Supplementary Material for Mitigating Perspective Distortion-induced Shape Ambiguity in Image Crops

Aditya Prakash, Arjun Gupta, and Saurabh Gupta

University of Illinois Urbana-Champaign
{adityap9,arjung2,saurabhg}@illinois.edu
https://bit.ly/AmbiguityEnc

In this supplementary document, we first provide additional analysis (Sec. A) of the results in the main paper. We then show visualizations of our approach and baseline on each of the three tasks (Sec. B).

A Additional Analysis

Dense metric depth prediction: We train ZoeDepth [1] with crop augmentation on NYU [8] at different input resolutions: 96×128, 384×512. We compare different variants of KPE with Cam-Conv [4] in Tab. A. Similar to the results in the main paper, KPE outperforms Cam-Conv [4], a past encoding specifically designed for depth estimation. We see gains in both settings, with & without pre-training of the MiDaS backbone, with larger gains at smaller resolution.
3D object detection: We add KPE to Cube R-CNN [2] and compare to sinusoidal encoding of image location & Cam-Conv, in terms of AP scores at different IoUs. We consider both single dataset (nuScenes [3], KITTI [6]) and multiple dataset (nuScenes [3] + KITTI [6]) settings for training in Tab. B. Overall, we see consistent improvements across all metrics in most settings. The gains are more prominent in multi-dataset setting, indicating that KPE is more effective with input from different cameras.
3D pose of articulated objects: We ablate different variants of KPE, *i.e.*, sparse, dense & sparse+dense, when added to ArcticNet-SF [5] on ARCTIC. We observe that sparse+dense performs the best in most settings (Tab. C). Note that all the results on validation set use ground truth bounding boxes. For evaluation on the ARCTIC leaderboard, we train a bounding box detector, by finetuning a MaskRCNN [7] model on the ARCTIC training set, since ground truth bounding boxes for the test set are not available. Our object bounding box predictor achieves a mIoU of 0.91 on the ARCTIC validation set. We also show results with predicted bounding box in Tab. C, which leads to a decrease in performance compared to ground truth boxes, but the trends remain consistent.
Variation in camera's field of view (FOV):. The parallelepipeds case study (Sec. 3 in the paper) uses a fixed FOV. We repeat the study with with different FOV values (Fig. A) and observe the same trends as in the paper, *i.e.*, predicting 3D shape from 2D image crops fails in the absence of information about the location of the crop in camera's field of view (evident by training loss saturating

at a high value due to the inherent ambiguity). Adding information about the location of crop in camera's field of view alleviates the ambiguity in predicting 3D shape. Moreover, the gap between using centered and absolute inputs increases as the ambiguity increases with wider fov.

Metrics *vs.* crop distance from image center:. We study the impact in performance of different models as the input crops are further away from the image center for the task of pose estimation on ARCTIC. We plot the variantion in CDev and MRRPE with the crops distance from center in Fig. B. As the crop goes farther away, metrics get worse for ArcticNet (baseline) whereas KPE retains the same performance, indicating better robustness.

Reliance on camera calibration:. We progressively add noise to the intrinsics from 1-5% at evaluation to test the robustness to intrinsics on pose estimation task on ARCTIC. From the results in Tab. D, we observe a slight decrease in the metrics, but KPE is still better than ArcticNet & Image location baselines in most cases. This indicates that KPE is not overly reliant on accurate intrinsics.

Table A: Dense metric depth prediction with crop augmentation. We test the effectiveness of adding KPE to ZoeDepth [1] trained with crop augmentation at different input resolutions: 96×128, 384×512. We see gains in both settings, with and without pre-training of the MiDaS backbone, with larger gains at smaller resolution.

