WIDER FACE: A Face Detection Benchmark

Shuo Yang Ping Luo Chen Change Loy Xiaoou Tang
Department of Information Engineering, The Chinese University of Hong Kong
{ys014, pluo, ccloy, xtang}@ie.cuhk,edu.hk

1 检测数据集wider face，对比现有benchmark数据集，在数据量、标定类型、场景多样性、人脸尺度、遮挡、姿态方面都提出了很高的挑战
2 提出wider face上sota人脸检测算法Multiscale Cascade CNN，对比四个benchmark算法，性能优势明显；
3 提出将wider face训练集用于模型训练，验证集测试集用于评估，发现性能提升明显（如5.2小节），以后也成为各类算法评估的一个标准：在wider face训练集上做训练，验证集测试集上做评估；

Abstract

Face detection is one of the most studied topics in the computer vision community. Much of the progresses have been made by the availability of face detection benchmark datasets. We show that there is a gap between current face detection performance and the real world requirements. To facilitate future face detection research, we introduce the WIDER FACE dataset, which is 10 *times larger than existing datasets. The dataset contains rich annotations, including occlusions, poses, event categories, and face bounding boxes. Faces in the proposed dataset are extremely challenging due to large variations in scale, pose and occlusion, as shown in Fig. 1. Furthermore, we show that WIDER FACE dataset is an effective training source for face detection. We benchmark several representative detection systems, providing an overview of state-of-the-art performance and propose a solution to deal with large scale variation. Finally, we discuss common failure cases that worth to be further investigated. Dataset can be downloaded at:* mmlab.ie.cuhk.edu.hk/projects/WIDERFace

1.当前人脸检测方法的性能与真实是世界的需要存在较大差距，本文提出更大的数据集，以期缩小这种差距

2.数据集包含丰富的标签，包括遮挡、尺度、时间分类、面部bbox等，远超当前数据集

3. wider face训练集用于模型训练，验证集测试集用于评估，性能提升明显

4. 复现了四种人脸检测算法，作为wider face的benchmark，同时提出Multiscale Cascade CNN，在wider face上是sota，能处理人脸的大尺度变化问题；

1. Introduction

Face detection is a critical step to all facial analysis algorithms, including face alignment [38, 24, 30], face recognition [27], face verification [26, 25], and face parsing [20]. Given an arbitrary image, the goal of face detection is to determine whether or not there are any faces in the image and, if present, return the image location and extent of each face [35]. While this appears as an effortless task for human, it is a very difficult task for computers. The challenges associated with face detection can be attributed to variations in pose, scale, facial expression, occlusion, and lighting condition, as shown in Fig. 1. Face detection has made significant progress after the seminal work by Viola and Jones [29]. Modern face detectors can easily detect near frontal faces and are widely used in real world applications, such as digital camera and electronic photo album. Recent research [3, 17, 21, 33, 36] in this area focuses on the unconstrained scenario, where a number of intricate factors

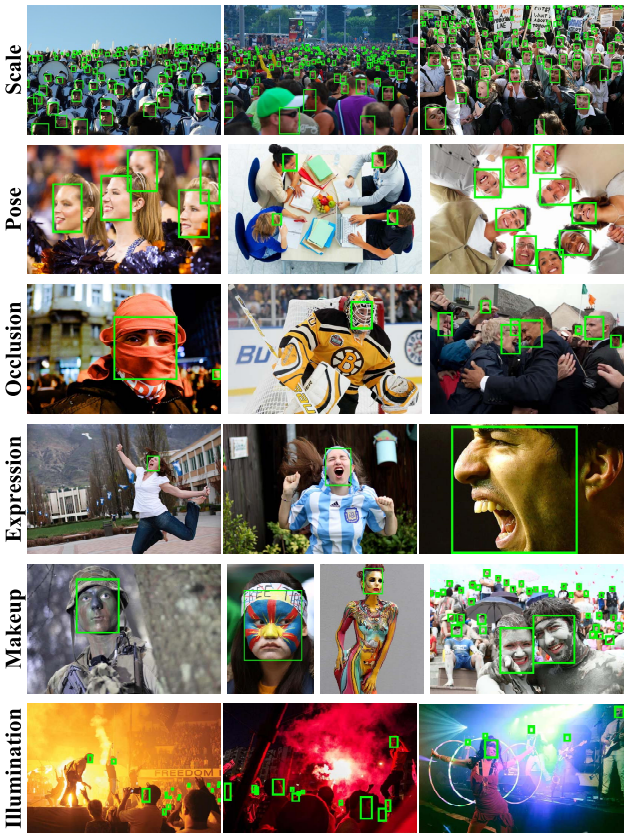


Figure 1. We propose a WIDER FACE dataset for face detection, which has a high degree of variability in scale, pose, occlusion, expression, appearance and illumination. We show example images (cropped) and annotations. The annotated face bounding box is denoted in green color. The WIDER FACE dataset consists of 393, 703 labeled face bounding boxes in 32, 203 images **(Best view in color).**

such as extreme pose, exaggerated expressions, and large portion of occlusion can lead to large visual variations in face appearance.

Publicly available benchmarks such as FDDB [13], AFW [39], PASCAL FACE [32], have contributed to spurring interest and progress in face detection research. However, as algorithm performance improves, more chal-

lenging datasets are needed to trigger progress and to inspire novel ideas. Current face detection datasets typically contain a few thousand faces, with limited variations in pose, scale, facial expression, occlusion, and background clutters, making it difficult to assess for real world performance. As we will demonstrate, the limitations of datasets have partially contributed to the failure of some algorithms in coping with heavy occlusion, small scale, and atypical pose.

In this work, we make three contributions. (1) We introduce a large-scale face detection dataset called WIDER FACE. It consists of $32,203$ images with $393,703$ labeled faces, which is 10 times larger than the current largest face detection dataset [14]. The faces vary largely in appearance, pose, and scale, as shown in Fig. 1. In order to quantify different types of errors, we annotate multiple attributes: occlusion, pose, and event categories, which allows in depth analysis of existing algorithms. (2) We show an example of using WIDER FACE through proposing a multi-scale two-stage cascade framework, which uses divide and conquer strategy to deal with large scale variations. Within this framework, a set of convolutional networks with various size of input are trained to deal with faces with a specific range of scale. (3) We benchmark four representative algorithms [21, 29, 33, 36], either obtained directly from the original authors or reimplemented using open-source codes. We evaluate these algorithms on different settings and analyze conditions in which existing methods fail.

2. Related Work

Brief review of recent face detection methods: Face detection has been studied for decades in the computer vision literature. Modern face detection algorithms can be categorized into four categories: cascade based methods [3, 12, 18, 19, 29], part based methods [23, 32, 39], channel feature based methods [2, 33], and neural network based methods [7, 17, 36]. Here we highlight a few notable studies. A detailed survey can be found in [35, 37]. The seminal work by Viola and Jones [29] introduces integral image to compute Haar-like features in constant time. These features are then used to learn AdaBoost classifier with cascade structure for face detection. Various later studies follow a similar pipeline. Among those variants, SURF cascade [18] achieves competitive performance. Chen *et al*. [3] learns face detection and alignment jointly in the same cascade framework and obtains promising detection performance.

