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

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

Input Image | Predicted Segmentation | Predicted Shape | 3D View

Figure 1. Models trained on our automatically generated data from time-lapse imagery can reliably estimate amodal 2D bounding box, segmentation as well as 3D shape and pose despite the complex occlusions presented in the input image.

Abstract

Current methods for 2D and 3D object understanding struggle with severe occlusions in busy urban environments, partly due to the lack of large-scale labeled ground-truth annotations for learning occlusion. In this work, we introduce a novel framework for automatically generating a large, realistic dataset of dynamic objects under occlusions using freely available time-lapse imagery. By leveraging off-the-shelf 2D (bounding box, segmentation, keypoint) and 3D (pose, shape) predictions as pseudo-groundtruth, unoccluded 3D objects are identified automatically and composited into the background in a clip-art style, ensuring realistic appearances and physically accurate occlusion configurations. The resulting clip-art image with pseudo-groundtruth enables efficient training of object reconstruction methods that are robust to occlusions. Our method demonstrates significant improvements in both 2D and 3D reconstruction, particularly in scenarios with heavily occluded objects like vehicles and people in urban scenes.

1. Introduction

In recent years, remarkable progress has been made in advancing scene understanding tasks such as object detection [49, 51, 63], tracking [14, 78, 81], segmentation [10, 31], and 3D reconstruction [35, 44, 70]. These achievements are mainly attributed to the availability of large-scale datasets [13, 15, 23, 50, 77] and architectural innovations [17, 30, 42]. Despite this progress, a notable challenge persists in scenarios with severe *occlusion*, where only a portion of the object is visible: an object may be partially occluded by other objects or truncated by the camera's field-of-view. This phenomenon is referred to as *amodal perception* [36], and it is hard to intuitively infer their complete shapes. Overcoming these challenges is important to advance many smart cities applications as well as robotics applications, where the number of cameras on vehicles and city infrastructure is rapidly increasing [3, 25, 55].

Efforts to learn holistic representations necessitate a substantially annotated and realistic dataset. While recent works such as KINS [58], COCO-Amodal [84], and Ithaca365 [16] have contributed by annotating some amodal ground-truth data, the available annotations remain limited. This scarcity is primarily due to the inherent difficulty in obtaining supervision for amodal representations, as labeling hidden parts of objects is a difficult task for people to accomplish consistently [58, 60, 84]. Moreover, 3D annotations under occlusions, including object shape and pose [7, 43, 74], are even meager due to the difficulty of annotating 3D data, posing challenges for 3D prediction tasks.

Expanding on WALT [62], we utilize time-lapse videos from stationary cameras to synthesize realistic occlusion scenarios by extracting unoccluded objects and composite them back into the background image at their original positions. Unlike WALT which focuses solely on compositing and learning 2D tasks, our approach extends to generating high-quality 3D pseudo-groundtruth data for robust 3D object reconstruction under occlusion. Additionally, our 3D-based compositing method, named **WALT3D**, in

*denotes equal contribution and joint first author

contrast to WALT's simplistic 2D compositing, produces physically accurate occlusion configurations, leading to increased training data efficiency and scalability.

We start with the observation that, although not perfect, existing off-the-shelf methods demonstrate good accuracy in both 2D (segmentation [31], keypoints [67, 68]) and 3D (pose, shape [37, 54, 73]) prediction tasks, especially on unoccluded objects. Thus, they can be used as *"pseudo-groundtruth"* to improve the robustness of existing approaches in occlusion scenarios. Next, we randomly select these unoccluded, non-intersecting 3D objects and put them back into the background image at their original positions (i.e., clip-art style). Specifically, we arrange them based on their distance from the camera, ensuring physically accurate and realistic occlusion configurations. Each such resulting clip-art image is accompanied by amodal bounding box, segmentation masks, and 3D poses and shapes – referred to as *"pseudo-groundtruth"* as these were predicted by off-the-shelf methods on the original unoccluded objects.

Through extensive experiments, we demonstrate the effectiveness of our data in both vehicle and human reconstruction, particularly in scenarios with heavy occlusions. It is important to note that our method does not require any human labeling and hence is easily scalable and serves as an effective method to automatically generate realistic training data for reconstructing dynamic objects under occlusion.

Our notable technical contributions are summarized below:

- We introduce a novel method that automatically generates 2D/3D supervision data from time-lapse imagery with realistic occlusion configurations without human labeling.
- We demonstrate that the utilization of our generated data significantly enhances training efficiency and the accuracy of 2D/3D object reconstruction on real-world data, particularly in scenarios with high occlusion.

2. Related Works

Occlusion Reasoning: Understanding and reasoning occlusions has been extensively studied for decades [21, 22, 65]. Bad predictions due to occlusions are dealt with as noise/outliers in robust estimators. On the other hand, occlusions are explicitly treated as missing parts in model fitting methods [71, 82]. But severe occlusions, such as when a large part of an object is blocked, can result in poor model fitting [28, 85], especially when they attempt to simultaneously estimate the model fit as well as the missing parts.

