IEEE Open Journal of Signal Processing

Received 11 September 2023; revised 2 December 2023; accepted 4 December 2023. Date of publication 18 December 2023; date of current version 29 December 2023. The review of this article was arranged by Associate Editor F. Guerrini.

Digital Object Identifier 10.1109/OJSP.2023.3344079

Zero-Shot Visual Sentiment Prediction via Cross-Domain Knowledge Distillation

YUYA MOROTO [1] (Graduate Student Member, IEEE), YINGRUI YE [1], KEISUKE MAEDA [2] (Member, IEEE), TAKAHIRO OGAWA [2] (Senior Member, IEEE), AND MIKI HASEYAMA [2] (Senior Member, IEEE)

[1]Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan
[2]Faculty of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan

CORRESPONDING AUTHOR: MIKI HASEYAMA (e-mail: mhaseyama@lmd.ist.hokudai.ac.jp).

This work was supported in part by JSPS KAKENHI under Grant JP21H03456 and in part by AMED under Grant JP23zf01270004h0003.

ABSTRACT There are various sentiment theories for categorizing human sentiments into several discrete sentiment categories, which means that the theory used for training sentiment prediction methods does not always match that used in the test phase. As a solution to this problem, zero-shot visual sentiment prediction methods have been proposed to predict unseen sentiments for which no images are available in the training phase. However, the training of these previous zero-shot methods relies on a single sentiment theory, which limits their ability to handle sentiments from other theories. Thus, this article proposes a more robust zero-shot visual sentiment prediction method that can handle cross-domain sentiments defined in different sentiment theories. Specifically, by focusing on the fact that sentiments are abstract concepts common to humans regardless of whether their theories are different, we incorporate knowledge distillation into our method to construct a teacher–student model that can train the implicit relationships between sentiments defined in different sentiment theories. Furthermore, to enhance sentiment discrimination capability and strengthen the implicit relationships between sentiments, we introduce a novel sentiment loss between the teacher and student models. In this way, our model becomes robust to unseen sentiments by exploiting the implicit relationships between sentiments. The contributions of this article are the introduction of knowledge distillation and a novel sentiment loss between the teacher and student models for zero-shot visual sentiment prediction, and improved performance of zero-shot visual sentiment prediction. Experiments on several open datasets demonstrate the effectiveness of the proposed method.

INDEX TERMS Visual sentiment prediction, zero-shot learning, knowledge distillation, cross-domain analysis, text analysis.

I. INTRODUCTION

With the rapid development of the Internet and social networking services (SNS), an increasing number of people share their lives and experiences by posting images and texts. These multimedia contents are mostly related to their feelings. Under these circumstances, the visual sentiment prediction (VSP) task, which predicts users' sentiment from visual contents, has attracted significant attention due to its numerous applications, such as affective image retrieval [1], [2] and comment support for SNS [3], [4]. There are many sentiment theories defining sentiments using different numbers and types of categories. For instance, Ekman's [5] and Parrott's sentiment theories [6] use six different sentiments, whereas Mikels' sentiment theory [7] consists of eight sentiments. Thus, previous VSP methods, trained using a specific sentiment theory [8], [9], [10], [11], struggle to predict new sentiments in different theories, which limits their applicability. Furthermore, with the development of sentiment theories, more fine-grained sentiments are explored in new sentiment theories [12]. Therefore, the development of a VSP method that can handle unseen sentiments in different sentiment theories is desired.

To predict unseen classes for which there are no images in the training data, zero-shot learning (ZSL) methods have been proposed [13], [14]. In zero-shot VSP methods [15], [16], ancillary information (e.g., adjective-noun pair (ANP)

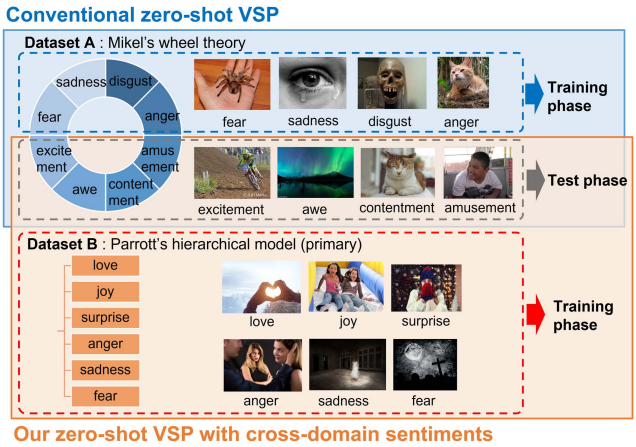

FIGURE 1. Novel problem settings in this article. We use knowledge distillation to train our zero-shot VSP method on one sentiment dataset and then test it on other datasets defined by different sentiment theories, whereas conventional zero-shot VSP methods are trained and tested on the same dataset.

[8]) is used to construct the embedding space that can bridge the affective gap between visual and semantic features related to sentiment labels. However, even these zero-shot VSP methods use a single sentiment dataset in the training phase, which means that their performance is strongly influenced by the domain gap between the sentiment theories used in the training and test phases. Thus, it is expected to reduce such domain gap by introducing a learning mechanism that can handle several datasets with different types of sentiment labels as shown in Fig. 1, and the robust VSP can be constructed by acquiring the high representation ability for the ambiguous sentiments. Here, although the discrete sentiments are often employed in the VSP task, it is reported that sentiments are continuous as Russell [17] has suggested. Therefore, previous VSP methods that focus on a single sentiment theory cannot deal with such complex sentiments, that is, they might be unrealistic approaches.

In this article, we propose a zero-shot VSP method, where knowledge distillation [18] is used to train the method with cross-domain sentiments defined in different sentiment theories. To bridge the domain gap between different sentiment theories, the knowledge distillation can be useful to obtain the implicit relationships between different sentiment theories, as sentiments are abstract concepts common to humans regardless of whether the sentiment theories are different. Specifically, we train the VSP method with a sentiment dataset as the teacher model and train the student model with another sentiment dataset that has different sentiments. By using cross-domain sentiments to train the two models based on knowledge distillation, our method can effectively optimize the embedding space. Furthermore, to improve the robustness of the embedding space, we introduce a sentiment loss between the teacher and student models to bridge the domain gap between different sentiment theories, enhancing the sentiment discrimination capability and strengthening the implicit relationships between sentiments. Consequently, a robust embedding space can be constructed that can help compensate for the affective gap between visual and semantic features. In this way, we improve the performance of zero-shot VSP, which can better handle new sentiments from different sentiment theories.