	Method	**REL** ↓	**RMSE** ↓	$\log_{10}$ ↓
96×128 crops	ZoeD-X-N	0.182	0.538	0.078
	ZoeD-X-N + Cam-Conv	0.182 -0.0%	0.542 +0.7%	0.078 -0.0%
	ZoeD-X-N + KPE (dense, after MiDaS backbne)	0.175 -3.8%	0.520 -3.3%	0.075 -3.8%
	ZoeD-X-N + KPE (dense, inside MiDAS backbne)	**0.175** -3.8%	**0.516** -4.1%	**0.074** -5.1%
	ZoeD-M12-N	0.187	0.536	0.079
	ZoeD-M12-N + Cam-Conv	0.187 -0.0%	0.553 +3.2%	0.081 +2.5%
	ZoeD-M12-N + KPE (dense, after MiDaS backbne)	0.186 -0.5%	0.538 -0.4%	0.078 -1.3%
	ZoeD-M12-N + KPE (dense, inside MiDAS backbne)	**0.182** -2.7%	**0.516** -3.7%	**0.074** -6.3%
384×512 crops	ZoeD-X-N	0.090	0.353	**0.039**
	ZoeD-X-N + Cam-Conv	0.100 +11.1%	0.405 +14.7%	0.043 +10.3%
	ZoeD-X-N + KPE (dense, after MiDaS backbne)	0.090 -0.0%	0.360 +1.9%	0.040 +2.6%
	ZoeD-X-N + KPE (dense, inside MiDaS backbne)	**0.089** -1.1%	**0.350** -0.8%	**0.039** -0.0%
	ZoeD-M12-N	0.084	0.337	0.037
	ZoeD-M12-N + Cam-Conv	0.098 +16.7%	0.399 +18.4%	0.042 +13.5%
	ZoeD-M12-N + KPE (dense, after MiDaS backbne)	0.084 -0.0%	0.333 -1.2%	0.036 -2.7%
	ZoeD-M12-N + KPE (dense, inside MiDaS backbne)	**0.082** -2.3%	**0.329** -2.4%	**0.036** -2.7%

B Visualizations

Dense metric depth prediction: We compare the depth predicted by adding KPE encoding to ZoeDepth [1] with the base ZoeDepth model. We show the depth predictions along with the squared error Δ w.r.t. to ground truth depth (ranging from dark blue: low to dark red: high). We consider two settings: crop

Table B: 3D Object Detection. We add KPE to the Cube R-CNN [2] model and compare with Cam-Conv & sinusoidal encoding of image location, in terms of AP scores at different IoUs. We consider both single dataset (nuScenes) and multiple dataset (nuScenes + KITTI) settings for training. Overall, we see consistent improvements across all metrics in most settings. The gains are more prominent in multi-dataset setting, indicating that KPE is more effective with input from different cameras. Green numbers denote relative improvements over not using any encodings.

Method	AP^{25}_{3D} ↑	AP^{50}_{3D} ↑	AP^{75}_{3D} ↑	AP_{3D} ↑
KITTI				
Cube R-CNN	72.59	44.80	14.68	56.11
Cube R-CNN + KPE (dense)	72.05	45.11	15.16	55.96
Cube R-CNN + Image location (dense)	**73.98**	**47.70**	**17.09**	**58.08**
Cube R-CNN + Image location (sparse)	71.81	45.11	14.61	56.00
Cube R-CNN + Cam-Conv [4]	73.76	47.23	16.79	57.80
Cube R-CNN + KPE (sparse)	73.02	47.08	16.34	57.50
nuScenes				
Cube R-CNN	72.83	50.45	**18.76**	58.53
Cube R-CNN + Image location (dense)	72.66	49.80	17.44	57.96
Cube R-CNN + Image location (sparse)	73.10	49.91	17.74	58.41
Cube R-CNN + Cam-Conv [4]	73.90	51.05	18.20	59.31
Cube R-CNN + KPE	**74.03**	**51.50**	18.27	**59.48**
nuScenes + KITTI				
Cube R-CNN	70.17	43.83	14.23	54.59
Cube R-CNN + Image location (dense)	71.97	45.03	15.16	56.86
Cube R-CNN + Image location (sparse)	72.08	45.93	15.32	56.30
Cube R-CNN + Cam-Conv [4]	72.62	46.51	14.59	56.82
Cube R-CNN + KPE (sparse)	**73.10** +4.2%	**47.51** +8.4%	**16.58** +16%	**57.54** +5.4%

Table C: 3D pose estimation of articulated objects. We ablate different variants of KPE, *i.e.*, sparse, dense and a combination of both, when added to ArcticNet-SF [5] on the ARCTIC validation set. We observe that the combination of both sparse and dense KPE performs the best in most settings.

