WALT: Watch And Learn 2D amodal representation from Time-lapse imagery

N Dinesh Reddy Robert Tamburo Srinivasa G. Narasimhan
Carnegie Mellon University
http://walt.cs.cmu.edu

Abstract

Current methods for object detection, segmentation, and tracking fail in the presence of severe occlusions in busy urban environments. Labeled real data of occlusions is scarce (even in large datasets) and synthetic data leaves a domain gap, making it hard to explicitly model and learn occlusions. In this work, we present the best of both the real and synthetic worlds for automatic occlusion supervision using a large readily available source of data: time-lapse imagery from stationary webcams observing street intersections over weeks, months, or even years. We introduce a new dataset, Watch and Learn Time-lapse (WALT), consisting of 12 (4K and 1080p) cameras capturing urban environments over a year. We exploit this real data in a novel way to automatically mine a large set of unoccluded objects and then composite them in the same views to generate occlusions. This longitudinal self-supervision is strong enough for an amodal network to learn object-occluder-occluded layer representations. We show how to speed up the discovery of unoccluded objects and relate the confidence in this discovery to the rate and accuracy of training occluded objects. After watching and automatically learning for several days, this approach shows significant performance improvement in detecting and segmenting occluded people and vehicles, over human-supervised amodal approaches.

1. Introduction

While there has been strong progress in data-driven methods for object detection [10, 14, 20, 40], tracking [7, 58, 59, 62], segmentation [4, 22, 30, 39, 50] and reconstruction [25, 27, 29, 53] with limited occlusions, most methods underperform in severely occluded scenarios. Severe occlusions are common in busy intersections and crowded places. Even in less dense scenes, pedestrians and vehicles often pass each other or pass behind other objects. As a result, objects are either not detected at all, or the 2D bounding boxes and segments are truncated and produce errors in downstream processes such as 3D reconstruction [5, 6, 25, 41, 42, 45].

Much of this state of affairs can be attributed to the fact that occlusions are treated as noise that must be overcome by robust measures [16, 17, 23, 36, 52, 57]. There are several challenges that make this strategy hard to succeed. First, it is much harder to label object bounding boxes or segments that are occluded, even for humans [47, 49, 63]. Thus, even large datasets like COCO [38] and ImageNet [34] have relatively few objects labeled that are severely occluded [47, 63]. This creates a strong bias against learning robustness to occlusions [11, 46, 56]. Further, the evaluation metrics are often reported on the entire datasets [9, 18, 38] that could hide problems in occluded scenarios.

Figure 1. We visualize the prediction of amodal representation of vehicles and people under severe occlusions trained using our longitudinal self-supervision framework. The method shows significant improvement in amodal detection and segmentation with images captured from different cameras.

As a result, there is growing recognition that occlusions must be explicitly modeled and learned [15, 19, 30, 30, 48, 61]. This has led to new efforts in labeling occlusions explicitly in multiple datasets [21, 47, 63]. Using such supervision, amodal, or holistic, representations (*e.g.* segmentations and bounding boxes) of objects are learned from partially occluded observations [28, 54, 60]. While producing significantly better results than before, these commendable efforts are still plagued by the same challenges - difficulty for humans to label occlusions in real scenes and the limited dataset size. To supplement such limited data, focus has turned toward synthesizing objects in occluded scenarios using synthetic inpainting [30, 31, 60] using computer graphics [1, 13, 24]. CG can generate a large amount of data for supervision (given today's cloud computing resources)

but even the best renderers [8,12,44] leave a notable domain gap to the real data, which needs to be bridged [33,51].

In this work, we present the best of both the real and synthetic worlds for automatic occlusion supervision using a large source of hitherto unexploited data: time-lapse imagery from stationary cameras observing street intersections over weeks, months, and even years[1]. We exploit this data in a novel way to first mine a large dataset of real *unoccluded* objects over time and then use them to synthesize a large number of occlusion scenarios. We develop a new method to classify unoccluded objects based on the idea that when objects on the same ground plane occlude one another, their bounding boxes overlap in a particular common configuration. Once unoccluded objects are discovered, they are composited in layers back into the same scene. These compositions have artifacts that perhaps do not make them too useful for visualization. But they are close enough to real data to reduce the domain gap for a deep network that explicitly predicts the object, its occluder, and the occluded.

Being patient pays off here. Over time, our method discovers tens of thousands of unoccluded objects at diverse positions, orientations, and appearances due to lighting and weather conditions, even in busy scenes. We speed up this discovery by combining sparse time sampling of the data with burst local tracking. This step reduces the required observation period from many months to several days (images captured every few mins.). The data enables us to analyze the performance of our approach over different durations and confidences of self-supervision. Specifically, we relate the confidence in *unoccluded* object prediction to the rate and accuracy of training *occluded* objects. In the beginning, including lower confidence predictions increases more supervision to speed up training, but is quickly passed by training only on high confidence supervisions.

