Supplementary Material for
Deep Fusion Transformer Network with Weighted Vector-Wise Keypoints Voting for Robust 6D Object Pose Estimation

A. Overview

In this part, we give more information about our method. We first describe the details of the prediction headers (Sec. B). and then provide more experimental results to validate the effectiveness of our method (Sec. C).

B. Details of the Prediction Headers Architecture

Given the dense fused RGB-D features from the deep fusion transformer network, three prediction headers are followed to perform semantic segmentation, vector field, and vector-wise confidence prediction. These three headers consist of shared MLPs. Specifically, the output dimension of each layer in these headers is as follows: 1) the semantic segmentation module: $MLP[\mathcal{F}_{dense}, 1024, 512, 128, N_{classes}]$, 2) the vector field prediction module: $MLP[\mathcal{F}_{dense}, 1024, 512, 128, N_{kps} * 3]$, and 3) the vector-wise confidence prediction module: $MLP[\mathcal{F}_{dense}, 1024, 512, 128, N_{kps} * 1]$, where $\mathcal{F}_{dense}$ denotes the dimension of output dense fused RGB-D features, $N_{classes}$ denotes the number of object classes and N_{kps} denotes the number of keypoints of each object.

C. More Results

C.1. Qualitative Results on the MP6D Dataset

We report more qualitative comparison results between our method and the SOTA RGB-D fusion-based method FFB6D [19] in Fig. 7. Our approach is more robust to these challenging scenarios.

C.2. Qualitative and Quantitative Results on the YCB-Video Dataset

The object-wise experimental quantitative results on YCB-Video dataset are reported in Table 8. We also give qualitative comparison results between our method and FFB6D [19] in Fig. 8.

C.3. Qualitative Results on the Occlusion-LineMOD Dataset

We provide qualitative comparison results between our method and the ground truth on the Occlusion-LineMOD dataset in Fig. 9.

Table 8. Quantitative evaluation results (ADD-S [61] and ADD(S) [22] AUC) without iterative refinement on the YCB-Video Dataset. Symmetric objects are in bold.