Method	**Object**		**Contact**		**Motion**	
	AAE (°) ↓	Success (%) ↑	$CDev_{ho}$ (mm) ↓	$MRRPE_{ro}$ (mm) ↓	$MDev_{ho}$ (mm) ↓	Acc (m/s^2) ↓
Validation Split						
No KPE	8.0	59.0	44.1	36.8	11.8	11.3
KPE (sparse)	5.9	71.5	39.4	29.7	**9.3**	**8.7**
KPE (dense)	6.2	71.4	37.7	29.3	9.9	9.4
KPE (sparse+dense)	**5.6**	**73.2**	**37.2**	**28.9**	10.0	9.3
Ground truth (GT) vs Predicted bounding box (pred)						
KPE (sparse, GT box)	5.9	71.5	39.4	29.7	9.3	8.7
KPE (sparse, pred box)	6.1	70.2	44.1	29.9	11.6	8.8
KPE (dense, GT box)	6.2	71.4	37.7	29.3	9.9	9.4
KPE (dense, pred box)	6.0	69.7	41.6	31.3	11.8	9.0

Table D: Perturbing camera intrinsics. We study the reliance of KPE on accurate intrinsics by progressively adding the noise levels from 1-5% on pose estimation task on ARCTIC. We observe slight decrease in the scores, but KPE is still better than ArcticNet & Image location baselines in most cases, indicating that model is not too reliant on intrinsics.

Method	Object		Contact		Motion	
	AAE (°) ↓	Success (%) ↑	CDev$_{ho}$ (mm) ↓	MRRPE$_{ro}$ (mm) ↓	MDev$_{ho}$ (mm) ↓	Acc (m/s^2) ↓
ArcticNet	8.02	58.96	44.13	36.77	11.82	11.27
Image location	6.16	67.96	39.10	29.99	10.13	9.25
No noise	5.92	71.55	39.38	29.68	9.28	8.72
1%	5.92	71.55	39.39	29.69	9.42	8.73
2%	5.92	71.55	39.42	29.73	9.59	8.74
5%	5.92	71.55	39.58	29.89	10.87	8.80

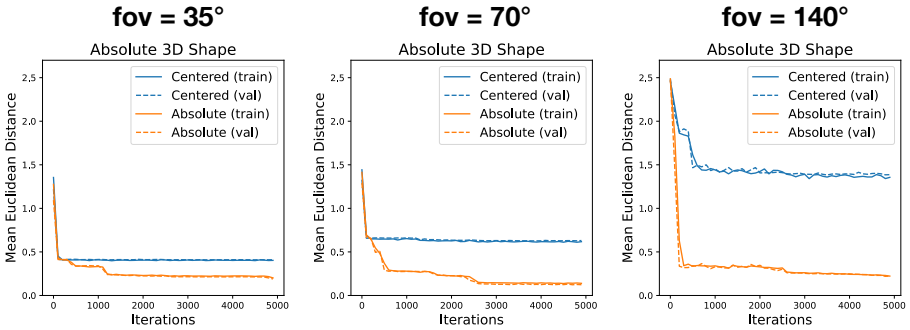

Fig. A: Variation in camera's field of view. In the parallelepipeds case study, we observe consistent trends with different FOV, *i.e.*, evident by training loss saturating at a high value due to the inherent ambiguity and the location of crop in camera's field of view helps alleviates the ambiguity. Moreover, the gap between using centered and absolute inputs increases as the ambiguity increases with wider FOV.

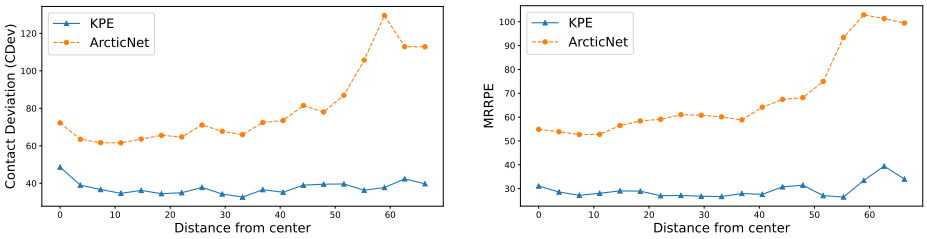

Fig. B: Metrics vs crop distance from image center. We plot the change in CDev (left) & MRRPE (right) as crops go farther from the image center for pose estimation task on ARCTIC. We observe deterioration in scores for ArcticNet (baseline) whereas adding KPE helps retain performance.

augmentation at high resolution 384×512 crops (Fig. D) & low resolution 96×128 crops (Fig. E) and scale augmentation at 384×512 input resolution (Fig. C). Our model predicts better depth as evident by lower Δ (lower intensity red areas). Gains are more prominent in the low resolution setting compared to 384×512 setting when using crop augmentation.

