Knowledge-Spreader: Learning Semi-Supervised Facial Action Dynamics by Consistifying Knowledge Granularity

Xiaotian Li Xiang Zhang Taoyue Wang Lijun Yin
State University of New York at Binghamton
{xli210, zxiang4, twang61, lyin}@Binghamton.edu

Abstract

Recent studies on dynamic facial action unit (AU) detection have extensively relied on dense annotations. However, manual annotations are difficult, time-consuming, and costly. The canonical semi-supervised learning (SSL) methods ignore the consistency, extensibility, and adaptability of structural knowledge across spatial-temporal domains. Furthermore, the reliance on offline design and excessive parameters hinder the efficiency of the learning process. To remedy these issues, we propose a lightweight and online semi-supervised framework, a so-called ***Knowledge-Spreader (KS)****, to learn AU dynamics with sparse annotations. By formulating SSL as a Progressive Knowledge Distillation (PKD) problem, we aim to infer cross-domain information, specifically from spatial to temporal domains, by consistifying knowledge granularity within Teacher-Students Network. Specifically, KS employs sparsely annotated key-frames to learn AU dependencies as the* privileged knowledge. *Then, the model spreads the learned knowledge to their unlabeled neighbours by jointly applying knowledge distillation and pseudo-labeling, and completes the temporal information as the* expanded knowledge. *We term the progressive knowledge distillation as "Knowledge Spreading", which allows our model to learn spatial-temporal knowledge from video clips with only* ***one label*** *allocated. Extensive experiments demonstrate that KS achieves competitive performance as compared to the state of the arts under the circumstances of using only* ***2%*** *labels on BP4D and* ***5%*** *labels on DISFA. In addition, we have tested it on our newly developed large-scale comprehensive emotion database BP4D++, which contains considerable samples across well-synchronized and aligned sensor modalities for alleviating the scarcity issue of annotations and identities.*

1. Introduction and Related Work

Facial action unit (AU) detection plays a vital role in automatic facial action analysis. Over the past few years, the deep neural networks [46, 10, 18, 19, 48] trained on large scale data have become the de facto model for AU detection. However, a lab-controlled AU video typically contains thousands of frames that need to be densely labeled by human annotators.

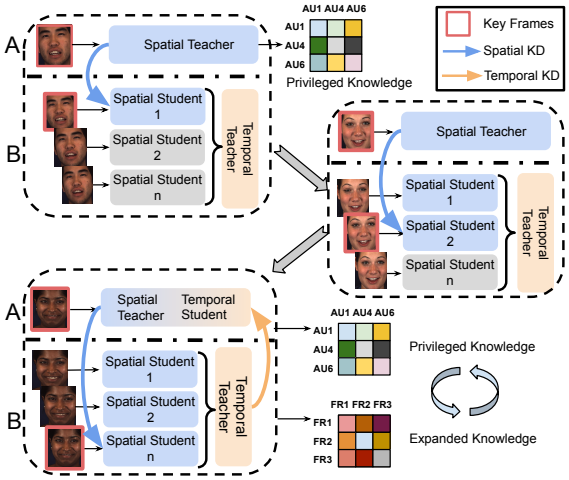

Figure 1: Overall pipeline of "Knowledge Spreader". The blue color indicates the *privileged knowledge* (AU dependencies) learned by Spatial Teacher on sparse annotations. After fully distilled to the Spatial Students, the *privileged knowledge* promotes to generate a sequential pseudo labels with high confidence, and complete the temporal relationship as the *expanded knowledge* (orange color). *expanded knowledge* can build a more powerful Temporal Teacher that feeds the learned knowledge to Temporal Student (Spatial Teacher) reversely.

Semi-supervised learning (SSL) aims to tackle the issue where the labeled instances are inadequate, yet a large amount of unlabeled data are easy to obtain. Some early SSL work [38, 44] summarized the distribution or co-occurrence statistics from existing ground-truth AU labels, as the knowledge constrain, to detect AUs using partially labeled data. These models often yield unsatisfactory performance due to lacking sufficient supervision information. Besides, only applying the prior distribution or manually defined knowledge may fall into sub-optimal due to lacking adaptation mechanism. For instance, some scholars

[35, 22] pointed out that the real world AU distribution or dependencies may vary in terms of individual's expressions, age, gender, and ethnicity. Most recent advances [28, 25, 5, 39] achieved excellent performance improvement by utilizing a larger external dataset as the auxiliary information. However, these improvements often come at the cost of consuming millions or larger sized data, which indicates the intensive demand of annotation and data is still not effectively alleviated. Besides, most of current research trends to design static SSL from the spatial perspective. Whereas, some recent work [3, 20] have demonstrated that explicitly relationship modeling of contiguous frames is also an important factor for robust AU detection. Applying these image-level SSL methods on dynamic corpus is challenging due to some nuisance factors such as motion blur, video defuse, and frequent pose occlusions. The aforementioned issues motivate us to explore a balanced SSL approach that can reduce the need of dense annotations, while maximize the knowledge acquisition by expanding information from spatial-wise to temporal-wise.

Knowledge Distillation (KD) was initially proposed by [9] to reduce the computational cost of the deep neural networks. After that, [40] proposed a new paradigm for training model with auxiliary knowledge named "Learning Using Privileged Information" (LUPI). Here the privileged information, which serve as the additional descriptions to the training data, can only be used by the teacher. By combining the privilege information and KD into a unified framework, [21] proposed a novel framework "Generalized Distillation" (GD). Following this, [6] is the first work to formulates SSL as a generalized distillation problem. They adopted the textual explanation in Wikipedia as the privileged information for guiding the student network to learn with unlabeled data. A recent work Noisy Student [45] improved the idea of self-training (pseudo-labeling) and knowledge distillation with the use of noise added to the student networks. Inspired by these work, we propose to formulate SSL as as a progressive knowledge distillation (PKD) problem. Here we define PKD as a process that the learning system can actively infer *expanded knowledge* by utilizing previously acquired *privileged knowledge*. In this work, the *privileged knowledge* indicates the spatial knowledge (i.e., inter-action relation) learned by self-attention on sparse labeled data, while the *expanded knowledge* is the temporal knowledge (i.e., inter-frame relation) learned on the combination of labeled and unlabeled data. This seamless progressive knowledge distillation enhances performance of dynamic semi-supervised learning. The presented dual-network exploits the gaps in their knowledge granularity (i.e., privileged versus non-privileged knowledge and expanded versus non-expanded knowledge) as a unique perturbation for consistency regularization.

