# Multiple Feature Refining Network for Visual Emotion Distribution Learning

**Qinfu Xu[1], Shaozu Yuan[2], Yiwei Wei[3], Jie Wu[1], Leiquan Wang[1], Chunlei Wu[1]***

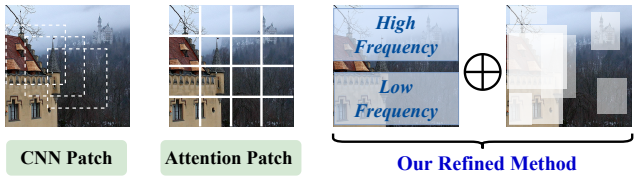
[1]Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China)
[2] JD AI Research
[3]China University of Petroleum (Beijing) at Karamay
xqfupc@163.com, wuchunlei@upc.edu.cn

## Abstract

The significance of visual emotion distribution learning (VEDL) has surged, particularly with the growing inclination to convey emotions through images. The key of VEDL lies in capturing both low- and high-level features within the same visual content, thus promoting the model for salient and subtle emotion awareness. To learn the distribution of emotions involved in images, most previous works learn coarse semantic knowledge with unbiased filtering. Consequently, they focus on the entire scene and suffer from the redundancy of semantic-irrelevant information, which diminishes the affective coherence, impeding the comprehension of emotional attributes within the treated features. In light of this, we reanalyze from the perspective of information filtering and propose a novel method called Multiple Feature Refining Network (MFRN). To minimize low-level feature redundancy, we design a wavelet-based separated frequency modeling, named Spectral Mixer, to learn invariant representations and enhance emotion saliency in low-level image features. At the higher semantic level, we design a Semantic Graph Prompt Learning for emotional semantic filtering, ensuring the purity of emotional information and providing the model with richer content semantics. Experiments conducted on three commonly used datasets have demonstrated the superiority of our MFRN model over cutting-edge methods.

## Introduction

Image emotion analysis has garnered substantial research interest due to its ability to effectively convey the diverse range of emotions expressed by individuals (Lu et al. 2019; Song et al. 2021; Yang et al. 2023b). Presently, this field finds application in diverse areas, including multimedia retrieval and social network analysis (Veltmeijer, Gerritsen, and Hindriks 2021; Zhang, He, and Lu 2019; Song et al. 2021). However, in practice, an image inherently encapsulates a blend of various emotions, rather than being characterized by a singular emotion. This is because individuals may exhibit distinct preferences and responses to the same image. The intricate interplay of these varied emotions within an image gives rise to the formidable challenge of emotion ambiguity (Plutchik 1982). Along this line, visual emotion distribution learning (VEDL) was proposed as a promising solution to tackle the issue, which advocates that visual content has more sentiment diversity. Previous works on VEDL can mainly be categorized into two types: (1) CNN-based methods (Yang, She, and Sun 2017; Xiong et al. 2019; He and Jin 2019) that employ Convolutional Neural Networks (CNNs) as feature extractor or progressive integration component. (2) Attention-based methods (Xu and Wang 2021; Zhao et al. 2019; Song et al. 2018) adopt an attention mechanism to participate in the feature representing of the emotion analysis. However, the semantic filtering of both two methods is unbiased from the emotional perspective. Therefore, the features they extracted may contain numerous disturbances that hinders the in-depth content learning.




(a) Comparison of Different Patches


(b) Illustration of Visual Semantic Diversity

Figure 1: Illustration of the difference of patches and the nature of visual semantics. (a) is the comparison of different patches in VEDL. (b) shows the visual semantic diversity.

Specifically, existing visual emotion distribution learning methods often suffer from two issues: **(1) Information redundancy.** As is shown in Figure 1 (a), CNN-based approaches tend to calculate at multiple scales through progressively deeper layers, leading to overlapping features (He et al. 2016; Lu et al. 2022). As a result, the redundancy in the treated features might introduce noise, making it challenging to distill essential emotional cues. Similarly, in Transformer-

*Corresponding Author.

based methods, the attention mechanism, while powerful in capturing long-term dependencies, may also inadvertently focus on redundant or irrelevant details in visual content, contributing to the sparse semantic density. This inefficient visual content processing can easily guide the modal with uncorrelated sentiment modeling. **(2) Limited contextual understanding.** While both CNN and Vision Transformer architectures are adept at capturing the local and global relationships, respectively, integrating these features to form a comprehensive contextual understanding remains a complex task (Yang et al. 2023a). The complicated and dynamic interactions of different emotions in an image require a more refined contextual analysis. However, existing methods either focus on learning detailed visual representations or strive to capture global relationships, lacking effective integration of features at different granularities. As is demonstrated in Figure 1 (b), visual information exhibits different semantics at different levels, which embodies the diversity of its content analysis.

To address these limitations, we reanalyze the nature of the VEDL task from the information filtration perspective and propose a novel Multiple Feature Refining Network, named MFRN, for visual emotion distribution learning. Our method aims to obtain better visual-emotion features that have less information redundancy and more semantic aggregation. Spectral Mixer module employs Dual-Tree Complex Wavelet Transform (DTCWT) (Selesnick, Baraniuk, and Kingsbury 2005) to encode images in the spatial dimension, which filters out superfluous information redundancy and therefore reduces the impact of inductive bias. The mutual conversion between the RGB domain and the frequency domain involves no information loss, which reduce the complexity of overall calculation. The superiority lies in the more compact distribution of energy in the frequency domain, where each channel clearly represents information from different frequency bands (Patro and Agneeswaran 2024). Concretely, it learns separated high and low-frequency representations for decoupled hierarchy features representations. For high-level information modeling, Semantic Graph Prompt Learning is designed to provide image objects co-relations and emotion attributes as semantic filtration when fusing with spectral features. In order to acquire effective graph representations, we explore different prompt methods for graph construction in terms of high-level semantics. After integrating the meta-information obtained by spectral features and emotion semantics obtained by graph prompt learning, our model strikes a delicate balance between visual representations and emotion distributions, thus promoting the final performance with acceptable computation complexity.

Our model achieves a new state-of-the-art (SOTA) performance for visual emotion distribution learning on three publicly available datasets: Emotion6, Flickr-LDL, and Twitter-LDL. To the best of our knowledge, we are the first to refine feature extraction towards emotional semantics to benefit the VEDL task, which introduces novel avenues for exploration in emotion analysis and other related fields.

