

Chapter 3

CMS-RCNN: Contextual Multi-Scale Region-Based CNN for Unconstrained Face Detection

Chenchen Zhu, Yutong Zheng, Khoa Luu and Marios Savvides

Abstract Robust face detection in the wild is one of the ultimate components to support various facial related problems, i.e., unconstrained face recognition, facial periocular recognition, facial landmarking and pose estimation, facial expression recognition, 3D facial model construction, etc. Although the face detection problem has been intensely studied for decades with various commercial applications, it still meets problems in some real-world scenarios due to numerous challenges, e.g., heavy facial occlusions, extremely low resolutions, strong illumination, exceptional pose variations, image or video compression artifacts, etc. In this paper, we present a face detection approach named Contextual Multi-Scale Region-based Convolution Neural Network (CMS-RCNN) to robustly solve the problems mentioned above. Similar to the region-based CNNs, our proposed network consists of the region proposal component and the region-of-interest (RoI) detection component. However, far apart of that network, there are two main contributions in our proposed network that play a significant role to achieve the state-of-the-art performance in face detection. First, the multi-scale information is grouped both in region proposal and RoI detection to deal with tiny face regions. Second, our proposed network allows explicit body contextual reasoning in the network inspired from the intuition of human vision system. The proposed approach is benchmarked on two recent challenging face detection databases, i.e., the WIDER FACE Dataset which contains high degree of variability, as well as the Face Detection Dataset and Benchmark (FDDB). The experimental results show that our proposed approach trained on WIDER FACE

C. Zhu (✉) · Y. Zheng (✉) · K. Luu · M. Savvides
CyLab Biometrics Center and the Department of Electrical and Computer Engineering,
Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: chenchez@andrew.cmu.edu

Y. Zheng
e-mail: yutongzh@andrew.cmu.edu

K. Luu
e-mail: kluu@andrew.cmu.edu

M. Savvides
e-mail: msavvid@ri.cmu.edu

Dataset outperforms strong baselines on WIDER FACE Dataset by a large margin, and consistently achieves competitive results on FDDB against the recent state-of-the-art face detection methods.

3.1 Introduction

Detection and analysis on human subjects using facial feature based biometrics for access control, surveillance systems, and other security applications have gained popularity over the past few years. Several such biometrics systems are deployed in security checkpoints across the globe with more being deployed every day. Particularly, face recognition has been one of the most popular biometrics modalities attractive to security departments. Indeed, the uniqueness of facial features across individuals can be captured much more easily than other biometrics. In order to take into account a face recognition algorithm, however, face detection usually needs to be done first.

The problem of face detection has been intensely studied for decades with the aim of ensuring the generalization of robust algorithms to unseen face images [2–13]. Although the detection accuracy in recent face detection algorithms [14–19] has been highly improved due to the advancement of deep Convolutional Neural Networks (CNN), they are still far from achieving the same detection capabilities as a human due to a number of challenges in practice. For example, off-angle faces, large occlusions, low-resolutions and strong lighting conditions, as shown in Fig. 3.1, are always the important factors that need to be considered.

This paper presents an advanced CNN-based approach named Contextual Multi-Scale Region-based CNN (CMS-RCNN) to handle the problem of face detection in digital face images collected under numerous challenging conditions, e.g., heavy facial occlusion, illumination, extreme off-angle, low resolution, scale difference, etc. Our designed region-based CNN architecture allows the network to simultaneously look at multi-scale features, as well as to explicitly look outside facial regions as the potential body regions. In other words, this process tries to mimic the way of face detection by human in a sense that when humans are not sure about a face, seeing the body will increase our confidence. Additionally this architecture also helps to synchronize both the global semantic features in high-level layers and the localization features in low-level layers for facial representation. Therefore, it is able to robustly deal with the challenges in the problem of unconstrained face detection.

Our CMS-RCNN method introduces the Multi-Scale Region Proposal Network (MS-RPN) to generate a set of region candidates and the Contextual Multi-Scale Convolution Neural Network (CMS-CNN) to do inference on the region candidates of facial regions. A confidence score and bounding-box regression are computed for every candidate. In the end, the face detection system is able to decide the quality of the detection results by thresholding these generated confidence scores in given face images. The architecture of our proposed CMS-RCNN network for unconstrained face detection is illustrated in Fig. 3.2.



Fig. 3.1 An example of face detection results using our proposed CMS-RCNN method. The proposed method can robustly detect faces across occlusion, facial expression, pose, illumination, scale and low-resolution conditions from WIDER FACE Dataset [1]

Our approach is evaluated on two challenging face detection databases and compared against numerous recent face detection methods. First, the proposed CMS-RCNN method is compared against four strong baselines [1, 11, 16] on the WIDER FACE Dataset [1], a large-scale face detection benchmark database. This experiment shows its capability to detect face images in the wild, e.g., under occlusions, illumination, facial poses, low-resolution conditions, etc. Our method outperforms the baselines by a huge margin in all easy, medium, and hard partitions. It is also benchmarked on the Face Detection Data Set and Benchmark (FDDB) [20], a dataset of face regions designed for studying the problem of unconstrained face detection. The experimental results show that the proposed CMS-RCNN approach consistently achieves highly competitive results against the other state-of-the-art face detection methods.

The rest of this paper is organized as follows. In Sect. 3.2, we summarize prior work in face detection. Section 3.3 reviews a general deep learning framework, the background as well as the limitations of the Faster R-CNN in the problem of face detection. In Sect. 3.4, we introduce our proposed CMS-RCNN approach for the problem of unconstrained face detection. Section 3.5 presents the experimental face detection results and comparisons obtained using our proposed approach on two challenging face detection databases, i.e., the WIDER FACE Dataset and the FDDB database. Finally, our conclusions in this work are presented in Sect. 3.6.

