Iterative Self-distillation for Precise Facial Landmark Localization

Shigenori Nagaer[1] and Yamato Takeuchi[1]

OMRON Corporation, 9-1 Kizugawadai, Kizugawa-City Kyoto, 619-0283 JAPAN
{shigenori.nagae,yamato.takeuchi}@omron.com

Abstract. In this paper, we propose a novel training method to improve the precision of facial landmark localization. When a facial landmark localization method is applied to a facial video, the detected landmarks occasionally jitter, whereas the face apparently does not move. We hypothesize that there are two causes that induce the unstable detection: (1) small changes in input images and (2) inconsistent annotations. Corresponding to the causes, we propose (1) two loss terms to make a model robust to changes in the input images and (2) self-distillation training to reduce the effect of the annotation noise. We show that our method can improve the precision of facial landmark localization by reducing the variance using public facial landmark datasets, 300-W and 300-VW. We also show that our method can reduce jitter of predicted landmarks when applied to a video.

1 Introduction

Facial landmark localization is widely used as a pre-process of many computer vision tasks, such as face recognition [1], face reconstruction [2], and measurement of biometrics from face, such as gaze estimation [3] and heart rate estimation [4]. High precision of the facial landmark localization is required for the reproducible results of these methods.

Although recent studies on facial landmark localization have significantly improved its accuracy, less attention has been paid to its precision. In fact, many landmark detectors output jittering landmarks when applied to a video (Figure 4). Our assumption is that the imprecise detection is caused by two factors: (1) small changes in the input images and (2) inconsistent annotations of facial landmark datasets.

Consider when landmark localization is carried out on a face in a video. Generally facial images in consecutive frames are similar but slightly different. If a landmark detector is not trained to be robust to such a small change, the exact position of the predicted landmarks will be different among the frames and the difference appears as jitter of the landmarks. We found that a loss term called Equivalent Landmark Transform (ELT), which is used in semi-supervised training, can make the model robust to the changes in the input images in supervised training.

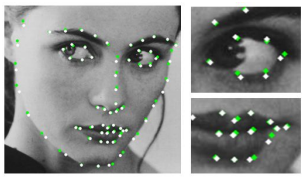

Fig. 1. In the 300-W dataset, some faces have two different annotations. An example of the two annotations are shown in the figure as the white and green circles. There are large differences between the two annotations at some landmarks (e.g., facial contours). Even at discriminative landmarks such as the tail of the eye, small differences are observed.

Inconsistent annotations are another cause of the unstable localization. We found that in the 300-W dataset, some faces are annotated twice and the a clear difference can be observed between the two annotations (Figure 1). Such noisy annotations will confuse a trained model and lead to unstable landmark localization. To reduce the effect of the noise contained in the annotated landmarks, we propose training a model in an unsupervised fashion using its output as a supervision instead of the annotated landmarks after the model is trained in a supervised way.

Our contributions are: (1) present a novel training method on the basis of iterative self-distillation and two loss terms to improve the precision of the facial landmark localization, (2) show that our method can reduce the variance of the detection result in a facial dataset 300-W and actually suppress jitter of predicted landmarks on a facial video dataset 300-VW.

2 Related Works

2.1 Facial Landmark Localization

Facial landmark localization has been intensively studied for over two decades. Active Appearance Models (AAM) is an early successful method [5]. The method solves the facial landmark localization problem by modeling a whole facial shape consistent with the appearance. More recently, Cao et al. used cascaded regressors that map facial appearances to the landmark coordinates [6]. Ren et al. used local binary features to improve the performance up to 3000 FPS [7].

Since convolutional neural networks (CNNs) were introduced in this field, significant progress has been achieved [8–10]. Sun et al. used three cascaded networks to gradually improve landmark prediction [8]. Zhang et al. improved localization robustness with a multi-task learning framework, in which landmark coordinates, facial attributes and head poses are predicted simultaneously [9]. Merget et al. proposed a global context network to complement local features extracted by fully CNN [10].

