Supplementary for WALT: Watch And Learn 2D Amodal Representation from Time-lapse Imagery

N Dinesh Reddy Robert Tamburo Srinivasa G. Narasimhan
Carnegie Mellon University
{dnarapur, rtamburo, srinivas}@cs.cmu.edu

1. Dataset Details:

WALT Dataset: We use [3] to detect faces in the images. We blur the detected faces to anonymous the people. We have captured the data from each camera for nearly an year. The cameras are located in multiple cities spanning different continents to generalize well to different situations. The dataset has not been included with the submission because of the size of the dataset. Similarly code as well is specific to the dataset and cannot be included as standalone code cannot be run without the dataset. We add multiple videos to showcase the dataset in the website and show results on different sequences to emphasis the advantage of our algorithm. We will be making the dataset and code to automatically capture and retrain networks public with the camera ready submission. We observe that the accuracy of segmentation is slightly lower when trained on blurred faces compared to non-blurred faces.

Clip Art WALT Dataset(CWALT): We generate 10000 images per camera for training and 1000 images for testing across all 10 cameras. Different layers of occlusions are captured using different labels. Further we can generate different representations using this methodology like keypoints, segmentation, 3D reconstuctions etc.

Rendered WALT Dataset(RWALT): The computer graphic dataset is used to cross verify our method choices and compute ablation analysis on different segments of the algorithm.

KINS and COCOA: These datasets are manually annotated for the task of amodal segmentation. KINS dataset [6] was built on top of KITTI Dataset [1] with amodal annotations of 14991 images with 7,474 training and 7517 testing images. The dataset is more oriented for the task of autonomous driving tasks. COCOA Dataset contains 3823 images with 1500 training and 1323 testing samples from images from crowd-sourcing. We used these datasets to pretrain the models used for comparision. We did not evaluate the method on these datasets because of the human labels used to evaluate. Clearly our self-Supervision is superior to the human annotated methods. This has even been cross validated using [7]. Since our framework is a continous framework, we outperform [7].

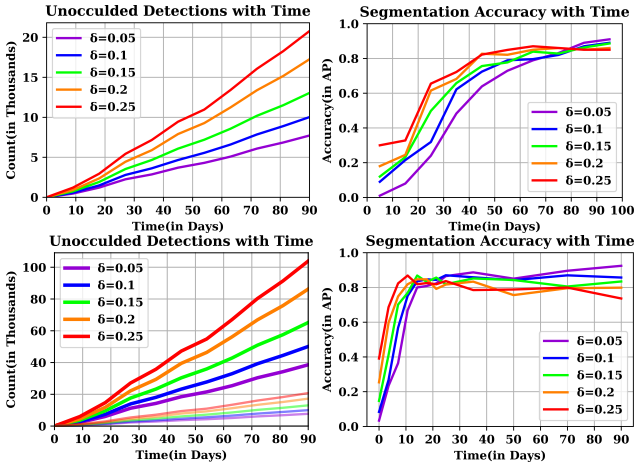

Figure 1. We Compare the time taken to train the network with the new strategy(bottom) vs naive sampling of unoccluded objects(top). Clearly the time taken for convergence of the accuracy is longer using the naive strategy compared to our tracking based approach.

2. Implementation Details:

Network Architecture: We use the mmdetection [2] based code base to train the network. The backbone [5] has been ported from [2]. All the materials and code used in the submission belong to MIT License. We replicate the Maskrcnn Head for each of the proposed heads .i.e Occluder Head, Occluded Head and Amodal Object Head. From the ROI we compute feature maps of 3 layers i.e. first layer is $14 \times 14 \times 256$, second layer is $14 \times 14 \times 256$, third layer is $28 \times 28 \times 256$. Finally we do a softmax to produce the mask heatmap of $28 \times 28 \times c$, where c is the number of classes.

