2018 13th IEEE International Conference on Automatic Face & Gesture Recognition

Enhancing Interior and Exterior Deep Facial Features for Face Detection in the Wild

Chenchen Zhu, Yutong Zheng, Khoa Luu, Marios Savvides
Department of Electrical & Computer Engineering, Carnegie Mellon University, Pittsburgh, USA

***Abstract*— Although face detection has been intensely studied for decades, it is still a challenging topic due to numerous conditions, e.g. heavy occlusions, low resolutions, extreme poses, non-face patterns that look like human faces, etc. This paper proposes a novel region-based ConvNet to address these issues. Our approach enhances the interior deep facial features and explicitly incorporates the exterior deep features. The enhanced interior features provide fine details for small faces. The exterior features capture the local information surrounding the face, supporting the detection under challenging conditions. Experiments show that our proposed components improve the baseline method significantly. Additionally, our approach consistently achieves competitive performance in four challenging databases, i.e. Wider Face, AFW, PASCAL Faces, and FDDB. We also introduce a new challenging non-face dataset[1] of 6,000 images to benchmark false positive rates for future research.**

I. Introduction

Face detection has been studied for decades to find robust solutions to detect faces in the wild [32], [44], [19], [15], [23], [16], [24], [3], [34], [21], [7]. In recent years, the use of Convolutional Neural Networks (ConvNets) has brought a huge performance improvement in face detection [17], [6], [37], [26], [35], [27], [42], [2], [20]. However, there is no current method that can match the capability of human in face detection, not to mention beating it. We argue that the difficulty mainly comes from two challenges: 1) **small faces** (Figure 1 top) and 2) **unclear faces** (Figure 1 bottom). By small it means the face consists of very few pixels, so that little information can be acquired from the face region. By unclear it means that the face lacks the explicit patterns due to occlusion, low resolution, or large pose. These two challenges are common in the wild and they often appear at the same time. One may argue that we can upscale a small face into a bigger one and do detection on top of it. But the enlarged small face becomes an unclear face due to low resolution. So naively upscaling before detection is not enough. Therefore we should address these two challenges simultaneously.

Instead of training a ConvNet that is invariant to all the challenging conditions, we argue the solution lies in the design of features which explicitly address these challenges. [41] shows that features in deep ConvNets have the hierarchical nature. Low-level layers respond to corners and other edge/color conjunctions. Middle-level layers have more complex invariances. And high-level layers show significant invariance and are more class-specific. Based on this property, we design a novel region-based ConvNet with challenge-oriented features constructed from given features in the base ConvNet.

[1]will be publicly available

Fig. 1. Detection examples of crowded small faces (top) and heavily occluded faces (bottom) in the wild using our proposed approach. Box colors indicate confidence given by the colorbar. **Best viewed in color.**

Dealing with small faces: The network should be able to extract good interior features with very fine details such as structures of eyes, mouth and nose. In ConvNets, the fine detailed features appear in the low-level layers with local receptive fields and have high resolutions. But they lack the class-specific information. On the other hand, many detection systems use coarse high-level features from layers with global receptive fields. But they are invariant to fine details. Therefore, we enhance the high-level features with

