Inference Network

4.1 Network Architecture

Our network consists of two parts: one feature extraction network that learns discriminative features for each query point in the 3D space, and one classification network that consumes the output of the preceding network and predicts the per-point possibilities of lying inside and outside the reconstructed body. Both networks are trained in an end-to-end manner.

Feature Extraction. The feature extraction network takes multi-view images along with their corresponding camera calibration parameters and 3D query points as input. The multi-view images are first passed to a shared-weight fully convolutional network, whose building block includes a convolutional layer, an *Relu* activation layer, and a pooling layer. Batch normalization [76] is utilized in each convolutional layer.

We then associate each query point p_i with its features by projecting it onto the multi-view image planes. Let q_{ij} denote p_i's projection onto image plane j.

As shown in Fig. 1, we track each q_{ij} throughout the feature maps at each level of convolutional layers. The features retrieved from each layer at the projected location are concatenated to obtain a single-view feature vector $\mathcal{F}_{ij}$.

Since the view projection has floating-point precision, ambiguity may arise for feature extraction if the projected point lies on the boundary between two adjacent pixels. To address this issue, at each level of the feature maps, we perform bilinear interpolation on the nearest four pixels according to the local coordinate of the projected location. It is worth mentioning that by applying bilinear interpolation, our method further increases the receptive field of the feature vector at each layer, and makes the network more robust against boundary points around the silhouette. If the projection of a query point is out of scope of the input image, we fill its feature vector with zeros and do not include it in the back propagation.

Scale-invariant Symmetric Pooling. After obtaining the feature vector $\mathcal{F}_{ij}$ from each view j, one key module must effectively aggregate these view-dependent signatures. However, the viewing distance and focal length may differ for each camera, and so the scales of projections of the same 3D volume may vary significantly from viewpoint to viewpoint. As a result, features on the same level of convolutional layers may have different 3D receptive fields across different views. Therefore, direct element-wise pooling on view-dependent features may not be effective, as it could be operated on mismatched scales.

To resolve this issue, we introduce shared-weight MLP layers *before* the pooling operation so that multi-scale features will be more uniformly distributed to all element entries, enabling the follow-up pooling module to be feature scale invariant. Then, we apply a permutation invariant pooling module on the output feature vectors of the MLP layers. The outputs of the pooling module are the final feature vector associated with each query point.

Classification Network. After obtaining a feature vector for a query point, we employ a classification network to infer its probability of being on the reconstructed surface. A simple structure consisting of multiple fully connected layers is used for this classification task. In particular, we predict two labels (P_{in}, P_{out}) for each point, where P_{in} and P_{out} stand for the possibility of the 3D point being inside and outside the reconstructed object, respectively. For a query point p and a ground-truth mesh $\mathcal{M}$, if p is inside $\mathcal{M}$, we mark its labels as $(1, 0)$; if p lies on the surface, it is marked as $(1, 1)$; otherwise, p is marked as $(0, 1)$.

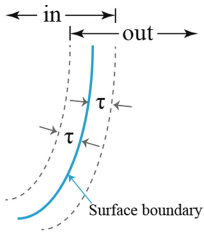

Fig. 2. Classification boundary.

In reality, only very few sample points lay exactly on the surface. To better capture the surface, we relax the criteria for determining the inside/outside labels. As shown in Fig. 2, in addition to the points inside the surface, we also include those outside points whose distance to the surface is below a threshold τ (τ is set as 1 cm) and mark their P_{in} label as 1. Similarly, we apply the same threshold to mark P_{out}. Therefore, points in the near-surface region are labeled as both $(1, 1)$. We predict the two labels independently and train the network

using sigmoid cross-entropy loss. Therefore, the predicted value of P_{in} and P_{out} ranges from 0 to 1, where a larger value indicates a higher probability. More details of network design are provided in the *supplementary materials*.

4.2 Network Training

As our approach aims to predict a dense probability field, for each 3D mesh, it is necessary to generate a large amount of query points for training the network. However, uniformly sampling the 3D space would be prohibitive in terms of computational cost. In fact, we only care about points that are near the final reconstructed surface. We therefore adopt an adaptive sampling strategy to emphasize our sampling on such points. For each ground-truth mesh $\mathcal{M}$, we first generate a regular point grid with resolution 256^3 filling the space of an enlarged (1.5 times) bounding box of $\mathcal{M}$. We compute signed distances of the grid points with the method given by [77]. We then calculate the largest distance l from the interior grid point to $\mathcal{M}$'s surface: $l = |\min_i \mathrm{dist}(\mathrm{t_i}, \mathcal{M})|$.

To select points that are more centered around the surface of $\mathcal{M}$, we utilize Monte Carlo sampling to keep those grid points t_i whose distance $|\mathrm{dist}(t_i, \mathcal{M})|$ satisfies the Gaussian distribution: $\mathrm{norm}(\mu = 0, \sigma = l)$. For each of the combinations of multi-view images and their camera matrices that will appear in the training, we augment the data by firstly reconstructing the visual hull from the input views; and then randomly sampling more points inside the visual hull but ensuring the newly added points achieve an equal distribution inside and outside the ground-truth mesh $\mathcal{M}$. We stop adding samples when the total number of query points for each $\mathcal{M}$ reaches $100,000$.

