Applied Intelligence (2022) 52:9320–9333
https://doi.org/10.1007/s10489-021-02848-8

# A robust occlusion-adaptive attention-based deep network for facial landmark detection

**Muhammad Sadiq**[1] · **D. Shi**[1] · **Junwei Liang**[2]

Accepted: 2 September 2021 / Published online: 4 January 2022


**Abstract**

The Internet of Things (IoT) has extensively transformed the industry. The innovation of 5G technology and its rapid growth have enabled fast communication between IoT devices and the cyber domain. Technological advancement and the desire for ease of life have resulted in the development of the concept of smart cities. Security is one of the prime objectives of smart cities. The surveillance video management system is rapidly expanding its scope and applications. The use of 5G technology in smart cities enables the integration of real-time video observations with access to specific locations. This allows facial recognition to detect known criminals or a person of interest in a crowd. Facial landmark detection (FLD) is an essential step in facial attribute analysis, the face recognition pipeline, and face verification. Currently, researchers are focusing on convolutional neural network (CNN) based facial landmark detection approaches, and they have attained substantial advancement. However, occlusion is still the leading cause of difficulty impeding the ability of convolutional neural networks to achieve accurate results. Because attention plays a vital role in the human visual system, the significance regarding rich feature representation in computer vision problems has been recently proved by researchers. In this paper, an occlusion-adaptive attentive deep network (OADN) is proposed for facial landmark detection. In short, we extend our already well-established occlusion-adaptive deep network (ODN) by modifying the geometry-aware module (GM) and distillation module (DM). The results of our experiments show that our proposed model outperforms the current state-of-the-art methods on the available benchmark datasets. It reduces the error from 4.17 to 3.82 for the 300W Full-set dataset. After training on the Menpo dataset, the error decreases to 3.63, this is a 13% decrease in error compared to that of the ODN. In addition, we perform a statistical analysis with a 95% confidence interval to validate the effectiveness of our proposed methodology. Our method reduces the total number of network parameters from 6.6 million to 5.46 million, an approximately 16% decrease in network parameters, effectively reducing the training time and cost. Hence, it is more suitable for scalable data processing. Furthermore, taking advantage of our proposed model's inherently low weight, we also propose a distributive facial recognition model for 5G camera-based surveillance systems.

**Keywords** Facial landmark detection · Cyber-physical system · Attention · Spatial attention · Channel-wise attention · Facial recognition

## 1 Introduction

Technological advancement, efficient management, and the desire for an easy life have enabled the development of smart cities. Security is one of the prime objectives of a smart city. Regarding physical security, cyber-security biometrics is one of the most reliable and robust methods for human identification. The objective is to monitor the city, classify trouble spots and persons, and take protective and corrective measures [1–3]. This method increases the need for 24/7 video surveillance of streets, public places such as passenger stations (bus/train/airport, etc.), shopping centres, etc., and the need to logically examine them to detect criminal activities. Even the ability to identify individuals in a crowd has become necessary. 5G in smart cities enables the integration of real-time video observations with access to specific locations [4, 5]. All of the above allows the use of facial recognition to detect known criminals or a person of interest in a crowd. IoT-based devices such as 5G video surveillance cameras can also be connected to a cyber-

✉ D. Shi
dshi@szu.edu.cn

Extended author information available on the last page of the article.

domain-enabled cyber-physical system (CPS). Increasing demand for face recognition technology in the cyber-security and physical security domains, such as in cyber-physical systems [6–12] renders face-based authentication and authorization as very attractive due to their high accuracy and ease of use. Facial landmarks are generally used as part of a pre-processing step of facial recognition for facial alignment. Incorrect landmark locations indicate incorrect semantic alignment among the faces or features, which can further result in matching or classification inaccuracies. Perfect alignment is often challenging without perfect facial landmark detection, as mention in Fig. 1. The detection of the landmarks is very challenging when the face is occluded as mentioned in Fig. 2, which affects the overall performance of the system. Accurate landmark detection ultimately helps to improve the accuracy of facial detection.

Previously, to address occlusion, we proposed the ODN [14]. The ODN consists of three modules: a GM to capture the geometric relation among facial components, a DM to model occlusion, and a low-rank learning module (LM) to recover missing features. The ODN works very well on profile facial images as mentioned in Fig. 3, but the progress in terms of extreme poses and expressions is not satisfactory because the geometrical structure helps the network recover missing features of profile photos but misleads the network in the case of extreme poses and expressions. In the case of exaggerated poses and emotions, the structure misleads the network into learning missing features. Second, the ODN uses two separate subnetworks to model occlusion and improve feature representation through attention, which is computationally inefficient and expensive. To address this problem, in this paper, we proposed the OADN by modifying the geometry-aware module (GM) and distillation module (DM) of the ODN. We replaced the GM and DM with an attention module.

To be more specific, our contributions can be summarized as follows: i) We modified the structure of the ODN in terms of model occlusion and capture the global facial appearance to achieve better performance. ii) To the best of our knowledge, we are the first to introduce the channel-wise attention (CA), and spatial attention (SA) for FLD to model occlusion and obtain rich feature representations simultaneously. iii) Our method reduces the total number of network parameters, which effectively reduces the training time and cost; hence, it is more suitable for scalable data processing. Furthermore, taking advantage of the inherently low weight of our proposed model, we proposed a distributive facial recognition model for 5G camera-based cyber-physical surveillance systems. The results of our experiments show that our proposed model outperforms the current state-of-the-art methods on the available benchmark datasets.

The remainder of the paper is organized as follows. Related work is elaborated in Section 2. We elaborate our proposed Occlusion-adaptive Attentive Deep Network (OADN) solution in Section 3. Detailed experiments of our proposed framework are spelled-out in Section 4. Section 5 is about Application for 5G Camera based Cyber-Physical Surveillance Systems. Section 6 draws the conclusion of this paper.

## 2 Related work

Facial points can be defined as the predetermined indication points on a given face graph. Mainly, these points are placed around the familiar components of the face, e.g., ear, eyes, nose, mouth, and chin. Usually, these points are located around or centre on some common facial components. The tasks related to facial analysis can differ based on the numbers, types, and required quantity of landmarks and the use of these landmarks. Localization of these landmarks has been done for facial analysis tasks, and it has drawn more attention from researchers during the last decade due to its importance. FLD is a key step for many facial analysis tasks, e.g., facial action unit detection, face recognition, face expression detection, face frontalisation, head pose estimation and 3D face modelling [16] etc. The objective behind FLD is to identify some precisely predefined key markers on facial components. There are several challenges regarding FLD, e.g., occlusion, illumination, expressions, and extreme poses.

Existing FLD methods can be categorized generally into three main groups: regression-based methods, template-based methods, and deep-learning-based methods. Regression-based techniques learn the mapping from the facial image appearance to landmark locations, and unlike template-based methods, regression-based methods do not usually build global shape models [17].

Regression-based methods predict all facial landmarks jointly, but the shape constraint and structure information are learned through the process [14]. Template-based

**Fig. 1** An example of misalignment from [13] by incorrect facial landmarks

**Fig. 2** Examples of the occlusion by glasses, food, masks, hair, and hands, etc. From COFW dataset [15]. It can be easily observed that it is difficult to identify facial landmarks in presence of occlusion

methods leverage information about the facial appearance and global facial shape, controlling facial appearance and shape variations through statistical models [17]. These models learn a parametric shape model through a dataset that is already labelled to model the changes in facial shape using principal component analysis (PCA). Furthermore, PCA is used to build the global facial shape and facial appearance. This entire process helps to refine the fitting algorithm. Some notable examples are active shape models (ASMs) [18], "Face detection, Pose estimation, and Landmark Localization"; (FPLL) [19], active appearance models (AAM) [20], and discriminative response map fitting (DRMF) [21]. The drawback of this kind of modes is, the reconstruction error effects whole face under occlusion, and as a result, all this leads model in hard circumstances, being unable to locate facial landmarks.

