Exploiting Self-Supervised and Semi-Supervised Learning for Facial Landmark Tracking with Unlabeled Data

Shi Yin[1], Shangfei Wang*[1,2], Xiaoping Chen[2] and Enhong Chen[3]
davidyin@mail.ustc.edu.cn,{sfwang,xpchen,cheneh}@ustc.edu.cn
[1]Key Lab of Computing and Communication Software of Anhui Province,
School of Computer Science and Technology, University of Science and Technology of China
[2]Anhui Robot Technology Standard Innovation Base, University of Science and Technology of China
[3]Anhui Province Key Lab of Big Data Analysis and Application,
School of Computer Science and Technology, University of Science and Technology of China

ABSTRACT

Current work of facial landmark tracking usually requires large amounts of fully annotated facial videos to train a landmark tracker. To relieve the burden of manual annotations, we propose a novel facial landmark tracking method that makes full use of unlabeled facial videos by exploiting both self-supervised and semi-supervised learning mechanisms. First, self-supervised learning is adopted for representation learning from unlabeled facial videos. Specifically, a facial video and its shuffled version are fed into a feature encoder and a classifier. The feature encoder is used to learn visual representations, and the classifier distinguishes the input videos as the original or the shuffled ones. The feature encoder and the classifier are trained jointly. Through self-supervised learning, the spatial and temporal patterns of a facial video are captured at representation level. After that, the facial landmark tracker, consisting of the pre-trained feature encoder and a regressor, is trained semi-supervisedly. The consistencies among the tracking results of the original, the inverse and the disturbed facial sequences are exploited as the constraints on the unlabeled facial videos, and the supervised loss is adopted for the labeled videos. Through semi-supervised end-to-end training, the tracker captures sequential patterns inherent in facial videos despite small amount of manual annotations. Experiments on two benchmark datasets show that the proposed framework outperforms state-of-the-art semi-supervised facial landmark tracking methods, and also achieves advanced performance compared to fully supervised facial landmark tracking methods.

CCS CONCEPTS

• **Computing methodologies → Biometrics**.

KEYWORDS

Facial landmark tracking, self-supervised learning, semi-supervised learning

ACM Reference Format:
Shi Yin[1], Shangfei Wang*[1,2], Xiaoping Chen[2] and Enhong Chen[3]. 2020. Exploiting Self-Supervised and Semi-Supervised Learning for Facial Landmark Tracking with Unlabeled Data. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3394171.3413547

1 INTRODUCTION

Facial landmark localization is crucial for face analysis tasks, such as face recognition [13], facial expression classification [10], facial action unit recognition [30] and face verification [12]. It can be divided into two sub-tasks, i.e., facial landmark detection on a static image, and landmark tracking in a dynamic video. The tracking task is more complex since a tracker has to integrate both spatial and temporal patterns existed in a video to make robust predictions.

Most works of facial landmark tracking [2, 19, 21, 25, 25, 26, 33–35, 38] adopt supervised learning approach, which requires a large amount of training videos fully annotated with landmarks frame by frame. Even for a short video clip lasting one minute with 30 frames a second and 68 landmarks a face, 122400 landmarks need annotating. As the number of training videos increases, the manual work for annotation is very expensive and time consuming, if not impossible.

To reduce the dependency on manually labeled landmarks, recently several works [7, 8, 11, 14, 22, 27, 29, 37] have tried to leverage unlabeled or partially labeled data for face alignment. However, these works have three disadvantages. First, some works only detect landmarks on a static image [7, 8, 11, 14, 22, 27, 29], totally ignoring temporal patterns, and others consider relations between two adjacent frames [37], ignoring the long-term sequential patterns existed in a facial video. Therefore, these methods have little benefit on sequential modeling and are sub-optimal for facial landmark tracking. Second, several works [11, 14, 27] require additional labels, such as the bounding boxes for facial areas [27], or expression labels and head poses [11]. These additional labels bring extra annotation burdens, although the landmark annotations are reduced. Third, few work exploits self-supervised mechanism to learn better representations for facial landmark tracking, while self-supervised learning has been proved effective in learning visual representations [4, 16, 18, 28].

*Corresponding author.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413547