We train the network using various combinations of camera views. For a certain number of views (3, 4 or 8), we train an individual model. We test each model using corresponding number of views. The combinations of views are selected such that every adjacent two of them have a wide baseline and all the views together cover the entire subject in a loop. The query points and their labels for each mesh are pre-computed so as to save training time.

During training, we randomly draw $10,000$ query points from the pre-computed set for each sample. We directly take color images of each view as inputs in their original resolutions, which varies from 1600×1200 to 1920×1080. For each batch, we only load images from one multi-view scene due to the limited GPU memory. The network is optimized using Adam optimizer. We start with a learning rate of 0.00001 and gradually decay it exponentially every $100,000$ batches with a factor of 0.7. We train the network for 20 epochs on a single NVIDIA GTX 1080Ti GPU.

5 Surface Reconstruction

At test time, we first use our network to generate a dense probability field from the input images. As the near-surface region only occupies little volume compared to the space it encloses, it is highly inefficient to apply uniform sampling

over the space. Therefore we employ an octree-based approach to achieve a high-resolution reconstruction with a low computational cost. In particular, we first compute the center of the scene according to the camera positions and their calibration parameters. A bounding box of length 3 meters on each side is placed at the scene center. We then fill the bounding box with a regular 3D grid. By traversing each cube in the grid, we subdivide those cubes whose centers are surface points, or whose vertices consist both inside and outside points, recognized by our network. As our network predicts two probabilities (P_{in}, P_{out}) per point, we propose to aggregate the two probabilities into one signed distance for surface point prediction and later reconstruction of the entire surface.

As discussed in Sect. 4.2 and illustrated in Fig. 2, P_{in} and P_{out} indicate the relaxed probabilities of being inside and outside the object, respectively. Since P_{in} and P_{out} are independent events, the probability of a point being near the surface can be simply computed as: $P_{surf} = P_{in} \times P_{out}$. By excluding the near-surface region (defined above), we define the probability of reliably staying inside the object as $P'_{in} = P_{in} \times (1 - P_{out})$. Similarly, the probability of lying in the outer region but having point-to-mesh distance larger than τ can be calculated as $P'_{out} = P_{out} \times (1 - P_{in})$. We compute all three probabilities $\{P_{surf}, P'_{in}, P'_{out}\}$ for each grid point. We then determine the signed distance value for each point by selecting the largest probability. In particular, we only assign three discrete signed distance values: $\{-1, 0, 1\}$, which represent inner, surface and outer points respectively. For instance, for one query point, if its P_{surf} is larger than the other probabilities, it will be assigned with 0 and treated as a surface point. A similar strategy is applied to determine inner and outer points and to assign their corresponding signed distances.

We then generate a dense signed distance field in a coarse-to-fine manner. As discussed previously, we subdivide those cubes marked by the network, further infer the signed distance for all the octant cubes, and iterate until a target resolution is achieved. Finally, after obtaining the signed distance field, we use the marching cubes algorithm to reconstruct the surface whose signed distance equals 0.

6 Results

6.1 Dataset

A good training set of sufficient size is key to a successful deep learning model. However, existing datasets of multi-view clothed body capture usually consist of only a few subjects, making them unsuitable for training a deep neural network. SURREAL dataset [78] has large amount of synthetic humans but it does not contain geometric details of clothes and thus is not suitable for our task.

We therefore generate a synthetic dataset by rendering rigged and animated human character models from Mixamo [79] as seen from multiple views. The characters share the same rig, and so a variety of animations and human poses could be rapidly synthesized of different figures dressed in many clothing types and styles. In total we render images with 50 characters and 13 animations,

for eight camera viewpoints with known projection matrices. We use 43 characters and 10 animations for training. The remaining seven characters and three animations are used for validation and testing.

6.2 Evaluations

In this section, we evaluate our model on various datasets, including [4,6,18], as well as our own synthetic data. For real-world datasets whose original backgrounds are removed, we composite a green background according to the provided segmentation.

Qualitative Results. We first reconstruct these results from four views on grids of resolution 1024^3 as shown in Fig. 3. All the results are generated directly from our pipeline without any post-processing except edge collapse to reduce file sizes. All the results are generated from test cases. To validate the accuracy of the reconstructed geometry, we colorize each vertex with the visible cameras with simple cosine weight blending. Our rendering results could be further improved via recent real-time [80] or offline [81] texturing approaches.

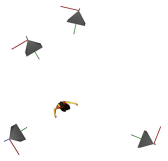

Fig. 3. Camera setting for reported four-view results.

Figure 6 shows the camera setting for the results. From only four-view inputs with limited overlap between each of them, our network reconstructs a watertight surface that resembles the subject's geometry and recovers reasonable local details. Even for ambiguous areas where no cameras have line-of-sight, our network can still predict plausible shapes. We also present results on a sequence of challenging motion performance from Vlasic et al. [6] in Fig. 4. Even for the challenging poses and extreme occlusion, our network can robustly recover a plausible shape.