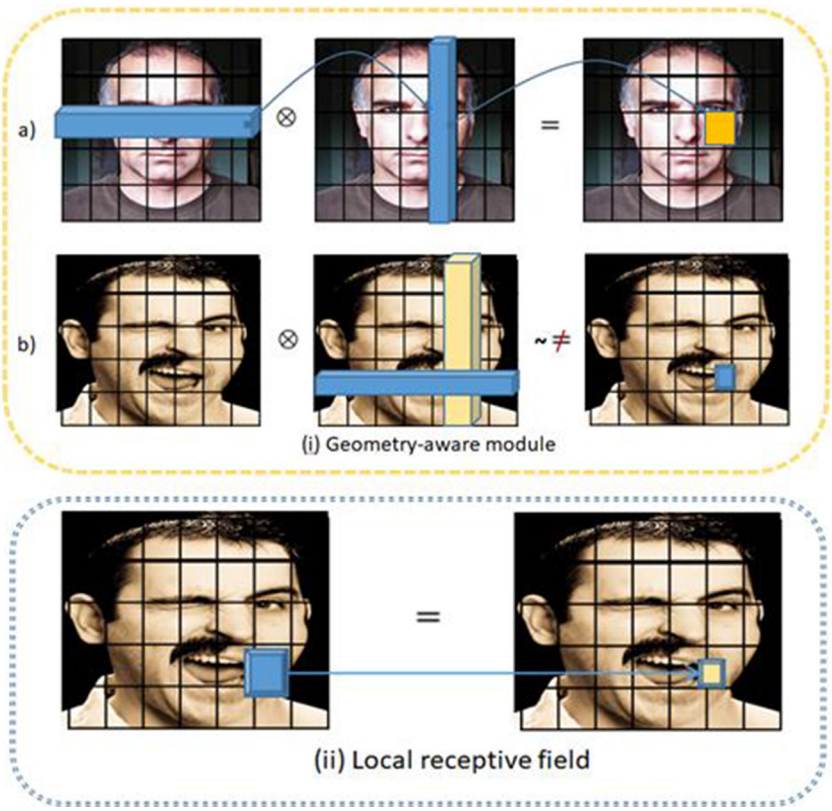



**Fig. 3** Feature extraction structure of Geometry-aware Module (GA) and Local receptive filed

To solve computer vision problems, deep learning (DL)-based methods are getting a prominent place. Inspired by the popularity and effectiveness of DL based methods, FLD researchers also started to apply DL techniques to solve FLD related problems [22–28]. In comparison to conventional methods, DL based methods gained higher performance [29, 30]. Lately, CNN based models gained a prominent place in DL based solutions to deal with FLD related tasks. But, to deal with occluded faces, is yet a challenge for CNN as well [14, 31], it is because of the decrease in localizing accuracy due to occlusion. When the face is partially occluded, it is still challenging to improve localizing accuracy because occlusion probably deceives CNN during the learning of features.

To address the occlusion problem, it is important to identify the occlusion and model it. This task is very challenging because the occurrence of occlusion is random, irregular, and complex, as shown in Fig. 2. In the literature, there have been several attempts [14, 30–34] to solve the occlusion problem. Wu et al. [32] proposed a supervised regression method to update the probabilities steadily of landmark visibility at each iteration. In 2016 Liu et al. [33] proposed adaptive cascade regression besides of adaptive exemplar-based shape model to estimate the occlusion level of each landmark. Later, Xing et al. [34] proposed an occlusion dictionary into the already existing face appearance dictionary. The occlusion dictionary is learned and updated automatically in a data-driven manner. Recently, in 2019 Zhu et al. [30] proposed BCNN-JDR. In BCNN-JDR, each part of the face is treated with a separate pipeline. The objective is to share minimum information with other components pipelines to avoid correlative impact. As different facial components have a different number of facial points as well as different levels of hardness to predict facial points. It is challenging to calculate the unbiased hardness due to a lack of balance benchmark dataset. For example, if during the prediction of mouth hardness if the dataset has more or fewer number of images having different expressions, occlusions, poses than those of other parts.

Attention plays a very significant role in capturing long-range dependencies. Usually, the objective behind attention is to tell the network precisely where to focus on by calculating the response for a particular location as a weighted sum of the features at all positions [35, 36].

Attention plays a vital role in the human visual system [37–39]. To improve the performance of CNNs by using attention, several attempts have been made in the literature. Huifang Li et al. [40] proposed the spatial alignment network (SAN) for FLD based on the hand-crafted method and learning-based method. The basic target problem of SAN is to deal, appearance and spatial variations. But the problem with SAN is, if it uses a handcrafted method, the efficiency is very low. If it uses a learning base method, it is not steady. To improve network performance Attention Alignment Network (AAN) [41] and Joint AU detection and face Alignment Network (JAA-Net) [42] also use spatial attention for FLD. As we already discussed, the purpose of SA is to tell network 'where' to focus, but still, the network is not aware of 'what' to focus. In image classification tasks CA and SA [31, 43–45] can yield a significant improvement.

## 3 Occlusion-adaptive attentive deep network

### 3.1 Occlusion-adaptive deep network (ODN)

The overall ODN structure can be divided into three modules. The LM is followed by the DM and GM. A shared structure matrix is generated by the GM and DM to help the LM recover missing features. The purpose of the GM is to capture the geometric structure of the facial image. The DM is used to model the occlusion probability. In the ODN, the occlusion probability is based on high-level features. In regular scenarios, irregular positions of the camera affect facial images by spatial and appearance variations because the images are collected in the wild [46].

We observe that the distillation module does not have enough rich feature representation and is not able to focus on occluded parts. Second, the GM works well on profile faces; in the case of extreme poses and expressions, it misleads the network into recovering missing features, as mentioned in Fig. 3. Motivated by the success achieved by deep learning, we propose in this paper a novel framework OADN to address this issue. In the OADN, we introduce the attention module instead of the distillation module. The experimental results show that our model performs better than the current state-of-the-art methods on benchmark datasets.

To be more specific, to obtain the OADN, we modify the last residual block of ResNet-18 [47]. The detailed structure of OADN is as per Fig. 4. For simplification, the OADN structure can be divided into two modules. Attention module (AM), and low-rank learning module (LM). AM consists of three subnetworks. To capture holistic facial features, we use subnetwork-A, consisting of three $3 \times 3$ convolutional layers with two short connections, aiming to assemble facial features more accurately in reverse order. To model occlusion, we use subnetwork-B. Subnetwork-B exploits a residual block to implement CA and SA. The main objective is to model occlusion more precisely and obtain a rich representation of facial features.The aim of subnetwork-C is to avoid input signal decay to obtain stable features. Furthermore, the output of B and C is integrated

