The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Multi-Scale Bidirectional FCN for Object Skeleton Extraction

Fan Yang,[1] **Xin Li,**[1*] **Hong Cheng,**[2] **Yuxiao Guo,**[1] **Leiting Chen,**[1] **Jianping Li**[1]
[1]School of Computer Science & Engineering, University of Electronic Science and Technology of China
[2]Center for Robotics, School of Automation Engineering, University of Electronic Science and Technology of China
fanyang_uestc@hotmail.com, xinli_uestc@hotmail.com, hcheng@uestc.edu.cn

Abstract

Object skeleton detection is a challenging problem with wide application. Recently, deep Convolutional Neural Networks (CNNs) have substantially improved the performance of the state-of-the-art in this task. However, most of the existing CNN-Based methods are based on a skip-layer structure where low-level and high-level features are combined and learned so as to gather multi-level contextual information. As shallow features are too messy and lack semantic knowledge, they may cause errors and inaccuracy. Therefore, we propose a novel network architecture, Multi-Scale Bidirectional Fully Convolutional Network (MSB-FCN), to better capture and consolidate multi-scale *high-level* context information for object skeleton detection. Our network uses only deep features to build multi-scale feature representations, and employs a bidirectional structure to collect contextual knowledge. Hence the proposed MSB-FCN has the ability to learn the semantic-level information from different sub-regions. Furthermore, we introduce dense connections into the bidirectional structure of our MSB-FCN to ensure that the learning process at each scale can directly encode information from all other scales. Extensive experiments on various commonly used benchmarks demonstrate that the proposed MSB-FCN has achieved significant improvements over the state-of-the-art algorithms.

Introduction

Object skeleton detection, also known as object symmetry extraction, is an important topic of great interest to computer vision researchers. It is aimed at localizing the object symmetry axes in an image. Object skeleton detection can be used in a variety of tasks, such as image segmentation (Teo, Fermuller, and Aloimonos 2016), text-line detection (Zhang et al. 2015), object proposal generation (Lee, Fidler, and Dickinson 2015), foreground extraction (Fu et al. 2014), and 3D object structure estimation (Gao and Yuille 2017).

To automatically extract object skeletons from natural images is a challenging vision task due to varied appearances, self occlusion, object diversity and scene variety. Traditional methods (Tsogkas and Kokkinos 2012) (Sironi, Lepetit, and Fua 2014) have difficulties in handling images with complex scenes, or objects with cluttered interior textures. This is mainly caused by the limitations of traditional handcrafted features, which are unable to accurately summarize semantic knowledge. In recent years, deep Convolutional Neural Networks (CNNs) have demonstrated a strong capability of learning semantic-level representations. This has motivated recent research efforts to employ CNNs for object skeleton detection. The CNN-Based methods (Shen et al. 2016c) (Ke et al. 2017) have substantially improved the performance, achieving accurate results even when handling complex "in-the-wild" scenes.

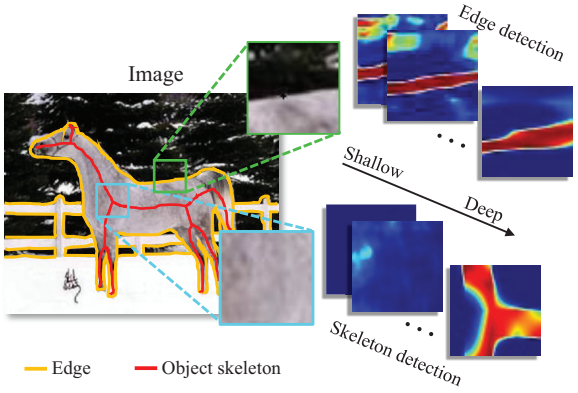


Figure 1: Object skeleton detection is a high-level vision task. It is inherently different from edge detection. Edge has some local properties, and thus shallow CNN features can benefit edge detection. In contrast, object skeleton is more of a semantic-level concept, and in object skeleton extraction, deep features play a much more important role.

In order to deal with the scale-space problem, existing CNN-Based methods (Shen et al. 2016c) (Ke et al. 2017) adopt the Holistically-Nested Edge Detector (HED) network architecture (Xie and Tu 2015) that was originally designed for edge detection. With its skip-layer structure, HED learns rich hierarchical representations to resolve ambiguity in edge. However, object skeleton extraction is an inherently different problem from edge detection (Wang, Zhao, and Huang 2017). As shown in Fig. 1, the edge has some local properties and unique local structures. Therefore, shallow CNN features, despite their lack of semantic-knowledge, are

*Corresponding author.

still capable of summarizing local information for edge detection. In contrast, object skeleton is more of a semantic-level concept, which can hardly be detected by using local or shallow features. This leaves open two questions. Is it a good strategy to use both shallow and deep features to learn and infer contextual relations for object skeleton detection? Is there a better strategy to consolidate multi-scale context information for this high-level vision task?

