Teacher Supervises Students How to Learn
From Partially Labeled Images for Facial Landmark Detection

Xuanyi Dong[1,2] and Yi Yang[1,3]

[1]SUSTech-UTS Joint Centre of CIS, Southern China University of Science and Technology

[2] Baidu Research, [3] ReLER, University of Technology Sydney

`xuanyi.dong@student.uts.edu.au`, `yi.yang@uts.edu.au`

Abstract

Facial landmark detection aims to localize the anatomically defined points of human faces. In this paper, we study facial landmark detection from partially labeled facial images. A typical approach is to (1) train a detector on the labeled images; (2) generate new training samples using this detector's prediction as pseudo labels of unlabeled images; (3) retrain the detector on the labeled samples and partial pseudo labeled samples. In this way, the detector can learn from both labeled and unlabeled data to become robust.

In this paper, we propose an interaction mechanism between a teacher and two students to generate more reliable pseudo labels for unlabeled data, which are beneficial to semi-supervised facial landmark detection. Specifically, the two students are instantiated as dual detectors. The teacher learns to judge the quality of the pseudo labels generated by the students and filter out unqualified samples before the retraining stage. In this way, the student detectors get feedback from their teacher and are retrained by premium data generated by itself. Since the two students are trained by different samples, a combination of their predictions will be more robust as the final prediction compared to either prediction. Extensive experiments on 300-W and AFLW benchmarks show that the interactions between teacher and students contribute to better utilization of the unlabeled data and achieves state-of-the-art performance.

1. Introduction

Facial landmark detection aims to find some pre-defined anatomical keypoints of human faces [44, 27, 43, 37]. These keypoints include the corners of a mouth, the boundary of eyes, the tip of a nose, etc [36, 35, 21]. It is usually a prerequisite of a large number of computer vision tasks [26, 39, 3]. For example, facial landmark coordinates are required to align faces to ease the visualization for users when people would like to sort their faces by time
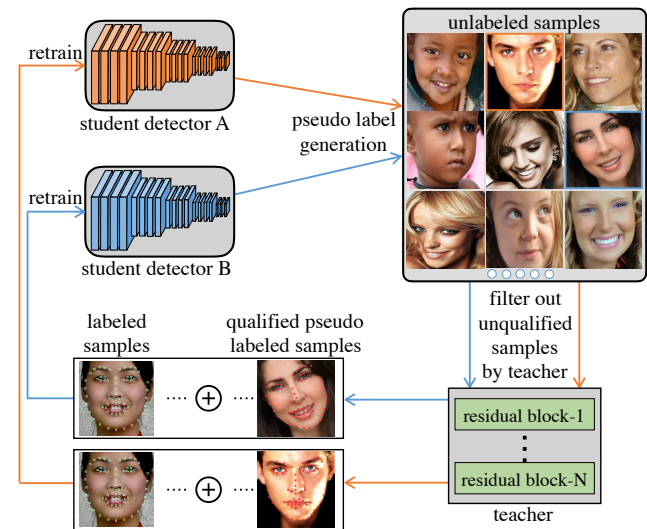

Figure 1. The interaction mechanism between teacher and students. Two student detectors learn to generate pseudo labels for unlabeled samples, among which qualified samples are selected by the teacher. These premium pseudo labeled data along with real labeled data is used for the retraining of the students detectors.

and see the changes over time [9]. Other examples include face morphing [3], face replacement [39], etc.

The main challenge in recent landmark detection literatures is how to obtain abundant facial landmark labels. The annotation challenge comes from two perspectives. First, a large number of keypoints are required for a single face image, e.g., 68 keypoints for each face in the 300-W dataset [35]. To precisely depict the facial features for a whole dataset, millions of keypoints are usually required. Second, different annotators have a semantic gap. There is no universal standard for the annotation of the keypoints, so different annotators give different positions for the same keypoints. A typical way to reduce such semantic deviations among various annotators is to merge the labels from several annotators. This will further increase the costs of the whole annotation work.

Semi-supervised landmark detection can to some extent alleviate the expensive and sophisticated annotations by utilizing the unlabeled images. Typical approaches [17, 2, 23] for semi-supervised learning use self-training or similar paradigms to utilize the unlabeled samples. For example, the authors of [23, 17, 28] adopt a heuristic unsupervised criterion to select the pseudo labeled data for the retraining procedure. This criterion is the loss of each pseudo labeled data, where its predicted pseudo label is treated as the ground truth to calculate the loss [17, 28]. Since no extra supervision is given to train the criterion function, this unsupervised loss criterion has a high possibility of passing inaccurate pseudo labeled data to the retraining stage. In this way, these inaccurate data will mislead the optimization of the detector and make it easier to trap into a local minimum. A straightforward solution to this problem is to use multiple models and regularize each other by the co-training strategy [4]. Unfortunately, even if co-training performs well in simple tasks such as classification [4, 28], in more complex scenarios such as detection, co-training requires extremely sophisticated design and careful tuning of many additional hyper-parameters [12], e.g., more than 10 hyper-parameters for three models in [28].

To better utilize the pseudo labeled data as well as avoid the complicated model tuning for landmark detection, we propose Teacher Supervises StudentS (TS^3). As illustrated in Figure 1, TS^3 is an interaction mechanism between one teacher network and two (or multiple) student networks. Two student detection networks learn to generate pseudo labels for unlabeled images. The teacher network learns to judge the quality of the pseudo labels generated from students. Consequently, the teacher can select qualified pseudo labeled samples and use them to retrain the students. TS^3 applies these steps in an iterative manner, where students gradually become more robust, and the teacher is adaptively updated with the improved students. Besides, two students can also encourage each other to advance their performances in two ways. First, predictions from two students can be ensembled to further improve the quality of pseudo labels. Second, two students can regularize each other by training on different samples. The interactions between the teacher and students as well as the students themselves help to provide more accurate pseudo labeled samples for retraining and the model does not need careful hyper-parameter tuning.