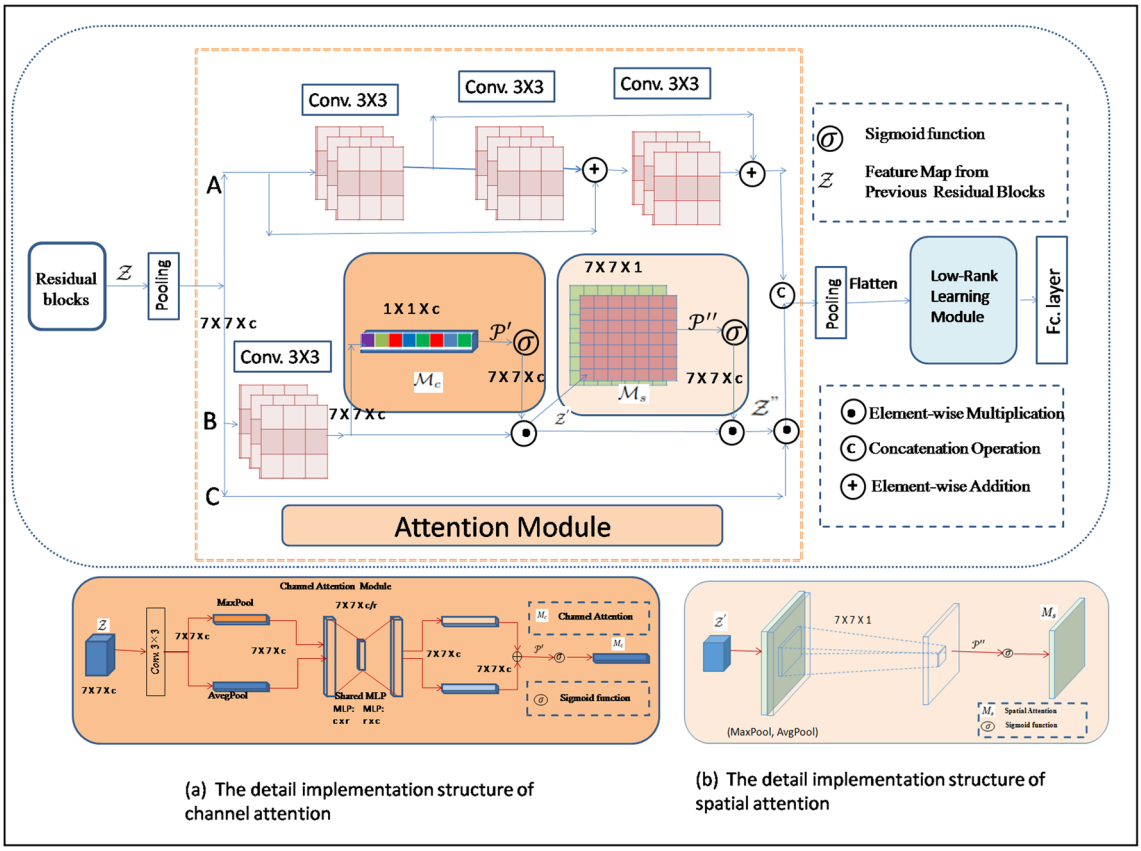



**Fig. 4** The diagram of the proposed Occlusion-adaptive Attentive Deep Network with detail implementation of channel attention, and spatial attention

by element-wise multiplication to assign small weights to the background and occluded parts.Finally, as the output from subnetwork-B and subnetwork-C, for a holistic face, a weighted feature map (clean features) can be obtained. Finally, the output from subnetwork-B and subnetwork-C is concatenated with the output of subnetwork-A to obtain a high-dimensional single-feature map to obtain a hybrid feature map of the facial graph. Furthermore, these hybrid feature maps are used as the input of the LM after down sampling and flattering. The objective of concatenation is to obtain a hybrid feature representation of facial appearance. The LM recovers the missing features of the face by modelling the inter-feature correlation. Mathematically in OADN, the given training set $\left\{\left(I_i, \check{S}_i\right)\right\}$ can be learned by (1).

$$min \frac{1}{N} \sum_{i=1}^{N} \left\| \check{S}_i - S \right\|_F^2 + \beta Rank\left(\mathcal{M}\right) \tag{1}$$

Where $\check{S}$ and S represents ground-truth and corresponding prediction. $\check{S} = \{s_1, s_2, ....s_L\}$ and $S = W_{fc}^T \mathcal{M}^T \mathcal{X}$. The output of the geometry-aware module is denoted as $\mathcal{X}$. L and s are the numbers of landmarks and facial landmarks, respectively. β is used as a regularization factor to adjust the rank of $\mathcal{M}$. $W_{fc}$ means, the parameters of a fully connected layer. The OADN trained in an end-to-end manner the same as ODN.

## 3.2 Attention module

As already discussed, feature representation plays a significant role in modelling occlusion more precisely and learning missing features more proficiently. The effectiveness of CA and SA for image classification related tasks has already been determined by [43–45]. Inspired by their work, we incorporated CA and SA to obtain rich feature representation and model occlusion more accurately. The detailed structure and implementation details of CA and SA are shown in Fig. 4a, and b, respectively. In OADN the attention module consists of subnetwork-B, and subnetwork-C. The subnetwork-B exploits a residual block to implement CA and SA, respectively. The main goal is to model occlusion more accurately and get the rich representation of facial features. The aim of subnetwork-C is to avoid input signal decay, to obtain stable features. Overall refined feature map can be defined as (2).

$$\mathcal{F} = \mathcal{Z}^{"} \bullet \mathcal{Z} \tag{2}$$

Where $\mathcal{F}$ represents the final feature map after merging the residual map and the attention map. The $\mathcal{Z}^{"}$ is a feature map obtained through the attention process, and $\mathcal{Z}$ represents the feature map of previous residual blocks. In simple words, $\mathcal{F}$ is output of element-wise multiplication between $\mathcal{Z}^{"}$ and $\mathcal{Z}$.

### 3.3 Channel-wise attention and spatial attention

Usually, researchers increase the width or depth of the corresponding network to obtain better feature representation during network engineering. In deep networks, the increase in network parameters affects the efficiency in terms of cost and time. Furthermore, the assembling of features accurately in reverse order is also a questing mark. To deal with this issue [43, 45] used attention and proved its effectiveness for image classification tasks. We also used CA to guide the network 'what' is meaningful in a given facial image, and SA guide the network 'where' to focus. The objective behind this attempt is to ensure the sensitivity of network to informative features. To simplify, CA and SA can be written, as mentioned in (3) and (4), respectively.

$$\mathcal{Z}' = M_c(\mathcal{Z}) \bullet \mathcal{Z} \tag{3}$$

$M_c$, $\mathcal{Z}$ denotes the channel-wise attention map, and the feature map of prior residual blocks, respectively. In short, $\mathcal{Z}'$ can be obtained by element-wise multiplication of $M_c(\mathcal{Z})$ and $\mathcal{Z}$

$$\mathcal{Z}^{"} = M_s(\mathcal{Z}') \bullet \mathcal{Z}' \tag{4}$$

$\mathcal{M}_s$ is the spatial attention map and $\mathcal{Z}'$ is the feature map obtained by channel-wise attention. The objective of CNN is to extract the features from a given image. If an image $W \times H \times 3$ passes over convolutional layer with C channels. CNN uses filters to scan the given image and produce a $\acute{W} \times \acute{H} \times c$ feature map as output, it can be input of further convolutional layers.

To get a channel attention map, we squeezed the spatial dimension of the input feature map. To get distinct features of the given object to refine CA, we used max-pooling along with average-pooling. As max-pooling guides network more precisely towards more distinct features. Mathematically, the channel attention map can be as per (5).

$$\begin{aligned} M_c(\mathcal{Z}) &= \sigma(MLP\,(AvgPool\,(\mathcal{Z})) \\ &\quad + MLP\,(MaxPool\,(\mathcal{Z}))) \\ &= \sigma(W_1\left(W_0\left(\mathcal{Z}^c_{avg}\right)\right) + W_1\left(W_0\left(\mathcal{Z}^c_{max}\right)\right)) \end{aligned} \tag{5}$$

We accumulated spatial information of feature maps through max-pooling and average-pooling to obtain channel attention. Later, two separate context descriptor: $\mathcal{Z}^c_{Max}$ and $\mathcal{Z}^c_{Avg}$, represents max-pooled and average-pooled features correspondingly. As mentioned in Fig. 4a, a shared network with both descriptors generates channel attention map $M_c \in \mathcal{R}^{1\times1\times c}$. The shared network consists of Multi-Layer Perceptron (MLP) with one hidden layer. We used hidden activation size $\mathcal{R}^{1\times1\times\frac{c}{r}}$ to reduce parameter overhead, where 'r' stands for reduction ratio, and 'c' is number of channels. We combined the output feature vectors by element-wise addition after applying the shared network to each descriptor. Where $\sigma$ means the sigmoid function and $W_0 \in \mathcal{R}^{\frac{c}{r}\times c}$ and $W_1 \in \mathcal{R}^{c\times\frac{c}{r}}$, where $W_0, W_1$ means the shared weights of MLP.

