Instance-wise Occlusion and Depth Orders in Natural Scenes

Hyunmin Lee
LG AI Research
hyunmin@lgresearch.ai

Jaesik Park
POSTECH GSAI & CSE
jaesik.park@postech.ac.kr

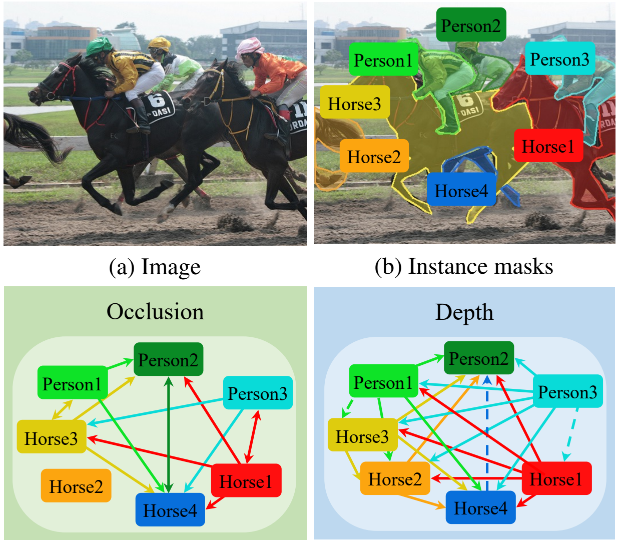
Abstract

*In this paper, we introduce a new dataset, named **InstaOrder**, that can be used to understand the geometrical relationships of instances in an image. The dataset consists of 2.9M annotations of geometric orderings for class-labeled instances in 101K natural scenes. The scenes were annotated by 3,659 crowd-workers regarding (1) occlusion order that identifies occluder/occludee and (2) depth order that describes ordinal relations that consider relative distance from the camera. The dataset provides joint annotation of two kinds of orderings for the same instances, and we discover that the occlusion order and depth order are complementary. We also introduce a geometric order prediction network called **InstaOrderNet**, which is superior to state-of-the-art approaches. Moreover, we propose a dense depth prediction network called **InstaDepthNet** that uses auxiliary geometric order loss to boost the accuracy of the state-of-the-art depth prediction approach, MiDaS [54].*

1. Introduction

Understanding a scene from an image is a fundamental problem in computer vision. Deep learning-based approaches have achieved great success in various tasks, such as object detection [2, 13, 17, 18, 23, 37, 55, 56], semantic segmentation [5, 6, 36, 42, 43, 48, 64, 65, 74, 81], instance segmentation [11, 12, 22, 32, 41, 50, 51, 80] and depth estimation [4, 14, 20, 29, 33, 39, 58, 61, 70, 78]. More recently, approaches have inferred high-level information, such as amodal segmentation [52, 73, 83, 84], physics [68], and 3D-property recognition [10, 16, 25, 47, 59, 60, 72, 75]. More importantly, many studies have emphasized the importance of understanding relationships between objects to learn high-level context [15, 27, 46, 49, 53, 85]. Given a natural image (Figure 1a, b), examples of such understanding would be ‘Horse3 occludes Person2.’, ‘Horse1 and Person3 occlude each other.’, or ‘Horse2 is closer than Person2.’.

For this purpose, we introduce a new large-scale dataset, called INSTAORDER, for geometric scene understanding. The dataset has extensive annotations on *geometric orderings between class-labeled instances in the natural scenes*. INSTAORDER provides (1) *Occlusion order* that determines objects that occlude others (occluders), and objects that are occluded (occludees), and (2) *Depth order* that describes which object is closer or farther to the camera. INSTAORDER is the first dataset that provides these two kinds of orders together from the same image.



(a) Image (b) Instance masks
(c) Occlusion order (d) Depth order

Figure 1. Overview of the proposed INSTAORDER dataset. (a and b) Example image of a cluttered scene and instance masks with class labels. (c) *Occlusion order*. Arrows run from occluder to occludee. (d) *Depth order*. Arrows point from close to far.

The two types of geometric relations can be expressed using directed graphs (Figure 1c, d). Occlusion order and depth order are *complementary* to each other, and *neither alone can fully depict the geometric relationship in the cluttered scene.* For example, Horse2 in the occlusion graph (Figure 1c) is isolated, so Horse2’s depth is not clear without looking at the depth order graph (Figure 1d). In contrast, looking only at the depth order graph does not demonstrate whether Horse1 occludes Horse3, whereas the occlusion order graph does provide such information. Compared with other datasets shown in Figure 2, InstaOrder is the only large-scale and comprehensive dataset that provides instance segmentation mask, instance class label, occlusion, and depth order with delicate annotation of ordering types as shown as bidirectional edges and dashed edges.

InstaOrder is built on the COCO 2017 [38] dataset. A total of 3,659 crowd-workers annotated geometric ordering for 100,623 images having 503,939 instances, for a total of 2,859,919 depth and occlusion orderings. Such large-scale annotation distinguishes InstaOrder from the prior datasets that only cover occlusion order [52, 84] or depth order [7]. In addition to its scale, InstaOrder introduces *richer annotation on ambiguous cases* that had not been addressed before [7, 52, 84]. For example, *bidirectional* order covers the case (Figure 1) in which Horse3 and Person1 occlude each other. For depth order, in addition to {closer, farther, or equal} orderings, we introduce *distinct* and *overlapping* depth orders. For example, some parts of Person1's left leg are closer than Horse3, whereas the right arm is farther (Figure 1a). This case is annotated as overlap depth and displayed as a dashed line (Figure 1d). The direction of the dashed line indicates that some part of Person1 (left leg) is nearer than any part of Horse3.

We also propose new networks called InstaOrderNet and InstaDepthNet. InstaOrderNet is used to recover instance-wise orders from an image. We show that InstaOrderNet achieves higher accuracy than state-of-the-art approaches, such as PCNet-M [77] and OrderNet$^{\text{M+I}}$ [84]. InstaDepthNet is used to predict a dense depth map from an image. With the proposed instance-wise disparity loss and the InstaOrder dataset, InstaDepthNet can boost the accuracy of MiDaS [54], a state-of-the-art depth estimation network.