In summary, the contributions of this work are as follows:

- We propose a novel Multiple Feature Refining Network (MFRN) for visual emotion distribution learning, which firstly considers emotional features as the orientation of information filtration and provides detailed analysis.
- To learn both low- and high-level semantics, we design a Spectral Mixer with separated wavelet modeling and Semantic Graph Prompt Learning for feature refining.
- Our MFRN achieves new state-of-the-art (SOTA) performance, even compared with large multimodal models. And it also provides insights for further exploration.

## Related Work

**Visual Emotion Distribution Learning.** Visual emotion distribution learning (VEDL) aims to distinguish the emotion category or distribution contained in an image. Previous methods can be divided into three categories: (1) Traditional machine learning methods. PT-Bayes, PT-SVM, AA-kNN, AA-BP, SA-BFGS, and SA-CPNN (Geng 2016) employed label distribution learning for visual emotion analysis task. (2) Deep learning methods, which mostly focus on the visual classification which simply employs traditional patterns (e.g., Convolution Neural Networks (He et al. 2016) and attention mechanism (Dosovitskiy et al. 2020)). JointLDL (Yang, She, and Sun 2017) designed a multi-task deep framework by jointly optimizing classification and distribution prediction. SSDL (Wang and Geng 2021) performed label distribution learning for the VEDL task by exploiting label distribution manifold. LDL-LDM (Xiong et al. 2019) designed the Structured and Sparse annotations method to capture the structured and sparse information naturally contained in the annotations of emotions. DIEDL (Wu, Huang, and Nan 2023) proposed a unified framework equipped with densely connected graph convolutional networks (DCGCN) for both coupling learning to tackle the challenging emotion ambiguity problem. StyleEDL (Jing et al. 2023) emphasized the importance of style information contained in images for visual emotion distribution learning task. Although promising, they are still facing the challenge of information filtration, which introduces additional visual noise and ambiguous semantics that hinder fine-grained visual-emotion learning. Therefore, we propose to refine the visual features and semantics by learning spectral features and graph representations for effective emotion understanding.

**Wavelet-based Frequency Modeling.** Wavelet Transformation has been proven as an effective method for visual representation and analysis. Considering that Wavelet-based modeling method is capable of learning decoupled frequency information, it has been exploited in many fundamental visual representation architectures. SDRL (Bae, Yoo, and Chul Ye 2017) introduced a novel feature space deep residual learning algorithm to obtain fine-grained visual features by considering the label manifolds mapping. DWSR (Guo et al. 2017) used low-resolution wavelet representations as inputs to recover the missing details and proposed a deep wavelet-based CNN model for image super-resolution task. MWCNN (Liu et al. 2018) presented a novel multi-level wavelet CNN model for image restoration task. Wavelet Pooling (Williams and Li 2018) leverage

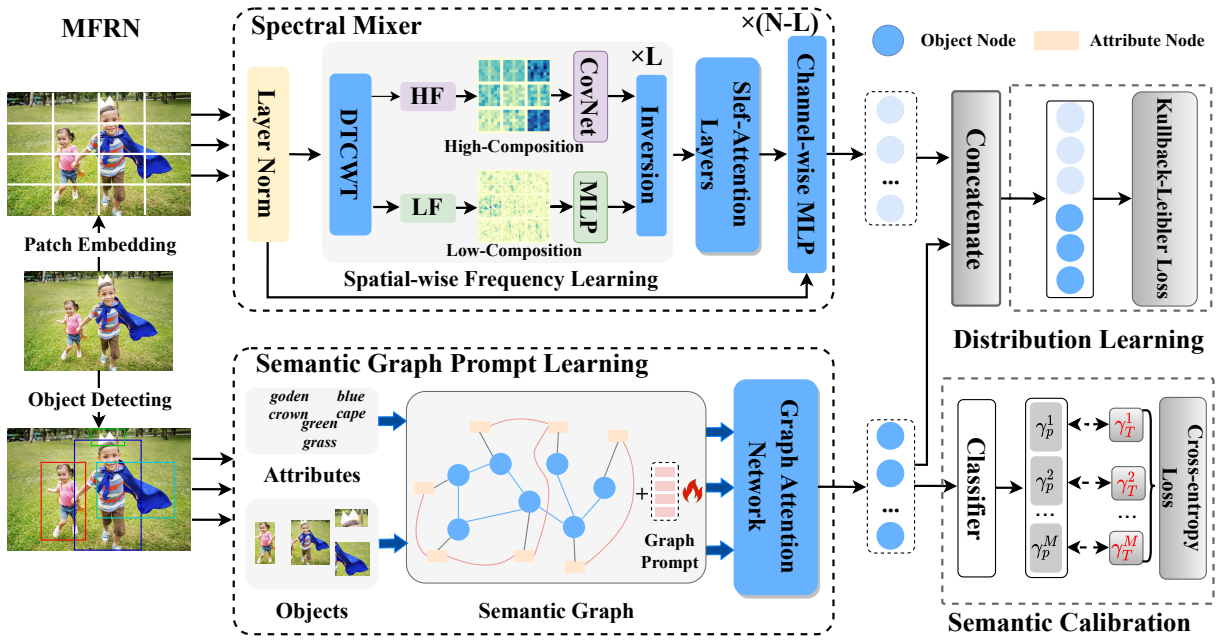



Figure 2: The architecture of MFRN. It consists of two vital modules: (1) Spectral Mixer comprises a wavelet transformation that processes Low-Frequency (LF) and High-Frequency (HF) components and a Channel-wise MLP. (2) Semantic Graph Prompt Learning module learns the sentimental information of image and offers semantic calibration for final prediction.

the wavelet information for pooling operation. WCNN (Fujieda, Takayama, and Hachisuka 2018) combined a multi-resolution analysis and CNNs into one model. SVT (Patro and Agneeswaran 2024) introduced wavelet-based modeling to reduce the computational complexity in Transformer. Although promising, these efforts have not been exploited in the field of visual or multimodal emotion analysis, which is sensitive to the semantics contained in features. In this work, we introduce frequency modeling for visual emotion analysis and provide insights and inspirations about how it works and why it brings better performance.

## Methodology

### Preliminaries

**Problem Definition.** Visual emotion distribution learning (VEDL) aims to identify the distribution of emotion categories to which a given sample image belongs. Formally given an input image $\mathbf{X_i} \in \mathbb{R}^{3\times H\times W}$, where $H$ and $W$ denote the height and width of the image, respectively, the goal of visual emotion distribution learning is to predict the emotion label distribution $\mathcal{Y}_i = \{\gamma_1^i, \gamma_2^i, ..., \gamma_N^i\}$, where $\gamma_j^i \in [0, 1]$ denotes the probability of $i$-th emotion label and $N$ represents the number of the true emotion labels.