Most of these studies focused on localizing facial landmarks on static images. When these detectors are applied to a video in a tracking-by-detection fashion,

the detected landmarks occasionally jitter. Although some studies used video datasets to train their models [11–14], they focused on difficulties specific to faces on video, such as the change of the illumination and large changes in facial pose. The jittering problem is often ignored.

To overcome the unstable detection, Dong et al. uses a temporal constraint among successive frames in an unsupervised way to track landmarks [15]. Although this method effectively reduces the jitter of the detected landmarks, it requires a video dataset to train a network. Our training framework requires only static images, showing that the temporal information is not required for precise landmark localization.

2.2 Self-Distillation

Knowledge distillation (KD) is commonly used to train a smaller student network effectively by using information acquired by a larger teacher network. Hinton et al. uses the final output of a teacher network as a soft target to train a student network [16]. Recently, Gao et al. applied the technique to facial landmark localization by transferring intermediate features of a teacher network to a smaller student network [17].

Whereas KD is used to train a small network by transferring knowledge from a larger network, student and teacher networks have the same architecture in self-distillation (SD). Furlanello et al. trained a series of students iteratively [18]. The output of a trained network in the previous iteration is used as a supervision to a student model in the next iteration. Finally, an ensemble of the students is used to obtain additional gains [18]. Interestingly, SD has been used to refine erroneous ground-truth labels [19, 20]. Bagherinezhad et al. iteratively trained a student model one-by-one and observed the knowledge from the previous iteration can help to refine the noisy labels [19]. Kato et al. used the output of the trained model to recover erroneous labels or missing labels in multi-person pose estimation [20]. Similar to these studies, we used the output of a trained model to remove the annotation noise.

3 Method

We assume that the unstable landmark localization is caused by two factors: small changes in input images and the annotation noise. Corresponding to the factors, our method consists of two components: (1) two loss terms and (2) iterative self-distillation. The loss terms are introduced to make a model robust to small changes of the input images. The self-distillation technique is used to reduce the effect of the annotation noise contained in ground-truth labels.

3.1 Loss terms

Equivariant Landmark Transformation (ELT) loss. When a landmark localization method is applied to a video, most successive frames are very similar

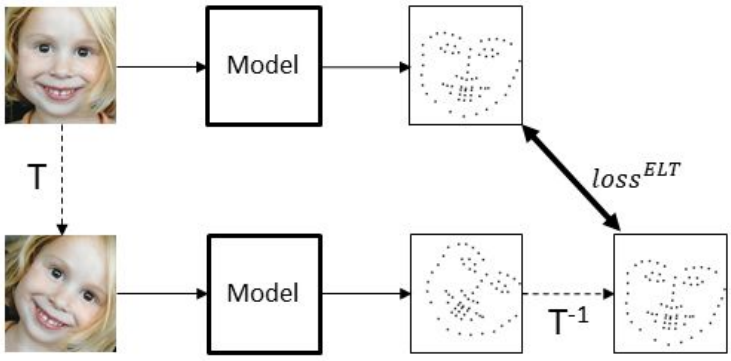

Fig. 2. An input image is transformed by an affine transformation T. The output to the transformed image is then transformed by the inverse transformation T^{-1}. The ELT loss is defined as a distance between the inversely transformed output and the output of the original image.

but not the same. The small difference in the input images is caused by some reasons and one of them is small facial movements. If a landmark detector cannot perfectly follow the movement, a slight difference occurs between the predicted and actual landmarks. The difference will appear as jittering landmarks. We believe that this is one of the main causes of unstable landmark localization.

To address the problem, we directly force the model to follow the movement of input images by adding a loss term. Let T be an arbitrary affine transformation, and the loss is expressed as below:

$$loss^{ELT} = \frac{1}{N} \sum_{n} g(f(I_n),\, T^{-1} \otimes f(T \otimes I_n)) \tag{1}$$

where f is a trained model, $I_n \in \mathbb{R}^{h \times w \times 3}$ is the n th input image, N is the number of training images in a batch, and g is a loss function.

The loss means that if an input image is transformed by an affine transformation, the output should be equally transformed (Figure 2). This loss is proposed by Honari et al. as Equivariant Landmark Transformation (ELT) [21]. The ELT loss is first introduced for images without annotations in semi-supervised training because the loss does not require any annotation. In this paper we found that the loss effectively reduces the variance of the detected landmarks as described in Section 4.