The contributions of this study are summarized as follows.

i) We propose a novel zero-shot VSP method robust to unseen sentiments in new sentiment theories by compensating for the domain gap between different sentiment theories through knowledge distillation.
ii) We enhance sentiment discrimination capability and strengthen the implicit relationships between sentiments by introducing a sentiment loss between the teacher and student models.
iii) The effectiveness of the proposed method is confirmed through experiments on several open datasets.

Specifically, the contribution (ii) is the novel and unique technique that can be effective for sentiment prediction based on knowledge distillation although we adopt the existing knowledge distillation framework.

II. RELATED WORKS

This section introduces several studies related to the proposed zero-shot VSP method via knowledge distillation.

A. VISUAL SENTIMENT PREDICTION

The purpose of the VSP task is to predict sentiment labels evoked by images [9]. Previous research in this area is similar to object recognition studies. Lately, deep learning-based VSP methods have demonstrated promising results. Tao et al. [8] employed a deep convolutional neural network (CNN) [19] to develop a VSP method. Yang et al. [10] proposed using image local information by implementing a weakly supervised coupled CNN for VSP. Lee et al. pay attention to semantic information obtained from objects for improving the performance of VSP, and the pre-trained word embedding succeed in extracting such information [20]. Although these approaches show performance improvement, they still require many labeled images, and annotating images with sentiment labels is challenging. Simultaneously, some VSP techniques attempt to apply few-shot learning approaches, which can be trained using a limited amount of labeled data. Few-shot VSP methods [11], [21] utilize prior knowledge of sentiment labels and effectively train the method. Nevertheless, these methods require at least one image of each sentiment to handle a new sentiment, indicating that they cannot handle unseen sentiments across different sentiment theories.

B. ZERO-SHOT LEARNING FOR VSP

ZSL methods aim to predict unseen classes for which there are no samples in the training data. In ZSL for the image classification task, ancillary information such as attributes or seen class labels are typically used to realize ZSL [13], [14]. Concretely, UMF [22] unified a multiplicative framework for attribute learning to solve the zero-shot problem. In

HAP [23], a hypergraph is constructed to incorporate information about attribute relationships in the training data. This approach transforms the attribute prediction problem into a regularized hypergraph partitioning problem, making it easier to understand and solve the zero-shot problem. HAT [24] utilizes the hierarchical structure of WordNet [25] to learn attribute classifiers for different categories independently to realize ZSL. However, these methods are designed for conventional image classification tasks for object recognition and do not capture the sentiment-related information required for sentiment recognition.

In the case of the zero-shot VSP task, it is important to extract sentiment-related information from training data. The EmotionGCN [26] applies graph convolutional networks [27] for sentiment distribution learning to capture the correlation between sentiments and realize ZSL. In addition, some methods use a common embedding space of visual features and semantic features (e.g., Word2vec features [28]). For example, in zero-shot VSP methods [15], [16], ancillary information such as ANP [8] and tweet representation (Tweet2vec [29]) are used to construct an embedding space that can bridge the affective gap between visual and non-visual semantic features. However, these zero-shot VSP methods mainly use a single sentiment dataset in the training phase, which means that their performance is strongly influenced by the domain gap between the sentiment theories in the training and test data. This limitation restricts their ability to adapt to new sentiment theories and unseen sentiment categories. Thus, to overcome this limitation, we introduce a novel sentiment loss to exploit the implicit relationships between seen and unseen sentiments.

C. KNOWLEDGE DISTILLATION FOR SENTIMENTS

Knowledge distillation [18], [30] is a technique for transferring knowledge from a teacher model to a smaller student model. In addition, knowledge distillation can effectively extract sentiment-related information for enhancing sentiment analysis. Albanie et al. [31] use cross-modal knowledge distillation for speech sentiment recognition without access to labeled audio data. Tang et al. [32] distill knowledge from the BERT model [33] into a single-layer bidirectional long short-term memory for natural language processing sentiment classification. DKDFMH [34] uses a fused multi-head attention mechanism to employ decoupled knowledge distillation in CNNs, helping the method focus on the distinctions between sentiment features.

In the context of VSP, knowledge distillation has great potential for learning the implicit relationships between sentiment theories and bridging the domain gap between them [21]. By using cross-domain sentiments to train the method based on knowledge distillation, the embedding space can be effectively optimized, allowing the method to better handle unseen sentiments from different sentiment theories. However, previous knowledge distillation approaches for cross-domain sentiments have not fully explored the potential of knowledge distillation for VSP. The use of a single sentiment dataset to train the method and the reliance on ancillary information limit the effectiveness of these methods in predicting unseen sentiments in new sentiment theories.

In this article, we address such limitations of existing ZSL methods and knowledge distillation approaches for cross-domain sentiments by proposing a novel zero-shot VSP method. Our method leverages knowledge distillation to learn the implicit relationships between sentiment theories and bridges the domain gap between them. Furthermore, we introduce a novel sentiment loss between the teacher and student models to enhance sentiment discrimination capability and strengthen the implicit relationships between sentiments. The details of our method are shown in Section III.

III. PROPOSED ZERO-SHOT VSP WITH CROSS-DOMAIN SENTIMENTS

This section presents the details of the proposed method as shown in Fig. 2. Concretely, the proposed method consists of two model training steps. In the first step, we train the teacher model with a sentiment dataset. Second, we use knowledge distillation and a novel sentiment loss to train the student model with another sentiment dataset, which can utilize the knowledge obtained in the teacher model. The teacher and student model are optimized with the loss functions L_{teacher} and L_{student} consisting of several losses, respectively, and the student model is optimized after the teacher model has been optimized. Hereafter, we explain the details of the proposed method by mainly focusing on each loss included in the teacher and student models in Section III-A and B, respectively, and the way to predict the sentiment label is explained in Section III-C.

