MT-ORL: Multi-Task Occlusion Relationship Learning

Panhe Feng [1,2]   Qi She [2]   Lei Zhu [1]   Jiaxin Li [2]   Lin ZHANG [2]   Zijian Feng [2]

Changhu Wang [2]   Chunpeng Li [1]   Xuejing Kang [*,1]   Anlong Ming [1]

[1]Beijing University of Posts and Telecommunications.   [2]ByteDance Inc.

{fengpanhe,sheqi1991}@gmail.com,zhulei@stu.pku.edu.cn,{lijx1992,1999forrestz,vincent.fung13}
@gmail.com,wangchanghu@bytedance.com,{chunpeng.li,kangxuejing,mal}@bupt.edu.cn

Abstract

*Retrieving occlusion relation among objects in a single image is challenging due to sparsity of boundaries in image. We observe two key issues in existing works: firstly, lack of an architecture which can exploit the limited amount of coupling in the decoder stage between the two subtasks, namely occlusion boundary extraction and occlusion orientation prediction, and secondly, improper representation of occlusion orientation. In this paper, we propose a novel architecture called **O**cclusion-shared and **P**ath-separated Network (**OPNet**), which solves the first issue by exploiting rich occlusion cues in shared high-level features and structured spatial information in task-specific low-level features. We then design a simple but effective orthogonal occlusion representation (**OOR**) to tackle the second issue. Our method surpasses the state-of-the-art methods by 6.1%/8.3% Boundary-AP and 6.5%/10% Orientation-AP on standard PIOD/BSDS ownership datasets. Code is available at https://github.com/fengpanhe/MT-ORL.*

1. Introduction

While the human visual system is capable of intuitively performing robust scene understanding and perception. The reasoning of occlusion relation is highly challenging for machines. Depending on the number, category, orientation, and position of objects, the boundaries of objects are elusive; no simple priors can be applied to recover the foreground and background in the scene. Occlusion relationship reasoning of objects from monocular images reveals relative depth differences among the objects in the scene, which is fundamental in computer vision applications, such as mobile robots [18, 28, 2], object detection [8, 4, 32, 21, 5], segmentation [8, 38, 6, 31, 14, 39, 22], monocular depth

estimation/ordering [23, 25, 24, 26, 17, 37], and 3D reconstruction [30, 20, 9].

Traditional methods [12, 33, 26, 29, 13] extract the occlusion boundary and infer occlusion relation by exploiting low-level visual cues with hand-crafted features, which is not effective due to the difficulty of defining boundaries and occlusion cues. Modern convolution neural networks (CNNs) [35, 34, 15] have significantly boosted the performance of occlusion relationship reasoning by commonly decomposing the task into two subtasks: occlusion boundary extraction and occlusion orientation prediction. The former aims to extract object boundaries from the image, while the latter targets are discovering the orientation relation. The occlusion relationship is then recovered by progressively accumulating the orientation information on the extracted boundaries. Despite the leap in occlusion relationship reasoning, we observe that two critical issues have rarely been discussed, which greatly impedes performances of previous work.

The first issue is the lack of exploitation of the limited amount of coupling between the two subtasks. On one hand, shared global properties of the image and abstract high-level semantic features act as an effective initialization for the networks to learn low-level visual cues. Shared high-level features help different branches to generate consistent outputs under the same semantic guidance while different features may produce semantically misaligned results in the fusion stage. On the other hand, the two subtasks need to learn diverse and concrete properties in the lower stage with larger spatial size. Specifically, occlusion boundary extraction emphasizes more on locating boundaries, while occlusion orientation attaches more importance to exploiting relations between areas. In a word, the two subtasks are inherently coupled in abstract global features and decoupled in task-specific low-level features. However, previous architectures adopt a fully shared decoder [34, 15], resulting in an initialization feature map from the lower network

---

[*]corresponding author.
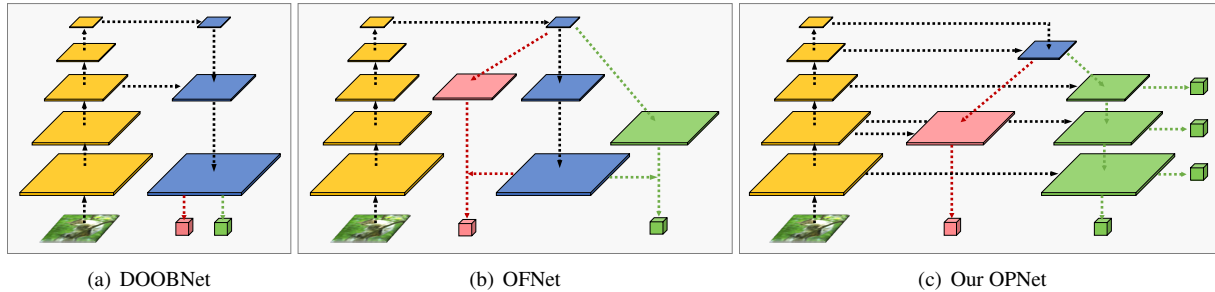
Figure 1: Comparisons of different network architectures. Yellow indicates encoder structure. Green indicates boundary extraction path. Pink refers to occlusion orientation path. Blue indicates shared structure in decoder. Two color cubes indicate the boundary map and the orientation map of network output, respectively. Different from DOOBNet and OFNet, which over-shares local spatial information in larger feature maps, our OPNet only shares feature maps in deep stages while remaining path-separated in the shallow stage.

stage.

The second issue is the lack of a good representation of occlusion orientation. DOC [35] first proposed to use pixel-wise continuous orientation variable as a representation, which however makes the design of loss function complex due to angle periodicity. DOOBNet [34] then proposed to truncate the variable into range $(-\pi, \pi]$, which caused serious endpoint errors: angles close to $-\pi$ and π produce large loss value while remaining relatively close orientation. This leaves the training process problematic. Therefore, designing a proper occlusion orientation representation remains challenging in the occlusion relationship reasoning.