978-1-5386-2335-0/18/$31.00 ©2018 IEEE
DOI 10.1109/FG.2018.00041

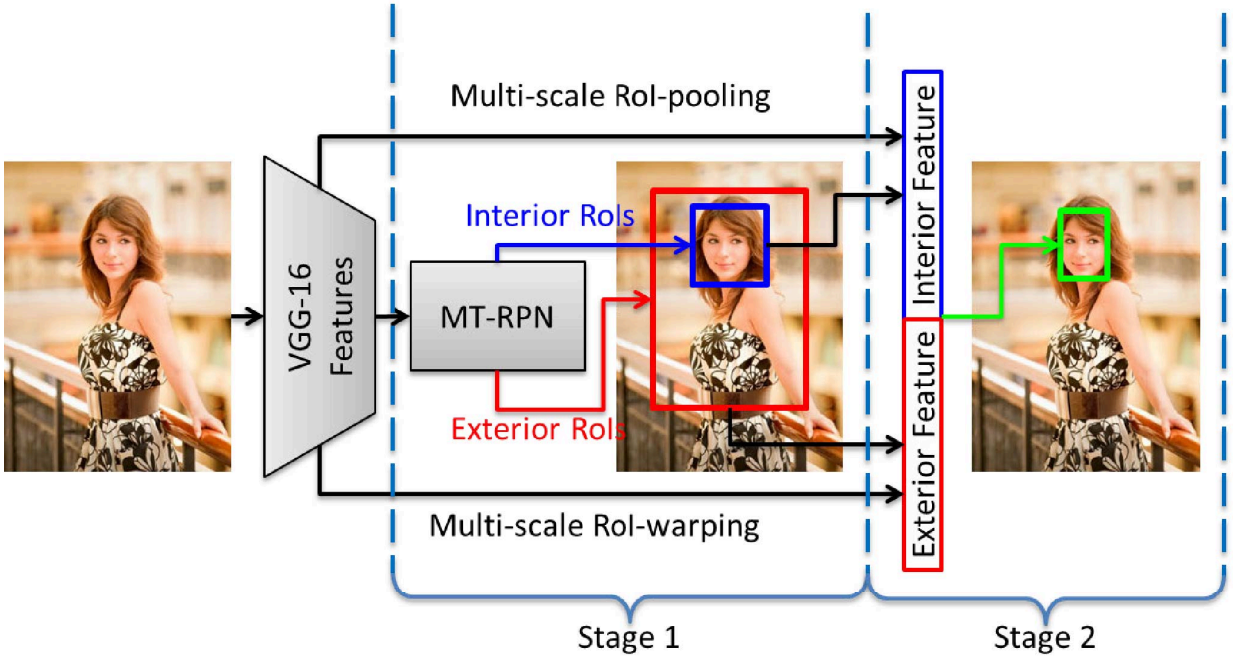


Fig. 2. The 2-stage network overview of our detection method.

fine detailed features by constructing multi-scale features.

Dealing with unclear faces: In this case, the interior features are somehow corrupted because of occlusion or blur. So we also need to look outside the face region for exterior features. This is actually the way humans detect faces. Because we have the prior knowledge of the human body, we not only look into the face region for facial components, but also outside the face for surrounding contextual cues like hair, shoulder and torso. Indeed, the contextual information helps verify the existence of challenging faces. On the other hand, it can also help to reject false positives. Based on this intuition, we let the network predict an exterior region for each interior region proposal and extract the features in exterior regions to support face detection.

Model design and evaluation: Our region-based ConvNet is illustrated in Figure 2. We first conduct ablation studies to verify the effect of proposed components. Then our best model is evaluated on four public face detection databases, the Wider Face [38], the Face Detection Dataset and Benchmark (FDDB) [12], the Annotated Faces in the Wild (AFW) [44] and the PASCAL Faces [36]. It is compared against many other recent face detection methods [24], [3], [34], [37], [17], [6], [26], [35], [38], [27], [42], [20]. Experimental results show that our approach outperforms these approaches by a considerable margin and is robust when dealing with faces under extreme conditions. We also evaluate our approach on a collected non-face dataset with around 6,000 images to benchmark the false positive rates.

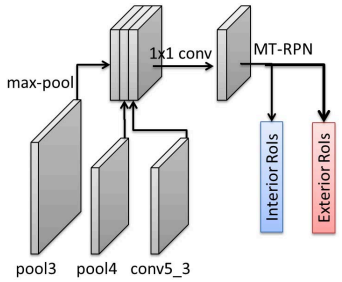
Our Contributions: **1)** We propose a novel region-based ConvNet method with enhanced interior and exterior deep facial features to find faces in the wild. **2)** We demonstrate state-of-the-art performance on numerous challenging datasets, i.e. Wider Face, AFW, PASCAL Faces and FDDB. **3)** We also introduce a new challenging non-face dataset to benchmark false positive rates in the wild.

II. Related Work

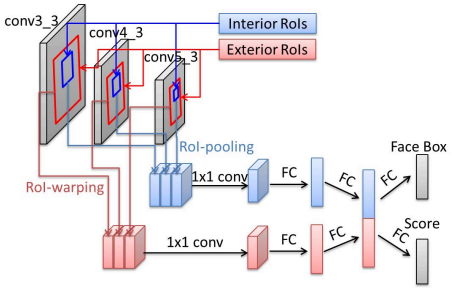
Cascade interior ConvNet features: Thanks to the development of deep ConvNets [14], [30], robust features can be extracted from large-scale raw data. Thus, the recent CNN-based face detection methods can address the wildness of environments. Some works adopt the boosting style methods with a cascade of ConvNets [17], [42]. Some studies follow Deformable Part Models approaches, splitting a face into several parts and training each part to give scores for the whole face [37], [26]. However, these methods require training multiple networks that are computational expensive.

