Stacked Multi-Target Network for Robust Facial Landmark Localisation

Yun Yang, Bing Yu, Xiaodong Li, and Bailan Feng
Huawei Technologies Co., Ltd.
Beijing, China
{yangyun18,yubing5,lixiaodong33,fengbailan}@huawei.com

Abstract

We thoroughly analyse regression-based face alignment methods and introduce a novel stacked multi-target network for robust facial landmark localisation. The primary heatmap regression-based network concentrates on locating the coarse position of pre-defined landmarks while the secondary coordinate regression-based network is responsible for modelling fine sub-pixel features. Specifically, we elaborate the differences among widely-used Cross Entropy related loss functions and propose a new Bilateral Inhibition Cross Entropy loss function, which enlarges the margin between elements in the output heatmaps.

Besides, in order to deal with the discrepancy between optimization and evaluation, we propose to dynamically adjust the radius of kernel function during the training process. We demonstrate that training with decreasing radius in temporal order performs much better than assigning it spatially, i.e. decreasing radius along the stages of stacked hourglass networks. Finally, we innovatively limit the output of the secondary coordinate regression network to a reasonable range by importing the hinge loss to refine the coarse coordinate locations for sub-pixel accuracy. Extensive experiments on public datasets such as 300-W, COFW, and AFLW demonstrate that our proposed method performs superiorly to the state-of-the-art approaches.

1. Introduction

Facial landmark localisation, also known as face alignment, refers to locating pre-defined facial landmarks, such as mouth corners and face contours on an input face image. It is the foundation of many interesting algorithms in human face related visual tasks, e.g., face recognition [27, 32, 42, 13], 3D face reconstruction [16, 38], face beautification [25], and emotion estimation [41, 26]. In face recognition algorithm, researchers usually utilize facial landmark localisation method to align input human face to an approximate canonical shape and reduce the image-embedding mapping complexity [37].

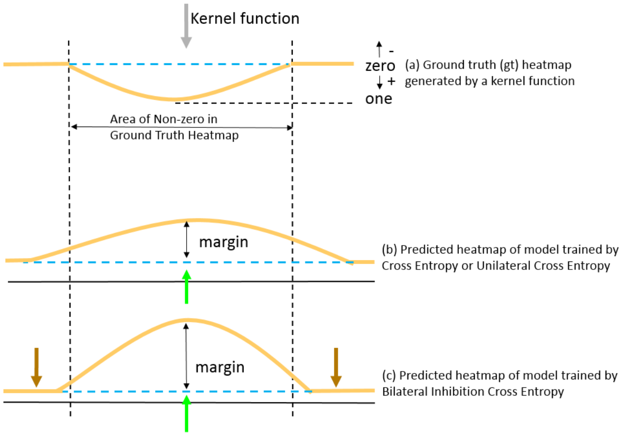
Kernel function
zero
one
-
+
(a) Ground truth (gt) heatmap generated by a kernel function
Area of Non-zero in Ground Truth Heatmap
margin
(b) Predicted heatmap of model trained by Cross Entropy or Unilateral Cross Entropy
margin
(c) Predicted heatmap of model trained by Bilateral Inhibition Cross Entropy

Figure 1. The margin of cross entropy and proposed Bilateral Inhibition Cross Entropy loss function. Figure (a) represents a ground truth heatmap generated by an arbitrary kernel function. Figure (b) shows the margin of predicted heatmap instructed only by the Non-zero area of ground truth heatmaps. Figure (c) depicts the margin of our proposed Bilateral Inhibition Cross Entropy loss, which adds extra penalty to predicted heatmap pixels whose corresponding ground truth probability equals zero. The green arrow indicates the impact from non-zero area in gt heatmap to fit predicted heatmap for ground truth, while the brown arrow shows the impact from zero area in gt heatmap to inhibit the probability error between corresponding pixels. (Best view in colour.)

In the past decades, researchers have devoted a lot of efforts to developing an accurate localisation algorithm for face images captured under large poses and illumination variations. Among the evolution of omnigenous methods, Active Shape Model (ASM) [11], Active Appearance Model (AAM) [10], and Constrained Local Model (CLM) [12] stand out for their simplification and effectiveness. It achieves satisfactory results for face landmark localisation under constrained condition. However, human face images in real-world scenarios appear with large poses, expressions, and illumination variations. To deal with these problems, cascaded-regression based approaches have been proposed to achieve refined localisation accuracy. Kazemi *et al*. [21] introduces an ensemble of essential regression trees and then incorporates them in a streamlined formulation into a cascade framework. Wu *et al*. [44] proposes a constrained joint cascade regression framework which updates landmark locations and the activation probabilities of action unit iteratively.

Recently, deep neural networks have achieved great progress in a variety of computer vision tasks, including pose estimation [30, 31], face recognition [27, 47], and image segmentation [35, 8]. As with face alignment task, different types of networks [5, 17, 14] have been put forward to achieve high performance in unconstrained scenarios. According to the format of network output, these alignment algorithms can be divided into two mainstream categories, coordinate regression and heatmap regression. In the case of coordinate regression-based facial landmark localisation methods [17], the network outputs a vector of $2L$ real numbers for the 2D coordinates of L landmarks. In practice, cascading strategy is usually integrated to ensemble multiple coordinate regression networks for improved performance in unconstrained scenarios. In heatmap regression-based facial landmark localisation approaches [29, 48], a network outputs a heatmap with the same size of input image for each target landmark. Generally, softmax manipulation is implemented on output heatmaps and the intensity of a pixel in the heatmap indicates the probability that it belongs to a target landmark [19, 34].