As our network is not limited by the number of views, we train and test our models with different numbers of views. We test our model with three-view, four-view, and eight-view settings, selecting input views incrementally. As shown in Fig. 5, with more views, our network can predict more details, e.g. facial shape and hairstyle. For most of the results shown in this paper, we use a four-view setting, which achieves the best balance between view-sparsity and reconstruction quality.

Quantitative Results. We evaluate our reconstruction accuracy by measuring Euclidean distance from reconstructed surface vertices to the reference scan. For real world data, we use results given by [6] and [18] as our references, which approximate the ground-truth surface using a much more advanced capturing setup. We show visualizations for the mesh-to-scan distances and evaluate the distance statistics.

As shown in Fig. 3, given inputs from various test sets, our network predicts accurate surface, with median mesh-to-scan distance of all examples less than 0.9 cm. As shown in Fig. 4, our network also predicts accurate reconstruction for the challenging input image sequences, with median mesh-to-scan distance below

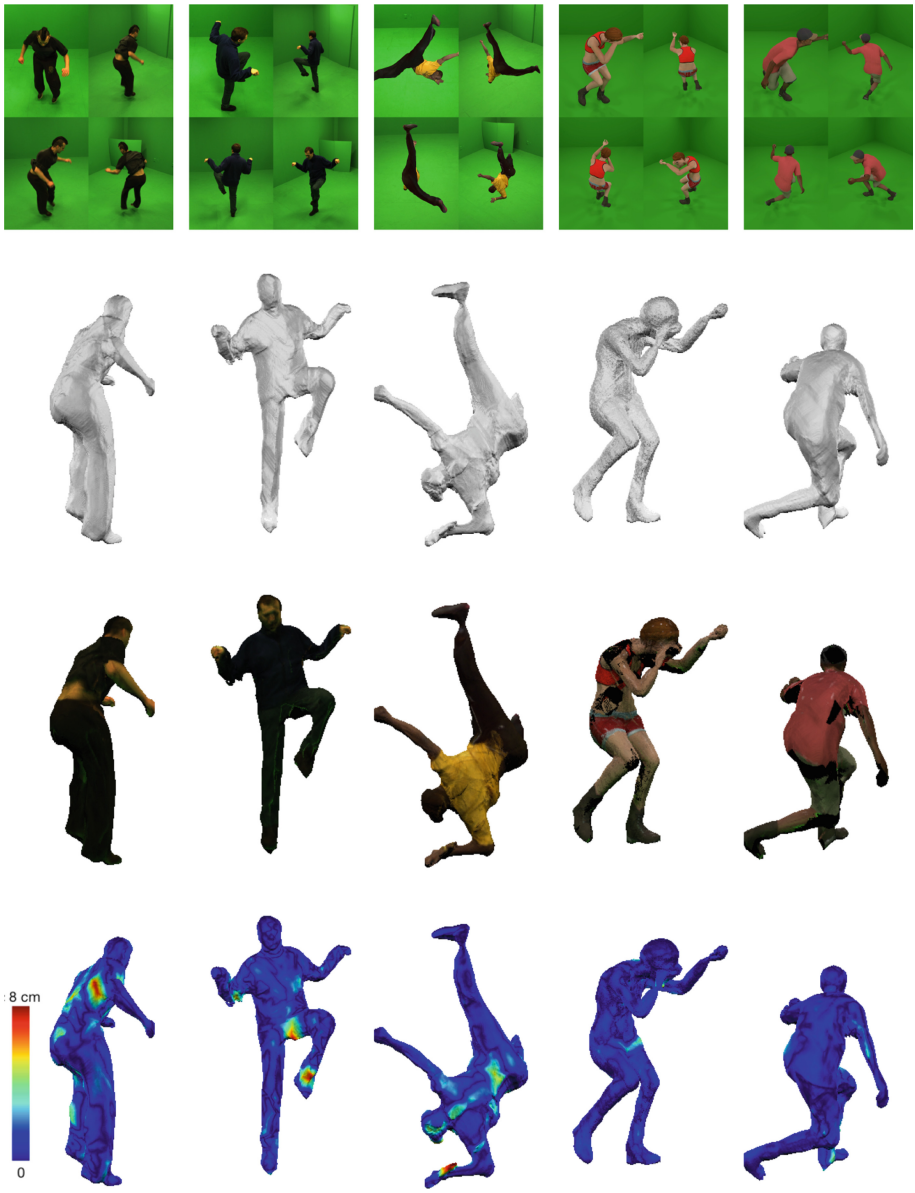

Fig. 4. Results reconstructed from four views Top to bottom rows: input multi-view images, reconstructed mesh, textured mesh, and error visualization. From left to right, median mesh-to-scan distance: 0.90 cm, 0.66 cm, 0.85 cm, 0.54 cm, 0.59 cm; mean mesh-to-scan distance: 1.18 cm, 0.88 cm, 1.10 cm, 0.65 cm, 0.76 cm.