One of the well-known part based methods is deformable part models (DPM) [8]. Deformable part models define face as a collection of parts and model the connections of parts through Latent Support Vector Machine. The part based methods are more robust to occlusion compared with cascade-based methods. A recent study [21] demonstrates state-of-the art performance with just a vanilla DPM, achieving better results than more sophisticated DPM variants [32, 39]. Aggregated channel feature (ACF) is first proposed by Dollar *et al*. [4] to solve pedestrian detection. Later on, Yang *et al*. [33] applied this idea on face detection. In particular, features such as gradient histogram, integral histogram, and color channels are combined and used to learn boosting classifier with cascade structure. Recent studies [17, 36] show that face detection can be further improved by using deep learning, leveraging the high capacity of deep convolutional networks. We anticipate that the new WIDER FACE data can benefit deep convolutional network that typically requires large amount of data for training.

Table 1. Comparison of face detection datasets.

	Training		Testing		Height			Properties		
Dataset	#Image	#Face	#Image	#Face	10-50 pixels	50-300 pixels	≤300 pixels	Occlusion labels	Event labels	Pose labels
AFW [39]	-	-	0.2k	0.47k	12%	70%	18%	-	-	✓
FDDB [13]	-	-	2.8k	5.1k	8%	86%	6%	-	-	-
PASCAL FACE [32]	-	-	0.85k	1.3k	41%	57%	2%	-	-	-
IJB-A [14]	16k	33k	8.3k	17k	13%	69%	18%	-	-	-
MALF [34]	-	-	5.25k	11.9k	N/A	N/A	N/A	✓	-	✓
WIDER FACE	16k	**199k**	16k	**194k**	50%	43%	7%	✓	✓	✓

Existing datasets: We summarize some of the well-known face detection datasets in Table 1. AFW [39], FDDB [13], and PASCAL FACE [32] datasets are most widely used in face detection. AFW dataset is built using Flickr images. It has 205 images with 473 labeled faces. For each face, annotations include a rectangular bounding box, 6 landmarks and the pose angles. FDDB dataset contains the annotations for $5,171$ faces in a set of $2,845$ images. PASCAL FACE consists of 851 images and $1,341$ annotated faces. Recently, IJB-A [14] is proposed for face detection and face recognition. IJB-A contains $24,327$ images and $49,759$ faces. MALF is the first face detection dataset that supports fine-gained evaluation. MALF [34] consists of $5,250$ images and $11,931$ faces. The FDDB dataset has helped driving recent advances in face detection. However, it is collected from the Yahoo! news website which biases toward celebrity faces. The AFW and PASCAL FACE datasets contain only a few hundred images and has limited variations in face appearance and background clutters. The IJB-A dataset has large quantity of labeled data; however, occlusion and pose are not annotated. The MAFL dataset labels fine-grained face attributes such as occlusion, pose and expression. The number of images and faces are relatively small. Due to the limited variations in existing datasets, the performance of recent face detection algorithms saturates on current face detection benchmarks. For instance, on AFW, the best performance is 97.2% AP; on FDDB, the highest recall is 91.74%; on PASCAL FACE, the best result is 92.11% AP. The best few algorithms have only marginal difference.

3. WIDER FACE Dataset

3.1. Overview

To our knowledge, WIDER FACE dataset is currently the largest face detection dataset, of which images are selected from the publicly available WIDER dataset [31]. We choose 32, 203 images and label 393, 703 faces with a high degree of variability in scale, pose and occlusion as depicted in Fig. 1. WIDER FACE dataset is organized based on 60 event classes. For each event class, we randomly select 40%/10%/50% data as training, validation and testing sets. Here, we specify two training/testing scenarios:

- **Scenario-Ext:** A face detector is trained using any external data, and tested on the WIDER FACE test partition.
- **Scenario-Int:** A face detector is trained using WIDER FACE training/validation partitions, and tested on WIDER FACE test partition.

We adopt the same evaluation metric employed in the PASCAL VOC dataset [6]. Similar to MALF [34] and Caltech [5] datasets, we do not release bounding box ground truth for the test images. Users are required to submit final prediction files, which we shall proceed to evaluate.

3.2. Data Collection

Collection methodology. WIDER FACE dataset is a subset of the WIDER dataset [31]. The images in WIDER were collected in the following three steps: 1) Event categories were defined and chosen following the Large Scale Ontology for Multimedia (LSCOM) [22], which provides around 1, 000 concepts relevant to video event analysis. 2) Images are retrieved using search engines like *Google* and *Bing*. For each category, 1, 000-3, 000 images were collected. 3) The data were cleaned by manually examining all the images and filtering out images without human face. Then, similar images in each event category were removed to ensure large diversity in face appearance. A total of 32, 203 images are eventually included in the WIDER FACE dataset.

Annotation policy. We label the bounding boxes for all the recognizable faces in the WIDER FACE dataset. The bounding box is required to tightly contain the forehead, chin, and cheek, as shown in Fig. 2. If a face is occluded, we still label it with a bounding box but with an estimation on the scale of occlusion. Similar to the PASCAL VOC dataset [6], we assign an 'Ignore' flag to the face which is very difficult to be recognized due to low resolution and small scale (10 pixels or less). After annotating the face bounding boxes, we further annotate the following attributes: pose (typical, atypical) and occlusion level (partial, heavy). Each annotation is labeled by one annotator and cross-checked by two different people.

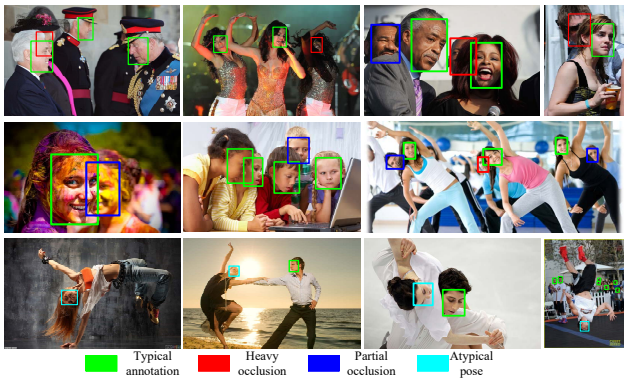


Figure 2. Examples of annotation in WIDER FACE dataset **(Best view in color).**

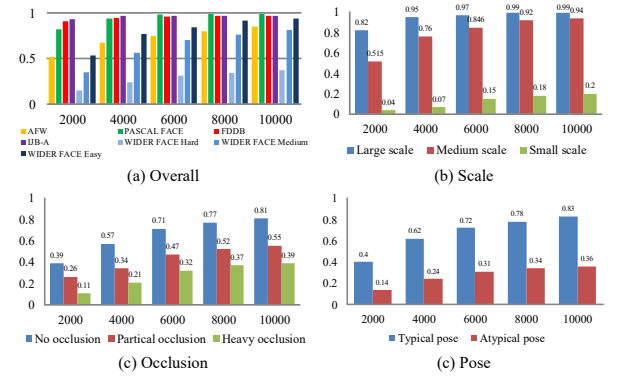


Figure 3. The detection rate with different number of proposals. The proposals are generated by using Edgebox [40]. Y-axis denotes for detection rate. X-axis denotes for average number of proposals per image. Lower detection rate implies higher difficulty. We show histograms of detection rate over the number of proposal for different settings (a) Different face detection datasets. (b) Face scale level. (c) Occlusion level. (d) Pose level.

3.3. Properties of WIDER FACE

WIDER FACE dataset is challenging due to large variations in scale, occlusion, pose, and background clutters. These factors are essential to establishing the requirements for a real world system. To quantify these properties, we use generic object proposal approaches [1, 28, 40], which are specially designed to discover potential objects in an image (face can be treated as an object). Through measuring the number of proposals vs. their detection rate of faces, we can have a preliminary assessment on the difficulty of a dataset and potential detection performance. In the following assessments, we adopt EdgeBox [40] as object proposal, which has good performance in both accuracy and efficiency as evaluated in [11].