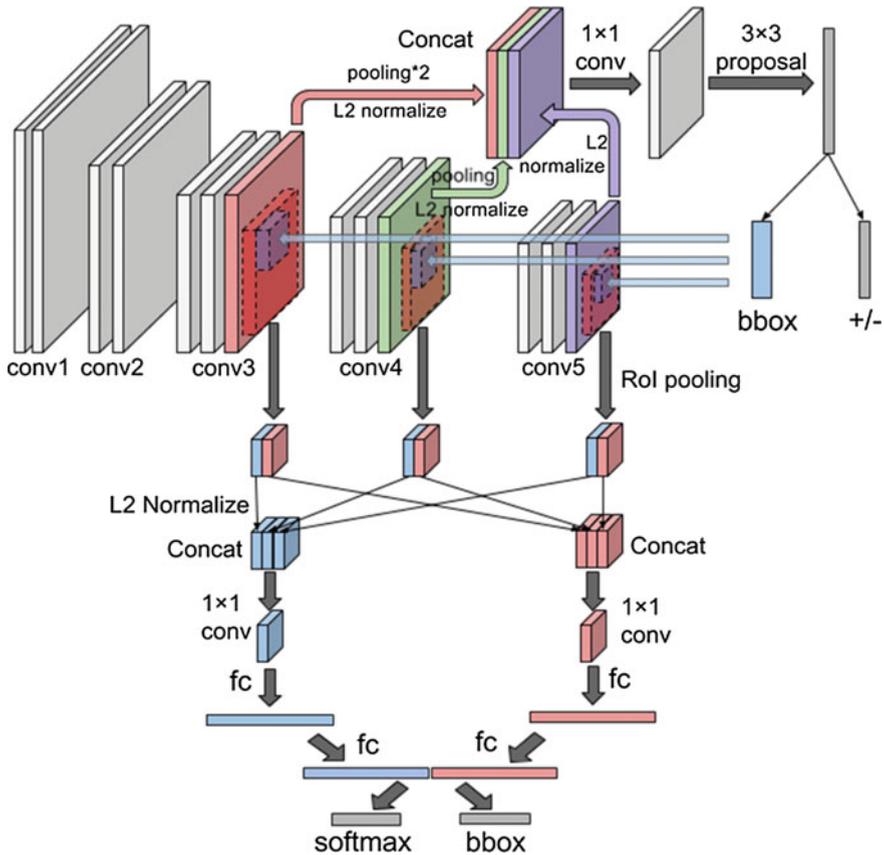


Fig. 3.2 Our proposed Contextual Multi-Scale Region-based CNN model. It is based on the VGG-16 model [21], with five sets of convolution layers in the middle. The upper part is the Multi-Scale Region Proposal Network (MS-RPN) and the lower part is the Contextual Multi-Scale Convolution Neural Network (CMS-CNN). In the CMS-CNN, the face features labeled as *blue blocks* and the body context features labeled as *red blocks* are processed in parallel and combined in the end for final outputs, i.e., confidence score and bounding box

3.2 Related Work

Face detection has been a well-studied area of computer vision. One of the first well-performing approaches to the problem was the Viola–Jones face detector [2]. It was capable of performing real-time face detection using a cascade of boosted simple Haar classifiers. The concepts of boosting and using simple features has been the basis for many different approaches [3] since the Viola–Jones face detector. These early detectors tended to work well on frontal face images but not very well on faces in different poses. As time has passed, many of these methods have been able to deal with off-angle face detection by utilizing multiple models for the various poses

of the face. This increases the model size but does afford more practical uses of the methods. Some approaches have moved away from the idea of simple features but continued to use the boosted learning framework. Li and Zhang [5] used SURF cascades for general object detection but also showed good results on face detection.

More recent work on face detection has tended to focus on using different models such as a Deformable Parts Model (DPM) [4, 22]. Zhu and Ramanan's work was an interesting approach to the problem of face detection where they combined the problems of face detection, pose estimation, and facial landmarking into one framework. By utilizing all three aspects in one framework, they were able to outperform the state of the art at the time on real-world images. Yu et al. [23] extended this work by incorporating group sparsity in learning which landmarks are the most salient for face detection as well as incorporating 3D models of the landmarks in order to deal with pose. Chen et al. [10] have combined ideas from both of these approaches by utilizing a cascade detection framework while simultaneously localizing features on the face for alignment of the detectors. Similarly, Ghiasi and Fowlkes [12] have been able to use heirarchical DPMs not only to achieve good face detection in the presence of occlusion but also landmark localization. However, Mathias et al. [9] were able to show that both DPM models and rigid template detectors similar to the Viola-Jones detector have a lot of potential that has not been adequately explored. By retraining these models with appropriately controlled training data, they were able to create face detectors that perform similarly to other, more complex state-of-the-art face detectors.

All of these approaches to face detection were based on selecting a feature extractor beforehand. However, there has been work done in using a ConvNet to learn which features are used to detect faces. Neural Networks have been around for a long time but have been experiencing a resurgence in popularity due to hardware improvements and new techniques resulting in the capability to train these networks on large amounts of training data. Li et al. [14] utilized a cascade of CNNs to perform face detection. The cascading networks allowed them to process different scales of faces at different levels of the cascade while also allowing for false positives from previous networks to be removed at later layers in a similar approach to other cascade detectors. Yang et al. [16] approached the problem from a different perspective more similar to a DPM approach. In their method, the face is broken into several facial parts such as hair, eyes, nose, mouth, and beard. By training a detector on each part and combining the score maps intelligently, they were able to achieve accurate face detection even under occlusions. Both of these methods require training several networks in order to achieve their high accuracy. Our method, on the other hand, can be trained as a single network, end to end, allowing for less annotation of training data needed while maintaining highly accurate face detection.

The ideas of using contextual information in object detection have been studied in several recent work with very high detection accuracy. Divvala et al. [24] reviewed the role of context in a contemporary, challenging object detection in their empirical evaluation analysis. In their conclusions, the context information not only reduces the overall detection errors, but also the remaining errors made by the detector are more reasonable. Bell et al. [25] introduced an advanced object detector method named

Inside-Outside Network (ION) to exploit information both inside and outside the region of interest. In their approach, the contextual information outside the region of interest is incorporated using spatial recurrent neural networks. Inside the network, skip pooling is used to extract information at multiple scales and levels of abstraction. Recently, Zagoruyko et al. [26] have presented the MultiPath network with three modifications to the standard Fast R-CNN object detector, i.e., skip connections that give the detector access to features at multiple network layers, a foveal structure to exploit object context at multiple object resolutions, and an integral loss function and corresponding network adjustment that improve localization. The information in their proposed network can flow along multiple paths. Their MultiPath network is combined with DeepMask object proposals to solve the object detection problem.

Unlike all the previous approaches that select a feature extractor beforehand and incorporate a linear classifier with the depth descriptor beside RGB channels, our method solves the problem under a deep learning framework where the global and the local context features, i.e., multi scaling, are synchronized to Faster Region-based Convolutional Neural Networks in order to robustly achieve semantic detection.