The contributions of this paper are:

- We introduce the **InstaOrder** dataset that provides 2.9M of comprehensive instance-wise geometric orderings for 101K natural scenes. InstaOrder is the first dataset of both occlusion and depth order from the same image, with bidirectional occlusion order and delicate depth range annotations.
- We discover that occlusion and depth order are complementary, and that instance-wise orders are helpful for the monocular depth prediction task.
- We introduce **InstaOrderNet** for geometric order prediction and show its superior accuracy over state-of-the-arts. In addition, we introduce **InstaDepthNet**, which demonstrates that the proposed auxiliary loss for geometric ordering can boost the depth prediction accuracy of the state-of-the-art approach, MiDaS [54].
- The InstaOrder dataset, pre-trained model, and toolbox are available at https://github.com/POSTECH-CVLab/InstaOrder

2. Related Work

Datasets for occlusion orders. Understanding occlusion is proven to improve the ability of scene understanding in various computer vision tasks, such as detection [66, 79], instance segmentation [30, 76], depth estimation [44], and optical flow estimation [26, 67]. Recently, the concept of amodal perception has been emphasized, estimating the whole physical structure from a partial observation. Knowing occlusion order is crucial when inferring amodal masks.

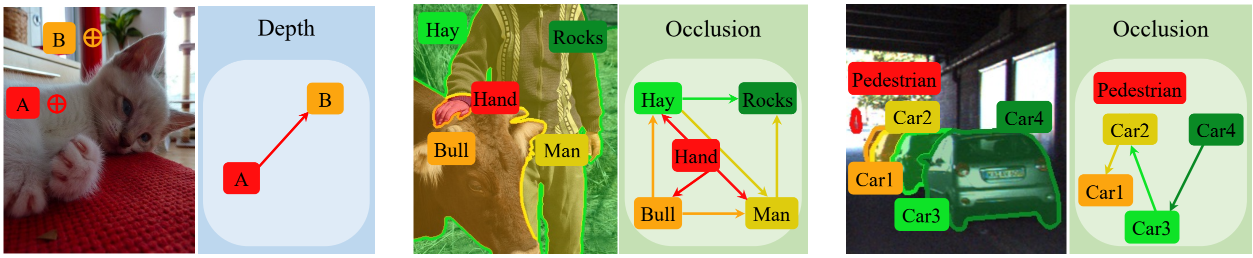
COCOA [84] is the first amodal dataset that contains both modal and amodal segmentation masks and their pair-wise occlusion orders. However, COCOA provides only 5,073 images, which is an insufficient number for data-driven approaches. Moreover, COCOA is designed with one-directional occlusion, and therefore splits instance masks if two instances occlude each other. In this procedure, annotators assign arbitrary labels to the new masks instead of assigning pre-defined class labels. KINS [52] provides modal and amodal segmentation masks with relative occlusion order. KINS consists of 14,991 images, and it is built upon KITTI [16] dataset. Therefore, all of its images are of driving scenes.

InstaOrder is much larger than COCOA [84] and KINS [52] (Table 1). In addition, InstaOrder provides *bidirectional occlusion*, whereas COCOA and KINS do not. We observed that bidirectional order facilitates the understanding of scenes.

Occlusion order prediction. Tighe *et al.* [62] build a histogram to predict occlusion overlap scores between two classes and solve quadratic integer programs. Zhu *et al.* [84] proposed OrderNet$^{\text{M+I}}$ that takes two masks and an image patch as input then produces pair-wise occlusion order in a supervised scheme. Zhan *et al.* [77] proposed PCNet-M that recovers occlusion order in a self-supervised manner.

The proposed InstaOrderNet$^{\text{o}}$ and InstaOrderNet$^{\text{o,d}}$ can identify instances of 'no occlusion', 'unidirectional occlusion', and 'bidirectional occlusion'. To the best of our knowledge, it is the first attempt in the field to identify three types simultaneously. We experimentally show that our InstaOrderNet$^{\text{o}}$, InstaOrderNet$^{\text{o,d}}$ is more accurate than OrderNet$^{\text{M+I}}$ [84] and PCNet-M [77] networks in COCOA [84], KINS [52] and InstaOrder datasets.

Datasets for depth maps. Advances in depth sensors have enabled extending 2D space to 3D space by collecting large-scale RGB-D datasets. For indoor environments, NYU Depth V2 [47], SUN3D [72], SUN RGB-D [59], SceneNN [25], ScanNet [25] datasets exist. For outdoor environments, KITTI [16], Cityscapes [10], Waymo Open Dataset [60], and BDD100K [75] datasets exist. Although these datasets have enabled rapid progress in scene understanding, the scene types are restricted to either indoor or driving scenes. As a result, those datasets do not cover *natural scenes*, in which diverse kinds of instances co-exist in



(a) DIW [7] (b) COCOA [84] (c) KINS [52]

Figure 2. Overview of prior work. (a) DIW [7] provides depth order of arbitrary points in a large-scale. (b) COCOA [84] and (c) KINS [52] provide instance segmentation mask, instance class label, and instance-wise occlusion order.

	# Img	Scene	Source	# of classes	Depth order	Occ. order	# of annotations	Year
DIW [7]	495K	Natural Scenes	Flickr	Not available	Two points per img	Not available	495K	2016
COCOA [84]	5K	Natural Scenes	COCO [38]	Not deterministic	Not available	✓ (instance)	311K	2017
KINS [52]	15K	Driving Scenes	KITTI [16]	8	Not available	✓ (instance)	1.6M	2019
INSTAORDER	101K	**Natural Scenes**	COCO [38]	**80**	✓ **(instance)**	✓ **(instance)**	**2.9M**	Proposed

Table 1. Summary of related datasets available to the community. The colored cells indicate weak (🟥), moderate (🟨), or strong (🟩) points of each dataset. The proposed INSTAORDER provides the largest amount of occlusion and depth annotations for various classes.

an unconstrained setting. In addition, the depth value of transparent, specular, or distant objects is often not reliable because of the limitations of depth sensors.

Recent datasets obtained geometric information for diverse scene types by crowd-sourcing [7, 8] or by applying photogrammetry methods [34, 35]. Stereo photos in the web [69, 71] or 3D movies [54] are used as another type of way to capture depth information. These datasets have been applied successfully to estimate a dense depth map from a single image, but they lack instance information and instance-wise relationships. In addition, the dataset for training a model is not publicly available [54].

In contrast, INSTAORDER covers instance masks, class labels, and instance-wise ordinal relationships. We experimentally show that instance-wise depth order improved modern monocular depth estimation networks like MiDaS [54].