	PoseCNN [61]		PointFusion [63]		DCF [40]		DF (per-pixel) [58]		PVN3D [20]		FFB6D [19]		Ours	
Object	ADDS	ADD(S)	ADDS	ADD(S)	ADDs	ADD(S)	ADDS	ADD(S)	ADDS	ADD(S)	ADDS	ADD(S)	ADDS	ADD(S)
002 master chef can	83.9	50.2	90.9	-	90.9	74.6	95.3	70.7	96.0	80.5	96.3	80.6	**97.0**	**92.3**
003 cracker box	76.9	53.1	80.5	-	87.1	79.3	92.5	86.9	96.1	**94.8**	**96.3**	94.6	95.9	93.9
004 sugar box	84.2	68.4	90.4	-	94.3	84.2	95.1	90.8	97.4	96.3	**97.6**	**96.6**	97.1	95.5
005 tomato soup can	81.0	66.2	91.9	-	90.5	79.8	93.8	84.7	**96.2**	88.5	95.6	89.6	95.6	**92.6**
006 mustard bottle	90.4	81.0	88.5	-	90.6	83.5	95.8	90.9	97.5	96.2	**97.8**	**97.0**	97.6	96.3
007 tuna fish can	88.0	70.7	93.8	-	91.7	73.8	95.7	79.6	96.0	89.3	96.8	88.9	**97.3**	**94.5**
008 pudding box	79.1	62.7	87.5	-	89.3	84.1	94.3	89.3	97.1	**95.7**	97.1	94.6	**97.4**	**95.7**
009 gelatin box	87.2	75.2	95.0	-	92.9	89.5	97.2	95.8	97.7	96.1	**98.1**	**96.9**	97.6	96.3
010 potted meat can	78.5	59.5	86.4	-	83.2	74.6	89.3	79.6	93.3	88.6	94.7	88.1	**95.9**	**92.1**
011 banana	86.0	72.3	84.7	-	84.8	71.0	90.0	76.7	96.6	93.7	**97.2**	94.9	97.1	**95.0**
019 pitcher base	77.0	53.3	85.5	-	89.5	80.3	93.6	87.1	97.4	96.5	**97.6**	**96.9**	96.0	93.1
021 bleach cleanser	71.6	50.3	81.0	-	88.4	79.8	94.4	87.5	96.0	93.2	**96.8**	94.8	**96.8**	**94.9**
024 bowl	69.6	69.6	75.7	75.7	80.3	80.3	86.0	86.0	90.2	90.2	96.3	96.3	**96.9**	**96.9**
025 mug	78.2	58.5	94.2	-	90.7	76.6	95.3	83.8	**97.6**	**95.4**	97.3	94.2	**97.6**	94.9
035 power drill	72.7	55.3	71.5	-	87.4	78.4	92.1	83.7	96.7	95.1	**97.2**	**95.9**	96.9	95.2
036 wood block	64.3	64.3	68.1	68.1	84.2	84.2	89.5	89.5	90.4	90.4	92.6	92.6	**96.2**	**96.2**
037 scissors	56.9	35.8	76.7	-	84.2	70.3	90.1	77.4	96.7	92.7	**97.7**	**95.7**	97.2	93.3
040 large marker	71.7	58.3	87.9	-	89.5	81.0	95.1	89.1	96.7	91.8	96.6	89.1	**96.9**	**92.7**
051 large clamp	50.2	50.2	65.9	65.9	63.6	63.6	71.5	71.5	93.6	93.6	**96.8**	**96.8**	96.3	96.3
052 extra large clamp	44.1	44.1	60.4	60.4	64.4	64.4	70.2	70.2	88.4	88.4	96.0	96.0	**96.4**	**96.4**
061 foam brick	88.0	88.0	91.8	91.8	83.1	83.1	92.2	92.2	96.8	96.8	**97.3**	**97.3**	**97.3**	**97.3**
ALL	75.8	59.9	83.9	-	85.7	77.9	91.2	82.9	95.5	91.8	96.6	92.7	**96.7**	**94.4**

