Feature Agglomeration Networks for Single Stage Face Detection

Jialiang Zhang[†‡*], Xiongwei Wu[†*], Jianke Zhu[‡], Steven C.H. Hoi[†§]
[†]School of Information Systems, Singapore Management University, Singapore
[‡]Colleage of Computer Science and Technology, Zhejiang University, Hangzhou, China
[§]DeepIR Inc., Beijing, China
{chhoi,jlzhang,xwwu.2015@phdis}@smu.edu.sg;{zjialiang,jkzhu}@zju.edu.cn

Abstract

Recent years have witnessed promising results of face detection using deep learning, especially for the family of region-based convolutional neural networks (R-CNN) methods and their variants. Despite making remarkable progresses, face detection in the wild remains an open research challenge especially when detecting faces at vastly different scales and characteristics. In this paper, we propose a novel framework of "Feature Agglomeration Networks" (FAN) to build a new single stage face detector, which not only achieves state-of-the-art performance but also runs efficiently. As inspired by the recent success of Feature Pyramid Networks (FPN) [15] for generic object detection, the core idea of our framework is to exploit inherent multi-scale features of a single convolutional neural network to detect faces of varied scales and characteristics by aggregating higher-level semantic feature maps of different scales as contextual cues to augment lower-level feature maps via a hierarchical agglomeration manner at marginal extra computation cost. Unlike the existing FPN approach, we construct our FAN architecture using a new Agglomerative Connection module and further propose a Hierarchical Loss to effectively train the FAN model. We evaluate the proposed FAN detector on several public face detection benchmarks and achieved new state-of-the-art results with real-time detection speed on GPU.

1. Introduction

Face detection is often the first key step towards face related applications, such as face alignment, face verification, face recognition, face tracking and facial expression analysis, etc. Despite being studied extensively, detecting faces in the wild remains an open research problem due to various challenges with real-world faces, such as varied scales of faces and diverse characteristics of real-world faces captured from different scenarios.

Early works of face detection in computer vision community were mainly focused on crafting effective features manually and then building powerful classifiers from the hand-crafted features [25], which are often sub-optimal and may not always achieve satisfactory results. Recent years have witnessed the successful applications of deep learning techniques for face detection tasks, inspired by the remarkable successes of deep convolutional neural networks (CNN) techniques for generic image recognition [23, 14] and object detection tasks [6, 22]. Despite being extensively studied, it remains an open challenge for building a fast face detector with high accuracy in any real-world scenario.

In general, face detection can be viewed as a special case of generic object detection [6, 22]. Many previous state-of-the-art face detectors inherited a lot of successful techniques from generic object detection, especially for the family of region-based CNN (R-CNN) methods and their variants [22, 5, 1]. Among the family of R-CNN based face detectors, there are two major categories of detection frameworks: (i) two-stage detectors (a.k.a. "proposal-based"), such as Fast R-CNN [5], Faster R-CNN [22], etc; and (ii) single-stage detectors (a.k.a. "proposal-free") , such as Region Proposal Networks (RPN) [22], Single-Shot Multibox Detector (SSD) [16], etc. The single-stage detection framework enjoys much higher inference efficiency, and thus has attracted increasing attention recently due to the high demand of real-time face detectors in real applications.

Despite enjoying significant computational advantages, single-stage detectors are not always effective in detecting faces of different scales and their performance can drop dramatically when handling small faces. In order to build a robust detector that can detect faces with a large range of scales, there are two major routes for improvement. One way is to train multi-shot single-scale detectors by using the idea of image pyramid to train multiple separate single-scale detectors each of which is tuned for one specific scale (e.g., the HR detector in [8] trained multiple scale-specific RPN detectors). However, such approach with the image pyramid is computationally expensive since it has to pass a

*The first two authors contributed equally to this work.

