Learning to Optimize Non-Rigid Tracking
Supplementary Material

Nonrigid Dataset Fig. 1 shows the examples in the collected non-rigid dataset. Similar to the ScanNet, the RGB-D videos are captured via iPad mounted Structure Sensor. The depth frames are recorded at a resolution of 640×480. A variety of dynamic objects are included: animals, adults, children, cloth, furniture, *etc*. The backgrounds vary from static scenes to dynamic ones with multiple moving objects. Refer to [1] for more details.

Tracking Results Fig. 2 and Fig. 3 show the frame-frame non-rigid tracking results on different ranges of motion, from small to large. We compare our method with two baselines, N-ICP-1 [3] and N-ICP-2 [4]. Though all three methods show artifacts when the motion is large, the results of our method are geometrically closer to the target frames.

Network Config Table. 1 shows the network structure of the non-rigid feature extractor. It is based on the fully convolutional networks [2]. The RGB-D images are resized to 128×96 before feeding to the feature extractor. All convolutions are followed by a batch normalization layer and a ReLU layer. We use spatial average pooling of size 2 to down-sample features between two feature pyramids. The input volume is [128, 96, 4], where 4 represents the 4 channels of the RGB-D frame. Down-sampling is applied three times in total. The output volume of the network is [16, 12, 5]. Table 2 shows the network structure of the ConditionNet. The number of input and output channels for ConditionNet-Dense and ConditionNet-Sparse is 1. ConditionNet-Diagonal has 36 channels for input and 21 channels for output.

References

[1] Aljaž Božič, Michael Zollhöfer, Christian Theobalt, and Matthias Nießner. Deepdeform: Learning non-rigid rgb-d reconstruction with semi-supervised data. In *CVPR*, 2020. 1

[2] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[3] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, pages 343–352, 2015. 1, 3, 4

[4] Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rehmann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, Christian Theobalt, et al. Real-time non-rigid reconstruction using an rgb-d camera. *ACM Transactions on Graphics (ToG)*, 33(4):156, 2014. 1, 3, 4

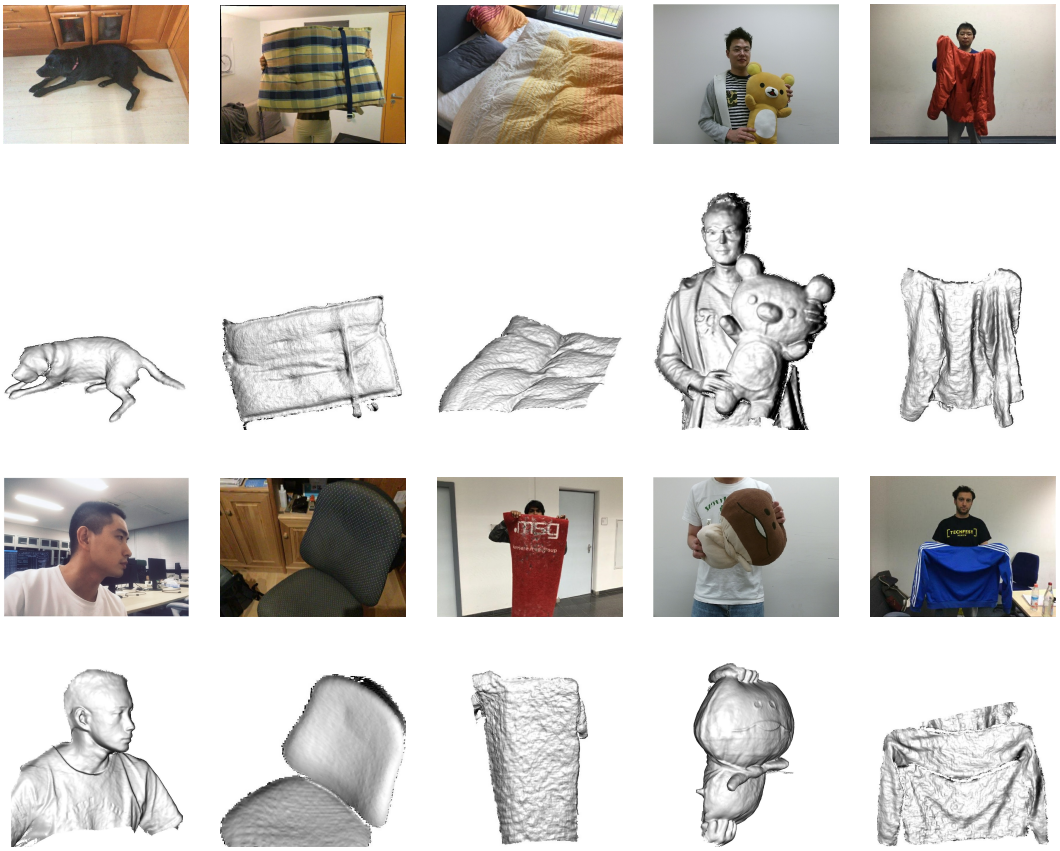


Figure 1. Examples of sequence in the non-rigid dataset. **Top row:** Color frame in the sequence. **Bottom row:** The reconstructed foreground object models using our non-rigid tracking and reconstruction method.

Figure 2. Non-Rigid tracking results on the ***kitty*** scene.

