Revisting Quantization Error in Face Alignment

Xing Lan[1,2] Qinghao Hu[1] Jian Cheng[1,2*]
[1]Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
{lanxing2019,huqinghao2014}@ia.ac.cn, jcheng@nlpr.ia.ac.cn

Abstract

Recently, heatmap regression models have become the mainstream in locating facial landmarks. To keep computation affordable and reduce memory usage, the whole procedure involves downsampling from the raw image to the output heatmap. However, how much impact will the quantization error introduced by downsampling bring? The problem is hardly systematically investigated among previous works. This work fills the blank and we are the first to quantitatively analyze the negative gain. The statistical results show the NME generated by quantization error is even larger than ***1/3*** *of the SOTA item, which is a serious obstacle for making a new breakthrough in face alignment. To compensate for the impact of quantization effect, we propose a novel method, called* Heatmap In Heatmap (HIH)*, which leverages two categories of heatmaps as label representation to encode the coordinate. And in HIH, the range of one heatmap represents a pixel of the other category of heatmap. Also, we even combine the face alignment with solutions of other fields to make a comparison. Extensive experiments on various benchmarks show the feasibility of HIH and superior performance than other solutions. Moreover, the mean error reaches to* ***4.18*** *on WFLW, which exceeds SOTA a lot.*

1. Introduction

Face alignment, or facial landmark detection, refers to detect a set of predefined landmarks on the human face. It is a fundamental step for many facial analysis tasks including face verification, facial attributes analysis, and 3D face reconstruction.

Owing to the rapid development of deep learning, very recent works based on heatmap regression have got much higher accuracy than direct coordinate regression. Heatmap-based methods use the heatmap as label representation to encode the coordinate of facial landmark labels, so that the supervised learning loss can be quantified. The heatmap is usually generated by the 2-dimension Gaussian distribution density function, and it is characterized by giving spatial support around the ground-truth location, considering not only the contextual clues but also the inherent target position ambiguity [42]. In this way, this label representation method effectively reduces the model overfitting phenomenon during the training procedure. And there is no doubt that heatmap-based methods [29, 20, 34] have dominated the state-of-the-art performance on face alignment.

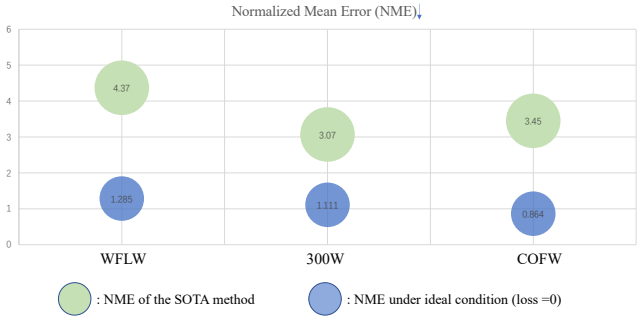

Figure 1. NME Comparison between the SOTA methods and quantization error under ideal condition (loss=0). The normalized mean error (NME,%) is tested on the various benchmark testsets.

However, there is a major obstacle for the heatmap representation, which is reflected in two aspects. One is downsampling the raw face images at arbitrarily large resolutions into a prefixed small resolution with a data preprocessing procedure. The other is that the model computation involves downsampling layers, which cause the size of the output is usually smaller than the input image. The prediction is considered as the location (integer type) with maximal activation, which is required to remap back to the original coordinate space. Both sides show that quantization error is introduced during the total resolution reduction, which will cause precision lost.

How much impact does the quantization error cause on the final results? Some works[42, 26, 23, 21] have considered the phenomenon, but no one answers this question. We are the first to conduct quantitative analysis, and the experiments use the most common setting from 256x256 in-

put images to 64x64 heatmaps and use round operation for raw label to input label. We calculate the distance between the ground-truth locations and the results from decoding the ground-truth heatmap, and the results (Tab.1) show that the NME generated by quantization error is even larger than 1/3 of the SOTA item. That is to say, how to make up for the quantization effect is a key problem in face alignment.

DARK [42] and G-RMI [26] leverage neighbor points information to softmax the result. Cornernet [21] and Centernet [13] consider another map to represent the offset error. Some works directly combine the solutions of other fields (pose estimation, object detection) to fix the issue. For example, LUVLI [20] trivially use the softmax way from [23]. However, the number of facial landmarks(68,98,106) is larger than the body skeletons(15,19), the proportion of the face($>$80%) in the image area is bigger than human body ($<$40%), and the key-point location is much denser. It is not appropriate to directly use the softmax solutions from pose estimation in face alignment.

In order to address the above problems, we propose a novel *Heatmap in Heatmap(HIH)*, which uses the two categories of heatmaps as label representation to encode coordinate of landmark labels. One keeps consistent with the original to represent integer value, called integer heatmap. And another one represents the remaining decimal value of location, named decimal heatmap. The range of one decimal heatmap represents one pixel in the corresponding integer heatmap, and the resolution of the decimal map means the degree of offset accuracy. Both heatmaps take the location of maximal activation as the coordinate result, and the final prediction is represented by the sum of the integer coordinate and the normalized decimal coordinate.

To verify the effectiveness of our method, we use two different architectural designs (CNN and Transformer) to implement HIH. Moreover, we even combine the face alignment with solutions of other fields to make a comparison. Extensive experiments on various benchmarks are conducted to demonstrate the superior performance of HIH than other solutions. Also to our best knowledge, it is the first attempt at extending success of Transformer to face alignment. Specifically, the NME even reaches 4.18 on WFLW, which outperforms SOTA a lot.

The main contributions in this work are summarized as follows.