**Framework Overview.** As discussed above, our method consists of two important modules: (1) Spectral Mixer $\mathcal{M}_{\mathbf{sm}}$ that consists of spatial learning with DTCWT transformation and channel-wise MLP for channel learning. (2) Semantic Graph Prompt Learning model $\mathcal{M}_{\mathbf{sg}}$, which considers the objects and attributes contained in an image in the view of graph nodes to learn high-level semantic features, therefore understanding fine-grained visual content relationships.

### Spectral Mixer

First, we introduce a frequency-based visual modeling method, Spectral Mixer. This method is motivated by the observations that (1) some visual content and regions are acting as jammers while learning emotional semantics. (2) Frequency domain transformation provides a new perspective to observe images, making certain characteristics of images more prominent and easier to analyze in the frequency domain. Therefore, inspired by the research conclusion, we utilize frequency information to find the low-level semantics, which play a key role in emotion analysis. In detail, given an image $\mathcal{I}_\mathbf{v} \in \mathbb{R}^{3\times H\times W}$, we first employ patch embedding to transfer image into patch sequence $\mathcal{I}_\mathbf{v}^\mathbf{p} = \{s_i\}_{i=0}^{N_p}$, where $\mathbf{s_i} \in \mathbb{R}^{p\times p}$, $N_p$, $p$ are the number of sequence and the dimension of patch, respectively. Then we obtain embedding of each patch token by utilizing positional encoder and token embedding module. To convert an image into spectral information, we use frequency transform using DTCWT to obtain the corresponding frequency representations $\mathbf{X_F}$, which has two separated components: low-frequency part $\mathbf{X}_l \in \mathbb{R}^{C\times H\times W}$ and high-frequency part $\mathbf{X}_h \in \mathbb{R}^{k\times C\times H\times W\times 2}$. Theoretically, low-frequency $\mathbf{X}_l$ encodes the global information, i.e., overall brightness, contrast, edges and contours. High-frequency $\mathbf{X}_h$ represents fine-grain information, like textures, patterns, and small features. In light of this, we further utilize two latent feature encoders MLP that can perform dense fusion and ConvNet (i.e., ResNet (He et al. 2016)) that aims to integrate hierarchical information to dispose of low and high-frequency representations, respectively. After concatenation and inversion, the spatial-wise features $\mathbf{X_F^s}$ are obtained as the dis-

criminative low-level semantic embeddings. We have provided a detailed conversion process along with the corresponding formulas in the Appendix.

$$\mathbf{X_F^s} = \mathcal{F}_{Inverse}([Proj(\mathbf{X}_l) \oplus Conv(\mathbf{X}_h)]) \quad (1)$$

Next, since the features have not been implicitly infused in the image level, we introduce an additional self-attention module (Vaswani et al. 2017) to mitigate the feature misalignment between different frequencies and therefore learn the low-level semantic correlations.

$$\mathbf{X_F^s} = \mathcal{F}_{attn}(Q, K, V, d_{sm}) = Softmax(\frac{\mathbf{QK}^\top}{\sqrt{d_{sm}}})V \quad (2)$$

where $\mathbf{Q} = \mathbf{W_q X_F^s}, \mathbf{K} = \mathbf{W_k X_F^s}, \mathbf{V} = \mathbf{W_v X_F^s}$.

As for channel modeling, we propose to integrate spectral features in the channel dimension. Channel-wise MLP $\mathcal{F}_c$ consists of a multi-layer MLP and GeLU activation functions, which are designed for channel dimension feature learning. We finally obtained the representations of the Spectral Mixer in the dimension of $d_s$:

$$\mathbf{X_V^I} = \mathbf{W}_c^i(\mathbf{W}_p \mathcal{I}_v^p + \xi \mathbf{X_F^s W}_s) + \mathbf{b_c^i} \in \mathbb{R}^{d_s} \quad (3)$$

where $\mathbf{W}_c^i$, $\mathbf{W}_p$, $\mathbf{W}_s$, and $\mathbf{b}_c^i$ denote weight matrices, $\xi$ is the hyper-paramters of fusion coefficient.

## Semantic Graph Prompt Learning

For prompting on the semantic graphs we have two major objectives: (1) Exploiting the emotional information by introducing information update between nodes, and (2) providing additional parameters to adapt the extracted features towards the VEDL dataset distribution.

In that regard, we explore the possibility of prompt injection in terms of semantic infusion. Since the spatial relationship among objects is likely to reflect their semantic correlations (Xu et al. 2024), we propose to employ object detection toolkit[1] that is built with Faster-RCNN to obtain object and attribute features for graph construction. For an image, we extract $M$ objects for graph node mapping. We leverage the object and position features $\mathcal{V}^{r+p} \in \mathbb{R}^{d_v}$ and attributes features (class name with object attribute) $\mathcal{H}^{a+c} \in \mathbb{R}^{d_h}$ to represent the visual information of an image. $d_v$ and $d_h$ are the dimensions of object and attribute features.

$$\begin{cases} \mathcal{H}_i^{a+c} &= \mathcal{M}_{roberta}\left(\left[w_a^i \ \oplus w_c^i\right]\right) \\ \mathcal{V}_i^{r+p} &= \mathbf{W_p}\left(\left[\mathbf{W_v} v_r^i \ \oplus v_p^i \mathbf{W_p}\right]\right) + \mathbf{b_p} \end{cases} \quad (4)$$

where $\mathbf{W_v}$ and $\mathbf{W_p}$ are the weight matrices, $t_a^i$ and $t_c^i$ denote the class name and attribute of the $i^{th}$ object, $w_a^i$ and $w_a^i$ denote the embedding of $t_a^i$ and $t_c^i$, $v_r^i$ and $v_p^i$ denote the object and position features, $||$ denotes the concatenation operation. Therefore, the final features can be defined as:

$$X^v = \begin{cases} \mathcal{O}_o &= [r_1, r_2, ..., r_M] \\ r_i &= \left[\mathcal{V}_i^{r+p}, \mathcal{H}_i^{a+c}\right] \end{cases} \quad (5)$$

where $\mathcal{O}_o$ denotes the object set that is formed with the object elements $r_i$. $M$ is a hyper-parameter of object number.

