Occlusion-shared and Feature-separated Network for Occlusion Relationship Reasoning

Rui Lu¹ Feng Xue¹ Menghan Zhou¹,³ Anlong Ming¹ Yu Zhou²,*
¹Beijing University of Posts and Telecommunications, Beijing, China
²Huazhong University of Science and Technology, Wuhan, China
³Lenovo Research, Beijing, China
{lurui,xuefeng,mal}@bupt.edu.cn zhoumh3@lenovo.com yuzhou@hust.edu.cn

Abstract

Occlusion relationship reasoning demands closed contour to express the object, and orientation of each contour pixel to describe the order relationship between objects. Current CNN-based methods neglect two critical issues of the task: (1) simultaneous existence of the relevance and distinction for the two elements, i.e, occlusion edge and occlusion orientation; and (2) inadequate exploration to the orientation features. For the reasons above, we propose the Occlusion-shared and Feature-separated Network (OFNet). On one hand, considering the relevance between edge and orientation, two sub-networks are designed to share the occlusion cue. On the other hand, the whole network is split into two paths to learn the high-level semantic features separately. Moreover, a contextual feature for orientation prediction is extracted, which represents the bilateral cue of the foreground and background areas. The bilateral cue is then fused with the occlusion cue to precisely locate the object regions. Finally, a stripe convolution is designed to further aggregate features from surrounding scenes of the occlusion edge. The proposed OFNet remarkably advances the state-of-the-art approaches on PIOD and BSDS ownership dataset. The source code is available at https://github.com/buptlr/OFNet.

1. Introduction

Reasoning the occlusion relationship of objects from monocular image is fundamental in computer vision and mobile robot applications, such as [11, 2, 24, 17, 29]. Furthermore, it can be regarded as crucial elements for scene understanding and visual perception[40, 42, 43, 39, 18], such as object detection, image segmentation and 3D reconstruction [6, 1, 37, 7, 26, 34]. From the perspective of the observer, occlusion relationship reflects relative depth difference between objects in the scene.

Previously, a number of influential studies infer the occlusion relationship by designing hand-crafted features, e.g. [5, 22, 25, 16, 38, 41]. Recently, driven by Convolutional Neural Networks (CNN), several deep learning based approaches outperform traditional methods at a large margin. DOC [21] specifies a new representation for occlusion relationship, which decomposes the task into the occlusion edge classification and the occlusion orientation regression. And it utilizes two networks for these two sub-tasks, respectively. DOOBNet [31] employs an encoder-decoder structure to obtain multi-scale and multi-level features. It shares backbone features with two sub-networks and simultaneously acquires the predictions.
In occlusion relationship reasoning, the closed contour is employed to express the object, and the orientation values of the contour pixels are employed to describe the order relationship between the foreground and background objects. We observe that two critical issues have rarely been discussed. Firstly, the two elements, i.e., occlusion edge and occlusion orientation, have the relevance and distinction simultaneously. They both need the occlusion cue, which describes the location of the occluded background, as shown in Fig. 1 (d). Secondly, the high-level features for orientation prediction are not fully revealed. It needs additional cues from foreground and background areas (shown in Fig. 1 (e)). Consequently, existing methods are limited in reasoning accuracy. Compared with our approach (shown in Fig. 1 (c)), previous works [21, 31] (shown in Fig. 1 (a)(b)) exist false positive and false negative detection of edge, as well as false positive prediction of orientation.

Aiming to address the two issues above, and boost the occlusion relationship reasoning, a novel Occlusion-shared and Feature-separated Network (OFNet) is proposed. As shown in Fig. 2 (c), considering the relevance and distinction between edge and orientation, our network is different from the other works (shown in Fig. 2 (a)(b)). Two separate network paths share the occlusion cue and encode different high-level features. Furthermore, a contextual feature for orientation prediction is extracted, which is called the bilateral feature. To learn the bilateral feature, a Multi-rate Context Learner (MCL) is proposed. The learner has different scales of receptive field so that it can fully sense the two objects, i.e., the foreground and background objects, fundamentally assisting the occlusion relationship reasoning. To extract the feature more accurately, the Bilateral Response Fusion (BRF) is proposed to fuse the occlusion cue with the bilateral feature from MCL, which can precisely locate the areas of foreground and background. To effectively infer the occlusion relationship by the special orientation features, a stripe convolution is designed to replace the traditional plain convolution, which elaborately integrates the bilateral feature to distinguish the foreground and background areas.