Scale Compensation Term (SCT). The ELT loss works well for making a model robust to small changes of input images. However, we found that the model's output slightly moves towards the center of images when the ELT loss is used. This may be because points near to the center of the image move less than points far from the center by affine transformation induced in calculation of the ELT loss, especially by rotation and resize transformation. Therefore, it may be

easier for the model to track points near to the center. To overcome the shrinking effect of the ELT loss, we introduce a second loss term, scale compensation term (SCT). The SCT loss penalizes the change of the scale of the output landmarks through training. The SCT loss is defined as

$$loss^{SCT} = \frac{1}{N}\sum_{n} |\sigma(f(I_n)) - \sigma(\boldsymbol{l_n})| \tag{2}$$

where $\boldsymbol{l_n} \in \mathbb{R}^{2L}$ is the ground-truth landmarks of the n th sample and σ is a scale function that measures a scale of predicted landmarks defined as

$$\sigma_x(\boldsymbol{l_n}) = \sum_{i} \|\boldsymbol{l_{n,i}} - \overline{\boldsymbol{l_n}}\|_1 \tag{3}$$

where $\boldsymbol{l_{n,i}} \in \mathbb{R}^2$ is a i th landmark in $\boldsymbol{l_n}$ and $\overline{\boldsymbol{l_n}}$ is the average of $\boldsymbol{l_{n,i}}$ (i.e. the centroid of the landmarks).

Overall Loss. In addition to the two loss terms, we also use a standard loss term:

$$loss^{GT} = \frac{1}{N}\sum_{n} g(f(I_n),\, \boldsymbol{l_n}) \tag{4}$$

The overall loss is the weighted sum of the loss terms:

$$loss = loss^{GT} + w^{ELT} loss^{ELT} + w^{SCT} loss^{SCT} \tag{5}$$

where w^{ELT} and w^{SCT} are fixed coefficients. In this paper, we use $w^{ELT} = 1$ and $w^{SCT} = 1$.

3.2 Self-distillation (SD).

We found that the ground-truth labels may contain noise (Figure 1). One source of the noise is a variance among annotators. It is common to hire multiple annotators to make a large dataset. The definition of landmarks is usually shared among the annotators, but it is difficult for the annotators to point the exactly same position in a facial image [15]. In fact, we found two different annotations to the same image in the 300-W dataset and the difference between the annotations can be observed (Figure 1). The inconsistency of the annotation may confuse a trained model and lead to unstable detection.

We observed that a trained model with the above loss terms output landmarks more consistently than the ground-truth (Supplementary Figure 1) and we hypothesize that the outputs of the model can be used as a ground-truth with less noise. Thus, we use a self-distillation (SD) method iteratively to reduce the effect of the annotation noise. In SD, a student model in an iteration is trained to fit the output of a teacher model trained in the previous iteration,

where the student and the teacher have the same architecture. Specifically, the student model in the i th iteration is trained to minimize $loss_i^{ST}$ defined as

$$loss_i^{SD} = \frac{1}{N} \sum_n g(f_i(I_n),\ f_{i-1}(I_n)) \tag{6}$$

where f_i is a trained model in i th iteration and f_{i-1} is a model trained in the previous iteration. The parameters of f_{i-1} are fixed in the i th iteration.

The ELT loss and the SCT loss are also used in the SD part. The ELT loss in the SD is the same as the one in the supervised training part except for using f_i instead of f in Equation 1. The SCT loss has more important role in the SD because the shrinking effect of the ELT loss becomes more significant with longer epochs by iterative training. To keep the scale of output landmarks through the iteration, we used the output of the model trained in the supervised training as a reference and is fixed throughout the SD training (Figure 3):

$$loss_i^{SCT} = \frac{1}{N} \sum_n |\sigma(f(I_n)) - \sigma(f_0(I_n))| \tag{7}$$

where f_0 is the model trained in the supervised training part. The parameters of the model f_0 are fixed during the SD training.

