Trend-Aware Supervision: On Learning Invariance for Semi-Supervised Facial Action Unit Intensity Estimation

Yingjie Chen[1*], **Jiarui Zhang**[1*], **Tao Wang**[1†], **Yun Liang**[2]

[1] School of Computer Science, Peking University, Beijing China
[2] School of Integrated Circuits, Peking University, Beijing China
chenyingjie@pku.edu.cn, zjr954@pku.edu.cn, wangtao@pku.edu.cn, ericlyun@pku.edu.cn

Abstract

With the increasing need for facial behavior analysis, semi-supervised AU intensity estimation using only keyframe annotations has emerged as a practical and effective solution to relieve the burden of annotation. However, the lack of annotations makes the spurious correlation problem caused by AU co-occurrences and subject variation much more prominent, leading to non-robust intensity estimation that is entangled among AUs and biased among subjects. We observe that trend information inherent in keyframe annotations could act as extra supervision and raising the awareness of AU-specific facial appearance changing trends during training is the key to learning invariant AU-specific features. To this end, we propose **T**rend-**A**ware **S**upervision (TAS), which pursues three kinds of trend awareness, including intra-trend ranking awareness, intra-trend speed awareness, and inter-trend subject awareness. TAS alleviates the spurious correlation problem by raising trend awareness during training to learn AU-specific features that represent the corresponding facial appearance changes, to achieve intensity estimation invariance. Experiments conducted on two commonly used AU benchmark datasets, BP4D and DISFA, show the effectiveness of each kind of awareness. And under trend-aware supervision, the performance can be improved without extra computational or storage costs during inference.

Introduction

Facial behavior analysis is a fundamental task in the field of affective computing with wide application scenarios such as driver fatigue monitoring in driver assistance systems and pain estimation in medical treatment. According to FACS (Friesen and Ekman 1978), facial action units (AU) are defined as subtle facial muscle movements, and almost all facial behaviors can be expressed as certain combinations of AUs. To describe more fine-grained facial behavior, not only the states of AUs (activated or not) but also their intensity values are required. Therefore, AU intensity estimation has become a popular solution for facial behavior analysis. In FACS, the intensity values of AU are quantified into six discrete ordinal intensity levels, including neutral, trace, slight, pronounced, extreme, and maximum (from 0 to 5).

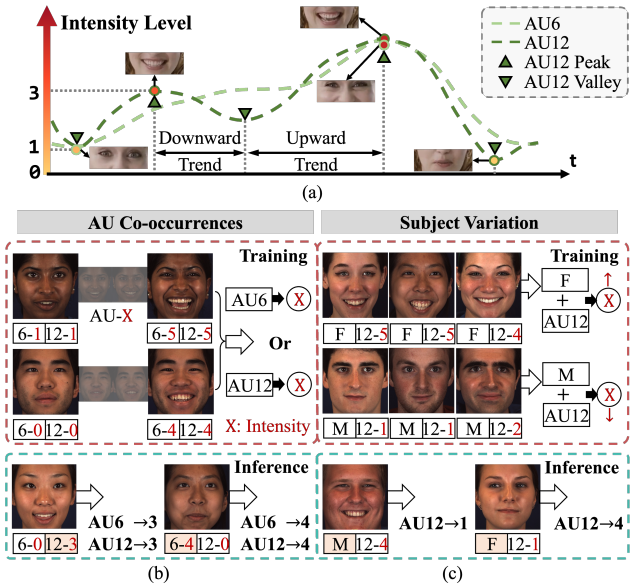

Figure 1: Motivation. (a) Illustration of keyframes and local trends. (b) Spurious correlation caused by AU co-occurrences. The limited annotations may lead the model to learn only AU features of one of the two AUs instead of both, resulting in non-robust intensity estimation entangled between them. (c) Spurious correlation caused by subject variation. Subject variation magnified by the lack of annotations may lead the model to learn non-causal gender features instead of AU features for estimating the intensity of AU12 (orange denotes the dominant features for estimation).

Automatic AU intensity estimation aims to learn a model using FACS-quantified annotations to estimate the intensity values of multiple AUs for a facial image.

The training data requires manual labeling by FACS-certificated experts, which is time-consuming and expensive. Therefore, training AU intensity estimation models using limited annotations, referred to as semi-supervised AU intensity estimation, has attracted increasing attention. To annotate the intensity of each AU for each frame in a video sequence, annotators usually skim through the whole video

*Equal Contribution.
†Corresponding author.

quickly and mark the frames containing a local maximum or minimum (called peak or valley) intensity value of one AU first, *i.e.*, **keyframes** as marked in Fig. 1 (a). Then, for each AU, with every two adjacent keyframes of this AU as references, they annotate frames between them. Considering the process, using only keyframe annotations for training has emerged as a practical and effective task setting, employed by previous works such as (Zhang et al. 2018b, 2019a).

Semi-supervised AU intensity estimation is challenging due to the prominent spurious correlation problem (Geirhos et al. 2018; Hu et al. 2022), *i.e.*, non-causal correlations learned from the limited training annotations. The trained model is supposed to only exploit AU-specific features representing the corresponding facial appearance changes, which is much harder to achieve with non-negligible spurious correlation. **First**, the model aims to estimate the intensity values for multiple AUs simultaneously, and thus AU co-occurrences may lead to spurious correlation. As in Fig. 1 (b), the model is supposed to capture both facial appearance changes from the cheek region for AU6 and lip corner region for AU12. However, their co-occurrence patterns (*e.g.*, both AUs with the same intensity) that frequently occur in keyframe annotations may mislead the model into learning features of only one of them (Shi et al. 2022), *e.g.*, using features from the lip corner region to represent both AUs and ignoring those from the cheek region. **Second**, as in Fig. 1 (c), the lack of annotations magnifies the spurious correlation caused by subject variation. *E.g.*, suppose the annotated frames are mostly women with AU12 at high-intensity levels and men with AU12 at low-intensity levels. In that case, the model may be misled to use non-causal gender features instead of AU features to represent the intensity of AU12.

