DOC: Deep Occlusion Recovering From A Single Image

Peng Wang, Alan Yuille
Department of Statistics
University of California, Los Angeles

Abstract

Recovering the occlusion relationship between stuffs and things is a fundamental ability of human vision, and yields information about the 3D world from 2D images. In this paper, we propose deep occlusion (DOC), which is an end-to-end deep learning strategy that recovers the occlusion boundaries from a single image. Firstly, for each pixel along the boundaries, we introduce an orientation variable (illustrated at the top of Fig. 1), indicating the occlusion relationship between two adjacent regions. Then, deep features, long range context and object-level knowledge are exploited by an end-to-end deep convolutional neural network (DCNN) to accurately locate the boundaries and recover the occlusion relationships. To evaluate and train the DOC network, we construct a large-scale instance occlusion boundary dataset from the PASCAL images, which we call the PASCAL instance occlusion dataset. It includes 20k images ($100\times$ bigger than the existing occlusion datasets for outdoor images) and so is big enough to train and evaluate DCNNs for occlusion learning. Finally, extensive experiments are performed to demonstrate that our approach has strong generalization ability and outperform state-of-the-art strategies.

1 Introduction

Humans have a strong ability to infer the occlusion relationships and depth orders of objects and backgrounds (stuff and things). This has long been recognized as a basis for human scene understanding and perception (Gibson [1968], Biederman [1981]). As visualized at the top left of Fig. 1, we can deduce that “the person is holding a dog”, based on the occlusion relationship between the person’s hand and the dog, and how the dog occludes the person’s body. This ability is crucial for us as well as for computer vision to understand any given image. Biologically, the importance of recovering occlusion boundaries for human scene perception has also long been realized. In particular, von der Heydt et al. (2005) studied the visual cortex with electrophysiology while Fang et al. (2009) performed fMRI studies, suggesting that humans have special neurons and cells at V2 that perform some forms of boundary ownership prediction indicating occlusion. Extensive psychological studies reveal that a range of cues convey occlusion information for human beings such as edge convexity (Kanizsa & Gerbino [1976]), edge-junctions, contrast and gradients in intensity and texture (Palmer & Ghose [2008]). In addition, Craft et al. (2007) suggest that occlusion detection may require feedback from higher level cortical regions, indicating that long-range context and semantic-level knowledge might be needed to infer occlusion.

Guided by these visual studies, computer vision researchers have explored multiple features such as convexity, T-junctions, Geometric context, HOG, spectral and Gestalt-like features etc, to model the occlusion cues, and applied machine learning techniques to infer occlusion relationships, e.g. (Ren et al. [2006], Hoiem et al. [2011], Teo et al. [2015]). Although carefully engineered, these methods only performed studies on small datasets, and depended on hand-crafted features. But in recent years, deep convolutional neural networks (DCNN) learned on big data delivers strikingly good results in many computer vision tasks. This is mostly due to the learned rich hierarchical representations contain information from low-level edge contours, junctions, to high-level object parts and shapes, as visualized and comprehended by (Zeiler & Fergus [2014], Mahendran & Vedaldi [2015]).

In addition it is also plausible that DCNN can be applied to infer occlusion, since the rich deep features are also capable of representing occlusion cues, as shown in visualization studies. Nonetheless, to the best of our knowledge, learning a network for occlusion recovering has not been explored, perhaps due to the following issues: (1) there are technical hurdles for designing a pixel-wise end-to-end
mapping from a image to occlusion relationship encoding both boundary detection and occlusion inference. (2) The lack of a large occlusion dataset, since big datasets are generally needed to train robust DCNN models. In this work, to solve the first issue, we propose a joint labelling strategy for each pixel, with a variable to indicate if a pixel is located on a boundary and an orientation variable which represents the occlusion relationship. This enables us to perform a pixel-wise end-to-end learning for occlusion boundary recovery making use of fully convolutional networks [Long et al., 2015]. The details will be elaborated in Sec. 4. Regarding the second issue, current occlusion boundary datasets, such as the BSDS ownership [Ren et al., 2006] (200 images), GeoContext [Hoiem et al., 2011] (100 images) and the NYU depth [Silberman et al., 2012] (1449 indoor images), are limited in size. This motivates us to construct a large-scale dataset containing many stuffs and things in real world scenarios. Specifically, we choose the PASCAL VOC [Everingham et al., 2010] images to perform our labelling task because many of the object edges have already been annotated in previous works [Hariharan et al., 2011; Chen et al., 2014b]. This reduces the difficulties of labelling edges from scratch and means that we only need to label the occlusion relations. Our proposed dataset contains 20k natural images, which is significantly larger than the traditional ones. In our experiments in Sec. 5, we will demonstrate the benefits obtained from the this rich data.