3.3 Background in Deep Convolution Nets

The recent studies in deep ConvNets have achieved significant results in object detection, classification and modeling [27]. In this section, we review various well-known Deep ConvNets. Then, we show the current limitations of the Faster R-CNN, one of the state-of-the-art deep ConvNet methods in object detection, in the defined context of the face detection.

3.3.1 *Region-Based Convolution Neural Networks*

One of the most important approaches for the object detection task is the family of Region-based Convolution Neural Networks (R-CNN).

R-CNN [28], the first generation of this family, applies the high-capacity deep ConvNet to classify given bottom-up region proposals. Due to the lack of labeled training data, it adopts a strategy of supervised pretraining for an auxiliary task followed by domain-specific fine-tuning. Then the ConvNet is used as a feature extractor and the system is further trained for object detection with Support Vector Machines (SVM). Finally, it performs bounding-box regression. The method achieves high accuracy but is very time-consuming. The system takes a long time to generate region proposals, extract features from each image, and store these features in a hard disk, which also takes up a large amount of space. At testing time, the detection process takes 47 s per image using VGG-16 network [21] implemented in GPU due to the slowness of feature extraction. In other words, R-CNN is slow because it processes each object proposal independently without sharing computation.

Fast R-CNN [29] solves this problem by sharing the features between proposals. The network is designed to only compute a feature map once per image in a fully convolutional style, and to use ROI pooling to dynamically sample features from the feature map for each object proposal. The network also adopts a multitask loss, i.e., classification loss and bounding-box regression loss. Based on the two improvements, the framework is trained end to end. The processing time for each image significantly reduced to 0.3 s. Fast R-CNN accelerates the detection network using the ROI pooling layer. However the region proposal step is designed out of the network hence still remains a bottleneck, which results in suboptimal solution and dependence on the external region proposal methods.

Faster R-CNN [30] addresses the problem with fast R-CNN by introducing the Region Proposal Network (RPN). An RPN is implemented in a fully convolutional style to predict the object bounding boxes and the objectness scores. In addition, the anchors are defined with different scales and ratios to achieve the translation invariance. The RPN shares the full-image convolution features with the detection network. Therefore the whole system is able to complete both proposal generation and detection computation within 0.2 s using very deep VGG-16 model [21]. With a smaller ZF model [31], it can reach the level of real-time processing.

3.3.2 *Limitations of Faster R-CNN*

The Region-based CNN family, e.g., Faster R-CNN and its variants [29], achieves the state-of-the-art performance results in object detection on the PASCAL VOC dataset. These methods can detect objects such as vehicles, animals, people, chairs, etc., with very high accuracy. In general, the defined objects often occupy the majority of a given image. However, when these methods are tested on the challenging Microsoft COCO dataset [32], the performance drops a lot, since images contain more small, occluded, and incomplete objects. Similar situations happen in the problem of face detection. We focus on detecting only facial regions that are sometimes small, heavily occluded and of low resolution (as shown in Fig. 3.1).

The detection network in designed Faster R-CNN is unable to robustly detect such tiny faces. The intuition point is that the regions of interest pooling layer, i.e., ROI pooling layer, builds features only from the last single high-level feature map. For example, the global stride of the ‘conv5’ layer in the VGG-16 model is 16. Therefore, given a facial region with the sizes less than 16×16 pixels in an image, the projected ROI pooling region for that location will be less than 1 pixel in the ‘conv5’ layer, even if the proposed region is correct. Thus, the detector will have much difficulty to predict the object class and the bounding-box location based on information from only one pixel.

3.3.3 *Other Face Detection Method Limitations*

Other challenges in object detection in the wild include occlusion and low resolution. For face detection, it is very common for people to wear stuffs like sunglasses, scarf and hats, which occlude the face. In such cases, the methods that only extract features from faces do not work well. For example, Faceness [16] consider finding faces through scoring facial parts responses by their spatial structure and arrangement, which works well on clear faces. But when facial parts are missing due to occlusion or when face itself is too small, facial parts become more hard to detect. Therefore, the body context information plays its role. As an example of context-dependent objects, faces often come together with human body. Even though the faces are occluded, we can still locate it only by seeing the whole human body. Similar advantages for faces at low resolution, i.e., tiny faces. The deep features cannot tell much about tiny faces since their receptive field is too small to be informative. Introducing context information can extend the area to extract features and make them meaningful. On the other hand, the context information also helped with reducing false detection as discussed previously, since context information tells the difference between real faces with bodies and face-like patterns without bodies.

3.4 Contextual Multi-Scale R-CNN

Our goal is to detect human faces captured under various challenging conditions such as strong illumination, heavily occlusion, extreme off-angles, and low resolution. Under these conditions, the current CNN-based detection systems suffer from two major problems, i.e., (1) tiny faces are hard to identify; (2) only face region is taken into consideration for classification. In this section, we show why these problems hinder the ability of a face detection system. Then, our proposed network is presented to address these problems by using the Multi-Scale Region Proposal Network (MS-RPN) and the Contextual Multi-Scale Convolution Neural Network (CMS-CNN), as illustrated in Fig. 3.2. Similar to Faster R-CNN, the MS-RPN outputs several region candidates and the CMS-CNN computes the confidence score and bounding box for each candidate.

3.4.1 *Identifying Tiny Faces*

Why tiny faces are hard to be robustly detected by the previous region-based CNNs? The reason is that in these networks both the proposed region and the classification score are produced from one single high-level convolution feature map. This representation does not have enough information for the multiple tasks, i.e., region proposal and RoI detection. For example, Faster R-CNN generates region candidates

and does RoI pooling from the ‘conv5’ layer of the VGG-16 model, which has a overall stride of 16. One issue is that the reception field in this layer is quite large. When the face size is less than 16-by-16 pixels, the corresponding output in ‘conv5’ layer is less than 1 pixel, which is insufficient to encode informative features. The other issue is that as the convolution layers go deeper, each pixel in the feature map gather more and more information outside the original input region so that it contains lower proportion of information for the region of interest. These two issues together make the last convolution layer less representative for tiny faces.