Experiments prove that we achieve SOTA performance on both PIOD [21] and BSDS ownership [22] dataset. The main contributions of our approach lie in:

- The relevance and distinction between occlusion edge and occlusion orientation are re-interpreted. The two sub-tasks share the occlusion cues, but separate the contextual features.
- The bilateral feature is proposed, and two particular modules are designed to obtain the specific features, i.e., Multi-rate Context Learner (MCL) and Bilateral Response Fusion (BRF).
- To elaborately infer the occlusion relationship, a stripe convolution is designed to further aggregate the feature from surrounding scenes of the contour.

2. Related Work


Multi-level Features are extracted from different layers, which are widely used in image detection [14, 23, 27, 35]. Peng et al. [20] fuse feature maps from multi-layer with refined details. Shrivastava et al. [28] adopt lateral connections to leverage top-down context and bottom-up details.

Occlusion Relationship Representation has evolved overtime from triple points and junctions representation [10, 22] to pixel-based representation [30, 21]. The latest representation [21] applies a binary edge classifier to determine whether the pixel belongs to an occluded edge, and a

Figure 2. The schematic demonstration of the high-level feature propagation process of the state-of-the-arts and ours. (a) indicates the two separate stream networks employing side-outputs of various layers. (b) presents the single stream network sharing decoder features. (c) shows our network which captures contextual features for specific tasks and shares decoder features.
It requires reserved resolution of the original image to provide the accurate location and large receptive field to perceive the mutual constraint of pixels on the boundary.

We adopt the module proposed in [15], which has a high capability to capture accurate location cue and sensitive perception of the entire object. In [15], the low-level cue from the first three side-outputs preserves the original size of the input image and encodes abundant spatial information. Without losing resolution, the large receptive field is achieved via dilated convolution [35] after res50 [9]. The Bilateral Response Fusion (BRF) shown in Fig. 3 is presented to compensate the precise position for high-level features and suppress the clutter of non-occluded pixels for low-level features. Different from [15], we employ an additional convolution block to refine the contour, and integrate specific task features provided by diverse channels. Besides, this well-designed convolution block eliminates the gridding artifacts [36] caused by the dilated convolution in high-level layers.

The resulting edge map embodies low-level and high-level features, which guarantees the consistency and accuracy of the occlusion edge. Specifically, the edge path provides complete and continuous contour, which makes up the object region. The object region is delineated by a set of occlusion edges.

3.2. Orientation Path

For the orientation path, we innovatively introduce the bilateral feature, which is conducive to describe the order relationship. Specifically, the bilateral feature represents information of surrounding scenes, which includes sufficient ambient context to deduce whether it belongs to the foreground or background areas.

3.2.1 Multi-rate Context Learner

Bilateral feature characterizes the relationship between the foreground and background areas. To infer the occlusion relationship between objects, the sufficient receptive field for the objects with different sizes is essential.

To perceive the object with various ranges and learn the bilateral feature, the Multi-rate Context Learner (MCL) is designed, which consists of three components, as shown in Fig. 4. Firstly, the high-level semantic cue is convolved by multiple dilated convolutions, which allows the pixels on the edge to perceive the foreground and background objects as completely as possible. The dilated convolutions have kernel size of $3 \times 3$ with various dilation rates. With various dilated rates of the dilated convolutions, the learner is able to perceive the scene cue at different scales from the foreground and background areas, which is beneficial to deduce which side of the region is in front. Secondly, an element-wise convolution module, i.e., $1 \times 1$ conv, is used to integrate
foreground area, the occlusion cue obtained through the decoder represents the location information of the boundary. As shown in Fig.6, after the bilinear upsampling, bilateral feature is hard to locate the exact location of the contour. Hence, to sufficiently learn the feature for occlusion relationship reasoning, more precise location of object region is demanded, which is provided by the occlusion cue from the decoder. Thus, it is necessary to introduce clear contour to describe the areas of the foreground and background objects, thereby extracting the object features more accurately.

