ATF:An Alternating Training Framework for Weakly Supervised Face Alignment

Xing Lan[1,2], Qinghao Hu[2], Jian Cheng[1,2]
[1] University of Chinese Academy of Sciences, Beijing, 100049 China
[2]Institute of Automation, Chinese Academy of Sciences, Beijing, 100190 China

In recent years, various face-landmark datasets have been published. Intuitively, it is significant to integrate multiple labeled datasets to achieve higher performance. Due to the different annotation schemes of datasets, it is hard to directly train models using them together. Although numerous efforts have been made in the joint use of datasets, there remain three shortages in previous methods, *i.e.*, additional computation, limitation of the markups scheme, and limited support for the regression method. To solve the above issues, we proposed a novel *Alternating Training Framework* (ATF), which leverages the similarity and diversity across multiple datasets for a more robust detector. ATF mainly contains two sub-modules: *Alternating Training with Decreasing Proportions* (ATDP) and *Mixed Branch Loss* ($\mathcal{L}_{MB}$). In particular, ATDP trains multiple datasets simultaneously via a weakly supervised way to take advantage of the diversity among them, and $\mathcal{L}_{MB}$ utilizes similar landmark pairs to constrain different branches of the corresponding datasets. Besides, we extend the framework to easily handle three situations: single target detector, joint detector, and novel detector. Extensive experiments demonstrate the effectiveness of our framework for both heatmap-based and direct coordinate regression. Moreover, we have achieved a joint detector that outperforms state-of-the-art methods on each benchmark.

***Index Terms*—Multi-task Learning, Weakly Supervised, Face Alignment.**

I. Introduction

The facial landmark detector or face alignment is designed to locate a set of pre-defined facial landmarks. These points always have specific semantics such as the tip of the nose, the center of the eye, and the corners of the mouth, which represent sufficiently geometric information. Face alignment is a basic step for numerous facial tasks including recognition [5], 3D face reconstruction [6], and face tracking [7].

Due to the rapid development of deep learning [8], [9], [10], recent works of facial landmark detector is considered to be more effective alternative than traditional methods [11], [12], [13], [14]. The methods are mainly divided into two kinds: direct coordinate regression [15], [16], [17], [18] and heatmap-based regression [19], [4], [20]. To achieve extensive assessments, a large number of datasets[4], [2], [3], [1] are published on wild or laboratory conditions. The above datasets have sufficient variations in angle, illumination, occlusion, and makeup. Intuitively, utilizing other variations from different datasets is beneficial to the accuracy of the target facial landmark detector. Because the single benchmark usually concerns only one or a few types of scenarios, combined with other datasets containing different variations can improve the robustness in unseen variation. However, the gap among different markup protocols leads to various landmarks, location difference still exists even if they represent the same meaning, and the artificial labeled variance is uncontrollable for each image. Ideally, all datasets are re-labeled as the target scheme, and a more robust model can be learned based on this solution. However, it is time-consuming for re-labeling other datasets and even impossible in some cases.

Recently, there are some methods[21], [22], [23], [4], [24] aim to obtain a robust model without re-labeling other annotations. However, there are some shortages in previous methods. **1.** Some methods [21], [22], [23] combines extra computational operations or repeated additional computation for dealing the extra data. For example, DVLN [22] designs CD-Net to generate two candidate predictions, which requires two more times computation. **2.** Some approaches are limited to the annotation protocols. For example, LAB [4] considers the boundary information from another dataset to improve the performance of the target landmark detector. However, it only supports the use of a detailed labeled dataset to help another one *e.g.* 68pt→19pt. The boundary generated from the sparse scheme like 19 points[1] has the negative impact on the 68-point[2] detector. **3.** Some methods have limited support for regression methods. LAB uses a heatmap to represent boundary information, which is not suitable for the direct coordinate regression. And some works [24] design multiple linear regressors for landmark union, while it is not suitable for heatmap-based regression.

To solve the above issues, we introduce a novel *Alternating Training Framework* (ATF) for the facial landmark detector. ATF makes full use of the similarity and diversity across multiple datasets, which is mainly implemented through *Alternating Training with Decreasing Proportions* (ATDP) and *Mixed Branch Loss* ($\mathcal{L}_{MB}$). Specifically, ATDP shares parameters across different datasets to obtain high-level representations, which trains multiple datasets simultaneously by alternating batch data. ATDP can integrate different datasets and take advantage of different variations that have no chance to be

Manuscript received August 15, 2021; revised February 22, 2022; accepted March 24, 2022. This work was supported in part by the National Key Research and Development Program of China (Grant No. 2021ZD0201504), National Natural Science Foundation of China (No.62106267), Jiangsu Key Research and Development Plan (No.BE2021012-2), Jiangsu Leading Technology Basic Research Project (BK20192004). Xing Lan, Qinghao Hu, and Jian Cheng are with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation Chinese Academy of Sciences (CASIA) and University of Chinese Academy of Sciences (UCAS), Beijing, China. (e-mail: {lanxing2019,huqinghao2014}@ia.ac.cn.; jcheng@nlpr.ia.ac.cn).

Corresponding author: Jian Cheng (email: jcheng@nlpr.ia.ac.cn).

(a) AFLW [1] 19 landmarks distribution

(b) 300W [2] 68 landmarks distribution

(c) COFW [3] 29 landmarks distribution

(d) WFLW [4] 98 landmarks distribution

Fig. 1: Facial landmarks' location distribution from four benchmarks. The yellow points in different datasets represent similar locations. The pink points represent different locations.

learned in the target dataset. In addition, the model gradually increases the fitting performance of the target branch by *decreasing the proportion*. As shown in Fig.1, inspired by the phenomenon [24] that some points are well labeled with the same geometric meaning across datasets, we propose a novel loss function ($\mathcal{L}_{MB}$) coupled with ATDP to constrain the different branches of corresponding datasets. The proposed loss function enables a branch to learn extra information from other datasets. We find that ATF can implicitly guide the network to learn the general feature of faces from different datasets, and the model trained by ATF performs better in the unseen domain than the original one.

Moreover, we extend ATF for three designed situations: the single detector, the joint detector including all source detectors, and the new novel detector. These situations are most commonly used in real scenarios: when one target detector is enough, when a joint detector is needed, when a new detector needs to be done quickly.

Extensive experiments on benchmarks [2], [4], [3], [1] are conducted to verify the effectiveness of the framework. The normalized mean error (NME) even reaches 3.17 on the 300W leveraging WFLW, which highly outperforms others. And not limited to the markup scheme and regression methods, ATF achieves relative ascension in each benchmark.

In summary, our **main contributions** include:

1. We propose *Alternating Training Framework* (ATF) in face alignment to fix the mentioned issues, which contains ATDP and $\mathcal{L}_{MB}$. ATDP trains multiple datasets with different annotation simultaneously, making it possible to get higher accuracy with the same structure. $\mathcal{L}_{MB}$ leverages the similarity between different datasets to constrain the branches.

2. We extend the framework to the most common scenarios, the single prediction, the joint prediction, and a novel prediction; these may bring some reference to the practical applications. Moreover, we have implemented a joint detector whose performance outperforms SOTA in each benchmark.

3. Detailed experiments show our framework performs feasible in solving the previous problems and make great improvement comparing with baselines in each benchmark, especially in tough variation up to 9.96% improvement.

Different from our prior work[25] focus on leveraging auxiliary datasets to improve the target detector, we extend ATF to three common scenarios and achieve a joint detector via multi-task learning. Moreover, the joint detector outperforms SOTA on each benchmark. We also conduct more detailed experiments and ablation studies, as well as implement more comprehensive metrics to analyze the experiments.

II. Related Work

In this section, we review relevant works on facial landmark detection(coordinate regression models, heatmap-based regression models), methods across datasets, and multi-task learning(hard-parameter sharing, soft-parameter sharing, auxiliary learning).

A. Generic Face Alignment

In the past decades, many classic face alignment methods [11], [12], [13], [14] were proposed in the literature. Recently, due to the power of deep learning, regression models generated by convolutional neural networks occupied the most advanced performance in this field. They were divided into two subcategories: direct coordinate regression [15], [16], [17], [18] and heatmap-based regression [19], [4], [20]. Because of the spatial support[26], the methods based on heatmap show better performance than the direct coordinate regression in all benchmarks.

Direct Coordinate Regression Methods Direct coordinate regression maps facial images to a set of pre-defined coordinates. CNN is usually used to extract the features of the input image and then map the features to coordinates through the fully connected layers. Sun *et al.* [15] first proposed a method of using CNN for face alignment and combining with cascaded regression. Zhang *et al.* [16] designed using other face tasks to achieve good performance. This work first considered multi-task learning in face alignment, while

these works were concerning face analysis. Feng *et al.* [17] introduced special Wing Loss for facial landmark detector and cascaded regression to achieve the higher accuracy of the network. Dong *et al.* developed a style-aggregated network (SAN) [18], which is accompanied by source images of face and style aggregated face images to enable the detector model training together. We take the network constructed of the bottlenecks from Mobilenet-v2[27] as our coordinate-method baseline.

