Attention-Guided Coarse-to-Fine Network for 2D Face Alignment in the Wild

XIN LIU,1 HUABIN WANG,1 JIAN ZHOU,1 AND LIANG TAO1

1Key laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University, Hefei 230031, China
Corresponding authors: Huabin Wang (wanghuabin@ahu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61372137 and in part by the Natural Science Foundation of Anhui Province, China, under Grant 1908085MF209, 1708085MF151.

ABSTRACT Methods based on stacked hourglass networks (SHNs) have achieved great progress in face alignment tasks. However, most of these algorithms have met with limited success in modelling correlations among features. Visual attention mechanisms have shown promise in terms of effectively understanding scenes in various computer vision tasks. In this paper, an attention-guided coarse-to-fine network (AGCFN) based on an attention mechanism is proposed for robust face alignment. Thus, the network is guided to emphasize key information while suppressing less important information. Meanwhile, the fusion of features from different levels is adopted to improve the information flow through the proposed network. Additionally, conditional random fields (CRFs) are introduced to model the spatial interactions between landmarks in the prediction maps. Experimental results obtained on the 300-W dataset, the 300-W private test set, and the WFLW dataset demonstrate the superiority of the proposed method in terms of accuracy and robustness.

INDEX TERMS Face alignment, Attention mechanisms, Corse-to-fine, CNNs, CRFs

I. INTRODUCTION

The task of face alignment or facial landmark detection [1], [2], defined as the problem of the localization of facial landmarks, has attracted substantial attention in the computer vision community. Understanding a person’s facial landmarks (e.g., the corners of the eyes, eyebrows, and corners of the mouth) is useful for higher-level tasks such as emotion recognition [3] and expression analysis [4]. Face alignment serves as a fundamental tool in many face applications, and has presented researchers with a variety of challenges over the years. In particular, achieving accurate facial landmark localization is challenging due to large variations in viewpoint, the variety of expressions that can appear on different faces, and partial occlusion.

Many algorithms have been proposed to alleviate the challenges mentioned above. Recently, convolutional neural networks (CNNs) [1], [2], [5] have achieved significantly enhanced performance in face alignment based on their strong ability to approximate complex and non-linear functions for mapping points from arbitrary face images to the facial landmark locations, even under the conditions of unconstrained facial appearances and background clutter. Diverse network architectures have been extensively studied, including the development of deep neural networks.

Stacked hourglass networks (SHNs) [6] have proven to be very effective for tasks involving human pose estimation. The architecture of the hourglass modules enables such a network to capture features from different scales as well as contextual information. The primary purpose of an hourglass module is to locate the precise pixel locations of key points on a human body. The goal of face alignment is to detect the locations of landmarks on a human face. Therefore, hourglass modules have been applied to face alignment tasks and have inspired many research works [1], [2], [7], [8]. For example, an SHN [7] has been proposed as the basis of a robust face alignment method by combining supervised face transformation with the SHN. In the work of [8], an SHN with a novel residual block was designed to cope with the problem of 2D and 3D facial landmark detection. Although great progress has been achieved with SHNs and their variants, most SHN-based methods have met with limited success in modelling spatial correlations among human facial landmarks. Because the outputs of the hourglass modules are independent heatmaps and the highest response value
in each heatmap corresponds to the location of the corresponding landmark, the relationships between the heatmaps are not considered. However, for the task of facial landmark detection, global contextual information and the relationships between facial components are indispensable for individual landmark prediction, providing crucial clues for addressing the challenges of facial landmark detection in real-world circumstances.

One crucial aspect of deep learning methods that allows them to alleviate the above problems is their ability to leverage visual attention mechanisms, which are usually neglected by researchers in face alignment tasks. Attention is the behavioural and cognitive process of selectively concentrating on a discrete aspect of information [9]. Visual attention in particular is an essential mechanism exploited by the human brain for effectively understanding scenes. Tsotsos et al. [10] reported that there are at least three main fundamental components of a visual attention mechanism. He et al. [11] summarized these components as follows: (1) the selection of regions of interest, (2) the selection of feature channels and values of interest and (3) the control of the information flow through the network. On this basis, an attention-guided coarse-to-fine network (AGCFN) which combines an attention mechanism with an SHN is proposed.

Specifically, the hourglass modules capture features from different resolutions, which provide sufficient scale clues for the human face, in a stage-by-stage manner. Feature fusion among levels is adopted to facilitate the flow of information and the passing of messages through the network. In addition to successive top-down and bottom-up modules, a channel attention module is designed for learning the relationships among the feature channels. Additionally, conditional random field (CRF) modelling is incorporated into the CNN for learning the spatial relationships among facial landmarks. Message passing among different feature maps provides valuable cues for inferring the locations of hidden parts from visible ones. In contrast, cluttered backgrounds and complex activities can often mislead a CNN in regard to feature representation learning without guiding principles between landmarks. Consequently, the performance of landmark localization can be severely reduced. Hence, both spatial attention and channel attention modules are incorporated into the model with the aim of improving the ability of the SHN to model the structural relationships between proximate predictions.