To tackle the first issue, we propose an Occlusion-shared and Path-separated Network (**OPNet**), which uses abstract feature maps with smaller spatial sizes for weight sharing and is then split into two separate task-specific paths, namely the occlusion boundary extraction path and occlusion orientation prediction path. Besides, multi-scale supervision [36] is adopted in the boundary extraction path to enhance the multi-scale features. To solve the second representation dilemma, we propose a simple and robust orthogonal occlusion representation (**OOR**), which represents orientation with two orthogonal vectors in horizontal and vertical directions, respectively. Experiments show our method is effective and boosts the performances of both object boundary extraction and occlusion orientation prediction.

In a nutshell, our contributions are three-fold:

- We rethink the inherent properties of occlusion relationship reasoning, associated with two sub-tasks: occlusion boundary extraction and occlusion orientation regression. This motivates our advocate of a novel Occlusion-shared and Path-separated Network (OPNet) embodied with discrimination capability and expressiveness for visual occlusion reasoning as an alternative.

- We further propose the robust orthogonal occlusion

representation(OOR) for predicting the occlusion orientation, which resolves the endpoint error and angle periodicity dilemma.

- The method works well on both PIOD [35] and BSDS ownership datasets [27], offering significantly better performances than the counterparts.

2. Related Work

Deep learning based occlusion relationship reasoning has achieved great success in recent years [35, 34, 15]. DOC [35] represents occlusion relation by using a binary boundary indicator. It exploits local and non-local image cues to learn this representation and hence recovers the occlusion relation. DOOBNet [34] proposes to address extreme boundary/non-boundary class imbalance by up-weighting the loss contribution of false negative and false positive examples with a novel Attention Loss function. OFNet [15] builds on top of the encoder-decoder structure and side-output utilization. A fusion module is designed to precisely locate the object regions from the occlusion cues.

However, these methods either use two separate architectures for boundary extraction and orientation prediction or over-share local spatial information in shallow stages. Instead, our OPNet only shares features in deep stages of the decoder and adopts a multi-scale structure, as shown in Figure 1. We also propose an orthogonal orientation representation to tackle endpoint error and angle periodicity dilemma.

3. The OPNet

In this section, we detail our network architecture, Occlusion-shared and Path-separated Network (**OPNet**), which is composed of a shared encoder and our proposed novel Occlusion-shared and Path-separated decoder (**OPDecoder**). Following previous methods, we choose a ResNet50 [10, 11] pretrained on ImageNet as our encoder.
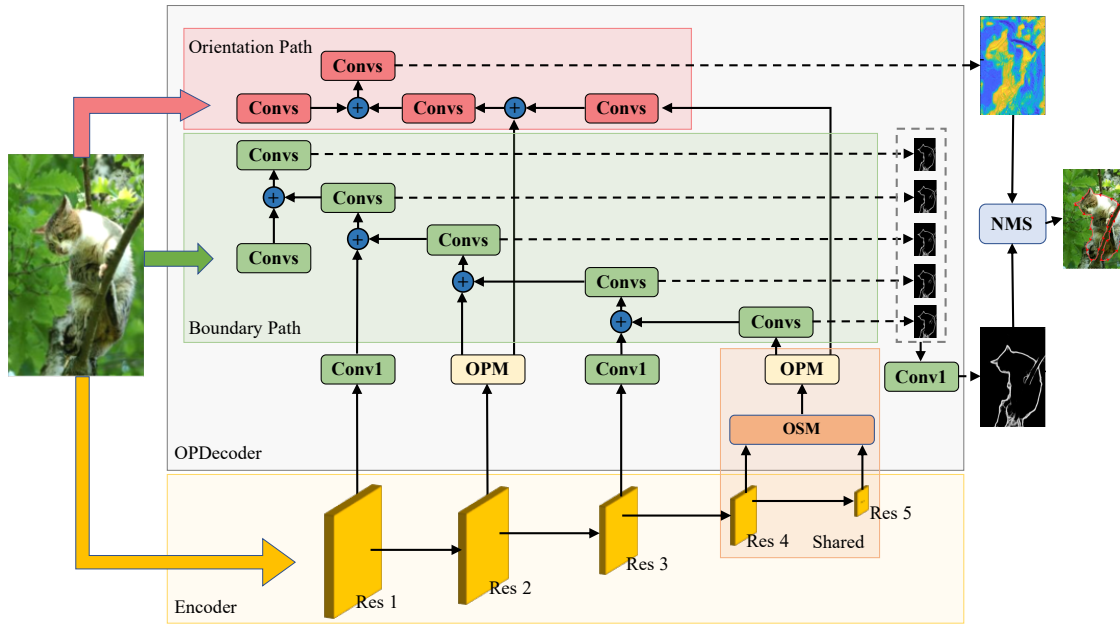
(a) Overall architecture
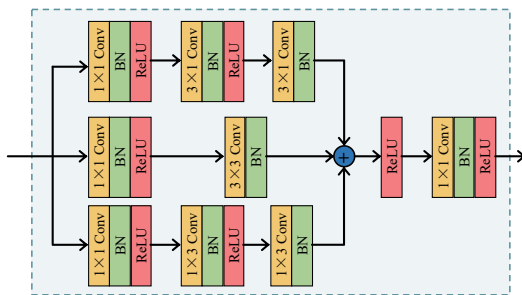
(b) OPM: Orthogonal Perception Module
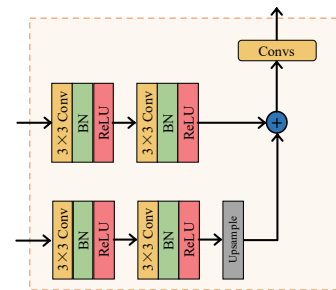
(c) OSM: Occlusion-shared Module

Figure 2: The diagram of the proposed OPNet. (a) Overall architecture. The input image is first encoded by an encoder and then input into the proposed decoder, which first aggregates shared features in the abstract high-level by OSM, and then generates boundary map and orientation map in two separated paths. NMS is used to fuse two outputs to produce the ultimate occlusion relationship map. (b) Orthogonal Perception Module (OPM). (c) Occlusion-shared Module (OSM). ⊕ represents addition operation. Details can be found in section 3.