However, there are several crucial issues which has not been considered and carefully studied in heatmap regression-based algorithms. Firstly, appropriate kernel function parameters need to be manually designed to generate proper ground truth heatmap labels. A careless choice of the parameters would result in a serious deteriorated performance. Secondly, we argue that more attention need to be paid to the metric discrepancy between training (usually Cross Entropy loss) and evaluation process (Euclidean distance) in heatmap regression methods. A wrongly targeted objective in the optimization (training) process would guide the network to a sub-optimal result. Besides, the commonly used Cross Entropy loss only brings penalty to the non-zero elements in the ground truth heatmaps and has no impact on the zero elements. This slows down the convergence of training procedure, and makes it difficult to find the optimal result. Last but not least, we have to redeem the localisation accuracy attenuation caused by the conversion from target sub-pixel floating-point label to integer coordinate outputs generated by argmax manipulation.

To address these issues, in this paper, we propose a new stacked multi-target network (SMT-Net) for robust facial landmark localisation and introduce solutions to the above three questions. The main contributions of our work are summarized as follows:

- We explore a multi-target network combined with heatmap and coordinate regression to learn the primary heatmap regression error and secondary subpixel residuals jointly.
- We propose two methods of dynamically adjusting a kernel radius to shrinking the discrepancy between probability heatmap similarity metric in training process and coordinate euclidean distance metric in evaluation process.
- We detailedly analyse the widely-used Cross Entropy (CE) related loss functions and introduce a novel Bilateral Inhibition Cross Entropy (BICE) loss (Fig.1) to penalize misaligned landmarks whose corresponding ground truth probability equals zero. We theoretically prove that BICE loss enlarges the margin of elements in network output probability heatmaps and matches subsequent argmax manipulation.

2. Related Works

Coordinate Regression model inputs a detected human face image and outputs the 2D coordinate vector consisting of $2L$ elements. It usually utilizes a classical Convolutional Neural Networks (CNN) model to extract hierarchical feature and mapping it to target label [17, 28]. Researchers usually combine cascade regression and neural networks to achieve accurate facial landmark localisation. In detail, a face image is first input to a stage-one network and output the rough initial facial landmarks, followed by an affine transformation operation, it is aligned to a canonical shape. Then we refine the landmark coordinate by feeding aligned image to a stage-two network [39, 28]. In [17], authors discuss different loss functions which are critical in the training process of a regression network. Feng *et al*. presents a systematic analysis of different loss functions (L1, L2, and smooth L1 loss functions) and proposes new wing loss function to improve the network training capability for small range errors. Besides, researchers usually utilize multi-task framework to jointly improve performance on all the associated computer vision tasks [51].

Heatmap Regression based methods have achieved promising results recently and showed the effectiveness for facial landmark detection in the wild. Rather than directly estimating coordinate vectors, heatmap regression methods output probability heatmaps for each pre-defined landmark. [48] proposes a Convolutional Experts Constrained Local Model (CE-CLM) algorithm that uses convolutional experts network to model complex landmark appearance that is affected by expression, makeup, and accessories. [29] points out that fully-convolutional neural networks are not able to aggregate global context due to their constrained receptive field. They introduce a subnet to bring in extra global-context information and achieve good performance. Although these methods have achieved promising results on

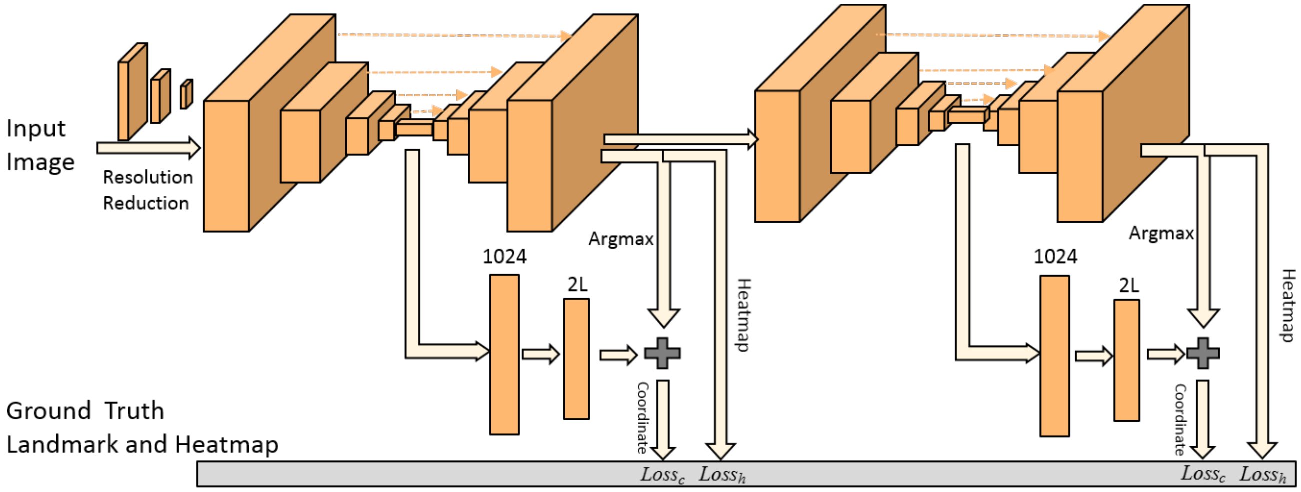

Figure 2. The architecture of our proposed SMT-Net. Two hourglass networks are stacked and supervised by ground truth heatmaps individually. For each hourglass network, sub-pixel coordinates regression is performed (length equals $2L$, L denotes number of landmarks) with the global context feature obtained from centre of hourglass network. The whole network is optimized with two loss functions. Specifically, $Loss_c$ (L2 loss + hinge loss) is used for Coordinate network and $Loss_h$ (BICE loss) is used for Heatmap network. Dotted lines represent the skip connections between feature maps of the same resolution.

public datasets, the calculation of two separated networks is usually computationally expensive.