We used the inter-spatial relationship of features to obtain spatial attention map. The objective of SA is to direct the network 'where' to emphasize more specifically, it is corresponding to CA. We compute SA by applying pooling beside the channel axis. The detailed implementation structure of SA is illustrated in Fig. 4b. We used max-pooling and average-pooling besides channel axis and concatenated both to obtain feature descriptors as mentioned in (6), and Fig. 4b. Furthermore, to attain spatial attention map $M_s(\mathcal{Z}) \in \mathcal{R}^{H\times W}$ we use a convolutional layer to direct the network 'where' to emphasize. We combined channel information by using max-pooling and average-pooling operations, to generate two 2D maps:$\mathcal{Z}^s_{avg} \in \mathcal{R}^{H\times W\times1}$ and $\mathcal{Z}^s_{max} \in \mathcal{R}^{H\times W\times1}$.

$$\begin{aligned} M_s(\mathcal{Z}) &= \sigma\left(f^{7\times7}([AvgPool(\mathcal{Z}); MaxPool(\mathcal{Z})])\right) \\ M_s(\mathcal{Z}) &= \sigma\left(f^{7\times7}([\mathcal{Z}^s_{avg}; \mathcal{Z}^s_{max}])\right) \end{aligned} \tag{6}$$

## 4 Experiments

In this section, we broadly assess the performance of the proposed framework on different benchmark datasets under various settings for the task of FLD. First, in Section 4.1, we present optimization details of our proposed OADN. In Section 4.2, we present some initial details about the benchmark datasets and used experimental settings, Section 4.3, elaborates the evaluation metric, and employment details for the training of OADN. Following that, we investigate the effect of several system parameters as well as the contribution of the different components to the FLD performance. In Section Section 4.4, we present the ablation study to validate our framework. We resized and cropped all images (224 × 224) and perform a flip, scale, rotation tasks and translation to do the data augmentation for the training set. Same as ODN, we also pre-trained all our models on the ImageNet dataset [48].

### 4.1 Optimization

Mathematically the OADN can be formulated as following minimization problem:

$$min \frac{1}{N} \sum_{i=1}^{N} \left\| \check{S}_i - S_i \right\|_F^2 + \beta \mathcal{R}ank(\mathcal{M}) + \gamma \left\| \mathcal{M} \right\|_F^2 + \alpha \left\| \mathcal{W}_c \right\|_F^2 + \lambda \left\| \mathcal{W}_{fc} \right\|_F^2 + \eta' \left\| \mathcal{P}_i' \right\|_F^1 + \eta'' \left\| \mathcal{P}_i'' \right\|_F^1, \quad (7)$$

Where $\mathcal{S} = \mathcal{F}_{OADN}(\mathcal{I}; \mathcal{W}_{fc}; \mathcal{M})$, and $\mathcal{F}_{OADN}(.)$ is our proposed OADN. $\mathcal{P}_i'$, and $\mathcal{P}_i''$ are feature map of channel attention and spatial attention respectively. $\mathcal{W}_c$ represents the parameter set of convolutional layer, and $\mathcal{W}_{fc}$ represents the fully connected layer. $\mathcal{M}$ is the parameter set of LM. Frobenius norms control the shrinkage of all three given parameter sets with the connected parameters $(\alpha, \gamma, \lambda)$, respectively. Feature map $\mathcal{P}$ from attention module with parameter $\eta$ is imposed by $L_1$. In (7), $\lambda, \gamma, \alpha$ are set to $\left(1 \times 10^{-5}\right)$, $\eta$ and $\beta$ are set as $\left(1 \times 10^{-6}\right)$. The value we used for $\sigma$ in our case is 0.5 for optimization of CA and SA.

### 4.2 Datasets

We used different benchmark datasets: 300W [49], AFLW [50], COFW [15], Menpo [51] and 300VW [52], to broadly assess the performance of the proposed framework under various settings for the task of FLD. All these datasets are benchmark datasets and publicly available for research purposes. We compare outcomes of our method with contemporary methods [14, 29–31, 41, 53, 54]. We trained our model on a 300W training set and tested our model on other datasets. A summary of datasets used for our experiments is given in Table 1.

- 300W is a re-known, widely used, and publicly available dataset for FLD, to measure the effectiveness of the method. It contains 3,837 images from the common standing datasets: AFW [19], LEN [55], LFPW [56] with annotation of 68 landmarks. To train our proposed model, we divided the 300W dataset into two parts – one for training and the other for testing purposes. For training purposes, we used 3,148 images, rest 689 images used for testing. We divided the testing samples into three further subgroups: (a) Common-set, having 554 images (330 images of HELEN and 224 images of LFPW dataset); (b) Challenging-set, having 135 images taken from IBUG dataset; (c) Full-set having, all 689 images of the testing part.
- COFW is a very famous publicly available dataset having a total of 1852 images (1345 images are for training purposes, and the rest of 507 are for testing purposes). We used COFW to measure the performance of our proposed model. So, we just used its testing part. We used the re-annotation of [57] (68 landmarks) because originally annotated with 29 subcategories.
- 300VW is a publicly available dataset with 114 videos. All Videos are extracted into corresponding frames and annotated with 68 landmarks. We divided the 300VW dataset into two parts – training and testing. For training purposes, we used 50 videos, rest 61 used for testing. We the testing samples into three further subcategories.
- Menpo dataset consists of 5,658 semi-frontal and 1,906 profile facial images for training. For testing purposes, it consists of 5,335 frontal and 1,946 profile facial images. The training set is publicly available, but testing data is not publicly released yet. Profile facial images are annotated with 39 profile landmarks, and 68 landmarks are used for near-frontal faces. We use face images, annotated with 68 landmarks for our training. Due to the unavailability of menpo test data, we used other publicly available datasets for testing purposes.

### 4.3 Evaluation metric and implementation details

We adopted two evaluation methods to evaluate the performance of OADN: the Cumulative Error Distribution (CED) curve [14, 29–31, 41, 53, 54], and Normalized Root

**Table 1** The summary of used dataset for performance evaluation

| | Daata set | Total images/Videos | Training images/Videos | Testing images/Videos | Size | Purpose | |
|---|---|---|---|---|---|---|---|
| | | | | | | Train. | Test. |
| 300W | Full-Set | 3837 | 3148 | 689 | 224 x 224 | Yes | Yes |
| | Common-Set | – | – | 554 | 224 x 224 | – | Yes |
| | Challenging-Set | – | – | 135 | 224 x 224 | – | Yes |
| COFW | 1852 | 1345 | 507 | 224 x 224 | – | Yes | |
| Menpo | 14845 | 7564 | 7281 | 224 x 224 | Yes | – | |
| 300VW(Video) | 114 | 50 | 61 | 224 x 224 | – | Yes | |

Mean Squared Error (NRMSE) [14, 25, 29–31, 53] . The NRMSE can be illustrated as:

$$NRMSE = \frac{1}{N}\sum_{i=1}^{N}\frac{\left\|\check{S}_i - S_i\right\|_2}{L\Omega_i} \tag{8}$$

where L represents the number of landmarks, and $\Omega$ is inter-ocular distance. In our case, $\Omega$ is the width of the bounding box of the AFLW dataset. We used different settings of parameters to measure the effectiveness, such as reduction ration r = 16, 32, 64. After detailed analysis, we found r = 64 is best in our case to have better results. We also tried different combinations of CA and SA and found better performance sequentially. For CED, we used ODN and other most recent methods as our benchmark to compare our results because these methods have proved their significance already in comparison to other methods. Secondly, we generated all CED curves in real-time on same data. Other results are taken from ODN as reference. We used all other required parameters the same as used in ODN.