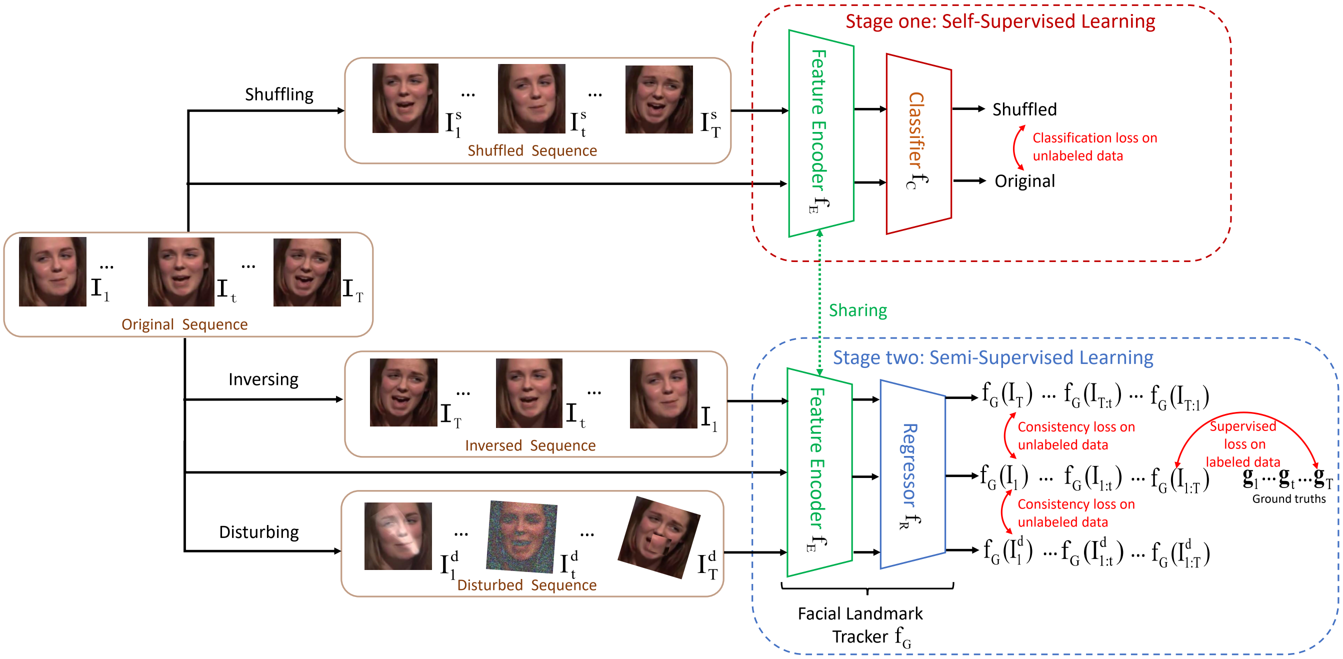

Stage one: Self-Supervised Learning
Shuffling
I_1^s ... I_t^s ... I_T^s
Shuffled Sequence
Feature Encoder f_E
Classifier f_C
Shuffled
Classification loss on unlabeled data
Original
I_1 ... I_t ... I_T
Original Sequence
Sharing
Stage two: Semi-Supervised Learning
Inversing
I_T ... I_t ... I_1
Inversed Sequence
Disturbing
I_1^d ... I_t^d ... I_T^d
Disturbed Sequence
Feature Encoder f_E
Regressor f_R
$f_G(I_T)$ ... $f_G(I_{T:t})$ ... $f_G(I_{T:1})$
Consistency loss on unlabeled data
Supervised loss on labeled data
$f_G(I_1)$ ... $f_G(I_{1:t})$ ... $f_G(I_{1:T})$
$g_1 \cdots g_t \cdots g_T$
Ground truths
Consistency loss on unlabeled data
$f_G(I_1^d)$ ... $f_G(I_{1:t}^d)$ ... $f_G(I_{1:T}^d)$
Facial Landmark Tracker f_G

Figure 1: The proposed two-stage learning framework. The first stage is self-supervised learning stage, as shown in the upper part of the figure. The second stage is semi-supervised learning stage, as shown in the lower part of the figure.

To address these, we propose a new framework combining self-supervised learning and semi-supervised learning in a two-stage learning paradigm to make full use of unlabeled data for facial landmark tracking, as shown in Fig. 1. The first stage is self-supervised learning stage, as depicted in the upper part of Fig. 1. In this stage, the feature encoder of the tracker is trained to extract order-aware sequential representations by a pretext classification task. Specifically, the long facial sequence and its shuffled version are fed into the feature encoder. Based on the extracted representations, a classifier distinguishes them as the original or the shuffled sequence. The feature encoder and the classifier are trained jointly by the classification loss. Such training process embeds informative temporal patterns into the extracted facial sequence representations. The second stage is semi-supervised learning stage. Based on the feature encoder pre-trained by self-supervised learning, the whole tracker, consisting of the feature encoder and a regressor, is trained on both labeled and unlabeled data. On the labeled data, the tracker is trained by label supervision. On the unlabeled data, we train the tracker by regression tasks guided by constraints on tracking consistency. Two consistency constraints are proposed. First, the landmark predictions from the original facial sequence should be as the same as those from the inverse one. Second, the tracking results on the original sequence and the sequence with disturbances should keep the geometric consistency between the two sequences. The disturbance operations include texture disturbances, i.e., occlusion, blurring, noises, illumination changes; and spatial transformation, i.e., translation, scaling and rotation. The first constraint forces the tracker to integrate the spatial and temporal patterns from both forward and backward directions, and captures ground truth patterns implicitly by the complementary information from the tracking results of two directions. The second constraint forces the tracker to capture the geometric consistency inherent in a facial sequence, and make robust predictions despite various disturbances that may exist "in the wild". Through the training process, diverse sequential patterns are captured despite the lack of manual annotations. Experimental results on the 300VW database and the TF database demonstrate the superiority of the proposed method.

2 RELATED WORK

2.1 Fully-Supervised Learning for Facial Landmark Tracking

A comprehensive survey of facial landmark tracking can be found in [5, 31]. The mainstream of facial landmark tracking is to train the tracker on fully labeled videos. Current works can be classified into two categories, i.e., modeling facial patterns from appearance level, or modeling landmark dependencies from target label level. The former directly maps facial appearances to landmark coordinates. For example, Xiong *et al.* [33] proposed a supervised decent method (SDM), which minimizes a nonlinear least square objective function to close the landmark prediction and the ground truth one. Cao *et al.* [2] designed a two-level boosted regression framework (ESR) to minimize the prediction errors in a cascaded way and capture the inherent shape dependency. Recently, deep neural models, such as convolutional neural networks (CNN) [25] or recurrent neural networks (RNN) [21], or the combination of CNN and RNN [17], are proposed to extract spatial and temporal patterns from facial appearances. The latter explicitly encodes the spatial and temporal constraints from target label level. For example, several probabilistic graphic models, such as Markov Random Field (MRF)

[6], Restricted Boltzmann Machine (RBM) [32] and Conditional Random Field (CRF) [3], are adopted to model the global constraints of shape deformation. Tai *et al.* [26] incorporated manually designed loss functions as constraints for shape deformation. Yin *et al.* [34] utilized adversarial learning to explore the inherent dependencies among the movement of facial landmarks. All these methods require fully labeled training data, which cost huge human labors. While the proposed approach fully explores the intrinsic temporal and spatial patterns in unlabeled facial videos, and reduce the dependency on labeled data.

2.2 Semi-Supervised Learning for Facial Landmark Detection and Tracking

To alleviate the dependency on labeled data, several semi-supervised learning methods are proposed for facial landmark detection [7, 8, 11, 14, 22, 27, 29] and tracking [37].

Early attempts [14, 29] of semi-supervised facial landmark detection are only adept at processing images captured under controlled conditions and could not well fits to “in-the-wild" images. Recently, Tang *et al.* [27] proposed to train a landmark detector semi-supervisedly on faces labeled with bounding boxes of facial components. Honari *et al.* [11] proposed to jointly predict several facial attributes, e.g., head poses and facial expressions as auxiliary tasks, which are beneficial to landmark detection. They also proposed to make landmark predictions equivariant to transformations through unsupervised learning on original-transformed image couples. Dong *et al.* proposed to generate pesudo landmark labels by optical flow [8] or a student network [20] as training samples for the detector. These methods are landmark detection methods, which predict landmarks on static images or frames and ignore sequence-level patterns in a video. Furthermore, some of them [11, 14, 27] require additional labels for training and brings extra burdens for annotation.

For facial landmark tracking, Zhu *et al.* [7] proposed a semi-supervised method with a circle loss produced by two adjacent frames. This method only considers temporal relations between two frames while the temporal dependencies in the long sequence are ignored.

To address these disadvantages, we propose a new semi-supervised learning strategy which trains the tracker by regression tasks from the consistency constraints on the long facial sequence instead of two adjacent frames, such that the long-term dependencies existed in a facial sequence are captured. The proposed semi-supervised learning strategy does not require any extra labels. Thus, large scale unlabeled data can be exploited for training.

