Context-Aware Feature and Label Fusion for Facial Action Unit Intensity Estimation With Partially Labeled Data

Yong Zhang[1]*, Haiyong Jiang[2]*, Baoyuan Wu[1]†, Yanbo Fan[1], Qiang Ji[3]
[1]Tencent AI Lab, [2]Nanyang Technological University, Singapore, [3]Rensselaer Polytechnic Institute
{zhangyong201303,haiyong.jiang1990,wubaoyuan1987,fanyanbo0124}@gmail.com, qji@ecse.rpi.edu

Abstract

Facial action unit (AU) intensity estimation is a fundamental task for facial behaviour analysis. Most previous methods use a whole face image as input for intensity prediction. Considering that AUs are defined according to their corresponding local appearance, a few patch-based methods utilize image features of local patches. However, fusion of local features is always performed via straightforward feature concatenation or summation. Besides, these methods require fully annotated databases for model learning, which is expensive to acquire. In this paper, we propose a novel weakly supervised patch-based deep model on basis of two types of attention mechanisms for joint intensity estimation of multiple AUs. The model consists of a feature fusion module and a label fusion module. And we augment attention mechanisms of these two modules with a learnable task-related context, as one patch may play different roles in analyzing different AUs and each AU has its own temporal evolution rule. The context-aware feature fusion module is used to capture spatial relationships among local patches while the context-aware label fusion module is used to capture the temporal dynamics of AUs. The latter enables the model to be trained on a partially annotated database. Experimental evaluations on two benchmark expression databases demonstrate the superior performance of the proposed method.

1. Introduction

Facial Action Coding System (FACS) [5] defines a set of AUs to depict the movements of facial muscles. Each AU is associated with one or a set of muscles. AUs can be treated as basic elements to encode nearly all anatomically possible human expressions [22]. AU intensity is used to describe the extent of muscle movement, which presents detailed information of facial behaviours. It is quantified into six-point ordinal scales in FACS. Automatic AU intensity estimation is valuable for facial behaviour analysis, but it is more challenging than AU detection since distinguishing subtle changes between neighbor intensities are more difficult than recognizing the existence of AU.

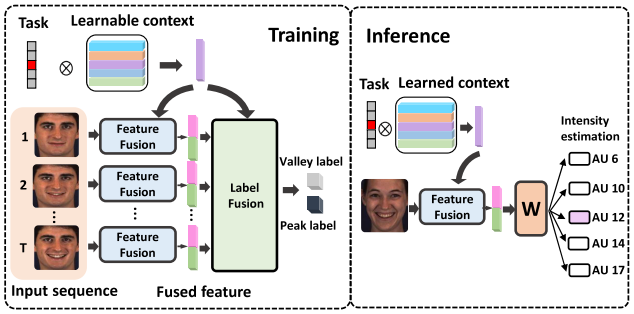
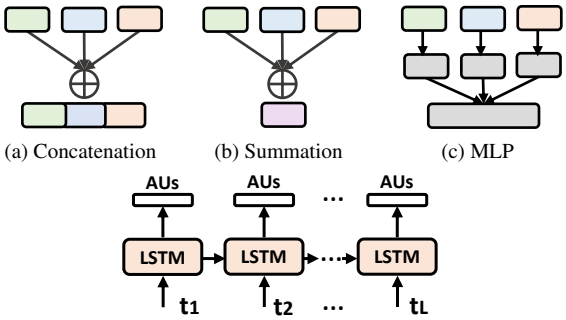


Figure 1. The training and inference phases of the proposed method. During training, we use a sequence as input and use the sequence level labels (intensity annotations of peak and valley frames) to provide supervision. Feature fusion and label fusion involve the enhanced attention mechanisms. During inference, we use a single frame as input as well as the learned task-related context to perform context-aware AU intensity estimation.

Most previous methods in AU detection [56, 54, 57, 19, 53] and AU intensity estimation [47, 50] focus on extracting features from a whole face image. A few region or patch-based methods [61, 12] extract features from local regions, since AUs are defined according to the facial appearance of local regions which contain informative patterns. Most deep learning methods such as [43, 48] directly feed a whole image into deep models, while only several methods [62, 20, 17, 25] consider extracting deep features from local patches. And these methods simply fuse features via concatenation, summation, or a multilayer perceptron (MLP) [17] (see Fig. 2). Note that these patch-based methods treat each patch equally during feature fusion without considering connections among patches or the relevance of a patch to the given AU. However, when annotating the intensity of an AU, we focus on the AU-related regions and ignore unrelated regions. Hence, patches should be treated

*Authors contributed equally
†Corresponding author



(a) Concatenation (b) Summation (c) MLP

(d) LSTMs for dynamics modeling

Figure 2. Existing strategies for feature fusion of multiple patches and dynamics modeling of sequence.

differently according to their associations to the given AU.

Though FACS provides the description of facial appearance of AUs at each intensity, it is quite laborious and expensive to annotate a large scale database [4]. A few methods leverage partially annotated databases to learn models for intensity estimation, including both shallow models [63, 35, 60, 59] and deep models [58]. These methods incorporate knowledge to provide additional supervision to compensate for the lack of intensity annotations. However, shallow models can only take pre-extracted features as inputs rather than raw images and they have limited representation capability. The end-to-end deep model [58] considers temporal relationships among only four frames randomly sampled from a video segment. It can not precisely reflect the temporal evolution of AUs in a consecutive segment. Besides, these methods train a model for each AU and can not perform joint intensity estimation of multiple AUs.

To alleviate the burden of labeling AU intensities and utilize informative local facial appearance, we propose a weakly supervised patch-based deep model on basis of two types of attention mechanisms to simultaneously estimate the intensities of multiple AUs. The whole pipeline is shown in Fig. 1. We only use intensity annotations on peak and valley frames (key frames) in videos instead of sparing efforts on labeling every frame. Then annotations of key frames are treated as sequence-level labels. Our model consists of a feature fusion module and a label fusion module, which are designed based on the following two observations. First, for each AU, its intensity label is only determined by local appearance of its related regions. Unrelated regions should be ignored. And each patch should contribute differently when analyzing different AUs. Second, each AU has its own rule to evolve temporally in sequences. When modeling the dynamics of AUs, each AU should be treated differently. Both the importance of a patch and the temporal dynamics should be modeled according to the given AU, rather than being the same for all AUs. To this end, we augment the two attention mechanisms with a learnable task-related (AU-related) context for feature and label fusion. The context-aware feature fusion is used to capture spatial relationships among local patches while the context-aware label fusion is used to capture temporal dynamics of AUs. The latter is the key that enables the model to be learned with partially labeled data. Enhanced attention mechanisms allow the model to predict feature attention and label attention adaptively according to the given AU.

Our contributions are two folds:

- We propose a novel weakly supervised patch-based deep model consisting of a feature fusion module and a label fusion module. The model can be trained on a partially annotated database, which greatly saves the effort of labeling AU intensities.