We argue that the spurious correlation problem hinders the achievement of intensity estimation invariance under semi-supervised settings mainly because features of each AU are entangled with those of other AUs or subject-related features, and the lack of annotations makes it much harder to learn AU-specific features that represent the corresponding facial appearance changes. We observe that there is rich trend information inherent in keyframe annotations, which could act as causal knowledge for capturing AU-specific facial appearance changes. Using such trend information as extra supervision for unannotated frames to raise awareness of facial appearance changing trends during training, is the key to learning invariant AU-specific features. To this end, we propose Trend-Aware Supervision (TAS) to make full use of the inherent trend information for pursuing both intra-trend and inter-trend awareness for AU feature learning, including intra-trend ranking awareness, intra-trend speed awareness, and inter-trend subject awareness.

Specifically, for each AU, every two adjacent keyframes divide a whole video sequence into several local trends, as shown in Fig. 1 (a), and facial appearance corresponding to the specific AU changes monotonically in the local trend. To alleviate the spurious correlation caused by AU co-occurrences, *intra-trend ranking awareness* is pursued to distinguish the specific AU from those with significantly different trends, *i.e.*, AUs corresponding to facial appearance that does not change monotonically. And for the rest of AUs with similar monotonic facial appearance changing trends, *intra-trend speed awareness* further distinguishes among them by encouraging the changing speed of AU features and that of the corresponding facial appearance as close as possible. To alleviate the spurious correlation caused by subject variation, *inter-trend subject awareness* strips subject information by encouraging AU features of the same class with the same intensity label but from different subjects to be as similar as possible. Note that all kinds of awareness supervise on AU features rather than directly on intensity values, to relieve the burden of regression layers by learning invariant AU-specific features. Combining all three awareness, here comes our trend-aware supervision scheme, which encourages the model to learn invariant correlations instead of spurious ones to achieve intensity estimation invariance.

Our main contributions are threefold.

- We inspect the keyframe-based semi-supervised AU intensity estimation problem and first identify the spurious correlation problem as the main challenge for achieving intensity estimation invariance.
- We observe the rich trend information inherent in keyframe annotations and propose Trend-Aware Supervision to raise intra-trend and inter-trend awareness during training to learn invariant AU-specific features.
- Experimental results show that without adding extra computational or storage costs in the inference stage, the performance of our model is improved and even exceeds several fully supervised ones.

Related Work

Supervised AU Intensity Estimation

To pursue a more precise way of describing facial behaviors, AU intensity estimation task has gradually emerged as a promising solution (Rudovic, Pavlovic, and Pantic 2014; Mohammadi, Fatemizadeh, and Mahoor 2017; Sánchez-Lozano, Tzimiropoulos, and Valstar 2018; Wang, Jiang, and Shen 2020; Song et al. 2021). Works such as (Li et al. 2015; Walecki et al. 2017a,b) take advantage of Probabilistic Graphical Models to model the latent dependencies among AUs. (Wang, Hao, and Ji 2018) proposed HBN that employs a hybrid Bayesian Network to estimate AU intensity values by capturing the global dependencies among AUs.

Works such as (Kaltwang, Todorovic, and Pantic 2015; Linh Tran et al. 2017; Fan et al. 2020; Fan, Lam, and Li 2020; Fan and Lin 2021) make efforts in learning AU-specific features by injecting prior knowledge into models. For using location information, (Fan et al. 2020) proposed a joint AU intensity prediction and localization method that works directly on the whole input image. And works such as (Fan, Lam, and Li 2020; Fan and Lin 2021) transfer AU intensity estimation task into a facial landmark heatmap regression problem. Instead of introducing explicit multi-task or joint learning, our model learns better AU-specific features implicitly via a trend-aware training process.

Semi-Supervised AU Intensity Estimation

In the face of the overwhelming pressure of manually annotating a large amount of training data, semi-supervised AU

intensity estimation has become a practical and effective solution (Zhao et al. 2016; Zhang et al. 2018c,b; Wang et al. 2019; Zhang et al. 2019a; Sanchez et al. 2020).

Due to the lack of annotations, prior knowledge plays an essential role in improving performance. OSVR (Zhao et al. 2016) exploits the ordinal information among different frames and intensity labels of selected frames, but it requires pre-extracted features as input and training models separately for each AU. BORMIR (Zhang et al. 2018c) learns an extra relevant value for each frame and not only constraints ordinal relevance but also constraints relevance smoothness and intensity smoothness via a regularization term. KBSS (Zhang et al. 2018b) uses segments containing only four sampled frames with extra neutral frames for training, resulting in insufficient use of ordinal information. To capture fine-grained AU features, (Zhang et al. 2019a) made efforts in learning context-aware features and proposed an LSTM-based method with learnable AU-related context for AU-specific feature learning. We argue that constraining feature or intensity value smoothness via a regularization term as previous works may lead to the over-smooth problem. In contrast, we propose intra-trend speed awareness to achieve smooth and stable estimation without breaking away from the corresponding facial appearance changing trend.

Despite their success, previous works failed in solving the spurious correlation problem, leading to non-robust estimation. In this paper, we propose Trend-Aware Supervision to achieve intensity estimation invariance.

Approach

Problem Formulation

AU intensity estimation aims to train a model $f(I, \Theta)$ that estimates the intensity values of C AU classes $\tilde{\mathbf{v}} = \{\tilde{v}_c \in \mathbb{R}\}_{c=1}^{C}$ for a given facial image simultaneously. Under the keyframe-based setting, video sequences with only keyframe annotations are used as training data, *i.e.*, for each frame in a video sequence, FACS-quantified intensity label $v_c \in \mathbb{N}$ is given for the current frame only if the intensity of the c^{th} AU reaches peak or valley. The goal is to find an optimal Θ that makes the estimated intensity $\tilde{\mathbf{v}}$ as close to the real trend of AU-specific facial appearance changes as possible, based on FACS-quantified keyframe annotations.

