Supplemental

A. Implementation Details

We adopt the original parameters used in DORN [12] for NYUv2 depth estimation, except that we output 13 dimensions at the final layer and use our loss to train the network. During our experiments, the input and output image resolution is 320x240. We use the fixed learning rate as 1e-5.

B. Additional Baselines

Necessity to predict both tangent directions and the surface normal We did an experiment on Scannet to test how the results change when only one tangent direction is predicted. This method yields mean angle errors of $16.34°$ for normals (versus $15.28°$ with ours) and $12.41°$ for tangents (versus $12.26°$ with ours). Although a normal and one tangent can define a coordinate frame, it is better to predict the second tangent directly rather than deriving it from the normal. This result corroborates the main thesis of the paper – predicting tangents helps predicting 3D coordinate frames.

Geometry prediction followed with canonical frames computation We did a baseline experiment where we predict depth and normals using the DORN architecture and then use the resulting surface reconstruction to compute the canonical 3D tangent frame with Quadriflow. The mean angle error of the tangent directions on the ScanNet test set is $35.84°$, while ours is $12.26°$. The difference is not surprising since our method is trained with the ground truth tangent supervision.

C. Surface Normal Estimation

Compare with the state-of-the-art We compare the performance of the surface normal estimation from our approach with the state-of-the-art methods on SunCG [38]. We use our approach to train four networks and evaluate them. Table 8 shows the results including UNet [31], SkipNet [3], GeoNet [28] and DORN [12]. With the assistance of the projected tangent principal directions, the normal prediction has been improved. Please refer to the experiments for ScanNet in the main paper in section 5.1.

Test on NYUv2 We test different versions of our network on NYUv2 [11] as a standard evaluation dataset. We train the network on SunCG datasets and directly test on NYUv2, as shown in Table 9. Specifically, GeoNet-origin trained and tested on NYUv2 [28], and is the current state-of-the-art method on normal estimation. Other rows are networks trained w/o. our joint losses on SunCG.

The joint loss in the training process results in better normal estimation. From the ScanNet experiment in section 5.1 of the main section, ScanNet gets better performance compared to SunCG, possibly due to the domain gap between synthetic (SunCG) and real (NYUv2).

SunCG	mean	median	rmse	11.25°	22.5°	30°
UNet	14.88	6.20	24.94	64.4	78.6	83.9
UNet-Ours	13.25	4.64	23.73	69.8	81.6	86.1
SkipNet	13.38	3.97	24.54	70.2	80.3	85.1
SkipNet-Ours	12.82	3.87	23.69	71.0	80.2	86.1
GeoNet	13.14	3.56	23.54	70.6	80.7	86.0
GeoNet-Ours	12.68	3.60	**22.73**	71.2	81.3	**86.6**
DORN	12.90	3.36	24.12	71.3	81.3	85.3
DORN-Ours	**12.38**	**3.33**	23.34	**72.3**	**82.3**	86.3

Table 8. Evaluation on Surface Normal Predictions. We train and test our algorithm with different network architectures on the SunCG [9]. Assisted by our joint loss, the performances of all networks are improved.

NYUv2	mean	median	rmse	11.25°	22.5°	30°
GeoNet-origin	**19.0**	**11.8**	**26.9**	**48.4**	**71.5**	**79.5**
SunCG	mean	median	rmse	11.25°	22.5°	30°
UNet	25.21	18.26	32.82	32.2	57.7	68.3
UNet-Ours	24.64	17.10	32.65	35.0	59.6	69.5
SkipNet	24.75	17.36	32.45	33.8	58.1	69.0
SkipNet-Ours	23.67	16.28	31.72	36.1	62.2	72.7
GeoNet	22.32	14.97	30.59	39.8	64.3	73.4
GeoNet-Ours	22.15	14.41	30.18	40.1	65.3	74.4
DORN	22.19	14.46	30.16	40.3	65.3	74.1
DORN-Ours	21.99	14.29	29.87	40.5	65.8	74.6

Table 9. Normal prediction on NYUv2 [11]. GeoNet-origin trained and tested on NYUv2 [28]. In other rows, we train network w/o. our joint loss on SunCG and tested on NYUv2. DORN-Ours trained on ScanNet performs best among all.

D. Visualization

Surface Normal Comparison Figure 13 visualizes the normal prediction using the best model w/o. our approach on both the datasets. With our approach, the errors are smaller especially at object boundaries, possibly because of the additional supervision given by the projected tangent principal directions. We show more accurate prediction, especially for small objects.

Visualize the tangent principal directions We show more visualization for the projected tangent principal directions in figure 14. The model is trained using the Dorn [12] with our joint loss on ScanNet [9]. The visualization shows a similar direction field compared to the ground truth and is consistent with human intuition.