A. TRAINING OF TEACHER MODEL

A sentiment dataset is used to train the teacher model in the training of the teacher model. To make the teacher model effectively learn affective structural information about images, we define the total objective function of the teacher model as follows:

$$L_{\text{teacher}} = L_{\text{as}} + L_{\text{vis}} + L_{\text{tweet}} + L_{\text{adv}}, \tag{1}$$

where L_{as} represents the total affective structural loss that helps embed the sentiment features. L_{vis} denotes the total embedding loss of visual features, and L_{tweet} represents the embedding loss of tweet features that helps embed visual and tweet features. L_{adv} denotes the adversarial constraint loss that optimizes visual and semantic feature embeddings dynamically.

1) Affective Structural Loss: The total affective structural loss is defined as follows:

$$L_{\text{as}} = L_{\text{re}} + L_{\text{w}}(z_y) + L_{\text{w}}(z_{y'}), \tag{2}$$

where L_{re} denotes the reconstruction loss calculated by the auto-encoder with fully connected layers, which embeds ANP features into the common embedding space $h(\boldsymbol{x})$

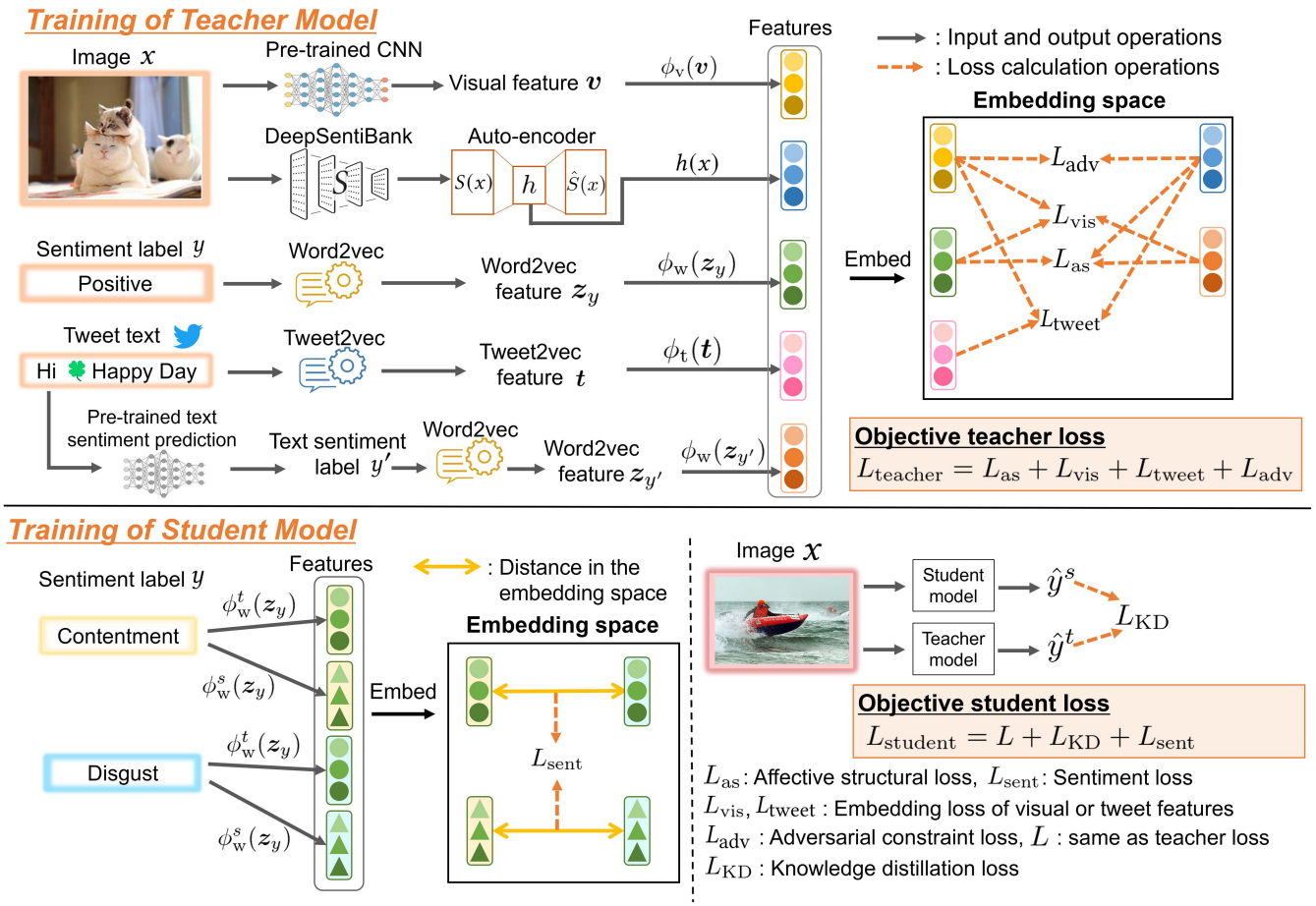

FIGURE 2. **Overview of the proposed zero-shot VSP method. We embed visual and several semantic features into the common embedding space to optimize the model. Concretely, by optimizing the teacher model using several losses with tweet texts, we effectively compensate for the affective gap between visual and sentiment features, making the teacher model more robust to the unseen sentiment labels. Here, we attempt to convert the knowledge included in the pretrained text sentiment prediction model to the VSP model for training with different sentiment theories in the ZSL framework. Moreover, in the training phase of the student model, the proposed method compensates for the domain gap between different sentiment theories and improves robustness to unseen sentiments in new sentiment theories by using knowledge distillation and the sentiment loss.**

as follows:

$$L_{\mathrm{re}} = \| \hat{S}(\boldsymbol{x}) - S(\boldsymbol{x}) \|_2^2, \tag{3}$$

$$\hat{S}(\boldsymbol{x}) = f(\boldsymbol{W}_2 h(\boldsymbol{x})),\ h(\boldsymbol{x}) = f(\boldsymbol{W}_1 S(\boldsymbol{x})), \tag{4}$$

where $S(\boldsymbol{x})$ denotes the ANP features of an image $\boldsymbol{x}$ calculated by the pretrained DeepSentiBank [8] model. $f(\cdot)$ denotes the activation function. $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ denote the weight matrixes of fully connected layers. $\hat{S}(\boldsymbol{x})$ denotes the reconstructed ANP features. In this way, we construct a common embedding space.