2.3 Self-Supervised Learning for Learning Representation

Self-supervised learning [15] aims to learn good spatial or temporal representation by self-producing supervised signals on unlabeled data. Recently, several approaches have been proposed on related tasks. Lu *et al.* [18] proposed to model inter-frames correlation by an unsupervised attention mechanism for object segmentation in a video. Chen at al. [4] incorporated self-supervised learning with adversarial learning for image synthesis by randomly rotating an image and training the discriminator to predict the rotation angle together with the realness of an image. Thewlis et al. [28] proposed a self-supervised key point discovering method by objectives from transformation consistency. Kim et al. [16] proposed to permute 3D spatio-temporal crops extracted from a video clip to the correct arrangement for action recognition tasks. In our method, the feature encoder of the tracker learns sequential representations by a pretext task, i.e., distinguishing between the original facial sequence and the shuffled one. To the best of our knowledge, we are the first exploiting self-supervised learning for facial landmark tracking task.

3 METHODOLOGY

A facial landmark tracker $f_G(\cdot)$ usually consists of a feature encoder $f_E(\cdot)$ and a regressor $f_R(\cdot)$. $f_E(\cdot)$ encodes representations from facial appearances, while $f_R(\cdot)$ decodes these representations as landmark coordinates. Formally, $f_G(\cdot)$ predicts the coordinates of M predefined landmarks from a facial video with T frames, as shown in Eq. (1).

$$\begin{aligned}\mathbf{c}_t &= f_G(\mathbf{I}_{1:t};\theta_G)\\ &= f_R(f_E(\mathbf{I}_{1:t};\theta_E);\theta_R), 1 \le t \le T\end{aligned} \tag{1}$$

where $\mathbf{I}_{1:t} = (\mathbf{I}_1, ..., \mathbf{I}_t)$ is a facial sequence from the video stream, $\mathbf{I}_t (1 \le t \le T)$ is the t th frame of the sequence. $\mathbf{c}_t \in \mathbb{R}^{2M} (1 \le t \le T)$ is the concatenation of x, y coordinates for all M landmarks predicted on the t th frame. θ_E and θ_R denote the parameters of $f_E(\cdot)$ and $f_R(\cdot)$, respectively. $\theta_G = \{\theta_E, \theta_R\}$ denotes the parameters of the whole tracker.

If the ground truth landmark coordinate $\mathbf{g}_t (1 \le t \le T)$ is available, $f_G(\cdot)$ is trained by supervised regression typically, as shown in Eq. (2)

$$\min_{\theta_G} L_S = \sum_{t=1}^{T} ||f_G(\mathbf{I}_{1:t};\theta_G) - \mathbf{g}_t||_2^2 \tag{2}$$

However, the annotations for ground truths may not be sufficient due to the high labor costs for labeling. To address this, we propose a two-stage learning framework making full use of unlabeled data to train the tracker. The first learning stage is self-supervised learning, which trains $f_E(\cdot)$ on unlabeled data by a pretext classification task. The second learning stage is semi-supervised learning stage. Based on the pre-training of $f_E(\cdot)$ in the first stage, $f_E(\cdot)$ and $f_R(\cdot)$ are trained jointly on both labeled and unlabeled data by regression tasks. Through these classification and regression tasks, the tracker can learn the intrinsic spatial and temporal patterns existed in the long facial sequence from small scale labeled data and large scale unlabeled data.

3.1 Self-Supervised Learning Stage

To capture sequential patterns from the facial sequence, the feature encoder $f_E(\cdot)$ is trained to distinguish the original facial sequence from the shuffled one by temporal clues, e.g., the deformation smoothness of a face, existed in the facial sequence. Let the shuffled sequence denoted as $\mathbf{I}^s_{1:T} = g_s(\mathbf{I}_{1:T})$, where $g_s(\cdot)$ is the shuffling function. $f_E(\cdot)$ encodes sequential representations from $\mathbf{I}_{1:T}$ and $\mathbf{I}^s_{1:T}$, as shown in Eq. (3):

$$\mathbf{f}_{1:T} = f_E(\mathbf{I}_{1:T};\theta_E),\ \ \mathbf{f}^s_{1:T} = f_E(\mathbf{I}^s_{1:T};\theta_E) \tag{3}$$

where $\mathbf{f}_{1:T} = (\mathbf{f}_1, ..., \mathbf{f}_t, ..., \mathbf{f}_T)$ and $\mathbf{f}^s_{1:T} = (\mathbf{f}^s_1, ..., \mathbf{f}^s_t, ..., \mathbf{f}^s_T)$ are the representations extracted from each frame of the two sequences, respectively. The encoded representations are fed into a binary classifier, denoted as $f_C(\cdot)$, which classifies $\mathbf{f}_{1:T}$ as 0 and $\mathbf{f}^s_{1:T}$ as 1. $f_E(\cdot)$ and $f_C(\cdot)$ are trained jointly by the self-supervised classification loss in Eq. (4):

$$\begin{aligned}\min_{\theta_E,\theta_C} L_C &= -(log(1-\sigma(f_C(\mathbf{f}_{1:T};\theta_C))) + log(\sigma(f_C(\mathbf{f}^s_{1:T};\theta_C))))\\ &= -(log(1-\sigma(f_C(f_E(\mathbf{I}_{1:T};\theta_E);\theta_C)))\\ &\quad + log(\sigma(f_C(f_E(g_s(\mathbf{I}_{1:T});\theta_E);\theta_C))))\end{aligned} \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function. Through self-supervised learning, $f_E(\cdot)$ can encode informative representations from the facial sequence, thus the sequential modeling capability is promoted.

3.2 Semi-Supervised Learning Stage

In the semi-supervised learning stage, the tracker is trained on both labeled and unlabeled data. For the labeled data, the tracker is trained by label supervision, as shown in Eq. (2). For the unlabeled data, we train the tracker by two consistency constraints.

3.2.1 *Consistency Constraint from the Original and the Inverse Sequence.*

The tracker is trained by the constraint that the tracking results should be invariant between the original facial sequence and its inverse sequence. The tracker firstly tracks landmarks on the original sequence, i.e., $\mathbf{I}_{1:T} = (\mathbf{I}_1, ..., \mathbf{I}_T)$. For the $t(1 \leq t \leq T)$ th frame, i.e., $\mathbf{I}_t$, the tracker locates its landmarks based on facial appearances from the current and the previous frames, i.e., $\mathbf{I}_{1:t} = (\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_t)$. The tracking result is represented as $f_G(\mathbf{I}_{1:t};\theta_G)$. Then, the tracker tracks landmarks on the inverse sequence , i.e., $\mathbf{I}_{T:1} = (\mathbf{I}_T, ..., \mathbf{I}_1)$. For $\mathbf{I}_t$, the tracker locates landmarks based on facial appearances from the current and the following frames, i.e., $\mathbf{I}_{T:t} = (\mathbf{I}_T, \mathbf{I}_{T-1}, ..., \mathbf{I}_t)$. The tracking result is represented as $f_G(\mathbf{I}_{T:t};\theta_G)$. $f_G(\mathbf{I}_{1:t};\theta_G)$ and $f_G(\mathbf{I}_{T:t};\theta_G)$ are expected to be the same. The difference between $f_G(\mathbf{I}_{1:t};\theta_G)$ and $f_G(\mathbf{I}_{T:t};\theta_G)$, is adopted as a training loss, as shown in Eq. (5).

$$\min_{\theta_G} L_I = \sum_{t=1}^{T} ||f_G(\mathbf{I}_{1:t};\theta_G) - f_G(\mathbf{I}_{T:t};\theta_G)||_2^2 \tag{5}$$

By Eq. (5), the tracker integrates the spatial and temporal patterns from both forward and backward directions, and label patterns are captured implicitly by the complementary information from the tracking results of two directions.

