S^2-VER: Semi-Supervised Visual Emotion Recognition

Guoli Jia[1 [0000−0002−9494−7013]] and Jufeng Yang[1 [0000−0003−0219−3443]]

Nankai University, Tianjin, China
exped1230@gmail.com, yangjufeng@nankai.edu.cn

Abstract. Visual emotion recognition (VER), which plays an important role in various applications, has attracted increasing attention of researchers. Due to the ambiguous characteristic of emotion, it is hard to annotate a reliable large-scale dataset in this field. An alternative solution is semi-supervised learning (SSL), which progressively selects high-confidence samples from unlabeled data to help optimize the model. However, it is challenging to directly employ existing SSL algorithms in VER task. On the one hand, compared with object recognition, in VER task, the accuracy of the produced pseudo labels for unlabeled data drops a large margin. On the other hand, the maximum probability in the prediction is difficult to reach the fixed threshold, which leads to few unlabeled samples can be leveraged. Both of them would induce the suboptimal performance of the learned model. To address these issues, we propose S^2-VER, the first SSL algorithm for VER, which consists of two components. The first component, reliable emotion label learning, aims to improve the accuracy of pseudo-labels. In detail, it generates smoothing labels by computing the similarity between the maintained emotion prototypes and the embedding of the sample. The second one is ambiguity-aware adaptive threshold strategy, which is dedicated to leveraging more unlabeled samples. Specifically, our strategy uses information entropy to measure the ambiguity of the smoothing labels, then adaptively adjusts the threshold, which is adopted to select high-confidence unlabeled samples. Extensive experiments conducted on six public datasets show that our proposed S^2-VER performs favorably against the state-of-the-art approaches. The code is available at *https://github.com/exped1230/S2-VER*.

1 Introduction

Visual emotion recognition (VER) aims at identifying human's emotions towards different visual stimuli [37]. With the popularization of multimedia, many people utilize images to record their feelings on social platforms, such as Instagram and Twitter. Therefore, visual emotion has drawn increasing attention from computer vision researchers [8, 24, 37] with its wide applications, *e.g.*, opinion mining [27, 52] and image captioning [1, 6]. Among them, recognizing the dominant emotion evoked by affective images is one of the most popular research directions [39, 45].

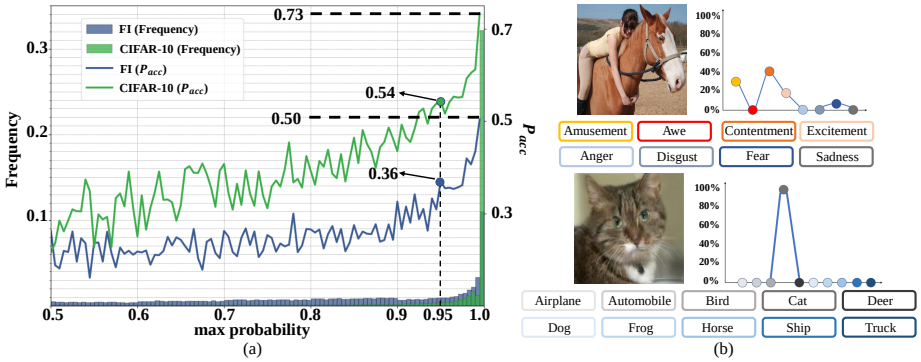


Fig. 1: Comparison between FI and CIFAR-10. (a) shows the frequency histogram and pseudo label accuracy P_{acc} when training FixMatch on FI and CIFAR-10, respectively. The maximum probability of the prediction is used to measure the confidence of the pseudo labels in SSL. FixMatch sets 0.95 as threshold to select high-confidence samples. (b) illustrates the label distribution of images from FI and CIFAR-10. The labels are represented by different colors.

In the past decades, many works make considerable improvements to VER [28, 39, 47]. However, most of these methods train networks in fully-supervised manner, which need a large amount of labeled data. It is extremely time-consuming to construct such datasets. Besides, due to the diversity of cultural backgrounds and personalities, different viewers may have different emotions induced by the same image [52]. Furthermore, a viewer may even have multiple emotions towards an image, *i.e.*, ambiguity [42]. Therefore, compared with object recognition, it is challenging to annotate a reliable large-scale dataset for FER. In this paper, we explore leveraging pseudo-labeling based semi-supervised learning (SSL) algorithms to address this issue. On the one hand, with the help of SSL algorithm, the cost of annotation can be significantly reduced. On the other hand, the algorithms progressively adopt high-confidence samples to train the model, which alleviates the impact of unreliable samples. We believe it is a promising direction to address the difficulty of emotion annotation.

SSL aims to address the need for labeled data by designing an algorithm to utilize unlabeled data [4]. As a representative method, FixMatch [30] selects high-confidence predictions from weakly augmented unlabeled instances, and then exploits them as the pseudo labels for the strongly augmented instances. To explore the performance of FixMatch in VER, we conduct a comparison experiment on FI [44] and CIFAR-10 [15]. Specifically, on both datasets we sample 100 labeled samples and 1,000 unlabeled samples from each class to train ResNet50 [9] with the same setting. The results are shown in Fig. 1 (a), and we have two observations. First, when the maximum probability is 0.95, compared with CIFAR-10, the accuracy of pseudo labels on FI drops a large margin. During the training process, due to the challenge of VER, the accumulated mistakes of pseudo labels result in confirmation bias, which is a common hazard in SSL [18,

32]. Second, only a few samples have a high maximum probability. As shown in Fig. 1 (b), different from the one-hot description for images in CIFAR-10, the probability of dominant emotion may be limited by other existing emotions. Therefore, the number of samples reaching the threshold is small, which limits the performance of the model [35].

To address these problems, we propose S^2-VER, the first semi-supervised VER algorithm, which consists of two components. First, the reliable emotion label learning module adopts label smoothing to improve the accuracy of pseudo labels. Label smoothing has been proven to implicitly calibrate the learned models so that the confidences are more aligned with the accuracies of their predictions [22]. Inspired by this, we generate smoothing labels for affective images. Specifically, we calculate the similarity between embeddings and the emotional prototypes. To capture the associations among emotions, the smoothing labels are multiplied with a maintained emotional relation matrix. Furthermore, since the quality of smoothing labels depends on the embeddings, we introduce an continuous contrastive loss to obtain emotionally discriminative representations. Second, we propose an ambiguity-aware adaptive threshold strategy, which aims to exploit more emotionally high-confidence unlabeled samples. For each sample, the strategy measures the ambiguity of the smoothing labels by information entropy and the polarity cue of emotions. Based on this strategy, the threshold is adaptively adjusted and more high-confidence unlabeled data can be leveraged.