Overall. Fig. 3(a) shows that WIDER FACE has much lower detection rate compared with other face detection datasets. The results suggest that WIDER FACE is a more challenging face detection benchmark compared to exist-

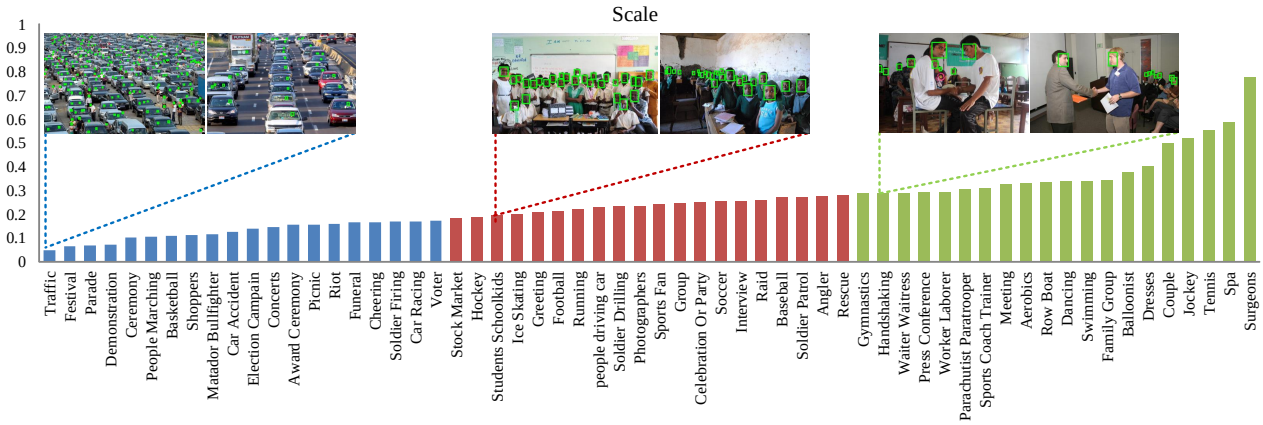


Figure 4. Histogram of detection rate for different event categories. Event categories are ranked in an ascending order based on the detection rate when the number of proposal is fixed at $10,000$. Top $1-20$, $21-40$, $41-60$ event categories are denoted in blue, red, and green, respectively. Example images for specific event classes are shown. Y-axis denotes for detection rate. X-axis denotes for event class name.

ing datasets. Following the principles in KITTI [9] and MALF [34] datasets, we define three levels of difficulty: 'Easy', 'Medium', 'Hard' based on the detection rate of EdgeBox [40], as shown in the Fig. 3(a). The average recall rates for these three levels are 92%, 76%, and 34%, respectively, with $8,000$ proposal per image.

Scale. We group the faces by their image size (height in pixels) into three scales: small (between 10-50 pixels), medium (between 50-300 pixels), large (over 300 pixels). We make this division by considering the detection rate of generic object proposal and human performance. As can be observed from Fig 3(b), the large and medium scales achieve high detection rate (more than 90%) with $8,000$ proposals per image. For the small scale, the detection rates consistently stay below 30% even we increase the proposal number to $10,000$.

Occlusion. Occlusion is an important factor for evaluating the face detection performance. Similar to a recent study [34], we treat occlusion as an attribute and assign faces into three categories: no occlusion, partial occlusion, and heavy occlusion. Specifically, we ask annotator to measure the fraction of occlusion region for each face. A face is defined as 'partially occluded' if 1%-30% of the total face area is occluded. A face with occluded area over 30% is labeled as 'heavily occluded'. Fig. 2 shows some examples of partial/heavy occlusions. Fig. 3(c) shows that the detection rate decreases as occlusion level increases. The detection rates of faces with partial or heavy occlusions are below 50% with $8,000$ proposals.

Pose. Similar to occlusion, we define two pose deformation levels, namely typical and atypical. Fig. 2 shows some faces of typical and atypical pose. Face is annotated as atypical under two conditions: either the roll or pitch degree is larger than 30-degree; or the yaw is larger than 90-degree. Fig. 3(d) suggests that faces with atypical poses are much harder to be detected.

Event. Different events are typically associated with different scenes. WIDER FACE contains 60 event categories covering a large number of scenes in the real world, as shown in Fig. 1 and Fig. 2. To evaluate the influence of event to face detection, we characterize each event with three factors: scale, occlusion, and pose. For each factor we compute the detection rate for the specific event class and then rank the detection rate in an ascending order. Based on the rank, events are divided into three partitions: easy (41-60 classes), medium (21-40 classes) and hard (1-20 classes). We show the partitions based on scale in Fig. 9. Partitions based on occlusion and pose are included in the **appendix**.

Effective training source. As shown in the Table 1, existing datasets such as FDDB, AFW, and PASCAL FACE do not provide training data. Face detection algorithms tested on these datasets are frequently trained with ALFW [15], which is designed for face landmark localization. However, there are two problems. First, ALFW omits annotations of many faces with a small scale, low resolution, and heavy occlusion. Second, the background in ALFW dataset is relatively clean. As a result, many face detection approaches resort to generate negative samples from other datasets such as PASCAL VOC dataset. In contrast, all recognizable faces are labeled in the WIDER FACE dataset. Because of its event-driven nature, WIDER FACE dataset has a large number of scenes with diverse background, making it possible as a good training source with both positive and negative samples. We demonstrate the effectiveness of WIDER FACE as a training source in Sec. 5.2.

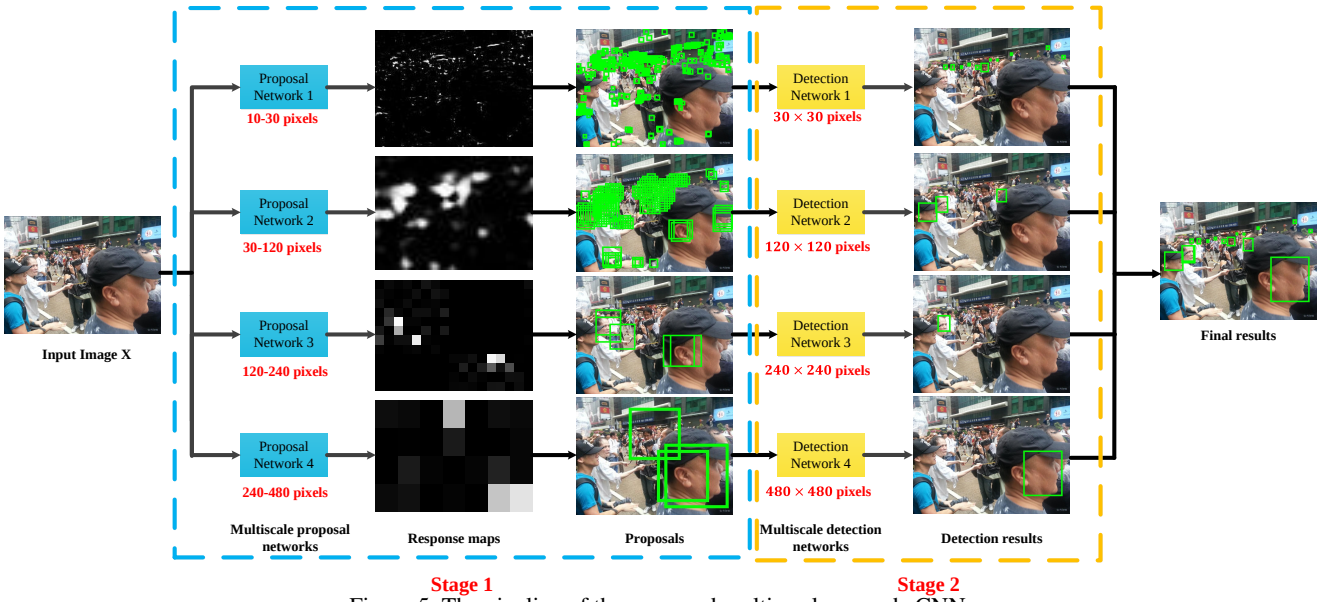


Figure 5. The pipeline of the proposed multi-scale cascade CNN.