1. We are the first to quantitatively analyze the quantization error from downsampling. The statistical results show the NME generated by quantization error is even larger than **1/3** of the SOTA item. We suppose it is a serious problem that needs to focus on.

2. We propose a novel Heatmap in Heatmap(HIH) to fix the mentioned issues, which uses two categories of heatmaps as feature representation to encode coordinate labels. We further use two designs (CNN and Transformer) implementing HIH independently to verify the effectiveness of our method. In the literature of face alignment, it is the first attempt on extending success of Transformer.

3. Detailed experiments show HIH performs feasible in compensating quantization effect and outperforms other solutions. Our approach outperforms the state-of-the-art algorithms by a significant margin on various benchmarks.

4. Results on various solutions have demonstrated the accuracy decline caused by the quantization error is a serious problem. And methods about compensating the error is a fast and effective way to improve performance.

2. Related Work

This work has a close connection with areas of facial landmark detection(face alignment) and quantization error.

2.1. Face Alignment

Facial alignment has been intensively studied for years, and people have proposed various methods in this area which has achieved great progress. Early methods were based on active shape models (ASMs) [31] and active appearance models (AAMs) [5]. Later constrained local models (CLMs) [6], SDM [38] were also proposed. And the subsequent cascaded regression methods [46] promote the accuracy of face alignment. Due to the success of deep learning, more and more deep learning-based methods have been applied in this area, and have dominated the state-of-the-art performance on face alignment. These methods mainly fall into two categories: direct coordinate regression [30, 43, 14, 11] and heatmap-based regression [18, 9, 39, 35, 29, 20]. Heatmap-based methods perform much better than direct coordinate regression due to their spatial support.

Heatmap-based Regression Models. Heatmap regression methods are used to indirectly map the input image to the probability heatmaps, each of which represents the respective probability of a landmark location. In the inference stage, the location with the highest response on each heatmap indicates the corresponding landmark. DAN [18] is the first method that combines heatmap with landmarks regression. It is coupled with the affine transformation to a standard shape. Benefit from Hourglass [25], heatmap regression methods have been successfully applied to landmark localization problems and have achieved state-of-the-art performance. JMFA [9] achieves high localization accuracy with a stacked hourglass network [25] in the Menpo competition [41] . [39] adopts a supervised face transformation to normalize the faces, then employed hourglass [25] to regress it. LAB [35] proposes to use additional boundary lines as the geometric structure of a face image to help facial landmark localization. HRNet [29] maintains multi-resolution representations in parallel and exchanges information between these streams to obtain a final representa-

tion with great semantics and precise locations. Recently, LUVLI [20] first introduces the concept of parametric uncertainty estimation as well as considers the visibility likelihood.

2.2. Quantization Error

Indirect inference through a predicted heatmap offers several advantages over direct prediction [1]. The landmarks are estimated by the highest response of each predicted heatmap. In general, due to the high computation demand for maintaining high resolution, the shape of output heatmap is smaller than the original input image by using downsampling operations. Specifically in face alignment networks [29, 20, 39], the shape of the input image is 256 pixel for width and height while the output heatmap is 64 pixel. That is to say, no matter how effective the above methods are, there will be the quantization error from 4 times down-sample operations influence the accuracy. However, they suffer from quantization error since the heatmap's argmax is only determined to the nearest pixel. What's more, the size of original face images is usually much larger than input, and they are transformed into input data by a series of preprocessing operations like rotation and resize. These operations will inevitably lead to the existence of a decimal part of the transformed keypoints' value. However, the decimal part of the keypoints cannot be directly obtained even if the heatmap is restored to the same size as input, which leads to another quantization error in the final results.

2.3. Solutions in Other Fields

At present, most of existing solutions about quantization error are based on human pose estimation [26, 25, 42, 23] and object detection[21, 13], which can also be divided into two categories: based on the heatmap result, and based on the offset branch.

Based on the heatmap result. These coordinate decoding methods are only based on the heatmap. G-RMI [26] also take the mask into consideration and use all the surrounding response to vote for the weighted value. Hourglass [25] first identifies the coordinates of the maximal and second maximal activation, and then uses the position and gradient to fine-tune the maximum response coordinate. DARK [42] first smooth the irregular heatmap by Gaussian smoothing, then expand Taylor's formula to fit the derivative of the Gaussian function, and finally calculate the position of the maximum. LUVLI [20] draws lessons from [23] and considers the irregular heatmap distribution . Thus, it subsequently carries out a softmax operation and then takes advantage of joint spatial mean to get final coordinates.

Based on the offset branch. These methods [21, 13, 44] propose to use a new branch to regress the offset caused by quantization, which will produce no bias in encoding. Recently, Cornernet [21], Centernet [13] and other keypoint-detection methods [44] in object detection propose a shared offset map to represent offset. PIPNet [17] and some pose estimation methods use convolution and linear operations to direct regress the remaining offset.

However, in the area of face alignment, there is hardly such solution to address the quantization error. Different from other tasks, human face accounts for a larger proportion of the image, the precision threshold requires to be more accurate, the distribution of facial landmarks is very dense, and the number of facial landmarks is much larger than others. Thus, it is not appropriate to copy the fixed design straightforwardly. Our experimental results show that directly combining these methods with face alignment is inferior to our proposed method, which demonstrate our contributions to reducing quantization errors for heatmap based regression methods.

