Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 2477605, 15 pages
https://doi.org/10.1155/2022/2477605

Research Article

Image Sentiment Analysis via Active Sample Refinement and Cluster Correlation Mining

Hongbin Zhang,[1] **Haowei Shi**,[1] **Jingyi Hou**,[1] **Qipeng Xiong**,[1] **and Donghong Ji**[2]

[1]*School of Software, East China Jiaotong University, Nanchang 330013, China*
[2]*School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China*

Correspondence should be addressed to Hongbin Zhang; zhanghongbin@whu.edu.cn

Received 15 November 2021; Revised 12 January 2022; Accepted 7 February 2022; Published 24 March 2022

Academic Editor: Yugen Yi

Training an effective image sentiment analysis model using high-quality samples and the implicit cross-modal semantics among heterogeneous features is still challenging. To address this problem, we propose an active sample refinement (ASR) strategy to obtain sufficient high-quality images with definite sentiment semantics. We mine the cluster correlation among the heterogeneous SENet features. Discriminative cross-modal semantics is generated to train an effective but robust image classifier. Ensemble learning is employed to further boost performance. Our method outperforms other competitive baselines, demonstrating its effectiveness and robustness. Meanwhile, the ASR strategy is a useful supplement to the current data augmentation method.

1. Introduction

In early research studies, text-based sentiment analysis [1–4] has been proposed to predict the implicit sentiment of texts. These methods have demonstrated evident progress in sentiment analysis. However, visual information accounts for about 80% of all perception in the human brain. With the rapid development of social media (e.g., WeChat, Twitter, and microblog), an increasing number of people would like to use all sorts of photographs to express their private emotions. Hence, besides texts, image (or any kind of visual information) is another important complementary way to characterize human emotions. Accurate assessment of the implicit emotions in images, called image sentiment analysis, has become an active and hot research topic in the field of computer vision (CV). More attractively, this research topic has many practical applications, including in education, entertainment, psychotherapy, and advertisements [5, 6]. As shown in Figure 1, we can make real-time fashion tendency prediction (left) or public opinion monitoring (right) based on the results of image sentiment analysis. In particular, if we want to know whether a commodity is popular, we typically look at how much the item has been purchased. Meanwhile, the proportion of positive comments is another important factor. As we know, there are many images uploaded by users in these comments. Accurate analysis of these images can evaluate the positive rating and then make fashion trendency prediction. For public opinion monitoring, we can make the real-time monitoring of the images transferred on the social platforms. If some images are detected to be offensive or negative, then we need to focus on this user and take the necessary processing step. Therefore, learning the visual sentimental information has the following advantages. (1) It helps explain human emotions in terms of the visual signal rather than textual signal. (2) It helps produce more personalized predictions that are consistent with human's real preferences. Owing to these advantages, its interpretability, and numerous practical applications, image sentiment analysis has attracted more and more attention from academia and industry.

However, owing to the subjective differences among different annotators, numerous semantically ambiguous images remain in the existing benchmark datasets. As we know, image sentiment annotation is more difficult than other traditional CV tasks because the emotions hidden in images are not independent in the aspect of semantics.

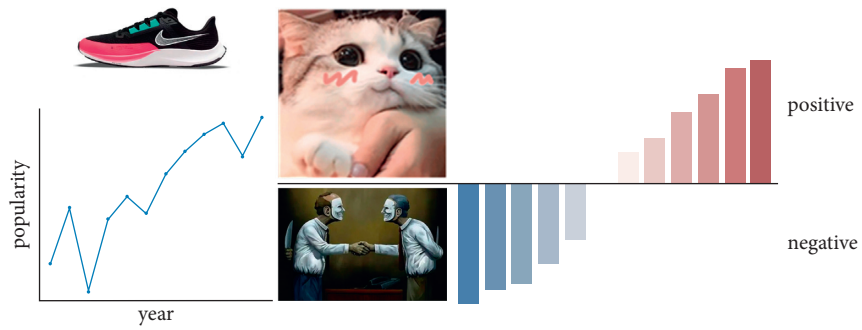


Figure 1: Practical applications of image sentiment analysis.

Contrarily, those concrete objects (e.g., person, cat, dog, bike, and bird) are independent [7]. Moreover, few studies have attempted to exploit the implicit correlation among the heterogeneous features from a homologous neural network. To address these problems, we propose a simple yet effective model for image sentiment analysis, which can not only adaptively augment the original public datasets but also make full use of the implicit complementary information among the heterogeneous SENet features. All these ideas can boost the final classification performance and strengthen the practicality of our model.

Conceptually and empirically, the main contributions of this study are summarized as follows:

(1) We proposed an active sample refinement (ASR) strategy that can adaptively "generate" (not really generating but mining) high-quality images with definite sentiment semantics, which builds a firm data foundation for model training. More importantly, the ASR strategy is a useful supplement to the current data augmentation method. It can be seamlessly incorporated into the state-of-the-art baselines to achieve better performance.

(2) We created a set of more discriminant but robust features by fully mining the implicit cluster correlation among the heterogeneous SENet features. These features can more accurately and comprehensively characterize the key sentimental semantics in images.

(3) We conducted extensive experiments using two benchmark datasets. The results demonstrate the superior performance of our model over other state-of-the-art baselines. The code for our method is available at https://github.com/Danicaghost/SHW.git.

2. Related Works

Image sentiment analysis is still a challenging CV task, which makes its automated classification trickier [8, 9]. Initially, many hand-crafted image features that focus on low-level visual information, including color, texture, and shape, were used to perform image sentiment analysis. For instance, Machajdik and Hanbury [10] used a set of color and textural features to represent the visual sentiment content in images. Lu et al. [11] investigated how the shape characteristics in images affect human emotions. Ko and Kim [12] utilized both color and scale-invariant feature transform (SIFT) features to train a pLSA-based model for image sentiment analysis. These hand-crafted features [10–12] usually require manual interventions and are insufficient to characterize the deep-level sentiment semantics in images. Moreover, any single image feature is hard to integrally describe an image. Hence, Zhang et al. [13] proposed a feature mid-fusion algorithm called gene selection XGBoost (GS-XGB) to mine the implicit correlation among a group of image features [14], which obtained satisfactory performance of image sentiment analysis. However, the feature mid-fusion algorithm requires numerous features and is intricate to realize. Furthermore, it separates the inherent dependencies between different classes. Unlike feature mid-fusion, Zhang et al. [14] employed discriminant correlation analysis (DCA) [15] to complete feature early-fusion and mine the correlation among a set of heterogeneous features. The corresponding ME^{2}M model [14] is effective and efficient. But the feature dimension after the DCA operation is too low, which loses some important information. The above studies [10–14] are fully supervised, which need massive accurate sentiment annotations.

Recently, convolutional neural networks (CNNs) have demonstrated the ability to learn more representative features. Some state-of-the-art methods employed CNN to complete image sentiment analysis. Das et al. [16] proposed a deep learning model including an attention mechanism for focusing local regions and determining the required sentiment. She et al. [17] proposed a model called weakly supervised coupled network (WSCNet) that aims to automatically select relevant regions to reduce the burden of sentiment annotation and improve performance. Zhu et al. [18] combined a weakly supervised learning strategy with a CNN model to complete end-to-end image sentiment prediction. To align image regions for gaining spatial invariance and learning strongly localized features, Durand et al. [19] designed a new method called weakly supervised learning of deep ConvNets for image classification, pointwise localization, and segmentation (WILDCAT). Evidently, these weakly supervised methods [16–19] are dedicated to easing the annotation burden. Similarly, Sun et al. [20] discovered the affective regions via a deep network. Rao et al. [21] designed a multilevel region-based framework to

explore the sentiment response of local regions. In [22], a model called MldrNet, which learns multilevel deep representations, was proposed for image sentiment classification. Furthermore, Wu et al. [23] developed a multiattention model for jointly discovering and localizing multiple relevant local regions that gave predicted attributes. Although the sentiment annotation burden has been alleviated to some extent, the corresponding classification performance of most weakly supervised methods [17–23] is unsatisfactory. Unlike the above studies, Simonyan and Zisserman [24] proposed SmileyNet using a novel sentiment-aligned image embedding to leverage the intricate relationship between emojis and images in large-scale available social media data. However, the SmileyNet model leverages additional emoji information, which may be scarce in most public available datasets.

Hence, how to make full use of the existing benchmark datasets (this helps alleviate sentiment annotation burden from another perspective) and reduce the impact of ambiguous annotations have become a hot concern. Recently, some researchers focused on utilizing state-of-the-art domain adaptation methods to address the ambiguous annotation problem caused by subjective differences. For example, Zhu et al. [25] translated each image from a source domain to a target domain in the absence of paired examples. Zhao et al. [26] proposed CycleEmotionGAN for image sentiment classification by adapting source domain images to have distributions like those of the target images by enforcing emotional semantic consistency. Lin et al. [27] proposed a multisource sentiment generative adversarial network (MSGAN), which uses a unified sentimental latent space to handle data from multiple domains. These methods [25–27] can make full use of the existing data resources. However, the GAN-based domain adaption methods [26, 27] are difficult to train and require additional computing resources. In contrast to these works [25–27], Zhang et al. [14] designed a static sample refinement method that is relatively easy to reproduce and obtain high-quality images. Valuable knowledge can be mined from the refined samples and adaptively merged into the classification model, which helps boost the final classification performance. Unlike those GAN-based methods, the goal of the refinement-based method is to fully refine the original samples. Additionally, the principle of the GAN-based and refinement-based methods has a large difference. The GAN-based methods train a network to minimize the generated image and real image. However, the refinement-based model usually employs the traditional machine learning method, and it refines samples based on the diverse predictions of different classifiers. Hence, our method belongs to the refinement-based method. Certainly, the refinement-based method can also be combined with the GAN-based method to further improve the image quality.