The Bilateral Response Fusion (BRF), shown in Fig.5, is proposed to fuse these two disparate streams of features, i.e. the bilateral map \( \{B\} \) and occlusion map \( \{D\} \). The unified orientation fused map of ample bilateral response and emphatic occlusion is formed, which is denoted as \( \{F\} \), where \( Conv_3() \) represents the \( 3 \times 3 \) conv:

\[
\{F\} = Conv_3(Conv_3(Concat(\{B\}, \{D\})))
\]  

(2)

\( \{F\} \) denotes the feature map generated by BRF module, and each element of the set is a feature map. Subsequently, \( \{F\} \) has \( 224 \times 224 \) spatial resolution and is taken as the input of the Occlusion Relationship Reasoning module (Sec.3.2.3), as shown in Fig.6. Through BRF, the occlusion feature is effectively combined with the bilateral feature. For occlusion relationship reasoning, the fused orientation map not only possesses the boundary location between two objects with occlusion relationship, but also own contextual information of each object. The BRF module provides adequate cues for the following feature learning module to infer the foreground and background relationship. Besides, by integrating bilateral feature, the scene cue near the contour is enhanced.

### 3.2.3 Occlusion Relationship Reasoning

By utilizing the MCL and BRF, the bilateral feature is learned and fused, an inference module is necessary to determine the order of the foreground and background areas, which makes full use of this feature. Existing method [31] utilizes \( 3 \times 3 \) conv to learn the features. This small convolution kernel only extracts the cues at the local pixel patch, which is not suitable to infer the occlusion relationship. The
reason is that the tiny perceptive field is unable to perceive the learned object cue. Thus, a large convolution kernel is necessary for utilizing the bilateral feature, which is able to perceive surrounding regions near the contour.

Nevertheless, large convolution kernels are computation demanding and memory consuming. Instead, two stripe convolutions are proposed, which are orthogonal to each other. Compared to the $3 \times 3$ conv, which captures only nine pixels around the center (shown in Fig. 7(a)), the vertical and horizontal stripe convolutions have $11 \times 3$ and $3 \times 11$ receptive field, as shown in Fig. 7(b). Specifically, for a contour pixel with arbitrary orientation, its tangent direction can be decomposed into vertical and horizontal directions. Context along orthogonal directions make varied amount of contributions in expressing the orientation representation. Thus, tendency of the extended contour and occluded relationship of bilateral scenes are recognized.

In addition, two main advantages are achieved. First, the large receptive field aggregates contextual information of object to determine the depth order, which is without large memory consuming. Second, although the slope of the edge is not exactly perpendicular or parallel to the ground, one of the stripe convolutions can successfully perceive the foreground and background objects. After the concatenation of the two orthogonal convolution modules, we apply the $3 \times 3$ conv to refine the features.

3.3. Loss Function

**Occlusion Edge:** Occlusion edge characterizes depth discontinuity between regions, reflecting as the boundary between objects. Given a set of training images $\Psi = \{I_1, I_2, \ldots, I_N\}$, the corresponding ground truth edge of the $k$-th input image at pixel $p$ is $E_k(p \mid I) \in \{0, 1\}$ and we denote $E_k(p \mid I, W) \in [0, 1]$ as its network output, indicating the computed edge probability.

**Occlusion Orientation:** Occlusion orientation indicates the tangent direction of the edge using the left rule (i.e., the foreground area is on the left side of the background area). Following the mathematical definition above, for the $k$-th input image, its orientation ground truth at pixel $p$ is $O_k(p \mid I) \in (-\pi, \pi]$. The regression prediction result of orientation path is $\hat{O}_k(p \mid I, W) \in (-\pi, \pi]$.