As mentioned above, coordinate regression and heatmap regression approaches exhibit different advantages and own different areas of expertise. Intuitively, we combine both approaches, heatmap network for coarse target and coordinate network for refined target, into a stacked multi-target framework. With the help of alignment residuals trained by euclidean distances, chen *et al*. [9] conclude that the discrepancy between training and testing is relieved. In [9], chen *et al*. also construct a small sub-network to produce residual features from the middle layer of the network and estimate the alignment errors of the current output heatmaps. However, in this case, the whole network is sometimes dominated by sub-network which weakens the contribution of powerful hourglass network. Instead, in this paper, we draw a hinge loss into the L2 loss function to limit the coordinate network output to a relative reasonable range. In addition, we also analyse the difference of cross entropy related loss functions and introduce a more effective BICE loss. Detailed discussion of primary heatmap network and secondary coordinate network will be illustrated in section 3.3 and 4.

How to choose a proper kernel function radius is also a problem that is worthy people's attention. In previous heatmap regression-based algorithms, authors [23] design a chessboard distance with power function following two rules: the probability of pixel located at grounding truth coordinate is the largest and pixel value becomes smaller if it is away from the grounding truth coordinate. Supposing the base of power function equals 0.5, the probability of pixel approximately equals zero (more accurate is $1e^{-3}$) if it is away from more than 10 pixels. The non-zero area that is useful for model optimization only occupies a little proportion. Considering ideal situation where the probability heatmap is a impulse response only activated at ground truth location, the model optimization becomes difficult it discards lots of available global structure information. In this paper, we suggest to dynamically adjust the kernel radius in training process, which reduces the Region-of-Interest gradually to balance training and testing.

In this paper, we creatively propose an Stacked Multi-Target Network (Fig.2) to jointly utilize heatmap-based multi-scale local feature and coordinate-based global context information. In addition, two loss functions are carefully designed for primary heatmap regression target and secondary coordinate regression target specifically. To our best knowledge, the proposed approach is the first work to study the margins of output heatmaps with different loss functions. This study brings in a novel BICE loss function for heatmap regression-based facial landmark localisation.

3. Stacked Multi-Target Network

3.1. Framework

As mentioned above, the target of heatmap regression-based facial landmark localisation is to find a nonlinear mapping that outputs probability heatmaps, followed by argmax manipulation to obtain the coordinates of predefined landmarks:

$$\Phi_h : I \rightarrow P \rightarrow \mathbf{t} \tag{1}$$

where $I \in \mathbb{R}^{H \times W \times 3}$ represents the input colour image cropped by a face detector, $P \in \mathbb{R}^{Hs \times Ws \times L}$ denotes the output heatmaps and $\mathbf{t} \in \mathbb{R}^{2L}$ denotes corresponding facial landmarks vector $[x_1, y_1, x_2, y_2, ..., x_L, y_L]^T$. L represents the number of landmarks and (x_l, y_l) denotes the coordinates of the lth landmark.

We usually utilize a CNN model with designed loss function $Loss_h()$ to learn the mapping Φ_h. The target of training process is to find a Φ which minimizes:

$$\sum_{i=1}^{N} Loss_h(\Phi(I_i), Q_i) \tag{2}$$

where $Q_i \in \mathbb{R}^{H_s \times W_s \times L}$ is the target heatmap generated by a kernel function $Q_i = K(t_i)$ of the ith sample.

We empirically utilize a stacked hourglass network which is able to extract high-level and pixel-wise features at different scales. The input is a $256 \times 256 \times 3$ colour image and the output P is a tensor with shape $64 \times 64 \times L$, where L=68 for 300-W and COFW dataset and L=19 for AFLW dataset. Since the shape of hourglass network output is the same as the input, a pixel located at (j, k) in the predicted heatmap can be mapped to the same position of corresponding input image. We adhere following principles to obtain the coordinates: the l-th channel of CNN output feature maps predicts the location of l-th facial key point and the coordinate of the largest value in the l-th feature map represents the l-th facial key point (x_l, y_l).

The ultimate goal of face alignment is to identify the coordinates of pre-defined landmarks. Nevertheless, both heatmap and coordinate regression network have drawbacks. Heatmap regression network outputs a heatmap evaluating each pixel's probability belonging to a pre-defined landmark and thus requires argmax manipulation which suffers precision drop from floating-point to integer coordinates. Coordinate regression based network is not able to capture pixel-wise information that is crucial for facial landmark detection. We intuitively incorporate two networks to SMT-Net and make improvements on individual network.