The overall pipeline of the proposed "Knowledge-Spreader" (KS) is presented in Figure 1. In term of the model designing, KS sparsely samples the annotations by every k frames in the training data pool, and feeds these labeled key frames to branch A. As the limited key frames can support fully-supervision, the Spatial Teacher in A utilizes a standard transformer [41] to encode the spatial AU dependencies as the initial *privileged knowledge*. At the same time, KS randomly pick $n-1$ neighbours around key frames as the unlabeled data. These neighbours, with the corresponding key frame, form an n-frame sequence as the input of B. KS sets several light-weighted Spatial Students to accommodate every single frame of the input sequence. By maximizing the representational similarity between the outputs from Spatial Teacher and Spatial Students, and shifting the active knowledge distillation target (Spatial Student), the privileged spatial knowledge can be gradually spread to every Spatial Students. Thus, through the learned *privileged knowledge*, the Spatial Students can infer more confident pseudo labels from unlabeled frames. The pseudo labels complete the supervision information of an entire video clip. Thus, another transformer-based Temporal Teacher in B is capable of learning the temporal dependencies as the *expanded knowledge*. We term the progressive knowledge distillation as the "Knowledge Spreading" which is the core component of KS. Intuitively, the basis behind this is supported by the two major aspects: (1) *privileged knowledge* is a better heuristic for getting pseudo labels with high-confidence; (2) integrating the sporadic power of students can pursue *expanded knowledge* in a higher dimension, and build a sound teacher network.

Our contribution lies in three-fold: (1) to the best of our knowledge, the paper is the first work to formulate semi-supervised learning as a Progressive Knowledge Distillation (PKD) problem, which focuses on knowledge extensibility, consistency of knowledge granularity, and model efficiency. (2) this work explores a novel semi-supervised setting for dynamic databases where the labeled data and unlabeled data is not independent. (3) We have built a new spontaneous emotion database by capturing 3D geometric facial sequences, 2D facial videos, thermal videos, and physiological data sequences from 233 participants across two years period. The new database will be released to the research community along with the paper being published.

2. Methodology

2.1. Overview

Our goal is to detect facial action units using sparse video clips that contains only single-frame labels. Specifically, we assume the facial action labels are available by every k frames in the training set. The ith video clip, which comprises n continuous frames, is represented as a set $V^i = \left\{Z^i_{U,1}, Z^i_{U,2}, \cdots, Z^i_{L,m}, \cdots, Z^i_{U,n}\right\}$ where Z^i_U de-

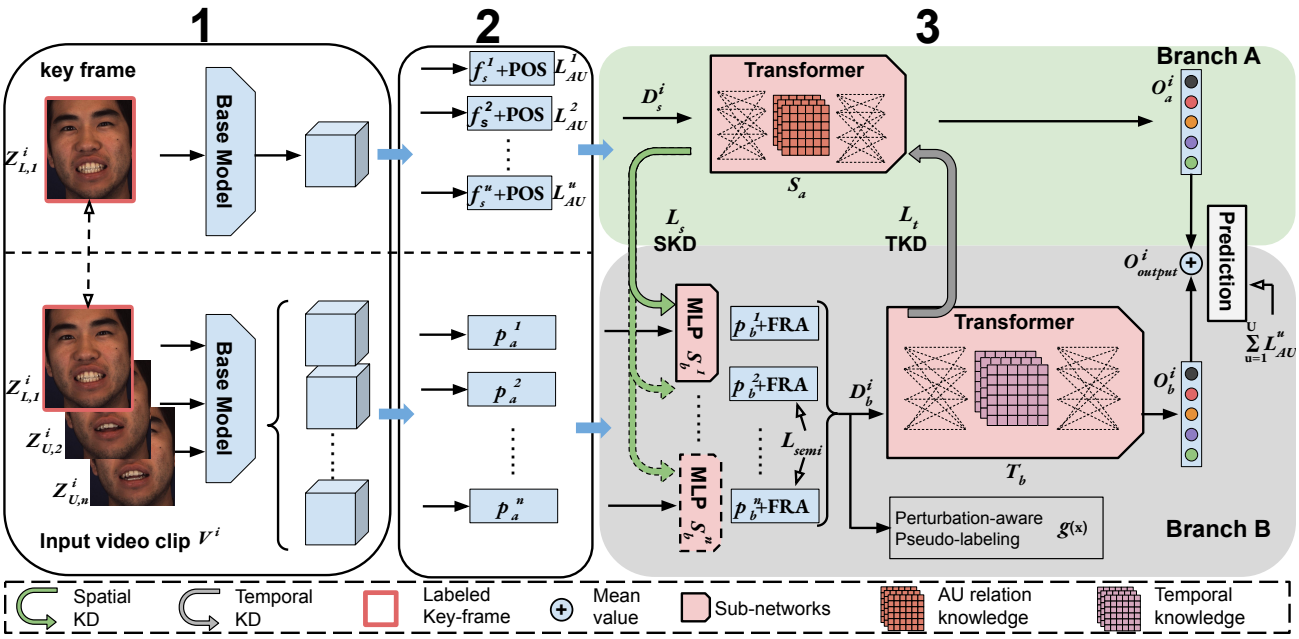

Figure 2: **Illustration of the detailed architecture for Knowledge-Spreader.** In stage 1, the base model extracts a sequential latent features. In stage 2, two group of the project heads are utilized to learn AU-specific features f_s^u and initial frame-specific features p_a^n. In stage 3, The Spatial Teacher S_a dynamically distills (green arrow) the privileged AU dependency knowledge to the Spatial Students S_b. The Temporal Teacher T_b completes the inter-frame relation knowledge, and distills (grey arrow) the expanded knowledge to S_a for optimal fusion. In the example, the key-frame of V^i is 1, and S_b^1 is the active student for spatial knowledge distillation (SKD).

notes unlabeled frames and $Z_{L,m}^i$ denotes the labeled key frame. Note that the location of the key frame m is not fixed, and it cyclically switches from 1 to n during training. We use the video clips and their key frames as the coupled inputs of our framework. The network structure of KS is depicted in Figure 2.