The two complementary strategies (the attention mechanism and the CRF modelling approach) are simple yet powerful. We show the effectiveness of the proposed framework on three widely used datasets: the 300-W dataset [12], the 300-W private test set [13] and the Wider Facial Landmarks in the Wild (WFLW) dataset [2].

In summary, our contributions are as follows:

1) An AGCFN is proposed, which is extremely powerful in extracting both local evidence and global context features to guide the model to focus on key regions of the human face.

2) A simple and yet computationally efficient channel attention module is proposed to improve the ability of our model to select and maintain the key channels among all feature channels.

3) CRFs are introduced to model the spatial structural relationships among proximate facial landmarks.

4) Our algorithm achieves state-of-the-art results in terms of the normalized mean error (NME), failure rate (FR), and area under the curve (AUC) on three popular challenging datasets: 300-W, the 300-W private test set, and WFLW.

The remainder of the paper is organized in the following manner. Section II presents a brief review of related works on face alignment, attention mechanisms and structure learning. The proposed framework is presented in Section III. The experimental details are the subject of Section IV; the advocated approach is experimentally validated in this section. Finally, the paper is concluded in Section V.

II. RELATED WORK

A. FACE ALIGNMENT

In the following, we review developments in the 2D face alignment task, which has been an active research field for decades. At present, the available methods fall into two main types: classical approaches and CNN-based methods.

Classical approaches. Early classical methods such as active shape models (ASMs) [14], active appearance models (AAMs) [15]–[17], constrained local models (CLMs) [18], [19] and cascaded regression models [5], [20]–[22] were widely used prior to the advent of neural networks. Classical cascaded regression approaches [4], [23]–[27] for face alignment are popular among researchers. Xiong et al. [27] proposed a method named the Supervised Descent Method (SDM) for learning cascade regressors based on strong hand-crafted features such as SIFT features [28]. Ren et al. [26] utilized a random forest model to learn discriminative binary features based on the “locality” principle, achieving a speed of 3000 fps. Zhu et al. [29] proposed a coarse-to-fine search method to weaken the influence of inaccurate shape initializations. Using a fuzzy membership weighting strategy, Feng et al. [30] trained a multi-view cascaded regression model to improve the fault tolerance of cascaded regression. Although great improvements have been achieved, face alignment is still a challenging task with these methods because of the limited representation capabilities of hand-crafted features.

CNN-based methods. In recent years, works based on CNNs have revolutionized landmark localization, yielding face alignment results of remarkable accuracy even on the most challenging datasets. These methods can be divided into two groups: coordinate regression and heatmap regression.

Coordinate regression. Conventional cascaded regression methods have been greatly changed by CNNs such as [31]. Zhou et al. [24] proposed a multi-stage deep network for detecting facial landmarks in a coarse-to-fine manner. Zhang et al. [32] modelled facial landmark localization and attribute classification as a multi-task problem. Zhang et al. [32] em-
employed coarse-to-fine auto-encoder networks (CFANs) that combined several stacked auto-encoder networks in a cascaded manner. Trigeorgis et al. [5] applied a recurrent neural network for face alignment by replacing the feature extraction stage with a convolutional network and the fitting stage with the recurrent neural network. Xiao et al. [22] employed an attention long short-term memory (A-LSTM) network and a refinement long short-term memory (R-LSTM) network; the A-LSTM sequentially selected the attention centre, and the R-LSTM refined the landmarks around the attention centre. Lv et al. [33] presented a deep regression architecture with two-stage reinitialization to explicitly address the initialization problem for face detection. Zhen et al. [1] presented a new loss function and achieved superior performance. More recently, Wayne et al. [2] proposed an algorithm in which boundary information is introduced to assist in key point regression. Roberto et al. [34] presented a real-time facial landmark regression method based on a coarse-to-fine ensemble of regression trees (ERT).

Heatmap regression. Newell et al. [6]. first proposed a U-shaped SHN by stacking several hourglass modules to generate predictions for human pose estimation. An SHN [7] can serve as the basis for a robust face alignment method that combines supervised face transformation with the SHN. A deep alignment network [35] consists of multiple stages, each of which refines the locations of the facial landmarks based on the landmark heatmaps and features from the previous stage. The face alignment network (FAN) proposed in [8] also employs an SHN with a novel residual block to solve the 2D and 3D face alignment problem. The authors of another previous work [36] formulated a novel multi-view hourglass model that attempts to jointly estimate both semi-frontal and profile facial landmarks. The score-guided FAN proposed by Xiang et al. [37] is an extension of an SHN that has proven efficient in cases of occlusion.

B. VISUAL ATTENTION MECHANISMS

Attention is the behavioural and cognitive process of selectively concentrating on a discrete aspect of information [9]. However, common algorithms usually ignore the influence of attention mechanisms; consequently, key information is not highlighted, and uncorrelated information is not suppressed [11]. Different features carry information with different impacts on prediction; thus, they should be treated appropriately. Attention mechanisms help to improve the ability of a network to represent different degrees of importance. Very recently, since visual attention models are computationally efficient and effective for understanding images, they have achieved great success in various computer vision tasks. A lightweight convolutional block attention module (CBAM) [38] was introduced by Sanghyun et al. for object recognition. An object-part attention model (OPAM) [39] has been proposed for fine-grained image classification and achieves superior performance compared with several state-of-the-art methods. A residual attention network [40], a deep neural network using an attention mechanism, is composed of multiple stacked attention modules for image classification. Two types of attention modules for semantic segmentation have been proposed in [41]. A multi-context attention mechanism has been incorporated into an SHN for human pose estimation [42].