Recently, some researchers advocated combining more than one modality to infer perceived sentiment. Compared to a single modality, multimodalities have richer semantic information. Mittal et al. [28] proposed a M3ER method, which combines the cues from multiple co-occurring modalities such as face, text, and speech. Lu et al. [29] borrowed the idea of bidirectional encoder representations from transformers (BERTs) and proposed a model for learning task-agnostic joint representations of image content and natural language, namely, short for vision-and-language BERT (ViLBERT). The ViLBERT model can process both visual and textual inputs in the separate streams that interact through co-attentional transformer layers. Yang et al. [30] proposed a multimodal sentiment analysis model based on the multiview attentional network (MVAN), which utilizes a memory network to obtain the deep semantic information of the image-text pair. These multimodal methods have some evident limitations. First, these multimodal datasets are hard to collect, especially for the audio modality. Second, how to fuse the hidden information between mutimodalities is still a big challenge. Hence, we perform our method on a single image modal.

Summarily, the above methods have achieved evident progress in the field of image sentiment analysis. Early works [16–23] focused on the analysis of local regions in images rather than the whole image. Most of them require massive high-quality data to train a robust model. Nevertheless, the corresponding performance is unsatisfactory. Although some domain adaption methods [25–27] have been proposed to alleviate this problem, the scarcity of high-quality images with definite sentiment semantics becomes more and more evident. Hence, some state-of-the-art data augmentation methods, including DADA [31] and AutoAugment [32], have been proposed to generate more high-quality samples. However, they need massive computing resources. On the other hand, few studies have attempted to exploit the implicit correlation among the heterogeneous features from a homologous neural network, such as the state-of-the-art SENet modules [33]. Thus, unlike recent works, our model is an organic combination of the feature early fusion and ASR. To a certain extent, our model can alleviate the issue of the lack of high-quality images. Meanwhile, our model tries to fully mine the cluster correlation among the heterogeneous SENet features. Hence, our method does not generate any new samples and is easier to reproduce.

3. Method

An overview of the proposed model is shown in Figure 2. First, we perform the ASR strategy to adaptively and gradually augment the original datasets. The existing high-quality images in a dataset (D_5) are used to train classifiers, and then, the semantically ambiguous images are predicted by the well-trained classifier. If the predicted results are consistent with the corresponding annotations, the images are chosen to augment the original high-quality dataset. For the high-quality dataset D_{fine} generated by the ASR strategy, the heterogeneous SE-ResNet features are extracted for the subsequent correlation mining. Because these features are derived from a homologous neural network, we attempt to mine the implicit but effective cross-modal semantics (CMS) among the heterogeneous SENet features by performing cluster correlation mining (CCM). Subsequently, several classifiers are trained to predict the sentiment of the images. Finally, an ensemble learning strategy is proposed to further

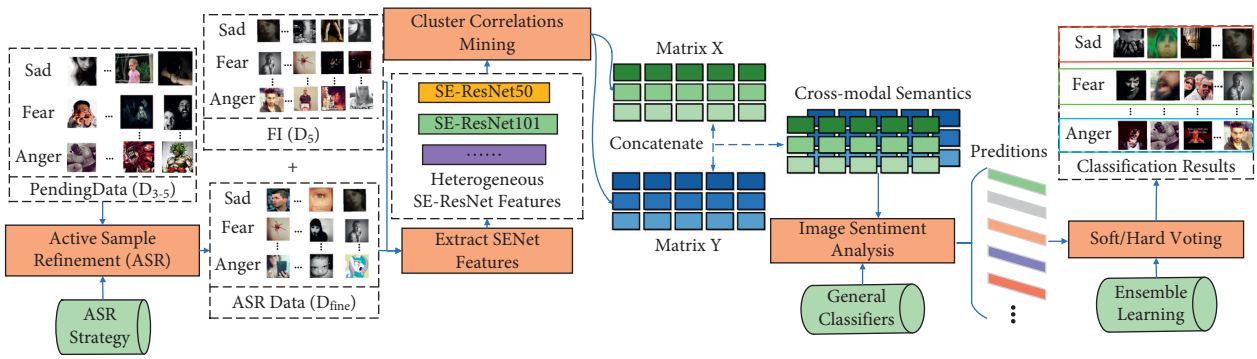


FIGURE 2: The pipeline of our model. We use the fine-grained Flickr and Instagram (FI) dataset to illustrate our model. High-quality images are obtained through ASR. Then, we extract a set of heterogeneous SENet features and perform CCM among these features to obtain more discriminative CMS. A simple voting strategy is designed to boost the final performance.

improve the final performance. Through ensemble learning, the predictions of different classifications are integrated together to get more objective results and further improve the final classification performance. Hence, in this section, we first present our ASR strategy. Next, we show how to perform CCM among the heterogeneous SENet features. Finally, the training procedure is described.

3.1. Active Sample Refinement. As described earlier, image sentiment analysis requires massive high-quality images with definite sentiment semantics. Based on the concept of active learning [34], we propose the ASR strategy to augment the original dataset. Active learning is a type of semi-supervised learning that is usually used to annotate unlabeled samples and improve the performance of classifiers [35]. In this study, we adopt active learning to retrieve high-quality images from the original dataset. Unlike other GAN-based methods [26, 27], we will not generate any new images and only make full use of the original benchmark datasets. Hence, our method can mitigate the absence of images with definite sentiment semantics and lay a firm data foundation for model training. The proposed ASR strategy is divided into two progressive steps: coarse-grained sample refinement and fine-grained sample refinement. The technology flowchat of the ASR strategy is shown in Figure 3.

As shown in Figure 3, the ASR strategy consists of two stages: coarse-grained sample refinement and fine-grained sample refinement. The left part illustrates the coarse-grained refinement procedure, while the right part illustrates the fine-grained refinement procedure. First, the classifiers trained by high-quality images are utilized to predict the semantically ambiguous images, and the coarse-grained refinement is completed by comparing the predicted label with the ground truth. So we obtain the D_{coarse} dataset. Then, the active learning idea is used for the subsequent fine-grained sample refinement. In this procedure, we use one classifier to predict the data in D_{coarse} each time, and the refined images will be directly added into D_5. We loop the previous step to realize the gradual screening of images based on active learning (please see the blue arrow). Finally, the final D_{fine} dataset is obtained by intersection or union of the images generated by all the classifiers.

Step 1: **Coarse-grained sample refinement.** First, we obtain a data subset called $D_{3\text{-}5}$ from the original dataset D_3 (D_3 represents that at least three Amazon Mechanical Turk (AMT) workers gave the same sentiment label to an image. Similarly, D_5 represents that all five AMT workers gave the same sentiment label to an image. We remove the D_5 data from the D_3 dataset and obtain $D_{3\text{-}5}$). Hence, $D_{3\text{-}5}$ can be regarded as the pending data for ASR. Our ultimate purpose is to achieve high-quality images from $D_{3\text{-}5}$ and use these samples to augment the original D_5 dataset. Therefore, we consider D_5 as the training set (because any sample in D_5 is high-quality) and $D_{3\text{-}5}$ as the testing set. Second, to promote real-time efficiency, similar to reference [14], we utilize the cross-modal information between SIFT and VGG19 to efficiently characterize the $D_{3\text{-}5}$ and D_5 datasets. Third, to avoid selective bias, we perform image sentiment analysis using a group of heterogeneous classifiers (nine classifiers for Twitter I and five classifiers for FI). Finally, if all the predictions are consistent with the real label, the corresponding testing sample in $D_{3\text{-}5}$ will be merged into D_{coarse}. Therefore, D_{coarse} is a data subset obtained from $D_{3\text{-}5}$. It contains lots of images with relatively definite sentiment semantics. This builds a solid foundation for the subsequent fine-grained sample refinement.