To answer these questions, we propose a novel *Multi-Scale Bidirectional Fully Convolutional Network* (MSB-FCN) to take full advantage of multi-scale context information for object skeleton detection. The proposed MSB-FCN can generate feature maps of different scales directly from the last convolutional layer of a pre-trained underlying model by using a pyramid pooling strategy (Zhao et al. 2017). Moreover, our MSB-FCN employs a bidirectional structure to capture and encode multi-context information. Different from existing CNN-Based models (Shen et al. 2016c) (Ke et al. 2017), our model focuses only on *high-level* contextual relationship. Therefore, semantic information loss caused by shallow features has little influence on our method. In addition, we introduce dense connections to the bidirectional network architecture within our MSB-FCN framework. All feature maps are densely connected in a feed-forward fashion so that the learning process at each contextual level can encode more useful information and context priors which can help further improve the performance.

Our MSB-FCN achieves the currently best performance on four widely-used datasets. Its performance is further improved by the proposed dense connections. Moreover, the proposed MSB-FCN is faster than most existing methods. In summary, the contributions of this work are three folds:

- We propose a novel Multi-Scale Bidirectional Fully Convolutional Network (MSB-FCN) for object skeleton detection in natural images. With the multi-scale bidirectional network, the proposed MSB-FCN is intrinsically able to capture and combine multi-level context information during learning and inferring.
- We introduce dense connections to strengthen feature propagation and enrich context information, which is useful to improve the final skeleton detection performance.
- The proposed method significantly improves the results of the state-of-the-art on four widely-used benchmarks, and also maintains high efficiency.

Related Work

Object skeleton extraction has evolved greatly over the past decades. Originally, object skeleton detection algorithms (Lam, Lee, and Suen 2002) (Saha, Borgefors, and Baja 2016) were designed for extracting symmetry axes from binary images. Because these methods require pre-segmented masks of objects, they cannot be applied in many real-life tasks. Recently, researchers have tried to extract the object skeletons directly from natural images. These attempts can be categorized into two groups: Traditional methods and CNN-Based methods. The following is a brief review of these works.

Traditional methods Traditional object skeleton detection methods mainly depend on different handcrafted image features. For example, by using shape context-like features, Levinshtein *et al.* (Levinshtein, Dickinson, and Sminchisescu 2010) propose to train a classifier to measure appearance affinity between two adjacent superpixels, and then group together medial branches that belong to the same object. Tsogkas *et al.* (Tsogkas and Kokkinos 2012) introduce a learning-based framework, namely Multiple Instance Learning (MIL) framework, that exploits low-level features of different scales to train a symmetry detector. However, as symmetry and non-symmetry pixels are quite confusing, training a global model may have difficulties in differentiating these pixels in clutters scenes. To solve this problem, Shen *et al.* (Shen et al. 2016b) propose to train a group of MIL classifiers on well partitioned subspaces. This method further improves the performance. Similarly, in (Lee, Fidler, and Dickinson 2014), a number of extensions are introduced to MIL framework. In general, by using different handcrafted features, traditional methods can accurately detect object skeletons in many simple cases. However, they are still unable to deal with some cluttered scenes or objects with a complex structure due to the lack of high-level semantic knowledge.

CNN-Based methods Recently, some CNN-Based methods have been proposed to learn high-level feature representations for object skeleton detection. Shen *et al.* (Shen et al. 2016c), for the first time, present a fully convolutional network with multiple scale-associated deep side outputs (FSDS) to detect the object skeletons in natural images. Ke *et al.* (Ke et al. 2017) introduce a Side-output Residual Network (SRN) to extract object skeletons in a deep-to-shallow manner. SRN can adapt to symmetry scales without needing any scale-level annotation, and it achieves better performance. Basically, all existing CNN-Based methods adopt the architecture of HED (Xie and Tu 2015), and use both shallow and deep features to learn contextual relationship for object skeleton detection. By employing the learned CNN features, these models have successfully broken the limits from traditional methods, and achieved new state-of-the-art performance. However, for shallow features, their lack of semantic knowledge may harm their representational capacity for object skeleton detection. Therefore, there is a great room for improvement over these existing methods.

Method

This section starts with an overview of the MSB-FCN, followed by a detailed description. Then, we describe how to improve detection accuracy by using a series of connections among multi-scale intermediate feature layers within our bidirectional network architecture.

Overview

Our method is based on our observation that high-level semantic knowledge is much more important than low-level local information in object skeleton detection (see Fig. 1). To accurately localize object symmetry axes in an image, the network should be capable of i) producing semantically