As shown in Fig. 2, our SMT-Net starts with a 7×7 convolution with stride $= 2$, pad $= 3$, a residual layer, a pooling layer, and three consecutive residual blocks to bring the resolution down from 256 to 64. All residual blocks are replaced with the parallel and multi-scale block [4]. Two hourglass network [30] is stacked to capture hierarchical representation from an input $64 \times 64 \times 256$ feature map. For each hourglass network, we design a regression branch educed from the centre-layer feature maps of hourglass network followed by two fully connected layers to regress coordinates gradually. The ultimate loss function is formulated as follows:

Method ($\sigma = 4$)	Normalized Mean Error (%)
Baseline network	6.23
Spatial, $\delta = 0.5$	6.15
Spatial, $\delta = 1$	6.31
Temporal, $\delta = 0.5$	6.17
Temporal, $\delta = 1$	6.12

Table 1. A comparison of spatial and temporal category between different δ values on 300-W challenge dataset. To simplification the experiments, a two stages stacked hourglass network and Adam optimizer with weight decay $= 5 \times 10^{-4}$ is applied for training a baseline network.

$$\sum_{i=1}^{N} (Loss_h(\Phi_h(I_i), Q_i) + Loss_c(\Phi_c(I_i), \mathbf{t}_{i,c})) \\ \mathbf{t}_{i,c} = \mathbf{t}_i - \mathbf{t}_{i,h} \tag{3}$$

Respectively, $Loss_h$ and $Loss_c$ represent the loss function for heatmap and coordinate regression-based networks. Φ_h and Φ_c denote the image-to-target mapping function parametrized by consecutive dense CNN layers. $\mathbf{t}_i$ and $\mathbf{t}_{i,h}$ denote the ground truth landmark and current heatmap network output coordinates generated by argmax operation. $\mathbf{t}_{i,c}$ denotes non-negligible target residuals for the coordinate regression network.

3.2. Dynamically adjusting kernel radius

In **Heatmap Regression** based networks, there is a crucial issue to quantify the similarity between the predicted probability heatmap P and the ground truth probability heatmap Q. In [23], Lai *et al.* suggests that a good ground truth possibility heatmap should meet the following requirements: (1) the probability for the index (x_l, y_l) is the largest one in a heatmap. (2) the probability should be smaller if the pixel is farther away from (x_l, y_l). Therefore, researchers usually utilize gaussian or power kernel function to generate a ground truth probability heatmap. Noted that, in previous methods, kernel function radius is usually fixed in training process. Researchers empirically choose possible optimal value of kernel radius σ to fit for different input image size. Considering the following two boundary conditions: 1) σ is large enough which makes all elements of target heatmap probability heatmap equal one. In this case, target label is not able to instruct the model updating and the optimization process collapses. 2) σ is small enough which makes all elements of target heatmap probability heatmap equal zero except the pixel located at (x_l, y_l). It degrades to one-hot label and discards structure information around ground truth landmarks which is effective to guide the optimization of model parameters.

To deal with these problems, we propose to adjust kernel

Method	NME (%)
only heatmap network	6.22
heatmap+coordinate, $\theta = 5$	6.33
heatmap+coordinate, $\theta = 0.5$	6.16

Table 2. A comparison of different θ values in Loss_c on 300-W challenge dataset. Results show that performance promotes if we restrict coordinate network output to a reasonable range.

radius dynamically in training process spatially or temporally. **Spatial category**: We assign decreasing radius σ with step size δ to N-stacked hourglass network from stack-one to stack-N. In the first few stacked networks, they output the coarse landmark locations of input images. The latter stacked networks take advantage of current location information and output fine-grained face landmark localisation results. **Temporal category**: We first utilize an initialized kernel radius σ, in this paper $\sigma = 4$, to train a stage-one network. Subsequently, we use a decreased $\sigma' = \sigma - \delta$ (δ denotes a step value) to continue training from a convergent epoch of last step. Finally, we repeatedly implement the continue training manipulation until the result of face alignment not improves. In our experiment, the mean error is no longer significantly decreasing after three iterations. Table 1 detailedly describes the results of different parameters used in our experiments on adjusting radius dynamically.

Experiments results demonstrate that either spatial category or temporal category is effective in training process. Noted that the Normalized Mean Error (NME) errors become worse if we change σ severely in spatial category, which indicates different stages of stacked hourglass network is closely related. A relatively large δ makes the latter hourglass network more difficult to utilize the information generated by previous networks.

3.3. Constrain on coordinate network output

In **Coordinate Regression** based methods, the output of CNN network is a real number that is able to measure the errors of sub-pixels. However, in heatmap regression based algorithms, the output is an integer coordinate that can not obtain the sub-pixel location precision. In [9], Chen *et al.* proposes to regress the sub-pixel residual from the centre output of a hourglass network. Though it is able to obtain the real localisation value consisting of integer heatmap output and floating-point regression layer output, it injures the backbone hourglass network because it does not draw restriction on the output of regression layers. We reimplement the experiments in [9] and find that the output of regression layer sometimes dominates the locations of coordinates, which means that the secondary sub-network squeezes the information for heatmap network optimization.

In this paper, we creatively propose to draw a penalty to coordinate network output with large values. An intuitive idea is to utilizing hinge loss as follows:

$$Loss_c = L_2(t_{i,c} - p_{i,c}) + Loss_{hg} \tag{4}$$

where

$$Loss_{hg} = \sum_{l=1}^{L} max(0, |p_{l,cx}| - \theta) + max(0, |p_{l,cy}| - \theta) \tag{5}$$

$[p_{1,cx}, p_{1,cy}, p_{2,cx}, p_{2,cy}, ..., p_{L,cx}, p_{L,cy}]^T$ is the output of coordinate network and θ is designed to restrain the range of coordinate network output between $-\theta$ and θ. The experimental results shown in table 2 demonstrate our method outperforms the others.