Our contributions are summarized as follows: 1) We address the difficulty of annotating emotion datasets by SSL. To the best of our knowledge, this is the first visual emotion work that focuses on learning in semi-supervised manner. 2) We propose S^2-VER, which can improve the accuracy of pseudo labels, and leverage more emotionally high-confidence unlabeled data for VER. 3) We conduct extensive experiments on six datasets and the results demonstrate the effectiveness of S^2-VER.

2 Related Work

2.1 Visual Emotion Recognition

The research on visual emotion recognition has developed for more than two decades [49, 52]. In the early years, researchers exploit handcrafted features to recognize the dominant emotion conveyed by an affective image. Inspired by the theory of psychology and art, Machajdik *et al.* [20] extract features from four aspects, containing color, texture, composition, and content. It is a representative low-level handcrafted feature. Zhao *et al.* [48] explore the research of principles-of-art, and propose a mid-level representation of visual emotion. To understand the visual concepts that are strongly related to emotion, Borth *et al.* [5] automatically collect adjective-none pairs (ANP) as the high-level representation.

Recently, many methods [34, 45] exploit convolutional neural network (CNN) for VER. Considering the localized information, Yang *et al.* [39] design a weakly supervised coupled network to integrate recognition and detection tasks. To ex-

tract various levels of related visual features, Rao *et al.* [28] construct a region-based CNN network with multi-level framework. [13, 14] analyze the relation between person and context scene to extract rich information about emotional states. Furthermore, [36] proposes a novel Scene-Object network, which leverages reasoning network to mine the relations among objects and the correlation between the objects and scene. Different from previous methods, [38] proposes a stimuli-aware visual emotion model consisting of Global-Net, Semantic-Net, and Expression-Net, which extracts three aspects of emotional stimulus simultaneously. Although these methods have made great progress in VER, a fundamental weakness of these deep models is that they typically require a lot of accurately annotated data to work well [35]. However, it is still one of the main challenges for VER. Therefore, we explore SSL algorithm to address this issue.

2.2 Semi-Supervised Learning

SSL trains models incorporating labeled and unlabeled samples. Many classic methods have been proposed, such as transductive models [10, 11], generative models [12, 29], and graph-based models [54, 55]. In addition, [17] proposes to generate pseudo labels by picking up the category which has the maximum probability in the predicted distribution. ICT [33] regularizes the model based on the liner interpolation assumption.

Consistency regularization is an effective method, which is based on the smoothness theory that slight perturbations on the data points will not change the output of the network [23]. UDA [35] and ReMixMatch [3] generate targets from weakly augmented images, and then enforce consistency against strongly augmented images. In recent years, pseudo-labeling combined with strong augmentations becomes a powerful method for SSL [43]. For the pseudo-labeling methods, the network predicts pseudo labels of unlabeled samples, then trains itself with these labels. FixMatch [30] leverages a confidence-based strategy to obtain reliable pseudo labels. Considering the different learning status and difficulties of each class, FlexMatch [46] proposes a curriculum learning approach to address this issue. For VER, these SSL algorithms suffer from the low accuracy of generated pseudo labels and the lack of high-confidence unlabeled data. In this paper, we leverage the label smoothing method to improve the accuracy of pseudo labels, and adaptively adjust this threshold based on the ambiguity.

3 Methodology

3.1 Overview

Our proposed S^2-VER is illustrated in Fig. 2. Given a batch of labeled samples $\mathcal{X} = \{(x_b, q_b)\}_{b=1}^{B}$, where B is the batch size, $x_b \in \mathbb{R}^{H\times W\times 3}$ denotes a sample in the batch, $q_b \in \mathbb{R}^{1\times 1\times C} \in$ is one-hot label, which contains C emotions. We optimize supervised loss on the labeled samples $\mathcal{X}$ as:

$$\mathcal{L}_x = \frac{1}{B}\sum_{b=1}^{B} H\left(q_b, p(y|x_b^w)\right), \tag{1}$$

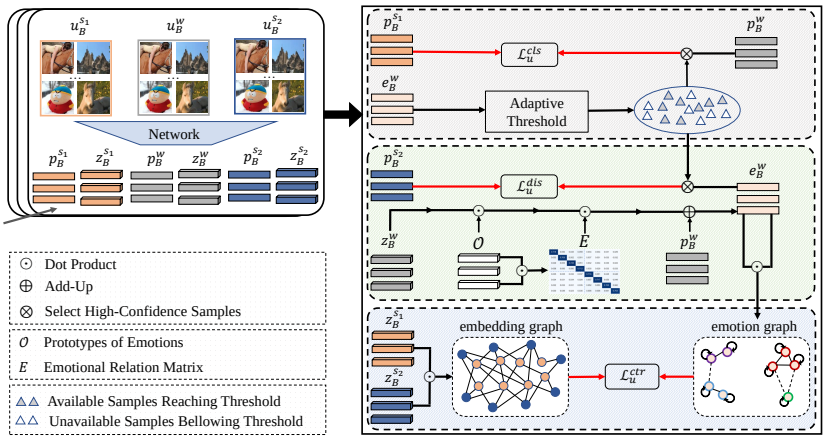


Fig. 2: Illustration of our proposed method. u_B^w, u_B^{s1}, u_B^{s2} represent the weakly augmented instances and two strongly augmented instances of a batch of unlabeled samples. The three modules on the right are ambiguity-aware adaptive threshold, label smoothing based calibration, and continuous contrastive learning, respectively. For each instance, the network outputs its prediction p and low-dimensional embedding z, and e_B^w denotes the generated smoothing labels.