- We introduce a new strategy for better feature and label fusion by incorporating a learnable task-related context into two attention mechanisms. The enhanced attention mechanisms allow to learn task-related features and capture task-related temporal evolution of AUs.

2. Prior Work

Supervised learning methods. Most existing methods of AU intensity estimation are supervised learning methods which require a large set of fully annotated samples to achieve good performance, including frame-based and sequence-based methods. Frame-based methods [12, 14, 13, 26, 26, 52, 28, 27] learn an estimator to predict AU intensity from a single image, including relevance vector machine [12, 14], latent tree [13], multi-kernel SVM [26], and copula ordinal regression [47]. Sequence-based methods [31, 38, 32, 1, 33, 18] model the dynamics by considering the relations among multiple frames. Probabilistic graphical models are effective tools to capture the spatial and temporal dependencies of AU intensities, including Hidden Conditional Ordinal Random Field (H-CORF) [31], Kernel CORF [32], context-sensitive CORF [33], and Dynamic Bayesian Network[18]. Recently, several deep learning methods [7, 48, 43, 64, 15, 37, 49] are proposed for AU intensity estimation, including CNN [7], CCNN [48], 2DC [43], and HBN [49]. These supervised learning methods require that the database should be fully annotated, *i.e.,* every frame of sequences has the annotation of AU intensity, so that a good performance can be achieved. However, annotating the intensity of AU is more difficult than annotating the existence of AU even for trained AU coders. It is expensive and laborious to annotate a large scale database.

Weakly and semi-supervised learning methods. A few weakly or semi-supervised methods use sequence level labels or labels of partial frames for model learning. Multi-instance learning (MIL) [65] is a commonly used strategy to use sequence level labels, which has been applied to facial event detection [36, 41] and key frame detection [42, 40].

There is also a set of deep MIL methods [51] that combine the idea of MIL and deep learning. However, these methods focus on binary classification problem, which can not generalize to AU intensity estimation due that AU intensity has six ordinal scales. Several attention-based MIL methods [46, 29, 11] are proposed to predict the bag-level label rather than the instance-level label, which are not applicable for frame-level AU intensity estimation. Only two MIL methods are proposed for frame-level AU intensity estimation, *i.e.,* MI-DORF [34] and BORMIR [60]. MI-DORF uses the intensities of peak frames for training and requires a sequence as input for inference. BORMIR uses only the annotations of peak and valley frames and exploits different types of domain knowledge to provide weak supervision. Except for MIL, there are also several methods that use partial annotations for learning. Fernando *et al.* [3] propagate the AU label of peak frames to unlabeled frames by computing the similarity between features of one frame and the peak. Zhao *et al.* [63] combine ordinal regression and SVM and train a linear model by using the annotations of key frames. Zhang *et al.* [58] use the annotations of key frames and the relationships among four frames in tuples to learn a deep model. However, only four frames are not sufficient to capture detailed dynamics in sequences. Similarly, our method uses key frames to learn an AU intensity estimator as [3, 60, 63, 58]. Differently, we model the temporal dynamics adaptively according to the given task by integrating a task-related context to the attention mechanism. A Long Short-Term Memory (LSTM) network, which is used to capture temporal relations among frames, is designed to predict the label attention of each frame (see Fig. 4), rather than the AU labels [16, 2] (see Fig. 2d).

Patch-based methods. Patch-based methods extract features from informative local regions to reduce side effects of unrelated regions. Features of patches are fused via concatenation, summation, or MLP. Zhao *et al.* [61] extract features from regions around landmarks to form the final feature vector by concatenation. They further propose a deep region multi-label learning (DRML) method [62] for the detection of multiple AUs. Feature maps of patches are fused by convolution. Li *et al.* [17] multiply pre-defined attention map with feature maps of VGG and fuse cropped local feature maps by MLP. The fusion of cropped feature maps is also used in [16] and [21], along with LSTMs to capture temporal dynamics. Li *et al.* [20] use the same strategy as [61] to extract features around landmarks. They use an attention mechanism (*i.e.,* an additional branch to predict a weight map) to extract features for each patch and then fuse local features via MLP. No relations among patches are explicitly considered. Unlike these methods, our method uses enhanced attention mechanisms to capture the spatial relations among patches for feature fusion rather than feature extraction. We introduce a learnable task-related context to augment the attention mechanism since each patch should be treated differently for different AUs. We also consider the fusion of local and global features.

3. Proposed Approach

In this work, we propose a novel weakly supervised patch-based deep model on basis of two attention mechanisms for AU intensity estimation. The framework is demonstrated in Fig. 1. We first present the problem statement and then the feature and label fusion modules in Sec. 3.1 and Sec. 3.2, respectively. The objective function is defined in Sec. 3.3

Problem statement. Given a set of expression sequences with only AU intensity annotations of peak and valley frames (key frames), our goal is to learn a frame-level intensity estimator for multiple AUs. Key frames are identified following [63, 58, 23]. Given the key frames of an AU, sequences can be split into segments. In each segment, the AU intensity monotonically increases, decreases, or stays the same. We invert the frame order of segments that have decreasing AU intensity. Then, each segment evolves from a valley frame to a peak frame. Since locations of key frames are different for each AU, sequences are split for each AU individually.

Let one-hot vector $\mathbf{v} \in \mathcal{R}^K$ specify the category of AU, which we call task. For example, $\mathbf{v} = [1, 0, ..., 0]$ represents the first AU. K is the number of AUs. Let $\mathbf{X_v} = \{X_1, ..., X_T\}$ denote a segment with T frames of AU $\mathbf{v}$. X_t is the raw image of the t-th frame. Let $y^v_{\mathbf{v}} \in R$ denote the intensity of the valley frame X_1 and $y^p_{\mathbf{v}} \in R$ that of the peak frame X_T. They are sequence-level labels of $\mathbf{X_v}$. Given a partially labeled database $\mathcal{D} = \{\mathbf{X}_{\mathbf{v},n}, y^v_{\mathbf{v},n}, y^p_{\mathbf{v},n}, \mathbf{v}_n\}_{n=1}^N$, we learn a frame-level intensity estimator for multiple AUs, where N is the number of training segments.

3.1. Context-aware feature fusion

Previous patch-based methods simply fuse local features via concatenation, summation, or MLP (see Fig. 2). They treat each patch equally without considering their spatial relationships, namely the patch importance. We improve this by designing an attention module to fulfill the spatial relation. However, AU intensity is ought to be annotated according to the AU-related local patches regardless of unrelated ones, and the related patches of different AUs are different. To this end, we incorporate a learnable task-related context to augment attention mechanism for capturing spatial relations among local patches.