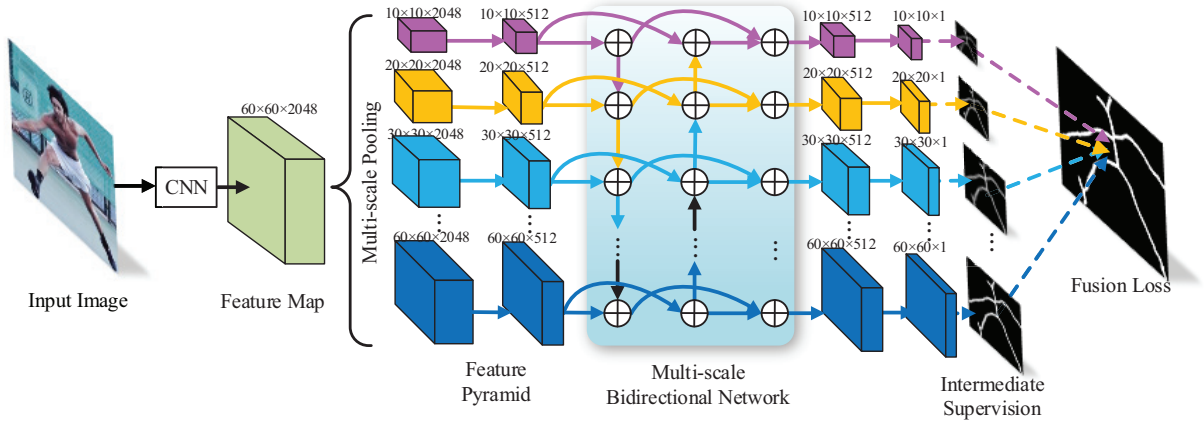


Figure 2: Overall of the Multi-Scale Bidirectional FCN. Our MSB-FCN is built upon a pre-trained ResNet-based feature map, and contains a feature pyramid for handling scale variations of object(s). A multi-scale bidirectional network with intermediate supervision is employed to learn and gather multi-scale features. Then, the multi-scale intermediate perditions are merged by fusion module to generate the final result.

strong features with high-resolution, ii) handling scale variations, and iii) gathering *high-level* context information from different sub-regions. To achieve this goal, we introduce a novel MSB-FCN that can meet all the requirements above.

As can be seen in Fig. 2, given an input RGB image, our model produces an object skeleton map of the same size to be the output. The MSB-FCN is built upon a ResNet-Based feature map, and employs a multi-scale bidirectional network architecture. We use the dilated network technique (Yu and Koltun 2015) to ensure that the last three groups of ResNet-101 (He et al. 2016) have the same resolution. The last convolutional layer of ResNet-101 is used as the underlying feature map to build a feature pyramid. The multi-level feature representations are then learned by using a bidirectional network architecture, in which local and global context knowledge can be effectively gathered and combined in both coarse-to-fine and fine-to-coarse manners. Supervision is imposed on the intermediate output layer to "guide" the prediction at every specific scale. Finally, a fusion module is used to automatically learn the combination weight of these intermediate predictions and generate the final result. Note that our model is a single-stage network, where all modules mentioned above are highly integrated and jointly learned under the proposed framework.

Multi-Scale Bidirectional FCN

The MSB-FCN adopts a ResNet-based feature map F of input image I, and produces a full-resolution object skeleton map $O(F, \Theta)$ where Θ represents all standard network layer parameters. As mentioned above, we use the dilated network technique (Yu and Koltun 2015) (Chen et al. 2016) to produce a high-resolution underlying feature map f, which can avoid spatial information loss and benefit the subsequent learning process. The MSB-FCN consists of three important modules: feature pyramid, multi-scale bidirectional network and fusion module.

Feature pyramid To accurately detect object skeletons in natural images, the first important problem is the scale-space problem. A straight-forward solution, as many previous methods used (Hu and Ramanan 2017) (He et al. 2017), is to build an image pyramid for feature extraction. However, the image pyramid is computationally complex due to redundant convolutional computation over multi-scale images. To address the computational bottleneck, we build a feature pyramid based on the ResNet-based feature map F for multi-scale feature learning and inferring by using the pyramid pooling strategy (Zhao et al. 2017). Specifically, the underlying feature map F is pooled into N feature maps of different bin sizes to form the feature pyramid. Furthermore, we decrease the dimensions of pooled features by using one 1×1 convolution layer (from 2048 to 512), so that the computation workload of the following learning process can be reduced. We denote the reduced feature map within the feature pyramid as $f_{s_i} \in \{f_{s_1}, f_{s_2} ... f_{s_N}\}$, and use them to learn multi-scale contextual relationship for object skeleton detection.

Multi-Scale bidirectional network We propose a multi-scale bidirectional network to learn and consolidate multi-level context information. Specifically, it uses multi-scale feature maps $\{f_{s_i}\}$ as input and produces intermediate predictions of multiple resolutions $\{o_{s_i}\}$. In the network, there are two directional pathways among multi-scale features $\{f_{s_i}\}$ so that the learning process at each scale can benefit from both more global and more detailed information.

The coarse-to-fine pathway starts from the coarsest feature map with the most global information, and ends at the finest feature map with the highest resolution. In this pathway, the feature learning for the i-th scale encodes the reduced feature f_{s_i} of its own scale and the learned feature representation $h^{CF}_{s_{i-1}}$ of its previous scale s_{i-1}. The feature representations can be written as:

$$h_{s_i}^{cf} = \begin{cases} f_{s_i} + \alpha_i \mathrm{R}_{s_i}(h_{s_{i-1}}^{CF}), i \geq 2 \\ f_{s_i}, i = 1 \end{cases} \quad (1)$$

where $h_{s_i}^{cf}$ denotes the updated feature map, and α represents the combination weight. $R_{s_i}(\cdot)$ is used to resize the updated feature map to the same size of current scale s_i through bilinear interpolation. The updated feature map $h_{s_i}^{cf}$ is further learned by using the convolution operation and RELU to form the learned feature map $h_{s_i}^{CF}$ which serves as the input for the next scale.