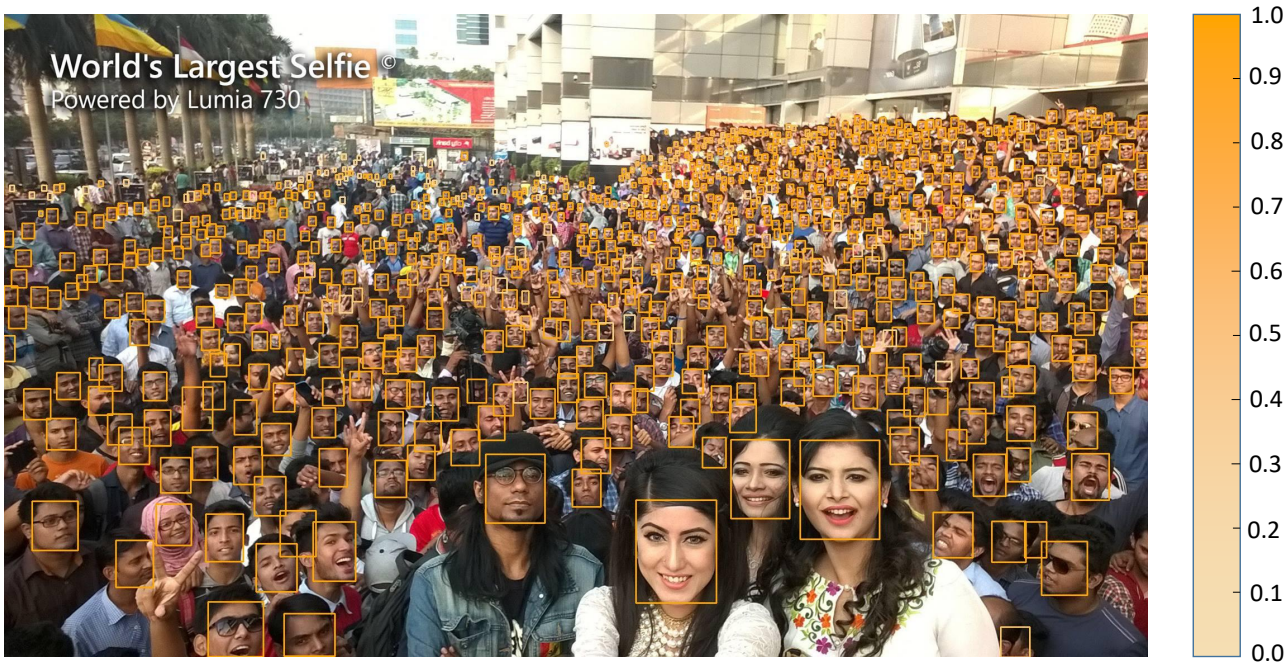


Figure 1. Example of face detection with the proposed method. In the above image, the proposed method can find 858 faces out of 1000 facial images present. The detection confidence scores are also given by the color bar as shown on the right. Best view in color.

very deep network multiple times during inference. Another way is to train a single-shot multi-scale detector by exploiting multi-scale representations in the feature hierarchy of a deep convolutional network, requiring only a single pass to the network during inference. For example, the Single-shot multi-scale Face Detector (S3FD) in [31] follows the second approach by extending SSD [16] for face detection.

Despite achieving promising performance, S3FD shares the similar drawback of SSD-style detection frameworks, where each of multi-scale feature maps is used alone for prediction and thus a high-resolution semantically weak feature map may fail to perform accurate predictions. Inspired by the recent success of Feature Pyramid Networks (FPN) [15] for generic object detection, we propose a novel detection framework of "Feature Agglomeration Networks" (FAN) to overcome the drawback of the single stage face detector "S3FD" by attempting to combine low-resolution semantically strong features with high-resolution semantically weak features. In particular, FAN aims to create a feature pyramid with rich semantics at all scales to boost the prediction performance of high-resolution feature maps using rich contextual cues from low-resolution semantically strong features. However, unlike the existing FPN for generic object detection that creates feature pyramid using the skip connection module, we propose a novel "Agglomerative Connection" module to create a new feature pyramid for FAN. Moreover, we introduce a new Hierarchical Loss to train the FAN model effectively in an end-to-end approach. We conduct extensive experiments on several public face detection benchmarks, in which our results show that FAN is more effective than a naive application of FPN [15] for face detection (though FPN has yet to be explored for face detection), and the proposed Hierarchical Loss is also critical to training the proposed FAN model.

As a summary, the main contributions of this paper include the following

- We propose a novel framework of Feature Agglomeration Networks (FAN) for single stage face detection, which creates a new effective feature pyramid with rich semantics at all scales by introducing a new agglomerative connection module to agglomerate multi-scale features via a hierarchical agglomerative manner;
- We introduce an effective Hierarchical Loss based training scheme to train the proposed FAN model in an end-to-end manner, which enables us to learn discriminative features of the feature pyramid effectively;
- We conducted comprehensive experiments on several public Face detection benchmarks to evaluate the effectiveness of the proposed FAN framework, in which promising results show that our FAN detector not only achieves the state-of-the-art performances but also runs efficiently with real-time speed on GPU.

2. Related Work

Generic Object Detection. As a special case of generic object detection, many face detectors inherit successful techniques for generic object detection [31, 16, 8, 24], which has been extensively studied using deep learning, especially for the family of region-based convolutional neural networks (R-CNN) methods and their variants[22, 5, 1]. In particular, there are two major categories of R-CNN variants for object detection: (i) two-stage detection systems where proposals are first generated in the first stage and further classified in the second stage; and (ii) single-stage detection systems where the object detection and classification are done simultaneously from the feature maps of a deep convolutional network without a separate proposal generation stage. The two-stage detection systems include Fast R-CNN[5], Faster R-CNN[22] and their variants, and the single-stage detection systems include YOLO [21], RPN [22], SSD [16], etc. Our detector essentially belongs to the single-stage detection framework.

Multi-shot single-scale Face Detector. To detect faces with a large range of scales, one way is to train multiple detectors each of which targets for a specific scale. Hu et al. [8] trained multiple separate RPN detectors for different scales and made inference using image pyramids. However, their method is very time-consuming since images are required to pass a very deep network multiple times during inference. Hao et al. [7] learned a Scale Aware Network by estimating face scales in images and built image pyramids according to the estimated values. Although to some extent they avoid computation cost from inputs with invalid scales, multiple passes are still required if faces of large ranges of scales presented in the same image. Due to high computational cost, such paradigm is not suitable for real-time applications.

Single-shot multi-scale Face Detector. Single-shot multi-box detector (SSD) [16] extracts multi-scale representations in the feature hierarchy of a network for different scales of faces and thus only a single pass is required. S3FD [31] inherits the SSD framework, and carefully designs scale-aware anchors in different feature hierarchy, according to the effective receptive fields [17]. However, S3FD shares the same limitation of SSD, where each feature map is used alone for prediction and as a consequence some high-resolution semantically weak features could fail to provide robust prediction. As inspired by FPN [15], we propose a new framework of FAN to effectively address the limitation of S3FD by combining low resolution semantically strong features with high resolution semantically weak features using the Agglomerative Connection module.