3. INSTAORDER dataset

3.1. Data Collection

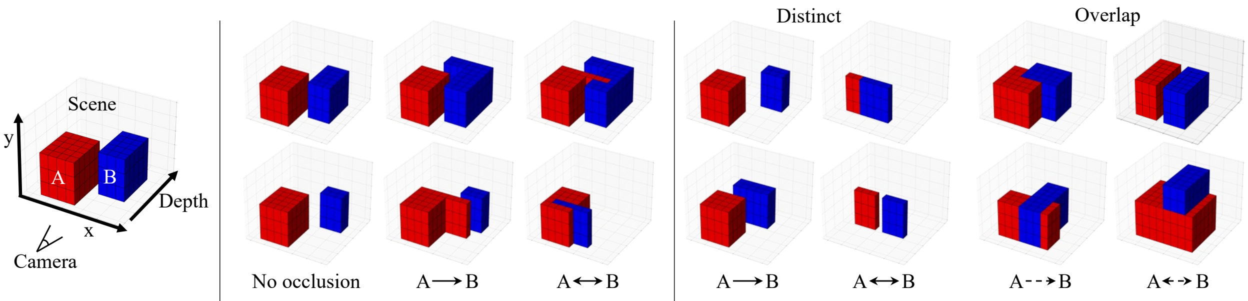
Parent dataset. We annotate occlusion and depth orders upon COCO 2017 [38] dataset to get a benefit from large-scale instance labeling of natural scenes. Several other datasets also provide instance segmentation, such as LVIS [21], ADE20K [82], Cityscapes [10]. We decided to use COCO 2017 due to the following strengths: large-scale image set, covering diverse natural scenes, many instances in each image, and providing instance masks. We omit instances smaller than 25 × 25 pixels from the annotation because orders are often difficult to discern for tiny objects. We also discard inappropriate images for annotation, such as images with a single instance and collage images. As a result, images and instance masks in COCO 2017 [38] train set (96,552 images) and validation set (4,071 images) are used for the annotation.

Annotation task. As Todd *et al.* [63] stated, humans are good at judging relative depths. Inspired by this, our annotation procedure is designed as the task of requesting pairwise depth ordering between two instances in the same image. Both occlusion and depth annotation tasks start with guidelines, and then real examples appear as quizzes. Only annotators who passed all quizzes were allowed to participate in the annotation. Moreover, if a worker gives a wrong answer multiple times, the worker is dismissed. We provide a guideline to crowd-workers to annotate only the semantically meaningful instances (Sec. A3.2 in the supplement).

Minimizing dataset biases. We build our dataset with the following consideration to minimize dataset biases.

(1) Class balance. Our INSTAORDER dataset reuses the images of the COCO 2017 [38] dataset, which was built using a careful image category decision and image collection mechanism to minimize dataset biases. The candidate image categories are gathered from frequently used words or PASCAL VOC [45], and the decision is made by a vote on how one category is distinguishable from others. The COCO dataset is a collection of images from Flickr; to avoid collecting iconic photos, the process searched for images that had multiple keywords, such as ’dog + car’.

(2) Crowdsourcing. INSTAORDER is collected using a sophisticated crowdsourcing engine. Thousands of people participated in the annotation of occlusion order or depth order, so the huge number of crowd-workers reduced the bias in the ordering annotations. To minimize diverged annotations, we asked two random workers to annotate every



(a) 3D scene (b) Occlusion order (c) Depth order

Figure 3. Ordering types defined for the InstaOrder dataset. (a) Camera looking at the scene of the two objects. (b) Occlusion ordering is denoted considering the occluder and occludee. (c) Depth ordering. Distinct object pairs do not have overlapping depth regions, whereas overlapping pairs do. See Sec. 3.2 for details.

pairwise ordering. If the annotation results from two workers did not match, we invited additional workers until two of the workers made the same decision. We use ***count*** to denote the number of participants per question, and our dataset provides count along with the occlusion and depth orderings to indicate the difficulty of the annotations.

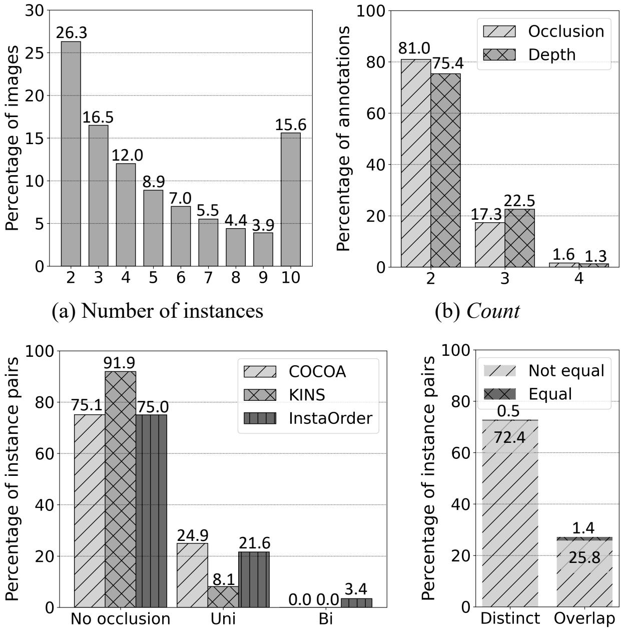
3.2. Ordering Types

Given a scene observed using a camera (Figure 3a), we identify occlusion and depth order. Occlusion order is determined by identifying the occluder and the occludee (Figure 3b). We utilize ‘no occlusion’ (no edge connection between A and B), ‘unidirectional occlusion’ (A occludes B: A→B; or B occludes A: B→A), and ‘bidirectional occlusion’ (A and B occlude each other: A⟷B).

Depth order denotes the relative distances of two objects from the camera. When an instance’s depth range covers the other instance’s depth range (Figure 3a), it is ambiguous to represent depth order with one of {closer, farther, equal}; thus, *distinct, overlap* label is needed. Depth order is annotated with a tuple of (x,y), where x∈{closer, farther, equal} and y∈{distinct, overlap}. Let’s denote instance A’s starting depth as A_S and ending depth as A_E, where $A_S <= A_E$. A→B (distinct) is when $A_E < B_S$. A⟷B (distinct) is when $A_S = B_S = A_E = B_E$. A pair of instances in the same plane belong here (e.g. instances shown on TV). A-→B (overlap) is when (i) $A_S < B_S < A_E$ or (ii) $A_S = B_S$ and $A_E < B_E$. A←-→B (overlap) is when $A_S = B_S$, $A_E = B_E$ and $A_S \neq A_E$. Figure 3c shows such examples.

3.3. Statistics

InstaOrder consists of 100,623 images with 503,939 instances that belong to 80 class categories, for a total of 2,859,919 instance-level occlusion and depth orders. Due to the limited annotation budget, we randomly selected ten instances if an image included more than ten instances. The histogram of instance number per image is shown in (Figure 4a).



(a) Number of instances (b) *Count*

(c) Occlusion order types (d) Depth order types

Figure 4. Statistics of the InstaOrder dataset. Sec. 3.3 for details.

The dataset was annotated by 1,549 crowd-workers for occlusion order, and by 2,110 for depth order. Of the instance pairs, 81% were accepted by the first two annotators for occlusion ordering, whereas 75.4% were accepted by the first two annotators for the depth ordering (Figure 4b). This comparison indicates that depth ordering was slightly harder to annotate than occlusion ordering.