3.2.2 *Consistency Constraint from the Original and the Disturbed Sequence.*

The inherent geometric consistency between the landmark coordinates of the original sequence $\mathbf{I}_{1:T}$ and its disturbed sequence $\mathbf{I}^d_{1:T}$ is another constraint to train the tracker. To build a disturbed sequence, a combination of the texture disturbance operations and spatial transformation operations are applied on the original sequence with a probability of δ that each operation is implemented. As shown in Fig. 2, texture disturbance operations include occlusion, blurring, noises and illumination changes, while spatial transformation operations include translation, rotation and scaling. An “in-the-wild" facial video usually deforms or zooms gradually and smoothly without sharp changes. Such smooth constraint should also be followed when applying transformation on the sequence, such that the trained tracker well fits to real world testing data. Therefore, the translating displacement Δ_t, the rotating angle ω_t, the scaling ratio r_t of the face in the t th fame, are constrained that:

$$\begin{aligned}|\Delta_t - \Delta_{t-1}| &\leq \alpha_1 * \sqrt{WH}\\ |\omega_t - \omega_{t-1}| &\leq \alpha_2 * \pi\\ |r_t - r_{t-1}| &\leq \alpha_3\end{aligned} \tag{6}$$

where W and H are the width and height of the facial bouding box, respectively. α_1, α_2 and α_3 are hyper-parameters to adjust the smoothness of face movement.

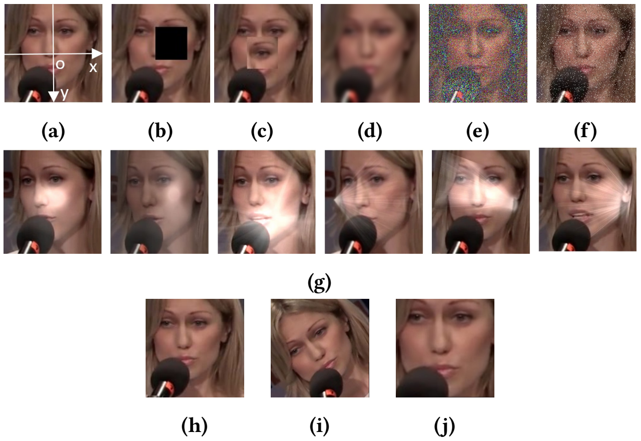


Figure 2: (a) The original face. (b) Occlusion by black. (c) Occlusion by part of the face. (d) Blurring. (e) Gaussian noises. (f) Salt noises. (g) Faces with changed illumination conditions. (h) Translation. (i) Rotation. (j) Shape Scaling.

Let $g_{te}(\cdot)$ be a uniform representation for texture disturbance operations, and $g_{tr}(\cdot)$ represent the spatial transformation operations. $g_{te}(\cdot)$ does not change the landmark positions, while $g_{tr}(\cdot)$ may convert the landmarks to new coordinates. The consistency constraints between landmark predictions of the original sequence and the disturbed sequence can be formalized as the objective function shown in Eq. (7):

$$\begin{aligned}\min_{\theta_G} L_D &= \sum_{t=1}^{T} ||f_G(\mathbf{I}_{1:t};\theta_G) - g_{tr}^{-1}(f_G(g_{tr}(g_{te}(\mathbf{I}_{1:t}));\theta_G))||_2^2\\ &= \sum_{t=1}^{T} ||f_G(\mathbf{I}_{1:t};\theta_G) - g_{tr}^{-1}(f_G(\mathbf{I}^d_{1:t};\theta_G))||_2^2\end{aligned} \tag{7}$$

where $g_{tr}^{-1}(\cdot)$ denotes the inverse operation of $g_{tr}(\cdot)$. After training by Eq. (7), the tracker can integrate representations from multiple frames to track the landmarks despite polluted facial textures and extreme head poses that may exist "in the wild".

Please note that the above data disturbance techniques are also frequently used in data augmentation. However, unlike data augmentation, which extends labeled data in need of their ground truths labels, the proposed semi-supervised learning strategy leverages the consistency between the tracking results of the original facial sequence and its disturbed sequence to explore the spatial and temporal patterns existed in unlabeled data.

3.3 Overall Loss Function

Suppose there are M labeled training videos and N unlabeled videos, loss functions from the self-supervised and semi-supervised learning stages are combined together as an overall loss function, as shown in Eq. (8):

$$\begin{aligned}\min L_o = \min_{\theta_E,\theta_C} \frac{1}{N}\sum_{n=1}^{N}\lambda_C \cdot L_C \\ + \min_{\theta_G}(\frac{1}{M}\sum_{m=1}^{M}\lambda_S \cdot L_S + \frac{1}{N}\sum_{n=1}^{N}(\lambda_I L_I + \lambda_D L_D))\end{aligned} \tag{8}$$

The first and second items of Eq. (8) correspond to the self-supervised loss and the semi-supervised loss, respectively. λ_C, λ_S, λ_I, and λ_D are four weighting hyper-parameters.

4 EXPERIMENTS

4.1 Experimental Conditions

The 300 Videos in the Wild (300VW) [24] is chosen as the training video set, which contains 50 training videos, a total of 95192 frames. Each frame is annotated with 68 pre-defined landmarks. The proposed method is evaluated on the 300VW testing set and the Talking Face (TF) [9] dataset. The 300VW testing set contains 60 videos from three challenging levels, i.e., well-lit (Scenario 1), mild unconstrained (Scenario 2) and challenging (Scenario 3). In the following parts of the paper, these three scenarios are simplified as S1, S2 and S3, respectively. The TF dataset contains one video of 5000 frames from a talking person. Due to the different landmark definitions between the TF and the 300VW dataset, on the TF dataset, only seven landmarks in common are applied for testing, as previous works [17, 26] did.