Visualize the feature matching We show more visualization for comparison between SIFT and SIFT with our

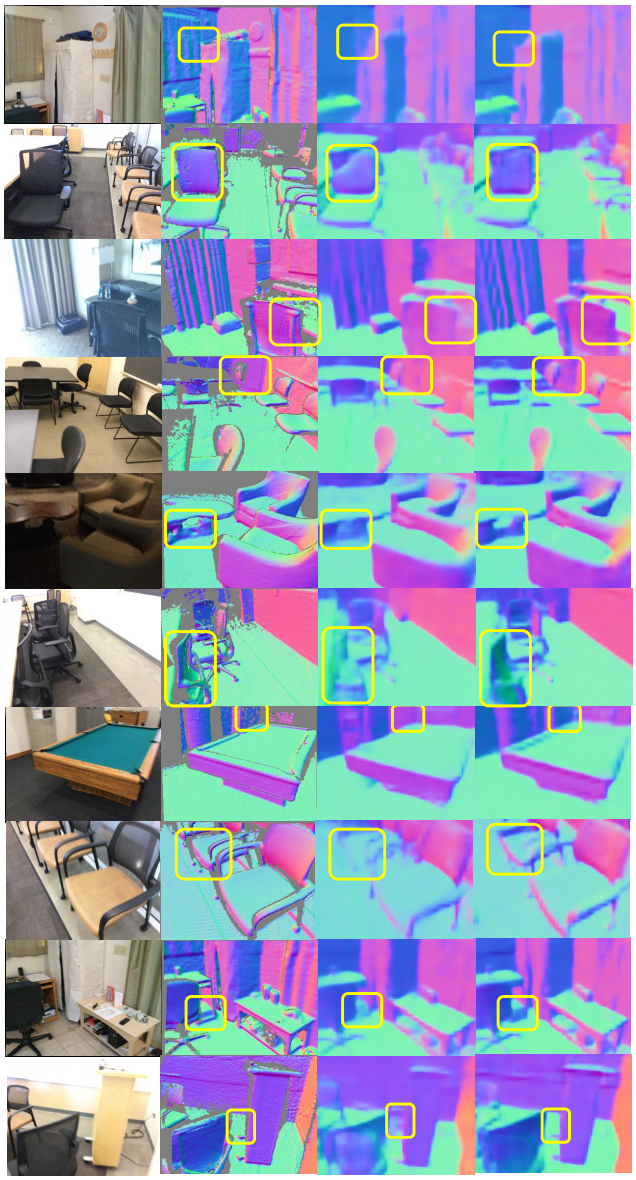

Figure 13. Visual comparison of the results. With our joint loss, the predicted surface normals produce less errors and more details. We show more accurate prediction especially for small objects.

perspective rectification on the DTU [1] in figure 15. We produce more correct matching than SIFT does.

Visualize the augmented reality results We show more examples of new elements insertion into the scene in figure 16. The perspectives are locally consistent with the canonical frames of the geometry.