2.1.1 Latent Feature Learning

We first project the raw data into the latent space based on several considerations: (1) capturing the underlying patterns and achieving high-level semantics of deep features. By doing so, it creates a compressed and more meaningful representation that facilitates better understanding and interpretation of the data; (2) compressing the data dimension for computational and memory efficiency, making the network more scalable and practical for large-scale datasets; and (3) disentangling latent representations that correspond to specific attributes (e.g., facial action units) for fine-grained feature learning.

As shown in the first stage of Figure 2 , the image-level and video-level features are extracted by the base models which share the same weights. We adopt ResNet-18 [8] pre-trained on ImageNet [29] as the feature extractor, obtaining a sequential $512 \times 7 \times 7$ feature maps from the last convolutional layer.

In the second stage, KS aims to disentangle AU-specific features and frame-specific features. To generate AU-specific features f_s^u, we employ global average pooling (GAP) to flatten the extracted feature of key frames from stage one, and project it into multiple linear heads. Each head is supervised by a binary AU label independently for activating the specific AU local region. The corresponding loss function L_{AU} can be found in Equation (3). Likewise, we use GAP to achieve the initial frame-specific embeddings p_a^n, where n is the length of input video clip. Each feature corresponds to one frame in the video clips. Due to the lack of supervision for unlabeled non-key frames, these initial frame-specific features are considered to be less confident, meaning that they cannot be directly forwarded to the Temporal Teach T_b for learning temporal relation. As the Spatial Teacher S^a is fully supervised with the key frames, KS sets a group of student networks S_b for accepting privileged knowledge from the Spatial Teacher, and further infers trustable pseudo labels to supervise the learning of unlabeled frames. The loss function can be found in Equation (4) and Equation (5). Finally, the post frame-specific embeddings p_b^n with higher confidence are obtained in the third stage.

2.1.2 Spatial and Temporal Relation Learning

As shown in the third stage of Figure 2, we design a transformer [41] based module to learn both spatial AU de-

pendencies by S_a and temporal inter-frame context by T_b. In this work, the inter-AU and inter-frame correlation are represented as the multi-head attention matrix. A self-diversified attention design [18, 19] can be optionally applied to address the redundancy and over-parameterization issues across multiple attention heads. First of all, We generate u learnable positional embeddings (POS in Figure 2), and add them with high-level AU-specific features f_s^u. These features is concatenated and represented as $D_s^i = \{f_s^1, f_s^2, \cdots, f_s^u\}$, where u refers to the number of AU classes. Then, D_s^i is fed to a standard transformer encoder S_a. The Spatial Student S_a learns the AU relation as privileged knowledge. The latent weight matrix for multi-head attention is defined as: $MultiHead(Q, K, V) = Concat(Head_1, ..., Head_h)W$, where the ith attention matrix is calculated as:

$$Head_i(Q_i, K_i, V_i) = softmax(\frac{Q_i K_i^T}{\sqrt{d_k}})V_i \quad (1)$$

where $Q_i = W_{q,i}X_i, K_i = W_{k,i}X_i$ and $V_i = W_{v,i}X_i$. Q_i, K_i, and V_i represent a set of query, key, and value respectively. This module constructs a set of linear transformations based on the normalized layer input $X_i = LayerNorm(D_s^i)$. d_k is the dimension of query and key. Consequently, $MultiHead(Q, K, V)$ is applied to value V. Along with the residual connection, the output of attention module is calculated as $O_{Attention} = MultiHead(Q, K, V)V + D_s^i$. The feed-forward layer contains two linear transformation and a GELU non-linearity. By adding its output with $O_{Attention}$, the output O_a^i in branch A is calculated. Similarly, The post frame-specific features p_b^n are extracted by The Spatial Students S_b^n. The module adds them with the corresponding frame-positional embeddings (FRA in Figure 2). These features is concatenated and represented as $D_b^i = \{p_b^1, p_b^2, \cdots, p_b^n\}$, and fed to the Temporal Teacher T_b for learning temporal context information. Note that the frame-positional embeddings are utilized to address the order invariance issue for standard Transformer, letting the model be aware of the impact of temporal perturbations on predictions. Although KS does not necessarily guarantee the perception of spatial action order, we adopt the same design to maintain consistency with the temporal branch, aligning high-frequency information.

2.1.3 Knowledge Spreading

Knowledge Spreading is the core module of the proposed framework. As shown in the stage 3 of Figure 2, we conduct five steps to (1) set n Spatial Students S_b; (2) process Spatial Knowledge Distillation (SKD) from the Spatial Teacher S_a to one active Spatial Student S_b. The active student is determined by the key-frame's location number m where m equals B mod n, and B is the Bth batch of training samples, n is the input clip length; (3) shift the location of key frame to activate Spatial Students alternatively; (4) use the Spatial Students to generate pseudo labels for non-key-frames, and train the Spatial Students on the combination of labeled images and pseudo labeled images; (5) process Temporal Knowledge Distillation (TKD) from the Spatial Teacher T_b to the Temporal Student S_a (former Spatial Teacher), and ensemble the output O_a^i and O_b^i for prediction.

We adopt dual-level and progressive knowledge distillation in our model. Since obtaining fine-grained privileged knowledge requires full supervision, a single network is inadequate for achieving semi-supervised learning directly. Therefore, our approach employs a teacher network as the proxy for gradually transferring learned knowledge to semi-supervised student networks. In the initial level of Spatial Knowledge Distillation, our framework forwards a single labeled key-frame along two distinct branches, introducing perturbations through data augmentations and adjustments in knowledge granularity. This unique treatment leads to disparate predictions from the same input, prompting KS to seek maximally coherence between these predictions with and without privileged knowledge. The manipulation of knowledge granularity introduces an innovative perspective on feature perturbation, central to consistency-driven learning. Spatial Teacher, fed with limited yet fully-supervised data, becomes the repository of fine-grained spatial action relationships as its privileged knowledge. Conversely, Spatial Students, fueled by abundant yet less assured dense samples, encompass broader spatial insights of a coarse-grained nature. By enforcing knowledge consistency across spatial Teacher-Students, a unified acquisition of privileged knowledge transpires, subsequently applicable to unlabeled data. Knowledge Distillation fully compresses the scale of Spatial Students, following the design philosophy of Noisy Student. These Students, denoted as S_b, comprise a set of n mini MLP networks, each smaller in scale compared to the Spatial Teacher. The integration of multiple Spatial Students is applied for enhancing system robustness. Inspired by recent work [7], we adopt an online distillation method using KL divergence for collaborative learning. Here, the KL divergence loss is used to minimize the probability distribution of teacher-student networks and their corresponding ensembles. The KL loss function of Spatial Knowledge Distillation is defined as:

$$L_{skd} = \frac{1}{b}\left(\sum_{i=1}^{n} T^2 KL\left(p_i, q_i\right) + \sum_{i=1}^{n} T^2 KL\left(w_i^m, q_i\right)\right) \quad (2)$$

where b is the batch size, T is the temperature parameter. p and w denote the soften probability distribution calculated by the Spatial Teacher S_a and the active Spatial Student, respectively. The soft target q is expressed as $q = softmax(z_s/T)$, where z_s is performed by the mean

pooling of the outputs from both teacher and active student. m is the key frame position used for determining the active Spatial Student. m equals B mod n, where B indicates the Bth batch of training samples, n is the clip length. This setting ensures each Spatial Student has almost the same chance to be selected as the active one. Here, the temperature parameter is set as 1. Due to the issue of data imbalance which may cause the performance degradation, we choose weighted BCE with logits as the multi-label classification loss. The function is defined as:

$$L_{AU} = \sum_{i=1}^{u} w_i[y_i log\sigma_i(x) + (1-y_i)\, log\,(1-\sigma_i(x))] \tag{3}$$

where w_i is calculated by the ith AU's occurrence ratio [30], and less likely occurred AUs have higher weights. $\sigma(x)$ is the corresponding predicted probability. y_i is the ground truth. We further employ the Pseudo-label [13] to train the inactive Spatial Students, injecting learned knowledge into unlabeled data. The loss function is defined as:

$$L_{pd} = \sum_{i=1}^{u} w_i[\hat{y}_i log\sigma_i(x) + (1-\hat{y}_i)\, log\,(1-\sigma_i(x))] \tag{4}$$

where x denotes p_a^n, and $\hat{y}_i$ denotes the learned pseudo labels by picking up the class which has the maximum predicted probability. The total loss function of semi-supervised learning is denoted as:

$$L_{semi} = \sum_{i=1}^{n-1} L_{pd}^i \tag{5}$$

In the subsequent phase of Temporal Knowledge Distillation, the enriched privileged knowledge within Spatial Students facilitates the assimilation of finer-grained spatial-temporal action relationships as expanded knowledge. Consequently, a new disparity in knowledge granularity arises between Branch A and B. Through the secondary-level consistency, the Temporal Student is guided to incorporate domain-expanded information (i.e., temporal knowledge), even when only sparse labels are available. The KL loss function of Temporal Knowledge Distillation is defined as

$$L_{tkd} = \frac{1}{b}\left(\sum_{i=1}^{n} T^2 KL\,(p_i, q_i) + \sum_{i=1}^{n} T^2 KL\,(w_i, q_i)\right) \tag{6}$$

where p and w denote the soften probability distribution calculated by the Temporal Teacher T_b and the Temporal Student S_a, respectively. The loss functions of Spatial Knowledge Distillation and Temporal Knowledge Distillation are expressed as:

$$L_s = \sum_{i=1}^{z} L_{AU}^i + \alpha L_{skd}, L_t = \sum_{i=1}^{z} L_{AU}^i + \alpha L_{tkd} \tag{7}$$

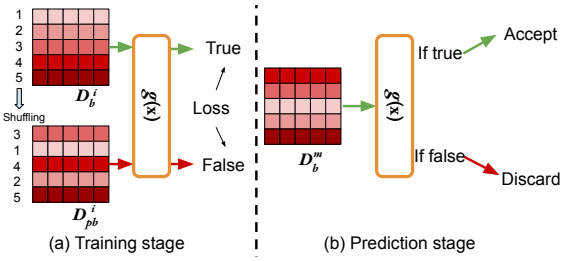

Figure 3: Pipeline of perturbation-aware pseudo-labeling. At the training stage, it feeds D_b^i (without shuffling) and D_{pb}^i (randomly shuffling D_b^i along the timeline of the features) into the model. At the prediction stage, we let the feature D_b^m (without shuffling) as the input. If it is classified as false, the corresponding pseudo labels will be discarded and vice versa. The training stage starts from the first epoch. By observation, we find the simple binary classifier takes only one or two epochs to convergence. Thus, we set the third epoch as the flag to start the prediction stage.

where α is the trade-off weight, z is the number of Teacher-Student networks. Although branch B contains multiple student networks $\{S_b^1, S_b^2, ..., S_b^n\}$, only one of them is selected as the active student for processing the loss function within the same iteration. Here, z is 2, and α is 0.5.

2.1.4 Perturbation-aware Pseudo-labeling (PPL)

We exploit a simple yet effective self-supervised module to determine confident pseudo labels by predicting if the features contain any temporal perturbation. Since facial muscles move gradually and smoothly over time, the pseudo labels generated by incorrect features signify the anomalies in the temporal domain. Inspired by the work [42], we use the temporal feature shuffling to simulate the sequential perturbation and generate negative feature samples. By applying this module, KS does not need to set any hard threshold for filtering low-confident pseudo labels. The auxiliary task is designed with a binary classifier $g(x)$. It is jointly trained with the AU detection task. Specifically, We first label the ith collection of sequential features D_b^i as True. We duplicate D_b^i as D_{pb}^i, and randomly shuffle it along the temporal axis. D_{pb}^i is labeled as False. Afterward, D_b^i and D_{pb}^i are fed into classifier $g(x)$ for training. L_{self} denotes the loss of the binary classification task. Figure 3 illustrates the pipeline of the proposed module. As shown in Figure 4, the negative samples show the irregular pattern of AU occurance in a short period of time. By detecting the temporal perturbation of the features, the proposed module is able to filter low-quality pseudo labels from a global perspective. Figure 4 illustrates how PPL senses and processes the incorrect pseudo labels.

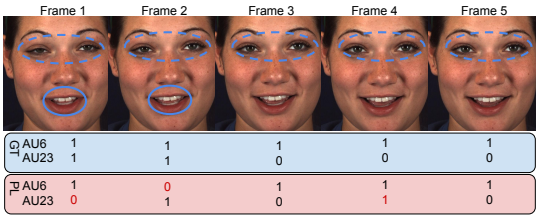

Figure 4: Pseudo labels discarded by PPL. The circles on the face indicate the occurrence of AU6 and AU23. "GT" and "PL" indicate the ground-truth and the pseudo labels. Red number means incorrectly generated pseudo labels. The ground-truth labels usually present stable changing patterns, while the low-quality pseudo labels with temporal perturbation present the abnormal pattern where AU occurs and disappears repeatedly in a short-term period.