3D object detection: We show the 3D bounding box predictions on Cars category for Cube R-CNN [2] and our model (Cube R-CNN + KPE) , both in image space and in top-down view, in Fig. F (KITTI) and Fig. G (nuScenes). Our model predicts better 3D bounding boxes, as evident by fewer collisions (*i.e.* intersections between car bounding boxes) and missed detections. We consider the models trained jointly on KITTI and nuScenes in these visualizations.

3D pose of articulated objects: We show the 3D pose predictions of our model and the baseline on ARCTIC in Fig. H. We observe that our model predicts better 3D poses in interaction scenarios (note the difference in the articulation angle and global pose). For each image, we show the projection of the object mesh with the predicted pose on the image and from 2 different camera views. We also show 2 *failure cases* of our approach in the last row.

Fig. C: Depth prediction on NYU [8] with scale augmentation. We compare the depth predicted by adding KPE to ZoeDepth [1] with the base ZoeDepth model. We show the depth predictions along with the squared error Δ w.r.t. to ground truth depth (ranging from dark blue: low to dark red: high. Invalid regions are indicated as grey). Our model predicts better depth as evident by lower Δ (lower intensity red areas).

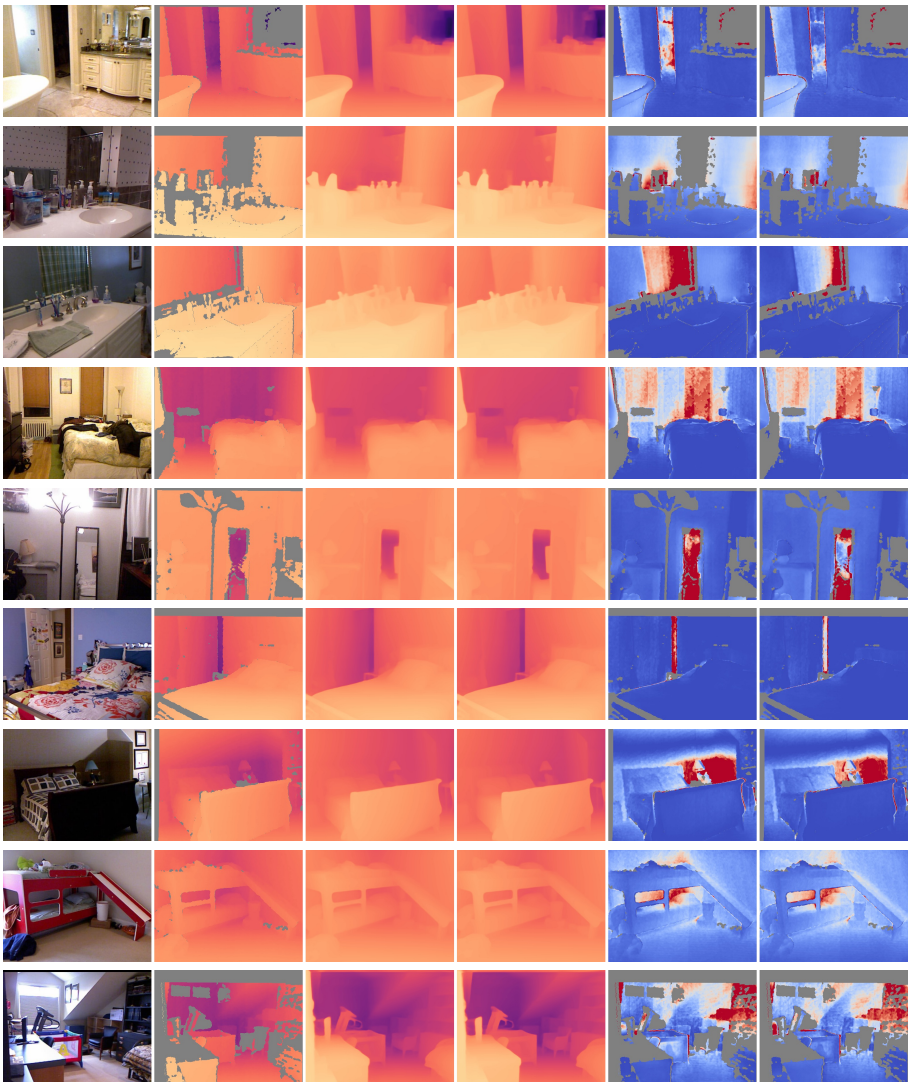

Fig. D: Depth prediction on NYU [8] with 384×512 **crops**. We compare the depth predicted by adding KPE encoding to ZoeDepth [1] with the base ZoeDepth model, when trained with crop augmentation. We show the depth predictions along with the squared error Δ w.r.t. to ground truth depth (ranging from dark blue: low to dark red: high. Invalid regions are indicated as grey). Our model predicts better depth as evident by lower Δ (lower intensity red areas).