4. Bilateral Inhibition Cross Entropy Loss

Researchers utilize hourglass network or Fully Convolutional Networks (FCN) to capture feature of different scales and achieve promising results on different public available datasets. In the case of training heatmap regression-based model with cross entropy loss, we pursuit to draw CNN output heatmap and target probability heatmap closer. In [23, 4], authors propose to utilize softmax loss and sigmoid cross-entropy pixel-wise loss to measure the probability-distance between the l-th feature map and the l-th ground truth coordinate. Noted that in [4], Bulat *et al.* claims that the use of sigmoid cross-entropy pixel-wise loss increases the gradients by 10-15x compared to the L2 loss and offers a noticeable improvement in localisation precision. The objective function of cross entropy loss is formulated as

$$CE(Q_i, P_i) = -\sum_{l} Q_{i,l} log(P_{i,l}) \tag{6}$$

where $P_{i,l}$ and $Q_{i,l}$ are the predicted and ground truth heatmap of lth landmark for ith face image. While cross entropy loss is widely used in existing heatmap regression-based facial landmark detection systems, it only obtains loss value when the element $Q_{i,l}^{j,k}$ of target heatmap is not equal to zero. (j, k) represents a pixel in the heatmap that is located at jth column and kth row. In the following content of this paper, we simply utilize $Q_l \in \mathbb{R}^{H_s \times W_s}$ denotes a CNN output heatmap.

In previous papers, the Gaussian kernel is widely used to obtain the ground truth heatmap $Q_l^{j,k} = N((j,k); (x_l, y_l), \sigma^2 I)$, where (x_l, y_l) is the ground truth location of the lth landmark, and σ is designed to control the variance of the response. It can be formulated as:

$$Q_l^{j,k} = \exp^{-\frac{d_{jk}}{2\sigma^2}}, d_{jk} = \sqrt{(j - x_l)^2 + (k - y_l)^2} \tag{7}$$

According to the three-sigma rule, the value is nearly equal to zero when $d_{jk} > 3\sigma$. In the case of a target heatmap with size of 64*64 and $\sigma = 4$ (shown in Fig.3),

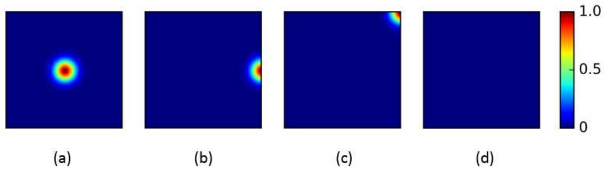

Figure 3. Example of a ground truth heatmap generated by gaussian kernel fucntion. Gaussian kernel radius is set to 4 and heatmap size equals 64×64. Different cases are shown in the figure. (a) All non-zero area of gaussian kernel is inside image. (b) Landmark locates on the boundary of the image, thus only half of non-zero area is inside. (c) Landmark locates on the corner with only quarter of non-zero area inside. (d) Landmark locates outside the image thus all heatmap pixels equal zero. (Best view in colour.)

neglecting landmarks which exceed face bounding box, the maximal proportion of non-zeros is $\frac{\pi(3*4)^2}{64*64} \approx 11.04\%$ (Figure 3(a)) and the minimal proportion of non-zeros is $\frac{\frac{1}{4}\pi(3*4)^2}{64*64} \approx 2.76\%$ (Figure 3(c)). We conclude that at least 90% elements in a ground truth heatmap is approximately equal to zero. However, CE loss only brings unilateral inhibition (from non-zero side) to samples and is not able to make an impact on pixels when $Q_l^{j,k} = 0$.

Inspired by lateral inhibition in neurobiology which depicts the capacity of an excited neuron to reduce the activity of its neighbours [43], we introduce a bilateral inhibition cross entropy loss function which brings penalty to both zero and non-zero pixels and further improves the accuracy of facial landmark detection systems. It can be written as the following equation:

$$\begin{aligned} Loss_h =& BICE(Q_i, P_i) \\ =& -\sum_l (Q_{i,l} log(P_{i,l}) + (1 - Q_{i,l}) log(1 - P_{i,l})) \end{aligned} \tag{8}$$

To illustrate the effectiveness of the proposed BICE loss, a theoretical analysis is conducted to study and compare the margins of both CE and BICE functions. Once we obtain the ground truth heatmaps, we utilize cross entropy loss to judge whether the hourglass net output is accurate. We usually implement softmax manipulation on CNN output heatmaps Q_l, $QS_l = softmax(Q_l)$, to ensure that the summation of a output heatmap is equal to one. However, another ground truth heatmap which is generated by a kernel function does not always satisfy this constraint. In this case, similarity evaluation between network outputs and ground truth heatmaps is not strictly a cross entropy loss. We utilize unilateral cross entropy (UCE) to denominate this loss function where only one side (i.e. the CNN output heatmaps) satisfies the definition of probability distribution and the other side (i.e. the ground truth heatmaps) is not satisfied.

As mentioned in [43], lateral inhibition increases the contrast and sharpness in visual response. Similarly, our proposed BICE loss also enlarges the margin between elements in predicted heatmaps. Let's start with a binary classification case as a toy example to analyse the margin of unilateral cross entropy loss firstly.