To highlight our contribution, we propose an easy-to-train interaction mechanism between teacher and students (TS^3) to provide more reliable pseudo labeled samples in semi-supervised facial landmark detection. To validate the performance of our TS^3, we do experiments on 300-W, 300-VW, and AFLW benchmarks. TS^3 achieves state-of-the-art semi-supervised performance on all three benchmarks. In addition, using only 30% labels, our TS^3 achieves competitive results compared to supervised methods using all labels on 300-W and AFLW.

2. Related Work

We will first introduce some supervised facial landmark algorithms in Section 2.1. Then, we will compare our algorithm with semi-supervised learning algorithms and semi-supervised facial landmark algorithm in Section 2.2. Lastly, we explain our algorithm in a meta learning perspective in Section 2.3.

2.1. Supervised Facial Landmark Detection

Supervised facial landmark detection algorithms can be categorized into linear regression based methods [44, 7] and heatmap regression based methods [41, 11, 9, 30]. Linear regression based methods learn a function that maps the input face image to the normalized landmark coordinates [44, 7]. Heatmap regression based methods produce one heatmap for each landmark, where the coordinate is the location of the highest response on this heatmap [41, 11, 9, 30, 5]. All above algorithms can be readily integrated into our framework, serving as different student detectors.

These supervised algorithms require a large amount of data to train deep neural networks. However, it is tedious to annotate the precise facial landmarks, which need to average different annotations from multiple different annotators. Therefore, to reduce the annotation cost, it is necessary to investigate the semi-supervised facial landmark detection.

2.2. Semi-supervised Facial Landmark Detection

Some early semi-supervised learning algorithms are difficult to handle large scale datasets due to the high complexity [8]. Others exploit pseudo-labels of unlabeled data in the semi-supervised scenario [1, 2, 23, 28]. Since most of these algorithms studied their effect on small-scale datasets [8, 1, 23, 28], a question remains open: can they be used to improve large-scale semi-supervised landmark detection? In addition, those self-training or co-training approaches [23, 28, 12] simply leverage the confidence score or an unsupervised loss to select qualified samples. For example, Dong et al. [12] proposed a model communication mechanism to select reliable pseudo labeled samples based on loss and score. However, such selection criterion does not reflect the real quality of a pseudo labeled sample. In contrast, our teacher directly learns to model the quality, and selected samples are thus more reliable.

There are only few of researchers study the semi-supervised facial landmark detection algorithms. A recent work [16] presented two techniques to improve landmark localization from partially annotated face images. The first technique is to jointly train facial landmark network with an attribute network, which predicts the emotion, head pose,

etc. In this multi-task framework, the gradient from the attribute network can benefit the landmark prediction. The second technique is a kind of supervision without the need of manual labels, which enables the transformation invariant of landmark prediction. Compared to using the supervision from transformation, our approach leverages a progressive paradigm to learn facial shape information from unlabeled data. In this way, our approach is orthogonal to [16], and these two techniques can complement our approach to further boost the performance.

Radosavovic et al. [31] applied the data augmentation to improve the quality of generated pseudo landmark labels. For an unlabeled image, they ensemble predictions from multiple transformations, such as flipping and rotation. This strategy can also be used to improve the accuracy of our pseudo labels and complement our approach. Since the data augmentation is not the focus of this paper, we did not apply their algorithms in our approach. Dong et al. [11] proposed a self-supervised loss by exploiting the temporal consistence on unlabeled videos to enhance the detector. This is a video-based approach and not the focus of our work. Therefore, we do not discuss more with those video-based approach [20, 11].

2.3. Meta Learning

In a meta learning perspective, our TS^3 learns a teacher network to learn which pseudo labeled samples are helpful to train student detectors. In this sense, we are related to some recent literature in "learning to learn" [25, 33, 13, 45]. For example, Ren et al. [33] learn to re-weight samples based on gradients of a model on the clean validation set. Xu et al. [45] suggest using meta-learning to tune the optimization schedule of alternative optimization problems. Jiang et al. [18] propose an architecture to learn data-driven curriculum on corrupted labels. Fan et al. [13] leverage reinforcement learning to learn a policy to select good training samples for a single student model. These algorithms are designed in the supervised scenarios and can not easily be modified in semi-supervised scenario.

Difference with other teacher-student frameworks and generative adversarial networks (GAN). Our TS^3 learns to utilize the output (pseudo labels) of the student model qualified by the teacher model to do semi-supervised learning. Other teacher-student methods [38, 15, 10, 24] aim to fit the output of the student model to that of the teacher model. The student and teacher in our work do similar jobs as the generator and discriminator in GAN [14], while we aim to predict/generate qualified pseudo labels in semi-supervised learning using a different training strategy.

3. Methodology

In this section, we will first introduce the scenario of the semi-supervised facial landmark detection in Section 3.1.
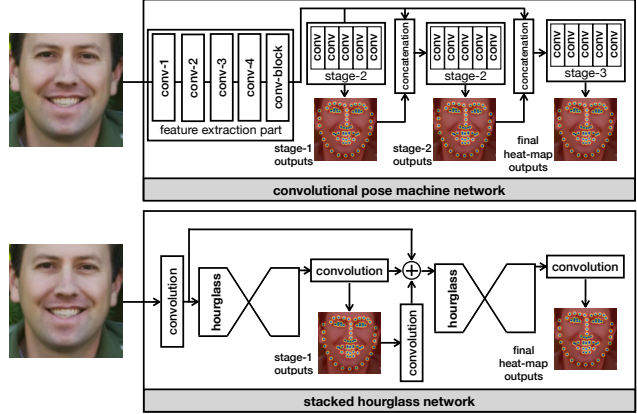

Figure 2. A brief overview of the structure between the two student detection networks in our TS^3. The first network is convolutional pose machine [41] and the second is stacked hourglass [30].

