CausVSR: Causality Inspired Visual Sentiment Recognition

Xinyue Zhang^{1,2} , **Zhaoxia Wang**^{4} , **Hailing Wang**^{2,3} , **Jing Xiang**^{2,3} , **Chunwei Wu**^{2,3}
and **Guitao Cao**^{2,3*}
[1]Shanghai Institute of Artificial Intelligence for Education, East China Normal University
[2]MoE Engineering Research Center of SW/HW Co-design Technology and Application, East China Normal University
[3]Shanghai Key Laboratory of Trustworthy Computing, East China Normal University
[4]School of Computing and Information Systems, Singapore Management University
xyzhang@stu.ecnu.edu.cn, zxwang@smu.edu.sg, {52215902004, 51215902108, 52215902005}@stu.ecnu.edu.cn, gtcao@sei.ecnu.edu.cn

Abstract

Visual Sentiment Recognition (VSR) is an evolving field that aims to detect emotional tendencies within visual content. Despite its growing significance, detecting emotions depicted in visual content, such as images, faces challenges, notably the emergence of misleading or spurious correlations of the contextual information. In response to these challenges, we propose a causality inspired VSR approach, called CausVSR. CausVSR is rooted in the fundamental principles of Emotional Causality theory, mimicking the human process from receiving emotional stimuli to deriving emotional states. CausVSR takes a deliberate stride toward conquering the VSR challenges. It harnesses the power of a structural causal model, intricately designed to encapsulate the dynamic causal interplay between visual content and their corresponding pseudo sentiment regions. This strategic approach allows for a deep exploration of contextual information, elevating the accuracy of emotional inference. Additionally, CausVSR utilizes a global category elicitation module, strategically employed to execute front-door adjustment techniques, effectively detecting and handling spurious correlations. Experiments, conducted on four widely-used datasets, demonstrate CausVSR's superiority in enhancing emotion perception within VSR, surpassing existing methods.

1 Introduction

With the exponential growth of visual social media content, Visual Sentiment Recognition (VSR) has become a critical task in human perception intelligence [You *et al.*, 2016; Yang *et al.*, 2017a; Xu *et al.*, 2020]. VSR analyzes visual content to recognize sentiments and emotions such as joy, sadness, anger, etc., providing personalized support by assessing human emotional states through the analysis of individual visual features [Zhao *et al.*, 2016]. The evolution of VSR has transformed it into an interdisciplinary field aimed at optimizing interactions between Artificial Intelligence (AI) and humans. Its applications span diverse areas, including but not limited to education support [Tonguç and Ozkara, 2020], assessment of mental health [Fei *et al.*, 2020], and protection of the young generation [Tan *et al.*, 2023].

Despite progress in existing methods, challenges in VSR persist. Prevailing VSR frameworks, predominantly rely on weakly supervised strategies [Zhou *et al.*, 2016; Durand *et al.*, 2017]. These frameworks entail an initial classification, succeeded by the generation of pseudo sentiment maps employing various methods [She *et al.*, 2019; Zhang and Xu, 2020]. Pseudo sentiment maps play a pivotal role in emotion perception. However, while pseudo sentiment maps provide valuable weak supervision and enhance feature representation during training, thus contributing to final predictions, their generation process can be inconsistent in yielding robust features [Zhang *et al.*, 2023]. Factors like hidden contextual information and spurious correlation often act as confounders, leading to misinterpretations and inaccurate predictions.

Grounded in the foundations of causal theory [Yang *et al.*, 2021b; Wang *et al.*, 2021] to tackle challenges, we build a novel VSR method named **Caus**ality inspired **V**isual **S**entiment **R**ecognition (CausVSR). CausVSR is designed to accurately predict emotions within visual content by emphasizing the causal relationships between emotional stimuli and the most emotionally evocative regions, effectively countering the challenges stemming from confounders. Our proposed method seeks to replicate the psychological concept of "Emotional Causality" [Coëgnarts and Kravanja, 2016; Mittal *et al.*, 2021]. This involves intricately embedding the sequence of human emotions evoked when perceiving visual content, such as images, into an emotional chain encompassing "External Events", "Emotional Perception", and "Emotional State", illustrated in Figure 1(a). A detailed description of the proposed CausVSR (see Figure 1(b)) is provided in Section 3. To excavate the implicit contextual information [Goh *et al.*, 2019] while eliminating the effects of the confounders in the weakly-supervised VSR, we utilize a structural causal model [Yang *et al.*, 2021a; Wang, 2022] to establish the pseudo sentiment recognition

[*] Corresponding Author

process, simulating the complete human emotional response from stimulus to perception. Furthermore, we propose a global category elicitation module, adopting the concept of causal front-door intervention, to block the confounding factors' influence on pseudo sentiment maps, thereby steering the final results towards the desired outcome. The proposed CausVSR method leverages the causality inspired framework to decode the intricate interplay of emotions, leading to a more comprehensive understanding of visual sentiments.

Our contributions are summarized as follows:

- We propose CausVSR, a novel VSR approach inspired by Emotional Causality theory. It effectively simulates the human transition from perceiving emotional stimuli to identifying emotional states in visual content.
- We develop a causal model for CausVSR, analyzing the interaction between visual content and pseudo sentiment regions to enhance contextual exploration and improve emotional inference accuracy.
- We utilize a global category elicitation module in the proposed CausVSR to facilitate front-door adjustment techniques, effectively handling spurious correlations in sentiment recognition.
- Extensive experiments conducted across four widely-used datasets demonstrate the effectiveness and superiority of the proposed CausVSR compared to existing methods.

