Supplementary Material for PlaneFormers

Samir Agarwala, Linyi Jin, Chris Rockwell, and David F. Fouhey

University of Michigan, Ann Arbor
{samirag,jinlinyi,cnris,fouhey}@umich.edu

The supplemental material consists of both video results and this PDF. The PDF portion of the supplemental material shows: detailed descriptions of model architectures (Section 1), details about the experimental setup (Section 2), and additional results (Section 3). Video results visualize qualitative reconstructions from the main paper from a variety of viewpoints.

1 Model Architecture

Table 1. Model Architecture. We define the number of planes from view i be M_i, $M = M_1 + M_2$, dimension $D = 899$. Embeddings are passed through a 5-layer transformer encoder which has 1 head, dropout probability of 0.1 and a feedforward network dimension of 2048. We create a pair-wise feature tensor of dimension $M \times M \times 4D$ and pass this tensor through 4 separate MLP heads to estimate the plane correspondence, camera correspondence, rotation residual and translation residual. We mask out entries in the MLP outputs such that only the pairwise predictions between planes across views are considered during average pooling in the camera correspondence and residual heads (note: the shape after masking in the table represents non-zero entries). Finally, we apply a sigmoid function to the plane correspondence and the camera correspondence scores, and extract plane correspondences across views.

Index	Inputs	Operation	Output Shape
(1)	Inputs	Input Embedding	$M \times D$
(2)	(1)	5-Layer Transformer Encoder	$M \times D$
(3)	(2)	Create Pair-wise Feature Tensor	$M \times M \times 4D$
(4)	(3)	Plane Correspondence: Linear($4D \to 2D$), Linear($2D \to D$), Linear($D \to D/2$), Linear($D/2 \to D/4$), Linear($D/4 \to 1$), Sigmoid($M \times M$), Extract Submatrix($M \times M \to M_1 \times M_2$)	$M_1 \times M_2$
(5)	(3)	Camera Correspondence: Linear($4D \to 2D$), Linear($2D \to D$), Linear($D \to D/2$), Linear($D/2 \to D/4$), Linear($D/4 \to 1$), Mask Matrix($M \times M \to M_1 \times M_2$), AveragePool($M_1 \times M_2 \to 1$), Sigmoid(1)	1
(6)	(3)	Rotation Residual: Linear($4D \to 2D$), Linear($2D \to D$), Linear($D \to D/2$), Linear($D/2 \to D/4$), Linear($D/4 \to 4$), Mask Matrix($M \times M \times 4 \to M_1 \times M_2 \times 4$), AveragePool($M_1 \times M_2 \times 4 \to 4$)	4
(7)	(3)	Translation Residual: Linear($4D \to 2D$), Linear($2D \to D$), Linear($D \to D/2$), Linear($D/2 \to D/4$), Linear($D/4 \to 3$), Mask Matrix($M \times M \times 3 \to M_1 \times M_2 \times 3$), AveragePool($M_1 \times M_2 \times 3 \to 3$)	3

2 Experimental Details

2.1 Multiview Dataset Creation

The two view dataset is the same as [1]. For 3-view and 5-view datasets, we use the single images sampled by [1], then randomly sample combinations of images within each floor of the house. We select sets of images where each image in any pair has ≥ 3 matches and ≥ 3 unique planes. The maximum number of sets per floor is 10. We finally get a three-view test set of size 258 and a five-view test set of size 76. We do not need training set or validation set for 3-view and 5-view cases since our network is not trained on the multiview dataset.

2.2 Multiview Evaluation

We compared our proposed approach with baselines on the same view graph that was built as discussed in the approach section. For multi-view evaluation, we consider all combinations of input views in a sample and independently compute the pair-wise IPAA, rotation error and translation error for each combination. For instance, in the 3-view case, we consider the IPAA and camera error metrics independently between view 1 and view 2, view 1 and view 3, and view 2 and view 3.

The relative camera transformations and plane correspondences between any combination of views is computed by chaining together relative camera transformations and plane correspondences across the created view graph, and is then compared to the ground-truth. Finally, we compute IPAA-X and camera error statistics over all combinations of views across samples in the test set (i.e. we consider each combination of views in a test sample as an independent datapoint while computing our metrics).