The training process in the i th iteration is shown in Figure 3. The overall loss in the i th iteration is

$$loss_i = loss_i^{SD} + w_i^{ELT} loss_i^{ELT} + w_i^{SCT} loss_i^{SCT} \tag{8}$$

where w_i^{ELT} and w_i^{SCT} are coefficients and fixed during each iteration. We use $w_i^{ELT} = w_i^{SCT} = i + 1$ for i th iteration ($i = 1$ at the first iteration).

4 Result

4.1 Dataset

We used 300-W facial landmark datasets [22–24] to train our models. The 300-W dataset re-annotated various datasets, including LFPW [25], AFW [26], HELEN [27], and XM2VTS [28] with 68 landmarks. We split the dataset into four subsets following [29]: training, common testing, challenging testing, and full testing. For facial landmark localization in video, we used the 300-VW dataset [30–32], which contains video clips of training subjects and testing subjects. We used the dataset only to evaluate our trained models. Therefore, we used only the testing clips in this paper. The test dataset contains 64 clips with 123,405 frames in total.

4.2 Model

We tested our method with two kinds of neural networks. One is based on residual networks [33]. We removed an average pooling layer and a softmax layer at the end of the network because the average pooling removes positional information

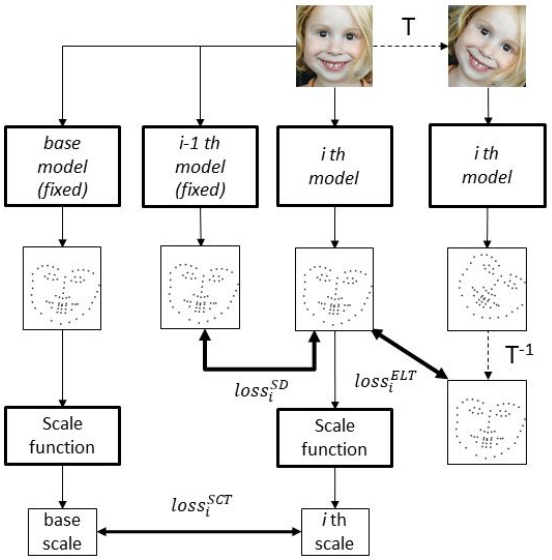

Fig. 3. Overview of our training method in the i th iteration. First, the base model is trained to minimize *loss* in Equation 5. The $i-1$ th model is a model trained in the previous iteration. Both models are fixed and only the i th model is trained in the i th iteration. Outputs from the $i-1$ th model are used instead of a ground truth and are compared with outputs from the i th model ($loss_i^{SD}$). Outputs from the base model are used to keep the scale of the outputs among iterations unchanged ($loss_i^{SCT}$). The input image is transformed with an affine transformation and fed into the i th model. The output is inversely transformed and compared with the output of the original image ($loss_i^{ELT}$).

and the softmax operation is not appropriate with our regression tasks. We also change the output dimension of the last fully connected layer from 1,000 to $2L$. We denote the model as "ResNet18". The other is based on the Face Alignment Network (FAN) [34]. Although the FAN has four Hour-Glass (HG) modules, we used two HG modules to reduce the computational burden and we observed little effect on the accuracy by reducing the HG modules. We denote the models with two HG as "FAN2HG".

We used different loss function g for the two kinds of models because the outputs from the models are different. Whereas ResNet18 outputs landmark coordinates $\boldsymbol{l_o} \in \mathbb{R}^{L\times 2}$, FAN2HG outputs a heatmap of the same size with the input image for each landmark. Therefore, the total output dimension of FAN2HG is $h \times w \times L$ where h and w is the height and the width of input images. We calculated the landmark coordinates as the centroids of the heatmaps as below:

$$\boldsymbol{l_{o,i}} = \sum_{\boldsymbol{p}} \frac{\boldsymbol{p} H_i(\boldsymbol{p})}{\sum_{\boldsymbol{p'}} H_i(\boldsymbol{p'})} \tag{9}$$

where H_i is the output heatmap for the i th landmark, $\boldsymbol{l_{o,i}} \in \mathbb{R}^2$ is the coordinate of the landmark and $\boldsymbol{p}, \boldsymbol{p'} \in \mathbb{R}^2$ is iterated over all the pixels in H_i.