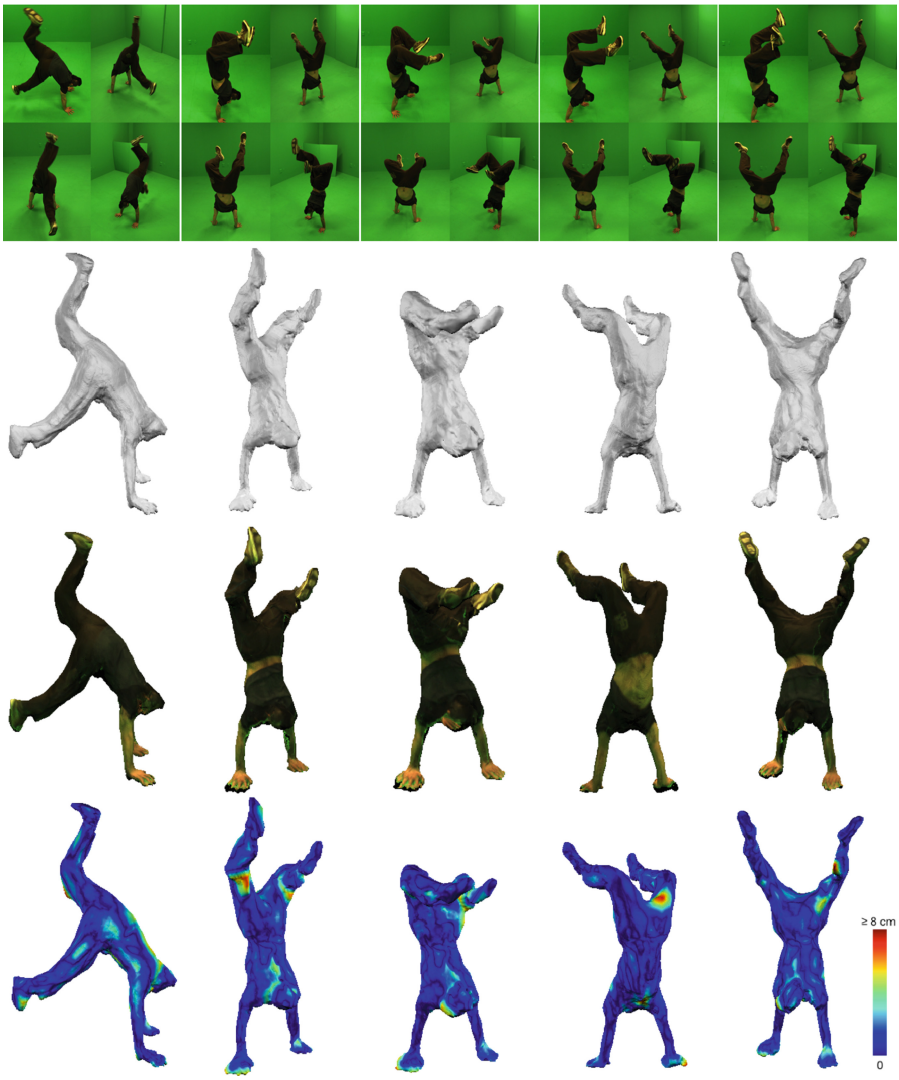

Fig. 5. Sequence results. Top to bottom rows: multi-view images, reconstructed mesh, textured mesh, and error visualization. From left to right, median mesh-to-scan distance: 0.94 cm, 0.86 cm, 0.82 cm, 0.76 cm, 0.85 cm; mean mesh-to-scan distance: 1.31 cm, 1.27 cm, 1.21 cm, 1.06 cm, 1.25 cm.

0.95 cm. In Fig. 5, we observe that the distance error decreases as more views are available during network training. The median distance for 8-view drops below half of the distance as for the three-view training setting.

Fig. 6. Reconstructions with different views. Top to bottom rows: reconstructed mesh, textured mesh, and error visualization. Left to right columns: three-view results, four-view results, and eight-view-view results, for both two test cases respectively. Median mesh-to-scan distance: left subject: 0.84 cm (three-view), 0.77 cm (four-view), 0.45 cm (eight-view); right subject: 1.38 cm (three-view), 1.06 cm (four-view), 0.59 cm (eight-view).

6.3 Comparisons

In this section we compare our approach with existing methods using four-view input in Fig. 7.

While traditional multi-view stereo PMVS [82] is able to reconstruct an accurate point cloud, it often fails to produce complete geometry with large baseline (four views to cover 360 degree in this case) and texture-less inputs. As a learning-based approach, SurfaceNet [74] reconstructs a more complete point cloud, but still fails at the region with fewer correspondences due to large baseline. It remains difficult to reconstruct a complete surface from sparse point clouds results of PMVS and SurfaceNet. Although visual hull [17] based approach can reconstruct a complete shape, the reconstruction deviates significantly from the true shape due to its incapability of capturing concavities. On the contrary, our method is able to reconstruct a complete model with the clothed human shape well sculpted given as few as four views.