Feature Pyramid. Feature pyramid is a structure which combines shallow semantically weak features with deep semantically strong features using skip-connection which has been successfully used in both two-stage and one-stage generic object detectors. IoN [1] extracts RoI feature vectors from different layer feature maps and concatenates them together. HyperNet [13] makes prediction on a Hyper Feature Map produced by aggregating multi-scale feature maps. FPN [15] builds feature pyramids by lateral connecting deep layer features with shallow layer features in a top-down architecture. SSD-style frameworks as DSSD [4] and RON [12] also apply the idea of feature pyramid and achieve promising performance. However, all the methods above create feature pyramids using the skip connection module. In this paper, we propose a new Agglomerate Connection module which can aggregate multi-scale features more effectively than the skip connection module. Besides, we also introduce a novel Hierarchical Loss on the proposed FAN framework which enables us to train this powerful detector effectively and robustly in an end-to-end approach.

3. Feature Agglomeration Networks

In this section, we present the proposed Feature Agglomeration Networks (FAN) framework for face detection.

3.1. General Architecture

Our goal is to create an effective feature hierarchy with rich semantics at all levels to achieve robust multi-feature detection. To this end, we propose a novel hierarchical feature agglomeration structure which agglomerates adjacent features sequentially to construct a Feature Agglomeration Network (FAN) for detection. Figure 2 shows an example of the proposed FAN with 3-level feature hierarchies.

The proposed FAN framework is general-purpose and applicable to any types of detectors and CNN architectures. In this paper, without loss of generality, we consider the widely used VGG16 model as the backbone CNN architecture and SSD as the single stage detector. Suppose detection is performed on n layers of feature maps (ranging from index l to $l+n-1$), then the network structure in m-level FAN ($m \leq n$) can be mathematically defined as follows:

$$\phi_l^k = \mathcal{F}_l(\phi_{l-1}^k) \quad k=1 \tag{1}$$

$$\phi_l^k = \mathcal{A}_l(\phi_l^{k-1}, \phi_{l+1}^{k-1}) \quad k=2,..,m \tag{2}$$

where ϕ_l^k denotes the feature maps in the l-th layer and the k-th hierarchy. Specifically, for $k=1$, i.e., the first-level hierarchy, ϕ_l^k is the original feature maps in vanilla SSD, and $\mathcal{F}_l(\cdot)$ is the non-linear function to transform the feature maps from l-th layer to $(l+1)$-th layer, which consists of Convolution, ReLU and pooling layers, etc. For simplicity, we omit the range of layer index l in Eq.(1) and (2). For $k>1$, Eq.(2) denotes that ϕ_l^k is generated through an agglomeration operation $\mathcal{A}_l$ to agglomerate two adjacent-layer feature maps in the same hierarchy ϕ_l^{k-1}, ϕ_{l+1}^{k-1}. This is called an "Agglomerative Connection" building block as shown in Figure 3, which is denoted as $\mathcal{A}$-block for short.

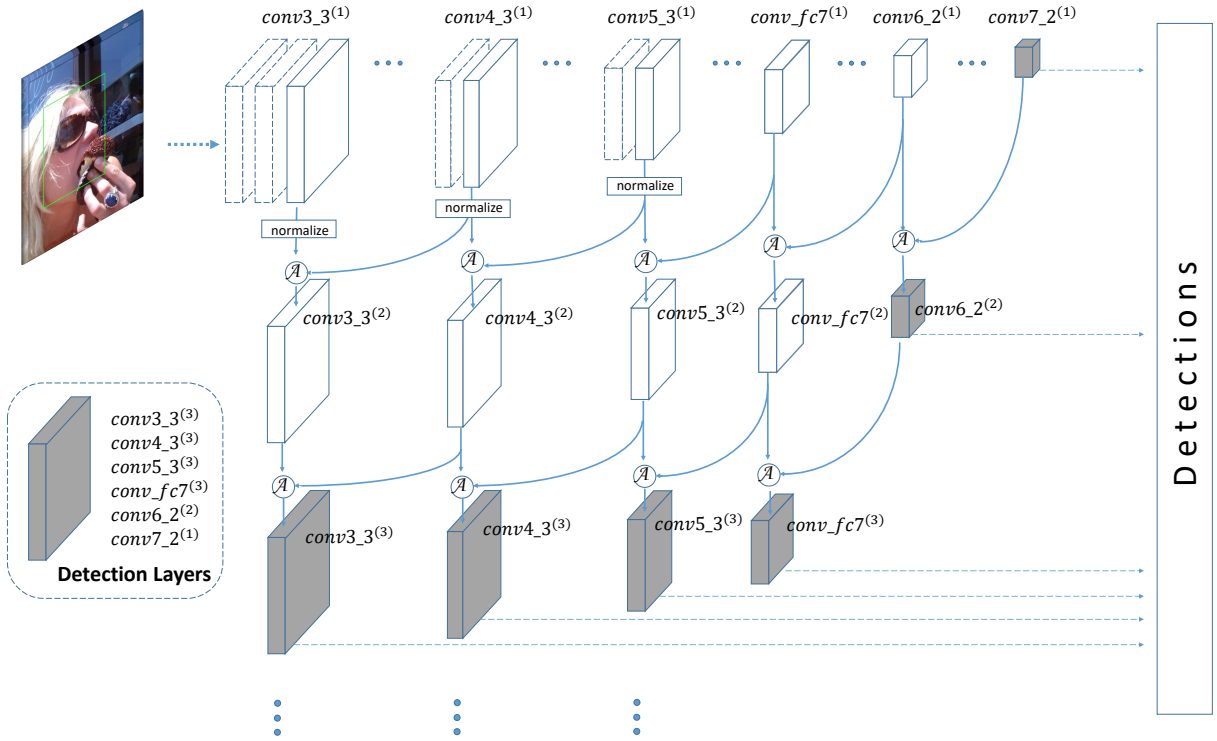


Figure 2. The network architecture of the proposed "Feature Agglomeration Networks" (FAN). Here demonstrates an example of three-level FAN architecture using the VGG-16 variant as the backbone CNN network.