This paper makes three main contributions: (1) We are (probably) the first to apply DCNNs to occlusion boundary estimation and our method obtains strong results outperforming the state-of-the-art while being computationally efficient (0.6s/image). (2) We created a large occlusion boundary dataset over the PASCAL VOC images, which is a resource for studying occlusion. It makes the description of natural objects beyond the self semantic category, as well as the inter-relationships. (3) Extensive experiments have been performed, which provide a starting point for other researchers to explore. We will release all our models, codes and dataset upon the publication of this paper.

2 RELATED WORK

In computer vision, the study of occlusion reasoning has been often confined to the context of stereo, motion (videos) and other multiple view problems [Stein & Hebert, 2009; He & Yuille, 2010; Sundberg et al., 2011]; Ayyaci et al., 2012; Weinzapefel et al., 2015], which depend on the availability of image sequences and technical issues, such as discovering corresponding pixels/patches.

For occlusion recovering within a single 2D image, in early vision, the recovering of depth order and occlusion relationship was first tackled by the block world of Roberts (1963), which is based on the single view geometry from lines. Later works extend the approach as well as handling curved objects [Nitzberg & Mumford, 1988; Cooper, 1997] with algebraic [Egiazarian, 1990; Sugihara, 1984] and markov random field (MRF)-based [Saund, 2005; Yu et al., 2001] formulations, while those techniques are limited to line drawings. For inference in natural images, Ren et al. (2006) first applied the two stage approach, i.e. extract features from the Pb boundary fragments (Martin et al., 2004) and learned a MRF model for classification. Leichter & Lindenbaum (2009) followed up by finding distributions of occlusion cues in ordinal depth: parallelity, image folds etc. Later, Hoiem et al. (2011) further extended the feature from local appearance to high-level semantic regions, making use of the extra learned geometric context models, e.g. ground, objects or sky etc. Additionally, a conditional random field (CRF) was designed to jointly infer the occlusion boundaries and geometric labels for classification. Most recently, Teo et al. (2015) proposed multiple features, e.g. HOG etc., for ownership cues and structure random forest (SRF) [Dollár & Zitnick, 2015] to recover the ownership and boundary simultaneously. To obtain the high level knowledge, they further proposed
the Gestalt-like grouping to access long range context. While carefully engineered, the system is somehow complex for training.

For 3D recovery from a single image, there also appeared a lot of works to estimate depth value from a single images with either MRF model (Hoiem et al., 2007; Saxena et al., 2009; Liu et al., 2010), deep learning (Eigen & Fergus, 2014) or combination of the both (Wang et al., 2015a; Liu et al., 2015a; Li et al., 2015). Implicitly, occlusion boundary can be recovered by finding the depth discontinuity from the estimated depth map. Nonetheless, these works normally require large amount of densely labelled depth ground truth images which is hard to obtain for nature images containing cluttered objects like that in PASCAL VOC. In addition, most works were focusing on scene layouts, making it not applicable to infer the complex occlusion interaction between objects.

Our method solves the previous difficulties by jointly detecting boundaries and recovering occlusion relationships using end-to-end DCNN learning which directly maps a given image to occlusion boundaries. This is easy to train and use, while producing strong performance. In our framework, the signal downsampling of DCNN is handled by jointly applying the multi-scale output, the "hole" algorithm (Chen et al., 2014a) and the deconvolution technique (Long et al., 2015), resulting in clear and accurate occlusion recovery.