Heatmap-based Regression Methods Heatmap regression methods are highly different from direct coordinate regression methods, especially in encoding and decoding ways. They use high-resolution heatmaps to encode the ground-truth coordinate labels, where each map represents a landmark prediction. During validation and test processes, the decoding method only considers the location in the heatmap with the largest activation, next resizes to the original coordinate system. There exist some examples of obtaining heatmaps which are high resolution. DAN [19] first combined heatmap information with landmark regressor and showed good performance coupled with the coarse-to-fine network. Wu *et al.* [4] designed utilizing extra heatmap to represent boundary information as coarse geometric prediction. HRNet [20] proposed a novel network that maintained multi-level resolutions and representations simultaneously and added the regression head, which had achieved SOTA in several tasks. Recently, LUVLI [28] used the hourglass[29] as the backbone and considered the visibility likelihood firstly. Based on this, it introduced the concept of parametric uncertainty estimation. Considering the powerful extraction capacity of the backbone., we take the HRNET[20] as our heatmap-method baseline.

B. Face Alignment Across Datasets

Comparing with general face alignment, only a few works focus on the performance across dataset. Smith *et al.* [21] was the first work that combined with other different datasets for robust detection. They introduced the method that taken different benchmarks as input and re-labeled a fractional labeled target benchmark Leveraging a union of keypoints defined in the source datasets. However, it caused high computational resources, suffered high relabeled cost, and could not address the single test scenes. Zhu *et al.* [24] transferring keypoint markup across multiple datasets, which were able to extend the original SDM to transductive SDM [14]. The approach was capable of transferring the markup schemes from one benchmark to the other benchmarks by utilizing common facial keypoints (same geometric meaning) as a guideline. It showed good performance across benchmark evaluation and rarely seen scenarios. However, there are still some shortages in the above solution. Especially, it could not address the situation that some samples unsuited for SDM (like one eye is unseen in AFLW [1]). Zhang *et al.* [23] exploited the view of the cascaded regression to output each kind of keypoints in each stage. The approach utilized a unified network which was together with the sparse shape regression. However, the head regressor learned its corresponding inductive bias independently, which their relationship among semantically similar images was not considered. Wu and Yang [22] designed Dataset-Across Network (DA-Net) and Candidate-Decision Network (CD-Net), which performed robust detection. DA-Net was inspired by multi-scale training, used the alternating training mechanism for the network robustness. Besides, CD-Net took the latent output from flipped images into the joint predictions. Nevertheless, because of the different iteration alternating in the late training period, the parameters of whole networks would keep fluctuated. Furthermore, the flipped version would cost double times computation due to the repeated computation. Wu *et al.* [4] introduced the two-stage model based on the hourglass[29] networks to produce the facial geometric map, Next, coupled with the boundary map, the feature was regressed to keypoint predictions via linear operations. However, this method only supported the dense landmark schemes assist the sparse schemes, for which the boundary information that dense landmark offered is more detailed, but not vice versa. Moreover, LAB[4] had limitations on the regression way, which is only applicable for the heatmap-based models. Comparing with the above methods, our framework ATF is not limited to the regression methods, markup schemes and adds no additional latency.

C. Multi-Task Learning

Multi-task Learning (MTL) [30] is widely used in various tasks including computer vision and natural language processing, which obtain great achievement among them. MTL enhances the robustness via utilizing the domain meaning contained in the training data of similar fields.

In the literature of this field, MTL always works with hard or soft parameter sharing [31], [32]. Hard parameter sharing is good at reducing the overfitting, which shares the parameter in a hard way. Moreover, the frequency of hard parameter sharing is higher for which is not difficult to implement. Hard parameter sharing is usually used by sharing the common layers among all tasks and retain their independent output layers. For soft parameter sharing, the different tasks have their independent output, and the corresponding parameters are also different. Next, the differences among the parameters of the different layers are constrained to enable the parameters to be related. The constraints for soft parameter sharing in independent networks are obviously inspired by MTL regularization applied in the different fields. In this work, ATDP leverages hard parameter sharing, and $\mathcal{L}_{MB}$ takes advantage of soft parameter sharing.

Auxiliary learning [33], [34], or learning with auxiliary tasks, uses the beneficial information contained in related tasks to provide a stronger inductive bias for the learning of the target task. Our framework leverages the auxiliary datasets to improve the target detector via a weakly supervised way.

III. Our Method

In this section, we firstly give a brief overviewIII-A of our *Alternating Training Framework* (ATF). Next, we discuss the novel designs in detail, which consists of alternating training with decreasing proportions (Section III-B), mixed branch loss (Section III-C) and extending to common situations (Section III-D).

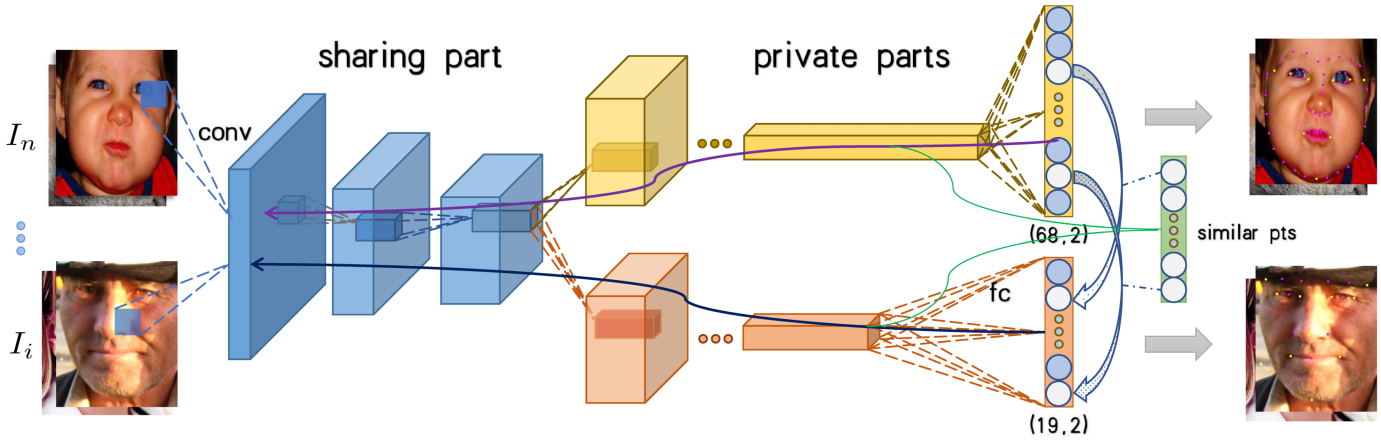


Fig. 2: The overall pipeline of *Alternating training framework* (ATF). The whole framework contains two parts: the sharing part and the private parts. In the private parts, each branch is supervised by the corresponding dataset. Furthermore, we add supervision on similar pts from different branches for different datasets, present in the green box.

A. Overview

Suppose we have datasets $D_1, D_2, \ldots, D_n$ for corresponding detectors $M_1, M_2, \ldots, M_n$. For simplicity, the all detectors are denoted as $\mathbf{M} = \{M_i\}_{i=1}^n$, datasets denoted as $\mathbf{D} = \{D_i\}_{i=1}^n$. For $\mathbf{D}$, the scheme of D_i defines a separate type of keypoint, denoted as $S_i \in \mathbf{R}^{p_i \times 2}$, where p_i is number of predefined keypoints for the scheme of D_i. Take the WFLW[4] benchmark as example, p_i equals 98 and one point consists of two-dimension coordinates. And samely we define n kinds of keypoints $(S_1, S_2, \ldots, S_n)$ which are corresponded to the $\mathbf{D}$ respectively. Also, we use $\mathbf{S} = \{S_i\}_{i=1}^n$ to represent the set of all keypoints.

In this framework, our goals are considered in three scenarios. The first and default purpose is to improve the performance of the target detector without additional cost. ATF takes advantage of the union of multiple datasets instead of the target dataset via a weakly supervised method. Suppose the target detector is M_i, ATF firstly splits M_i into two submodules: sharing part and private part, denoted as $Part_s$ and $Part_p$ respectively. As shown in Fig 2, ATF next generates other private parts from corresponding M_j, and all detectors share the same $Part_s$, where j is the index of auxiliary tasks. In other words, ATF achieves a great model including multiple detectors via the multi-task learning method. In particular, all detectors use the same sharing part $Part_s$ and possess their independent private parts $Part_p$. We take $\{Part_{p_i}\}_{i=1}^n$ to represent these private parts, and $Part_{p_i}$ is the corresponding i-th private part.