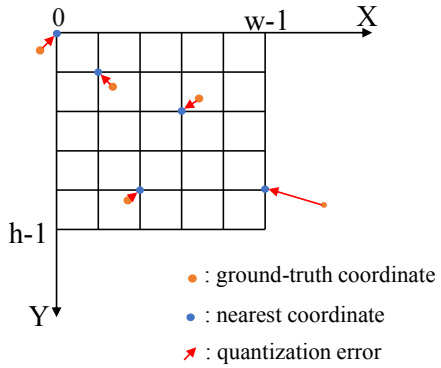

Figure 2. Illustration of quantisation error in heatmap. The yellow points are ground-truth coordinates, and the blue points are nearest coordinates after $round$ operation. Each red arrow is the distance between them, which represents the quantization error introduced by mentioned operation.

3. Methodology

In this section, we first discuss the quantization error on the based-heatmap face alignment. And then, we describe in detail our proposed solution, which consists of the encoding and decoding method and two architecture designs. Finally, we discuss the differences from previous solutions of object detection and pose estimation.

3.1. Quantization error in heatmap-based methods

Heatmap is characterized by giving spatial support around the ground-truth location, which effectively reduces the model overfitting risk in training.

However, there is a major obstacle for the heatmap label representation, which is reflected into two sides. One is that the computational cost is a quadratic function of the input image resolution, preventing the CNN models from processing the typically high-resolution raw imagery data.

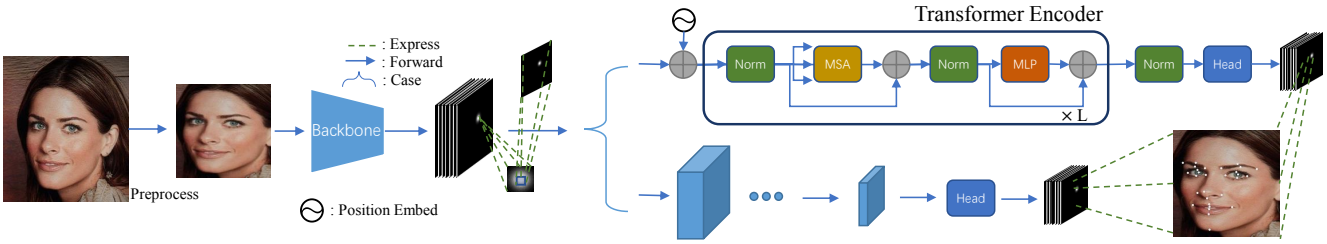

Figure 3. The overall pipeline of *Heatmap In Heatmap* (HIH). The whole quantization error is introduced from two sides: data preprocessing and model downsampling. There exist two kinds of heatmaps that jointly represent coordinate labels: one identifies integer coordinate and the other refines its float location. The decimal heatmap refines the pixel of maximal activation in integer heatmap. We have designed two different schemes to realize the subsequent structure: upper one is based Transformer and the below is based CNN.

To be computationally affordable, a standard strategy is to downsample the face images into a prefixed small resolution with a data preprocessing procedure. The other involves down-sample layers in networks to gather global information and to reduce memory usage. When we remap the location output from the lower coordinate space to the higher resolution space, precise offsets will be lost. The generation process of the quantization error is shown in Fig.2 The distance between ground-truth value and nearest integer value is the corresponding quantization error. Since the location cannot recover precisely, Both down-sample operations cause the quantization error introduced.

None of previous work has systematically investigated how much impact the quantization error has on the final results. We are the first to conduct quantitative analysis, and the statistical results (Tab.1) show that the NME generated by quantization error is even larger than 1/3 of the SOTA item. That is to say, How to compensate for the quantization effect is a key problem in face alignment.

3.2. Heatmap in Heatmap

In order to address the problem, we propose a novel *Heatmap in Heatmap*(HIH) method, which uses two categories of heatmaps to jointly represent coordinate labels. As shown in Fig.3, one identifies the integer location, named *integer heatmap*, and the other locates its float location, called *decimal heatmap*. At the integer-heatmap coordinate space, integer heatmap represents the integer part from 0 to the resolution, and the decimal heatmap represents the remaining offset from 0 to 1. In other words, the range of one decimal heatmap represents the range of one pixel in integer heatmap. And to verify the effectiveness of HIH, we have designed two different schemes to realize the structure.

3.2.1 Encoding and Decoding

Different from common heatmaps only representing the integer part, we also propose a novel decimal heatmap to represent the decimal part as the offset. HIH's integer and decimal location are represented by the same heatmap method but with different parameters(sigma, resolution, *etc.*). They both use the maximal response location as the coordinate value. The range of a decimal heatmap represents the range of one pixel of the integer heatmap. Specifically, given a task for detecting N facial landmarks, an original image is transformed into the input image (h_i, w_i) after n_p times downsampling during preprocessing. And the input image is transformed into an integer heatmap (h, w) and a decimal heatmap (h_o, w_o) during model computation.

Encoding design. The integer heatmap is computed as same as before methods[29, 25]. The relative offset value o_k, corresponding to the integer-heatmap coordinate space, is computed as

$$o_k = (\frac{x_k}{n} - \lfloor \frac{x_k}{n} \rfloor, \frac{y_k}{n} - \lfloor \frac{y_k}{n} \rfloor) \tag{1}$$

where n is the down-sample factor (contain preprocessing and model computation), x_k and y_k are the x- and y-ground-truth coordinates of the k-th landmark on the original image. $(\lfloor \frac{x_k}{n} \rfloor, \lfloor \frac{y_k}{n} \rfloor)$ denotes the corresponding integer value on the integer heatmap. While the final offset value $o_k^\star$ needs to be enlarged by the resolution of the decimal heatmap. $o_k^\star$ represents the coordinates of maximal activation in decimal heatmap and is computed as

$$o_k^\star = \lfloor o_k * (w_o, h_o) \rceil \tag{2}$$

And the decimal-heatmap generation is computed as similar as Two-dimensional Gaussian distribution density function

$$O = \begin{cases} e^{\frac{-||o-o_k^\star||_2^2}{2\sigma^2}} & ||o - o_k^\star||_\infty \leq 3\sigma \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where o is the coordinate of each pixel in the k-th decimal heatmap, and O is the probability value corresponding to location o. There are $N * h_o * w_o$ such probability values in decimal heatmaps.

