



Registration-Free Face-SSD: Single Shot Analysis of Smiles, Facial Attributes, and Affect in The Wild

Youngkyoon Jang^{a,**}, Hatice Gunes^b, Ioannis Patras^c

^a*University of Bristol, 1 Cathedral Square, Trinity Street, Bristol BS1 5DD, UK*

^b*University of Cambridge, William Gates Building, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK*

^c*Queen Mary University of London, Mile End Road, London E1 4NS, UK*

ABSTRACT

In this paper, we present a novel single shot face-related task analysis method, called Face-SSD, for detecting faces and for performing various face-related (classification / regression) tasks including smile recognition, face attribute prediction and valence-arousal estimation in the wild. Face-SSD uses a Fully Convolutional Neural Network (FCNN) to detect multiple faces of different sizes and recognise / regress one or more face-related classes. Face-SSD has two parallel branches that share the same low-level filters, one branch dealing with face detection and the other one with face analysis tasks. The outputs of both branches are spatially aligned heatmaps that are produced in parallel – therefore Face-SSD does not require that face detection, facial region extraction, size normalisation, and facial region processing are performed in subsequent steps. Our contributions are threefold: 1) Face-SSD is the first network to perform face analysis without relying on pre-processing such as face detection and registration in advance – Face-SSD is a simple and a single FCNN architecture simultaneously performing face detection and face-related task analysis – those are conventionally treated as separate consecutive tasks; 2) Face-SSD is a generalised architecture that is applicable for various face analysis tasks without modifying the network structure – this is in contrast to designing task-specific architectures; and 3) Face-SSD achieves real-time performance (21 FPS) even when detecting multiple faces and recognising multiple classes in a given image (300×300). Experimental results show that Face-SSD achieves state-of-the-art performance in various face analysis tasks by reaching a recognition accuracy of 95.76% for smile detection, 90.29% for attribute prediction, and Root Mean Square (RMS) error of 0.44 and 0.39 for valence and arousal estimation.

© 2019 This manuscript version is made available under the CC-BY-NC-ND 4.0 license.

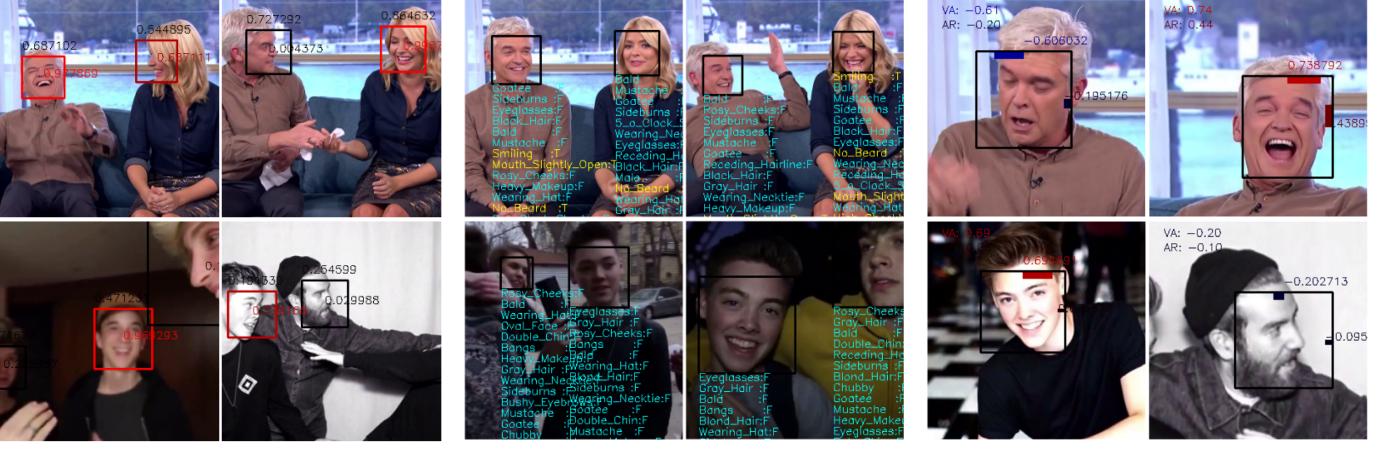
1. Introduction

Face analysis is one of the most studied areas in various research communities including Computer Vision (CV) and Affective Computing (AC). Cutting edge results are constantly obtained for various face-related analysis and recognition tasks including face detection [60, 62, 63], face recognition [55], expression recognition [26], valence-arousal estimation [24], action unit detection [57, 26], face attribute recognition [30, 12], age estimation [5, 15, 1], landmark detection [31, 45] and face alignment [19]. However, in order to get the best performance,

recent studies design specific architectures for each individual face analysis task. Although some works propose unified frameworks for handling multiple face-related tasks [56, 3, 35], several open issues remain yet to be explored:

- **Unconstrained conditions:** Most of the existing approaches require a detected and normalised face input.
- **Scalability:** Most methods design separate networks for different tasks. However, networks that are specifically designed to maximise the performance for certain tasks cannot be easily adapted to do other types of face analysis tasks.
- **Real-time performance:** Existing methods do not achieve real-time performance because they require time-

**Corresponding author: Tel.: +44-(0)752-214-2643;
 e-mail: youngkyoon.jang@bristol.ac.uk (Youngkyoon Jang)



(a) Smile Recognition

(b) Facial Attribute Prediction

(c) Valence-Arousal Estimation

Fig. 1. Our system, which we refer to here as Face-SSD, detects faces and smiles, recognises facial attributes, and predicts affect along the valence and arousal dimensions, in the wild. (a) When detected faces are determined as smiling faces, the colour of the black bounding box changes to red. The probability that appears at the top of the box indicates the face confidence score, and the one appearing in the middle of the box is the smile confidence score. The intensity of red corresponds to the level of confidence. (b) 40 attribute labels that are sorted in descending order (using prediction score) are displayed below the corresponding face bounding box. Attributes classified as “True” are displayed in yellow text, or displayed in another colour otherwise. For sorting, we use a modified value ($1.0 - \text{predicted score}$) for the attributes that are classified as “False”. (c) Horizontal and vertical bars indicate the degree of valence (VA) and arousal (AR), respectively. A bar starts in the middle (representing a value of zero) and ends at the corner of the bounding box along both positive (right / top) and negative (left / bottom) directions. The intensity of the colour (red for positive, blue for negative) corresponds to the level of the predicted score between -1 and 1 . According to the content of the dataset, Face-SSD has been trained for a limited range of face sizes (see Sec. 4.3.3 for details). (Best viewed in colour)

consuming preprocessing steps such as face detection and registration before performing face analysis.

In order to address the above mentioned challenges, we propose Face-SSD, a network that performs simultaneously face detection and one or more face analysis tasks (see Fig. 1) in a single architecture. Face-SSD aims to not only detect faces in a given colour image (upper part in Fig. 2 (a)), but also to perform several other face analysis tasks (lower part in Fig. 2 (a)) associated with the detected faces. Similar to the SSD used for object detection [29], the proposed Face-SSD uses a pre-trained VGG16 network [46] to extract low level features as shown in Fig. 2 (a) [G1:G5]. Then, multi-scaled convolution layers are added after the convolutional layers of the VGG16 to perform both classification (face classification and face analysis task) and regression (bounding box localisation) tasks (see Fig. 2 (a) [G6:G10]). To the best of our knowledge, Face-SSD is the first single face network that can handle several face analysis tasks without a pre-normalisation step.

The proposed architecture is trained and evaluated using well-known benchmark datasets for face detection (AFLW [23]), smile recognition (GENKI-4K [16], CelebA [30]), facial attribute prediction (CelebA [30]) and valence-arousal estimation (AffectNet [32]). As discussed in Sec. 3.1.1, we first obtain a set of matched default boxes as proposed by Liu et al. [29]. Then, we train the Face-SSD by optimising multiple losses (associated with face classification, bounding box regression, and a face-related task such as smile recognition, multiple facial attributes prediction or valence-arousal estimation). We adopt data augmentation and Hard Negative Mining (HNM) strategies, and achieve state-of-the-art or very competitive performance in various face analysis tasks without modifying the

structure and while maintaining real-time performance.

The main contributions of our work are three-fold:

1. **Unconstrained processing:** Face-SSD does not rely on a pre-normalisation step, it requires neither face detection nor registration in advance. Most of the existing approaches to face analysis require a cropped or normalised face in advance.
2. **Universal architecture:** Face-SSD can be applied to most face analysis tasks with a simple modification (the number of final prediction channels), and achieve state-of-the-art or very competitive results. Most of the existing approaches use separate networks.
3. **Real-time processing:** Face-SSD can be trivially extended to perform several face analysis tasks at negligible additional processing time.

The remainder of the paper is organised as follows: In Section 2, related work in face analysis that rely on registration, require task-specific model design, or handle multiple tasks is reviewed. In Section 3, we present the proposed Face-SSD framework and explain how to apply Face-SSD to several face applications. Experimental results of the applications using the proposed Face-SSD are provided in Section 4. Finally, conclusions are drawn and discussed in Section 5.