To train our AU intensity estimation model, for each AU class, we split those video sequences into segments based on the locations of keyframes, the same as (Zhang et al. 2019a). In each segment for a specific AU, the intensity of this AU evolves either from peak to valley or from valley to peak. For each segment, only the linearly sampled T frames are used. In this way, the training data is organized into segments $\{\mathbf{I}_i, v_{i,c}^1, v_{i,c}^T\}_{i=1}^{N}$, where N denotes the number of segments, $c \in \{1, 2, \ldots, C\}$ denotes the c^{th} AU out of the total C AU classes, and $\mathbf{I} = \{I^t \in \mathbb{R}^{H \times W \times 3}\}_{t=1}^{T}$ is the T images contained in each segment. Note that only the intensity label of the c^{th} AU in the first and the last images are given, *i.e.*, $\{v_c^1, v_c^T\}$. For convenience, the intensity labels are uniformly normalized from range $[0, 5]$ to range $[0, 1]$.

Overview

As shown in Fig. 2, our goal is to learn an optimal Θ for an image-based AU intensity estimation model $f(I, \Theta)$. In the training stage, a batch of pre-processed segments is taken as input. Each image is fed into a backbone network to extract global feature $F^{\text{global}} \in \mathbb{R}^{d \times H_{\text{g}} \times W_{\text{g}}}$. Then, to obtain AU features $F^{\text{au}} = \{f_c \in \mathbb{R}^d\}_{c=1}^{C}$, C separate spatial attention layers (Zhao and Wu 2019) are applied to the extracted F^{global}, and through the corresponding c^{th} branch, AU feature f_c is obtained after an average pooling operation and an l_2 normalization operation. After that, AU feature f_c is fed into the corresponding MLP consisting of two linear layers to estimate the intensity value $\tilde{v}_c \in \mathbb{R}$ for the c^{th} AU class, which is clipped to range $[0, 1]$. In this way, the final intensity estimation results $\tilde{\mathbf{v}} = \{\tilde{v}_c\}_{c=1}^{C}$ are obtained.

During training, a commonly used regression loss function acts as direct supervision for the two annotated keyframes in each segment, for which we employ Mean-squared Error (MSE) loss function as shown in Eq. 1. Besides, we apply the proposed trend-aware supervision to the AU features F^{au} extracted from each image, to make up for the lack of annotations by raising trend awareness for invariant AU-specific feature learning.

$$\mathcal{L}_{\text{reg}} = \sum_{i=1}^{N} \big((v_{i,c}^1 - \tilde{v}_{i,c}^1)^2 + (v_{i,c}^T - \tilde{v}_{i,c}^T)^2 \big). \tag{1}$$

In the inference stage, *note that segment input is not required anymore*. The obtained model $f(I, \Theta)$ takes one facial image as input and outputs estimation result $\tilde{\mathbf{v}}$ directly.

Trend-Aware Supervision

We propose trend-aware supervision to achieve intensity estimation invariance, which pursues three kinds of awareness. Different from previous works (Zhang et al. 2018c, 2019b) that directly constrain smoothness or ordering of intensity values, we first focus on learning invariant AU-specific features for more robust intensity estimation.

Intra-Trend Ranking Awareness As aforementioned, keyframe annotations of each AU divide a whole video sequence into several local trends, and between pairs of adjacent keyframes, facial appearance corresponds to the specific AU changes monotonically, which is essential supervision for distinguishing the AU from others that correspond to non-monotonic changing trends of AU-specific facial appearance. Instead of focusing on constraining intensity values (Zhang et al. 2018c, 2019b), we first focus on constraining AU-specific features. We argue that constraining intensity values directly will make the MLP layers do a lot of heavy lifting, leading to less specificity in the extracted AU features. Therefore, instead of directly forcing the estimated results to be monotonic, we enforce the changing amplitudes of AU features relative to the first frame to be monotonic, to make it correspond to the changing amplitudes of AU-specific facial appearance, which benefits the model to capture facial appearance features from regions related to the specific AU, *i.e.*, making AU features more sensitive to the intensity changes of the corresponding facial appearance.