Decoding design. The decimal and integer heatmaps both are decoded by identifying the location of the maximal activation. Suppose the predicted integer heatmaps $\hat{H}_c$ and

decimal heatmap $\hat{H}_o$, which are both composed of the location's activation. The final normalized coordinate is predicted by

$$\hat{P_k} = \frac{\text{argmax}(\hat{H_c^k}) + \frac{\text{argmax}(\hat{H_o^k})}{(w_o,h_o)}}{(w,h)} \quad k = 1 \dots N \qquad (4)$$

where $\hat{H_c^k}$ is the predicted k-th integer heatmap, and $\hat{H_o^k}$ is the k-th decimal heatmap. In general, $(h_o, w_o) > \frac{(h_i,w_i)}{(h,w)}$. The quantization error can be controlled to a small degree according to the decimal-heatmap resolution.

3.2.2 Architecture design

As shown in Fig.3, the whole network consists of three components: the backbone, integer part and offset part. Following previous works[34, 35, 9], our backbone is based on the stacked HG architecture[25]. For each HG, the output heatmap (integer heatmap) is supervised with ground truth. And the offset part uses the fusion of the backbone and integer heatmap as input, we propose two implementation designs of structure (CNN and Transformer) to verify the effectiveness of HIH. For simplicity, we use the symbols HIH_C and HIH_T to represent the CNN-based and Transformer-based networks.

CNN-based structure. CNN has powerful spatial information extraction capabilities. To keep the same computation as before solutions, we use directly the above fusion input in the remaining part. The first layer is a convolution with batch normalization and ReLu activation, which stride equals 1. The maxpool layer is able to reduce computation by downsampling feature maps, which is behind the first convolution layer. The subsequent operation is a list of convolution blocks to learn spatial information, which uses the basicblock of resnet[16]. In the above basicblock, we set the stride into 2 in the first convolution, and to increase the receptive field quickly, we modify the kernel size of the second convolution to 3. The final regression head is composed of one convolution similar to the first layer and another single convolution, which is as same as Eq.6.

Transformer-based structure. Transformers[33, 10] have proven its dominated power in various natural language processing tasks. Inspired by the success of Transformers in the NLP field, there have been interesting attempts on extending the success of the Transformers to the computer vision field. DETR [4] use a transformer architecture for end-to-end object detection. The subsequent Deformable DETR [48] modified the decoder structure to improve computational efficient. Most recently, ViT [12] treated the input images as 16x16 words and obtained excellent results on image classification.

We follow the standard Transformer architecture [33] as closely as possible. The overall architecture of this network is inspired by the DETR[4] and Vit[12], which remains the same backbone as before and leverages only the encoder transformer. Before inputting features into the transformer encoder, we split the given features into patches and each patch represents a word in NLP. Given an input image $I \in \mathbb{R}^{3\times h_i \times w_i}$, we assume that the CNN backbone outputs a integer heatmaps $H_c \in \mathbb{R}^{N\times h\times w}$ and the subsequent fusion features $H_u \in \mathbb{R}^{d\times h\times w}$. Then, the fusion features are reshaped into a sequence of flatten 2D patches $X_p \in \mathbb{R}^{N_p\times(p^2\cdot d)}, i.e.$, where p is the patch size, and $N_p = \frac{h\times w}{p^2}$ is the number of patch images. Actually, the resolution product of one decimal heatmap should be $N_p = h_o \times w_o$. We flatten the patches and map to D dimension with a trainable linear projection $E \in \mathbb{R}^{(p^2\cdot d)\times D}$. To maintain the position information of each patch, we add learnable position $E_{pos} \in \mathbb{R}^{N_p\times D}$ for each patch. The calculation of the transformer encoder can be formulated as

$$\begin{aligned}
&y_0 = X_pE + E_{pos} \\
&q_i = k_i = v_i = LN(y_{i-1}) && i = 1 \dots L \\
&y_i' = MSA(q_i, k_i, v_i) + y_{i-1} && i = 1 \dots L \\
&y_i = MLP(LN(y_i')) + y_i' && i = 1 \dots L \\
&y_t = LN(y_L)
\end{aligned} \qquad (5)$$

where L denotes the number of encoder layers, MSA represents the multi-head self-attention module in the conventional transformer model[33]. The MLP consists of two fully connected layers, which represents the feed forward network.

And then the transformer output feature $y_t \in \mathbb{R}^{Np\times D}$ are reshaped to 3D feature $X_t \in \mathbb{R}^{D\times h_o\times w_o}$. The final regression head reset the channel from D to N, which is computed as

$$y = Conv(Relu(BN(Conv(X_t)))) \qquad (6)$$

where both convolution layers set kernel size and stride to 1, and the output y represents the decimal heatmap.

3.3. Differences from solutions in other fields

In addition to our HIH, there are a few works of other fields that also consider the problem of quantization errors. The works contains pose estimation [26, 25, 42] and object detection [21, 13, 44], which can be divided into three categories: *with second maximal*(WSM), *with offset value*(WOV), *with offset map*(WOM). As shown in Tab.1, we have counted these NMEs under the ideal state (loss=0). In this paper, we are inspired by them to try to solve the impact of quantization error. For more details on these trivial solutions, please see the supplementary material.