Specifically, each $\mathcal{A}$-block consists of two input feature maps, a shallower ϕ_l^{k-1} and a deeper ϕ_{l+1}^{k-1}. We first use a 1×1 convolution to change the channel of the shallower feature ϕ_l^{k-1} to a fixed number N (e.g., 256). Then the dimensionality of the deeper feature ϕ_{l+1}^{k-1} is reduced via a 1×1 convolution to $\frac{1}{8}$ of N (e.g., 32) followed by a $2\times$ bilinear upsampling in order to achieve the same size as ϕ_l^{k-1}. The final agglomerative feature ϕ_l^k is obtained by the concatenation of these two features.

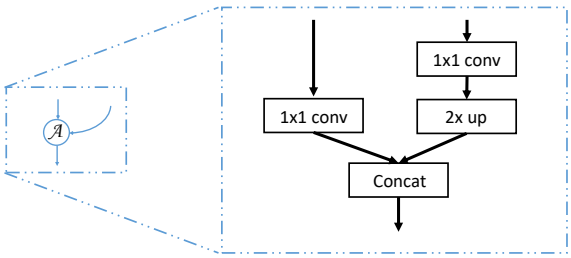


Figure 3. The Agglomerative Connection block ($\mathcal{A}$-block).

The final detection exploits the m-th hierarchy of feature maps, e.g., if $m = 3$, the detection layers of features are $\{\text{conv3\_3}^{(3)}, \text{conv4\_3}^{(3)}, \text{conv5\_3}^{(3)}, \text{conv\_fc7}^{(3)}, \text{conv6\_2}^{(2)}, \text{conv7\_2}^{(1)}\}$, and the detection result denotes as

$$Result = \mathcal{D}(\phi_l^m, \phi_{l+1}^m, ..., \phi_{l+n-1}^m) \quad (3)$$

where $\mathcal{D}$ denotes the final detection process including bounding box regression and class prediction followed by Non-Maximum Suppression to obtain the final detection results. To make the notation consistent, we use superscript m to denote all the feature maps in the m-th level hierarchy. However, $\text{conv7\_2}^{(3)}$ and $\text{conv7\_2}^{(2)}$ is actually identical to $\text{conv7\_2}^{(1)}$, and $\text{conv6\_2}^{(3)}$ is identical to $\text{conv6\_2}^{(2)}$, etc.

It is worth noting that the proposed network degenerates to the vanilla SSD detector if the agglomeration in Eq.(2) is excluded, and use the first hierarchy for detection, i.e., $m = 1$ in Eq.(3). The insight behind hierarchical agglomerative design is that in the vanilla SSD, shallower features despite being semantically weak are important to detect small faces. The $\mathcal{A}$-block hierarchically aggregates semantics information from deeper layers to form a stronger set of hierarchical multi-feature maps. The ratio of deeper feature and shallower feature in one $\mathcal{A}$-block is set to $\frac{1}{8}$ which ensures that the shallower feature dominates the composition. The deeper (semantically stronger) feature generally plays a role of providing extra contextual cues. Besides, we note that the receptive field largely impacts the performance of detecting small faces. As shown in [8], too large receptive field can hurt the performance, and so as if it is too small. The superiority of our design is that the agglomerative connection only incorporates semantics from a deeper layer feature, and thus

we can easily control the receptive field of each feature map through our hierarchical design. This is in contrast to FPN [15] where a feature map incorporates information from all the deeper layers. We found our new design attributes to achieving stable and effective training.

3.2. Detailed Configurations

We discuss more details with the proposed FAN framework, which can be designed in a flexible way. In practice, assume we perform detection on n feature maps of a CNN model (specifically n=6 in VGG-16 as used in our experiment), we can design a FAN structure with m-level hierarchies where $m \leq 6$. Figure 2 showed a 3-level hierarchy FAN structure. The detection layers of feature maps $\{\text{conv3\_3}^{(3)}, \text{conv4\_3}^{(3)}, \text{conv5\_3}^{(3)}, \text{conv\_}fc7^{(3)}, \text{conv6\_2}^{(2)}, \text{conv7\_2}^{(1)}\}$ have strides of $\{4, 8, 16, 32, 64, 128\}$, respectively. We follow the settings of [31], each of the six detection feature maps is associated with a specific scale anchor $\{16, 32, 64, 128, 256, 512\}$ to detect corresponding scale faces. The aspect ratio of each anchor is 1:1 since the size of a face is roughly 1:1.

For anchor-based detectors, we need to match each anchor as a positive or negative sample according to the ground truth bounding boxes. We adopt the following matching strategy: (i) for each face, the anchor with best jaccard overlap is matched; and (ii) each anchor is matched to the face that has jaccard overlap larger than 0.35. Max-out background label for the lowest feature map $\text{conv3\_3}^{(3)}$ is also adopted [31]. Specifically, for each anchor of $\text{conv3\_3}^{(3)}$, $M = 3$ scores are predicted and then the highest score is chosen as the final background score.