2. Related Work

General pipeline for facial analysis. Sariyanidi et al. [42] discusses the state-of-the-art methods for face registration, representation, dimensionality reduction and recognition, which

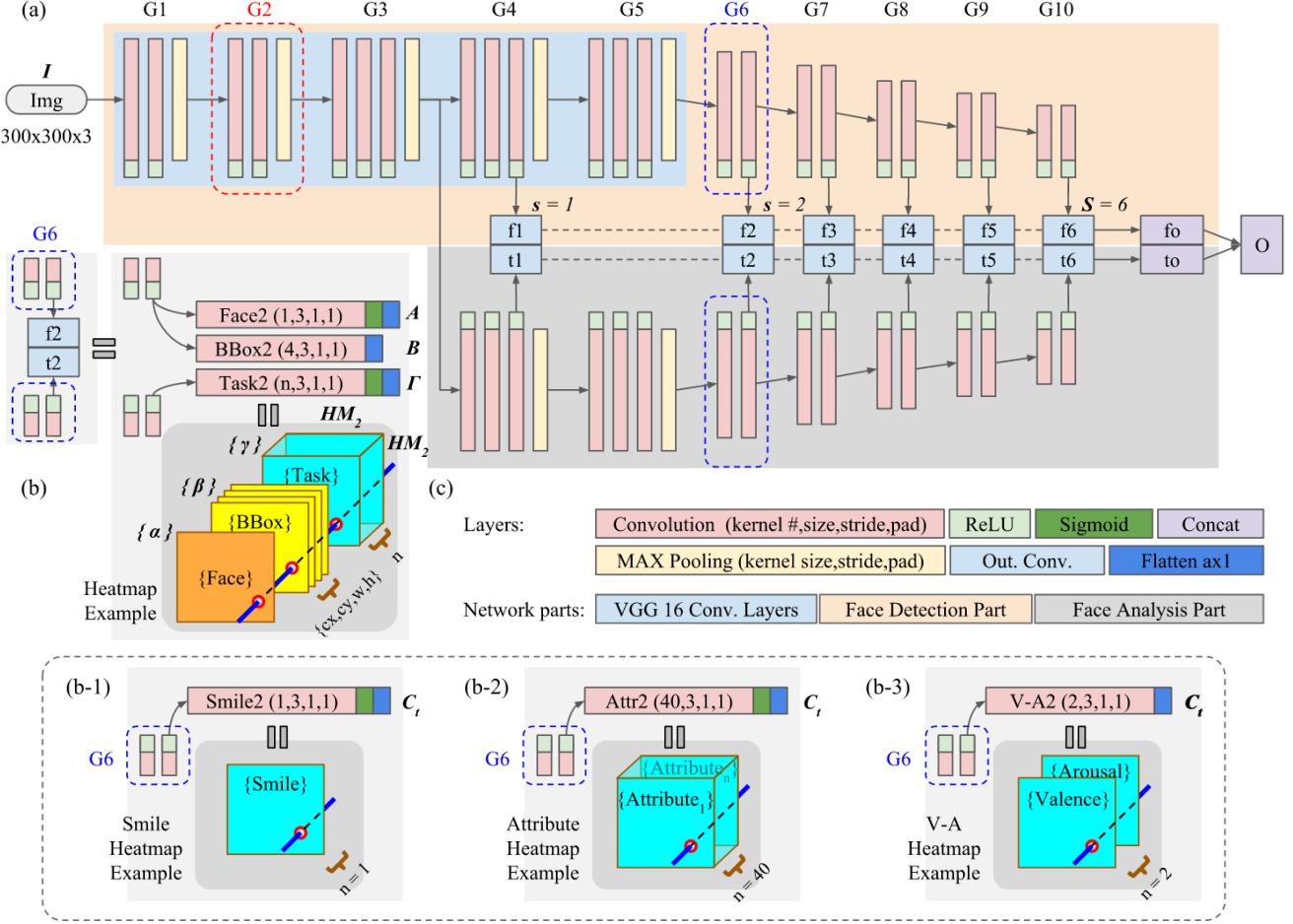


Fig. 2. The architecture of Face-SSD. (a) The entire architecture of Face-SSD consisting solely of convolution and pool layers. (b) Example of the concatenated output convolution layer for the second scale ($s = 2$) that produces a heatmap volume. (b-1) through (b-3) show how to modify Face-SSD for binary classification (smile recognition), multi-class recognition (facial attribute prediction), and multi-task regression (valence-arousal estimation). (c) Legend for layers and parts of Face-SSD.

are the common components of a generic pipeline for performing automatic facial affect analysis. Depending on the target application, the generic pipeline might have to be changed to some degree. Nonetheless, the first two steps of face localisation and 2D / 3D registration steps have been necessary for most of the face analysis tasks such as smile recognition, facial attribute prediction, valence-arousal estimation, gender recognition, age prediction, and head pose estimation. See [42, 50, 53, 58] for details.

Registration-based face analysis. Despite significant advances in deep learning, automatic face analysis tasks, such as smile detection (a comparative review is provided in Table 3), attribute prediction [12, 11, 43] and valence-arousal estimation [32], still face major challenges caused by occlusions and variances of head pose, scale, and illumination. These challenges are the main reason why every state-of-the-art approach to face analysis requires a pre-normalisation step involving face detection and registration (rotation, scaling, and 2D/3D transformation).

Approaches without pre-normalisation. There exist some works that process the original input image without pre-normalisation steps. Liu et al. [30] combines LNet localising

a face and ANet to predict facial attributes. However, they use EdgeBox [65] that proposes a number of candidate windows to determine the final facial region among the multiple predicted positions scattered by LNet. Before feeding the output of LNet to ANet, this process for narrowing the potential face region is performed several times through several LNet stages. This method uses an actual image as input, but the processes inside the architecture operate as a sequential pipeline.

Ranjan et al. [35] proposed a deep neural network consisting of multiple branches to handle various face-related tasks. The proposed network uses Selective Search [41] to generate multi-region proposals. Although the proposed network deals with face classification and face-related tasks in different branches, the face classification and multiple face analysis tasks are performed as separate continuous pipelines. In summary, most of the previous works that do not require a pre-normalisation step follow similar mechanisms to [30, 35], which require region proposal steps in the middle of the process. These region proposal steps typically increase the overall processing time.

Task-specific model design. Several recent approaches address the problem of facial attributes prediction [11, 43]. Some propose to use successful face-specific feature representations

[64], modelling class distributions [8] and balancing attributes [37], indirect guiding the categorisation of similar features [52, 43], or direct grouping the relevant attributes [12, 11]. The best performances (more than 90% accuracy) are obtained by specifically designing a model structure that utilises the relations between relevant attributes [12, 11]. For this purpose, MCNN-AUX [12] uses implicit and explicit attribute relationships while DMTL [11] relies on attribute correlation and heterogeneity. R-Codean [43] proposes a new loss function that incorporates both magnitude and direction of image vectors in feature learning and proposes a framework incorporating a patch-based weighting mechanism. By assigning higher weights to relevant patches for each attribute, the method has similar advantages to grouping relevant attributes.

Compared to the previous studies proposing a task-specific model, we propose a generalised architecture that can be used for attribute prediction as well as other face analysis tasks. Unlike the state-of-the-art methods (specifically in face attribute prediction), the proposed Face-SSD uses similar face-size categories associated with each output layer, and incorporates the Hide-and-Seek [47] data augmentation method which forces the network to explore the entire face area more extensively during training. This size-based categorisation and simple data augmentation strategy enabled us to achieve performance close to the state of the art (more than 90% accuracy in attribute prediction) without customising the network.

Multi-task facial analysis. While most works on facial analysis use a specially designed architecture to tackle a single application, some works attempt to use an integrated single architecture with multiple branches to address multiple tasks. Ranjan et al. [35] categorised face-related tasks into two groups: subject-independent tasks (e.g., keypoint detection, pose and smile) and subject-dependent tasks (e.g., gender and facial identity). The all-in-one network [35] learns multiple tasks in a single architecture, but first uses subject-independent class results to register faces. Then the network performs subject-dependent classification tasks sequentially.

Chang et al. [3] proposed FATAUVA-Net to learn multiple tasks related to affect behaviours in a single deep neural network. Similarly to the all-in-one network [35], FATAUVA-Net categorised similar tasks that share a feature layer. For example, the network branches eye-related tasks (attributes: eyeglasses, narrow eyes / Action Unit (AU): AU6, AU7, AU45) from the same previous layer. The network branches mouth-related tasks (attributes: mouth slightly open, smile / AU: AU23, AU24, AU25) from other layers extracting mouth-related features. The network branches the valence and arousal prediction layer from the associated AU layers. Although these architectures predict multiple face analysis tasks, it is still difficult to generalise or use them in other tasks that take advantage of large patterns for face detection and use small patterns for other face analysis tasks.

Object (face) detection in the wild. The proposed Face-SSD is inspired by SSD [29] considering the face as a specific type of object. SSD [29] has been applied and extended in many research domains, including text detection [14], face detection [62], object pose estimation [34, 21] and temporal action de-

tection [27]. Similar to the latest methods Face-SSD uses the concept of default box [29] or anchor box [36]. Using a baseline architecture that has been successfully applied to various detection tasks, we propose the first SSD-inherited architecture that tackles both continuous large-pattern-leveraged tasks (e.g., face detection) and small-pattern-leveraged tasks (e.g., face analysis) in parallel.

3. The Proposed Framework: Face-SSD

The proposed Face-SSD framework is shown in Fig. 2. Face-SSD is a fully convolutional neural network consisting solely of convolutional and pooling layers. The input to the Face-SSD is a colour image I . There are six layers (i.e. $S = 6$), each corresponding to a certain scale, that is size of face. At each scale $s \in [1 \dots 6]$ the output is a heatmap A containing at each spatial position i the confidence score that a face is present at that location, a heatmap B with the parameters of the bounding box of the face associated with that position i , and a heatmap Γ with the face analysis task confidence score(s) at each position i , that is $\{\alpha_i, \beta_i, \gamma_i\}_s$ at every spatial location i , as shown in Fig. 2(b). At test time, a threshold at the face detection confidence score heatmap A selects candidate faces at several spatial locations i . Subsequently, Non-Maximum Suppression (NMS) [33] is used to derive the bounding boxes and in each of them calculate scores for the face analysis tasks and for the face detection.