WSM is the most common practice in decode the heatmap, which costs little computation comparing to DARK[42]. Under ideal conditions, WSM still exists big

bottlenecks. Thus, we use another heatmap information to better fix the bottleneck caused by the quantization error. Different from WOV, we use the map to regress indirectly, which can elegantly address the limitations of the lack of spatial and contextual information. Different from WOM, we use the same as original heatmaps instead of shared x-offset and y-offset maps to address the conflicting phenomenon.

4. Experiments

4.1. Experiments Settings

Datasets Our experiments use the datasets WFLW[35], 300W[28], COFW[3]. These datasets are challenges due to images with large head pose, occlusion, and illumination variations.

WFLW[35] dataset is a challenging one, which contains 7,500 faces for training and 2,500 faces for testing, based on WIDER Face[40] with 98 manually annotated landmarks. The faces in WFLW introduce large variations in pose, expression, and occlusion. The testing set is further divided into six subsets for a detailed evaluation, namely, pose (326 images), expression (314 images), illumination (698 images), make-up (206 images), occlusion (736 images) and blur (773 images).

300W[28] dataset provides 68 landmarks for each face, where the face images are collected from LFPW[2], AFW[47], HELEN[22], XM2VTS[24] and IBUG. Following the protocol used in [47], all 3148 training images are from the training set of LFPW and HELEN, and the full set of AFW. The 689 testing images are from the testing set of LFPW and HELEN, and the full set of IBUG. The testing images are further divided into three sets: the test samples (554 images) from LFPW and HELEN as the common subset, the 135-image IBUG as the challenging subset, and the union of them as the full set.

COFW[3] dataset consists of 1345 images for training and 507 faces for testing, where the face images have large variations and occlusions. 29 landmarks are provided for each face.

Evaluation Metrics Following previous works[29, 19, 35], We use the standard metrics NME, AUC and FR, which are the most authoritative for face alignment. In each table, we report results using the same metric adopted in respective baselines.

Normalized Mean Error (NME) is defined as:

$$NME(\%) = \frac{1}{N}\sum_{k=1}^{N}\frac{||P_k - \hat{P}_k||_2}{d} \times 100 \qquad (7)$$

where we use the distance between the outer corners of the two eyes as normalized factor d. P_k and $\hat{P}_k$ denote the ground-truth and prediction location of the k-th landmark. For simplicity, we magnify 100 times to omit the % symbol. Lower NME is better to facial landmark detector.

Area Under the Curve (AUC) is calculated based on the cumulative error distribution (CED) curve. The AUC for a testset is computed as the area under the curve, up to the cut off NME value. Higher AUC represents better detector.

Failure Rate (FR) is another metric to evaluate localization quality, which refers to the percentage of images in the testset when NME is larger than a perfixed threshold. And we also omit the % symbol to simplify variables. The Lower FR is corresponding to the better performance.

Implementation Details The raw face images are cropped according to a retina-face [8] detector (except COFW by itself bounding box). The cropped images then are processed by a set of data augmentation operations for the training dataset. Following previous works, these augmentation operations contain horizontal flip (50%), rotation ($\pm 30°, 50\%$), occlusion (50%) and Gaussian blur (30%). And then the cropped images of training or testing are resized to input data with 256×256 resolution. The output heatmap from backbone is 64×64, we use no more than 2 stacks of HG. And for the recover part, They control the number of channels to make the parameters the same. About baseline without recovery part, we expand the channel from 256 to 280, which is used to counteract the effect of increasing the recovery part.

All experiments are conducted on the RTX 2080Ti device without pretraining. For training procedure, we used the Adam optimizer with init learning rate 1e-5, which decreases 10 times every 50 epoch. For transformer structure in HIH, depth is 4, embed dim is 512, the number of head is 2 and mlp ratio is 3. All parts are supervised by L2 loss, except the recover part of WOM uses SmoothL1 followed by [21, 13, 44] More training details are shown in the supplementary material.

Method	Full	Com.	Cha.
LAB [35]	3.49	2.98	5.19
Wing [14]	3.60	3.01	6.01
AS w. SAN [27]	3.86	3.21	6.49
HRNET [29]	3.32	2.87	5.15
AWing [34]	**3.07**	**2.72**	**4.52**
LUVLI [20]	3.23	2.76	5.16
HIH_C(ours)	3.36	2.95	5.04
HIH_T(ours)	3.33	2.93	5.00

Table 3. Comparison in NME on the 300W testsets.

Method	NME(↓)	AUC_{10}(↑)	FR_{10}(↓)
RAR [37]	6.03	-	4.14
DAC-CSR [15]	6.03	-	4.74
LAB [35]	5.58	-	2.76
PCD-CNN [19]	5.77	-	3.73
Wing [14]	5.07	-	3.16
HRNET [29]	3.45	-	0.19
AWing [34]	4.94	0.489	0.99
HIH_C(ours)	3.29	0.671	0
HIH_T(ours)	**3.28**	**0.672**	**0**

Table 4. Comparisons in NME, AUC and FR on the COFW.