The framework of feature fusion module is shown in Fig. 3. The input face image is decomposed into M patches which contain local appearance of AUs. Features of patches are extracted and then fused by the enhanced attention mechanism along with the task-related context. Our method

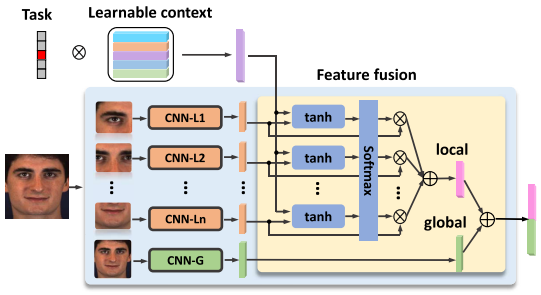


Figure 3. Context-aware feature fusion module

captures spatial relations by placing different importance on different spatial patches. Then, the fused features are concatenated with the global features which are extracted by feeding the whole face image into a dedicated CNN. Let $X = \{P_1, P_2, ..., P_M\}$ denote local patches of a frame X and P_0 denotes the resized image of X. The extraction of local and global features can be represented as

$$\mathbf{h}_m = g_m(P_m; \Theta_{g,m}), m \in \{0, 1, 2..., M\}, \quad (1)$$

where g_m is the m-th CNN and $\Theta_{g,m}$ is its parameters.

Previous methods such as [46] make the assumption that the bag-level features are the weighted summation of features of all instances, *i.e.,* $\mathbf{h} = \sum_{m=1}^{M} a_m \mathbf{h}_m$ where $\{a_m\}_{m=1}^{N}$ is a set of variables to learn. The variables do not depend on image representation. Then, methods such as [55] directly use latent features to predict attention weights as $\mathbf{h} = \sum_{m=1}^{M} a(P_m; X) \mathbf{h}_m$. The patch attention $a(P_m; X)$ depends on only the image. Attention values predicted by these methods are the same for different tasks. However, each patch plays different roles for analyzing different AUs. The attention value of a patch should be different when predicting different AUs. The attention mechanisms of these methods do not properly model the task-related spatial relationships among local patches.

To alleviate this issue, we improve the attention mechanism by incorporating a learnable task-related context. Let $\mathbf{C} \in \mathcal{R}^{K \times d_c}$ denote task-related context, which is a variable to learn. Each row is a context vector for one AU and d_c is the dimension of the context. The context for AU $\mathbf{v}$ is $\mathbf{c} = \mathbf{C}\mathbf{v}$. The fusion of local features can be represented as

$$\mathbf{h} = \sum_{m=1}^{M} a(P_m; X, \mathbf{C}, \mathbf{v}) \mathbf{h}_m, \quad (2)$$

where $a(P_m; X, \mathbf{C}, \mathbf{v})$ is a function to compute the spatial attention of P_m with consideration of all patches in X, the task $\mathbf{v}$, and the task-related context $\mathbf{C}$. The function of our context enhanced attention mechanism is

$$a(P_m; X, \mathbf{C}, \mathbf{v}) = \frac{\exp\{\mathbf{w}^T \tanh(\mathbf{W}_c \mathbf{C}\mathbf{v} + \mathbf{W}_h \mathbf{h}_m)\}}{\sum_j \exp\{\mathbf{w}^T \tanh(\mathbf{W}_c \mathbf{C}\mathbf{v} + \mathbf{W}_h \mathbf{h}_j)\}},$$

where $\mathbf{w}$, $\mathbf{W}_c$, and $\mathbf{W}_h$ are learnable parameters. Note that spatial relationships among patches are reflected in the computation of the attention which involves all patches, the task, and the context. When the task $\mathbf{v}$ varies, the attention of each patch changes accordingly. This design is consistent with the annotation process of AU intensity, where the same patch plays different roles in the intensity estimation of different AUs. The above fused features are local, but global features are also important. So we combine local and global features via concatenation since they are extracted under two different scales, *i.e.,*

$$\mathbf{f} = [\mathbf{h}, \mathbf{h}_0]. \quad (3)$$

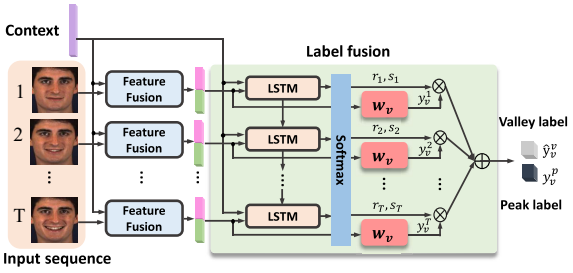


Figure 4. Context-aware label fusion module

3.2. Context-aware label fusion

Facial muscles cooperate with each other to perform meaningful expressions. Each AU has its own rule to evolve temporally. For example, appearance of some AUs changes rapidly during a period while appearance of some others changes slowly and smoothly. The temporal dynamics should be modeled according to the given AU, rather than using the same way for all AUs. Hence, we incorporate the learnable task-related context in the attention mechanism for label fusion to learn task-related dynamics of AUs. The framework of context-aware label fusion is shown in Fig. 4.

Given a segment of AU $\mathbf{v}$ and the intensity annotations of its peak frame and valley frame, *i.e.,* $\{\mathbf{X}_\mathbf{v}, y_\mathbf{v}^v, y_\mathbf{v}^p, \mathbf{v}\}$, features are first extracted for each frame of the input segment by using the feature fusion module (see Fig. 3). We denote output features of all frames as $\mathbf{F}_\mathbf{v} = \{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_T\}$, where $\mathbf{f}_t \in \mathcal{R}^{d_f}$ and d_f is the dimension of fused features. Features $\mathbf{F}_\mathbf{v}$ are then fed into a one-layer Long Short-term Memory (LSTM) network along with the context $\mathbf{C}\mathbf{v}$. For each time stamp, LSTM takes the concatenation of $\mathbf{f}_t$ and $\mathbf{C}\mathbf{v}$ as input, and outputs are a set of two-element pairs,

$$\{\hat{r}_t, \hat{s}_t\}_{t=1}^{T} = g_{\text{lstm}}(\mathbf{F}, \mathbf{C}, \mathbf{v}; \Theta_{\text{lstm}}), \quad (4)$$

where Θ_{lstm} denotes the parameters of LSTM, $\hat{r}_t = \hat{r}(\mathbf{f}_t; \mathbf{F}, \mathbf{C}, \mathbf{v})$, and $\hat{s}_t = \hat{s}(\mathbf{f}_t; \mathbf{F}, \mathbf{C}, \mathbf{v})$. We normalize the outputs through a softmax function, *i.e.,* $r_t = \exp\{\hat{r}_t\}/\sum_j \exp\{\hat{r}_j\}$ and $s_t = \exp\{\hat{s}_t\}/\sum_j \exp\{\hat{s}_j\}$. r_t is the temporal label attention with respect to the peak label at time stamp t and s_t is the temporal label attention with respect to the valley label. Note that each segment of T frames has only the intensity annotations of sequence-level

labels (*i.e.,* the annotations of the peak frame and the valley frame). Other frames are unlabeled. Inspired by [29] and [60], we use the assumption that the sequence-level label is a linear combination of frame-level labels for weakly supervised learning. The estimation of the sequence-level label is defined as