We introduce a new dataset, Watch and Learn Time-lapse (WALT), consisting of 12 (4K or 1080p) cameras capturing urban environments over a year. The cameras view a diverse set of scenes from traffic intersections to boardwalks. The performances of pedestrian and vehicle detection and segmentation improve significantly on all cameras. Like in [32,49,57], we report performances at different levels of occlusion and show that the performance drops more slowly as occlusion increases, compared to methods that do not use longitudinal self-supervision. Because of this, we achieve strong results in detecting and tracking objects as they pass each other - a common failure mode of existing approaches. The methods we present are simple but provide an effective baseline to inspire future work on exploiting longitudinal supervision for computer vision under strong occlusions.

[1]In the past decades, much analysis on time-lapse data was conducted for illumination and weather understanding [35] [43], object insertion and rendering, from thousands of webcams all over the world [2,26,37].

Figure 2. Illustrating the region used to classify unoccluded (Blue) and occluded objects (Red) using planar based IOU (Green) for different categories of objects like vehicles and people.

2. Watch and Learn Amodal Representation

We address the problem of layer representation of objects in a scene under severe occlusions. We propose a continuous learning framework to resolve occlusion ambiguities from images. Initially, given a time-lapse stream of data from a stationary camera, we detect and mine all the unoccluded objects over a long duration of time. These unoccluded objects collected over time automatically act as supervision that we term *longitudinal self-supervision*. We follow a clip art-based integration method to place these unoccluded objects within the scene at the same detected location but overlapping with another unoccluded object from the database. This generates many realistic occlusion configurations for training a network to disentangle holistic object segmentation from a cluttered scene. We further show how to speed up the training for learning amodal representations by tracking around unoccluded detections.

2.1. Unoccluded Object Mining

We exploit the time-lapse data in a novel way to mine a large dataset of real unoccluded objects over time. We develop a new method to classify unoccluded objects based on the idea that when objects on the same ground plane occlude one another, their bounding boxes overlap in a particular common configuration.

Preprocessing Videos: On the time lapse feed from a camera, we run instance segmentation [40] on each frame. We use Intersection-Over-Union based tracker [3] to track the detected bounding box and segmentation. We represent the detections as $D^{t_0,\ldots,t_N}_{m=0,\ldots,M}$, where t_N represents time, while N represents the number of images and m corresponds to the index of the object from a total of M detections.

Occlusion Classification: We locate and segment unoccluded objects in the scene from time lapse video sequences. The unoccluded objects are detected by exploiting overlap between objects detected in an image as shown in Fig 2. For every detection D_i at time instance t_j, we com-

Figure 3. We illustrate generated training images(top) from Clip Art WALT dataset. The synthesized Ground-Truth amodal segmentation map(bottom) captures multiple layers(darker represents higher order of occlusion) of occlusions for training. The Clip Art images have realistic occlusions because the inpainting is performed by superimposing the object at the same location as it was observed but from varying time instances.

pute the occlusion indicator $O(D_i^{t_j})$ using

$$O(D_i^{t_j}) = \begin{cases} 0, & \text{if } D_i^{t_j} \cap D^{t_j} = 0 \; or \; B(D_i^{t_j}) \cap D^{t_j} < \delta \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

We use two hypotheses to classify the detected objects as occluded or fully visible. The first constraint is that the bounding box should not intersect any other detected objects D^{t_j} from the same time instance. Secondly, for every overlapping bounding box, we disentangle the occluded object and the occluder assuming planar constraints. When both objects are on the same plane, we observe that the bottom of the occluded bounding box always intersects with another bounding box from the scene. We exploit this observation and find the intersection of the occluding bounding box with the bottom of the occluded bounding box $B(D_i^{t_j})$. If the intersection is larger than a threshold δ, we classify the object as occluded. This classification is computed iteratively over all the detections $D_{m=0,\ldots,M}^{t_0,\ldots,t_N}$ and unoccluded object detections and segmentations are extracted.

2.2. Clip-Art based Self-Supervision

Once unoccluded objects are discovered, they are composited in layers back into the same scene as shown in Fig 3. These are close enough to real data to reduce the domain gap for a deep network that explicitly predicts the object, its occluder, and the occluded.

Background Computation: Given a sequence of images from a stationary camera, we compute the median image by finding the median RGB value per pixel from a collection of images. Since the camera is captured throughout the day and in different weather computing a single median image is unrealistic. To create realistic background images, we generate median images for varying imaging conditions like time of the day or different weather i.e. sunny, rainy, etc. This is computed by sampling the images under different conditions. We also compute the spatial distribution of the object occurrence for each median image to simulate the occlusion patterns similar to the real-world images.

Generating Layered Representation: We randomly select a background image and its object occurrence data distribution. We sample P unoccluded objects from the data distribution $D_{m=0,\ldots,M}^{t_0,\ldots,t_N}$ where $O(D_i^{t_j}) == 0; i \in P$. These sampled objects and their segmentation masks are segregated into different layers for generating varied occlusions of the scene. We iterate through each layer and composite the objects onto the background image using the segmentation masks. Since they are composited layer-wise onto the image, an amodal segmentation map is automatically generated using the segmentation mask for all the objects in the scene. Since we use longitudinal information (images over a long period of time) to generate these objects the network learns from large variations of objects as well as different occlusion configurations. The composited image and the amodal segmentation map are passed to the network for training the Amodal Representation.