Shared interior ConvNet features: Feature sharing is the key to alleviate computational inefficiency. For example, DDFD [6] combines sliding window approaches with fine-tuned CNN models to detect faces. Chen et al. [2] and Li et al. [20] incorporate the region proposal step into the networks. However, these methods extract features purely from the last convolution feature map, which may fail to deal with faces in different scales.

Multi-scale interior ConvNet features: For tiny faces, low-level features are more discriminative than high-level features. ScaleFace [39] splits a large range of target scales into a set of sub-ranges and addresses each sub-range with a feature map of specialized scale. HyperFace [27] follows the framework of R-CNN [9], but concatenates features from low to high levels. However, the region proposals are generated



(a) Stage 1



(b) Stage 2

Fig. 3. Visualization of interior and exterior features for each stage.

by external modules, which can be inconsistent and is the bottleneck to the performance. Also, purely using interior features makes the system not robust if faces are occluded or in poor quality.

Integrating exterior ConvNet features: Features outside the region proposal is the only clue for the system under some circumstances, e.g. a person wearing a mask. It is also helpful to use additional exterior information in ordinary cases. Inside-Outside Networks (ION) [1] apply 4-way within-layer recurrent structure to make feature maps learn global information. However, for any region, it uses global features, which may bring noise. MultiPath [40] network concatenates features from enlarged RoI regions to capture exterior information. However, the exterior region is binded to the RoI, making it agnostic to specific objects.

III. Our Approach

A. Method Overview

Our network is a generalization of Faster R-CNN [28]. The overall architecture is illustrated in Figure 2. Given an input image with faces, it first passes through a set of convolution layers to generate the convolutional feature maps. In this work we use the convolution layers fine-tuned from the VGG-16 network [30]. Then the feature maps are shared by three stages to construct features for their tasks.

Stage 1 is a Multi-Task Region Proposal Network (MT-RPN). It constructs a multi-scale feature map from the convolutional feature maps to predict a set of candidate interior regions, as well as associated exterior regions simultaneously. MT-RPN is in fully convolutional style so it can take an input image of arbitrary size. In **Stage 2** the interior and exterior features are extracted from the convolutional feature

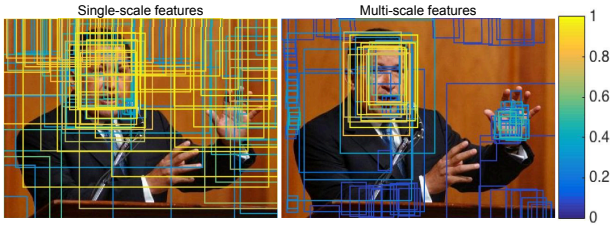


Fig. 4. Region proposals generated rely solely on the last convolution feature map **(left)** and multi-scale feature maps **(right)**. The high confidence RoIs become much more consistent by introducing features from multiple layers. Box colors indicate confidence scores given by the colorbar on the right.

maps given each interior region and exterior region. They are fused (concatenated) together to predict the final face bounding box.

We discuss about how to enhance interior features and incorporate exterior features in the following sections. The visualization of interior and exterior features are illustrated in Figure 3.

B. Enhancing Interior Features

Traditional region-based ConvNets extract interior features from a single coarse high-level feature map. We argue that this is insufficient to detect challenging faces, especially the small ones. Because small faces require extraction of very fine details but the high-level features are invariant to fine details.

In order to solve this problem, our proposed network employs both high-level and low-level features, i.e. multi-scale features. The feature maps are incorporated from lower level convolution layers with the last convolution layer for all three stages as shown in Figure 3. Therefore, both low-level features with localization capability and high-level feature with semantic information are fused together [10]. Specifically, **Stage 1** concatenates "pool3", "pool4", and "conv5_3" from the convolutional feature maps and reduce the dimension to construct the multi-scale RPN feature (Figure 3 (a)). Notice that "pool3" has twice the width and height as "pool4" and "conv5_3", it is down sampled using max-pooling before concatenation. In **Stage 2** the proposed interior regions are projected into not only "conv5_3" but also "conv3_3" and "conv4_3". Then features with fixed size are extracted from these layers using RoI-pooling [8] and are concatenated followed by dimension reduction (Figure 3 (b)). The multi-scale features can help to generate more accurate region proposals as visualized in Figure 4.

To construct the multi-scale features, naive concatenation of feature maps is infeasible since the feature maps from different layers have different properties in terms of number of channels, scale of values and norm of feature map pixels. Generally, the values in higher layers are usually much smaller than the ones in lower layers, which leads to the dominance of low-level features. This problem is addressed by performing normalization right before the concatenation (not visualized in Figure 3) using ℓ-2 normalization layer [22].

