

# A Tri-Layer Plugin to Improve Occluded Detection

Guanqi Zhan<sup>1</sup>

guanqi@robots.ox.ac.uk

Weidi Xie<sup>1,2</sup>

weidi@robots.ox.ac.uk

Andrew Zisserman<sup>1</sup>

az@robots.ox.ac.uk

<sup>1</sup> Visual Geometry Group

University of Oxford

Oxford, UK

<sup>2</sup> Coop. Medianet Innovation Center

Shanghai Jiao Tong University

Shanghai, China

## Abstract

Detecting occluded objects still remains a challenge for state-of-the-art object detectors. The objective of this work is to improve the detection for such objects, and thereby improve the overall performance of a modern object detector.

To this end we make the following four contributions: (1) We propose a simple ‘plugin’ module for the detection head of two-stage object detectors to improve the recall of partially occluded objects. The module predicts a tri-layer of segmentation masks for the target object, the occluder and the occludee, and by doing so is able to better predict the mask of the target object. (2) We propose a scalable pipeline for generating training data for the module by using amodal completion of existing object detection and instance segmentation training datasets to establish occlusion relationships. (3) We also establish a COCO evaluation dataset to measure the recall performance of partially occluded and separated objects. (4) We show that the plugin module inserted into a two-stage detector can boost the performance significantly, by only fine-tuning the detection head, and with additional improvements if the entire architecture is fine-tuned. COCO results are reported for Mask R-CNN with Swin-T or Swin-S backbones, and Cascade Mask R-CNN with a Swin-B backbone.

## 1 Introduction

Occlusion frequently occurs in images of real scenes and is both a benefit and a problem. It is a benefit, because it reveals depth orderings and so contributes to a 3D perception of the scene [14]. However, it is a problem because despite the continual increase in performance of object detectors over the last decade, detection of occluded objects is still a significant deficiency [15, 16, 68].

In this paper, our objective is to improve the detection for objects under occlusion, and thereby improve the overall performance of the object detector. Specifically, we develop a lightweight ‘plugin’ module, that can be inserted into the detection head of any two-stage object detector, *e.g.* Mask R-CNN [17], to improve the recall of occluded objects. The module simultaneously infers *three* segmentation ‘layers’: the mask for the target object; the mask of the occluder (the object in front that occludes the target); and the mask of



**Figure 1. Improving occluded object detection and instance segmentation.** Left: Occlusion is very common in the 3D world, where one object is in front of another and a portion of the scene disappears behind the non-transparent object that is closer to the viewer. Right: For this example from COCO val, a Swin-T + Mask R-CNN Baseline incorrectly detects and segments the target object (the middle of the three cases). However, if the detection head is replaced with our tri-layer plugin, a correct segmentation mask is obtained. The plugin is tasked with predicting the masks of the target object occluder and occludee, and this leads to a better modelling for occluded detection.

the occludee (the object behind that is occluded by the target) within the same detection box. This *tri-layer prediction head* forces the detection model to explicitly understand the existence of occlusion relationships, and thereby is better able to predict the target object mask. Additionally, we show that the process can be iterated, using the better prediction of the target mask to adapt the detection box for the next tri-layer prediction.

One challenge for training our proposed plugin module lies in the lack of suitable training data – almost all large-scale detection datasets provide annotations on the visible part of objects, but no occlusion information is available. To this end, we propose a scalable pipeline for automatically discovering occluded objects (and their occluders), by running an amodal completion model on publicly available detection datasets, *e.g.*, COCO. Additionally, to verify the occlusion relationships between objects, *i.e.*, occluder or occludee, we adopt an off-the-shelf monocular depth estimator [24] and determine the occlusion relationship based on their relative depth. With this pipeline, we acquire a reliable large-scale data source for training our tri-layer model. We also establish an evaluation dataset to measure the recall performance for occluded objects, distinguishing between the two cases where the target object is *partially occluded* but the segmentation mask is connected, or where the target object segmentation mask is *separated* into distinct regions by the occluder.

To summarise, in this paper, we make the following four contributions: (i) We propose a simple ‘plugin’ module that can be inserted into two-stage object detectors, to improve their performance on occluded objects. (ii) We establish a scalable pipeline to determine occlusion relationships between objects, which is used to train the plugin module on publicly available detection datasets. (iii) We set up an evaluation benchmark with real images to measure the recall performance of partially occluded and separated objects. (iv) We show that the plugin module inserted into two-stage detectors can boost the performance significantly, by only fine-tuning the detection head, and with additional improvements if the entire architecture is fine-tuned. We show detection results for Mask R-CNN with Swin-T, Swin-S backbones, and Cascade Mask R-CNN with a Swin-B backbone.