Inspired by these works, we investigate the relationship between channels by using a visual attention mechanism, with the goal of explicitly modelling the interdependencies between the convolutional features in different channels. To this end, we introduce a novel channel attention module, which allows the network to perform feature recalibration to selectively emphasize informative features while suppressing less useful ones. Specifically, we wish to guide the network to focus on key features while suppressing unnecessary ones.

C. STRUCTURE MODELLING

Structure modelling has shown great power in many tasks. To summarize, in recent studies, multi-scale processes have been incorporated into deep neural network modules to help capture spatial correlations between features, as in the Inception family of networks [43]–[46]. A structured feature learning method was proposed by Chu et al. [47] for inferring the relationships among human joints at the feature level for human pose estimation. Prior domain knowledge is directly incorporated into the network proposed by Yang et al. [48], thus greatly accelerating the learning process. A facial boundary heatmap was first used by Wu et al. [2] to help regress landmarks for facial landmark detection. A novel attention model was introduced by Chu et al. [42] for spatial correlation modelling rather than adopting con-
inputs to the

where the
g

bottom-up process,
respectively, of the
x
(ReLU), and pooling.
f

where
n
module. The mathematical processes can be formulated as

same resolution in this hourglass module have shortcut con-

is referred to as an hourglass module [6]. Features with the

an illustration of a top-down and bottom-up module, which

Figure 2 shows the architectures of the naive hourglass mod-

to refine the results.

how the attention modules work together within the network

modules: a basic U-shaped module, a channel attention mod-

III. THE PROPOSED METHOD

In this section, the attention-guided coarse-to-fine network
(AGCFN) is proposed. Figure 1 shows an overview of the

proposed architecture. The AGCFN mainly consists of three
modules: a basic U-shaped module, a channel attention mod-
ule, and a spatial attention module. First, we discuss the

stratey of the methods above, CRFs are incorporated into

convolutional CRFs (ConvCRFs) [49]. Following the general

stragegy, in a cascaded manner to refine the predictions

semantic meanings are also shared among the U-Nets. More

but in the stacked U-Nets structure, features with the same

among the hourglasses would improve the reuse of features

later staging encoding the global context, skip connectivity

glasses focus on representing different aspects of an image,

together to refine the detected results in a coarse-to-fine

face alignment. Several hourglasses could simply be stacked

learning the complex relationships among the landmarks for

In a single hourglass, the connections and encoder-decoder
architecture improve the ability to represent features of dif-

cient attention schemes such as global softmax. The

strong and valid assumption of conditional independence
has been incorporated into an existing framework, namely,

convolutional CRFs (ConvCRFs) [49]. Following the general

strategy of the methods above, CRFs are incorporated into

our network as geometric constraints to achieve better facial

landmark detection.

A. BASIC U-SHAPE MODULE

Figure 2 shows the architectures of the naive hourglass mod-
ule and the coupled hourglass module. Figure 2 (a) shows
an illustration of a top-down and bottom-up module, which
is referred to as an hourglass module [6]. Features with the
same resolution in this hourglass module have shortcut con-

nections. These skip connections help the information to flow

efficiently across different levels within a single hourglass
module. The mathematical processes can be formulated as
shown below. For the top-down process,

\[ x_l^n = f_l^n(x_{l-1}^n), \]

where \( n \) is the stage of the hourglass, \( l \) denotes the level
of the top-down and bottom-up blocks in one hourglass, and
\( f_l^n(\cdot) \) denotes a combination of operations consisting con-

volution (Conv), batch normalization (BN), rectified linear units
(Relu), and pooling. \( x_{l-1}^n \) and \( x_l^n \) are the input and output,
respectively, of the \( l^{th} \) block of the top-down process. For
the bottom-up process,

\[ y_l^n = g_l^n(y_{l-1}^n + x_l^n), \]

where \( g_l^n(\cdot) \) is the same as the \( f_l^n(\cdot) \). \( y_{l-1}^n \) and \( x_l^n \) are
inputs to the \( l^{th} \) bottom-up block, and \( y_l^n \) denotes the output.

B. CHANNEL ATTENTION MODULE

We first briefly introduce conventional soft attention mech-

anisms, which are commonly used in computer vision, and
then present the proposed channel attention module.