2.1.5 Overall loss function and algorithm

The total loss function is expressed as follow:

$$L_{total} = L_{AU} + L_s + \lambda w_{ramp} \left(L_t + L_{self} + L_{semi} \right) \quad (8)$$

where λ is the trading-off hyper-parameter, and w_{ramp} is a ramp-up function [12] to make sure the semi-supervised modules and self-supervised module converge relatively slow compared with the fully-supervised module. w_{ramp} is a simple Gaussian curve function:

$$w_{ramp} = exp\left(-\omega (1 - \frac{(x-\mu)^2}{\sigma^2}) \right) \quad (9)$$

where x is the epoch number, and ω indicates the height of the Gaussion curve's peak. In this paper, we set $\omega = 2$, $\mu = 0$, and $\sigma = 5$. Here, the w_{ramp} increases to 1 after 5 epochs' warming-up.

In this work, we adopts the inductive strategy for the semi-supervised learning. At inference time, all modules are used except the pseudo-labeling. The algorithm of KS is shown in Algorithm 1. More details such as transformer design, and feature dimension are included in the supplementary material.

Algorithm 1 Pseudocode of Knowledge-Spreader

Require: The input frame F_l^i, the input clip V^i and its frame number N, the position of key frame m. B means the Bth batch of training sample. Functions of the models in branch A: base model $b_\theta(x)$, Model S_a $f_\theta(x)$. Functions of the models in branch B: base model $b_\sigma(x)$, sub-networks in S_b^n with supervision $l_\sigma(x)$ and pseudo-labeling $k_\sigma(x)$, The AU detection classifier of Model T_b $f_\sigma(x)$ and the binary classifier for self-supervised learning $g_\sigma(x)$.

for each epoch *E* **do**
  for each mini-batch *b* **do**
    $O_{abase}^i \leftarrow b_\theta(F_l^i); O_{bbase}^{(i,n)} \leftarrow b_\sigma(V^i)$
    $O_a^i \leftarrow f_\theta(O_{abase}^i)$
    for $q = 1, \ldots, N$ **do**
      if $q == B \bmod N$ **then**
        $O_k^i \leftarrow l_\sigma(O_{bbase}^{(i,q)})$
        process spatial KD with O_a^i and O_k^i
      else
        process pseudo-labeling with $O_{ps}^{(i,n-1)} \leftarrow k_\sigma(O_{bbase}^{(i,q)})$
      end if
    end for
    $O_{bti}^i \leftarrow$ Concatenate$(O_{ps}^{(i,n-1)}, O_k^i)$
    $O_b^i \leftarrow f_\sigma(O_{bti}^i)$
    $O_{pb}^i \leftarrow f_\sigma($Shuffle$(O_{bti}^i))$
    $O_{ssl}^i \leftarrow g_\sigma(O_b^i = 0, O_{pb}^i = 1)$
    if $E <= 2$ or $O_{ssl}^i == 1$ **then**
      Let the weight of L_{semi}, $\lambda_4 = 0$
    end if
    process temporal KD with O_b^i and O_a^i
    $O_{output}^i \leftarrow$ MeanValue(O_b^i, O_a^i)
    Update θ and σ via SGD of equation 8
  end for
end for

3. Experiments

3.1. Datasets

BP4D [47] and DISFA [23] are are widely used benchmark databases for dynamic AU detection. We followed the experimental setting of the previous work [15] to evaluate our approach for a fair comparison.

New MME: The existing facial action datasets are limited in terms of subjects number, diversity, and metadata. Thanks to the existing available multi-modal datasets [47] [49], we extend to develop a new larger-scale multi-modal emotion (MME) database, which consists of **233** participants (132 females and 101 males). The data is significantly expanded in terms of participants number as compared to the existing databases: DISFA (27 subjects) [23], MMI (44 subjects) [26], BP4D (41 subjects) [47], BP4D+ (140 subjects) [49]. Following ethical principles, our data collection was approved by the institutional review board (IRB). Each subject signed an informed consent form. A professional performer/interviewer applied a procedure containing 10 seamlessly-integrated tasks as [47, 49] that resulted in effective elicitation of spontaneous emotions. The dataset was well-synchronized and aligned with multi-modalities including 3D geometric facial model, 2D facial videos, thermal videos, and physiology data sequences (*e.g.* heart rate, blood pressure, skin conductance (EDA), and respiration rate). Around 94,000 frames were well-annotated by three expert FACS coders for AU coding. More details are described in the supplemental material. The new database is ready for public and will be released to the research community by the time of the paper being published.

3.2. Implementation Details

We process the image by cropping off redundant area which is not relevant to face recognition. Then the images are resized as $224 \times 224 \times 3$ to fit the model. Each of the training images is randomly rotated, flipped horizontally, and with color jitters (saturation, contrast, and brightness) for data augmentation. We choose SGD as the optimizer with a learning rate of 0.01 for 50 epochs. The model was implemented with Pytorch framework. The hyper-

parameters in Equation (8) are set as $\lambda = 0.5$. Each video clips contains 5 frames.

3.3. Model analysis

3.3.1 Comparison with semi-supervised methods

Figure 5 shows the performance compared with semi-supervised methods from two areas (AU detection and general action recognition). We carefully investigated the existing works that adopt limited labels for AU detection. BGCS [44] and DAUR [38] are selected for comparison. Figure 5 (d), (e), and (f) shows the proposed model achieves significant performance improvement. Considering our foundation model may have advantages in generalization ability, we can compare the performance trend. With the available labels decreasing (from 90% to 50%), Knowledge-Spreader shows no obvious performance attenuation. It worth noting KS is trained from scratch. Thus, some models from [28, 39, 25, 5, 1] are not considered for comparison, as they use full annotation pools, extra data, fine-tuning, or other jointly trained tasks. We further report the comparison results with some semi-supervised methods from general action recognition for a comprehensive evaluation, including Pseudo-label [13], FixMatch [34], and a video-level TCL [33]. Compared with the conventional setting of previous AU works [44, 38, 27, 28], the percentages of available AU annotations are significantly reduced (1%, 2%, 5%, 10%, 20%, 50%, 60%, 70%, 80%, 90%, and 100%) to explore where the limit of KS is. Figure 5 (a), (b), and (c) shows prominent improvement of KS, especially when extremely limited annotations are available (1%, 2%, 5% , 10%, 15% and 20% on BP4D; 1%, 2%, and 5% on DISFA; 1%, 2%, 5% , 10%, 15% and 20% on MME). The quantitative results with different label ratios are shown in Table 1.