3 PASCAL INSTANCE OCCLUSION DATASET

As indicated in Sec. 1, a large dataset is one of the keys to train a robust deep learning model. Thus, we build the PASCAL VOC instance occlusion boundary dataset by utilizing the existed instance segmentations. Specifically, we first directly derive the boundaries from the ground truth segments, and our task is to label the occlusion relationship between each object instance and its adjacent instances or backgrounds. We separate the labelling into the following two steps.

Label with line segments. Given an image, we provide two label maps to a labeller. The first one is the boundary map, and another is the semantic instance map. By default, we assume all the object instances are occluding the background, thus the user only need to label the boundaries between any two adjacent instances or the boundaries where an instance is occluded by the background. For each boundary needing labelling, a labeller labels it using piecewise directed line segments to indicate the occlusion relationship along the boundary, and the direction is following the “left” hand rules, i.e. the left side of a directed line segment occludes the right.

Line to contour matching. After the directed line segments are labelled, we developed a correspondence matching tool which maps all the labelled line segments to the boundaries of all instances. Similar to Hoiem et al. (2011), our ground truth occlusion boundaries of an image can be firstly represented by a set of boundary fragments. Each fragment is associated with a starting point, an ending point and a binary label id indicating the occlusion relationship. Finally, we map such a representation to an orientation map where each pixel on the instance boundary is assigned an orientation value indicating the local occlusion direction. At the left of Fig. 1, we show the images with our labelled results overlaid.

Statistics. At the right of Fig. 1 we show statistics of occlusion frequency between semantic categories. Each row indicates the frequency of a particular class occluding others, where we can see first, large amount of pixels are objects occluding the backgrounds, and in addition, the most frequently appeared object in the pascal is “person” as human commonly interacts with others. This helps us understand the PASCAL data biases which might be meaningful for us to use the data.

4 END-TO-END DEEP OCCLUSION RECOVERING

Typically, a successful end-to-end DCNN system contains two key components: 1) a pixel-wise loss function which penalizes reasonably for the error cases. 2) a well designed network structure that robustly capture the target information with reasonable level-of-context. In this section, we will introduce how we design the two components to solve the occlusion reasoning.

4.1 JOINT EDGE AND OCCLUSION INFERENCE

Existing representation for occlusion boundary reasoning. First, we start by discussing the representations of occlusion boundary in order to distinguish our contribution. Historically, researchers have proposed several labelling methods for learning the occlusion relationship over boundaries, which we summarize as follows (illustrated in Fig. 2): (a) label with binary foreground/background with the “left” rule (Ren et al., 2006; Leichter & Lindenbaum, 2009). For instance, in the figure, left of R_{t2} is background, thus we label $r_{t2} = 0$. This maps the occlusion problem to a binary labelling for each fragment of edge in-between detected junctions. (b) label with the semantic type (Hoiem et al., 2011) of foreground region. This method additionally infers the adjacent region types of a detected edge segment (e.g., $R_{t1}$ and $R_{t2}$ in Fig. 2), and then labels the edge segment to the foreground region type. (c) label with structure output as quantized orientation (Teo et al., 2015), which
Figure 2: Illustration of four ways to represent the occlusion boundary for learning, as elaborated in Sec. 4.1. In the figure, \( Fg \) and \( Bkg \) are short of foreground and background respectively. 1, 2, 3 are junction points and \( e_{12}, e_{23} \) are edges. \( R_{t1} \) and \( R_{t2} \) are regions of semantic type \( t1 \) and \( t2 \) respectively. \( L_q, L_o \) are the structure output for occlusion composed of pixels-wise labels. (a) Label with binary foreground/background. (b) Label with foreground region type. (c) Structure edge label with quantized orientations. (d) Structure edge label with continuous orientations.

uses a label patch with 8 quantized orientation values as structured output. (d) label with continuous orientation variable as we proposed, in which at each pixel we use the exact orientation value.