4.2. Comparsion with State-of-the-art Methods

We compare our networks HIH_T and HIH_C against the state-of-the-art methods on various benchmarks. Note again the above networks are based on 1-stack and 2-stack HG[25], which are slight enough compare to SOTAs. Tab.1 shows the NME results and network information on WFLW, and Tab.2 shows the comparisons in AUC and FR on these networks. From the tables, we observe that our lightweight network HIH(1-stack) is even better in NME metric than SOTA, even with quarter flops. And for the other two met-

Method	Year	Params	Flops	Fullset	Pose	Exp.	Ill.	Mu.	Occ.	Blur
LAB [35]	2018	12.26M	18.96G	5.27	10.24	5.51	5.23	5.15	6.79	6.32
Wing [14]	2018	25M	-	4.99	8.43	5.21	4.88	5.26	6.21	5.81
DCFE [32]	2018	-	-	4.69	8.63	6.27	5.73	5.98	7.33	6.88
DeCaFA [7]	2019	∼10M	-	4.62	8.11	4.65	4.41	4.63	5.74	5.38
HRNET [29]	2019	9.66M	4.75G	4.60	7.94	4.85	4.55	4.29	5.44	5.42
AS w. SAN [27]	2019	35.02M	33.87G	4.39	8.42	4.68	**4.24**	4.37	5.60	4.86
AWing [34]	2019	24.15M	26.8G	4.36	7.38	4.58	4.32	4.27	5.19	4.96
LUVLI [20]	2020	-	-	4.37	-	-	-	-	-	-
HIH_C(1-stack)	ours	10.37M	6.99G	4.31	7.40	4.36	4.52	4.08	5.17	5.00
HIH_T(1-stack)	ours	24.08M	6.91G	4.29	7.41	4.34	4.58	4.12	5.14	4.91
HIH_C(2-stack)	ours	14.47M	10.38G	**4.18**	**7.20**	**4.19**	4.45	**3.97**	**5.00**	4.81
HIH_T(2-stack)	ours	28.18M	10.29G	4.21	**7.20**	4.28	4.42	4.03	**5.00**	**4.79**

Table 1. Comparison in inter-ocular normalized mean error (NME) on the WFLW (fullset and all subsets).

Metric	Method	Fullset	Pose	Exp.	Ill.	Mu.	Occ.	Blur
$FR_{10}(\downarrow)$	CFSS [45]	20.56	66.26	23.25	17.34	21.84	32.88	23.67
	DVLN [36]	10.84	46.93	11.15	7.31	11.65	16.30	13.71
	LAB [35]	7.56	28.83	6.37	6.73	7.77	13.72	10.74
	Wing [14]	6.00	22.70	4.78	4.30	7.77	12.50	7.76
	DeCaFA [7]	4.84	21.40	3.73	3.22	6.15	9.26	6.61
	AS w. SAN [27]	4.08	18.10	4.46	2.72	4.37	7.74	4.40
	AWing [34]	**2.84**	**13.50**	2.23	2.58	2.91	5.98	3.75
	LUVLI [20]	3.12	-	-	-	-	-	-
	HIH_C(1-stack)	3.08	15.64	1.91	2.29	2.91	6.79	**3.36**
	HIH_T(1-stack)	3.28	15.95	2.55	2.43	2.91	6.52	3.88
	HIH_C(2-stack)	2.96	15.03	**1.59**	2.58	**1.46**	6.11	3.49
	HIH_T(2-stack)	**2.84**	14.41	2.55	**2.15**	**1.46**	**5.71**	3.49
$AUC_{10}(\uparrow)$	CFSS [45]	0.366	0.063	0.316	0.385	0.369	0.269	0.303
	DVLN [36]	0.456	0.147	0.389	0.474	0.449	0.379	0.397
	LAB [35]	0.532	0.235	0.495	0.543	0.539	0.449	0.463
	Wing [14]	0.554	0.310	0.496	0.541	0.558	0.489	0.492
	DeCaFA [7]	0.563	0.292	0.546	0.579	0.575	0.485	0.494
	AS w. SAN [27]	0.591	0.311	0.549	**0.609**	0.581	0.516	0.551
	AWing [34]	0.572	0.312	0.515	0.578	0.572	0.502	0.512
	LUVLI [20]	0.577	-	-	-	-	-	-
	HIH_C(1-stack)	0.587	0.321	0.576	0.598	0.598	0.514	0.538
	HIH_T(1-stack)	0.587	0.324	0.574	0.596	0.593	0.516	0.540
	HIH_C(2-stack)	**0.597**	**0.342**	**0.590**	0.606	**0.604**	**0.527**	**0.549**
	HIH_T(2-stack)	0.593	0.332	0.583	0.605	0.601	0.525	0.546

Table 2. Comparisons in area under curve (AUC) and failure rate (FR) on the WFLW (fullset and all subsets).

rics, it is very comparable to SOTA. To clearly discover the superiority of HIH, we extend the network depth and enlarge the total computation to the one-third flops of SOTA models. The experimental result shows no matter use which metric, our HIH(2-stack) exceeds SOTA a lot. Specifically, NME reaches 4.18 on the WFLW benchmark.

Tab.4 shows the comparisons in NME, AUC and FR on COFW benchmark. Our HIH greatly surpasses the SOTA methods once again, the NME reaches 3.28, and the FR even declines to 0. The Tab.3 shows the comparison in NME on 300W benchmarks, which contain fullset, challenge set and common set. Our method performs well, which is comparable to the SOTA, especially in the challenge set.