In terms of runtime, PMVS takes 3.2 seconds with 12 threads using four views. Since SurfaceNet is not designed to reconstruct object in 360 degrees, we run on neighboring views for four times and then fuse them to obtain a

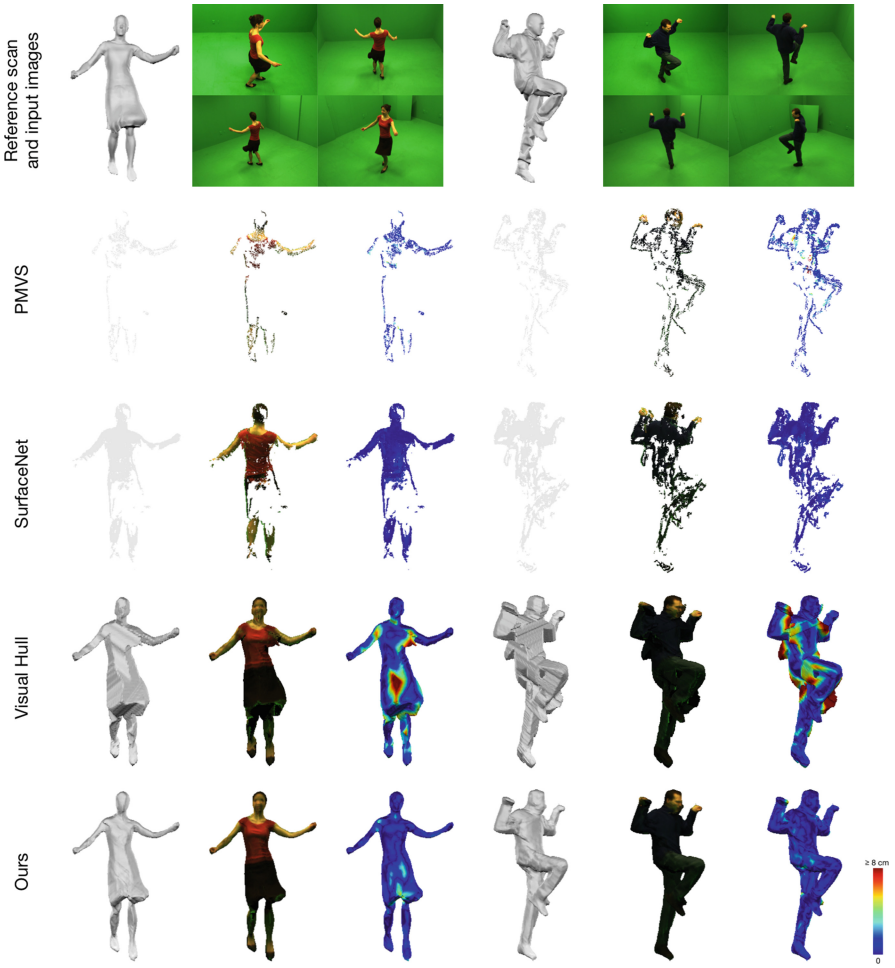

Fig. 7. Comparisons. Top to bottom rows: input 4-view images, PMVS, SurfaceNet, visual hull, and ours.

complete reconstruction. This process takes 15 minutes with one Titan X GPU. For visual hull, it takes 30 ms on Titan X GPU at 512^3 resolution with an octree implementation. Our multi-view network takes 4.4 seconds for 256^3 resolution, and 18 seconds for 512^3 resolution with an octree implementation on GTX 1080 Ti. Since operations for image feature extraction, pooling, point query, octree transverses, and marching cubes can all be done distributively in parallel, the performance of our method could be potentially further boosted.

7 Discussion and Conclusion

In this work, we present a fully-automatic lightweight solution to the challenging problem of dynamic human body performance capture, without requiring active lighting, explicit foreground segmentation, specialized tracking hardware, or a human body template. Using only sparse-view RGB images as input, our novel multi-view CNN encodes the probability of a point lying on the surface of the capture subject, enabling subsequent high resolution surface reconstruction. Our network architecture, including the scale-invariant symmetric pooling, ensures the robustness of our approach, even under as few as three input views.

Since only trained on synthetic data where all training subjects were rendered in a virtual green screen room, our current implementation does not generalize to handle input images with arbitrarily complex backgrounds. We have experimented with training our network using data composited with a small set of random backgrounds. However, the results are not satisfactory (Fig. 8). Also, results of using unseen camera views that are significantly different from our training data can be less ideal.

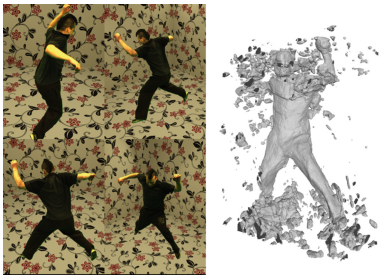

Fig. 8. Failure case.

It would be a future avenue to introduce larger variations into the training data including complex backgrounds, additional camera viewpoints from which to sample, and various lighting conditions. It is also interesting to explore the problem of unconstrained reconstruction, i.e., how to faithfully capture human motion from highly sparse viewpoints even when the camera calibration parameters are not available.