From the above descriptions, the first two settings (in Fig. 2(a) and (b)) are similar. The problems of such representation is the results heavily rely on the quality of junction and edge detection. An error detected junction or edge might negatively influence the whole model a lot as pixels are hardly grouped into edge segments which are locally connected in the models. To achieve flexibility and consistency of edge and occlusion prediction, the third form (Fig. 2(c)) use a structured label map that is robust to local confusion. However, in practice, the quantized orientation is less accurate, i.e. two very closed orientations can be separated as two quantized labels, yielding inaccurate distance measurement during learning. In our case (Fig. 2(d)), we perform structured regression to continuous orientation value with no loss of precision, while keeping the edge and occlusion consistent depended on a larger field of context.

**Formulation.** Formally, given an image \( I \), for each pixel, we assign it a joint label, \( l = \{e, \theta\} \), where \( e \in \{1, 0\} \) is an indicator, and \( e = 1 \) means the pixel is located on an occlusion boundary. \( \theta \in (-\pi, \pi] \) is an orientation variable along the boundary, i.e. when \( e = 1 \), indicating the direction of occlusion. As illustrated in Fig. 2(d) or Fig. 1 left, we follow the commonly used “left” rule [Ren et al. 2006; Hoiem et al. 2011], where the left side of the pixel, indicated by the orientation \( \theta \), occludes the right. For \( e = 0 \), we set \( \theta = NA \) and do not use them for the occlusion loss computation.

For training, we denote the set of training data as \( S = \{(I_i, \mathcal{L}_i)\}_{i=1}^{N} \), where \( N \) is the number of training images, and \( \mathcal{L}_i = \{L_{ei}, L_{oi}\} \) is the ground truth labelling maps, where \( L_{ei} \) is the boundary ground truth map and \( L_{oi} \) is the ground truth occlusion orientation map. Our goal is to design a DCNN that can learn a mapping function parametrized by \( W \), i.e. \( f(I, W) \), that can approximate the ground truth maps \( \mathcal{L}_i \).

To learn the parameters \( W \), our loss function for the training set is defined as:

\[
l_{doc}(S : W) = \frac{1}{N} \left( \sum_{i} l_e(I_i, L_{ei} : W) + \lambda \sum_{i} l_o(I_i, L_{oi} : W) \right)
\]

where \( l_e(I, L_e : W) \) is the loss for edge detection, and \( l_o(I, L_o : W) \) is the loss for occlusion recovering. The edge loss is defined as a balanced sigmoid cross entropy loss borrowed from the state-of-the-art edge detector [Xie & Tu 2015]. Here, we introduce the formulation of the loss for self-completeness:

\[
l_e(I, L_e : W) = -\beta \sum_{j: e_j = 1} \log P(e_j^* = 1 | I, W) - (1 - \beta) \sum_{j: e_j = 0} \log P(e_j^* = 0 | I, W)
\]

where \( e_j \) and \( e_j^* \) are the edge ground truth and edge prediction of pixel \( j \) respectively, and \( \beta = |L_e^*| / |L_e| \) is the percentage of pixels labelled as non-edge for image \( I \). \( P(e_j^* = 1 | I, W) \) is computed from the sigmoid output from the activation value on pixel \( j \).

Our occlusion loss is defined by penalizing strongly on the direction reversing of the predicted orientation \( \theta^* \) w.r.t. the ground truth orientation \( \theta \), while penalizing weakly if the prediction indicates
Choosing an effective network structure is another key for achieving good performance. In this work, we alternatively explored two multi-scale networks derived from the fully convolutional neural network (FCN) \cite{Long2015} structures to perform end-to-end learning of DOC, i.e. the holistic-nested edge detector (HED) \cite{Xie2015} network and the deeplab multi-scale large field of view DMLFOV \cite{Chen2014} network. We select to explore two networks for two reasons: (1) focusing on boundaries and occlusion recovery, the multi scale output technique is important to recover the spatial information. (2) the two networks are performing impressively over their corresponding tasks. The HED is the current best performer for the BSDS 500 edge detection \cite{Arbelaez2011}, and the DMLFOV is the state-of-the-art network for PASCAL semantic segmentation. The two networks represent the best for two level of semantics, which is valuable for us to explore for our occlusion boundary recovery. In our experiments, we separately train and compare them, yielding different rank of performance in different dataset varied with level of semantics required. Thus, we keep both of the networks in our description.