2D Amodal Representation: Although the effects of occlusion on visual reasoning has been widely studied, estimating the amodal representation (i.e. both the occluded and visible regions) has only been recently explored. Initial attempts [20, 29, 58, 84] use a supervised learning paradigm using small datasets [58, 84] where humans have annotated occlusions to the best of their abilities. Some methods [59, 60, 66] have explored using multiple views to provide accurate supervision for occluded parts but are not scalable due to capture limitations. To expand supervision, several methods synthesize occlusions to varying degrees of realism. But pure CG renderings [2, 18, 19, 32, 38, 47, 80] suffer from a wide domain-gap [41, 65]. While driving datasets [6, 8, 23, 79] provide 2D amodal bounding boxes, annotating additional representations like segmentation and keypoints under occlusion remains challenging. Some methods, like Ghiasi et al. [24], address this by randomly copy-pasting objects onto diverse background images. However, as these methods solely rely on 2D information, they generate unrealistic occlusion configurations leading to poor performance and limited training efficiency.

3D Reconstruction Under Occlusion: Reconstruction under severe occlusion is still in the nascent stages of research. Most algorithms developed focus on self-occluded objects with shape completion from partial observations [12, 56, 83]. On the other hand, shape models fitting for objects only with the visible regions either from images [28, 33, 45, 61, 64] or depth sensors [1, 11, 57, 76] have been explored. Due to the inherent challenges in annotating 3D information, ground-truth data for 3D object understanding is limited, often confined to specific object sets [43, 74], indoor [5], or driving scenarios [6, 23, 67], and struggles to generalize well to novel viewpoints such as stationary traffic cameras. Recent methods addressing pose estimation for vehicles [54, 73] and people [26, 40] have demonstrated robustness in handling occlusion, either implicitly or explicitly. Thus, our dataset generation method can significantly enhance the robustness of these approaches to occlusion.

3. Generating Realistic Supervision Data

An overview of our approach is shown in Figure 2. From a time-lapse video, we identify unoccluded objects and extract their 2D attributes such as segmentation and keypoints. We then use off-the-shelf 3D object reconstruction methods to obtain the pose and shape of these objects, constrained by the camera intrinsics and ground plane. After that, we re-insert these non-intersecting 3D objects into the background image at their original positions in a "clip-art" style, arranged based on their distance from the camera to ensure the occlusion configurations are physically accurate and realistic. Finally, we use the clip-art composited image together with its pseudo-groundtruth supervision data to learn robust 2D/3D object reconstruction under occlusion.

Mining Unoccluded Objects: Given a stream of time-lapse data from a camera, our goal is to mine for unoccluded objects. In this context, "unoccluded" denotes instances where a detected object is not obstructed by any other object or truncated out of the field-of-view. On the time-lapse feed from a camera, we run instance segmentation [51] on each frame and use a simple object tracker [75] to track the detected bounding box and segmentation that belongs to "per-

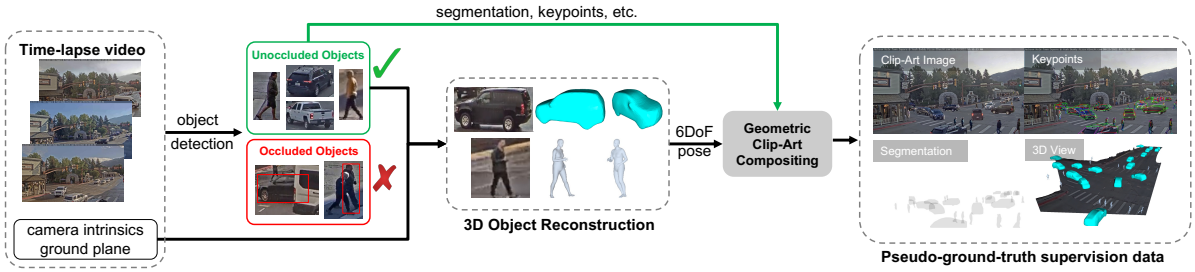


Figure 2. Given a time-lapse video, we automatically generate 2D/3D training data under severe occlusions. We start by detecting each object in the video, and unoccluded (fully visible) objects are identified. Each unoccluded object is then reconstructed using the ground plane and camera parameters. With the 3D pose, unoccluded objects are composited back into the same location (i.e., clip-art style) in a geometrically consistent approach. The composited image and its pseudo-groundtruth from off-the-shelf methods (e.g., segmentation, keypoints, shapes) are utilized to train a model that can produce accurate 2D/3D object reconstruction under severe occlusions.

son" and "car" classes. The simplest heuristic for detecting unoccluded objects involves calculating the Intersection over Union (IOU) of detected objects. In previous methods like [62], objects were considered unoccluded if the intersection of their bounding box bottom with any other box was smaller than a threshold δ. However, this strategy becomes less reliable in scenarios where dynamic objects are occluded by static elements, such as vehicles being occluded by buildings, trees, poles, etc.

To address this limitation, we train a simple Occlusion Classifier (OC) to categorize each detected object as either unoccluded or occluded, which is trained using a supervised approach with human annotators who label objects as unoccluded/occluded. Despite the OC module outperforming the heuristic filter in classifying objects' occlusion status, we did not observe a significant improvement in downstream evaluations (less than ± 0.5 AP) due to the small amount of outliers overall in the training data. Therefore, it is not a mandatory component in our pipeline: the simple heuristics proposed in WALT [62] can be employed for the unoccluded object detection task without additional training. Nevertheless, we believe this data can be useful for occlusion reasoning tasks and we will publicly release it.