On the other hand, the fine-to-coarse pathway is completely opposite. This pathway collects the multi-scale information from local to global, and the feature map of the i-th scale can be computed below:

$$h_{s_i}^{fc} = \begin{cases} f_{s_i} + \beta_i \mathrm{R}_{s_i}(h_{s_{i+1}}^{FC}), i \leq (N-1) \\ f_{s_i}, i = N \end{cases} \quad (2)$$

where $h_{s_i}^{fc}$ denotes the updated feature map produced by combining the multi-level features in a finer-to-coarse manner, where β represents the combination weight. We adopt the same procedure as that used in the coarse-to-fine pathway to learn the feature representation $h_{s_i}^{FC}$.

The learned feature maps (*i.e.*, $h_{s_i}^{CF}$ and $h_{s_i}^{FC}$) of each scale are then merged to form the final feature representations M_{s_i} for a specific scale s_i,

$$M_{s_i} = \sigma(cat(h_{s_i}^{CF}, h_{s_i}^{FC}) \otimes W_{s_i} + b_{s_i}), \quad (3)$$

where $\otimes$ represents the convolution operation, and $cat(\cdot)$ is used to combine the two learned feature maps of different directions. W_{s_i} and b_{s_i} represent the convolutional filters and biases, respectively. RELU serves as the non-linear function $\sigma(\cdot)$.

For the output of any specific scale (or resolution), the intermediate prediction error is computed by using the standard cross-entropy loss function,

$$\begin{aligned} l_{s_i}(\Theta, \theta_{s_i}) = -\sum_{p=1}^{|I_{s_i}|} & G_{s_i}(p) \log \Pr(o_{s_i}(p) = 1 | I_{s_i}; \Theta, \theta_{s_i}) \\ & + (1 - G_{s_i}(p)) \log \Pr(o_{s_i}(p) = 0 | I_{s_i}; \Theta, \theta_{s_i}). \end{aligned} \quad (4)$$

where $G_{s_i}(p)$ denotes the groundtruth label at pixel p. Note that the loss function is computed over all pixels in every training image with a different scale s_i.

Fusion module Given an input RGB image I, the multi-scale bidirectional network produces several intermediate object skeleton maps $\{o_{s_i}\}$ with different scales $\{s_i\}$. We use a fusion module to efficiently combine these object skeleton maps and then generate the final result O.

The fusion module is composed of a connection layer, two convolutional layers, and a loss layer. It takes N resized intermediate object skeleton maps by the connection layer to generate a N-channel map. Then the resulting map is forwarded through two convolutional layers. Finally, the following loss function is used to compute the errors between the final prediction O and the ground-truth G,

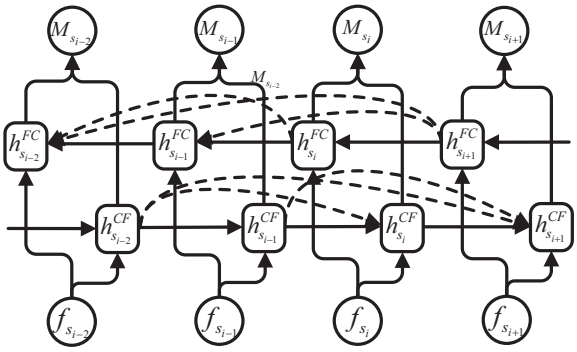


Figure 3: Illustration of dense connections. With dense connections, the feature learning in each scale can directly encode features from all the other scales.

$$\mathcal{L}_{fuse}(\Theta, \theta, h) = Dist(G, O) \quad (5)$$

where h denotes the layer parameters of the fuse module, and $Dist(\cdot)$ represents the distance function used to compute the pixel-wise distance between the final object skeleton map O and the groundtruth map G.

Objective function Our model is a single-stage network, where all the modules above are integrated into one framework. The loss function of the whole MSB-FCN is given as:

$$\mathcal{L}_{all}(\Theta, \theta, h) = \xi_{s_f} \mathcal{L}_{fuse}(\Theta, \theta, h) + \sum_{m=1}^{N} \gamma_{s_m} l_{s_m}(\Theta, \theta_{s_m}), \quad (6)$$

where ξ_{s_f} and γ_{s_m} represent balance weights.

We formulate the object skeleton detection task as per-pixel regression to groundtruth annotations. In the training phase, the objective of MSB-FCN is to minimize error between ground-truths and estimated object skeleton maps, and can be written as:

$$(\Theta, \theta, h)^* = \arg\min(\mathcal{L}_{all}(\Theta, \theta_{s_i}, h)) \quad (7)$$

In the inferring phase, the input image I is simply forwarded through the MSB-FCN to generate a full-resolution object skeleton map $O(F, \Theta)$.