4.2 Network Structures

Choosing an effective network structure is another key for achieving good performance. In this work, we alternatively explored two multi-scale networks derived from the fully convolutional neural network (FCN) \cite{Long2015} structures to perform end-to-end learning of DOC, i.e. the holistic-nested edge detector (HED) \cite{Xie2015} network and the deeplab multi-scale large field of view DMLFOV \cite{Chen2014} network. We select to explore two networks for two reasons: (1) focusing on boundaries and occlusion recovery, the multi scale output technique is important to recover the spatial information. (2) the two networks are performing impressively over their corresponding tasks. The HED is the current best performer for the BSDS 500 edge detection \cite{Arbelaez2011}, and the DMLFOV is the state-of-the-art network for PASCAL semantic segmentation. The two networks represent the best for two level of semantics, which is valuable for us to explore for our occlusion boundary recovery. In our experiments, we separately train and compare them, yielding different rank of performance in different dataset varied with level of semantics required. Thus, we keep both of the networks in our description.

Preliminaries. For self-completeness, we first briefly introduce the structure of the two networks, illustrated in Fig. 4 for our task. The HED network is showed at the bottom row. It is a trimmed network from the VGG \cite{Simonyan2014} network without the fully connected layers, yielding better capturing of low-level local detailed edges required by the BSDS natural images. The final output of the network is a re-weighted fusion of all side outputs. At the top row of Fig. 4, we show the DMLFOV network. This network targets at high-level semantic segmentations, which contains the two fully connected layers while replacing the original (4096 dimension) with a much lighter parameter space (1024 dimension). In addition, the network introduced stronger side outputs by utilizing a two layer small network, which further help recover the semantic and segmentation details. We refer the readers to the original papers for more details.

Training phase. In our case, we perform several modifications in order to achieve optimal performance from each network: (1) to handle the signal down sampling brought by network pooling, we jointly use the “hole” algorithm \cite{Chen2014} and the deconvolution up-sampling strategy \cite{Long2015}. This allows us to perform high resolution loss comparison unlike the low resolution loss used by the DMLFOV when training segmentation, yielding precise localization for the boundary and occlusion predictions. (2) for joint prediction, i.e. boundary detection and occlusion recovery in this work, we adopt a two-stream network structure where each stream perform one task. As experimented by previous works \cite{Wang2015c,Liang2015}, usually two
stream network adapts better than a single stream networks when requiring the performance over joint tasks. For edge network, we keep the original network structures. For occlusion network, due to that occlusion is a higher level knowledge requiring longer range of context, we found lower-level predictions from the first two layers brings noisy output most of the time. Thus, for HED network, we drop the prediction before “conv3”, and for DMLFOV network, we drop the prediction before “pool4”, which produces robust orientation estimation and well localization for occlusion recovery.

Testing phase. Given an input image, a boundary map and an orientation map can be obtained from the two stream network. To combine the results, we first perform non-maximum suppression proposed in (Dollár & Zitnick, 2015), then the orientation value of each edge pixel can be retrieved from the orientation map at the corresponding place. Finally, we do the orientation rectification because at some pixels of the boundary, the predicted orientation is not exactly aligned with the tangent direction at the correspondent locations. In particular, we align the orientation to the tangent line, resulting in a reasonable prediction. Formally, at a pixel $j$, the predicted orientation and one direction of the tangent line is $\theta_j$ and $\theta_{tj}$ respectively. We set $\theta_j$ to $\theta_{tj}$ if $|\theta_j - \theta_{tj}| \mod 2\pi \in [0, \pi/2) \cup (3\pi/2, 2\pi]$, and to the reversed direction of $\theta_{tj}$ otherwise. Finally, we take $|\cos(|\theta_j - \theta_{tj}|)|$ as the confidence score for the occlusion prediction at pixel $j$.