We explain how to design our student detectors and the teacher network in Section 3.2. Lastly, we demonstrate our overall algorithm in Section 3.3.

3.1. The Semi-Supervised Scenario

We introduce some necessary notations for the presentation of the proposed method. Let $\mathcal{L} = \{(x_1, y_1), (x_2, y_2), ..., (x_{n_l}, y_{n_l})\}$ be the labeled data in the training set and $\mathcal{U} = \{(x_{n_l+1}), (x_{n_l+2}), ..., (x_{n_l+n_u})\}$ be the unlabeled data in the training set, where x_i denotes the i-th image, and $y_i \in \mathcal{R}^{2 \times K}$ denotes the ground-truth landmark label of x_i. K is the number of the facial landmarks, and the k-th column of y_i indicates the coordinate of the k-th landmark. n_l and n_u denote the number of labeled data and unlabeled data, respectively. The semi-supervised facial landmark detection aims to learn robust detectors from both $\mathcal{L}$ and $\mathcal{U}$.

3.2. Teacher and Students Design

The Student Detectors. We choose the convolutional pose machine (CPM) [41] and stacked hourglass (HG) [30] models as our student detectors. These two landmark detection architectures are the cornerstone of many facial landmark detection algorithms [30, 9, 6, 37]. Moreover, their architectures are quite different, and can thus complement each other to achieve a better detection performance compared to using two similar neural architectures. Therefore, we integrate these two detectors in our TS^3 approach. In this paragraph, we will give a brief overview of these two facial landmark detectors. We illustrate the structures of CPM and HG in Figure 2. Both CPM and HG are the heatmap regression based methods and utilize the cascaded structure. Formally, suppose there are M convolutional stages in CPM, the output of CPM is:

$$f_1(x_i|\boldsymbol{w}_1) = \{\boldsymbol{H}_i^m | 1 \leq m \leq M\}, \quad (1)$$
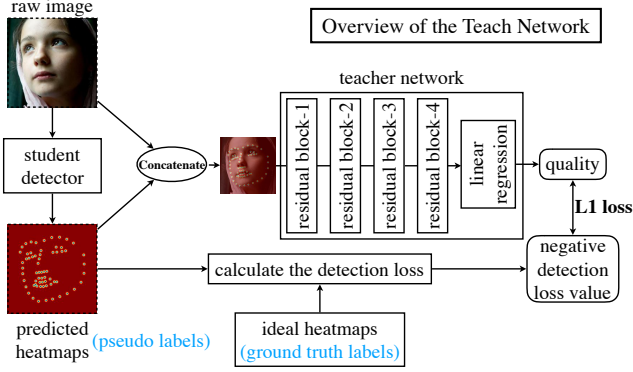
Figure 3. **The illustration of our teacher network.** The input of the teacher is the concatenation of the original RGB face image and the heatmap (pseudo label) predicted by the student detector. The output of teacher is a scalar, representing the quality of the input pseudo labeled face image. During training, we can calculate a detection loss using the ideal heatmap and the predicted heatmap. The teacher aims to fit the negative value of this detection loss by an L1 loss. During evaluation, a higher value of the quality represents a lower detection loss, which means this pseudo labeled image is reliable.

where f_1 indicates the CPM student detector whose parameters are $\boldsymbol{w}_1$. x_i is the RGB image of the i-th data-point and $\boldsymbol{H}_i^m \in \mathcal{R}^{(K+1) \times h' \times w'}$ indicates the heatmap prediction of the m-th stage. h' and w' denote the spatial height and width of the heatmap. Similarly, we use f_2 indicates the HG student detector whose parameters are $\boldsymbol{w}_2$. The detection loss function of the CPM student is:

$$\ell(f_1(x_i|\boldsymbol{w}_1), y_i) = \sum_m^M ||\boldsymbol{H}_i^m - \boldsymbol{H}_i^*||_F^2$$

$$= \sum_m^M ||\boldsymbol{H}_i^m - p(y_i)||_F^2, \quad (2)$$

where p is a function taking the label $y_i \in \mathcal{R}^{2 \times K}$ as inputs to generate the the ideal heatmap $\boldsymbol{H}_i^* \in \mathcal{R}^{(K+1) \times h' \times w'}$. Details of p can be found in [41, 30]. During the evaluation, we take the argmax results over the first K channel of the last heatmap $\boldsymbol{H}_M$ as the coordinates of landmarks, and the $(K+1)$-th channel corresponding to the background will be omitted.

The Teacher Network. Since our student detectors are based on heatmap, the pseudo label is in the form of heatmap and ground truth label is the ideal heatmap. We build our teacher network using the structure of discriminators adopted in CycleGAN [46]. As shown in Figure 3, the input of this teacher network is the concatenation of a face image and its heatmap prediction $\boldsymbol{H}_i^{M1}$. The output of this teacher network is a scalar representing the quality of a pseudo labeled facial image. Since we train the teacher on

---

[1] $\boldsymbol{H}_i^M$ will be resized into the same spatial size as its face image