Step 2: **Fine-grained sample refinement.** We attempt to refine further to obtain high-quality images with definite sentiment semantics based on the D_{coarse} subset. Therefore, we consider D_5 as the training set and D_{coarse} as the testing set, and the active learning strategy is seamlessly incorporated into the fine-grained sample refinement procedure. The fine-grained sample refinement process requires multiple iterations to gradually absorb effective knowledge into our model. In iterations, each classifier predicts every sample in D_{coarse}. Then, we use the ranked batch mode queries [36] to rank the sample. The rank score function is expressed as

$$\text{score} = \partial\left(1 - \Phi\left(x, D_5\right)\right) + (1 - \partial)U(x), \quad \partial = \frac{D_5}{\left|D_{3-5}\right| + \left|D_5\right|}, \tag{1}$$

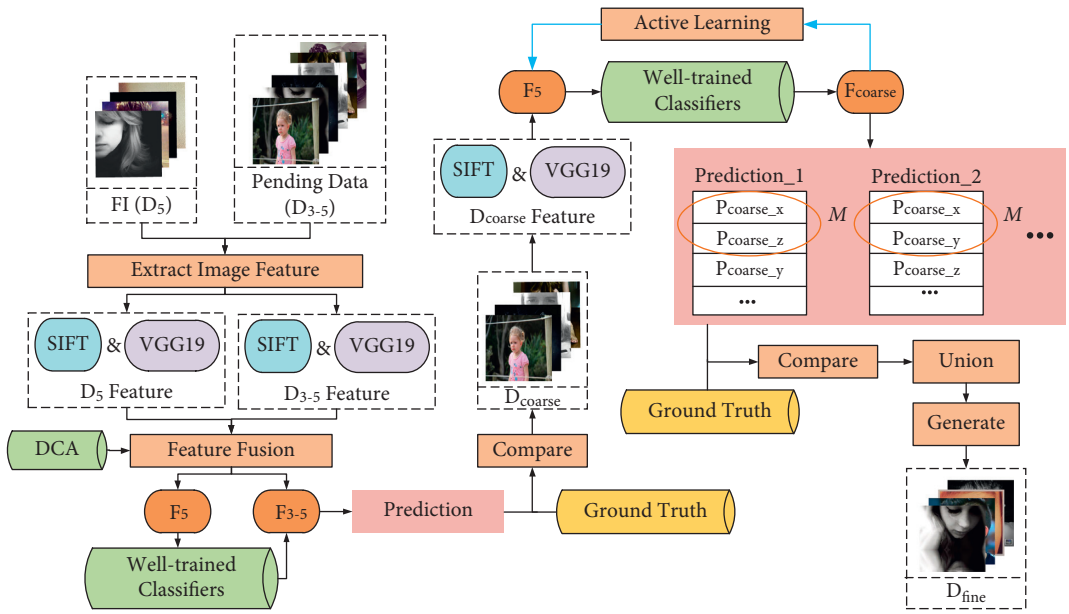


Figure 3: The technology flowchart of the ASR strategy.

where x denotes an image from D_{coarse}, $U(x)$ is the uncertainty of the predictions for x, and Φ is the Euclidean function used to calculate the distance between two images. High-scoring samples are placed at the top of a list. We select the top three samples and compare the predicted label of each sample with its real label. The sample will be chosen if it is completely consistent. Thus, we augment the original training dataset by adding this new sample. Otherwise, the sample will be removed from the list. After multiple iterations, much more valuable knowledge from these high-quality samples is adaptively and gradually merged into the ASR algorithm, making it more efficient, robust, and capable of autonomous learning. To prevent selective bias, we employ a group of heterogeneous classifiers to complete this fine-grained sample refinement procedure (two classifiers for Twitter I and four classifiers for FI (please refer to Figure 5). Hence, each classifier outputs a set of candidate high-quality images from its own perspective, and then, we perform union processing on these sets to build D_{fine}, which is the final augmentation data of the original D_5 dataset. Each sample in D_{fine} has definite sentiment semantics. Hence, the augmented data builds a firm foundation for model training. Unlike those GAN-based methods, we only make full use of the existing benchmark datasets. We do not generate any new samples. So, it is easier to reproduce the ASR strategy.

3.2. Cluster Correlation Mining. Robust but effective features help improve the final accuracy. However, a single image feature is insufficient for image sentiment analysis. Hence, we first extract a set of heterogeneous yet effective features from the SENet modules, namely, SE-ResNet50, SE-ResNet101, SE-ResNet152, SE-ResNeXT-50, and SE-ResNeXT101 (why do we make this choice? please refer to Table 5). Then, we mine the implicit cluster correlation among these SENet features because they all derive from a homologous neural network. We believe that some valuable information hides in these heterogeneous features. Meanwhile, we try to mine the inherent dependencies between different classes. Finally, we use this clustering correlation to generate CMS for characterizing the core sentimental semantics.

Let x_i^c denote the i-th sample in the c-th class, and y_i^c denotes the i-th label in the c-th class. X and Y denote two heterogeneous SENet features for training, respectively, whereas X^* and Y^* represent the corresponding features for testing, respectively. The cluster correlation between X and Y is calculated as follows:

$$\rho = \max_{w,v} \frac{w' V_{XY} v}{\sqrt{w' V_{XX} w} \sqrt{v' V_{YY} v}}, \tag{2}$$

where w and v are the two projection matrices of X and Y, respectively. The covariance matrices, namely, V_{xy}, V_{xx}, and V_{yy}, are defined as follows:

$$\begin{aligned} V_{XY} &= \frac{1}{N} \sum_{c=1}^{C} \sum_{i=1}^{|X_c|} \sum_{j=1}^{|Y_c|} x_i^c y_j^{c\prime}, \\ V_{XX} &= \frac{1}{N} \sum_{c=1}^{C} \sum_{i=1}^{|X_c|} |Y_c| x_i^c x_i^{c\prime}, \\ V_{YY} &= \frac{1}{N} \sum_{c=1}^{C} \sum_{j=1}^{|Y_c|} |X_c| y_j^c y_j^{c\prime}, \end{aligned} \tag{3}$$

where $N = \sum_{c=1}^{C} |X_c||Y_c|$. We aim to maximize the correlation between the projections of X and Y on w and v. After we obtain w and v, we transform X, X^*, Y, and Y^* into X_α, X_α^*, Y_α, and Y_α^*, respectively, as follows:

$$\begin{cases} X_\alpha = w'X, & X_\alpha^* = w'X^* \\ Y_\alpha = v'Y, & Y_\alpha^* = v'Y^* \end{cases}. \quad (4)$$

Using the formula (4), we construct U and V, respectively, as follows:

$$U = \begin{pmatrix} X_\alpha \\ X_\alpha^* \end{pmatrix}, V = \begin{pmatrix} Y_\alpha \\ Y_\alpha^* \end{pmatrix}. \quad (5)$$

Finally, we concatenate U with V and generate CMS among heterogeneous SENet features as follows. The implicit and valuable cluster correlation among the SENet features is obtained. The inherent dependencies between different classes are mined in turn.

$$F = \begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} X_\alpha & Y_\alpha \\ X_\alpha^* & Y_\alpha^* \end{pmatrix}. \quad (6)$$

3.3. Training Procedure. Based on the augmented data "$D_5 + D_{fine}$," we used four general classifiers, namely, logistic regression (LR), support-vector machine (SVM), adaptive boosting (AdaBoost), and extreme gradient boosting (XGBoost), to train the proposed model. Then, ensemble learning (hard voting or soft voting) is performed on the top three classification results. This helps further improve the final classification performance.

4. Experiments and Discussion

In this section, we first compare our model against state-of-the-art methods on Twitter I [37] and FI [38]. The Twitter I dataset is collected from social websites and labeled with two sentiment polarity categories (positive and negative) by AMT participants. Like D_5 or D_3 (please refer to Section 3.1), D_4 indicates that at least four AMT workers gave the same sentiment label to the same image. The Twitter I dataset consists of 1,269 images. It is a coarse-grained dataset. Unlike Twitter I, the FI dataset is collected by querying eight sentiment categories, namely, anger, amusement, awe, contentment, disgust, excitement, fear, and sadness, as keywords from social websites. It consists of 23,308 images, each with at least three agreements. The FI is a fine-grained dataset. Making accurate sentiment predictions on this fine-grained dataset is still a big challenge. In our experiments, each benchmark dataset is randomly divided into 70% for training and 30% for testing. The details of the two datasets are provided in Table 1. We also exhibit the number of refined images in this table.

Second, we evaluate the effectiveness and robustness of the proposed ASR and CCM strategies, respectively. We want to know whether the ASR strategy is a general but effective data augmentation method and the CMSs generated by the CCM strategy are effective for image sentiment classification.

Third, we compare the computational efficiency against some baseline models. Finally, we show the corresponding details of parameter tuning.

Table 1: The Twitter I and FI datasets with refined data.

Dataset	D_5	D_4	D_3	D_{coarse}	D_{fine}
Twitter I	882	1116	1269	122	74
FI	5238	12644	21508	1214	610

In this study, we use a relatively high-performance laptop with the following hardware configuration: Intel (R) Core (TM) i7-8550, MX150, and 8 GB RAM, to implement our experiments.

4.1. Comparisons with State-of-the-Arts. In this section, we compare our model with the following state-of-the-art methods:

(1) Classical fine-tuned deep learning models, namely, VGG16 [39], AlexNet [40], CAM-Res101 [41], and Res101 [42].

(2) State-of-the-art feature fusion methods, namely, DCA [15], canonical correlation analysis (CCA) [43], gradKCCA [44], GS-SVM [13], GS-LR [13], and GS-XGB [13].