Conventional attention mechanisms. A soft attention
mechanism, in a nutshell, can be subdivided into three steps:
squeezing, excitation, and scaling. The computation process
is as follows: (1) The spatial dimensions of the input feature
maps are squeezed by means of pooling. The squeezing op-

eration produces a channel descriptor by aggregating feature
maps across their spatial dimensions \((H \times W)\). (2) Fully
connected (FC) layers are usually employed for the excitation

\[ x_l^n = f_l^n(C(x_{l-1}^n, X)), \]

where \( X \) denotes the outputs (e.g., \( x_l^{n-1} \)) of top-down and
bottom-up blocks which share the same resolution with \( x_{l-1}^n \)
in the preceding blocks, \( C \) denotes the concatenation oper-

ation. For bottom-up process of the second stage,

\[ y_l^n = g_l^n(C(y_{l-1}^n, Y)), \]

where similarly, \( Y \) represents the connections (e.g., \( y_{l-1}^{n-1} \))
from previous blocks which share the same resolution with
\( y_{l-1}^n \).
operation, which takes the form of a simple self-gating mechanism and produces a collection of per-channel modulation weights. (3) The generated attention maps are normalized by using the sigmoid function, and the weights are then applied to \( f_0 \) (the raw features) to obtain \( F \) (the refined features). The mathematical formulation is summarized as follows:

\[
F = \sigma(f(Pooling(f_0))) \odot f_0,
\]

where \( \sigma(\cdot) \) denotes the sigmoid function, \( f(\cdot) \) denotes the operations of BN, ReLU, and Conv, \( Pooling(\cdot) \) represents the pooling operation, and \( \odot \) denotes element-wise multiplication.

### Channel attention module

The relationships between channels are essential for facial landmark detection. Therefore, a channel attention module is designed to explicitly model the interdependencies between channels. Specifically, let \( F \) be the input feature maps for the channel attention module, where \( F \in \mathbb{R}^{C \times H \times W} \) denotes the features extracted by previous residual blocks, with \( H, W, \) and \( C \) denoting the height, width, and number of channels, respectively.

A global pooling or encoding layer is usually adopted for the squeezing process, as described in previous work [51]. However, we directly calculate the channel attention map \( A \in \mathbb{R}^{C \times C} \) from the original features \( F \in \mathbb{R}^{C \times H \times W} \) without using any convolution layers. Following the strategy in [41], \( F \in \mathbb{R}^{C \times H \times W} \) is reshaped to \( F \in \mathbb{R}^{C \times N} \), and then, a matrix multiplication is performed between and the transpose of \( F \). Here, \( N \) denotes \( H \times W \). Finally, a softmax layer is applied to obtain the channel attention map \( A \in \mathbb{R}^{C \times C} \):

\[
a_{ij} = \frac{e^{F_i \cdot F_j}}{\sum_{i=1}^C e^{F_i \cdot F_j}},
\]

where \( a_{ij} \) is designed to measure the impact of the \( i \)th channel on the \( j \)th channel. In addition, we perform a matrix multiplication between the transpose of \( A \) and \( F \), and then reshape the result to \( \mathbb{R}^{C \times H \times W} \). An element-wise sum operation is then conducted with \( A \) to obtain the refined feature set \( R \in \mathbb{R}^{C \times H \times W} \):

\[
R_j = \alpha \sum_{i=1}^C (a_{ij}F_i) + F_i,
\]

where \( \alpha \) is a weight, initially \( \alpha = 0 \) and gradually learns a weight from 0. The output feature from each channel is a weighted sum of the features from all channels, thus modelling the dependencies between channels. For the task of facial landmark detection, the relationships between facial components are indispensable for individual landmark prediction, which provides crucial clues for addressing the challenges of facial landmark detection. The proposed channel-wise attention module is introduced to capture the channel dependencies between any two channel maps, and update each channel map with a weighted sum of all channel maps.
In this way, the network is guided to learn the relationships among the channels. The output of the channel attention module can be fed directly into subsequent layers of the network.

C. SPATIAL ATTENTION MODULE

Deep neural networks have a limited ability to model accurate spatial relationships of the human face. Moreover, the output heatmaps produced by the hourglass modules are independent of each other during the inference process, and the spatial relationships between the heatmaps are ignored. Specifically, \( k \) is the total number of facial landmarks, and the location of a predicted landmark \( L_k \) is decoded from the predicted heatmap \( H_k(\cdot) \) by using the argmax function to obtain the location with the maximum value as follows:

\[
L_k = \text{argmax}(H_k(\cdot))
\] (8)

To overcome the issue that the hourglass modules lack the ability to model interactions among the predicted facial landmarks, the structure modelling capabilities of CRFs are incorporated into our proposed network. Spatial information from face images is beneficial for improving face alignment performance because the locations of most landmarks are closely related to those of other landmarks nearby. Concretely, we choose the mean-field approximation method to model these correlations in a recursive manner. In the CRF module, the energy function of a label assignment is as follows:

\[
E(l) = \sum \psi_u(l_i) + \sum \psi_p(l_i, l_j),
\] (9)

\[
\psi_p(l_i, l_j) = \sum_{i,j} l_i W_{i,j} l_j,
\] (10)

where \( \psi_u(l_i) \) is a unary term that measures the likelihood that the label of position \( l_i \) is correct; \( \psi_p(l_i, l_j) \) is the second paired potential term of the energy function, which aims to model the correlations between predicted landmarks; and \( W_{i,j} \) is a weight representing the compatibility between \( l_i \) and \( l_j \).