**Occlusion Relationship:** During the testing phase, we first refine the $\hat{E}_k$ by conducting non-maximum suppression $\hat{E}_k = NMS(\hat{E}_k)$. The nonzero pixels of sharpened $\hat{E}_k$ form the binary matrix $M_k = \text{sign}(\hat{E}_k)$. We then perform element-wise product of $M_k$ and orientation map $\hat{O}_k$, obtain refined orientation map $\hat{O}_k = M_k \circ \hat{O}_k$. Finally, following [11], we adjust the $\hat{O}_k$ to the tangent direction of $\hat{E}_k$ and gain the final occlusion edge map.

**Loss Function:** Following [31], we use the following loss function to supervise the training of our network.

$$l(W) = \frac{1}{M} \sum_j \sum_i AL(y_i, \epsilon_i) + \lambda \sum_j \sum_i SL(f(\pi_i, a_i))$$

(3)

The parameters include: collection of all standard network layer parameters ($W$), predicted edge value at pixel $i$ ($y_i \in [0, 1]$), mini-batch size ($M$), image serial number in a mini-batch ($j$), the Attention Loss ($AL$), the Smooth $L_1$ Loss ($SL$) [31].

4. Experiments

In this section, abundant experiments are demonstrated to validate the performance of the proposed OFNet. Further, we present some ablation analyses for discussions of the network design choices.

4.1. Implementation Details

**Dataset:** Our method is evaluated on two challenging datasets: PIOD [21] and BSDS ownership [22]. The PIOD dataset is composed of 9,175 training images and 925 testing images. Each image is annotated with ground truth object instance edge map and its corresponding orientation map. The BSDS ownership dataset includes 100 training images and 100 testing images of natural scenes. Following [31], all images in the two datasets are randomly cropped to $320 \times 320$ during training while retaining their original sizes during testing.

**Initialization:** Our network is implemented in Caffe [12] and finetuned from an initial pretrained Res50 model. All convolution layers added are initialized with “msra” [8].

Figure 7. The schematic illustration of how orientation information propagates in the feature learning phase. (a) the plain convolution. (b) the stripe convolution.
Table 1. OPR results on PIOD (left) and BSDS ownership dataset (right). ①-⑤ represent SRF-OCC [30], DOC-HED [21], DOC-DMLFOV [21], DOOBNet [31] and ours, respectively. † refers to GPU running time. Red bold type indicates the best performance, blue bold type indicates the second best performance (the same below).

<table>
<thead>
<tr>
<th></th>
<th>ODS</th>
<th>OIS</th>
<th>AP</th>
<th>FPS</th>
<th>ODS</th>
<th>OIS</th>
<th>AP</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>①</td>
<td>.268</td>
<td>.286</td>
<td>.152</td>
<td>0.018</td>
<td>.419</td>
<td>.448</td>
<td>.337</td>
<td>0.018</td>
</tr>
<tr>
<td>②</td>
<td>.460</td>
<td>.479</td>
<td>.405</td>
<td>18.3†</td>
<td>.522</td>
<td>.545</td>
<td>.428</td>
<td>19.6†</td>
</tr>
<tr>
<td>③</td>
<td>.601</td>
<td>.611</td>
<td>.585</td>
<td>18.9†</td>
<td>.463</td>
<td>.491</td>
<td>.369</td>
<td>21.1†</td>
</tr>
<tr>
<td>④</td>
<td>.702</td>
<td>.712</td>
<td>.683</td>
<td>26.7†</td>
<td>.555</td>
<td>.570</td>
<td>.440</td>
<td>25.8†</td>
</tr>
<tr>
<td>⑤</td>
<td>.718</td>
<td>.728</td>
<td>.729</td>
<td>28.3†</td>
<td>.583</td>
<td>.607</td>
<td>.501</td>
<td>27.2†</td>
</tr>
</tbody>
</table>

Table 2. EPR results on PIOD (left) and BSDS ownership dataset (right).