Reconstructing Unoccluded Objects: Once unoccluded objects are identified, we describe the 3D reconstruction process of two primary classes: vehicles and humans.

Each mined unoccluded vehicle is reconstructed following Li et al. [48]. We parameterize each vehicle's 3D keypoints $\mathbf{X}$ by a linear combination of the mean shape $\bar{\mathbf{Q}}$ and K principal components $\mathbf{Q}_1, \ldots, \mathbf{Q}_K$ computed from an object CAD model dataset [47]: $\mathbf{X} = \bar{\mathbf{Q}} + \sum_{k=1}^{K} \alpha_k \mathbf{Q}_k$, where α_k is the shape coefficient that needs to be optimized. Starting from the mean shape $\bar{\mathbf{Q}}$, we first detect the 2D keypoints for each unoccluded vehicle and initialize the 6-DoF poses using EPnP [46]. Subsequently, for each track of detected vehicles, we optimize for the 6-DoF object poses while regularizing the shape parameter α_k to be constant for the same vehicle in different frames by minimizing reprojection errors. Additionally, we also constrain the objects to lie on the constant global ground plane, ensuring physically plausible reconstructions.

In the context of human reconstruction, we employ the Human Mesh Recovery (HMR) method HMR 2.0 from Goel et al. [26]. Using HMR 2.0, we predict the SMPL pose and shape parameters for each detected unoccluded human [52]. To determine camera translation, we find the intersection between the backprojected ray from the bottom of the 2D bounding box with the ground plane. Given the camera's intrinsic matrix $\mathbf{K}$, ground plane equation $\begin{bmatrix}\mathbf{n} & d\end{bmatrix}$, and $\mathbf{p}_b$ as the pixel coordinate of the bottom of the box, its depth is computed as $z_b = -\frac{d}{\mathbf{n}^T(\mathbf{K}^{-1}\mathbf{p}_b)}$. For camera intrinsic parameters and ground plane equation, we employ a recent method by Vuong et al. [72] that utilizes Google Street View [27] to automatically calibrate stationary traffic cameras and infer the ground plane. The mined unoccluded object image with its corresponding segmentation mask, keypoint locations, and 3D poses are used in a clip-art based framework to generate 2D and 3D supervision data in severe occlusions as illustrated in Fig. 2.

Pseudo-groundtruth Data Generation: To obtain supervision for occluded objects, we use image-based compositing to generate composited images, serving as input for training a network to learn 2D/3D reconstruction in severe occlusions. During the compositing process, various 2D/3D representations for each unoccluded object, including keypoints, segmentation, pose, and shape obtained from off-the-shelf methods, serve as pseudo-groundtruth supervision signals even after compositing. By using the clip-art-based compositing approach described below, we automatically generate a large number of realistic supervision signals in severe occlusions. Leveraging the 3D poses of mined unoccluded objects, we geometrically compose the object's image in a clip-art manner. Non-intersecting 3D objects are randomly sampled and reinserted into the background image (i.e., median image) at their original positions, ranging from the farthest to the closest. This approach ensures the creation of physically accurate and realistic occlusion configurations (see Fig. 3). Crucially, our geometry-based

