WALT3D: Generating Realistic Training Data from Time-Lapse Imagery for Reconstructing Dynamic Objects under Occlusion

SUPPLEMENTARY MATERIAL

Khiem Vuong[1,*] N Dinesh Reddy[2,*] Robert Tamburo[1] Srinivasa G. Narasimhan[1]

[1]Carnegie Mellon University [2]Amazon

https://www.cs.cmu.edu/~walt3d

A. Summary and More Results

For a brief summary of our method and additional results, we highly encourage the readers to check out the included short video.

B. Vehicle 2D Keypoints Dataset

As mentioned in the main paper, although existing datasets like PASCAL3D+ [11], KITTI-3D [3], Carfusion [6], and ApolloCar3D [8] provide annotated 2D vehicle keypoints, they mostly focus on driving scenes [3, 6, 8] or have limited training examples [11], lacking the necessary appearance diversity. To increase the dataset diversity, we prioritized the number of different cameras and viewpoints rather than the number of images per camera. A summary and comparison of our proposed Vehicle 2D Keypoints dataset with other publicly available datasets are detailed in Table 1. On average, we extracted 120 images per camera source for more than 60 different cameras spanning a wide variety of viewpoints, appearances, sensor types, etc. For each image, we run an off-the-shelf object detector to extract the car instances with high confidence score. This set of car instances are manually annotated by the trained annotators from a commercial annotation service. We utilized a web-based interface annotation tool from DeepLabCut [5] where the annotators were asked to select 12 keypoint locations and its corresponding occlusion category (visible/self-occluded/occluded-by-others) for every car. Note that we also asked the annotators to filter out erroneous instances such as bad quality images and/or wrong detections. As of the time of paper submission, we have annotated a total of 42,547 car instances in 7,018 images.

C. Camera Intrinsics and Ground Plane

We follow Vuong et al. [9] to obtain the intrinsic parameters and ground plane equation for each of the stationary traffic camera. Specifically, we used the panorama images from Google Street View (GSV) [2] to build a metric 3D scene reconstruction (at the desired camera location), then the stationary camera is registered within the reconstruction to determine its intrinsic and extrinsic parameters. We also geo-register the scene to a metric scale using the GPS coordinates, and the road plane equation is estimated by fitting a plane to the set of 3D points whose 2D pixel projections belong to the *road* category obtained from off-the-shelf semantic segmentation method [1]. The camera poses and plane equation are used in 3D reconstruction pipeline to reconstruct unoccluded objects as described in the main paper. Thanks to the ground plane geometry constraint, we can reconstruct the accurate 3D geometry of cars and pedestrians, generating realistic occlusion configurations. This method enables us to obtain accurate calibration for more than 100 stationary cameras worldwide, thus allowing for a significant expansion of our clip-art dataset.

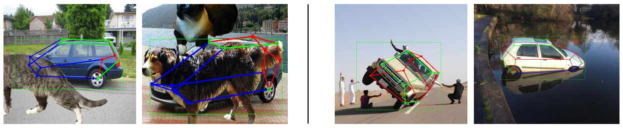

Figure 1. Qualitative Results on OccludedPascal3D+ (left) and OOD-CV (right) dataset.

D. Benchmarking on Additional Datasets

Evaluation on OccludedPascal3D+ dataset: Table 2 shows that our method performs better than NeMo [10] and Ma et al. [4] on the OccludedPascal3D+ [10] dataset.
Evaluation on OOD-CV dataset: Quantitative results on OOD-CV [12] dataset are shown in Table 3. Although our method has never been trained on the anomalous scenarios in this dataset, our approach shows higher performances on many testing subsets. Please see qualitative results in Fig. 1.
Mining Unoccluded Objects: To identify unoccluded objects, we evaluate two methods: a simple heuristic based on bounding box IOU threshold δ (as used in WALT [7]) and training an Occlusion Classifier (OC) using human-

Dataset	Image source	Appearance diversity in terms of				# images	# car instances	Occ. keypoint annotations	Per-keypoint occ. type
		Cities	Times of Day	Weathers	Viewpoints				
PASCAL3D+	Natural	Yes	Yes	Yes	No	6,704	7,791	No	No
KITTI-3D	Self-driving	No	No	No	No	2,040	2,040	No	No
Carfusion	Handheld	No	No	No	No	53,000	100,000	Yes	No
ApolloCar3D	Self-driving	No	No	No	No	5,277	60,000	No	No
Ours	Handheld Self-driving Traffic cameras	Yes	Yes	Yes	Yes	7,018	42,547	Yes	Yes

Table 1. Summary and comparison of our **Vehicle 2D Keypoints dataset** to other publicly available datasets.