Table 1: Quantitative comparison with semi-supervised methods using F1 score. Underlines indicate the best results.

Model	BP4D	DISFA	MME
Pseudo-label (1%)	54.3	40.4	45.8
Pseudo-label (2%)	57.8	50.8	47.5
Pseudo-label (5%)	59.7	51.5	52.1
Pseudo-label (10%)	60.7	56.8	54.2
Pseudo-label (15%)	61.2	57.1	54.9
Pseudo-label (20%)	62	58.5	55.2
Pseudo-label (50%)	63.6	57.9	55.3
Pseudo-label (60%)	62.7	56.7	55.3
Pseudo-label (70%)	63.3	57.9	55.3
Pseudo-label (80%)	62.4	58.3	56.6
Pseudo-label (90%)	62.3	57.5	55.5
Pseudo-label (100%)	62.7	58.8	56.9
FixMatch (1%)	49.9	35.6	41.6
FixMatch (2%)	55.1	46.2	46.5
FixMatch (5%)	59.2	52.7	52.6
FixMatch (10%)	60.5	55	55.4
FixMatch (15%)	62.1	57.7	55.6
FixMatch (20%)	62	58.4	56.4
FixMatch (50%)	62	57.9	58.3
FixMatch (60%)	62.1	56	56.4
FixMatch (70%)	61.9	57.8	57.2
FixMatch (80%)	62.2	56.9	55.5
FixMatch (90%)	61.9	57.5	55.3
FixMatch (100%)	62.7	58.8	56.9
TCL (1%)	55.6	42.3	43.3
TCL (2%)	58.9	51.2	48.2
TCL (5%)	60.5	53.6	53.4
TCL (10%)	61.7	55.8	55.7
TCL (15%)	62.3	56.7	56.2
TCL (20%)	62.7	57.9	55.6
TCL (50%)	63.2	59.2	57.6
TCL (60%)	62.8	60.1	57.9
TCL (70%)	63.0	59.6	57.9
TCL (80%)	62.9	60.4	58.3
TCL (90%)	62.7	58.3	57.8
TCL (100%)	63.1	59.7	58.1
Our KS (1%)	59.9	49.4	51.2
Our KS (2%)	62.5	52.8	54.8
Our KS (5%)	63.9	56.9	57.6
Our KS (10%)	64.4	58	58.4
Our KS (15%)	64.5	58.8	58.7
Our KS (20%)	64.4	59.5	58.9
Our KS (50%)	64.5	61.6	59.5
Our KS (60%)	64.4	62.9	59.4
Our KS (70%)	64.5	61.9	59.5
Our KS (80%)	64.4	62	59.4
Our KS (90%)	64.6	62.2	59.6
Our KS (100%)	64.7	62.8	59.7

3.3.2 Comparison with supervised methods

We report the results under two training setups by following [43]: (1) Compare KS against the fully-supervised state-of-the-art methods with 100% labeled data. (2) Compare KS against a supervised counterpart under different training label ratios. As shown in Table 2, a collection of recent and strong benchmark algorithms including JAA [31], DSIN [4], LP [24], ARL [32], SRERL [16], SRERL [16], UGN [37], SEV [46], HMP-PS [36], HMP-PS [36], FAUDT [10], and EAC (1%) [17] are selected for a comprehensive evaluation. Knowledge-Spreader outperforms all other advances using only 10% labels on BP4D and 50% labels on DISFA. KS still performs competitively using only 2% labels on BP4D and 5% labels on DISFA. In addition, the experiments conducted with 100% labels show the effeteness of Knowledge-Spreader in a supervised manner, where the main contribution comes from Spatial-Temporal relation learning module. Especially, it shows that KS surpasses the best benchmark (FAUDT) by 1.3 f1-score on DISFA. Table 3 further shows the comparison results in terms of individual AUs. The proposed method performs best on 9 out of 12 AUs on BP4D and 3 out of 8 AUs on DISFA.

3.4. Data Structure Analysis

Through a experiment, we find the non-overlapping AU annotations (different AU combinations) account for only a small proportion of the overall frames number (1693 out of 140,000 frames on BP4D, 102 out of 130,000 frames on

Figure 5: Comparison with other advanced semi-supervised algorithms using different percentages of labels on BP4D, DISFA, and MME.

Table 2: Comparison with state-of-the-art methods using F1 score. The first table indicates the results of other methods using 100% labeled data. The second table indicates the results of a baseline model (left) and the proposed KS (right) using different percentages of labels. Underlines indicate the best result of other methods. Bold numbers indicate KS surpasses others' best performance.

Model	Reference	BP4D	DISFA
JAA	ECCV'18	60.0	56.0
DSIN	ECCV'18	58.9	53.6
LP	CVPR'19	61.0	56.9
ARL	AC'19	61.1	58.7
SRERL	AAAI'19	62.1	55.9
UGN	AAAI'21	63.3	60.0
SEV	CVPR'21	63.9	58.8
HMP-PS	CVPR'21	63.4	61.0
FAUDT	CVPR'21	<u>64.2</u>	<u>61.5</u>

Model	BP4D	DISFA	Model	BP4D	DISFA
EAC (1%)	43.8	31.8	KS (1%)	59.9	49.4
EAC (2%)	48.7	33.3	KS (2%)	62.5	52.8
EAC (5%)	52.2	39.4	KS (5%)	63.9	56.9
EAC (10%)	54.8	43.9	KS (10%)	**64.4**	58
EAC (50%)	55.6	48.0	KS (50%)	**64.5**	**61.6**
EAC (100%)	<u>56.3</u>	<u>51.2</u>	KS (100%)	**64.7**	**62.8**