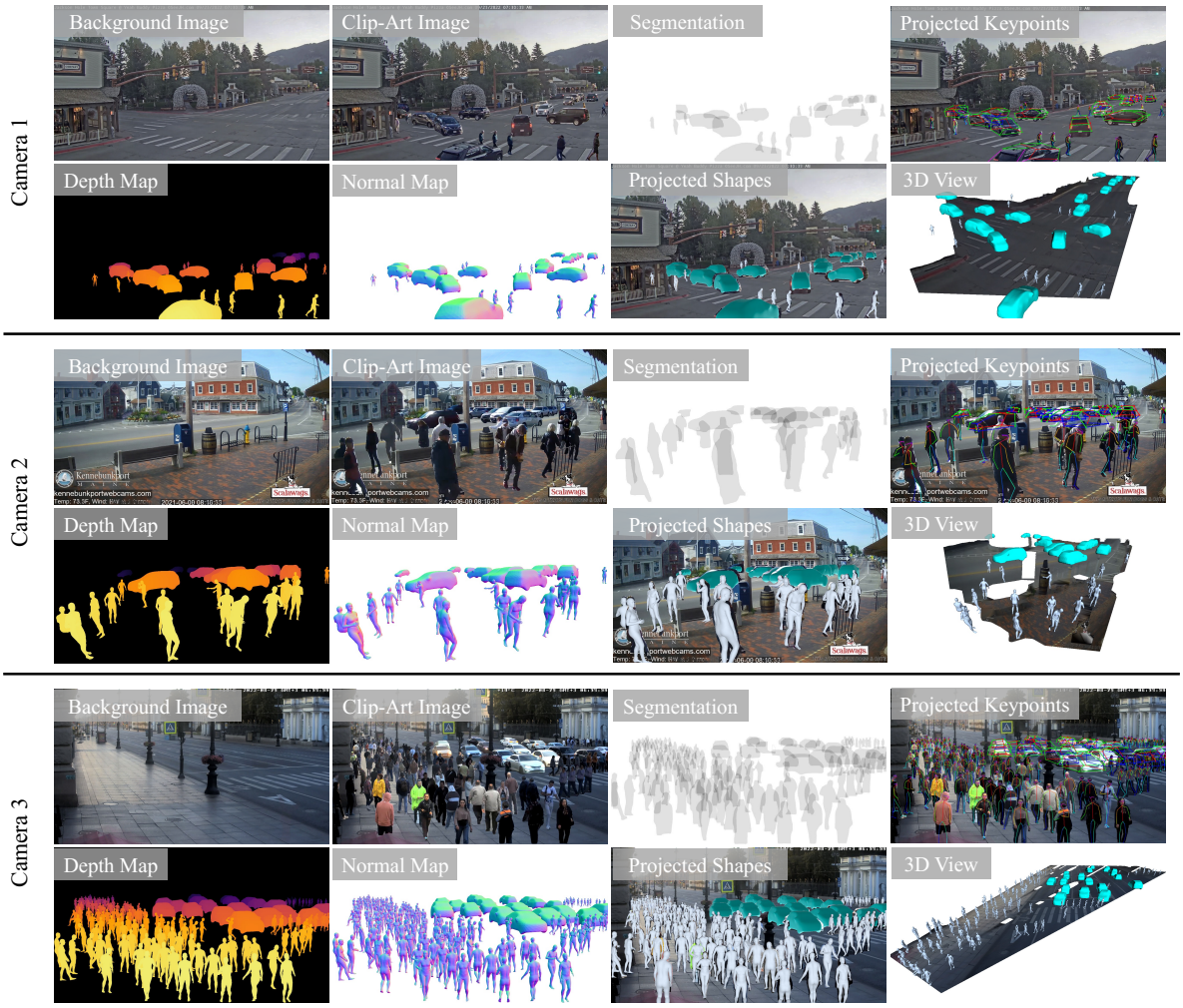


Figure 3. **Automatically generated 2D and 3D Clip-Art to supervise our network:** Unoccluded objects are first mined using time-lapse imagery of WALT dataset [62]. Non-intersecting unoccluded objects are composited back into the background image in their respective original positions to preserve correct appearances. The resulting clip-art images, along with their corresponding amodal pseudo-groundtruth information, such as segmentation, keypoints, depth/normal maps, and 3D shapes, are shown. Our method generates realistic appearances from any stationary camera, incorporating diverse viewing geometries, weather conditions, lighting, and occlusion configurations.

approach distinguishes itself from the mere compositing of 2D images such as WALT [62], which can result in unrealistic composition (see the comparison in Fig. 4). Our data with physically accurate and realistic occlusion configurations contributes to more efficient and scalable training, as demonstrated in our experimental results. Each such generated clip-art image is accompanied by amodal segmentation masks, keypoints, 3D poses and shapes, provided by off-the-shelf methods that serve as pseudo-groundtruth supervision to learn object reconstruction under occlusion.

4. Learning Reconstruction under Occlusion

We have generated an extensive clip-art image dataset with corresponding 2D/3D pseudo-groundtruth representations (see Fig. 3). Using this supervision signal, we train a model capable of inferring holistic object representations in the presence of severe occlusions.

Base Network and 2D Amodal Representations: Using the Swin Transformer [51] backbone with MaskRCNN-based [31] detection heads for its simplicity, our goal is to show improvement irrespective of the base model used. We train the default network with a multi-task loss $\mathcal{L}_{\texttt{2D}}$ for each representation. These losses, computed on each sampled Region of Interest (RoI), include label classification loss $\mathcal{L}_{\texttt{cls}}$, bounding-box loss $\mathcal{L}_{\texttt{box}}$, binary cross-entropy loss on the mask branch $\mathcal{L}_{\texttt{mask}}$ and keypoints $\mathcal{L}_{\texttt{kp}}$: $\mathcal{L}_{\texttt{2D}} = \mathcal{L}_{\texttt{cls}} + \mathcal{L}_{\texttt{box}} + \mathcal{L}_{\texttt{mask}} + \mathcal{L}_{\texttt{kp}}$. Crucially, with our clip-art dataset including *amodal* 2D bounding box, segmentation, and keypoints, our method effectively learns *amodal* instance segmentation and keypoint detection.

3D Object Reconstruction: To regress object 3D pose and shape under occlusion, we extend the base network with a

Composited using WALT

Composited using our method

Figure 4. Comparison between images composited using the 2D-based method WALT2D [62] (left) and our 3D-based method WALT3D (right). It is evident that our 3D-based compositing method generates realistic and geometrically accurate occlusion configurations, in contrast to the 2D-based method (e.g., cars and people overlapping in an unfeasible way).

3D regression branch. Each vehicle N semantic 3D keypoints $\mathbf{X}$ are represented by a linear combination of the mean shape $\bar{\mathbf{Q}}$ and K principal components $\mathbf{Q}_{1\ldots K}$, where α_k is the shape coefficients: $\mathbf{X} = \bar{\mathbf{Q}} + \sum_{k=1}^{K} \alpha_k \mathbf{Q}_k$. Thus, we want to regress for its 6-DoF pose $\{(\mathbf{R}, \mathbf{t}) \in \mathbf{SE}(3)\}$ and shape coefficients α_k. Given the predicted 2D keypoints and the corresponding 3D keypoints $\mathbf{X}$, we can obtain its 6-DoF pose through a differentiable PnP layer [9]. Defining $\pi(\cdot)$ as the projection function with known intrinsic $\mathbf{K}$, we can optimize for the keypoint reprojection loss $\mathcal{L}_{\texttt{kp2D}}$, shape loss $\mathcal{L}_{\texttt{shape}}$, and pose loss $\mathcal{L}_{\texttt{pose}}$ as:

Figure 5. Sample images from our new vehicle 2D keypoints dataset. The dataset contains a wide range of appearance variations including day and night and various traffic scenarios.