The performance of the tracker is evaluated by the tracking accuracy and stability. Accuracy reflects the closeness of the predicted landmark coordinates to the ground truths. Stability reflects the moving consistency between predicted landmarks and ground truths. A tracking result with good stability usually implies smooth landmark predictions without sharp jumps. NRMSE and AUC@0.08 are adopted as the accuracy metrics. NRMSE is the *N*ormalized *R*oot *M*ean *S*quared *E*rror between the predicted landmark coordinates and the ground truths. AUC@0.08 is defined as the *A*rea *U*nder the error *C*urve calculated for a threshold of 0.08. The stability metric is defined as the error of landmark displacement between the tracking results and the ground truths. A formal definition of these metrics can be found in previous papers [23, 26]. A lower NRMSE, higher AUC@0.08 and lower stability metric correspond to better accuracy and stability, respectively.

The optimal values for all hyper-parameters, i.e., δ, λ_C, λ_S, λ_I, λ_D, α_1, α_2 and α_3, are determined by 10-fold cross validations on the 300VW training set. Their optimal values are $\delta^* = 0.4$, $\lambda_C^* = 0.7$, $\lambda_S^* = 0.8$, $\lambda_I^* = 0.5$, $\lambda_D^* = 0.6$, $\alpha_1^* = 0.02$, $\alpha_2^* = 0.03$, and $\alpha_3^* = 0.02$. After cross validation, we assign these hyper-parameters with their optimal values and re-train the tracker on the whole training set. The training loss is optimized by Adam optimizer with a learning rate of $1e-4$.

To evaluate the proposed framework on different ratios of labeled training data, we randomly drop the labels of some training video frames in the 300VW dataset as unlabeled samples. A training video is then splitted as a labeled video and an unlabeled video. The ratio of remaining labeled data to the total is denoted as γ. As shown in Table 1, γ is set as 2%, 5%, 10%, 25%, 50% in turn. Since randomly dropping labels may cause randomness on the experimental results, we conduct ten repeated experiments under the same conditions. In table 1, average values of the NRMSE and stability performance from the ten experiments are shown outside the bracket, while their standard deviations are shown inside the bracket. For each γ, we make ablation study on every individual training loss, i.e., the self-supervised loss L_C and the semi-supervised loss L_I and L_D, by setting their weight $\lambda \in \{\lambda_C, \lambda_I, \lambda_D\}$ in Eq. (8) as 0 and their optimal values $(\lambda_C^*, \lambda_I^*, \lambda_D^*)$ in turn. When λ_C, λ_I, λ_D are all assigned as 0.0, the tracker is only trained on the labeled data by supervised learning. When λ_C, λ_I, λ_D are all assigned as their optimal values, the tracker is trained on both labeled and unlabeled data by the proposed framework.

4.2 Implementation Details

4.2.1 Details for Texture Disturbance Operations. The texture disturbance operations include occlusion, blurring, noises and illumination changes. Two types of occlusions are used. The first one is occlusion by black, i.e., the occluded area is covered by black, as shown in Fig. 2b. The second one is occlusion by part of the face. As shown in Fig. 2c, part of the nose of the woman is occluded by the facial area copied from the eye. The second category of occlusion is more challenging, since the appearance feature from the copied area of the face may extremely disturb the tracker. For example, there are three eyes in Fig. 2c and the tracker has to integrate previous frames by temporal relations to decide the true location of the eyes. Fig. 2d shows Gaussian blurring on the face with a random size of Gaussian kernel. Fig. 2e and 2f show Gaussian and salt noises on the face, respectively. We imitate different light sources and superpose them with the original face as a new illumination sample, as shown in Fig. 2g. The position, orientation and intensity of the light source are randomly determined, making more diverse illumination conditions.

4.2.2 Structure of the tracker. Following Yin et al. [34], the feature encoder $f_E(\cdot)$ consists of a stacked hourglass network and a CNN network. The stacked hourglass network encodes each facial frame as a high-dimensional representation, while the CNN network compresses the facial representation as a feature vector. We choose a two-layers LSTM as the regressor $f_R(\cdot)$. $f_R(\cdot)$ takes the feature vectors from the facial frame sequence as the input, and predicts the landmark coordinates for each frame. Faces in a video are firstly cropped from the bounding box, then scaled to 256 × 256 pixels as the input for the tracker. The self-supervised classifier $f_C(\cdot)$ is instantiated as a two-layer bidirectional LSTM (BLSTM) followed by a fully-connected (FC) network. The FC network converts the average of the BLSTM hidden states to the likelihood that a sequence is shuffled.