<table>
<thead>
<tr>
<th></th>
<th>ODS</th>
<th>OIS</th>
<th>AP</th>
<th></th>
<th>ODS</th>
<th>OIS</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>①</td>
<td>.345</td>
<td>.369</td>
<td>.207</td>
<td>0.207</td>
<td>.511</td>
<td>.544</td>
<td>.442</td>
</tr>
<tr>
<td>②</td>
<td>.509</td>
<td>.532</td>
<td>.468</td>
<td>.658</td>
<td>.685</td>
<td>.602</td>
<td></td>
</tr>
<tr>
<td>③</td>
<td>.669</td>
<td>.684</td>
<td>.677</td>
<td>.579</td>
<td>.609</td>
<td>.519</td>
<td></td>
</tr>
<tr>
<td>④</td>
<td>.736</td>
<td>.746</td>
<td>.723</td>
<td>.647</td>
<td>.668</td>
<td>.539</td>
<td></td>
</tr>
<tr>
<td>⑤</td>
<td>.751</td>
<td>.762</td>
<td>.773</td>
<td>.662</td>
<td>.689</td>
<td>.585</td>
<td></td>
</tr>
</tbody>
</table>

Evaluation Criteria: Following [21], we compute precision and recall of the estimated occlusion edge maps (i.e. OPR) by performing three standard evaluation metrics: fixed contour threshold (ODS), best threshold of each image (OIS) and average precision (AP). Notably, the orientation recall is only calculated at the correctly detected edge pixels. Besides, the above three metrics are also used to evaluate the edge map after NMS.

4.2. Evaluation Results

Quantitative Performance: We evaluate our approach with comparisons to the state-of-the-art algorithms including SRF-OCC [30], DOC-HED [21], DOC-DMLFOV [21] and DOOBNet [31].

As shown in Table.1 and Fig.8, our approach outperforms all other state-of-the-art methods for OPR results. Specifically, in terms of the PIOD dataset, our method performs the best, outperforming the baseline DOOBNet of 4.6% AP. This is due to the efficiency of extracting high semantic features of the two paths separately. Edge path succeeds in enhancing contour response and orientation path manages to perceive foreground and background relationship. Splitting these two tasks into two paths enables the promotion of the previous algorithms. For the BSDS ownership dataset, which is difficult to train due to the small training samples, the proposed OFNet obtains the gains of 2.8% ODS, 3.7% OIS and 6.1% AP compared with the baseline DOOBNet. Specifically, our approach increases bilateral cue between the foreground and background objects, and fuses them with high semantic features to introduce clear contour, which describes the areas of the foreground and background better. Besides, stripe convolution in our network plays an important role in harvesting the surrounding scenes of the contour. The improvement in orientation proves the effectiveness of the module.

EPR results are presented in Table.2 and Fig.9. For the PIOD dataset, our approach superiorly performs against the other evaluated methods, surpassing DOOBNet by 5.0% AP. We take the distinction between edge and orientation into consideration, and extract specific features for sub-networks, respectively. For edge path, by utilizing the contextual features, which reflect pixels constraint on the occlusion edge, our network outputs edge maps with augmented contour and less noise around. With location cue extracted from low-level layers, the predicted edge in our method fits the contour better, thus avoiding false positive detections compared to others. For the BSDS ownership dataset, our approach achieves the highest ODS as well.

Qualitative Performance: Fig.10 shows the qualitative results on the two datasets. The top four rows show the results of the PIOD dataset [21], and the bottom four rows represent the BSDS ownership dataset [22]. The first col-
Figure 10. Example results on PIOD (first four rows) and BSDS ownership dataset (last four rows). 1\textsuperscript{st} column: input images; 2\textsuperscript{nd}–4\textsuperscript{th} columns: visualization results of ground truth, baseline and ours; 5\textsuperscript{th}–6\textsuperscript{th} columns: edge maps and orientation maps of ours. Notably, "red" pixels with arrows: correctly labeled occlusion boundaries; "cyan": correctly labeled boundaries but mislabeled occlusion; "green": false negative boundaries; "orange": false positive boundaries (Best viewed in color).