Our OPDecoder is equipped with Occlusion-shared Module (**OSM**), Orthogonal Perception Module (**OPM**), and two separated paths, i.e, a boundary path and an orientation path. Extracted boundary map from boundary path and predicted orientation map from orientation path are then merged by Non-Maximum Suppression(NMS) to produce the final occlusion relationship map in the evaluation stage.

3.1. The Occlusion-shared and Path-separated Decoder

We define the deep stages in an encoder/decoder architecture as the ones where the feature map is deeper but spatially smaller, such as res4 or res5 layer in a ResNet architecture. Such deep stages contain less parameters and spa-

tial details, thus are rich in abstract and global information that can be well adapted to downstream tasks. In contrast, shallow stages produce spatially-large feature maps with reduced channels. As a result, they are beneficial for learning task-specific low-level details. Our OPDecoder fully utilizes such distinct features at multiple stages.

Shared Deep Stage and Separated Shallow Stages Both boundary extraction and occlusion orientation prediction are dense prediction tasks aiming to recover pixel-wise spatial details with the understanding of high-level occlusion information. The former focuses on locating, while the latter expresses the relationship between regions where occlusion occurs. We thus propose our OPDecoder with a deep-

to-shallow structure. Specifically, in order to better consider the connection and distinction between the two tasks, we divide the decoder into two parts, the shared deep stage with joint occlusion information and the separate shallow stages with concrete spatial details. The shared high-level semantic features in the deep stage learn more abstract occlusion cues by accepting supervision signals from both the boundary path and the orientation path. The boundary extraction path pays more attention to the spatial details, while the orientation regression path focuses on recovering regional relationships. Excessive sharing could result in mutual interference between the two tasks. Therefore, we also adopt separate structures in the shallow stages.

As shown in Figure 2 (a), our decoder contains a deep occlusion-shared phase and a shallow spatial information branching phase. In the occlusion-shared phase, we use OSM to decode and aggregate two deepest stages' features and then pass them to the subsequent boundary path and orientation path. In the shallow branching phase, considering the possible interference between the two tasks, we use OPM to enhance the features from shallow layers of the encoder, which will be fed to the following two paths.

In the boundary extraction path, the shared feature map from OSM is progressively aggregated with transformed encoder features from deep to shallow stages. To better utilize spatial information at different scales, multi-scale boundary maps are generated, which are then fused together to extract a merged boundary map. Multi-scale supervision is applied to each generated boundary map.

Different from the boundary path, the orientation path is deliberately designed to only accept skip-connected features from the res2 layer to exploit less shared spatial details in shallow layers. It only keeps one supervision signal in the original scale due to the ambiguity of scaled orientation maps.
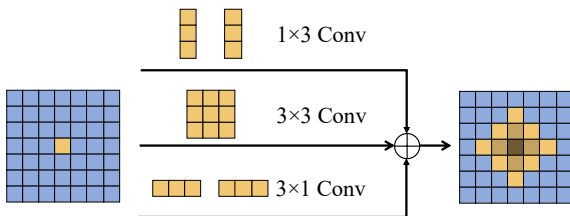

Figure 3: Receptive fields of the OPM. Two consecutive convolutions of size 1×3 produce a receptive field of 1×5 size in the vertical. And two convolutions of size 3×1 produce a receptive field of 5×1 size in the horizontal. Resulted convolution layers thus pay more attention along with the orthogonal directions, adhering to our orthogonal occlusion representation design.

Orthogonal Perception Module (OPM) As shown in Figure 2 (b), OPM contains three branches. The center

branch consists of two 3×3 convolution blocks, while the other two are composed of two 1×3 and 3×1 strip convolution blocks, respectively. Figure 3 shows the receptive fields of module output features. The two stripe convolution paths simultaneously increase the receptive field and feature weight of the center pixel in orthogonal directions, and hence accumulate more informative orthogonal signals for further prediction of the OOR.

Occlusion Shared Module (OSM) As shown in Figure 2 (c), OSM aggregates shared high-level features by fusing the upsampled res5 feature map and res4 feature map. It is designed to only merge features from deep stages, retaining abstract high-level shared occlusion features. OSM accepts supervision signals from both the boundary path and the orientation path to act on the high-level semantic features with larger receptive fields.

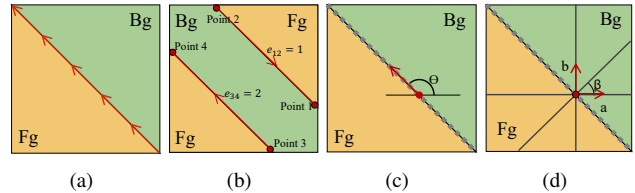## 3.2. Orthogonal Occlusion Representation

Figure 4: The left rule and three ways to represent occlusion relations based on boundary. (a) is the "left" rule. (b) is the background label representation. (c) is the occlusion orientation representation. (d) is the orthogonal Occlusion representation.

In this section, we first point out the disadvantages of existing occlusion representations, i.e. angle periodicity and endpoint error. Different representation methods are then visualized for intuitive illustration. Finally, we present the novel orthogonal occlusion representation, which avoids the aforementioned two disadvantages by using two orthogonal vectors. Figure 4 shows a schematic diagram of four boundary-based representations of occlusion orientation, the lower left and upper right corners are the foreground and the background, respectively, while the diagonal line is the occlusion boundary between the two regions.