(3) Recently proposed baseline methods for image sentiment analysis, namely, ResNet-MldrNet [24], SPN [18], WSCNet [17], WILDCAT [19], ME^2M (M) [27], SR-w-DCA [27], CycleGAN [25], CycleEmotionGAN [26], MSGAN [27], Sun's model [20], Rao's model [21], and SmileyNet [24].

We also created two variants of our model by using different correlation mining methods, namely, DCA [15] and CCA [43]. We wanted to demonstrate the generalization of the proposed idea. So, we used "$Ours_{CCA}$" and "$Ours_{DCA}$" to represent the two variants. Meanwhile, in order to demonstrate that D_{fine} contains more high-quality image samples than D_{coarse}, we also perform our model on D_{coarse} and compare the performance on D_{coarse} with that on D_{fine}. Hence, "$Ours_{CCM}$ (coarse)" means that we perform our model on D_{coarse}. The corresponding experimental results are shown in Table 2. Herein, "D_5" represents that we only used the D_5 data to evaluate the classification model. "$D_5 + D_{fine}$" means that we executed each model on the augmented dataset. Others only used D_3 to evaluate each model. Hence, we trained all the methods under the same setting. Moreover, in the "Type" column, "D" means **d**eep learning-based method. "F" means **f**eature fusion method. "B" means state-of-the-art **b**aseline. "O" means **o**ur proposed model.

First, our model outperforms the classical fine-tuned deep learning models. For example, compared to the most powerful VGG16 and Res101 models, our model improves 3.34% on the Twitter I dataset and 2.21% on the FI dataset, respectively. As we know, these classical deep learning models use very complex network structures, which make them more prone to overfitting when high-quality images are very scarce. This phenomenon is more evident on the FI dataset because this fine-grained dataset needs more high-quality samples to train a robust model that can discriminate

Table 2: Performance comparisons. The best value of each dataset is shown as **89.90**. Unit: %.

Dataset	Method	Type	Accuracy	Method	Type	Accuracy
Twitter I	VGG16	D	76.75	WILDCAT (D_5+D_{fine})	B	73.85
	VGG16 (D_5)	D	86.56	WSCNet	B	84.25
	VGG16 (D_5+D_{fine})	D	82.69	WSCNet (D_5)	B	87.62
	AlexNet	D	73.24	WSCNet (D_5+D_{fine})	B	89.40
	CAM-Res101	D	82.67	MSGAN	B	63.58
	GS-SVM (D_5)	F	88.72	CycleGAN	B	61.59
	GS-LR (D_5)	F	87.22	CycleEmotionGAN	B	62.38
	GS-XGB (D_5)	F	86.47	SmileyNet	B	89.16
	GS-SVM (D_5+D_{fine})	F	82.04	Sun's model	B	88.91
	GS-LR (D_5+D_{fine})	F	84.51	ME^2M (M)	B	87.15
	GS-XGB (D_5+D_{fine})	F	83.10	SPN	B	81.67
	CCA (D_5)	F	80.08	$Ours_{CCA}$	O	85.71
	gradKCCA (D_5)	F	77.07	$Ours_{CCA}$ (coarse)	O	77.81
	DCA (D_5)	F	87.59	$Ours_{DCA}$	O	82.23
	WILDCAT	B	79.53	$Ours_{CCM}$ (coarse)	O	89.07
	WILDCAT (D_5)	B	71.43	$Ours_{CCM}$	O	**89.90**
FI	Res101	D	66.16	WILDCAT (D_5+D_{fine})	B	72.23
	Res101 (D_5)	D	75.61	WSCNet	B	70.07
	Res101 (D_5+D_{fine})	D	78.11	WSCNet (D_5)	B	72.22
	AlexNet	D	58.13	WSCNet (D_5+D_{fine})	B	74.72
	CAM-Res101	D	68.54	MSGAN	B	70.63
	GS-SVM (D_5)	F	73.52	CycleGAN	B	63.87
	GS-LR (D_5)	F	73.59	CycleEmotionGAN	B	67.78
	GS-XGB (D_5)	F	72.89	MldrNet	B	67.75
	GS-SVM (D_5+D_{fine})	F	74.79	Rao's model	B	75.46
	GS-LR (D_5+D_{fine})	F	75.30	SR-w-DCA	B	75.72
	GS-XGB (D_5+D_{fine})	F	75.98	SPN	B	66.57
	CCA (D_5)	F	50.29	$Ours_{CCA}$	O	61.89
	gradKCCA (D_5)	F	61.84	$Ours_{DCA}$	O	75.93
	DCA (D_5)	F	73.71	$Ours_{DCA}$ (coarse)	O	73.86
	WILDCAT	B	67.03	$Ours_{CCM}$ (coarse)	O	79.16
	WILDCAT (D_5)	B	70.09	$Ours_{CCM}$	O	**80.32**

more sentiment categories. Contrarily, our model handles the above problem well. It uses the ASR strategy to obtain sufficient high-quality images (please refer to Table 1), which can augment the original benchmark datasets and build a firm data foundation for the subsequent classification. Meanwhile, the new CMS fits the number of training samples well, which helps reduce the risk of overfitting to a certain degree. More importantly, owing to a simpler structure (please refer to Figure 3), our model is relatively easier to reproduce. Moreover, we found another interesting phenomenon. The ResNet 101 model obtains about 2.50% performance improvement on the FI dataset when the augmented "$D_5 + D_{fine}$" data are used. Contrarily, the VGG 16 model gets some performance decline when the augmented "$D_5 + D_{fine}$" data are used. We guess the proposed ASR strategy enriches the original benchmark dataset from the fine-grained sentiment perspective. This offers sufficient training samples for the fine-tuned deep learning models. However, for the coarse-grained Twitter I dataset, the refined samples tend to be consistent with the existing samples in sentimental semantics, which has a little positive influence on performance improvement. Hence, more robust performance improvements can be observed on the fine-grained FI dataset.

Second, our model beats the state-of-the-art feature fusion methods. For example, compared to "GS-XGB ($D_5 + D_{fine}$)," a 4.34% performance improvement is observed on the FI dataset because the ASR strategy can enrich the original dataset from the fine-grained sentiment perspective. Additionally, our model improves 6.61% compared to the DCA model on the FI dataset. This is another evidence of the effectiveness of our ASR strategy. We also found that the performance of CCA is lower than that of DCA on each dataset. Evidently, the corresponding sample distribution of DCA is better than that of CCA (please see Figure 7). The proposed ASR strategy plays an important role in our model. It offers us plentiful high-quality image samples (Table 1). Meanwhile, our model performs CCM, which is a relatively simple feature early-fusion strategy only considering interclass dependence. Contrarily, the GS-based methods require numerous features and need to perform a very complex feature mid-fusion procedure. Thus, our model is easier to reproduce, demonstrating its practicality. Similarly, obvious performance improvements can be observed on the FI dataset when the "$D_5 + D_{fine}$" data are used. Our ASR strategy offers more valuable samples for the fine-grained FI dataset. So, more robust performance improvements can be observed on this dataset.

Last, our model exhibits superior performance over the state-of-the-art baselines. The ME^2M (M) and SR-w-DCA models perform static sample refinement. Unlike them, our model performs ASR to adaptively and gradually augment the original datasets, which can bring us more valuable samples. The SPN and WILDCAT methods focus on obtaining the local regions with very strong sentiment tendency, which mostly relies on additional manual annotations. The corresponding performance is unsatisfactory. However, the proposed model only uses the whole image and does not need any additional annotations. This helps alleviate the sentiment annotation burden. The SmileyNet leverages additional private emoji information, whereas our model only uses the publicly available images ($D_{3\text{-}5}$). Compared with the domain adaptation-based methods (MSGAN and CycleGAN), our model exhibits outstanding performance. These domain adaptation-based approaches should face a significant gap between the source and target domains. Hence, it is hard to train them. Contrarily, owing to a simple but clear structure, it is easier to train the proposed model. Moreover, we observe that the CCM strategy is effective and robust, which beats other correlation analysis methods by a large margin. More importantly, larger performance improvements can be observed on the fine-grained FI dataset, demonstrating that our model is highly practical. In our daily life, we usually present more fine-grained emotions rather than "like" or "dislike." Finally, to our surprise, we found that the refined samples can improve the final performance of the state-of-the-art baselines (please compare "WSCNet ($D_5 + D_{fine}$)" and "WSCNet (D_5)"). This can firmly demonstrate the generalization of the ASR strategy. Therefore, this strategy is an effective and robust data augmentation method. We will make further analysis in the following section.

In summary, the proposed model is effective and robust for image sentiment analysis. On the one hand, the ASR strategy can alleviate the data scarcity problem and augment the original benchmark datasets. On the other hand, we created a set of more discriminant but robust features called CMS by mining the cluster correlation among the heterogeneous SENet features. Meanwhile, our two variants ($Ours_{CCA}$ and $Ours_{DCA}$) also get very competitive performance, demonstrating the generalization ability of the proposed idea. Moreover, Ours $_{CCM}$ beats $Ours_{CCM}$ (coarse) in a large performance margin, which proves that D_{fine} really contains more high-quality samples than D_{coarse} and this fine-grained data are effective for image sentiment analysis. Based on the ASR and CCM strategies, our model beats the state-of-the-art baseline models.