2.3. Watch and Learn Time-lapse Network

We learn the amodal representation of the scene by training a network using the composite image and its amodal segmentation map as shown in Fig 4. The input image is passed through a backbone network [40] to produce feature maps. The feature map produced from the backbone is passed through the box head [55] to produce an amodal bounding box. The amodal bounding box is combined with the feature map to produce the amodal segmentation by learning Object-Occluder-Occluded interaction.

Amodal Bounding Box: The feature map from the backbone is passed through the box head to compute the amodal bounding box hypothesis. We train this box head using

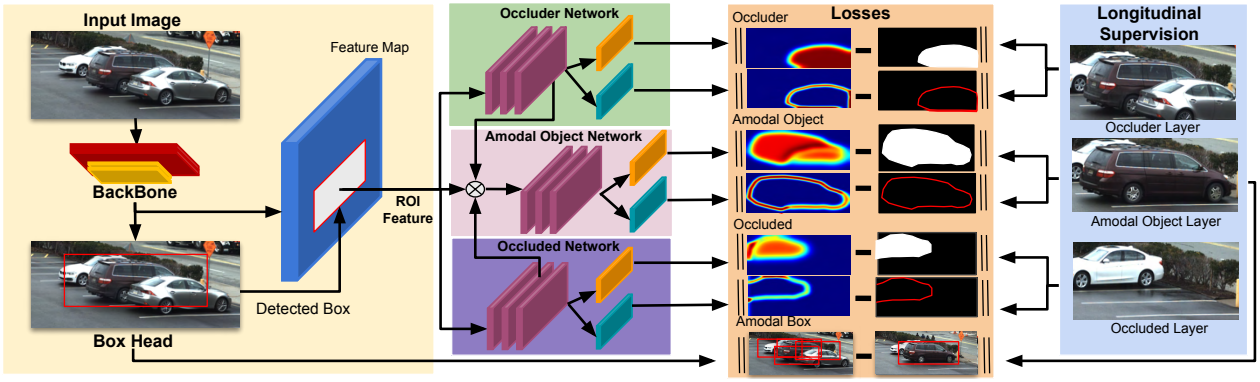

Figure 4. The composite images are passed through our Network to train for amodal representations of the scene. The feature map from the backbone is passed through the box head to produce the amodal bounding box. This bounding box is combined with the feature map from the backbone to produce an ROI feature. The ROI feature is used to train for amodal segmentation. The key to predicting holistic object representation is to understand the occluder and the occluded objects in the amodal bounding box. The features from occluder and occluded are concatenated with the ROI feature to produce accurate amodal segmentation. We supervise this network with a segmentation map generated using Clip-Art based Self-Supervision.

FCOS [55] based losses as:

$$L_{AmodalBox} = L_{Regression} + L_{Centerness} + L_{Class} \quad (2)$$

The ground truth bounding box is computed using the amodal segmentation map obtained by compositing. Bounding box hypotheses are combined with the backbone feature map to learn the amodal segmentation network.

Object-Occluder-Occluded Interaction: We learn the interaction between the object and other layers present in the bounding box. Every amodal bounding box has three components *i.e.* the object we want to detect (amodal object(AO)), object occluding the amodal object (occluder(OR)), objects occluded other than background(occluded(OD)). To learn a holistic representation of the object, the interaction of the object with both the occluder and occluded must be exploited by the learning framework. To train for such interactions we propose using different modules for each of the categories. The occluder network takes as input the ROI features and predicts the occluder layer in the amodal bounding box. The occluded network predicts the occluded layer of the amodal bounding box from the ROI features. The object network predicts the amodal object segmentation by robustness to the occluder and the occluded. We combine the occluder and occluded features with the object features to make the network robust to different occlusions. We use both the boundary and segmentation mask to learn the amodal segmentation. We train the boundary for each component using the loss function L^B:

$$L^B_M = L_{BCE}(W_B F^B_M, GT^B_M) \quad (3)$$

We train the segmentation for each component using the loss function L^S:

$$L^S_M = L_{BCE}(W_S F^S_M, GT^S_M) \quad (4)$$

Here, $M \in [AO, OR, OD]$ denotes different network components, and L_{BCE} denotes binary cross-entropy loss between the Ground-Truth GT and the predicted heatmap. W_S and W_B denote the weights trained for segmentation and boundary respectively. F^S_M and F^S_M are the computed feature map for segmentation and boundary respectively for each M. To make the amodal segmentation robust, we combine the occluder F_{OC}, occluded F_{OD} and input feature maps to produce the amodal object feature map F_{AO}.

End-to-End Parameter Learning: The whole amodal representation framework can be trained in an end-to-end manner defined by a multi-task loss function L as,

$$L = \lambda_b L_{AmodalBox} + L_{AO} + L_{OR} + L_{OD} \quad (5)$$

$$L_{Object} = \lambda^S_{AO} L^S_{AO} + \lambda^B_{AO} L^B_{AO} \quad (6)$$

where, L_{AO},L_{OR},L_{OD} are losses for Amodal object, Occluder and Occluded networks, respectively. As shown in Eq(6), for each layer the loss is a summation of the boundary loss and the segmentation loss. Similar to Eq(6), we compute the boundary and segmentation loss for both the occluder and occluded layers. Finally the network is trained with an end-to-end framework optimizing all the losses.