Next, we embed the sentiment feature z, extracted from the sentiment label y by the Word2vec model, into the common embedding space by a nonlinear embedding function $\phi_{\mathrm{w}}(\cdot)$. To minimize the distance between sentiment and ANP features in the common embedding space, the embedding loss is used as follows:

$$L_{\mathrm{w}}(z) = \| h(\boldsymbol{x}) - \phi_{\mathrm{w}}(z) \|_2^2. \tag{5}$$

In addition, we use not only the sentiment feature z_y but also the sentiment feature $z_{y'}$ extracted from the text sentiment label y', which is predicted from the texts of tweets using the pretrained text sentiment prediction model [35]. That is, we use $L_{\mathrm{w}}(z_y)$ and $L_{\mathrm{w}}(z_{y'})$ to calculate the embedding loss of the sentiment features in (2). In this way, we use the affective information extracted from tweet texts as the auxiliary data.

To introduce the affective structural loss into the training of the teacher model using the sentiment features extracted from the sentiment labels, we need to prepare images with the sentiment labels. Although there is an open dataset containing images, texts, and sentiment labels [36], its kind of sentiments is limited. Hence, we apply the pretrained text sentiment prediction model to the tweet texts to increase the kind of sentiments. Here, text sentiment prediction achieves the high accuracy in the field of natural language processing [37], [38], and we attempt to convert the knowledge included in the text sentiment prediction model to the VSP model for ZSL. In this way, our training framework of the teacher model

can be regarded as weakly supervised learning because we use only simple sentiments (positive, negative, and neutral) for supervision.

2) Embedding Loss of Visual Features: To embed visual features into the common embedding space, we combine visual and sentiment features using the embedding function $\phi_{\mathrm{v}}(\cdot)$, and then the embedding loss between visual and sentiment features is defined as follows:

$$L_{\mathrm{v}}(z) = \| \phi_{\mathrm{v}}(\boldsymbol{v}) - \phi_{\mathrm{w}}(z) \|_2^2, \tag{6}$$

where $\boldsymbol{v}$ denotes the visual feature extracted from the image $\boldsymbol{x}$ using the pretrained CNN. We define the embedding loss of visual features as follows:

$$L_{\mathrm{vis}} = L_{\mathrm{v}}(z_y) + L_{\mathrm{v}}(z_{y'}). \tag{7}$$

3) Embedding Loss of Tweet Features: We introduce the tweet features $\boldsymbol{t}$, extracted from tweet texts using the pretrained Tweet2vec model [29], as the auxiliary data to better compensate for the affective gap. The tweet features $\boldsymbol{t}$ are also embedded into the common embedding space by using a nonlinear embedding function $\phi_{\mathrm{t}}(\cdot)$. We calculate the embedding loss of tweet features as follows:

$$L_{\mathrm{tweet}} = \| h(\boldsymbol{x}) - \phi_{\mathrm{t}}(\boldsymbol{t}) \|_2^2 + \| \phi_{\mathrm{v}}(\boldsymbol{v}) - \phi_{\mathrm{t}}(\boldsymbol{t}) \|_2^2. \tag{8}$$

By optimizing our model using the constraint of tweet features, the common embedding space can improve the preservability of affective structural information.

4) Adversarial Constraint Loss: As both visual and semantic feature embeddings are dynamically optimized, we need to apply an adversarial constraint to fool the discriminator network D. We define the adversarial constraint loss as follows:

$$L_{\mathrm{adv}} = \mathbb{E}_{\boldsymbol{x}}(\log D(h(\boldsymbol{x}))) + \mathbb{E}_{\boldsymbol{x}}(\log[1 - D(\phi_{\mathrm{v}}(\boldsymbol{v}))]). \tag{9}$$

Note that we adopt the strategy of WGAN [39] to optimize adversarial learning.

The affective structural loss L_{as} makes semantic features have sentiment-related representations, whereas the visual feature loss L_{vis} and the adversarial constraint loss L_{adv} make the visual features better fit semantic features with the help of the constraint of tweet features L_{tweet}. By optimizing the teacher model with these losses, we effectively compensate for the affective gap between visual and sentiment features, making the teacher model more robust to unseen sentiment labels. In this way, semantic features can acquire sentiment-related representations, and semantic features of unseen sentiments more associated with visual features.

B. TRAINING OF STUDENT MODEL

To train the student model, we first initialize the student model by creating a clone of the teacher model, which is trained in Section III-A. In the training phase of the student model, we use another sentiment dataset to train the student model to improve the generalization ability for zero-shot VSP. To optimize the student model, we define the total objective function as follows:

$$L_{\mathrm{student}} = L + L_{\mathrm{KD}} + L_{\mathrm{sent}}, \tag{10}$$

where L represents the same loss used in the teacher model for the student model, L_{KD} represents the knowledge distillation loss and L_{sent} represents the sentiment loss between the teacher and student models.