Given an image \( I \), the probability of the label assignment for position \( l_i \) is \( P(l_i) = \frac{1}{Z} \exp(-E(l_i)) \), where \( P(l_i) \) denotes the probability map, which corresponds to the probability that each pixel is given the correct label, and \( Z \) is the partition function, which can be considered equivalent to a softmax layer. The probability for position \( l_i \) is iteratively obtained using the mean-field approximation as follows:

\[
\phi(l_i)_t = \sigma(\psi_u(l_i) + \sum_{j=1}^C W_{i,j} \phi(l_i)_{t-1}),
\] (11)

where \( \sigma(a) = \frac{1}{1 + \exp(-a)} \) is the sigmoid function, \( C \) is the number of channels, \( \psi_u(l_i) \) is obtained from the previous features through convolution, \( \sum_{j} W_{i,j} \phi(l_i) \) is implemented by convolving the estimated feature map \( \phi_{t-1} \) from \( t-1 \) stage with the filters, and \( W_{i,j} \) is shared across different stages, which learns a weight between \( l_i \) and \( l_j \). In our network, we use three stages of recursive convolution. Initially, \( \phi(l_i)_1 = \sigma(\psi_u(l_i)) \). The whole process of the proposed AGCFN is summarized in Algorithm 1.

Algorithm 1 The training process of AGCFN

Require: Face image \( I \), face rectangle \( R \), landmark label \( L \).

1: while epoch <= n do
2: Obtain partial facial image \( I_R \) in accordance with face rectangle \( R \) of face image \( I \)
3: Feed \( I_R \) into the AGCFN and obtain the preliminary facial landmark detection result \( L^* \)
4: Feed to preliminary result \( L^* \) into the channel attention module to calculate weight \( W \), using Equation 6
5: Feed the refined channel to the spatial attention module to model the spatial relationships among the predicted landmarks
end
6: Obtain the final shape \( P \).
7: Perform an affine to original resolution by Procrustes analysis.

IV. EXPERIMENT RESULTS

In this section, the proposed method is evaluated on the 300-W dataset [12], the 300-W private test set [13] and the Wider Facial Landmarks in the Wild (WFLW) dataset [2]. We first introduce our implementation details and experimental settings. Then, the proposed algorithm is compared with other state-of-the-art approaches on 300-W, the 300-W private test set and WFLW. Finally, extensive ablation studies are conducted to show the effectiveness of the proposed approach.

A. DATASETS

1) 300-W

300-W [12] is the most widely used in-the-wild dataset for 2D face alignment. It covers a large range of variations of identity, expression, illumination conditions, and occlusion conditions. The training set consists of 3148 face images in total. We present results obtained on both the common subset and the challenging subset. All face images in 300-W are labelled with 68 landmarks.

2) 300-W private

The 300-W private dataset [13] is a test set that is usually regarded as the challenging benchmark of 300-W. It consists of both indoor images and outdoor images, and all images in 300-W private are annotated with 68 landmarks.

3) WFLW

WFLW was introduced in [2] to facilitate future research on face alignment. WFLW contains 10000 faces (7500 for training and 2500 for testing) with 98 fully manually annotated
FIGURE 4: Results on both helen and ibug subset of the 300-W test set.