4. Multi-scale Detection Cascade

We wish to establish a solid baseline for WIDER FACE dataset. As we have shown in Table 1, WIDER FACE contains faces with a large range of scales. Fig. 3(b) further shows that faces with a height between 10-50 pixels only achieve a proposal detection rate of below 30%. In order to deal with the high degree of variability in scale, we propose a multi-scale two-stage cascade framework and employ a divide and conquer strategy. Specifically, we train a set of face detectors, each of which only deals with faces in a relatively small range of scales. Each face detector consists of two stages. The first stage generates multi-scale proposals from a fully-convolutional network. The second stage is a multi-task convolutional network that generates face and non-face prediction of the candidate windows obtained from the first stage, and simultaneously predicts for face location. The pipeline is shown in Fig. 5. The two main steps are explained as follow.

Multi-scale proposal. In this step, we joint train a set of fully convolutional networks for face classification and scale classification. We first group faces into four categories by their image size, as shown in the Table 4 (each row in the table represents a category). For each group, we further divide it into three subclasses. Each network is trained with image patches with the size of their upper bound scale. For example, Network 1 and Network 2 are trained with 30×30 and 120×120 image patches, respectively. We align a face at the center of an image patch as positive sample and assign a scale class label based on the predefined scale subclasses in each group. For negative samples, we randomly cropped patches from the training images. The patches should have an intersection-over-union (IoU) of smaller than 0.5 with any of the positive samples. We assign a value -1 as the scale class for negative samples, which will have no contribution to the gradient during training.

Table 2. Summary of face scale for multi-scale proposal networks.

Scale	Class 1	Class 2	Class 3
Network 1	10-15	15-20	20-30
Network 2	30-50	50-80	80-120
Network 3	120-160	160-200	200-240
Network 4	240-320	320-400	400-480

We take Network 2 as an example. Let $\{\mathbf{x}_i\}_{i=1}^N$ be a set of image patches where $\forall\mathbf{x}_i \in \mathbb{R}^{120\times120}$. Similarly, let $\{\mathbf{y}_i^f\}_{i=1}^N$ be the set of face class labels and $\{\mathbf{y}_i^s\}_{i=1}^N$ be the set of scale class label, where $\forall\mathbf{y}_i^f \in \mathbb{R}^{1\times1}$ and $\forall\mathbf{y}_i^s \in \mathbb{R}^{1\times3}$. Learning is formulated as a multi-variate classification problem by minimizing the cross-entropy loss. $L = \sum_{i=1}^N \mathbf{y}_i \log p(\mathbf{y}_i = 1|\mathbf{x}_i) + (\mathbf{1} - \mathbf{y}_i) \log \big(\mathbf{1} - p(\mathbf{y}_i = 1|\mathbf{x}_i)\big)$, where $p(\mathbf{y}_i|\mathbf{x}_i)$ is modeled as a sigmoid function, indicating the probability of the presence of a face. This loss function can be optimized by the stochastic gradient descent with back-propagation.

Face detection. The prediction of proposed windows from the previous stage is refined in this stage. For each scale category, we refine these proposals by joint training face classification and bounding box regression using the same CNN structure in the previous stage with the same input size. For face classification, a proposed window is assigned with a positive label if the IoU between it and the ground truth bounding box is larger than 0.5; otherwise it is negative. For bounding box regression, each proposal is predicted a position of its nearest ground truth bounding box.

If the proposed window is a false positive, the CNN outputs a vector of $[-1, -1, -1, -1]$. We adopt the Euclidean loss and cross-entropy loss for bounding box regression and face classification, respectively. More details of face detection can be found in the **appendix**.

5. Experimental Results

5.1. Benchmarks

As we discussed in Sec. 2, face detection algorithms can be broadly grouped into four representative categories. For each class, we pick one algorithm as a baseline method. We select VJ [29], ACF [33], DPM [21], and Faceness [36] as baselines. The VJ [29], DPM [21], and Faceness [36] detectors are either obtained from the authors or from open source library (OpenCV). The ACF [33] detector is reimplemented using the open source code. We adopt the Scenario-Ext here (see Sec. 3.1), that is, these detectors were trained by using external datasets and are used 'as is' without re-training them on WIDER FACE. We employ PASCAL VOC [6] evaluation metric for the evaluation. Following previous work [21], we conduct linear transformation for each method to fit the annotation of WIDER FACE.

Overall. In this experiment, we employ the evaluation setting mentioned in Sec. 3.3. The results are shown in Fig. 6 (a.1)-(a.3). Faceness [36] outperforms other methods on three subsets, with DPM [21] and ACF [33] as marginal second and third. For the easy set, the average precision (AP) of most methods are over 60%, but none of them surpasses 75%. The performance drops 10% for all methods on the medium set. The hard set is even more challenging. The performance quickly decreases, with a AP below 30% for all methods. To trace the reasons of failure, we examine performance on varying subsets of the data.

Scale. As described in Sec. 3.3, we group faces according to the image height: small (10-50 pixels), medium (50-300 pixels), and large (300 or more pixels) scales. Fig. 6 (b.1)-(b.3) show the results for each scale on un-occluded faces only. For the large scale, DPM and Faceness obtain over 80% AP. At the medium scale, Faceness achieves the best relative result but the absolute performance is only 70% AP. The results of small scale are abysmal: none of the algorithms is able to achieve more than 12% AP. This shows that current face detectors are incapable to deal with faces of small scale.

Occlusion. Occlusion handling is a key performance metric for any face detectors. In Fig. 6 (c.1)-(c.3), we show the impact of occlusion on detecting faces with a height of at least 30 pixels. As mentioned in Sec. 3.3, we classify faces into three categories: un-occluded, partially occluded (1%-30% area occluded) and heavily occluded (over 30% area occluded). With partial occlusion, the performance drops significantly. The maximum AP is only 26.5% achieved by Faceness. The performance further decreases in the heavy occlusion setting. The best performance of baseline methods drops to 14.4%. It is worth noting that Faceness and DPM, which are part based models, already perform relatively better than other methods on occlusion handling.

Pose. As discussed in Sec. 3.3, we assign a face pose as atypical if either the roll or pitch degree is larger than 30-degree; or the yaw is larger than 90-degree. Otherwise a face pose is classified as typical. We show results in Fig. 6 (d.1)-(d.2). Faces which are un-occluded and with a scale larger than 30 pixels are used in this experiment. The performance clearly degrades for atypical pose. The best performance is achieved by Faceness, with a recall below 20%. The results suggest that current face detectors are only capable of dealing with faces with out-of-plane rotation and a small range of in-plane rotation.

Summary. Among the four baseline methods, Faceness tends to outperform the other methods. VJ performs poorly on all settings. DPM gains good performance on medium/large scale and occlusion. ACF outperforms DPM on small scale, no occlusion and typical pose settings. However, the overall performance is poor on WIDER FACE, suggesting a large room of improvement.