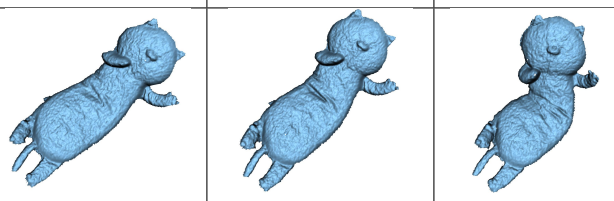
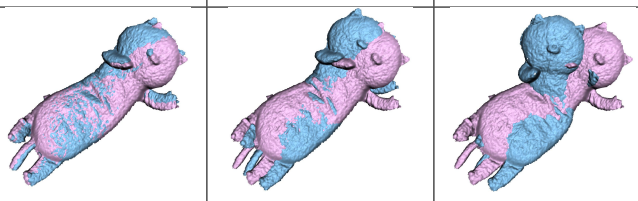
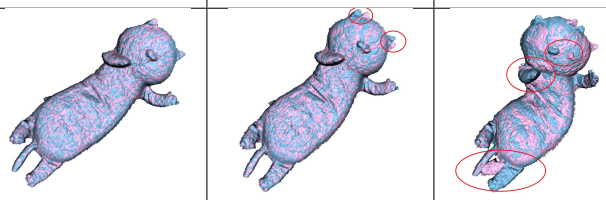
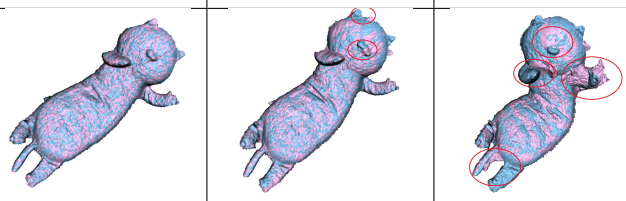
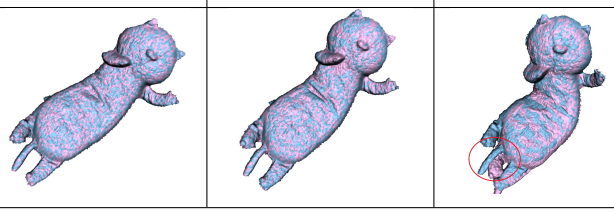
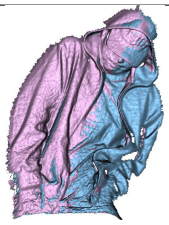
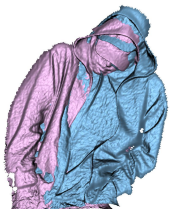
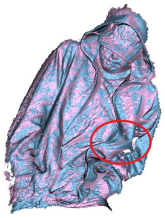
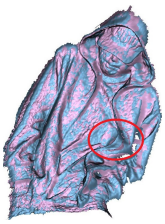
Motion Size (Source → target)	Small (0 → 3)	Medium (0 → 10)	Large (0 → 19)
Target color image			
Target Mesh			
Initial Alignment			
N-ICP-1[3]			
N-ICP-2 [4]			
Ours			

Figure 3. Non-Rigid tracking results on the ***hoody*** scene.

Motion Size (Source → Target)	Small (0 → 2)	Medium (0 → 5)	Large (0 → 12)
Target color image			
Initial Alignment			
N-ICP-1[3]			
N-ICP-2 [4]			
Ours			

Module Name	Channels	Stride	Batch Norm.	Activation	Dilation
Conv2d	16	1	True	ReLU	1
Conv2d	32	1	True	ReLU	2
Conv2d	32	1	True	ReLU	2
Average Polling					
Conv2d	32	1	True	ReLU	1
Conv2d	64	1	True	ReLU	2
Conv2d	64	1	True	ReLU	2
Average Polling					
Conv2d	64	1	True	ReLU	1
Conv2d	96	1	True	ReLU	2
Conv2d	96	1	True	ReLU	2
Average Polling					
Conv2d	96	1	True	ReLU	1
Conv2d	128	1	True	ReLU	2
Conv2d	128	1	True	ReLU	2
Conv2d	5	1	True	ReLU	1

Table 1. The network configuration of the non-rigid extractor.

Module Name	Channels	Stride	Batch Norm.	Activation	padding
Input	n_{in}				
Conv2D-1.1	32	1	true	PReLU	1
Conv2D-1.2	32	1	true	PReLU	1
Max pooling					
Conv2D-2.1	64	1	true	PReLU	1
Conv2D-2.2	64	1	true	PReLU	1
Max pooling					
Conv2D-3.1	128	1	true	PReLU	1
Conv2D-3.2	128	1	true	PReLU	1
Max pooling					
Conv2D-4.1	128	1	true	PReLU	1
Conv2D-4.2	128	1	true	PReLU	1
NearestUpSample1					
Concat (w/ Conv2d-3.2)	128+128	1	true	PReLU	1
Conv2D	128	1	true	PReLU	1
Conv2D	128	1	true	PReLU	1
NearestUpSample2					
Concat (w/ Conv2d-2.2)	128+64	1	true	PReLU	1
Conv2D	64	1	true	PReLU	1
Conv2D	64	1	true	PReLU	1
NearestUpSample3					
Concat (w/ Conv2d-1.2)	64+32	1		PReLU	1
Conv2D	32	1	true	PReLU	1
Conv2D	n_{out}	1	true	PReLU	1

Table 2. The network configuration of ConditionNet. $n_{in} = 1, n_{out} = 1$ for ConditionNet-Dense and ConditionNet-Sparse. $n_{in} = 36, n_{out} = 21$ for ConditionNet-Diagonal. The kernel size is 2x2 for ConditionNet-Diagonal and alternate between 1x1 and 2x2 for ConditionNet-Dense and ConditionNet-Sparse.