The following sections describe how to configure Face-SSD (Sec. 3.1), how to train face detection and face analysis task in a single architecture (Sec. 3.2), and how to combine the face detection and analysis results during testing (Sec. 3.3).

3.1. Model Construction

Face-SSD consists of layers performing at various stages feature extraction (VGG16 Conv. Layers), face detection, and face analysis as shown in Fig. 2(a). $G[1 : 10]$ represents convolution and pooling layer groups with the same input resolution. For example, G2 consists of two convolution layers and one pooling layer, whereas G6 consists of two convolution layers. Similarly to SSD [29], Face-SSD outputs six-scale ($S = 6$) heatmap volumes generated by multiple output convolution layers [(f1, t1):(f6, t6)]. f[1:6] is produced by the face detection part, while t[1:6] is produced by the face analysis part. The output convolution layers of the two different parts are aligned and concatenated at the end.

Each concatenated output convolution layer outputs a pixel-wise heatmap volume consisting of $(1 + 4 + n)$ heatmap planes. For example, the concatenated output convolution layer for the second scale ($s = 2$) outputs a three-dimensional volume $(HM_2 \times HM_2 \times (1 + 4 + n))$ consisting of $(1 + 4 + n)$ heatmap planes having the same resolution $(HM_2 \times HM_2)$ of the second scale, as shown in Fig. 2(b). The first plane indicates the existence of a face. The next four heatmap planes at each spatial position i contain the centre $(cx, cy) \in R^2$ of the face bounding box and its width w , and height h . The former is relative to the location i (i.e., (cx, cy) are actually offsets) and the latter is relative to the current heatmap scale s . The remaining set of n heatmap planes are the confidences for the n face analysis tasks

Table 1. The detailed parameters of Face-SSD layers (see text)

Group ID	Conv. ID: Parameters	Pool
G1	[1:2]: (64, 3, 1, 1)	(2, 2, 0)
G2	[1:2]: (128, 3, 1, 1)	(2, 2, 0)
G3	[1:3]: (256, 3, 1, 1)	(2, 2, 0)
G4	[1:3]: (512, 3, 1, 1)	(2, 2, 0)
G5	[1:3]: (512, 3, 1, 1)	(3, 1, 1)
G6	1: (1024, 3, 1, 1) 2: (1024, 1, 1, 0)	.
G7	1: (256, 1, 1, 0) 2: (512, 3, 2, 1)	.
G8	1: (128, 1, 1, 0) 2: (256, 3, 2, 1)	.
G9	1: (128, 1, 1, 0) 2: (256, 3, 1, 0)	.
G10	1: (128, 1, 1, 0) 2: (256, 3, 1, 0)	.
Out. Conv.	C_f : (1, 3, 1, 1) B : (4, 3, 1, 1) C_t : (n, 3, 1, 1)	.

– note that these are also heatmaps, that is, they have spatial dimensions as well.

All of the convolution layers are followed by ReLU activation function except for the output convolution layer. For the output convolution layer, for binary classification tasks, such as face classification, smile recognition and attribute prediction, we use the sigmoid function (see Fig. 2(b), (b-1) and (b-2), respectively). For regression tasks such as bounding box offsets and valence-arousal estimation, we use linear functions similarly to SSD [29] (see Fig. 2(b) and (b-3)). The parameters for the layers in Face-SSD are summarised in Table 1. The parameters of the convolution layer are denoted in the order of number of kernels, kernel size, stride and padding, while the parameters of the pool layer follow the order of kernel size, stride and padding.

During training, the output (prediction) values that appear in heatmaps responsible for the bounding box and tasks are examined only when the corresponding face label exists in the pixel (see details in Sec. 3.2.1). During testing, the values for the bounding box and the task-related output are examined only when the corresponding face confidence score exceeds a threshold. The face detection threshold is determined by selecting the optimal value that provides the best performance on the face detection task.

3.1.1. Implementation details

Single aspect ratio: We utilise only one aspect ratio (square) configuring a default box to assign a ground truth label to a pixel position in a heatmap, as shown in Fig. 3. This is because face deformations, caused by expression and pose, result in similar aspect ratios. This is in accordance with the related work in the literature – e.g., Hao et al. [13] proposed Single-Scale RPN utilising one anchor box and Zhang et al. [62] proposed S³FD utilising one default box.

Usage of pre-trained models: Several works including Liu et al. [30] demonstrate that models pre-trained on object recogni-

tion (e.g., ImageNet [6]) are useful for face localisation. Similarly, networks pre-trained on face recognition (e.g., CelebFaces [49]) are useful for capturing face attributes at a more detailed level. For this reason, we selectively use pre-trained parameters (trained with an object dataset [38, 46] and a face dataset [23]) to initialise the convolution filters for face detection and analysis tasks (see details in Sec. 3.2). This usage of pretrained models helps with improving the Face-SSD performance for both face detection (utilising large patterns) and analysis (utilising relatively smaller patterns) tasks.

3.2. Training

Training of Face-SSD follows the following four steps:

1. Copying parameters of the VGG16 network [46] (convolution layers) to the VGG16 (feature extraction) part $G[1 : 5]$ of Face-SSD and subsampling¹ the parameters from fully connected layers ($fc6$ and $fc7$) of VGG16 network to the $G6$ layers of Face-SSD, as described in SSD [29].
2. Freezing the face analysis part and finetuning the face detection part by using the AFLW (face) dataset [23].
3. Copying the parameters of the layers $G[4 : 10]$ constituting the face detection part to the corresponding layers of the face analysis part.
4. Freezing the face detection part and finetuning the layers $G[4 : 10]$ constituting the face analysis part by using task-related datasets (e.g., CelebA [30] or GENKI-4K [16] for smile recognition, CelebA [30] for facial attribute prediction, AffectNet [32] for valence-arousal estimation).

The first and second steps are similar to the initialisation and end-to-end learning process of SSD network [29]. We use the same cost function as the SSD to finetune the face detection part of Face-SSD.

3.2.1. Face Detection

As described above, finetuning of the face detection part is based on the use of an objective loss function L_{face} , which is a weighted sum of the face classification loss L_{cls} and the bounding box regression loss L_{reg} defined as:

$$L_{face}(x_f, c, l, g) = \frac{1}{N} (L_{cls}(x_f, c) + \lambda x_f L_{reg}(l, g)), \quad (1)$$

where N is the total number of matched default boxes. For the regression loss L_{reg} , smooth L1 loss [10] is used for calculating the distance between the predicted $l = \{l_{cx}, l_{cy}, l_w, l_h\}$ and

¹For example, the first fully connected layer $fc6$ of the VGG16 network [46] connects all the positions of a $T_i = (f_{vi}, m, m) = (512, 7, 7)$ dimensional input feature map, where f_{vi} is the feature (kernel) dimension at each of the m^2 spatial locations, to a $f_{vo} = 4096$ dimension output vector T_o . Let us organise the weights in a tensor W_{vgg} with dimensions $(f_{vo}, f_{vi}, m, m) = (4096, 512, 7, 7)$. On the other hand, Face-SSD takes an input feature map with dimensions $(512, 18, 18)$ and outputs a feature map with dimensions $T'_o = (a, m', m') = (1024, 18, 18)$ using filters with kernel size 3×3 . The weight tensor W_{fssd} is then of dimensions $(1024, 512, 3, 3)$. In order to initialise the W_{fssd} , we uniformly subsample the W_{vgg} along each of its modes – in our case by a factor $(4, 1, 3, 3)$. This corresponds to subsampling by a factor of 4 along the dimension of the output feature vector T_o and by a factor of 3 along each spatial dimension of the input tensor T_i of the VGG16 network – we copy the corresponding weights.

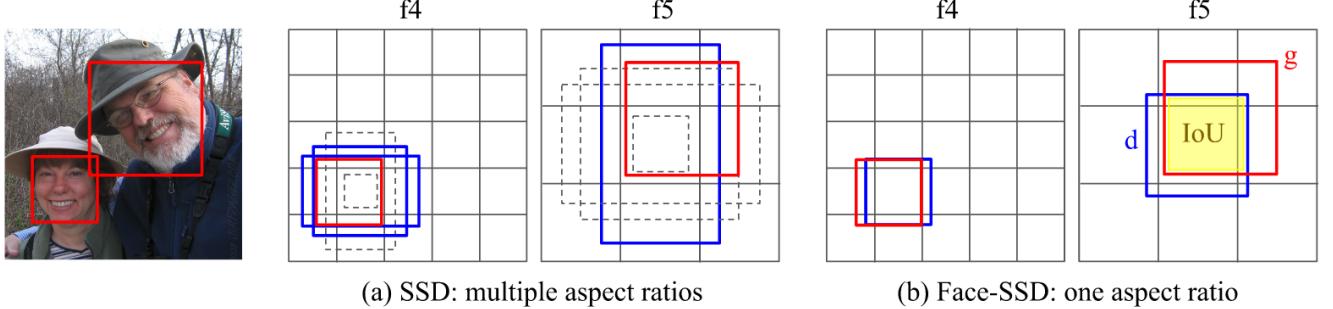


Fig. 3. Example of matched default box for the face confidence heatmaps $C_{f[4:5]}$, produced by $f4$ and $f5$ output convolution layers (see Fig. 2). (a) Dotted boxes (grey) represent multiple candidate default boxes with multiple different aspect ratios. Face-SSD (b) uses only one aspect ratio in the matching process of the default box d . The example image is one of the sample images of AFLW dataset [23].

the ground truth $g = \{g_{cx}, g_{cy}, g_w, g_h\}$ bounding boxes [29], as shown in Eq. 2 and 3. Specifically,

$$\begin{aligned} L_{reg}(l, g) &= \sum_{m \in \{cx, cy, w, h\}} smooth_{L_1}(l_m - \hat{g}_m), \\ \hat{g}_{cx} &= (g_{cx} - d_{cx})/d_w, \quad \hat{g}_{cy} = (g_{cy} - d_{cy})/d_h, \\ \hat{g}_w &= \log(g_w/d_w), \quad \hat{g}_h = \log(g_h/d_h), \end{aligned} \quad (2)$$

where

$$smooth_{L_1}(k) = \begin{cases} 0.5k^2, & \text{if } \|k\| < 1 \\ \|k\| - 0.5, & \text{otherwise} \end{cases} \quad (3)$$

The face classification loss L_{cls} is based on binary cross entropy over face confidence scores c , as shown in Eq. 4.

$$L_{cls}(x_f, c) = -x_f \log(c) - (1 - x_f) \log(1 - c) \quad (4)$$

The flag $x_f \in \{1, 0\}$, used in the equations above is set to 1 when the overlap between the ground truth and the default bounding box $d = \{d_{cx}, d_{cy}, d_w, d_h\}$ exceeds a threshold. Note that the regression loss is only used when $x_f = 1$, and is disabled otherwise.