Considering UCE loss, the optimization objective function can be written as:

$$\arg\min_{p_1, p_2} -(q_1 \log p_1 + q_2 \log p_2),\ p_1 + p_2 = 1 \tag{9}$$

where q_1, q_2 denote two items of the ground truth heatmaps, p_1, p_2 denote two elements of the network softmax layer output probability which satisfies $p_1 + p_2 = 1$.

We can obtain:

$$\begin{aligned} p_1 &= \frac{q_1}{q_1 + q_2} \\ p_2 &= \frac{q_2}{q_1 + q_2} \end{aligned} \tag{10}$$

Obviously, the margin between p_1 and p_2 is:

$$M_{CE} = \frac{|q_1 - q_2|}{q_1 + q_2} \tag{11}$$

In the case of $q_1 + q_2 = 1$, which means q satisfies the definition of probability distribution, UCE loss degrades to widely-used CE loss.

As for the proposed BICE loss, the optimization objective is:

$$\begin{aligned} \arg\min_{p_1, p_2} -(&q_1 \log p_1 + (1 - q_1) \log(1 - p_1) \\ &+ q_2 \log p_2 + (1 - q_2) \log(1 - p_2)) \end{aligned} \tag{12}$$

We can obtain:

$$\begin{aligned} p_1 &= \frac{1 + (q_1 - q_2)}{2} \\ p_2 &= \frac{1 + (q_2 - q_1)}{2} \end{aligned} \tag{13}$$

Thus the margin between p_1 and p_2 is:

$$M_{BICE} = |q_1 - q_2| \tag{14}$$

In the case of $q_1 + q_2 >= 1$, where the summation of target heatmap elements is greater than or equal to one, the margin of BICE loss is obviously larger than or equal to conventional UCE loss. This implies that the proposed BICE loss would perform better in most cases where $q_1 + q_2 > 1$.

5. Experimental Results

5.1. Datasets

The 300 Faces in the Wild (**300-W**) dataset [36] is a widely used 68-point benchmark dataset, which consists of 3,148 train images and 689 test images. We utilize all the 3,148 samples for training and perform testing on *(i)* Common set, *(ii)* Challenge set, and *(iii)* Full set. *(i)* Common set includes 554 images come from subsets of LFPW [3] and HELEN [24]. *(ii)* Challenge set consists of 135 samples from IBUG. (iii) Full set is a mixture of the above two datasets that includes 689 images.

The Caltech Occluded Faces in the Wild (**COFW**) dataset [6] consists of 1,345 images which are annotated with 29 facial landmarks. To evaluate our algorithm precisely, we use the re-annotated test set with 68 facial landmarks for comparison [18]. Specifically, we conduct experiments on 507 images in the test subset which are occluded to different degrees.

The Annotated Facial Landmarks in the Wild (**AFLW**) dataset [22] contains 24,386 in-the-wild faces with large head pose which is a challenging dataset for facial landmark localisation. Each image is annotated with 19 landmarks (without two ears). We follow AFLW-Full protocol [53], and use 4,386 images for a cross-dataset test to evaluate our proposed method.

5.2. Evaluation metrics

We adopt a widely used normalized error metric to evaluate our proposed method and compare with state-of-the-art methods. It measures normalized distance errors between model-generated and ground truth locations. The calculation of the normalized error E_i for the ith given sample is formulated as:

$$E_i = \frac{\frac{1}{L}\sum_{l=1}^{L} |t_{i,l} - p_{i,l}|_2}{d_i} \tag{15}$$

where $p_{i,l}$ and $t_{i,l}$ represent the predicted and target coordinates respectively. l represents the lth landmark of total L landmarks and d_i denotes the normalization term that ensures the NME scores across faces of different size are fairly weighted. We utilize the inter-ocular distance (or outer eye corner distance) suggested in [28] as the normalization term of 300-W and COFW datasets. For the AFLW dataset, we follow [28] to use face size as the normalization term.

For Normalized Mean Error, it is calculated as follows:

$$\text{NME}_i = \frac{1}{N}\sum_{i=1}^{N} E_i \tag{16}$$

where N is the total numbers of all test samples. For Normalized Median Error (NMDE), it is calculated as follows:

$$\text{NMDE}_i = \text{median}(E_i) \tag{17}$$

Method	Challenge	Common
CLNF[2]	6.37 / 4.93 / 1.44	3.47 / 2.51 / 0.96
SDM[46]	$-/10.73/-$	$-/3.31/-$
CFAN[49]	8.38 / 6.99 / 1.39	$-/-/-$
DRMF[1]	10.36 / 8.64 / 1.72	4.97 / 4.22 / 0.75
CFSS[52]	5.97 / 4.49 / 1.48	3.20 / 2.46 / 0.74
TCDCN[50]	6.87 / 5.56 / 1.31	4.11 / 3.32 / 0.79
3DDFA[54]	12.31 / 8.34 / 3.97	7.27 / 5.17 / 2.1
PO-CR[40]	$-/$**3.33**$/-$	$-/2.67/-$
CE-CLM[48]	5.62 / 4.05 / 1.57	3.13 / **2.23** / 0.9
FC-LGCN[29]	5.55 / 4.36 / 1.19	**3.04** / 2.34 / 0.7
Ours	**5.41** / 4.34 / **1.07**	3.35 / 2.87 / **0.48**

Table 3. Normalized median mean error with (68 points)/without (49 points) face outline for 300-W Challenge and Common datasets. The difference of errors between 68 and 49 points is also evaluated to demonstrate the robustness of methods.