Method	$Acc\,(\frac{\pi}{6})$			$Acc\,(\frac{\pi}{18})$			Med Pose Err			Med ADD		
	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3
Occluded PASCAL3D+ Dataset (car)												
NeMo	48.3	34.3	18.2	17.4	9.6	3.3	0.5	1.0	2.4	1.9	2.0	2.3
Ma et al.	66.6	47.9	27.4	30.8	16.2	5.3	0.3	0.6	1.1	0.8	1.2	1.9
Ours	**70.4**	**56.5**	**35.3**	**36.8**	**25.4**	**15.3**	**0.2**	**0.4**	**0.8**	**0.6**	**1.0**	**1.4**

Table 2. Baseline comparisons across object pose metrics on Occluded-PASDAL3D+ [10] for vehicle category.

$Acc - \frac{\pi}{6}$	i.i.d	shape	pose	texture	context	weather
NeMo	66.7	**51.7**	**56.9**	52.6	**51.3**	49.8
Ours	**75.4**	48.6	50.8	**56.7**	49.1	**55.6**

Table 3. Comparisons on the OOD-CV [12] dataset (car).

Metric	$\delta = 0.01$	$\delta = 0.1$	$\delta = 0.2$	$\delta = 0.5$	**OC (ours)**
Recall	0.60	0.42	0.17	0.01	**0.81**
Precision	0.32	0.41	0.52	0.57	**0.70**

Table 4. Accuracy of our OC module compared with baseline using bbox IOU threshold δ in detecting unoccluded objects.

annotated data (using images from our new vehicle keypoints dataset). Table 4 demonstrates that our OC module is more effective than the heuristic, particularly in inter-category occlusion scenarios (e.g., vehicles occluded by people or background objects). This allows us to efficiently filter out unwanted occluded objects in the training data, improving data purity. While not essential for our method, we believe this human-annotated dataset is important for future research on understanding and handling occlusion.

E. Additional 2D/3D Clip-Art Data Examples

More examples from our 2D/3D Clip-Art pseudo-groundtruth supervision data, including the clip-art image with corresponding amodal segmentation, keypoints, and 3D object reconstruction, are shown in Fig. 2.

F. Additional Qualitative Results

Additional results are shown in Fig. 3, with various occlusion configurations, including self-occlusion, truncation, and occlusion-by-others. Notably, training with our clip-art data yields a substantial improvement over baseline methods, particularly in scenarios with heavy occlusion.

References

[1] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 1

[2] Google. Google Street View. https://www.google.com/streetview/. 1

[3] Chi Li, M Zeeshan Zia, Quoc-Huy Tran, Xiang Yu, Gregory D Hager, and Manmohan Chandraker. Deep supervision with shape concepts for occlusion-aware 3d object parsing. In *CVPR*, 2017. 1

[4] Wufei Ma, Angtian Wang, Alan Yuille, and Adam Kortylewski. Robust category-level 6d pose estimation with coarse-to-fine rendering of neural features. In *ECCV*, 2022. 1

[5] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie W. Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 2018. 1

[6] N. Dinesh Reddy, Minh Vo, and Srinivasa G. Narasimhan. Carfusion: Combining point tracking and part detection for dynamic 3d reconstruction of vehicles. In *CVPR*, 2018. 1

[7] N. Dinesh Reddy, Robert Tamburo, and Srinivasa G. Narasimhan. Walt: Watch and learn 2d amodal representation from time-lapse imagery. In *CVPR*, 2022. 1, 3, 4

[8] Xibin Song, Peng Wang, Dingfu Zhou, Rui Zhu, Chenye Guan, Yuchao Dai, Hao Su, Hongdong Li, and Ruigang Yang. Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving. In *CVPR*, 2019. 1

[9] Khiem Vuong, Robert Tamburo, and Srinivasa G. Narasimhan. Toward planet-wide traffic camera calibration. In *WACV*, 2024. 1

[10] Angtian Wang, Adam Kortylewski, and Alan Yuille. Nemo: Neural mesh models of contrastive features for robust 3d pose estimation. In *ICLR*, 2021. 1, 2