2.4. Speeding Up Amodal Learning

The accuracy of the amodal representation is affected by the quality and quantity of the unoccluded objects. We speed up the discovery of unoccluded objects by combining sparse time sampling of the data with burst local tracking. This step reduces the required observation period from many months to several days (images captured every few mins.). We discover nearly 3 times more unoccluded objects with different thresholds of detection using this strategy, as shown by the thin transparent lines on the left of Fig

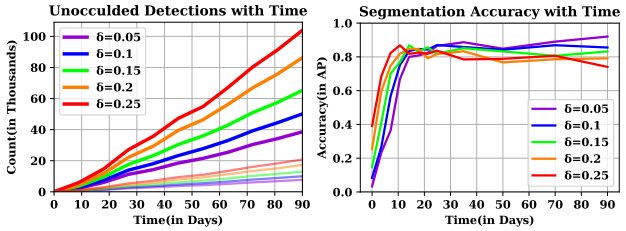

Figure 5. We compare the number of detected unoccluded objects (bold) using our unoccluded tracking framework compared to uniform sampling (transparent) on the left image. Using the new module, achieving high accuracy faster(within 15 days) compared to uniform sampling for nearly all thresholds of γ (right).

5. These additional mined unoccluded objects speed up the training by more than 5 times and plateau in just 14 days of observation as shown in Fig 5 for different thresholds δ. Another important insight is that the network learns faster with higher δ but loses accuracy as the mined unoccluded objects are erroneous. On the other hand, lower δ shows that the network takes longer to learn but gains accuracy with the addition of more samples. We reduce δ linearly with the number of days captured for faster training.

3. Dataset and Metrics

We introduce a new dataset, called WALT, of 12 (4K and 1080p) cameras capturing data over a year in short bursts. Further, we propose a novel evaluation method using stationary objects to improve on the shortcomings of human-annotated or synthetic datasets on real images.

Watch And Learn Time-lapse (WALT) Dataset: The dataset consists of 6 4K resolution cameras setup by us and 6 1080p YouTube public live streams. The cameras overlook public urban settings analyzing the flow of traffic and people with severe occlusions, as shown in Fig 6. We used 4 cameras from our setup and 6 cameras from YouTube for training. Data captured from 2 cameras are used for testing. The data is captured for 3-second bursts at 30 FPS every few minutes. Only the images with notable changes from the previous image are stored. This results in storing approximately 5000 images per day for a year. We will be releasing months of data captured from cameras set up by us and publish a live stream video of the cameras on YouTube for research purposes. The code to automatically capture and process data from YouTube live streams will be released.

Potential Societal Impact: We do not perform any human subject studies from these cameras. To discourage any human subject study and preserve the privacy of the object captured in the images, we blur the faces and license plates in all the images to be released. The data is captured in short bursts around random time instances to discourage identification of movement patterns of particular persons or vehicles. This study is designated as non-human subjects research by our Institutional Review Board (IRB).

Figure 6. Sample visualizations from the WALT(Right) and Rendered WALT(Left) dataset. The dataset contains diverse objects with severe occlusions captured over years. The results show significant performance in amodal representation learning on such large scale real data for the first time.

Rendered WALT Dataset(RWALT): We replicate the WALT Dataset using computer graphics rendering [8]. We use a parking lot 3D model and simulate object trajectories similar to the real-world parking lot. We render 1000 time-lapse images of the scene from multiple viewpoints. The cameras for rendering are placed on the dashboard of the vehicles or on infrastructure around the parking lot. Sample rendered images from the dataset are shown in Fig 6. We use rendering from 100 cameras for training and 20 cameras for testing. We use the dataset to compute the ablation study of the network using Ground-Truth from rendering.

Metrics: We use average precision (AP) for evaluating bounding box and segmentation accuracy throughout our experiments unless specified otherwise. We evaluate our method on three different categories of data generated from the WALT Dataset: the Rendered WALT Dataset (RWALT), Clip Art WALT Dataset (CWALT), and Stationary Objects WALT Dataset (SWALT). For the Rendered WALT Dataset, the amodal representation is computed on the synthetic image and compared to the Ground-Truth silhouette produced from rendering. For Clip Art WALT Dataset, we compute the unoccluded objects for 90 days on the test and train cameras of the WALT Dataset and synthesize 10000 composite images per camera using the method from Sec 2.2. We pass the layered image through the network and compare the results with generated Ground-Truth for test images.

Stationary Object-Based Evaluation (SWALT): Since human annotators can only hallucinate the object extent in the occluded region, their labeling is not reliable. To circumvent this problem, we propose using consistency in stationary object segmentation and detection under occlusions as a metric to quantify the accuracy of the algorithm. From the test set of WALT, we mine unoccluded stationary objects by clustering objects detected at the same location. We use unoccluded bounding box and segmentation of the stationary object as ground truth to compare predictions when the object is occluded by another object at a different time instance. The mean Intersection-over-union (IOU) between the Ground Truth and prediction is computed for the station-

Dataset	Amodal Object(AO)			Occluder(+OR)			Occluded(+OD)		
	B	M	BM	B	M	BM	B	M	BM
RWALT	55.3	60.5	61.4	64.2	65.5	66.3	66.2	67.9	**68.1**
CWALT	62.3	65.5	66.1	70.2	71.2	73.2	73.9	74.2	**75.3**

Table 1. Ablation analysis of the proposed learning architecture on Rendered and CWALT Dataset. Note that each component .i.e Occluder (+OR) and Occluded (+OD) network improves the accuracy of segmentation. Training with Boundary(B) and Segmentation Mask(M) consistently outperforms models trained only with Boundary or Segmentation Mask.

ary object when it is occluded by greater than a threshold of γ. γ is computed as the overlap between the Ground-Truth bounding box and the bounding box of other objects in the scene. Using this strategy, we extracted 536 stationary objects observed over 60k frames for evaluation.