In the training period, ATF takes iterations alternately from random datasets as inputs, which enables $Part_s$ to implement hard parameter sharing among these detectors. Besides, ATF can easily set the master-servant relationship by adjusting the ratios of iterations in the single epoch. However, the convergence speed and the ideal accuracy of each detector are different because of the differences in markup variance, dataset capacity, and variation distribution. During the later training, parameters of the network constantly fluctuate if the next random batch comes from a different dataset. To fix the problem, we propose a simple but effective *Decreasing Proportions* in each epoch, which enables the frequency of auxiliary iteration to decrease in the next epoch. In other words, the special ability on the target benchmark is gradually improved in the next epoch. Furthermore, inspired by the phenomenon that there always exist some points representing the same meaning, we further propose $\mathcal{L}_{MB}$ to constrain the parameters of different $Part_p$. We use S_c to represent these same or similar points, which can bring reference to the constraints of other branches. $\mathcal{L}_{MB}$ enables the target branch to learn other features, which cannot be learned from the target dataset. Besides the focus on the first scenario, we also extend ATF to two practical situations: joint detector and novel detector. In both scenarios, we add special optimizations for better performance.

B. Alternating Training with Decreasing Proportions

To support the different labels, we should generate the networks consists of the sharing part $Part_s$ and multiple private parts $\{Part_{p_i}\}_{i=1}^n$. If ATF is in the first situation, the network structure becomes the same as before when pruning its extra $Part_p$. In other words, after pruning the needless private parts, the latency and computation keep the same during the inference period.

The network deformation first splits the original model into two sub-modules: sharing part $Part_s$ and private part $Part_p$. Next ATF produces extra $Part_{p_1}, Part_{p_2}, \ldots, Part_{p_{n-1}}$ corresponded to extra datasets $D_1, D_2, \ldots, D_{n-1}$ (Suppose the original $Part_p$ is $Part_{p_n}$). As shown in Fig. 2, these private parts $\{Part_{p_i}\}_{i=1}^n$ are connected by a parallel way, and all detector are shared with the parameters of $Part_s$. The mainstream methods of face alignment have been classified into two subcategories: direct coordinate regression and heatmap-based regression. The heatmap-based network deformation is showed in Fig.3, and coordinate network deformation in Fig.4. They both add the additional regression head corresponding with the original one.

In this work, we take an ordinary convolutional network (*OCN*) which consists of mobilenet-v2 [27] blocks for direct regression, the state-of-the-art network HRNET [20] for heatmap regression. To keep consistency with the original network structure, we set their regression head as the private part and backbone as the sharing part. In particular, the regression head of HRNet contains two 1x1 convolutional layers and OCN uses only a single fully-connected layer. It should be noted that sharing part is also applicable to other structures. To verify the effectiveness of ATF, we consider the different settings including network depth, input resolution, loss function, and computation. OCN couples with L1 loss as well as 112x112 input and has 1.228M parameters as well as 279.96M FLOPs. HRNet is L2 loss, 256x256 input, 9.663M parameters and 4.734G FLOPs.

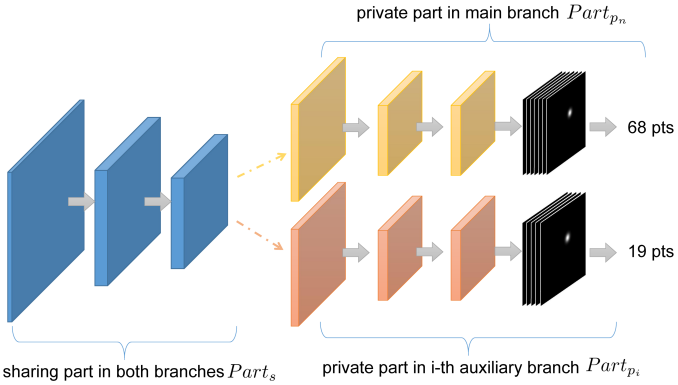


Fig. 3: The structure of the heatmap-based network with network deformation.

During the training process, ATF utilizes random iterations from different datasets to enable the $Part_s$ to share the parameters, which implements the hard parameter sharing. We also exploit decreasing proportion to enhance the ability to fit the target task. In the first situation, our evaluation is on only target benchmark even though training on all datasets $\{D_i\}_{i=1}^{n}$. Thus, we take validation data or test data of the target dataset as test, which is the same as common training. For the first scenario, the goal of ATF is to minimize the normalized mean error between the ground-truth points and predictions as below:

$$\arg\min_{\{\omega_s,\omega_m\}} \sum_{i=1}^{N_t} ||S_{n_i} - \Phi_m(\Phi_s(I_{n_i},\omega_s),\omega_m)|| \qquad (1)$$

where N_t represents the number of images in testset, and S_{n_i}, I_{n_i} are corresponding landmarks and input image for D_n. Φ_s represents the network structure of $Part_s$, and Φ_m is corresponded to the network structure of $Part_n$ (suppose $Part_n$ is corresponded to the target detector M_n). $\Phi_m(\Phi_s(I_{n_i},\omega_s),\omega_m)$ represents the feature vector of one sample $I_{n_i} \in D_n$. We use ω_s to represent the parameters of $Part_s$, and ω_m is the parameters of $Part_m$. When only inference in the target detector, we may prune other private parts $\{Part_{p_i}\}_{i=1}^{n-1}$ for minimizing latency.

During the training process, suppose I_{ij} denotes the j-th iteration whose batch data is from D_i. Alternating training leverages sufficient variations of facial features from multiple datasets. Suppose $r_0, r_1, \ldots, r_n$ is the proportion in one epoch corresponded to relative dataset. We adjust the ratios decreasing to enhance the fitting ability of the target branch $Part_{p_i}$. The ratios vary as follows:

$$(r_0, r_1, \ldots, r_n) = (\alpha, \beta, \ldots, 1)^{ce} \times (r_0, r_1, \ldots, r_n) \qquad (2)$$

Where α and β are less than 1 and represent the coefficient of descent, considered by dataset capacity and markup variance. ce is the number of the current epoch, and ratios decline in every epoch.

The following pseudo-code shows the training process of the whole method.

Algorithm 1 The training pipeline of our proposed framework

Require: A network with sharing part Φ_s and multiple branches Φ_m, Φ_a, different kinds of iterations I_m, I_a corresponded to datasets.

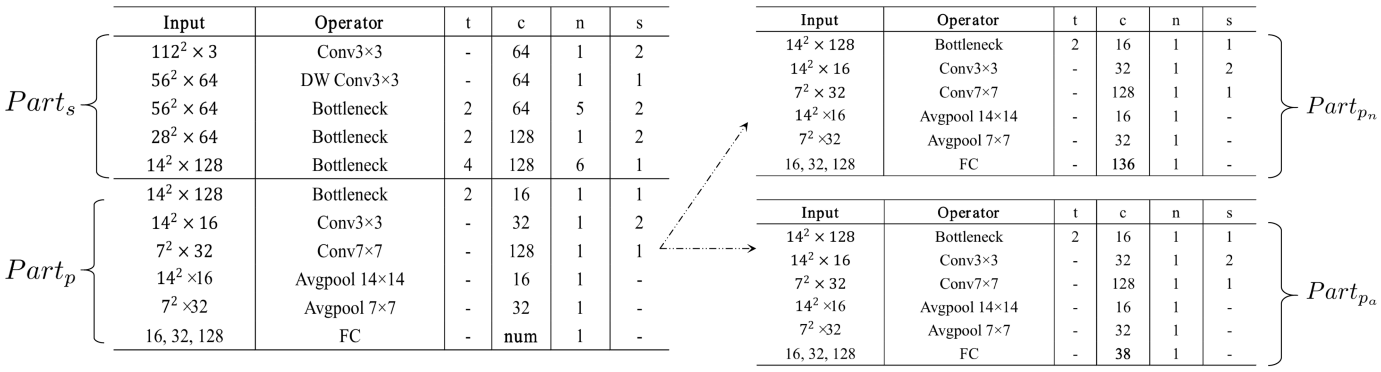
Set initial ratios of iterations in $Random$
for epoch = 1,2,..., End **do**
  while current iteration not End **do**
    $I = next(Random(I_m, I_a))$
    Forward sharing part $F_s = \Phi_s(I, \omega_s)$
    Forward branches $P_m = \Phi_s(F_s, \omega_m), P_a = \Phi_a(F_s, \omega_a)$
    Optimize $\omega_s, \omega_a, \omega_m$ by $\mathcal{L}_{MB}$ defined in Eq.3
  end while
  Test goal for validation according to Eq.1
  Change random ratios of iterations via Eq.2
end for

In the training process, suppose we have five random iterations $(a_1, b_1, a_2, c_1, b_2)$ are from different datasets (A, B, C) where a_1, a_2 are corresponded to A. When training on b_1, the model leverages the checkpoint from training on a_1 as pretraining and serves the pretraining checkpoint for the next iteration a_2. ATDP makes the networks shared their parameters via randomly finetuning on different datasets.