4.3. Comparsion with Compensating Solutions of Other Tasks

Besides comparing the SOTA methods, we also compare HIH with other solutions from object detection and human pose estimation. The detailed solutions are showed in the

Metric	Method	Fullset	Pose	Exp.	Ill.	Mu.	Occ.	Blur
NME(↓)	Baseline	4.53	7.91	4.72	4.74	4.32	5.48	5.23
	WSM	4.35	7.68	4.54	4.56	4.14	5.33	5.07
	WOV	4.34	7.50	4.50	4.56	4.12	**5.17**	4.98
	WOM	4.44	7.83	4.61	4.79	4.15	5.33	5.20
	HIH	**4.31**	**7.40**	**4.36**	**4.52**	**4.08**	**5.17**	**5.00**
$FR_{10}(\downarrow)$	Baseline	4.00	19.02	3.82	3.15	3.88	7.74	4.27
	WSM	3.64	18.10	3.18	3.01	3.88	7.61	3.62
	WOV	3.20	**15.33**	2.55	2.58	**1.94**	6.39	3.88
	WOM	3.28	15.95	2.55	3.01	2.91	**6.25**	4.27
	HIH	**3.08**	15.64	**1.91**	**2.29**	2.91	6.79	**3.36**
$AUC_{10}(\uparrow)$	Baseline	0.568	0.291	0.547	0.579	0.574	0.497	0.516
	WSM	0.586	0.310	0.565	0.596	0.591	0.511	0.531
	WOV	0.585	0.320	0.570	0.594	0.593	**0.517**	**0.538**
	WOM	0.579	0.308	0.559	0.586	0.590	0.507	0.526
	HIH	**0.587**	**0.321**	**0.576**	**0.598**	**0.598**	0.514	**0.538**

Table 5. Comparisons in NME, AUC and FR on the WFLW.

Method	Parmas, Flops	Val	Com.	Cha.	Test.
Baseline	5.33M, 6.67G	3.581	3.157	5.321	4.132
WSM	5.33M, 6.67G	3.420	2.988	5.193	3.999
WOV	5.03M, 6.91G	3.448	3.014	5.228	4.010
WOM	4.76M, 6.85G	3.640	3.127	5.745	4.445
HIH	10.34M,6.93G	**3.356**	**2.946**	**5.038**	**3.835**

Table 6. Comparison in NME on the 300W.

supplementary material. We have controlled the same backbone (1-stack) and the same fusion feature method. And the last parts, which are designed to compensate for the quantization effect, keep about the same amount of calculation. Considering other solutions all use CNN structure as components, we also use HIH_C as the comparison object.

We have conducted statistical experiments on various benchmarks to reveal the seriousness of the influence of quantization error. At the same time, the experiments quantitatively analyze the remaining quantization errors of all methods. We also have conducted experiments on WFLW and 300W, all results show HIH's superior performance than others. The baselines consist of backbone and integer heatmap, without recovery part. And it extends corresponding channels to keep the same flops as other solutions.

The Tab.6 shows the comparisons in NME and the parameter information on 300W. Note that WSM recovers the quantization error based on the original heatmap, so it has

Metric	Network	Fullset	Pose	Exp.	Ill.	Mu.	Occ.	Blur
NME(↓)	HIH^1_C	4.90	8.10	5.00	5.11	4.63	5.68	5.54
	HIH^1_T	4.88	8.11	4.96	5.18	4.66	5.65	5.45
	HIH^2_C	**4.77**	**7.89**	**4.83**	5.04	**4.50**	**5.50**	5.37
	HIH^2_T	4.82	7.92	4.94	**5.03**	4.56	5.52	**5.35**
FR_{10}(↓)	HIH^1_C	3.72	17.79	2.23	**2.58**	2.91	7.20	4.40
	HIH^1_T	3.56	17.79	2.23	2.87	3.88	7.20	4.27
	HIH^2_C	**3.32**	**15.64**	**1.91**	3.01	**1.94**	**5.98**	4.27
	HIH^2_T	3.40	16.87	2.86	**2.58**	2.43	6.25	**4.14**
AUC_{10}(↑)	HIH^1_C	0.531	0.265	0.515	0.541	0.546	0.466	0.486
	HIH^1_T	0.531	0.267	0.515	0.537	0.542	0.468	0.489
	HIH^2_C	**0.541**	**0.284**	**0.529**	**0.548**	**0.552**	**0.480**	**0.497**
	HIH^2_T	0.535	0.272	0.520	0.546	0.550	0.475	0.493

Table 7. Comparisons in NME, AUC and FR on the WFLW. Tested only on the integer heatmap.

Resolution	Ideal	Fullset	Pose	Exp.	Ill.	Mu.	Occ.	Blur
4x4	0.420	4.54	7.92	4.71	4.76	4.36	5.39	5.26
8x8	0.182	**4.31**	**7.40**	**4.36**	4.52	**4.08**	5.17	5.00
16x16	0.091	4.36	7.53	4.42	**4.50**	4.12	**5.15**	**4.98**

Table 8. Comparisons about decimal's resolution in the NME on the WFLW. The Ideal item represent the errors under ideal conditions.

the same parameters as the baseline. From the table, WSM, WOV and our HIH have a positive gain effect, while WOM is negative. The improvement effect of HIH is significantly higher than others, and the effect of WSM is slightly higher than that of WOV. The Tab.5 shows the comparisons in NME, AUC and FR on WFLW. From the table, these four solutions have all achieved positive results, and our HIH performs better than others again. WOV is very close to WSM with a little gap. WOM has achieved the lowest improvement among these methods.

We have discussed in detail the effect of WOM in the supplementary material. And for all solutions, we have analyzed the differences among the heatmaps generated from the network backbone. The detailed discussions are also shown in the supplementary material.

The above results have proved that the accuracy decline caused by quantization error is a serious problem, and methods about compensating the error is a fast and effective approach to improve performance.