[11] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, 2014. 1

[12] Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. Ood-cv: a benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. In *ECCV*, 2022. 1, 2

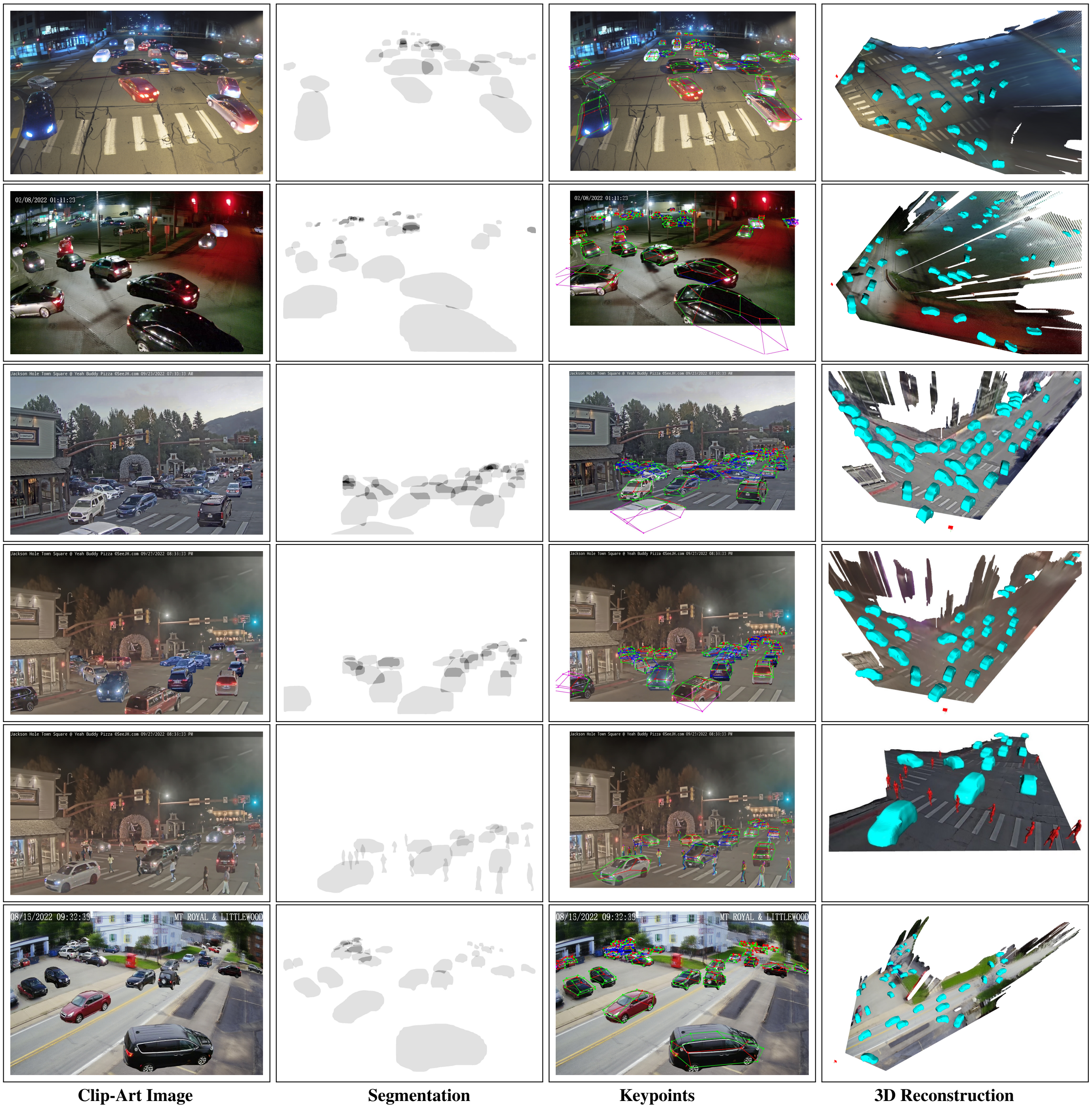

Figure 2. **Automatically generated 2D and 3D Clip-Art to supervise our network:** Unoccluded objects are first mined using time-lapse imagery of the WALT dataset [7]. Random non-intersecting unoccluded objects are composited back into the background image in their respective original positions to preserve correct appearances. The resulting Clip-Art images and their respective amodal segmentation masks, keypoint locations, and 3D meshes are shown. Our method generates realistic appearances from any camera, incorporating diverse viewing geometries, weather conditions, lighting, and occlusion configurations.