Acknowledgement. We would like to thank the authors of [74] who helped testing with their system. This work was supported in part by the ONR YIP grant N00014-17-S-FO14, the CONIX Research Center, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA, the Andrew and Erna Viterbi Early Career Chair, the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005, Adobe, and Sony. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

1. Collet, A., et al.: High-quality streamable free-viewpoint video. ACM Trans. Graph. (TOG) **34**(4), 69 (2015)
2. Orts-Escolano, S., et al.: Holoportation: Virtual 3d teleportation in real-time. In: Proceedings of the 29th Annual Symposium on User Interface Software and Technology, pp. 741–754. ACM (2016)

3. Joo, H., et al.: Panoptic studio: a massively multiview system for social motion capture. In: The IEEE International Conference on Computer Vision (ICCV) (2015)
4. Vlasic, D., et al.: Dynamic shape capture using multi-view photometric stereo. ACM Trans. Gr. (TOG) **28**(5), 174 (2009)
5. Li, H., et al.: Temporally coherent completion of dynamic shapes. ACM Trans. Gr. (TOG) **31**(1), 2 (2012)
6. Vlasic, D., Baran, I., Matusik, W., Popović, J.: Articulated mesh animation from multi-view silhouettes. ACM Trans. Gr. (TOG) **27**, 97 (2008). ACM
7. De Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.P., Thrun, S.: Performance capture from sparse multi-view video. ACM Trans. Gr. (TOG) **27**, 98 (2008). ACM
8. Xu, W., et al.: Monoperfcap: Human performance capture from monocular video. arXiv preprint arXiv:1708.02136 (2017)
9. Matusik, W., Buehler, C., Raskar, R., Gortler, S.J., McMillan, L.: Image-based visual hulls. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, pp. 369–374. ACM Press/Addison-Wesley Publishing Co. (2000)
10. Furukawa, Y., Ponce, J.: Carved visual hulls for image-based modeling. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 564–577. Springer, Heidelberg (2006). https://doi.org/10.1007/11744023_44
11. Esteban, C.H., Schmitt, F.: Silhouette and stereo fusion for 3d object modeling. Comput. Vis. Image Underst. **96**(3), 367–392 (2004)
12. Cheung, G.K., Baker, S., Kanade, T.: Visual hull alignment and refinement across time: A 3d reconstruction algorithm combining shape-from-silhouette with stereo. In: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on. Volume 2, IEEE (2003) II-375
13. Song, D., Tong, R., Chang, J., Yang, X., Tang, M., Zhang, J.J.: 3d body shapes estimation from dressed-human silhouettes. In: Computer Graphics Forum, Vol. 35, pp. 147–156 (2016). Wiley Online Library
14. Zuo, X., Du, C., Wang, S., Zheng, J., Yang, R.: Interactive visual hull refinement for specular and transparent object surface reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2237–2245 (2015)
15. Liu, Y., Dai, Q., Xu, W.: A point-cloud-based multiview stereo algorithm for free-viewpoint video. IEEE Trans. Vis. Comput. Gr. **16**(3), 407–418 (2010)
16. Franco, J.S., Lapierre, M., Boyer, E.: Visual shapes of silhouette sets. In: Third International Symposium on 3D Data Processing, Visualization, and Transmission, pp. 397–404. IEEE (2006)
17. Loop, C., Zhang, C., Zhang, Z.: Real-time high-resolution sparse voxelization with application to image-based modeling. In: Proceedings of the 5th High-Performance Graphics Conference, pp. 73–79. ACM (2013)
18. Starck, J., Hilton, A.: Surface capture for performance-based animation. IEEE Comput. Gr. Appl. **27**(3) (2007)
19. Zitnick, C.L., Kang, S.B., Uyttendaele, M., Winder, S., Szeliski, R.: High-quality video view interpolation using a layered representation. ACM Trans. Gr. (TOG) **23**, 600–608 (2004) . ACM
20. Waschbüsch, M., Würmlin, S., Cotting, D., Sadlo, F., Gross, M.: Scalable 3d video of dynamic scenes. Vis. Comput. **21**(8), 629–638 (2005)
21. Wu, C., Varanasi, K., Liu, Y., Seidel, H.P., Theobalt, C.: Shading-based dynamic shape refinement from multi-view video under general illumination. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 1108–1115. IEEE (2011)