3.4.1.1 Multiple Scale Faster-RCNN

Our solution for this problem is a combination of both global and local features, i.e., multiple scales. In this architecture, the feature maps are incorporated from lower level convolution layers with the last convolution layer for both MS-RPN and CMS-CNN. Features from lower convolution layer help get more information for the tiny faces, because stride in lower convolution layer will not be too small. Another benefit is that both low-level feature with localization capability and high-level feature with semantic information are fused together [33], since face detection needs to localize the face as well as to identify the face. In the MS-RPN, the whole lower level feature maps are down-sampled to the size of high-level feature map and then concatenated with it to form a unified feature map. Then we reduce the dimension of the unified feature map and use it to generate region candidates. In the CMS-CNN, the region proposal is projected into feature maps from multiple convolution layers. And RoI pooling is performed in each layer, resulting in a fixed-size feature tensor. All feature tensors are normalized, concatenated and dimension-reduced to a single feature blob, which is forwarded to two fully connected layers to compute a representation of the region candidate.

3.4.1.2 L2 Normalization

In both MS-RPN and CMS-CNN, concatenation of feature maps is done with L2 normalization layer [34], shown in Fig. 3.2, since the feature maps from different layer have generally different properties in terms of numbers of channels, scale of value and norm of feature map pixels. Generally, comparing with values in shallower layers, the values in deeper layers are usually too small, which leads to the dominance of shallower layers. In practice, it is impossible for the system to readjust and tune value from each layer for best performance. Therefore, L2 normalization layers before concatenation are crucial for the robustness of the system because it keeps the value from each layer in roughly the same scale.

The normalization is performed within each pixel, and all feature map is treated independently

$$\hat{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2} \|\mathbf{x}\|_2 = \left(\sum_{i=1}^d |x_i| \right)^{\frac{1}{2}} \quad (3.1)$$

where the \mathbf{x} and $\hat{\mathbf{x}}$ stand for the original pixel vector and the normalized pixel vector respectively. d stands for the number of channels in each feature map tensor.

During training, scaling factors γ_i will be updated to readjust the scale of the normalized features. For each channel i , the scaling factor follows:

$$y_i = \gamma_i \hat{x}_i \quad (3.2)$$

where y_i stand for the rescaled feature value.

Following the backpropagation and chain rule, the update for scaling factor γ is

$$\frac{\partial l}{\partial \hat{\mathbf{x}}} = \frac{\partial l}{\partial \mathbf{y}} \cdot \gamma \frac{\partial l}{\partial \mathbf{x}} = \frac{\partial l}{\partial \hat{\mathbf{x}}} \left(\frac{\mathbf{I}}{\|\mathbf{x}\|_2} - \frac{\mathbf{xx}^T}{\|\mathbf{x}\|_2^3} \right) \frac{\partial l}{\partial \gamma_i} = \sum_{y_i} \frac{\partial l}{\partial y_i} \hat{x}_i \quad (3.3)$$

where $\mathbf{y} = [y_1, y_2, \dots, y_d]^T$.

3.4.1.3 New Layer in Deep Learning Caffe Framework

The system integrate information from lower layer feature maps, i.e., third and fourth convolution layers, to extract determinant features for tiny faces. For both parts of our system, i.e., MS-RPN and CMS-CNN, the L2 normalization layers are inserted before concatenation of feature maps from the three layers. The features were rescaled to proper values and concatenated to a single feature map. We set the initial scaling factor in a special way, following two rules. First, the average scale for each feature map is roughly identical; second, after the following 1×1 convolution, the resulting tensor should have the same average scale as the conv5 layer in the work of Faster R-CNN. As implied, after the following 1×1 convolution, the tensor should be the same as the original architecture in Faster R-CNN, in terms of its size, scale of values, and function for the downstream process.

3.4.2 Integrating Body Context

When humans are searching for faces, they try to look for not only the facial patterns, e.g., eyes, nose, mouth, but also the human bodies. Sometimes a human body makes us more convinced about the existence of a face. In addition, sometimes human body

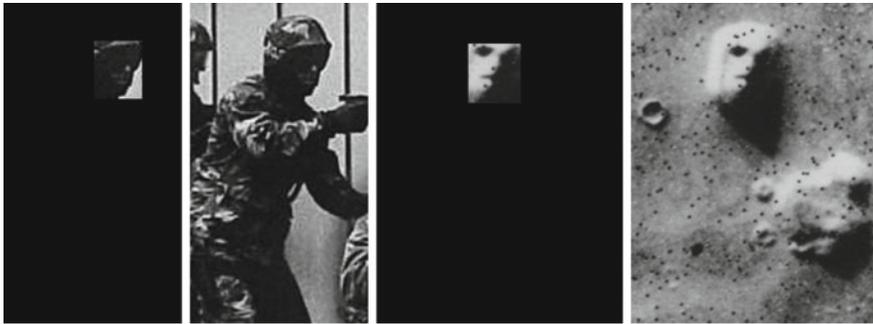


Fig. 3.3 Examples of body context helping face identification. The first two figures show that existence of a body can increase the confidence of finding a face. The last two figures show that what looks like a face turns out to be a mountain on the planet surface when we see more context information

helps to reject false positives. If we only look at face regions, we may make mistakes identifying them. For example, Fig. 3.3 shows two cases where body region plays a significant role for correct detection. This intuition is not only true for human but also valid in computer vision. Previous research has shown that contextual reasoning is a critical piece of the object recognition puzzle, and that context not only reduces the overall detection errors, but, more importantly, the remaining errors made by the detector are more reasonable [24]. Based on this intuition, our network is designed to make explicit reference to the human body context information in the RoI detection.

In our proposed network, the contextual body reasoning is implemented by explicitly grouping body information from convolution feature maps shown as the red blocks in Fig. 3.2. Specifically, additional RoI pooling operations are performed for each region proposal in convolution feature maps to represent the body context features. Then same as the face feature tensors, these body feature tensors are normalized, concatenated, and dimension-reduced to a single feature blob. After two fully connected layers the final body representation is concatenated with the face representation. They together contribute to the computation of confidence score and bounding-box regression.

With projected region proposal as the face region, the additional RoI pooling region represents the body region and satisfies a predefined spatial relation with the face region. In order to model this spatial relation, we make a simple hypothesis that if there is a face, there must exist a body, and the spatial relation between each face and body is fixed. This assumption may not be true all the time but should cover most of the scenarios since most people we see in the real world are either standing or sitting. Therefore, the spatial relation is roughly fixed between the face and the vertical body.