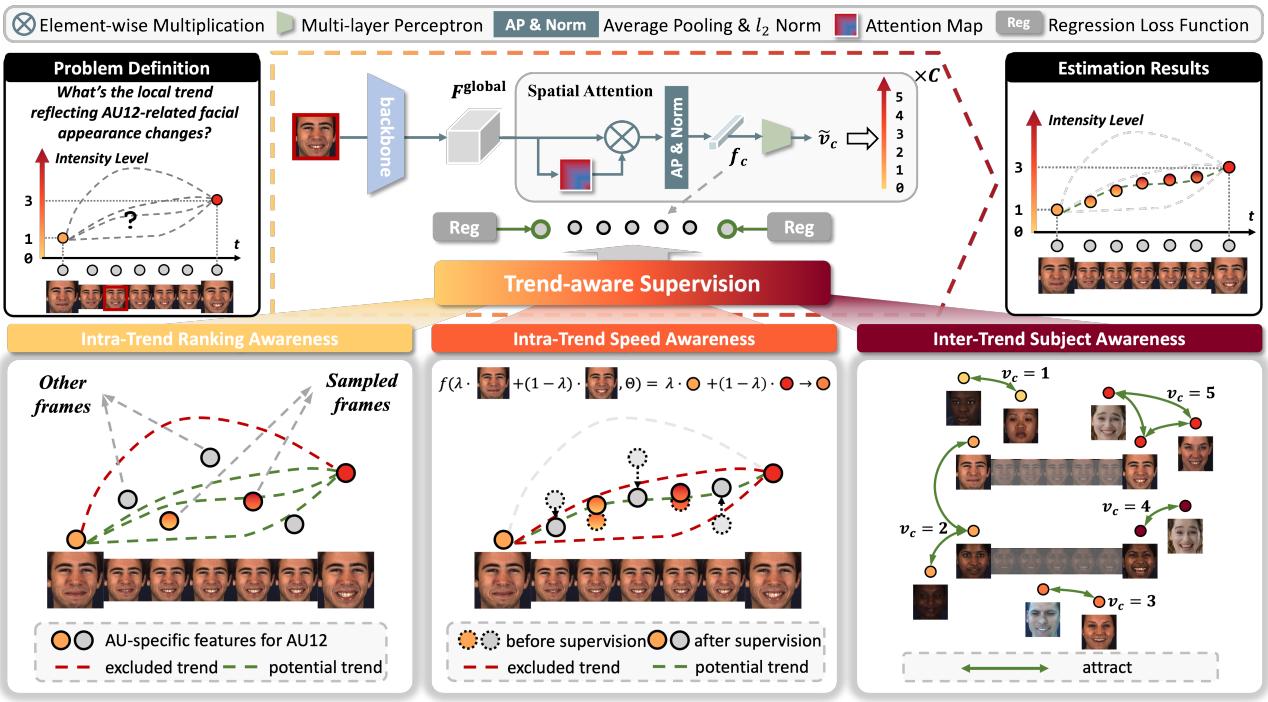

Figure 2: Overview. Our method takes a batch of segments as input and each image (taken the one with red box as an example) is first fed into a backbone network for global feature extraction. Then, C separate spatial attention layers are applied to F_{global} for AU feature extraction. After that, each AU feature f_c is fed into an MLP to estimate the intensity value $\tilde{v}_c$ for the c^{th} AU class. During training, trend-aware supervision is applied to AU features $\{f_c^t\}_{t=1}^T$ in each segment. And regression loss function is applied to the estimated intensity results of the two annotated keyframes only, *i.e.*, the first and the last ones in each segment.

To measure the changing amplitudes of AU features, we compute the Euclidean distance between the AU feature of each frame and that of the first frame, $\Delta\mathbf{f}_c = \{\Delta f_c^t = \|f_c^t - f_c^1\|_2\}_{t=1}^T$, and enforce a monotonic increasing trend among them, *i.e.*, $\Delta f_c^1 \leq \Delta f_c^2, \ldots, \leq \Delta f_c^T$. Note that whether the trend is upward or downward, the corresponding facial appearance changing amplitudes relative to the first frame keep getting larger. In this way, the loss function for intra-trend ranking awareness is defined as in Eq. 2.

$$\mathcal{L}_{\text{rank}} = \sum_{i=1}^{N} \sum_{t=1}^{T-1} \text{Max}(\mathbf{0}, \mathbf{A}\Delta\mathbf{f}_{i,c}), \tag{2}$$

where $\mathbf{A}_{T\times T}$ is for computing the difference of first order of $\Delta\mathbf{f}_{i,c}$ by setting $\{a_{t,t}\}_{t=1}^T$ to 1, $\{a_{t,t+1}\}_{t=1}^{T-1}$ to -1, and others to 0. Note that for each segment, only AU features of the annotated AU class are constrained.

Intra-Trend Speed Awareness Intra-trend ranking awareness is not enough to distinguish the AU from those that also correspond to a monotonic changing trend of AU-specific facial appearance. Therefore, intra-trend speed awareness is pursued to raise the awareness of facial appearance changing speed by encouraging the model to behave linearly in between frames in a local trend. In other words, if the AU-specific facial appearance changes rapidly or slowly in a local trend, the changing amplitudes of AU features relative to the first frame are also supposed to change rapidly or slowly. The lack of annotations increases the difficulty of raising such speed awareness since there is no direct supervision for frames between adjacent keyframes. We make up for that by modeling the vicinity relation across images in one local trend. Given *mixup* operation (Zhang et al. 2018a): $\mathcal{M}_\lambda(x_i, x_j) = \lambda \cdot x_i + (1-\lambda)\cdot x_j$, we sample virtual image-target pairs from the *mixup* vicinal distribution and encourage the model to behave linearly in-between the sampled images in each segment, *i.e.*, encouraging the model to estimate an intermediate intensity $\mathcal{M}_\lambda(f(I^i,\Theta), f(I^j,\Theta))$ for input $\mathcal{M}_\lambda(I^i, I^j)$ (I^i and I^j are from the same segment).

The benefits lie in two folds: 1) for limited annotations, the augmented virtual image-target pairs encourage the model to learn decision boundaries that transit linearly between quantified intensity labels, providing a smoother estimation. 2) for AU-specific feature learning, since $\mathcal{M}$ is only applied between images in a segment of one subject, the features relevant to subject identity are almost unchanged and the features relevant to AU-specific facial appearance changes are highlighted. Thus, the model benefits from learning AU-specific features coupling with intensity changes, promoting stable estimation reflecting the real AU-specific facial appearance changes.

In each iteration, for an input segment with the estimated intensity results $\{\tilde{\mathbf{v}}^t = f(I^t,\Theta)\}_{t=1}^T$, we

