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

Aditya Prakash Matthew Chang Matthew Jin Saurabh Gupta
University of Illinois Urbana-Champaign
bit.ly/AmbiguityEnc

In this document, we provide visualizations of our approach and baseline on each of the three tasks. The video summarizes our key ideas, contributions and results.

1. Visualizations

3D pose of articulated objects on ARCTIC [4]: We show the 3D pose predictions of our model and the baseline on ARCTIC in Fig. 1. 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.

Depth prediction on NYU [6]: 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: high resolution 384×512 crops (Fig. 2) and low resolution 96×128 crops (Fig. 3). 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.

3D Object Detection on KITTI [5] & 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, in Fig. 4 (KITTI) and Fig. 5 (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.

References

[1] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. 2023. 1, 3, 4

[2] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. 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)*, pages 13154–13164, 2023. 1, 5, 6

[3] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 6

[4] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 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. 1, 5

[6] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 746–760. Springer, 2012. 1, 3, 4

Figure 1. **3D pose visualizations on ARCTIC [4]**. Our proposed modification of intrinsics-aware positional encoding (KPE) improves over the ArcticNet-SF [4] 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.

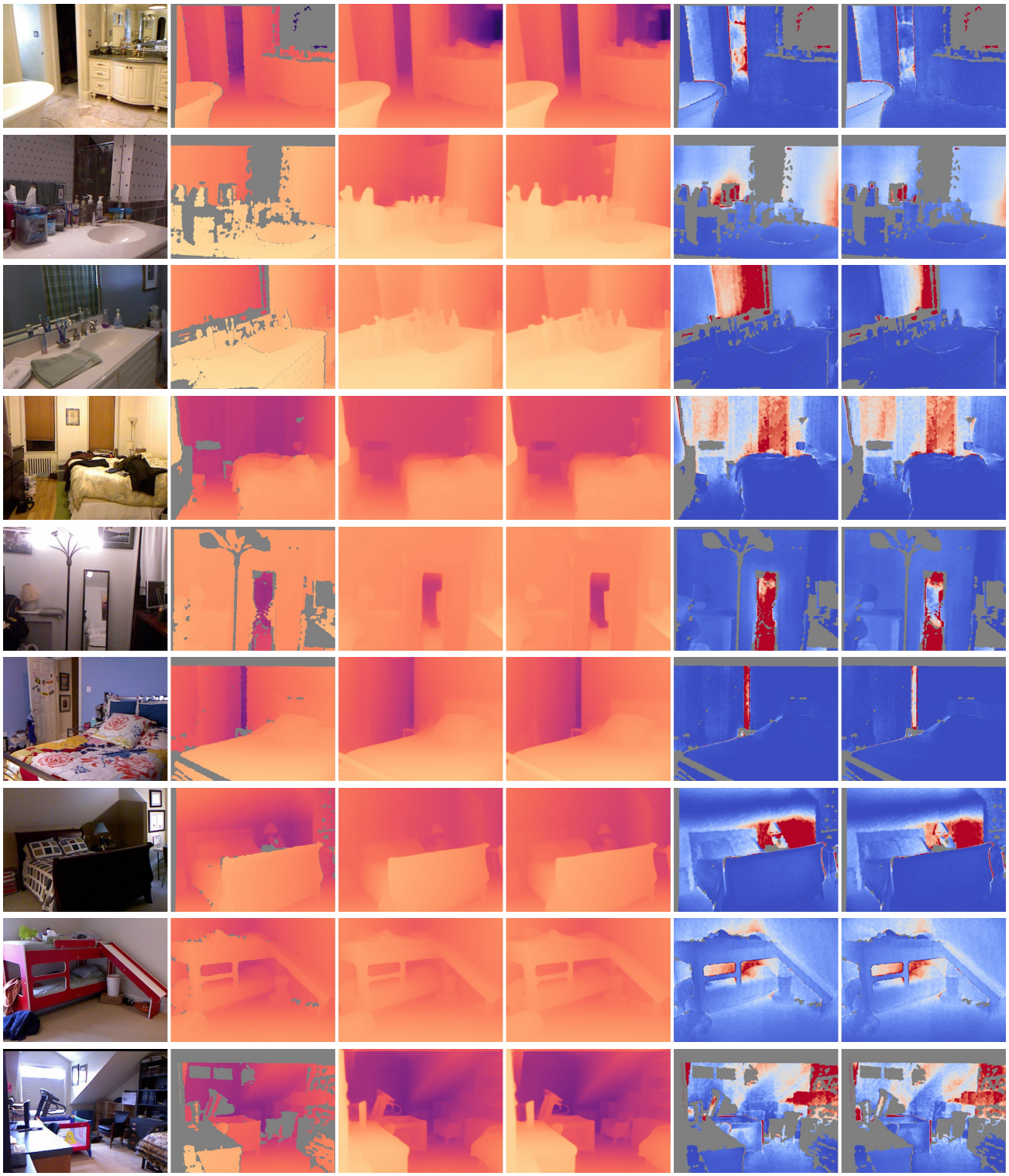

Figure 2. **Depth prediction on NYU [6] with** 384×512 **crops**. We compare the depth predicted by adding KPE encoding to ZoeDepth [1] with the base ZoeDepth model. We show the depth predictions along with the squared error Δ w.r.t. to ground truth depth (ranging from dark blue: low to dark red: high. Invalid regions are indicated as grey). Our model predicts better depth as evident by lower Δ (lower intensity red areas).

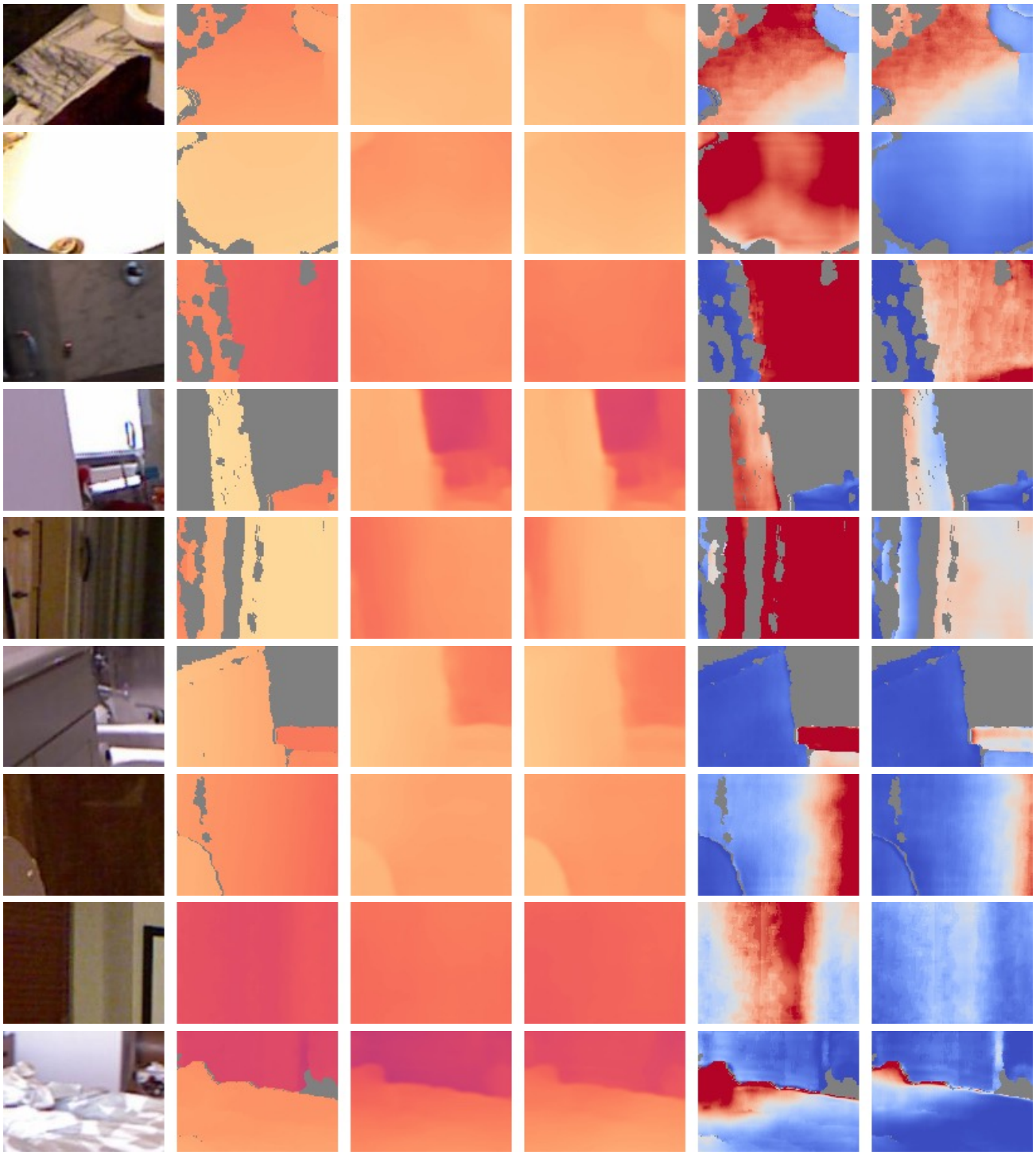

Figure 3. **Depth prediction on NYU [6] with** 96×128 **crops**. We compare the depth predicted by adding KPE encoding to ZoeDepth [1] with the base ZoeDepth model. We show the depth predictions along with the squared error Δ w.r.t. to ground truth depth (ranging from dark blue: low to dark red: high. 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. 2).