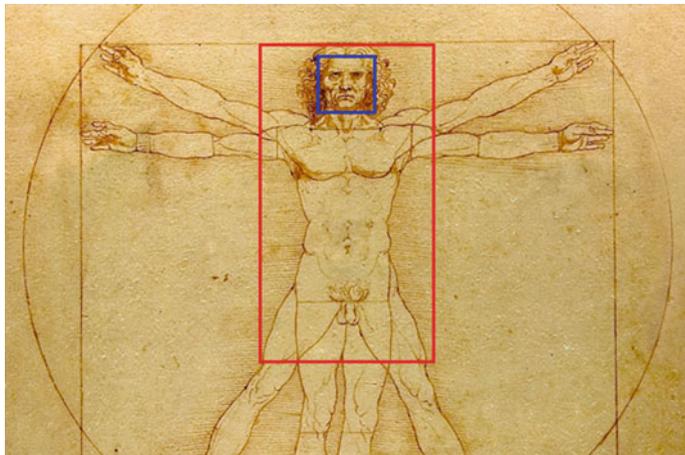


Fig. 3.4 The Vitruvian Man: spatial relation between the face (*blue box*) and the body (*red box*)

Mathematically, this spatial relation can be represented by four parameters presented in Eq. 3.4.

$$t_x = (x_b - x_f)/w_f t_y = (y_b - y_f)/h_f t_w = \log(w_b/w_f) t_h = \log(h_b/h_f) \quad (3.4)$$

where $x_{(*)}$, $y_{(*)}$, $w_{(*)}$, and $h_{(*)}$ denote the two coordinates of the box center, width, and height respectively. And b and f stand for body and face respectively. t_x , t_y , t_w , and t_h are the parameters. Through out this paper, we fix the for parameters such that the two projected RoI regions of face and body satisfies a certain spatial ratio illustrated in the famous drawing in Fig. 3.4.

3.4.3 Information Fusion

It's worth noticing that in our deep network architecture we have multiple face feature maps and body context feature maps for each proposed region. A critical issue is how we effectively fuse these information, i.e., what computation to apply and in which stage.

In our network, features extracted from different convolution layers need to be fused together to get a uniform representation. They cannot be naively concatenated due to the overall differences of the numbers of channels, scales of values, and norms of feature map pixels among these layers. The detailed research shows that the deeper layers often contain smaller values than the shallower layers. Therefore, the larger values will dominate the smaller ones, making the system rely too much on shallower features rather than a combination of multiple scale features causing the system to no

longer be robust. We adopt the normalization layer from [34] to address this problem. The system takes the multiple scale features and apply L2 normalization along the channel axis of each feature map. Then, since the channel size is different among layers, the normalized feature map from each layer needed to be reweighted, so that their values are at the same scale. After that, the feature maps are concatenated to one single feature map tensor. This modification helps to stabilize the system and increase the accuracy. Finally, the channel size of the concatenated feature map is shrunk to fit right in the original architecture for the downstream fully connected layers.

Another crucial question is whether to fuse the face information and the body information at a early stage or at the very end of the network. Here we choose the late fusion strategy in which face features and body context features are extracted in two parallel pipelines. At the very end of the network two representations for face and body context are concatenated together to form a long feature vector. Then this feature vector is forwarded to compute confidence score and bounding-box regression. The other strategy is the early fusion, in which face feature maps and body context feature maps get concatenated right after RoI pooling and normalization. These two strategies combine the information from face and body context, but we prefer the late fusion. The reason is that we want the network to make decisions in a more semantic space. We care more about the existence of the face and the body. The localization information is already encoded in the predefined spatial relation mentioned in Sect. 3.4.2. Moreover empirical experiments also show that late fusion strategy works better.

3.4.4 *Implementation Details*

Our CMS-RCNN is implemented in the Caffe deep learning framework [35]. The first five sets of convolution layers have the same architecture as the deep VGG-16 model, and during training their parameters are initialized from the pretrained VGG-16. For simplicity we refer to the last convolution layers in set 3, 4 and 5 as ‘conv3’, ‘conv4’, and ‘conv5’ respectively. All the following layers are connected exclusively to these three layers. In the MS-RPN, we want ‘conv3’, ‘conv4’, and ‘conv5’ to be synchronized to the same size so that concatenation can be applied. So ‘conv3’ is followed by pooling layer to perform down-sampling. Then ‘conv3’, ‘conv4’, and ‘conv5’ are normalized along the channel axis to a learnable reweighting scale and concatenated together. To ensure training convergence, the initial reweighting scale needs to be carefully set. Here we set the initial scale of ‘conv3’, ‘conv4’, and ‘conv5’ to be 66.84, 94.52, and 94.52 respectively. In the CMS-CNN, the RoI pooling layer already ensure that the pooled feature maps have the same size. Again we normalize the pooled features to make sure the downstream values are at reasonable scales when training is initialized. Specifically, features pooled from ‘conv3’, ‘conv4’, and ‘conv5’ are initialized with scale to be 57.75, 81.67, and 81.67 respectively, for both face and body pipelines. The MS-RPN and the CMS-CNN share the same parameters

for all convolution layers so that computation can be done once, resulting in higher efficiency. Additionally, in order to shrink the channel size of the concatenated feature map, a 1×1 convolution layer is then employed. Therefore the channel size of final feature map is at the same size as the original fifth convolution layer in Faster R-CNN.

3.5 Experiments

This section presents the face detection benchmarking using our proposed CMS-RCNN approach on the WIDER FACE dataset [1] and the Face Detection Data Set and Benchmark (FDDB) [20] database. The WIDER FACE dataset is experimented with high degree of variability. Using this database, our proposed approach robustly outperforms strong baseline methods, including Two-stage CNN [1], Multi-scale Cascade CNN [1], Faceness [16] and Aggregate Channel Features (ACF) [11], by a large margin. We also show that our model trained on WIDER FACE dataset generalizes well enough to the FDDB database. The trained model consistently achieves competitive results against the recent state-of-the-art face detection methods on this database, including HyperFace [19], DP2MFD [17], CCF [18], Faceness [16], NPDFace [13], MultiresHPM [12], DDFD [15], CascadeCNN [14], ACF-multiscale [11], Pico [7], HeadHunter [9], Joint Cascade [10], Boosted Exemplar [8], and PEP-Adapt [6].

3.5.1 Experiments on WIDER FACE Dataset

Data Description

WIDER FACE is a public face detection benchmark dataset. 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. However, in the testing set, only the testing images (not the ground truth labels) are available online. All detection results are sent to the database server for evaluating and receiving the Precision-Recall curves.