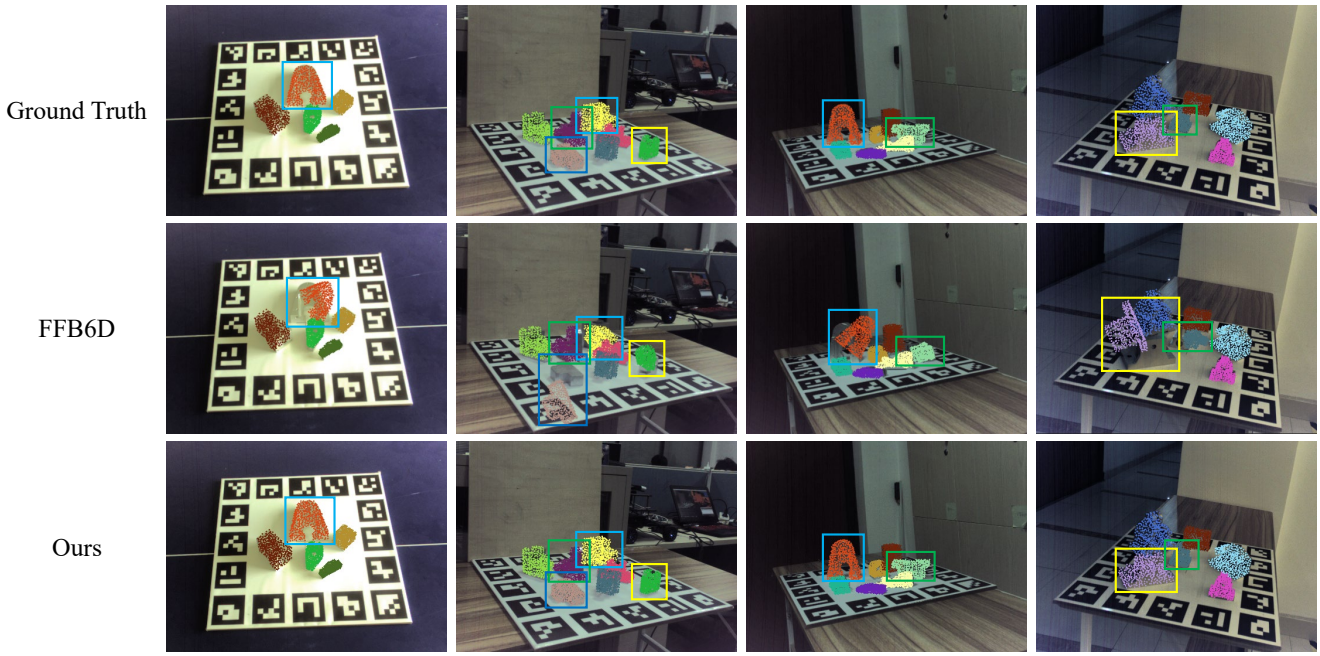

Figure 7. Qualitative comparison results on MP6D dataset.

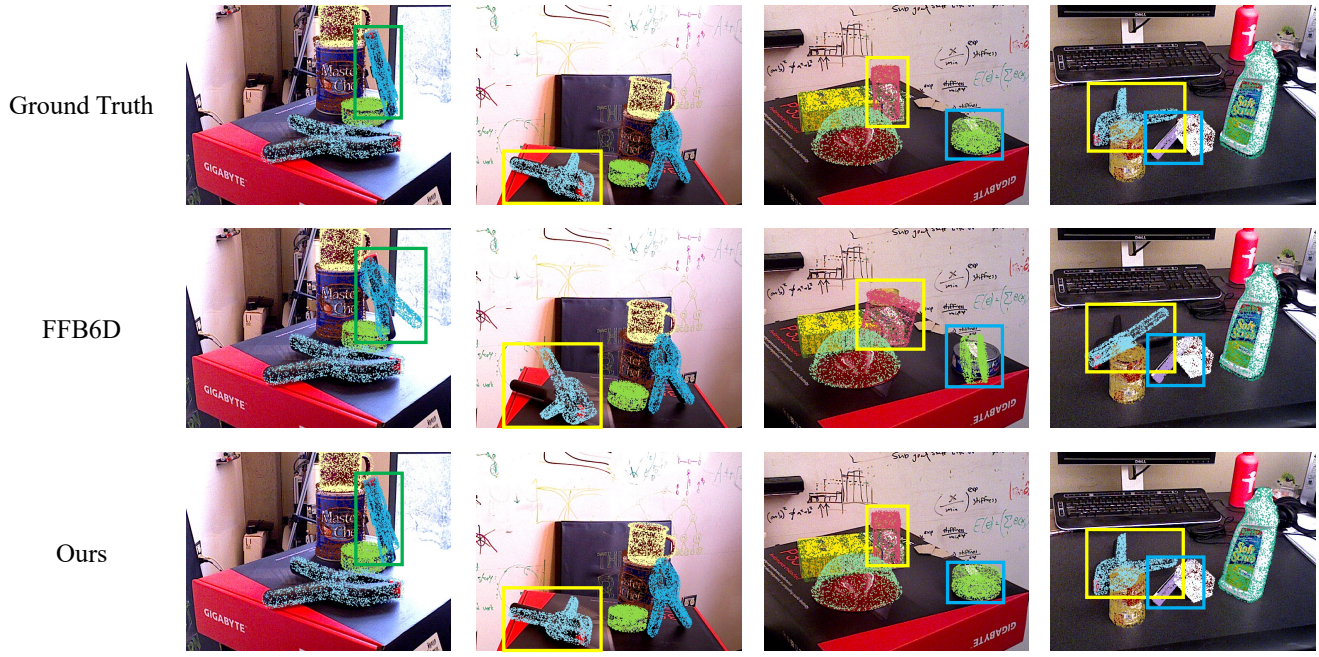

Figure 8. Qualitative comparison results on YCB-Video dataset.