TABLE 1: Evaluation of AGCFN on test set and 6 typical subsets of WFLW (98 landmarks) in terms of NME (%), FR (%), and AUC. Note that error is normalised by outer eye corner distance.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Method</th>
<th>Fullset</th>
<th>Pose</th>
<th>Expression</th>
<th>Illumination</th>
<th>Make-up</th>
<th>Occlusion</th>
<th>Blur</th>
</tr>
</thead>
<tbody>
<tr>
<td>NME (%)</td>
<td>ESR [52]</td>
<td>11.13</td>
<td>25.88</td>
<td>11.47</td>
<td>10.49</td>
<td>11.05</td>
<td>13.75</td>
<td>12.70</td>
</tr>
<tr>
<td></td>
<td>SDM [27]</td>
<td>10.29</td>
<td>24.10</td>
<td>11.45</td>
<td>9.32</td>
<td>9.38</td>
<td>13.03</td>
<td>11.28</td>
</tr>
<tr>
<td></td>
<td>DVLN [53]</td>
<td>6.08</td>
<td>11.54</td>
<td>6.78</td>
<td>5.73</td>
<td>5.98</td>
<td>7.33</td>
<td>6.88</td>
</tr>
<tr>
<td></td>
<td>LAB [2]</td>
<td>5.27</td>
<td>10.24</td>
<td>5.51</td>
<td>5.23</td>
<td>5.15</td>
<td>6.79</td>
<td>6.32</td>
</tr>
<tr>
<td></td>
<td>AGCFN</td>
<td>4.90</td>
<td>8.78</td>
<td>5.00</td>
<td>4.93</td>
<td>4.85</td>
<td>6.26</td>
<td>5.73</td>
</tr>
<tr>
<td>Failure (%)</td>
<td>ESR [52]</td>
<td>35.24</td>
<td>90.18</td>
<td>42.04</td>
<td>30.80</td>
<td>38.84</td>
<td>47.28</td>
<td>41.40</td>
</tr>
<tr>
<td></td>
<td>SDM [27]</td>
<td>29.40</td>
<td>84.36</td>
<td>33.44</td>
<td>26.22</td>
<td>27.67</td>
<td>41.85</td>
<td>35.32</td>
</tr>
<tr>
<td></td>
<td>CFSS [29]</td>
<td>20.56</td>
<td>66.26</td>
<td>23.25</td>
<td>17.34</td>
<td>21.84</td>
<td>32.88</td>
<td>23.67</td>
</tr>
<tr>
<td></td>
<td>DVLN [53]</td>
<td>10.84</td>
<td>46.93</td>
<td>11.15</td>
<td>7.31</td>
<td>11.65</td>
<td>16.30</td>
<td>13.71</td>
</tr>
<tr>
<td></td>
<td>AGCFN</td>
<td>5.92</td>
<td>24.23</td>
<td>5.41</td>
<td>4.72</td>
<td>5.82</td>
<td>11.00</td>
<td>8.79</td>
</tr>
<tr>
<td>AUC</td>
<td>ESR [52]</td>
<td>0.2774</td>
<td>0.0177</td>
<td>0.1981</td>
<td>0.2953</td>
<td>0.2485</td>
<td>0.1946</td>
<td>0.2204</td>
</tr>
<tr>
<td></td>
<td>SDM [27]</td>
<td>0.3002</td>
<td>0.0226</td>
<td>0.2293</td>
<td>0.3237</td>
<td>0.3125</td>
<td>0.2060</td>
<td>0.2398</td>
</tr>
<tr>
<td></td>
<td>CFSS [29]</td>
<td>0.3659</td>
<td>0.0632</td>
<td>0.3157</td>
<td>0.3854</td>
<td>0.3691</td>
<td>0.2688</td>
<td>0.3037</td>
</tr>
<tr>
<td></td>
<td>DVLN [53]</td>
<td>0.4551</td>
<td>0.1474</td>
<td>0.3899</td>
<td>0.4743</td>
<td>0.4494</td>
<td>0.3794</td>
<td>0.3973</td>
</tr>
<tr>
<td></td>
<td>LAB [2]</td>
<td>0.5323</td>
<td>0.2345</td>
<td>0.4951</td>
<td>0.5433</td>
<td>0.5394</td>
<td>0.4490</td>
<td>0.4630</td>
</tr>
<tr>
<td></td>
<td>AGCFN</td>
<td>0.5452</td>
<td>0.2860</td>
<td>0.5267</td>
<td>0.5511</td>
<td>0.5547</td>
<td>0.4621</td>
<td>0.4823</td>
</tr>
</tbody>
</table>

landmarks. Compared with previous datasets, the faces in this dataset exhibit larger variations in occlusion, pose, makeup, illumination, blur and expression.

B. PERFORMANCE EVALUATION METRICS

We use the normalized mean error (NME), the cumulative error distribution (CED) curve, the area under the curve (AUC), and the failure rate (FR) to measure the landmark location error. Concretely, different evaluation metrics are calculated on different datasets for comparison with existing popular methods. For the 300-W dataset, we utilize both the inter-pupil distance (the distance between the centres of the eyes) and the inter-ocular distance (the distance between the outer corners of the eyes), and when the NME of an image is greater than 0.08, this case is considered a failure. For the 300-W private test set, we employ the NME normalized with respect to the inter-pupil distance, and when the NME of an image is greater than 0.08, this case is considered a failure. For the WFLW dataset, we use the inter-ocular distance, and when the NME of an image is greater than 0.1, this case is deemed a failure. The NME for the inter-pupil (or inter-ocular) distance is computed as follows:

\[ NME_i = \frac{1}{N} \sum_{j=1}^{N} \frac{|P_{ij} - G_{ij}|}{|L - R|_2} \]  

where \( N \) is the total number of landmarks; \( i \) and \( j \) are the indices representing different face images and landmarks, respectively; \( P_{ij} \) and \( G_{ij} \) are the predicted and ground-truth landmarks.
locations, respectively, of landmark j in face image i; L and R are the locations of the centres (outer corners) of the left and right eyes, respectively, in the $i^{th}$ face image.