4. Evaluations and Ablation Analysis

The performances of pedestrian and vehicle detection and segmentation improve significantly in all of the cameras. we report performances at different levels of occlusion and show that the performance drops more slowly as occlusion increases, compared to methods that do not use Clip-Art Based self-supervision.

Notations: Modal represents a model trained using visible segmentations or bounding boxes, while Amodal uses our amodal supervision. In Amodal methods, just using the Amodal object network is represented as AO, while adding just occluder network as +OR. +OD is given as a combination of final layers from both occluder and occluded networks. B and M represent boundary and segmentation Mask respectively, while BM represents training jointly.

Occluder and Occluded Networks Analysis: We observe that adding features from the occluder and occluded networks to the amodal object prediction network increases the accuracy of amodal segmentation for the Rendered WALT Dataset and the Clip Art WALT Dataset as shown in Fig 1. We observe robust segmentation accuracy with an increase in occlusion percentage when using the occluder and occluded networks in Fig 7 for both vehicles and people.

Boundary and Mask Prediction Analysis: Segmentation based methods are observed to be better than boundary based methods. We observe that combining the object boundary with segmentation mask consistently improves accuracy on both the Datasets as shown in Tab 1.

Robustness to Occlusions: We evaluate the accuracy of our algorithm with different percentages of occlusions using CWALT Dataset. We use the Ground-Truth segmentation masks from the dataset to group objects based on the percentage of occlusion. Fig 7 shows the accuracy of detection and segmentation on the Clip Art WALT Dataset with different occlusion percentages. Clearly, we observe that the proposed method is very robust to occlusion compared to other methods for both people and vehicles.

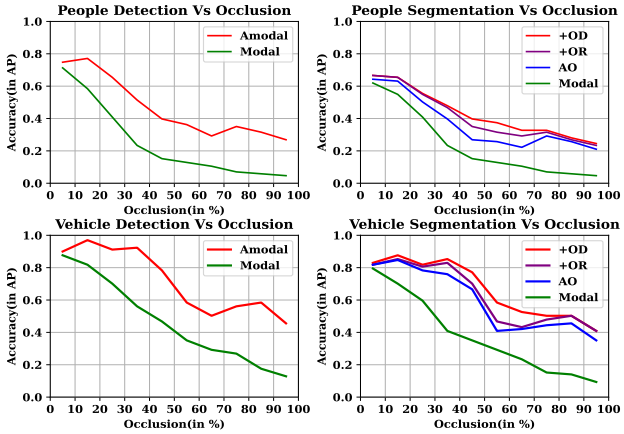

Figure 7. Comparative analysis of Segmentation and Detection accuracy of people and vehicles. Clearly Amodal(Holistic Representation) based methods outperform Modal(only visible representation) based methods in detection and segmentation. Addition of each Network(AO, +OD, +OR) to amodal training improves accuracy of segmentation for severely occluded scenarios. At 50 % occlusion we observe nearly 90 % and 60 % improvement in detection accuracy compared to modal based for people and vehicle respectively. Similarly, at 50 % occlusion we observe 20 % and 12 % improvement in segmentation accuracy compared to Occluder(+OR) for people and vehicles respectively.

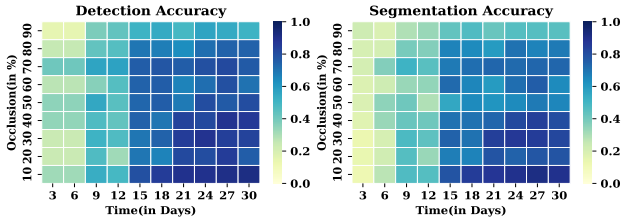

Figure 8. Heatmap of accuracy with different occlusion levels over time on the CWALT Dataset. Observe that the accuracy improves drastically with time for severe occlusions(.i.e >50%) emphasising that our framework learns robust amodal segmentation.

Occlusions Over Time: We analyze the accuracy of amodal representation with respect to training data from different lengths of the Clip Art WILD dataset, in Fig 8. The N-th day plot corresponds to a model trained with N days of unoccluded object detection. We observe from the heatmap that the accuracy increases with time as more unoccluded objects are used to train but decrease with occlusion percentage. We further observe that accuracy improves over time for more severe occlusions, emphasizing that longitudinal learning is important to handle severe occlusions.

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 SWALT based evaluation methodology produces more accurate ground truth. Further, SWALT methodology generates much larger test sets compared to any existing human

Figure 9. Accurate amodal segmentation of vehicles during occlusion while passing each other(Top) or when a vehicle is parking. Our method is able to provide consistent segmentation and detection of all the vehicles in severe occlusions and motions. This can lead to a drastic improvement in tracking objects with occlusions.