Fig. E: Depth prediction on NYU [8] with 96×128 crops. We compare the depth predicted by adding KPE encoding to ZoeDepth [1] with the base ZoeDepth model, when trained with crop augmentation. We show the depth predictions along with the squared error Δ w.r.t. to ground truth depth (ranging from dark blue: low to dark red: high. Invalid regions are indicated as grey). Our model predicts better depth as evident by lower Δ (lower intensity red areas). Gains are more prominent in this low resolution setting compared to 384×512 setting (Fig. D).

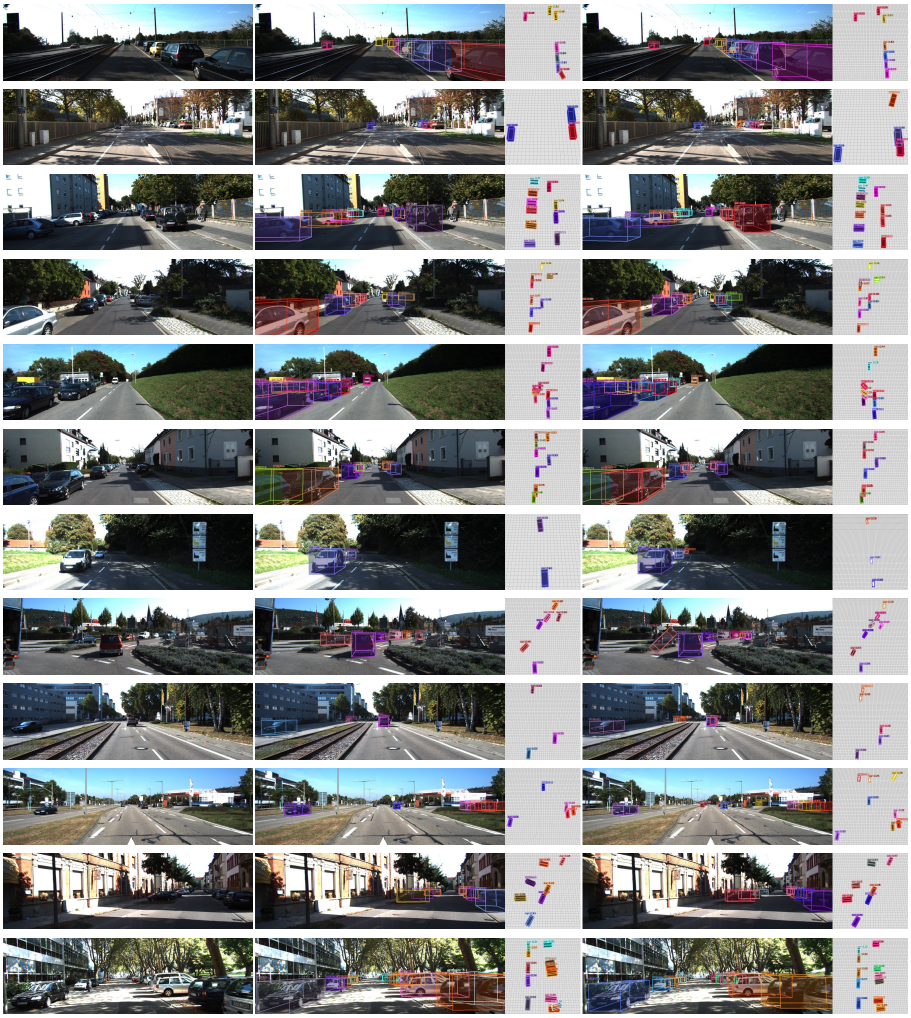

Image Cube R-CNN Ours

Fig. F: 3D Object Detection on KITTI [6]. We show the 3D bounding box predictions on Cars category for Cube R-CNN [2] and our model (Cube R-CNN + KPE), both in image space and in top-down view. Our model predicts better 3D bounding boxes, as evident by fewer collisions (*i.e.* intersections between car bounding boxes) and missed detections.