2 Related Work

Emotion Causality. Emotion Causality, as illustrated in Figure 1(a), delves into the traditional understanding of emotions, suggesting that our emotional experiences are part of a broader causal chain [Coëgnarts and Kravanja, 2016]. This chain typically comprises three distinct stages: (i) an external event, (ii) an emotional perception process, and (iii) a resulting emotional state [Young and Suri, 2019; He *et al.*, 2019]. Emotion Causality underscores the dynamic and multifaceted nature of emotional responses, emphasizing that these are shaped not solely by the immediate visual stimuli but also by the individual's cognitive and emotional framework. Human emotions are intricately intertwined with our perception of the surrounding environment, be it immediate, imagined, or rooted in memories [Brown, 2023]. The domain of affective science has been actively exploring diverse methodologies to quantify these complex emotional states. In previous research, the various approaches in this field have been categorized into two distinct methodologies: some examine the emotional journey as a holistic process, while others adopt a more segmented perspective, dissecting emotions into a series of steps from the initial trigger to the final orientation [Herzberg, 2009]. A prominent example of the latter approach is the theory of Emotion Causality.

Visual Sentiment Recognition (VSR). VSR has gained prominence for interpreting emotions from visual content [You *et al.*, 2016]. Initially, VSR methods predicted sentiments from entire images [Zhao *et al.*, 2014; Rao *et al.*, 2020], but later research highlighted the benefits of focusing on local emotional regions for improved accuracy [Yang *et al.*, 2018]. Current VSR approaches in VSR mainly involve attention-based methods, enhancing relevant regions [Xu *et al.*, 2020; Zhang *et al.*, 2023], and multi-dimensional aggregation-based methods [Zhang *et al.*, 2022; Yang *et al.*, 2018], identifying emotional areas through mathematical combinations of different dimensions. Although existing VSR methods have successfully directed the macro-level identification of emotional regions, they encounter difficulties in accurately detecting emotions [Saxena *et al.*, 2020]. This is primarily due to the emergence of misleading or spurious correlations within contextual information [Yang *et al.*, 2023a]. Consequently, there is a pressing need to confront and resolve these challenges.

3 CausVSR: Our Approach

Our proposed CausVSR, depicted in Figure 1(b), addresses the challenges, presents an innovative approach by integrating Emotion Causality as its fundamental framework. It incorporates human experiences of visual emotion within a causal model, enabling a more profound exploration of sentiment analysis in visual content. The essence of the proposed CausVSR can be conceptualized into the following core components: Emotion-Stimuli Feature Representation, Causality-based Emotional Perception, and Discrete Emotion State Prediction.

3.1 Emotion-Stimuli Feature Representation

This component corresponds to the External Events in Emotion Causality. Human emotions are triggered by environmental stimuli. Images, as vital emotional stimuli, contain elements ranging from low-level cues like color and contrast to higher-level context. Recognizing these stimuli's nature is vital for understanding emotions. CausVSR introduces a stimulus generation source that elicits emotional perceptions and aids in the development of emotional states.

We utilize a deep feature extractor to construct a multi-scale feature representation. Given an emotional image dataset $(x_i, y_i)_{i=1}^{N}$, where x_i represents the i-th image, N denotes the number of samples, and y_i represents the corresponding emotional label, among M emotion labels. The architecture of the feature extractor includes convolutional blocks $\{B_1, B_2, ..., B_n\}$, here n represents the number of convolutional blocks. Specifically, we utilize the Res2Net-101 network pretrained on ImageNet as the backbone of the feature extractor [Gao *et al.*, 2019], calculating multiple-scale feature maps through forward propagation, therefore n is set to 4. Considering that high-level semantics in computer vision are typically human-understandable and expressible descriptors used to represent the content of images [Dou *et al.*, 2023], we select the feature map generated by the last convolutional block, denoted as $P_{B_n} \in R^{w \times h \times c}$, where w and h represent the spatial dimensions (width and height) of the feature map, and c represents the number of channels.

3.2 Causality-based Emotional Perception

This component aligns with the Emotional Perception in Emotion Causality. The eyes transform light into neural im-

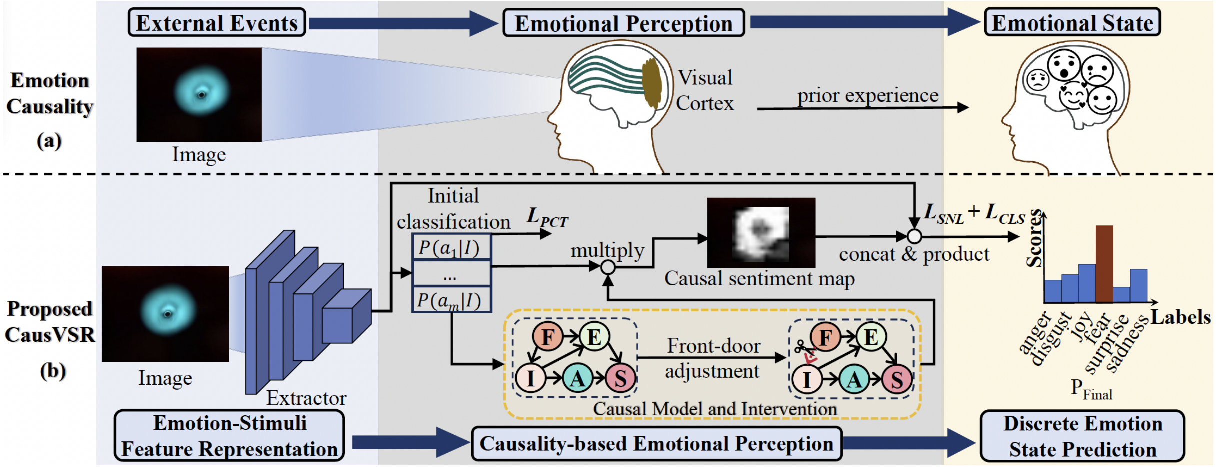


Figure 1: The pipeline of the proposed causality inspired VSR. Inspired by (a) Emotion Causality, (b) CausVSR uniquely extracts emotion-stimuli features from images, acting as external events. CausVSR bridges the visual sentiment recognition process using a structural causal model with front-door adjustment.

pulses processed by the brain's visual cortex, playing a crucial role in bridging external events and emotional states. CausVSR introduces a structural causal graph to model this emotional perception process, identifying causal links between visual stimuli and sentiment regions. Also, it implements a deep-learning front-door adjustment model for causal intervention.