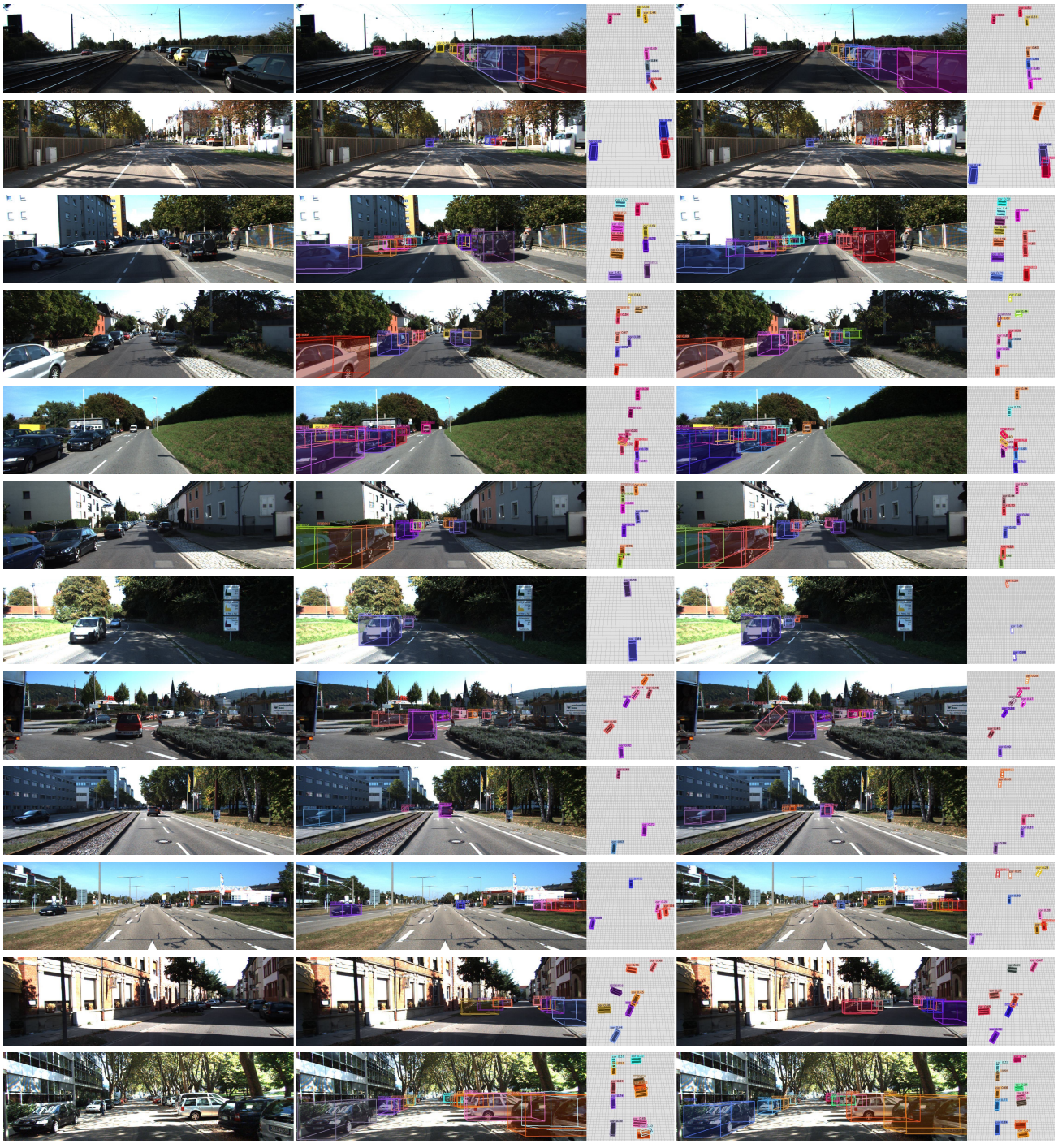

Figure 4. **3D Object Detection on KITTI [5].** We show the 3D bounding box predictions on Cars category for Cube R-CNN [2] and our model (Cube R-CNN + KPE), both in image space and in top-down view. Our model predicts better 3D bounding boxes, as evident by fewer collisions (*i.e.* intersections between car bounding boxes) and missed detections.