## 2 Related Work

**Object Detection & Instance Segmentation.** Methods for object detection and instance segmentation have made great progress by training deep neural networks on large-scale datasets [8, 16, 18]. Two-stage detectors like Faster R-CNN [26] and Mask R-CNN [12] first train region proposal networks to propose candidate objects. These candidates are then fur-

ther refined in their location with a regression head, and categorised by a classifier head. With the development of stronger transformer-based [19, 20] architectures, replacing the original CNN backbones, there have been further improvements in the detection performance [19]. There has also been a similar development in single-stage detectors such as [9, 2, 40] but we do not consider those in this work. However, both single and two-stage detectors are still suffering when detecting occluded objects [27].

**Amodal Segmentation & Occlusion-Related Datasets.** Amodal segmentation refers to the task of segmenting the object as whole, including the portions that are partially occluded [17, 30, 39]. In the recent literature, various datasets have been collected to study occlusion [8, 6, 23, 32, 40]. However, most of these datasets either focus on certain domains/tasks or do not provide a large number of images for training and evaluating models in terms of occlusion for a large variety of categories.

**Layered Scene Understanding.** Layered representation was originally proposed by Wang and Adelson [13] to represent a video as a composition of layers with simpler motions. Since then, layered representations have been widely adopted in computer vision, *e.g.*, to decompose videos into layers [2, 14], to improve scene segmentation by explicitly modelling occlusions, temporal consistency, depth ordering [21, 33, 36, 37], to estimate optical flow [28, 29, 34, 35], and for novel view synthesis [40].

**Occlusion Handling.** To improve object detection and instance segmentation under occlusion scenarios, novel architectures, *e.g.*, compositional networks [61, 38], have been proposed. Rather than develop a new architecture, ASN [23] and ORCNN [9] ask modern detectors to output both modal and amodal masks of target objects. Additionally, BCNet [15] exploits two-layer graph neural networks to modern detectors for better instance segmentation under occlusion, inferring both the target object and the surrounding objects, while in this paper, we go one step further, and propose a simple, automatic pipeline for estimating objects' occlusion order, which enables training of the tri-layer plugin with explicit supervision on the occlusion ordering for objects.

### 3 Detector Architecture and Application

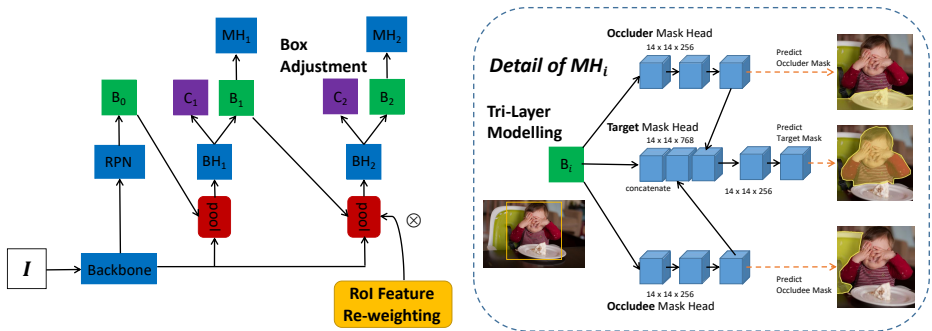
In this section we describe our lightweight plugin module, and its application within an object detector. The module is designed to improve the detection performance on occluded objects. We start by introducing the standard two-stage detector, and then describe the tri-layer plugin architecture and functionality.

**Two-stage detector.** Given an image detection dataset,  $\mathcal{D} = \{(I_1, y_1), \dots, (I_n, y_n)\}$ , a standard two-stage Mask R-CNN detector can be parametrized as:

$$y_j = \{(b_j, c_j, m_j)\}^K = \Phi_{\{\text{CLS+BOX;SEG}\}} \circ \Phi_{\text{ALIGN}} \circ \Phi_{\text{RPN}} \circ \Phi_{\text{ENC}}(I_j) \quad (1)$$

where an input image ( $I_j \in \mathbb{R}^{H \times W \times 3}$ ) with a total of  $K$  objects is sequentially processed by a set of operations: an image encoder,  $\Phi_{\text{ENC}}(\cdot)$ ; a region proposal network,  $\Phi_{\text{RPN}}(\cdot)$ ; a region of interest feature alignment,  $\Phi_{\text{ALIGN}}(\cdot)$ ; after predicting the class and box offset ( $\Phi_{\text{CLS+BOX}}(\cdot)$ ), a binary mask for each RoI is also predicted,  $\Phi_{\text{SEG}}(\cdot)$ . As a result,  $y_j = \{(b_j, c_j, m_j)\}^K$  denotes the box coordinates ( $b_j^k \in \mathbb{R}^4$ ), category label ( $c_j^k \in \mathbb{R}^C$ ), and **modal** segmentation mask ( $m_j^k \in [0, 1]^{H \times W}$ ) of the object, which has been converted from the spatial resolution of RoI align to the original image.