[1]https://github.com/peteanderson80/bottom-up-attention

To establish the visual graph based on the above features, we first define the nodes and then introduce a new edge definition that considers more information associated with image objects. We build the visual graph $\mathcal{G}_v$ that has $2M$ nodes, denoted as $\mathbf{X}_n^{\mathcal{G}} = \{g_1, g_2, g_3, ..., g_{2M-1}, g_{2M}\}$. The nodes are initialized with the visual representations $\{\mathcal{V}_1^{r+p}, \mathcal{H}_1^{a+c}, \mathcal{V}_2^{r+p}, \mathcal{H}_2^{a+c}, ..., \mathcal{V}_M^{r+p}, \mathcal{H}_M^{a+c}\}$. We then define the edge $\mathcal{A}_{ij}^v \in \mathbb{R}^{2M \times 2M}$ as below, where $i, j \in \{1, 2, ..., 2M\}$, $Sim$ and $S_{ij}^{IoU}$ denote the cosine similarity function that calculates the relationships between object attributes and the IoU score function that gives the geometrical similarity between different visual regions.

$$\mathcal{A}_{ij}^v = \begin{cases} Sim(\mathcal{H}_i^{a+c}, \mathcal{H}_j^{a+c}), if\ i\%2 = j\%2 \\ S_{ij}^{IoU}(\mathcal{V}_i^{r+p}, \mathcal{V}_j^{r+p}), if\ i\%2 = 1, j\%2 = 1 \\ 0, otherwise \end{cases} \quad (6)$$

After that, we integrate the semantic graph with $\mathcal{R}$-layer adaptive graph attention networks with dynamic node infusion attention score, which automatically adjust the degree of fusion between nodes with different types.

$$\alpha_{ij}^k = \frac{\sigma_{ReLU}(\alpha_{\mathbf{k}}^\top [g_i^k \cdot \mathbf{W}_{\mathbf{proj}}^{(\mathbf{n1})} \oplus g_j^k \cdot \mathbf{W}_{\mathbf{proj}}^{(\mathbf{n2})}]))}{\sum_{l=1}^{\mathcal{R}} exp(\sigma_{ReLU}(\alpha_{\mathbf{k}}^\top [g_i^k \cdot \mathbf{W}_{\mathbf{proj}}^{(\mathbf{d1})} \oplus g_l^k \cdot \mathbf{W}_{\mathbf{proj}}^{(\mathbf{d2})}]))} \quad (7)$$

$$\mathbf{X}_{\mathbf{k}}^{\mathcal{G}'} = \frac{1}{\mathcal{K}} \sum_{l=1}^{\mathcal{K}} \sum_{m \in \mathcal{N}_i} \alpha_{ij}^k \cdot \mathbf{g}_{\mathbf{m}}^{\mathbf{k}} \quad (8)$$

where $\mathbf{h}_{\mathbf{k}}'$ denotes the final graph note features in the $k^{th}$ layer and $\mathcal{K}$ is the number of attention heads.

However, given the specificity of the VEDL task, we expect the semantic graph can learn more emotion bias beyond common semantics. So we design a novel graph prompt learning method to introduce additional parameters for emotion calibration. One straightforward strategy is providing a set of parameters $\mathbf{Z}_p = \{p_1, p_2, ..., p_{2M}\}$ for all the node:

$$\mathbf{X}_V^S = FFN(LN(\mathbf{X}_n^{\mathcal{G}'} + \mathbf{Z}_p)) + \mathbf{X}_n^{\mathcal{G}'} + \mathbf{Z}_p \quad (9)$$

We then apply the feed-forward network (FFN) on $\mathbf{X}_n^{\mathcal{G}'}$ to obtain $\mathbf{X}_V^S$ as final representations. Since the prompt injection method can be various, we conduct experiments and analysis for empirical study in the Experiments section.

## Learning Objective

In addition to deploying KL loss used in most of VEDL methods, as the emotion distribution is inherently indeterminate, we introduce a semantic calibration loss for emotion semantic learning during image graph modeling.

We incorporate a classification loss to strengthen the model's ability to identify the dominant emotion, i.e., the emotion with the highest percentage in the label distribution. For each image, we identify the primary emotion category based on the highest percentage and apply an additional cross-entropy loss focused on this dominant emotion. This

strategy ensures that the model not only learns the overall distribution but also emphasizes accurate recognition of the predominant emotion, thus promoting the final performance.

$$\begin{cases} \mathcal{L}_{SC} = -\frac{1}{N}\sum_{k=1}^{N}\gamma_k \cdot \log \mathcal{P}(\gamma_p^k) + \zeta||\theta_1||_2^2 \\ \mathcal{L}_{DL} = \mathcal{D}_{KL}([\mathcal{M}_{\mathbf{sm}}(\mathbf{X}_i) \oplus \mathcal{M}_{\mathbf{sg}}(\mathbf{X}_i)], \mathcal{Y}_{\mathbf{gt}}) \end{cases} \tag{10}$$

To balance the difference in the algebraic scale of the two losses, we adopt an adaptive balance loss formula:

$$\mathcal{L} = (1-\mu)\mathcal{L}_{DL} + \mu\mathcal{L}_{SC} + \lambda||\theta_2||_2^2 \tag{11}$$

where $\mu \in [0, 1]$ is the tunable hyper-parameter, $\zeta$ and $\lambda$ is the $L_2$ regularization weight, $\theta_1$ and $\theta_2$ are the $L_2$ regularization parameters.

# Experiments

## Datasets and Evaluation Metrics

We conduct our experiments on three VEDL datasets, including Emotion6 (Peng et al. 2015), Flickr-LDL, and Twitter-LDL (Yang, She, and Sun 2017). **Emotion6** has 1,980 images, which are collected from Flickr and annotated with 7 classes. **Flickr-LDL** and **Twitter-LDL** contain 11,150 and 10,045 images respectively. They are annotated with 7 categories. In the Appendix, we show detailed statistics of datasets. Following previous works (Yang, She, and Sun 2017), we evaluate the performance of MFRN with six metrics (Chebyshev distance, Clark distance, Canberra metric, Kullback-Leibler divergence, cosine coefficient, and intersection similarity), which compute the similarity or distance between the predicted and true distribution.