The feature maps are re-scaled such that they have learnable norms along the channel direction after normalization. The initial learnable norms are set following two rules. Firstly, the average norm for each feature map is roughly identical. Secondly, after dimension reduction the resulting feature map should have the same average norm as the original feature map in Faster R-CNN.

C. Incorporating Exterior Features

When people search for faces, we look for not only the facial patterns, e.g. eyes, nose, mouth, etc., but also surrounding information outside the faces, such as hair, shoulders, torso, etc, as the supporting information. On one hand, these exterior information helps to confirm the existence of a face. On the other hand, exterior information also helps to reject confusing false positives as shown in Figure 5. Based on this intuition, our network is designed to make explicit reference to the exterior information.

To be more specific, in **Stage 1** the MT-RPN generates both interior and exterior region candidates in a multi-task fashion from the shared mutli-scale features (Figure 3 (a)). The multi-task fashion allows two tasks benefit each other [27]. Then in **Stage 2** the exterior region candidates are projected into "conv3_3", "conv4_3" and "conv5_3" to extract feature maps from these layers using RoI-warping [5]. Similar to the extraction of interior feature, these feature maps are then normalized, concatenated and dimension-reduced to a single feature blob. After two fully connected layers, the final exterior feature vector is concatenated with the interior feature vector. They together contribute to the computation of confidence scores and bounding box regression (Figure 3 (b)).

In order to model the one-to-one spatial relation between interior and exterior regions, the exterior regions are parameterized as box offsets w.r.t. their associated interior regions. Mathematically, there are four spatial parameters t_x, t_y, t_w, and t_h defined in Equation (1).

$$\begin{aligned} t_x &= (x_e - x_i)/w_i \qquad & t_w &= \log(w_e/w_i) \\ t_y &= (y_e - y_i)/h_i \qquad & t_h &= \log(h_e/h_i) \end{aligned} \tag{1}$$

where $x_{(*)}$, $y_{(*)}$, $w_{(*)}$, and $h_{(*)}$ denote the two coordinates of the box center, width, and height respectively. e and i stand for exterior and interior respectively. The back-propagation rules are straightforward as presented in Equation (2) derived from Equation (1).

$$\begin{aligned} \frac{\partial l}{\partial t_x} &= \frac{\partial l}{\partial x_e} w_i \qquad & \frac{\partial l}{\partial t_w} &= \frac{\partial l}{\partial w_e} w_i \exp(t_w) \\ \frac{\partial l}{\partial t_y} &= \frac{\partial l}{\partial y_e} h_i \qquad & \frac{\partial l}{\partial t_h} &= \frac{\partial l}{\partial h_e} h_i \exp(t_h) \end{aligned} \tag{2}$$

where $\frac{\partial l}{\partial x_e}$, $\frac{\partial l}{\partial y_e}$, $\frac{\partial l}{\partial w_e}$, and $\frac{\partial l}{\partial h_e}$ are back-propagated from the RoI-warping layer. During training, the spatial parameters are initialized such that the interior region and exterior region roughly satisfy the spatial relation between the face and the surrounding region of a standard human body. In this work we set the initial spatial parameters as $t_x = 0.07$, $t_y = 1.53$, $t_w = 0.95$, and $t_h = 1.34$. These parameters are computed using the PASCAL-Part Dataset [4]. Specifically, we go through all the person instances with visible face and torso parts and retrieve the face boxes and body boxes which tightly enclose their part segmentation. Note that face box encloses only the face part while body box encloses both face and torso part. We ignore the arms and legs because they may be self-occluded. For each face and body box pair, parameters are computed using Equation (1). In the end all parameters are averaged across all the instances.

Fig. 5. Examples of images in our collected non-face dataset. These images contain objects similar to human faces. In these cases, exterior information is important to reject fake faces.

D. Network Training

Our network is implemented in the Caffe [13] framework. We train the whole network jointly. Here we adopt the same hyper-parameters in [28]. In Stage 1 the smooth ℓ-1 loss is applied on proposed interior bounding box regression and the softmax loss is applied on interior region scores. Since there is no ground-truth for proposed exterior bounding boxes, they are supervised by the gradients back-propagated from Stage 2. In Stage 2 the final face scores are supervised by softmax loss and face bounding box regression are again supervised using smooth ℓ-1 loss.