In the following, we refer to a pair of objects that have an occlusion relationship as **occluder** (the object that is in front of and thus occludes the other one), and **occludee** (the



**Figure 2. Architecture and function of the plugin module.** There are three functions: (a) The tri-layer mask head ( $MH_i$  shown in detail on the right) predicts the mask of the target object (the infant wiping food on their face), the occluder (the dining table), and the occludee (the chair) within the detection box  $B_i$ . The feature embeddings of the occluder/occludee branch are concatenated to the target mask embedding as cues to help better predict the target object mask; (b) As shown on the left, the process of predicting the target mask is iterated (index  $i$ ), such that the second iteration is able to adjust the initial box predictions and better detect partially occluded / separated objects; (c) After the first iteration, RoI features are pooled according to the predicted target mask to guide the model to focus more on the partially occluded / separated object itself. The notation used is: “ $I$ ” for the input image, and “BH”, “MH”, “B”, “C” refer to bbox head, mask head, bounding box, and classification respectively.

object at the back and being occluded), as illustrated in Figure 1. Note that, the role of each object is often relative, thus it can be both occluder and occludee at the same time, depending on its paired objects. In Section 4, we detail the procedure for acquiring the occlusion information, including their estimated **amodal** segmentation masks and the occlusion orderings.

### 3.1 Architecture

In this section, we introduce the architecture details of our tri-layer plugin for better detecting objects under occlusion. Specifically, in Section 3.1.1, we augment the detection module to simultaneously output both occluder and occludee, along with the target object itself; in Section 3.1.2, we introduce the idea of box adjustment, that facilitates the model to reason about the full object coverage with only partial observation, due to partial occlusion or separation; and in Section 3.1.3, we improve the RoI pooling procedure by weighting the feature map with the inferred object mask, effectively preventing the model from concentrating on the occluder/occludee. Figure 2 illustrates the architecture of the plugin.

### 3.1.1 Tri-Layer Modelling

Here, we augment the instance segmentation head with a tri-layer module that accepts the RoI aligned feature map as input, and outputs the modal segmentation masks of the occluder and occludee for the target object respectively (denoted as **Occluder Mask Head** and **Occludee Mask Head**). Specifically, we pass the RoI aligned feature maps into three different mask prediction heads:

$$\{\hat{b}_j, \hat{c}_j, \hat{m}_j, \hat{m}_{j1}, \hat{m}_{j2}\}^K = \Phi_{\{\text{CLS+BOX;SEG}\}}(\mathcal{F}_j) \quad (2)$$

where  $\mathcal{F}_j$  refers to the feature map from RoI align, and  $\hat{m}_{j1}, \hat{m}_{j2}$  refer to the inferred modal segmentation masks for the object’s occluder and occludee respectively. Note that, an object may not have occluder or occludee, and the predicted mask in this case is all zero.

As shown in Figure 2 (right), the feature embeddings from occluder/occludee branches are further integrated into the target object segmentation branch ( $\hat{m}_j$ ), providing cues to better infer the modal mask of the target object. To act as a proper layering model, we distinguish the order of occluder and occludee by concatenating the feature embeddings in order. Note that, this is in contrast to the previous approach [15], where the embeddings are simply element-wise added, leading to ambiguous occlusion ordering from the commutative rule.

### 3.1.2 Box Adjustment

After predicting the box and segmentation, we carry out a second iteration, to refine the initial predictions. Intuitively, if the network can detect part of an occluded object in the first iteration, an extra iteration will provide the opportunity to adjust the box accordingly, to include any part of the occluded object missed in the first iteration. For the  $k$ th instance in image  $j$ ,

$$\{\hat{b}_j, \hat{c}_j, \hat{m}_j, \hat{m}_{j1}, \hat{m}_{j2}\}^k = \Phi_{\{\text{CLS+BOX;SEG}\}}(\Phi_{\text{ALIGN}}(\mathcal{V}_j, \hat{b}_j^k)), \forall k \in [1, K] \quad (3)$$

where  $\mathcal{V}_j$  denotes the feature map from the visual encoder. The inferred object box ( $\hat{b}_j^k$ ) from the first iteration is used for RoI align to generate refined boxes ( $\hat{b}_j^k$ ), and then we can do tri-layer modelling in the refined boxes.