---

Algorithm 1 The Algorithm Description of Our TS^3

Input: Labeled data $\mathcal{L} = \{(x_i, y_i)|1 \leq i \leq n_l\}$
1:   Unlabeled data $\mathcal{U} = \{(x_i^u)|n_l + 1 \leq i \leq n_u + n_l\}$
2:   Two student detectors f_1 with $\boldsymbol{w}_1$ and f_2 with $\boldsymbol{w}_2$
3:   The teacher network g with parameters $\boldsymbol{w}_g$
4:   The selection ratio r and the maximum step S
5: Initialize the $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ by minimizing Eq. (2) on $\mathcal{L}$
6: **for** $i = 1; i \leq S; i{+}{+}$ **do**
7:     Predict $\boldsymbol{H}_i^M$ on both $\mathcal{L}$ and $\mathcal{U}$ using Eq. (5), and denote $\mathcal{U}$ with its pseudo labels as $\mathcal{U}^1$ ▷ update the first student
8:     Optimize teacher with $\boldsymbol{w}_g$ by minimizing Eq. (4) on $\mathcal{L}$ with prediction $\boldsymbol{H}_i^M$ and ground truth label $\boldsymbol{H}_i^*$
9:     Compute the quality scalar of each sample in $\mathcal{U}^1$ using the optimized teacher via Eq. (3)
10:     Pickup the top $r \times i \times |\mathcal{U}|$ samples from $\mathcal{U}^1$, named as $\mathcal{L}_{\text{ex}}^1$
11:     Retrain $\boldsymbol{w}_1$ on $\mathcal{L}^1 = \mathcal{L} \cup \mathcal{L}_{\text{ex}}^1$ by minimizing Eq. (2)
12:     Predict $\boldsymbol{H}_i^M$ on both $\mathcal{L}$ and $\mathcal{U}$ using Eq. (5), and denote $\mathcal{U}$ with its pseudo labels as $\mathcal{U}^2$ ▷ update the second student
13:     Optimize teacher with $\boldsymbol{w}_g$ by minimizing Eq. (4) on $\mathcal{L}$ with $\boldsymbol{H}_i^M$ and $\boldsymbol{H}_i^*$
14:     Compute the quality scalar of each sample in $\mathcal{U}^2$ using Eq. (3)
15:     Pickup the top $r \times i \times |\mathcal{U}|$ samples from $\mathcal{U}^2$, named as $\mathcal{L}_{\text{ex}}^2$
16:     Retrain $\boldsymbol{w}_2$ on $\mathcal{L}^2 = \mathcal{L} \cup \mathcal{L}_{\text{ex}}^2$ by minimizing Eq. (2)
17: **end for**
Output: Students with optimized parameters $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$

---

the trustworthy labeled data, we could obtain a supervised detection loss by calculating $||\boldsymbol{H}_i^M - \boldsymbol{H}_i^*||_F^2$. We consider the negative value of this detection loss as the ground truth label of the quality, because a high negative value of the detection loss indicates a high similarity between the predicted heatmap and the ideal heatmap. In another word, a higher quality scalar corresponds to a more accurate pseudo label.

Formally, denote the teacher network as g, we have:

$$g(x_i \frown \boldsymbol{H}_i^M | \boldsymbol{w}_g) = q_i, \quad (3)$$

$$\ell_t(g(x_i \frown \boldsymbol{H}_i^M | \boldsymbol{w}_g), y_i) = |q + ||\boldsymbol{H}_i^M - \boldsymbol{H}_i^*||_F^2|, \quad (4)$$

where the parameters of the teacher is $\boldsymbol{w}_g$. "$x \frown \boldsymbol{H}$" first resizes the tensor $\boldsymbol{H}$ into the same spatial shape as x and then concatenates the resized tensor with x to get a new tensor. This new tensor is regarded as pseudo labeled image and will be qualified by the teacher later. The teacher outputs a scalar q_i representing the quality of the i-th sample associated with its pseudo label $\boldsymbol{H}_i^M$. We optimize the teacher on the trustworthy labeled data by minimizing Eq. (4).

3.3. The TS³ Algorithm

Our TS³ aims to progressively improve the performance of the student detector. The key idea is to learn a teacher network that can teach students which pseudo labeled sample is reliable and can be used for training. In this procedure, we define the pseudo label of a facial image is as follows:

$$
\begin{aligned}
f(x_i) &= \frac{1}{2}(f_1(x_i|\boldsymbol{w}_1) + f_2(x_i|\boldsymbol{w}_2)) \\
&= \{\frac{1}{2}(\boldsymbol{H}_i^{(1,m)} + \boldsymbol{H}_i^{(2,m)})|1 \le m \le M\}, \\
&= \{\boldsymbol{H}_i^m|1 \le m \le M\}, \tag{5}
\end{aligned}
$$

where $\boldsymbol{H}_i^{(1,m)}$ indicates the heatmap prediction from the first student at the m-th stage for the i-th sample. $\boldsymbol{H}_i^m$ in Eq. (5) indicates the ensemble result from both two students detection networks. It will be used as the prediction during the inference procedure.

We show our overall algorithm in Algorithm 1. We first initialize the two detectors f_1 and f_2 on the labeled facial images $\mathcal{L}$. Then, in the first round, our algorithm applies the following procedures: (1) generate pseudo labels on $\mathcal{L}$ via Eq. (5) and train the teacher network from scratch with these pseudo labels; (2) generate pseudo labels on $\mathcal{U}$ and estimate the quality of these pseudo labeled using the learned teacher; (3) select some high-quality pseudo labeled samples to retrain one student network from scratch. (4) repeat the first three steps to update another student detection network. In the next rounds, each student can be improved and generate more accurate pseudo labels. In this way, we will select more pseudo labeled samples when retraining the students. As the rounds go, students will gradually become better, and the teacher will also be adaptive with the improved students. Our interaction mechanism helps to obtain more accurate pseudo labels and select more reliable pseudo labeled samples. As a result, our algorithm achieves better performance in the semi-supervised facial landmark detection.

3.4. Discussion

Can this algorithm generalize to other tasks? Our algorithm relies on the design of the teacher network. It requires the input pseudo label to be a structured prediction. Therefore, our algorithm is possible to be applied to tasks with structured predictions, such as segmentation and pose estimation, but is not suitable other tasks like classification.