Build the Structural Causal Graph

Our objective in capturing emotion-related regions is to improve VSR models by investigating the process of generating sentiment regions. Taking the advantages of causal theory specialties, we utilize a structural causal graph in this section to model human emotion perception. Considering the diverse attributes of emotional objects across different categories, the confounders may contribute to a range of spurious correlations that impact the initial sentiment categories. For instance, when context acts as a general confounder, it introduces non-causal features into initial classifier. This is evident when "joy" is frequently associated with flowers, potentially misleading the classifier to focus on floral features. Moreover, subtle differences between images of different categories present additional challenges for the model, particularly when common features are shared. For example, "fear" and "sadness" are both negative emotions, and their visual representations often include dark tones and overlapping elements, like dilapidated windows, complicating the distinction between them.

Figure 2(a) depicts the proposed structural causal model for human-like emotion perception ($I \rightarrow A \rightarrow S$). Here, I represents feature representations (external events), A represents the initial classification results obtained by the first classifier in the weakly-supervised framework, S represents causal sentiment maps. There exists a causal relationship $I \rightarrow A$, since the initial classification A is generated from features I. The initial A guides the arousal of the emotional region, thereby obtaining the causal sentiment map S ($A \rightarrow S$). F is referred to as the confounder, and E represents the confounder set related to external events I. It can be observed that F is a common cause of I and E ($F \rightarrow E \rightarrow S$, $F \rightarrow I \rightarrow A \rightarrow S$). F leads to the contamination of $P(S|I)$, and failing to reflect the true causal relationship between I and S. In conclusion, we initially establish the causal relationship from $I \rightarrow S$, which can be formulated as follows.

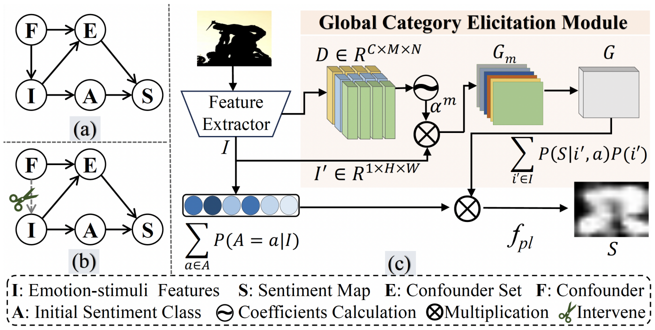

F
E
I
A
S
(a)
(b)
Global Category Elicitation Module
$D \in R^{C \times M \times N}$
G_m
G
Feature Extractor
I
α^m
$I' \in R^{1 \times H \times W}$
$\sum_{i' \in I} P(S|i', a)P(i')$
$\sum_{a \in A} P(A = a|I)$
f_{pl}
S
(c)
I: Emotion-stimuli Features
S: Sentiment Map
E: Confounder Set
F: Confounder
A: Initial Sentiment Class
⊖ Coefficients Calculation
⊗ Multiplication
✂ Intervene

Figure 2: Details of Causality-based Emotional Perception in CausVSR. (a) Structural causal diagram mapping visual feature representation to pseudo causal sentiment maps. (b) Front-door adjustment intervention. (c) Deep learning architecture based on the causal model in (a) and interventions in (b).

$$P(S|I) = \sum_{a \in A} P(A = a|I)P(S|A = a) \qquad (1)$$

Front-door Adjustment

In the presence of the confounder F, it is challenging to fully uncover the causal relationship from I to S. We observe that the causal diagram $I \rightarrow A \rightarrow S$ can be formulated in a front-door model [Pearl, 2000]. The causal effect of $I \rightarrow A$ can be ascertained from the data, as all other paths are isolated by the collider structure $I \rightarrow E \leftarrow F$, free from confounding and backdoor paths [Yang *et al.*, 2021b]. The causal effect of A on S has confounding due to a common cause F. Even though we lack the specific data of F, we have information on features *I*. By cutting off $F \rightarrow I$ in the causal graph (scissors in Figure 2(b)), we eliminate the paths that previously caused confounding, which helps to learn the desired causal relationship. Hence, we perform an intervention operation $do(\cdot)$ as $P(S|do(I))$, and analyze the causal effect of $I \rightarrow S$ through front-door adjustment [Pearl, 2000; Wang *et al.*, 2021]:

$$P(S|do(I)) = \sum_{a \in A} P(a|i) \sum_{i' \in I} P(S|i', a)P(i') \tag{2}$$

where $P(i')$ represents the probability of features $i \in I$ belongs to class a occurring. Assuming the training samples $P(I = i)$ follow a uniform distribution [Amrani *et al.*, 2022; Wang, 2022], that the probability of an input i of class a occurs is approximately $1/N$, where N represents the total number of training samples. The variable $\sum_{i' \in I} P(S|I = i', A)P(I = i')$ represents the expectation of the predicted causal sentiment maps S for class a across the entire training set. Inspired by [Cheng *et al.*, 2023], we build a Global Category Elicitation Module (GCEM) to compute the expectation $\sum_{i' \in I} P(S|i', a)P(i')$. As shown in the light orange region of Figure 2(c), suppose $D \in R^{C \times M \times N}$ is a global dictionary for the entire training set, where C represents the number of channels, M represents the categories of emotions, and N represents the number of atoms in each category in *D*. According to the previous analysis on $\sum_{i' \in I} P(S|i', a)P(i')$, we aim to express the feature map I through a sparse linear combination of atoms associated with the class dictionary *D*, with the intention of invoking a global category map. Through Equation (3), we obtain the category map G_m for class $m \in M$.