### 3.1.3 RoI Feature Re-weighting

To guide the model to focus more on the target objects, rather than the occluders/occludees that may take up a large proportion of the box, we also introduce RoI feature re-weighting with the inferred object mask from the previous iteration:

$$\{\hat{b}_j, \hat{c}_j, \hat{m}_j, \hat{m}_{j1}, \hat{m}_{j2}\}^k = \Phi_{\{\text{CLS+BOX;SEG}\}}(\Phi_{\text{ALIGN}}(\mathcal{V}_j, \hat{b}_j^k) \otimes \hat{m}_j^k), \forall k \in [1, K] \quad (4)$$

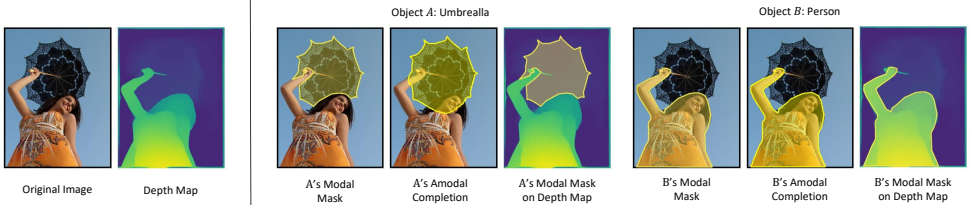
where  $\hat{m}_j^k$  denotes the inferred segmentation mask from the previous iteration,  $\otimes$  denotes the element-wise product between the RoI aligned feature map and the object segmentation from the previous iteration, and  $\hat{b}_j, \hat{c}_j, \hat{m}_j$  refer to the output bounding box, object category and instance mask, respectively.

### 3.1.4 Training

We start from publicly released pre-trained models [10]. The fine-tuning procedure is conducted progressively from “Tri-Layer Modelling” to “BBox Adjustment” and then “RoI Feature Re-weighting”, *i.e.*, proposed modules are gradually added after the previous one converges. For tri-layer modelling, both occluder and occludee mask heads use the same structures as the main mask head, except that they are class-agnostic while the main mask head is not. They are trained with binary cross-entropy loss using the inferred ground truth masks, as will be detailed in Section 4.2. The implementation and training are based on MMDet [9]. Please refer to the supplementary for more training details.

## 4 Data Preparation

To properly train our plugin, ground truth occluder/occludee masks for each object are required, as discussed in Section 3.1.4. In this section, we describe the procedure for determining objects’ occlusion order based on their **amodal** segmentation and depth ordering.



**Figure 3. The process of occlusion reasoning via amodal completion and depth estimation.** Left: The original image and its depth map. Right: the amodal completion and depth maps for the two objects: Umbrella and Person. In this case we conclude that the Person occludes the Umbrella since: (i) the amodal umbrella has an overlap with the modal person, but no overlap vice versa; and (ii) the average depth map indicates that the depth of the umbrella is greater than that of the person. In this way, we predict that **“The umbrella is occluded by the person”**.

In general, acquiring amodal segmentations can be very costly, due to the requirement of annotating the occluded parts of the object. To our knowledge, none of the existing large-scale datasets, *e.g.* COCO [18], LVIS [11], provide amodal segmentation masks. Here, we start by describing a simple, automatic, thus scalable pipeline to approximate object amodal segmentation masks (Section 4.1); in Section 4.2, we further exploit these amodal segmentations to infer the occlusion relationships between paired objects; and in Section 4.3, we detail two different types of occlusion, namely, separation and partial occlusion, and define an evaluation benchmark for occlusion, based on the COCO2017 val set.

## 4.1 Amodal Completion

Amodal completion aims to infer the amodal mask for an object, given its modal mask, and we want to do amodal completion on COCO. Here, we adopt a similar approach to that of [39], but re-train the amodal completion model on COCO to reduce the domain gap. In detail, instance masks are selected from the COCO dataset and randomly pasted onto training images in order to create artificial occlusions in the images. The amodal completion model is then trained on these images with artificial occlusions to predict the occluded parts, conditioned on the partial (modal) observation. During inference time, for each object, we slightly dilate its modal mask to reduce the gap between connected COCO masks, and the model can thus infer the object’s mask to its full extent, *i.e.*, amodal mask.

While evaluating for amodal completion on different datasets, as shown in Table 1, our model brings a significant improvement over previous approach on COCO. We refer the readers to supplementary for more evaluation details.