## Implementation Details

For a fair comparison, we follow previous work (Jing et al. 2023) to pre-process the datasets. The hidden dimension of projection layer is set to 1024, which is a quarter of the original embedding dimensionality. During the training stage, we train the MFRN model for 15000 steps, utilizing AdamW (Loshchilov, Hutter et al. 2017) as the optimizer with an initial learning rate of $2 \times 10^{-3}$. We set weight decay as 0.05, batch size as 64, and dropout rate as 0.2 to train the model. We use 16 prompt vectors as the default setting. The learnable prompt is randomly initialized. All the experiments are carried out on NVIDIA Tesla P100s with totally of 32GB CUDA memory. More details about experimental implementation are provided in the Appendix.

## Comparison Methods

To ensure a fair comparison, we evaluate our model against various methods. We categorize the previous methods into three categories according to the mechanism they utilized: (1) Machine learning methods: PT-Bayes, PT-SVM, AA-kNN, AA-BP, and SA-BFGS (Geng 2016), SA-CPNN (Geng, Yin, and Zhou 2013). (2) CNN-based methods: SSDL (Wang and Geng 2021), LDL-LDM (Xiong et al. 2019), DIEDL (Wu, Huang, and Nan 2023), JointLDL (Yang, She, and Sun 2017) and StyleEDL (Jing et al. 2023). (3) Attention-based methods: EAD (Xu and Wang 2021), PDANet (Zhao et al. 2019).

## Comparison with the State-of-the-Art Methods

We compare our MFRN model with existing state-of-the-art methods to evaluate its performance. To assess the effectiveness of spectral mixing and graph prompt learning, we conduct experiments on three public datasets. The results are presented in Table 1. From Table 1, we derive the following observations: (1) Our method achieves competitive results on all three datasets, highlighting the effectiveness of integrating spectral representation with semantic graph calibration. However, the performance improvement on the Twitter-LDL dataset is less pronounced compared to the other datasets. This could be due to the abstract nature of the Twitter-LDL images, which have more complex and less explicit semantics. Consequently, there is less prior knowledge available for analyzing emotional relationships, leading to a more modest improvement in performance. (2) When compared to the most relevant methods that focus on learning representations through Convolutional Neural Networks (CNNs) or attention-based approaches, our method shows superior performance. This demonstrates the advantages of our graph prompt enhancement technique, which effectively leverages additional contextual information and refines spectral analysis. The improvement underscores the benefits of incorporating graph prompts into semantic modeling, leading to more accurate and insightful understanding.

## Ablation Study

To investigate the effectiveness of the components, we introduce several variants of our method for comparison on the Flickr-LDL and Twitter-LDL datasets.

***Effect of Different Modules.*** To explore the roles of different model designs, we compared MFRN with the following derivations. Regarding Mixer, **S** denotes that we only use spatial representations. **S+C** denotes that we introduce a channel-wise learning process. In the view of different modules, **SM** indicates that we only employ Spectral Mixer for final prediction. **GPL** denotes we only use Semantic Graph Prompt Learning method. As for loss function design **w/o SC**, we remove the Semantic Calibration component to verify its effectiveness. Table 2 reports the result of variants. We have the following conclusions: (1) Different from the attention mechanism, channel-wise learning is more vital for Spectral Mixer, which indicates that frequency transformation lacks stacked representation. (2) From the view of modules, we can infer that spectral modeling or graph modeling alone is not sufficient to produce satisfactory results, demonstrating the necessity of the cooperative relation between Spectral Mixer and Semantic Graph Prompt Learning. (3) Performance is Slightly down without the participation of semantic calibration loss, indicating that distinguishing the dominant emotion is vital for distribution learning task.

***Effect of Different Graph Prompt.*** As mentioned in Section , prompt learning in graph modeling is still not fully exploited and there is no recognized prompt injection methodology. Therefore, we conducted the experiments to explore the effectiveness of different prompt learning methods in terms of VEDL task. As is shown in Figure 3, we design four graph prompt methods. **Object Prompt**: we introduce