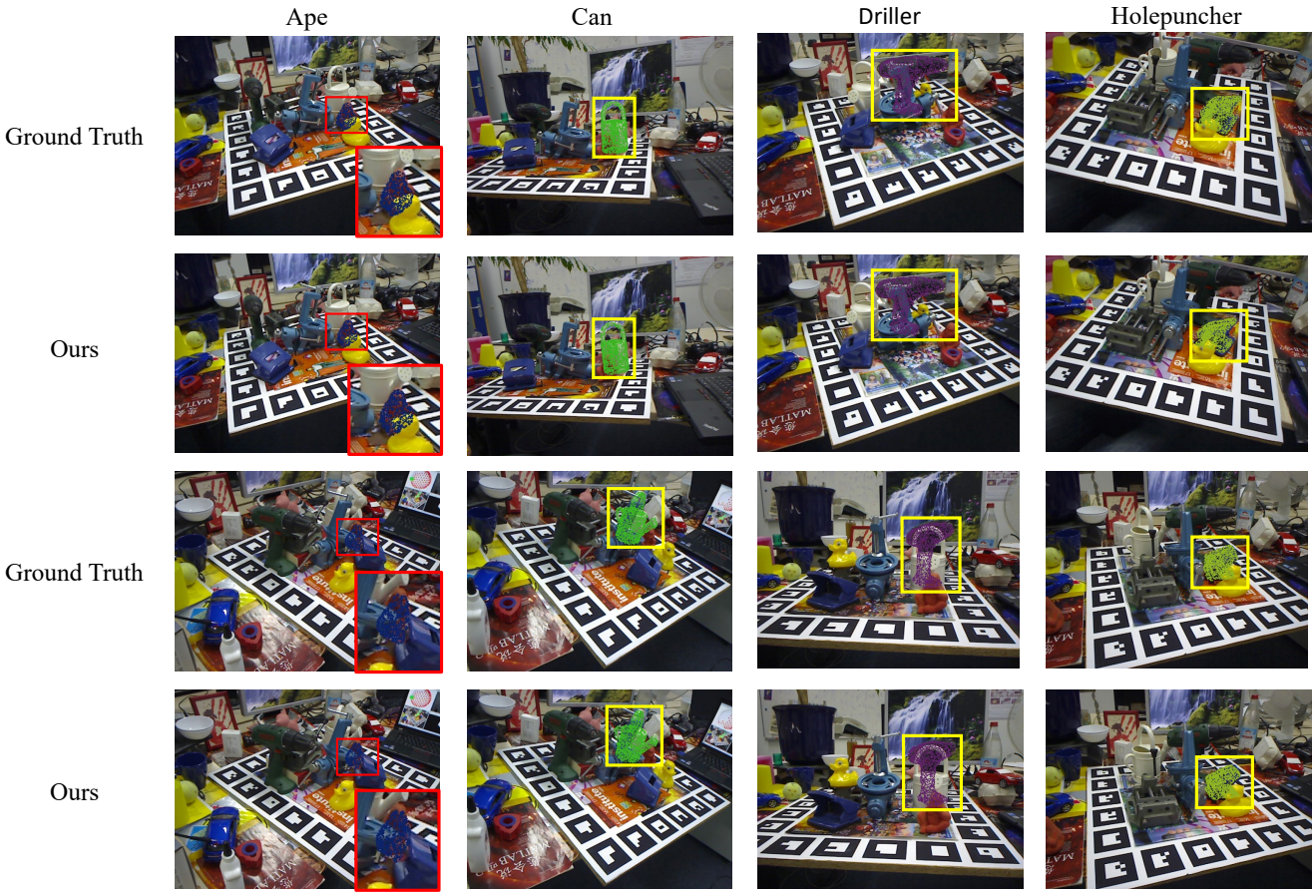

Figure 9. Qualitative comparison results on Occlusion-LineMOD dataset.