In our experiments, the proposed CMS-RCNN 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.

With Context Versus Without Context

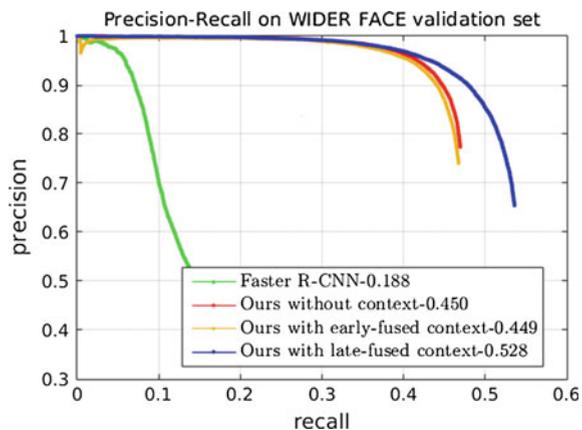
As we show in Sect. 3.4.2 that human vision can benefit from additional context information for better detection and recognition, we show in this section how does explicit contextual reasoning in the network help improve the model performance.

To prove this, we test our models with and without body context information on the validation set of WIDER FACE dataset. The model without body context is implemented by removing the context pipeline and only use the representation from face pipeline to compute the confidence score and the bounding-box regression. We compare their performances as illustrated in Fig. 3.5. The Faster R-CNN method is setup as a baseline. We also compare our models with different strategies of information fusion discussed in Sect. 3.4.3, which we call early-fused context and late-fused context respectively.

Experiment results show that the model with late-fused context has significant improvement over the one without context, which means that they can handle the challenging conditions better. Moreover, we observe that early-fused context dose help detecting more faces because its curve is longer than the one without context. However, early-fused context gets fired on too many false positives which results in low AP score and a drop in the beginning of the curve. Figure 3.6 illustrates the false positives compared to late-fused context. This is probably because early-fused context concatenates all feature maps at a very early stage, resulting in a high-dimensional feature map. This can cause over-fitting due to the curse of dimension. Therefore, we fix our model to be late-fused context for other experiments.

To find out the minimum resolution at which our proposed network can detect a face, we set the threshold such that our detector can reach 90% precision in Fig. 3.5. Then we collect all the correct faces detected by our method in the WIDER FACE validation set. It turns out the minimum size of faces we can detect is around 10×10 pixels.

Fig. 3.5 Precision-Recall curves on the WIDER FACE validation set. The baseline (green curve) is generated by the Faster R-CNN [30] model trained on WIDER FACE training set. We show that our model without context (red curve) outperforms baseline by a large gap. With late-fused context, the performance gets boosted even further (blue curve). The numbers in the legend are the average precision values



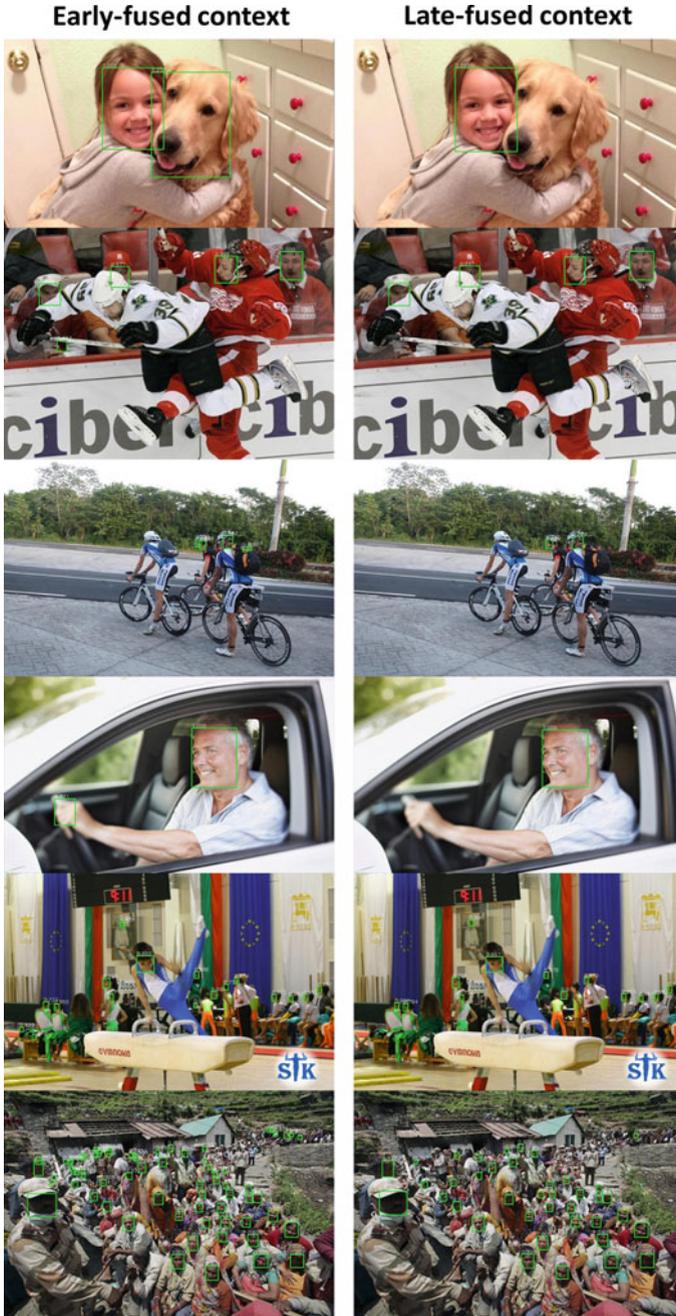


Fig. 3.6 Comparison between early-fused context and late-fused context on some examples in WIDER FACE validation set. It is clear that early-fused context is easier getting fired on some false positives, resulting in a lower AP score

Testing and Comparison

During the testing phase, the face images in the testing set are divided into three parts based on their detection rates on EdgeBox [36]. In other words, face images are divided into three levels according to the difficulties of the detection, i.e., Easy, Medium, and Hard [1]. The proposed CMS-RCNN model is compared against recent strong face detection methods, i.e., Two-stage CNN [1], Multi-scale Cascade CNN [1], Faceness [16], and Aggregate Channel Features (ACF) [11]. All these methods are trained on the same training set and tested on the same testing set.