Model	Eval Dataset	mIoU
[39]	COCOA val	81.35
[39]	COCO2017 val	69.32
Ours	COCO2017 val	<b>81.55</b>

**Table 1. Amodal Completion on COCO:**  
The comparison between our model and [39]

Dataset	# Total Objects
Separated COCO	3522
Occluded COCO	5550
Occluder Masks	345169
Occludee Masks	328561

**Table 2. Statistics of our generated training and evaluation datasets.**

## 4.2 Occlusion Reasoning

Once we get the amodal mask for each object, the occlusion ordering between a pair of connected objects is inferred in two stages (most easily understood by first looking at the example in Figure 3): *First*, we compute the intersection between one object’s amodal segmentation and the other object’s modal segmentation. Specifically, for each pair of connected





**Figure 4. Examples in generated training and testing datasets.** Left: Example of a target object with its occluder and occludee. Right: Examples of automatically picked objects for Separated COCO and Occluded COCO. More examples are in supplementary.

objects  $A$  and  $B$ , the intersection of  $A$ 's amodal mask with  $B$ 's modal mask can be denoted as  $I_A$ , and the intersection of  $B$ 's amodal mask with  $A$ 's modal is denoted as  $I_B$ . Then  $A$  is likely to be occluded by  $B$ , if  $I_A > I_B$ . *Second*, we verify the results by depth estimation, *i.e.*, the occludee should have a greater depth than the occluder. We adopt an off-the-shelf depth estimator [47]. After conducting inference on all COCO images, we compute the average depth of each object over all pixels. Denoting object  $A$ 's average depth as  $d_A$ ,  $B$ 's average depth as  $d_B$ . If  $d_A > d_B$  also holds, the result of occlusion ordering is then verified. In this way, we determine the occlusion relationship between objects with greater confidence. For those cases with inconsistent amodal completion and depth verification, we do not assign any occluder-occludee relationship.

With such occlusion reasoning, for each confirmed object pair we have the pseudo ground truth relative occlusion order and ground truth modal masks, to train the tri-layer module (described in Section 3.1.1). Note that, when there are multiple occluders/occludees for one target object, we merge all their modal masks to form the final occluder/occludee mask.

### 4.3 Occluded COCO & Separated COCO for Evaluation

To monitor progress of object detection and instance segmentation under occlusion, we create a benchmark based on the COCO2017 val dataset. We start by defining two different splits for occluded objects, namely, **Separated COCO** and **Occluded COCO**, consisting of **separated** and **partially occluded** objects respectively. Unlike previous work [61] that manually selects occluded objects in COCO, we collect the data splits automatically, based on the occlusion reasoning. Specifically, for each object, we can easily check its ground truth modal mask for connectivity. If the mask is split into pieces, then the object is "separated" and should be in the Separated COCO split, for example, the car behind the horses in Figure 4; otherwise, the object is put into the Occluded COCO split as "partially occluded" object, if it is confirmed to be occluded by some other objects, *e.g.*, the airplane at the back in the last image of Figure 4. In this way, we can easily collect two datasets with different occlusion types. Detailed statistics are given in Table 2.

## 5 Experiments

### 5.1 Datasets and Implementation Details

**Datasets:** We train all models on the COCO2017 training split, together with the occluder/occludee masks obtained in the way described in Section 4.2. We evaluate the performance to detect occluded objects on Separated COCO and Occluded COCO generated in

Section 4.3, and the overall detection performance on both COCO2017 val and COCO2017 test-dev. As for evaluating generalisability, we also test on other datasets like KINS [23] (7517 images for testing), OVIS [22] and OpenImages [16] (details in supplementary).

**Baselines:** We compare the following state-of-the-art architectures, Swin-T + Mask R-CNN, Swin-S + Mask R-CNN and Swin-B + Cascade Mask R-CNN, with or without our designed plugin. In addition, we also compare our tri-layer plugin with the bi-layer modelling approach (inspired by BCNet [15], set all connected objects to be occluders), and compositional network [58] that is specifically designed to handle object occlusion with instance masks.

**Evaluation Metrics:** In addition to the standard metric for COCO2017 val and COCO2017 test-dev, *i.e.*, mAP, we also calculate the recall on Occluded COCO and Separated COCO, *i.e.*, the number of partially occluded / separated objects that are recalled with a fixed number of RPN proposals. Specifically, we treat a partially occluded / separated object being recalled if and only if there is a detection whose confidence  $> 0.3$  and mask IOU  $> 0.75$  with it. mIOU on KINS is calculated, in accordance with that in [58].

## 5.2 Ablation Study

Here, we conduct thorough experiments to validate the effectiveness of different modules in our plugin. As shown in Table 3, we make the following observations: 1) **Tri-layer modelling**: brings a significant improvement in terms of Recall on Occluded COCO (B1 to B3, +75, when only fine-tuning the mask heads; B1 to C2, +96, when fine-tuning the whole network); 2) **Box adjustment**: gives a significant performance boost for BBox mAP, for example, from B1 to B2, +1.9 for only fine-tuning the head, and from B1 to C1, +2.3 for fine-tuning the network; 3) **RoI feature re-weighting**: further improves Mask mAP, Recall on Occluded COCO and Recall on Separated COCO (+0.3/+10/+21 for only fine-tuning the head and +0.1/+7/+15 for fine-tuning the whole network), which is shown by the models, from B4 to B5, and C3 to C4; 4) **Fine-tuning the whole network**: can generally bring improvement on all evaluation metrics, for example, comparing B2-B5 with C1-C4. Notably, only fine-tuning the head could already contribute the majority of the improvement, validating the effectiveness of our proposed module as a general ‘plugin’, which can be inserted into pre-trained detectors, and give quick performance improvement.