C. Mixed Branch Loss

However, the regression parts are still independently learning their own inductive bias. To address the issue, we are inspired by the following phenomenon.

Despite the gap that exists in the markup protocol, annotation variance, and capacity among the labels of landmark datasets, some keypoints still are well labeled with the same geometric meaning across datasets, *i.e.* eye corners, nose tip. Although they are a little different in the annotation schemes, their distributions of location are similar. As shown in figure 1, similar landmarks can usually be found among the datasets though markup protocols are different. There are always 8 to 16 similar landmarks annotated on a pair of datasets (300W[2], WFLW[4], COFW[3], AFLW[1]). These keypoint pairs provide a chance to transfer information from one dataset to the other detector.



$Part_s$

Input	Operator	t	c	n	s
$112^2 \times 3$	Conv3×3	-	64	1	2
$56^2 \times 64$	DW Conv3×3	-	64	1	1
$56^2 \times 64$	Bottleneck	2	64	5	2
$28^2 \times 64$	Bottleneck	2	128	1	2
$14^2 \times 128$	Bottleneck	4	128	6	1

$Part_p$

Input	Operator	t	c	n	s
$14^2 \times 128$	Bottleneck	2	16	1	1
$14^2 \times 16$	Conv3×3	-	32	1	2
$7^2 \times 32$	Conv7×7	-	128	1	1
14^2 ×16	Avgpool 14×14	-	16	1	-
7^2 ×32	Avgpool 7×7	-	32	1	-
16, 32, 128	FC	-	num	1	-

$Part_{p_n}$

Input	Operator	t	c	n	s
$14^2 \times 128$	Bottleneck	2	16	1	1
$14^2 \times 16$	Conv3×3	-	32	1	2
$7^2 \times 32$	Conv7×7	-	128	1	1
14^2 ×16	Avgpool 14×14	-	16	1	-
7^2 ×32	Avgpool 7×7	-	32	1	-
16, 32, 128	FC	-	136	1	-

$Part_{p_a}$

Input	Operator	t	c	n	s
$14^2 \times 128$	Bottleneck	2	16	1	1
$14^2 \times 16$	Conv3×3	-	32	1	2
$7^2 \times 32$	Conv7×7	-	128	1	1
14^2 ×16	Avgpool 14×14	-	16	1	-
7^2 ×32	Avgpool 7×7	-	32	1	-
16, 32, 128	FC	-	38	1	-

Fig. 4: The structure of the direct-regression network with network deformation.

Our proposed loss function ($\mathcal{L}_{MB}$) takes advantage of the above information of these similar landmarks to constrain irrelevant branches, especially private parts $\{Part_p\}_{i=1}^{n}$. In addition to the native loss, $\mathcal{L}_{MB}$ also constrain the difference between a pair of similar landmarks predicted by different branches. It can be formulated as follows:

$$\mathcal{L}_{MB} = \begin{cases} \mathcal{L}_m + \alpha \times \mathcal{L}_{ps} & I_i \in D_m \\ \beta \times \mathcal{L}_a + \gamma \times \mathcal{L}_{ps} & \text{otherwise} \end{cases} \tag{3}$$

where I_i is the random iteration from a dataset in the current stage, and the proportionality in an epoch corresponds to equation 2. We use $\mathcal{L}_m$ to denote the distance between the ground-truth coordinate S_n from target dataset D_n and prediction from target branch $Part_s + Part_{p_n}$, $\mathcal{L}_a$ is denoted as the error between the ground-truth coordinate $\{S_i\}_{i=1}^{n-1}$ from auxiliary datasets $\{D_i\}_{i=1}^{n-1}$ and prediction from auxiliary branches $Part_s + \{Part_i\}_{i=1}^{n-1}$, and $\mathcal{L}_{ps}$ represents the error between the similar landmarks which are predicted by corresponding $Part_p$. The weight hyperparameters α, β, γ depend on database capacity, markup protocol, and annotation scheme.

Following the previous works[17], [20], [35], the optimization methods in heatmap-based and direct coordinate are different. The heatmap-based method needs to calculate the error of each heatmap. Taking the most common second normal form in heatmap[19], [20], the detailed representation of $\mathcal{L}_{MB}$ in HRNet with multiple parts (HRNETMP) is as follows:

$$\begin{cases} \mathcal{L}_m^H = \dfrac{1}{N_I^m \cdot N_L^m} \sum\limits_{i}^{N_I^m} \sum\limits_{j}^{N_L^m} \sum\limits_{k}^{H} \sum\limits_{l}^{W} ||P_{(i,j,k,l)}^m - V_{(i,j,k,l)}^{gt_m}||_2^2 \\ \mathcal{L}_a^H = \dfrac{1}{N_I^a \cdot N_L^a} \sum\limits_{i}^{N_I^a} \sum\limits_{j}^{N_L^a} \sum\limits_{k}^{H} \sum\limits_{l}^{W} ||P_{(i,j,k,l)}^a - V_{(i,j,k,l)}^{gt_a}||_2^2 \\ \mathcal{L}_{ps}^H = \dfrac{1}{N_I^{m/a} \cdot N_L^s} \sum\limits_{i}^{N_I^{m/a}} \sum\limits_{j}^{N_L^s} \sum\limits_{k}^{H} \sum\limits_{l}^{W} ||P_{(i,j,k,l)}^{sm} - P_{(i,j,k,l)}^{sa}||_2^2 \end{cases} \tag{4}$$

where N_I^m and N_I^a represent the capacities of D_n and $\{D_i\}_{i=1}^{n-1}$, N_L^m, N_L^a and N_L^s denote the number of landmarks in S_n , $\{S_i\}_{i=1}^{n-1}$ and their similar pairs. H, W are the height and width of heatmap image, P^m, P^a represent the landmark value predicted by target branch and auxiliary branch. V^{gt_i} is the ground truth value in S_i. $N_I^{m/a}$ is N_I^m or N_I^a, which is depend on the current iteration. P^{sm} and P^{sa} represent the value in the set of similar pairs, $P^{sm} \in P^m$ and $P^{sa} \in P^a$.

For direct coordinate regression, the calculation of error is the distance between relative landmarks, And Feng *et al.* [17] proposes 1st NF are better suitable for direct regression than 2nd NF, the formula of $\mathcal{L}_{MB}$ in OCN with multiple parts (OCNMP) is as follows:

$$\begin{cases} \mathcal{L}_m^D = \dfrac{1}{N_I^m \cdot N_L^m} \sum\limits_{i}^{N_I^m} \sum\limits_{j}^{N_L^m} ||(P_{(i,j)}^{m_x}, P_{(i,j)}^{m_y}) - (V_{(i,j)}^{m_x}, V_{(i,j)}^{m_y})||_1 \\ \mathcal{L}_a^D = \dfrac{1}{N_I^a \cdot N_L^a} \sum\limits_{i}^{N_I^a} \sum\limits_{j}^{N_L^a} ||(P_{(i,j)}^{a_x}, P_{(i,j)}^{a_y}) - (V_{(i,j)}^{a_x}, V_{(i,j)}^{a_y})||_1 \\ \mathcal{L}_{ps}^D = \dfrac{1}{N_I^{m/a} \cdot N_L^s} \sum\limits_{i}^{N_I^{m/a}} \sum\limits_{j}^{N_L^s} ||(P_{(i,j)}^{sm_x}, P_{(i,j)}^{sm_y}) - (P_{(i,j)}^{sa_x}, P_{(i,j)}^{sa_y})||_1 \end{cases} \tag{5}$$

where $N_I^m, N_I^a, N_L^m, N_L^a, N_I^{m/a}$ and N_L^s are the same definition as before. P^{sm} and p^{sa} are similar with the above definition, which are changed from a pixel value to a couple coordinate. $P_{(i,j)}^{m_x}, P_{(i,j)}^{m_y}$ are the x- and y-coordinate of j-th in i-th image, which is predicted by main branch P^m. $P_{(i,j)}^{a_x}, P_{(i,j)}^{a_y}$ are the x- and y-coordinate of j-th in i-th image, which is predicted by auxiliary branch P^a. $(V_{(i,j)}^{m_x}, V_{(i,j)}^{m_y})$ is the ground truth coordinate of j-th in i-th image of D_n. $(V_{(i,j)}^{a_x}, V_{(i,j)}^{a_y})$ is the ground truth coordinate of j-th in i-th image of the auxiliary dataset.