We used a L2 loss function as the loss function for ResNet18. For FAN2HG, we have two kinds of output: landmark coordinates and heatmaps. We applied the L2 loss function to the both outputs. The overall loss is the sum of the values. The ground-truth heatmaps are not provided by the datasets, so we generated them from the ground-truth landmarks as below:

$$H_i^{GT}(\boldsymbol{p}) = A exp(-\frac{\|\boldsymbol{p} - \boldsymbol{l_i}\|_2^2}{\sigma^2}) \tag{10}$$

where A and σ are constants and $\boldsymbol{l_i}$ is the i th landmark position of ground-truth. We used A = 4096 and $\sigma = 1$ in this paper.

The ResNet18 and FAN2HG models were trained with the 300-W training set for experiments using 300-W and 300-VW. First, the models were trained using ground-truth landmarks. Then six and three iterations of self-distillation were carried out for ResNet18 and FAN2HG, respectively. Note that at the beginning of each SD iteration, parameters of the student network were initialized with the parameters of the teacher network (i.e. the same parameters at the end of the previous iteration were used), because we found that initializing the student network with random values leads to worse result.

4.3 Implementation Detail

We used Chainer as a deep learning framework [35–37]. We cropped faces and resized to 256 × 256. The crop size is determined by 1.4 × the size of a bounding box of the landmarks. They were then normalized by the mean and the variance of 300-W and augmented by random flips and scaling (0.5-1.1). For the affine transformation of the ELT loss, we used a combination of random translation (±8 pixels), scaling (0.8-1.0) and rotation (±30 degree). The Adam optimizer was used for training with a weight decay of 5.0×10^{-4}, β_1 of 0.9, and β_2 of 0.999 with a mini-batch size of 12. The initial learning rate was 10^{-2} and decreased by one tenth at 150 and 230 epochs. The training ended at 250 epochs. In subsequent SD iterations, the learning rate was reset to 10^{-2} and decreased by one tenth at 100 and 130 epochs. The training in each SD iteration ended at 150 epochs.

4.4 Experimental Result

Evaluation Metric. To analyze the variance of detected landmarks, we calculated normalized root mean square error (NRMSE) and decomposed it into bias

and standard deviation (std). Bias, std, and NRMSE are calculated as below:

$$\boldsymbol{d_{im}} = \frac{\boldsymbol{l_{o,im}} - \boldsymbol{l_{im}}}{D_m} \tag{11}$$

$$bias = \frac{1}{LM} \| \sum_{i,m} \boldsymbol{d_{im}} \|_2 \tag{12}$$

$$std = \sqrt{\frac{1}{LM} \sum_{i,m} \| \boldsymbol{d_{im}} - \overline{\boldsymbol{d_{im}}} \|_2^2} \tag{13}$$

$$NRMSE = \sqrt{\frac{1}{LM} \sum_{i,m} \| \boldsymbol{d_{im}} \|_2^2} \tag{14}$$

where L is the number of landmarks, M is the number of test samples, $\boldsymbol{l_{o,im}} \in \mathbb{R}^2$ is the i th predicted landmark of the m th sample, $\boldsymbol{l_{im}} \in \mathbb{R}^2$ is the ground-truth landmark, D_m is a normalized factor (i.e., distance between outer corners of the eye) of the m th sample, and $\boldsymbol{d_{im}}$ is the normalized displacement between the predicted and the ground-truth landmarks.

Mislocalized landmarks generally have large errors and dominant effects on the NRMSE. We are interested in measuring the small vibrations of correctly localized landmarks, so we removed such incorrect detections by rejecting landmarks with a large error. Specifically, if an error of a predicted landmark $\| \boldsymbol{d_{im}} \|_2 > \alpha$, the landmark is not used in calculating bias, std and NRMSE. We used $\alpha = 0.05$ in this paper but qualitatively the same results were obtained with other values of α.