| | Criterion | PT-Bayes | PT-SVM | AA-kNN | AA-BP | SA-BFGS | SA-CPNN | SSDL | LDL-LDM | DIEDL | JointDL | StyleEDL | EAP† | PDANet‡ | MFRN |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| E | Cheb ↓ | 0.35(7) | 0.39(9) | 0.29(5) | 0.30(6) | 0.38(8) | 0.30(6) | 0.24(4) | 0.26(5) | 0.26(5) | 0.24(4) | 0.22(3) | 0.21(2) | 0.26(5) | **0.19(1)** |
| | Clark ↓ | 0.73(7) | 0.69(6) | 0.62(3) | 0.64(5) | 0.74(8) | 0.63(4) | 0.62(3) | 1.65(10) | 0.62(3) | 1.62(9) | 0.59(2) | 0.63(4) | 0.62(3) | **0.57(1)** |
| | Canber ↓ | 0.66(8) | 0.62(7) | 0.51(4) | 0.54(6) | 0.67(9) | 0.54(5) | 0.51(4) | 3.64(11) | 0.52(5) | 3.58(10) | 0.47(2) | 0.49(3) | 0.51(4) | **0.46(1)** |
| | KLdiv ↓ | 2.32(12) | 1.07(10) | 0.85(9) | 0.63(8) | 1.16(11) | 0.56(7) | 0.40(4) | 0.44(5) | 0.40(4) | 0.53(6) | 0.36(2) | 0.37(3) | 0.40(4) | **0.33(1)** |
| | Cosine ↑ | 0.69(8) | 0.48(12) | 0.75(6) | 0.68(9) | 0.63(11) | 0.66(10) | 0.79(5) | 0.72(7) | 0.81(4) | 0.82(3) | 0.84(2) | 0.79(5) | 0.48(12) | **0.87(1)** |
| | Intersec ↑ | 0.56(8) | 0.42(10) | 0.62(5) | 0.59(7) | 0.52(9) | 0.60(6) | 0.66(3) | 0.65(4) | 0.66(3) | 0.65(4) | 0.70(2) | 0.70(2) | 0.56(8) | **0.74(1)** |
| | Avg Rank | 8.00(11) | 8.67(12) | 5.00(6) | 6.50(9) | 9.00(13) | 6.17(9) | 3.33(4) | 6.50(10) | 3.50(5) | 5.55(7) | 2.17(2) | 3.17(3) | 6.00(8) | **1.00(1)** |
| F | Cheb ↓ | 0.44(10) | 0.55(11) | 0.28(6) | 0.36(8) | 0.37(9) | 0.30(7) | 0.23(3) | 0.25(5) | 0.23(3) | 0.24(4) | 0.21(2) | 0.22(3) | 0.24(4) | **0.20(1)** |
| | Clark ↓ | 0.89(9) | 0.87(8) | 0.57(2) | 0.82(6) | 0.86(7) | 0.82(6) | 0.78(4) | 2.14(10) | 0.79(5) | 2.19(11) | 0.76(3) | 0.77(4) | 0.76(3) | **0.73(1)** |
| | Canber ↓ | 0.85(10) | 0.83(9) | **0.41(1)** | 0.75(7) | 0.82(8) | 0.74(6) | 0.69(4) | 5.26(11) | 0.70(5) | 5.55(12) | 0.66(3) | 0.68(4) | 0.69(4) | 0.58(2) |
| | KLdiv ↓ | 1.88(9) | 1.69(8) | 3.28(10) | 0.82(6) | 1.06(7) | 1.06(7) | 0.46(3) | 0.49(4) | 0.46(3) | 0.53(5) | 0.39(2) | 0.44(3) | 0.49(4) | **0.38(1)** |
| | Cosine ↑ | 0.63(10) | 0.32(11) | 0.79(7) | 0.72(8) | 0.70(9) | 0.70(9) | 0.85(4) | 0.84(5) | 0.86(3) | 0.82(6) | 0.88(2) | 0.86(3) | 0.84(5) | **0.91(1)** |
| | Intersec ↑ | 0.49(10) | 0.29(11) | 0.64(6) | 0.53(9) | 0.56(8) | 0.60(7) | 0.68(3) | 0.66(4) | 0.70(2) | 0.65(5) | 0.71(1) | 0.69(3) | 0.65(5) | **0.71(1)** |
| | Avg Rank | 9.67(11) | 9.67(11) | 5.33(4) | 7.33(9) | 8.00(10) | 7.00(7) | 3.50(4) | 6.50(6) | 3.50(3) | 7.17(8) | 2.17(2) | 3.33(3) | 4.17(5) | **1.17(1)** |
| T | Cheb ↓ | 0.53(8) | 0.63(9) | 0.28(5) | 0.37(7) | 0.37(7) | 0.36(6) | 0.25(3) | 0.27(4) | 0.24(2) | 0.25(3) | **0.22(1)** | 0.24(2) | 0.28(5) | 0.25(3) |
| | Clark ↓ | 0.85(5) | 0.91(7) | **0.58(1)** | 0.89(6) | 0.89(6) | 0.85(5) | 0.84(3) | 2.35(8) | 0.84(3) | 2.36(8) | 0.84(3) | 0.83(4) | 0.84(3) | 0.82(2) |
| | Canber ↓ | 0.77(4) | 0.88(7) | **0.41(1)** | 0.84(6) | 0.84(6) | 0.78(5) | 0.76(3) | 6.05(8) | 0.77(4) | 6.05(8) | 0.77(4) | 0.75(2) | 0.76(3) | 0.75(2) |
| | KLdiv ↓ | 1.31(9) | 1.65(10) | 3.89(11) | 1.19(8) | 1.19(8) | 0.85(7) | 0.51(5) | 0.53(6) | 0.47(3) | 0.53(5) | **0.42(1)** | 0.50(4) | 0.51(5) | 0.45(2) |
| | Cosine ↑ | 0.53(9) | 0.25(10) | 0.82(6) | 0.71(8) | 0.82(6) | 0.75(7) | 0.86(4) | 0.85(5) | 0.87(3) | 0.85(5) | 0.89(2) | 0.87(3) | 0.86(4) | **0.92(1)** |
| | Intersec ↑ | 0.40(10) | 0.21(11) | 0.66(6) | 0.59(7) | 0.57(8) | 0.56(9) | 0.69(3) | 0.67(5) | 0.67(5) | 0.68(4) | 0.73(2) | 0.69(3) | 0.67(5) | **0.74(1)** |
| | Avg Rank | 7.17(12) | 8.67(13) | 4.83(6) | 6.67(11) | 6.50(10) | 6.17(9) | 3.33(4) | 5.67(8) | 3.00(3) | 5.50(7) | 2.17(2) | 3.00(3) | 4.17(5) | **1.83(1)** |

Table 1: Experimental Results on three datasets: Emotion6 (E), Flickr-LDL (F), and Twitter-LDL (T), are shown as mean(rank). Since each measure reflects a certain aspect of an algorithm, "Avg Rank" is used to indicate the overall performance of distribution prediction. † denotes that we re-implement EAP with Emotion6. ‡ indicates the results of PADNet on VEDL datasets

| Category | Method | Param(M) | Flickr-LDL | | | | Twitter-LDL | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | KL ↓ | Cheb ↓ | Cosine ↑ | Intersec ↑ | KL ↓ | Cheb ↓ | Cosine ↑ | Intersec ↑ |
| Mixer | S-only | 43 | 0.4352 | 0.2914 | 0.8653 | 0.5962 | 0.4972 | 0.3024 | 0.8911 | 0.6695 |
| | S+C | 68 | 0.3808 | 0.2036 | 0.9142 | 0.7164 | 0.4547 | 0.2575 | 0.9288 | 0.7454 |
| Module | SM-only | 68 | 0.5914 | 0.4163 | 0.6874 | 0.5237 | 0.6483 | 0.4516 | 0.7275 | 0.6848 |
| | GPL-only | 54 | 0.7142 | 0.5373 | 0.7035 | 0.4908 | 0.6266 | 0.5274 | 0.6645 | 0.5687 |
| | **MFRN (ours)** | 122 | 0.3898 | 0.2009 | 0.9168 | 0.7142 | 0.4517 | 0.2562 | 0.9262 | 0.7472 |
| Loss | w/o SC | 109 | 0.4528 | 0.3256 | 0.7628 | 0.6969 | 0.6137 | 0.3842 | 0.7788 | 0.6293 |

Table 2: Effectiveness of proposed components. We compare different variants on two datasets. S, S+C, SM, GPL, and SC indicate Spatial-only, Spatial-Channel, Spectral Mixer, Graph Prompt Learning, and Semantic Calibration, respectively.