Although we conduct experiments with 1st NF and 2nd NF loss functions, it should be noted that our method is also applicable to other common loss functions as the original one.

D. Extending to Common Scenarios

In practical applications, it will cost more computations and GPU memory if we use ATF directly. Thus, we have done

some optimizations for ATF to extend to three situations: when the target detector is enough, when a joint detector including all source detectors is needed, when a new novel detector is coming.

For the situation where only the target detector is needed, auxiliary detectors are useless and consume resources in the test stage. We keep the sharing part $Part_s$ and the target branch $Part_{p_n}$ unchanged, pruning all auxiliary branches $Part_{p_1}, Part_{p_2}, \ldots, Part_{p_{n-1}}$, so that the optimized networks structure is consistent with the original model.

Due to the different annotation schemes, there always exists the meet that a joint detector including more than one even all detectors aims to get as many points as possible. Our framework uses only one shared backbone, and multiple regression branches, which greatly reduces the amount of calculation and GPU memory compared with other methods [21], [24]. Compared with the original detector, ATF adds a little computing resource and realizes multiple detectors. However, the immediate use of ATF will leave a gap in its peak performance on the auxiliary datasets. To bridge this gap, it is recommended to finetune the network on its data independently. Moreover, limited by the resources and training time, we freeze the parameters of the sharing part and only finetune the parameters of the private branches with each dataset. The results are shown in the subsection.IV-D, which demonstrate that our backbone is robust enough for the rarely or unseen variations.

It is common to construct an accurate detector on a new facial landmark dataset, and the novel landmark is with different schemes. Intuitively, restarting training from zero to be finished is time-consuming. We propose a simple and fast method to obtain an accurate enough detector via our framework. Add another private part to the original ATF, and this branch should correspond to the new labels. Freeze the parameters of the sharing backbone, and finetune the new regression branch. Comparing with re-training, ATF takes less than 1/10 of the time and obtains comparable or even higher accuracy. Although we only finetune on regressional parts, it should be noted that finetuning the entire network will achieve higher performance.

IV. Experiments

A. Experimental Settings

Datasets We conduct extensive experiments on widely used benchmarks [2], [1], [4], [3]. These datasets are challenges because facial images focus on variations in makeup, extreme pose, and wide occlusion.

300W[2] data set offers 68 keypoints for each 2D face, which are collected from AFW[48], HELEN[49], LFPW[50], XM2VTS[51], and IBUG. According to the protocol used in [48], all 3148 training images are from the training set of LFPW and HELEN and the full set of AFW. The 689 test images come from the test collection of LFPW and HELEN, as well as the full set of IBUG. The test images are further divided into three subsets: 554 samples from LFPW and HELEN as the common subset, the 135-image IBUG as the challenging subset, and their union as the full set.

WFLW[4] dataset is the most challenging benchmark, which contains 7,500 face images for training and 2,500 images for the test. And WFLW is the newest benchmark based on WIDER Face[52] with 98 manually annotated keypoints. The faces images introduce many types of variations concerning large pose, expression, occlusion, etc. The test set is further divided into six subsets for comprehensive evaluation: pose (326 images), expression (314 images), illumination (698 images), make-up (206 images), occlusion (736 images), and blur (773 images).

COFW[3] dataset contains 1345 images for training and 507 images for the test, and each face image has been annotated with 29 landmarks. In this dataset, the face data focus on large variations and occlusions.

AFLW[1] dataset is annotated with 19 landmarks, which contains 24386 wild faces with large head pose up to 120° for yaw and 90° for pitch and roll. According to the previous protocol [36], [53], the AFLW-full uses 20000 images for training and 4386 images for the test.

Baseline networks We take OCN and HRNet as baselines for direct coordinates regression and heatmap-based regression. OCN exploits the bottleneck [27] like [35] to construct the network. It uses the $112 \times 112 \times 3$ as the input resolution and coupled with $L1$ loss following previous practice [17], [35]. HRNET uses the backbone of HR-18[20] and take the $256 \times 256 \times 3$ as input shape, which combine $L2$ loss for optimization. We transform the baseline networks to get new networks with multiple private parts, noted as OCNMP and HRNETMP. For simplicity, we add the first letter of dataset name behind the network name, denoted as multiple $Part_p$. For example, $OCNMP_{W3}$ means OCNMP has two private parts, one corresponding to WFLW[4] and the other corresponding to 300W[2].

Evaluation Metrics Following most previous works[20], [54], [4], [18], we take normalized mean error (NME), area under the curve (AUC) and failure rate (FR) as metrics to comprehensive verify our method.

Normalized Mean Error (NME) is the most authoritative and used for face alignment. We take the outer-eye-corner distance as the normalizing factor for WFLW[4], COFW[3], and 300W[2], and the face size as the normalizing factor for AFLW because of various profile faces[1]. NME is defined as:

$$NME(\%) = \frac{1}{N} \sum_{k=1}^{N} \frac{||P_k - \hat{P}_k||_2}{d} \times 100 \tag{6}$$

where the normalized factor d represents the distance between the outer corners of the left and right eyes. P_k and $\hat{P}_k$ represent the ground-truth coordinate and prediction of the k-th keypoint. For simplicity, we magnify 100 times to omit the $\%$ symbol. Lower NME is for better performance.

Area Under the Curve (AUC) is calculated based on the cumulative error distribution (CED) curve. The AUC for a test set is computed as the area under the curve, up to the cut-off NME value. Higher AUC represents a better detector. To compare with previous methods, the AUC and FR are used on WFLW fullset and all subsets.

Metric	Method	Fullset	Pose	Exp.	Ill.	Mu.	Occ.	Blur
NME(↓)	SDM [14]	10.29	24.10	11.45	9.32	9.38	13.03	11.28
	CFSS [36]	9.07	21.36	10.09	8.30	8.74	11.76	9.96
	DVLN [22]	6.08	11.54	6.27	5.73	5.98	7.33	6.88
	LAB [4]	5.27	10.24	5.51	5.23	5.15	6.79	6.32
	PDB [17]	5.11	8.75	5.36	4.93	5.41	6.37	5.81
	DCFE [37]	4.69	8.63	6.27	5.73	5.98	7.33	6.88
	DeCaFA [38]	4.62	8.11	4.65	4.41	4.63	5.74	5.38
	AS w. SAN [39]	**4.39**	8.42	4.68	**4.24**	4.37	5.60	**4.86**
	OCN	4.93	8.94	5.05	5.13	5.00	5.95	5.62
	OCNMP	4.60	8.05	4.69	4.66	4.60	5.54	5.20
	HRNET [20]	4.60	7.94	4.85	4.55	4.29	5.44	5.42
	HRNETMP	4.50	**7.54**	**4.65**	4.45	**4.20**	**5.30**	5.19
FR_{10}(↓)	CFSS [36]	20.56	66.26	23.25	17.34	21.84	32.88	23.67
	DVLN [22]	10.84	46.93	11.15	7.31	11.65	16.30	13.71
	LAB [4]	7.56	28.83	6.37	6.73	7.77	13.72	10.74
	Wing [17]	6.00	22.70	4.78	4.30	7.77	12.50	7.76
	DeCaFA [38]	4.84	21.40	3.73	3.22	6.15	9.26	6.61
	AS w. SAN [39]	4.08	18.10	4.46	2.72	4.37	7.74	4.40
	OCN	5.92	27.00	5.10	5.16	6.80	10.46	5.69
	OCNMP	4.52	20.15	3.18	3.58	5.83	8.42	4.53
	HRNET	3.12	16.26	2.55	2.72	1.94	5.71	4.53
	HRNETMP	**2.52**	**13.19**	**2.23**	**2.44**	**0.49**	**5.03**	**3.88**
AUC_{10}(↑)	CFSS [36]	0.366	0.063	0.316	0.385	0.369	0.269	0.303
	DVLN [22]	0.456	0.147	0.389	0.474	0.449	0.379	0.397
	LAB [4]	0.532	0.235	0.495	0.543	0.539	0.449	0.463
	Wing [17]	0.554	0.310	0.496	0.541	0.558	0.489	0.492
	DeCaFA [38]	0.563	0.292	0.546	0.579	0.575	0.485	0.494
	AS w. SAN [39]	**0.591**	**0.311**	**0.549**	**0.609**	**0.581**	**0.516**	**0.551**
	OCN	0.5388	0.2407	0.5121	0.5544	0.5242	0.4604	0.4885
	OCNMP	0.5615	0.2897	0.5339	0.5753	0.5449	0.4853	0.5156
	HRNET	0.5536	0.2867	0.5360	0.5623	0.5770	0.4965	0.4969
	HRNETMP	0.5604	0.3010	0.5460	0.5663	**0.5806**	0.4874	0.4893

TABLE I: Comparisons in inter-ocular normalized mean error (NME), area under curve (AUC) and failure rate (FR) on the WFLW (fullset and all subsets). Note HRNet's paper does not offer the information about AUC and FR, and we take its checkpoint from the offical code.