#### 4.3.1 Empirical analysis under normal circumstances

To evaluate our proposed method under normal circumstances, we used two benchmark subsets of 300W (Common-set, and Full-set). Both subsets are benchmark datasets and have very fewer variations in illumination, pose, and occlusion. Table 2, consists of comparison results in terms of NRMSE $\left(\times 10^{-2}\right)$. We compared our results with the current best models. We can see the results for the Common-set; OADN improves result upto 3.22, where as it is 3.56 in ODN. Which proves the significant change of results for the Common-set. We can see in the CED curve OADN has significant improvement over other current state-of-the-art methods. The same as the Common-set, we analyzed the proposed model for the Full-set. Substantial change can be seen in Table 2, for the Full-set also. Figure 5 is about the CED curve for the Full-set and has vigilant improvement in results in comparison to other given methods.

**Table 2** The NRMSE $\left(\times 10^{-2}\right)$ comparison results on Common-set and Full-set of 300W

| Method | Year | Common-set | Full-set |
|---|---|---|---|
| Seq_MT [53] | 2018 | 4.20 | 4.90 |
| PCD-CNN [29] | 2018 | 3.67 | 4.44 |
| BCCN-JDR [30] | 2019 | 3.68 | 4.36 |
| **ODN** [14] | **2019** | **3.56** ±0.0172 | **4.17** ±0.0261 |
| ADN [31] | 2019 | 3.52 | 4.14 |
| LGSA [25] | 2020 | 3.36 | 4.06 |
| **OADN** | **2020** | **3.22** ±0.0136 | **3.82** ±0.0213 |
| **OADN-Menpo** | **2020** | **3.12** ±0.0128 | **3.63** ±0.0197 |

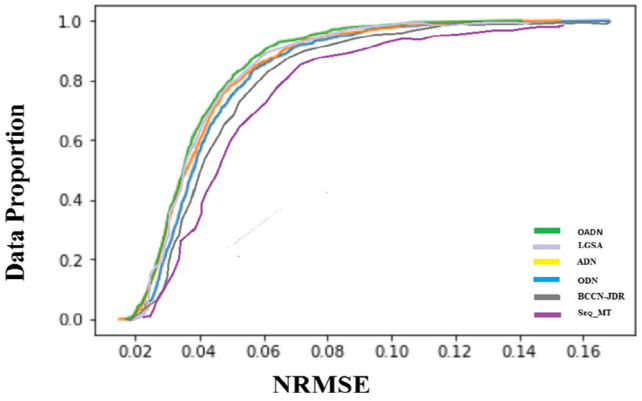



**Fig. 5** The CED curve on 300W Full-set

#### 4.3.2 Empirical analysis for robustness against occlusion

It is hard to deal with occluded faces than regular faces. We performed several experiments to check the robustness of OADN against occlusion. We used two diverse benchmark datasets: the Challenging-set of 300W, and COFW to measure the effectiveness of OADN against occlusion.

The comparison results for challenging set are given in Table 3. We compare our results with the current state-of-the-art methods. It can be easily observed through the comparison table, OADN performs better and improves $\left(\times 10^{-2}\right)$ from 6.60 to 6.23, which is a significant improvement for any model. The results of the COFW are in Fig. 6. The results of OADN for COFW are also awe-inspiring and have a significant improvement. Figure 7 is about the CED curve of Challenging-set and COFW, respectively. Significant improvement in both CED curves can also be seen.

#### 4.3.3 Empirical analysis for videos

To measure the effectiveness of our model, we use the 300VW dataset, which is a benchmark dataset for FLD

**Table 3** The NRMSE $\left(\times 10^{-2}\right)$ comparison results on Challenging set of 300W

| Method | Year | Challeenging-set |
|---|---|---|
| DSRN [58] | 2018 | 9.68 |
| SBR [54] | 2018 | 7.58 |
| SAN [59] | 2018 | 7.55 |
| BCCN-JDR [30] | 2019 | 7.16 |
| **ODN** [14] | **2019** | **6.67** ±0.0107 |
| ADN [31] | 2019 | 6.60 |
| HORNet [60] | 2020 | 6.36 |
| OADN | **2020** | **6.23** ±0.0084 |
| OADN-Menpo | **2020** | **6.11** ±0.0079 |

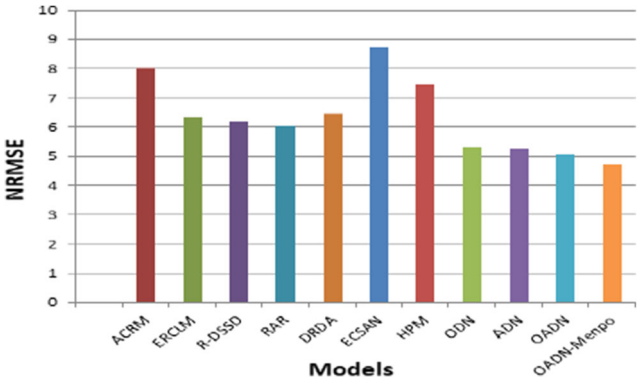



**Fig. 6** The NRMSE $(\times 10^{-2})$ comparison results on COFW dataset

on video. We used the testing part or 300VW, as already mentioned. For the training of our model, we used 300W for OADN. Experimental results of all three categories in comparison to other current state-of-the-art methods are shown in Table 4. It can be easily observed that our proposed method outer performs in comparison with other methods. For Cat. 1, it improves performance from 4.75 to 4.63 for OADN. Same as Cat. 1, for Cat. 2, and Cat. 3, the improvement in performance on video dataset is significant, which proves the significance of our proposed method over other methods in terms of all three categories.

### 4.4 Ablation study

After implementing the proposed changes, the OADN has a significant difference in the number of network parameters; it reduced the number of parameters from 6.60 million to 5.46 million, which is comparatively very less. This is also helpful in minimizing the computation time and cost, especially in the case of scalable computing. Figure 8 shows the comparison based on the number of network parameters in millions between OADN, ADN, ODN, and CU-Net-8 [62]. We just selected the most recent state-of-the-art methods for comparison purposes. Figure 9 is about the

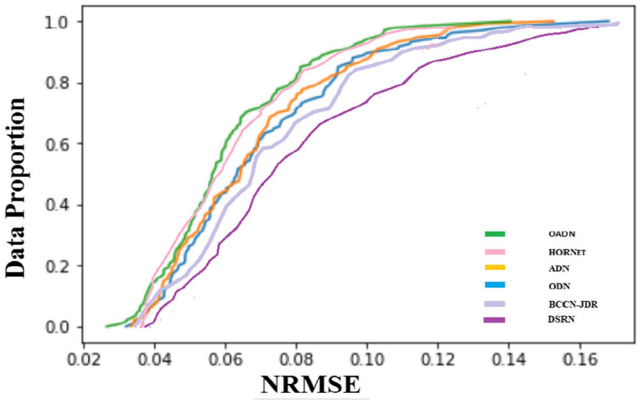



**Fig. 7** The CED curve on 300W Challenging-set

**Table 4** The NRMSE $(\times 10^{-2})$ comparison results on 300VW video dataset for all three categories

| Method | Year | Cat.1 | Cat.2 | Cat.3 |
|---|---|---|---|---|
| TSTN [61] | 2017 | 5.36 | 4.51 | 12.84 |
| AAN [41] | 2018 | 5.03 | 4.82 | 7.98 |
| ADN [31] | 2019 | 4.75 | 4.34 | 6.72 |
| OADN | **2020** | **4.63** | **4.20** | **6.61** |
| OADN-Menpo | **2020** | **4.46** | **4.06** | **6.53** |

qualitative detection results of the proposed approach for some sample faces from the 300W full dataset. The ground-truth landmarks are marked in green color (top row), while our predicted landmarks with estimated NRMSE are in blue color (bottom row). We used different settings of parameters to measure the effectiveness, such as reduction ration r = 16, 32, 64. After detailed analysis, we found r = 64 is best in our case to have better results.The importance of each module in illustrated in Table 5. We performed all experiments on Pytorch [63] DL framework and Nvidia's K80 platform. The average FLD time for OADN per image is $\sim$ 3 ms per image with Core(TM) i7 CPU@3.40GHz and 8GB of RAM. The speed of our proposed OADN clearly illustrate the realtime requirement satisfaction.