$$\tilde{y}^p_{\mathbf{v}} = \sum_{t=1}^{T} r(\mathbf{f}_t; \mathbf{F}, \mathbf{C}, \mathbf{v})\ y^t_{\mathbf{v}}, \tag{5}$$

$$\tilde{y}^v_{\mathbf{v}} = \sum_{t=1}^{T} s(\mathbf{f}_t; \mathbf{F}, \mathbf{C}, \mathbf{v})\ y^t_{\mathbf{v}}, \tag{6}$$

where $y^t_{\mathbf{v}}$ is the estimated intensity of AU $\mathbf{v}$ for the t-th frame by $y^t_{\mathbf{v}} = \mathbf{f}_t^T \mathbf{W}_o \mathbf{v}$. $\mathbf{W}_o \in \mathcal{R}^{K \times d_f}$ is an output matrix that maps the features to intensity labels of multiple AUs. $\mathbf{w}_{\mathbf{v}} = \mathbf{W}_o \mathbf{v}$ is the vector that corresponds to AU $\mathbf{v}$. The label attention $r(\mathbf{f}_t; \mathbf{F}, \mathbf{C}, \mathbf{v})$ and $s(\mathbf{f}_t; \mathbf{F}, \mathbf{C}, \mathbf{v})$ depend on the features of all frames and the task-related context. Here, the context-aware attention mechanism plays a similar role as feature fusion part, *i.e.,* allowing to model the temporal dynamics of AUs accordingly with respect to the given task. Note we use LSTMs to capture temporal dynamics in a novel way by predicting the temporal label attention, rather than predicting AU labels [16]. This enables the weakly supervised learning on partially labeled data. Previous works [29, 60] capture temporal information by optimizing weights to sum intensities from all frames. However, these weights do not depend on its corresponding image and do not explicitly encode the temporal relationships among multiple frames. Unlike them, our label attention function not only involves the image and the relationships among frames, but also incorporates the task-related context to model the dynamics according to the given task.

3.3. Objective functions

The proposed method requires only the intensity annotations of key frames. Given a partially labeled database $\mathcal{D} = \{\mathbf{X}_{\mathbf{v},n}, y^v_{\mathbf{v},n}, y^p_{\mathbf{v},n}, \mathbf{v}_n\}_{n=1}^N$, we define the loss of sequence-level labels for one sequence by computing L2 loss between the estimated sequece-level labels and the ground-truth, *i.e.,*

$$L_0 = (\tilde{y}^p_{\mathbf{v}} - y^p_{\mathbf{v}})^2 + (\tilde{y}^v_{\mathbf{v}} - y^v_{\mathbf{v}})^2. \tag{7}$$

Since AU intensity evolves from a valley frame to a peak frame in each training sequence, frames that are closer to the peak frame should have larger label attention values with respect to the peak label. Thus, the predicted label attention should satisfy $r_1 \leq r_2 ... \leq r_T$ and $s_1 \geq s_2 ... \geq s_T$. Let $\mathbf{r} = [r_1, r_2, ..., r_T]$ and $\mathbf{s} = [s_1, s_2, ..., s_T]$. The loss is defined as

$$L_1 = \sum_j [\max\{\mathbf{Ar}, 0\} + \max\{-\mathbf{As}, 0\}]_j, \tag{8}$$

where $\mathbf{A} \in \mathcal{R}^{T-1 \times T}$ is a matrix with $A_{i,i} = 1$, $A_{i,i+1} = -1$, and other elements being 0's.

In expression sequences, facial appearance changes smoothly, thus neighbor frames have similar facial appearance. We constrain label attention with a smoothness regularization,

$$L_2 = \frac{1}{2}(\mathbf{r}^T \mathbf{L} \mathbf{r} + \mathbf{s}^T \mathbf{L} \mathbf{s}), \tag{9}$$

where $\mathbf{L} = \mathbf{B} - \mathbf{C}$ is a Laplacian matrix. $\mathbf{C}$ is an adjacent matrix with $C_{i,j} = 1$ if $|i - j| = 1$. Other elements are 0's. $\mathbf{B}$ is a matrix with $B_{i,i} = \sum_j C_{i,j}$ with other elements being 0's. The objective function is defined as

$$L = L_0 + \lambda_1 L_1 + \lambda_2 L_2. \tag{10}$$

3.4. Training and Inference

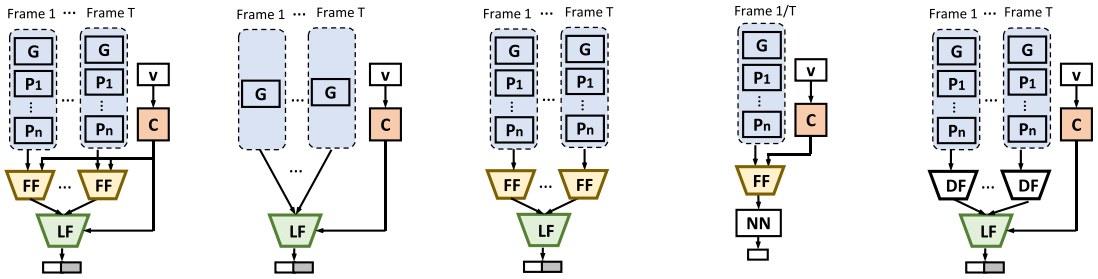
As shown in Fig. 1, **during training**, we use a segment as input of the patch-based deep model. The annotations of key frames are used as sequence-level labels to provide supervison. We capture temporal dynamics of AUs by a LSTM based context-aware attention mechanism. It builds the connection between labeled key frames and unlabeled frames, and enables the model to be trained with partially labeled data. **During testing**, the network takes a single frame as input and outputs its intensity of the given AU. Given a task (AU) $\mathbf{v}$, the learned task-related context ($\mathbf{c} = \mathbf{Cv}$) and the input frame X are fed into the feature fusion module. The fused feature vector $\mathbf{f}$ is mapped to the corresponding AU intensity through the learned output matrix $\mathbf{W}_o$, *i.e.,*

$$y_v = \mathbf{f}^T \mathbf{W}_o \mathbf{v}.$$

4. Experiments

4.1. Experimental setup

Datasets. FERA 2015 [44] and DISFA [24] are currently the two largest spontaneous expression databases for AU intensity estimation. FERA 2015 contains about 140,000 images from 41 subjects. Intensities are annotated for 5 AUs. Following the protocol of [44], we use 21 subjects for training and the other 20 subjects for testing. DISFA contains about 130,000 images from 27 subjects. Intensities are annotated for 12 AUs. We perform 3-fold subject independent cross validation, *i.e.,* 18 subjects for training and 9 subjects for testing. AU intensity has 6 ordinal scales in both databases. The distributions of the two databases are shown in Fig. 6. The percentage of key frames is about 2% in FERA 2015 and 1% in DISFA. Using only the annotations of key frames for learning would greatly save the effort of intensity annotation. Note that FERA 2017 [45], another expression database, is built for AU intensity estimation under different poses. It is not used here since we

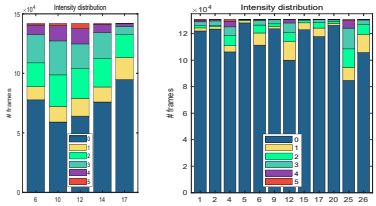


(a) CFLF (ours) (b) CFLF-S (c) CFLF-NC (d) CFLF-NL (e) CFLF-DF

Figure 5. Comparison between our method (CFLF) and its variants. 'G' represents the whole image and 'P_n' represents the n-th patch. 'v' represents the task index and 'C' represents the context. 'FF' is the feature fusion module and 'LF' is the label fusion module. 'DF' represents the feature fusion by directly concatenating feature vectors. Note that CFLF-NL is a supervised learning method which use one frame as input for training. Others are weakly supervised learning methods which use a sequence of T frames as input for training.

