Supplemental Materials: 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
[2] USC Institute for Creative Technologies
[3] Snap Inc.
[4] Pinscreen

1 Network Details

We provide details of our proposed network in Fig. 1. Our network consists of two parts: (1) a weight-sharing fully convolutional neural network for extracting features from a given view and a 3D query point (Fig. 1a), and (2) a classification network (Fig. 1b) that consumes multi-view features from the preceding networks and predicts per-point probabilities of lying inside and outside the reconstructed object.

The feature extraction network aggregates features of multiple scales. The input image is first processed by a convolutional layer and then passed to six down-scaling units. Each unit is composed of one max pooling layer and two convolutional layers, halving the size of feature map and, in the meanwhile, doubling the feature channels. At each level of feature map, we compute the local coordinate of the projection of query point and apply bilinear interpolation to retrieve the feature vector at the projected location. Feature vectors from different channels are concatenated to obtain the signature at each scale. The multi-scale features are further concatenated and then enhanced by a 3-layer MLP network to obtain scale-invariant per-view features. Note that each convolutional layer contains a ReLU activation layer at its output. In addition, batch normalization is applied to all convolutional layers. The parameter settings for each layer can be found in Fig. 1.

The classification network first concatenates feature vectors from all input views and then apply both max and average poolings. The outcome of the two pooling layers are concatenated again and passed to a 2-layer MLP network for inferring the values of P_{in} and P_{out}, which stand for the probability of staying inside and outside the object surface, respectively.

2 Quantitative Results

We quantitatively evaluate the reconstruction errors of visual hull based reconstruction and our approach. We test both methods on the captured multi-view

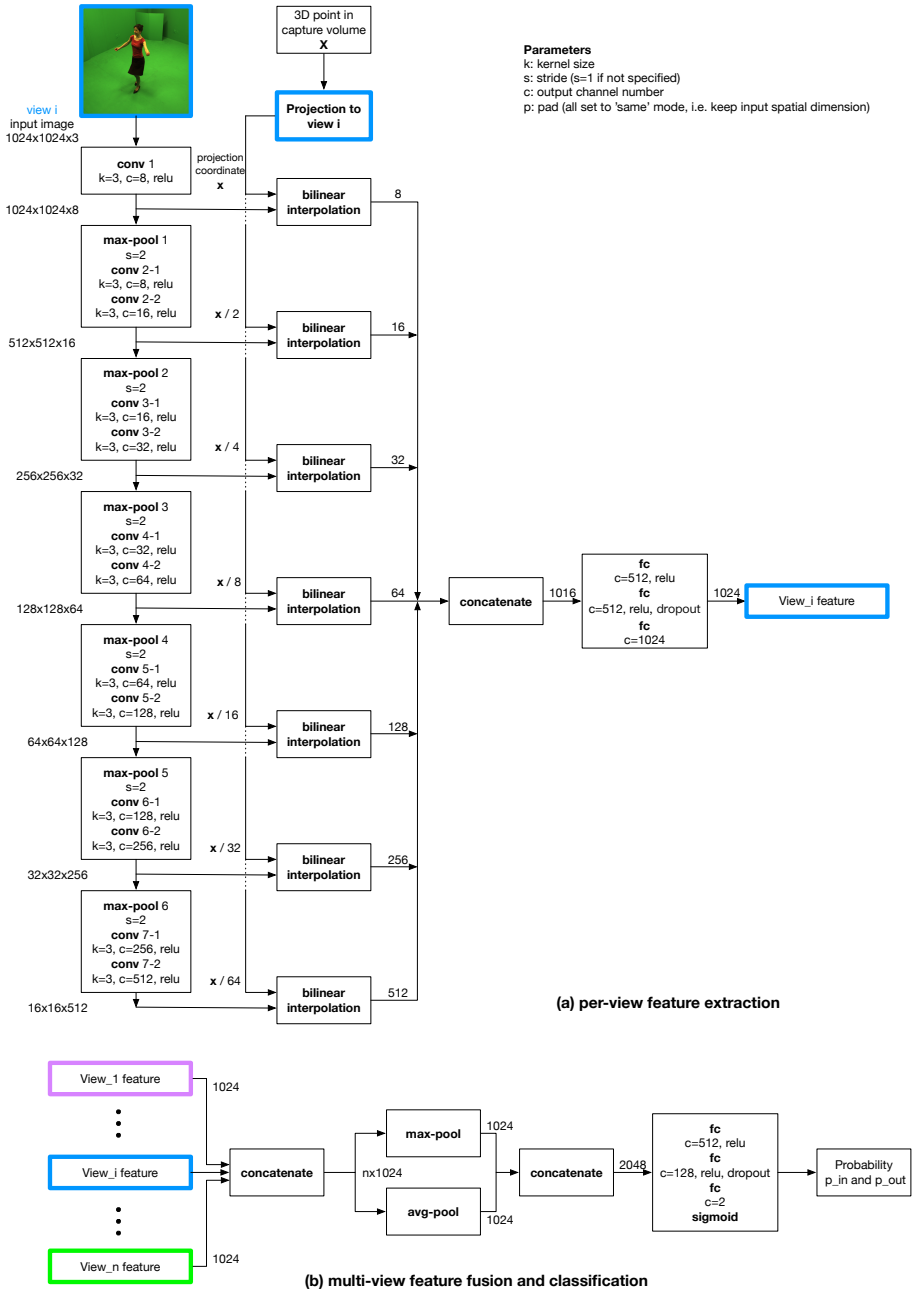

Fig. 1: Detailed Network Architecture

sequences from [1]. For quantitative metric, we first compute the mesh-to-scan (M2S) distance, defined as the median of the Euclidean distances from all vertices on the reconstructed mesh to the closest points on the ground truth scan. We then calculate the reconstruction error by computing the median of the M2S distances of a whole sequence. A lower reconstruction error indicates better reconstruction quality. As shown in Table 1, when only three or four views are used as input, our method outperforms visual hull by a large margin in terms of reconstruction accuracy.