At the later stages of the training, similar to [29] we use Hard Negative Mining (HNM), that is, we sort calculated losses only in the background region ($\neg(x_f = 1)$) in descending order and select and backpropagate only from the highest ones. Following [29], we set the loss-balancing weight λ (in Eq. 1) to 1.

3.2.2. Face Analysis

This section describes how to apply Face-SSD to various face analysis tasks. We address three problems: smile recognition as binary classification, facial attribute prediction as multi-class recognition and valence-arousal estimation as multi-task regression. In all three problems, the architecture of the network differs only in terms of the number n of the facial task heatmaps. For datasets that have multiple annotations for the same image, Face-SSD supports multi-task learning by defining a multi-task loss function as in Eq. 5.

$$L_{total} = \sum_{t=1}^T \|w_t L_t(g_t, p_t)\|_2, \quad (5)$$

That is, the multi-task loss L_{total} is defined as the $L2$ norm of multiple weighted individual face analysis task losses $\{w_t L_t\}$.

L_t is used to calculate errors using a ground truth g_t and a prediction p_t for a given task t . T denotes the total number of face analysis tasks. In what follows we define the loss functions used for different problems we address.

Smile Recognition. The smile classification loss L_{smile} , is the binary cross entropy over smile confidence scores e and the ground truth $x_e = \{1, 0\}$ as defined in Eq. 6.

$$L_{smile}(x_e, e) = -x_e \log(e) - (1 - x_e) \log(1 - e) \quad (6)$$

The ground truth $x_e = \{1, 0\}$ at each location is set using the default box matching strategy [29]. The loss is defined at each spatial location of the output heatmap, and in this case, we do not use Hard Negative Mining (HNM), which was required to select negative samples for face detection (see Sec. 3.2.1).

Finetuning the network for face analysis tasks (i.e., smile recognition) does not impair the face detection performance due to freezing the parameters for the face detection part of Face-SSD.

Facial Attribute Prediction. Facial attribute prediction is treated as multiple binary classification problems where a number of attributes may exist simultaneously. For example, a face attribute (such as smiling) can appear independently of other attributes (such as the gender or hair colour). Therefore, we define the facial attribute prediction loss L_{att} as the average of independent attribute losses, that is

$$L_{att}(G, P) = -\frac{1}{N_a} \sum_{a=1}^{N_a} (g_a \log(p_a) + (1 - g_a) \log(1 - p_a)), \quad (7)$$

where N_a denotes the total number of attributes. $g_a \in G$ and $p_a \in P$ denote the ground truth (1 or 0) label and a predicted attribute confidence score of the a -th attribute, respectively. For calculating a single attribute prediction loss associated with an individual attribute a , we use the binary cross entropy over attribute confidence scores p_a .

Valence and Arousal Estimation. Similar to several other previous works (e.g. [22], [32]), we treat arousal and valence prediction as a regression problem. Valence is related to the degree of positiveness of the affective state, whereas arousal is related

to the degree of excitement [39, 40]. We used the Euclidean (L2) distance between the predicted value \hat{y}_n and ground truth value of valence/arousal y_n , as shown in Eq. 8. The loss is then defined as the sum of the valence E_v and the arousal E_a losses, that is

$$L_{emo} = E_v + E_a,$$

$$E = \frac{1}{2N} \sum_{n=1}^N ||\hat{y}_n - y_n||_2^2, \quad (8)$$

where N is the number of image samples in a mini-batch.

3.2.3. Data Augmentation in Training

Face-SSD uses a 300×300 resolution and 3 channel colour input image. Prior to data augmentation, all pixel values for the R, G, and B channels of a sample image are normalised based on the mean and standard deviation values of the entire dataset. Each sample image is first flipped in the horizontal direction with a probability of 0.5. In the training session, we randomly select one of data augmentation mechanisms (shrinking, cropping, gamma correction and Hide-and-Seek (H-a-S) [47]) to create noisy data samples for each epoch.

Both shrinking and cropping maintain the aspect ratio. Gamma correction is applied separately to the individual R, G, B channels. In Hide and Seek (H-a-S) [47] we hide image sub-areas and force a network to seek more context in areas that are not as discriminative as key distinctive areas such as lip corners. We first randomly select a division number among 3, 4, 5 or 6. If we select 3, the image region will be divided into 9 (3×3) sub-image patches. Each sub-image patch is then hidden (filled with the mean R, G, B values of all data samples in a dataset) with a probability of 0.25.

3.3. Testing

The registration-free Face-SSD for a specific face analysis task (e.g., smile recognition) is based on both face and task (e.g., smile) confidence scores. First, the locations in the face confidence heatmap, for which the score exceeds a threshold ($th_{face} = 0.1$), are selected. Then Non-Maximum Suppression (NMS) method (with jaccard overlap value 0.35 as in S³FD [62]) is used to extract the final bounding boxes. Subsequently, a task-specific threshold th_t is applied on the task related score of the final bounding boxes (Fig. 4). In the case of the regression (e.g., valence-arousal estimation), the output value of the final bounding box is used.

As mentioned in Sec. 3.1, each output layer of Face-SSD generates several heatmaps: one for face detection, four for the offset coordinates of face bounding box and n for the n number of face analysis tasks, as shown in Fig. 2(b). Specifically, Fig. 4(a) and (b) visualise the heatmaps generated by Face-SSD's second and third-scale output layers ($s = 2, 3$), which handle the second and third smallest sizes of the face that appears in the image, respectively. Thus, activations in the heatmap are high when a specific size of face is detected. For the given example of smile recognition, as shown in Fig. 4(a), the forefront heatmap shows two clusters of pixels, indicating the existence of two faces. The rearmost heatmap highlights the corresponding pixel only when a task is detected. In this example the

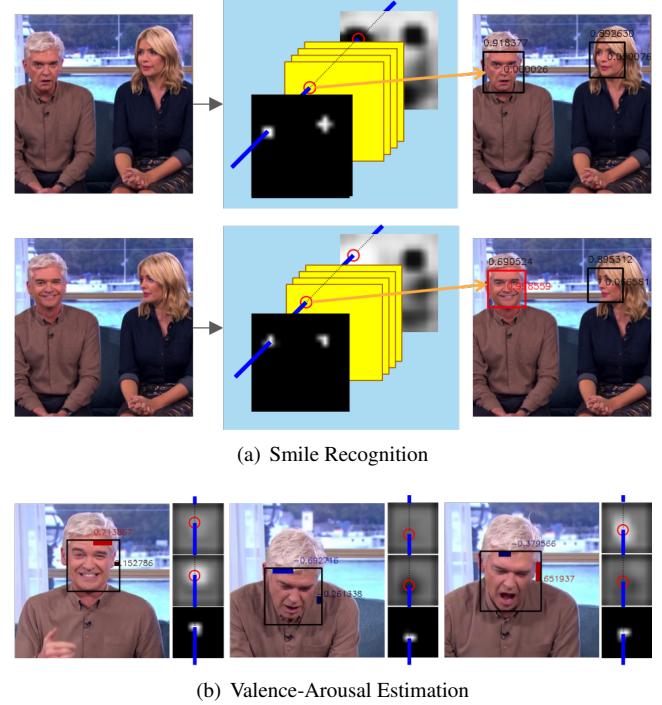


Fig. 4. Examples of face detection and face analysis tasks. As a representative example of classification and regression, we visualised the output heatmaps for smile recognition and valence-arousal estimation. (a) The heatmaps represent face classification, bounding box regression and smile recognition results. (b) For the valence-arousal example, we only visualise the output heatmaps for face classification, valence and arousal estimation from the bottom row. We rescaled the range of output values at the valence-arousal estimation heatmap from $[-1 : 1]$ to $[0 : 255]$ for the visualisation. The median (127) in this example represents the neutral valence or arousal value (0).

heatmap has high values when the detected face is a smiling face.

4. Experiments and Results

4.1. Datasets

In this paper, we show the performance of the proposed Face-SSD on three representative face analysis applications such as smile recognition (binary classification), facial attribute prediction (multiple class recognition), and valence-arousal estimation (multiple task regression). We stress that the structure of the network, including the number of filters and filter sizes remain the same – the only change is the number of output layer heatmaps. We used GENKI-4K [16], CelebA [30], and AffectNet [32] datasets to test the three representative applications using Face-SSD.

Beginning with [54], which performed the first extensive smile detection study, most of the subsequent studies used the GENKI-4K² dataset for performance evaluation [16]. In this paper, the smiling face detection experiments were performed

²The GENKI-4K [16] dataset is a subset of the GENKI dataset used in [54]. This dataset consists of 4,000 face images, each labelled with smile and head pose (yaw, pitch, roll). Only the GENKI-4K dataset is publicly available.