IV. Experiments

This section presents the experimental results in face detection[2]. We first conduct the ablative study to analyze the effectiveness of each of our proposed components. Then we evaluate the final model on common face detection benchmarks. Finally we also provide the reference time.

A. Ablation Study

To investigate the effect of each component in our network, we conduct several ablation experiments. All models are trained on the Wider Face training set and evaluated on the validation set with 3,226 images. The validation images are divided into three parts based on their detection rates on EdgeBox [45]. In other words, face images are divided into three levels according to the difficulties of the detection, i.e. Easy, Medium and Hard. The Hard level includes more challenging faces such as small faces or heavily occluded faces. The performance comparison are presented in Table I.

Baseline. Our detector is a generalization of Faster R-CNN [28], so we directly train a slightly modified version as our baseline method termed FRCNN-face with no enhanced interior features nor exterior features. Different from [28],

[2]More qualitative results can be seen in the supplementary materials

Fig. 6. Visualization of exterior regions (magenta) predicted by the network. Green boxes are the detection of faces.

TABLE I

ABLATION STUDIES ON THE EASY, MEDIUM, AND HARD LEVEL OF **WIDER FACE VALIDATION SET**. NUMBERS ARE THE AVERAGE PRECISION SCORES. IF: ENHANCED INTERIOR FEATURES; EF: EXTERIOR FEATURES; MST: MULTI-SCALE TESTING.

Methods	Easy	Medium	Hard
FRCNN-face	84.3%	73.1%	40.9%
Ours(IF)	88.7%	86.1%	61.2%
Ours(IF+EF)	**90.6%**	**88.3%**	**66.1%**
HR-VGG16 [11]	86.2%	84.4%	74.9%
Ours(IF+EF+MST)	**91.6%**	**89.8%**	**79.2%**

we only use square anchors. All other hyper-parameters are kept the same. It can give decent performance on Easy and Medium cases, but unsatisfying results on Hard level (shown as the first row of Table I).

The effect of enhanced interior features. Then we add enhanced interior features (IF). This is implemented by adding multi-scale features in Stage 1 and 2 as mentioned in Section III-B, which improves the performance on all three levels (shown as the second row of Table I). Especially, we observe about 20% absolute improvement on Hard level. This suggests that enhanced interior features help finding more challenging faces.

The effect of exterior features. Next we incorporate the exterior features (EF). This is implemented by predicting additional exterior region proposals in Stage 1 and fusing the interior and exterior features in Stage 2. We observe another improvement on all three levels, especially the Hard case (shown as the third row of Table I). This is a clear evidence that explicit reference to exterior features can support detection of challenging faces. To better understand how exterior features help improve the performance, we visualize all the interior and exterior region pairs, i.e. the face and body box pairs, predicted by the network. Some examples of challenging faces and their corresponding body regions are illustrated in Fig. 6. It shows the exterior regions roughly focus on the face and torso parts and adapt across instances. This means the features from the exterior regions (torso part) provide visual cues for finding hard faces.

The effect of multi-scale testing. Inspired by [11], we apply the multi-scale testing technique (MST) to our method and compare with the VGG version of [11] termed as HR-VGG16. We rescale each image to 0.5X, 1.0X and 2.0X and run our detector. The detection results are gathered from three images and combined together. We use non-maximum suppression with a threshold of 0.3 to remove redundant detections. This gives another performance boost, outperforming [11] on all three cases. This indicates that our method can find more faces on various scale level.

B. Benchmark Evaluation

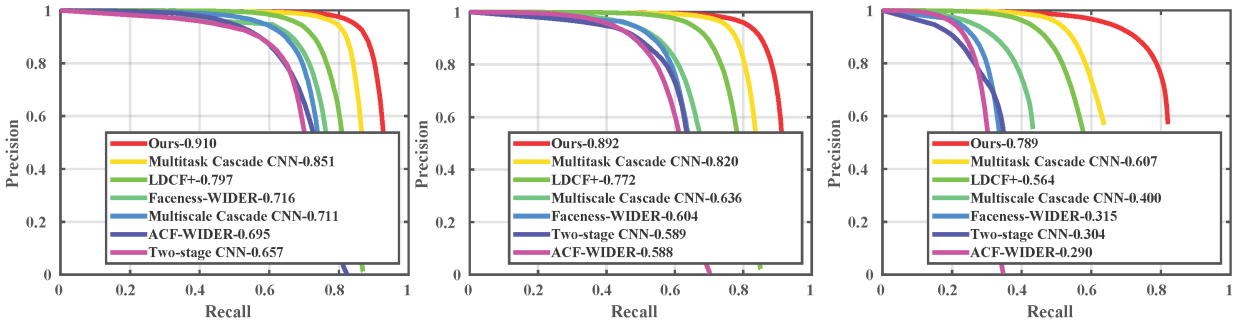
We run our model with IF, EF and MST on the Wider Face dataset [38]. Under this database, our approach robustly outperforms strong baseline methods, including Two-stage CNN [38], Multiscale Cascade CNN [38], Faceness [37] and Aggregate Channel Features (ACF) [34], Multitask Cascade CNN [42] by a considerable margin.