## 5 Application for 5G camera based cyber-physical surveillance systems

As per Cisco forecast, the Global IP video traffic will be 82 percent of all consumer internet traffic by 2021 [1]. As mentioned earlier, technological advancement and the desire for ease of life enable the concept of smart cities.Security is one of the prime objectives of a smart city. City-wide video surveillance systems applied with video analysis technology are increasing to enhance the efficiency of city monitoring and prompt field response as well as concurrent identification of persons using an intelligent surveillance system. The objective is to monitor the city,

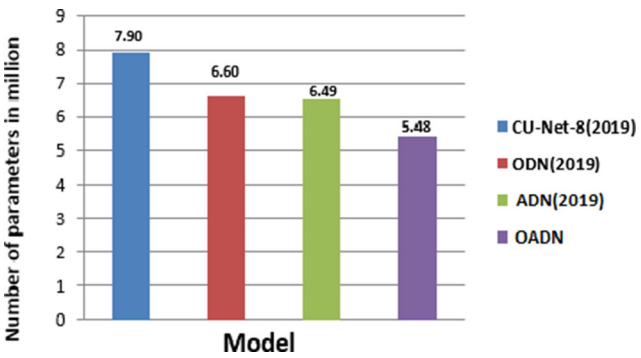



**Fig. 8** The comparison of number of total parameters in millions

**Table 5** The ablation analysis on 300W Challenging set

| Model | NRMSE |
|---|---|
| BRNet | 7.21 |
| BRNet+LM | 6.90 |
| BRNet+AM+LM (without L1 ) | 6.37 |
| BRNet+AM+LM(r=32) | 6.27 |
| BRNet+AM+LM(r=16) | 6.25 |
| BRNet+AM+LM(r=64) | 6.23 |

classify trouble spots and persons, and take protective and corrective measures. It enables the need for 24/7 video surveillance of streets, public places such as passenger stations (bus/train/airport, etc.), shopping centres, etc., and logically examining them to detect criminal activities. Even the ability to identify individuals in a crowd becomes necessary. There is also a need to integrate these data with data from other sources (such as driver's licences, passports, and student IDs) for comparison and to take quick action. The surveillance video management system is rapidly expanding its scope of application at the request of citizens and the development of related technologies.

With the help of advanced technologies, such as cloud computing, mobile edge computing and artificial intelligence, the scope and function of surveillance systems are being enhanced to meet the needs of intelligent surveillance systems beyond simple monitoring.5G in smart cities enables the integration of real-time video observations with access to specific locations. This allows the carrying out of facial recognition to detect known criminals or a person of interest in a crowd. The facial recognition system is one of the prime and reliable models for biometric authentication and authorization. Face alignment is an essential pre-processing step of the facial recognition pipeline. It is most important and challenging to align face images appropriately with either the reference faces or a pre-defined general face model with the help of landmarks. The detection of landmarks is very challenging when the face is occluded, which affects the overall performance of the system. Accurate landmark detection ultimately helps to improve the accuracy of facial detection.

In traditional systems, each entity has separate isolated surveillance systems, such as banks, shopping malls, and offices. 5G technology makes it possible to connect all systems through a central database to monitor the whole city very intelligently. Connecting thousands or millions of cameras to a central server is also very challenging because it affects the performance of the system and the processing cost. It becomes more complicated when all processing is performed at a central place. Deep-learning-based intelligent systems require a large amount of computation; thus, the need is to build a lightweight and efficient system. Second, the computation should be distributed, and the server should perform lightweight tasks to increase efficiency. Considering the above mentioned problems and taking advantage of our lightweight system, we established a distributed facial recognition model based on three parts. 5G cameras capture the video and send it directly to the locally available mobile edge server to undergo required tasks such as face detection, landmark detection, alignment, and required feature extraction. As mentioned in Fig. 10, the database server is just responsible for matching the features with the available database to identify the person. The workload on server and time delay can be reduced by performing partial computation operations on edge devices. The central server can be located anywhere in cyber-space.

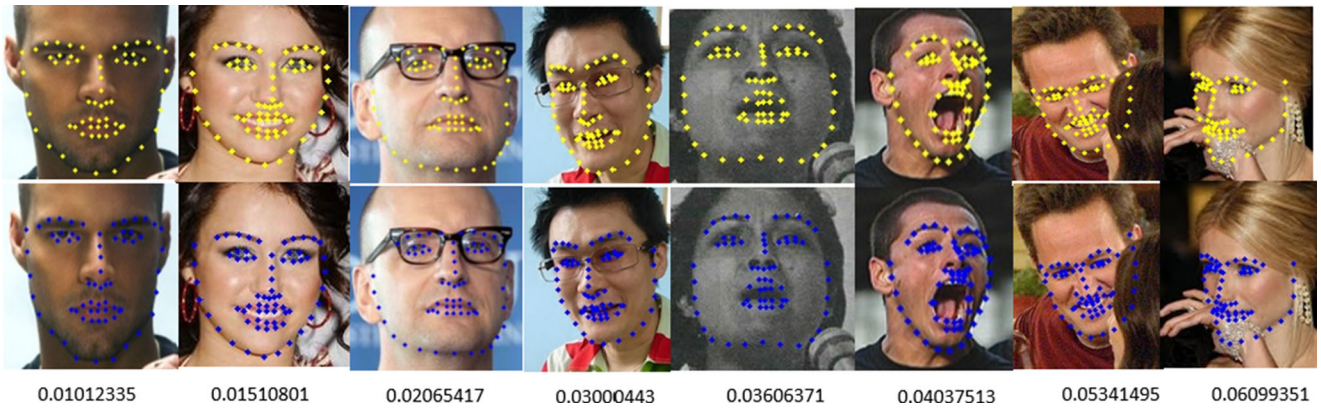



**Fig. 9** Qualitative detection results of the proposed approach for some sample faces from 300W full dataset. The ground-truth landmarks are marked in green color (top row), while our predicted landmarks are in blue color with estimated NRMSE (bottom row)