The Precision-Recall curves and AP values are shown in Fig. 3.7. Our method outperforms those strong baselines by a large margin. It achieves the best average precision in all level faces, i.e., AP = 0.902 (Easy), 0.874 (Medium) and 0.643 (Hard),

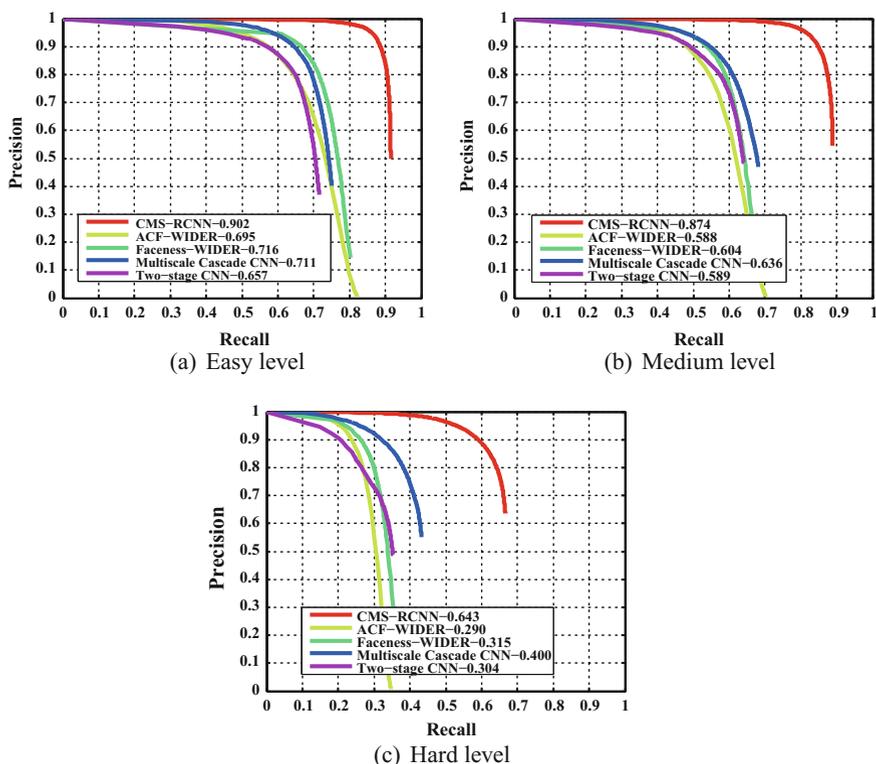


Fig. 3.7 Precision-Recall curves obtained by our proposed CMS-RCNN (*red*) and the other baselines, i.e. Two-stage CNN [1], Multi-scale Cascade CNN [1], Faceness [16], and Aggregate Channel Features (ACF) [11]. All methods trained and tested on the same training and testing set of the WIDER FACE dataset. **a** Easy level, **b** Medium level, and **c** Hard level. Our method achieves the state-of-the-art results with the highest AP values of 0.902 (Easy), 0.874 (Medium), and 0.643 (Hard) among the methods on this database. It also outperforms the second best baseline by 26.0% (Easy), 37.4% (Medium) and 60.8% (Hard)



Fig. 3.8 Some examples of face detection results using our proposed CMS-RCNN method on WIDER FACE database [1]

and outperforms the second best baseline by 26.0% (Easy), 37.4% (Medium), and 60.8% (Hard). These results suggest that as the difficulty level goes up, CMS-RCNN can detect challenging faces better. So it has the ability to handle difficult conditions hence is more closed to human detection level. Figure 3.8 shows some examples of face detection results using the proposed CMS-RCNN on this database.

Visualization of False Positives

As it is well known that precision-recall curves get dropped due to the false positives, we are interested in the false positives produced by our CMS-RCNN 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 the CMS-RCNN 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 Fig. 3.9. 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, which is a problem of the dataset itself. 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.

3.5.2 Experiments on Fddb Face Database

To show that our method generalizes well to other database, the proposed CMS-RCNN is also benchmarked on the Fddb database [20]. It is a standard database for testing and evaluation of face detection algorithms. It contains annotations for



Fig. 3.9 Examples of the top 20 false positives from our CMS-RCNN 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

5,171 faces in a set of 2,845 images taken from the Faces in the Wild dataset. Most of the images in the Fddb database contain less than three faces that are clear or slightly occluded. The faces generally have large sizes and high resolutions compared to WIDER FACE. We use the same model trained on WIDER FACE training set presented in Sect. 3.5.1 to perform the evaluation on the Fddb database.

The evaluation is performed based on the discrete criterion following the same rules in PASCAL VOC Challenge [37], i.e., if the ratio of the intersection of a detected region with an annotated face region is greater than 0.5, it is considered as a true positive detection. The evaluation is proceeded following the Fddb evaluation protocol and compared against the published methods provided in the protocol, i.e. HyperFace [19], DP2MFD [17], CCF [18], Faceness [16], NPdFace [13], MultiresHPM [12], DDFD [15], CascadeCNN [14], ACF-multiscale [11], Pico [7], HeadHunter

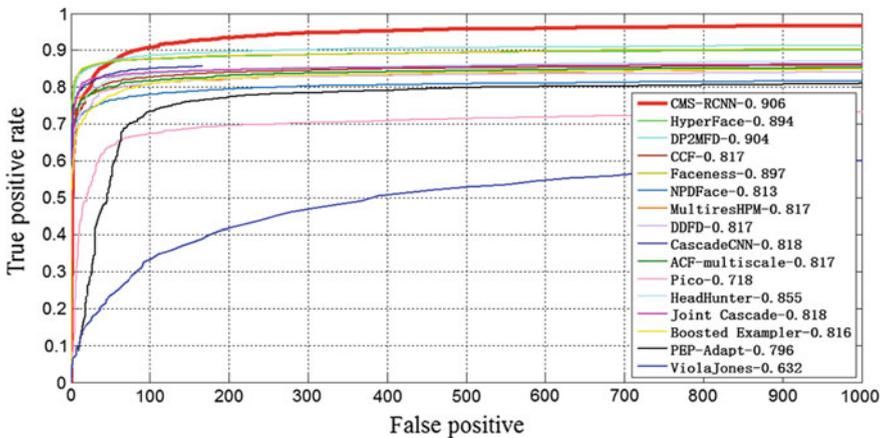


Fig. 3.10 ROC curves of our proposed CMS-RCNN and the other published methods on Fddb database [20]. Our method achieves the best recall rate on this database. Numbers in the legend show the average precision scores



Fig. 3.11 Some examples of face detection results using our proposed CMS-RCNN method on FDDB database [20]

[9], Joint Cascade [10], Boosted Exemplar [8], and PEP-Adapt [6]. The proposed CMS-RCNN approach outperforms most of the published face detection methods and achieves a very high recall rate comparing against all other methods (as shown Fig. 3.10). This is concrete evidence to demonstrate that CMS-RCNN robustly detects unconstrained faces. Figure 3.11 shows some examples of the face detection results using the proposed CMS-RCNN on the FDDB dataset.