where p denotes the conditional probability of the sample, H is the cross-entropy between q_b and p, x_b^w means the weakly augmented instance of x_b. Let $\mathcal{U} = \{u_b\}_{b=1}^{\mu B}$, where $\mathcal{U}$ denotes a batch of unlabeled samples, μB means that the batch size of $\mathcal{U}$ is μ times of X. For each $u_b \in \mathbb{R}^{H\times W\times 3}$, we perform a weak augmentation Aug_w and two random strong augmentations Aug_s on it, obtaining the transformed instances u_b^w, $u_b^{s_1}$, $u_b^{s_2}$. The network outputs the prediction p_b and normalized embedding $z_b \in \mathbb{R}^{1\times 1\times D}$ of each instance, where D denotes the dimension of the embedding. Then, the unlabeled samples are optimized with losses $\mathcal{L}_u^{cls}$, $\mathcal{L}_u^{dis}$, and $\mathcal{L}_u^{ctr}$. The single label classification loss $\mathcal{L}_u^{cls}$ is defined the same as previous works [18, 30, 46]:

$$\mathcal{L}_u^{cls} = \frac{1}{\mu B}\sum_{b=1}^{\mu B} \mathbb{1}(q_b^w \geq \tau) H(q_b^w, p(y|u_b^{s_1})), \tag{2}$$

where $q_b^w = argmax(p_b^w)$ denotes the pseudo label from weakly augmented instance. The τ means the threshold to select high-confidence prediction. The $\mathcal{L}_u^{dis}$ and $\mathcal{L}_u^{ctr}$ are elaborated in Sec. 3.2. Let λ^{cls}, λ^{dis}, λ^{ctr} denote the weight of the three losses for unlabeled data respectively, the overall loss function can be defined as:

$$\mathcal{L} = \mathcal{L}_x + \lambda^{cls}\mathcal{L}_u^{cls} + \lambda^{dis}\mathcal{L}_u^{dis} + \lambda^{ctr}\mathcal{L}_u^{ctr}. \tag{3}$$

3.2 Reliable Emotion Label Learning

This component consists of two modules. Specifically, we design the label smoothing based calibration module to improve the accuracy of pseudo labels. Due to

the quality of smoothing labels determined by the extracted embeddings, we further adopt continuous contrastive learning module to optimize the network. In the following parts, we first describe the process of label smoothing based calibration, and then introduce the continuous contrastive learning strategy.

Label smoothing based calibration aims to improve the accuracy of pseudo labels. Label smoothing is a regularization method that maintains a reasonable ratio among the logits of the incorrect classes [26]. This regularization method can implicitly calibrate the over-confidence of the learned models so that their predictions are more aligned with the accuracies [22]. Here, we utilize learnable embeddings to dynamically generate emotional smoothing labels.

Each emotion is represented by the maintained prototype $\mathcal{O}^i \in \mathbb{R}^{1\times1\times D}$, $\mathcal{O} = \{\mathcal{O}^1, \mathcal{O}^2, ...\mathcal{O}^C\}$. The prototypes $\mathcal{O}$ are updated by the momentum moving average of embeddings from $\mathcal{X}$, with $\lambda = 0.9$. To be specific, the prototypes are initialized as zero-vectors, and during the training stage, the $\mathcal{O}^i$ of the i-th class is calculated as:

$$\mathcal{O}^i = \frac{\sum_{b=1}^{B} z_b^w \cdot \mathbb{1}(y_b = i)}{\sum_{b=1}^{B} \mathbb{1}(y_b = i)}, \tag{4}$$

where y_b is the label of the b-th labeled sample. Then, we generate initial smoothing label $\hat{d}_b^w$ by calculating the similarity between the prototypes $\mathcal{O}$ and the embedding z_b^w extracted from u_b^w. The probability of i-th class in distribution $\hat{d}_b^w$ is calculated as:

$$\hat{d}_{b^i}^w = \frac{exp(z_b^w \cdot \mathcal{O}^i/t)}{\sum_{i=1}^{C} exp(z_b^w \cdot \mathcal{O}^i/t)}, \tag{5}$$

where the t is a scalar which denotes the temperature. Here we use softmax to ensure $\sum_{i=1}^{C} \hat{d}_{b^i}^w = 1$. In addition, since the distances between emotions are different [51], we maintain an emotional relation matrix $E \in \mathbb{R}^{C\times C}$. For instance, the distance between amusement and contentment is relatively smaller than the distance between amusement and awe. In detail, the emotional relation matrix E is initialized with $\frac{1}{C}$, and updated by the distance of prototypes $\mathcal{O}$. Here we exploit L_2 distance as the metric, and the value of the i-th row and j-th column is formally defined as:

$$E^{ij} = \frac{exp(-\left\|(\mathcal{O}^i - \mathcal{O}^j)\right\|_2^2)}{\sum_{k=1}^{C} exp(-\|(\mathcal{O}^i - \mathcal{O}^k)\|_2^2)}, \tag{6}$$

The emotional relation matrix E is also updated on the moving average with the same λ. Next, the emotional distribution is adjusted by the relation matrix as $d_b^w = \hat{d}_b^w \cdot E$. In order to control the degree of smoothing, we leverage θ as the weight, and combine the model's prediction and the generated distribution

as the smoothing label e_b^w . It can be formally defined as $e_b^w = (1-\theta)p_b^w + \theta d_b^w$. Finally, the smoothing label is leveraged to calculate the Kullback-Leibler (KL) loss between two distributions e_b^w and p:

$$\mathcal{L}_u^{dis} = -\frac{1}{\mu B}\sum_{b=1}^{\mu B} \mathbb{1}(q_b^w \geq \tau) KL(e_b^w, p(y|u_b^{s_2})). \tag{7}$$

Continuous contrastive learning aims to learn emotionally discriminative embeddings, which could improve the quality of emotional smoothing labels. Recently, many SSL algorithms exploit contrastive learning to learn better representations [2,53]. Among them, CoMatch [18] designs a graph-based contrastive algorithm, which has been proved effective for SSL. However, unlike other classification tasks, emotions are closely related to each other [37], and the distances between emotions are different. Therefore, it is suboptimal to simply identify whether the images are from the same class. Inspired by this, we introduce continuous contrastive to regularize the emotional embeddings.