InstaOrder has a similar distribution of occlusion order types to those of COCOA [84] or KINS [52] (Figure 4c). In addition to unidirectional occlusion type, InstaOrder provides bidirectional occlusion order. On depth order annotations, 72.9% belonged to a distinct type and 27.2% belonged to an overlap type (Figure 4d). The majority of depth orders belong to a ‘distinct’ and ‘not equal’ category (A→B or B→A), composing 72.4% of total depth orders.

3.4. Key Findings

Here we provide interesting observations from the InstaOrder. These findings were observed in the comprehensive annotation of occlusion and depth order. Therefore, we highlight that these are new findings not discussed in previous literature [7, 52, 84].

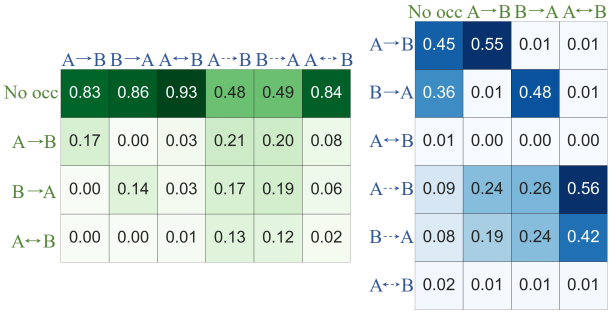
(1) *Occlusion order and depth order should be annotated independently, because neither can indicate the other.* We cannot perfectly infer depth order from occlusion order and vice versa. We demonstrate the claim with the correlation (Figure 5) between occlusion and depth order in the InstaOrder dataset. A pair of instances have occlusion and depth orders, so we can calculate P(occ. order | depth order); the proportions of occlusion order types given depth order types (Figure 5a). For example, P(No occ | A→B) is 83%. We can see that no "must happen correlation" occurs between two order types. For example, all types of occlusion order can occur when the depth order is A⇢B. Similar results are obtained using the proportions of depth order types given occlusion order type (Figure 5b). For reference, LabelMe [57] uses heuristics to infer occlusion order from instance masks. However, such heuristics are only applicable when masks intersect. In the InstaOrder dataset, only 16.4% of mask pairs intersect, where we determine 'intersect' if two masks overlap more than ten pixels. Therefore, we cannot apply the heuristics to 83.6% of non-intersecting instance pairs.

(2) *Occlusion order and depth order are complementary to each other.* We demonstrate this relationship in experiments (Tables 2, 3). A network trained with both occlusion and depth order is more accurate than baselines trained with individual orders. We think utilizing both types of orders eliminates cases that cannot happen. For example, (first column of Figure 5a) if the depth order is A→B, the occlusion orders B→A and A⟷B cannot occur. Therefore, joint use of occlusion and depth orders provide rich supervision for comprehensive scene understanding.

(3) *Bidirectional occlusion order is helpful.* We conduct experiments with the InstaOrder dataset to verify the effect of bidirectional occlusion orders (Sec. A2.1 in the supplement). The result indicates methods that can determine bidirectional order distinguishes ambiguous cases better than methods that cannot determine bidirectional order.

4. Methods

This section introduces neural networks and a loss function that can be applied to the InstaOrder dataset. We present **InstaOrderNet**, which predicts instance-wise orders. Then we introduce a depth map prediction network **InstaDepthNet**, which gains accuracy with the proposed **instance-wise disparity loss**. The details of the network architecture are described in Sec. A4 in the supplement.



	A→B	B→A	A↔B	A⇢B	B⇢A	A⇠⇢B
No occ	0.83	0.86	0.93	0.48	0.49	0.84
A→B	0.17	0.00	0.03	0.21	0.20	0.08
B→A	0.00	0.14	0.03	0.17	0.19	0.06
A↔B	0.00	0.00	0.01	0.13	0.12	0.02

(a) P(occ. order | depth order)

	No occ	A→B	B→A	A↔B
A→B	0.45	0.55	0.01	0.01
B→A	0.36	0.01	0.48	0.01
A↔B	0.01	0.00	0.00	0.00
A⇢B	0.09	0.24	0.26	0.56
B⇢A	0.08	0.19	0.24	0.42
A⇠⇢B	0.02	0.01	0.01	0.01

(b) P(depth order | occ. order)

Figure 5. (a) Proportion of occlusion order given depth order and (b) vice versa. Each column is summed to one.

4.1. Order Prediction

Occlusion order. The proposed InstaOrderNet$^{\mathrm{o}}$ takes pairwise instance masks and an image patch as input, then outputs occlusion order. InstaOrderNet$^{\mathrm{o}}$ is largely inspired by OrderNet$^{\mathrm{M+I}}$ [84], so InstaOrderNet$^{\mathrm{o}}$ uses the pre-trained ResNet-50 [24] backbone as OrderNet$^{\mathrm{M+I}}$ does.

OrderNet$^{\mathrm{M+I}}$ [84] classifies three types of occlusion order: {No occlusion, A→B, B→A}. The output dimension of OrderNet$^{\mathrm{M+I}}$ is [`batch_size`, 3], trained with cross-entropy loss. On the other hand, the output dimension of InstaOrderNet$^{\mathrm{o}}$ is [`batch_size`, 2] and it is trained with binary cross-entropy loss, $\mathcal{L}_{oo}$. InstaOrderNet$^{\mathrm{o}}$ solves two simple tasks: (1) 'does A occludes B?' and (2) 'does B occludes A?', answering two questions expresses four types of occlusion order: {No occlusion, A→B, B→A, A⟷B}.

Depth order. We introduce InstaOrderNet$^{\mathrm{d}}$ to predict depth order. InstaOrderNet$^{\mathrm{d}}$ takes pairwise instance masks and an image as input then produces depth orders. InstaOrderNet$^{\mathrm{d}}$ uses pre-trained ResNet-50 backbone, and it is trained with cross-entropy loss $\mathcal{L}_{do}$. The output dimension of the network is [`batch_size`, 3], and three channels stand for {A→B, B→A, A⟷B}.

Occlusion and depth order. To demonstrate the effectiveness of jointly using occlusion and depth order, we introduce InstaOrderNet$^{\mathrm{o,d}}$, which takes pairwise instance masks along with an image and produces *both* occlusion and depth order. For a fair comparison, we build InstaOrderNet$^{\mathrm{o,d}}$ with the same neural architecture that was used in InstaOrderNet$^{\mathrm{o}}$ and InstaOrderNet$^{\mathrm{d}}$ except for the last fully-connected (FC) layer. Specifically, InstaOrderNet$^{\mathrm{o,d}}$ consists of two FC layers placed in parallel; one predicts occlusion order, and one predicts depth order (Figure A2 in the supplement).