DoF vehicle pose during compositing, we require high-quality 2D vehicle keypoints. Existing datasets [47, 59, 67, 77] offer human-annotated 2D vehicle keypoints but mainly focus on driving scenes or have limited training examples, lacking necessary appearance diversity for novel viewpoints like our stationary traffic cameras. Thus, we propose a new dataset with 7,018 images and 42,547 annotated instances from diverse viewpoints (see Fig. 5), each keypoint containing 2D location and occlusion status. With our dataset, we observe significant improvement in keypoint localization accuracy (66.41% to 80.12% on PCK@0.1), and consequently, an improvement in pseudo-groundtruth data quality. Further details are in the Supplementary.

6. Baselines, Metrics, and Evaluations

6.1. Baselines

Detection and Instance Segmentation: All baselines share the same network architecture, employing a Swin [51] backbone as detailed in Section 4. The **SWIN** baseline is initially trained on the COCO dataset [50]. Subsequently, we perform further finetuning on the **WALT2D** and **WALT3D (Ours)** datasets, as detailed in Section 5.

Vehicle 3D Reconstruction: We compare our approach with **Occ-Net** [62] and **3DRCNN** [45] by changing their feature extraction backbone to be as close as possible to ours for fair comparison. They are trained on a combination of PASCAL3D+ [77], KITTI-3D [47], Carfusion [59], and ApolloCar3D [67], where we further finetune our baselines on the **WALT2D** and our **WALT3D** dataset, respectively.

Human 3D Reconstruction: Our architecture mirrors HMR 2.0 [26] as described in Section 4. We initialize from the pretrained model, then finetune on our **WALT3D** data.

6.2. Evaluation Metrics

2D Metrics: We report Average Precision (AP) on bounding box detection (AP^{box}) and instance segmentation (AP^{mask}) following [50]. For keypoints, we use Percentage of Correct Keypoints (PCK) metric where a keypoint

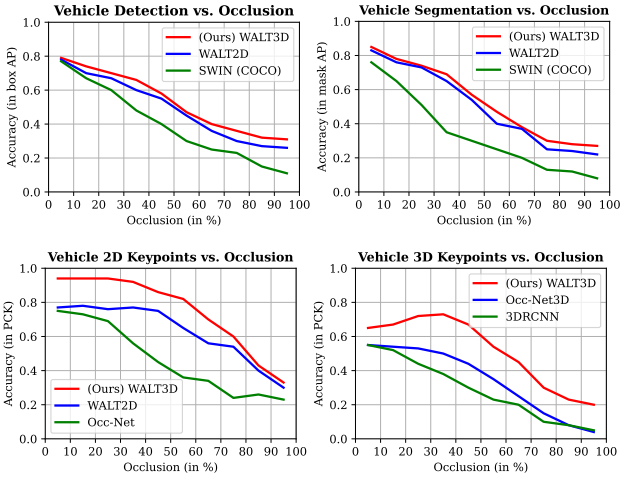


Figure 6. We show the accuracy of our method with respect to increasing percentage of occlusion on multiple tasks like amodal vehicle detection, segmentation, 2D and 3D keypoint estimation. Observe that our method consistently performs better than other baselines showing robustness to increasing occlusion percentage.

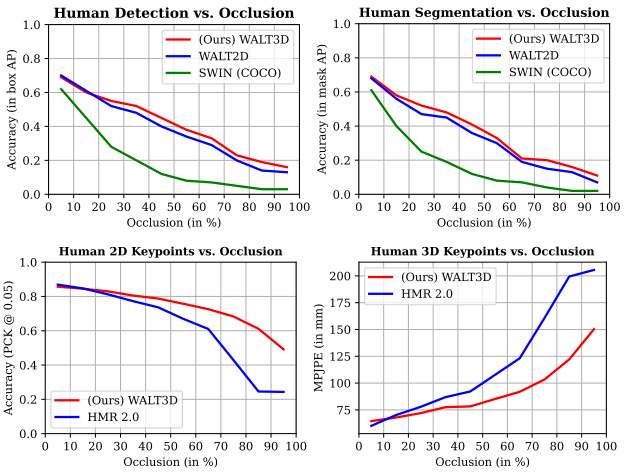


Figure 7. We show the accuracy of our method with respect to increasing percentage of occlusion on multiple tasks like amodal human detection, segmentation, and human mesh recovery (HMR). Observe that our method consistently performs better than other baselines showing robustness to increasing occlusion percentage.

is considered correct if it lies within the radius α (with $0 < \alpha < 1$) of the ground-truth keypoint.

3D Metrics: We use the standard metrics of previous work [34], reporting the Mean Per Joint Position Error (MPJME) between predicted and ground-truth 3D keypoints (aligned using the root joint) as well as 3D PCK.