Figure 4 (a) shows the left rule of indicating foreground and background in an image. The occlusion boundary is represented with an arrow, the left-hand side of which is the foreground. Figure 4 (b) is the label classification method based on complete boundaries commonly adopted by the traditional methods [19, 1, 12, 7, 3], which classifies the obtained occlusion boundaries into two types. Label 1 indicates that the orientation of the boundary is from the starting point to the ending point, while label 2 means the opposite. This method is considered ineffective due to the fact that de-

tected boundaries are sparse and sometimes discontinuous. Figure 4 (c) illustrates pixel-level orientation variable representation proposed in DOC [35], which predicts boundary orientation by predicting a continuous orientation variable $\theta \in (-\pi, \pi]$ for each pixel in the image. This method can well adapt to convolution blocks with dense prediction attributes and avoid the dependency on the quality of boundaries. It brings significant performance improvement on occlusion relationship reasoning and is widely adopted by CNN-based methods [34, 16, 15, 35].

However, pixel-level orientation variable representation still has two critical issues. Firstly, it requires the network to regress an accurate angle for each boundary pixel, whereas predicting accurate angle variables brings an unnecessary burden. As long as the predicted angle has an error less than $\pi/2$, the foreground-background relationship can be correctly recovered. The second issue is the negative impact of angle periodicity on loss functions. The period of the angle variable is 2π, meaning that two angles whose difference is a multiple of 2π are equivalent. However, such equivalence is hard to maintain in the loss function. In order to solve this problem, DOC [35] defines a periodic segmentation function to represent the difference between two angles. But experiments show that this discontinuous loss function hinders the learning of the network. DOOB-Net adopts the strategy of truncating the predicted value to $(-\pi, \pi]$, which however brings serious endpoint errors, i.e. angles near $-\pi$ and π could result in the higher loss while in fact staying close, introducing unnatural supervision signals near the endpoints.

We hereby propose our orthogonal occlusion representation (**OOR**): using a pair of orthogonal vectors, $\vec{a}$ along the horizontal axis and $\vec{b}$ along the vertical axis, to represent occlusion orientation. As shown in Figure 4(d), both $\vec{a}$ and $\vec{b}$ point to the background near the occlusion boundary. OOR simplifies the prediction of occlusion orientation by instead pointing out the background both horizontally and vertically, greatly enhancing the robustness of prediction. Compared with previous the orientation variable representation, our OOR is simpler and works better, bypassing both angle periodicity and endpoint error dilemma.

Figure 5 visualizes the feature maps extracted from the Res2 layer trained with different orientation representations. Compared with the feature maps obtained by using DOOBNet's orientation representation and loss, the contrast between foreground and background of our result is more obvious. Experiments in Table. 1 show that our method improves all three occlusion-related metrics (O-ODS, O-ODS, O-AP) by at least 2% upon orientation representation of DOOBNet, which shows stronger occlusion perception potential of our OOR. Figure 6 shows the comparisons of different methods on occlusion orientation recall towards different boundary directions. We divide $(-\pi, \pi]$
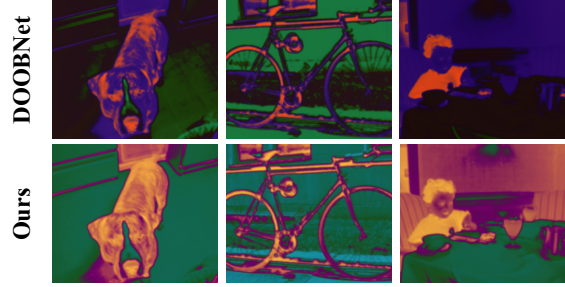

Figure 5: Visualization of the feature maps from networks trained with **different Orientation Representation**, which is produced by reducing feature dimension to 3 via PCA. The first row is the result of using the DOOBNet[34] Orientation Representation, and the second row is Ours. Our result obviously shows clearer boundaries and occlusion relation, validating the effectiveness of the proposed OOR.
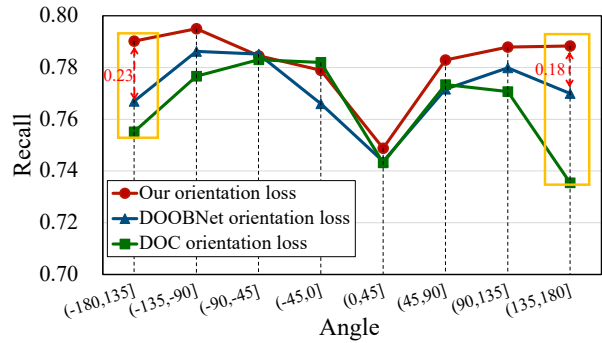

Figure 6: Recall vs. Angle curve on PIOD dataset. Angle is classified into eight bins and recall is calculated in each bin. Our proposed orientation representation shows obvious improvement near two endpoints $-\pi$ and π (the yellow box).

into eight bins and calculate the recall of orientation prediction in each bin. Results show that DOC's and DOOB-Net's orientation representations and loss pay too much attention to the vertical direction and produce large errors near two endpoints due to the aforementioned endpoint error dilemma.