4.2. Depth Map Prediction

We also propose InstaDepthNet to show how instance-wise orderings can improve the state-of-the-art monocular depth estimation approach, MiDaS [54]. InstaDepthNet consists of two heads, one for instance-wise order prediction

Methods	COCOA [84] dataset Recall ↑	 Prec. ↑	 F1 ↑	KINS [52] dataset Recall ↑	 Prec. ↑	 F1 ↑
PCNet-M [77]	82.33	84.58	82.80	94.62	91.60	92.59
OrderNet^{M+I} [84]	**89.12**	83.91	85.63	98.33	93.45	95.19
InstaOrderNet^{o}	88.60	**85.38**	**86.16**	**98.70**	**94.56**	**96.07**

INSTAORDER dataset	Input			Output		Occlusion acc. ↑		
Methods	Mask	Image	Category	Occ. order	Depth order	Recall	Prec.	F1
Area	✓			✓		56.33	71.55	59.67
Y-axis	✓			✓		44.84	57.34	47.30
PCNet-M [77]	✓	✓		✓		59.19	76.42	63.02
OrderNet^{M+I}(ext.)	✓	✓		✓		84.93	78.21	77.51
InstaOrderNet^{o}(M)	✓			✓		87.35	79.07	78.98
InstaOrderNet^{o}(MC)	✓		✓	✓		88.70	78.21	79.18
InstaOrderNet^{o}(MIC)	✓	✓	✓	✓		89.38	79.00	79.98
InstaOrderNet^{o}	✓	✓		✓		**89.39**	79.83	80.65
InstaOrderNet^{o,d}	✓	✓		✓	✓	82.37	**88.67**	**81.86**

Table 2. Occlusion order prediction results. We use COCOA [84] and KINS [52] (top), and we use INSTAORDER (bottom) for the experiments. We discuss methods highlighted in yellow in Sec. 5.1.

and the other for depth map prediction. The instance-wise order prediction head is composed of ResNet-50 [24]. The depth map prediction head is composed of MiDaS-v2 [54], adopting pre-trained weights provided by authors.

The InstaDepthNet architecture (Figure A3 in the supplement) has a modular design for the tasks. Therefore, the ordering prediction heads can be used for the training, and InstaDepthNet can produce a dense disparity map without instance masks during test time. We propose two versions of InstaDepthNet, such as InstaDepthNet^{d} and InstaDepthNet^{o,d} depending on the ordering types for the supervision.

We apply four loss functions to train InstaDepthNet. For the order prediction heads, we use binary cross-entropy loss $\mathcal{L}_{oo}$ for the occlusion order prediction head or cross-entropy loss $\mathcal{L}_{do}$ to the depth order prediction head.

For the depth map prediction head, we introduce ***instance-wise disparity loss*** $\mathcal{L}_{disp}$. We denote depth order as d_{AB} and set it to $\{1, 0, \text{-}1\}$ when depth order is $\{$closer, equal, farther$\}$, respectively. Disparity is inversely proportional to depth, so when A is closer than B ($d_{AB} = 1$), the disparity should be bigger for A than for B. $\mathcal{L}_{disp}$ penalizes violations of this relation by applying the proposed loss function:

$$\begin{aligned}\mathcal{L}_{disp} = \frac{1}{2N}\sum_{i\in A\cup B}\Big\{&\mathbb{1}\Big(d_{AB}D'_A(i) \le d_{AB}\max(D'_B)\Big) + \\ &\mathbb{1}\Big(d_{AB}D'_B(i) \ge d_{AB}\min(D'_A)\Big)\Big\},\end{aligned} \qquad (1)$$

where i is a pixel in the area $A \cup B$, D'_A is a predicted disparity map of A, $\mathbb{1}(\cdot)$ is an indicator function, and N is the number of pixels in $A \cup B$. We apply $\mathcal{L}_{disp}$ to distinct pairs because these orders are clear to supervise. We also use edge-aware smoothness loss [19]: $\mathcal{L}_s = \frac{1}{N}\sum_i |\partial_x D'(i)|\, e^{-\|\partial_x I(i)\|} + |\partial_y D'(i)|\, e^{-\|\partial_y I(i)\|}$, where I is an image, and ∂_x and ∂_y respectively are x- and y-directional image gradient operators.

The final loss is $\lambda_0\mathcal{L}_{oo} + \lambda_1\mathcal{L}_{do} + \lambda_2\mathcal{L}_{disp} + \lambda_3\mathcal{L}_s$. $\{\lambda_0, \lambda_1, \lambda_2, \lambda_3\}$ is set to $\{0,\ 1,\ 1,\ 0.1\}$ for InstaDepthNet^{d}, and $\{1,\ 1,\ 1,\ 0.1\}$ for InstaDepthNet^{o,d}. We conduct an ablation study on loss functions in Sec. A2.2 in the supplement.

5. Experiments

5.1. Occlusion Order Recovery

Baselines. We compare the performance of the proposed InstaOrderNet^{o} with OrderNet^{M+I} [84] and PCNet-M [77]. InstaOrderNet^{o} can process bidirectional occlusion order, whereas others cannot. Therefore we extend OrderNet^{M+I} to be able to predict bidirectional order and named it as OrderNet^{M+I}(ext.). For evaluating PCNet-M on COCOA [84] and KINS [52] dataset, we use pre-trained weights provided by the authors. We use the official source code of PCNet-M for training and testing with INSTAORDER dataset. We implement OrderNet^{M+I} from scratch, because the official code is not available[1].

Simple approaches. We also conduct experiments using a simple heuristic proposed by Zhu *et al*. [84]. Specifically, given a pair of masks, the simple approach determines the occluder as the larger instance (the 'Area' method) or as the instance that is closer to the image bottom (the 'Y-axis' method). This experiment is intended to show that the simple prior cannot determine occlusion order well.

Datasets. The experiments were conducted with three representative datasets that contain instance-wise occlusion order: COCOA [84], KINS [52], and INSTAORDER. For fairness, we train and test each method with the same dataset. For example, the second column in Table 2 (top) indicates all methods are trained and tested with the KINS dataset.

Results. To evaluate the occlusion order of every instance pair, we use *Recall*, *Precision* and *F1* score (Table 2). In particular, we report the accuracy of the prediction of which of the two instances is an occluder, as done by OrderNet^{M+I} [84] and PCNet-M [77]. PCNet-M is a network trained in a self-supervised manner, whereas OrderNet^{M+I} and our InstaOrderNet^{o} are trained in a supervised manner. Interestingly, PCNet-M[2] showed comparable accuracy to the supervised methods on both the COCOA [84] and KINS [52] datasets (Table 2, top).

When the INSTAORDER dataset was used (Table 2,

[1] On the COCOA dataset, our OrderNet^{M+I} implementation achieves 89.1 recall, which is higher than the originally reported recall, 88.3.