γ	weights	300VW S1		300VW S2		300VW S3		TF	
		N	S	N	S	N	S	N	S
2%	$\lambda_C = 0.0, \lambda_I = 0.0, \lambda_D = 0.0$	8.11 (0.45)	2.58 (0.22)	7.01 (0.36)	1.87 (0.22)	8.09 (0.42)	4.88 (0.29)	3.95 (0.18)	1.03 (0.12)
	$\lambda_C = \lambda_C^*, \lambda_I = 0.0, \lambda_D = 0.0$	7.62 (0.32)	2.27 (0.17)	6.56 (0.32)	1.72 (0.20)	7.25 (0.34)	4.63 (0.27)	3.40 (0.17)	0.98 (0.07)
	$\lambda_C = \lambda_C^*, \lambda_I = \lambda_I^*, \lambda_D = 0.0$	6.90 (0.28)	1.85 (0.14)	6.14 (0.29)	1.66 (0.19)	6.44 (0.28)	4.53 (0.30)	3.12 (0.19)	0.95 (0.08)
	$\lambda_C = \lambda_C^*, \lambda_I = \lambda_I^*, \lambda_D = \lambda_D^*$	5.03 (0.27)	1.46 (0.12)	5.12 (0.25)	1.39 (0.18)	5.52 (0.24)	2.90 (0.21)	2.46 (0.16)	0.83 (0.07)
5%	$\lambda_C = 0.0, \lambda_I = 0.0, \lambda_D = 0.0$	6.24 (0.34)	1.84 (0.26)	6.31 (0.35)	1.69 (0.28)	6.76 (0.47)	3.43 (0.32)	3.01 (0.18)	0.97 (0.13)
	$\lambda_C = \lambda_C^*, \lambda_I = 0.0, \lambda_D = 0.0$	5.92 (0.31)	1.72 (0.23)	5.96 (0.34)	1.41 (0.25)	6.60 (0.42)	3.02 (0.29)	2.74 (0.17)	0.88 (0.10)
	$\lambda_C = \lambda_C^*, \lambda_I = \lambda_I^*, \lambda_D = 0.0$	5.42 (0.33)	1.59 (0.21)	5.20 (0.31)	1.28 (0.14)	6.25 (0.37)	2.84 (0.28)	2.58 (0.12)	0.80 (0.09)
	$\lambda_C = \lambda_C^*, \lambda_I = \lambda_I^*, \lambda_D = \lambda_D^*$	4.64 (0.28)	1.24 (0.18)	4.55 (0.27)	0.98 (0.07)	5.29 (0.34)	2.44 (0.25)	2.30 (0.09)	0.67 (0.05)
10%	$\lambda_C = 0.0, \lambda_I = 0.0, \lambda_D = 0.0$	5.14 (0.32)	1.74 (0.27)	5.31 (0.38)	1.45 (0.20)	5.62 (0.40)	2.97 (0.31)	2.67 (0.25)	0.92 (0.12)
	$\lambda_C = \lambda_C^*, \lambda_I = 0.0, \lambda_D = 0.0$	4.95 (0.26)	1.51 (0.23)	5.13 (0.32)	1.32 (0.15)	5.47 (0.25)	2.76 (0.26)	2.50 (0.23)	0.91 (0.08)
	$\lambda_C = \lambda_C^*, \lambda_I = \lambda_I^*, \lambda_D = 0.0$	4.62 (0.23)	1.43 (0.18)	4.70 (0.26)	1.15 (0.13)	5.38 (0.26)	2.53 (0.24)	2.46 (0.18)	0.89 (0.07)
	$\lambda_C = \lambda_C^*, \lambda_I = \lambda_I^*, \lambda_D = \lambda_D^*$	4.12 (0.19)	1.12 (0.10)	4.07 (0.22)	0.96 (0.05)	4.98 (0.18)	2.13 (0.13)	2.11 (0.14)	0.64 (0.04)
25%	$\lambda_C = 0.0, \lambda_I = 0.0, \lambda_D = 0.0$	4.57 (0.28)	1.42 (0.24)	4.72 (0.35)	1.20 (0.23)	5.48 (0.39)	2.52 (0.20)	2.16 (0.15)	0.79 (0.06)
	$\lambda_C = \lambda_C^*, \lambda_I = 0.0, \lambda_D = 0.0$	4.44 (0.26)	1.34 (0.20)	4.23 (0.22)	1.12 (0.20)	5.26 (0.35)	2.41 (0.18)	2.14 (0.12)	0.74 (0.05)
	$\lambda_C = \lambda_C^*, \lambda_I = \lambda_I^*, \lambda_D = 0.0$	4.34 (0.23)	1.31 (0.19)	4.20 (0.25)	1.05 (0.12)	4.93 (0.28)	2.18 (0.15)	2.10 (0.10)	0.68 (0.05)
	$\lambda_C = \lambda_C^*, \lambda_I = \lambda_I^*, \lambda_D = \lambda_D^*$	4.01 (0.15)	0.94 (0.07)	3.77 (0.20)	0.91 (0.08)	4.52 (0.23)	1.73 (0.11)	2.03 (0.06)	0.61 (0.04)
50%	$\lambda_C = 0.0, \lambda_I = 0.0, \lambda_D = 0.0$	4.02 (0.22)	1.21 (0.17)	4.04 (0.24)	1.13 (0.09)	4.95 (0.32)	2.14 (0.12)	2.09 (0.04)	0.69 (0.04)
	$\lambda_C = \lambda_C^*, \lambda_I = 0.0, \lambda_D = 0.0$	3.93 (0.19)	1.17 (0.14)	3.88 (0.23)	1.06 (0.07)	4.83 (0.28)	1.96 (0.11)	2.08 (0.03)	0.66 (0.04)
	$\lambda_C = \lambda_C^*, \lambda_I = \lambda_I^*, \lambda_D = 0.0$	3.76 (0.18)	0.97 (0.07)	3.75 (0.22)	0.97 (0.07)	4.77 (0.26)	1.68 (0.10)	2.05 (0.02)	0.64 (0.03)
	$\lambda_C = \lambda_C^*, \lambda_I = \lambda_I^*, \lambda_D = \lambda_D^*$	3.34 (0.14)	0.70 (0.05)	3.51 (0.18)	0.75 (0.06)	4.33 (0.16)	1.37 (0.08)	2.01 (0.02)	0.58 (0.03)

Table 1: The NRMSE (N) and stability (S) performance under different parameter settings. Average values of these metrics from ten repeated experiments are shown outside the bracket, while their standard deviations are shown inside the bracket. γ is the ratio of labeled data to the whole data. λ_C, λ_I, and λ_D are the weights of the self-supervised loss L_C and the semi-supervised losses L_I and L_D, respectively. λ_C^*, λ_I^*, and λ_D^* are their optimal values.

4.3 Experimental Results and Analysis of Facial Landmark Tracking on Different Ratios of Labeled Data

From Table 1, we have the following observations.

First, compared to supervised learning ($\lambda_C = 0.0$, $\lambda_I = 0.0$, $\lambda_D = 0.0$) which only trains the tracker on the remaining labeled data, the proposed learning framework ($\lambda_C = \lambda_C^*, \lambda_I = \lambda_I^*, \lambda_D = \lambda_D^*$) significantly promotes the tracking accuracy and stability under the same amount of labeled data. For example, when γ is 10%, our framework decreases the average NRMSE by 19.84%, 23.35%, 11.39% and 20.97% on the 300VW S1, S1, S3 and the TF dataset, respectively; and decreases the average value of the stability metric by 34.88%, 33.79%, 28.28% and 30.43% on the respective datasets. From another perspective, our framework can achieve comparable performance with less labeled data. For example, when γ = 2%, the tracker trained by our framework outperforms the tracker trained by supervised learning when γ = 10%. Furthermore, the tracker trained by the proposed framework achieves smaller standard deviations from repeated experiments compared to supervised learning, which means our framework brings a better convergence. Supervised learning only trains the tracker on labeled data, while the proposed framework can make good use of unlabeled data to enhance facial sequence modeling, and well alleviate the sparsity of labeled data. Therefore, our framework can achieve better accuracy and stability performances.

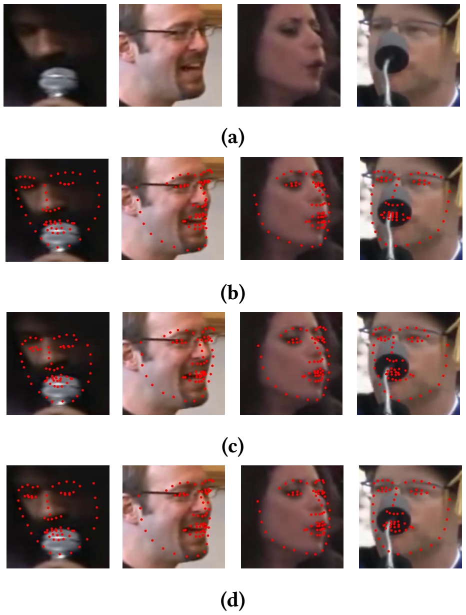
Second, we find each of the proposed training losses, i.e., self-supervised loss L_C and semi-supervised loss L_I and L_D, contributes to the overall performance. When we set their weights as the optimal values, the average value of NRMSE and the stability metric is smaller than the case when each of the weights is set as 0.