5 Experiments

We experimented our DOC approach over two popular natural images datasets for occlusion reasoning, i.e. the BSDS boundary ownership dataset labelled by Ren et al. (2006) and our collected PASCAL instance occlusion dataset. The two dataset varies in terms of the level of semantics, where BSDS images contain sub-object level boundaries from object parts and backgrounds. Our dataset contains only boundaries around object instances. In this experiment session, we first propose a more reliable evaluation criteria, and then extensive experiments are conducted using the networks as elaborated in Sec. 4.2 As a result, our DOC approach outperforms the state-of-the-arts. However, there is not a single network dominants all the data, and we further performed extensive experiments which provides many insights for researchers to do further development.

Implementation details For line to contour matching in building our PASCAL instance occlusion dataset (Sec.3), we utilize the edge processing toolbox released by by Hoiem et al. (2011) to propose the boundary fragments based on the instance mask. We transfer the ground truth to each boundary fragment based on the corresponding labelled directed line segments. Additionally, we also labelled the 10k testing images from detection task on PASCAL, but we will keep it for an evaluation server later with respect to the PASCAL VOC test benchmark.

For the orientation loss function as Eqn. 3, we set $\alpha = 4$ and $\delta = 0.05$ respectively. We select $\alpha$ based on making the probability at 0 difference larger than 99%. For orientation loss in Eqn. 3, one may consider using $\cos$. However, we found it penalizes better of perpendicular cases but not for direction reversing. For the $\lambda$ in Eqn. 1, we do not need to tune it since a two-stream learning. For learning both of the networks, we also apply the deep supervision strategy, while keep the learning rate and stage-wise training the same with that of HED and DMLFOV. For initializing the models for the two stream network, we utilize the respective modified VGG model released by the authors.

For learning BSDS data, we follow the HED, and use adaptive input size for training and testing by setting the “batchsize” to 1 and “itersize” to 10. For learning PASCAL data, as the number of images are very large, we resize all the images to $386 \times 386$ by keeping the aspect ratio and padding.
With zeros, and we set the “batchsize” to 15 and “itersize” to 2. For both dataset, we augmented the data in the same way as the HED.

We implement all our model based on the published parsenet (Liu et al., 2015b) fork of Caffe (Jia et al., 2014), which includes both the “hole” algorithm and deconvolution layers.

5.1 Evaluation Criteria

For evaluation of the occlusion boundary, one utilized criteria is proposed by Ren et al. (2006). It computes the percentage of the pixels having occlusion correctly inferred inside pixels matching with the ground truth boundaries. However, such criteria is not standardized, which is also argued by Leichter & Lindenbaum (2009) that the evaluation results can vary in terms of the selected pixel matching method and the chosen of thresholding the edge detector, e.g. a higher threshold for edge selection may result in a higher accuracy. This leads to a non-reliable comparison between different algorithms of occlusion boundary recovering. Teo et al. (2015) released their code of evaluation, however, as illustrated in Fig. 5, we found two hidden problems that might result in non-reliable results. The first is the ground truth quantization error, as indicated in the white rectangle at left of Fig. 5. The continuous ground truth orientation is not only quantized into only 8 values, but also computed based on a local pixel-wise gradient (relying on a pair of neighboured pixels within 8 connections). This may produce error ground truth for a pixel. The second is results are evaluated based on a single fused map of multiple ground truth maps, where inconsistent ground truth edges labelled by different labellers are discarded. this may bias the evaluation positively since the error cases from label inconsistent are dropped.