Given a batch of unlabeled data U, we utilize the smoothing labels e^w to construct emotion graph $W^e \in \mathbb{R}^{B\times B}$, and utilize the embeddings z^w to construct embedding graph $W^z \in \mathbb{R}^{B\times B}$. Specifically, we use the samples as the vertex, and adopt the cosine similarity to represent the weight of the edge. In this way, the W^e and W^z can be easily obtained in each batch. Then, we adjust the emotion graph based on two priors: (i) Samples should have the same emotion with themselves, thus we set the value of the diagonal element to 1. (ii) We observe that there are many pairs with small similarities in a batch, lots of such weak associations will impact the performance, so we set the values which below T to 0. This process can be defined as:

$$W_{ij}^e = \begin{cases} 1, & i = j, \\ e_i^w \cdot e_j^w, & i \neq j, e_i^w \cdot e_j^w > T, \\ 0, & otherwise. \end{cases} \tag{8}$$

We empirically set the T to 0.3. Note that emotion datasets usually contain few classes, such as 6 or 8. Therefore, the T with 0.3 encourages models to learn rich emotion relations. Next, both the emotion graph and embedding graph are normalized by softmax. Finally, the contrastive loss between W^e and W^z is calculated by KL loss:

$$\mathcal{L}_u^{ctr} = KL(W^e, W^z). \tag{9}$$

3.3 Ambiguity-Aware Adaptive Threshold

Most SSL algorithms leverage the pseudo-labels exceeding the fixed threshold [18, 30]. However, due to the ambiguity, an affective image may contain not only one emotion [37,41,42]. Although some unlabeled samples already have correct pseudo labels, the probability of dominant emotion of the images would be limited by other existing emotions, making it difficult to reach the threshold.

Throughout the training process, the ambiguity of emotion will result in the lack of available unlabeled data. To make better use of these data, we propose a strategy to adaptively adjust the threshold. We adopt β as the lower bound of the threshold, ω controls the extent of the adjustment, the strategy can be formally defined as:

$$\tau = \beta + (1 - \beta)\omega. \tag{10}$$

Specifically, we use information entropy $\mathcal{A} = \sum_{i=1}^{C} -e_{b_i}^{w} \cdot lne_{b_i}^{w}$ to measure the ambiguity of the smoothing label. A large $\mathcal{A}$ means a low threshold is needed. However, it is difficult to distinguish whether the prediction with large $\mathcal{A}$ is caused by the ambiguity of the emotion or the poor performance of the model. As an extreme example, the $\mathcal{A}$ of distribution [0.25, 0.25, 0.25, 0.25] is large, but it is more like a random prediction caused by insufficient training. Thanks to the polarity that emotions can be divided into positive and negative, we can select the ambiguous predictions in accord with the rule of emotion. Based on [41], the emotional ambiguity often exists between emotions from the same polarity. To be specific, we add the probabilities having the same polarity with the dominant emotion, which can be seen as the reliability of $\mathcal{A}$. Therefore, we can adaptively calculate the extend ω of the i-th sample as:

$$\omega = \frac{1}{(\mathcal{A} + a) \cdot \sum_{j=1}^{C} \mathbb{1}(P(j) == P(argmax_j(e_{b_i}^{w}))e_{b_i^j}^{w}}. \tag{11}$$

The constant a aims to leverage more unlabeled data, here we set a as 1 empirically. $P(j)$ means the polarity of the j-th emotion. In practice, such an algorithm is simple and effective.

4 Experiments

4.1 Datasets

We evaluate our proposed S^2-VER on seven public emotion datasets, including FI [44], SE30K8 [34], FlickrLDL, TwitterLDL [42], Emotion-6 [24], UnBiasedEmo [24], and WEBEmo [24]. The images of FI are collected from Flickr and Instagram by querying Mikel's eight emotions as search keywords. A total of 225 AMT workers assess the emotions of images resulting in 23,308 images receiving at least three agreements. SE30K8 contains 33K images, which are annotated in eight emotions (anger, happiness, surprise, disgust, sadness, fear, neutral, surprise-positive, and surprise-negative). Following [44], we leave 22,866 images that receive more than half of the agreements. FlickrLDL and TwitterLDL consist of 11,150 and 10,045 images respectively, which are annotated by Mikel's emotion too. Due to the lack of manually annotated large-scale datasets, we merge these two datasets. Same as SE30K, we leave 15,816 images with high consistency. In the rest of the paper, we call it LDL dataset.

Table 1: Accuracy (%) of 5-folds on FI, SE30K8, and LDL datasets. We evaluate S^2-VER against four representative VER methods and ten classic SSL methods. Note that Pseu-Lab, Mean-Tea, and Remix denote Pseudo-Label, Mean-Teacher, and ReMixMatch, respectively. To ensure a fair comparison, we adopt ResNet50 as backbone for all the 15 methods.

Method	FI			SE30K8			LDL		
	80	800	1600	80	400	800	80	800	1600
Yang *et al.* [41]	19.9±0.36	25.4±0.37	30.1±0.33	19.8±0.41	22.7±0.32	26.0±0.55	21.4±0.26	26.5±0.29	32.3±0.41
RCA [40]	18.4±0.33	25.9±0.39	31.4±0.17	18.6±0.29	21.9±0.33	26.5±0.33	23.8±0.48	29.2±0.19	33.2±0.21
WSCNet [39]	20.2±0.37	27.5±0.41	31.2±0.39	18.4±0.25	23.2±0.28	27.4±0.36	22.3±0.46	29.2±0.31	35.2±0.45
PDANet [50]	21.4±0.26	26.6±0.22	33.2±0.31	20.6±0.18	23.4±0.25	27.7±0.45	23.5±0.29	30.5±0.30	33.5±0.33
π-Model [16]	22.9±0.54	28.3±0.36	31.7±0.21	22.1±1.71	23.7±0.69	26.9±0.31	24.3±0.61	31.8±0.59	34.4±0.27
Pseu-Lab [17]	22.9±0.48	31.3±0.43	33.5±0.31	23.4±1.10	25.9±0.50	27.6±0.16	24.2±0.49	32.3±0.44	35.8±0.13
VAT [21]	23.6±0.78	31.5±0.77	35.1±0.37	24.4±0.69	27.2±0.39	28.9±0.25	26.3±0.58	34.5±0.49	38.9±0.36
Mean-Tea [32]	23.8±0.51	29.3±0.48	33.9±0.33	24.3±0.67	26.7±0.53	28.2±0.22	26.6±0.54	33.8±0.42	38.6±0.20
MixMatch [4]	26.3±1.53	35.1±0.74	38.0±0.32	26.6±0.87	28.3±0.62	29.6±0.40	28.1±0.78	34.2±0.52	38.9±0.23
ReMix [3]	29.7±0.68	35.4±0.53	38.3±0.42	26.4±1.10	29.9±0.98	31.9±0.63	29.1±0.67	35.3±0.54	39.2±0.35
UDA [35]	28.5±0.87	37.7±0.56	40.3±0.38	27.3±0.89	29.6±0.64	32.2±0.37	30.7±0.76	40.9±0.58	43.4±0.47
FixMatch [30]	28.2±0.78	37.4±0.51	42.2±0.29	29.7±0.70	32.2±0.57	32.7±0.46	32.4±0.84	39.4±0.45	43.2±0.24
FlexMatch [46]	29.7±0.90	38.2±0.49	42.9±0.17	28.5±1.03	33.2±0.60	33.9±0.26	33.2±0.93	41.3±0.71	46.7±0.42
CoMatch [18]	36.7±0.87	43.5±0.39	47.9±0.26	29.9±0.65	32.5±0.47	35.3±0.26	**38.1**±0.53	42.1±0.31	45.3±0.27
S^2-VER	**39.1**±0.66	**46.9**±0.46	**51.8**±0.21	**30.1**±0.73	**33.3**±0.62	**36.2**±0.49	37.9±0.80	**43.6**±0.47	**47.4**±0.43