We also show that our model trained on the Wider Face dataset generalizes well to other standard face detection datasets including the AFW [44], the PASCAL faces [36], and the FDDB [12]. Our network *without MST* alone is able to consistently achieve state-of-the-art results against other popular face detection methods, including MTCNN [42], Conv3D [20], HyperFace [27], DP2MFD [26], CCF [35], Faceness [37], NPDFace [21], MultiresHPM [7], DDFD [6], CascadeCNN [17], ACF [34], Pico [23], HeadHunter [24], Joint Cascade [3], Boosted Exemplar [16], and PEP-Adapt [15], Face++ [43], SURF Cascade multiview [18], XZJY [29].

Additionally, we collect a new challenging Non-Face dataset with 6,000 images containing objects that look like human faces but are not faces in the wild, which will be released. It is used for evaluating the detector's false positive rates under challenging conditions. A good detector should not only find as many faces as possible, but also reject non faces even they look like faces.

Wider Face dataset. Wider Face is a public face detection benchmark dataset [38] released recently. It contains 393,703 labeled human faces from 32,203 images collected based on 61 event classes from Internet. The database has many human faces with a high degree of pose variation, large occlusions, low-resolutions and strong lighting conditions. The images in this database are organized and split into three subsets, i.e. training, validation and testing. Each contains 40%, 10% and 50% respectively of the original databases. The images and the ground-truth labels of the training and the validation sets are available online for experiments. In the testing set, only the testing images are available online. All detection results are sent to the database server for evaluating and receiving the Precision-Recall curves. In our face detection experiments, our model is trained on the training set of the Wider Face dataset containing 159,424 annotated faces collected in 12,880 images. The trained model on this database are used in testing of all databases *without further fine-tuning*.

Testing and Comparison. We run our detector with IF, EF and MST on the Wider Face testing set. Our proposed method is compared against all published methods , i.e.



(a) Easy level (b) Medium level (c) Hard level

Fig. 7. Precision-Recall curves obtained by our proposed method (red) and the other published strong baselines on the **Wider Face testing set**. Numbers show the average precision scores.

Fig. 8. Examples of the top 20 false positives from our model tested on the Wider Face validation set. In fact these false positives include many human faces not in the dataset due to mislabeling, which means that our method is robust to the noise in the data.

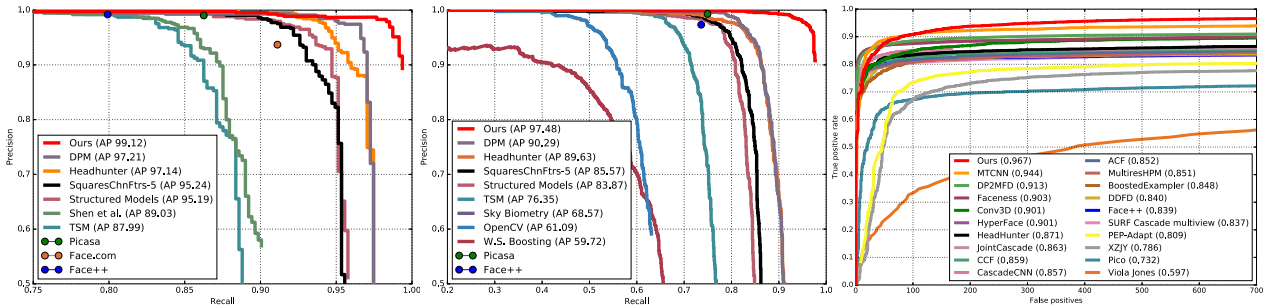
Multitask Cascade CNN [42], Two-stage CNN [38], Multiscale Cascade CNN [38], Faceness [37], and Aggregate Channel Features (ACF) [34]. All these methods are trained and tested following the same evaluation protocols. We don't compare with [11] because it doesn't report the performance of VGG version detector on the testing set. The Precision-Recall curves and average precision scores are shown in Figure 7. Our method outperforms those strong baselines by a considerable margin. It achieves the best average precision in all level faces, and outperforms the second best baseline by 7.06% (Easy), 8.78% (Medium) and 29.98% (Hard). These results suggest that as the difficulty level goes up, our model can detect challenging faces better. So it has the ability to handle difficult conditions hence is more closed to human detection performance.