5.2. WIDER FACE as an Effective Training Source

In this experiment, we demonstrate the effectiveness of WIDER FACE dataset as a training source. We adopt Scenario-Int here (see Sec. 3.1). We train ACF and Faceness on WIDER FACE to conduct this experiment. These two algorithms have shown relatively good performance on WIDER FACE previous benchmarks see (Sec. 5.1). Faces with a scale larger than 30 pixels in the training set are used to retrain both methods. We train the ACF detector using the same training parameters as the baseline ACF. The negative samples are generated from the training images. For the Faceness detector, we first employ models shared by the authors to generate face proposals from the WIDER FACE training set. After that, we train the classifier with the same procedure described in [36]. We test these models (denoted as ACF-WIDER and Faceness-WIDER) on WIDER FACE testing set and FDDB dataset.

WIDER FACE. As shown in Fig. 7, the retrained models perform consistently better than the baseline models. The average AP improvement of retrained ACF detector is 5.4% in comparison to baseline ACF detector. For the Faceness, the retrained Faceness model obtain 4.2% improvement on WIDER hard test set.

FDDB. We further evaluate the retrained models on FDDB dataset. Similar to WIDER FACE dataset, the retrained models achieve improvement in comparison to the baseline methods. The retrained ACF detector achieves a recall rate of 87.48%, outperforms the baseline ACF by a considerable margin of 1.4%. The retrained Faceness detector obtains a

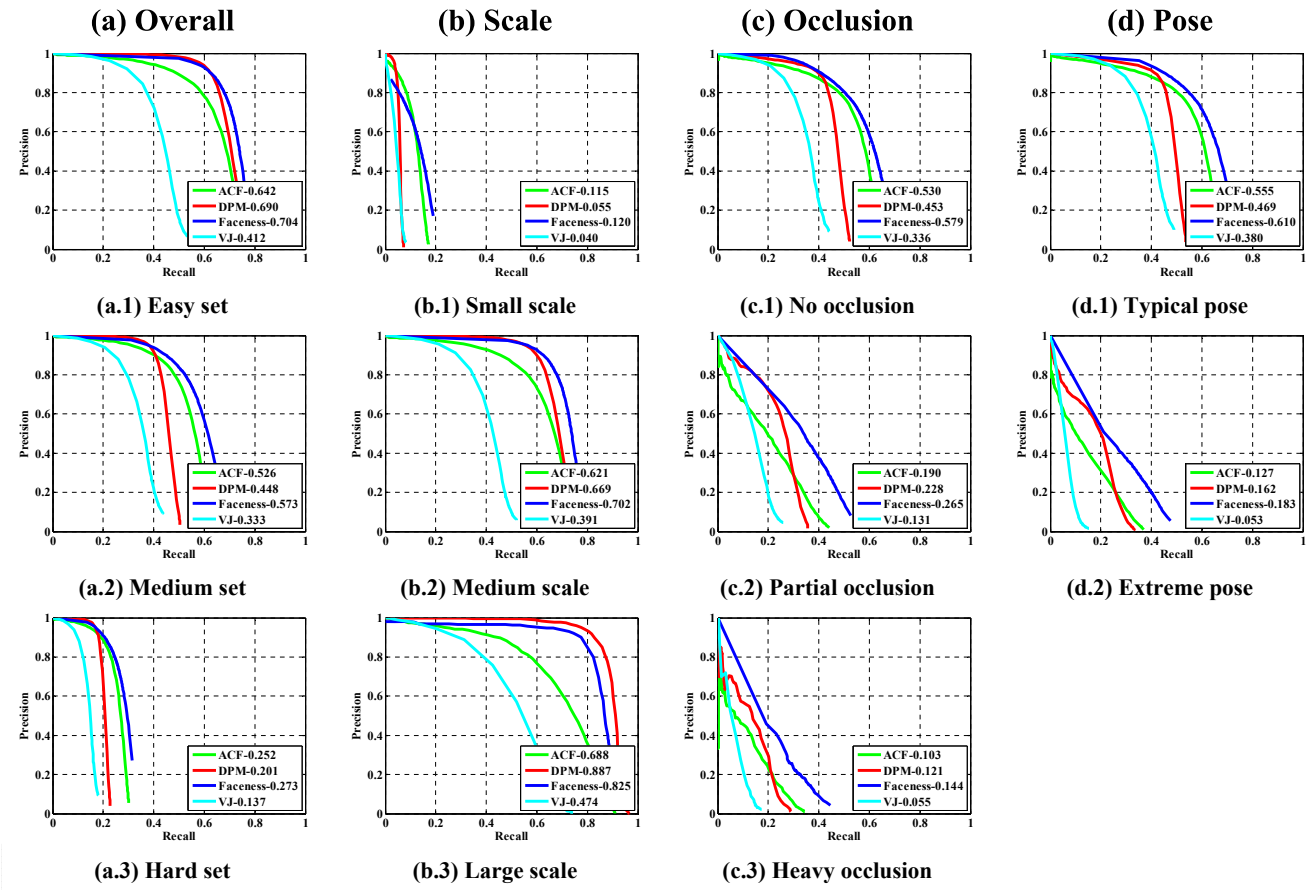


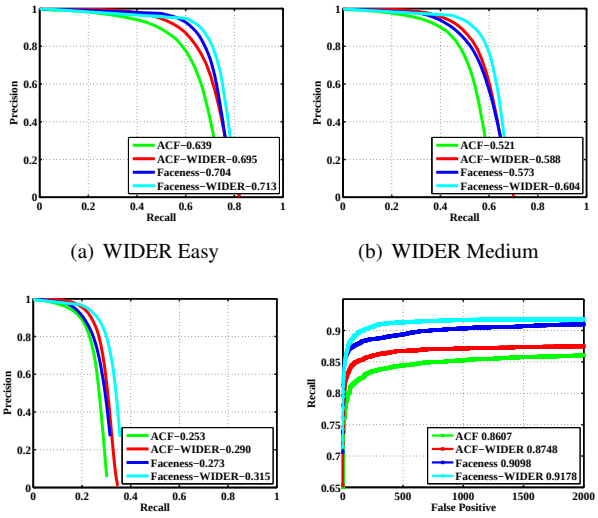
Figure 6. Precision and recall curves of different subsets of WIDER FACES: (a.1)-(a.3) Overall Easy/Medium/Hard subsets. (b.1)-(b.3) Small/Medium/Large scale subsets. (c.1)-(c.3) None/Partial/Heavy occlusion subsets. (d.1)-(d.2) Typical/Atypical pose subsets.

Table 3. Comparison of per class AP. To save space, we only show abbreviations of category names here. The event category is organized based on the rank sequence in Fig. 9 (from hard to easy events based on scale measure). We compare the accuracy of Faceness and ACF models retrained on WIDER FACE training set with the baseline Faceness and ACF. With the help of WIDER FACE dataset, accuracies on 56 out of 60 categories have been improved. The re-trained Faceness model wins 30 out of 60 classes, followed by the ACF model with 26 classes. Faceness wins 1 medium class and 3 easy classes.