6.3. Ablation Analysis

Robustness to Occlusions: We evaluate the effectiveness of our algorithm with varying occlusion levels. Similar to WALT [62], we use the pseudo-groundtruth segmentation masks from the evaluation set to group objects based on occlusion percentage. In 2D tasks such as detection and

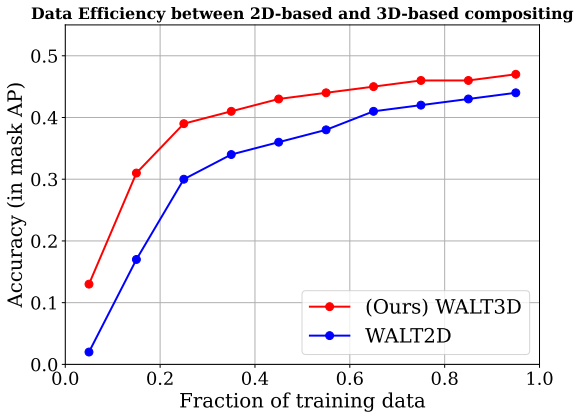
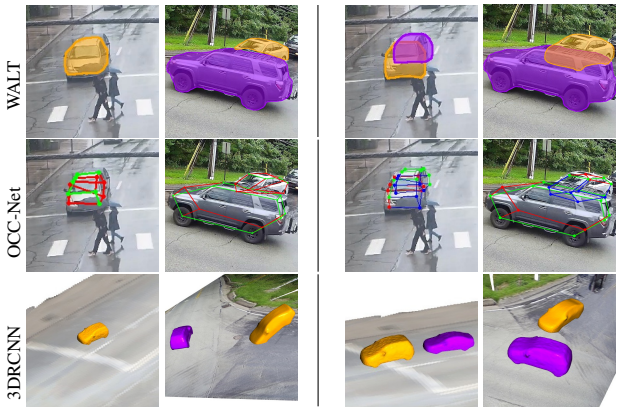


Figure 8. Comparison between WALT2D [62]) and our WALT3D approach. Since our method produces higher quality training data with more realistic occlusion configuration, our approach is especially useful in low-data regime.



Previous SOTA Ours

Figure 9. Qualitative results showing that our data improves amodal segmentation, keypoint prediction and 3D reconstruction compared to previous SOTA.

segmentation, the model trained with our **WALT3D** data significantly outperforms the baseline **SWIN** model trained on COCO [50], which only includes modal masks. This improvement is particularly evident in high occlusion percentages for both vehicle (Fig. 6) and human (Fig. 7). Compared to **WALT2D** [62], our data demonstrates slightly better accuracy across all occlusion levels, thanks to our 3D-based compositing method **WALT3D** producing physically accurate occlusion configurations with the same amount of training data. For 3D tasks such as vehicle (Fig. 6) and human (Fig. 7) 3D pose estimation, using our **WALT3D** data consistently outperforms other baselines, particularly in high occlusion percentages. Thus, models trained with our **WALT3D** data, even with imperfect pseudo-groundtruth from off-the-shelf methods, are robust to occlusion.

3D-based Compositing helps Data Efficiency: Fig. 8 shows that our **WALT3D** data significantly improves performance across all training data fractions, including at low data regime (at 15% of training data with +13.9 AP^{mask} improvement compared to training with **WALT2D** data). Essentially, a model trained on 25% of **WALT3D** data achieves comparable AP to one trained on 60% of **WALT2D** data. This improvement stems from our **WALT3D** approach generating higher quality and physically accurate occlusion configurations with the same training data volume, aligning more closely with the real-world distribution as in Fig. 4. Importantly, this data efficiency enables scaling to a larger number of scenes more efficiently.

Cross-evaluations on other datasets: Table 1 presents quantitative results on COCO-Amodal [84] (COCO-A) (natural images) and KINS [58] (driving images), each containing human-annotated ground-truth 2D amodal instance masks. It shows that additional fine-tuning with our **WALT3D** data improves results in both KINS and COCO-A, highlighting the cross-evaluation performance of our data, with potential for even better generalization with the increasing amount of automatically generated data.

Training Data \ Testing Data	Box AP		Mask AP	
	KINS	COCO-A	KINS	COCO-A
COCO	27.8	38.0	20.8	31.2
COCO + WALT2D	28.2	40.4	22.4	36.4
COCO + WALT3D	**28.8**	**43.2**	**23.5**	**38.7**
COCO + COCO-A + KINS	36.8	46.9	32.9	42.7
COCO + COCO-A + KINS + WALT2D	37.4	47.9	33.3	43.5
COCO + COCO-A + KINS + WALT3D	**38.8**	**48.7**	**35.6**	**45.3**

Table 1. Finetuning with our **WALT3D** data improves generalization across KINS and COCO-A (vehicle & people). Additionally, combining our data with domain-specific data like KINS and COCO-A further enhances performance in their domains.

Cross-evaluations on pedestrian tracking: We evaluate how our **WALT3D** data helps pedestrian tracking in Table 2. To highlight how better detections lead to better tracking, we use a simple tracker SORT [4] with two detectors: one pretrained on COCO and one finetuned with our **WALT3D** data. On MOT17-train, the model finetuned with **WALT3D** data produces high-quality detections under strong occlusions, resulting in improvements over the baseline in all metrics, particularly those favoring better detection quality like MOTA and DetA (see metrics details in [53]).