3.3. Hierarchical Loss

In order to train the proposed FAN model effectively, we propose a new loss function called the hierarchical loss defined on the proposed FAN structure. The key idea is to define a loss function that accounts for all the hierarchies of feature maps, and meanwhile allows to train the entire network effectively in an end-to-end approach. To this end, we propose the hierarchical loss as follows

$$\text{HL}(\phi_l^m, ..., \phi_{l+n-1}^m) = \sum_{i=1}^{m} \omega_i L(\phi_l^i, ..., \phi_{l+n-1}^i) \qquad (4)$$

where ω_i is a weight parameter for the loss of the i-th hierarchy. $\text{L}(\phi_l^i, ..., \phi_{l+n-1}^i)$ accounts for the loss on the i-th hierarchy, which is defined as follows

$$\text{L}(\phi_l^i, ..., \phi_{l+n-1}^i) = \frac{1}{N_i} \sum_{j=1}^{N_i} \Big(\lambda L_{cls}(y_j, y_j^*) + L_{loc}(\mathbf{p}_j, \mathbf{p}_j^*) \Big) \qquad (5)$$

where y_j denotes if the corresponding anchor is a face or not, $y_j^* \in \{0, 1\}$ is the ground truth label, $\mathbf{p}_j = [p_x, p_y, p_w, p_h]_j$ denotes the 4 coordinates of a predicted bounding box, $\mathbf{p}_j^*$ denotes the ground-truth box, N_i denotes the total number of matched bounding boxes, and λ is a parameter to tradeoff between classification loss and localization loss. In particular, the classification loss is based on a standard softmax loss, i.e.,

$$L_{cls}(y_j, y_j^*) = y_j^* \log(y_j) + (1 - y_j^*) \log(1 - y_j) \qquad (6)$$

and the localization of bounding box is based on a standard regression loss proposed in [5] defined as follows

$$L_{loc}(\mathbf{p}_j, \mathbf{p}_j^*) = \sum_{k \in \{x,y,w,h\}} \text{smooth}_{L_1}(\mathbf{p}_j - \mathbf{p}_j^*)_k \qquad (7)$$

where

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1, \\ |x| - 0.5 & \text{otherwise.} \end{cases} \qquad (8)$$

Using the proposed hierarchical loss, we can train the FAN detector end-to-end. Specifically, during training, all the losses are simultaneously computed, and the gradients are back propagated to each hierarchy of feature maps.

Remarks. In contrast to the standard loss, the proposed hierarchical loss enjoys some key advantages. First, the use of hierarchical loss plays a crucial role in training the FAN model robustly and effectively. This is because, in contract to the network of a vanilla SSD, FAN has more parameters for optimization, which is not easy to be directly trained with the existing loss in vanilla SSD training. With multiple hierarchies, hierarchical loss gradually increases the power of feature maps representation, and thus allows us to supervise the training process hierarchically to obtain more robust features. Finally, compared with the standard single loss, after the model has been trained, the use of hierarchical loss does not incur extra computation cost during inference.

3.4. Training

Next we introduce more details of our training method.

Training data and data augmentation. Our model was trained on $12,880$ images of the WIDER FACE training set [29]. We use several data augmentation strategies as follows. First of all, we follow a similar color distortion strategy [16] to preprocess training images, e.g., brightness, contrast, hue, saturation, etc. Random crop is adopted to generate training images [31]. Specifically, to keep the face aspect ratio which is important due to our anchor scale design, instead of directly resizing the whole image to a squared patch (e.g., we use 640×640 as input size of training), we first crop a squared patch from original image whose scale ranges from 0.3 to 1 of the short size of original image. The overlapped part of the face box is discarded if and only if its center is out of the sampled patch. After random cropping, the final patch is resized to 640×640 and horizontally flipped with probability of 0.5.

Hard negative mining. After the anchor matching, most of the anchors will be assigned as negative samples, which will result in a significant imbalance between positive and negative samples. Instead of using all negative samples for training, we use an online hard negative mining strategy [16] during training. In particular, all negative anchors are sorted by their classification loss values, and then the top ones are selected as negative samples to ensure the ratio between negative and positive anchors is at most $3:1$.

Other implementation details. In our experiments, we choose $\lambda = 3$ in Eq.(5) and the weight parameter ω_i in Eq.(4) as uniform for simplicity. The training starts from fine-tuning VGG16 backbone network using SGD optimizer with momentum of 0.9, weight decay of 0.0005, and a total batch size of 12 on two GPUs. The newly added layers are initialized with "xavier", the initial learning rate is 10^{-3} and becomes 10 times smaller at iteration $80k$ and $100k$. The training ended at $120k$ iterations. Our implementation is based on Caffe [10].

4. Experiments

In this section, we conduct extensive experiments and ablation studies to evaluate the effectiveness of the proposed FAN framework in two folds. First, we examine the impact of several key components including the proposed hierarchical agglomeration connection module, the layer-wise hierarchical loss, and other techniques used in our solution. Second, we compare the proposed FAN face detector with the state-of-the-art face detectors on several popular face detection benchmarks and finally evaluate the inference speed of the proposed face detector.

4.1. Model Analysis.

Dataset. We conduct model analysis on the WIDER FACE dataset [29], which has 32,203 images with about 400k faces for a large range of scales. It has three subsets: 40% for training, 10% for validation, and 50% for test. The annotations of training and validation sets are online available. According to the difficulty of detection tasks, it has three splits: Easy, Medium and Hard. The evaluation metric is mean average precision (mAP) with Interception-of-Union (IoU) threshold as 0.5. We train FAN on the training set of WIDER FACE, and evaluate it on the validation set.

Baseline. We adopt the closely related detector S3FD [31] as the baseline to validate the effectiveness of our technique. S3FD achieved the previous state-of-the-art results on several well-known face detection benchmarks. It inherited the standard SSD framework with carefully designed scale-aware anchors according to effective receptive fields. We follow the same experimental setup in [31].