	Traf.	Fest.	Para.	Demo.	Cere.	March.	Bask.	Shop.	Mata.	Acci.	Elec.	Conc.	Awar.	Picn.	Riot.	Fune.	Chee.	Firi.	Raci.	Vote.
ACF	.421	.368	.431	.330	.521	.381	.452	.503	.308	.254	.409	.512	.720	.475	.388	.502	.474	.320	.552	.457
ACF-WIDER	.385	.435	.528	**.464**	.595	.490	**.562**	**.603**	.334	.352	**.538**	.486	**.797**	.550	.395	.568	**.589**	.432	**.669**	.532
Faceness	.497	.376	.459	.410	.547	.434	.481	.575	.388	.323	.461	.569	.730	.526	.455	.563	.496	.439	.577	.535
Faceness-WIDER	**.535**	**.451**	**.560**	.454	**.626**	**.495**	.525	.593	**.432**	**.358**	.489	**.576**	.737	**.621**	**.486**	**.579**	.555	**.454**	.635	**.558**

	Stoc.	Hock.	Stud.	Skat.	Gree.	Foot.	Runn.	Driv.	Dril.	Phot.	Spor.	Grou.	Cele.	Socc.	Inte.	Raid.	Base.	Patr.	Angl.	Resc.
ACF	.549	.430	.557	.502	.467	.394	.626	.562	.447	.576	.343	.685	.577	.719	.628	.407	.442	.497	.564	.465
ACF-WIDER	.519	**.591**	**.666**	**.630**	.546	**.508**	**.707**	.609	**.521**	.627	**.430**	.756	.611	.727	.616	**.506**	.583	.529	**.645**	.546
Faceness	**.617**	.481	.639	.561	.576	.475	.667	.643	.469	.628	.406	.725	.563	.744	.680	.457	.499	.538	.621	.520
Faceness-WIDER	.611	.579	.660	.599	**.588**	.505	.672	**.648**	.519	**.650**	.409	**.776**	**.621**	**.768**	**.686**	.489	**.607**	**.607**	.629	**.564**

	Gymn.	Hand.	Wait.	Pres.	Work.	Parach.	Coac.	Meet.	Aero.	Boat.	Danc.	Swim.	Fami.	Ball.	Dres.	Coup.	Jock.	Tenn.	Spa.	Surg.
ACF	.749	.472	.722	.720	.589	.435	.598	.548	.629	.530	.507	.626	.755	.589	.734	.621	.667	.701	.386	.599
ACF-WIDER	.750	**.589**	**.836**	**.794**	**.649**	.492	**.705**	**.700**	**.734**	.602	.524	.534	**.856**	.642	**.802**	.589	**.827**	.667	.418	.586
Faceness	.756	.540	.782	.732	.645	.517	.618	.592	.678	.569	.558	**.666**	.809	.647	.774	**.742**	.662	.744	.470	**.635**
Faceness-WIDER	**.768**	.577	.740	.746	.640	**.540**	.637	.670	.718	**.628**	**.595**	.659	.842	**.682**	.754	.699	.688	**.759**	**.493**	.632



(a) WIDER Easy (b) WIDER Medium

(c) WIDER Hard (d) FDDB

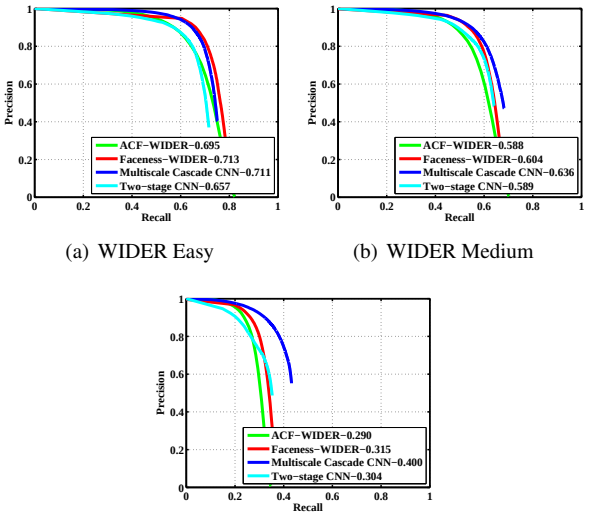
Figure 7. WIDER FACE as an effective training source. ACF-WIDER and Faceness-WIDER are retrained with WIDER FACE, while ACF and Faceness are the original models. (a)-(c) Precision and recall curves on WIDER Easy/Medium/Hard subsets. (d) ROC curve on FDDB dataset.

high recall rate of 91.78%. The recall rate improvement of the retrained Faceness detector is 0.8% in comparison to the baseline Faceness detector. It worth noting that the retrained Faceness detector performs much better than the baseline Faceness detector when the number of false positive is less than 300.

Event. We evaluate the baseline methods on each event class individually and report the results in Table 3. Faces with a height larger than 30 pixels are used in this experiment. We compare the accuracy of Faceness and ACF models retrained on WIDER FACE training set with the baseline Faceness and ACF. With the help of WIDER FACE dataset, accuracies on 56 out of 60 event categories have been improved. It is interesting to observe that the accuracy obtained highly correlates with the difficulty levels specified in Sec. 3.3 (also refer to Fig. 9). For example, the best performance on "Festival" which is assigned as a hard class is no more than 46% AP.

5.3. Evaluation of Multi-scale Detection Cascade

In this experiment we evaluate the effectiveness of the proposed multi-scale cascade algorithm. Apart from the ACF-WIDER and Faceness-WIDER models (Sec. 5.2), we establish a baseline based on a "Two-stage CNN". This model differs to our multi-scale cascade model in the way it handles multiple face scales. Instead of having multiple networks targeted for different scales, the two-stage CNN adopts a more typical approach. Specifically, its first stage consists only a single network to perform face classification. During testing, an image pyramid that encompasses different scales of a test image is fed to the first stage to generate multi-scale face proposals. The second stage is similar to



(a) WIDER Easy (b) WIDER Medium

(c) WIDER Hard

Figure 8. Evaluation of multi-scale detection cascade: (a)-(c) Precision and recall curves on WIDER Easy/Medium/Hard subsets.

our multi-scale cascade model – it performs further refinement on proposals by simultaneous face classification and bounding box regression.

We evaluate the multi-scale cascade CNN and baseline methods on WIDER Easy/Medium/Hard subsets. As shown in Fig. 8, the multi-scale cascade CNN obtains 8.5% AP improvement on the WIDER Hard subset compared to the retrained Faceness, suggesting its superior capability in handling faces with different scales. In particular, having multiple networks specialized on different scale range is shown effective in comparison to using a single network to handle multiple scales. In other words, it is difficult for a single network to handle large appearance variations caused by scale. For the WIDER Medium subset, the multi-scale cascade CNN outperforms other baseline methods with a considerable margin. All models perform comparably on the WIDER Easy subset.

6. Conclusion

We have proposed a large, richly annotated WIDER FACE dataset for training and evaluating face detection algorithms. We benchmark four representative face detection methods. Even considering an easy subset (typically with faces of over 50 pixels height), existing state-of-the-art algorithms reach only around 70% AP, as shown in Fig. 8. With this new dataset, we wish to encourage the community to focusing on some inherent challenges of face detection – small scale, occlusion, and extreme poses. These factors are ubiquitous in many real world applications. For instance, faces captured by surveillance cameras in public spaces or events are typically small, occluded, and atypical poses. These faces are arguably the most interesting yet crucial to detect for further investigation.

7. Appendix

7.1. Multi-scale Detection Cascade

Multi-scale detection cascade CNN consists of a set of face detectors, with each of them only deals with faces in a relatively small range of scale. Each face detector consists of two stages. The first stage generates multi-scale proposal from a fully-convolutional network. The second stage gives face and non-face prediction of the candidate windows generate from first stage. If the candidate window is classified as face, we further refine the location of the candidate window.