3.3. Network Training

For an input image I, the ground truth label is represented as a pair of object boundary map and occlusion orientation map $\{Y, O\}$, where $Y = (y_j, j = 1, \ldots, |I|), y_j \in \{0, 1\}$. $O = (\theta_j, j = 1, \ldots, |I|), \theta_j \in (-\pi, \pi]$ and θ_j is invalid when $y_j = 0$. In CNN-based methods of occlusion relationship reasoning, occlusion boundary extraction and occlusion orientation prediction are trained simultaneously. Thus the overall loss $\mathcal{L}$ is formulated as:

$$\mathcal{L} = \mathcal{L}_B + w_O \cdot \mathcal{L}_O, \tag{1}$$

where w_O is the weight for occlusion orientation loss. $\mathcal{L}_B$ and $\mathcal{L}_O$ represent the loss of boundary path and orientation path, respectively. In boundary path, we can get a collection of five side boundary maps and a fused boundary map $\{\hat{Y}_{s-1}, \hat{Y}_{s-2}, \cdots, \hat{Y}_{s-K}, \hat{Y}_f\}$. We formulate the loss of boundary extraction as, where in this work $K = 5$:

$$\mathcal{L}_B = w_f \cdot \Gamma(\hat{Y}_f, Y) + \sum_{i=1}^{K} w_{s-i} \cdot \Gamma(\hat{Y}_{s-i}, Y), \qquad (2)$$

where w_{s-i} and w_f are weights for the side boundary map loss and fused boundary map loss, respectively. $\hat{Y} = (y_j, j = 1, \ldots, |I|), y_j \in (0, 1)$ represents the predicted boundary map . The function $\Gamma(\hat{Y}, Y)$ is computed with all pixels. Since the distribution of boundary/non-boundary pixels is heavily biased, we employ a class-balanced cross-entropy loss. we define Y_- and Y_+ as the non-boundary and boundary ground truth label sets in a mini-batch of images. We then define $\Gamma(\hat{Y}, Y)$ as:

$$\Gamma(\hat{Y}, Y) = -\lambda\alpha \sum_{j \in Y_-} \log(1 - \hat{y}_j) - (1 - \alpha) \sum_{j \in Y_+} \log(\hat{y}_j).$$
$$(3)$$

Here, $\alpha = |Y_+| / (|Y_+| + |Y_-|)$ can balance the boundary/non-boundary pixels, and λ controls the weight of positive over negative samples.

The function $\mathcal{L}_O$ calculates the loss of the orientation only at the groundtruth boundary pixels. Based on our representation, we proposed a novel loss called **Orthogonal Orientation Regression loss (OOR)**, which is defined as:

$$\mathcal{L}_O = \text{OOR}(\hat{a}, \hat{b}; \theta) = \text{smooth}_{L_1}\left(\Phi(\hat{a}, \hat{b}; \theta)\right), \quad (4)$$

in which we design the novel formulation as:

$$\Phi(\hat{a}, \hat{b}; \theta) = (\frac{\hat{a}}{\sqrt{\hat{a}^2 + \hat{b}^2}} - \cos\theta_j)^2 + (\frac{\hat{b}}{\sqrt{\hat{a}^2 + \hat{b}^2}} - \sin\theta_j)^2.$$
$$(5)$$

$\hat{a}$ and $\hat{b}$ represent prediction value of the network. We define the $\vec{a} = (\hat{a}, 0)$ and $\vec{b} = (0, \hat{b})$. The θ is the ground truth of occlusion orientation.

4. Experiment

4.1. Datasets and Implementation Details

We evaluate the proposed approach on two public datasets: PIOD [35] and BSDS ownership dataset [27].
PIOD dataset contains 9,175 images for training and 925 images for testing. Each image is annotated with groundtruth object instance boundary map and corresponding orientation map.

BSDS ownership dataset includes 100 training images and 100 testing images of natural scenes. Each image is annotated with groundtruth object instance boundary map and corresponding orientation map.

Implementation details: Our network is implemented in PyTorch. We use a ResNet50 [10, 11] architecture pretrained on ImageNet as the backbone encoder. AdamW optimizer is adopted to train our network. The learning rate is 3e-6. Training images are randomly cropped into size 320×320 and form mini-batches of size 8. We train our network on the PIOD dataset for 40k iterations and set hyperparameter $w_{s-[2-5]}$, w_{s-1}, w_O, w_f and λ to 0.5, 1.1, 1.1, 2.1, 1.7, respectively. we train our network on the BSDS ownership dataset for 20k iterations and set λ as 1.1. Other hyper-parameters are kept the same as in PIOD.

Evaluation metrics: We compute the precision and recall (**PR**) for boundary extraction (**BPR**) and occlusion orientation (**OPR**). Three standard evaluation metrics are then calculated from PR: fixed contour threshold (**ODS**), best threshold of image (**OIS**) and average precision (**AP**). In the following sections, we use **B-Metric** to represent metrics calculated from BPR and **O-Metric** to represent metrics calculated from OPR. Note that, OPR is only calculated at the correctly detected boundary pixels.

4.2. Comparison with the state-of-the-art methods

We compare our approach with recent methods including SRF-OCC [33], DOC-HED [35], DOC-DMLFOV [35], DOOBNet [34], OFNet [15].

Performance on PIOD dataset: The Precision-Recall curves of both boundary extraction and occlusion orientation tasks on PIOD dataset are shown in Figure 7 (a) and (b). Clear improvements upon OFNet can be observed on both tasks. Table 1 shows that even though trained alone on each subtask, our method still outperforms others by a large margin(3.5% B-ODS and 4.3% O-ODS).

Performance on BSDS ownership dataset: The BSDS ownership dataset is difficult to train due to the small number of training samples. Nonetheless, our method can still achieve superior performance to others in both BPR and OPR as shown in Figure 7(c)(d). As shown in Table 1, benefited from our proposed OOR, our method achieves a huge 10% O-AP gain and an 8.3% B-AP gain compared to state-of-the-art, showing strong potential in both orientation prediction and boundary extraction.

By substituting OOR with dor/dbr (OPNet+dor/dbr in Table 1), OPNet with joint training strategy outperforms previous works but underperforms even the single boundary branch training of OPNet (Ours †), which verifies the superiority of our network structure and OOR, respectively.