Column to the sixth column show the original RGB image from datasets, ground truth, the result predicted by DOOBNet [31], the result predicted by the proposed OFNet, the detected occlusion edge and the predicted orientation, respectively. In the resulting image, the right side of the arrow direction is the background, and the left side corresponds to the foreground area.

In detail, the two occluded buses in the first row have similar appearances. Thus, it is hard to detect the dividing line between them, just as our baseline DOOBNet fails. However, our method detects the occlusion edge consistently. In the second row, the occlusion relationship between the wall and the sofa is easy to predict failure. Instead of the small receptive field, which is difficult to perceive objects with large-area pure color, our method with sufficient receptive field correctly predicts the relationship. The third scene is similar to the second row. Compared with the baseline, our method predicts the relationship between the sofa and the ground correctly. In the fourth row, the color of the cruise ship is similar to the hill behind, which is not detected by the baseline. By using the low-level edge cues, our method accurately locates the contour of the ship. The fifth row shows people under the wall, and the orientation cannot be correctly detected due to the low-level features in the textureless areas. Our method correctly infers the relationship by using the high-level bilateral feature. The last three scenes have the same problem as the third row, i.e., the object with a large region of pure color. Our method outperforms others in this situation by a large margin, which proves the effectiveness of our designed modules.

4.3. Ablation Analysis

One-branch or Multi-branch Sub-networks: To evaluate that our method provides different high-level features for different sub-tasks, an existing method [31], which adopts a single flow architecture by sharing high-level fea-
Table 3. Experimental results of baseline DOOBNet [31], baseline with split decoder, baseline with single stream sub-network and our approach. The experiments are conducted on the PIOD dataset (the same below).

<table>
<thead>
<tr>
<th>Methods</th>
<th>ODS</th>
<th>OIS</th>
<th>AP</th>
<th>ODS</th>
<th>OIS</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>.736</td>
<td>.746</td>
<td>.723</td>
<td>.702</td>
<td>.712</td>
<td>.683</td>
</tr>
<tr>
<td>Baseline(ssplit decoder)</td>
<td>.720</td>
<td>.735</td>
<td>.694</td>
<td>.702</td>
<td>.712</td>
<td>.683</td>
</tr>
<tr>
<td>Single edge stream</td>
<td>.739</td>
<td>.750</td>
<td>.685</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Single ori stream</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>.705</td>
<td>.716</td>
<td>.674</td>
</tr>
<tr>
<td>Ours</td>
<td>.751</td>
<td>.762</td>
<td>.773</td>
<td>.718</td>
<td>.728</td>
<td>.729</td>
</tr>
</tbody>
</table>

Table 4. Experimental results of our model without low-cues, without edge high-cues, without orientation high-cues and our model.

<table>
<thead>
<tr>
<th>Methods</th>
<th>ODS</th>
<th>OIS</th>
<th>AP</th>
<th>ODS</th>
<th>OIS</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours(w/o low-cues)</td>
<td>.746</td>
<td>.758</td>
<td>.764</td>
<td>.715</td>
<td>.722</td>
<td>.715</td>
</tr>
<tr>
<td>Ours(w/o edge high-cues)</td>
<td>.742</td>
<td>.753</td>
<td>.758</td>
<td>.709</td>
<td>.717</td>
<td>.698</td>
</tr>
<tr>
<td>Ours(w/o ori high-cues)</td>
<td>.743</td>
<td>.756</td>
<td>.759</td>
<td>.711</td>
<td>.719</td>
<td>.703</td>
</tr>
<tr>
<td>Ours</td>
<td>.751</td>
<td>.762</td>
<td>.773</td>
<td>.718</td>
<td>.728</td>
<td>.729</td>
</tr>
</tbody>
</table>

Table 5. Experimental results of bilateral feature and occlusion feature with various fusion ratio in BRF module.