4.2. Effectiveness Evaluation of ASR and CCM. In this section, we simultaneously evaluate the ASR and CCM strategies from another view. To obtain statistical interpretations, five SENet features and ten CMSs are chosen to complete the corresponding experiments on the D_5 and "$D_5 + D_{fine}$" datasets, respectively. Then, the mean accuracies are presented in Figure 4.

Figure 4 shows that any CMS outperforms the corresponding SENet feature. For example, a 2.59% performance improvement of the AdaBoost classifier can be observed on the FI dataset. This phenomenon is more evident in the boosting-based classifiers, such as AdaBoost and XGBoost. A similar phenomenon can be found on each dataset. All these demonstrate the effectiveness and robustness of the CCM strategy. The implicit valuable discriminative information among the heterogeneous SENet layers is fully mined and used to promote the final performance. Compared with D_5, the augmented dataset ("$D_5 + D_{fine}$") performs better. For example, a 1.58% improvement of the AdaBoost classifier can be observed on the FI dataset. All these demonstrate the effectiveness and robustness of the ASR strategy. Sufficient high-quality images with definite sentiment semantics "generated" by the ASR strategy build a firm data foundation for model training. Surprisingly, the single ASR strategy combined with SVM (or LR) outperforms CMS without ASR on the FI dataset. More high-quality data play a positive role in image sentiment analysis. More importantly, our model seamlessly combines ASR and CCM to achieve the best performance.

Herein, we make an ablation analysis to evaluate the real contribution of each component of our model. We calculate the average improvement of four classifiers on both datasets. Compared to the original SENet feature, the corresponding performance improvement of "ORI (ASR)," "CMS," and "CMS (ASR)" is 0.24%, 1.20%, and 1.34% on Twitter I whereas 1.41%, 1.52%, and 2.57% on FI. This is a very interesting phenomenon. Meanwhile, to validate the real contribution of ASR and CMS, we calculate the average performance difference between "ORI" and "ORI (ASR)" on the two datasets. A similar difference between "CMS" and "ORI" can also be obtained. The average performance difference is 0.83% and 1.36%, respectively. This indicates that CMS contributes more than the proposed ASR strategy. We can say the feature is more important than the refined samples. Certainly, the combination of the proposed CMS and ASR strategies plays a positive role in image sentiment analysis.

To further demonstrate the effectiveness and versatility of the ASR strategy, we performed four traditional data augmentations (TDAs), including rotate, crop, saturate, and scale operations, and the ASR strategy on the D_5 data subset of Twitter I. The corresponding results are provided in Table 3. We selected the fine-tuned VGG16 [39], fine-tuned Res101 [42], WILDCAT [19], and WSCNet [17] to complete comparisons. In Table 3, "NDA" represents "no data augmentation." "TDA + ASR" means that we performed the ASR strategy after "TDA." Our experimental details are as follows: first, we chose a basic data augmentation operation among all the four operations. Then, we combined this basic operation with the other three random augmentation operations to generate new images. Hence, the "TDA" approach generated 3,528 images based on the D_5 data. Furthermore, we implemented the ASR strategy on these images. Thus, 2,121 refined image samples are obtained. Finally, we used the refined image samples to train each model introduced above and made performance

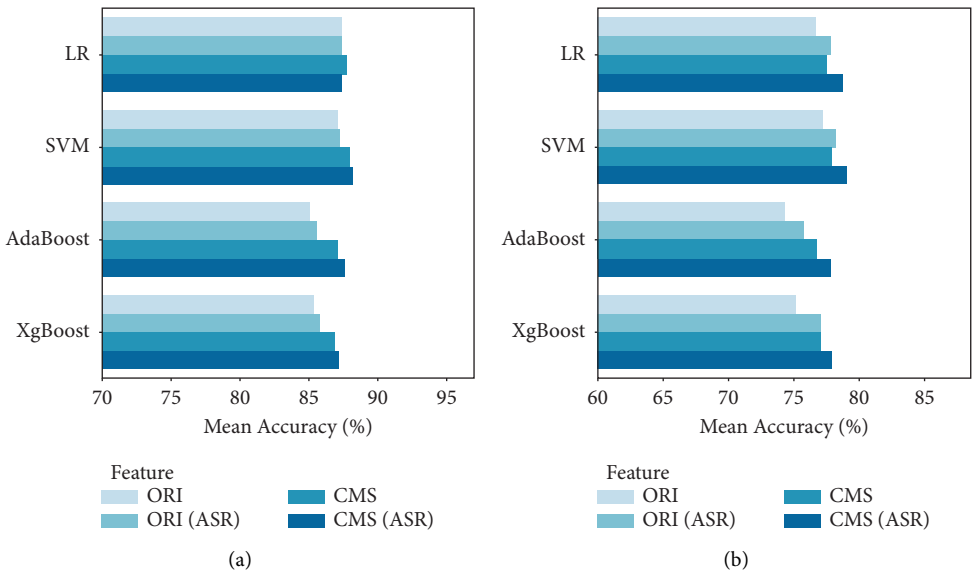


Figure 4: Effectiveness evaluations of the ASR and CCM strategies. "ORI" represents the original SENet feature. CMS represents the proposed cross-modal semantics. "ORI (ASR)" means that we use the SENet feature on "$D_5 + D_{fine}$." "CMS (ASR)" means that we use CMS on "$D_5 + D_{fine}$." (a) Twitter I (b) FI.

Table 3: Effectiveness and versatility evaluation of ASR. The best value is shown as **92.57**. Unit: %.

Methods	Accuracy	Methods	Accuracy	Methods	Accuracy
$VGG16_{NDA}$	86.56	$Res101_{ASR}$	86.93	$WSCNet_{NDA}$	87.62
$VGG16_{TDA}$	88.75	$Res101_{TDA+ASR}$	90.18	$WSCNet_{TDA}$	88.23
$VGG16_{ASR}$	82.69	$WILDCAT_{NDA}$	71.43	$WSCNet_{ASR}$	89.40
$VGG16_{TDA+ASR}$	89.46	$WILDCAT_{TDA}$	85.00	$WSCNet_{TDA+ASR}$	89.64
$Res101_{NDA}$	87.96	$WILDCAT_{ASR}$	73.85	$Ours_{FAA+ASR}$	88.30
$Res101_{TDA}$	88.37	$WILDCAT_{TDA+ASR}$	**92.57**	$Ours_{CCM}$	89.90

comparisons. We also combined the proposed ASR strategy with the popular adaptive data augmentation method called Fast AutoAugment (FAA). We implemented ASR after the application of the FAA. The corresponding model is Ours $_{FAA+ASR}$.

As shown in Table 3, apparent performance improvements can be observed when using the "TDA" method. For example, a 2.2% performance improvement is obtained on the VGG16 model. This demonstrates the effectiveness of the "TDA" method. Hence, most CV tasks employ the "TDA" method to augment their own datasets. However, we got further performance improvements by combining the proposed ASR strategy with "TDA." For example, a 1.81% performance improvement is obtained on the Res101 model by using "TDA + ASR." We found that owing to more scarce data, only using the ASR strategy on the fine-tuned deep learning models brings performance degradation (similar to that of Table 2). Surprisingly, unlike the fine-tuned deep learning models, more robust performance occurs on the state-of-the-art WILDCAT and WSCNet models. Hence, our sample refinement strategy can be regarded as a useful supplement to the current "TDA" method, which can be seamlessly incorporated into the state-of-the-art baselines to obtain better performance (please see $WILDCAT_{TDA+ASR}$). However, when we combine the ASR strategy with FAA, there is a certain performance degradation. The underlying reason is that there are still a small amount of samples after the combination of FAA and ASR.

In conclusion, the ASR strategy is effective and versatile. It can be combined with the "TDA" method to obtain better performance. Hence, it is a useful supplement to current data augmentation methods. More importantly, it can be incorporated into the state-of-the-art baselines, such as WILDCAT and WSCNet, to achieve better performance.

4.3. Real-Time Efficiency Comparison. Table 4 shows the average time that our model needs to test one image. The testing procedure can be roughly divided into three stages: feature extraction, CCM, and final classification. Since the feature selection procedure has been completed in the training stage, we only need to extract two features here. Meanwhile, we save our classifier models at the training stage. Therefore, we only require to use them for the final classification. Because it is too fast, the time cost of ensemble learning can be negligible here. The entire time required to

Table 4: Real-time efficiency comparison. The best value is shown as 0.170. Unit: *s*.

Method	Test time	Method	Test time
VGG16	0.053	ME^2M (M)	2.170
Res101	0.098	Ours	**0.170**

test one image of our model is 0.17 s, demonstrating that our model is lightweight and easy to deploy.

4.4. Parameters Tuning

4.4.1. Feature Selection. In this section, we need to choose effective but robust image features for CCM. We selected SIFT [45], GIST [46], LBP [47], DenseNet, VGG [39], ResNet, SENet [33], and SENetXT as our candidate features. We implemented principal component analysis on each traditional feature. Thus, each traditional feature was reduced to 500. For the deep learning features, we chose the bottleneck features of the corresponding CNN model. For example, we used the 16th and 19th layers of the VGG model in our experiments. Owing to a large amount of valuable distributed information, we did not implement feature reduction on these deep learning features. Detailed experimental results are shown in Table 5. "Dim" means dimension.