Result on 300-W. We trained ResNet18 and FAN2HG models with the 300-W training set and calculated bias, std and NRMSE on the full test set. The results are shown in Table 1. In both models, our loss terms reduced the std. The iterative SD further improves the std and achieved the lowest value compared with the recent models, indicating that the FAN2HG with our method was most precise among the compared models. Although we did not observe a consistent effect of the loss terms on the bias, SD iterations decreased the bias in both cases. The loss terms and the SD decreased NRMSE in both models, indicating that our method also improves the accuracy in addition to the precision.

Result on 300-VW. To evaluate whether our method can actually reduce the jitter of landmarks, we used a video dataset, 300-VW. The FAN2HG trained with 300-W with or without our method was used as landmark detectors. An example of the detected landmarks is shown in Figure 4. In the figure, detected landmarks during one second (25 frames) are plotted to show how the detected landmarks move during the period. Figure 4(a) and (b) shows the movement of detected landmarks was actually smaller with our method than without it. Figure 4(c) and (d) shows that the movement of the landmark is not directional, indicating that the movement is not driven by a movement of the face.

Table 1. Bias, standard deviation (std) and NRMSE of landmark localization by ResNet18 and FAN2HG on 300-W. '+ loss' means that the model was trained with the ELT and the SCT losses. '+ SD' means that SD iterations were carried out. All the values in the table are scaled by 10^3.

Method	bias	std	NRMSE
SAN [38]	0.87	20.8	22.6
SBR [15]	**0.50**	19.1	**19.1**
HRNet [39]	5.14	23.7	24.3
ResNet18	5.92	25.4	26.1
ResNet18 + loss	7.28	20.3	21.6
ResNet18 + loss + SD	7.07	20.1	21.3
FAN2HG	6.41	20.0	21.0
FAN2HG + loss	5.64	19.2	20.0
FAN2HG + loss + SD	4.00	**18.7**	**19.1**

Table 2. Bias, standard deviation (std) and NRMSE of landmark localization on 300-VW by FAN2HG trained with 300-W. '+ loss' means that the model was trained with the ELT and the SCT losses. '+ SD' means that SD iterations were carried out. All the values in the table are scaled by 10^3.

Method	bias	std	NRMSE
SAN [38]	9.20	28.0	29.5
SBR [15]	5.51	26.9	27.4
HRNet [39]	17.9	26.5	32.0
FAN2HG	4.86	26.9	27.3
FAN2HG + loss	7.62	25.6	26.7
FAN2HG + loss + SD	**4.49**	**25.3**	**25.7**

The bias, standard deviation, and NRMSE on 300-VW with FAN2HG are shown in Table 2. The proposed loss terms reduces the std but increases the bias. Iterative SD cancelled the decrease of the bias and further improved the std, indicating that our method can improve the precision of the model. As in the case of 300-W, the FAN2HG with our method achieved the lowest std.

Ablation Study. We showed that our method decreased the variance of the localization result and reduced the jitter of predicted landmarks. Our method has three key components: iterative self-distillation, ELT loss and SCT loss. To clarify the effect of each component, we trained the FAN2HG model with the 300-W dataset with four conditions of loss terms:

- only $loss^{GT}$
- $loss^{GT}$ and $loss^{ELT}$
- $loss^{GT}$ and $loss^{SCT}$
- all losses