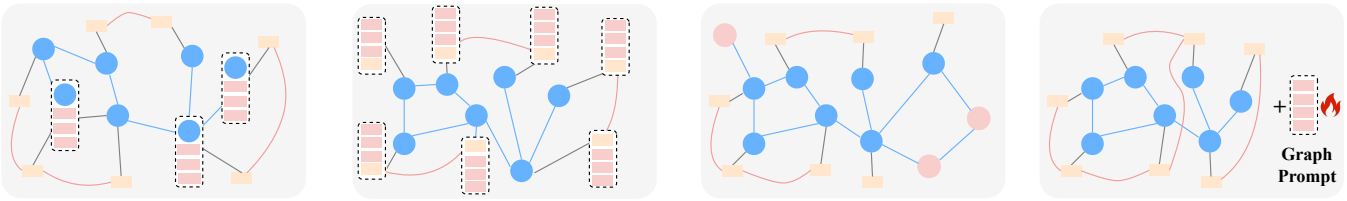
| Method | Flickr-LDL | | | |
|---|---|---|---|---|
| | KL ↓ | Cheb ↓ | Cosine ↑ | Intersec ↑ |
| OP | 0.3543 | 0.2636 | 0.9075 | 0.6583 |
| AP | 0.4771 | 0.3507 | 0.7939 | **0.8014** |
| EP | **0.3247** | 0.2876 | 0.8421 | 0.6353 |
| UP | 0.38 | **0.2001** | **0.9142** | 0.7187 |

Table 3: Effectiveness of graph prompt methods. OP, AP, EP, and UP indicate Object Prompt, Attribute Prompt, External Prompt, and Unified Prompt, respectively.

additional prompt parameters for object nodes to help adjust object features in graph learning. **Attribute Prompt**: denotes prompting attribute node with parameter optimization. **External Prompt** is to design a new node apart from the existing two types and is tunable during training. **Unified Prompt** provides all nodes with prompt parameters, which is adopted in our MFRN model. Table 3 reports the result of variants. We can obtain that: (1) Unified Prompt has the best performance across out of four metrics, indicating the advancement of appendant graph prompt method. (2) Object Prompt obtains secondary result and Attribute Prompt performs unsatisfactory, which gives us the insight that prompting object node is more befitting. (3) External Prompt is the only one that introduces new graph nodes, and it gets equally favorable results compared with Object Prompt.

## Semantic Effectiveness Analysis

From the experimental results, we can infer that our MFRN method can capture the emotional semantics within different levels. Thanks to the help of frequency modeling, it becomes a hierarchical process to capture the emotional semantics from visual features. Further, considering the redundancy in image content, we design object-centered graph reasoning method to implicitly build the interaction between objects and attributes. Different from previous methods that employ unbiased semantic filtering (Xiong et al. 2019; Jing et al. 2023), MRFN aims to learn emotion-oriented visual features that contains sparse but clean semantics, which alleviates the ambiguity in modeling visual sentiment information. Table 2 reveals that the results achieved when taking frequency features into consideration. However, as the paper explains, obtaining semantics only from Spectral Mixer cannot fully satisfy the information required for VEDL task, indicating that frequency is helpful but not dominant in distribution




(a) Object Prompt (b) Attribute Prompt (c) External Prompt (d) Unified Prompt

Figure 3: Ablation of different choices for graph prompt learning.

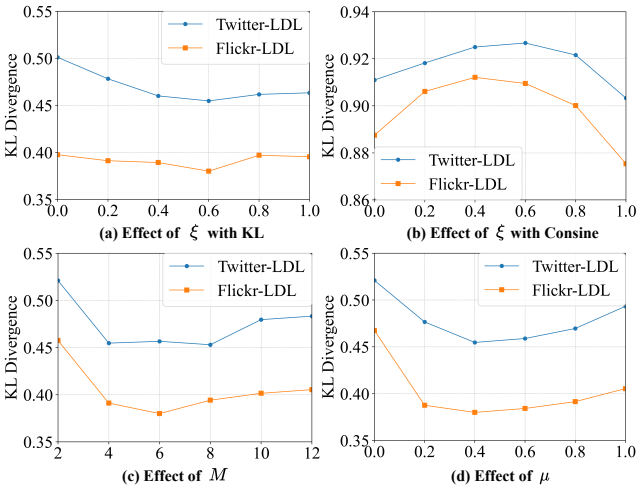



Figure 4: Effectiveness of hyper-parameters.

modeling. Next, we find that graph prompt influences much more than context. This is because the original semantics have less emotional bias that can be mined and additional prompt brings parameter update space for emotions.

## Emotion Classification Performance

For the emotion analysis task, it is vital to distinguish the emotion classes contained in images. Therefore, we perform a dual-loss training strategy to address both the accurate representation of emotion distributions and the precise identification of dominant emotions.By focusing on the emotion with the highest percentage, it reinforces the model's ability to accurately predict the dominant emotion. This targeted approach not only improves the model's sensitivity to the most prevalent emotion but also ensures that the classification of the dominant emotion is robust and reliable.

## Parameter Sensitivity Analysis

We evaluate the effect of the number of several hyper-parameters: infusion coefficient $\xi$, object number $M$, and loss balance coefficient $\mu$. The results are shown in Figure 4. We find that: (1) As the infusion proportion $\xi$ increases, the KL divergence decreases and then starts to rise, and the Cosine metric first rises and goes down after $\xi = 0.6$. It confirms that the low-level semantics may not be able to reason emotional information alone. (2) Tendency in Figure 4(c) reveals that it works best when the number of object $M$ is set to 6. (3) Loss balance $\mu$ is used to find suitable loss aggregation. The best performance is achieved when the distribution loss dominates, and the model performance deteriorates when increasing the proportion of semantic calibration loss.

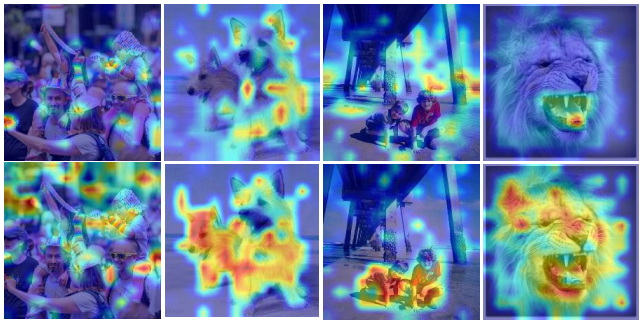


Figure 5: CAM visualizations. MFRN (second line) captures sentiment-related fine-grained details that the baseline (Style-LDL (Jing et al. 2023)) (first line) misses.