Limitation. It is challenging for a teacher to judge the quality of a pseudo label for an image, especially when the spatial shape of this image becomes large. Therefore, in this paper, we use an input size of 64×64. If we increase the input size to 256×256, the teacher will fail and need to be modified accordingly. There are two main reasons: (1) the larger resolution requires a deeper architecture or di-

lated convolutions for the teacher network and (2) the high-resolution faces bring high-dimensional inputs, and consequently, the teacher needs much more training data. This drawback limits the extension of our algorithm to high-resolution tasks, such as segmentation. We will explore to solve this problem in the future.

Further improvements. (1) In our algorithm, during the retraining procedure, a part of unlabeled samples are not involved during retraining. To utilize these unlabeled facial images, we could use self-supervised techniques such as [16] to improve the detectors. (2) In this framework, we use only two student detectors, while it is easy to integrate more student detectors. More student detectors are likely to improve the prediction accuracy, but this will introduce more computation costs. (3) The specifically designed data augmentation [31, 42] is another direction to improve the accuracy and precision of the pseudo labels.

Will the teacher network over-fit to the labeled data? In Algorithm 1, since labeled data set $\mathcal{L}$ is used to optimize both teacher and students, the teacher's judgment could suffer from the over-fitting problem. Most of the students' predictions on the labeled data can be similar to the ground truth labels. In other words, most pseudo labeled samples on $\mathcal{L}$ are "correctly" labeled samples. If the teacher is optimized on $\mathcal{L}$ with those pseudo labels, it might only learn what a good pseudo labeled sample is, but overlook what a bad one is. It would be more reasonable to let students predict on the unseen validation set, and then train the teacher on this validation set. However, having an additional validation set during training is different from the typical setting of previous semi-supervised facial landmark detection. We would explore this problem in our future work.

4. Empirical Studies

We perform experiments on three benchmark datasets to investigate the behavior of the proposed method. The datasets and experiment settings are introduced in Section 4.1 and Section 4.2. We first compare the proposed semi-supervised facial landmark algorithm with other state-of-the-art algorithms in Sec. 4.3. We then perform ablation studies in Sec. 4.4 and visualize our results at last.

4.1. Datasets

The 300-W dataset [35] annotates 68 landmarks from five facial landmark datasets, i.e., LFPW, AFW, HELEN, XM2VTS, and IBUG. Following the common settings [11, 9, 27], we regard all the training samples from LFPW, HELEN and the full set of AFW as the training set, in which there is 3148 training images. The common test subset consists of 554 test images from LFPW and HELEN. The challenging test subset consists of 135 images from IBUG to construct . The full test set the union of the common and challenging subsets, 689 images in total.

Ratio	Method	Common	Challenging	Full
100%	MDM [40]	4.83	10.14	5.88
100%	Two-Stage [27]	4.36	7.42	4.96
100%	RDR [43]	5.03	8.95	5.80
100%	Pose-Invariant [19]	5.43	9.88	6.30
100%	HF-ResNet [32]	-	8.18	-
100%	SAN [9]	3.34	6.60	3.98
100%[†]	SBR [11]	3.28	7.58	4.10
100%	PCD-CNN [22]	3.67	7.62	4.44
10%	RCN$^+$ [16]	-	10.35	6.32
10%	TS^3	4.67	9.26	5.64
20%	RCN$^+$ [16]	-	9.56	5.88
20%	TS^3	4.31	7.97	5.03
100%	TS^3	3.17	6.41	3.78
100%[‡]	TS^3	2.91	5.91	3.49

Table 1. Comparisons of the NME results on the 300-W dataset. "Ratio" indicates the annotation ratio of the whole training set. A "Ratio" value of 10% means that only 10% of the training face images have the landmark coordinate labels. [†] indicates that SBR [11] used additional unlabeled video data during training. When we use partially labeled training images, our TS^3 outperforms other semi-supervised algorithm [16]. [‡] indicates we use 100% labeled 300-W training data and unlabeled AFLW training data for our TS^3.

The AFLW dataset [21] contains 21997 real-world images with 25993 faces in total. They provide at most 21 landmark coordinates for each face, but they exclude invisible landmarks. Faces in AFLW usually have a different head pose, expression, occlusion or illumination, and therefore it causes difficulties to train a robust detector. Following the same setting as in [27, 47], we do not use the landmarks of two ears. There are two types of AFLW splits, i.e., AFLW-Full and AFLW-Frontal following [47, 9]. AFLW-Full contains 20000 training samples and 4386 test samples. AFLW-Front uses the same training samples as in AFLW-Full, but only use the 1165 samples with the frontal face as the test set.

The 300-VW dataset [36] is a video-based facial landmark benchmark. It contains 50 training videos with 95192 frames. Following [20, 11], we report the results for the 49 inner points on the category C subset of the 300-VW test set, which has 26338 frames.

4.2. Experimental Settings

Training student detection networks. The first student detector is CPM [41]. We follow the same model configuration as the base detector used in [41, 9], and the number of cascaded stages is set as three. Its number of parameters is 16.70 MB and its FLOPs is 1720.98 M. To train CPM, we apply the SGD optimizer with the momentum of 0.9 and the weight decay of 0.0005. For each stage, we train the CPM

for 50 epochs in total. We start the learning rate of 0.00005, and reduce it by 0.5 at 20-th, 25-th, 30-th, and 40-th epoch.

The second student detector is HG [30]. We follow the same model configuration as [6] but use the number of cascaded stages of four to build our HG model, where the number of parameters is 24.97 MB and FLOPs is 1600.85 M. To train HG, we apply the RMSprop optimizer with the alpha of 0.99. For each stage, we train the HG for 110 epochs in total. We start the learning rate of 0.00025, and reduce it by 0.5 at 50-th, 70-th, 90-th, and 100-th.