$$G_m = \alpha_m \otimes I^{'} \tag{3}$$

where $I^{'} \in R^{1 \times H \times W}$ is obtained through the channel-wise global average pooling operation, reducing the channel dimension to alleviate computational complexity. $\otimes$ represents the entrywise product. $\alpha_m \in R^{M \times H \times W}$ represents the coefficients of D, which reflects the response for each atom vector on the input, and can be calculated as follows:

$$\alpha_m = \frac{\exp(-(d_k)^\top p_i)}{\sum_{j=1}^{M \times N} \exp(-(d_j)^\top p_i)} \tag{4}$$

where $exp(\cdot)$ is used to calculated the similarity of pixel vector $p_i \in I$ and the k-th atom vector $d_k \in D$. According to maps G_m for each category, we obtain the global category map G through Equation (5):

$$G = concat(G_1, G_2, ...G_m, ..., G_M) \tag{5}$$

where $concat(\cdot)$ denotes the concatenation operation.

Our goal is the localization of affective-rich sentiment regions, which are specific areas in the image directly relevant to emotional experiences. To train the initial classifier, we introduce the pooling strategy [She *et al.*, 2019] followed by a network $g(\cdot)$, which approximates the distribution $P(S|do(I))$ in Equation (2), to obtain the causal pseudo sentiment maps S:

$$S = f_{pl}(g(I, A)) \tag{6}$$

The pooling strategy f_{pl} evokes the probability of different emotions in each receptive field, thereby sketching emotion-specific regions as Equation (7):

$$f_{pl} = \sum_{m=1}^{M} w_m(\frac{1}{j}\sum_{i=1}^{j} z_{m,i}), m \in \{1, 2, ..., M\} \tag{7}$$

where f_{pl} treats pseudo sentiment maps S as the prediction score, training a classifier that has been adjusted using the causal front-door adjustment. w_c represents a single feature map-level score for each emotional class $m \in M$, obtained through a cross-spatial pooling strategy regardless of the input size. The specific calculation formula is $w_m = \frac{1}{j}\sum_{i=1}^{j} f_{GAP}(z_{m,i})$. Here, f_{GAP} represents the global average pooling operation. $z_{m,i}$ refers to the i-th feature map of the m-th emotional label. j represents j emotional class-related detectors.

3.3 Discrete Emotion State Prediction

CausVSR mimics Emotion State through a weakly supervised framework, predicting discrete emotional labels by combining prior knowledge from an initial classifier with causal sentiment regions. This process emulates how human predictions of emotional states in visual content rely on past experiences and interpretations, distinguishing them from non-emotional material perception.

As shown in Figure 1, in the component of Discrete Emotion State Prediction, we derive the ultimate prediction P_{Final} with causal pseudo sentiment map S through the classifier:

$$P_{Final} = Softmax(f_{GAP}(concat(S, P_{B_n}))) \tag{8}$$

where the concatenate operation combining the S with the rich-semantic features B_n.

CausVSR generates two outputs: a causal sentiment map and sentiment classification predictions, supervised by two loss functions in an end-to-end manner. L_{PCT} oversees sentiment map perception, while L_{CLS} focuses on sentiment classification, utilizing Cross Entropy Loss to measure the similarity between predicted pseudo maps and ground truth for L_{PCT}, and the discrepancy in predicted classification and labels for L_{CLS}. Additionally, a surface normal loss L_{SNL} [Hu *et al.*, 2019] is integrated into L_{PCT} to refine the accuracy of pseudo sentiment regions by assessing the divergence between the surface normals of the pseudo sentiment region map and the ground truth, thereby enhancing the map's detail and structure as shown in Equation (9).

$$L_{SNL} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(1 - \frac{\langle m_{ij}^p, m_{ij}^g \rangle}{\sqrt{\langle m_{ij}^p, m_{ij}^p \rangle}\sqrt{\langle m_{ij}^g, m_{ij}^g \rangle}}) \tag{9}$$

where (i, j) denotes the specific spatial location, the surface normal of the pseudo sentiment map is $m_{ij}^p \equiv$

$[-\bigtriangledown_x (p_{ij}), -\bigtriangledown_y (p_{ij}), 1]^\top$, and the surface normal of the ground truth is $m_{ij}^g \equiv [-\bigtriangledown_x (g_{ij}), -\bigtriangledown_y (g_{ij}), 1]^\top$, and $\langle \cdot, \cdot \rangle$ is the inner product that multiply vectors m_{ij}^p and m_{ij}^g together. Ultimately, we train the proposed deep model CausVSR using the following loss function in Equation (10):

$$L_F = \lambda_1 L_{PCT} + \lambda_2 L_{SNL} + \lambda_3 L_{CLS} \tag{10}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the alternatively balanced parameters, which are experientially set on Res2Net-101 to 1, 0.5, and 1. And we employ stochastic gradient descent (SGD) to optimize the loss function L_F.

4 Experiments

4.1 Experiments Setup

The implementation of CausVSR is carried out using the widely adopted PyTorch framework [Paszke *et al.*, 2019]. The training input images are standardized to a size of 448 × 448. To diversify the training data, we employ random resized cropping initially, followed by random horizontal flips. These techniques are designed to address overfitting issues in scenarios with limited data and enhance overall model generalization. For model training, we utilize Stochastic Gradient Descent (SGD) as the optimization algorithm, with momentum decay and weight decay set to 0.9 and $5E-4$, respectively, for improved computational efficiency. The initial learning rate is set at $1E-4$ and is reduced by a factor of 100 every 10 iterations. All experiments are conducted on Nvidia Tesla P100-PCIE with a total memory capacity of 16 GB.