1) Knowledge Distillation Loss: By optimizing the student model using the knowledge distillation loss, the generalization ability can be improved. We define the knowledge distillation loss using $\mathrm{KL}(\cdot, \cdot)$, the Kullback–Leibler divergence [40], as follows:

$$L_{\mathrm{KD}} = \mathrm{KL}(\sigma(\hat{y}^s/T), \sigma(\hat{y}^t/T)), \tag{11}$$

where T represents the temperature parameter of knowledge distillation, and $\sigma(\cdot)$ denotes the softmax function. $\hat{y}^s$ and $\hat{y}^t$ denote the sentiment predictions of the image $\boldsymbol{x}$ calculated by the student and teacher models as follows:

$$\hat{y}^m = \underset{y \in Y_s^{\mathrm{train}}}{\arg\min} \| \phi_{\mathrm{v}}^m(\boldsymbol{v}) - \phi_{\mathrm{w}}^m(z_y) \|_2^2, \quad m \in \{s, t\}, \tag{12}$$

where Y_s^{train} are the sentiment labels used in the training of the student model. $\{\phi_{\mathrm{v}}^s(\cdot), \phi_{\mathrm{w}}^s(\cdot)\}$ and $\{\phi_{\mathrm{v}}^t(\cdot), \phi_{\mathrm{w}}^t(\cdot)\}$ denote the embedding functions of the student and teacher models, respectively.

2) Sentiment Loss: The sentiment loss L_{sent} helps the student model bridge the domain gap between different sentiment theories using the relationship between the same sentiments in different sentiment theories. The sentiment loss is defined as follows:

$$L_{\mathrm{sent}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \Big\| \|\phi_{\mathrm{w}}^t(z_{y_i}) - \phi_{\mathrm{w}}^t(z_{y_j})\|_2^2 - \|\phi_{\mathrm{w}}^s(z_{y_i}) - \phi_{\mathrm{w}}^s(z_{y_j})\|_2^2 \Big\|, \tag{13}$$

where $y_i, y_j \in Y^{\mathrm{com}}$ represent sentiment labels. Y^{com} represent common sentiments used in the training of both the student and teacher models ($Y^{\mathrm{com}} = Y_s^{\mathrm{train}} \cap Y_t^{\mathrm{train}}$, and Y_t^{train} being the sentiment labels used in the teacher model).

In the training phase of the student model, the proposed method compensates for the domain gap between different sentiment theories and improves robustness to unseen sentiments in new sentiment theories by transferring the knowledge distilled from the teacher model. Besides, the novel sentiment loss between the teacher and student models can enhance sentiment discrimination capability and strengthen the implicit relationships between sentiments.

C. TEST PHASE

In the test phase, we simply adopt the nearest-neighbor search in the common embedding space for predicting the sentiment label of the test image as follows:

$$\hat{y} = \underset{y \in Y^{\mathrm{test}}}{\arg\min} \| \phi_{\mathrm{v}}^s(\boldsymbol{v}) - \phi_{\mathrm{w}}^s(z_y) \|_2^2, \tag{14}$$

where $\boldsymbol{v}$ denotes the visual feature of the test image $\boldsymbol{x}$, z_y denotes the semantic feature corresponding to the unseen sentiment label y, and Y^{test} denotes the list of sentiment labels in the sentiment theory for test data. $\hat{y}$ denotes the predicted sentiment label of the test image $\boldsymbol{x}$. In this way, the label of the test image $\boldsymbol{x}$ is predicted using the robust embedding space constructed in the training phase of the student model.

IV. EXPERIMENT

In this section, we confirm the effectiveness of the proposed zero-shot VSP method on several open datasets.

A. EXPERIMENTAL SETUPS

In this experiment, we used several open datasets based on different sentiment theories as training and test sets. It should be noted that the test set is different from the training set used in the training phase of both student and teacher models.

Training set: We selected the Twitter for Sentiment Analysis (T4SA) dataset [36], which includes randomly collected English tweets accompanied by images. The T4SA dataset contains 1,179,957 tweets and 1,473,394 associated images (some tweets have multiple associated images). These tweets and images are labeled with one of three sentiments: negative, neutral, or positive. We used 974,053 images from the T4SA dataset after removing corrupted and near-duplicate images. In addition, we apply the text sentiment prediction model [35] to tweet texts in the T4SA dataset to train the teacher and student models. Concretely, in the training set of the teacher model, we used Ekman's six sentiments (joy, fear, anger, disgust, surprise, and sadness) [5] as the text sentiment labels y'. In the training set of the student model, we used Plutchik's eight sentiments (joy, fear, anger, trust, disgust, sadness, anticipation, and surprise) [41] as the text sentiment labels y'.

Test set: We selected four datasets commonly used in VSP tasks: the Flickr and Instagram (FI) dataset [42], ArtPhoto (ART) dataset [43], Abstract Paintings (ABST) dataset [43], and WEBEmo dataset [12]. The FI dataset was collected from SNS using sentiment labels as keywords and contains 21,829 images labeled according to Mikels' sentiment theory (amusement, anger, awe, contentment, disgust, excitement, fear, and sadness). The ART dataset, sourced from an art-sharing website, contains 807 artistic images. The ABST dataset contains 280 abstract paintings. Both the ABST and ART datasets have images labeled using Mikels' sentiment theory, similar to the FI dataset. The WEBEmo dataset is a large-scale web containing 267,438 images and is labeled by 25 fine-grained sentiments according to Parrott's sentiment theory [6] (joy, lust, envy, rage, zeal, neglect, horror, shame, suffering, disgust, confusion, pride, anger, sympathy, zest, relief, optimism, sadness, gratitude, cheerfulness, exasperation, disappointment, nervousness, surprise, enthrallment, and contentment).

As the zero-shot settings in the testing phase using the FI, ART, and ABST datasets, we used four unseen sentiments (amusement, excitement, awe, and contentment) that are not appeared in the training phase of both student and teacher models. In the test phasing using the ABST dataset, we also used all eight sentiments (including seen and unseen sentiments) for comparison with the ideal accuracy obtained from human prediction.[1] In the testing phase using the WEBEmo dataset, we used 20 unseen sentiments, which were grouped by five sentiments with more fine-grained sentiments (Joy: pride, zest, relief, optimism, enthrallment, contentment, and cheerfulness. Love: affection, gratitude, and lust. Anger: exasperation, irritability, rage, and envy. Sadness: disappointment, neglect, shame, suffering, and sympathy. Fear: horror and nervousness).