Dense Connections

To make our MSB-FCN more powerful, we introduce a series of connections among multi-level feature layers, which are called dense connections. Dense connections can enrich context information and strengthen underlying representations for feature learning. With dense connections, the learning process at each scale is able to directly encode context knowledge from all the other scales (not limited to its neighboring scale). In this way, the learning process can benefit more from multi-level information.

As can be seen in Fig. 3, the feature layer of each scale within the bidirectional architecture is connected to all the other layers in a feed-forward manner. With dense connections, the updated feature maps for the i-th scale in the

coarse-to-fine pathway and the fine-to-coarse pathway have been changed. Then, Eq.1 and Eq.2 become

$$\tilde{h}_{s_i}^{cf} = \begin{cases} f_{s_i} + \sum_{j=1}^{i-1} \alpha_i{}^{(j)} \mathrm{R}_{s_i}(\tilde{h}_{s_j}^{CF}), i \geq 2 \\ f_{s_i}, i = 1 \end{cases} \quad (8)$$

and

$$\tilde{h}_{s_i}^{fc} = \begin{cases} f_{s_i} + \sum_{j=i+1}^{N} \beta_i^{(j)} \mathrm{R}_{s_i}(\tilde{h}_{s_j}^{FC}), i \leq (N-1) \\ f_{s_i}, i = N \end{cases} \quad (9)$$

where $\alpha_i^{(j)}$ and $\beta_i^{(j)}$ are the weights of connections from the feature layer with j-th scale to feature layer with i-th scale. $\tilde{h}_{s_i}^{CF}$ and $\tilde{h}_{s_i}^{FC}$ denote the newly learned feature maps which are generated by performing the convolution operation and RELU on their corresponding updated maps.

Also, we rewrite Eq. 3 to generate the final feature representation $\tilde{M}_{s_i}$ for each scale s_i,

$$\tilde{M}_{s_i} = \sigma(cat(\tilde{h}_{s_i}^{CF}, \tilde{h}_{s_i}^{FC}) \otimes \tilde{W}_{s_i} + \tilde{b}_{s_i}), \quad (10)$$

where $\tilde{W}_{s_i}$ and $\tilde{b}_{s_i}$ are convolutional filters and biases for the i-th scale, respectively. The new fusion loss function can be represented by

$$\tilde{\mathcal{L}}_{fuse}(\Theta, \tilde{\theta}, h) = Dist(G, \tilde{O}), \quad (11)$$

and therefore the loss function for the whole MSB-FCN becomes:

$$\tilde{\mathcal{L}}_{all}(\Theta, \tilde{\theta}, h) = \xi_{sf} \tilde{\mathcal{L}}_{fuse}(\Theta, \tilde{\theta}, h) + \sum_{m=1}^{N} \gamma_{s_m} \tilde{l}_{s_m}(\Theta, \tilde{\theta}_{s_m}), \quad (12)$$

where $\tilde{l}_{s_m}$ denotes the cross-entropy loss which has been defined in Eq. 4.

Our model is inherently able to capture and gather multi-scale *high-level* context information for object skeleton detection. Dense connections enable our MSB-FCN to be more powerful in consolidating the multi-level feature representations. It is different from all existing deep object skeleton networks (Shen et al. 2016c) (Ke et al. 2017) that do not explicitly deal with the high-level contextual relationship. In terms of architecture, our model can be considered as a novel *coarse-to-fine & fine-to-coarse* learning-based approach: multi-scale features are jointly learned and combined with a directed cycle for object skeleton detection.

Experiments

In this section, we describe implementation details of our model, introduce the datasets and evaluation metrics, give ablation analysis and provide exhaustive comparison results over four widely-used datasets.

Implementation

In general, our network is based on the publicly available platform Caffe (Jia et al. 2014). We use ResNet-101 (He et al. 2016) as the pre-trained model. The dilated network strategy (Yu and Koltun 2015) is adopted to ensure the last three groups of ResNet-101 have the same resolution (*i.e.*, 60×60). Under our MSB-FCN, we use four scales (*i.e.*, 10×10, 20×20, 30×30 and 60×60) to handle scale variations of object(s). Then, multi-level features are jointly learned to generate the final result.

Training details The input image is resized such that its resolutions become 480×480 pixels. In Eq. 12, we define $\xi_{s_f} = 2$ and $\gamma_{s_m} = 1$ so as to emphasize the final output. We use the "poly" learning rate policy (Liu, Rabinovich, and Berg 2015), where the learning rate is automatically controlled by $(1 - \frac{iter}{max_{iter}})^{power}$. The initial learning rate is set to 10^{-8}, and the $power$ is set to 0.9. We set the maximum number of iterations to $60K$. The Stochastic Gradient Descent (SGD) is employed for optimization. The outputs of different scales are also resized to 480×480 pixels to compute the loss. To reduce overfitting, the training data is augmented by rotating all the training images by every 90 degrees, flipping them with different axes, and resizing them to three different scales (0.8, 1.0, 1.2), following (Shen et al. 2016c).