5.3. Implementation details

In our experiments, two stacked hourglass network [30] is applied as the backbone network. To enhance the ability of representing hierarchical features, we enlarge the channel number from 256 to 512 in down-sampling process and back to 256 in up-sampling process. The experiments are conducted with Pytorch framework on a server equipped with 8 × Nvidia Tesla V100 GPU cards. Adam solver with weight decay 5×10^{-4} is applied for network training. Initial learning rate is set to 0.01 and later drops to 1×10^{-4} at epoch 120 and epoch 200.

In data augmentation, we crop human faces with default tight bounding box and randomly rotate each training image between $[-30°, 30°]$. Besides, with the probability of 50% , we randomly flip each training image, translate the bounding box between [-5%, 5%] of the box size, scaling with proportion between [-7%, 7%] and inject a random Gaussian blur with $\sigma = 1$ for each training image.

For end-to-end training, we resize the cropped face images to $256 \times 256 \times 3$. After several convolution, residual, and pooling blocks, the input shape of hourglass network is resized to $64 \times 64 \times 256$ as shown im Fig.2. In details, we set the stack number equals two, $\sigma = 4$ and $\delta = 1$ for kernel radius and $\theta = 0.5$ for hinge loss.

5.4. Comparison with state-of-the-art

Following [48], we evaluate our method using the inter-ocular distance normalized median per image error. Since coordinate error is very sensitive to outliers, the median is a more robust evaluation metric than the mean of errors. Results of landmark localisation on the 300-W dataset is shown in Table 3. Our method outperforms all previous baselines in the most difficult 68-point Challenge scenario. The Challenge scenario contains images with large poses and illumination variations and is not yet well tackled in

Method	Common	Challenge	Full Set
SDM[46]	5.57	15.40	7.52
ESR[7]	5.28	17.00	7.58
LBF[33]	4.95	11.98	6.32
CFSS[52]	4.73	9.98	5.76
MDM[39]	4.83	10.14	5.88
TCDCN[50]	4.80	8.60	5.54
Two-StageOD[28]	4.36	7.56	4.99
Two-StageGT[28]	4.36	7.42	4.96
RDR[45]	5.03	8.95	5.80
Pose-Invariant[20]	5.43	9.88	6.30
SBR[15]	**3.28**	7.58	4.10
SANOD[14]	3.41	7.55	4.24
SANGT[14]	3.34	6.60	3.98
Ours	3.44	**5.75**	**3.89**

Table 4. A comparison of different methods on 300-W dataset. Normalized Mean Error(%) metric is used for evaluation.

Method	COFW-68	AFLW-Full
RCPR[6]	8.76	3.73
TCDCN[50]	7.66	–
HPM[18]	6.72	–
CFSS[52]	6.28	3.92
Two-StageOD(CVPR17)[28]	–	2.33
Two-StageGT(CVPR17)[28]	–	2.17
SBR(CVPR18)[15]	–	2.14
SAN(CVPR18)[14]	–	1.91
Ours	**5.32**	**1.87**

Table 5. A comparison of different methods on COFW dataset. Normalized Mean Error(%) metric is used for evaluation

some approaches.

For 300-W Common set, we obtain comparable results without the help of model fitting in a post-processing step, as adopted in [29]. Compared with FC-LGCN model without model fitting, we outperforms it in challenge dataset with or without face outlines. It suggests that our proposed method outperforms FC-LGCN in wild scenarios. We also calculate the performance drop from 68 points to 49 points without face outline [46], and the results show that our method achieves superior performance and depicts good generalization ability to both facial outline points and interior points.

To compare with other methods which is evaluated by Normalized Mean Error metric, we also conduct related experiments on three public available datasets (300-W, COFW, and AFLW). In our experiments, inter-ocular distance [14, 28] is still applied as the normalization term for 300-W and COFW-68 and width (or height) of the face bounding box as the normalisation term for AFLW.

As shown in Table 4, our approach achieves 5.75% mean error and outperforms all the other approaches in 300-W Challenge and Full set which demonstrates the effectiveness of out method towards large pose human face images.

Additionally, to comprehensively evaluating the robustness of our method, we conduct cross-dataset experiments utilizing the model we trained on 3,148 300-W training images without any data from other datasets. We evaluate our algorithm on COFW-68 dataset and AFLW dataset respectively. Since original COFW dataset contains less than 68 landmarks, we utilize re-annotated COFW-68 dataset produced by [18]. Thanks to the generalization ability of our proposed model, we achieve the best result on COFW-68 and AFLW-Full dataset. Superior results to SAN [14] and SBR [15] on AFLW dataset which has different annotation protocols demonstrate the robustness of our method on a large scale dataset. The results is shown in Table 5. Noted that our method is not trained with AFLW dataset and we use two-stacked hourglass network which is half size of most hourglass based networks.

6. Conclusion

In this paper, we propose a Stacked Multi-Target network and suggest to dynamically adjust kernel function radius temporally or spatially during training process to compensate the discrepancy between heatmap optimization and coordinate evaluation. Besides, we introduce to utilize hinge loss and constrain the output of sub-network coordinate regression in a reasonable range to avoid damage to primary hourglass network. Finally, we theoretically analyse different loss functions used in heatmap regression-based facial landmark localisation methods and propose a new loss function named Bilateral Inhibition Cross Entropy loss. The introduced loss function not only brings bilateral inhibition into CNN output heatmaps but also enlarges the margin between elements in output maps. Empirical evaluations on three datasets show that our proposed method performs effectively and robustly.