Dataset		BP4D						DISFA												
Metric	Method	AU6	10	12	14	17	**Avg.**	AU1	2	4	5	6	9	12	15	17	20	25	26	**Avg.**
ICC↑	Ladder	.67	.62	.79	.07	.44	.52	-.01	.06	.04	.03	.46	.09	.60	-.02	.01	.00	.58	.37	.18
	OSVR	.65	.58	.78	.27	.45	.54	.21	.04	.25	.15	.23	.15	.31	.12	.07	.09	.62	.09	.19
	BORMIR	.73	.68	.86	.37	.47	.62	.20	.25	.30	.17	.39	.18	.58	.16	.23	.09	.71	.15	.28
	KJRE (6%)	.71	.61	.87	.39	.42	.60	.27	.35	.25	.33	.51	.31	.67	.14	.17	.20	.74	.25	.35
	KBSS	.76	.73	.84	.45	.45	.65	.14	.12	.48	.17	.43	.35	.71	.15	.25	.09	.78	.54	.35
	CFLF	.77	.70	.83	.41	.60	.66	.26	.19	.46	.35	.52	.36	.71	.18	.34	**.21**	.81	.51	.41
	RandCon	.77	.75	.86	**.48**	.55	.68	.33	.33	.65	-.02	.60	.34	.78	.18	.24	.08	.88	.58	.41
	Ours	**.78**	**.77**	**.88**	**.48**	**.62**	**.71**	**.47**	**.57**	**.70**	**.38**	**.62**	**.37**	**.80**	**.25**	**.41**	.14	**.92**	**.59**	**.52**
MAE↓	Ladder	.69	.84	.60	1.20	.64	.79	.65	.34	1.26	**.11**	**.28**	.33	**.35**	.19	.30	**.15**	.76	.39	.43
	OSVR	1.02	1.13	.95	1.35	.93	1.08	1.65	1.87	2.94	1.38	1.56	1.69	1.64	1.10	1.61	1.37	1.33	1.79	1.66
	BORMIR	.85	.90	.68	1.05	.79	.85	.88	.78	1.24	.59	.77	.78	.76	.56	.72	.63	.90	.88	.79
	KJRE (6%)	.82	.95	.64	1.08	.85	.87	1.02	.92	1.86	.70	.79	.87	.77	.60	.80	.72	.96	.94	.91
	KBSS	.74	.77	.69	.99	.90	.82	.53	.49	.82	.24	.39	.38	.43	.32	.50	.36	.61	.44	.46
	CFLF	.62	.83	.62	1.00	**.63**	.74	**.33**	**.28**	**.61**	.13	.35	.28	.43	**.18**	**.29**	.16	.53	.40	**.33**
	RandCon	.65	.91	.83	**.98**	**.63**	.80	.43	.36	.65	.19	.38	.37	.46	.25	.38	.21	.45	.39	.38
	Ours	**.53**	**.67**	**.51**	.99	.73	**.69**	.52	.36	**.61**	.19	.33	**.27**	.39	.23	.46	.27	**.33**	**.35**	.36

Table 1: Comparison with the state-of-the-art semi-supervised methods. The best and second results are indicated using bold and underline, respectively. Note that KJRE uses 6% of the annotations while others use 2%.

shuffle images to construct virtual image-target pairs, $(\mathcal{M}_\lambda(I^i, I^j), \mathcal{M}_\lambda(\tilde{\mathbf{v}}^i, \tilde{\mathbf{v}}^j))$, and define the loss function as:

$$\mathcal{L}_{\text{spd}} = \sum_{n=1}^{N} \sum_{i,j} \| f\left(\mathcal{M}_\lambda(I^i, I^j), \Theta\right) - \mathcal{M}_\lambda(\tilde{\mathbf{v}}^i, \tilde{\mathbf{v}}^j)) \|_2^2, \tag{3}$$

where $\lambda \sim Beta(\alpha, \alpha)$ and α is a hyper-parameter, and for i, j, frame indexes are shuffled and matched with the original ones to construct T pairs for each segment.

Inter-Trend Subject Awareness By raising the above two awareness during training, we focus on promoting the coupling between intra-trend AU-specific facial appearance changes and AU feature changes, making the learned AU-specific features reflect the real trend of the corresponding facial appearance changes. However, due to spurious correlation caused by subject variation, the model has no incentive to learn invariant AU-specific features for estimation as learning some spurious features (such as gender, race, *etc*.) suffices to estimate target AU intensity. Therefore, AU features of the same class with the same intensity labels still vary among local trends from different subjects, which makes the model hard to generalize well on unseen subjects. To tackle the problem, AU-specific features that are invariant among subjects are required to be learned.

To learn invariant AU-specific features that are not entangled with subject-related ones, it is important to strip subject-related features by aligning AU features of one class from different subjects but with the same intensity label, *i.e*., only retaining AU-specific facial appearance features invariant among subjects for achieving intensity estimation invariance. To this end, we pursue inter-trend subject awareness during training for AU-specific feature learning. The object function is clear that AU features of the same class with the same intensity label are supposed to be similar, no matter from which subjects they are extracted. And considering that only keyframes are with reliable annotations, the loss function for inter-trend subject awareness is only applied to keyframes of different segments, as shown in Eq. 4:

$$\mathcal{L}_{\text{sub}} = \frac{1}{C} \sum_{c=1}^{C} \mathbb{E}_{v_c} \left[(1 - f_{i,c}^t \cdot f_{j,c}^k) \right], \tag{4}$$

where $t, k \in \{1, T\}$ are annotated keyframes in segment i and j with the same intensity label v_c.

The overall loss function for training is as follows:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{reg}} + \lambda_{\text{rank}} \mathcal{L}_{\text{rank}} + \lambda_{\text{spd}} \mathcal{L}_{\text{spd}} + \lambda_{\text{sub}} \mathcal{L}_{\text{sub}}, \tag{5}$$

where λ_{rank}, λ_{spd}, and λ_{sub} are for balancing.

Experiments

Experimental Settings

We use two AU benchmark datasets for evaluation, BP4D (Zhang et al. 2014) and DISFA (Mavadati et al. 2013). **BP4D** for FERA 2015 challenge (Valstar et al. 2015), involves 23 female and 18 male subjects. Around 140,000 frames are annotated by two FACS-certified coders with quantified AU intensity labels for 5 AU classes. As the same as works (Zhang et al. 2018b, 2019a), the official split is used, in which 21 subjects are selected for training and the rest 20 subjects are for testing. **DISFA** is a spontaneous AU dataset consisting of 26 adult subjects. Around 130,000 image frames are annotated with quantified AU intensity labels for 12 AU classes. For a fair comparison, we follow the same experimental setting mentioned in (Zhang et al. 2018b, 2019a), conducting a subject-exclusive 3-fold cross-validation. FACS-quantified AU intensity labels of both datasets are within a discrete scale from 0 to 5. **Only about 2% annotations in BP4D and 1% in DISFA** — keyframe