For solving the two problems, we first propose to compute the orientation based on a local boundary fragment, as suggested by Hoiem et al. (2011), yielding a smoother and intuitively reasonable ground truth orientation for evaluation, as showed at the right of Fig. 5. Secondly, for evaluating the occlusion boundary, we propose a standardized criteria called the Occlusion accuracy w.r.t. boundary recall Curve (AOR curve), which borrows the idea from the PR curve utilized by edge detection. Formally, given the occlusion boundary recovering result under a threshold \( t \), we first find the correctly detected boundary pixels and their corresponding ground truth pixels by matching it to a ground truth map using the standard edge correspondence method (Arbelaez et al., 2011). For each pixel \( i \) on the recalled boundaries, its predicted orientation \( e_i \) if further compared to the corresponding ground truth orientation, i.e. \( e_i^* \). We keep the match if \( |e_i - e_i^*| \in [0, \pi/2) \cup (3\pi/2, 2\pi] \), while drop it as a false positive otherwise. From the matching of all the pixels, we obtain two values: the recall rate of the ground truth boundary \( R_e \), and the accuracy of orientation prediction given the recalled ground truth boundaries \( A_o \). By tuning the threshold, we can summarize the relationship between \( R_e \) and \( A_o \) into a curve comparing accuracy of the occlusion recovery under certain amount of boundary recall. In our experiments, for drawing the curves, we uniformly sample 33 thresholds and normalize the edge ranking scores in \([0, 1]\). Generally for AOR curve, the accuracy at high recall rate is more valued since more test data used for evaluation yields more reliable indication for the model’s general ability.

5.2 Performance Comparisons

In this session, we extensively compare our deep occlusion (DOC) approaches with different settings and configurations from the HED Xie & Tu (2015) and DMLFOV Chen et al. (2014a) networks.

---

1 We use the toolbox from the BSDS benchmark website.
Figure 6: Quantitative comparison between different models. SRF \citep{teo2015semantically} is the our baseline model. On the left we show the edge performance and we show the performance of DOC under various network configurations. Details can be found in Sec. 5.2 (best viewed in color).

We also compare the deep occlusion to the state-of-the-art occlusion recovering algorithm using structure random forests (SRF) \citep{teo2015semantically} over two popular datasets, in Fig. 6 we can see our best models are largely outperform SRF over both datasets over 6\%, which demonstrates the effectiveness of our approach.

BSDS occlusion dataset. The BSDS dataset contains 100 training images and 100 testing natural images. At the left of Fig. 6(a), we show the edge comparisons. HED performs best, DMLFOV is the runner up and SRF performs worst. For HED, this curve is not as good as that reported by \cite{xie2015holistically} due to two reasons: (1) we train merely on 100 images. (2) each image only contains two ground truth maps, yielding many more false positives and lower precision in this case. At the right of Fig. 6(a), we show the comparison for occlusion using our proposed AOR curve. Trained on just 100 images with single scale input, the DOC HED network (green line) performs the best, outperforming the SRF when the edge recall rate is higher than 0.3, and the margin goes above 4\% at high recall rate of 0.7. For the DOC DMLFOV network (pink line), due to the network contains a complex structure, it can hardly learn well with small amount of images to capture the local boundaries and orientations. Thus, the performance is lower than DOC HED network, but still competitive with SRF when the recall goes up to 0.7. Finally, as showed by DOC HED-PAS (blue line), if we fine-tune the DOC HED network based on the model pre-trained using our PASCAL instance occlusion data, the performance further boosts 3\%, yielding averagely 6\% improvement over the SRF model on BSDS ownership data. This proofs the general ability of a model learned from a large dataset like PASCAL.

PASCAL instance occlusion dataset. Our PASCAL instance occlusion dataset contains 10100 images, and we take 925 images from the VOC 2012 validation set for testing. At left of Fig. 6(b), we list the performance for semantic instance boundary detection. This is a high level boundary detection task where a lot of low-level edges are excluded. In the figure, we can see the DMLFOV provides the best performance, due to the fact that it captures stronger object-level context, while HED perform weaker in this case due to the false positive edges. In addition, we also show the models trained only using BSDS data, i.e. the "HED-BSDS" and "SRF-BSDS", where HED provides much better general ability in capturing semantic boundaries than the SRF.
Under review as a conference paper at ICLR 2016

Figure 7: Qualitative comparison between different models (best viewed in color). We use a threshold of 0.1 for better visualization. The “red” pixels with arrows plotted is correctly labelled occlusion boundaries, the “green” pixels are correct boundaries but error occlusion directions and the “blue” pixels are false positive boundaries. (a) Comparisons on BSDS ownership data between SRF (Teo et al., 2015) and our DOC HED model. (b) Comparisons on PASCAL occlusion data between DOC-HED model and DOC-DMLFOV model.