Figure 10. Accurate prediction of amodal segmentation of people when a person passes by another(top) or when they walk occluding throughout the video(bottom). Such representation directly extrapolates to improved tracking of people in generic videos.

	KINS	COCOA	SWALT	
			$\gamma = 0.01$	$\gamma = 0.5$
ASN	24.9	29.6	79.4	76.91
BCN	27.3	32.7	82.79	77.44
Ours	**27.9**	**33.1**	**83.6**	**78.2**

(a) Trained on KNIS [47]+COCOA [63]

	CWALT	SWALT	
		$\gamma = 0.01$	$\gamma = 0.5$
ASN	66.1	83.1	81.9
BCN	73.2	89.9	88.3
Ours	**75.3**	**92.19**	**91.7**

(b) Trained on CWALT

Table 2. Amodal Segmentation comparisons trained on Human annotated datasets (a) and Clip-Art WALT Dataset (CWALT) (b) with respect to three different network architectures ASN [47], BCNet [63] and Ours. Tab. 2a shows that Human annotated dataset training only achieves around 78% accuracy on SWALT. On the other hand, Tab.2b reports 91.7% accuracy on SWALT showing the advantage of training on CWALT. In fact, all methods show improvement on SWALT by training on CWALT. γ represents the percentage of occlusion for each object in SWALT but needs further study to report for human-annotated datasets.

annotated datasets (60K images from WALT dataset compared to 6157 images in KINS dataset) and is expected to grow significantly 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. Nonetheless, we report accuracy of our method when trained on Human annotated datasets and tested on KINS, COCOA and SWALT in Tab 2a. Our method slightly outperforms previous methods here.

Comparisons to other Networks: We analyze the advantage of training/testing different methods on our data (CWALT/SWALT). The test scores show improvement in amodal accuracy as compared to other methods. In fact, all methods improve by training on CWALT and testing on SWALT as shown in Tab 2b. We show a qualitative comparison of these methods on multiple real-world images with severe occlusions in Fig 11.

Robust Tracking Using Amodal Representations: We demonstrate that learning robust amodal representation automatically improves tracking of severely occluded objects, as shown in Fig 10 for people and Fig 9 for vehicles. Specifically, observe that the objects are well-segmented and consistent across frames with various levels of occlusions. See supplementary material for more results and videos.

5. Conclusion and Limitations

Limitations: Generalization of the amodal segmentation on new cameras that view significantly different scenes needs to be analyzed. Speeding up learning rate even further needs to be investigated for broader application of our approach.

Conclusion: This work demonstrates that real longitudinal data can be used effectively to self-supervise amodal learning. The key insight is that it is easier to discover unoccluded objects accurately and quickly (over several days) and use them to learn amodal segmentations from any stationary camera observing a scene over time. The confidence of this discovery can be used as a quasi-learning rate to speed up amodal training of occluded objects. We introduce a new dataset, called WALT, of 12 (4K and 1080p) cameras capturing data over a year in short bursts every 5 minutes or so. The data will be released with faces and license plates anonymized to help preserve privacy. The results show significant performance in amodal representation learning on large scale real data for the first time. In the future, we will extend our approach to learn from cameras placed on vehicles for self-driving applications.

Acknowledgments: This work was sponsored in part by an ARL Grant W911QX20F016, NSF CNS-2038612, and DOT RITA Mobility-21 Grant 69A3551747111, and a Qualcomm Innovation Fellowship.