Third, from these training losses, L_D contributes the most by bringing the largest decrease on the average value of NRMSE and the stability metric. This may be because that the disturbed sequences contain more abundant patterns through diverse forms of disturbance operations, making the tracker more robust to challenging conditions.

We sample some challenging frames from the 300VW dataset and visualize their tracking results in Fig. 3. From Fig. 3, we find that the tracker trained by our framework makes more accurate predictions than the tracker trained by supervised learning when labeled data are limited.

4.4 Comparison with Semi-Supervised Detection Methods

We compare our method with state-of-the-art semi-supervised detection methods, including RCN$^+$ [11], SBR [8] and TS^3 [7]. We follow the experimental conditions of Dong et al. [7] to train our approach on the labeled 300W [23] dataset and unlabeled 300VW dataset, then evaluate it on the 49 inner landmarks of 300VW S3 by AUC@0.08. There are 3148 labeled images in the 300W training set and 95192 frames in the 300VW training set, which means the ratio (γ) of labeled training data to the whole is 3%. Since the 300W dataset is an image dataset with no temporal information, each image sample is taken as a one-frame video to pre-train the tracker.


(a)

(b)

(c)

(d)

Figure 3: (a) Some testing frames from 300VW S3. (b) Ground truths. (c) Tracking results of supervised learning (γ = 10%). (d) Tracking results of the proposed learning framework (γ = 10%).

After pre-training, the whole tracker is trained on the unlabeled 300VW training set by the proposed learning framework. The comparison is shown in Table 2. Results of the compared methods are directly copied from their papers, except for RCN$^+$. Since there are no published results of RCN$^+$ on video datasets, we just re-implement it under the same training condition of ours and report the result.

From Table 2, our framework outperforms all these semi-supervised learning methods. This may be because that all these methods detect landmarks on static images or separated frames and ignore the temporal information of the face, while our framework enhances the tracker at sequential modeling by the self-supervised classification task and semi-supervised learning regression tasks, making the tracker more adept at spatial and temporal modeling from a facial sequence.

4.5 Comparison with Semi-Supervised Tracking Methods

We compare our method with STRRN [37], a semi-supervised learning based tracking method, which trains the tracker by the tracking consistency within a circle composed of two adjacent frames. The NRMSE performance of our method and STRRN under different labeled data ratios (γ) are listed in Table 3, where the results of STRRN are directly copied from Zhu et al.'s work [37].

From Table 3, we find that our method outperforms STRRN when γ is 25% and 50%, and achieves comparable performance when γ is 10%. We may conclude that our method has a better overall performance than STRRN. The reason may be that STRRN only considers temporal relations between two adjacent frames, while the proposed method builds classification and regression tasks on the long facial sequence to fully integrate sequential patterns from the facial video.

Dataset	γ = 3%			
	RCN$^+$	SBR	TS^3	Ours
300VW S3	58.81	59.39	59.65	**60.80**

Table 2: Comparison on AUC@0.08 performance between our framework and semi-supervised detection methods.

Dataset	γ = 10%		γ = 25%		γ = 50%	
	STRRN	Ours	STRRN	Ours	STRRN	Ours
300VW S1	4.67	4.12	4.49	4.01	4.21	**3.34**
300VW S2	4.00	4.07	4.05	3.77	4.18	**3.51**
300VW S3	5.93	4.98	5.88	4.52	5.16	**4.33**

Table 3: Comparison on the average NRMSE performance between our framework and STRRN under different ratios (γ) of labeled data.

4.6 Comparison with Fully-Supervised Learning Methods

The proposed framework is also compared with state-of-the-art fully-supe-rvised learning methods. For comparison with fully-supervised learning methods, the tracker is trained on partially labeled 300VW training set (γ = 10% or γ = 50%). The compared methods include facial landmark detection methods, i.e., CFSS [38], TCDCN [36], FAN [1], DSRN [19], FHR [26]; and facial landmark tracking methods, i.e., SDM [33], TSCN [25], TSTN [17], STA [26], and GAN [34]. These methods are trained on fully annotated 300VW dataset. Comparisons on tracking accuracy (NRMSE) and stability are shown in Tables 4. Results of the compared methods are directly copied from Yin et al. [34].

From Tables 4, when γ = 10% that only 10% training data are labeled, the proposed framework outperforms most fully supervised learning method. This result demonstrates that the proposed approach can achieve advanced performance even with a dramatic label reduction. This is because our learning framework can capture the intrinsic temporal and spatial patterns from unlabeled data, and therefore reducing the requirement for labels significantly. When γ = 50%, the proposed approach outperforms all the compared works. It outperforms GAN, which performs the best among the compared methods on both accuracy and stability. From Table 4 , our framework brings an NRMSE decrease by 6.86%, 4.36%, 5.87% and 0.99% on the 300VW S1, S2, S3 and the TF dataset, respectively. With respective to stability, our framework also outperforms GAN with a decrease of the stability metric by 21.35%, 10.71%, 24.73 and 5.08% on these testing datasets.

5 CONCLUSION

Facing the huge cost to obtain labeled training data for facial landmark tracking, we propose a new method combining self-supervised and semi-supervised learning in a two-stage learning framework to make thorough use of unlabeled video data. In the self-supervised

Dataset	SDM		TSCN		CFSS		TCDCN		FAN		TSTN		DSRN		FHR		FHR+STA		GAN		Ours			
																					$\gamma = 10\%$		$\gamma = 50\%$	
	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S
300VW S1	7.41	-	12.54	-	7.68	-	7.66	-	5.58	-	5.36	-	5.33	-	4.82	2.67	4.21	1.58	3.50	0.89	4.12	1.12	**3.34**	**0.70**
300VW S2	6.18	-	7.25	-	6.42	-	6.77	-	4.87	-	4.51	-	4.92	-	4.23	1.77	4.02	1.09	3.67	0.84	4.07	0.96	**3.51**	**0.75**
300VW S3	13.04	-	13.13	-	13.67	-	14.98	-	7.75	-	12.84	-	8.85	-	7.09	4.43	5.64	2.62	4.43	1.82	4.98	2.13	**4.33**	**1.37**
TF	4.01	-	-	-	2.36	-	-	-	2.31	-	2.13	-	-	-	2.07	0.97	2.10	0.69	2.03	0.59	2.11	0.64	**2.01**	**0.58**

Table 4: Comparison on the average NRMSE (N) and stability (S) performance between our framework and fully-supervised leaning methods.

learning stage, supervised signals are self-produced from a facial sequence and its shuffled sequence in a binary classification task. In the semi-supervised learning stage, the tracker is trained by consistency constraints on tracking results of a facial sequence and its temporally inverse sequence as well as its disturbed sequence. Experimental results on the 300VW and the TF dataset show the superiority of the proposed framework over other semi-supervised learning methods and fully supervised learning methods.