4.2 Comparison with State-of-the-art Methods

We conducted experiments on four widely used VSR datasets, which include a large-scale dataset, Flickr and Instagram (FI-8) [You *et al.*, 2016], and three small-scale datasets: EmotionROI (6 classes) [Panda *et al.*, 2018], IAPS-Subset (2 classes) [Machajdik and Hanbury, 2010], and Twitter II (2 classes) [Borth *et al.*, 2013]. Given the balanced class distribution in the datasets, we employ accuracy as the evaluation metric, consistent with previous research work.

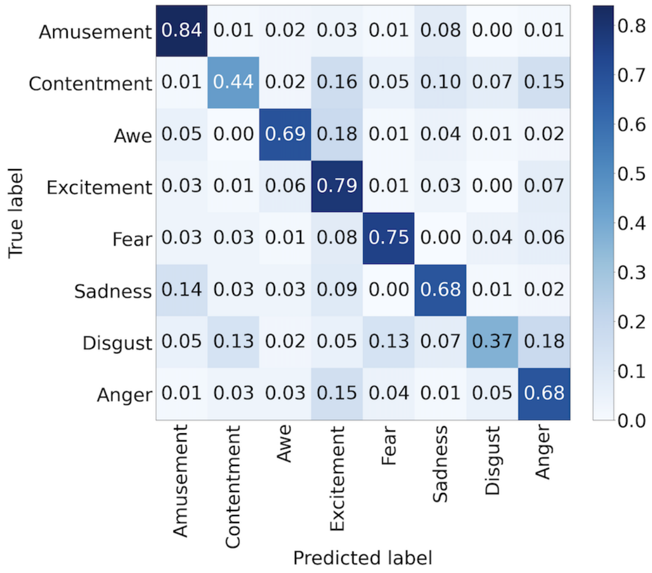
Results on Large-scale FI-8

To validate the efficacy of CausVSR in visual sentiment recognition, we conduct a comparative analysis with state-of-the-art methods on widely recognized image sentiment datasets. Table 1 presents the experimental results for the FI-8 dataset, and our proposed model achieves an accuracy of 72.57%, surpassing other VSR methods.

We conducted a comprehensive performance analysis, comparing CausVSR with well-established models such as Class Activation Mapping (CAM) [Zhou *et al.*, 2016], WSCNet [She *et al.*, 2019], Yang's [Yang *et al.*, 2023b], and DCNet [Zhang *et al.*, 2023]. CAM, a leader in visual interpretability, utilizes features from the final convolutional layer for information visualization. WSCNet resolves emotional label ambiguities by weighing all class activation maps, identifying emotion-inducing regions. Yang's leverages psychology's gradual emotion cognition mechanism, organizing the relationship between images and emotions in a knowledge graph to identify image emotions visually. DCNet uses a saliency prior for sentiment region generation. Despite their good performance, CausVSR outperforms these models in sentiment region perception, showing significant improvements of 4.03%, 2.5%, 1.44% and 0.92%.

Methods	FI-8
Self-Attention [Vaswani *et al.*, 2017]	24.01
Zhao's [Zhao *et al.*, 2014]	46.13
Sentibank [Borth *et al.*, 2013]	49.23
DeepSentibank [Chen *et al.*, 2014]	51.54
ImageNet-AlexNet [Krizhevsky *et al.*, 2017]	38.26
ImageNet-VGG16	41.22
ImageNet-ResNet101 [He *et al.*, 2016]	50.01
Yang's [Yang *et al.*, 2017a]	66.79
SPN [Zhu *et al.*, 2017]	66.57
WILDCAT [Durand *et al.*, 2017]	67.03
MAP [He *et al.*, 2019]	68.13
CAM [Zhou *et al.*, 2016]	68.54
WSCNet [She *et al.*, 2019]	70.07
Yamamoto's [Yamamoto *et al.*, 2021]	70.46
Yang's [Yang *et al.*, 2023b]	71.13
DCNet [Zhang *et al.*, 2023]	71.65
CausVSR	**72.57**

Table 1: Comparison on FI-8 dataset



	Amusement	Contentment	Awe	Excitement	Fear	Sadness	Disgust	Anger
Amusement	0.84	0.01	0.02	0.03	0.01	0.08	0.00	0.01
Contentment	0.01	0.44	0.02	0.16	0.05	0.10	0.07	0.15
Awe	0.05	0.00	0.69	0.18	0.01	0.04	0.01	0.02
Excitement	0.03	0.01	0.06	0.79	0.01	0.03	0.00	0.07
Fear	0.03	0.03	0.01	0.08	0.75	0.00	0.04	0.06
Sadness	0.14	0.03	0.03	0.09	0.00	0.68	0.01	0.02
Disgust	0.05	0.13	0.02	0.05	0.13	0.07	0.37	0.18
Anger	0.01	0.03	0.03	0.15	0.04	0.01	0.05	0.68

Figure 3: Confusion matrix on FI-8 dataset.

In the classification aspect of VSR tasks, CausVSR's performance was assessed against the FI-8 dataset using a confusion matrix (Figure 3). CausVSR performs well in expressing the majority of sentiment categories. However, it exhibits suboptimal performance in conveying the emotions of "Contentment" and "Disgust". We speculate that this might be attributed to the overlapping feature distributions of "Contentment" and "Disgust" compared to other categories, making it more challenging for the model to predict these emotions accurately.