Finally, we visualize the boundary maps and occlusion relationship maps, as shown in Figure 8. It can be observed that our method extracts more complete and clearer

Table 1: Comparison between our method with others on PIOD dataset and BSDS ownership dataset. † indicates training boundary extraction branch only. ‡ indicates training orientation prediction branch only. dor indicates using the orientation loss function proposed by DOC [35]. dbr indicates the orientation loss function proposed by DOOBNet [34]. The red/green superscripts ($\uparrow$ / $\downarrow$) indicate increasing/decreasing upon the state-of-the-art method (OFNet[15]). All numbers are represented in percentage. Details of different evaluation metrics can be found in Section 4.1.

Method	PIOD Dataset						BSDS ownership Dataset					
	B-ODS	B-OIS	B-AP	O-ODS	O-OIS	O-AP	B-ODS	B-OIS	B-AP	O-ODS	O-OIS	O-AP
SRF-OCC	34.5	36.9	20.7	26.8	28.6	15.2	51.1	54.4	44.2	41.9	44.8	33.7
DOC-HED	50.9	53.2	46.8	46.0	47.9	40.5	65.8	68.5	_60.2_	52.2	54.5	42.8
DOC-DMLFOV	66.9	68.4	67.7	60.1	61.1	58.5	57.9	60.9	51.9	46.3	49.1	36.9
DOOBNet	73.6	74.6	72.3	70.2	71.2	68.3	64.7	66.8	53.9	55.5	57.0	44.0
OFNet †	73.9	75.0	68.5	-	-	-	-	-	-	-	-	-
OFNet ‡	-	-	-	70.5	71.6	67.4	-	-	-	-	-	-
OFNet	_75.1_	_76.2_	_77.0_	_71.8_	_72.8_	_72.9_	_66.2_	_68.9_	58.5	_58.3_	_60.7_	_50.1_
OPNet + dor	76.6$^{\uparrow 1.5}$	77.8$^{\uparrow 1.6}$	76.1$^{\downarrow 0.9}$	73.7$^{\uparrow 1.9}$	74.6$^{\uparrow 1.8}$	72.2$^{\downarrow 0.7}$	66.2$^{\uparrow 0}$	68.7$^{\downarrow 0.2}$	63.2$^{\uparrow 4.7}$	55.5$^{\downarrow 2.8}$	57.5$^{\downarrow 3.2}$	47.0$^{\downarrow 3.1}$
OPNet + dbr	77.8$^{\uparrow 2.7}$	79.0$^{\uparrow 2.8}$	80.6$^{\uparrow 3.6}$	74.7$^{\uparrow 1.9}$	75.8$^{\uparrow 3.0}$	75.9$^{\uparrow 3.0}$	66.8$^{\uparrow 0.6}$	68.9$^{\uparrow 0}$	62.4$^{\uparrow 3.9}$	58.1$^{\downarrow 0.2}$	60.0$^{\downarrow 0.7}$	50.5$^{\uparrow 0.4}$
Ours †	78.6$^{\uparrow 3.5}$	79.6$^{\uparrow 3.4}$	79.5$^{\uparrow 2.5}$	-	-	-	-	-	-	-	-	-
Ours ‡	-	-	-	76.1$^{\uparrow 4.3}$	77.0$^{\uparrow 4.2}$	76.1$^{\uparrow 3.2}$	-	-	-	-	-	-
Ours	**79.5**$^{\uparrow 4.4}$	**80.4**$^{\uparrow 4.2}$	**83.1**$^{\uparrow 6.1}$	**77.1**$^{\uparrow 5.3}$	**77.8**$^{\uparrow 5.0}$	**79.4**$^{\uparrow 6.5}$	**67.2**$^{\uparrow 1.0}$	**70.3**$^{\uparrow 1.4}$	**66.8**$^{\uparrow 8.3}$	**61.9**$^{\uparrow 3.6}$	**64.8**$^{\uparrow 3.4}$	**60.1**$^{\uparrow 10}$
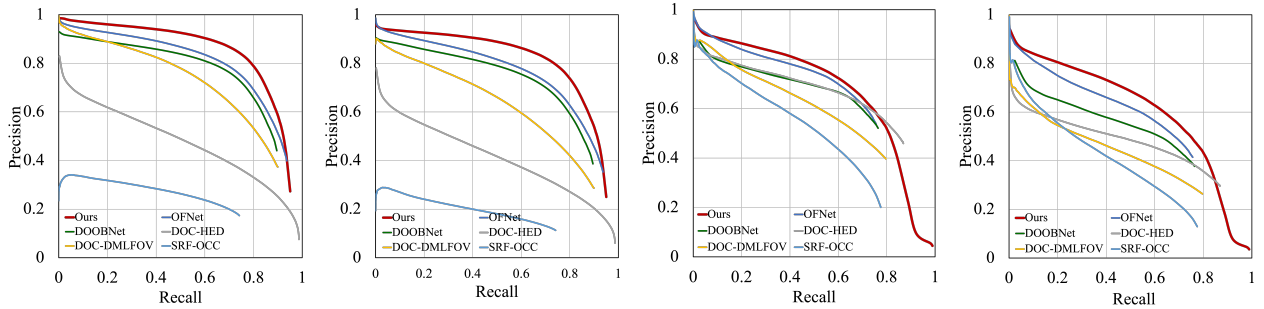
(a) PRC of boundary on PIOD. (b) PRC of occlusion on PIOD. (c) PRC of boundary on BSDS ownership. (d) PRC of occlusion on BSDS ownership.

Figure 7: The precision-recall curves(PRC) of our method and others on PIOD and BSDS ownership dataset.

boundaries on multiple images than DOOBNet[34] and OFNet[15]. In complex and indistinguishable scenes, such as images on the right-hand side, we are still ahead of others, showing stronger generalization potential. As shown in the last image sets on the left-hand side, our model can even discern the leg which is not labeled on the groundtruth map. In summary, our method can outperform others by a large margin both in boundary extraction and occlusion orientation prediction, which validates its effectiveness.