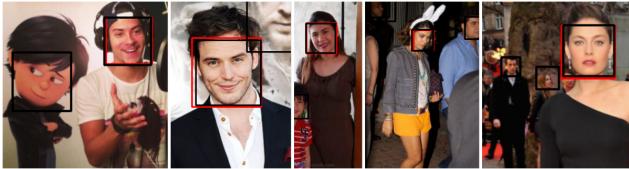


Fig. 5. If there were multiple faces detected (black boxes), only the annotated faces with the ground truth label (red box) were evaluated for attribute prediction. The face detected in the background was not used for accuracy measurement.

not only on the GENKI-4K dataset but also on the CelebA dataset [30] which also contains smile labels. For facial attribute prediction experiments, we used the CelebA dataset [30] which is the most representative dataset. Finally, for the valence-arousal estimation experiment we used the AffectNet [32] dataset consisting of continuous level (valence-arousal) labels and face images captured in the wild.

The AFLW dataset [23] used for face detection and other datasets used for face analysis tasks (i.e., GENKI-4K [16], CelebA [30], AffectNet [32]) have different bounding box positions and shapes. To solve this problem, we empirically adjusted the bounding box position of these datasets to create a square box that surrounds the entire face area centred on the nose (similar to the bounding box of the AFLW dataset). To do this, we first used the trained Face-SSD to detect a face bounding box. Then, we double-checked whether the detected bounding box is correct. If it was incorrect, we modified the bounding box manually.

In particular, when using the CelebA [30] dataset, we only examined smile recognition and facial attribute prediction performance for annotated faces. Each image sample in the CelebA dataset has only one bounding box with its corresponding attribute labels, even if the image contains multiple faces. Therefore, when multiple bounding boxes were detected (black boxes in Fig. 5) during the test time, we only calculated the accuracy for the detected bounding box that matched the ground truth position (red box in Fig. 5). If there is no bounding box detected for the ground truth location, it is considered as a false negative when calculating the accuracy.

4.2. Face Detection

First, we evaluate the face detection performance. Although Face-SSD performs face detection in parallel with one or more tasks, the face analysis task results appearing in the output heatmap are only examined at the corresponding pixel positions that indicate successful face detection (as discussed in Sec. 3.3).

Here, we evaluate the face detection performance of Face-SSD on face analysis task datasets, including GENKI-4K [16], CelebA [30], and AffectNet [32]. According to [7]’s experimental results, the visual recognition ability of a human is degraded when image resolution falls below 20×30 pixels. For this reason, the face detection of Face-SSD aims to support face analysis tasks rather than detecting tiny faces, which is beyond the scope of this work. To this end, we evaluate the face detection performance on the face analysis task (e.g., smile, attribute,

valence-arousal) datasets that do not include severe occlusion or very small faces. Instead, these datasets consist of images that typically contain high-resolution faces compared to 20×30 pixels and are captured in the wild (with naturalistic variations in pose, occlusion, and/or scale).

The face detection results are shown in Table 2 in terms of Equal Error Rate (EER) and Average Precision (AP) [9]. First, we investigated face detection performance using the same strategy as the SSD [29] called Face-SSD Baseline (Face-SSD-B) [18]. The AFLW dataset [23] was used for training face detection part of Face-SSD. For data augmentation, Face-SSD-B used shrinking, cropping, and gamma correction (see details in Sec. 3.2.3). Using the data augmentation, Face-SSD-B trained on the non-challenging face dataset AFLW did not achieve a competitive performance (EER=05.42% and AP=99.50) in comparison to using other face detection datasets. However, unlike general face detection evaluation, we used the simplest face analysis task dataset (GENKI-4K [16]) to provide a performance comparison between different strategy combinations.

To improve the face detection performance we first lowered the IoU threshold from 0.50 to 0.35 when assigning ground truths, similarly to S³FD [62]. Lowering the IoU threshold when matching default box increases the number of positive examples. By doing so the accuracy was improved from EER=05.42% and AP=99.50 to EER=03.68% and AP=99.91.

In order to improve the performance further, we applied a Hard Negative Mining (HNM) strategy on the training data samples in a minibatch. Specifically, we extracted 30% of the data samples that currently output the largest loss in a mini-batch, and then re-used the data samples in the next mini-batch. By doing so, we further reduced the detection error from EER=03.68% and AP=99.91 to EER=01.72% and AP=99.88.

Finally, we applied H-a-S [47] as one of our data augmentation strategies. However, unlike what is reported in the original H-a-S paper [47], when the H-a-S method was applied to all training samples, the detection performance dropped significantly to EER=34.83% and AP=93.54. Applying the H-a-S method randomly to approximately half of the training samples reduced the error to EER=01.95% and AP=99.89. In addition, as shown in Table 2, our results indicate that for face detection it is better to hide coarsely divided patches (EER=01.16% and AP=99.91) than to hide finely divided ones (EER=01.95% and AP=99.89) because face detection relies on relatively large continuous patterns. In Table 2, for H-a-S, the coarse patch division process randomly selects the patch size from 3, 4, 5 and 6 (see Sec. 3.2.3), whereas the fine patch division process randomly selects the patch splitting size from 16, 32, 44 and 56 as proposed originally in [47].

By applying the training strategies of low IoU threshold, HNM and H-a-S, we achieved EER=0.66% and AP=99.88 on the GENKI-4K dataset. For the CelebA dataset, we achieved EER=0.56% and AP=99.88, as shown in Fig. 6. For the AffectNet dataset, we achieved EER=0.65% and AP=99.41. These results indicate that Face-SSD can robustly detect faces in unconstrained environments, and the Face-SSD can be used for further face analysis tasks such as facial attribute and affect prediction along the dimensions of valence and arousal. The op-

Table 2. Effects of using Hard Negative Mining (HNM) and Hide-and-Seek (H-a-S) methods when training face detection in Face-SSD. (See text for more details about abbreviations and description)

	IoU for GTs		HNM	H-a-S for All		H-a-S for Half		GENKI-4K Test Results	
	0.50	0.35		Fine	Coarse	Fine	Coarse	EER (%)	AP
Face-SSD-Baseline [18]	✓	05.42	99.50
Face-SSD-B with More GTs	.	✓	03.68	99.91
Face-SSD-B with HNM	.	✓	✓	01.72	99.88
Face-SSD-B with H-a-S	.	✓	.	✓	.	.	.	34.83	93.54
	.	✓	.	.	✓	.	.	08.26	97.79
	.	✓	.	.	.	✓	.	01.95	99.89
	.	✓	✓	01.16	99.91
Face-SSD	.	✓	✓	.	.	.	✓	00.66	99.88

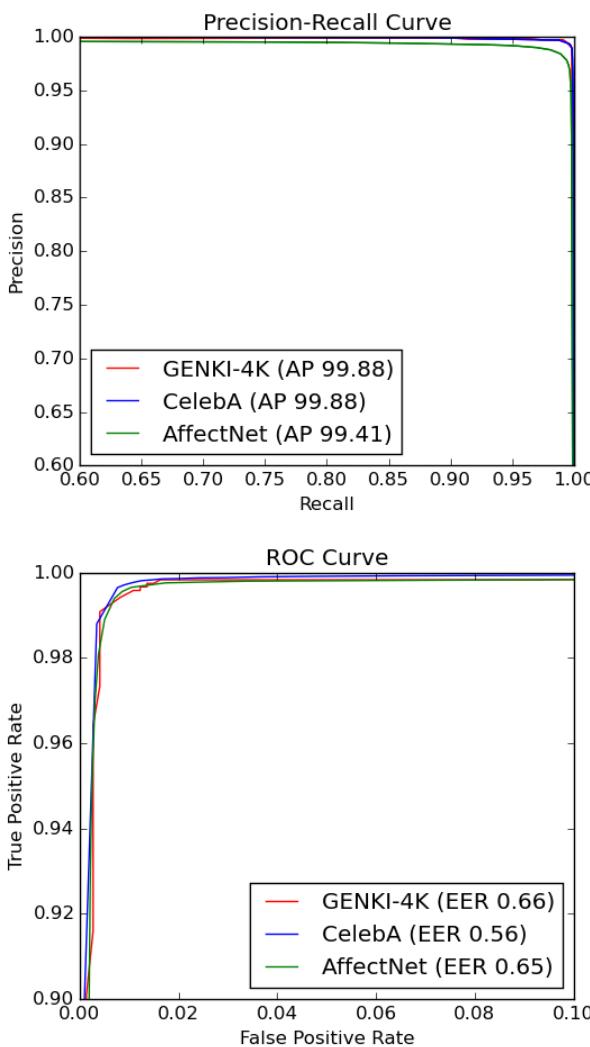


Fig. 6. Experimental curves for face detection performance on GENKI-4K [16], CelebA [30] and AffectNet [32] datasets: Precision-Recall curves and Receiver Operating Characteristic (ROC) curves.

Optimal thresholds for the best face detection accuracy were 0.20 for the GENKI-4K dataset, 0.16 for CelebA dataset, and 0.11 for AffectNet dataset.

4.3. Face Analysis

Face-SSD is inspired by SSD [29], which promises real-time detection performance. Thus, the parameter values used in the process of finetuning the face detection and the face analysis parts of Face-SSD are initialised with the values used for training the base network of SSD [29]. We used SGD with initial learning rate= 10^{-3} , momentum=0.9, weight decay=0.0005, and batch size=16. We used learning rate= 10^{-3} for the first 40K iterations, then continued training for 40K with learning rate= 10^{-2} . We continuously reduced the learning rate every 40K iterations until it reached learning rate= 10^{-5} . Increasing the learning rate for the second 40K iterations speeds up the optimisation process. However, we first started the training process with learning rate= 10^{-3} , because the optimisation process tends to diverge if we use a larger learning rate in the beginning.