For occlusion recovering, as showed at right of Fig. 6(b), the performance of DOC DMLFOV is the best, while the improvement is marginal comparing to DOC HED model (i.e. around 1.5%) and worse than DOC HED when recall is higher than 0.78. This is because for the boundaries correctly recalled, DOC HED orientation also provide quit accurate estimations. More importantly, we evaluated the general ability of SRF and DOC HED models solely trained on BSDS. As showed by HED-BSDS and SRF-BSDS (black line and pink line), DOC HED network outperforms SRF significantly with a margin of 5% and is higher at every level of recall, indicating a better general ability of deep model. Several examples visualizing our results are showed in Fig. 7(b). Notice many of the false positive in the DOC predictions are actually intuitively correct but not labelled.

Additional comparisons from the two datasets. Here, to get more insights about the ability of different models, we perform further experiments.

Tuning of $\alpha$. As show in Fig. 6(a) (HED-$\alpha$: red curve), we set $\alpha$ to 8 and the performance slightly dropped because it penalizes too on the closeness between $\theta$ and $\theta^*$ as showed in Fig. 3.

Exchanging the boundary detector subnet. Due to the AOR curve is a joint evaluation of boundary and occlusion, one might be interested in solely replacing one information to see the effect of individual factor. In Fig. 6(a) (HED+DL-ORI: black curve), we use HED boundary detector but DOC DMLFOV for orientation map, in which the performance improved by 2%, and in Fig. 7(b) (DMLFOV+HED-ORI: green curve), we apply the same strategy but using DMLFOV boundary and HED orientation, and the results is similar. This means correct boundary detection is important for retrieving the right occlusion orientation.

Deeper network. In our experiments, the DMLFOV model is deeper and provides large field of view than the HED model, which generally performs better in recognition tasks. However, though based on multi-scale technique, deeper model does not success on the BSDS due to low-level edges required, as showed in Fig. 6(a). However, human can perform the best over both data set, thus we believe further efforts is needed in improving the scale of data or the network structures. While for orientation, in general, the DOC HED model always gives strong results, indicating the occlusion can be well inferred using only 5 layers that can help us reducing the network size.

Multi-scales network vs. Single scale network. We perform a comparison of fused output vs. single side output (from the “fc8” layer) of orientation based on the DMLFOV model. As showed at right of Fig. 6(b) (DLFov-Fc8: cyan curve), it gives weaker performance since multi-scale can provide stronger localization of orientation, resulting in better result.

High resolution loss vs. Low resolution loss. Compared with the original loss based on downsampled ground truth used by DMLFOV for training semantic segmentation (Chen et al., 2014a), our loss is computed from the label map in original resolution. Although the training time is increased, as compared at left of Fig. 6(b) with low resolution model (DMLFOV-LR: pink curve), such a strategy
boots the semantic boundary detection accuracy. Same improvement is also obtained for occlusion inference.

Scales of input images. Using multi-scale training and testing usually gives better results. At left of Fig. 6(a) (HED mean: cyan curve), we show the results from averaging of three images scales ([0.5, 1.0, 1.5]) output from the DOC HED network. In our case, multi-scale provides marginal improvement. This indicates that for boundaries detection, multi-scale network and multi-scale input contains similar information.

6 CONCLUSION AND FUTURE WORK
In this paper, we propose an end-to-end deep occlusion (DOC) system recovering the occlusion boundaries from a single image, yielding large improvement over the state-of-the-art methods. Additionally, a large scale instance occlusion boundary dataset from the PASCAL VOC images is established, which significantly benefit the visual learning. From our experiments, currently there is not a dominant network structure that can perform best over both the high level and low level boundary detections simultaneously, which asks for either improvement on the scale of edge data beyond BSDS, or optimizing the network structure.

REFERENCES


Liu, Wei, Rabinovich, Andrew, and Berg, Alexander C. Parsenet: Looking wider to see better. *CoRR*, abs/1506.04579, 2015b.