For both of these two detectors, we use the batch size of eight on two GPUs. To generate the heatmap ground truth labels, we apply the Gaussian distribution with the sigma of 3. Each face image is first resized into the size of 64×64, and then randomly resized between the scale of 0.9 and 1.1. After the random resize operation, the face image will be randomly rotated with the maximum degree of 30, and then randomly cropped with the size of 64×64^2. We set selection ratio r as 0.1 and the maximum step S as 6 based on cross-validation.

Training the teacher network[3]**.** We build our teacher network using the structure of discriminators adopted in CycleGAN [46]. Given a 64×64 face image, we first resize the predicted heatmap into the same spatial size of 64×64. We use the Adam to train this teacher network. The initial learning rate is 0.01, and the batch size is 128. Random flip, random rotation, random scale and crop are applied as data argumentation.

Evaluation. Normalized Mean Error (NME) is usually applied to evaluate the performance for facial landmark predictions [27, 34, 47, 9]. For the 300-W dataset, we use the inter-ocular distance to normalize mean error following the same setting as in [35, 27, 11, 9]. For the AFLW dataset, we use the face size to normalize mean error [27]. Area Under the Curve (AUC) @ 0.08 error is also employed for evaluation [6, 40]. When training on the partially labeled data, the sets of $\mathcal{L}$ and $\mathcal{U}$ are randomly sampled. During evaluation, we use Eq. (5) to obtain the final heatmap and follow [41, 30] to generate the coordinate of each landmark. We repeat each experiment three times and report the mean result. The codes will be public available upon the acceptance.

4.3. Comparison with state-of-the-art

Comparisons on 300-W. We compare our algorithm with several state-of-the-art algorithms [44, 43, 27, 43, 19, 16], as shown in Table 1. In this table, [9, 22, 11] are very recent methods, which represent the state-of-the-art supervised facial landmark algorithms. By using 100% fa-

---

[2]Different input image resolution can cause different detection performance. We choose 64×64 to ease the training of our teacher network.

[3]Model codes are publicly available on GitHub: https://github.com/D-X-Y/landmark-detection

788

Methods	SDM [44]	LBF [34]	CCL [47]	Two-Stage [27]	SBR [9]†	SAN [9]	DSRN [29]
AFLW-Full	4.05	4.25	2.72	2.17	2.14	1.91	1.86
AFLW-Front	2.94	2.74	2.17	-	2.07	1.85	-
Methods	RCN$^+$ [16] (5%)	TS^3 (5%)		TS^3(10%)		TS^3(20%)	
AFLW-Full	2.17	2.19		2.14		1.99	
AFLW-Front	-	2.03		1.94		1.86	

Table 2. Comparisons of NME normalized by face size on the AFLW dataset. † indicates that SBR [11] used additional unlabeled video data during training. The ratio number in the brackets represents the portion of the labels that we use. Compared to the semi-supervised algorithm [16], our TS^3 obtains a similar NME result (2.19 vs. 2.17). Compared to supervised algorithms which use 100% labels, our TS^3 obtains competitive NME when using only 20% labels.

Method	DGCM [20]	SBR [11]	TS^3
AUC@0.08	59.38	59.39	59.65

Table 3. AUC @ 0.08 error on 300-VW category C. Note that all compared algorithms [20, 11] use all labels on the 300-VW training data and 300-W training data, whereas our TS^3 only uses the unlabeled 300-VW training data and labeled 300-W training data.

Figure 4. We compare three different algorithms, which can train two detectors in a progressive manner: SPL [23, 17], SPaCo [28], and our TS^3. All these algorithms iteratively improve detectors one round by another round. The x-axis shows the results of the first five rounds. The y-axis indicates the NME results on the 300-W full test set.

cial landmark labels on 300-W training set and unlabeled AFLW, our algorithm achieves competitive 3.49 NME on the 300-W common test set, which is competitive to other state-of-the-art algorithms. In addition, even though our approach utilizes two detectors, the number of parameters is much lower than SAN [9]. The robust detection performance of ours can be mainly caused by two reasons. First, the proposed teacher network can effectively sample the qualified pseudo labeled data, which enables the model to exploit more useful information. Second, our framework leverages two advanced CNN architectures, which can complement each other.

We also compare our TS^3 with a recent work on semi-supervised facial landmark detection [16] in Table 1. When using 10% of labels, our TS^3 obtains a lower NME result on the challenging test set than RCN$^+$ [16] (5.64 NME *vs.* 6.32 NME). When using 20% of labels, our TS^3 is also superior to it (5.03 NME *vs.* 5.88 NME). Note that [16] utilizes a transformation invariant auxiliary loss function. This auxiliary loss can also be easily integrated into our framework. Therefore, [16] is orthogonal to our work, combining two methods can potentially achieve a better performance.