Table 1. Ablation study. This table presents the comparison between the proposed method and its variants.

Database		FERA 2015						DISFA												
AU		6	10	12	14	17	Avg	1	2	4	5	6	9	12	15	17	20	25	26	Avg
ICC	CFLF-S	.697	.637	.808	.421	.501	.614	.132	.152	.361	.199	.501	.314	.641	.105	.269	.189	.690	.390	.329
	CFLF-NL	.607	.492	.682	.282	.356	.484	.176	.161	.277	.173	.390	.186	.532	.069	.191	.162	.615	.406	.278
	CFLF-NC	.759	**.719**	.816	.364	.487	.629	.186	.176	.367	.342	.448	.326	.657	.194	.329	.229	.753	.446	.371
	CFLF-DF	.740	.701	.790	**.439**	.537	.641	.241	**.217**	.403	.211	.456	.315	.646	**.223**	.338	**.241**	.672	.480	.370
	CFLF	**.766**	.703	**.827**	.411	**.600**	**.661**	**.263**	.194	**.459**	**.354**	**.516**	**.356**	**.707**	.183	**.340**	.206	**.811**	**.510**	**.408**
MAE	CFLF-S	.835	.906	.666	1.036	.702	.829	.462	.329	.702	.134	.388	.316	.471	.227	.347	.197	.734	.488	.400
	CFLF-NL	.872	1.049	.895	1.100	.789	.941	.527	.493	.825	.295	.519	.399	.690	.359	.456	.358	.879	.552	.529
	CFLF-NC	.701	**.781**	**.621**	1.032	.621	.751	.347	.286	.655	.130	**.346**	**.258**	.438	.198	.304	.188	.610	.444	.350
	CFLF-DF	.691	.791	.720	1.151	**.608**	.792	.442	.355	.811	.178	.416	.319	.499	.245	.356	.248	.699	.471	.420
	CFLF	**.624**	.830	.624	**1.000**	.626	**.741**	**.326**	**.280**	**.605**	**.126**	.350	.275	**.425**	**.180**	**.290**	**.164**	**.530**	**.398**	**.329**



(a) FERA 2015 (b) DISFA

Figure 6. AU intensity distribution

focus on weakly supervised AU intensity estimation. EmotionNet [6] is a large database that contains annotations of AU occurrence and no intensity labels of sequences are provided. It is not applicable for AU intensity estimation.

Training. Sequences are split into segments according to the key frames. Given a segment, we sample a set of sub segments to construct training segments. Each training segment contains T frames including the peak frame and the valley frame. By using the provided facial landmarks, we crop M local regions around facial components rather than small regions around landmarks [17] since each component involves multiples AUs which are closely related. Detailed locations of patches are presented in the supplementary material. Each region is resized to the size of 32x32. The whole face is cropped out and resized into the size of 32x32. In feature fusion module, we use an individual ResNet18 [9] to extract features for each region and the whole face. In label fusion module, we use a one-layer LSTM net [10] to predict label attention. Both fusion modules are jointly trained from scratch. The number of patches is $M = 8$. The length of training segment is $T = 10$. We set the hyperparameters as $\lambda_1 = 0.1$ and $\lambda_2 = 0.01$. The batchsize is 128 and the learning rate is 0.01 with the decay rate of 0.95.

Evaluation. Intra-class Correlation [39] (ICC(3,1)) and Mean Absolute Error (MAE) are two commonly used evaluation metrics for AU intensity estimation [43, 48, 58]. We use the two metrics to evaluate the performance of the proposed method and competitive methods.

4.2. Ablation study

We first performe an ablation study to verify the effectiveness of the feature fusion module, the label fusion module, and the task-related context. Fig. 5 shows the composition of our method and its four variants. Our method (CFLF) contains both fusion modules and the context. CFLF-S has no feature fusion module and uses only the whole face. CFLF-NC drops the task-related context. It uses the standard attention mechanism for feature fusion, which can be treated as the weighted summation of features (see Fig. 2b). CFLF-NL drops the label fusion module. It becomes a supervised learning method and can not use unlabeled frames. CFLF-DF replaces the feature fusion module with straightforward feature concatenation (see Fig. 2a).

The results are shown in Table 1. Our method achieves the best average performance on both databases under two metrics. We analyze the results as follows. Firstly, our method outperforms CFLF-S which uses only the whole face image. This shows that features extracted from lo-

Table 2. Comparison with the state-of-the-art weakly-supervised and semi-supervised methods. The best results are shown in bold and in brackets. The second best results are shown in bold only. (*) Indicates results taken from the reference.

Database		FERA 2015						DISFA												
AU		6	10	12	14	17	Avg	1	2	4	5	6	9	12	15	17	20	25	26	Avg
ICC	Ladder [30]	.670	.619	.793	.073	.444	.520	-.012	.058	.040	.027	**.463**	.089	.596	-.015	.011	.000	.575	.369	.183
	LBA [8]	.706	.642	.812	.230	**.502**	.578	.080	.085	.363	.041	.379	.150	**[.738]**	.075	.242	.084	**[.830]**	.459	.294
	OSVR [63]*	.646	.577	.780	.269	.449	.544	**.208**	.038	.248	.151	.229	.152	.313	.115	.066	**.094**	.618	.093	.194
	BORMIR [60]*	.725	.675	**[.861]**	.368	.469	.620	.198	**[.248]**	.302	**.173**	.385	.181	.583	**.157**	.225	.088	.707	.148	.283
	KBSS [58]	**.760**	**[.725]**	**.840**	**[.445]**	.454	**.645**	.136	.116	**[.480]**	.169	.433	**.353**	**.710**	.154	**.248**	.085	.778	**[.536]**	**.350**
	CFLF (ours)	**[.766]**	**.703**	.827	**.411**	**[.600]**	**[.661]**	**[.263]**	**.194**	**.459**	**[.354]**	**[.516]**	**[.356]**	.707	**[.183]**	**[.340]**	**[.206]**	**.811**	**.510**	**[.408]**
MAE	Ladder [30]	.685	.838	**.599**	1.195	.640	.791	.647	.343	1.259	**.114**	**[.283]**	.327	**.354**	.187	**.304**	**.148**	.755	**[.390]**	.426
	LBA [8]	**.636**	**.802**	**[.560]**	1.097	**[.616]**	**.742**	**.357**	**[.258]**	**.786**	**[.078]**	**.313**	**[.169]**	**[.292]**	**[.138]**	.311	**[.106]**	**[.384]**	.422	**[.301]**
	OSVR [63]*	1.024	1.126	.953	1.354	.928	1.077	1.648	1.873	2.943	1.378	1.556	1.690	1.636	1.101	1.614	1.371	1.329	1.789	1.661
	BORMIR [60]*	.848	.895	.678	1.046	.791	.852	.875	.783	1.240	.589	.769	.777	.757	.564	.716	.628	.898	.875	.789
	KBSS [58]	.738	**[.773]**	.694	**[.990]**	.895	.818	.532	.489	.818	.237	.389	.375	.434	.321	.497	.355	.613	.440	.458
	CFLF (ours)	**[.624]**	.830	.624	**1.000**	**.626**	**[.741]**	**[.326]**	**.280**	**[.605]**	.126	.350	**.275**	.425	**.180**	**[.290]**	.164	**.530**	**.398**	**.329**