Methods	EmotionROI	Methods	IAPS-Subset	Methods	Twitter II
[Yang *et al.*, 2017b]	45.40	[Borth *et al.*, 2013]	81.79	[Chen *et al.*, 2014]	70.23
[Yang *et al.*, 2017a]	52.40	[Chen *et al.*, 2014]	85.63	[Simonyan *et al.*, 2014]	71.79
[Zhu *et al.*, 2017]	52.70	[You *et al.*, 2015]	88.84	[Durand *et al.*, 2017]	78.81
[Durand *et al.*, 2017]	55.05	[Simonyan *et al.*, 2014]	89.37	[Zhou *et al.*, 2016]	79.13
[Zhou *et al.*, 2016]	55.72	[Yang *et al.*, 2018]	92.39	[Sun *et al.*, 2016]	80.91
[She *et al.*, 2019]	58.25	[Zhang and Xu, 2020]	95.83	[She *et al.*, 2019]	81.35
[Zhang *et al.*, 2023]	59.60	[Zhang *et al.*, 2023]	95.90	[Zhang *et al.*, 2023]	82.50
CausVSR	**59.82**	CausVSR	**95.97**	CausVSR	**82.86**

Table 2: Classification accuracy comparison on three small-scale datasets

Results on Other Three Datasets

We conducted experiments on the other three datasets as well, with results in Table 2 showing CausVSR outperforming other methods across all datasets. WSCNet [She *et al.*, 2019] introduced a weakly-supervised framework for sentiment analysis, while DCNet [Zhang *et al.*, 2023] leveraged visual saliency for sentiment classification. CausVSR advances this by using causal modeling to enhance sentiment region expression, leading to a 1.57% and 0.22% performance increase over WSCNet and DCNet on the EmotionROI dataset, and 0.07% and 0.36% on the IAPS-Subset dataset and Twitter II dataset compared with DCNet, respectively, as shown in Table 2.

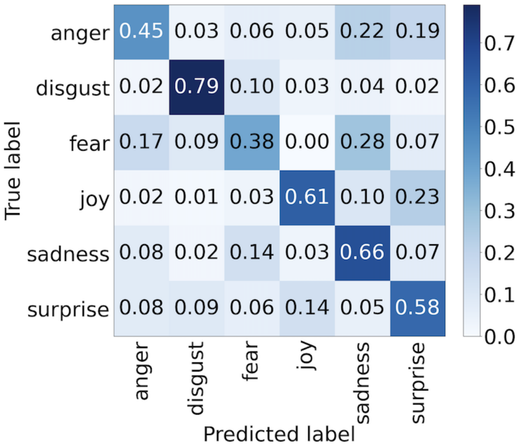
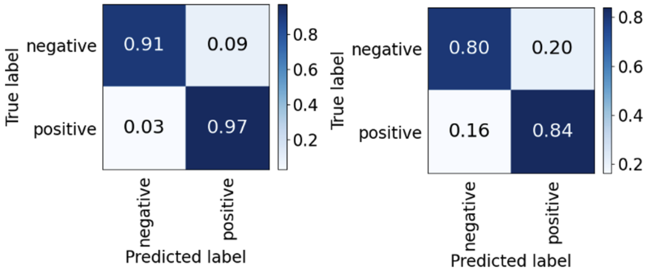


Figure 4: Confusion matrix on EmotionROI dataset.

Confusion matrices for three datasets are presented in Figures 4 to Figure 5. Figures 4 display the results obtained by the proposed CausVSR for the 6-class EmotionROI dataset. It is observed that CausVSR generally performs well in recognizing most emotions in the EmotionROI datasets, yet it tends to confuse "anger" and "fear" with "sadness". We analyze that beyond the subjective nature of dataset annotations, the dynamic nature of real-world emotional experiences, where individuals frequently transition between feelings of fear, anger, and sadness, could contribute to the model's difficulty in distinguishing between these emotions.

Figures 5(a) and 5(b) show the results obtained on two binary classification datasets. Visual sentiment recognition performance on two binary datasets is notably better than on the other two larger datasets. We analyze that the smaller datasets simplify the recognition task, making it easier for the model to distinguish between two distinct categories. Moreover, their focus on clear and consistent features of positive and negative sentiments allows for improved recognition outcomes.



(a) IAPS-Subset (b) Twitter II

Figure 5: Confusion matrices on binary classification dataset.

4.3 Ablation Studies

Causality Importance Analysis. To assess the impact of causality in VSR, we use DCNet as the baseline, and integrate the Causality-based Emotional Perception (CEP) process with DCNet. Table 3 demonstrates that the generated causal pseudo sentiment maps improve DCNet's performance on the FI-8 and EmotionROI datasets. Additionally, as shown in Table 3, CEP enhances CausVSR's capability on the same datasets.

	CEP	FI-8	EmotionROI
DCNet	*w/o*	71.65	59.60
	w	**72.18**	**59.79**
CausVSR	*w/o*	71.58	59.63
	w	**72.57**	**59.82**

Table 3: Impact of causality

Ablation Study on Integration for Emotion State Prediction. As CausVSR adopts a weakly supervised strategy to mimic human emotion state prediction, we investigated integrating causal sentiment maps S with various stages of Emotion-Stimuli Feature Representation. The feature maps generated by different convolutional blocks in Emotion-Stimuli Feature Representation are denoted as P_{B_n}. Table 4 presents results indicating the impacts of integration with each convolutional block individually. The results in Table 4 demonstrate that the integration of causal sentiment maps S with the convolutional block P_{B_4} yields the best accuracy. This observation aligns with our initial hypothesis in Section 3.1, suggesting that higher-level semantic features closely correspond to human-understandable descriptors.

P_{B_1}	P_{B_2}	P_{B_3}	P_{B_4}	FI-8	EmotionROI
$\surd$				65.96	35.71
	$\surd$			67.62	36.05
		$\surd$		69.75	32.91
			$\surd$	**72.57**	**59.82**

Table 4: Impact of integration of S with different P_{B_n}

4.4 Qualitative Results

Figure 6 displays the original image (column 1), saliency maps [Chen *et al.*, 2020] (column 2), pseudo sentiment maps generated through CAM [Zhou *et al.*, 2016] (column 3), DCNet [Zhang *et al.*, 2023] (column 4), and CausVSR (column 5) on the EmotionROI dataset.