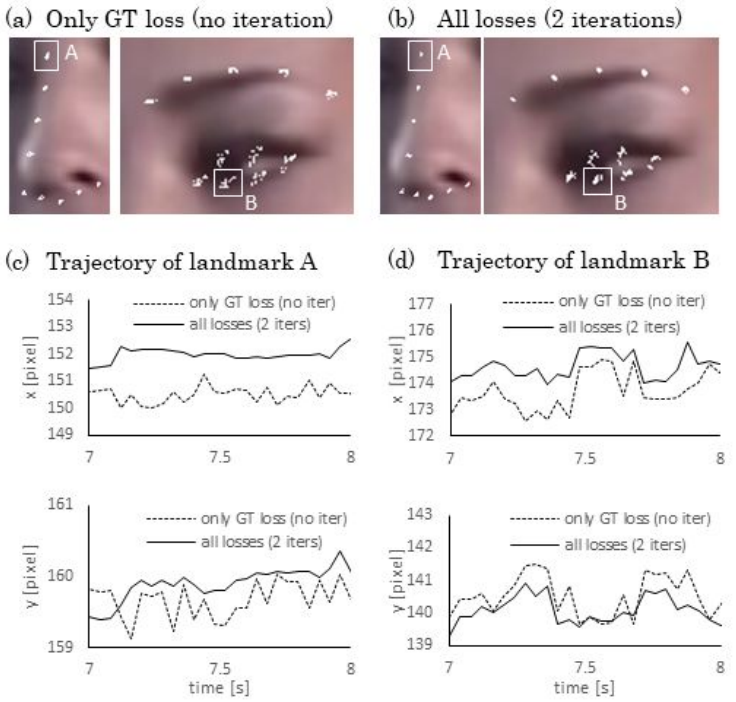

Fig. 4. An example of localization result in 300-VW. (a, b) Detected landmarks around the nose and the left eye during one second (25 frames) are plotted. The detector was FAN2HG trained with 300-W without (a) or with (b) SD iterations, ELT loss and SCT loss. (c) Trajectories of a representative landmark (specified as A in (a) and (b)). (d) Trajectories of other representative landmark (specified as B in (a) and (b)).

The result is shown in Figure 5. When $loss^{ELT}$ was used ('GT + ELT loss' and 'all losses' in Figure 5), the bias was initially high but rapidly decreased to values comparable to other conditions when SD is applied. In contrast, the standard deviation (std) consistently decreased by using $loss^{ELT}$. $loss^{SCT}$ slightly improves the std but we did not observe a consistent effect on the bias. SD improved the std and also the bias, but more than two iterations was harmful in the case of FAN2HG.

In conclusion:

- $loss^{ELT}$ can decrease the variance but may increases the bias.
- $loss^{SCT}$ slightly improves the variance.
- SD improves the variance. It also improves the bias with small iterations.

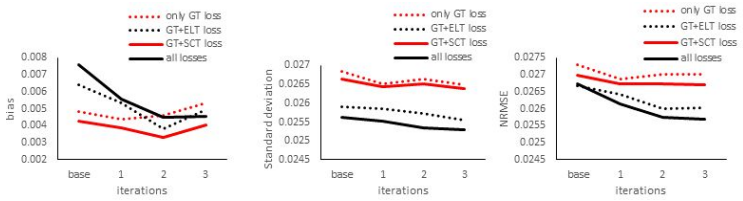

Fig. 5. Bias, standard deviation and NRMSE on 300-VW was measured. The detector was FAN2HG trained on 300-W with some combination of loss functions.

4.5 Conclusion

In this paper, we propose a novel training method to improve the stability of landmark localization. We assume that there are two causes of the instability: (1) the small changes in input images and (2) annotation noise. Corresponding to the causes, we proposed a training method using (1) ELT and SCT losses and (2) self-distillation to stabilize the localization result. We showed our method successfully reduces the variance of the localization result and suppresses the jitter of the predicted landmarks in videos.

We introduced the ELT loss to make a model robust to the small changes in input images. In calculation of the ELT loss, The affine transformation is used to mimic the small changes in input images. However, the changes in input images are caused by various reasons, such as camera noises and local facial movements, which are not considered in this paper. Incorporating these reasons to our framework might lead to a better training method and it is an interesting future direction.

References

1. Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: 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) 2144–2151
2. Gilani, S.Z., Mian, A., Eastwood, P.: Deep, dense and accurate 3d face correspondence for generating population specific deformable models. Pattern Recognition **69** (2017) 238–250
3. Wang, K., Zhao, R., Ji, Q.: A hierarchical generative model for eye image synthesis and eye gaze estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 440–448
4. Wang, Z., Yang, X., Cheng, K.T.: Accurate face alignment and adaptive patch selection for heart rate estimation from videos under realistic scenarios. PloS one **13** (2018) e0197275
5. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Transactions on Pattern Analysis & Machine Intelligence (2001) 681–685
6. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. International Journal of Computer Vision **107** (2014) 177–190

7. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 fps via regressing local binary features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 1685–1692
8. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2013) 3476–3483
9. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: European conference on computer vision, Springer (2014) 94–108
10. Merget, D., Rock, M., Rigoll, G.: Robust facial landmark detection via a fully-convolutional local-global context network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 781–790
11. Liu, H., Lu, J., Feng, J., Zhou, J.: Two-stream transformer networks for video-based face alignment. IEEE transactions on pattern analysis and machine intelligence **40** (2017) 2546–2554
12. Sánchez-Lozano, E., Tzimiropoulos, G., Martinez, B., De la Torre, F., Valstar, M.: A functional regression approach to facial landmark tracking. IEEE transactions on pattern analysis and machine intelligence **40** (2017) 2037–2050
13. Belmonte, R., Ihaddadene, N., Tirilly, P., Bilasco, I.M., Djeraba, C.: Video-based face alignment with local motion modeling. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE (2019) 2106–2115
14. Guo, M., Lu, J., Zhou, J.: Dual-agent deep reinforcement learning for deformable face tracking. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 768–783
15. Dong, X., Yu, S.I., Weng, X., Wei, S.E., Yang, Y., Sheikh, Y.: Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 360–368
16. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
17. Gao, P., Lu, K., Xue, J.: Efficientfan: Deep knowledge transfer for face alignment. In: Proceedings of the 2020 International Conference on Multimedia Retrieval. (2020) 215–223
18. Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., Anandkumar, A.: Born-again neural networks. In: International Conference on Machine Learning. (2018) 1602–1611
19. Bagherinezhad, H., Horton, M., Rastegari, M., Farhadi, A.: Label refinery: Improving imagenet classification through label progression. arXiv preprint arXiv:1805.02641 (2018)
20. Kato, N., Li, T., Nishino, K., Uchida, Y.: Improving multi-person pose estimation using label correction. arXiv preprint arXiv:1811.03331 (2018)
21. Honari, S., Molchanov, P., Tyree, S., Vincent, P., Pal, C., Kautz, J.: Improving landmark localization with semi-supervised learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 1546–1555
22. Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: Database and results. Image and vision computing **47** (2016) 3–18
23. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2013) 397–403

24. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: A semi-automatic methodology for facial landmark annotation. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. (2013) 896–903
25. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: CVPR 2011, IEEE (2011) 545–552
26. Ramanan, D., Zhu, X.: Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE conference on computer vision and pattern recognition, IEEE (2012) 2879–2886
27. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: European conference on computer vision, Springer (2012) 679–692
28. Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G.: Xm2vtsdb: The extended m2vts database. In: Second international conference on audio and video-based biometric person authentication. Volume 964. (1999) 965–966
29. Zhu, S., Li, C., Change Loy, C., Tang, X.: Face alignment by coarse-to-fine shape searching. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 4998–5006
30. Chrysos, G.G., Antonakos, E., Zafeiriou, S., Snape, P.: Offline deformable face tracking in arbitrary videos. In: Proceedings of the IEEE international conference on computer vision workshops. (2015) 1–9
31. Shen, J., Zafeiriou, S., Chrysos, G.G., Kossaifi, J., Tzimiropoulos, G., Pantic, M.: The first facial landmark tracking in-the-wild challenge: Benchmark and results. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2015) 50–58
32. Tzimiropoulos, G.: Project-out cascaded regression with an application to face alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 3659–3667
33. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
34. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 1021–1030
35. Tokui, S., Okuta, R., Akiba, T., Niitani, Y., Ogawa, T., Saito, S., Suzuki, S., Uenishi, K., Vogel, B., Yamazaki Vincent, H.: Chainer: A deep learning framework for accelerating the research cycle. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM (2019) 2002–2011
36. Tokui, S., Oono, K., Hido, S., Clayton, J.: Chainer: a next-generation open source framework for deep learning. In: Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS). (2015)
37. Akiba, T., Fukuda, K., Suzuki, S.: ChainerMN: Scalable Distributed Deep Learning Framework. In: Proceedings of Workshop on ML Systems in The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS). (2017)
38. Dong, X., Yan, Y., Ouyang, W., Yang, Y.: Style aggregated network for facial landmark detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 379–388
39. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence (2020)