22. Ahmed, N., Theobalt, C., Dobrev, P., Seidel, H.P., Thrun, S.: Robust fusion of dynamic shape and normal capture for high-quality reconstruction of time-varying geometry. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008, CVPR 2008, pp. 1–8. IEEE (2008)
23. Stoll, C., Gall, J., De Aguiar, E., Thrun, S., Theobalt, C.: Video-based reconstruction of animatable human characters. ACM Trans. Gr. (TOG) **29**(6), 139 (2010)
24. Bradley, D., Popa, T., Sheffer, A., Heidrich, W., Boubekeur, T.: Markerless garment capture. ACM Trans. Gr. (TOG) **27**, 99 (2008). ACM
25. Wu, C., Varanasi, K., Theobalt, C.: Full Body performance capture under uncontrolled and varying illumination: a shading-based approach. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7575, pp. 757–770. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33765-9_54
26. Gall, J., Stoll, C., De Aguiar, E., Theobalt, C., Rosenhahn, B., Seidel, H.P.: Motion capture using joint skeleton tracking and surface estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009, CVPR 2009, pp. 1746–1753. IEEE (2009)
27. Liu, Y., Stoll, C., Gall, J., Seidel, H.P., Theobalt, C.: Markerless motion capture of interacting characters using multi-view image segmentation. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1249–1256. IEEE (2011)
28. Bray, M., Kohli, P., Torr, P.H.S.: PoseCut: simultaneous segmentation and 3D pose estimation of humans using dynamic graph-cuts. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 642–655. Springer, Heidelberg (2006). https://doi.org/10.1007/11744047_49
29. Brox, T., Rosenhahn, B., Cremers, D., Seidel, H.-P.: high accuracy optical flow serves 3-D pose tracking: exploiting contour and flow based constraints. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 98–111. Springer, Heidelberg (2006). https://doi.org/10.1007/11744047_8
30. Brox, T., Rosenhahn, B., Gall, J., Cremers, D.: Combined region and motion-based 3d tracking of rigid and articulated objects. IEEE Trans. Pattern Anal. Mach. Intell. **32**(3), 402–415 (2010)
31. Mustafa, A., Kim, H., Guillemaut, J.Y., Hilton, A.: General dynamic scene reconstruction from multiple view video. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 900–908 (2015)
32. Wu, C., Stoll, C., Valgaerts, L., Theobalt, C.: On-set performance capture of multiple actors with a stereo camera. ACM Trans. Gr. (TOG) **32**(6), 161 (2013)
33. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM Trans. Gr. (TOG) **34**(6), 248 (2015)
34. Loper, M., Mahmood, N., Black, M.J.: Mosh: motion and shape capture from sparse markers. ACM Trans. Gr. (TOG) **33**(6), 220 (2014)
35. Hasler, N., Ackermann, H., Rosenhahn, B., Thormählen, T., Seidel, H.P.: Multilinear pose and body shape estimation of dressed subjects from image sets. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1823–1830. IEEE (2010)
36. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. ACM Trans. Gr. (TOG) **24**, 408–413 (2005). ACM
37. Balan, A.O., Sigal, L., Black, M.J., Davis, J.E., Haussecker, H.W.: Detailed human shape and pose from images. In: IEEE Conference on Computer Vision and Pattern Recognition, 2007, CVPR 2007, pp. 1–8. IEEE (2007)

38. Plänkers, R., Fua, P.: Tracking and modeling people in video sequences. Comput. Vis. Image Underst. **81**(3), 285–302 (2001)
39. Sminchisescu, C., Triggs, B.: Estimating articulated human motion with covariance scaled sampling. Int. J. Robot. Res. **22**(6), 371–391 (2003)
40. Tan, J.K.V., Budvytis, I., Cipolla, R.: Indirect deep structured learning for 3d human body shape and pose prediction
41. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: automatic estimation of 3D human pose and shape from a single image. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 561–578. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_34
42. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. arXiv preprint arXiv:1712.06584 (2017)
43. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3d and 2d human representations. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
44. Guan, P., Weiss, A., Balan, A.O., Black, M.J.: Estimating human shape and pose from a single image. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1381–1388. IEEE (2009)
45. Dou, M., Fuchs, H., Frahm, J.M.: Scanning and tracking dynamic objects with commodity depth cameras. In: 2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp. 99–106. IEEE (2013)
46. Dou, M., et al.: Fusion4d: real-time performance capture of challenging scenes. ACM Trans. Gr. (TOG) **35**(4), 114 (2016)
47. Ye, G., Liu, Y., Hasler, N., Ji, X., Dai, Q., Theobalt, C.: Performance capture of interacting characters with handheld kinects. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, pp. 828–841. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33709-3_59
48. Zollhöfer, M., et al.: Real-time non-rigid reconstruction using an rgb-d camera. ACM Trans. Gr. (TOG) **33**(4), 156 (2014)
49. Wang, R., Wei, L., Vouga, E., Huang, Q., Ceylan, D., Medioni, G., Li, H.: Capturing dynamic textured surfaces of moving targets. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 271–288. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_17
50. Tylecek, R., Sara, R.: Refinement of surface mesh for accurate multi-view reconstruction. Int. J. Virtual Real. **9**(1), 45–54 (2010)
51. Wu, C., Liu, Y., Dai, Q., Wilburn, B.: Fusing multiview and photometric stereo for 3d reconstruction under uncalibrated illumination. IEEE Trans. Vis. Comput. Gr. **17**(8), 1082–1095 (2011)
52. Hernández, C., Vogiatzis, G., Brostow, G.J., Stenger, B., Cipolla, R.: Non-rigid photometric stereo with colored lights. In: IEEE 11th International Conference on Computer Vision, 2007, ICCV 2007, pp. 1–8 IEEE (2007)
53. Robertini, N., Casas, D., De Aguiar, E., Theobalt, C.: Multi-view performance capture of surface details. Int. J. Comput. Vis. **124**, 1–18 (2017)
54. Pons-Moll, G., Pujades, S., Hu, S., Black, M.: Clothcap: Seamless 4d clothing capture and retargeting. ACM Trans. Gr. (Proc. SIGGRAPH) [to appear] 1 (2017)
55. Zhang, C., Pujades, S., Black, M., Pons-Moll, G.: Detailed, accurate, human shape estimation from clothed 3D scan sequences. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) Spotlight
56. Yang, J., Franco, J.-S., Hétroy-Wheeler, F., Wuhrer, S.: Estimation of human body shape in motion with wide clothing. In: Leibe, B., Matas, J., Sebe, N., Welling, M.