cal patches contain useful information that is not covered by the global features. Secondly, CFLF-NL achieves the poorest performance. It does not contain label fusion and can use only annotated key frames to perform supervised learning. It overfits the limited number of training samples. The label fusion is the key to enable weakly supervised learning with partially labeled data. Thirdly, our method achieves better performance than CFLF-NC which does not use the task-related context. CFLF-NC uses the standard attention mechanism for feature fusion and uses the LSTMs without context for label attention prediction. Therefore, it is unable to capture the spatial and temporal relations in different tasks and each AU is treated equally. The comparison shows the importance of the task-related context. Fourthly, our method outperforms CFLF-DF which replaces the attention-based feature fusion with straightforward feature concatenation. CFLF-DF treats local patches equally while our method treats patches differently according to the given task. Our method is more consistent with the way that we annotate AU intensity since we focus on AU-related local regions and ignore unrelated regions. These results demonstrate the effectiveness of the feature fusion module, the label fusion module, and the task-related context.

4.3. Comparison with the state-of-the-art

Comparison with weakly and semi-supervised learning methods. We compare the proposed method with several state-of-the-art weakly-supervised learning methods (OSVR [63], BORMIR [60] and KBSS [58]) and semi-supervised learning methods (Ladder [30] and LBA [8]). Ladder uses unlabeled samples by designing a denoising loss. LBA propagates labels of labeled samples to unlabeled samples based on the assumption that samples with similar labels have similar latent features. OSVR combines ordinal regression with SVM for expression intensity estimation. BORMIR uses the idea of multi-instance regression and uses domain knowledge to provide weak supervision. KBSS uses knowledge-based losses based on a four-element tuple to train a deep model. Note that OSVR, BORMIR, and KBSS train one model for each AU. Unlike them, we train one model to jointly predict the intensities of multiple AUs. Our method and competitive methods require only the intensity annotations of key frames for training.

The results are shown in Table 2. We analyze the results as follows. Firstly, on FERA 2015, our method achieves superior average performance over other methods under both metrics. On DISFA, our method achieves the best average performance under ICC and the second best under MAE. Note that ICC and MAE should be jointly considered to evaluate one method. Though Ladder and LBA can get good MAEs, their ICCs are much worse than KBSS and our method. Because they do not consider temporal relations among frames for regularization and overfit the labeled samples of the training set. As intensity distributions of two databases are imbalanced and the majority intensity is 0, they always predict the majority intensity for testing frames. It results in that they have good MAEs, but low ICCs. Secondly, our method outperforms KBSS on both databases. During the training phase, our method uses a segment consisting of T frames as the input while KBSS uses a four-element tuple sampled from a training segment. However, four frames are not enough to capture the temporal dynamics in segments, especially when facial appearance changes rapidly. Our method can capture the dynamics better than KBSS since we consider more frames. Besides, our method uses two types of context augmented attention mechanisms to capture the spatial relations among patches and temporal dynamics of AUs. Another advantage of our method over KBSS is that we train one model for the joint intensity estimation of multiple AUs while KBSS trains one model for each AU. Thirdly, compared to OSVR and BORMIR, our method performs much better, especially on DISFA. They are two linear models using hand-craft features while our model is a deep model that can model more complex data distribution. These results show the superior performance of the proposed method over competitive weakly and semi-supervised learning methods.

Comparison with patch-based methods. We compare our method with two state-of-the-art patch-based methods of facial behaviour analysis, *i.e.,* EAC [17] and DRML [62]. EAC extracts cropped feature maps around facial landmarks and fuses them through a MLP (see Fig. 2c). DRML di-

Table 3. Comparison with the state-of-the-art patch-based methods under two scenarios.

Database		FERA 2015						DISFA												
AU		6	10	12	14	17	Avg	1	2	4	5	6	9	12	15	17	20	25	26	Avg
*Using the intensity annotations of **all frames** for training*																				
ICC	EAC [17]	.705	.643	.844	.328	.452	.594	.088	.077	.302	.144	.462	.150	.705	.090	.273	.141	.820	.367	.301
	DRML [62]	.731	.676	.813	.366	.476	.612	.093	.057	.415	.157	.408	.266	.718	.175	.189	.113	.805	.547	.329
MAE	EAC [17]	.762	.866	.612	1.067	.723	.806	.483	.464	.858	.099	.406	.416	.445	.246	.370	.238	.508	.517	.421
	DRML [62]	.731	.863	.675	1.279	.717	.853	.446	.380	.808	.079	.357	.299	.360	.165	.281	.142	.535	.382	.353
*Using only the intensity annotations of **key frames** for training*																				
ICC	EAC [17]	.496	.597	.754	.030	.018	.379	.000	-.004	.000	-.002	**.524**	-.002	.438	.000	.000	.001	.497	-.001	.121
	DRML [62]	.606	.521	.620	.089	.243	.416	-.055	-.073	.335	.044	.427	.179	.531	.001	.124	.001	.757	.413	.224
	CFLF (ours)	**.766**	**.703**	**.827**	**.411**	**.600**	**.661**	**.263**	**.194**	**.459**	**.354**	.516	**.356**	**.707**	**.183**	**.340**	**.206**	**.811**	**.510**	**.408**
MAE	EAC [17]	.898	.890	.735	1.156	.822	.900	.493	.380	.782	.200	.400	.337	.624	.249	.466	.248	.798	.622	.467
	DRML [62]	.874	1.040	.902	1.037	.864	.944	.546	.598	.858	.189	.356	.490	.454	.229	**.278**	.306	.552	.507	.447
	CFLF (ours)	**.624**	**.830**	**.624**	**1.000**	**.626**	**.741**	**.326**	**.280**	**.605**	**.126**	**.350**	**.275**	**.425**	**.180**	.290	**.164**	.530	**.398**	**.329**

vides feature maps into patches and applies the proposed region layer on each patch. The resulting feature maps of patches are then fused via a convolutional layer. Note that both methods are proposed for AU recognition, rather than AU intensity estimation. We adapt them for intensity estimation by replacing the classification loss with a regression loss. We evaluate EAC and DRML under two scenarios, *i.e.,* using the intensity annotations of all frames for training and using the intensity annotations of only key frames.