Agglomerative Connection. We first validate the contribution of *Agglomerative Connection* module with different hierarchical levels. When hierarchical level equals to 1, FAN degenerates to Vanilla S3FD. In Table 1, FAN with 2-level Agglomerative Connection outperforms baseline with a large margin in all three difficulty levels. More specifically, we notice the performance of 2-level FAN with single-scale inference is even comparable to S3FD with multi-scale inference, which validates the effectiveness of the agglomerative connection module. As increasing the hierarchical level from 2 to 3, the results become slightly worse than before, but still outperforms the baseline consistently. We argue this is because the complexity of FAN increases as the hierarchical level becomes deeper, traditional learning scheme suffers from the training. Next we will show that Hierarchical Loss is necessary in effectively training FAN with high hierarchical-level Agglomerative Connection.

Hierarchical Loss. We optimize 2-level FAN and 3-level FAN with Hierarchical Loss. In Table 1, the performance of 3-level FAN with Hierarchical Loss gains significant improvement compared with its single loss setting in all difficulty levels (+1.0% in Easy, +1.0% in Medium and +1.1% in Hard), while 2-level FAN with Hierarchical Loss also gains slight improvement. We argue this is because 2-level FAN training is not very difficult so that traditional optimization methods can still handle. In Hierarchical Loss optimization scheme, 3-level FAN outperforms 2-level FAN consistently, which indicates high level Agglomerative Connection is crucial to improve detection accuracy with Hierarchical Loss optimization method.

Robust Feature Learning. We compare agglomerative connection with skip connection which was widely used in building feature pyramids. We build "S3FD w/ FPN" based on S3FD with skip-connection in top-down structure as FPN does [15]. In Table 2, compared with vanilla S3FD, "S3FD w/ FPN" gains improvement in Hard level, which validates the efficacy of feature pyramid for improving shallow features. Our FAN outperforms "S3FD w/ FPN" with large margin, which indicates the superiority of our agglomerative connection over the skip connection.

Moreover, to further validate the robust features learned by FAN, we remove all the Agglomerative Connection Module in 3-level FAN trained with the Hierarchical Loss, which shares the same network structure as the Vanilla S3FD. We use this model ("S3FD w / FAN") to make inference as S3FD does. The results in Table 2 show this "truncated" model achieves a high detection accuracy and outperforms both Vanilla S3FD and "S3FD w/ FPN". This proves that the proposed structure of FAN enables us to learn robust and discriminative features.

Multi-scale Inference. Multi-scale testing is a widely used technique in object detection, which can further boost the detection accuracy especially for small objects. In Table 1, both vanilla S3FD and FAN gain improvements of detection accuracy. We conduct multi-scale method during

	S3FD [31]						FAN (ours)
1-Level Agglomeration	√	√					
2-Level Agglomeration			√		√		
3-Level Agglomeration				√		√	√
Hierarchical Loss?					√	√	√
Multi-scale inference?		√					√
WIDER FACE mAP (Easy)	92.9	93.7	93.8	93.0	93.8	94.0	94.8
WIDER FACE mAP (Medium)	91.8	92.5	92.6	91.9	92.7	92.9	93.8
WIDER FACE mAP (Hard)	83.4	85.9	85.8	85.3	86.0	86.4	87.6

Table 1. Ablation studies of FAN. All settings are trained on the training set of WIDER FACE and then tested on the validation set.

Loss	Easy	Medium	Hard
Vanilla S3FD	92.9	91.8	83.4
S3FD w/ FPN	92.9	91.8	84.7
S3FD w/ FAN	93.4	92.4	85.2
FAN	94.0	92.9	86.4

Table 2. Evaluation of our FAN with agglomerative connection and hierarchical loss (HL) for learning discriminative features in contrast to vanilla S3FD and a simple FPN with skip connection.

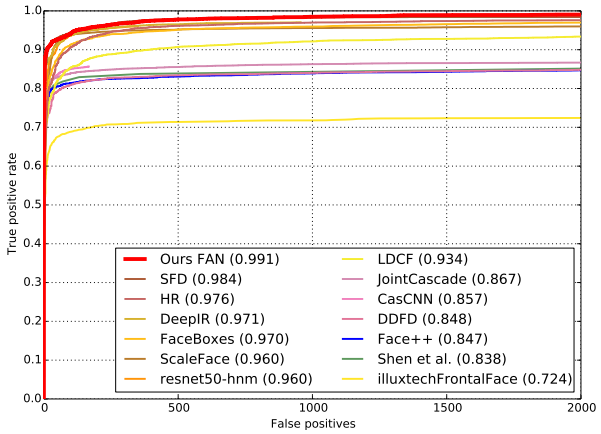
inference with fixing the aspect ratio of images.

Comparisons with the State of the Art. We use a 3-level FAN network trained by the Hierarchical Loss as our final face detector to compare with various state-of-the-art detectors on the WIDER FACE datasets. Figure 5 and Figure 6 show the precision-recall curves and Table 3 summarizes the overall results on the WIDER Face validation set.