We used ResNet-50 [44] pretrained on ImageNet [45] as our backbone CNN, with the embedding feature dimension set to 1,024. A fully connected layer was added before the ReLU layer to embed visual and semantic features into the common embedding space. Here, the pretrained ResNet-50 model is one of the most fundamental image encoder models, and the various studies on the application of deep learning adopt this model. By the same token, in the other studies on zero-shot VSP [15], [16], only the ResNet50 model has been adopted. Moreover, in the study on VSP [46], the authors have adopted the ResNet50 model. In this way, these previous studies demonstrated that the outputs of the ResNet-50 can be enough to have the discrimination ability as the visual features, and thus, we performed our experiments on the pretrained ResNet-50.

To demonstrate the effectiveness of sentiment-related information in the proposed zero-shot VSP method, we adopted conventional ZSL methods, DEM [47], RN [48], and SAE [49] for comparison. Furthermore, as there are some ZSL methods specific to VSP, we adopted ASE [15] and AEF (state-of-the-art) [16] as the comparison methods. Note that AEF is good at predicting the sentiments of images on SNS, whereas ASE outperforms AEF in predicting the sentiments of abstract paintings difficult to describe in human words. We evaluated the methods in terms of average accuracy.

B. RESULTS AND DISCUSSION

1) RESULTS OF SENTIMENTS IN MIKELS' SENTIMENT THEORY

We experimented with the FI, ART, and ABST datasets with four unseen sentiments to prove the effectiveness of the proposed method. In addition, we used eight sentiments, including four seen sentiments in the ABST dataset, to evaluate all methods, including human prediction results for reference. Here, by comparing the human prediction results with these methods, we cannot only observe the difference between each method and human intuition but also demonstrate the difficulty of the VSP task. We present the experimental results in Table 1.

[1]Human prediction in the ABST dataset was based on the voting results of annotators. The ground truth was determined by a majority vote of approximately 10 annotators per image. The human prediction accuracy was the proportion of annotators who voted for the same sentiment as the ground truth. Note that this result was originally included in the ABST dataset.

TABLE 1. Accuracy on FI, ART, and ABST Datasets. The Bolded Text Represents the Highest Accuracy, Whereas the Underlined Text Represents the Second Highest Accuracy. The "number of Sentiments" With 2, 3, or 4 Refers to Testing With 2, 3, or 4 Sentiments of the 4 Unseen Sentiments (Amusement, Excitement, Awe, and Contentment). Specifically, We Conducted Experiments on All Possible Combinations (E.g., When Considering 3 Sentiments, There are 4 Possibilities: "amusement, Excitement, Awe," "amusement, Excitement, Contentment," "amusement, Awe, Contentment," and "excitement, Awe, Contentment") and Then Calculated the Average Results for Each Case. This Experimental Setting Was Referred to [15], [16]

	FI			ART			ABST			
Number of sentiments	2	3	4	2	3	4	2	3	4	8
DEM [47]	56.51%	52.20%	36.21%	50.87%	40.12%	26.43%	59.41%	35.62%	25.50%	18.57%
RN [48]	56.77%	51.91%	35.62%	51.33%	38.54%	24.23%	58.64%	34.09%	24.39%	17.14%
SAE [49]	53.48%	49.02%	35.17%	51.59%	39.73%	25.37%	58.73%	36.81%	26.11%	18.21%
ASE [15]	69.15%	61.19%	43.38%	54.07%	44.80%	30.91%	61.75%	40.73%	33.20%	21.79%
AEF [16]	68.79%	61.83%	44.26%	54.13%	**45.32%**	31.64%	61.21%	40.49%	32.88%	21.43%
human prediction	-	-	-	-	-	-	-	-	-	**38.61%**
PM	**70.72%**	**62.31%**	**46.39%**	**55.00%**	45.15%	**33.14%**	**62.94%**	**40.95%**	**34.13%**	22.14%

The effectiveness of incorporating knowledge distillation and sentiment loss into our method was proved by comparing the proposed method (PM) with AEF, which used the same semantic features but without employing knowledge distillation or sentiment loss. According to the results, classic ZSL methods such as DEM, RN, and SAE are remarkably outperformed by methods that utilize sentiment-related information such as ASE, AEF, and PM. When comparing ASE, AEF, and PM, although the state-of-the-art method AEF performed well on datasets related to images on SNS, especially on the ART dataset, AEF was still inferior to ASE for abstract paintings in the ABST dataset. Conversely, although PM performed slightly worse than AEF when the sentiment setting was three in the ART dataset, it outperformed AEF in all other experimental settings, including the ABST dataset. Therefore, comparing PM with other methods, PM outperformed other methods on these datasets, which proved its effectiveness and robustness.

Regarding the FI dataset, PM achieved an accuracy of 70.72% for two sentiments, 62.31% for three sentiments, and 46.39% for four sentiments. These results were significantly higher than those achieved by other methods, including the state-of-the-art zero-shot VSP method, AEF, which achieves 68.79%, 61.83%, and 44.26%, respectively. For the ART dataset, PM achieved an accuracy of 55.00% for two sentiments and 33.14% for four sentiments, which were also higher than those achieved by other methods. Besides, PM achieved an accuracy of 45.15% for three sentiments, which was close to the 45.32% achieved by the state-of-the-art VSP method, AEF. Similar trends can be observed in the ABST dataset for two- and four-sentiment settings.

For eight sentiments (4 unseen sentiments and 4 seen sentiments) on the ABST dataset, PM also outperformed the other methods in all sentiment settings. Specifically, PM achieved an accuracy of 22.14% for eight sentiments, which was most close to the human prediction results. It can be seen that the average human prediction result is only 38.61%, indicating the difficulty of the VSP task because each person's intuition varies. Furthermore, because the absolute value of the human prediction results were low, small improvements in accuracy for the PM are meaningful. Thus, although PM was only 0.35% higher than the second-best method, we proved the effectiveness of PM.