The results are shown in Table 3. When using only the annotations of key frames, our method achieves much better results than EAC and DRML on both databases. Even when EAC and DRML use the annotations of all frames, our method still outperforms them. The reason is that they overfit the training data even using all frames. They have higher training accuracies than our method, but have lower testing accuracies. The training sets of both databases have less than 90,000 images while their models have millions of parameters to train. This leads to the overfitting. Once fewer annotations are used, their performance drops sharply. Differently, we perform weakly supervised learning and use two regularization terms in the objective, which avoids the overfitting to a certain extent.

Comparison with supervised learning methods. We compare with several state-of-the-art supervised learning methods of AU intensity estimation, including CNN [7], ResNet18 [9], 2DC [43], CCNN-IT [48], HBN [49], and Heatmap [37]. CNN [7] uses a four-layer CNN for intensity estimation. HBN uses the hybrid Bayesian networks. ResNet18 is the starndand Resnet with 18 layers. 2DC combines Gaussian Process and variational autoencoder. CCNN-IT combines copula funcitons, CRF, and CNN. Heatmap jointly predicts the locations of AUs and their intensities. These supervised methods require annotating AU intensity of each frame in sequences while ours needs only the annotations of key frames. The key frames occupy only about 2% in FERA 2015 and 1% in DISFA.

The results are shown in Table 4. Our method achieves better performance under MAE on both databases. On FERA 2015, our ICC is better than CNN and CCNN-IT, and is comparable to HBN, Heatmap, and 2DC. On DISFA, our ICC is better than CNN, ResNet18, and CCNN-IT. Please note that we only use the intensity annotations of key frames in sequences while other methods use that of all frames. CNN and ResNet18 have the same overfitting problem as EAC and DRML. The results show that our method can still achieve comparable or even better performance than the competitive supervised learning methods though we use much fewer annotations.

Table 4. Comparison with the state-of-the-art supervised methods. Note that competing methods use every annotated frame in the training set while our method uses the intensity annotations of partial frames which occupy a very small portion. (*) Indicates results taken from the reference.

Database	FERA 2015		DISFA	
Method	ICC	MAE	ICC	MAE
HBN [49]*	**.700**	-	-	-
Heatmap [37]*	.680	-	-	-
2DC [43]*	.660	-	**.494**	-
CCNN-IT [48]*	.630	1.260	.377	.663
CNN [7]	.596	.817	.328	.423
ResNet18 [9]	.580	.882	.270	.483
CFLF (ours)	.661	**.741**	.408	**.329**

5. Conclusion

This paper proposes a novel weakly supervised patch-based deep model on basis of two types of attention mechanisms for the joint intensity estimation of multiple AUs. We explore spatial relationships among local patches with a feature fusion module, while incorporating temporal dynamics of AUs with a label fusion module to enable weakly supervised learning. The attention mechanisms of both modules are further enhanced with a learnable task-related context, which facilitates joint analysis of multiple AUs and boosts performances. Evaluations on two public benchmark databases demonstrate the effectiveness of the feature fusion module and the label fusion module.

Acknowledgments: This work is partially funded by the National Natural Science Foundation of China (61802362, 61620106003), Beijing Nova Program (Z171100001117048), and Youth Innovation Promotion Association CAS (2015361). Qiang Ji's involvement in this work is supported in part by the US National Science Foundation award CNS No. 1629856.

References

[1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Continuous conditional neural fields for structured regression. In *ECCV*, 2014. 2

[2] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. In *FG*, 2017. 3

[3] Fernando De la Torre, Tomas Simon, Zara Ambadar, and Jeffrey F Cohn. Fast-facs: A computer-assisted system to increase speed and reliability of manual facs coding. In *ACII*, 2011. 3

[4] Gianluca Donato, Marian Stewart Bartlett, Joseph C Hager, Paul Ekman, and Terrence J Sejnowski. Classifying facial actions. *TPAMI*, 1999. 2

[5] Paul Ekman. Facial action coding system (facs). *A human face*, 2002. 1

[6] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *CVPR*, 2016. 6

[7] Amogh Gudi, H Emrah Tasli, Tim M Den Uyl, and Andreas Maroulis. Deep learning based facs action unit occurrence and intensity estimation. In *FG*, 2015. 2, 8

[8] Philip Haeusser, Alexander Mordvintsev, and Daniel Cremers. Learning by association-a versatile semi-supervised training method for neural networks. In *CVPR*, 2017. 7

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6, 8

[10] Sepp Hochreiter and Schmidhuber Jrgen. Long short-term memory. In *Neural computation*, 1997. 6

[11] Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*, 2018. 3

[12] Sebastian Kaltwang, Ognjen Rudovic, and Maja Pantic. Continuous pain intensity estimation from facial expressions. In *IVC*, 2012. 1, 2

[13] Sebastian Kaltwang, Sinisa Todorovic, and Maja Pantic. Latent trees for estimating intensity of facial action units. In *CVPR*, 2015. 2

[14] Sebastian Kaltwang, Sinisa Todorovic, and Maja Pantic. Doubly sparse relevance vector machine for continuous facial behavior estimation. *TPAMI*, 2016. 2

[15] Liandong Li, Tadas Baltrusaitis, Bo Sun, and Louis-Philippe Morency. Edge convolutional network for facial action intensity estimation. In *FG*, 2018. 2

[16] Wei Li, Farnaz Abtahi, and Zhigang Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *CVPR*, 2017. 3, 5

[17] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. In *FG*, 2017. 1, 3, 6, 7, 8

[18] Yongqiang Li, S Mohammad Mavadati, Mohammad H Mahoor, Yongping Zhao, and Qiang Ji. Measuring the intensity of spontaneous facial action units with dynamic bayesian network. *PR*, 2015. 2

[19] Yongqiang Li, Baoyuan Wu, Bernard Ghanem, Yongping Zhao, Hongxun Yao, and Qiang Ji. Facial action unit recognition under incomplete data based on multi-label learning with missing labels. *PR*, 2016. 1

[20] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *TIP*, 2019. 1, 3

[21] Chen Ma, Li Chen, and Junhai Yong. Au r-cnn: Encoding expert prior knowledge into r-cnn for action unit detection. *arXiv preprint arXiv:1812.05788*, 2018. 3

[22] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. *TAC*, 2017. 1

[23] Mohammad Mavadati, Peyten Sanger, and Mohammad H Mahoor. Extended disfa dataset: Investigating posed and spontaneous facial expressions. In *CVPR Workshops*, pages 1–8, 2016. 3