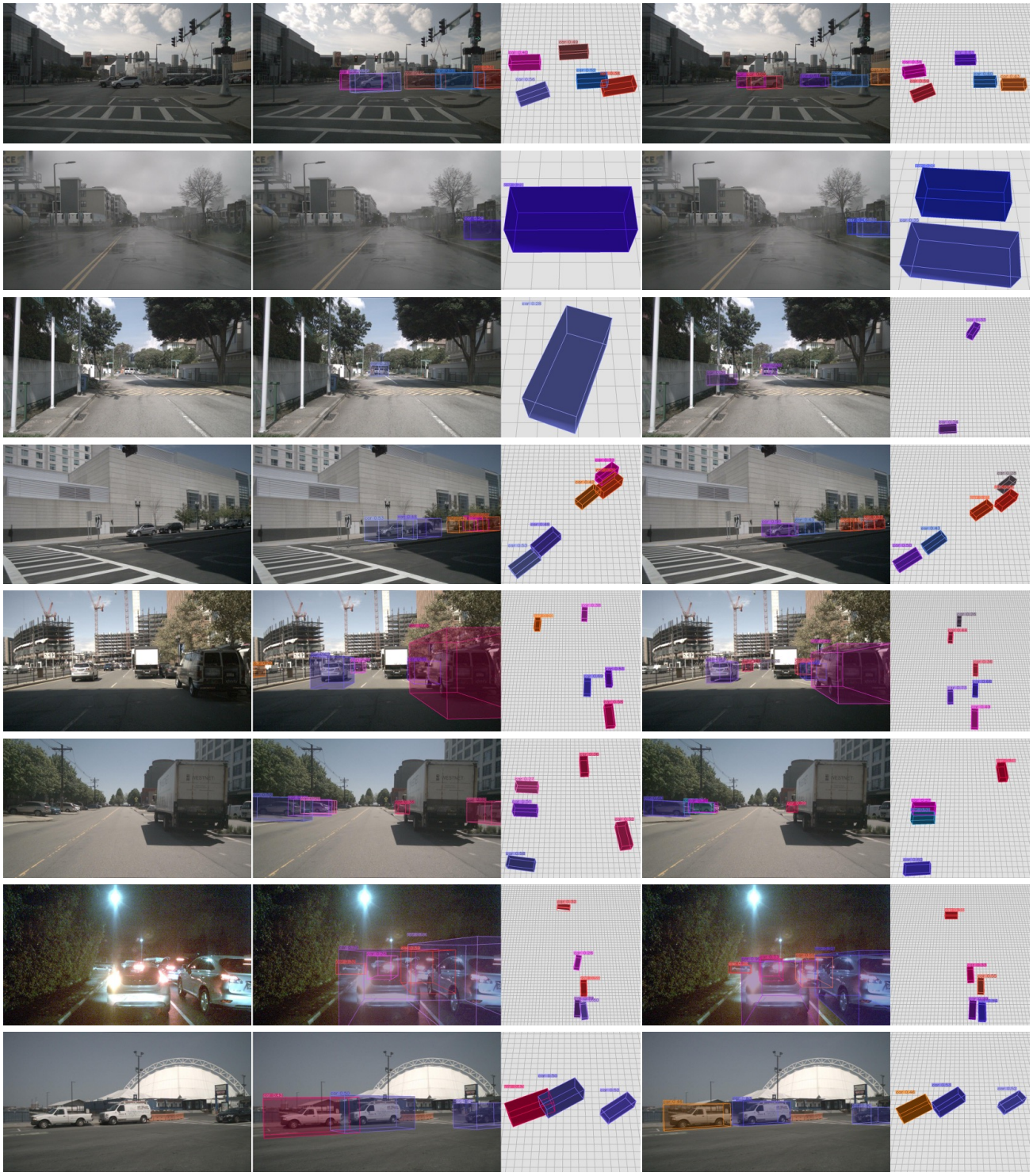

Figure 5. **3D Object Detection on nuScenes [3]**. We show the 3D bounding box predictions on Cars category for Cube R-CNN [2] and our model (Cube R-CNN + KPE), both in image space and in top-down view. Similar to the results on KITTI (Fig. 4), we observe better 3D bounding box prediction of our model, as evident by fewer collisions (*i.e.* intersections between car bounding boxes) and missed detections.