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 prefixed threshold. And we also omit the % symbol to simplify variables. The Lower FR is corresponding to better performance.

Implementation Details When conduct experiments on OCN and OCNMP, we crop and resize all image data of the target detector to $112 \times 112 \times 3$ resolution via the facial boxes from retina-face detector[55]. We set the initial learning rate to 1e-3 with Adam optimization. The data augmentation contains random crop(0.2), horizontal flipped(0.5), and rotation($\pm$30). The learning rate will decrease by 0.1 times if the loss of validation stops decrease in 10 epochs. For a fair comparison, the experiment settings on HRNETMP are the same as HRNet[20]. In the experiments concerning alternating training, we also adjust iteration ratios for one epoch. We set the initial proportion to 2:1 or 3:1 and set ratios from 0.8 to 0.95, depending on the dataset capacity and markup schemes. And loss weights α, β, γ are decided by the annotation protocol and dataset capacity, which also exist a little distinguish in each pair. Especially, when leveraging AFLW as auxiliary and WFLW as main, we set αequals$0.2, \beta$equals$0.8, \gamma$equals0.1.

B. Comparison with State-of-the-art Methods

We firstly conduct experiments with OCN on each benchmark to obtain the baseline result. Next, we use the union database of multiple datasets to train OCNMP and HRNETMP, which take OCN and HRNET[20]'s results as baselines. To verify the effectiveness, we further compare OCNMP, HRNETMP with the state-of-the-art works on each benchmark. Note OCN again is a slight network with 1.228M parameter and 279.96M FLOPs, and the model size of HRNET is 9.663M parameters and 4.734G FLOPs. As shown in Tab.I, the result about WFLW fullset and subsets are impressive. Note that the FR and AUC of HRNet are evaluated from the official checkpoint. The Tab. II shows the normalized mean error on 300W, AFLW, and COFW, including their fullsets and subsets. At all fullset and subsets, ATF improves the performance of the detector in each network.

Compared with HRNet, HRNETMP leveraging ATF has obtained obvious improvement in various benchmarks. And on experiments on 300W, COFW, and AFLW, HRNETMP outperforms other SOTA methods by a lot. Especially, the NME of ATF reaches 3.17, 2.75, and 4.89 for fullset and subsets of 300W[2], 3.32 for COFW fullset, 1.55 for AFLW. Compared with the HRNet baseline, it achieves a huge enhancement on 300W, 4.5%, 4.1%, and 5.04% relative ascension. Moreover, we extend three metrics on WFLW, HRNETMP

Method	300W Full	300W Com.	300W Cha.	COFW Full	AFLW Full
RAR [40]	-	-	-	6.03	-
RCN [41]	5.41	4.67	8.44	-	5.6
DAN [19]	3.59	3.19	5.24	-	-
DAC-CSR [42]	-	-	-	6.03	2.27
SAN [18]	3.98	3.34	6.60	-	1.91
RCN+ [43]	4.90	4.20	7.78	-	1.61
PDB [17]	3.60	3.01	6.01	-	**1.47**
LAB [4]	3.49	2.98	5.19	5.58	1.85
DCFE [37]	3.24	2.76	5.22	-	2.17
TS^3 [44]	3.78	3.17	6.41	-	1.99
LaplaceKL [45]	4.01	3.28	7.01	-	1.97
ODN [46]	4.17	3.56	6.67	5.3	1.63
AS w. SAN [39]	3.86	3.21	6.49	-	-
HG-HSLE [47]	3.28	2.85	5.03	-	-
OCN	3.87	3.36	5.84	4.18	2.00
OCNMP	3.55	3.12	5.32	3.91	1.91
HRNET [20]	3.32	2.87	5.15	3.45	1.57
HRNETMP	**3.17**	**2.75**	**4.89**	**3.32**	1.55

TABLE II: Comparison in inter-ocular normalized mean error (NME) on the 300W (common subset, Challenging subset, and fullset), COFW and AFLW.

always performs best at FR and NME, particularly 2.52 for failure rate. The experiments on WFLW fullset and all subsets demonstrate that our method is useful for all benchmarks. This phenomenon is noticeable on the challenge dataset. Especially, the large-pose subset includes data with relatively large Euler angles and severe self-occlusion. The ascension on this subset reaches 5.41%, which is higher than other improvements. And experiments on 300W, the enhancement of the challenge subset (5.04%) is more significant than the common subset (4.1%). The above phenomenon demonstrates that our framework makes it possible to learn the data variations which is rarely or unseen in the original dataset.

On the experiments with the slight network, OCNMP obtains bigger relative improvement than HRNETMP. Leveraging ATF again makes obvious improvement on each benchmark than original training. As shown in Tab.4, OCNMP reaches 4.60 on the WFLW fullset and 3.55 on 300W, which is comparable to the SOTA result. Thus, OCNMP provides a great trade-off between accuracy and computation. Comparing with the OCN baseline, OCNMP achieves 4.5% relative ascension for AFLW, 8.24% for 300W, 6.81% for WFLW, and COFW 6.55%. Moreover, ATF works well on not only the fullset but also all subsets, 7.14% for 300W Common, 8.90% for 300W Challenge, 7.12% for WFLW expression etc. In particular, the enhancements of the challenge subsets are higher, 8.9% for cha subset and 9.96% for pose subset. Again, ATF achieves obvious enhancement comparing with original training, and the performance outperforms SOTA by a large margin.

We also have visualized some cases on WFLW as shown in Fig.5. The pictures from left to right are raw image (input image), HRNet[20] output, ATF output (ours) and ground-truth.

C. Evaluation ATF on Cross Datasets

We conduct experiments on the cross dataset to verify the feasibility of the framework. To demonstrate the robustness of each benchmark, we also experiment on each pair of datasets with OCNMP and HRNETMP.

We firstly experiment on WFLW fullset and all subsets, which the auxiliary dataset is from 300W, COFW and AFLW independently. The result is showed in Tab.III, with AFLW assistant, both HRNetMP and OCNMP achieve the best result, 4.60 for ONCMP and 4.50 for HRNETMP. For convenience in network abbreviation, we take C to denote COFW, W for WFLW, A for AFLW, and 3 for 300W. Specifically, ATF_{WC} denotes COFW assist the target dataset WFLW. It also obtains significant ascension on all subsets of WFLW, HRNETMP reaches 5.29 for occlusion, OCNMP reaches 4.69 for expression. Furthermore, makeup and pose subsets are improved, which means the auxiliary dataset contains additional representation focus on pose and makeup. Although 300W and COFW are with smaller capacities, ATF still improves the performance by leveraging them, which also demonstrates the effectiveness of our framework.

HRNET/OCN	test
baselines	3.45/4.18
ATF_{C3}	3.42/3.97
ATF_{CA}	3.36/3.95
ATF_{CW}	3.32/3.91

TABLE V: Comparison in NME of networks on COFW

HRNET/OCN	test
baselines	1.57/1.99
ATF_{A3}	1.56/1.91
ATF_{AC}	1.56/1.92
ATF_{AW}	1.55/1.91

TABLE VI: Comparison in NME of networks on AFLW

Verified on 300W[2] benchmark, leveraging WFLW achieves the most significant improvement among datasets in both HRNETMP and OCNMP, which is shown in Tab. IV. And the NME reaches to the **3.17** on fullset, 2.75 on common set, and 4.89 on challenge set, which outperforms SOTA a lot. Also, utilizing COFW or AFLW to assist 300W independently performs better than baseline, and ATF again demonstrates its

Network	Method	test	blur	exp.	ill.	pose	mu.	occ.
OCN	baseline	4.93	5.62	5.05	5.13	8.94	5.00	5.95
	ATF_{WC}	4.64	5.26	4.66	4.79	8.40	4.79	5.59
	ATF_{W3}	4.72	5.35	4.86	4.88	8.62	4.86	5.69
	ATF_{WA}	4.60	5.20	4.69	4.66	8.05	4.60	5.54
HRNET	baseline	4.60	5.42	4.85	4.55	7.94	4.29	5.44
	ATF_{WC}	4.54	5.30	4.74	4.50	7.56	4.25	5.32
	ATF_{W3}	4.52	5.28	4.63	4.44	7.53	4.24	5.29
	ATF_{WA}	4.50	5.19	4.65	4.45	7.54	4.20	5.30

TABLE III: Comparison in NME of HRNET with variants and OCN with variants on WFLW benchmark (fullset, and all subsets). ATF_{WC} denotes COFW assist the target dataset WFLW.