As shown in Table 5, the five SENet features outperform any other feature. In particular, the two SENetXT features obtain the best trade-off between accuracy and efficiency. As we know, the SENet features pay more attention to those local regions with strong sentiment tendencies, including Ferris wheel, smiling face, shabby house, and so on. These sentimental signals can accurately depict the visual content and create a powerful foundation for CCM. In contrast, SIFT mainly describes the key gradient variations in images. However, the gradient variations between background and foreground sometimes are not obvious. GIST and LBP mainly describe the global and local textures in images. Texture only represents a little part of emotion, which is weaker than other features. Summarily, SENet is the best choice for our model. Meanwhile, compared with other features, SENet requires fewer resources and is efficient for extraction. Hence, we chose the five SENet features to complete CCM.

4.4.2. Classifier Tuning for Fine-Grained Sample Refinement. As described above, to avoid selective bias, we employed several heterogeneous classifiers to complete ASR. Therefore, classifier type and classifier number should be appropriately tuned to obtain the best performance. Here, we exhibit the corresponding tuning procedure for fine-grained sample refinement. For the Twitter I dataset, we chose LR and k-nearest neighbor (KNN) algorithms to perform fine-grained sample refinement. For the FI dataset, we used CatBoost, AdaBoost, random forest (RF), and LR to complete the same procedure. The refined samples of each classifier were used to perform the union operation for augmenting the existing benchmark datasets. The corresponding experimental results are shown in Figure 5. As analyzed above, the DCA feature has a lower dimension than the proposed CCM. As shown in Table 2, the DCA feature obtains very competitive accuracy. Hence, to save resources, we used the DCA model to fuse the original features, and then, we performed classification using CatBoost.

As shown in Figure 5, using multiple heterogeneous but complementary classifiers can prevent selective bias to a certain degree. More detailed, using multiple classifiers to implement ASR obtains a 1.38% performance improvement compared to using LR on the Twitter I dataset. For the FI dataset, compared with a single classifier, using multiple classifiers obtains the corresponding performance improvements of 1.99%, 1.37%, and 0.64%, respectively. In Figure 5(a), the best accuracy on the testing dataset is obtained when all the classifiers are used. Owing to the lack of (or imbalance of image samples), slight overfitting is observed on the Twitter I dataset. Therefore, training relatively small datasets is still challenging. In the future, we plan to introduce the well-known zero-shot learning method [48] to address this problem. In Figure 5(b), our model performs better when all the classifiers are used. The complementarity among multiple heterogeneous classifiers is fully utilized to boost the final performance. Summarily, we achieved the optimal classifier combination based on the corresponding accuracy and fitting status of the trained model.

4.4.3. Parameters Tuning of Ensemble Learning. Based on the classification results of different classifiers, we employed the ensemble learning strategy to further improve the final performance. Different ensemble learning strategies (hard voting or soft voting) combined with different classifier numbers result in different results. Hence, we should carefully tune the corresponding ensemble learning strategy combined with the top two (or three) classifiers to obtain the best performance. To obtain statistical interpretations, we used ten CMSs with different ensemble learning strategies to complete the tuning procedure. Figure 6 exhibits the corresponding experimental results.

As shown in Figure 6, the top three classifiers combined with the hard voting strategy perform best on any benchmark dataset. The top three classifiers play a decisive role in the final image sentiment prediction. The implicit complementarity among multiple heterogeneous classifiers is fully mined by the ensemble learning strategy to improve the final performance. In addition, the hard voting strategy is fairer than the weighted soft voting strategy. Hence, we used this setting to complete our experiments.

4.4.4. Visualization Results of Different Features. To intuitively demonstrate the effectiveness of the CCM strategy, in this section, we employ the well-known t-SNE [49] tool to visualize the sample distribution of each kind of CCM. Figure 7 exhibits the corresponding visualization results of different image features. Owing to space limitations, only the results of the Twitter I dataset are provided here. Similar experimental results can be observed on the FI dataset.

Table 5: Classification performance of each single image feature. Unit: %.

Datasets	Feature	Accuracy	Dim	Feature	Accuracy	Dim
Twitter I	SIFT	83.02	500	VGG16	82.26	4096
	GIST	69.43	500	VGG19	83.77	4096
	LBP	67.55	500	SENet101	87.80	2048
	DenseNet121	67.17	1024	SENet152	88.50	2048
	DenseNet161	65.66	2208	SENet50	84.32	2048
	ResNet50	70.19	2048	SENetXT101	87.11	2048
	ResNet152	69.06	2048	SENetXT50	89.20	2048
FI	SIFT	55.47	500	VGG16	71.18	4096
	GIST	50.51	500	VGG19	70.99	4096
	LBP	46.85	500	SENet101	78.32	2048
	DenseNet121	29.13	1024	SENet152	78.21	2048
	DenseNet161	42.29	2208	SENet50	77.24	2048
	ResNet50	57.44	2048	SENetXT101	79.46	2048
	ResNet152	58.72	2048	SENetXT50	77.98	2048

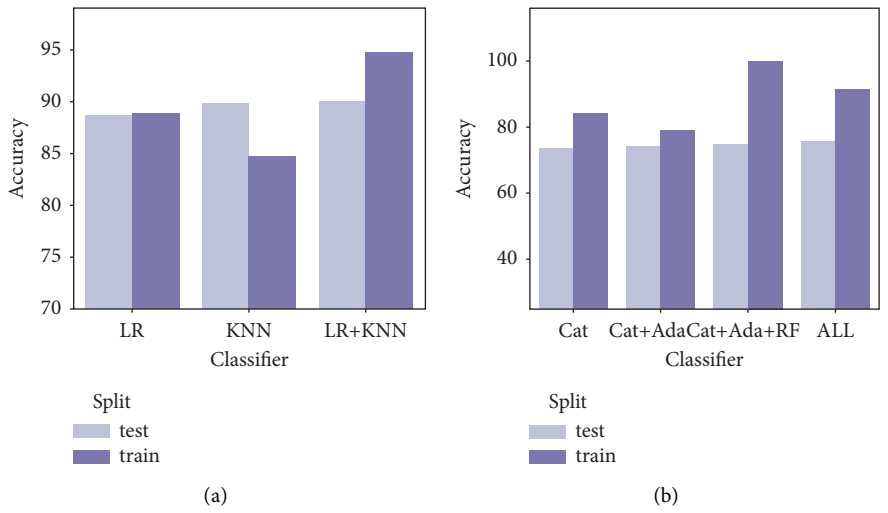


Figure 5: Classifier tuning for fine-grained sample refinement. (a) Twitter I. (b) FI.

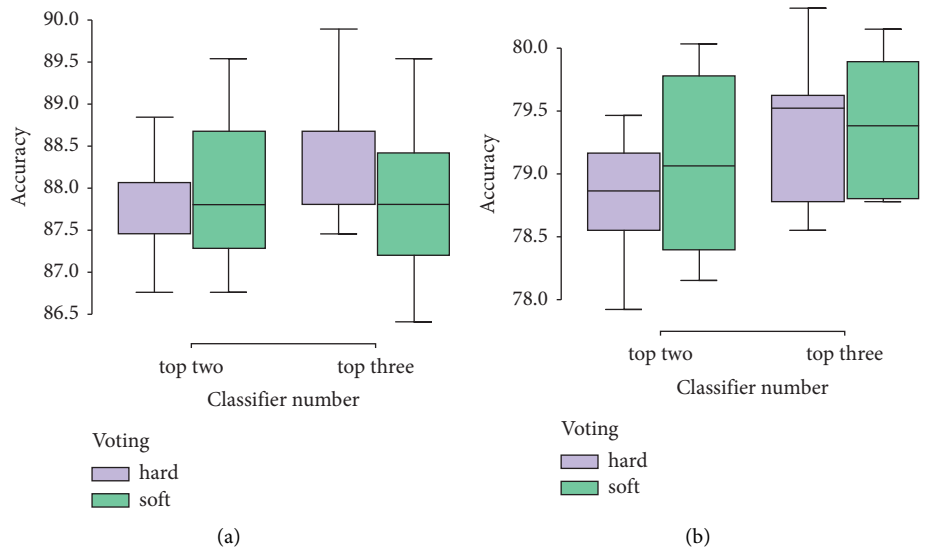


Figure 6: Parameters tuning of ensemble learning. (a) Twitter I. (b) FI.