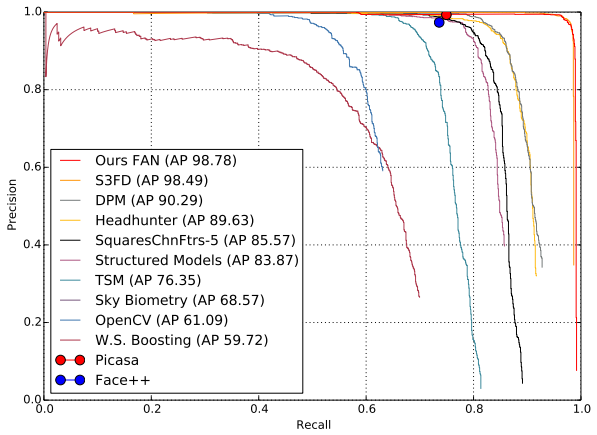
Algorithms	Backbone	Easy	Med	Hard
ACF-WIDER [20]	-	65.9	54.1	27.3
LDCF+ [20]	-	79.0	76.9	52.2
ScaleFace [30]	ResNet50	86.8	86.7	77.2
HR [8]	ResNet101	92.5	91.0	80.6
Face R-FCN* [27]	ResNet101	94.7	93.5	87.4
CMS-RCNN [32]	VGG16	89.9	87.4	62.4
SSD-face [31]	VGG16	92.1	89.5	71.6
RPN-face [31]	VGG16	91.0	88.2	73.7
Face-RCNN* [26]	VGG19	93.7	92.1	83.1
SSH [19]	VGG16	93.1	92.1	84.5
S3FD[31]	VGG16	93.7	92.5	85.9
FAN(ours)	VGG16	**94.8**	**93.8**	**87.6**

Table 3. Evaluation (mAP) on the validation set of WIDER FACE. "*" denote methods of arXiv papers (unpublished).

FAN outperforms all VGG-based detectors with large margin, especially in Hard difficulty level. Compared with ResNet-based detectors which utilize much stronger backbone architecture, FAN still outperforms all submitted results, while enjoying a clear advantage of high-inference speed. WIDER FACE is a very challenging face benchmark and the results strongly prove the effectiveness of FAN in handling high scale variances, especially for small faces.



(a) FDDB Discrete ROC Curves

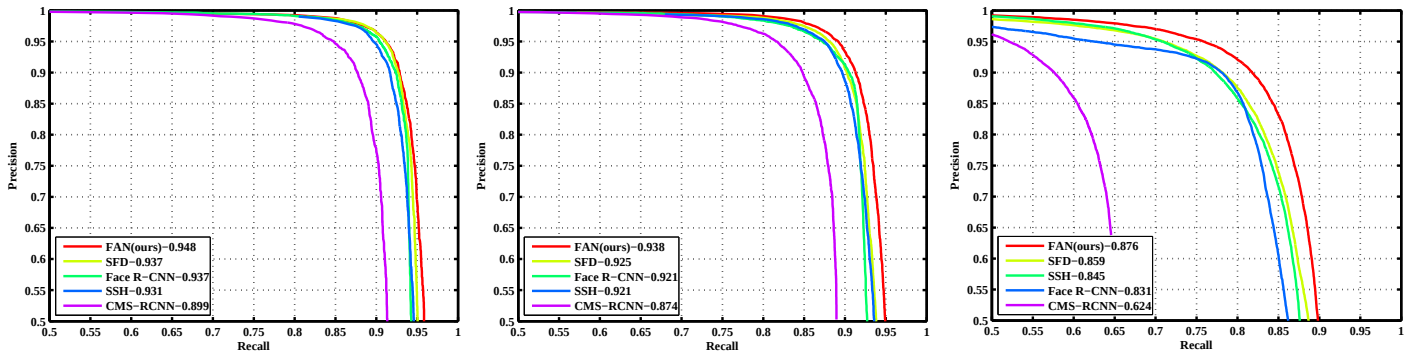


(b) PASCAL Face Results

Figure 4. Evaluation on two popular face detection benchmarks.

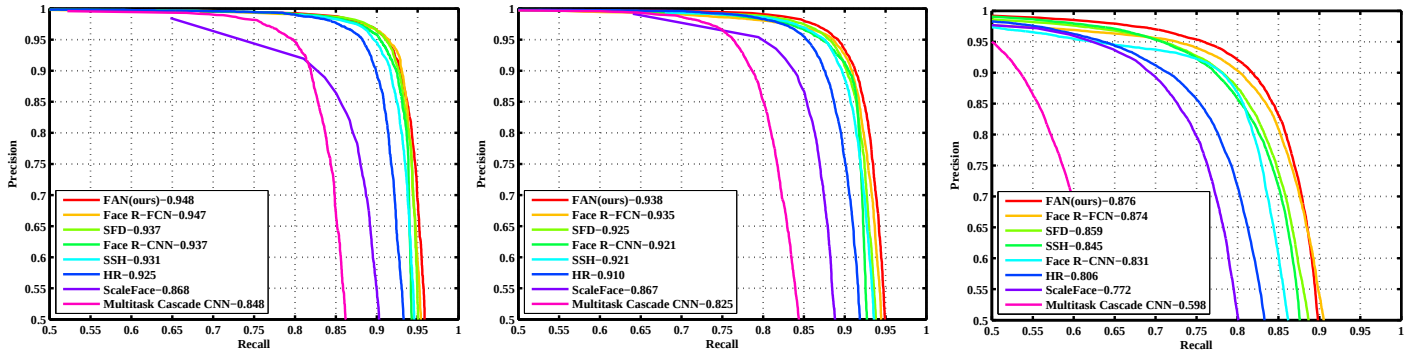
4.2. Evaluation on Other Public Face Benchmarks

FDDB. The Face Detection Data Set and Benchmark (FDDB)[9] is a well-known benchmark with 5,171 faces in 2,845 images. We compare our FAN detector trained on the WIDER FACE training set with other published results on FDDB. Figure 4(a) shows the evaluation results, in which our FAN detector achieves very promising results.