DISFA, 748 out of 94,000 frames on MME). A large number of similar labels and data densely exist across adjacent frames. It reveals that why using only a few **sparsely sampled** clips and annotations can achieve competitive or even better performance, which is consistent with the "less is better" principle from [14]. Different from existing video-level semi-supervised works [2, 42, 33, 11] that adopt continuous annotations, we sparsely sample the annotations and allocate only one annotation by every k frames. Figure 6 demonstrates that applying our method can keep more non-

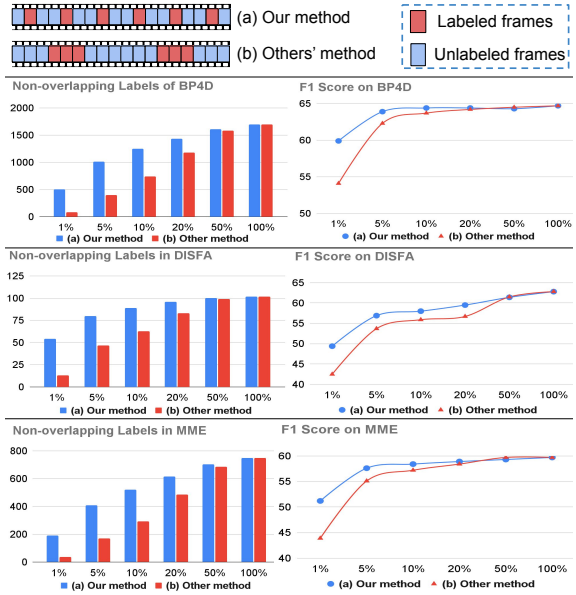

Figure 6: Evaluation of the label sampling methods. It shows the quantitative statistics of non-overlapping labels and the performance comparison (F1 score) using different methods on BP4D, DISFA, and MME. X-axis indicates the percentage of used labels.

overlapping AU annotations than conventional approaches using the same percentage of annotations.

3.5. Ablation Study

In this section, we justify the effectiveness of the key components in our proposed Knowledge-Spreader under the semi-supervised condition. All experiments are con-

Table 3: Comparison with state-of-the-art methods using F1 score in terms of individual AUs. The upper part is the F1 score on BP4D; The bottom part is the F1 score on DISFA. Bold numbers indicate the best performance.

Model	Used labels	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	Avg.
ARL	100%	45.8	39.8	55.1	75.7	77.2	82.3	86.6	58.8	47.6	62.1	47.4	55.4	55.4
SRERL	100%	46.9	45.3	55.6	77.1	78.4	83.5	**87.6**	63.9	52.2	**63.9**	47.1	53.3	62.9
UGN	100%	54.2	46.4	56.8	76.2	76.7	82.4	86.1	64.7	51.2	63.1	48.5	53.6	63.3
HMP-PS	100%	53.1	46.1	56.0	76.5	76.9	82.1	86.4	64.8	51.5	63.0	49.9	54.5	63.4
FAUDT	100%	51.7	49.3	61.0	77.8	79.5	82.9	86.3	**67.6**	51.9	63.0	43.7	56.3	64.2
Our KS	15%	**58.7**	**50.3**	**62.0**	**79.5**	75.4	**84.9**	87.1	65.9	45.5	62.9	48.3	53.3	64.5
Our KS	100%	55.3	48.6	57.1	77.5	**81.8**	83.3	86.4	62.8	**52.3**	61.3	**51.6**	**58.3**	**64.7**

Model	Used labels	AU1	AU2	AU4	AU6	AU9	AU12	AU25	AU26	Avg.
ARL	100%	43.9	42.1	63.6	41.8	40.0	**76.2**	**95.2**	66.8	58.7
SRERL	100%	45.7	47.8	59.6	47.1	45.6	73.5	84.3	43.6	55.9
UGN	100%	43.3	48.1	63.4	49.5	48.2	72.9	90.8	59.0	60.0
HMP-PS	100%	38.0	45.9	65.2	50.9	**50.8**	76.0	93.3	**67.6**	61.0
FAUDT	100%	46.1	48.6	**72.8**	**56.7**	50.0	72.1	90.8	55.4	61.5
Our KS	15%	41.7	53.5	69.7	41.3	46.2	72.0	92.3	54.0	58.8
Our KS	100%	**53.8**	**59.9**	69.2	54.2	**50.8**	75.8	92.2	46.8	**62.8**

ducted with using 50% labels.

Effect of the Relation Learning module: We replace the Transformer-based relation encoders S_a and T_b with the vanilla MLP to learn the spatial and temporal information. The results shows that the F1 Score drops to 62.8% from 64.5% on BP4D and 60.2% from 61.6% on DISFA. Removing the two knowledge encoders separately results different performance degradation. The F1 score decreases by a margin of 1.3% and 1.1% without S_a, while it decreases by a margin of 0.9% and 1.2% without T_b. This demonstrates that the impact of the two sub-modules is not uniform on different databases.

Effect of the Knowledge Spreading: We perform an experiment by removing the loss L_s and L_t for knowledge distillation module. We observe that the F1 score decreases by 2.0% and 1.5% on BP4D and DISFA. By removing SKD individually, the result shows the performance degradation by a margin of 1.3% and 1.1%. Without TKD, it decreases by 0.9% and 1.2%. The strategy of combining them as an integrated module can achieve the optimal effect.

Effect of the Perturbation-aware Pseudo-labeling The module is consisted of two parts including Pseudo-labeling and a self-supervised module with loss function L_{self}. By removing the whole module, we observe the performance degradation by a margin of 0.8% and 1.0% on BP4D and DISFA. By only removing Pseudo-labeling, the F1 score decreases by 0.6% and 0.7%. By replacing the self-supervised module with the hard threshold as the standard of confirming high-confident pseudo labels, we observe a performance drop by a margin of 0.3% and 0.4%. In addition, we compare the accuracy of pseudo labels generated by PPL and naïve pseudo-labeling [13] on BP4D using 10% labels. The result shows 76.35% accuracy on PPL and 73.36% on naïve pseudo-labeling. That demonstrates the performance of PPL improves by filtering the low quality pseudo labels with temporal perturbation. Another interesting finding is that if we only keep the loss L_{self} of PPL (not for label selection), the experimental results are also be reduced. That shows the auxiliary task in PPL benefit KS to learn better feature representation and inter-frame relation by identifying temporal disturbances.