## Visualization Study

To further understand the process of our MFRN model, we conduct a visualization analysis using grad-CAM (Bishop and Lewith 2010) to present a qualitative comparison of some examples sampled from Filckr-LDL dataset. As is shown in Figure 5, it is effortless to observe that our model pays more attention to the key emotions that reflect the main semantics and focuses more on regions that may influence the distributions. In the third column examples, previous SOTA method Style-LDL (Jing et al. 2023) concentrates incorrectly due to failed visual-emotion grounding, while our MRFN performs accurately due to its strong ability to capture specific objects in emotional visual scenarios.

## Conclusion and Future Work

In this work, we propose a novel approach named MFRN for visual emotion distribution learning. We utilize spectral information for spatial token mixing, enhancing the interpretability of final prediction. To better understand the emotional attribute contained in an image, we design a semantic graph prompt learning for semantic calibration. Experiments on public datasets demonstrate the best performance of our MFRN method. In the future, we will improve visual emotion distribution learning with large multimodal models to obtain better data annotations (Zhong et al. 2022).

## Acknowledgments

This work is partially supported by the grants from the Natural Science Foundation of Shandong Province (ZR2024MF145), the National Natural Science Foundation of China (62072469), and the Qingdao Natural Science Foundation (23-2-1-162-zyyd-jch).

## References

Bae, W.; Yoo, J.; and Chul Ye, J. 2017. Beyond deep residual learning for image restoration: Persistent homology-guided manifold simplification. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 145–153.

Bishop, F. L.; and Lewith, G. T. 2010. Who uses CAM? A narrative review of demographic characteristics and health factors associated with CAM use. *Evidence-Based Complementary and Alternative Medicine*, 7(1): 11–28.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fujieda, S.; Takayama, K.; and Hachisuka, T. 2018. Wavelet convolutional neural networks. *arXiv preprint arXiv:1805.08620*.

Geng, X. 2016. Label distribution learning. *IEEE-TKDE*, 1734–1748.

Geng, X.; Yin, C.; and Zhou, Z.-H. 2013. Facial age estimation by learning from label distributions. *IEEE-TAPMI*, 2401–2412.

Guo, T.; Seyed Mousavi, H.; Huu Vu, T.; and Monga, V. 2017. Deep wavelet prediction for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 104–113.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

He, T.; and Jin, X. 2019. Image emotion distribution learning with graph convolutional networks. In *ACM MM*, 382–390.

Jing, P.; Liu, X.; Wang, J.; Wei, Y.; Nie, L.; and Su, Y. 2023. StyleEDL: Style-Guided High-order Attention Network for Image Emotion Distribution Learning. In *ACM MM*, 853–861.

Liu, P.; Zhang, H.; Zhang, K.; Lin, L.; and Zuo, W. 2018. Multi-level wavelet-CNN for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 773–782.

Loshchilov, I.; Hutter, F.; et al. 2017. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5.

Lu, X.; Zhu, L.; Cheng, Z.; Li, J.; Nie, X.; and Zhang, H. 2019. Flexible online multi-modal hashing for large-scale multimedia retrieval. In *ACM MM*, 1129–1137.

Lu, Z.; Xie, H.; Liu, C.; and Zhang, Y. 2022. Bridging the gap between vision transformers and convolutional neural networks on small datasets. *Advances in Neural Information Processing Systems*, 35: 14663–14677.

Patro, B.; and Agneeswaran, V. 2024. Scattering Vision Transformer: Spectral Mixing Matters. *Advances in Neural Information Processing Systems*, 36.

Peng, K.-C.; Chen, T.; Sadovnik, A.; and Gallagher, A. C. 2015. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *CVPR*, 860–868.

Plutchik, R. 1982. A psychoevolutionary theory of emotions.

Selesnick, I. W.; Baraniuk, R. G.; and Kingsbury, N. C. 2005. The dual-tree complex wavelet transform. *IEEE signal processing magazine*, 22(6): 123–151.

Song, K.; Yao, T.; Ling, Q.; and Mei, T. 2018. Boosting image sentiment analysis with visual attention. *Neurocomputing*, 218–228.

Song, T.; Zheng, W.; Liu, S.; Zong, Y.; Cui, Z.; and Li, Y. 2021. Graph-embedded convolutional neural network for image-based EEG emotion recognition. *IEEE Transactions on Emerging Topics in Computing*, 10(3): 1399–1413.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Veltmeijer, E. A.; Gerritsen, C.; and Hindriks, K. V. 2021. Automatic emotion recognition for groups: a review. *IEEE Transactions on Affective Computing*, 14(1): 89–107.

Wang, J.; and Geng, X. 2021. Label distribution learning by exploiting label distribution manifold. *IEEE TNNLS*.

Williams, T.; and Li, R. 2018. Wavelet pooling for convolutional neural networks. In *International conference on learning representations*.

Wu, H.; Huang, Y.; and Nan, G. 2023. Doubled coupling for image emotion distribution learning. *Knowledge-Based Systems*, 110107.

Xiong, H.; Liu, H.; Zhong, B.; and Fu, Y. 2019. Structured and sparse annotations for image emotion distribution learning. In *AAAI*, 363–370.

Xu, Q.; Wei, Y.; Yuan, S.; Wu, J.; Wang, L.; and Wu, C. 2024. Learning emotional prompt features with multiple views for visual emotion analysis. *Information Fusion*, 108: 102366.

Xu, Z.; and Wang, S. 2021. Emotional attention detection and correlation exploration for image emotion distribution learning. *IEEE Trans. Affective Computing*.

Yang, J.; She, D.; and Sun, M. 2017. Joint Image Emotion Classification and Distribution Learning via Deep Convolutional Neural Network. In *IJCAI*, 3266–3272.

Yang, T.; Zhu, Y.; Xie, Y.; Zhang, A.; Chen, C.; and Li, M. 2023a. Aim: Adapting image models for efficient video action recognition. *arXiv preprint arXiv:2302.03024*.

Yang, Y.; Panagopoulou, A.; Zhou, S.; Jin, D.; Callison-Burch, C.; and Yatskar, M. 2023b. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19187–19197.

Zhang, W.; He, X.; and Lu, W. 2019. Exploring discriminative representations for image emotion recognition with CNNs. *IEEE Transactions on Multimedia*, 22(2): 515–523.

Zhao, S.; Jia, Z.; Chen, H.; Li, L.; Ding, G.; and Keutzer, K. 2019. PDANet: Polarity-consistent deep attention network for fine-grained visual emotion regression. In *ACM MM*, 192–201.

Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L. H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16793–16803.