Visualization of False Positives. As it is well known that precision-recall curves degrade due to the false positives, we are interested in the false positives produced by our model. We are curious about what object can fool our model to treat it as a face. Is it due to over-fitting, data bias, or miss labeling? In order to visualize the false positives, we test our model on the Wider Face validation set and pick all the false positives according to the ground truth. Then those positives are sorted by the confidence score in a descending order. We choose the top 20 false positives as illustrated in Figure 8. Because their confidence scores are high, they are the objects most likely to cause our model making mistakes. It turns out that most of the false positives are actually human faces caused by miss labeling in the dataset. For other false positives, we find the errors made by our model are rather reasonable. They all have the pattern of human face as well as the shape of human body.

AFW, PASCAL Faces and FDDB datasets. To show that our method generalizes well to other databases, the proposed method is also benchmarked on three challenging face detection datasets, i.e. AFW [44], PASCAL Faces [36] and FDDB database [12]. We choose the model with IF and EF only. *Without* MST, it can already achieve the state-of-the-art performance. The performance curves are generated using the tool from [24].

On the PASCAL Faces and AFW datasets we compare with DPM [24], HeadHunter [24], SquaresChnFtrs-5 [24], Structured Models [33], Shen et al. [29], TSM [44], Picasa, Face++ [43].

On FDDB dataset we compare with several published methods including MTCNN [42], Conv3D [20], HyperFace [27], DP2MFD [26], CCF [35], Faceness [37], NPDFace [21], MultiresHPM [7], DDFD [6], CascadeCNN [17], ACF [34], Pico [23], HeadHunter [24], Joint Cascade [3], Boosted Exemplar [16], and PEP-Adapt [15], Face++ [43], SURF



(a) AFW (b) PASCAL Faces (c) FDDB

Fig. 9. Comparison between our method (red) and popular state-of-the-art methods on the AFW, PASCAL Faces, and FDDB datasets.

(a) AFW (b) PASCAL Faces (c) FDDB

Fig. 10. Some examples of face detection results using our approach on the AFW, PASCAL Faces, and FDDB databases.

Cascade multiview [18], XZJY [29].

The proposed method outperforms all previous methods as shown in Figure 9. This is a concrete evidence to demonstrate that our method robustly detects unconstrained faces. Figure 10 shows some examples of the face detection results on three datasets.

Non-Face dataset. This experiment is designed to test the tolerance of our system to face-like non-face images. Though we have different databases for testing the precision and recall of the system, the testing images usually contain faces, which renders the current face databases samples from a particular distribution. To test whether our system is over-fitting to that particular distribution and whether it is generally robust to other kind of images, we establish a new non-face database. The 6,000 images are selected from the Internet and contain face-like objects in different circumstances (Figure 5). The background are generally in the wild. The result of our system is shown in Table II, along with some baselines. Each face detector is first benchmarked by setting a reasonable threshold based on its precision and recall on the AFW dataset [44]. Then we record the number of false alarms on the non-face dataset, the precision and recall on AFW dataset under the benchmarked setting. Our method triggers the fewest false alarms while achieving the highest precision and recall on AFW dataset.

C. Inference Time

During inference, our method is running on a single Titan X GPU machine with Intel Core i7-6700 CPU @ 3.40GHz in batch size of 1. It takes 0.168s per frame running on WiderFace or AFW (XGA quality) and 0.045s per frame on FDDB or PASCAL Faces (VGA quality).

TABLE II

THE NUMBER OF FALSE ALARMS ON OUR NON-FACE DATASET UNDER REASONABLE PRECISION AND RECALL ON THE AFW [44] DATASET.

Methods	#False-alarms	Precision	Recall
Pittpatt [25]	2353	0.76	0.92
Viola-Jones [32]	117	0.96	0.45
Joint Cascade [3]	325	0.89	0.76
Deep Cascade [31]	231	0.97	0.81
HR-ResNet101 [11]	48	0.98	0.99
Ours	**10**	**0.98**	**0.98**

V. CONCLUSIONS

This paper has presented our proposed region-based ConvNet for unconstrained face detection. The superior performance on four face detection datasets shows its capability to extract robust facial feature representation which is invariant to various challenging conditions in the real-world scenario. In addition we also introduced a Non-Face database for testing the system's tolerance to images containing face-like but non-face objects, which will be published.