References

[1] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3451, 2013. 7

[2] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Continuous conditional neural fields for structured regression. In *European Conference on Computer Vision (ECCV)*, pages 593–608, 2014. 7

[3] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940, 2013. 7

[4] A. Bulat and G. Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face

alignment with limited resources. In *IEEE International Conference on Computer Vision (ICCV)*, page 4, 2017. 4, 5

[5] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *IEEE International Conference on Computer Vision (ICCV)*, page 4, 2017. 2

[6] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1513–1520, 2013. 7, 8

[7] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision (IJCV)*, 107(2):177–190, 2014. 8

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018. 2

[9] W. Chen, Q. Zhou, and R. Hu. Face alignment by combining residual features in cascaded hourglass network. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 196–200, 2018. 3, 5

[10] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):681–685, 2001. 1

[11] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995. 1

[12] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *British Machine Vision Conference (BMVC)*, page 3, 2006. 1

[13] J. Deng, J. Guo, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*, 2018. 1

[14] X. Dong, Y. Yan, W. Ouyang, and Y. Yang. Style aggregated network for facial landmark detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 8

[15] X. Dong, S.-I. Yu, X. Weng, S.-E. Wei, Y. Yang, and Y. Sheikh. Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 360–368, 2018. 8

[16] P. Dou, S. K. Shah, and I. A. Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21–26, 2017. 1

[17] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[18] G. Ghiasi and C. C. Fowlkes. Occlusion coherence: Detecting and localizing occluded faces. *arXiv preprint arXiv:1506.08347*, 2015. 7, 8

[19] S. Honari, P. Molchanov, S. Tyree, P. Vincent, C. Pal, and J. Kautz. Improving landmark localization with semi-supervised learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1546–1555, 2018. 2

[20] A. Jourabloo, M. Ye, X. Liu, and L. Ren. Pose-invariant face alignment with a single cnn. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3219–3228, 2017. 8

[21] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1867–1874, 2014. 2

[22] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2144–2151, 2011. 7

[23] H. Lai, S. Xiao, Y. Pan, Z. Cui, J. Feng, C. Xu, J. Yin, and S. Yan. Deep recurrent regression for facial landmark detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(5):1144–1157, 2018. 3, 4, 5

[24] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *European Conference on Computer Vision (ECCV)*, pages 679–692, 2012. 7

[25] J. Li, C. Xiong, L. Liu, X. Shu, and S. Yan. Deep face beautification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 793–794, 2015. 1

[26] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593, 2017. 1

[27] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[28] J.-J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3691–3700, 2017. 2, 7, 8

[29] D. Merget, M. Rock, and G. Rigoll. Robust facial landmark detection via a fully-convolutional local-global context network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 781–790, 2018. 2, 7, 8

[30] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 483–499, 2016. 2, 4, 7

[31] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy. Towards accurate multi-person pose estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 6, 2017. 2

[32] R. Ranjan, C. D. Castillo, and R. Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017. 1

[33] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment via regressing local binary features. *IEEE Transactions on Image Processing*, 25(3):1233–1245, 2016. 8

[34] J. Robinson, Y. Li, N. Zhang, Y. Fu, and S. Tulyakov. Laplace landmark localization. *arXiv preprint arXiv:1903.11633*, 2019. 2

[35] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015. 2

[36] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 397–403, 2013. 7

[37] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. 1

[38] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Pérez, and C. Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *IEEE International Conference on Computer Vision (ICCV)*, page 5, 2017. 1

[39] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4177–4187, 2016. 2, 8

[40] G. Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3659–3667, 2015. 7

[41] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic. Copula ordinal regression for joint estimation of facial action unit intensity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4902–4910, 2016. 1

[42] H. Wang, Y. Wang, Z. Zhou, X. Ji, Z. Li, D. Gong, J. Zhou, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. *arXiv preprint arXiv:1801.09414*, 2018. 1

[43] Wikipedia contributors. Lateral inhibition — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Lateral_inhibition&oldid=859550316`, 2018. [Online; accessed 13-November-2018]. 6

[44] Y. Wu and Q. Ji. Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3400–3408, 2016. 2

[45] S. Xiao, J. Feng, L. Liu, X. Nie, W. Wang, S. Yan, and A. A. Kassim. Recurrent 3d-2d dual learning for large-pose facial landmark detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1642–1651, 2017. 8

[46] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 532–539, 2013. 7, 8

[47] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 7, 2017. 2

[48] A. Zadeh, Y. C. Lim, T. Baltrusaitis, and L.-P. Morency. Convolutional experts constrained local model for 3d facial landmark detection. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2519–2528, 2017. 2, 7

[49] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European Conference on Computer Vision (ECCV)*, pages 1–16, 2014. 7

[50] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision (ECCV)*, pages 94–108, 2014. 7, 8

[51] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):918–930, 2016. 2

[52] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4998–5006, 2015. 7, 8

[53] S. Zhu, C. Li, C.-C. Loy, and X. Tang. Unconstrained face alignment via cascaded compositional learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3409–3417, 2016. 7

[54] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 146–155, 2016. 7