[2] The recall reported in this paper is slightly different from the numbers appearing in PCNet-M because PCNet-M only consider neighboring instance mask pairs. In contrast, we used all pairs for the evaluation.

bottom), all InstaOrderNet versions achieved significantly higher accuracy than the other methods. InstaOrderNet^{o} is a simple extension of OrderNet^{M+I}, but achieves higher accuracy because it converts multi-class classification to the multi-label classification problem. InstaOrderNet^{o,d} is more accurate than InstaOrderNet^{d}. We can infer that occlusion and depth order are not independent, but provide *complementary* information for scene understanding. We observed that the accuracy of PCNet-M depends on instance mask quality and thus resulted in a low accuracy on the INSTAORDER dataset. We show qualitative results on occlusion order prediction (Figure 6a), and InstaOrderNet^{o} showed the best accuracy. Even though OrderNet^{M+I}(ext.) is an extended network to predict bidirectional occlusion order, but still missed most bidirectional occlusion orders.

Various input/output configurations. We conduct experiments with different types of inputs, such as image and category labels (Table 2, bottom). When we provide a category label to the mask, we assign appropriate category IDs to the masked areas. For the occlusion order prediction task, a network that uses the mask as a sole input (InstaOrderNet^{o}(M)) is even comparable to the network that uses both mask and image (InstaOrderNet^{o}); a similar result was reported by Zhu *et al*. [84]. We speculate that the mask provides enough clues to determine occlusion order. Moreover, we conduct an ablation study on bidirectional occlusion order (Sec. A2.1 in the supplement).

5.2. Depth Order Recovery

Baselines. To the best of our knowledge, no existing method directly analyzes depth ordering for instances. Therefore, we propose two baselines that use a state-of-the-art depth map estimation network, MiDaS-v2 [54], for the comparison of depth order prediction accuracy. The baseline approach, MiDaS(Mean), uses the instance-wise mean of the disparity predicted by MiDaS-v2. Similarly, MiDaS(Median) uses the instance-wise median value. This choice is guided by an assumption that instance-wise mean or median values represent instance-wise distance[3]. We compare baselines with the proposed InstaOrderNet^{d} (Table 3). We also evaluate simple approaches (Area, Y-axis).

WHDR. We evaluate the results using Weighted Human Disagreement Rate (WHDR) [3], which represents the percentage of weighted disagreement between ground truth d and predicted depth order d'. The weights are proportional to the confidence of each annotation. Here, we use the inverse of *count* multiplied by the minimum number of participants. We evaluate WHDR on each of {distinct, overlap, all} categories separately; which is defined as follows: WHDR = $\frac{\sum_{AB} w_{AB} \cdot \mathbb{1}(d'_{AB} \neq d_{AB})}{\sum_{AB} w_{AB}}$, where $w_{AB} = \frac{2}{count_{AB}}$.

Methods	Input			Output			WHDR ↓		
	Mask	Image	Category	Occ. order	Depth order	Disp. map	Distinct	Overlap	All
Area	✓				✓		30.90	35.66	32.19
Y-axis	✓				✓		22.19	39.04	29.20
MiDaS(Mean) [54]		✓				✓	10.42	37.67	21.70
MiDaS(Median) [54]		✓				✓	10.31	36.08	20.92
InstaOrderNet^{d}(M)	✓				✓		22.96	30.46	25.23
InstaOrderNet^{d}(MC)	✓		✓		✓		23.19	28.56	36.45
InstaOrderNet^{d}(MIC)	✓	✓	✓		✓		13.33	26.60	17.89
InstaOrderNet^{d}	✓	✓			✓		12.95	25.96	17.51
InstaOrderNet^{o,d}	✓	✓		✓	✓		11.51	25.22	15.99
InstaDepthNet^{d}(Mean)	✓	✓			✓	✓	9.80	37.97	21.46
InstaDepthNet^{d}(Median)	✓	✓			✓	✓	9.29	36.07	20.41
InstaDepthNet^{d}	✓	✓			✓	✓	7.25	23.34	12.94
InstaDepthNet^{o,d}	✓	✓		✓	✓	✓	**7.00**	**23.29**	**12.72**

Table 3. Depth order prediction results. We train and test networks with various input and output configurations. We discuss methods highlighted in yellow in Sec. 5.2 and Sec. 5.3.

Results. Both MiDaS(Mean, Median) achieved notable WHDR (Table 3) even though INSTAORDER is an unseen dataset to MiDaS [54]. Although trained with the INSTAORDER dataset, InstaOrderNet^{d} had inferior (higher) WHDR than MiDaS(Mean, Median) for the distinct instances. We observe that instance-wise depth ordering must involve the disparity map prediction task, as InstaDepthNet does. InstaDepthNet is trained for dense disparity map prediction as well as order predictions. We believe that such comprehensive tasks give plentiful supervision for distinct instances and bring significant accuracy gain. As observed in occlusion order recovery results, InstaOrderNet^{o,d} is superior to InstaOrderNet^{d}; this indicates occlusion and depth order are complementary information.

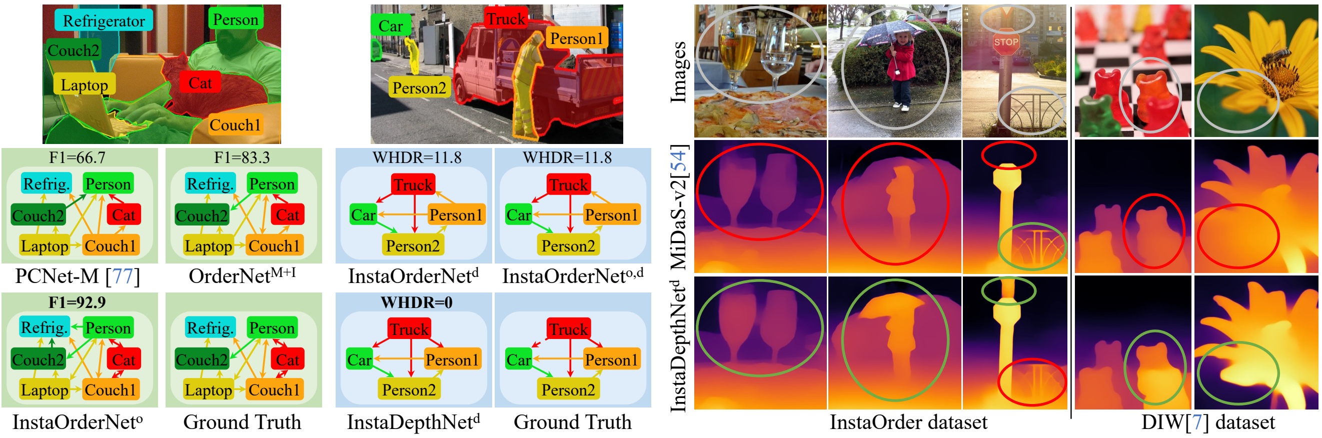
When the image is not used as an input, the accuracy degrades (InstaOrderNet^{d}(M, MC)). The results are different from the previous experiment with occlusion order. We think the image gives a global context to determine relative depth order correctly. Qualitative results on depth order prediction (Figure 6b) indicate InstaDepthNet^{d} is better at figuring out tricky relations, such as Truck⇢Person.