Detector Train Data	MOTA↑	DetA↑	IDF1↑	HOTA↑
COCO	42.2	40.3	46.7	40.1
COCO + WALT3D	**47.2**	**43.2**	**49.5**	**42.0**

Table 2. MOT17-train results using SORT with two detectors: one pretrained on COCO and one fine-tuned with our WALT3D data.

Qualitative Results: Results of our method can be seen in Fig. 9 and Fig. 10. Comparing to previous methods, observe that our method can produce robust 2D/3D predictions, even under challenging high-occlusion scenarios.

7. Discussion and Conclusion

Our data generation framework is method-agnostic, accommodating various robust object pose estimation alterna-

Figure 10. We show qualitative results of our method on real data. Our method produces accurate amodal segmentation, keypoints, as well as 3D poses and shapes across diverse poses and occlusion configurations. In particular, we show results on different level and types of occlusions like vehicle-vehicle **(row 1, 2, 3)**, vehicle-people **(row 3)**, and people-people occlusion **(row 4, 5)**.

tives [54, 73] as well as methods like VIBE [39] for high-quality human reconstructions from videos. Moreover, as our pseudo-groundtruth data includes metric-scale depth information using ground plane, we can augment existing datasets like Relative Human [69] to enhance the robustness of 3D multi-human reconstruction methods to occlusion.

Limitations: Further research is required for generalization across diverse views for it to be used as a generic solution. It also assumes a mean shape or a parametric object model is available, posing challenges for rare objects. Addressing appearance inconsistencies (e.g., variations in lighting) in clip-art images is a promising research direction.

In conclusion, our method introduces an automated approach to generate a realistic dataset for reconstructing dynamic objects under occlusions from time-lapse imagery. It demonstrates significant improvements in both 2D and 3D object reconstruction, particularly in busy urban scenes with diverse occlusion configurations.

Potential Societal Impact. We do not perform any human subjects research from these cameras. For privacy, we blur faces and license plates in all images intended for release.

Acknowledgements: This work was supported in part by an NSF Grant CNS-2038612, a US DOT grant 69A3551747111 through the Mobility21 UTC and grants 69A3552344811 and 69A3552348316 through the Safety21 UTC.

References

[1] William Agnew, Christopher Xie, Aaron Walsman, Octavian Murad, Yubo Wang, Pedro Domingos, and Siddhartha Srinivasa. Amodal 3d reconstruction for robotic manipulation via stability and connectivity. In *CoRL*, 2021. 2

[2] 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 *BMVC*, 2017. 2

[3] Asra Aslam. Detecting objects in less response time for processing multimedia events in smart cities. In *CVPR*, 2022. 1

[4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*, pages 3464–3468, 2016. 7

[5] 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 *CVPR*, 2023. 2

[6] 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 *CVPR*, 2020. 2

[7] Berk Calli, Arjun Singh, James Bruce, Aaron Walsman, Kurt Konolige, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. Yale-cmu-berkeley dataset for robotic manipulation research. *IJRR*, 2017. 1

[8] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *CVPR*, 2019. 2

[9] Bo Chen, Alvaro Parra, Jiewei Cao, Nan Li, and Tat-Jun Chin. End-to-end learnable geometric vision by backpropagating pnp optimization. In *CVPR*, 2020. 5

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

[11] Falak Chhaya, Dinesh Reddy, Sarthak Upadhyay, Visesh Chari, M. Zeeshan Zia, and K. Madhava Krishna. Monocular reconstruction of vehicles: Combining slam with shape priors. In *ICRA*, 2016. 2

[12] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *ECCV*, 2016. 2

[13] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, 2023. 1

[14] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 1

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 1

[16] Carlos A Diaz-Ruiz, Youya Xia, Yurong You, Jose Nino, Junan Chen, Josephine Monica, Xiangyu Chen, Katie Luo, Yan Wang, Marc Emond, et al. Ithaca365: Dataset and driving perception under repeated and challenging weather conditions. In *CVPR*, 2022. 1

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[18] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi. Segan: Segmenting and generating the invisible. In *CVPR*, 2018. 2

[19] 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 *ECCV*, 2018. 2

[20] Patrick Follmann, Rebecca König, Philipp Härtinger, Michael Klostermann, and Tobias Böttger. Learning to see the invisible: End-to-end trainable amodal instance segmentation. In *WACV*, 2019. 2

[21] Rik Fransens, Christoph Strecha, and Luc Van Gool. A mean field em-algorithm for coherent occlusion handling in map-estimation prob. In *CVPR*, 2006. 2

[22] Tianshi Gao, Benjamin Packer, and Daphne Koller. A segmentation-aware object detection model with occlusion handling. In *CVPR*, 2011. 2

[23] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1, 2

[24] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR*, 2021. 2

[25] Holger Glasl, David Schreiber, Nikolaus Viertl, Stephan Veigl, and Gustavo Fernandez. Video based traffic congestion prediction on an embedded system. In *ITSC*, 2008. 1

[26] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa*, and Jitendra Malik*. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 2, 3, 5, 6

[27] Google. Google Street View. https://www.google.com/streetview/. 3

[28] Can Gümeli, Angela Dai, and Matthias Nießner. Roca: robust cad model retrieval and alignment from a single image. In *CVPR*, 2022. 2