Comparisons on AFLW. We also show the NME comparison on the AFLW dataset in Table 2. Compared to semi-supervised facial landmark detection algorithm [16], we achieve a similar performance. RCN$^+$ [16] can learn transformation invariant information from a large amount of unlabeled images, while ours does not consider this information as it is not our focus. On the AFLW-Full test set, using 20% annotation, our framework achieves 1.99 NME, which is competitive to other supervised algorithms. On the AFLW-Front test set, using only 10% annotation, our
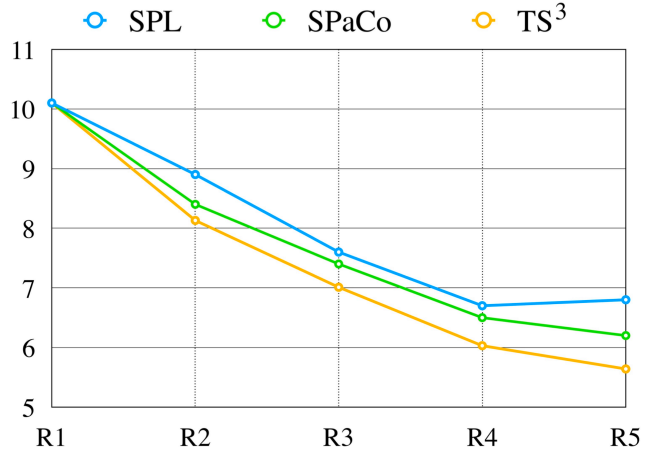
framework achieves competitive NME results to [9]. The above results demonstrate our framework can train a robust detector with much less annotation effort.

Comparisons on 300-VW. We experiment our algorithm to leverage a large amount of unlabeled facial video frames on 300-VW. We use the labeled 300-W training set and the unlabeled 300-VW training set to train our TS^3. We evaluate the learned detectors on the 300-VW C test subset w.r.t. AUC @ 0.08. Some video-based facial landmark detection algorithms [20, 11] utilize the labeled 300-VW training data to improve the base detectors. Compared with them, without using any label on 300-VW, our TS^3 obtains a higher AUC result than them, i.e., 59.65 vs. 59.39, as shown in Table 3.

4.4. Ablation Study

The key contribution of our TS^3 lies on two components: (1) the teacher supervising the training data selection of students. (2) the complementary effect of two students. In this subsection, we validate the contribution of these two com-

Ratio	Method	Common	Challenging	Full
10%	CPM	6.86	14.69	8.28
	HG	5.16	11.28	6.25
	TS^3	4.67	9.26	5.64
20%	CPM	5.36	11.31	6.68
	HG	5.84	10.15	6.68
	TS^3	4.31	7.97	5.03

Table 4. Comparisons of the NME results on the 300-W test sets for different configuration and models. CPM and HG indicate using only one CPM student or only one HG student in our framework. When using a single detector, we use the heatmap of the last stage in Eq. (1) as prediction. When using two students (TS^3), we use $\boldsymbol{H}_i^M$ in Eq. (5) as prediction. "Ratio" indicates the proportion of labeled data in our semi-supervised setting.

ponents to the final detection performance.

The effect of the teacher. Compared to other progressive pseudo label generation strategies [23, 17, 28], our designed teacher can sample pseudo labeled with higher quality. In Figure 4, we show the detection results after the first five training rounds (only 10% labels are used). We use SPL [23, 17] to separately train CPM and HG, and then ensemble them together as Eq. (5). We use SPaCo [28] to jointly optimize CPM and HG in a co-training strategy. To make a fair comparison, at each round, we control the number of pseudo labels is the same across these three algorithms. From Figure 4, several conclusions can be made: (1) TS^3 obtains the lowest NME, because the quality of selected pseudo labels is better than others. (2) SPL falls into a local trap at $round_4$ and results in a higher error at $round_5$, whereas SPaCo and our TS^3 not. This could be caused by that the interaction between two students can help regularize each other. (3) Our TS^3 converges faster than SPaCo and achieves better results. The pseudo labeled data selection in SPaco is a heuristic unsupervised criterion, whereas our criterion is a supervised teacher. Since no extra supervision is given in SPaCo, their criterion might induce inaccurate pseudo labeled samples. Besides, as discussed in Section 3.4, our TS^3 can utilize validation set to further improve the performance by avoid over-fitting, but the compared methods may not effectively utilize validation set.

The effect of the interaction between students. From Table 4, we show the ablative studies on the complementary effect of multiple students. In these experiments, we use the same teacher structure, while "CPM" and "HG" are trained without the interaction between students. Using 10% labels, CPM achieves 8.28 NME, and HG achieves 6.25 NME on 300-W. Leveraging from their mutual benefits, our TS^3 can boost the performance to 5.64, which is higher than CPM by about 30% and than HG by 9%. Under different portion of annotations, we can conclude similar observations. This ablation study demonstrates the contribution of student interaction to the final performance. Note that, our algorithm
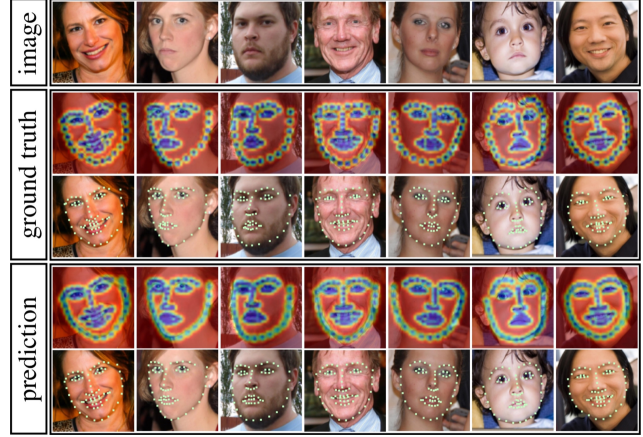

Figure 5. **Qualitative results on images in the 300-W test set.** We train our TS^3 with 314 labeled facial images and 2834 unlabeled facial images in the 300-W training set.

can be readily applied to multiple students without introducing additional hyper-parameters. In contrast, the number of hyper-parameters in other co-training strategies [28, 12] is quadratic to the number of detectors.

4.5. Qualitative Analysis

On the 300-W training set, we train our TS^3 using only 10% labeled facial images, and we show some qualitative results of the 300-W test set in Figure 5. The first row shows seven raw input facial images. The second row shows the ground truth background heatmaps, and the third row shows the faces with ground truth landmarks of these images. We visualize the predicted background heatmap in the fourth row and the predicted coordinates in the fifth row. As we can see, the predicted landmarks of our TS^3 are very close to the ground truth. These predictions are already robust enough, and human may not be able to distinguish the difference between our predictions (the third line) and the ground truth (the fifth line).