2) RESULTS OF SENTIMENTS IN WEBEMO DATASET

We experimented with the WEBEmo dataset to further validate the effectiveness and robustness of PM. To demonstrate that PM is capable of performing more detailed classifications of sentiments, we conducted experiments using more fine-grained sentiments of five groups. We show the experimental results in Table 2. On this dataset, PM outperformed the other methods for most sentiments. Specifically, for the "Joy" category, PM achieved the highest accuracy of 18.29%, outperforming the second-best method, AEF, by 0.72%. For the" Love" category, PM also achieved the best accuracy of 39.17%, which was 1.45% higher than that of the second-best method, ASE. For the "Anger" category, PM was demonstrated its superiority with an accuracy of 31.41%, surpassing the second-best method, AEF, by 0.75%. For the "Sadness" category, PM achieved the highest accuracy of 22.10%, outperforming the second-best method by 0.16%.

According to the results, PM, which was trained with Ekman's six sentiments and Plutchik's eight sentiments, can effectively leverage the sentiment-related information learned through knowledge distillation and sentiment loss for more fine-grained sentiment prediction with more than two sentiments. Conversely, when there are two candidate sentiments in the Fear sentiment, the number of sentiments can be too small for PM to effectively exploit the distance relationships between different sentiments. Thus, the performance of PM for the Fear sentiment is inferior to that of ASE and AEF. Although PM did not achieve the highest accuracy for the Fear sentiment, its performance of 60.93% was very close to the best-performing method, ASE achieving 62.20%.

By analyzing and comparing the zero-shot VSP results, PM demonstrated higher accuracy in predicting various sentiment labels compared with the other methods, which tend to confuse these sentiments with other sentiments. It is confirmed that PM effectively using cross-domain knowledge is robust to cross-domain sentiments when the target sentiments are fine-grained.

TABLE 2. Accuracy on the WEBEmo Dataset. We Grouped 20 Unseen Sentiments From the WEBEmo Dataset Into Their Corresponding Sentiment Categories and Then Tested the Methods on Them (E.g., We Predicted the Sentiment Label From Three Sentiments in the "love" Group and From Two Sentiments in the "fear" Group)

	Joy (pride, zest, relief, optimism, enthrallment, contentment, cheerfulness)	Love (affection, gratitude, lust)	Anger (exasperation, irritability, rage, envy)	Sadness (disappointment, neglect, shame, suffering, sympathy)	Fear (horror, nervousness)
DEM [47]	15.39%	34.51%	27.25%	21.10%	53.65%
RN [48]	14.11%	33.29%	26.66%	19.16%	54.76%
SAE [49]	14.80%	35.02%	27.31%	20.85%	51.86%
ASE [15]	17.34%	37.72%	29.92%	21.94%	**62.20%**
AEF [16]	17.57%	37.40%	30.66%	21.81%	61.77%
PM	**18.29%**	**39.17%**	**31.41%**	**22.10%**	60.93%

TABLE 3. Ablation Studies of Each Part of the Losses of the Student Model. Note That AS1 Represents Ablation Study 1, and AS2 Represents Ablation Study 2. Note That 2, 3, or 4 of FI, ART, and ABST Refer to Testing With 2, 3, or 4 Sentiments Out of the 4 Unseen Sentiments Like the Settings in Table 1

Loss	L	L_{KD}	L_{sent}	FI			ART			ABST		
				2	3	4	2	3	4	2	3	4
PM	✓	✓	✓	**70.72%**	**62.31%**	**46.39%**	**55.00%**	45.15%	33.14%	**62.94%**	**40.95%**	**34.13%**
AS1	✓	✓		69.12%	61.72%	46.10%	54.28%	44.49%	32.38%	61.72%	39.79%	33.37%
AS2	✓		✓	69.48%	61.20%	45.93%	54.78%	45.18%	**33.62%**	62.89%	40.57%	33.82%
Teacher model [16]	✓			68.79%	61.83%	44.26%	54.13%	**45.32%**	31.64%	61.21%	40.49%	32.88%

C. ABLATION STUDY

Table 3 shows the experimental results of the ablation studies to demonstrate the effectiveness of each part of the student model's losses. In this experiment, we performed ablation studies on knowledge distillation and sentiment losses since the other losses in the teacher model have already been shown to be effective in [16]. PM used the sum of L, L_{KD}, and L_{sent} as the total objective loss in the training of the student model. Conversely, in teacher model that is AEF [16], we used only the sum of L, in the ablation study 1 (AS1), we used the sum of L and L_{KD} as the total objective loss, and in the ablation study 2 (AS2), we used the sum of L and L_{sent} as the total objective loss. As compared above, PM is superior to teacher model [16], which indicates that the use of several sentiment theories for training through the knowledge distillation framework is effective.

For the FI dataset, AS1 and AS2 demonstrated varying levels of accuracy for different numbers of sentiments. Specifically, AS1 outperformed AS2 for three and four sentiments, whereas AS2 outperformed AS1 for two sentiments. This difference indicated that using L_{KD} and L_{sent} individually can improve the performance to some extent, but the combination of both loss functions in PM leads to the best results. Similar trends can be observed in the ABST dataset, where PM outperformed the ablation studies.

For the ART dataset, AS2 achieved an accuracy of 54.78% for two sentiments, 45.18% for three sentiments, and 33.62% for four sentiments, which outperformed PM for three and four sentiments. These results show that L_{sent} is more effective than L_{KD} for images with artistic styles such as the images in the ART dataset (consisting mainly of art photographs). As a result, PM achieved an accuracy of 55.00% for two sentiments, which outperformed the accuracy of AS2 (54.78%).

V. CONCLUSION

This article has proposed a zero-shot VSP method based on cross-domain knowledge distillation. By introducing knowledge distillation and a new sentiment loss between the teacher and student models using cross-domain sentiments, the proposed method can compensate for the domain gap between sentiment theories and improve sentiment discrimination capability. In the experiment on several open datasets, the proposed method outperformed other methods, including the state-of-the-art method.

The importance of each loss component in the proposed method cannot be inferred or adjusted although there are several loss components. Therefore, in future work, such importance should be adjusted by introducing hyperparameters corresponding to each loss component, and a detailed examination when changing them is needed.