Network	Method	full	com.	cha.	test
OCN	baseline	3.87	3.39	5.85	4.52
	ATF_{3C}	3.69	3.24	5.53	4.33
	ATF_{3A}	3.56	3.11	5.39	4.19
	ATF_{3W}	3.55	3.12	5.32	4.15
HRNET	baseline	3.32	2.87	5.15	3.85
	ATF_{3C}	3.28	2.85	5.08	3.87
	ATF_{3A}	3.24	2.82	4.98	3.77
	ATF_{3W}	3.17	2.75	4.89	3.65

TABLE IV: Comparison in NME of networks on 300W benchmark (fullset, common subset and challenge subset).

robustness for each dataset. As shown in Tab.V, leveraging WFLW also achieves best among experiments on COFW. The Tab.VI reveals the result of leveraging auxiliary datasets for AFLW. The joint performance ($OCNMP_{WA}$, $OCNMP_{C3}$ *et al.*) utilizing multiple datasets outperforms 5% higher than baseline, even often 8% higher. And due to WFLW is the newest public dataset, exploiting WFLW to assist other detectors always obtains the best performance, which also shows that the performance of ATF is related to the markup protocol.

It is demonstrated that ATF is effective in leveraging a union of multiple datasets via the weakly supervised approach. The above results illustrate that the proposed ATF framework based on either Heatmap-based or direct coordinate regression can exploit sufficient variations in different datasets, which enhances the variation generalization and produces a highly robust model.

D. Extending to Common Scenarios

We have extended ATF to three common situations, and optimized the single detector aims to save GPU memory and cost. Due to the first situation is already introduced in the above contents, we detailed discuss the results of the late two scenarios: when the joint detector including all source detectors is needed, when a new novel detector is coming. We conduct the experiment on WFLW[4] leveraging COFW[3] and 300W[2] independently, which contains two pair settings. For the first setting, we use 300W[2] as the auxiliary dataset, and COFW[3] as the novel dataset, And for the second, COFW[3] is auxiliary, and 300W[2] is novel. The heatmap sigma of 300W[2] sets 1.0, others[4], [3] set 1.5 in both experiments.

The results are shown in Tab.VIII and Tab.IX. And novel dataset is never trained during previous training. The baselines are HRNet[20]'s results on each benchmark. Leveraging COFW as auxiliary improves detector from 4.60 to 4.563 on WFLW, and then finetune COFW on its independent data and regressional part, which leads the joint detector to reaches 3.26

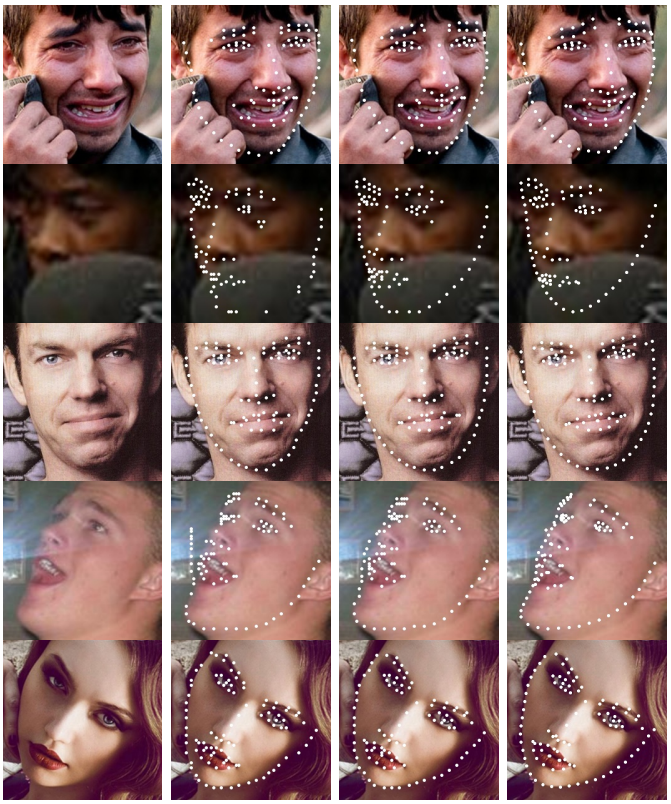


Fig. 5: Visualization comparison on WFLW. From left to right, they represent the input image, HRNET output, ATF output (ours), and ground-truth (reference).

Metric	Auxiliary	test	blur	exp.	ill.	pose	mu.	occ.
NME(↓)	COFW$_{wod}$	4.57	5.32	4.75	4.47	7.70	4.36	5.38
	COFW$_{wd}$	4.54	5.30	4.74	4.50	7.56	4.25	5.32
	300W$_{wod}$	4.56	5.29	4.72	4.52	7.67	4.22	5.41
	300W$_{wd}$	4.52	5.28	4.63	4.44	7.53	4.24	5.29
	AFLW$_{wod}$	4.53	5.27	4.74	4.44	7.65	4.31	5.32
	AFLW$_{wd}$	4.50	5.19	4.65	4.45	7.54	4.20	5.30
FR$_{10}$(↓)	COFW$_{wod}$	2.84	4.01	2.23	2.44	15.03	2.91	5.43
	COFW$_{wd}$	2.92	4.27	2.87	2.58	14.42	2.91	5.30
	300W$_{wod}$	3.00	4.27	2.87	2.72	15.64	0.49	5.84
	300W$_{wd}$	2.92	4.40	1.91	2.15	15.03	1.46	5.57
	AFLW$_{wod}$	2.92	4.40	2.55	2.72	15.95	1.94	5.43
	AFLW$_{wd}$	2.52	3.88	2.23	2.44	13.19	0.49	5.03
AUC$_{10}$(↑)	COFW$_{wod}$	0.5548	0.4890	0.5426	0.5657	0.2890	0.5677	0.4861
	COFW$_{wd}$	0.5565	0.4917	0.5396	0.5632	0.2998	0.5772	0.4917
	300W$_{wod}$	0.5566	0.4919	0.5416	0.5628	0.2963	0.5778	0.4927
	300W$_{wd}$	0.5574	0.4889	0.5454	0.5665	0.2989	0.5763	0.4921
	AFLW$_{wod}$	0.5576	0.4931	0.5439	0.5689	0.2924	0.5715	0.4913
	AFLW$_{wd}$	0.5604	0.4969	0.5460	0.5663	0.3010	0.5806	0.4965

TABLE VII: Comparison between with decreasing proportion and without decreasing proportion on WFLW (fullset and all subsets), assisted by 300W, COFW and AFLW independently.

Method	WFLW(main)	COFW(auxiliary)	300W(novel)
baseline[20]	4.60	3.45	3.32
ATF	4.563	3.260	3.407

TABLE VIII: Comparison in NME of WFLW, COFW and 300W with HRNet. WFLW is main, COFW is auxiliary and 300W is novel.

Method	WFLW(main)	300W(auxiliary)	COFW(novel)
baseline[20]	4.60	3.32	3.45
ATF	4.535	3.309	3.439

TABLE IX: Comparison in NME of WFLW, COFW and 300W with HRNet. WFLW is main, 300W is auxiliary and COFW is novel.

Pair / Method	3W	3A	3C	W3	WA	WC
baseline	3.32			4.60		
$\mathcal{L}_{AT}$	3.19	3.26	3.28	4.58	4.53	4.56
$\mathcal{L}_{MB}$	3.17	3.24	3.28	4.52	4.50	4.54

TABLE X: Comparison inter-ocular normalized mean error of different Loss function in various combination. left is based 300W, right is based WFLW

on COFW, better than SOTA. When finetune on the novel dataset 300W, it also can reach 3.407, comparable to SOTA. And for the second experiment, leveraging 300W as auxiliary improves detector from 4.60 to 4.535 on WFLW, and then finetune 300W on own data and branch, which leads the joint detector to reach 3.309 on 300W, better than SOTA. When finetune on the novel COFW, NME reaches 3.439, even better than SOTA.

Again, both experiments demonstrate that ATF outperforms better than SOTA not only on the main dataset but also on the auxiliary benchmark. It is amazing that leveraging 300W for WFLW and finetune a novel dataset COFW, which all get higher accuracy than SOTA on each benchmark. It is demonstrated again that ATF improves robustness for rarely or never appeared scenarios in the original dataset.

E. Ablation Study

ATF contains 2 pivotal components, ATDP and $\mathcal{L}_{MB}$. To emphasize the advantages of two sub-parts, we conduct comprehensive ablation studies on both 300W[2] and WFLW[4] benchmarks. We firstly compare ATDP coupled with $\mathcal{L}_{MB}$ with the single ATDP, which represents the random iteration only works in forward and backward of its independent private part. In other words, we conduct comparative experiments to evaluate the performance promoted or descend without utilizing similar landmark pairs. It also represents $\mathcal{L}_{MB}$ with $\alpha = 0$, $\beta = 1, \gamma = 0$ in eq.3. We use $\mathcal{L}_{AT}$ to denote the above loss. To verify the robustness, we leverage each dataset independently to assist the target task for extensive experiments. Take 300W as an example, we exploit the AFLW, COFW, and WFLW independently in carrying out experiments.