We also evaluate S^2-VER on two small-scale datasets and one large-scale dataset. Emotion-6 consists of 8,350 images, which initially collected 150K images from Google and labeled by five people. UnBiasedEmo contains 3,045 affective images from a collection of about 60,000 images. Both Emotion-6 and UnBiasedEmo are annotated by Ekman's emotion taxonomy. The WEBEmo is a large-scale web dataset searched by Parrott's hierarchical emotion model [25]. After removing duplicate images, about 268,000 stock samples are reserved.

4.2 Evaluation Settings

We first compare S^2-VER against representative VER and SSL methods on FI, SE30K8, and LDL. We report the results of four VER methods: Yang [41], RCA [40], WSCNet [39], and PDANet [50]. In particular, we make a simple transform to PDANet according to [52], making it suitable for classification. For SSL, we compare with ten representative methods, π-Model [16], Pseudo-Label [17], VAT [21], Mean-Teacher [32], MixMatch [4], ReMixMatch [3], UDA [35], FixMatch [30], CoMatch [18], and FlexMatch [46]. Furthermore, we conduct experiments on UnBiasedEmo, Emotion-6, and WEBEmo. These experiments aim to validate the effectiveness of S^2-VER on small-scale emotion datasets with the help of large-scale unlabeled web dataset. Specifically, we compare S^2-VER with three most powerful methods FixMatch, FlexMatch, and CoMatch.

4.3 Implementation Details

To ensure fairness, we adopt ResNet-50 [9] as the backbone for all experiments. The images are resized to 256×256 followed by a center 224×224 cropping. The batch size of unlabeled data is 64, which is 4 times that of labeled data. Our network is optimized by stochastic gradient descent. The momentum and weight decay are set to 0.9 and 0.0005 respectively. The total number of epochs is 512, each epoch contains 1024 iterations. Following [30,46], the learning rate is initialized as 0.03 with a cosine learning rate decay schedule [19]. The weak augmentation is adopted as standard crop-and-flip, and the strong augmentation is implemented by the RandAugment [7] the same as [30,46]. For the experiments whose results are shown in Table 1, following [18, 46], we randomly sample labeled data in a class-balanced way. Due to the lack of surprise-positive samples in SE30K8, we conduct experiments with 80, 400, and 800 labeled samples respectively. In addition, the neural is considered as a polarity different from positive and negative for ambiguity-aware adaptive threshold strategy. Same as FI, SE30K8 and LDL are randomly split into 80% training, 5% validation, and 15% testing sets. Emotion-6 and UnBiasedEmo are split into 90% training and 10% testing sets. Following previous SSL algorithms, we present the accuracy of the Exponential Moving Average (EMA) model [46].

4.4 Comparison with the State-of-the-art Methods

We conduct extensive experiments to compare S^2-VER with the state-of-the-art methods on VER datasets. The methods include VER models and SSL algorithms. The VER models are trained using labeled samples. This comparison aims to demonstrate the effectiveness of SSL algorithms with limited labeled data. We also adopt SSL algorithms to improve the performance on small datasets. To be specific, we utilize two annotated small datasets as labeled data, large-scale WEBEmo as unlabeled data.

The comparison results are shown in Table 1. SSL algorithms are divided according to whether they use strong augmented anchors. Overall, the SSL algorithms outperform the emotion recognition models. This suggests that leveraging SSL algorithms and adopting a large amount of unlabeled data is beneficial for emotion recognition. In general, the SSL algorithms enforce the consistency between weakly augmented anchors and strongly augmented instances achieving better performance. In addition, we compare S^2-VER with the competitive approaches. S^2-VER improve about 3% on FI with different settings. On SE30K8 and LDL, S^2-VER also performs favorably against the representative methods.

For emotion analysis, many methods can be used to automatically acquire large-scale web data which has not been annotated by humans [5,24,34]. Therefore, we have an intrinsic assumption that SSL algorithms can be effective on small-scale datasets with help of the large-scale web data. Inspired by this, we conduct experiments on Emotion-6, UnBiasedEmo, and WEBEmo. Specifically, we leverage WEBEmo as unlabeled data, training FixMatch, FlexMatch, CoMatch, and S^2-VER on the labeled datasets. In order to verify our assumption, we first train ResNet-50 on two small-scale datasets, the accuracy is 43.6%

Table 2: Accuracy (%) on Emotion-6 and UnBiasedEmo. For both datasets, we leverage WEBEmo as unlabeled data. We compare S^2-VER with current-best SSL methods, *i.e.*,FixMatch, FlexMatch, and CoMatch.