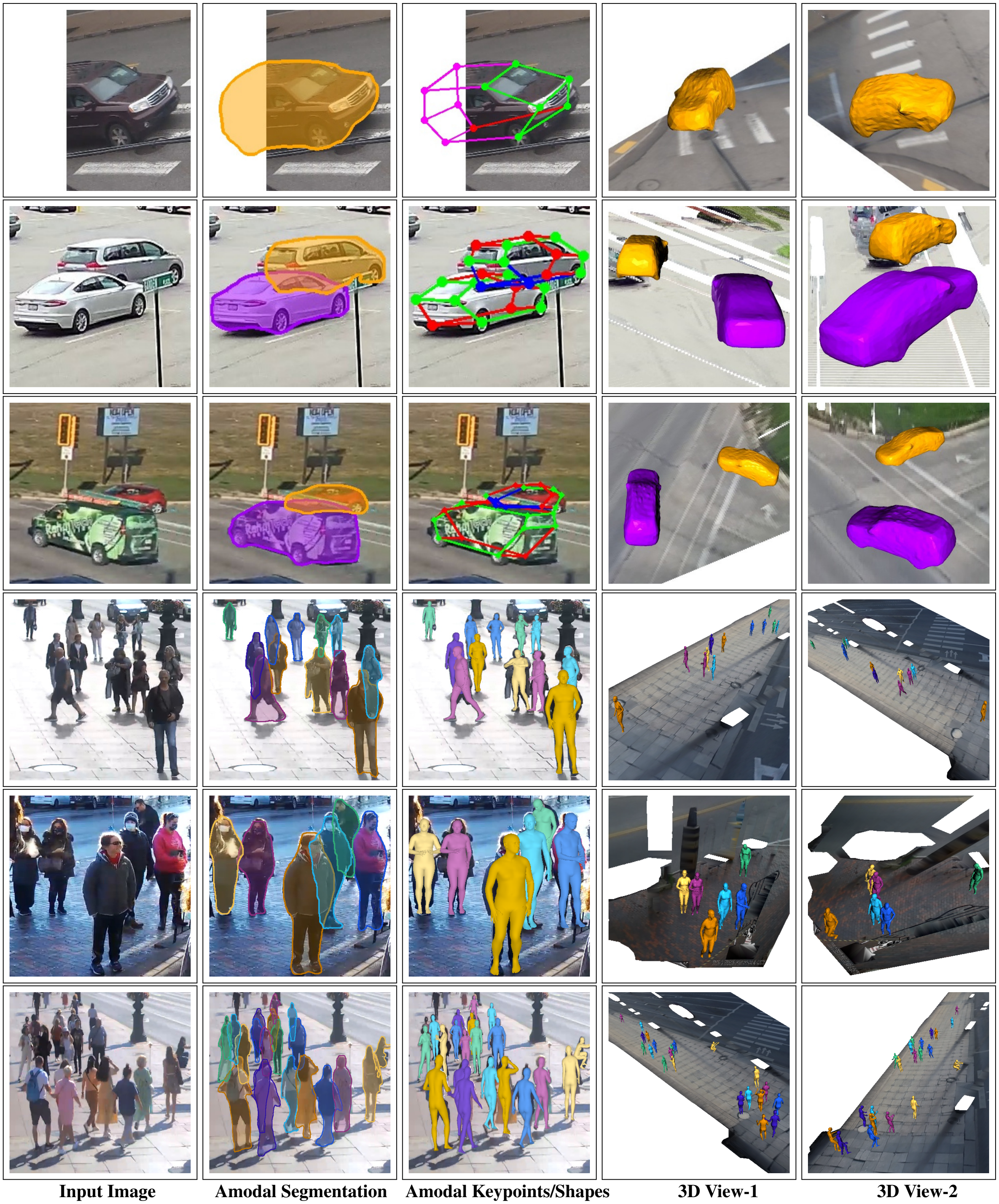

Figure 3. We show additional qualitative results on multiple sequences of the WALT [7] dataset. Our method produces accurate amodal segmentation, keypoints, as well as 3D poses and shapes across diverse poses and occlusion configurations.