REFERENCES

[1] L. Pang, S. Zhu, and C.-W. Ngo, "Deep multimodal learning for affective analysis and retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2008–2020, Nov. 2015.
[2] W. Wang and Q. He, "A survey on emotional semantic image retrieval," in *Proc. IEEE Int. Conf. Image Process.*, 2008, pp. 117–120.
[3] Y.-Y. Chen, T. Chen, W. H. Hsu, H.Y.M. Liao, and S.-F. Chang, "Predicting viewer affective comments based on image content in social media," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2014, pp. 233–240.
[4] Y.-Y. Chen, T. Chen, T. Liu, H.Y.M. Liao, and S.-F. Chang, "Assistive image comment robot—A novel mid-level concept-based representation," *IEEE Trans. Affect. Comput.*, vol. 6, no. 3, pp. 298–311, Jul.-Sep. 2015.
[5] P. Ekman and D. Cordaro, "What is meant by calling emotions basic," *Emotion Rev.*, vol. 3, no. 4, pp. 364–370, 2011.
[6] W P. Gerrod, *Emotions in Social Psychology: Essential Readings*, England, U.K.: Psychol. press, 2001.
[7] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, "Emotional category data on images from the international affective picture system," *Behav. Res. Methods*, vol. 37, no. 4, pp. 626–630, 2005.

[8] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks," 2014, *arXiv:1410.8586*.

[9] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdiscipl. Rev.: Data Mining Knowl. Discov.*, vol. 8, no. 4, 2018, Art. no. e1253.

[10] J. Yang, D. She, Y.-K. Lai, P. L. Rosin, and M.-H. Yang, "Weakly supervised coupled networks for visual sentiment analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7584–7592.

[11] L. Wang, X. Xu, F. Liu, X. Xing, B. Cai, and W. Lu, "Robust emotion navigation: Few-shot visual sentiment analysis by auxiliary noisy data," in *Proc. IEEE Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos*, 2019, pp. 121–127.

[12] R. Panda, J. Zhang, H. Li, J.-Y. Lee, X. Lu, and A. K. Roy-Chowdhury, "Contemplating visual emotions: Understanding and overcoming dataset bias," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 1–17.

[13] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–37, 2019.

[14] Y.anwei Fu, T. Xiang, Y.-G. Jiang, X.iangyang Xue, L. Sigal, and S. Gong, "Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 112–125, Jan. 2018.

[15] C. Zhan, D. She, S. Zhao, M.-M. Cheng, and J. Yang, "Zero-shot emotion recognition via affective structural embedding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1151–1160.

[16] Y. Ye, Y. Moroto, K. Maeda, T. Ogawa, and M. Haseyama, "Affective embedding framework with semantic representations from tweets for zero-shot visual sentiment prediction," in *Proc. ACM Int. Conf. Multimedia Asia*, 2022, pp. 1–7.

[17] J. A. Russell, "A circumplex model of affect," *J. Pers. Social Psychol.*, vol. 39, no. 6, 1980, Art. no. 1161.

[18] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[19] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022.

[20] S. Lee, C. Ryu, and E. Park, "OSANet: Object semantic attention network for visual sentiment analysis," *IEEE Trans. Multimedia*, vol. 25, pp. 7139–7148, 2023.

[21] Y. Ye, Y. Moroto, K. Maeda, T. Ogawa, and M. Haseyama, "Visual sentiment prediction using cross-way few-shot learning based on knowledge distillation," in *Proc. IEEE Int. Conf. Image Process.*, 2022, pp. 3838–3842.

[22] K. Liang, H. Chang, S. Shan, and X. Chen, "A unified multiplicative framework for attribute learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2506–2514.

[23] S. Huang, M. Elhoseiny, A. Elgammal, and D. Yang, "Learning hypergraph-regularized attribute predictors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 409–417.

[24] Z. Al-Halah and R. Stiefelhagen, "How to transfer? zero-shot object recognition via hierarchical transfer of semantic attributes," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2015, pp. 837–843.

[25] C. Fellbaum, "WordNet," in *Theory and Applications of Ontology: Computer Applications*. Berlin, Germany: Springer, 2010, pp. 231–243.

[26] T. He and X. Jin, "Image emotion distribution learning with graph convolutional networks," in *Proc. Int. Conf. Multimedia Retrieval*, 2019, pp. 382–390.

[27] S. Zhang, H. Tong, J. Xu, and R. Maciejewski, "Graph convolutional networks: A comprehensive review," *Comput. Social Netw.*, vol. 6, no. 1, pp. 1–23, 2019.

[28] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[29] B. Dhingra, Z. Zhou, D. Fitzpatrick, M. Muehl, and W. W. Cohen, "Tweet2Vec: Character-based distributed representations for social media," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 269–274.

[30] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3048–3068, Jun. 2022.

[31] S. Albanie, A. Vedaldi, A. Nagrani, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 292–301.

[32] R. Tang, Y. Lu, L. Liu, L. Mou, O. Vechtomova, and J. Lin, "Distilling task-specific knowledge from BERT into simple neural networks," 2019, *arXiv:1903.12136*.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.

[34] Z. Zhao, H. Wang, H. Wang, and B. Schuller, "Hierarchical network with decoupled knowledge distillation for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[35] N. Colnerič and J. Demšar, "Emotion recognition on Twitter: Comparative study and training a unison model," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 433–446, Jul.-Sep., 2020.

[36] L. Vadicamo et al., "Cross-media learning for image sentiment analysis in the wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2017, pp. 308–317.

[37] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100943–100953, 2019.

[38] E. Batbaatar, M. Li, and Keun Ho Ryu, "Semantic-emotion neural network for emotion recognition from text," *IEEE Access*, vol. 7, pp. 111866–111878, 2019.

[39] M. Arjovsky, S. Chintala, and Léon Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[40] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[41] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of Emotion*. Cambridge, MA, USA: Academic press 1980, pp. 3–33.

[42] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 308–314.

[43] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 83–92.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[46] Z. Wei et al., "Learning visual emotion representations from web data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13106–13115.

[47] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2021–2030.

[48] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. Torr, and T. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1199–1208.

[49] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3174–3183.