Method	Emotion-6			UnBiasedEmo		
	20%	50%	100%	20%	50%	100%
FixMatch [30]	46.6	48.3	49.0	65.6	69.2	71.5
FlexMatch [46]	48.1	50.1	51.2	67.2	71.1	73.4
CoMatch [18]	50.2	51.2	52.4	68.5	70.8	73.8
S^2-VER	**51.7**	**53.5**	**54.0**	**70.8**	**76.7**	**78.7**

and 61.3% respectively. We conduct experiments with 20%, 50%, 100% data of WEBEmo, the results of SSL algorithms are reported in Table 2. As we can see, these models perform much better than the model trained only with small-scale datasets. Even though there exists bias between datasets, SSL algorithms can still improve the performance with the help of unlabeled data. Moreover, our method achieves competitive performance compared with the representative SSL algorithms.

4.5 Ablation Study

In order to prob the effectiveness of different components in S^2-VER, we display ablation results here. Note that all the experiments are conducted on FI with 1,600 labeled samples. First, we show the effect of each component in Table 3, and draw the following conclusions: 1) Since the quality of smoothing labels particularly depends on the embeddings, combining both $\mathcal{L}_u^{dis}$ and $\mathcal{L}_u^{ctr}$ surpasses only using $\mathcal{L}^{dis}$ a large margin. 2) Although leveraging $\mathcal{L}_u^{dis}$ and $\mathcal{L}_u^{ctr}$ achieves high accuracy, the performance can be further improved by optimizing $\mathcal{L}_u^{cls}$ simultaneously. 3) The model achieves the best test accuracy by utilizing all the components, which shows the complementarity of our proposed S^2-VER.

Furthermore, we conduct detailed experiments to illustrate the contribution of the proposed reliable emotion label learning and adaptive threshold strategy, respectively. We compare the M_{acc} for FixMatch, FlexMatch, CoMatch, the base model with standard label smoothing [31], and proposed label smoothing for emotion. The base model with $\mathcal{L}_u^{cls}$ downgrads to FixMatch. Since the warmup training strategy used in FlexMatch brings a large number of unreliable samples [46], we also report the results of FlexMatch without warmup. As can be seen in Table 4, label smoothing can significantly improve the M_{acc}, and our proposed label smoothing performs better for VER.

To evaluate the effect of the proposed adaptive threshold, we show the proportion of high-confidence unlabeled samples reaching the threshold. As shown in Fig. 3, our method using the adaptive threshold can utilize more unlabeled

Table 3: Ablation study to prob the $\mathcal{L}_u^{cls}$, $\mathcal{L}_u^{dis}$, $\mathcal{L}_u^{ctr}$, and AT used in our S^2-VER. The $\mathcal{L}_u^{smo}$ here denotes the combination of $\mathcal{L}_u^{dis}$ and $\mathcal{L}_u^{ctr}$. Note that ✓denotes only $\mathcal{L}_u^{dis}$ is used, ✓* denotes both $\mathcal{L}_u^{dis}$ and $\mathcal{L}_u^{ctr}$ are leveraged. AT denotes the adaptive threshold strategy.

$\mathcal{L}_u^{cls}$	$\mathcal{L}_u^{smo}$	AT	Acc
✓			42.3
	✓		41.5
	✓*		49.4
✓	✓*		50.3
✓		✓	45.1
	✓*	✓	50.7
✓	✓*	✓	**51.8**

Table 4: Evaluating the effect of the smoothing label. M_{acc} means the accuracy of pseudo labels reaching threshold. W denotes training FlexMatch with warmup strategy. S and L represent the standard label smoothing strategy [31] and our proposed label smoothing for VER, respectively. We report the M_{acc} every 150 epoch and their average.

M_{acc}	150	300	450	Avg
FixMatch	44.9	59.7	61.8	55.5
FlexMatch	36.6	44.2	47.4	42.7
FlexMatch(W)	31.4	37.4	40.6	36.5
CoMatch	48.1	67.4	66.9	60.8
Base + S	**59.8**	70.7	67.8	66.1
Base + L	58.9	**72.3**	**69.3**	**66.8**

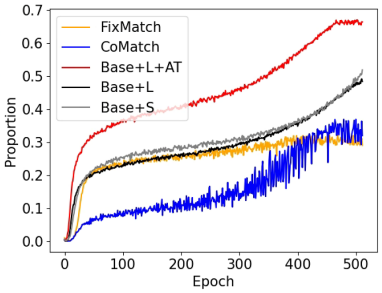


Fig. 3: Evaluating the effect of the proposed AT in detail. We show the proportion of high-confidence unlabeled samples during training.

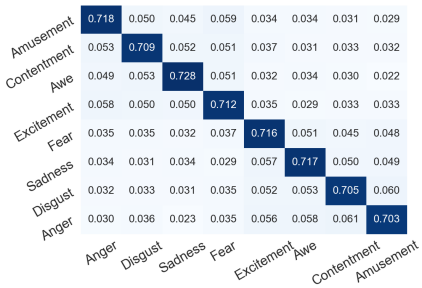
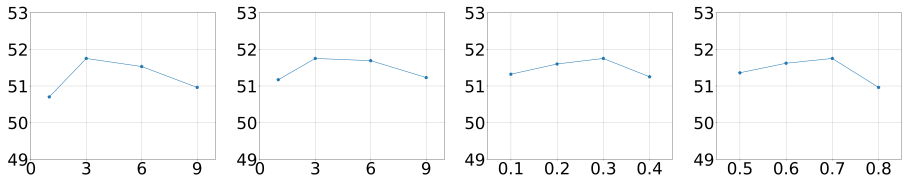


Fig. 4: The final emotional relation matrix in our model. We can find the relation between emotions from the same polarity is closer.

samples compared with other methods. Besides, we also provide the learned final emotional relation matrix in Fig. 4. The matrix is not completely symmetric. This is because we perform softmax in rows when updating the matrix at each iteration. As we can see, the values on the diagonal are large, and the emotions from the same polarity have relatively closer relations.