4.4. Ablation Study

4.4.1 Comparison with different resolution

Our method involves the decimal heatmap resolution, we further carry out ablation study on the resolution setting, from 4x4, 8x8 to 16x16. The result on WFLW is shown in Tab.8. Although 16x16 will make lower ideal erorr in its high resolution, it regress accurate heatmap more difficult. And 8x8 setting achieves higher performance in fullset and most subsets, which is also set to other experiments.

4.4.2 Comparison between CNN and Transformer

In this subsection, we discuss it in two parts: the precision of whole heatmaps, the precision of integer heatmap. From the Tab.1 and Tab.2, we observe that CNN-based performance is very close to the Transformer-based with a little superiority. For simplicity, HIH^2_T denotes the 2-stack HIH with Transformer structure. Others, and so on. In NME details, HIH_T performs a little better than HIH_C in occlusion and blur situations. And the HIH_C has achieved higher performance in the make-up subset. For other subsets, they have the same detection performance. The result on FR represents the number of the wrong prediction, which has revealed a fact that HIH_T is less likely to make mistakes at high dimension. The result on AUC reflects the distance between prediction and ground-truth, which has revealed a fact that HIH_c performs better in the coordinate accuracy.

Besides the whole heatmaps, we also testify the precision on the integer heatmap with above models, which is showed in the Tab.7. On the only integer heatmap, the normalized mean error and AUC remain the same conclusion and distribution as whole heatmaps. While the failure rates are different, the results on the integer-heatmap FR show that HIH_T's backbone is more likely to make mistakes at high dimension. That is to say, with help of Transformer-based structure, the number of points that are regressed correct will be more.

In summary, HIH_C performs very close to HIH_T with a little superiority. HIH_C is good at locating accurate points with high precision degree and causing lower error in total. HIH_T is expert in finding more points that do not require so high precision. The combination of the two networks may yield greater benefits.

5. Conclusion

In this paper, we identified a serious issue about that downsampling operations bring quantization error in heatmap regression methods. To investigate the impact, we were the first to quantitatively analyze the negative gain. Based on the issue, we proposed a novel Heatmap in Heatmap (HIH) method, which took advantage of two categories of heatmaps: the range of one heatmap represents the range of a pixel on the other heatmap. We even gave two architecture designs with CNN or transformer to verify the effectiveness of our method. Moreover, we also combined the face alignment with solutions of other fields to compare with HIH. Extensive experiments show that our method is feasible in face alignment and outperforms other solutions. Furthermore, as evaluated on various benchmarks, HIH is able to exceed SOTA performance a lot. In the future work, we will extend our HIH design to more fields such as object detection, pose estimation, *etc*.

References

[1] Vasileios Belagiannis and Andrew Zisserman. Recurrent human pose estimation. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 468–475. IEEE, 2017.

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

[3] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE international conference on computer vision*, pages 1513–1520, 2013.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[5] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

[6] David Cristinacce and Tim Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054–3067, 2008.

[7] Arnaud Dapogny, Kevin Bailly, and Matthieu Cord. Decafa: Deep convolutional cascade for face alignment in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6893–6901, 2019.

[8] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019.

[9] Jiankang Deng, George Trigeorgis, Yuxiang Zhou, and Stefanos Zafeiriou. Joint multi-view face alignment in the wild. *IEEE Transactions on Image Processing*, 28(7):3636–3648, 2019.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[11] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 379–388, 2018.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[13] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6569–6578, 2019.

[14] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2235–2245, 2018.

[15] Zhen-Hua Feng, Josef Kittler, William Christmas, Patrik Huber, and Xiao-Jun Wu. Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2481–2490, 2017.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Haibo Jin, Shengcai Liao, and Ling Shao. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *arXiv preprint arXiv:2003.03771*, 2020.

[18] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 88–97, 2017.

[19] Amit Kumar and Rama Chellappa. Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 430–439, 2018.

[20] Abhinav Kumar, Tim K Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8236–8246, 2020.

[21] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018.

[22] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer, 2012.

[23] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5137–5146, 2018.

[24] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, and Gilbert Maitre. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pages 965–966, 1999.

[25] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[26] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4903–4911, 2017.

[27] Shengju Qian, Keqiang Sun, Wayne Wu, Chen Qian, and Jiaya Jia. Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10153–10163, 2019.

[28] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013.

[29] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019.

[30] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.

[31] F Timothy. Active shape models-their training and application. *Computer Vision And Understanding*, 61, 1995.

[32] Roberto Valle, Jose M Buenaposada, Antonio Valdes, and Luis Baumela. A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 585–601, 2018.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[34] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6971–6981, 2019.

[35] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2129–2138, 2018.

[36] Wenyan Wu and Shuo Yang. Leveraging intra and inter-dataset variations for robust face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 150–159, 2017.

[37] Shengtao Xiao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, and Ashraf Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *European conference on computer vision*, pages 57–72. Springer, 2016.

[38] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

[39] Jing Yang, Qingshan Liu, and Kaihua Zhang. Stacked hourglass network for robust facial landmark localisation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 79–87, 2017.

[40] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016.

[41] Stefanos Zafeiriou, George Trigeorgis, Grigorios Chrysos, Jiankang Deng, and Jie Shen. The menpo facial landmark localisation challenge: A step towards the solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 170–179, 2017.

[42] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7093–7102, 2020.

[43] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pages 94–108. Springer, 2014.

[44] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 850–859, 2019.

[45] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4998–5006, 2015.

[46] Shizhan Zhu, Cheng Li, Chen-Change Loy, and Xiaoou Tang. Unconstrained face alignment via cascaded compositional learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3409–3417, 2016.

[47] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2879–2886. IEEE, 2012.

[48] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.