7.1.1 Training Multi-scale Proposal Network

We provide details of the training process for multi-scale proposal networks. In this step, we train four fully convolutional networks for face classification and scale classification. The network structures are summarized in Table 5, Table 6, Table 7, and Table 8. As we described in the paper, we group faces into four categories by their image size, as shown in the Table 4 (each row in the table represents a category). For each group, we further divide it into three subclasses. Each network is trained with image patches with the size of their upper bound scale. For example, Proposal Network 1 and Proposal Network 2 are trained with 30×30 and 120×120 image patches respectively. For the Proposal Network 1, Proposal Network 2, and Proposal Network 3, we initialize the layers from Conv 1 to Conv 5 using the Imagenet $1,000$ categories pre-trained Clarifai net. For the Proposal Network 4, we initialize the layers from Conv 2 to Conv 5 using pre-trained Clarifai net. The remaining layers in each network are randomly initialized with weights drawn from a Gaussian distribution of $\mu = 0$ and $\sigma = 0.01$. To account for the multi-label scenario, cross-entropy loss is adopted as shown below:

$$L = \sum_{i=1}^{|D|}(\mathbf{y_i}\log p(\mathbf{y_i}|\mathbf{I_i}) + (\mathbf{1}-\mathbf{y_i})\log(\mathbf{1}-p(\mathbf{y_i}|\mathbf{I_i}))), \quad \text{and}$$
$$p(\mathbf{y_i}=c|\mathbf{I_i}) = \frac{\mathbf{1}}{\mathbf{1}+\exp(-f(\mathbf{I_i}))}, \tag{1}$$

Back propagation and SGD are also employed here for optimizing Eqn. (1). Similar to [10, 16], we set the initial fine-tuning learning rate as one-tenth of the corresponding pre-training learning rate and drop it by a factor of 10 throughout training. After training, we conduct hard negative mining on the training set and further tune the proposal networks using hard negative samples.

7.1.2 Training Face Detector

In this section, we provide more details of the training process for face detection. As mentioned at beginning, we train four fully convolutional neural networks for face detection. Each detection network is trained to conduct face classification and bounding box regression simultaneously. We fine-tune the detection networks in this stage using respective proposal network in the previous stage. For example, the Detection Network 1 is trained by fine-tuning the Proposal Network 1 in the previous stage with the same structure. The detection networks are trained with face proposals generated from respective proposal networks. For face classification, we assign a proposed bounding box to a ground truth bounding box based on the minimal Euclidean distances between the center of the proposed bounding box and the center of ground truth bounding box. The proposed bounding box is assigned as positive, if the IoU between proposed

Table 4. Summary of face scale for multi-scale proposal networks.

Scale	Class 1	Class 2	Class 3
Network 1	10-15	15-20	20-30
Network 2	30-50	50-80	80-120
Network 3	120-160	160-200	200-240
Network 4	240-320	320-400	400-480

Table 5. Model structure of Network 1.

Layer name	Filter Number	Filter Size	Stride	Padding	Activation Function
Conv 1	96	7×7	2	0	RELU
Pool 1	-	3×3	2	-	-
LRN 1	-	5×5	-	-	-
Conv 2	256	5×5	1	1	RELU
LRN 2	-	5×5	-	-	-
Conv 3	384	3×3	1	1	RELU
Conv 4	384	3×3	1	1	RELU
Conv 5	256	3×3	1	1	RELU
Conv 6	256	4×4	1	0	RELU
Conv 7	256	1×1	1	0	RELU

Table 6. Model structure of Network 2.

Layer name	Filter Number	Filter Size	Stride	Padding	Activation Function
Conv 1	96	7×7	2	0	RELU
Pool 1	-	3×3	2	-	-
LRN 1	-	5×5	-	-	-
Conv 2	256	5×5	2	1	RELU
Pool 2	-	3×3	2	-	-
LRN 2	-	5×5	-	-	-
Conv 3	384	3×3	1	1	RELU
Conv 4	384	3×3	1	1	RELU
Conv 5	256	3×3	1	1	RELU
Conv 6	256	3×3	1	1	RELU
Conv 7	256	3×3	1	0	RELU
Conv 8	1024	4×4	1	0	RELU
Conv 9	1024	1×1	1	0	RELU

Table 7. Model structure of Network 3.

Layer name	Filter Number	Filter Size	Stride	Padding	Activation Function
Conv 1	96	7×7	2	0	RELU
Pool 1	-	3×3	2	-	-
LRN 1	-	5×5	-	-	-
Conv 2	256	5×5	2	1	RELU
Pool 2	-	3×3	2	-	-
LRN 2	-	5×5	-	-	-
Conv 3	384	3×3	1	1	RELU
Conv 4	384	3×3	1	1	RELU
Conv 5	256	3×3	1	1	RELU
Pool 5	-	3×3	3	-	-
Conv 6	4096	5×5	1	0	RELU
Conv 7	4096	1×1	1	0	RELU

Table 8. Model structure of Network 4.

Layer name	Filter Number	Filter Size	Stride	Padding	Activation Function
Conv 1	96	11×11	4	0	RELU
Pool 1	-	3×3	2	-	-
LRN 1	-	5×5	-	-	-
Conv 2	256	5×5	2	1	RELU
Pool 2	-	3×3	2	-	-
LRN 2	-	5×5	-	-	-
Conv 3	384	3×3	1	1	RELU
Conv 4	384	3×3	1	1	RELU
Conv 5	256	3×3	1	1	RELU
Pool 5	-	3×3	2	-	-
Conv 6	4096	4×4	1	0	RELU
Conv 7	4096	1×1	1	0	RELU

bounding box and the assigned ground truth bounding box is larger than 0.5; otherwise it is negative. For bounding box regression, we train the multi-task deep convolutional neural network to regress each proposal to predict the positions of its assigned ground truth bounding box.

During the training process, if the number of positive samples is less than 10% of the total samples, we randomly cropped the ground truth faces and add these samples as positive samples. We adopt Euclidean loss and cross-entropy loss for bounding box regression and face classification, respectively.

7.2. WIDER FACE Dataset

Event. We measure each event with three factors: scale, occlusion and pose. For each factor, we compute the detection rate for the specific event class and then rank the detection rate in the ascending order. Three classes are divided: easy (top 41-60 class), medium (top 21-40 class) and hard (top 1-20 class), as shown in the Fig. 9. The corresponding relationship between abbreviations of the events categories and full name of the event categories are shown in Table 10.

References

[1] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *CVPR*, 2014. 3

[2] Z. L. Bin Yang, Junjie Yan and S. Z. Li. Convolutional channel features. In *ICCV*, 2015. 2

[3] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In *ECCV*. 2014. 1, 2

[4] P. Dollar, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *BMVC*, 2009. 2

[5] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR*, 2009. 3

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal visual object classes VOC challenge. *IJCV*, 2010. 3, 6

[7] S. S. Farfade, M. Saberian, and L. Li. Multi-view face detection using deep convolutional neural networks. In *ICMR*, 2015. 2

[8] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2010. 2

[9] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012. 4

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 9

[11] J. Hosang, R. Benenson, and B. Schiele. How good are detection proposals, really? In *BMVC*, 2014. 3

[12] C. Huang, H. Ai, Y. Li, and S. Lao. High-performance rotation invariant multiview face detection. *TPAMI*, 2007. 2

[13] V. Jain and E. Learned-Miller. FDDB: A benchmark for face detection in unconstrained settings. Technical report, University of Massachusetts, Amherst, 2010. 1, 2

[14] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark A. In *CVPR*, 2015. 2

[15] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011. 4

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 9

[17] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *CVPR*, 2015. 1, 2

[18] J. Li and Y. Zhang. Learning SURF cascade for fast and accurate object detection. In *CVPR*, 2013. 2

[19] S. Liao, A. K. Jain, and S. Z. Li. A fast and accurate unconstrained face detector. *TPAMI*, 2015. 2