4.6 Hyperparameter Analysis

In this section, we present experimental results to demonstrate the effect of hyperparameters. All the experiments are conducted on the FI with 1600 labeled



(a) λ^{dis} (b) λ^{ctr} (c) θ (d) β

Fig. 5: Variation of accuracy with different hyperparameters. The accuracy of best settings achieves 51.8%. (a) shows the effect of λ^{dis}. (b) shows the effect of λ^{ctr}. (c) shows the effect of θ. (d) shows the effect of β.

samples. First, we explore the effect of λ^{dis} and λ^{ctr}, which denote the weights of $\mathcal{L}_u^{dis}$ and $\mathcal{L}_u^{ctr}$. With the increasing of λ^{dis}, S^2-VER will first perform better. The small λ^{dis} limits the importance of the smoothing label based calibration, so the performance drops significantly. The algorithm achieves the best performance when $\lambda^{dis} = 3$. As for $\mathcal{L}_u^{ctr}$, we find S^2-VER achieves highest accuracy when $\lambda^{ctr} = 3$ as well. Since the $\mathcal{L}_u^{ctr}$ is adopted to obtain more discriminative embeddings for $\mathcal{L}_u^{dis}$, it may be better to keep the λ^{ctr} consistent with the λ^{dis}.

We also present the results of θ, which denotes the weight of generated emotional smoothing label e_b^w. The larger θ would make the label more smoothing. As shown in Fig. 5(c), with the increasing of θ, the accuracy becomes higher, which demonstrates the effectiveness of smoothing label. However, the proportion of available unlabeled samples is decreasing. Our method achieves the best performance when $\theta = 0.3$. At this time, the trade off between the quality of pseudo labels and the amount of leveraged unlabeled data reaches the best balance. The results with different β are shown in Fig. 5(d). β controls the range of τ, which influences the trade off between pseudo labels and utilized unlabeled data as well. In particularly, we find that only about 55% of the samples can reach the threshold when β=0.8, which leads to a relatively large drop in performance.

4.7 Visualization

We further present some visualization of predictions on FI. As shown in Fig. 6 (a), we show some examples that both our method and CoMatch perform well. We can find that emotions with the same polarity may have a relatively high probability, such as amusement and contentment. In (b), we show some examples that our method outperforms CoMatch. Take the middle image as an example, focusing on the emotion of the person, the model predicts it as anger. However, we can easily find the image shows an exciting competition. Training with smoothing labels can effectively alleviate over-confidence problems like these examples. In (c), we present some examples showing thresholds, which are adopted to select high-confidence pseudo labels. With the ambiguity-aware adaptive threshold strategy, our method can exploit more reliable unlabeled samples.

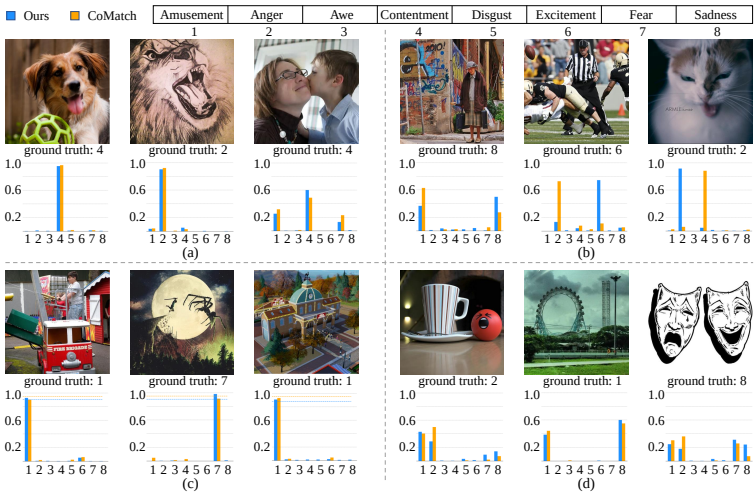


Fig. 6: Visualization of S^2-VER and CoMatch. (a) shows examples that both S^2-VER and CoMatch perform well. (b) shows examples that S^2-VER has correct predictions, while CoMatch has incorrect predictions. (c) shows examples that the predictions of S^2-VER reach the adaptive threshold, while CoMatch is below the fixed threshold. The blue and yellow line represent the threshold for S^2-VER and CoMatch, respectively. (d) shows some failure cases of our method.

In addition, we also present some failure cases in (d). Looking at the left image, focusing on the delicate cup may bring positive emotions, but the red toy additionally shows negative emotions. Due to the complexity of emotions, cases like (d) are inevitable. Therefore, we think that it is necessary to combine more psychological knowledge to alleviate this problem in the future.

5 Conclusion

In this paper, we propose S^2-VER, which is the first work exploring visual emotion in semi-supervised setting. We design a label smoothing method in the light of the characteristic of the emotion, improving the accuracy of high-confidence pseudo labels. Then, we propose an adaptive threshold strategy. With the help of polarity, this strategy is able to leverage more unlabeled samples effectively. Extensive experiments and comparisons indicate that S^2-VER has advantages compared with the state-of-the-art methods for VER.

6 Acknowledgments

This work was supported by the National Key Research and Development Program of China Grant (NO. 2018AAA0100403), NSFC (NO.61876094, U1933114), Natural Science Foundation of Tianjin, China (NO.20JCJQJC00020).

References

1. Achlioptas, P., Ovsjanikov, M., Haydarov, K., Elhoseiny, M., Guibas, L.J.: Artemis: Affective language for visual art. In: CVPR (2021)
2. Alonso, I., Sabater, A., Ferstl, D., Montesano, L., Murillo, A.C.: Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In: ICCV (2021)
3. Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Sohn, K., Zhang, H., Raffel, C.: Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. In: ICLR (2020)
4. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.: Mixmatch: A holistic approach to semi-supervised learning. In: NeurIPS (2019)
5. Borth, D., Ji, R., Chen, T., Breuel, T., Chang, S.F.: Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: ACM MM (2013)
6. Chen, T., Zhang, Z., You, Q., Fang, C., Wang, Z., Jin, H., Luo, J.: "factual" or "emotional": Stylized image captioning with adaptive learning and attention. In: ECCV (2018)
7. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: CVPRW (2020)
8. Fan, S., Shen, Z., Jiang, M., Koenig, B.L., Xu, J., Kankanhalli, M.S., Zhao, Q.: Emotional attention: A study of image sentiment and visual attention. In: CVPR (2018)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
10. Joachims, T.: Transductive learning via spectral graph partitioning. In: ICML (2003)
11. Joachims, T., et al.: Transductive inference for text classification using support vector machines. In: ICML (1999)
12. Kingma, D.P., Mohamed, S., Rezende, D.J., Welling, M.: Semi-supervised learning with deep generative models. In: NIPS (2014)
13. Kosti, R., Alvarez, J.M., Recasens, A., Lapedriza, A.: Emotion recognition in context. In: CVPR (2017)
14. Kosti, R., Alvarez, J.M., Recasens, A., Lapedriza, A.: Context based emotion recognition using emotic dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(11), 2755–2766 (2019)
15. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Mater's thesis, University of Toronto (2009)
16. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: ICLR (2017)
17. Lee, D.H.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: ICMLW (2013)
18. Li, J., Xiong, C., Hoi, S.C.: Comatch: Semi-supervised learning with contrastive graph regularization. In: ICCV (2021)
19. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. In: ICLR (2016)
20. Machajdik, J., Hanbury, A.: Affective image classification using features inspired by psychology and art theory. In: ACM MM (2010)
21. Miyato, T., Maeda, S.i., Koyama, M., Ishii, S.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(8), 1979–1993 (2018)