Model	Tri-Layer Modelling	BBox Adjustment	RoI Feature Re-weighting	Fine-tuning Whole Network?	Recall Occluded	Recall Separated	BBox mAP	Mask mAP
B1					3264(58.81%)	1125(31.94%)	46.0	41.6
B2		✓			3296(59.39%)	1141(32.40%)	47.9	42.2
B3	✓				3339(60.16%)	1157(32.85%)	46.0	41.9
B4	✓	✓			3400(61.26%)	1187(33.70%)	48.1	42.5
B5	✓	✓	✓		<b>3410(61.44%)</b>	<b>1208(34.30%)</b>	<b>48.2</b>	<b>42.8</b>
C1		✓		✓	3367(60.67%)	1170(33.22%)	48.3	42.5
C2	✓			✓	3360(60.54%)	1159(32.91%)	46.3	42.2
C3	✓	✓		✓	3434(61.87%)	1208(34.30%)	48.3	42.9
C4	✓	✓	✓	✓	<b>3441(62.00%)</b>	<b>1223(34.72%)</b>	<b>48.5</b>	<b>43.0</b>

**Table 3. Ablation study** for adding our plugin to Swin-T + Mask R-CNN. Note that B1-B2 only mask heads are fine-tuned while B2-B3, B3-B4 only bbox heads and mask heads are fine-tuned.



### 5.3 Comparison with State-of-the-Art

**Comparison on COCO.** As shown in Table 4, we inject our plugin into a series of popular strong architectures, *i.e.* Mask R-CNN / Cascade Mask R-CNN with different Swin Transformers as backbone. In all cases, the plugin can always improve recalls for both Occluded COCO and Separated COCO, as well as detection performance on BBox and Mask mAP, sometimes by over 2.5/1.4 (val) and 2.4/1.4 (test-dev) mAP on box and mask predictions. In particular, when compared with bi-layer modelling, our plugin on Swin-T + Mask R-CNN can recall 126 more objects on Occluded COCO and 76 more on Separated COCO, and boost BBox/Mask mAP by 2.2/1.0 (val) and 2.2/1.1 (test-dev) respectively, which shows the effectiveness of our plugin.

Detector	Backbone	Plugin	Recall Occluded	Recall Separated	val mAP		test-dev mAP	
					BBox	Mask	BBox	Mask
Mask R-CNN	Swin-T [10]	–	3264(58.81%)	1125(31.94%)	46.0	41.6	46.3	42.0
Mask R-CNN	Swin-T	bi-layer	3315(59.73%)	1147(32.57%)	46.3	42.0	46.5	42.3
Mask R-CNN	Swin-T	ours	<b>3441(62.00%)</b>	<b>1223(34.72%)</b>	<b>48.5</b>	<b>43.0</b>	<b>48.7</b>	<b>43.4</b>
Mask R-CNN	Swin-S [10]	–	3393(61.14%)	1186(33.67%)	48.5	43.3	49.0	44.1
Mask R-CNN	Swin-S	ours	<b>3473(62.58%)</b>	<b>1261(35.80%)</b>	<b>50.3</b>	<b>44.2</b>	<b>50.6</b>	<b>44.9</b>
Cascade Mask R-CNN	Swin-B [10]	–	3491(62.90%)	1279(36.31%)	51.9	45.0	52.6	45.6
Cascade Mask R-CNN	Swin-B	ours*	<b>3532(63.64%)</b>	<b>1299(36.88%)</b>	<b>52.1</b>	<b>45.4</b>	<b>52.7</b>	<b>45.9</b>

**Table 4. Comparison with state-of-the-art on different architectures.** The plugin gives a performance boost across all the architectures, even for the strongest detector (Swin-B + Cascade Mask R-CNN). \* Only Tri-Layer Modelling is applied as Cascade Mask R-CNN has already used multiple iterations.

**Comparison on other benchmarks.** Here, we directly evaluate the model on the KINS [23] dataset in terms of mIoU. To handle the problem that KINS classes and COCO classes are different, we make a mapping from COCO classes to KINS classes, as detailed in the supplementary. Note that, these models are only trained on COCO, thus resembling a cross-domain generalisation. Specifically, while evaluating the baseline Swin-T + Mask R-CNN, it only achieves 66.6 mIoU, which is slightly lower than the CompositionalNet instance segmentation work [68] with a 67.2 mIoU. However, with our plugin module, the performance can be largely improved, getting 68.5 mIoU and becoming better than [68] on KINS dataset. We refer the reader to the supplementary for more detailed results on KINS, OVIS, and Open-Images.