The following sections detail the experiments we have conducted to evaluate the two main performance factors of Face-SSD, namely prediction accuracy and processing time, for two tasks: smile recognition and facial attribute prediction.

4.3.1. Smile Recognition

Accuracy for this task refers to the smile recognition performance including the face detection results. If face detection fails, the result of smile recognition is considered to be a non-smile.

Testing on the GENKI-4K dataset: Experiments that use this dataset are conventionally based on four-fold validation procedures. However, as GENKI-4K dataset contains a relatively small number of data samples (4,000), for training we initially utilised the CelebA dataset that contains a rich set of images. When Face-SSD was trained on the CelebA dataset, we used the entire GENKI-4K dataset for testing. We obtained a smile recognition accuracy of 95.23%, as shown in Fig. 7. Despite being trained on a completely different dataset with different characteristics, Face-SSD has already surpassed all the latest methods that used the GENKI-4K dataset for testing, as shown in Table 3.

To provide a fair comparison with other methods that use the four-fold validation strategy, we used the GENKI-4K dataset together with the bounding box annotations obtained with our method (as explained in Sec. 4.1) to finetune the Face-SSD, which was trained on the CelebA dataset. In this case, the smile recognition accuracy is improved further. This is due to the fact that the training samples in GENKI-4K dataset are relatively

Table 3. A detailed comparison with the state-of-the-art methods on the GENKI-4K dataset [16]. We summarise the features, classifiers, detection / registration methods and input image resolution (width, height, and channel) that were used in previous studies in published order. All previous studies require a normalised (cropped and aligned) input image, which necessarily require face detection and registration steps in advance (except [4]-II and III). Some works [44, 17, 59, 25, 4] do not specify how to detect and align a face (in this case, ‘?’), while [20] mentions that the original image is used if the face detection fails.

Method	Feature	Classifier	Detection	Registration	Input ($W \times H \times C$)	Accuracy (%)
[44]	Pixel comparison	AdaBoost	?	Eyes (manual)	$48 \times 48 \times 1$	89.70 ± 0.45
[28]	HOG	SVM	VJ*	Eyes	$48 \times 48 \times 1$	92.26 ± 0.81
[17]	Multi-Gaussian	SVM	VJ*	?	$64 \times 64 \times 1$	92.97
[20]	LBP	SVM	VJ*+Sun* / ori.	5 + 6 Pts	$96 \times 96 \times 1$	93.20 ± 0.92
[2]	HOG	ELM	VJ*	Flow-based*	$100 \times 100 \times 1$	88.20
[59]	CNN	Softmax	?	Face Pts	$90 \times 90 \times 1$	94.60 ± 0.29
[25]	Gabor-HOG	SVM	VJ* / manual	?	$64 \times 64 \times 1$	91.60 ± 0.89
[4]-I	CNN	SVM	Liu*	?	$64 \times 64 \times 1$	92.05 ± 0.74
[4]-II	CNN	SVM	Liu*	.	$64 \times 64 \times 1$	90.60 ± 0.75
[4]-III	CNN	SVM	.	.	$64 \times 64 \times 1$	78.10 ± 0.56
Face-SSD	CNN	Sigmoid	.	.	$300 \times 300 \times 3$	95.76 ± 0.56

* VJ: [51], Liu: [30], Sun: [48], Flow-based: [2]

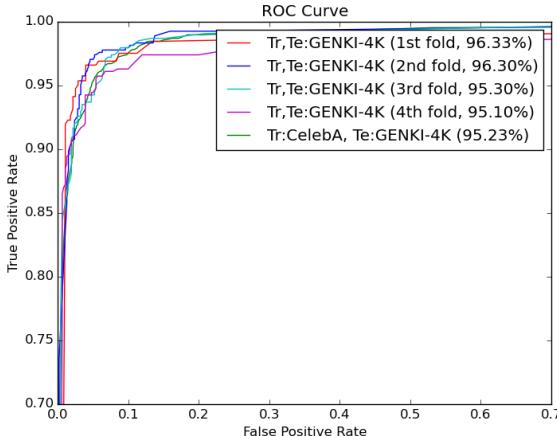


Fig. 7. Receiver Operating Characteristic (ROC) curve for smiling face detection accuracy using GENKI-4K [16] dataset. Tr and Te represent training and testing, respectively.

similar to the testing samples as compared to CelebA dataset. Although the training and testing samples do not overlap, using the same dataset (GENKI-4K) for training helps Face-SSD learn the test sample characteristics of the same (GENKI-4K) dataset. Our four-fold validation results were 96.33%, 96.30%, 95.30% and 95.10%, as shown in Fig. 7. Compared to the accuracies reported by existing works listed in Table 3, our method obtains the best results with mean=95.76% and standard deviation=0.56%.

Although Face-SSD does not require separate steps for face detection and registration, Face-SSD’s smile recognition results rely on the face detection performed in parallel on the same architecture (as explained in Sec. 4.2). Among the existing works listed in Table 3, Chen’s work ([4]-II) reports testing accuracy when the registration process is not used. We therefore compare Face-SSD’s smile recognition performance more closely to the method of Chen ([4]-II). Our experimental results show that Face-SSD outperforms (95.76%) the most recently reported smile recognition result of Chen (90.60%) based on a

Table 4. Comparison to the state-of-the-art methods on the CelebA dataset in terms of accuracy (%) and time (ms.). RP, EB and SS refer to Region Proposal, EdgeBox [65] and Selective Search [41], respectively.

Method	RP	Acc. (%)	Time (ms.)
Liu et al. [30]	EB [65]	92.00	139.00
Ranjan et al. [35]	SS [41]	93.00	3,500.00
Face-SSD	.	92.81	47.28

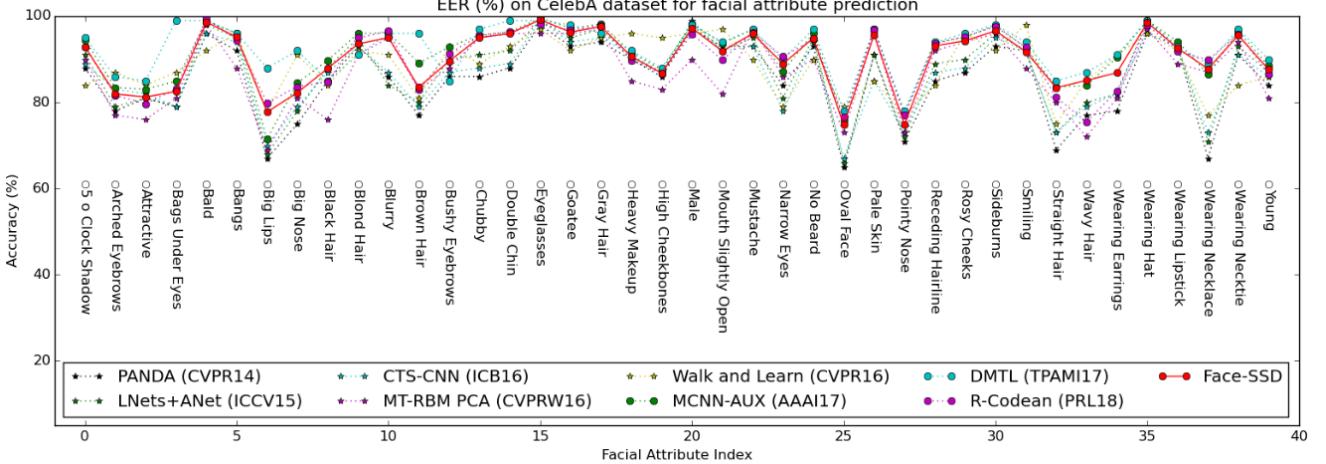
deep learning architecture ([4]-II).

Testing on the CelebA dataset: In the second experiment, we used the CelebA dataset to train and test Face-SSD. In this experiment, we randomly selected 75% of the dataset for training and used the remaining 25% for the testing. We performed several experiments using different combinations of randomly selected training and test samples. Our experimental results show that Face-SSD detects smiling faces accurately (mean=92.81%), similarly to the state-of-the-art methods ([30]: 92.00% and [35]: 93.00%), as shown in Table 4. However, Face-SSD is much faster (**47.28 ms**) than the other methods ([30]: 139 ms, [35]: 3,500 ms) that require region proposal methods for smile recognition (see Table 4).

4.3.2. Facial Attribute Prediction

In this section, we evaluated the performance of attribute prediction using Face-SSD for the prediction of 40 attributes such as gender, age, etc. Our framework treats this problem as multiple binary classification problems using 40 heatmaps at the output layers. The only difference with the smile recognition case is the number of filter kernels used at the final layer – everything else remains the same, including the learning hyperparameters. The effects of modifying various settings during training are presented in Table 6.