Training Details: We train the network using 4 A100 GPUs with a batch size of 11 for 12 epochs for all the trained models in the paper. We used 0.001 learning rate to train the network. We generate the Clip-art automatically while training and are extensively dependent on the CPU computation for

superimposing the objects and generating ground-truth.

Occluded Layer: Every amodal bounding box has three components *i.e*. the object we want to detect (amodal object), object occluding the amodal object (occluder), background or object occluded (occluded). Previous methods like BCNet[30] have not been supervised for the occluded objects making it difficult to distinguish amodal object from occluded object, when multiple objects lie in the bounding box. Our additional supervision of the occluded objects helps distinguish these objects and improve accuracy as shown in Tab. 1(in paper).

Comparison to Human Annotated Datasets: We reiterate that human annotations, especially for strong occlusions, are imprecise to learn amodal representations. Compared to human annotated datasets *i.e*. KINS or COCOA, our stationary object (SWALT) based evaluation methodology produces more accurate ground truth. Further, our SWALT methodology generates much larger test sets compared to any existing human annotated datasets (60K images from WALT dataset compared to 6157 images in KINS dataset) and will grow 100x as data is captured from more cameras in the following years. Scaling human annotations on such expanding datasets is costly and infeasible and our self-supervision based methodology automatically generates accurate and large training and testing datasets for amodal evaluation.

Composite images and Depth: We composite unoccluded objects in their original positions so their perspectives are correct. We generate training images using random overlap between composited objects that could cause physical interaction in real world. This strategy does not effect the testing accuracy with real world occlusions (SWALT) as they are a subset of the composite distribution in CWALT. Further research in compositing using depth can speed up training by reducing unrealistic occlusions in CWALT.

Comparison With SOTA: We use three baselines - MaskRCNN [5], ASN [6], BCNet [4] - to compare our amodal predictions. All methods were trained by combining the KINS [6], COCOA [8] and Synthetic Occlusion [4] datasets using the same backbone [5]. Our model is trained using Clip Art WALT Dataset. Here KINS and COCOA are hand-annotated amodal segmentation datasets. We evaluate the only vehicle and people segmentation in all the datasets. MaskRCNN is trained only on the visible segmentation from the above-specified dataset. Since our method uses ground truth from longitudinal supervision of unlabeled data, we cannot use hand-annotated datasets like KINS and COCOA to compare the accuracy of the method. We use the stationary object-based metric to evaluate the accuracy of the predicted amodal representation from the above methods. Tab 1 shows 12% improvement in the amodal prediction compared to BCNet for amodal segmentation and 6.3 % improvement for amodal bounding box detection on the stationary object WALT Dataset. This clearly shows that longitudinal supervision outperforms both supervised, synthetic COCO-based methods in predicting occlusions in the real world. Observe that our method consistently outperforms other baselines in predicting the amodal segmentation in severe occlusions.

Method	Amodal Box(IOU)		Amodal Seg(IOU)	
	$\gamma = 0.01$	$\gamma = 0.5$	$\gamma = 0.01$	$\gamma = 0.5$
MaskRCNN [5]	76.23	68.62	55.44	37.29
ASN [6]	82.72	79.38	79.45	76.91
BCNet [4]	86.47	82.23	82.79	77.44
WALT-Net	**91.9**	**91.71**	**92.19**	**91.70**

Table 1. We compare accuracy of amodal segmentation on stationary object WALT Dataset with respect to other SOTA. We observe 12% improvement in amodal segmentation compared to BCNet and 6.3 % improvement for amodal bounding box detection.

References

[1] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1

[2] Github. *https://github.com/microsoft/Swin-Transformer*, 2018. 1

[3] Github. *https://github.com/peiyunh/tiny*, 2018. 1

[4] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4019–4028, June 2021. 2

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021. 1, 2

[6] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2019. 1, 2

[7] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene de-occlusion. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, June 2020. 1

[8] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1464–1472, 2017. 2