<table>
<thead>
<tr>
<th>Scale</th>
<th>ODS</th>
<th>OIS</th>
<th>AP</th>
<th>ODS</th>
<th>OIS</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>scale = 16:16</td>
<td>.742</td>
<td>.752</td>
<td>.749</td>
<td>.710</td>
<td>.719</td>
<td>.703</td>
</tr>
<tr>
<td>scale = 32:16</td>
<td>.741</td>
<td>.754</td>
<td>.759</td>
<td>.712</td>
<td>.722</td>
<td>.709</td>
</tr>
<tr>
<td>scale = 48:16</td>
<td>.744</td>
<td>.758</td>
<td>.765</td>
<td>.715</td>
<td>.726</td>
<td>.717</td>
</tr>
<tr>
<td>scale = 64:16</td>
<td>.751</td>
<td>.762</td>
<td>.773</td>
<td>.718</td>
<td>.728</td>
<td>.729</td>
</tr>
</tbody>
</table>

Table 6. Experimental results of stripe convolutions with different aspect ratios.

<table>
<thead>
<tr>
<th>Scale</th>
<th>ODS</th>
<th>OIS</th>
<th>AP</th>
<th>ODS</th>
<th>OIS</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>conv = 3×3</td>
<td>.746</td>
<td>.753</td>
<td>.754</td>
<td>.712</td>
<td>.719</td>
<td>.694</td>
</tr>
<tr>
<td>conv = 3×5</td>
<td>.747</td>
<td>.755</td>
<td>.760</td>
<td>.712</td>
<td>.720</td>
<td>.696</td>
</tr>
<tr>
<td>conv = 3×7</td>
<td>.747</td>
<td>.754</td>
<td>.758</td>
<td>.713</td>
<td>.721</td>
<td>.699</td>
</tr>
<tr>
<td>conv = 3×9</td>
<td>.750</td>
<td>.759</td>
<td>.767</td>
<td>.716</td>
<td>.723</td>
<td>.712</td>
</tr>
<tr>
<td>conv = 3×11</td>
<td>.751</td>
<td>.762</td>
<td>.773</td>
<td>.718</td>
<td>.728</td>
<td>.729</td>
</tr>
</tbody>
</table>

Necessity for Each Feature: In order to verify the role of various low-level and high-level features, each feature is removed to construct an independent variant for evaluation, as shown in Table 4. Intuitively, even if the slope of the edge is not in a horizontal or vertical direction, the convolution kernels possess large receptive field and tend to learn the cues of both directions, respectively. Nevertheless, the larger convolution layer takes up too much computation cost, which increases the number of parameters. Consequently, the stripe convolutions in orthogonal directions extract the tendency of edges and bilateral cue around the contour.

5. Conclusion

In this paper, we present a novel OFNet, which shares the occlusion cue from the decoder and separately acquires the contextual features for specific tasks. Our algorithm builds on top of the encoder-decoder structure and side-output utilization. For learning the bilateral feature, an MSL is proposed. Besides, a BRF module is designed to apply the occlusion cue to precisely locate the object regions. In addition, we utilize a stripe convolution to further aggregate features from surrounding scenes of the contour. Significant improvement of the state-of-the-art through numerous experiments on PIOD and BSDS ownership dataset demonstrates the effectiveness of our network.

Acknowledgement. This work was supported by the National Natural Science Foundation of China Nos. 61703049, 61876022, 61872047, and the Fundamental Research Funds for the Central Universities No. 2019kfyRCPY001.
References

[22] Xiaofeng Ren, Charless Fowlkes, and Jitendra Malik. Figure/ground assignment in natural images. In *ECCV*, 2006.

[38] Menghan Zhou, Jianxiang Ma, Anlong Ming, and Yu Zhou. Objectness-aware tracking via double-layer model. In ICIP, 2018. 1


[41] Yu Zhou, Jianxiang Ma, Anlong Ming, and Xiang Bai. Learning training samples for occlusion edge detection and its application in depth ordering inference. In ICPR, 2018. 1

[42] Yu Zhou and Anlong Ming. Human action recognition with skeleton induced discriminative approximate rigid part model. PRL, 2016. 1