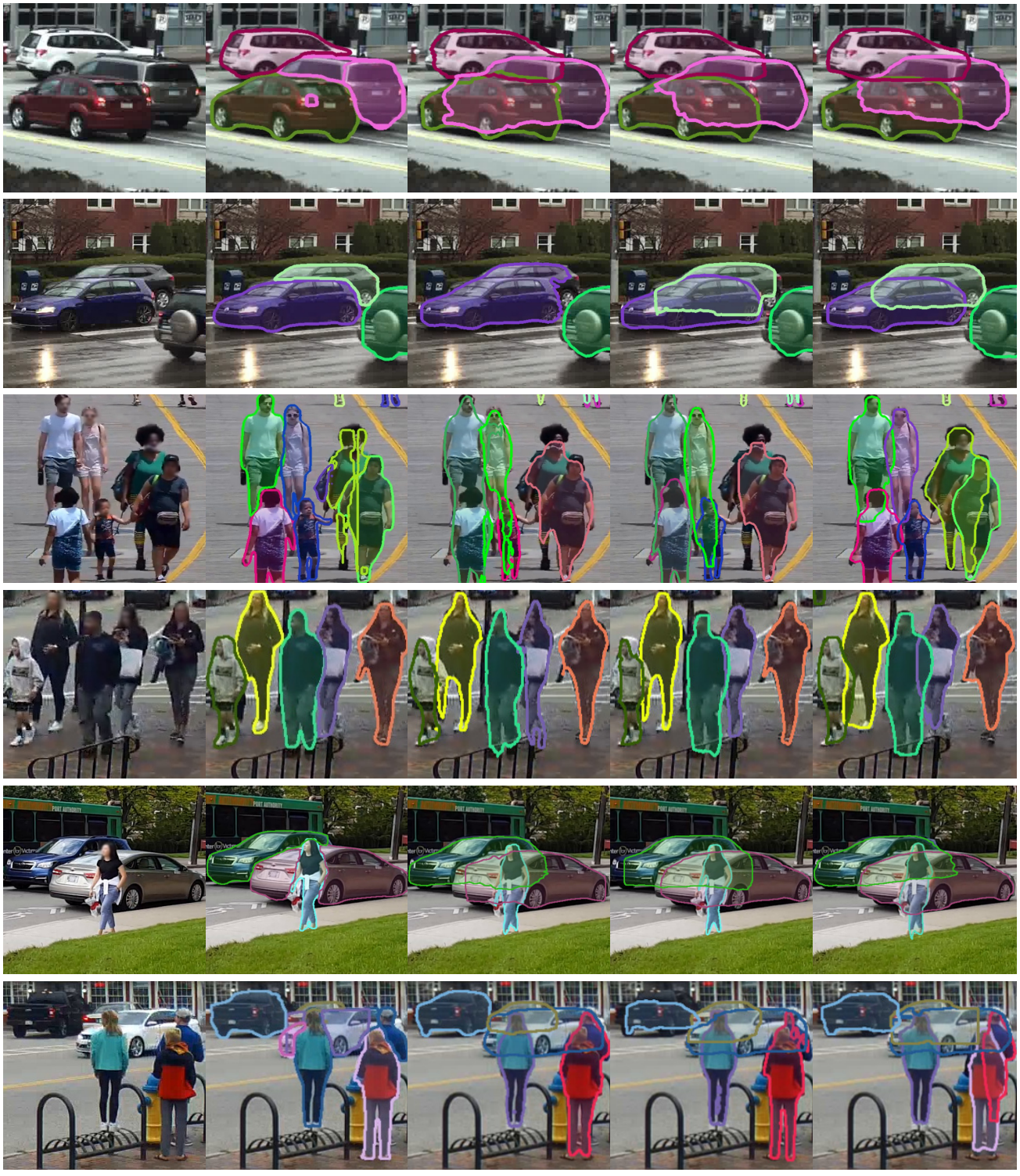

Figure 11. Quantitative results comparing our method to the state-of-the-art images captured from different datasets. The first two rows show vehicles occluding vehicles scenarios while the next two show people occluding people. Finally, we also show examples of people and vehicles occluding each other in the bottom two rows. Observe that our method consistently outperforms other baselines in predicting the amodal segmentation due to longitudinal self-supervision formulation. We perform accurate segmentation in difficult occlusions scenarios like objects having similar colors (Second Row) or large occlusions(Third Row, Sixth Row) or multiple layers of occlusions(First Row, Fifth Row). Our method even works with low-resolution images(Fourth Row) and inter-object interactions(Fifth Row, Sixth Row).

References

[1] Hassan Abu Alhaija, Siva Karthik Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. Augmented reality meets deep learning for car instance segmentation in urban scenes. In *British machine vision conference*, volume 1, page 2, 2017. 1

[2] Georges Baatz, Olivier Saurer, Kevin Köser, and Marc Pollefeys. Large scale visual geo-localization of images in mountainous terrain. In *European conference on computer vision*, pages 517–530. Springer, 2012. 2

[3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, 2016. 2

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1

[5] Yu Cheng, Bo Yang, Bo Wang, and Robby T Tan. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. *arXiv preprint arXiv:2004.11822*, 2020. 1

[6] Yu Cheng, Bo Yang, Bo Wang, Yan Wending, and Robby Tan. Occlusion-aware networks for 3d human pose estimation in video. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 723–732. IEEE. 1

[7] Wongun Choi. Near-online multi-target tracking with aggregated local flow descriptor. In *ICCV*, 2015. 1

[8] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 2, 5

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[10] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005. 1

[11] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi. Segan: Segmenting and generating the invisible. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6144–6153, 2018. 1

[12] Epic Games. Unreal engine. 2

[13] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. Learning to detect and track visible and occluded body joints in a virtual world. In *Proceedings of the European conference on computer vision (ECCV)*, pages 430–446, 2018. 1

[14] Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. *International journal of computer vision*, 61(1):55–79, 2005. 1

[15] Patrick Follmann, Rebecca König, Philipp Härtinger, Michael Klostermann, and Tobias Böttger. Learning to see the invisible: End-to-end trainable amodal instance segmentation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1328–1336. IEEE, 2019. 1

[16] Rik Fransens, Christoph Strecha, and Luc Van Gool. A mean field em-algorithm for coherent occlusion handling in map-estimation prob. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 300–307. IEEE, 2006. 1

[17] Tianshi Gao, Benjamin Packer, and Daphne Koller. A segmentation-aware object detection model with occlusion handling. In *CVPR 2011*, pages 1361–1368. IEEE, 2011. 1

[18] 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

[19] Golnaz Ghiasi, Yi Yang, Deva Ramanan, and Charless C Fowlkes. Parsing occluded people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2401–2408, 2014. 1

[20] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 1

[21] Ruiqi Guo and Derek Hoiem. Beyond the line of sight: labeling the underlying surfaces. In *European Conference on Computer Vision*, pages 761–774. Springer, 2012. 1