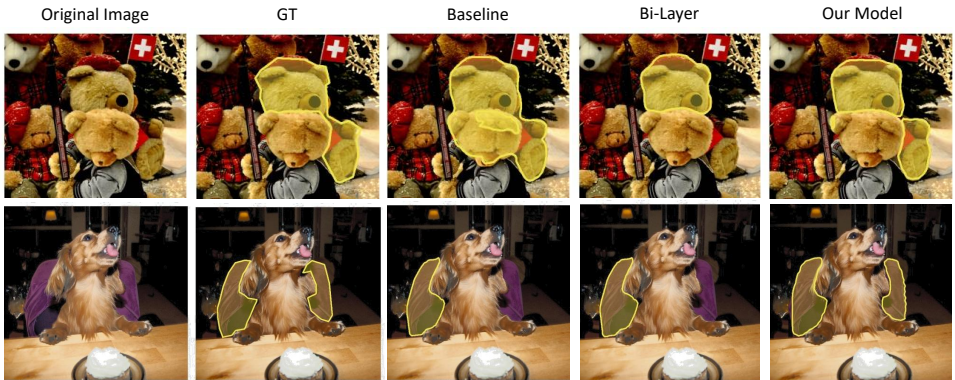
**Comparison of number of parameters and FLOPs.** Table 5 compares the number of parameters and FLOPs for different models with/without our plugin. “Tri-Layer Modelling” is a lightweight plugin, only introducing 9.3%, 13.4%, 13.3% more parameters for Swin-S + Mask R-CNN, Swin-T + Mask R-CNN, and Swin-B + Cascade Mask R-CNN, respectively. “BBox Adjustment” introduces more parameters for each model, where the majority of increase comes from the extra bbox head (14M) in the extra iteration. “RoI Feature Re-weighting” does not introduce any extra parameters since it only re-weights the RoI feature using the inferred segmentation mask from the previous iteration. When only fine-tuning the heads, only a small proportion of parameters need to be adjusted, so the training speed is fast.

Detector	Backbone	Tri-Layer Modelling	BBox Adjustment	RoI Feature Re-weighting	# Parameters	FLOPs
Mask R-CNN	Swin-T				47.8M	263.78G
Mask R-CNN	Swin-T	✓			54.2M	389.87G
Mask R-CNN	Swin-T	✓	✓		77.6M	583.33G
Mask R-CNN	Swin-T	✓	✓	✓	77.6M	583.33G
Mask R-CNN	Swin-S				69.1M	353.77G
Mask R-CNN	Swin-S	✓			75.5M	479.86G
Mask R-CNN	Swin-S	✓	✓		98.9M	673.32G
Mask R-CNN	Swin-S	✓	✓	✓	98.9M	673.32G
Cascade Mask R-CNN	Swin-B				145.0M	975.44G
Cascade Mask R-CNN	Swin-B	✓			164.3M	1353.68G

**Table 5. Comparison of number of parameters and FLOPs.** “Tri-Layer Modelling” introduces two extra mask heads, but only increases the number of parameters by a small proportion. The parameter increase brought by “BBox Adjustment” is mainly due to the extra bbox head (14M) in the extra iteration. The increase in FLOPs is approximately proportional to the increase in the number of parameters.

## 5.4 Qualitative Results

In Figure 5, we show qualitative results for inserting our plugin into Swin-T + Mask R-CNN. As can be seen, the baseline model tends to fail in challenging occlusion cases, either over-segmenting the partially occluded (row 1) or under-segmenting the separated (row 2) objects. While our proposed model has largely improved the detection, for example, disambiguating the teddy bears (row 1), and inferring the two separated pieces of the chair that is heavily occluded by the dog (row 2). See supplementary for more examples.



**Figure 5. Qualitative results on COCO.** Please see the text for more discussion. More qualitative results are provided in supplementary.

## 6 Conclusion and Future Work

We have proposed a simple ‘plugin’ module for two-stage object detectors that can improve their performance in detecting objects under challenging occlusions. Additionally, we describe a scalable pipeline for automatically identifying occluded and occluding objects in existing benchmarks, to provide training data and an evaluation dataset. Adding the module to a series of popular strong detectors, Mask R-CNN / Cascade Mask R-CNN with different Swin Transformer backbones, leads to consistent performance improvements.

A possible avenue of future work is to improve detection performance of occluded objects in videos, where multiple views of the objects and temporal cues are potentially available to help disambiguate the occlusions.