4.3. Ablation Study

In this section, we conduct experiments on the PIOD dataset to study the impact of different architectural designs and verify each component in our network.

Comparison on number of shared layers: We compared sharing top-1, top-2, top-3, and top-4 stages at high

stages. As shown in Table 2, the best result appears when sharing the top-2 stages, surpassing the second best(top-1) by 2.3% in B-AP. When the sharing extends to the lower stages, the result begins to deteriorate, which empirically proves our argument that the two branches should share occlusion information existing in higher stages and remain separated in lower stages to recover task-specific spatial information.

Comparison of orientation losses: We change the output of our occlusion orientation path so that it can regress the continuous orientation value, as in previous work. We then compare the performance of predicting continuous orientation value(DOOBNet) and predicting orthogonal vectors(OOR) with our network structure. As shown in Table 1, the proposed OOR and OOR loss surpasses DOOBNet's orientation representation by 6.5% and 10% O-AP on PIOD
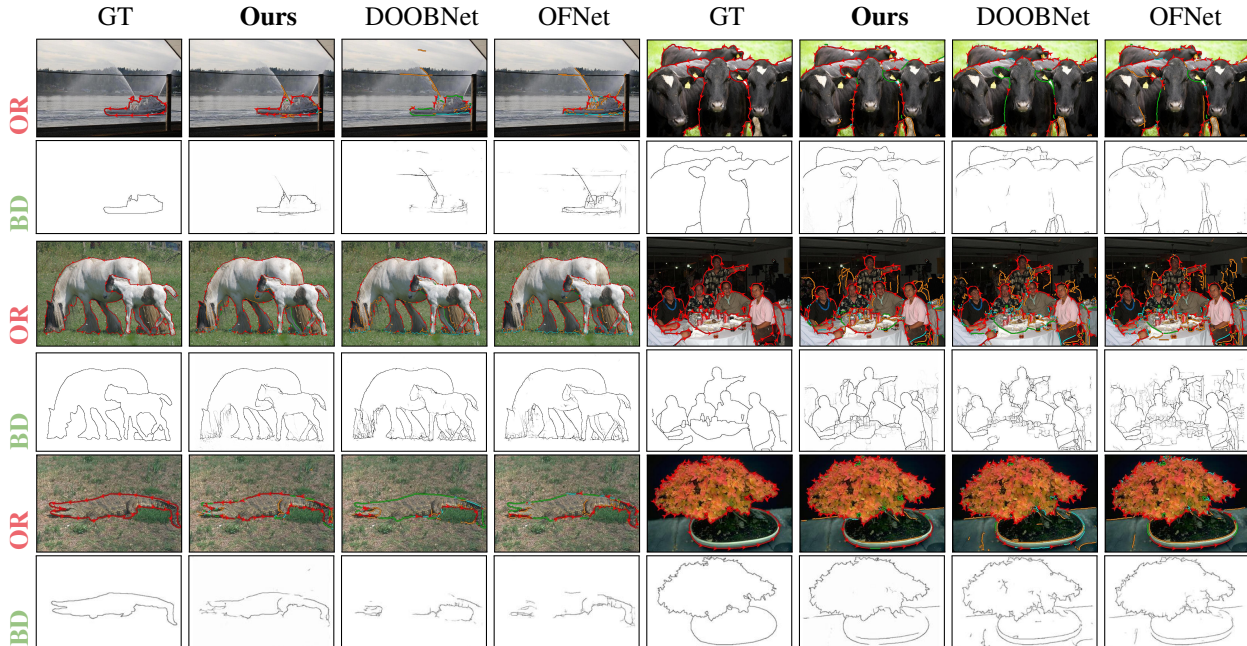
Figure 8: Occlusion reasoning(OR) map and boundary(BD) map predicted on PIOD (first four rows) and BSDS ownership datasets (last two rows). We compare our method with DOOBNet and OFNet. GT refers to groundtruth map. red pixels with arrows: correctly labeled occlusion boundaries; cyan: correctly labeled boundaries but mislabeled occlusion; green: false negative boundaries; orange: false positive boundaries. Our prediction results are visually much more accurate compared to DOOBNet and OFNet (Best viewed in color).

Table 2: Comparison on number of shared layers and multi-scale supervision on PIOD.

		BPR			OPR		
		ODS	OIS	AP	ODS	OIS	AP
Layer Shared	top-1	78.7	79.8	80.8	76.2	77.1	77.1
	top-2	**79.5**	**80.4**	**83.1**	**77.1**	**77.8**	**79.4**
	top-3	77.6	78.6	80.6	75.4	76.3	77.5
	top-4	77.0	78.4	79.7	74.8	76.0	76.4
Number of Scales	1	75.4	76.5	77.4	73.1	74.0	74.1
	2	78.6	79.5	82.6	76.3	76.4	78.9
	3	78.1	79.8	82.0	76.4	77.3	78.6
	4	78.6	79.5	81.1	76.1	76.9	77.6
	5	**79.5**	**80.4**	**83.1**	**77.1**	**77.8**	**79.4**

and BSDS ownership dataset, respectively, which validates the effectiveness of our orthogonal orientation representation and loss.

Multi-scale supervision effect: Our boundary path outputs five side boundary maps with multi-scale supervision signals. To validate the effectiveness of such multi-scale supervision, we compared results of supervising the lowest one to the lowest five side boundary maps, where a lower boundary map has a larger spatial size. As shown in Ta-

ble 2, the performance of only supervising the lowest side is significantly worse than other settings. Performance is also greatly improved by using five-scale supervision, surpassing the second best(using 4 side maps) by 2.0% B-AP, which demonstrates the importance of multi-scale supervision and multi-result fusion.