The result is shown in Tab.X, and the performance without considering private parts also outperforms better than baseline, which is demonstrated that the single ATDP is also beneficial to enhance the robustness. ATDP shows the advantages of learning a general detector with extra data variation which the original dataset rarely contains. Furthermore, $\mathcal{L}_{MB}$ works better in cross datasets than independent loss $\mathcal{L}_{AT}$ without considering private parts. Combined ATDP with $\mathcal{L}_{MB}$ can further enhance the performance. It is demonstrated that our $\mathcal{L}_{MB}$ is effective in utilizing constrained $\{Part_p\}_{i=1}^{n}$ via similar keypoint pairs.

Besides the $\mathcal{L}_{MB}$, we also conduct the ablation study on decreasing proportions to verify its effectiveness. We also have conducted the experiments on WFLW auxiliary with 300W, COFW, AFLW, and the results are showed in Tab.VII. The decreasing ratios set 0.97 and 0.95, and total epochs are the same as previous settings. We denote with decreasing proportion as wd, and without decreasing proportion as wod. For example, COFW$_{wod}$ means using COFW as an auxiliary

to assist WFLW, which is without decreasing proportion. According to the value of NME, FR, and AUC, the simple approach can effectively improve the performance on the target benchmark in the late period.

V. Conclusion

In this paper, we propose a novel framework ATF to address limitations in face alignment. ATF enable one target detector to learn other inductive bias that is rarely even impossible exist in the target dataset. ATF consists of two sub-modules, ATDP and $\mathcal{L}_{MB}$. Moreover, we also extend the framework to three common scenarios in practical application: single detector, joint detector, novel detector. As evaluated, ATF achieves relative performance ascension on each benchmark, up to 9.96%. We also have implemented the joint detector, which outperforms SOTA on not only main and auxiliary benchmarks but also the novel benchmark. In the future, we will extend ATF to other fields for higher performance and without extra calculation.

References

[1] 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 *2011 IEEE international conference on computer vision workshops (ICCV workshops)*. IEEE, 2011, pp. 2144–2151.

[2] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 397–403.

[3] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1513–1520.

[4] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2129–2138.

[5] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[6] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3d face reconstruction and dense alignment with position map regression network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 534–551.

[7] M. H. Khan, J. McDonagh, and G. Tzimiropoulos, "Synergy between face alignment and tracking via discriminative global consensus optimization," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 3811–3819.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[9] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[10] L. Deng, D. Yu *et al.*, "Deep learning: methods and applications," *Foundations and Trends® in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

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

[12] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[13] D. Cristinacce and T. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognition*, vol. 41, no. 10, pp. 3054–3067, 2008.

[14] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 532–539.

[15] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3476–3483.

[16] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *European conference on computer vision*. Springer, 2014, pp. 94–108.

[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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2235–2245.

[18] X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Style aggregated network for facial landmark detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 379–388.

[19] M. Kowalski, J. Naruniec, and T. 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*, 2017, pp. 88–97.

[20] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-resolution representations for labeling pixels and regions," *arXiv preprint arXiv:1904.04514*, 2019.

[21] B. M. Smith and L. Zhang, "Collaborative facial landmark localization for transferring annotations across datasets," in *European Conference on Computer Vision*. Springer, 2014, pp. 78–93.

[22] W. Wu and S. Yang, "Leveraging intra and inter-dataset variations for robust face alignment," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 150–159.

[23] J. Zhang, M. Kan, S. Shan, and X. Chen, "Leveraging datasets with varying annotations for face alignment via deep regression network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3801–3809.

[24] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Transferring landmark annotations for cross-dataset face alignment," *arXiv preprint arXiv:1409.0602*, 2014.

[25] X. Lan, Q. Hu, F. Xiong, C. Leng, and J. Cheng, "Atf: Towards robust face alignment via leveraging similarity and diversity across different datasets," *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[26] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware coordinate representation for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7093–7102.

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[28] A. Kumar, T. K. Marks, W. Mou, Y. Wang, M. Jones, A. Cherian, T. Koike-Akino, X. Liu, and C. Feng, "Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8236–8246.

[29] J. Yang, Q. Liu, and K. Zhang, "Stacked hourglass network for robust facial landmark localisation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 79–87.

[30] Y. Zhang and Q. Yang, "A survey on multi-task learning," *arXiv preprint arXiv:1707.08114*, 2017.

[31] R. Caruana, "Multitask learning: A knowledge-based source of inductive bias icml," *Google Scholar Google Scholar Digital Library Digital Library*, 1993.

[32] L. Duong, T. Cohn, S. Bird, and P. Cook, "Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2015, pp. 845–850.

[33] Y. Liu, B. Zhuang, C. Shen, H. Chen, and W. Yin, "Auxiliary learning for deep multi-task learning," 2019.

[34] L. Liebel and M. Körner, "Auxiliary tasks in multi-task learning," 2018.

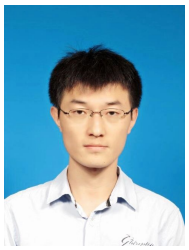
[35] X. Guo, S. Li, J. Yu, J. Zhang, J. Ma, L. Ma, W. Liu, and H. Ling, "Pfld: a practical facial landmark detector," *arXiv preprint arXiv:1902.10859*, 2019.

[36] S. Zhu, C. Li, C. Change Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4998–5006.

[37] R. Valle, J. M. Buenaposada, A. Valdes, and L. Baumela, "A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 585–601.

[38] A. Dapogny, K. Bailly, and M. Cord, "Decafa: Deep convolutional cascade for face alignment in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6893–6901.
[39] S. Qian, K. Sun, W. Wu, C. Qian, and J. Jia, "Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 10 153–10 163.
[40] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim, "Robust facial landmark detection via recurrent attentive-refinement networks," in *European conference on computer vision*. Springer, 2016, pp. 57–72.
[41] S. Honari, J. Yosinski, P. Vincent, and C. Pal, "Recombinator networks: Learning coarse-to-fine feature aggregation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5743–5752.
[42] Z.-H. Feng, J. Kittler, W. Christmas, P. Huber, and X.-J. 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*, 2017, pp. 2481–2490.
[43] S. Honari, P. Molchanov, S. Tyree, P. Vincent, C. Pal, and J. Kautz, "Improving landmark localization with semi-supervised learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1546–1555.
[44] X. Dong and Y. Yang, "Teacher supervises students how to learn from partially labeled images for facial landmark detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 783–792.
[45] J. P. Robinson, Y. Li, N. Zhang, Y. Fu, and S. Tulyakov, "Laplace landmark localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 10 103–10 112.
[46] M. Zhu, D. Shi, M. Zheng, and M. Sadiq, "Robust facial landmark detection via occlusion-adaptive deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3486–3496.
[47] X. Zou, S. Zhong, L. Yan, X. Zhao, J. Zhou, and Y. Wu, "Learning robust facial landmark detection via hierarchical structured ensemble," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 141–150.
[48] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2879–2886.
[49] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *European conference on computer vision*. Springer, 2012, pp. 679–692.
[50] 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*, vol. 35, no. 12, pp. 2930–2940, 2013.
[51] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "Xm2vtsdb: The extended m2vts database," in *Second international conference on audio and video-based biometric person authentication*, vol. 964, 1999, pp. 965–966.
[52] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5525–5533.
[53] S. Zhu, C. Li, C.-C. Loy, and X. Tang, "Unconstrained face alignment via cascaded compositional learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3409–3417.
[54] A. Kumar and R. 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*, 2018, pp. 430–439.
[55] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5203–5212.

Xing Lan received the B.E degree in computer science from Beijing University of Chemical Technology, Beijing, China, in 2019. He is currently pursuing his master degree in the Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include face alignment, multi-view geometry and 3D vision.



Qinghao Hu received the B.E. degree in computer science from Northwestern Polytechnical University, Xi'an, China, in 2014. He received the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He now works as an associate professor in Institute of Automation, Chinese Academy of Sciences, and his current research interests include deep neural network compression and acceleration, hashing, and quantization.

Jian Cheng received the B.S. and M.S. degrees from Wuhan University, Wuhan, China, in 1998 and 2001, respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2004. From 2004 to 2006, he held a post-doctoral position at Nokia Research Center, Beijing. He has been with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, since 2006. His current research interests include machine learning methods and their applications for image processing and social network analysis.