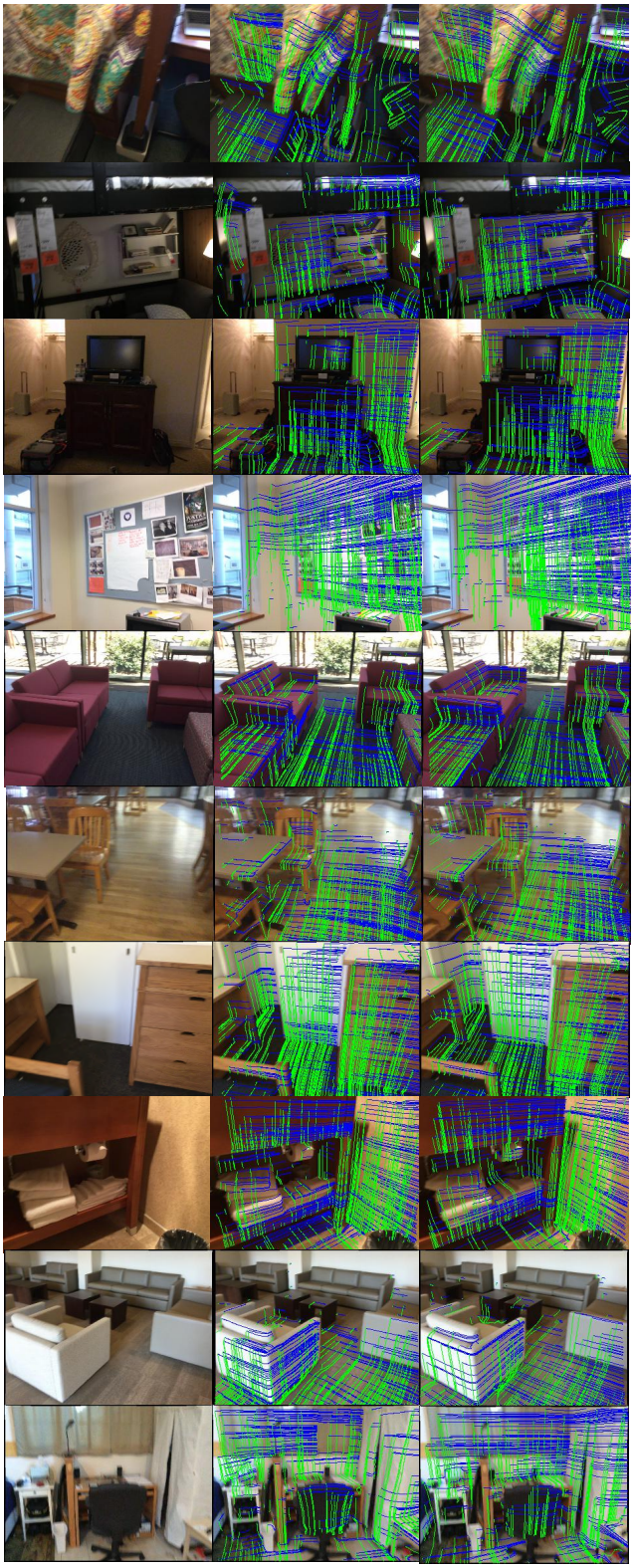

Figure 14. Visualization of the projected tangent principal directions. The visualization shows similar direction field compared to the ground truth, and is consistent with human intuition.

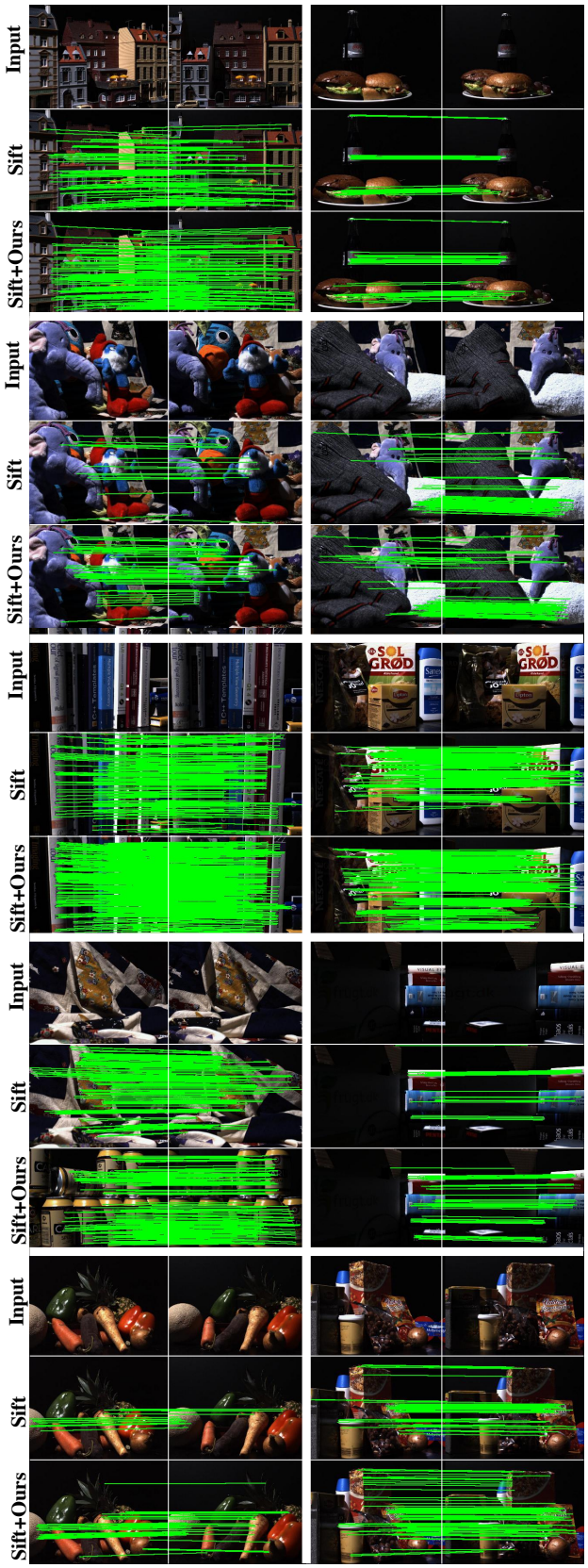

Figure 15. Visualization of the feature matching using SIFT and SIFT with our perspective rectification. We produce more correct matching than SIFT does.

Figure 16. Visualization of augmented reality results. We attach images in a rigid or deformable way (highlighted with yellow square), or 3D objects into the scenes. The perspectives are locally consistent with the canonical frames.