22. Müller, R., Kornblith, S., Hinton, G.E.: When does label smoothing help? In: NeurIPS (2019)
23. Oliver, A., Odena, A., Raffel, C., Cubuk, E.D., Goodfellow, I.J.: Realistic evaluation of deep semi-supervised learning algorithms. In: NeurIPS (2018)
24. Panda, R., Zhang, J., Li, H., Lee, J.Y., Lu, X., Roy-Chowdhury, A.K.: Contemplating visual emotions: Understanding and overcoming dataset bias. In: ECCV (2018)
25. Parrott, W.G.: Emotions in social psychology: Essential readings. Psychology Press (2001)
26. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., Hinton, G.: Regularizing neural networks by penalizing confident output distributions. In: ICLRW (2017)
27. Qian, S., Zhang, T., Xu, C.: Multi-modal multi-view topic-opinion mining for social event analysis. In: ACM MM (2016)
28. Rao, T., Li, X., Zhang, H., Xu, M.: Multi-level region-based convolutional neural network for image emotion classification. Neurocomputing **333**, 429–439 (2019)
29. Rasmus, A., Berglund, M., Honkala, M., Valpola, H., Raiko, T.: Semi-supervised learning with ladder networks. NIPS (2015)
30. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. NeurIPS (2020)
31. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR (2016)
32. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NIPS (2017)
33. Verma, V., Lamb, A., Kannala, J., Bengio, Y., Lopez-Paz, D.: Interpolation consistency training for semi-supervised learning. In: IJCAI (2019)
34. Wei, Z., Zhang, J., Lin, Z., Lee, J.Y., Balasubramanian, N., Hoai, M., Samaras, D.: Learning visual emotion representations from web data. In: CVPR (2020)
35. Xie, Q., Dai, Z., Hovy, E., Luong, M.T., Le, Q.V.: Unsupervised data augmentation for consistency training. In: NeurIPS (2019)
36. Yang, J., Gao, X., Li, L., Wang, X., Ding, J.: Solver: Scene-object interrelated visual emotion reasoning network. IEEE Transactions on Image Processing **30**, 8686–8701 (2021)
37. Yang, J., Li, J., Li, L., Wang, X., Gao, X.: A circular-structured representation for visual emotion distribution learning. In: CVPR (2021)
38. Yang, J., Li, J., Wang, X., Ding, Y., Gao, X.: Stimuli-aware visual emotion analysis. IEEE Transactions on Image Processing **30**, 7432–7445 (2021)
39. Yang, J., She, D., Lai, Y.K., Rosin, P.L., Yang, M.H.: Weakly supervised coupled networks for visual sentiment analysis. In: CVPR. pp. 7584–7592 (2018)
40. Yang, J., She, D., Lai, Y.K., Yang, M.H.: Retrieving and classifying affective images via deep metric learning. In: AAAI (2018)
41. Yang, J., She, D., Sun, M.: Joint image emotion classification and distribution learning via deep convolutional neural network. In: IJCAI (2017)
42. Yang, J., Sun, M., Sun, X.: Learning visual sentiment distributions via augmented conditional probability neural network. In: AAAI (2017)
43. Yang, X., Song, Z., King, I., Xu, Z.: A survey on deep semi-supervised learning. arXiv preprint arXiv:2103.00550 (2021)
44. You, Q., Luo, J., Jin, H., Yang, J.: Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In: AAAI (2016)
45. Zhan, C., She, D., Zhao, S., Cheng, M.M., Yang, J.: Zero-shot emotion recognition via affective structural embedding. In: ICCV (2019)

46. Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., Shinozaki, T.: Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In: NIPS (2021)
47. Zhang, H., Xu, M.: Weakly supervised emotion intensity prediction for recognition of emotions in images. IEEE Transactions on Multimedia (2020)
48. Zhao, S., Gao, Y., Jiang, X., Yao, H., Chua, T.S., Sun, X.: Exploring principles-of-art features for image emotion recognition. In: ACM MM (2014)
49. Zhao, S., Jia, G., Yang, J., Ding, G., Keutzer, K.: Emotion recognition from multiple modalities: Fundamentals and methodologies. IEEE Signal Processing Magazine **38**(6), 59–73 (2021)
50. Zhao, S., Jia, Z., Chen, H., Li, L., Ding, G., Keutzer, K.: Pdanet: Polarity-consistent deep attention network for fine-grained visual emotion regression. In: ACM MM (2019)
51. Zhao, S., Yao, H., Gao, Y., Ji, R., Xie, W., Jiang, X., Chua, T.S.: Predicting personalized emotion perceptions of social images. In: ACM MM (2016)
52. Zhao, S., Yao, X., Yang, J., Jia, G., Ding, G., Chua, T.S., Schuller, B.W., Keutzer, K.: Affective image content analysis: Two decades review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
53. Zhong, Y., Yuan, B., Wu, H., Yuan, Z., Peng, J., Wang, Y.X.: Pixel contrastive-consistent semi-supervised semantic segmentation. In: ICCV (2021)
54. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: NIPS (2004)
55. Zhu, X., Ghahramani, Z., Lafferty, J.D.: Semi-supervised learning using gaussian fields and harmonic functions. In: ICML (2003)