3.6 Conclusion and Future Work

This paper has presented our proposed CMS-RCNN approach to robustly detect human facial regions from images collected under various challenging conditions, e.g., highly occlusions, low resolutions, facial expressions, illumination variations, etc. The approach is benchmarked on two challenging face detection databases, i.e., the WIDER FACE Dataset and the FDDB, and compared against recent other face detection methods. The experimental results show that our proposed approach outperforms strong baselines on the WIDER FACE and consistently achieves very competitive results against state-of-the-art methods on the FDDB.

In our implementation, the proposed CMS-RCNN consists of the MS-RPN and the CMS-CNN. During training, they are merged together in an approximate joint training style for each SGD iteration, in which the derivatives w.r.t. the proposal boxes' coordinates are ignored. In the future, we want to go to the fully joint training so that the network can be trained in end-to-end fashion (Fig. 3.12).

4. X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2012), pp. 2879–2886
5. J. Li, Y. Zhang, Learning surf cascade for fast and accurate object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 3468–3475
6. H. Li, G. Hua, Z. Lin, J. Brandt, J. Yang, Probabilistic elastic part model for unsupervised face detector adaptation, in *Proceedings of the IEEE International Conference on Computer Vision* (2013), pp. 793–800
7. N. Markuš, M. Frljak, I.S. Pandžić, J. Ahlberg, R. Forchheimer, A method for object detection based on pixel intensity comparisons organized in decision trees (2013). [arXiv:1305.4537](https://arxiv.org/abs/1305.4537)
8. H. Li, Z. Lin, J. Brandt, X. Shen, G. Hua, Efficient boosted exemplar-based face detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 1843–1850
9. M. Mathias, R. Benenson, M. Pedersoli, L. Van Gool, Face detection without bells and whistles, in *Computer Vision-ECCV 2014* (Springer, Berlin, 2014), pp. 720–735
10. D. Chen, S. Ren, Y. Wei, X. Cao, J. Sun, Joint cascade face detection and alignment, in *Computer Vision-ECCV 2014* (Springer, Berlin, 2014), pp. 109–122
11. B. Yang, J. Yan, Z. Lei, S.Z. Li, Aggregate channel features for multi-view face detection, in *2014 IEEE International Joint Conference on Biometrics (IJCB)* (IEEE, 2014), pp. 1–8
12. G. Ghiasi, C.C. Fowlkes, Occlusion coherence: detecting and localizing occluded faces (2015). [arXiv:1506.08347](https://arxiv.org/abs/1506.08347)
13. S. Liao, A. Jain, S. Li, A fast and accurate unconstrained face detector (2014)
14. H. Li, Z. Lin, X. Shen, J. Brandt, G. Hua, A convolutional neural network cascade for face detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 5325–5334
15. S.S. Farfade, M.J. Saberian, L.-J. Li, Multi-view face detection using deep convolutional neural networks, in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval* (ACM, 2015), pp. 643–650
16. S. Yang, P. Luo, C.-C. Loy, X. Tang, From facial parts responses to face detection: a deep learning approach, in *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 3676–3684
17. R. Ranjan, V.M. Patel, R. Chellappa, A deep pyramid deformable part model for face detection, in *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)* (IEEE, 2015), pp. 1–8
18. B. Yang, J. Yan, Z. Lei, S.Z. Li, Convolutional channel features, in *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 82–90
19. R. Ranjan, V.M. Patel, R. Chellappa, Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition (2016). [arXiv:1603.01249](https://arxiv.org/abs/1603.01249)
20. V. Jain, E. Learned-Miller, Fddb: a benchmark for face detection in unconstrained settings. University of Massachusetts, Amherst, Technical Report UM-CS-2010-009 (2010)
21. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition (2014). [arXiv:1409.1556](https://arxiv.org/abs/1409.1556)
22. P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models. *IEEE Trans. PAMI* **32**(9), 1627–1645 (2010)
23. X. Yu, J. Huang, S. Zhang, W. Yan, D. Metaxas, Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model, in *Proceedings of the IEEE International Conference on Computer Vision* (2013), pp. 1944–1951
24. S.K. Divvala, D. Hoiem, J.H. Hays, A.A. Efros, M. Hebert, An empirical study of context in object detection, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009* (IEEE, 2009), pp. 1271–1278
25. S. Bell, C.L. Zitnick, K. Bala, R. Girshick, Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks (2015). [arXiv:1512.04143](https://arxiv.org/abs/1512.04143)

26. S. Zagoruyko, A. Lerer, T.-Y. Lin, P.O. Pinheiro, S. Gross, S. Chintala, P. Dollár, A multipath network for object detection (2016). [arXiv:1604.02135](https://arxiv.org/abs/1604.02135)
27. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems* (2012), pp. 1097–1105
28. R. Girshick, J. Donahue, T. Darrell, J. Malik, Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(1), 142–158 (2016)
29. R. Girshick, Fast R-CNN, in *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 1440–1448
30. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in *Advances in Neural Information Processing Systems* (2015), pp. 91–99
31. M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in *Computer Vision-ECCV 2014* (Springer, Berlin, 2014), pp. 818–833
32. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in *ECCV* (2014), pp. 740–755
33. B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, Hypercolumns for object segmentation and fine-grained localization, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 447–456
34. W. Liu, A. Rabinovich, A.C. Berg, Parsenet: looking wider to see better (2015). [arXiv:1506.04579](https://arxiv.org/abs/1506.04579)
35. Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in *Proceedings of the ACM International Conference on Multimedia* (ACM, 2014), pp. 675–678
36. C.L. Zitnick, P. Dollár, Edge boxes: locating object proposals from edges, in *ECCV* (Springer, Berlin, 2014), pp. 391–405
37. M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.* **88**(2), 303–338 (2010)