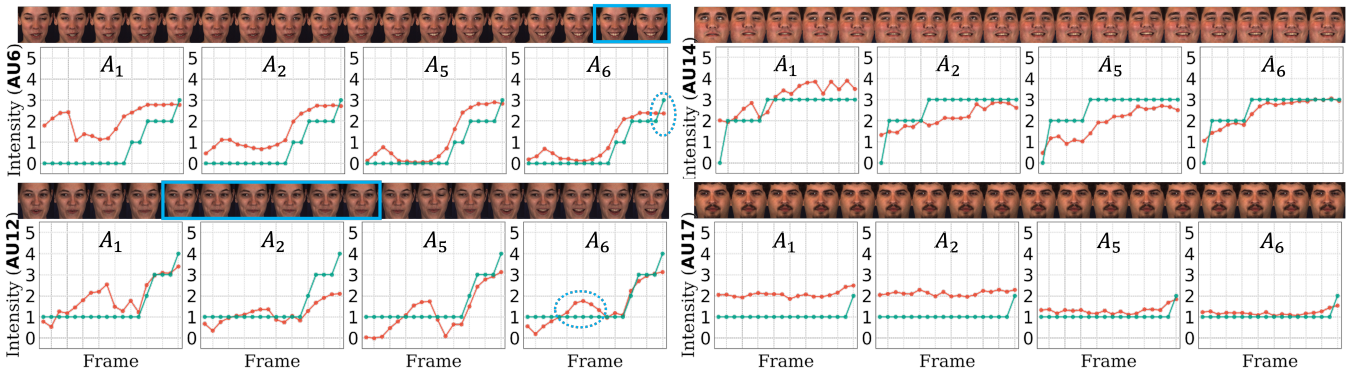

Figure 3: Case study. In each tuple, from left to right, four line charts show the intensity values estimated by model A_1, A_2, A_5 and A_6 (red line) for the given sequences on the top, respectively, and the FACS-quantified intensity labels (green line).

annotations for each AU are used for our semi-supervised setting, the same as (Zhang et al. 2019a). Intra-class Correlation (ICC(3,1)) (Shrout and Fleiss 1979) and Mean Absolute Error (MAE) are reported, the same as previous works such as (Zhao et al. 2016; Zhang et al. 2018c, 2019a).

Implementation Details For each input image, Dlib (King 2009) is used for aligning facial images. Each image is cropped and resized to 256×256. We use ResNet34 (He et al. 2016) without final layers as the backbone network. For each spatial attention layer, the kernel size is set to 9. $H_g = W_g = 8, d = 512$ for the extracted global feature. Each MLP for intensity estimation consists of two linear layers ($512 \rightarrow 64, 64 \rightarrow 1$) with LeakyReLU activation function between them, and a clipping function is applied to its output to clip values into range $[0, 1]$. Empirically, T, α, λ_{rank} and λ_{spd} are set to 16, 0.5, 0.1, and 0.05, respectively. λ_{sub} is set to 0.005 on BP4D and 0.05 on DISFA due to the different number of subjects.

We implement our method on Pytorch (Paszke et al. 2017) platform. SGD is used for optimization with weight decay of 0.0005 and learning rate of 0.005. We set batch size to 16 for both datasets. Each model is trained for 20 epochs, and early stopping strategy is adopted. All the experiments are conducted on one NVIDIA A100 Tensor Core GPU.

Comparison to the State-of-the-arts

Comparison to semi-supervised methods. First, we compare our method with top-performing semi-supervised AU intensity estimation methods, including Ladder (Rasmus et al. 2015), OSVR (Zhao et al. 2016), BORMIR (Zhang et al. 2018c), KJRE (Zhang et al. 2019b), KBSS (Zhang et al. 2018b), and CFLF (Zhang et al. 2019a), as shown in Table 1. In Table 1, our method achieves the highest ICC and the lowest MAE on both datasets, and both metrics outperform the compared methods by a large margin.

Comparison to fully-supervised methods. We also compare our method to several state-of-the-art fully-supervised methods, including HBN (Wang, Hao, and Ji 2018), Heatmap (Sánchez-Lozano, Tzimiropoulos, and Valstar 2018), 2DC (Linh Tran et al. 2017), CCNN-IT (Walecki et al. 2017a), CNN (Gudi et al. 2015), RE-Net (Yang and Yin 2020), and SCC (Fan, Lam, and Li 2020), All of them use full annotations for training, which is a much larger amount of annotations than that used for our setting. As in Table 2, our method achieves comparable ICC and MAE to other methods, considering the limited annotations.

Dataset	BP4D		DISFA	
Method	ICC	MAE	ICC	MAE
HBN	.70	-	-	-
Heatmap	.68	-	-	-
2DC	.66	-	.49	-
CCNN-IT	.63	1.26	.38	.66
CNN	.60	.82	.33	.42
RE-Net	.64	<u>.65</u>	**.54**	<u>.22</u>
SCC	**.72**	**.58**	.47	**.20**
Ours	<u>.71</u>	.69	<u>.52</u>	.36

Table 2: Comparison with supervised methods. The best and second results are indicated using bold and underline.

Ablation Study

To validate the effectiveness of all kinds of awareness, we conduct an ablation study in Table 3. Model A_0 is vanilla ResNet34 without spatial attention layers, and other mod-

	Dataset				BP4D		DISFA	
	Sep	$\mathcal{L}_{\text{rank}}$	$\mathcal{L}_{\text{spd}}$	$\mathcal{L}_{\text{sub}}$	ICC	MAE	ICC	MAE
A_0					.59	.92	.37	.89
A_1	✓				.68	.84	.44	.63
A_2	✓	✓			.69	.78	.48	.60
A_3	✓		✓		.70	.75	.48	.57
A_4	✓			✓	.68	.79	.46	.55
A_5	✓	✓	✓		.70	.71	.50	.47
A_6	✓	✓	✓	✓	**.71**	**.69**	**.52**	**.36**

Table 3: Ablation study on $\mathcal{L}_{\text{rank}}$, $\mathcal{L}_{\text{spd}}$, and $\mathcal{L}_{\text{sub}}$. *Sep* denotes separate spatial attention layers.

