Implementation Details of Our Network

Component	Name	Input	Channels	Setting
Location Encoder	Conv1	$\widehat{H}_{t-32}$, $\widetilde{H}_{t-24}$, ..., $\widetilde{H}_t$, I_{t-32}, ..., I_t	128	
	Conv2	Conv1	128	D=2
	Conv3	Conv2	128	S=2
	Side1	Conv3	128	
	Conv4	Conv3	128	D=4
	Conv5	Conv4	128	D=2
	Conv6	Conv5	256	S=2
	Side2	Conv6	512	
	Conv7	Conv6	256	D=8
	Conv8	Conv7	256	D=4
	Conv9	Conv8	256	S=2
	Side3	Conv9	512	
Decoder	Skip1	f_{out}^2	256	K=1
	Conv1	Skip1+f_{out}^3	256	
	Conv2	Conv1	128	
	Conv3	Conv2	64	
	Skip2	f_{out}^1	64	K=1
	Conv4	Skip2+Conv3	128	
	Conv5	Conv4	64	
	Conv6	Conv5	64	
	Conv7	f_{out}^0+Conv6	64	
	Conv8	Conv7	64	
	Conv9	Conv8	4	

Supplementary Table1. Detailed configuration of our model. We list the convolution layers in the location encoder and decoder with their input data, output channels and special settings respectively. The convolution kernel is 3×3 with $stride = 1$ and $dilation = 1$ by default. If there is a different setting, we point out the modified kernel size (K), dilation (D) or stride (S). In the location encoder and decoder, each convolution layer except the output layers (including the side output) is followed by batch normalization and ReLU activation.

More Details of Making ScanDPT

In fact, cameras move very fast in the videos of ScanNet as they are handheld. Nevertheless, for our task, only when cameras move very slow, viewers can see the inserted content clearly. To alleviate this problem, we use video frame interpolation algorithms to slow down the motion speed to 1/4 of the original one. The locations of the ground truth designated points in the interpolated middle frames are approximated by the mean value of the ground truth of the original neighbor frames.

Implementation Details of Other Methods

For the best homography estimation method SuperPoint+SuperGlue, the author

gave two pretrained model including outdoor models and indoor models. We adopted the indoor models pretrained on ScanNet which was the basis of our dataset ScanDPT. Consequently, we did not need to finetune the models of SuperPoint+SuperGlue on ScanDPT. For the best optical flow method RAFT, we chose the model pretrained on synthetic datasets which also has the best performance among all the pretrained models given by the authors. We did not finetune the RAFT model on the ScanDPT for two reasons. Firstly, RAFT needs very precise motion vectors for all the image pixels between adjacent frames. However, the depthmaps from RGB-D sensors are noisy, and thus ScanDPT based on RGB-D videos ScanNet is not appropriate for finetuning optical flow models. In other words, the finetuned result by the supervision of noisy ground truth could be worse. Secondly, due to the video frame interpolation procedure as described in the details of making ScanDPT, we are unable to acquire precise depthmaps for the interpolated frames.

For each other learning-based method, we also report the best performance among all the pretrained models given by the authors.