5.3. Depth Map Prediction

Testing with InstaOrder dataset. We further demonstrate that occlusion and depth orders can be used to increase the accuracy of a depth estimation network. We compare the disparity map predicted by MiDaS [54] and InstaDepthNet^{d} using (Sec. 5.2 Baselines) the mean and median scheme. (Table 3) InstaDepthNet^{d}(Mean, Median), were both more accurate than MiDaS(Mean, Median).

We compare the qualitative result of the disparity map estimated by MiDaS-v2 with InstaDepthNet^{d} (Figure 6c).

[3] Instance segmentation masks of COCO 2017 are not perfect, so for reliability of comparison, we ignore the top and bottom 5% disparity values.



(a) Occlusion order (b) Depth order (c) Disparity map

Figure 6. Qualitative results on the (a) occlusion, (b) depth order prediction. (c) Disparity maps generated by MiDaS-v2 [54] and our InstaDepthNet^{d} on INSTAORDER (left) and DIW [7] (right). Red ellipses: unreasonable predictions; green ellipses: reasonable predictions.

DIW [7]	Input			Output			Accuracy		
Methods	Mask	Image	Category	Occ. order	Depth order	Disp. map	# Correct ↑	# Wrong ↓	WHDR ↓
MiDaS-v2 [54]		✓				✓	64,723	9,718	13.06
InstaDepthNet^{d}		✓				✓	**65,317**	**9,124**	**12.26**

KITTI [16]	Error ↓			Accuracy ↑		
Methods	Abs Rel	Sq Rel	RMSE log	δ <1.25	δ <1.25^{2}	δ <1.25^{3}
MiDaS-v2 [54]	0.16	1.47	**0.20**	**0.81**	0.95	0.98
InstaDepthNet^{d}	**0.15**	**1.27**	**0.20**	0.80	**0.95**	**0.99**

Table 4. Evaluation of predicted disparity maps using unseen datasets (top table: DIW [7], bottom table: KITTI [16]).

Human annotation in INSTAORDER for challenging objects (transparent glasses) and instance-wise depth order helps to correct wrong disparity prediction.

Testing with unseen datasets. First, we compare the depth order accuracy on the DIW [7] dataset (Table 4, top). We compare InstaDepthNet^{d}, which was trained on the INSATAORDER dataset, whereas MiDaS-v2 was trained using numerous 3D movies. InstaDepthNet^{d} showed superior disparity maps (Figure 6c, right); this result supports the value of the proposed INSTAORDER dataset and the instance-wise disparity loss (Eq. 1).

We also test two approaches with the KITTI dataset [16] (Table 4, bottom). The predicted disparity maps of both approaches are not on a metric scale, so we adopt per-image median ground truth scaling [20], and report statistical metrics [14] (details in Sec. A1 in the supplement). Overall, InstaDepthNet^{d} was slightly better than MiDaS.

Limitation. InstaDepthNet^{d} did yield some problematic results (Figure 6c, red ellipse in the middle example). The confusion occurs because instance masks in our parent dataset, COCO 2017 [38] do not fully segment objects that have holes. We leave this problem to future work.

6. Discussion

We introduce INSTAORDER dataset and propose various order prediction networks. Our dataset has several benefits compared to DIW [7], COCOA [84], or KINS [52] in terms of scale, classes, and order types. We demonstrate the effectiveness of jointly using occlusion order and depth order. We show that the state-of-the-art depth map prediction approach can be improved by using the proposed auxiliary loss for instance-wise ordering.

We plan to study the benefit of INSTAORDER for the tasks beyond depth estimation. For example, as panoptic segmentation studies [9, 31, 40] gain accuracy by figuring out the occlusion order between objects, INSTAORDER can benefit the task by explicitly reasoning the occlusion order. In addition, INSTAORDER can help image captioning or VQA tasks. Specifically, (Figure1c,d) "Horse1" and "Person3" are occluding each other, then we can infer they are interacting. Moreover, we can make a question and answer like "Who is behind Person 1?". Image generation studies [1, 28] use scene graphs for generating images. It would be interesting to create images by considering occlusion, depth order, and scene graph. Moreover, as Zhan *et al*. [77] manipulate images by controlling occlusion order, INSTAORDER can also be used for image manipulation.

Acknowledgments. We thank Qualcomm for the generous support and SELECTSTAR for the data collection help. This work was supported by IITP grants 2021-0-00537 (Visual common sense) and 2019-0-01906 (Artifcial Intelligence Graduate School Program (POSTECH)) funded by the Korea government (MSIT).

References

[1] Oron Ashual and Lior Wolf. Specifying Object Attributes and Relations in Interactive Scene Generation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 8

[2] Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Context R-CNN: Long Term Temporal Context for Per-Camera Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[3] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (SIGGRAPH)*, 2014. 7

[4] Jia-Ren Chang and Yong-Sheng Chen. Pyramid Stereo Matching Network. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[5] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D. Collins, Ekin D. Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-Student: Leveraging Semi-Supervised Learning in Video Sequences for Urban Scene Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1

[6] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L. Yuille. Attention to Scale: Scale-Aware Semantic Image Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[7] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-Image Depth Perception in the Wild. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 2, 3, 5, 8

[8] Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, and Jia Deng. OASIS: A Large-Scale Dataset for Single Image 3D in the Wild. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[9] Yifeng Chen, Guangchen Lin, Songyuan Li, Omar El Farouk Bourahla, Yiming Wu, Fangfang Wang, Junyi Feng, Mingliang Xu, and Xi Li. BANet: Bidirectional Aggregation Network With Occlusion Handling for Panoptic Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 8

[10] 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 International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 3

[11] Jifeng Dai, Kaiming He, Yi Li, Shaoqing Ren, and Jian Sun. Instance-Sensitive Fully Convolutional Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1

[12] Jifeng Dai, Kaiming He, and Jian Sun. Instance-Aware Semantic Segmentation via Multi-task Network Cascades. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[13] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: Object Detection via Region-based Fully Convolutional Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 1

[14] David Eigen, Christian Puhrsch, and Rob Fergus. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 1, 8

[15] Carolina Galleguillos, Andrew Rabinovich, and Serge J. Belongie. Object categorization using co-occurrence, location and appearance. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 1

[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1, 2, 3, 8

[17] Ross B. Girshick. Fast R-CNN. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015. 1