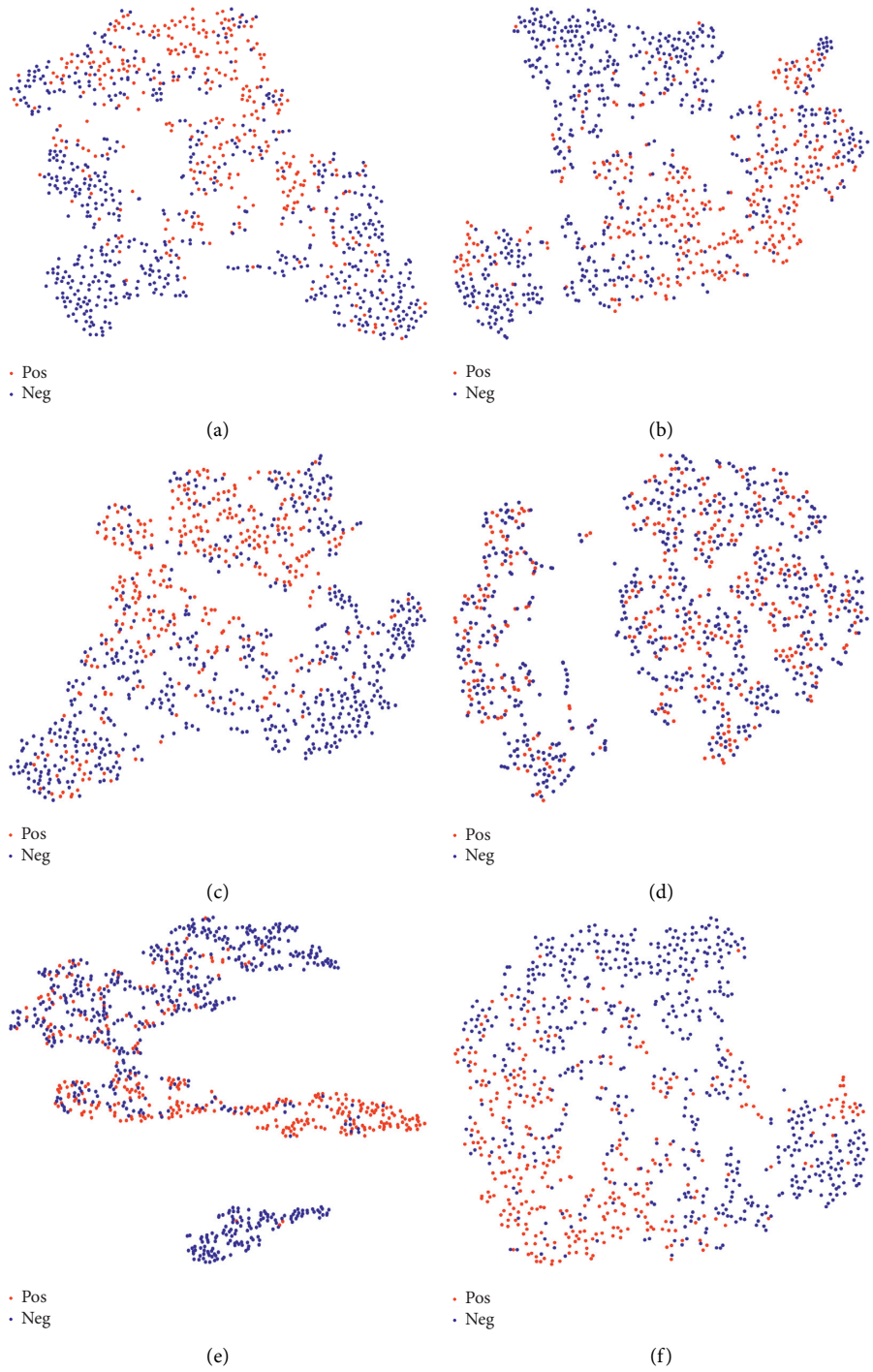
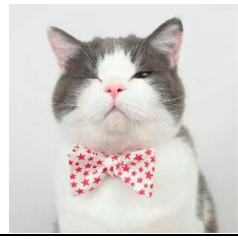


Figure 7: Visualization results of different features. (a) SENetXT50. (b) SENet 152. (c) SENet101. (d) CCA. (e) DCA. (f) CMS.

In Figures 7(a)~7(c), the corresponding sample distributions of the original SENet features show slight chaos. The aggregation degree of the image samples in the same class is low while that in different classes mixes up each other. Hence, it is difficult for a traditional classifier to fit the sample distribution (please refer to Table 5, the corresponding accuracy of the SENet152 feature is 88.50%. It builds a solid foundation for the subsequent CCM) using the original SENet feature. A relatively complex decision boundary of the traditional classifier is needed to address this problem. However, owing to the lack of high-quality images with definite sentiment semantics, overfitting may be prone to occur. So, a single image feature is insufficient for effective image sentiment analysis. As shown in Figure 7(d), the corresponding sample distribution of the CCA feature is worse than that of any SENet feature. This will bring a challenge to the traditional classifiers (in Table 2, the corresponding accuracy of the CCA feature is 80.08%).

Table 6: Evaluation on WeChat and microblog.

Web image				
True label	😃	😃	😕	😕
Our prediction	😃	😃	😕	😃

Similarly, heavy chaos can also be observed in Figure 7(e). But the result is better than Figure 7(d). As shown in Table 2, the corresponding accuracy of the DCA feature is 87.59%. Unlike other methods (CCA and DCA), as shown in Figure 7(f), the corresponding sample distribution of CMS is better than that of a single SENet feature. The proposed CCM strategy can increase the corresponding distance among different classes. Hence, a relatively simple decision boundary of the traditional classifier is needed to complete classification (in Table 3, the corresponding accuracy of our model is 89.90%). Summarily, the above visualization results intuitively validate the effectiveness of the proposed CCM strategy.

4.5. Empirical Analysis. We selected some open images from WeChat and microblog to prove the practicality of our model. We performed our model (the coarse-grained sentiment classification model) on these images. Experimental results are shown in Table 6. As shown in Table 6, we classified these images into two categories (positive and negative), and the smiling face usually represents positive sentiment tendency (😃), while the depressed face usually represents negative sentiment tendency (😕).

The first and second images have a riot of color, whereas the third and fourth images are generally in dark color. Moreover, the first image has a smiling face, but the third image has a sad face. Our model performs well in predicting the sentiment polarity of the first to the third images. This firmly demonstrates the practicality of the proposed model. Certainly, our model makes a prediction bias on the fourth image. Negative images are far fewer than positive images in the Twitter I dataset. Hence, we speculate that the prediction bias is mainly due to the lack of negative images. This further illustrates the importance of the data augmentation operation from another perspective. Summarily, besides the publicly available datasets, our model also performs well on the open images, demonstrating its high practicality.

5. Conclusions and Future Work

We presented a novel model for image sentiment analysis with the newly designed ASR and CCM strategies. We demonstrated its effectiveness and robustness on two benchmark datasets. Incorporating the ASR strategy into our model helps augment the original benchmark datasets, whereas CCM, which encourages robust and effective feature learning, has the potential to improve the final performance. More importantly, the ASR strategy is a useful supplement to the current data augmentation method. It can be seamlessly incorporated into the state-of-the-art baselines to achieve better performance.

However, there are still two limitations in our method. First, our method is based on the traditional machine learning idea. Although it has achieved very competitive performance, it is not end-to-end, which may limit its practicality. Second, the performance on the Twitter I dataset needs to be further improved in our future research.

In the future, we plan to use the state-of-the-art attention mechanism [16] or object detection systems [50–52] to retrieve the corresponding affective regions. Then, we will combine the whole image and the affective regions to obtain better performance and more powerful interpretability. Second, we plan to fuse the proposed ASR strategy with other popular data augmentation methods, such as DADA [31] and AutoAugment [32]. We hope to obtain more and more high-quality images with definite sentiment semantics. Finally, we will employ the well-known deep mutual learning [53] method to mine much more sentimental knowledge between heterogeneous neural networks for improving the final accuracy.

Data Availability

The data underlying the results presented in this paper are available in Refs. [37, 38].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the National Natural Science Foundation of China, grant nos. 62161011 and 61861016, the Natural Science Foundation of Jiangxi Provincial Department of Science and Technology, grant nos.

20202BABL212006, 20202BABL202044, and 20212BAB202006, the Key Research and Development Plan of Jiangxi Provincial Science and Technology Department, grant no. 20192BBE50071, the Humanity and Social Science Fund of Ministry of Education of China, grant no. 20YJAZH142, the Science and Technology Projects of Jiangxi Provincial Department of Education, grant nos. GJJ190323, GJJ200627, and GJJ200644, and the Humanity and Social Science Foundation of Jiangxi University, grant nos. TQ19101, TQ20108, and TQ21203.

References

[1] S. Y. Tseng, S. Narayanan, and P. Georgiou, "Multimodal embeddings from language models for emotion recognition in the wild," *IEEE Signal Processing Letters*, vol. 28, pp. 608–612, 2019.

[2] H. Tang, D. H. Ji, C. L. Li, and Q. J. Zhou, "Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification," in *Proceedings of the Fifty Eigth Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6578–6588, July 2020.

[3] M. Phan and P. Ogunbona, "Modelling context and syntactical features for aspect-based sentiment analysis," in *Proceedings of the Fifty Eigthth Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3211–3220, July 2020.

[4] N. Charlton, C. Singleton, and D. V. Greetham, "In the mood: the dynamics of collective sentiments on Twitter," *Royal Society Open Science*, vol. 3, no. 6, Article ID 160162, 2016.