**Acknowledgements.** This research is supported by EPSRC Programme Grant VisualAI EP/T028572/1, a Royal Society Research Professorship RP\R1\191132, and a China Oxford Scholarship. We thank Prannay Kaul, Tengda Han and Gyungin Shin for proof-reading.

## References

- [1] Swin-transformer-object-detection official github page. <https://github.com/SwinTransformer/Swin-Transformer-Object-Detection>. 25 July, 2022.
- [2] Gabriel J Brostow and Irfan A Essa. Motion based decompositing of video. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1, pages 8–13. IEEE, 1999.
- [3] Zhongang Cai, Junzhe Zhang, Daxuan Ren, Cunjun Yu, Haiyu Zhao, Shuai Yi, Chai Kiat Yeo, and Chen Change Loy. Messytable: Instance association in multiple camera views. In *European Conference on Computer Vision*, pages 1–16. Springer, 2020.
- [4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.
- [5] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.
- [6] Xiangyu Chen, Zelin Ye, Jiankai Sun, Yuda Fan, Fang Hu, Chenxi Wang, and Cewu Lu. Transferable active grasping and real embodied dataset. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3611–3618. IEEE, 2020.
- [7] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.
- [8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.
- [9] Patrick Follmann, Rebecca König, Philipp Härtinger, Michael Klostermann, and Tobias Böttger. Learning to see the invisible: End-to-end trainable amodal instance segmentation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1328–1336. IEEE, 2019.
- [10] James J Gibson. *The ecological approach to visual perception*. Psychology press, 1979.
- [11] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019.

- [12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.
- [13] Edward H. Adelson John Y.A. Wang. Representing moving images with layers. *The IEEE Transactions on Image Processing Special Issue: Image Sequence Compression*, 3(5):625–638, September 1994.
- [14] Nebojsa Jojic and Brendan J Frey. Learning flexible sprites in video layers. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.
- [15] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4019–4028, 2021.
- [16] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.
- [17] Ke Li and Jitendra Malik. Amodal instance segmentation. In *European Conference on Computer Vision*, pages 677–693. Springer, 2016.
- [18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.
- [19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- [20] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [21] M Pawan Kumar, Philip HS Torr, and Andrew Zisserman. Learning layered motion segmentations of video. *International Journal of Computer Vision*, 76(3):301–319, 2008.
- [22] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 2022.
- [23] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2019.
- [24] 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.

- [25] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021.
- [26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.
- [27] Kaziwa Saleh, Sándor Szénási, and Zoltán Vámosy. Occlusion handling in generic object detection: A review. In *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 000477–000484. IEEE, 2021.
- [28] Deqing Sun, Erik B Sudderth, and Michael J Black. Layered segmentation and optical flow estimation over time. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1768–1775. IEEE, 2012.
- [29] Deqing Sun, Jonas Wulff, Erik B Sudderth, Hanspeter Pfister, and Michael J Black. A fully-connected layered model of foreground and background flow. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2451–2458, 2013.
- [30] Yihong Sun, Adam Kortylewski, and Alan Yuille. Amodal segmentation through out-of-task and out-of-distribution generalization with a bayesian model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1215–1224, 2022.
- [31] Angtian Wang, Yihong Sun, Adam Kortylewski, and Alan L Yuille. Robust object detection under occlusion with context-aware compositionalnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12645–12654, 2020.
- [32] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. Ua-detrac: A new benchmark and protocol for multi-object detection and tracking. *Computer Vision and Image Understanding*, 193:102907, 2020.
- [33] John Winn and Jamie Shotton. The layout consistent random field for recognizing and segmenting partially occluded objects. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, volume 1, pages 37–44. IEEE, 2006.
- [34] Jonas Wulff and Michael J Black. Efficient sparse-to-dense optical flow estimation using a learned basis and layers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 120–130, 2015.
- [35] Jonas Wulff and Michael Julian Black. Modeling blurred video with layers. In *European Conference on Computer Vision*, pages 236–252. Springer, 2014.
- [36] Yi Yang, Sam Hallman, Deva Ramanan, and Charless Fowlkes. Layered object detection for multi-class segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3113–3120. IEEE, 2010.

- [37] Yi Yang, Sam Hallman, Deva Ramanan, and Charless C Fowlkes. Layered object models for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1731–1743, 2011.
- [38] Xiaoding Yuan, Adam Kortylewski, Yihong Sun, and Alan Yuille. Robust instance segmentation through reasoning about multi-object occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11141–11150, 2021.
- [39] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene de-occlusion. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, June 2020.
- [40] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *International Conference on Learning Representations*, 2021.
- [41] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1464–1472, 2017.
- [42] C Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. *ACM transactions on graphics (TOG)*, 23(3):600–608, 2004.