Complexity analysis The proposed model (15.6 million) stands in stark contrast to the models with ResNet50 backbone (23 million) and ViT base (86.9 million) commonly used in existing AUD works, resulting in a highly compact architecture. To compare the training/testing speed, we run the models (e.g., our model, ResNet50-based model, and ViT-based model) using samples from the BP4D dataset with image size 224x224. The training/testing speed in terms of samples per second (S/s) is 156.3 (ours), 111.8 (ResNet50-based model), and 46.5 (ViT-based model). More detailed ablation studies can be found in the supplementary material.

4. Conclusion

In this paper, we have proposed a deep unified semi-supervised framework "Knowledge-Spreader". We formulate semi-supervised learning as a Progressive Knowledge Distillation (PKD) problem, which aims to infer domain-expanded information by consistency learning of knowledge granularity. By spreading the knowledge from the spatial domain to the temporal domain, KS can effectively alleviate the demand of dense annotation for dynamic action recognition. Results show that the proposed model using extremely limited annotations achieves superior performance than existing methods. This work hopes to serve as an inspiration for alleviating the intensive annotation of dynamic databases in the future. In addition, a large-scale 3D dataset for spontaneous and dynamic facial action analysis is introduced to alleviate the scarcity issue of subject samples. The material is based on the work supported in part by the NSF under grant CNS-1629898 and the Center of Imaging, Acoustics, and Perception Science (CIAPS) of the Research Foundation of Binghamton University.

References

[1] Yanan Chang et al. Knowledge-driven self-supervised representation learning for facial action unit recognition. In *Proceedings of the IEEE/CVF Conference on CVPR*, 2022. 7

[2] Jinwoo Choi et al. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *WACV*, 2020. 8

[3] Wen-Sheng Chu et al. Learning spatial and temporal cues for multi-label facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG)*, 2017. 2

[4] Ciprian Corneanu et al. Deep structure inference network for facial action unit recognition. In *Proceedings of ECCV*, 2018. 7

[5] Zijun Cui et al. Knowledge augmented deep neural networks for joint facial expression and action unit recognition. In *NIPS*, 2020. 2, 7

[6] Chen Gong et al. Teaching semi-supervised classifier via generalized distillation. In *IJCAI*, 2018. 2

[7] Qiushan Guo et al. Online knowledge distillation via collaborative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[8] Kaiming He et al. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[9] Geoffrey Hinton et al. Distilling the knowledge in a neural network, 2015. 2

[10] Geethu Miriam Jacob et al. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 7

[11] Longlong Jing et al. Videossl: Semi-supervised learning for video classification. In *WACV*, 2021. 8

[12] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *5th International Conference on Learning Representations, ICLR*, 2017. 6

[13] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 2013. 5, 7, 9

[14] Jie Lei et al. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 8

[15] Guanbin Li et al. Semantic relationships guided representation learning for facial action unit recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 6

[16] Guanbin Li et al. Semantic relationships guided representation learning for facial action unit recognition. In *AAAI*, 2019. 7

[17] Wei Li et al. Eac-net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 7

[18] Xiaotian Li et al. Your "attention" deserves attention: A self-diversified multi-channel attention for facial action analysis. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, 2021. 1, 4

[19] Xiaotian Li, Zheng Zhang, Xiang Zhang, Taoyue Wang, Zhihua Li, Huiyuan Yang, Umur Ciftci, Qiang Ji, Jeffrey Cohn, and Lijun Yin. Disagreement matters: Exploring internal diversification for redundant attention in generic facial action analysis. *IEEE Transactions on Affective Computing*, pages 1–12, 2023. 1, 4

[20] Zhihua Li, Xiang Deng, Xiaotian Li, and Lijun Yin. Integrating semantic and temporal relationships in facial action unit detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 2

[21] D. Lopez-Paz et al. Unifying distillation and privileged information. In *International Conference on Learning Representations (ICLR)*, 2016. 2

[22] Cheng Luo et al. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. In *Proceedings of IJCAI*, 2022. 2

[23] S. M. Mavadati et al. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 6

[24] Xuesong Niu et al. Local relationship learning with person-specific shape regularization for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[25] Xuesong Niu et al. Multi-label co-regularization for semi-supervised facial action unit recognition. In *NeurIPS*, 2019. 2, 7

[26] M. Pantic et al. Web-based database for facial expression analysis. In *2005 IEEE International Conference on Multimedia and Expo*, pages 5 pp.–, 2005. 6

[27] Guozhu Peng et al. Weakly supervised facial action unit recognition through adversarial training. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2188–2196, 2018. 7

[28] Guozhu Peng et al. Dual semi-supervised learning for facial action unit recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 2, 7

[29] Olga Russakovsky et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2014. 3

[30] Zhiwen Shao et al. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of ECCV*, 2018. 5

[31] Zhiwen Shao et al. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of ECCV*, 2018. 7

[32] Z. Shao et al. Facial action unit detection using attention and relation learning. *IEEE Transactions on Affective Computing*, page 1–1, 2019. 7

[33] Ankit Singh et al. Semi-supervised action recognition with temporal contrastive learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 7, 8

[34] Kihyuk Sohn et al. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NIPS*, 2020. 7

[35] Tengfei Song et al. Hybrid message passing with performance-driven structures for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[36] Tengfei Song et al. Hybrid message passing with performance-driven structures for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 7

[37] Tengfei Song et al. Uncertain graph neural networks for facial action unit detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5993–6001, 2021. 7

[38] Yale Song et al. Exploiting sparsity and co-occurrence structure for action unit recognition. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2015. 1, 7

[39] Yang Tang et al. Piap-df: Pixel-interested and anti person-specific facial action unit detection net with discrete feedback learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2, 7

[40] Vladimir Vapnik et al. Learning using privileged information: Similarity control and knowledge transfer. *J. Mach. Learn. Res.*, 2015. 2

[41] Ashish Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2, 3

[42] Xiang Wang et al. Self-supervised learning for semi-supervised temporal action proposal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5, 8

[43] Xiang Wang et al. Self-supervised learning for semi-supervised temporal action proposal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1905–1914, June 2021. 7

[44] Shan Wu et al. Deep facial action unit recognition from partially labeled data. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 7

[45] Qizhe Xie et al. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 2

[46] Huiyuan Yang et al. Exploiting semantic embedding and visual feature for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 7

[47] Xing Zhang et al. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 6

[48] Xiang Zhang and Lijun Yin. Multi-modal learning for au detection based on multi-head fused transformers. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8, 2021. 1

[49] Zheng Zhang et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6