[24] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *TAC*, 2013. 5

[25] Chuanneng Mei, Fei Jiang, Ruimin Shen, and Qiaoping Hu. Region and temporal dependency fusion for multi-label action unit detection. In *ICPR*, 2018. 1

[26] Zuheng Ming, Aurélie Bugeau, Jean-Luc Rouas, and Takaaki Shochi. Facial action units intensity estimation by the fusion of features with multi-kernel support vector machine. In *FG*, 2015. 2

[27] Mohammad Reza Mohammadi, Emad Fatemizadeh, and Mohammad H Mahoor. Intensity estimation of spontaneous facial action units based on their sparsity properties. *IEEE transactions on cybernetics*, 2015. 2

[28] Mohammad R Mohammadi, Emad Fatemizadeh, and Mohammad H Mahoor. An adaptive bayesian source separation method for intensity estimation of facial aus. *TAC*, 2017. 2

[29] Nikolaos Pappas and Andrei Popescu-Belis. Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis. In *EMNLP*, 2014. 3, 5

[30] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *NIPS*, 2015. 7

[31] Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation. In *CVPR*, 2012. 2

[32] Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. Automatic pain intensity estimation with heteroscedastic conditional ordinal random fields. In *IVC*, 2013. 2

[33] Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. *TPAMI*, 2015. 2

[34] Adria Ruiz, Ognjen Rudovic, Xavier Binefa, and Maja Pantic. Multi-instance dynamic ordinal random fields for weakly-supervised pain intensity estimation. In *ACCV*, 2016. 3

[35] Adria Ruiz, Ognjen Rudovic, Xavier Binefa, and Maja Pantic. Multi-instance dynamic ordinal random fields for weakly supervised facial behavior analysis. *TIP*, 2018. 2

[36] Adria Ruiz, Joost Van de Weijer, and Xavier Binefa. Regularized multi-concept mil for weakly-supervised facial behavior categorization. In *BMVC*, 2014. 2

[37] Enrique Sanchez, Georgios Tzimiropoulos, and Michel Valstar. Joint action unit localisation and intensity estimation through heatmap regression. *BMVC*, 2018. 2, 8

[38] Georgia Sandbach, Stefanos Zafeiriou, and Maja Pantic. Markov random field structures for facial action unit intensity estimation. In *ICCV workshop*, 2013. 2

[39] Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 1979. 6

[40] Karan Sikka, Abhinav Dhall, and Marian Bartlett. Weakly supervised pain localization using multiple instance learning. In *FG*, 2013. 2

[41] Karan Sikka, Gaurav Sharma, and Marian Bartlett. Lomo: Latent ordinal model for facial analysis in videos. In *CVPR*, 2016. 2

[42] David MJ Tax, E Hendriks, Michel François Valstar, and Maja Pantic. The detection of concept frames using clustering multi-instance learning. In *ICPR*, 2010. 2

[43] Dieu Linh Tran, Robert Walecki, Ognjen Rudovic, Stefanos Eleftheriadis, Bjørn Schuller, and Maja Pantic. Deepcoder: Semi-parametric variational autoencoders for automatic facial action coding. *ICCV*, 2017. 1, 2, 6, 8

[44] Michel F Valstar, Timur Almaev, Jeffrey M Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F Cohn. Fera 2015-second facial expression recognition and analysis challenge. In *FG*, 2015. 5

[45] Michel F Valstar, Enrique Sánchez-Lozano, Jeffrey F Cohn, László A Jeni, Jeffrey M Girard, Zheng Zhang, Lijun Yin, and Maja Pantic. Fera 2017-addressing head pose in the third facial expression recognition and analysis challenge. In *FG*, 2017. 5

[46] Kiri L Wagstaff and Terran Lane. Salience assignment for multiple-instance regression. 2007. 3, 4

[47] Robert Walecki, Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. Copula ordinal regression for joint estimation of facial action unit intensity. In *CVPR*, 2016. 1, 2

[48] Robert Walecki, Ognjen Rudovic, Vladimir Pavlovic, Björn Schuller, and Maja Pantic. Deep structured learning for facial action unit intensity estimation. In *CVPR*, 2017. 1, 2, 6, 8

[49] Shangfei Wang, Longfei Hao, and Qiang Ji. Facial action unit recognition and intensity estimation enhanced through label dependencies. *TIP*, 2019. 2, 8

[50] Shangfei Wang, Bowen Pan, Shan Wu, and Qiang Ji. Deep facial action unit recognition and intensity estimation from partially labelled data. *TAC*, 2019. 1

[51] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *PR*, 2018. 3

[52] Philipp Werner, Sebastian Handrich, and Ayoub Al-Hamadi. Facial action unit intensity estimation and feature relevance visualization with random regression forests. In *ACII*, 2017. 2

[53] Baoyuan Wu, Siwei Lyu, Bao-Gang Hu, and Qiang Ji. Multi-label learning with missing labels for image annotation and facial action unit recognition. *PR*, 2015. 1

[54] Shan Wu, Shangfei Wang, Bowen Pan, and Qiang Ji. Deep facial action unit recognition from partially labeled data. In *ICCV*, 2017. 1

[55] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 4

[56] Jiabei Zeng, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Zhang Xiong. Confidence preserving machine for facial action unit detection. In *ICCV*, pages 3622–3630, 2015. 1

[57] Yong Zhang, Weiming Dong, Bao-Gang Hu, and Qiang Ji. Classifier learning with prior probabilities for facial action unit recognition. In *CVPR*, 2018. 1

[58] Yong Zhang, Weiming Dong, Bao-Gang Hu, and Qiang Ji. Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation. In *CVPR*, 2018. 2, 3, 6, 7

[59] Yong Zhang, Baoyuan Wu, Weiming Dong, Zhifeng Li, Wei Liu, Bao-Gang Hu, and Qiang Ji. Joint representation and estimator learning for facial action unit intensity estimation. In *CVPR*, 2019. 2

[60] Yong Zhang, Rui Zhao, Weiming Dong, Bao-Gang Hu, and Qiang Ji. Bilateral ordinal relevance multi-instance regression for facial action unit intensity estimation. In *CVPR*, 2018. 2, 3, 5, 7

[61] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Honggang Zhang. Joint patch and multi-label learning for facial action unit detection. In *CVPR*, 2015. 1, 3

[62] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *CVPR*, 2016. 1, 3, 7, 8

[63] Rui Zhao, Quan Gan, Shangfei Wang, and Qiang Ji. Facial expression intensity estimation using ordinal information. In *CVPR*, 2016. 2, 3, 7

[64] Yuqian Zhou, Jimin Pi, and Bertram E Shi. Pose-independent facial action unit intensity regression based on multi-task deep transfer learning. In *FG*, 2017. 2

[65] Zhi-Hua Zhou. Multi-instance learning: A survey. *Department of Computer Science & Technology, Nanjing University, Tech. Rep*, 2004. 2