[20] P. Luo, X. Wang, and X. Tang. Hierarchical face parsing via deep learning. In *CVPR*, 2012. 1

[21] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*. 2014. 1, 2, 6

[22] M. Naphade, J. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *MultiMedia*, 2006. 3

[23] R. Ranjan, V. M. Patel, and R. Chellappa. A deep pyramid deformable part model for face detection. *CoRR*, 2015. 2

[24] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *CVPR*, 2014. 1

[25] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 1

[26] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, 2014. 1

[27] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 1

[28] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *IJCV*, 2013. 3

[29] P. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 2004. 1, 2, 6

[30] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013. 1

[31] Y. Xiong, K. Zhu, D. Lin, and X. Tang. Recognize complex events from static images by fusing deep channels. 2015. 3

[32] J. Yan, X. Zhang, Z. Lei, and S. Z. Li. Face detection by structural models. *IVC*, 2014. 1, 2

[33] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Aggregate channel features for multi-view face detection. *CoRR*, 2014. 1, 2, 6

[34] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Fine-grained evaluation on face detection in the wild. In *FG*, 2015. 2, 3, 4

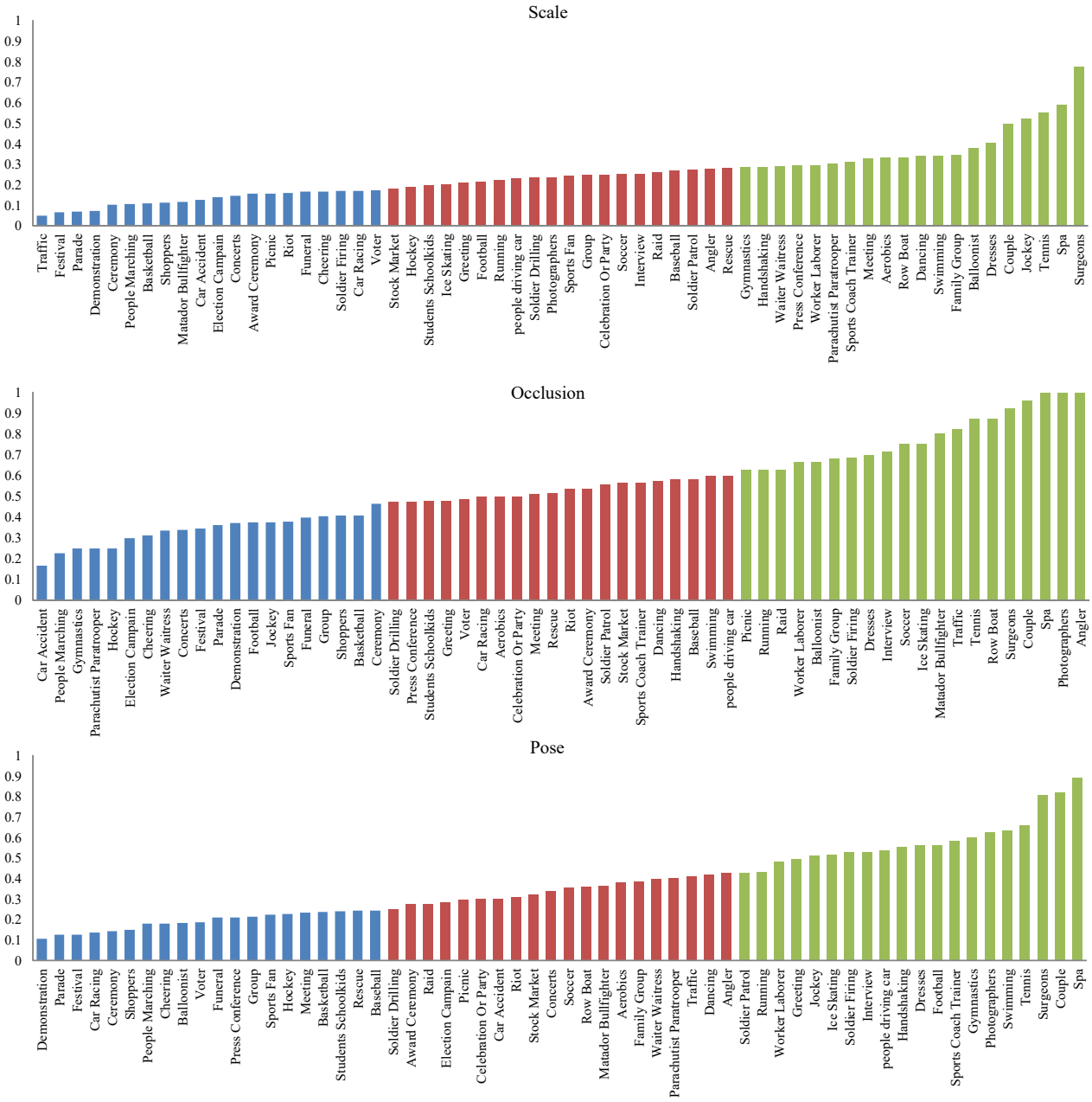


Figure 9. Histogram of detection rate for different event categories. Event categories are ranked based on detection rate when number of proposal is $10,000$ in the ascending order. Top $1-20$, $21-40$, $41-60$ event categories are denote in blue, red, green respectively. Example images for specific event class are shown. Y-axis denotes for detection rate. X-axis denotes for event class name.

[35] M.-H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: a survey. *TPAMI*, 2002. 1, 2

[36] S. Yang, P. Luo, C. C. Loy, and X. Tang. From facial parts responses to face detection: A deep learning approach. In *ICCV*, 2015. 1, 2, 6

[37] C. Zhang and Z. Zhang. A survey of recent advances in face detection. Technical report, Tech. rep., Microsoft Research, 2010. 2

[38] Z. Zhang, P. Luo, C. Loy, and X. Tang. Learning deep representation for face alignment with auxiliary attributes. *TPAMI*, 2015. 1

[39] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012. 1, 2

[40] C. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 3, 4

Table 9. Full name of event categories.

Traf.	Fest.	Para.	Demo.	Cere.	March.	Bask.	Shop.	Mata.	Acci.	Elec.	Conc.	Awar.	Picn.	Riot.	Fune.	Chee.	Firi.	Raci.	Vote.
Traffic	Festival	Parade	Demonstration	Ceremony	People Marching	Basketball	Shoppers	Matador Bullfighter	Car Accident	Election Campain	Concerts	Award Ceremony	Picnic	Riot	Funeral	Cheering	Soldier Firing	Car Racing	Voter
Stoc.	Hock.	Stud.	Skat.	Gree.	Foot.	Runn.	Driv.	Dril.	Phot.	Spor.	Grou.	Cele.	Socc.	Inte.	Raid.	Base.	Patr.	Angl.	Resc.
Stock Market	Hockey	Students Schoolkids	Ice Skating	Greeting	Football	Running	People Driving Car	Soldier Drilling	Photographers	Sports Fan	Group	Celebration Or Party	Soccer	Interview	Raid	Baseball	Soldier Patrol	Angler	Rescue
Gymn.	Hand.	Wait.	Pres.	Work.	Parach.	Coac.	Meet.	Aero.	Boat.	Danc.	Swim.	Fami.	Ball.	Dres.	Coup.	Jock.	Tenn.	Spa.	Surg.
Gymnastics	Handshaking	Waiter Waitress	Press Conference	Worker Laborer	Parachutist Paratrooper	Sports Coach Trainer	Meeting	Aerobics	Row Boat	Dancing	Swimming	Family Group	Balloonist	Dresses	Couple	Jockey	Tennis	Spa	Surgeons

Table 10. Full name of abbreviation of event categories.