REFERENCES

[1] S. Bell, C. Lawrence Zitnick, K. Bala, and R. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2874–2883, 2016.

[2] D. Chen, G. Hua, F. Wen, and J. Sun. Supervised transformer network for efficient face detection. In *European Conference on Computer Vision*, pages 122–138. Springer, 2016.

[3] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In *European Conference on Computer Vision*, pages 109–122. Springer, 2014.

[4] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[5] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3150–3158, 2016.

[6] S. S. Farfade, M. J. Saberian, and L.-J. Li. Multi-view face detection using deep convolutional neural networks. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 643–650. ACM, 2015.

[7] G. Ghiasi and C. C. Fowlkes. Occlusion coherence: Detecting and localizing occluded faces. *arXiv preprint arXiv:1506.08347*, 2015.

[8] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 38(1):142–158, 2016.

[10] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 447–456, 2015.

[11] P. Hu and D. Ramanan. Finding tiny faces. *arXiv preprint arXiv:1612.04402*, 2016.

[12] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[15] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic part model for unsupervised face detector adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 793–800, 2013.

[16] H. Li, Z. Lin, J. Brandt, X. Shen, and G. Hua. Efficient boosted exemplar-based face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1843–1850, 2014.

[17] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5325–5334, 2015.

[18] J. Li, T. Wang, and Y. Zhang. Face detection using surf cascade. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2183–2190. IEEE, 2011.

[19] J. Li and Y. Zhang. Learning surf cascade for fast and accurate object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3468–3475, 2013.

[20] Y. Li, B. Sun, T. Wu, and Y. Wang. Face detection with end-to-end integration of a convnet and a 3d model. In *European Conference on Computer Vision*, pages 420–436. Springer, 2016.

[21] S. Liao, A. K. Jain, and S. Z. Li. A fast and accurate unconstrained face detector. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):211–223, 2016.

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

[23] N. Markuš, M. Frljak, I. S. Pandzic, J. Ahlberg, and R. Forchheimer. A method for object detection based on pixel intensity comparisons. In *2nd Croatian Computer Vision Workshop*, 2013.

[24] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *Computer Vision–ECCV 2014*, pages 720–735. Springer, 2014.

[25] M. C. Nechyba, L. Brandy, and H. Schneiderman. Pittpatt face detection and tracking for the clear 2007 evaluation. In *Multimodal Technologies for Perception of Humans*, pages 126–137. Springer, 2008.

[26] R. Ranjan, V. M. Patel, and R. Chellappa. A deep pyramid deformable part model for face detection. In *Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on*, pages 1–8. IEEE, 2015.

[27] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *arXiv preprint arXiv:1603.01249*, 2016.

[28] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[29] X. Shen, Z. Lin, J. Brandt, and Y. Wu. Detecting and aligning faces by image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3460–3467, 2013.

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

[31] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483, 2013.

[32] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.

[33] J. Yan, X. Zhang, Z. Lei, and S. Z. Li. Face detection by structural models. *Image and Vision Computing*, 32(10):790–799, 2014.

[34] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Aggregate channel features for multi-view face detection. In *Biometrics (IJCB), 2014 IEEE International Joint Conference on*, pages 1–8. IEEE, 2014.

[35] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Convolutional channel features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 82–90, 2015.

[36] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Fine-grained evaluation on face detection in the wild. In *Automatic Face and Gesture Recognition (FG), 11th IEEE International Conference on*. IEEE, 2015.

[37] S. Yang, P. Luo, C.-C. Loy, and X. Tang. From facial parts responses to face detection: A deep learning approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3676–3684, 2015.

[38] S. Yang, P. Luo, C. C. Loy, and X. Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[39] S. Yang, Y. Xiong, C. C. Loy, and X. Tang. Face detection through scale-friendly deep convolutional networks. *arXiv preprint arXiv:1706.02863*, 2017.

[40] S. Zagoruyko, A. Lerer, T.-Y. Lin, P. O. Pinheiro, S. Gross, S. Chintala, and P. Dollár. A multipath network for object detection. *arXiv preprint arXiv:1604.02135*, 2016.

[41] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer vision–ECCV 2014*, pages 818–833. Springer, 2014.

[42] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.

[43] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 386–391, 2013.

[44] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.

[45] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, pages 391–405. Springer, 2014.