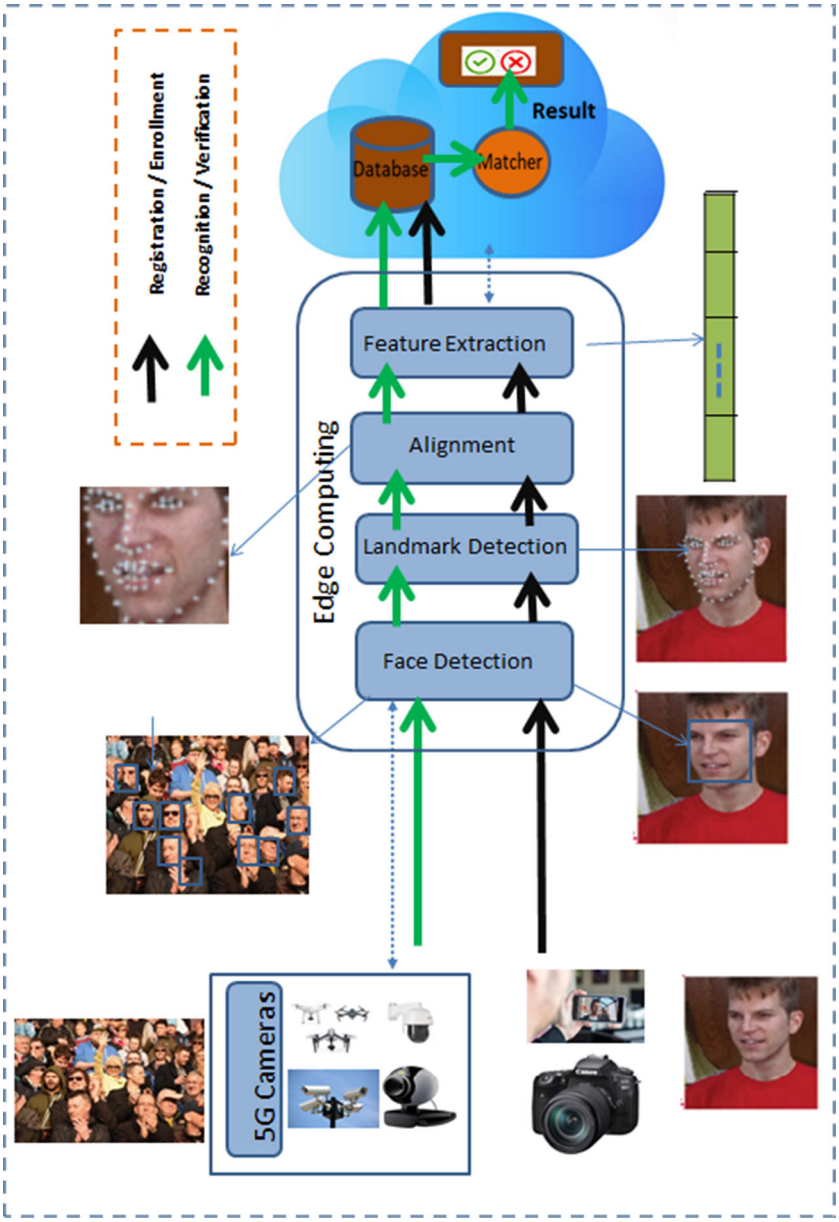



**Fig. 10** A comprehensive facial recognition model for 5G Camera based Cyber-Physical Surveillance Systems

## 6 Conclusion and future work

The innovation of 5G technology and its rapid growth have enabled fast communication between IoT devices and the cyber domain. The surveillance video management system is rapidly expanding its scope and applications. The use of 5G technology in smart cities enables the integration of real-time video observations with access to specific locations. This paper has demonstrated an occlusion-adaptive attentive deep network as another way to address FLD problems. Specifically, we introduced channel-wise attention and spatial attention in an already established attention-adaptive deep network model to improve its performance. The objective is to enhance the representation of interest, model occlusion, capture holistic facial features, and handle spatial distortion. The whole framework was tested on various benchmark datasets with different settings and compared against many state-of-the-art methods. The proposed method reduces the error from 4.17 to 3.82 for the 300W Full-set dataset. After training on Menpo dataset, error reduces up to 3.63, which is 13% decrease in error than ODN. Results proved that our proposed approach detects facial landmarks more accurately than existing methods. Considering the proposed model's property of lightweight parameters, we proposed an efficient distributed facial recognition model. Taking advantage of our model's robustness, the implementation of facial expression recognition is in our next research plan. At last, we also intend to implement a parallel computing version for more efficient and fast processing.

**Acknowledgements** This work is supported by Ministry of Science and Technology China (MOST) Major Program on New Generation of Artificial Intelligence 2030 No. 2018AAA0102200. It is also supported by Natural Science Foundation China (NSFC) Major Project No. 61827814 and Shenzhen Science and Technology Innovation Commission (SZSTI) Project No. JCYJ20190808153619413. The experiments in this work was conducted at the National Engineering Laboratory for Big Data System Computing Technology, China.

## Declarations

**Conflict of Interests** I would like to undertake that: • All authors of this research paper have directly participated in the planning, execution, or analysis of this study; • All authors of this paper have read and approved the final version submitted; • The contents of this manuscript have not been copyrighted or published previously; • Research not involving human participants and/or animals. The research is performed on publicly available data-sets and all data-sets are properly cited. • All funding information is mentioned in acknowledgement section.

## References

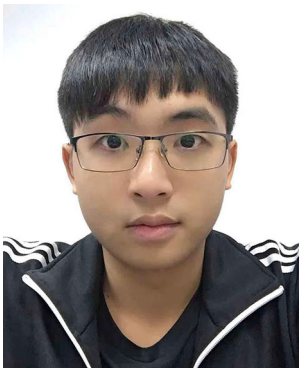
1. Rao SK, Prasad R (2018) Impact of 5g technologies on smart city implementation. Wirel Pers Commun 100(1):161–176
2. Gautam K, Thangavel SK (2019) Video analytics-based intelligent surveillance system for smart buildings. Soft Comput 23(8):2813–2837
3. Liang J, Ma M, Sadiq M, Yeung K-H (2019) A filter model for intrusion detection system in vehicle ad hoc networks: a hidden markov methodology. Knowl-Based Syst 163:611–623
4. Marabissi D, Mucchi L, Fantacci R, Spada MR, Massimiani F, Fratini A, Cau G, Yunpeng J, Fedele L (2019) A real case of implementation of the future 5g city. Fut Internet 11(1):4
5. Kim H, Cha Y, Kim T, Kim P (2020) A study on the security threats and privacy policy of intelligent video surveillance system considering 5g network architecture. In: 2020 International conference on electronics, information, and communication (ICEIC). IEEE, pp 1–4
6. Dsouza J, Elezabeth L, Mishra VP, Jain R (2019) Security in cyber-physical systems. In: 2019 Amity international conference on artificial intelligence (AICAI). IEEE, pp 840–844
7. Pricop E (2019) Biometrics the secret to securing industrial control systems. Biometr Technol Today 2019(4):8–10
8. Obaidat MS, Traore I, Woungang I (2019) Biometric-Based Physical and Cybersecurity Systems, vol 368. Springer
9. Karim ME, Phoha VV (2014) Cyber-physical systems security. In: Applied cyber-physical systems. Springer, pp 75–83
10. Chen S, Shi D, Sadiq M, Cheng X (2020) Image denoising with generative adversarial networks and its application to cell image enhancement. IEEE Access 8:82 819–82 831
11. Ni JJ (2020) Web based security system. uS Patent 10,694,149
12. Ghimire S, Lee B (2020) A data integrity verification method for surveillance video system. Multimed Tools Appl 79(41):30 163–30 185
13. Shan S, Chen X, Gao W (2015) Face misalignment problem
14. Zhu M, Shi D, Zheng M, Sadiq M (2019) Robust facial landmark detection via occlusion-adaptive deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3486–3496
15. Burgos-Artizzu XP, Perona P, Dollár P (2013) Robust face landmark estimation under occlusion. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1513–1520
16. Kemelmacher-Shlizerman I, Basri R (2010) 3D face reconstruction from a single image using a single reference face shape. IEEE Trans Pattern Anal Mach Intell 33(2):394–405
17. Wu Y, Ji Q (2019) Facial landmark detection: a literature survey. Int J Comput Vis 127(2):115–142
18. Cootes TF, Taylor CJ, Cooper DH, Graham J (1995) Active shape models-their training and application. Comput Vis Image Understand 61(1):38–59
19. Zhu X, Ramanan D (2012) Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE Conference on computer vision and pattern recognition. IEEE, pp 2879–2886
20. Tzimiropoulos G, Alabort-i Medina J, Zafeiriou S, Pantic M (2012) Generic active appearance models revisited. In: Asian conference on computer vision. Apringer, pp 650–663
21. Asthana A, Zafeiriou S, Cheng S, Pantic M (2013) Robust discriminative response map fitting with constrained local models. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3444–3451
22. Wu W, Qian C, Yang S, Wang Q, Cai Y, Zhou Q (2018) Look at boundary: a boundary-aware face alignment algorithm. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2129–2138
23. Zou X, Zhong S, Yan L, Zhao X, Zhou J, Wu Y (2019) Learning robust facial landmark detection via hierarchical structured ensemble. In: Proceedings of the IEEE International Conference on Computer Vision, pp 141–150
24. Chen L, Su H, Ji Q (2019) Deep structured prediction for facial landmark detection. In: Advances in neural information processing systems, pp 2447–2457
25. Gao P, Lu K, Xue J, Shao L, Lyu J (2020) A coarse-to-fine facial landmark detection method based on self-attention mechanism. IEEE Transactions on Multimedia
26. Zhang J, Hu H, Feng S (2020) Robust facial landmark detection via heatmap-offset regression. IEEE Trans Image Process 29:5050–5064
27. Dapogny A, Bailly K, Cord M (2019) Decafa: Deep convolutional cascade for face alignment in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, pp 6893–6901
28. Liu Z, Zhu X, Hu G, Guo H, Tang M, Lei Z, Robertson NM, Wang J (2019) Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3467–3476
29. Kumar A, Chellappa R (2018) Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 430–439
30. Zhu M, Shi D, Gao J (2019) Branched convolutional neural networks incorporated with jacobian deep regression for facial landmark detection. Neural Networks
31. Sadiq M, Shi D, Guo M, Cheng X (2019) Facial landmark detection via attention-adaptive deep network. IEEE Access 7:181 041–181 050
32. Wu Y, Ji Q (2015) Robust facial landmark detection under significant head poses and occlusion. In: Proceedings of the IEEE International Conference on Computer Vision, pp 3658–3666
33. Liu Q, Deng J, Yang J, Liu G, Tao D (2016) Adaptive cascade regression model for robust face alignment. IEEE Trans Image Process 26(2):797–807
34. Xing J, Niu Z, Huang J, Hu W, Zhou X, Yan S (2017) Towards robust and accurate multi-view and partially-occluded face alignment. IEEE Trans Pattern Anal Mach Intell 40(4):987–1001
35. Bringmann A, Syrbe S, Görner K, Kacza J, Francke M, Wiedemann P, Reichenbach A (2018) The primate fovea:

structure, function and development. Progress Retinal Eye Res 66:49–84
36. Tschulakow AV, Oltrup T, Bende T, Schmelzle S, Schraermeyer U (2018) The anatomy of the foveola reinvestigated. PeerJ 6:e4482
37. Gao P, Yuan R, Wang F, Xiao L, Fujita H, Zhang Y (2020) Siamese attentional keypoint network for high performance visual tracking. Knowl-Based Syst 193:105448
38. Wu Y, Jiang X, Fang Z, Gao Y, Fujita H (2021) Multi-modal 3d object detection by 2d-guided precision anchor proposal and multi-layer fusion. Appl Soft Comput 108:107405
39. Gao P, Zhang Q, Wang F, Xiao L, Fujita H, Zhang Y (2020) Learning reinforced attentional representation for end-to-end visual tracking. Inf Sci 517:52–67
40. Li H, Li Y, Xing J, Dong H (2019) Spatial alignment network for facial landmark localization. World Wide Web 22(4):1481–1498
41. Yue L, Miao X, Wang P, Zhang B, Zhen X, Cao X (2018) Attentional alignment networks. BMVC 2(6):7
42. Shao Z, Liu Z, Cai J, Ma L (2018) Deep adaptive attention for joint facial action unit detection and face alignment. In: Proceedings of the European Conference on Computer Vision (ECCV), 'pp 705–720
43. Chen L, Zhang H, Xiao J, Nie L, Shao J, Liu W, Chua T-S (2017) Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5659–5667
44. Woo S, Park J, Lee J-Y, So Kweon I (2018) Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 3–19
45. Park J, Woo S, Lee J-Y, Kweon IS (2018) Bam: Bottleneck attention module. arXiv:1807.06514
46. Li H, Li Y, Xing J, Dong H (2018) Spatial alignment network for facial landmark localization. Springer Sci Media
47. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
48. Deng J, Dong W, Socher R, Li L.-J., Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on computer vision and pattern recognition. IEEE, pp 248–255
49. Sagonas C, Tzimiropoulos G, Zafeiriou S, Pantic M (2013) 300 Faces in-the-wild challenge the first facial landmark localization challenge. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp 397–403
50. Koestinger M, Wohlhart P, Roth PM, Bischof H (2011) Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. In: 2011 IEEE International conference on computer vision workshops (ICCV workshops). IEEE, pp 2144–2151
51. Zafeiriou S, Trigeorgis G, Chrysos G, Deng J, Shen J (2017) The menpo facial landmark localisation challenge: a step towards the solution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 170–179
52. Tzimiropoulos G (2015) Project-out cascaded regression with an application to face alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3659–3667
53. Honari S, Molchanov P, Tyree S, Vincent P, Pal C, Kautz J (2018) Improving landmark localization with semi-supervised learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1546–1555
54. Dong X, Yu S-I, Weng X, Wei S-E, Yang Y, Sheikh Y (2018) Supervision-by-registration: an unsupervised approach to improve the precision of facial landmark detectors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 360–368
55. Le V, Brandt J, Lin Z, Bourdev L, Huang TS (2012) Interactive facial feature localization. In: European conference on computer vision. Springer, pp 679–692
56. Belhumeur PN, Jacobs DW, Kriegman DJ, Kumar N (2013) Localizing parts of faces using a consensus of exemplars. IEEE Trans Pattern Anal Mach Intell 35(12):2930–2940
57. Ghiasi G, Fowlkes CC (2014) Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2385–2392
58. Miao X, Zhen X, Liu X, Deng C, Athitsos V, Huang H (2018) Direct shape regression networks for end-to-end face alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5040–5049
59. Dong X, Yan Y, Ouyang W, Yang Y (2018) Style aggregated network for facial landmark detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 379–388
60. Zhen X, Yu M, Xiao Z, Zhang L, Shao L (2020) Heterogenous output regression network for direct face alignment. Pattern Recogn:107311
61. Liu H, Lu J, Feng J, Zhou J (2017) Two-stream transformer networks for video-based face alignment. IEEE Trans Pattern Anal Mach Intell 40(11):2546–2554
62. Tang Z, Peng X, Li K, Metaxas Dn (2019) Towards efficient u-nets: A coupled and quantized approach. IEEE transactions on pattern analysis and machine intelligence
63. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A (2017) Automatic differentiation in pytorch



**Muhammad Sadiq** received his MS(CS) degree from Riphah International University Pakistan in 2015. Mr. Sadiq is currently a Ph.D. Scholar in the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China from 2017. His research interests are Artificial Intelligence, Cloud Computing, Cloud Security, Computer Vision, etc. Mr. Sadiq has several publications in the last few years.

**D. Shi** received the PhD degree in mechanical engineering from Harbin Institute of Technology, China, in 1997, and the PhD degree in computer science from University of Southampton, United Kingdom, in 2002. Before he joined Shenzhen University as a Distinguished Professor in 2016, he had been serving as a Reader / Professor at Middlesex University, UK, since 2010, and Assistant Professor at Nanyang Technological University, Singapore, during 2002-2009. He has also held an appointment of Adjunct Professor at Harbin Institute of Technology. Prof. Shi chaired the technical committee on Intelligent Internet System, IEEE SMC Society from 2005-2010. His current research interests include machine learning, image processing, and computer vision. He has published one book and over 150 academic papers, which appear in reputable journals such as IEEE Transactions on Pattern Analysis and Machine Intelligence and IEEE Transactions on Image Processing, etc. He has completed quite a number of projects funded by national-level research councils, such as Agency of Science and Technology Research (A*STAR) Singapore, Natural Science Foundation Council (NSFC) China, and European Council FP7.



**Junwei Liang** received the B.Sc. degree from the Guangdong University of Petrochemical Technology in 2014, and the M.Sc. degree from Shenzhen University in 2017. He is currently pursuing the Ph.D. degree with the School of Electronic and Electrical Engineering, Nanyang Technological University, Singapore. His current research interests include intrusion detection systems, evolutionary computation, vehicle ad hoc network, and artificial intelligence.

## Affiliations

**Muhammad Sadiq[1] · D. Shi[1] · Junwei Liang[2]**

Muhammad Sadiq
muhammad.sadiq@szu.edu.cn

Junwei Liang
junwei001@e.ntu.edu.sg

1 College of Computer Science and Software Engineering, Shenzhen University, Guangdong Province, China

2 Nanyang Technological University, Singapore, Singapore