[5] Y. Zhao, B. Qin, T. Liu, and D. Tang, "Social sentiment sensor: a visualization system for topic detection and topic sentiment analysis on microblog," *Multimedia Tools and Applications*, vol. 75, no. 15, pp. 8843–8860, 2016.

[6] J. F. Si, A. Mukherjee, B. Liu, Q. Li, H. Li, and X. Deng, "Exploiting topic-based twitter sentiment for stock prediction," in *Proceedings of the Fifty First Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 24–29, Sofia, Bulgaria, August 2013.

[7] J. F. Yang, D. Y. She, Y. K. Lai, and M. H. Yang, "Retrieving and classifying affective images via deep metric learning," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 491–498, New Orleans, LA, USA, February 2018.

[8] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, "Analyzing and predicting sentiment of images on the social web," in *Proceedings of the ACM International Conference on Multimedia*, pp. 715–718, ACM MM), Firenze, Italy, October 2010.

[9] D. Borth, R. R. Ji, T. Chen, T. Breuel, and S. F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pp. 223–232, Barcelona, Spain, October 2013.

[10] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the ACM International Conference on Multimedia*, pp. 83–92, ACM MM), Firenze, Italy, October 2010.

[11] X. Lu, P. Suryanarayan, R. B. Adams, J. Li, M. G. Newman, and J. Z. Wang, "On shape and the computability of emotions," in *Proceedings of the ACM International Conference on Multimedia*, pp. 229–238, MM), Nara, Japan, November 2012.

[12] E. Ko and E. Y. Kim, " ∗ Recognizing the sentiments of web images using hand-designed features," in *Proceedings of the IEEE 14th International Conferences on Cognitive Informatics and Cognitive Computing. (ICCI CC)*, pp. 156–161, Beijing, China, July 2015.

[13] H. Zhang, D. Qiu, R. Wu, Y. Deng, D. Ji, and T. Li, "Novel framework for image attribute annotation with gene selection XGBoost algorithm and relative attribute model," *Applied Soft Computing*, vol. 80, pp. 57–79, 2019.

[14] H. Zhang, J. Wu, H. Shi et al., "Multidimensional extra evidence mining for image sentiment analysis," *IEEE Access*, vol. 8, Article ID 103619, 2020.

[15] M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi, "Discriminant correlation analysis for feature level fusion with application to multimodal biometrics," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1866–1870, ICASSP), Shanghai, China, March 2016.

[16] P. Das, A. Ghosh, and R. Majumdar, "Determining attention mechanism for visual sentiment analysis of an image using svm classifier in deep learning-based architecture," in *Proceedings of the Eighth International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)*, pp. 339–343, ICRITO), Noida, India, June 2020.

[17] D. She, J. Yang, M.-M. Cheng, Y.-K. Lai, P. L. Rosin, and L. Wang, "WSCNet: weakly supervised coupled networks for visual sentiment classification and detection," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1358–1371, 2020.

[18] Y. Zhu, Y. Z. Zhou, Q. X. Ye, Q. Qiu, and J. B. Jiao, "Soft proposal networks for weakly supervised object localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1841–1850, Venice, Italy, October 2017.

[19] T. Durand, T. Mordan, N. Thome, and M. Cord, "WILDCAT: weakly supervised learning of deep ConvNets for image classification, pointwise localization and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5957–5966, CVPR), Honolulu, HI, USA, July 2017.

[20] M. Sun, J. F. Yang, K. Wang, and H. Shen, "Discovering affective regions in deep convolutional neural networks for visual sentiment prediction," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, Seattle, WA, USA, July 2016.

[21] T. Rao, X. Li, H. Zhang, and M. Xu, "Multi-level region-based convolutional neural network for image emotion classification," *Neurocomputing*, vol. 333, pp. 429–439, 2019.

[22] T. Rao, X. Li, and M. Xu, "Learning Multi-level deep representations for image emotion classification," *Neural Processing Letters*, vol. 51, no. 3, pp. 2043–2061, 2019.

[23] Z. Wu, M. Meng, and J. Wu, "Visual sentiment prediction with attribute augmentation and multi-attention mechanism," *Neural Processing Letters*, vol. 51, no. 3, pp. 2403–2416, 2020.

[24] K. Simonyan and A. Zisserman, "Smile, be Happy :) emoji embedding for visual sentiment analysis," in *Proceedings of the IEEE International Conference on Computer Vision Workshop*, pp. 4491–4500, ICCVW), Seoul, Republic of Korea, October 2019.

[25] J. Y. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, Venice, Italy, October2017.

[26] S. Zhao, C. Lin, P. Xu et al., "CycleEmotionGAN: emotional semantic consistency preserved CycleGAN for adapting image emotions," *Proceedings of the AAAI Conference on*

Artificial Intelligence in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 2620–2627, AAAI), Honolulu, HI, USA, February 2019.

[27] C. Lin, S. Zhao, L. Meng, and T.-S. Chua, "Multi-source domain adaptation for visual sentiment classification," *Proceedings of the AAAI Conference on Artificial Intelligence* in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 3, pp. 2661–2668, AAAI), New York, NY, USA, February 2020.

[28] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3ER: multiplicative multimodal emotion recognition using facial, textual, and speech cues," *Proceedings of the AAAI Conference on Artificial Intelligence* in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 2, pp. 1359–1367, AAAI), New York, NY, USA, February 2020.

[29] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: pretraining task-agnostic visiolinguistic representationfor vision-and-language tasks," in *Proceedings of the Thirty Third International Conference on Neural Information Processing Systems(NIPS)*, pp. 12–23, Montreal, Canada, December 2019.

[30] X. Yang, S. Feng, D. Wang, and Y. Zhang, "Image-text multimodal emotion classification via multi-view attentional network," *IEEE Transactions on Multimedia*, vol. 23, pp. 4014–4026, 2021.

[31] Y. Li, G. Hu, Y. Wang, T. Hospedales, N. M. Robertson, and Y. Yang, "Differentiable automatic data augmentation, Computer Vision - ECCV 2020," in *Proceedings of the European Conference on Computer Vision*, pp. 580–595, ECCV), Glasgow, UK, August 2020.

[32] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le, "AutoAugment: learning augmentation strategies from data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 113–123, Long Beach, CA, USA, June 2019.

[33] J. Hu, L. Shen, and S. Albanie, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 2011–2023, 2019.

[34] A. Yu and K. Grauman, "Thinking outside the pool: active training image creation for relative attributes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 708–718, Long Beach, CA, USA, June 2019.

[35] X. Xie, "Sampling active learning based on non-parallel support vector machines," *Neural Processing Letters*, vol. 53, no. 3, pp. 2081–2094, 2021.

[36] N. C. C. Thiago, M. S. Rodrigo, C. Sérgio, M. M. Mirella, and A. G. Marcos, *Ranked batch-mode active learning*, Information Sciences, vol. 379, , pp. 313–337, 2017.

[37] Q. Z. You, J. B. Luo, H. L. Jin, and J. C. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 381–388, AAAI), Austin, TX, USA, anuary 2015.

[38] Q. Z. You, J. B. Luo, H. L. Jin, and J. C. Yang, "Building a large-scale dataset for image emotion recognition: the fine print and the benchmark," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 308–314, AAAI), Phoenix, AZ, USA, February 2016.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representation (ICLR)*, pp. 1409–1556, San Diego, Chile, May 2015.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[41] B. L. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, Las Vegas, NV, USA, June 2016.

[42] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[43] T. W. Anderson, "Introduction to hotelling (1936) relations between two sets of variates," *Springer Series in Statistics*, Springer, New York, NY, USA, pp. 151–161, 1992.

[44] V. Uurtio, S. Bhadra, and J. Rousu, "Large-scale sparse kernel canonical correlation analysis," in *Proceedings of the Thirty Sixth International Conference on Machine Learning (ICML)*, pp. 6383–6391, Long Beach, CA, USA, June 2019.

[45] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[46] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[47] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[48] Z. Liu, Y. Li, L. Yao, X. Z. Wang, and G. D. Long, "Task aligned generative meta-learning for zero-shot Learning," in *Proceedings of the Thirty Fifth AAAI Conference on Artificial Intelligence*, AAAI), Columbia, Canada, February 2021.

[49] L. V. D. Maaten and G. E. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[50] Q. Z. You, H. L. Jin, and J. B. Luo, "Visual sentiment analysis by attending on local image regions," in *Proceedings of the The Thirty Firrst AAAI Conference on Artificial Intelligence*, pp. 231–237, AAAI), San Francisco, CA, USA, February 2017.

[51] A. Bochkovskiy, C. Y. Wang, and H. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," 2020, https://arxiv.org/abs/2004.10934.

[52] K. Y. Sohn, Z. Z. Zhang, C. L. Li, H. Zhang, C.-Y. Lee, and T. Pfister, "A Simple Semi-supervised Learning Framework for Object Detection," 2020, https://arxiv.org/abs/2005.04757.

[53] Y. Zhang, T. Xiang, T. M. Hospedales, and H. C. Lu, "Deep mutual learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4320–4328, Salt Lake City, UT, USA, June 2018.