Our experiment focuses specifically on the effects of using the Gamma Correction (GC) and Hide-and-Seek (H-a-S) strategies used in the data augmentation process. Depending on the attribute label, there are two possible data augmentation strategies that might affect the accuracy of facial attribute prediction. Gamma correction (colour value adjustment) affects the accuracy of predicting colour-related attributes, such as hair colour



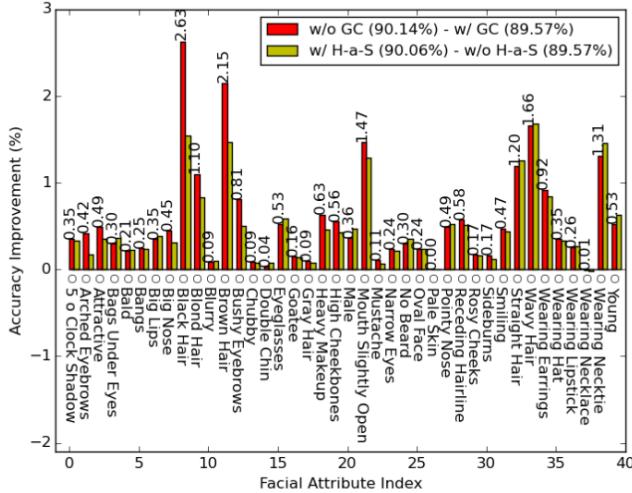


Fig. 9. Removing Gamma Correction (GC) during training Face-SSD (Case C in Table 6) improves the accuracy of predicting color-related attributes comparing to using GC (Case A in Table 6). Using Hide-and-Seek (H-a-S) (Case B in Table 6) does not improve overall face area-related attributes as expected.

Table 6. The effect of using Gamma Correction (GC) and Hide-and-Seek (H-a-S) in the data augmentation process when training Face-SSD for attribute prediction using CelebA dataset.

	Using GC	Using H-a-S	Accuracy (%)
Face-SSD A	✓	.	89.57
Face-SSD B	✓	✓	90.06
Face-SSD C	.	.	90.15
Face-SSD D	.	✓	90.29

rection has had more impact on improving the accuracy of the same attributes, as shown in Fig. 9. The results of the Face-SSD shown in Table 5 are obtained by training Face-SSD using of Hide-and-Seek (Case D in Table 6), but not using Gamma Correction. Although we use the generalised Face-SSD architecture as opposed to using a specially designed architecture for facial attribute prediction, we achieved state-of-the-art accuracy (in the top three among the performances of related works).

4.3.3. Valence and Arousal Estimation

In this section, we investigated the performance of valence-arousal estimation using Face-SSD. Unlike the previous sections that address binary classification (smile recognition) and multi-class recognition (facial attribute prediction) problems, Face-SSD for valence-arousal solves a regression problem. To this end we used a state-of-the-art dataset called AffectNet [32]. AffectNet consists of face images captured in the wild and its corresponding annotations of valence-arousal and emotion. To confirm the regression ability of Face-SSD, we only investigated the valence-arousal estimation performance.

Note that, as AffectNet consists only of cropped face images, we trained Face-SSD using a data augmentation strategy that allows only minor variations in terms of face size. Therefore, during testing, Face-SSD typically handles large faces for valence-arousal estimation. Despite this limitation during train-

Table 7. Experimental results of valence and arousal estimation using AffectNet [32] dataset. Experimental results are reported using Root Mean Square Error (RMSE), Pearson's Correlation Coefficient (CoRR), Sign Agreement Metric (SAGR) and Concordance Correlation Coefficient (CCC) (see [32] for the detailed description of the metrics).

	Valence		Arousal	
	AffectNet	Face-SSD	AffectNet	Face-SSD
RMSE	0.37	0.4406	0.41	0.3937
CORR	0.66	0.5750	0.54	0.4953
SAGR	0.74	0.7284	0.65	0.7129
CCC	0.60	0.5701	0.34	0.4665

ing, however, Face-SSD is able to handle not only large faces but also faces of medium size during testing, as shown in Fig. 1(c).

The performance of the valence-arousal estimation is shown in Table 7. For valence estimation, AffectNet yields slightly better results than Face-SSD. On the other hand, in terms of arousal, Face-SSD provides better results. Overall, Face-SSD provides close to the state-of-the-art performance without any modification to the original architecture of the Face-SSD network. See [32] for a detailed description of the units in the Table 7.

4.4. Computational Speed and Complexity

For all of the Face-SSD applications presented in this paper, we obtained an average processing time of **47.39 ms (21.10 FPS)** during testing, with an experimental environment consisting of an Intel Core i7-6700HQ CPU processor and an NVIDIA GeForce GTX 960M GPU with 23.5GB of DRAM. We used Theano for Face-SSD implementation. As shown in Table 8, most Face-SSD applications achieve near real-time processing speed. Smile recognition (binary classification), facial attribute prediction (40-class recognition) and valence-arousal estimation (multiple task regression) take **47.28 ms (21.15 FPS)**, **47.55 ms (21.03 FPS)** and **47.37 ms (21.11 FPS)**, respectively. Using the proposed generic Face-SSD for face analysis, the number of model parameters indicating complexity does not increase linearly even when the number of facial analysis tasks and classes increases. Although facial attribute prediction performs 40 times more tasks than smile recognition, the processing time by the attribute prediction task increases only by 0.27 ms and requires a small number of additional parameters (0.09 M).

As shown in Table 4, the proposed Face-SSD is significantly faster than traditional methods that use the steps of region proposal and task prediction to analyse faces. For example, the work of Liu et al. [30] requires 35 ms to generate the face confidence heatmap and 14 ms to classify the attributes. In addition, this method requires another 90 ms to find the candidate bounding box (EdgeBox [65]) for localising the final bounding box that ends up with a total processing time of 139 ms (7.19 FPS). The work of Ranjan et al. [35] takes an average of 3,500 ms (0.29 FPS) to process an image. Ranjan et al. [35] explains that the main bottleneck for speed is the process of proposing regions (Selective Search [41]) and the repetitive CNN process for every individual proposal.

Table 8. The total number of parameters and processing time for various face analysis applications using Face-SSD.

Face Analysis Task	Parameter Number	ms (FPS)
Face Detection Part (only)	$2.31 M$	25.57 (39.11)
Smile Recognition	$4.44 M$	47.28 (21.15)
Facial Attribute Prediction	$4.53 M$	47.55 (21.03)
Valence-Arousal Estimation	$4.46 M$	47.37 (21.11)
Average of All Applications	$4.48 M$	47.39 (21.10)

To ensure a fair comparison of the processing times, we should measure the time in the same experimental environment. However, Liu et al. [30] does not provide detailed information about the experimental environment, except that they use GPUs. Ranjan et al. [35] implemented their all-in-one network using 8 CPU cores and GTX TITAN-X GPUs. The processing speed of the proposed Face-SSD is 74 times faster than the all-in-one network, even in a less powerful experimental environment.

Although Face-SSD is faster than other face analysis methods, the processing speed is lower than the base object detection (SSD) model [29] as the complexity of Face-SSD is nearly twice that of SSD, as shown in Table 8. Placing more layers to perform face analysis tasks increased the number of parameters in Face-SSD. However, the structure of the all-in-one network [35] shows that sharing more convolutional features does not degrade the performance of various tasks. Capitalising on this idea, we expect to further reduce the complexity of Face-SSD by sharing more layers and assigning a relatively small number of layers to other face analysis tasks.

5. Conclusions

In this paper, we tackled the problem of multiple face analysis tasks, namely smile recognition, facial attribute prediction and valence-arousal estimation in the wild, without the traditional pre-normalisation steps of face detection and registration. To this end, we proposed Face-SSD which performs face detection and face analysis simultaneously in a single framework. For fast and scale-invariant detection, Face-SSD builds upon the state-of-the-art object detection network SSD. In addition, we used pre-trained parameters of two different networks, trained for object classification and for face detection, to learn the face and task-relevant patterns. Consequently, we built a single framework that enables real-time scale-invariant face analysis in the wild. By exploring various data augmentation strategies for face analysis while maintaining the same Face-SSD architecture, we achieved state-of-the-art performance for various face analysis tasks without increasing model complexity. Our experimental results show that Face-SSD achieves state-of-the-art performance (accuracy of 95.76% for smile recognition, and 90.29% for attribute prediction, RMSE of 0.44 and 0.39 for valence and arousal estimation) while maintaining real-time speed (21.15 *FPS* for smile recognition, 21.03 *FPS* for attribute prediction, 21.11 *FPS* for valence-arousal estimation). For our future work, we plan to investigate a way of using facial attributes to improve the face detection performance. The

challenge for doing this involves using heterogeneous annotations contained in separate datasets.

Acknowledgments

This work has been supported by the Technology Strategy Board, UK / Innovate UK project Sensing Feeling (project no. 102547). This work was undertaken while Youngkyoon Jang was a research associate affiliated with Queen Mary University of London and University of Cambridge.