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1

[23] Edward Hsiao and Martial Hebert. Occlusion reasoning for object detectionunder arbitrary viewpoint. *IEEE transactions on pattern analysis and machine intelligence*, 36(9):1803–1815, 2014. 1

[24] Y.-T. Hu, H.-S. Chen, K. Hui, J.-B. Huang, and A. G. Schwing. SAIL-VOS: Semantic Amodal Instance Level Video Object Segmentation – A Synthetic Dataset and Baselines. In *Proc. CVPR*, 2019. 1

[25] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7718–7727, 2019. 1

[26] Nathan Jacobs, Nathaniel Roman, and Robert Pless. Consistent temporal variations in many outdoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–6, June 2007. Acceptance rate: 23.4%. 2

[27] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. *CoRR*, abs/1803.07549, 2018. 1

[28] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Amodal completion and size constancy in natural scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 127–135, 2015. 1

[29] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. In *CVPR*, 2015. 1

[30] 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. 1, 8

[31] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Occlusion-aware video object inpainting. In *The IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[32] Adam Kortylewski, Ju He, Qing Liu, and Alan L Yuille. Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8940–8949, 2020. 2

[33] Philipp Krähenbühl. Free supervision from video games. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2955–2964, 2018. 2

[34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*. 1

[35] Jean-François Lalonde, Alexei A Efros, and Srinivasa G Narasimhan. Webcam clip art: Appearance and illuminant transfer from time-lapse sequences. *ACM Transactions on Graphics (TOG)*, 28(5):1–10, 2009. 2

[36] 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. *arXiv preprint arXiv:1612.02699*, 2016. 1

[37] Fangyu Li, N. Dinesh Reddy, Xudong Chen, and Srinivasa G. Narasimhan. Traffic4d: Single view reconstruction of repetitious activity using longitudinal self-supervision. In *Proceedings of IEEE Intelligent Vehicles Symposium (IV '21)*. IEEE, July 2021. 2

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[39] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[40] 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, 3, 8

[41] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 1

[42] Minh Vo N Dinesh Reddy and Srinivasa G. Narasimhan. Carfusion: Combining point tracking and part detection for dynamic 3d reconstruction of vehicle. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2018*. IEEE, June 2018. 1

[43] Srinivasa G Narasimhan, Chi Wang, and Shree K Nayar. All the images of an outdoor scene. In *European conference on computer vision*, pages 148–162. Springer, 2002. 2

[44] Merlin Nimier-David, Delio Vicini, Tizian Zeltner, and Wenzel Jakob. Mitsuba 2: A retargetable forward and inverse renderer. *ACM Transactions on Graphics (TOG)*, 38(6):1–17, 2019. 2

[45] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[46] Bojan Pepikj, Michael Stark, Peter Gehler, and Bernt Schiele. Occlusion patterns for object class detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3286–3293, 2013. 1

[47] 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, 7, 8

[48] N Dinesh Reddy, Laurent Guigues, Leonid Pishchulin, Jayan Eledath, and Srinivasa G. Narasimhan. Tessetrack: End-to-end learnable multi-person articulated 3d pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15190–15200, June 2021. 1

[49] N. Dinesh Reddy, Minh Vo, and Srinivasa G. Narasimhan. Occlusion-net: 2d/3d occluded keypoint localization using graph networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2

[50] Mengye Ren and Richard S Zemel. End-to-end instance segmentation with recurrent attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6656–6664, 2017. 1

[51] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016. 2

[52] Samuel Schulter, Menghua Zhai, Nathan Jacobs, and Manmohan Chandraker. Learning to look around objects for top-view representations of outdoor scenes. In *The European Conference on Computer Vision (ECCV)*, September 2018. 1

[53] Nathan Silberman, Lior Shapira, Ran Gal, and Pushmeet Kohli. A contour completion model for augmenting surface reconstructions. In *European Conference on Computer Vision*, pages 488–503. Springer, 2014. 1

[54] Yihong Sun, Adam Kortylewski, and Alan Yuille. Weakly-supervised amodal instance segmentation with compositional priors. *arXiv preprint arXiv:2010.13175*, 2020. 1

[55] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *Proc. Int. Conf. Computer Vision (ICCV)*, 2019. 3, 4

[56] Joseph Tighe, Marc Niethammer, and Svetlana Lazebnik. Scene parsing with object instances and occlusion ordering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3748–3755, 2014. 1

[57] Angtian Wang, Yihong Sun, Adam Kortylewski, and Alan L Yuille. Robust object detection under occlusion with context-aware compositionalnets. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 12645–12654, 2020. 1, 2

[58] Shaofei Wang and Charless C Fowlkes. Learning optimal parameters for multi-target tracking with contextual interactions. *International Journal of Computer Vision*, 122(3):484–501, 2017. 1

[59] Yu Xiang, Alexandre Alahi, and Silvio Savarese. Learning to track: Online multi-object tracking by decision making. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4705–4713, 2015. 1

[60] Xiaoding Yuan, Adam Kortylewski, Yihong Sun, and Alan Yuille. Robust instance segmentation through reasoning about multi-object occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11141–11150, June 2021. 1

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

[62] Li Zhang, Yuan Li, and Ramakant Nevatia. Global data association for multi-object tracking using network flows. In *CVPR*, 2008. 1

[63] 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. 1, 7