2.3 Ablation Details

For fair comparison, we use our same training setup for ablations. Ablations are trained until validation accuracy plateaus, which in practice is 40k iterations; the same as for the full model.

Feature Ablations. Ablating features results in a smaller input feature space. For fair comparison, we therefore use a linear layer to project this smaller input feature to features the same size as the full model. Transformer layers then operate at the same size as in the case of the full model.

Model Ablations. In our full model, we classify camera pose into clusters from [1], and predict a corrective residual camera pose to the predicted cluster. In the *without residual* ablation, we remove this residual camera pose; this tests if the corrective residual improves predicted pose. Our *without transformer* ablation takes input features as direct input to final camera and plane MLP layers, testing if the transformer improves plane features for final prediction.

3 Additional Results

3.1 Quantitative Results

Additional Plane Correspondence Results We report IPAA-100 and IPAA-80 for the feature and network ablations in addition to IPAA-90 that was provided in our experiments in Tables 2 and 3, respectively.

Table 2. IPAA-100, IPAA-90 and IPAA-80 for feature ablations.

Feature	IPAA-100 ↑	IPAA-90 ↑	IPAA-80 ↑
Proposed	**19.6**	**40.6**	**71.0**
- Appearance	11.3	26.9	57.1
- Plane	16.5	35.2	65.6
- Mask	15.1	34.5	67.2

Table 3. IPAA-100, IPAA-90 and IPAA-80 for network ablations. The IPAA-X results for the model without residual are the same as the proposed method since the camera residual affects the relative camera pose prediction but not the plane correspondences.

Network	IPAA-100 ↑	IPAA-90 ↑	IPAA-80 ↑
Proposed	**19.6**	**40.6**	**71.0**
- Transformer	13.8	32.7	64.3
- Residual	19.6	40.6	71.0

Additional Relative Camera Pose Estimation Results We report median error and % error $\leq$ 1m or 30° for the predicted relative translation and rotation for the feature and network ablations in addition to mean error that was provided in our experiments in Tables 4 and 5 respectively.

3.2 Qualitative Results

Video Results. Video results bring to life the reconstructions from the main paper. As stated, PlaneFormer planar reconstructions are often quite close to the ground truth even faced with large view changes and challenging coplanar settings. If 3 or 5 views are available, the model continues to produce coherent results.

Additional Examples. We include additional outputs for 2 input views (Figure 2), and 3 and 5 views (Figure 3). These results are consistent with those in the

Table 4. Median error, mean error and % error $\leq$ 1m or 30° for translation and rotation for the feature ablations.

Method	Translation			Rotation		
	Med.	Mean	(≤1m)	Med.	Mean	(≤30°)
Proposed	**0.66**	**1.19**	**66.8**	5.96	22.20	83.8
- Appearance	0.69	1.23	65.2	6.01	22.78	83.3
- Plane	0.81	1.32	59.6	10.34	25.92	81.4
- Mask	0.75	1.26	61.9	**4.65**	**21.21**	**84.3**

Table 5. Median error, mean error and % error $\leq$ 1m or 30° for translation and rotation for the network ablations.

Method	Translation			Rotation		
	Med.	Mean	(≤1m)	Med.	Mean	(≤30°)
Proposed	**0.66**	**1.19**	**66.8**	**5.96**	**22.20**	**83.8**
- Transformer	1.02	1.48	49	10.54	26.43	80.8
- Residual	0.88	1.34	57.7	6.22	22.38	83.7