Inferring During the inferring phase, the input RGB image is simply forwarded through the MSB-FCN to produce a full-resolution object skeleton map. The resulting object skeleton map is then used to generate object skeletons through a non-maximal suppression (NMS) algorithm (Dollar and Zitnick 2014), as many previous methods did (Shen et al. 2016c) (Ke et al. 2017).

Datasets and Evaluation Metrics

Datasets We evaluate the proposed MSB-FCN model on four widely-used benchmark datasets: SK-SMALL (Shen et al. 2016c), WH-SYMMAX (Shen et al. 2016b), SK-LARGE (Shen et al. 2016a) and Sym-PASCAL (Ke et al. 2017).

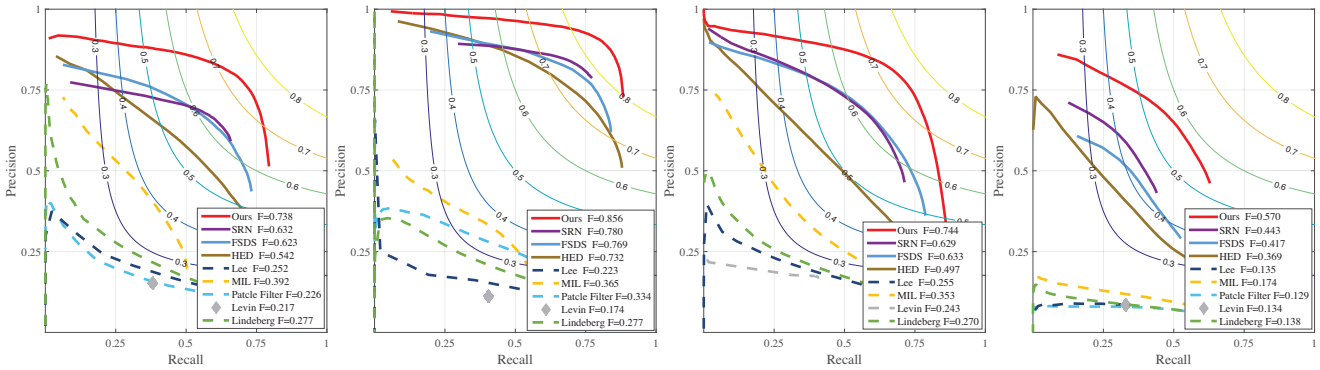
SK-SMALL is one of the most popular datasets, which includes 300 training images and 206 testing images. It includes object skeletons with large variances in both structure and scale.

WH-SYMMAX is derived from the well-known Weizmann Horse dataset (Borenstein and Ullman 2002). It contains 328 images. The first 228 images serve as training images and the rest are used for testing.

SK-LARGE is a recently introduced benchmark with annotation of object skeletons. It includes 746 images for training and 745 images for testing.

Sym-PASCAL is a very challenging dataset. It is converted from the PASCAL-VOC-2011 dataset (Everingham et al.), and contains 648 training images and 787 testing images in 20 object classes.

Evaluation metrics To evaluate skeleton detection performance, we adopt the precision-recall metric with F-measure by following the steps in (Shen et al. 2016a) (Ke



(a) SK-SMALL (b) WH-SYMMAX (c) SK-LARGE (d) Sym-PASCAL

Figure 4: The Precision-Recall curves (PR curves) on SK-SMALL, WH-SYMMAX, SK-LARGE and Sym-PASCAL.

Table 1: Quantitative comparison of the state-of-the-art approaches on the SK-SMALL, WH-SYMMAX, SK-LARGE and Sym-PASCAL dataset. The top three results are shown in Red, Green, and Blue, respectively.

Method	Levin	Lee	Lindeberg	Filter	MIL	HED	FSDS	SRN	Ours
SK-SMALL	0.217	0.252	0.227	0.226	0.392	0.542	0.623	0.632	0.738
WH-SYMMAX	0.174	0.223	0.277	0.334	0.365	0.732	0.769	0.780	0.856
SK-LARGE	0.243	0.255	0.270	—	0.353	0.497	0.633	0.629	0.744
Sym-PASCAL	0.134	0.135	0.138	0.129	0.174	0.369	0.418	0.443	0.570

Table 2: Analysis of the MSB-FCN. Our results are obtained on Sym-PASCAL benchmark. "MSB" refers to the multi-scale bidirectional architecture we designed; "DC" denotes the dense connections; "*" represents the method used in this paper.

Method	F_β
HED (Xie and Tu 2017)	0.369
FSDS (Shen et al. 2016c)	0.418
SRN (Ke et al. 2017)	0.443
Faster-RCNN (Ren et al. 2017)+FSDS	0.343
YOLO (Redmon et al. 2016)+FSDS	0.354
ResNet-FCN (baseline)	0.515
ResNet-FCN+MSB (ours)	0.557
*ResNet-FCN+MSB+DC (ours)	**0.570**

et al. 2017). The Precision-Recall curve (PR-curve) is obtained by matching the binary maps which are thresholded by using different values with the groundtruth skeleton map. The maximum F-measure (F_β) is defined as: $F_\beta = \frac{2 \cdot precision \cdot recall}{precision + recall}$.