5. Conclusion

In this paper, we propose an interaction mechanism between a teacher and multiple students for semi-supervised facial landmark detection. The students learn to generate pseudo labels for the unlabeled data, while the teacher learns to judge the quality of these pseudo labeled data. After that, the teacher can filter out unqualified samples; and the students get feedback from the teacher and improve itself by the qualified samples. The teacher is adaptive along with the improved students. Besides, multiple students can not only regularize each other but also be ensembled to predict more accurate pseudo labels. We empirically demonstrate that the proposed interaction mechanism achieves state-of-the-art performance on three facial landmark benchmarks.

790

References

[1] Phil Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. In *NeurIPS*, 2014.

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, 2009.

[3] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *IEEE TPAMI*, 2003.

[4] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *CLT*, 1998.

[5] Adrian Bulat and Georgios Tzimiropoulos. Convolutional aggregation of local evidence for large pose face alignment. In *BMVC*, 2016.

[6] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *ICCV*, 2017.

[7] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *IJCV*, 2014.

[8] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-supervised learning*. MIT press Cambridge, 2006.

[9] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *CVPR*, 2018.

[10] Xuanyi Dong and Yi Yang. Network pruning via transformable architecture search. *arXiv preprint arXiv:1905.09717*, 2019.

[11] Xuanyi Dong, Shoou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang, and Yaser Sheikh. Supervision-by-Registration: An unsupervised approach to improve the precision of facial landmark detectors. In *CVPR*, 2018.

[12] Xuanyi Dong, Liang Zheng, Fan Ma, Yi Yang, and Deyu Meng. Few-example object detection with model communication. *IEEE TPAMI*, 2018.

[13] Yang Fan, Fei Tian, Tao Qin, Xiang-Yang Li, and Tie-Yan Liu. Learning to teach. In *ICLR*, 2018.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Workshop*, 2014.

[16] Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher Pal, and Jan Kautz. Improving landmark localization with semi-supervised learning. In *CVPR*, 2018.

[17] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G Hauptmann. Self-paced curriculum learning. In *AAAI*, 2015.

[18] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, 2018.

[19] Amin Jourabloo, Xiaoming Liu, Mao Ye, and Liu Ren. Pose-invariant face alignment with a single cnn. In *ICCV*, 2017.

[20] Muhammad Haris Khan, John McDonagh, and Georgios Tzimiropoulos. Synergy between face alignment and tracking via discriminative global consensus optimization. In *ICCV*, 2017.

[21] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *ICCV Workshop*, 2011.

[22] Amit Kumar and Rama Chellappa. Disentangling 3D pose in a dendritic CNN for unconstrained 2D face alignment. In *CVPR*, 2018.

[23] M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *NeurIPS*, 2010.

[24] Hong Joo Lee, Wissam J Baddar, Hak Gu Kim, Seong Tae Kim, and Yong Man Ro. Teacher and student joint learning for compact facial landmark detection network. In *ICMM*, 2018.

[25] Lu Liu, Tianyi Zhou, Guodong Long, Jing Jiang, Lina Yao, and Chengqi Zhang. Prototype propagation networks (PPN) for weakly-supervised few-shot learning on category graph. In *IJCAI*, 2019.

[26] Yu Liu, Fangyin Wei, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Exploring disentangled feature representation beyond face identification. In *CVPR*, 2018.

[27] Jiangjing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, and Xi Zhou. A deep regression architecture with two-stage reinitialization for high performance facial landmark detection. In *CVPR*, 2017.

[28] Fan Ma, Deyu Meng, Qi Xie, Zina Li, and Xuanyi Dong. Self-paced co-training. In *ICML*, 2017.

[29] Xin Miao, Xiantong Zhen, Xianglong Liu, Cheng Deng, Vassilis Athitsos, and Heng Huang. Direct shape regression networks for end-to-end face alignment. In *CVPR*, 2018.

[30] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[31] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *CVPR*, 2018.

[32] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE TPAMI*, 2019.

[33] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, 2018.

[34] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment via regressing local binary features. *IEEE TIP*, 2016.

[35] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV Workshop*, 2013.

[36] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *ICCV Workshop*, 2015.

[37] Zhiqiang Tang, Xi Peng, Shijie Geng, Lingfei Wu, Shaoting Zhang, and Dimitris Metaxas. Quantized densely connected u-nets for efficient landmark localization. In *ECCV*, 2018.

[38] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017.

[39] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016.

[40] George Trigeorgis, Patrick Snape, Mihalis A Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *CVPR*, 2016.

[41] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[42] Yue Wu, Tal Hassner, KangGeon Kim, Gerard Medioni, and Prem Natarajan. Facial landmark detection with tweaked convolutional neural networks. *IEEE TPAMI*, 2017.

[43] Shengtao Xiao, Jiashi Feng, Luoqi Liu, Xuecheng Nie, Wei Wang, Shuicheng Yan, and Ashraf Kassim. Recurrent 3d-2d dual learning for large-pose facial landmark detection. In *CVPR*, 2017.

[44] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013.

[45] Haowen Xu, Hao Zhang, Zhiting Hu, Xiaodan Liang, Ruslan Salakhutdinov, and Eric Xing. AutoLoss: Learning discrete schedules for alternate optimization. In *ICLR*, 2019.

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

[47] Shizhan Zhu, Cheng Li, Chen-Change Loy, and Xiaoou Tang. Unconstrained face alignment via cascaded compositional learning. In *CVPR*, 2016.