[29] Ruiqi Guo and Derek Hoiem. Beyond the line of sight: labeling the underlying surfaces. In *ECCV*, 2012. 2

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[31] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 2, 4

[32] 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 *CVPR*, 2019. 2

[33] Vladislav Ishimtsev, Alexey Bokhovkin, Alexey Artemov, Savva Ignatyev, Matthias Niessner, Denis Zorin, and Evgeny Burnaev. Cad-deform: Deformable fitting of cad models to 3d scans. In *ECCV*, 2020. 2

[34] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 6

[35] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 1

[36] Gaetano Kanizsa. Organization in vision: Essays on gestalt perception. *Praeger Publishers*, 1979. 1

[37] Lei Ke, Shichao Li, Yanan Sun, Yu-Wing Tai, and Chi-Keung Tang. Gsnet: Joint vehicle pose and shape reconstruction with geometrical and scene-aware supervision. In *ECCV*, 2020. 2

[38] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *CVPR*, 2021. 2

[39] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 8

[40] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, 2021. 2

[41] Philipp Krähenbühl. Free supervision from video games. In *CVPR*, 2018. 2

[42] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[43] Alexander Krull, Eric Brachmann, Frank Michel, Michael Ying Yang, Stefan Gumhold, and Carsten Rother. Learning analysis-by-synthesis for 6d pose estimation in rgb-d images. In *ICCV*, 2015. 1, 2

[44] Nilesh Kulkarni, Abhinav Gupta, David F Fouhey, and Shubham Tulsiani. Articulation-aware canonical surface mapping. In *CVPR*, 2020. 1

[45] Abhijit Kundu, Yin Li, and James M. Rehg. 3d-rcnn: Instance-level 3d object reconstruction via render-and-compare. In *CVPR*, 2018. 2, 6

[46] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o (n) solution to the pnp problem. *IJCV*, 2009. 3

[47] 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. 2, 3, 6

[48] Fangyu Li, N Dinesh Reddy, Xudong Chen, and Srinivasa G Narasimhan. Traffic4d: Single view longitudinal 4d reconstruction of repetitious activity using self-supervised experts. In *IV*, 2021. 3

[49] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, 2022. 1

[50] 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 *ECCV*, 2014. 1, 6, 7

[51] 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 *ICCV*, 2021. 1, 2, 4, 6

[52] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 2015. 3, 5

[53] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *IJCV*, 2021. 7

[54] 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. 2, 8

[55] M. Naphade, S. Wang, D. C. Anastasiu, Z. Tang, M. Chang, Y. Yao, L. Zheng, M. Shaiqur Rahman, A. Venkatachalapathy, A. Sharma, Q. Feng, V. Ablavsky, S. Sclaroff, P. Chakraborty, A. Li, S. Li, and R. Chellappa. The 6th ai city challenge. In *CVPRW*, 2022. 1

[56] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 2

[57] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 2017. 2

[58] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *CVPR*, 2019. 1, 2, 7

[59] 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. 2, 6

[60] N. Dinesh Reddy, Minh Vo, and Srinivasa G. Narasimhan. Occlusion-net: 2d/3d occluded keypoint localization using graph networks. In *CVPR*, 2019. 1, 2

[61] 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, 2021. 2

[62] 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, 5, 6, 7

[63] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1

[64] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *CVPR*, 2017. 2

[65] Samuel Schulter, Menghua Zhai, Nathan Jacobs, and Manmohan Chandraker. Learning to look around objects for top-view representations of outdoor scenes. In *ECCV*, 2018. 2

[66] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 2

[67] 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. 2, 6

[68] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 2

[69] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J. Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, 2022. 8

[70] Shubham Tulsiani and Jitendra Malik. Viewpoints and keypoints. In *CVPR*, 2015. 1

[71] Andrea Vedaldi and Andrew Zisserman. Structured output regression for detection with partial truncation. In *NeurIPS*, 2009. 2

[72] Khiem Vuong, Robert Tamburo, and Srinivasa G. Narasimhan. Toward planet-wide traffic camera calibration. In *WACV*, 2024. 3

[73] Angtian Wang, Adam Kortylewski, and Alan Yuille. Nemo: Neural mesh models of contrastive features for robust 3d pose estimation. In *ICLR*, 2021. 2, 8

[74] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *CVPR*, 2019. 1, 2

[75] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017. 2

[76] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *CVPR*, 2019. 2

[77] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, 2014. 1, 6

[78] Yu Xiang, Alexandre Alahi, and Silvio Savarese. Learning to track: Online multi-object tracking by decision making. In *ICCV*, 2015. 1

[79] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 2

[80] Xiaoding Yuan, Adam Kortylewski, Yihong Sun, and Alan Yuille. Robust instance segmentation through reasoning about multi-object occlusion. In *CVPR*, 2021. 2

[81] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, 2022. 1

[82] Xiaowei Zhou, Spyridon Leonardos, Xiaoyan Hu, and Kostas Daniilidis. 3d shape estimation from 2d landmarks: A convex relaxation approach. In *CVPR*, 2015. 2

[83] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *CVPR*, 2018. 2

[84] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *CVPR*, 2017. 1, 2, 7

[85] M Zeeshan Zia, Michael Stark, and Konrad Schindler. Towards scene understanding with detailed 3d object representations. *IJCV*, 2015. 2