Ablation Study

In this section, we first explore two aspects of our design: the effectiveness of multi-scale bidirectional architecture and the necessity of dense connections. The ResNet-101 architecture with the dilated network technique (ResNet-FCN) is used as the baseline to show the value of our design. Against the baseline, we analyze the proposed components including Multi-Scale Bidirectional architecture (MSB) and Dense Connections (DC) by comparing the F-measures. The overall result on Sym-PASCAL benchmark is shown in Tab. 2. The experiments show that the ResNet101-based FCN with the proposed multi-scale bidirectional architecture can largely improve the accuracy of skeleton detection. It achieves 8.16% improvements on F-measure over the baseline. Dense connections enable MSB-FCN to result in a further 2.3% improvement according to F-measure, achieving the highest accuracy ($F_\beta = 0.570$). We find that the improvements should be attributed to the multi-scale bidirectional architecture and dense connections, which are the main contributions of this work.

In addition, to show the superiority of the end-to-end pipeline in complex scenes, we compare our single-stage MSB-FCN as well as other end-to-end solutions (Xie and Tu 2017) (Shen et al. 2016c) (Ke et al. 2017) with some two-stage methods. Specifically, we first detect all object(s) from an image by using the current best detection methods including Faster R-CNN (Ren et al. 2017) and YOLO (Redmon et al. 2016). Then, we use the skeleton method FSDS (Shen et al. 2016c) to extract the object skeletons based on the detection results. As can be seen in Tab. 2, the single-stage framework is more effective in object skeleton detection.

It also should be noted that a ResNet-Base model can achieve very accurate results for object skeleton detection. Our baseline model, ResNet-101 with the dilated network technique, achieves a higher accuracy ($F_\beta = 0.515$) than previous best result ($F_\beta = 0.443$) produced by the VGG-Based model (Ke et al. 2017). This can be explained by the fact that a very deep ResNet-101 can produce semantically stronger features than the VGG-16 network, and these semantically strong features are very important in this high-

Figure 5: Qualitative comparisons of our method and the state-of-the-art CNN-Based methods on some challenging scenes. Failed examples are shown in the last two columns.

Table 3: Comparison of running times. Mean run-times were measured on Sym-PASCAL.

Method	Levin	Lee	Lindeberg	Filter	MIL	HED	FSDS	SRN	Ours
Runtime(s)	181.87	647.94	6.79	28.32	82.35	0.10	0.12	0.12	0.13

level vision task.

Comparison to Other Methods

We compare the proposed method with three leading CNN-Based methods including HED (Xie and Tu 2017), FSDS (Shen et al. 2016c) and SRN (Ke et al. 2017). We also compare our approach with five top-ranked traditional methods: Lindeberg (Lindeberg 1996), Levinshtein (Levinshtein, Dickinson, and Sminchisescu 2010), MIL (Tsogkas and Kokkinos 2012), Lee (Lee, Fidler, and Dickinson 2014), and Partical Filter (Widynski, Moevus, and Mignotte 2014).

Quantitative comparison We compare the performance of our model with the state-of-the-art methods on four challenging benchmarks. As can be seen in Tab. 1, our MSB-FCN achieves the best performance according to F-measure. Specifically, on SK-SMALL, WH-SYMMAX, SK-LARGE and Sym-PASCAL, our MSB-FCN significantly improves the current best F-measure by $\mathbf{16.8}\%$, $\mathbf{9.7}\%$, $\mathbf{17.5}\%$ and $\mathbf{28.7}\%$, respectively. Furthermore, we compare our approach with the other methods in terms of PR curve. As is shown in Fig.4, our model consistently outperforms all the state-of-the-art methods. Because of the multi-scale bidirectional network architecture, the learning and inferring process can benefit from both more global and more detailed information. As a result, the generated object skeleton maps are much closer to the groundtruth annotations, which results in a significantly better F-measure and a much higher PR curve. It is also worth mentioning that thanks to the dense connections, our method is able to encode more useful contextual knowledge, yielding more accurate results.

Qualitative comparison A qualitative comparison is shown in Fig.5. As can be seen, our method achieves more accurate object skeleton detection results than all other methods, which are the closest to the ground truth. When handling images containing a cluttered background and objects with a complex structure, the proposed MSB-FCN can still generate very reliable results while the other methods fail in these cases.

Speed performance As is shown in Tab.3, it takes our method only about 0.13 seconds to generate an object skeleton map for one input image. The traditional methods are tested on a PC with an i7 2.50 GHz CPU and 8 GB RAM, while the CNN-based methods are accelerated by a NVIDIA GTX 1080ti GPU X 11G. We observe that the proposed MSB-FCN and other CNN-Based methods achieve similar speed performance. They are much faster than all traditional methods.

Conclusions

In this paper, we have proposed MSB-FCN, a novel CNN-based model for object skeleton detection. Our network employs a multi-scale bidirectional network architecture that can better gather and capture multi-scale *high-level* contextual relationship. Dense connections are used to enable the feature learning at each scale encode the knowledge directly from all other feature maps of different scales. Our MSB-FCN is end-to-end trainable, highly integrated and very powerful. Extensive experiments on four widely-used benchmarks demonstrate that the proposed method significantly outperforms the state-of-the-art methods.