[18] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[19] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised Monocular Depth Estimation with Left-Right Consistency. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6

[20] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging Into Self-Supervised Monocular Depth Estimation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 1, 8

[21] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017. 1

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 1

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5, 6

[25] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. SceneNN: A Scene Meshes Dataset with aNNotations. In *Fourth International Conference on 3D Vision, (3DV)*, 2016. 1, 2

[26] Junhwa Hur and Stefan Roth. Iterative Residual Refinement for Joint Optical Flow and Occlusion Estimation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[27] Arpit Jain, Abhinav Gupta, and Larry S. Davis. Learning What and How of Contextual Models for Scene Labeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010. 1

[28] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image Generation From Scene Graphs. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 8

[29] Adrian Johnston and Gustavo Carneiro. Self-Supervised Monocular Trained Depth Estimation Using Self-Attention and Discrete Disparity Volume. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[30] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep Occlusion-Aware Instance Segmentation with Overlapping BiLayers. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[31] Justin Lazarow, Kwonjoon Lee, Kunyu Shi, and Zhuowen Tu. Learning Instance Occlusion for Panoptic Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 8

[32] Ke Li, Bharath Hariharan, and Jitendra Malik. Iterative Instance Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[33] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T. Freeman. Learning the Depths of Moving People by Watching Frozen People. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[34] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the Depths of Moving People by Watching Frozen People. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[35] Zhengqi Li and Noah Snavely. MegaDepth: Learning Single-View Depth Prediction from Internet Photos. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[36] Guosheng Lin, Chunhua Shen, Anton van den Hengel, and Ian D. Reid. Efficient Piecewise Training of Deep Structured Models for Semantic Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[37] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature Pyramid Networks for Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[38] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 2, 3, 8

[39] Chao Liu, Jinwei Gu, Kihwan Kim, Srinivasa G. Narasimhan, and Jan Kautz. Neural RGB->D Sensing: Depth and Uncertainty From a Video Camera. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[40] Huanyu Liu, Chao Peng, Changqian Yu, Jingbo Wang, Xu Liu, Gang Yu, and Wei Jiang. An End-To-End Network for Panoptic Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 8

[41] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path Aggregation Network for Instance Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[42] Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen Change Loy, and Xiaoou Tang. Semantic Image Segmentation via Deep Parsing Network. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015. 1

[43] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[44] Xiaoxiao Long, Lingjie Liu, Christian Theobalt, and Wenping Wang. Occlusion-Aware Depth Estimation with Adaptive Normal Constraints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[45] Everingham Mark, Gool Luc, Williams Christopher K., Winn John, and Zisserman Andrew. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision (IJCV)*, 2010. 3

[46] Heesoo Myeong, Ju Yong Chang, and Kyoung Mu Lee. Learning object relationships via graph-based context model. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1

[47] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor Segmentation and Support Inference from RGBD Images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012. 1, 2

[48] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning Deconvolution Network for Semantic Segmentation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015. 1

[49] Devi Parikh, C. Lawrence Zitnick, and Tsuhan Chen. From appearance to context-based recognition: Dense labeling in small images. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 1

[50] Pedro H. O. Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to Segment Object Candidates. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 1

[51] Pedro Oliveira Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to Refine Object Segments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1

[52] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal Instance Segmentation With KINS Dataset. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3, 4, 5, 6, 8

[53] Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge J. Belongie. Objects in Context. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2007. 1

[54] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 1, 2, 3, 5, 6, 7, 8

[55] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[56] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 1

[57] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. LabelMe: A Database and Web-Based Tool for Image Annotation. *IJCV*, 2008. 5

[58] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-Metric Loss for Self-supervised Learning of Depth and Egomotion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1

[59] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2

[60] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[61] Zachary Teed and Jia Deng. DeepV2D: Video to Depth with Differentiable Structure from Motion. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 1

[62] Joseph Tighe, Marc Niethammer, and Svetlana Lazebnik. Scene Parsing with Object Instances and Occlusion Ordering. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[63] James T. Todd and J. Farley Norman. The visual perception of 3-D shape from multiple cues: Are observers capable of perceiving metric structure? In *Perception & Psychophysics*, 2003. 3

[64] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. MTI-Net: Multi-scale Task Interaction Networks for Multi-task Learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1

[65] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. FEELVOS: Fast End-To-End Embedding Learning for Video Object Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[66] Angtian Wang, Yihong Sun, Adam Kortylewski, and Alan L. Yuille. Robust Object Detection Under Occlusion With Context-Aware CompositionalNets. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[67] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. Occlusion Aware Unsupervised Learning of Optical Flow. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[68] Jiajun Wu, Erika Lu, Pushmeet Kohli, Bill Freeman, and Josh Tenenbaum. Learning to See Physics via Visual De-animation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 1

[69] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular Relative Depth Perception With Web Stereo Data Supervision. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[70] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-Guided Ranking Loss for Single Image Depth Prediction. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[71] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-Guided Ranking Loss for Single Image Depth Prediction. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[72] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. SUN3D: A Database of Big Spaces Reconstructed Using SfM and Object Labels. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2013. 1, 2

[73] Xiaosheng Yan, Yuanlong Yu, Feigege Wang, Wenxi Liu, Shengfeng He, and Jia Pan. Visualizing the Invisible: Occluded Vehicle Segmentation and Recovery. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 1

[74] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a Discriminative Feature Network for Semantic Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[75] 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 *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[76] Xiaoding Yuan, Adam Kortylewski, Yihong Sun, and Alan Yuille. Robust Instance Segmentation through Reasoning about Multi-Object Occlusion. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[77] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-Supervised Scene De-Occlusion. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 6, 8

[78] Haokui Zhang, Ying Li, Yuanzhouhan Cao, Yu Liu, Chunhua Shen, and Youliang Yan. Exploiting Temporal Consistency for Real-Time Video Depth Estimation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 1

[79] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Occlusion-Aware R-CNN: Detecting Pedestrians in a Crowd. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[80] Ziyu Zhang, Sanja Fidler, and Raquel Urtasun. Instance-Level Segmentation for Autonomous Driving with Deep Densely Connected MRFs. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[81] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid Scene Parsing Network. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[82] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene Parsing through ADE20K Dataset. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[83] Qiang Zhou, Shiyin Wang, Yitong Wang, Zilong Huang, and Xinggang Wang. Human De-occlusion: Invisible Perception and Recovery for Humans. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[84] Yan Zhu, Yuandong Tian, Dimitris N. Metaxas, and Piotr Dollár. Semantic Amodal Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 3, 4, 5, 6, 7, 8

[85] Daniel Zoran, Phillip Isola, Dilip Krishnan, and William T. Freeman. Learning Ordinal Relationships for Mid-Level Vision. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015. 1