CausVSR exhibits advantages over other methods. Firstly, CausVSR generates pseudo sentiment maps that distinctively differ from saliency maps. While saliency maps primarily delineate foreground and background boundaries, methods like CAM reveals that human emotional perception in sentiment classification concentrates near specific regions, which may not always align with the outlined shapes. In contrast, CausVSR accurately identifies and emphasizes emotionally significant areas, as demonstrated by its focus on the red spots near the bird's tail in the third row as shown in Figure 6. Secondly, CausVSR's superiority becomes evident in its ability to prioritize emotionally relevant regions rather than focusing solely on prominent objects highlighted by saliency maps. For example, in the fifth row, while saliency maps highlight the stamens, CausVSR captures not only the stamens but also surrounding areas linked to emotion, like the yellow petals. Furthermore, CausVSR captures the details overlooked by other methods, such as CAM and DCNet. For instance, in the fourth row, while other methods struggle to extract certain relevant areas, CausVSR's emotion regions intelligently include both the left railing and the right wall. Lastly, a notable strength of CausVSR is its reduced generation of false activations corresponding to unrelated backgrounds, as observed in the first and second rows. This reinforces its superiority in discerning emotionally relevant regions while minimizing erroneous activations.

5 Conclusion

The pioneering approach presented in this research, CausVSR, rooted in Emotional Causality theory, adeptly addresses the challenges prevalent in VSR tasks. By employing an intricate structural causal model and strategically utilizing a global category elicitation module, CausVSR adeptly navigates the complexities within visual content, resulting in a significant enhancement of emotional inference accuracy.

The comprehensive experiments conducted on four widely-used datasets showcase CausVSR's effectiveness and superiority for VSR tasks, surpassing existing methods. The proposed CausVSR, mimicking human emotional stimulus processing and integrating front-door adjustment techniques, not only demonstrates CausVSR's potential but also signifies a promising advancement in the VSR domain.

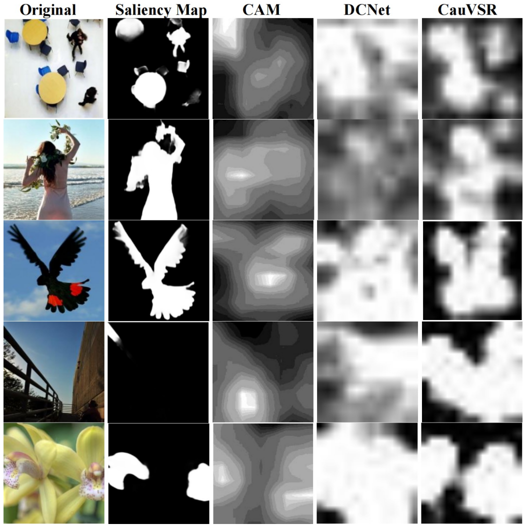


Figure 6: The visualization of pseudo sentiment maps from different methods for EmotionROI dataset. The pseudo sentiment maps from DCNet and CausVSR, generated through random horizontal flipping and random patch extraction for data augmentation, represent only a section of the original image. To enable effective comparison, all the figures are cropped to the same dimensions.

Looking forward, our future pursuits will involve exploring CausVSR's application in real-world scenarios, particularly within educational settings. Additionally, we aim to delve into its potential as a metric for measuring psychological states. As an innovative contribution, CausVSR presents a significant leap in overcoming the challenges inherent in detecting emotions within visual content, thereby opening avenues for further research and practical applications in the domain of visual sentiment recognition. These endeavors are geared toward amplifying CausVSR's practical utility and making substantive contributions to the evolving landscape of visual sentiment recognition.

Acknowledgments

This research was supported by National Natural Science Foundation of China under Grant Nos. 61871186, 61771322. Xinyue Zhang also acknowledges the support of the China Scholarship Council program (Project ID: 202306140115).

References

[Amrani *et al.*, 2022] Elad Amrani, Leonid Karlinsky, and Alex Bronstein. Self-supervised classification network. In *European Conference on Computer Vision*, pages 116–132. Springer, 2022.

[Borth *et al.*, 2013] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232, 2013.

[Brown, 2023] Teneille R Brown. Minding accidents. *U. Colo. L. Rev.*, 94:89, 2023.

[Chen *et al.*, 2014] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*, 2014.

[Chen *et al.*, 2020] Shuhan Chen, Xiuli Tan, Ben Wang, Huchuan Lu, Xuelong Hu, and Yun Fu. Reverse attention-based residual network for salient object detection. *IEEE Transactions on Image Processing*, 29:3763–3776, 2020.

[Cheng *et al.*, 2023] Gong Cheng, Pujian Lai, Decheng Gao, and Junwei Han. Class attention network for image recognition. *Science China Information Sciences*, 66(3):132105, 2023.

[Coëgnarts and Kravanja, 2016] Maarten Coëgnarts and Peter Kravanja. Perceiving emotional causality in film: a conceptual and formal analysis. *New Review of Film and Television Studies*, 14:1–27, 03 2016.

[Dou *et al.*, 2023] Hui Dou, Furao Shen, Jian Zhao, and Xinyu Mu. Understanding neural network through neuron level visualization. *Neural Networks*, 168:484–495, 2023.

[Durand *et al.*, 2017] Thibaut Durand, Taylor Mordan, Nicolas Thome, and Matthieu 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*, pages 642–651, 2017.

[Fei *et al.*, 2020] Zixiang Fei, Erfu Yang, David Day-Uei Li, Stephen Butler, Winifred Ijomah, Xia Li, and Huiyu Zhou. Deep convolution network based emotion analysis towards mental health care. *Neurocomputing*, 388:212–227, 2020.