5. Conclusion

In this paper, we present the novel OPNet, which shares occlusion features in the deep stage and splits into two separate decoder paths with larger spatial sizes. Besides, we apply multi-scale supervision to the boundary extraction. We also present a robust orthogonal occlusion representation, which avoids the endpoint prediction errors and negative impact of angle periodicity. Our network achieves new state-of-the-art performances on both PIOD and BSDS ownership datasets.

Acknowledgement

References

[1] Radhakrishna Achanta and Sabine Susstrunk. Superpixels and polygons using simple non-iterative clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4660, 2017.

[2] Mary B Alatise and Gerhard P Hancke. A review on challenges of autonomous mobile robot and sensor fusion methods. *IEEE Access*, 8:39830–39846, 2020.

[3] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010.

[4] Alper Ayvaci and Stefano Soatto. Detachable object detection: Segmentation and depth ordering from short-baseline video. *IEEE transactions on pattern analysis and machine intelligence*, 34(10):1942–1951, 2011.

[5] Deng-Ping Fan, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Huazhu Fu, and Ming-Ming Cheng. Taking a deeper look at co-salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2919–2929, 2020.

[6] Doron Feldman and Daphna Weinshall. Motion segmentation and depth ordering using an occlusion detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7):1171–1185, 2008.

[7] Charless C Fowlkes, David R Martin, and Jitendra Malik. Local figure–ground cues are valid for natural images. *Journal of vision*, 7(8):2–2, 2007.

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

[9] Santiago González Izard, Ramiro Sánchez Torres, Oscar Alonso Plaza, Juan Antonio Juanes Mendez, and Francisco José García-Peñalvo. Nextmed: automatic imaging segmentation, 3d reconstruction, and 3d model visualization platform using augmented and virtual reality. *Sensors*, 20(10):2962, 2020.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 558–567, 2019.

[12] Derek Hoiem, Alexei A Efros, and Martial Hebert. Recovering occlusion boundaries from an image. *International Journal of Computer Vision*, 91(3):328–346, 2011.

[13] Zhaoyin Jia, Andrew Gallagher, Yao-Jen Chang, and Tsuhan Chen. A learning-based framework for depth ordering. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 294–301. IEEE, 2012.

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

[15] Rui Lu, Feng Xue, Menghan Zhou, Anlong Ming, and Yu Zhou. Occlusion-shared and feature-separated network for occlusion relationship reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10343–10352, 2019.

[16] Rui Lu, Menghan Zhou, Anlong Ming, and Yu Zhou. Context-constrained accurate contour extraction for occlusion edge detection. *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1522–1527, 2019.

[17] Michael Maire, Takuya Narihira, and Stella X Yu. Affinity cnn: Learning pixel-centric pairwise relations for figure/ground embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 174–182, 2016.

[18] Jonathan A Marshall, Christina A Burbeck, Dan Ariely, Jannick P Rolland, and Kevin E Martin. Occlusion edge blur: a cue to relative visual depth. *JOSA A*, 13(4):681–688, 1996.

[19] Anlong Ming, Tianfu Wu, Jianxiang Ma, Fang Sun, and Yu Zhou. Monocular depth-ordering reasoning with occlusion edge detection and couple layers inference. *IEEE Intelligent Systems*, 31(2):54–65, 2015.

[20] Shayan Nikoohemat, Abdoulaye A Diakité, Sisi Zlatanova, and George Vosselman. Indoor 3d reconstruction from point clouds for optimal routing in complex buildings to support disaster management. *Automation in construction*, 113:103109, 2020.

[21] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9413–9422, 2020.

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

[23] Xuchong Qiu, Yang Xiao, Chaohui Wang, and Renaud Marlet. Pixel-Pair occlusion relationship map (P2ORM): Formulation, inference & application. In *European Conference on Computer Vision (ECCV)*, 2020.

[24] Michael Ramamonjisoa, Yuming Du, and Vincent Lepetit. Predicting sharp and accurate occlusion boundaries in monocular depth estimation using displacement fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[25] Michael Ramamonjisoa and Vincent Lepetit. Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. *The IEEE International Conference on Computer Vision (ICCV) Workshops*, 2019.

[26] Xiaofeng Ren, Charless C Fowlkes, and Jitendra Malik. Figure/ground assignment in natural images. In *European Conference on Computer Vision*, pages 614–627. Springer, 2006.

[27] Xiaofeng Ren, Charless C Fowlkes, and Jitendra Malik. Figure/ground assignment in natural images. In *European Conference on Computer Vision*, pages 614–627. Springer, 2006.

[28] Mehmet Emre Sargin, Luca Bertelli, Bangalore S Manjunath, and Kenneth Rose. Probabilistic occlusion boundary detection on spatio-temporal lattices. In *2009 IEEE 12th International Conference on Computer Vision*, pages 560–567. IEEE, 2009.

[29] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. Learning depth from single monocular images. In *Advances in neural information processing systems*, pages 1161–1168, 2006.

[30] Qi Shan, Brian Curless, Yasutaka Furukawa, Carlos Hernandez, and Steven M Seitz. Occluding contours for multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4002–4009, 2014.

[31] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods*, 18(1):100–106, 2021.

[32] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

[33] Ching Teo, Cornelia Fermuller, and Yiannis Aloimonos. Fast 2d border ownership assignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5117–5125, 2015.

[34] Guoxia Wang, Xiaochuan Wang, Frederick WB Li, and Xiaohui Liang. Doobnet: Deep object occlusion boundary detection from an image. In *Asian Conference on Computer Vision*, pages 686–702. Springer, 2018.

[35] Peng Wang and Alan Yuille. Doc: Deep occlusion estimation from a single image. In *European Conference on Computer Vision*, pages 545–561. Springer, 2016.

[36] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015.

[37] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene de-occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3784–3792, 2020.

[38] Ziyu Zhang, Alexander G Schwing, Sanja Fidler, and Raquel Urtasun. Monocular object instance segmentation and depth ordering with cnns. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2614–2622, 2015.

[39] 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.