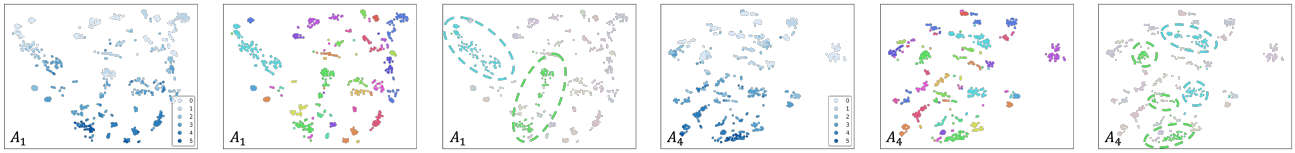

Figure 4: t-SNE visualization for AU features of AU12 on BP4D. From left to right, every three t-SNE results are colored according to FACS-quantified intensity labels (light blue to dark blue), subject identities (bright colors), and highlighted subject identities (bright blue and green). The first three are for Model A_1, and the last three are for Model A_4.

Dataset	BP4D		DISFA	
Method	ICC	MAE	ICC	MAE
S_1: segments, all frames, multi AU labels				
Baseline	.65	.68	.49	.31
+ $\mathcal{L}_{\text{all}}$	**.68**	**.63**	**.52**	**.29**
S_2: segments, randomly selected 2 frames, single AU label				
Baseline	.67	.80	.42	.55
+ $\mathcal{L}_{\text{all}}$	**.68**	**.73**	**.48**	**.49**
S_3: randomly selected 2% (1%) frames, multi AU labels				
Baseline	.67	.67	.44	.27
+ $\mathcal{L}_{\text{sub}}$	**.69**	**.63**	**.47**	**.25**

Table 4: Ablation study on other semi-supervised settings.

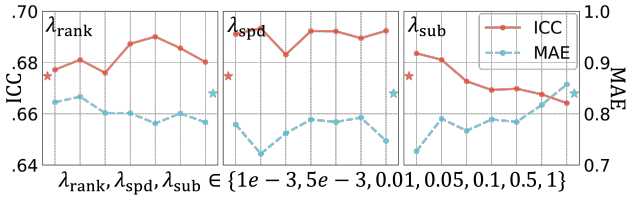

Figure 5: Empirical study for λ_{rank}, λ_{spd}, and λ_{sub} on BP4D. Red and blue stars show ICC and MAE of A_1.

els (A_1 to A_6) are with the same network architecture as in Fig. 2 and only differ in the applied loss function. It can be observed that each awareness contributes to performance improvement, and by adding them one by one (A_2, A_5, A_6), Model A_6 achieves the best performance.

To further demonstrate the superiority of trend-aware supervision, we apply it to baseline models trained under other semi-supervised settings (S_1, S_2, S_3), as shown in Table 4. We can observe that under all settings, by pursuing trend awareness, the model performance is further improved.

Empirical Study

We investigate the influence of hyper-parameters (Eq. 5) on BP4D. As in Fig. 5, by setting each λ to the given seven values, we can observe that for all kinds of awareness, the performance is not very sensitive to the corresponding trade-off hyper-parameter. When $\lambda_{\text{rank}} = 0.1$, $\lambda_{\text{spd}} = 0.05$ and $\lambda_{\text{sub}} = 0.005$, the model achieves better performance.

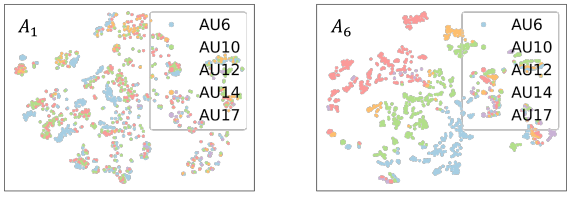

Figure 6: t-SNE visualization for AU features of one subject. Left: AU features from A_1. Right: AU features from A_6.

Qualitative Results

Fig. 3 shows that by pursuing ranking awareness, there are fewer jitters in the results of Model A_2. By pursuing speed awareness, the changing trend of intensity estimated by Model A_5 is more consistent with the corresponding facial appearance changes. With inter-trend subject awareness, for AUs with lower occurrence frequency such as AU14 and AU17, on which the model tends to overfit, some overall offsets on the changing trend of the estimated intensity are rectified by Model A_6, as in the right two tuples. And compared to FACS annotations, the estimated results are more accurate in some cases. *E.g.*, in the top-left tuple, there is no obvious difference between the last two frames, and thus the estimated intensity of AU6 barely changed. And in the bottom-left tuple, the woman pulls her lip corner a little bit with her dimple getting deeper in the marked frames, corresponding to a slight increase in the intensity of AU12.

Fig. 4 shows that AU features with different intensity labels extracted by Model A_4 (only with extra $\mathcal{L}_{\text{sub}}$) are more compact, and AU features from one subject but with different intensity are far away from each other, which indicates that the learned AU-specific features are consistent among subjects. Fig. 6 shows that AU features of Model A_6 are more disentangled with each other than those of Model A_1, which demonstrates the superiority of trend-aware supervision in learning invariant AU features.

Conclusion

In this paper, we inspect the keyframe-based semi-supervised AU intensity estimation and identify the spurious correlation problem as the main challenge for achieving intensity estimation invariance. To this end, we propose trend-aware supervision to raise trend awareness during training. Extensive experiments show that all kinds of awareness are essential and help in learning invariant AU-specific features.

References

Fan, Y.; Lam, J.; and Li, V. 2020. Facial action unit intensity estimation via semantic correspondence learning with dynamic graph convolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12701–12708.

Fan, Y.; and Lin, Z. 2021. G2RL: geometry-guided representation learning for facial action unit intensity estimation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 731–737.

Fan, Y.; Shen, J.; Cheng, H.; and Tian, F. 2020. Joint Facial Action Unit Intensity Prediction And Region Localisation. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Friesen, E.; and Ekman, P. 1978. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3(2): 5.

Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2018. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Gudi, A.; Tasli, H. E.; Den Uyl, T. M.; and Maroulis, A. 2015. Deep learning based facs action unit occurrence and intensity estimation. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 6, 1–5. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, Z.; Zhao, Z.; Yi, X.; Yao, T.; Hong, L.; Sun, Y.; and Chi, E. H. 2022. Improving Multi-Task Generalization via Regularizing Spurious Correlation. *arXiv preprint arXiv:2205.09797*.

Kaltwang, S.; Todorovic, S.; and Pantic, M. 2015. Doubly sparse relevance vector machine for continuous facial behavior estimation. *IEEE transactions on pattern analysis and machine intelligence*, 38(9): 1748–1761.

King, D. E. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10: 1755–1758.

Li, Y.; Mavadati, S. M.; Mahoor, M. H.; Zhao, Y.; and Ji, Q. 2015. Measuring the intensity of spontaneous facial action units with dynamic Bayesian network. *Pattern Recognition*, 48(11): 3417–3427.

Linh Tran, D.; Walecki, R.; Eleftheriadis, S.; Schuller, B.; Pantic, M.; et al. 2017. Deepcoder: Semi-parametric variational autoencoders for automatic facial action coding. In *Proceedings of the IEEE International Conference on Computer Vision*, 3190–3199.

Mavadati, S. M.; Mahoor, M. H.; Bartlett, K.; Trinh, P.; and Cohn, J. F. 2013. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2): 151–160.

Mohammadi, M. R.; Fatemizadeh, E.; and Mahoor, M. H. 2017. An adaptive bayesian source separation method for intensity estimation of facial aus. *IEEE Transactions on Affective Computing*, 10(2): 144–154.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in PyTorch. In *NIPS Workshop*.

Rasmus, A.; Berglund, M.; Honkala, M.; Valpola, H.; and Raiko, T. 2015. Semi-supervised learning with ladder networks. *Advances in neural information processing systems*, 28.

Rudovic, O.; Pavlovic, V.; and Pantic, M. 2014. Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. *IEEE transactions on pattern analysis and machine intelligence*, 37(5): 944–958.

Sanchez, E.; Bulat, A.; Zaganidis, A.; and Tzimiropoulos, G. 2020. Semi-supervised facial action unit intensity estimation with contrastive learning. In *Proceedings of the Asian Conference on Computer Vision*.

Sánchez-Lozano, E.; Tzimiropoulos, G.; and Valstar, M. 2018. Joint action unit localisation and intensity estimation through heatmap regression. *arXiv preprint arXiv:1805.03487*.

Shi, Y.; Daunhawer, I.; Vogt, J. E.; Torr, P.; and Sanyal, A. 2022. How robust is unsupervised representation learning to distribution shift? In *The Eleventh International Conference on Learning Representations*.

Shrout, P. E.; and Fleiss, J. L. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2): 420.

Song, T.; Cui, Z.; Wang, Y.; Zheng, W.; and Ji, Q. 2021. Dynamic Probabilistic Graph Convolution for Facial Action Unit Intensity Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4845–4854.

Valstar, M. F.; Almaev, T.; Girard, J. M.; McKeown, G.; Mehu, M.; Yin, L.; Pantic, M.; and Cohn, J. F. 2015. Fera 2015-second facial expression recognition and analysis challenge. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, 1–8. IEEE.

Walecki, R.; Pavlovic, V.; Schuller, B.; Pantic, M.; et al. 2017a. Deep structured learning for facial action unit intensity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3405–3414.

Walecki, R.; Rudovic, O.; Pavlovic, V.; and Pantic, M. 2017b. Copula ordinal regression framework for joint estimation of facial action unit intensity. *IEEE Transactions on Affective Computing*, 10(3): 297–312.

Wang, C.; Jiang, F.; and Shen, R. 2020. Facial Action Units Intensity Estimation via Graph Relation Network. In Yang, H.; Pasupa, K.; Leung, A. C.; Kwok, J. T.; Chan, J. H.; and King, I., eds., *Neural Information Processing - 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 23-27, 2020, Proceedings, Part II*, volume 12533 of *Lecture Notes in Computer Science*, 345–356. Springer.

Wang, S.; Hao, L.; and Ji, Q. 2018. Facial action unit recognition and intensity estimation enhanced through label dependencies. *IEEE Transactions on Image Processing*, 28(3): 1428–1442.

Wang, S.; Pan, B.; Wu, S.; and Ji, Q. 2019. Deep facial action unit recognition and intensity estimation from partially labelled data. *IEEE Transactions on Affective Computing*, 12(4): 1018–1030.

Yang, H.; and Yin, L. 2020. RE-Net: A Relation Embedded Deep Model for AU Occurrence and Intensity Estimation. In *Proceedings of the Asian Conference on Computer Vision*.

Zhang, H.; Cissé, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018a. mixup: Beyond Empirical Risk Minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Zhang, X.; Yin, L.; Cohn, J. F.; Canavan, S.; Reale, M.; Horowitz, A.; Liu, P.; and Girard, J. M. 2014. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10): 692–706.

Zhang, Y.; Dong, W.; Hu, B.-G.; and Ji, Q. 2018b. Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2314–2323.

Zhang, Y.; Jiang, H.; Wu, B.; Fan, Y.; and Ji, Q. 2019a. Context-aware feature and label fusion for facial action unit intensity estimation with partially labeled data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 733–742.

Zhang, Y.; Wu, B.; Dong, W.; Li, Z.; Liu, W.; Hu, B.-G.; and Ji, Q. 2019b. Joint representation and estimator learning for facial action unit intensity estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3457–3466.

Zhang, Y.; Zhao, R.; Dong, W.; Hu, B.-G.; and Ji, Q. 2018c. Bilateral ordinal relevance multi-instance regression for facial action unit intensity estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7034–7043.

Zhao, R.; Gan, Q.; Wang, S.; and Ji, Q. 2016. Facial expression intensity estimation using ordinal information. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3466–3474.

Zhao, T.; and Wu, X. 2019. Pyramid feature attention network for saliency detection. In *CVPR*, 3085–3094.