TABLE 2: Results on 300-W dataset in terms of the NME (%) of the proposed AGCFN. Note that error is normalised by the inter-pupil and inter-ocular distance respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>Year</th>
<th>Common Subset</th>
<th>Challenge Subset</th>
<th>Fullset</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESR [52]</td>
<td>2012</td>
<td>5.28</td>
<td>17.00</td>
<td>7.58</td>
</tr>
<tr>
<td>SDM [27]</td>
<td>2013</td>
<td>5.60</td>
<td>15.40</td>
<td>7.52</td>
</tr>
<tr>
<td>RCPR [54]</td>
<td>2013</td>
<td>6.18</td>
<td>17.26</td>
<td>8.35</td>
</tr>
<tr>
<td>CFAN [55]</td>
<td>2014</td>
<td>5.50</td>
<td>16.78</td>
<td>7.69</td>
</tr>
<tr>
<td>LBF [26]</td>
<td>2014</td>
<td>4.95</td>
<td>11.98</td>
<td>6.32</td>
</tr>
<tr>
<td>TCDCN [32]</td>
<td>2014</td>
<td>4.80</td>
<td>8.60</td>
<td>5.54</td>
</tr>
<tr>
<td>CFSS [29]</td>
<td>2015</td>
<td>4.73</td>
<td>9.98</td>
<td>5.76</td>
</tr>
<tr>
<td>DAR [22]</td>
<td>2016</td>
<td>4.12</td>
<td>8.35</td>
<td>4.94</td>
</tr>
<tr>
<td>DAN [35]</td>
<td>2017</td>
<td>4.42</td>
<td>7.57</td>
<td>5.03</td>
</tr>
<tr>
<td>TSR [33]</td>
<td>2017</td>
<td>4.36</td>
<td>7.56</td>
<td>4.99</td>
</tr>
<tr>
<td>SHN [7]</td>
<td>2017</td>
<td>4.12</td>
<td>7.00</td>
<td>4.68</td>
</tr>
<tr>
<td>DCFE [34]</td>
<td>2018</td>
<td>3.83</td>
<td>7.54</td>
<td>4.55</td>
</tr>
<tr>
<td>3DDE [56]</td>
<td>2019</td>
<td>3.73</td>
<td>7.10</td>
<td>4.39</td>
</tr>
<tr>
<td>AGCFN</td>
<td></td>
<td>3.73</td>
<td>7.24</td>
<td>4.42</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Year</th>
<th>Inter-ocular (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAN [35]</td>
<td>2017</td>
<td>3.19</td>
</tr>
<tr>
<td>PCD-CNN [57]</td>
<td>2018</td>
<td>3.67</td>
</tr>
<tr>
<td>SAN [58]</td>
<td>2018</td>
<td>3.34</td>
</tr>
<tr>
<td>DCFE [34]</td>
<td>2018</td>
<td>2.76</td>
</tr>
<tr>
<td>3DDE [56]</td>
<td>2019</td>
<td>2.69</td>
</tr>
<tr>
<td>AGCFN</td>
<td>2019</td>
<td>2.69</td>
</tr>
</tbody>
</table>

TABLE 3: Inter-ocular NME (%), FR (%) and AUC values of face alignment results obtained on the 300-W private test set

<table>
<thead>
<tr>
<th>Method</th>
<th>NME (%)</th>
<th>FR (%)</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESR [52]</td>
<td>-</td>
<td>17.00</td>
<td>0.3225</td>
</tr>
<tr>
<td>CFSS [29]</td>
<td>-</td>
<td>12.30</td>
<td>0.4132</td>
</tr>
<tr>
<td>MDM [5]</td>
<td>5.05</td>
<td>6.80</td>
<td>0.4532</td>
</tr>
<tr>
<td>DAN [35]</td>
<td>4.30</td>
<td>2.67</td>
<td>0.4700</td>
</tr>
<tr>
<td>SHN [7]</td>
<td>4.05</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DCFE [34]</td>
<td>3.88</td>
<td>1.83</td>
<td>0.5242</td>
</tr>
<tr>
<td>AGCFN</td>
<td>3.82</td>
<td>1.60</td>
<td>0.5252</td>
</tr>
</tbody>
</table>

C. TRAINING DETAILS

For training, we performed image flipping, scaling (by factors between 0.7 and 1.3), rotation (through angles between -30 and 30 degrees) and the addition of colour jitter to augment the data. The inputs were RGB images with a resolution of 128×128. The network was optimized using the Adam optimizer [59] with an initial learning rate of 2.5e-4, which decayed by a factor of two after every 40000 iterations. The experiments were conducted on a computer running Ubuntu 16.04 with an Nvidia 1080 GPU with a mini-batch size of 10 for 100000 iterations using PyTorch [60].

For model training, the mean-squared error (MSE) loss function is widely used for face alignment. To represent the ground-truth landmark labels, we generated a confidence map $m_k$ for each single landmark $k$ ($k \in \{1, ..., K\}$) by centring a Gaussian kernel on the labelled position $z_k = (x_k, y_k)$. More specifically, the Gaussian confidence map $m_k$ for the $k^{th}$ landmark label can be expressed as follows:

$$m_k(x, y) = \frac{1}{2\pi\sigma} \exp\left(-\frac{[(x-x_k)^2+(y-y_k)^2]}{2\sigma^2}\right),$$  \hspace{1cm} (13)

where $(x, y)$ specifies a pixel location and the hyperparameter $\sigma$ denotes a pre-fixed spatial variance. The MSE loss function is then obtained as follows:

$$Loss = \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{K} \|H_p - H_G\|^2,$$  \hspace{1cm} (14)

where $N$ is the number of training samples, $K$ is the number of landmarks, and $H_p$ and $H_G$ are the predicted and ground-truth confidence maps, respectively, corresponding to each hourglass. Intermediate supervision is introduced for better information flow through the network.

D. COMPARATIVE EXPERIMENTS

We compare the performance of our model with the performance of other state-of-the-art networks for 2D face alignment on 300-W, the 300-W private test set and WFLW.