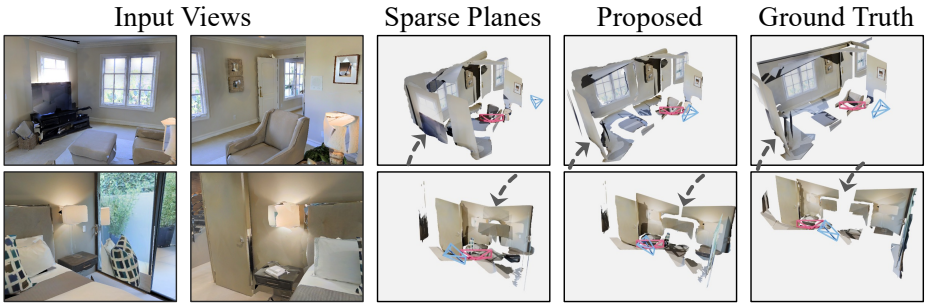

Fig. 1. Additional reconstruction comparison, extending Fig. 4 Sparse Plane reconstructions are a good baseline, but PlaneFormer yields superior results. It produces both better stitched planes (top), and more accurate camera (bottom).

paper: plane correspondences tend to be accurate even in challenging cases, and reconstructions are reasonable in very large view change cases and accurate in smaller view change cases.

Additional reconstruction comparison, extending Fig. 4. See Figure 1.

Limitations and Failure Cases. We also include limitations and failure cases in Figure 4. One limitation of a plane representation is that planes struggle to model small details in scenes, which sometimes leads to incomplete reconstructions (top two examples). The model may also perform poorly in some circumstances. Plane correspondences struggle when many small, similar objects are visible across large view change (second two examples). Predicting camera can sometimes be difficult given large view change leading to significant difference

Fig. 2. Additional 2 View Results.

in appearance (bottom two examples in two-view case). Sometimes both camera and correspondence are poor (bottom example, two-view case). When more views are present, planes are not always fused cleanly, leading to intersections (final two examples, multi-view case).

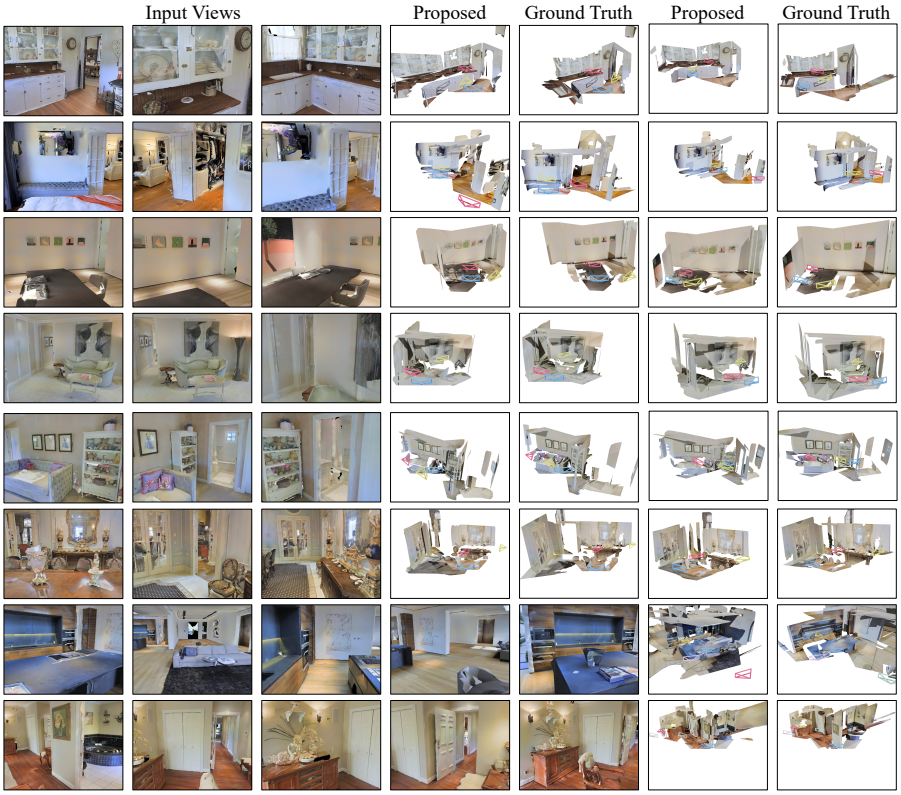

Fig. 3. Additional 3 and 5 View Results.

Fig. 4. Limitations and Failure Cases.

References

1. Jin, L., Qian, S., Owens, A., Fouhey, D.F.: Planar surface reconstruction from sparse views. In: ICCV (2021) 2