(eds.) ECCV 2016. LNCS, vol. 9908, pp. 439–454. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_27
57. Kalogerakis, E., Averkiou, M., Maji, S., Chaudhuri, S.: 3d shape segmentation with projective convolutional networks. In: Proceedings of CVPR, 2. IEEE (2017)
58. Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E.: Multi-view convolutional neural networks for 3d shape recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 945–953 (2015)
59. Shi, B., Bai, S., Zhou, Z., Bai, X.: Deeppano: deep panoramic representation for 3-d shape recognition. IEEE Signal Process. Lett. **22**(12), 2339–2343 (2015)
60. Qi, C.R., Su, H., Nießner, M., Dai, A., Yan, M., Guibas, L.J.: Volumetric and multi-view cnns for object classification on 3d data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5648–5656 (2016)
61. Huang, H., Kalogerakis, E., Chaudhuri, S., Ceylan, D., Kim, V.G., Yumer, E.: Learning local shape descriptors from part correspondences with multiview convolutional networks. ACM Trans. Gr. (TOG) **37**(1), 6 (2018)
62. Su, H., Wang, F., Yi, L., Guibas, L.: 3d-assisted image feature synthesis for novel views of an object. arXiv preprint arXiv:1412.0003 (2014)
63. Zhou, T., Tulsiani, S., Sun, W., Malik, J., Efros, A.A.: View synthesis by appearance flow. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 286–301. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_18
64. Park, E., Yang, J., Yumer, E., Ceylan, D., Berg, A.C.: Transformation-grounded image generation network for novel 3d view synthesis. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 702–711. IEEE (2017)
65. Rezende, D.J., Eslami, S.A., Mohamed, S., Battaglia, P., Jaderberg, M., Heess, N.: Unsupervised learning of 3d structure from images. In: Advances In Neural Information Processing Systems, pp. 4996–5004 (2016)
66. Sinha, A., Unmesh, A., Huang, Q., Ramani, K.: Surfnet: Generating 3d shape surfaces using deep residual networks. In: Proceedings of CVPR (2017)
67. Choy, C.B., Xu, D., Gwak, J.Y., Chen, K., Savarese, S.: 3D-R2N2: a unified approach for single and multi-view 3D object reconstruction. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 628–644. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_38
68. Lun, Z., Gadelha, M., Kalogerakis, E., Maji, S., Wang, R.: 3d shape reconstruction from sketches via multi-view convolutional networks. arXiv preprint arXiv:1707.06375 (2017)
69. Soltani, A.A., Huang, H., Wu, J., Kulkarni, T.D., Tenenbaum, J.B.: Synthesizing 3d shapes via modeling multi-view depth maps and silhouettes with deep generative networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1511–1519 (2017)
70. Tatarchenko, M., Dosovitskiy, A., Brox, T.: Multi-view 3D models from single images with a convolutional network. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 322–337. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_20
71. Tulsiani, S., Zhou, T., Efros, A.A., Malik, J.: Multi-view supervision for single-view reconstruction via differentiable ray consistency. In: CVPR, Vol. 1, p. 3 (2017)
72. Kar, A., Häne, C., Malik, J.: Learning a multi-view stereo machine. In: Advances in Neural Information Processing Systems, pp. 364–375 (2017)
73. Hartmann, W., Galliani, S., Havlena, M., Van Gool, L., Schindler, K.: Learned multi-patch similarity. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 1595–1603. IEEE (2017)

74. Ji, M., Gall, J., Zheng, H., Liu, Y., Fang, L.: Surfacenet: an end-to-end 3d neural network for multiview stereopsis. arXiv preprint arXiv:1708.01749 (2017)
75. Dibra, E., Jain, H., Oztireli, C., Ziegler, R., Gross, M.: Human shape from silhouettes using generative hks descriptors and cross-modal neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vol. 5 (CVPR), Honolulu, HI, USA (2017)
76. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456 (2015)
77. Xu, H., Barbič, J.: Signed distance fields for polygon soup meshes. Graphics Interface 2014 (2014)
78. Varol, G., et al.: Learning from synthetic humans. In: CVPR (2017)
79. Adobe: Mixamo (2013). https://www.mixamo.com/
80. Du, R., Chuang, M., Chang, W., Hoppe, H., Varshney, A.: Montage4d: interactive seamless fusion of multiview video textures. In: Proceedings of ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games (I3D), pp. 124–133. ACM (May 2018)
81. Prada, F., Kazhdan, M., Chuang, M., Collet, A., Hoppe, H.: Spatiotemporal atlas parameterization for evolving meshes. ACM Trans. Gr. (TOG) **36**(4), 58 (2017)
82. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. IEEE Trans. Pattern Anal. Mach. Intell. **32**(8), 1362–1376 (2010)