Further, we evaluate the completeness of reconstruction by measuring scan-to-mesh (S2M) distances, defined in a similar way as mesh-to-scan distances, but to change the vertex matching direction. A lower scan-to-mesh distance indicates a more complete reconstruction. As shown in Table 2, when only three or four views are used as input, our method outperforms visual hull in terms of reconstruction completeness for most of the sequences.

Sequence name	Samba	Crane	Bouncing	Jumping	Handstand
Visual hull (3 views)	1.728	2.532	2.535	2.372	2.272
Visual hull (4 views)	0.972	1.244	1.221	1.264	1.257
Ours (3 views)	0.847	0.902	1.106	1.084	1.374
Ours (4 views)	0.565	0.582	0.844	0.742	0.871

Table 1: Quantitative comparison of reconstruction errors between visual hull and our method. Here we show the median distance for each test data sequence, in centimeters.

Sequence name	Samba	Crane	Bouncing	Jumping	Handstand
Visual hull (3 views)	0.877	1.293	1.684	1.224	1.220
Visual hull (4 views)	0.627	0.851	0.904	0.882	0.877
Ours (3 views)	0.735	0.825	0.999	0.993	1.122
Ours (4 views)	0.578	0.646	0.896	0.781	0.892

Table 2: Quantitative comparison of reconstruction completeness between visual hull and our method. Here we show the median distance for each test data sequence, in centimeters.

In Figure 2, we show the cumulative histogram of point-to-scan distances computed over all 5 test sequences, which shows more details about the error distributions. For visual hull based reconstruction, there are 47.3% and 64.2% reconstructed vertices staying within 2cm from the ground truth mesh, given 3 and 4 views respectively. In comparison, our approach provides more accurate

reconstruction as 72.0% and 87.6% vertices of the reconstructed mesh are within 2cm from ground truth surface, when 3 and 4 views are available respectively.

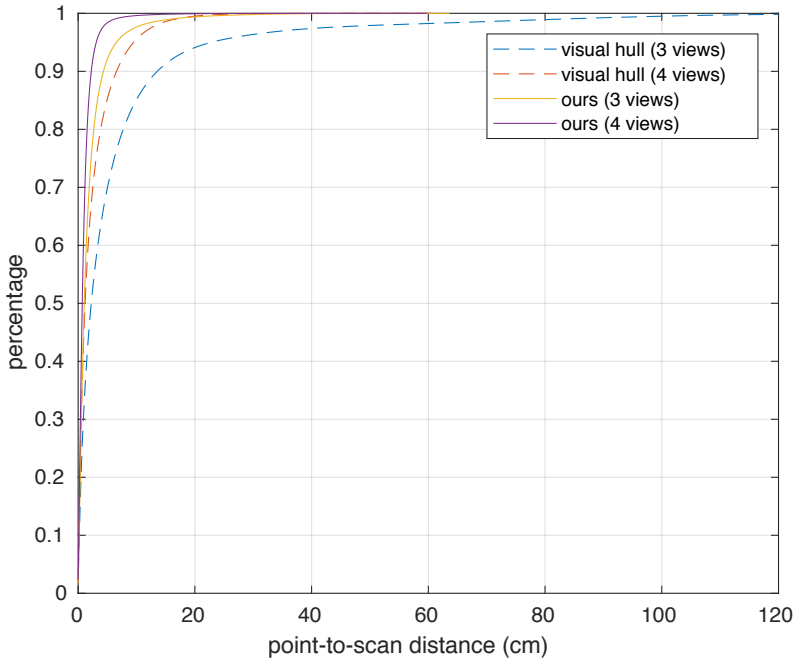

Fig. 2: Cumulative error histogram for point-to-scan distances computed over all 5 test sequences from [1].

3 Video

We provide more results on animation input in the accompanying video [2]. Please refer to the video for the following results:

- Reconstruction results using 4 views from real-world data;
- Reconstruction results using 4 views from synthetic data;
- Comparison of reconstructions using different number of views;
- Comparison of reconstructions against visual hull and Vlasic et al.[1];
- Comparison of reconstructions against visual hull and Starck et al.[3];
- Results reconstructed from novel viewpoints that not appeared in training;
- Examples of synthetic training data.

References

1. Vlasic, D., Baran, I., Matusik, W., Popović, J.: Articulated mesh animation from multi-view silhouettes. In: ACM Transactions on Graphics (TOG). Volume 27., ACM (2008) 97
2. Huang, Z.: Supplemental video: Deep volumetric video from very sparsemulti-view performance capture (2018) https://youtu.be/xjTVECIqZfc.
3. Starck, J., Hilton, A.: Surface capture for performance-based animation. IEEE computer graphics and applications **27**(3) (2007)