References

- [1] Agustsson, E., Timofte, R., Escalera, S., Baró, X., Guyon, I., Rothe, R., 2017. Apparent and real age estimation in still images with deep residual regressors on appa-real database, in: FG, IEEE Computer Society. pp. 87–94.
- [2] An, L., Yang, S., Bhanu, B., 2015. Efficient smile detection by extreme learning machine. Neurocomput. 149, 354–363.
- [3] Chang, W., Hsu, S., Chien, J., 2017. Fatauva-net: An integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1963–1971.
- [4] Chen, J., Ou, Q., Chi, Z., Fu, H., 2017a. Smile detection in the wild with deep convolutional neural networks. Machine Vision Applications 28, 173–183.
- [5] Chen, S., Zhang, C., Dong, M., Le, J., Rao, M., 2017b. Using ranking-cnn for age estimation, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- [6] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. ImageNet: A Large-Scale Hierarchical Image Database, in: CVPR.
- [7] Du, S., Martinez, A.M., 2011. The resolution of facial expressions of emotion. Journal of Vision 11, 24.
- [8] Ehrlich, M., Shields, T.J., Almaev, T., Amer, M.R., 2016. Facial attributes classification using multi-task representation learning, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 752–760.
- [9] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. International Journal of Computer Vision 88, 303–338.
- [10] Girshick, R., 2015. Fast R-CNN, in: Proceedings of the International Conference on Computer Vision (ICCV).
- [11] Han, H., Jain, A.K., Wang, F., Shan, S., , Chen, X., 2017. Heterogeneous face attribute estimation: A deep multi-task learning approach. IEEE Trans. on PAMI .
- [12] Hand, E.M., Chellappa, R., 2017. Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification, in: AAAI, AAAI Press. pp. 4068–4074.
- [13] Hao, Z., Liu, Y., Qin, H., Yan, J., Li, X., Hu, X., 2017. Scale-aware face detection, in: CVPR.
- [14] He, P., Huang, W., He, T., Zhu, Q., Qiao, Y., Li, X., 2017. Single shot text detector with regional attention, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22–29, 2017, pp. 3066–3074.
- [15] Hsu, G.S.J., Cheng, Y.T., Ng, C.C., Yap, M.H., 2017. Component biologically inspired features with moving segmentation for age estimation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 540–547.
- [16] <http://mplab.ucsd.edu>, 2009. The MPLab GENKI Database, GENKI-4K Subset.
- [17] Jain, V., Crowley, J.L., 2013. Smile Detection Using Multi-scale Gaussian Derivatives, in: 12th WSEAS International Conference on Signal Processing, Robotics and Automation, Cambridge, United Kingdom. URL: <https://hal.inria.fr/hal-00807362>.
- [18] Jang, Y., Gunes, H., Patras, I., 2017. SmileNet: Registration-Free Smiling Face Detection in the Wild, in: The IEEE International Conference on Computer Vision (ICCV) Workshops.
- [19] Jourabloo, A., Liu, X., 2017. Pose-invariant face alignment via cnn-based dense 3d model fitting. Int. J. Comput. Vision 124, 187–203.

- [20] Kahou, S.E., Froumenty, P., Pal, C.J., 2014. Facial expression analysis based on high dimensional binary features, in: Computer Vision - ECCV 2014 Workshops - Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II, pp. 135–147. doi:10.1007/978-3-319-16181-5_10.
- [21] Kehl, W., Manhardt, F., Tombari, F., Ilic, S., Navab, N., 2017. SSD-6D: making rgb-based 3d detection and 6d pose estimation great again, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pp. 1530–1538.
- [22] Koelstra, S., Patras, I., 2013. Fusion of facial expressions and eeg for implicit affective tagging. *Image Vision Comput.* 31, 164–174.
- [23] Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H., 2011. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization, in: First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies.
- [24] Kossaifi, J., Tzimiropoulos, G., Todorovic, S., Pantic, M., 2017. Afewva database for valence and arousal estimation in-the-wild. *Image Vision Comput.* 65, 23–36.
- [25] Li, J., Chen, J., Chi, Z., 2016. Smile detection in the wild with hierarchical visual feature, in: 2016 IEEE International Conference on Image Processing (ICIP), pp. 639–643. doi:10.1109/ICIP.2016.7532435.
- [26] Li, S., Deng, W., Du, J., 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- [27] Lin, T., Zhao, X., Shou, Z., 2017. Single shot temporal action detection, in: Proceedings of the 2017 ACM on Multimedia Conference, ACM, New York, NY, USA. pp. 988–996.
- [28] Liu, M., Li, S., Shan, S., Chen, X., 2012. Enhancing expression recognition in the wild with unlabeled reference data, in: Computer Vision - ACCV 2012, 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part II, pp. 577–588.
- [29] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C., 2016. SSD: Single shot multibox detector, in: Proceedings of the European Conference on Computer Vision.
- [30] Liu, Z., Luo, P., Wang, X., Tang, X., 2015. Deep learning face attributes in the wild, in: Proceedings of International Conference on Computer Vision (ICCV).
- [31] Lv, J., Shao, X., Xing, J., Cheng, C., Zhou, X., 2017. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- [32] Mollahosseini, A., Hasani, B., Mahoor, M.H., 2017. Affectnet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*.
- [33] Neubeck, A., Van Gool, L., 2006. Efficient non-maximum suppression, in: Proceedings of the 18th International Conference on Pattern Recognition - Volume 03, IEEE Computer Society, Washington, DC, USA. pp. 850–855.
- [34] Poirson, P., Ammirato, P., Fu, C., Liu, W., Kosecka, J., Berg, A.C., 2016. Fast single shot detection and pose estimation, in: 3DV, IEEE Computer Society. pp. 676–684.
- [35] Ranjan, R., Sankaranarayanan, S., Castillo, C.D., Chellappa, R., 2017. An all-in-one convolutional neural network for face analysis, in: 12th IEEE International Conference on Automatic Face and Gesture Recognition FG 2017, Washington, DC, USA, May 30-June 3.
- [36] Ren, S., He, K., Girshick, R., Sun, J., 2016. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.
- [37] Rudd, E.M., Gnther, M., Boult, T.E., 2016. Moon: A mixed objective optimization network for the recognition of facial attributes., in: ECCV, Springer. pp. 19–35.
- [38] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 211–252. doi:10.1007/s11263-015-0816-y.
- [39] Russell, J.A., 2003. Core affect and the psychological construction of emotion. *Psychological review* 110, 145–72.
- [40] Russell, J.A., Barrett, L.F., 1999. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of Personality and Social Psychology* 76, 805–819.
- [41] van de Sande, K., Uijlings, J., Gevers, T., Smeulders, A., 2011. Segmentation as Selective Search for Object Recognition, in: IEEE International Conference on Computer Vision.
- [42] Sariyanidi, E., Gunes, H., Cavallaro, A., 2015. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 1113–1133.
- [43] Sethi, A., Singh, M., Singh, R., Vatsa, M., 2018. Residual codean autoencoder for facial attribute analysis. *Pattern Recognition Letters*.
- [44] Shan, C., 2012. Smile detection by boosting pixel differences. *Trans. Img. Proc.* 21, 431–436.
- [45] Shen, J., Zafeiriou, S., Chrysos, G., Kossaifi, J., Tzimiropoulos, G., Pantic, M., 2015. The first facial landmark tracking in-the-wild challenge: Benchmark and results, in: 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), pp. 50–58.
- [46] Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations (ICLR).
- [47] Singh, K.K., Lee, Y.J., 2017. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization, in: International Conference on Computer Vision (ICCV).
- [48] Sun, Y., Wang, X., Tang, X., 2013. Deep convolutional network cascade for facial point detection, in: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Washington, DC, USA. pp. 3476–3483.
- [49] Sun, Y., Wang, X., Tang, X., 2014. Deep learning face representation from predicting 10,000 classes, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Washington, DC, USA. pp. 1891–1898.
- [50] Tam, G.K.L., Cheng, Z.Q., Lai, Y.K., Langbein, F.C., Liu, Y., Marshall, D., Martin, R.R., Sun, X.F., Rosin, P.L., 2013. Registration of 3d point clouds and meshes: A survey from rigid to nonrigid. *IEEE Transactions on Visualization and Computer Graphics* 19, 1199–1217.
- [51] Viola, P., Jones, M.J., 2004. Robust real-time face detection. *Int. J. Comput. Vision* 57, 137–154.
- [52] Wang, J., Cheng, Y., Feris, R.S., 2016. Walk and learn: Facial attribute representation learning from egocentric video and contextual data, in: CVPR, IEEE Computer Society. pp. 2295–2304.
- [53] Wang, N., Gao, X., Tao, D., Li, X., 2014. Facial feature point detection: A comprehensive survey. arXiv .
- [54] Whitehill, J., Littlewort, G., Fasel, I., Bartlett, M., Movellan, J., 2009. Toward practical smile detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 2106–2111.
- [55] Wu, W., Kan, M., Liu, X., Yang, Y., Shan, S., Chen, X., 2017a. Recursive spatial transformer (rest) for alignment-free face recognition, in: The IEEE International Conference on Computer Vision (ICCV).
- [56] Wu, Y., Gou, C., Ji, Q., 2017b. Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- [57] Zeng, J., Chu, W.S., la Torre Frade, F.D., Cohn, J., Xiong, Z., 2016. Confidence preserving machine for facial action unit detection. *IEEE Transactions on Image Processing* .
- [58] Zhang, C., Zhang, Z., 2010. A Survey of Recent Advances in Face Detection. Microsoft Research, Technical Report .
- [59] Zhang, K., Huang, Y., Wu, H., Wang, L., 2015. Facial smile detection based on deep learning features, in: 3rd IAPR Asian Conference on Pattern Recognition, ACPR 2015, Kuala Lumpur, Malaysia, November 3-6, 2015, pp. 534–538. doi:10.1109/ACPR.2015.7486560.
- [60] Zhang, K., Zhang, Z., Wang, H., Li, Z., Qiao, Y., Liu, W., 2017a. Detecting faces using inside cascaded contextual cnn, in: The IEEE International Conference on Computer Vision (ICCV).
- [61] Zhang, N., Paluri, M., Ranzato, M., Darrell, T., Bourdev, L.D., 2014. Panda: Pose aligned networks for deep attribute modeling, in: CVPR, IEEE Computer Society. pp. 1637–1644.
- [62] Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., Li, S.Z., 2017b. S^3FD : Single Shot Scale-invariant Face Detector, in: The IEEE International Conference on Computer Vision (ICCV).
- [63] Zhang, S., Zhu, X., Lei, Z., Wang, X., Shi, H., Li, S.Z., 2018. Detecting face with densely connected face proposal network. *Neurocomputing* 284, 119–127.
- [64] Zhong, Y., Sullivan, J., Li, H., 2016. Face attribute prediction using off-the-shelf cnn features, in: IEEE International Conference on Biometrics (ICB), pp. 1–7.
- [65] Zitnick, L., Dollar, P., 2014. Edge boxes: Locating object proposals from

edges, in: ECCV, European Conference on Computer Vision.