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

Figure 5. Evaluation of VGG-based methods on the validation set of WIDER FACE



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

Figure 6. Evaluation of various state-of-the-art methods on the validation set of WIDER FACE

PASCAL FACE. This dataset was collected from PASCAL person layout test set [3], with 1,335 labeled faces in 851 images. Figure 4(b) shows the evaluation results of the precision-recall curves. Among all the existing methods, the proposed FAN achieved the best mAP (98.78%), which outperforms the previous state-of-the-art detectors S3FD (98.45%) and SSH (98.27%)[1] [19], and significantly beats the other submitted methods [33, 28, 18, 11, 2].

4.3. Inference Time

Our FAN detector is a single-stage detector and thus enjoys high inference speed. FAN runs 32 FPS in GTX 1080ti and CuDNN v5.1 with a VGA-resolution input image. The majority of the time cost is spent on the VGG16 backbone network, while the Agglomerative Connection module is computationally efficient and has little extra cost.

4.4. Qualitative Results

In addition to the above quantitative results, we also observe promising qualitative results on the benchmark databases, including PASCAL FACE[3] as shown in Figure 7, FDDB[9] as shown in Figure 8, and WIDER FACE [29] as shown in Figure 9 and Figure 10.

[1]We cannot plot their curve as their result file is not available.

5. Conclusion

This paper proposed a novel framework of "Feature Agglomeration Networks" (FAN) for building single stage face detectors. The proposed FAN based face detector not only achieves the state-of-the-art performance in various common face detection benchmarks, but also runs very fast and enjoys real-time inference speed on GPU. FAN introduces two key novel components: (i) the proposed feature "Agglomerative Connection" module agglomerates multi-scale features and contextual information by hierarchical structure, which effectively handles scale variance in face detection; and (ii) the proposed Hierarchical Loss allows to train the FAN model robustly in an end-to-end manner. Our future direction is to extend and apply the Feature Agglomeration Networks (FAN) framework for more computer vision tasks, including generic object detection or specialized object detection tasks in other niche domains, such as pedestrian detection, car detection, etc.

Acknowledgements

The authors would like to acknowledge the assistance and collaboration with colleagues from DeepIR Inc.

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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 3

[2] D. Chen, G. Hua, F. Wen, and J. Sun. Supervised transformer network for efficient face detection. In *European Conference on Computer Vision*, 2016. 8

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *Intl. Journal of Computer Vision (IJCV)*, 2010. 8

[4] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017. 3

[5] R. Girshick. Fast r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015. 1, 3, 5

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[7] Z. Hao, Y. Liu, H. Qin, J. Yan, X. Li, and X. Hu. Scale-aware face detection. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[8] P. Hu and D. Ramanan. Finding tiny faces. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 3, 4, 7

[9] 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. 7, 8

[10] 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 22nd ACM intl. conference on Multimedia*, 2014. 6

[11] Z. Kalal, J. Matas, and K. Mikolajczyk. Weighted sampling for large-scale boosting. In *BMVC*, 2008. 8

[12] T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, and Y. Chen. Ron: Reverse connection with objectness prior networks for object detection. *arXiv preprint arXiv:1707.01691*, 2017. 3

[13] T. Kong, A. Yao, Y. Chen, and F. Sun. Hypernet: Towards accurate region proposal generation and joint object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012. 1

[15] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 3, 5, 6

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *European Conference on Computer Vision*, 2016. 1, 2, 3, 5, 6

[17] W. Luo, Y. Li, R. Urtasun, and R. Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2016. 3

[18] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *European Conference on Computer Vision*, 2014. 8

[19] M. Najibi, P. Samangouei, R. Chellappa, and L. Davis. Ssh: Single stage headless face detector. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 7, 8

[20] E. Ohn-Bar and M. M. Trivedi. To boost or not to boost? on the limits of boosted trees for object detection. In *Pattern Recognition (ICPR)*, 2016. 7

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[22] 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*, 2015. 1, 3

[23] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[24] X. Sun, P. Wu, and S. C. Hoi. Face detection using deep learning: An improved faster rcnn approach. *arXiv preprint arXiv:1701.08289*, 2017. 3

[25] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 2004. 1

[26] H. Wang, Z. Li, X. Ji, and Y. Wang. Face r-cnn. *arXiv preprint arXiv:1706.01061*, 2017. 7

[27] Y. Wang, X. Ji, Z. Zhou, H. Wang, and Z. Li. Detecting faces using region-based fully convolutional networks. *arXiv preprint arXiv:1709.05256*, 2017. 7

[28] S. Yang, P. Luo, C.-C. Loy, and X. Tang. From facial parts responses to face detection: A deep learning approach. In *IEEE Intl. Conference on Computer Vision (ICCV)*, 2015. 8

[29] S. Yang, P. Luo, C.-C. Loy, and X. Tang. Wwider face: A face detection benchmark. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5, 6, 8

[30] S. Yang, Y. Xiong, C. C. Loy, and X. Tang. Face detection through scale-friendly deep convolutional networks. *arXiv preprint arXiv:1706.02863*, 2017. 7

[31] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. S3fd: Single shot scale-invariant face detector. In *IEEE Intl. Conference on Computer Vision (ICCV)*, 2017. 2, 3, 5, 6, 7

[32] C. Zhu, Y. Zheng, K. Luu, and M. Savvides. Cms-rcnn: Contextual multi-scale region-based cnn for unconstrained face detection. In *Deep Learning for Biometrics*. 2017. 7

[33] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 8

Figure 7. Qualitative results on Pascal Face. Our model can handle faces in various conditions.

Figure 8. Qualitative results on FDDB. Our model is robust to occlusion and scale variance

Figure 9. Qualitative results on WIDER FACE. Our model is able to handle faces with a wide range of face scales, even with extremely small faces .

Figure 10. Qualitative results on Pascal Face. Our model is robust to blur, occlusion, pose, expression, makeup, illumination, etc.