Acknowledgement. This work was supported in part by the National Key Research and Development Program of China (NO.2017YFB0102500), and the NSFC (No.U1613223 and No.61573084). Y. Fan' participation was supported by the NSFC (No.61370073).

References

Borenstein, E., and Ullman, S. 2002. Class-specific, top-down segmentation. In *European Conference on Computer Vision*, 109–124.

Chen, L. C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2016. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP(99):1–1.

Dollar, P., and Zitnick, C. L. 2014. Fast edge detection using structured forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(8):1558–1570.

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html.

Fu, H.; Cao, X.; Tu, Z.; and Lin, D. 2014. Symmetry constraint for foreground extraction. *IEEE Transactions on Cybernetics* 44(5):644.

Gao, Y., and Yuille, A. L. 2017. Exploiting symmetry and/or manhattan properties for 3d object structure estimation from single and multiple images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, D.; Yang, X.; Liang, C.; Zhou, Z.; Ororbi, II, A. G.; Kifer, D.; and Lee Giles, C. 2017. Multi-scale fcn with cascaded instance aware segmentation for arbitrary oriented word spotting in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hu, P., and Ramanan, D. 2017. Finding tiny faces. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jia; Yangqing; Shelhamer; Evan; Donahue; Jeff; Karayev; Sergey; Long; and Jonathan. 2014. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*.

Ke, W.; Chen, J.; Jiao, J.; Zhao, G.; and Ye, Q. 2017. Srn: Side-output residual network for object symmetry detection in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lam, L.; Lee, S. W.; and Suen, C. Y. 2002. Thinning methodologies-a comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14(9):869–885.

Lee, T. S. H.; Fidler, S.; and Dickinson, S. 2014. Detecting curved symmetric parts using a deformable disc model. In *IEEE International Conference on Computer Vision (ICCV)*, 1753–1760.

Lee, T.; Fidler, S.; and Dickinson, S. 2015. Learning to combine mid-level cues for object proposal generation. In *IEEE International Conference on Computer Vision (ICCV)*, 1680–1688.

Levinshtein, A.; Dickinson, S.; and Sminchisescu, C. 2010. Multiscale symmetric part detection and grouping. In *IEEE International Conference on Computer Vision (ICCV)*, 2162–2169.

Lindeberg, T. 1996. Edge detection and ridge detection with automatic scale selection. In *Computer Vision and Pattern Recognition(CVPR)*, 117–156.

Liu, W.; Rabinovich, A.; and Berg, A. C. 2015. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Computer Vision and Pattern Recognition(CVPR)*, 779–788.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2017. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(6):1137.

Saha, P. K.; Borgefors, G.; and Baja, G. S. D. 2016. A survey on skeletonization algorithms and their applications. *Pattern Recognition Letters* 76:3–12.

Shen, W.; Zhao, K.; Jiang, Y.; Wang, Y.; Bai, X.; and Yuille, A. 2016a. Deepskeleton: Learning multi-task scale-associated deep side outputs for object skeleton extraction in natural images. *IEEE Trans Image Process* PP(99):1–1.

Shen, W.; Bai, X.; Hu, Z.; and Zhang, Z. 2016b. Multiple instance subspace learning via partial random projection tree for local reflection symmetry in natural images. *Pattern Recognition* 52(C):306–316.

Shen, W.; Zhao, K.; Jiang, Y.; Wang, Y.; Zhang, Z.; and Bai, X. 2016c. Object skeleton extraction in natural images by fusing scale-associated deep side outputs. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sironi, A.; Lepetit, V.; and Fua, P. 2014. Multiscale centerline detection by learning a scale-space distance transform. In *Computer Vision and Pattern Recognition*, 2697–2704.

Teo, C. L.; Fermuller, C.; and Aloimonos, Y. 2016. Detection and segmentation of 2d curved reflection symmetric structures. In *IEEE International Conference on Computer Vision (ICCV)*, 1644–1652.

Tsogkas, S., and Kokkinos, I. 2012. Learning-based symmetry detection in natural images. In *ECCV 2012: 12th European Conference on Computer Vision (ECCV)*.

Wang, Y.; Zhao, X.; and Huang, K. 2017. Deep crisp boundaries. In *CVPR*.

Widynski, N.; Moevus, A.; and Mignotte, M. 2014. Local symmetry detection in natural images using a particle filtering approach. *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society* 23(12):5309.

Xie, S., and Tu, Z. 2015. Holistically-nested edge detection. In *The IEEE International Conference on Computer Vision (ICCV)*.

Xie, S., and Tu, Z. 2017. Holistically-nested edge detection. *International Journal of Computer Vision* 1–16.

Yu, F., and Koltun, V. 2015. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations(ICLR)*.

Zhang, Z.; Shen, W.; Yao, C.; and Bai, X. 2015. Symmetry-based text line detection in natural scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.