WFLW. The WFLW dataset contains various kinds of challenges, such as exaggerated expressions, occlusion, makeup, illumination, and blur. Since WFLW is a newly released dataset, the proposed method is compared with fewer but competitive methods, including ESR [52], SDM [27], CFSS [29], DVLN [53], and LAB [2]. Instead of using bounding boxes cropped from the ground-truth facial landmarks, we employ the facial bounding boxes provided by the WFLW dataset. We compare our approach against the other...
state-of-the-art methods on WFLW in Table 1. Our method achieves an NME of 4.90% on the test set, which consists of 6 typical subsets. The performance of our method on the WFLW dataset is obviously better than that of LAB. Notably, our method employs only a coupled hourglass module as its basic structure, whereas the LAB method employs 8 stacked hourglasses as its backbone. Consequently, achieving performance comparable to that of our method with the LAB method would be inefficient in both time and cost. By contrast, our proposed model is simple yet effective. In addition, we can see from the second and third rows of the table that our method greatly outperforms the LAB method in terms of both the FR and the AUC. Examples of qualitative results obtained on the WFLW dataset can be seen in Figure 3.

300-W. The 300-W dataset is commonly used for face alignment tasks. A series of competitive results obtained on this dataset using many methods have been reported. Extensive comparative evaluations are carried out with most of these methods, and the corresponding experimental results obtained on 300-W are shown in Table 2. Our method achieves comparable performance relative to the state-of-the-art results. Note that our method achieves an NME of 7.24% on the challenging (IBUG) subset, which reflects its effectiveness in handling large view variations and exaggerated expressions. In addition to these results on the challenging set of 300-W, our method achieves an NME of 3.73% on the common test set. This performance indicates that our proposed AGCFN is robust against extreme variations in head pose and exaggerated expressions. Examples of qualitative outputs obtained on the 300-W dataset are shown in Figure 4.

300-W private. The 300-W private test set is also referred to as the challenging benchmark of 300-W. We evaluated our method on the 300-W private test set to test the generalization ability of our proposed method on different datasets. The 300-W private test set is designed for the purpose of evaluating robustness against exaggerated facial expressions, which are very challenging for current face alignment methods to address. As illustrated by the CED curves in Figure 5, our method outperforms the previous methods by a large margin on the 300-W private test set. Note that our method achieves an NME of 3.82% and that both the FR and AUC results on...
TABLE 4: Inter-ocular NME (%), FR (%) and AUC results for different combinations on 300-W

<table>
<thead>
<tr>
<th>Method</th>
<th>NME (%)</th>
<th>FR (%)</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>HG</td>
<td>4.45</td>
<td>2.56</td>
<td>0.4692</td>
</tr>
<tr>
<td>HG+CAM</td>
<td>3.92</td>
<td>2.34</td>
<td>0.5140</td>
</tr>
<tr>
<td>HG+CAM+CRF</td>
<td>3.83</td>
<td>2.15</td>
<td>0.5213</td>
</tr>
</tbody>
</table>

Note that the coupled hourglass module is denoted by CU-HG and that the structure consisting of 2 stacked hourglasses is denoted by 2-HG.

V. CONCLUSION

In this paper, an attention-guided coarse-to-fine network (AGCFN) for 2D face alignment is proposed, which is endowed with an enhanced ability to model feature correlations through the introduction of an attention mechanism. Specifically, a channel attention module is designed to learn the relationships among the predicted channels. Furthermore, conditional random fields (CRFs) are employed to model spatial correlations. The experimental results show that the proposed channel attention and CRF modules can guide the network to capture structural information and generate more robust results. Our proposed network achieves outstanding performance on the 300-W private test set, the 300-W private test set, and the WFLW dataset.

REFERENCES


XIN LIU received the B.S. degree in Information and Computing Science from Hubei University of arts and science, China, in 2017, and he is currently pursuing the M.S. degree in computer science and technology from Anhui University. His research interests include human pose estimation and face alignment.

WANG HUABIN received the B.S. degree in computer science and technology from Anhui University of Finance and Economics, China, in 2005, and the M.S. degree in signal and information processing and the Ph.D. degree in computer application technology from Anhui University, China, in 2008 and 2011, respectively. He is currently the deputy director of computer science and technology, Anhui University, Hefei, China. His research interests include face recognition, virtual reality and signal processing.

JIAN ZHOU received the B.S. degree in computer science and technology and the M.S. degree in computer application technology from Southwest Jiaotong University, Chengdu, China, in 2004 and 2007, respectively, and the Ph.D. degree in information and communication engineering from Southeast University, Nanjing, China, in 2013. He is currently an Associate Professor with the College of Computer Science and Technology, Hefei, China, where he is also the Director of the Computer Education Research Association of Anhui higher education institutions. His current research interests focus on speech intelligibility enhancement, speech conversion, and object tracking.

TAO LIANG received the B.S. degree in radio technology in 1985 and circuit and system in 1988 from Anhui University, Hefei, China, respectively, and the Ph.D. degree in information and communication engineering in 2013 from the University of Science and Technology of China (USTC), Hefei, China. He is currently the academic and technical leader of Anhui Province and a professor of Anhui University, Hefei, China. He has chaired a number of National Natural Science Foundation projects, the Anhui Provincial Natural Science Foundation Project and the key research projects of Natural Science in Anhui Provincial Department of education. His research interests include pattern recognition, intelligent information processing and multimedia signal processing.