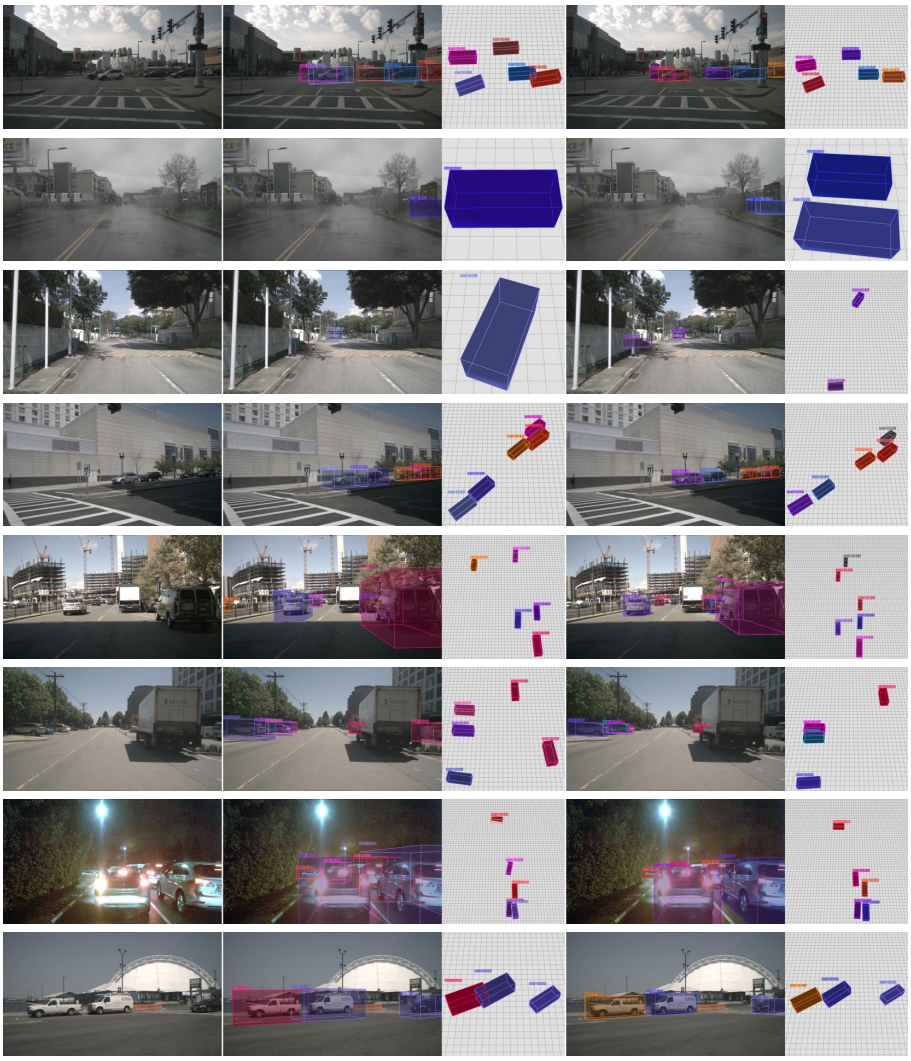

Image Cube R-CNN Ours

Fig. G: 3D Object Detection on nuScenes [3]. We show the 3D bounding box predictions on Cars category for Cube R-CNN [2] and our model (Cube R-CNN + KPE), both in image space and in top-down view. Similar to the results on KITTI (Fig. F), we observe better 3D bounding box prediction of our model, as evident by fewer collisions (*i.e.* intersections between car bounding boxes) and missed detections.

Fig. H: 3D pose visualizations on ARCTIC [5]. Our proposed modification of intrinsics-aware positional encoding (KPE) improves over the ArcticNet-SF [5] model by predicting better 3D poses in interaction scenarios (note the difference in the articulation angle and global pose). For each image, we show the projection of the object mesh with the predicted pose on the image and from 2 different camera views. We also show 2 *failure caes* of our approach in the last row.

References

1. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv (2023)
2. Brazil, G., Kumar, A., Straub, J., Ravi, N., Johnson, J., Gkioxari, G.: Omni3d: A large benchmark and model for 3d object detection in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13154–13164 (2023)
3. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
4. Facil, J.M., Ummenhofer, B., Zhou, H., Montesano, L., Brox, T., Civera, J.: Camconvs: Camera-aware multi-scale convolutions for single-view depth. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11826–11835 (2019)
5. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
6. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
7. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
8. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 746–760. Springer (2012)