ACKNOWLEDGMENTS

This work has been supported by the National Key R & D program of China (2018YFB1307102), and the National Natural Science Foundation of China (61727809).

REFERENCES

[1] Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *ICCV*. 1021–1030.

[2] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. 2014. Face alignment by explicit shape regression. *IJCV* 107, 2 (2014), 177–190.

[3] Lisha Chen, Hui Su, and Qiang Ji. 2019. Deep Structured Prediction for Facial Landmark Detection. In *NeurIPS*.

[4] Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. 2019. Self-Supervised GANs via Auxiliary Rotation Loss. In *CVPR*. 12154–12163.

[5] Grigorios G Chrysos, Epameinondas Antonakos, Patrick Snape, Akshay Asthana, and Stefanos Zafeiriou. 2018. A comprehensive performance evaluation of deformable face tracking "In-the-Wild". *IJCV* 126, 2-4 (2018), 198–232.

[6] Serhan Cosar and Mujdat Cetin. 2011. A graphical model based solution to the facial feature point tracking problem. *Image Vision Comput.* 29 (2011), 335–350.

[7] Xuanyi Dong and Yi Yang. 2019. Teacher Supervises Students How to Learn From Partially Labeled Images for Facial Landmark Detection. In *ICCV*.

[8] Xuanyi Dong, Shoou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang, and Yaser Sheikh. 2018. Supervision-by-Registration: An Unsupervised Approach to Improve the Precision of Facial Landmark Detectors. In *CVPR*. 360–368.

[9] FGNET. 2014. Talking Face Video. http://www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/talking_face.html.

[10] Quan Gan, Siqi Nie, Shangfei Wang, and Qiang Ji. 2017. Differentiating Between Posed and Spontaneous Expressions with Latent Regression Bayesian Network. In *AAAI*. 4039–4045.

[11] Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher J. Pal, and Jan Kautz. 2018. Improving Landmark Localization With Semi-Supervised Learning. In *CVPR*. 1546–1555.

[12] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. 2014. Discriminative Deep Metric Learning for Face Verification in the Wild. In *CVPR*. 1875–1882.

[13] Zhiwu Huang, Xiaowei Zhao, Shiguang Shan, Ruiping Wang, and Xilin Chen. 2013. Coupling alignments with recognition for still-to-video face recognition. In *ICCV*. 3296–3303.

[14] Xuhui Jia, Heng Yang, Kwok-Ping Chan, and ioannis Patras. 2014. Structured Semi-supervised Forest for Facial Landmarks Localization with Face Mask Reasoning. In *BMVC*.

[15] L. Jing and Y. Tian. 2020. Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. *TPAMI* (2020).

[16] Dahun Kim, Donghyeon Cho, and In So Kweon. 2019. Self-Supervised Video Representation Learning with Space-Time Cubic Puzzles. In *AAAI*. 8545–8552.

[17] Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou. 2018. Two-stream transformer networks for video-based face alignment. *IEEE transactions on pattern analysis and machine intelligence* 40, 11 (2018), 2546–2554.

[18] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. 2019. See More, Know More: Unsupervised Video Object Segmentation With Co-Attention Siamese Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[19] Xin Miao, Xiantong Zhen, Xianglong Liu, Cheng Deng, Vassilis Athitsos, and Heng Huang. 2018. Direct Shape Regression Networks for End-to-End Face Alignment. In *CVPR*. 5040–5049.

[20] Marco Pedersoli, Radu Timofte, Tinne Tuytelaars, and Luc Van Gool. 2014. Using a Deformation Field Model for Localizing Faces and Facial Points under Weak Supervision. In *CVPR*. 3694–3701.

[21] Xi Peng, Rogerio S Feris, Xiaoyu Wang, and Dimitris N Metaxas. 2016. A recurrent encoder-decoder network for sequential face alignment. In *ECCV*. 38–56.

[22] Ilija Radosavovic, Piotr Dollár, Ross B. Girshick, Georgia Gkioxari, and Kaiming He. 2018. Data Distillation: Towards Omni-Supervised Learning. In *CVPR*. 4119–4128.

[23] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 2016. 300 faces in-the-wild challenge: Database and results. *Image and vision computing* 47 (2016), 3–18.

[24] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. 2015. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *ICCV Workshops*. 50–58.

[25] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *NIPS*. 568–576.

[26] Ying Tai, Yicong Liang, Xiaoming Liu, Lei Duan, Jilin Li, Chengjie Wang, Feiyue Huang, and Yu Chen. 2019. Towards highly accurate and stable face alignment for high-resolution videos. In *AAAI*. 8893–8900.

[27] Xin Tang, Fang Guo, Jianbing Shen, and Tianyuan Du. 2018. Facial landmark detection by semi-supervised deep learning. *Neurocomputing* 297 (2018), 22–32.

[28] James Thewlis, Hakan Bilen, and Andrea Vedaldi. 2017. Unsupervised Learning of Object Landmarks by Factorized Spatial Embeddings. In *ICCV*.

[29] Yan Tong, Xiaoming Liu, Frederick W. Wheeler, and Peter H. Tu. 2012. Semi-supervised facial landmark annotation. *Computer Vision and Image Understanding* 116, 8 (2012), 922–935.

[30] Yue Wu and Qiang Ji. 2016. Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In *CVPR*. 3400–3408.

[31] Yue Wu and Qiang Ji. 2019. Facial Landmark Detection: A Literature Survey. *International Journal of Computer Vision* 127, 2 (2019), 115–142.

[32] Yue Wu, Zuoguan Wang, and Qiang Ji. 2013. Facial Feature Tracking Under Varying Facial Expressions and Face Poses Based on Restricted Boltzmann Machines. In *CVPR*. 3452–3459.

[33] Xuehan Xiong and Fernando De la Torre. 2013. Supervised descent method and its applications to face alignment. In *CVPR*. 532–539.

[34] Shi Yin, Shangfei Wang, Guozhu Peng, Xiaoping Chen, and Bowen Pan. 2019. Capturing Spatial and Temporal Patterns for Facial Landmark Tracking through Adversarial Learning. In *IJCAI*. 1010–1017.

[35] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2016. Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence* 38, 5 (2016), 918–930.

[36] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2016. Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence* 38, 5 (2016), 918–930.

[37] Congcong Zhu, Hao Liu, Zhenhua Yu, and Xuehong Sun. 2020. Towards Omni-Supervised Face Alignment for Large Scale Unlabeled Videos. In *AAAI*.

[38] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. 2015. Face alignment by coarse-to-fine shape searching. In *CVPR*. 4998–5006.