[Gao *et al.*, 2019] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):652–662, 2019.

[Goh *et al.*, 2019] Kam Meng Goh, Usman Ullah Sheikh, and Tomás H Maul. Recognizing hidden emotions from difference image using mean local mapped pattern. *Multimedia Tools and Applications*, 78:21485–21520, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[He *et al.*, 2019] Xiaohao He, Huijun Zhang, Ningyun Li, Ling Feng, and Feng Zheng. A multi-attentive pyramidal model for visual sentiment analysis. In *2019 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2019.

[Herzberg, 2009] Larry A Herzberg. Direction, causation, and appraisal theories of emotion. *Philosophical psychology*, 22(2):167–186, 2009.

[Hu *et al.*, 2019] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1043–1051. IEEE, 2019.

[Krizhevsky *et al.*, 2017] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[Machajdik and Hanbury, 2010] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 83–92, 2010.

[Mittal *et al.*, 2021] Trisha Mittal, Puneet Mathur, Aniket Bera, and Dinesh Manocha. Affect2mm: Affective analysis of multimedia content using emotion causality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5661–5671, 2021.

[Panda *et al.*, 2018] Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K Roy-Chowdhury. Contemplating visual emotions: Understanding and overcoming dataset bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 579–595, 2018.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[Pearl, 2000] Judea Pearl. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3, 2000.

[Rao *et al.*, 2020] Tianrong Rao, Xiaoxu Li, and Min Xu. Learning multi-level deep representations for image emotion classification. *Neural processing letters*, 51:2043–2061, 2020.

[Saxena *et al.*, 2020] Anvita Saxena, Ashish Khanna, and Deepak Gupta. Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems*, 2(1):53–79, 2020.

[She *et al.*, 2019] Dongyu She, Jufeng Yang, Ming-Ming Cheng, Yu-Kun Lai, Paul L Rosin, and Liang Wang. Wscnet: Weakly supervised coupled networks for visual sentiment classification and detection. *IEEE Transactions on Multimedia*, 22(5):1358–1371, 2019.

[Sun *et al.*, 2016] Ming Sun, Jufeng Yang, Kai Wang, and Hui Shen. Discovering affective regions in deep convolutional neural networks for visual sentiment prediction. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2016.

[Tan *et al.*, 2023] Yee Sen Tan, Nicole Anne Teo Huiying, Ezekiel En Zhe Ghe, Jolie Zhi Yi Fong, and Zhaoxia

Wang. Video sentiment analysis for child safety. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 783–790. IEEE, 2023.

[Tonguç and Ozkara, 2020] Güray Tonguç and Betul Ozaydın Ozkara. Automatic recognition of student emotions from facial expressions during a lecture. *Computers & Education*, 148:103797, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2021] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3091–3100, 2021.

[Wang, 2022] Yiping Wang. Causal class activation maps for weakly-supervised semantic segmentation. In *UAI 2022 Workshop on Causal Representation Learning*, 2022.

[Xu *et al.*, 2020] Jie Xu, Zhoujun Li, Feiran Huang, Chaozhuo Li, and S Yu Philip. Visual sentiment analysis with social relations-guided multiattention networks. *IEEE Transactions on Cybernetics*, 52(6):4472–4484, 2020.

[Yamamoto *et al.*, 2021] Takahisa Yamamoto, Shiki Takeuchi, and Atsushi Nakazawa. Image emotion recognition using visual and semantic features reflecting emotional and similar objects. *IEICE TRANSACTIONS on Information and Systems*, 104(10):1691–1701, 2021.

[Yang *et al.*, 2017a] Jufeng Yang, Dongyu She, and Ming Sun. Joint image emotion classification and distribution learning via deep convolutional neural network. In *IJCAI*, pages 3266–3272, 2017.

[Yang *et al.*, 2017b] Jufeng Yang, Ming Sun, and Xiaoxiao Sun. Learning visual sentiment distributions via augmented conditional probability neural network. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[Yang *et al.*, 2018] Jufeng Yang, Dongyu She, Ming Sun, Ming-Ming Cheng, Paul L Rosin, and Liang Wang. Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Transactions on Multimedia*, 20(9):2513–2525, 2018.

[Yang *et al.*, 2021a] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9593–9602, 2021.

[Yang *et al.*, 2021b] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9847–9857, 2021.

[Yang *et al.*, 2023a] Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, et al. Context deconfounded emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19005–19015, 2023.

[Yang *et al.*, 2023b] Hansen Yang, Yangyu Fan, Guoyun Lv, Shiya Liu, and Zhe Guo. Exploiting emotional concepts for image emotion recognition. *The Visual Computer*, 39(5):2177–2190, 2023.

[You *et al.*, 2015] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 381–388. AAAI Press, 2015.

[You *et al.*, 2016] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao 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*, volume 30, 2016.

[Young and Suri, 2019] Gerald Young and Gaurav Suri. Emotion regulation choice: A broad examination of external factors. *Cognition and Emotion*, 2019.

[Zhang and Xu, 2020] Haimin Zhang and Min Xu. Weakly supervised emotion intensity prediction for recognition of emotions in images. *IEEE Transactions on Multimedia*, 23:2033–2044, 2020.

[Zhang *et al.*, 2022] Jing Zhang, Xinyu Liu, Mei Chen, Qi Ye, and Zhe Wang. Image sentiment classification via multi-level sentiment region correlation analysis. *Neurocomputing*, 469:221–233, 2022.

[Zhang *et al.*, 2023] Xinyue Zhang, Jing Xiang, Hanxiu Zhang, Chunwei Wu, Hailing Wang, and Guitao Cao. Dcnet: Weakly supervised saliency guided dual coding network for visual sentiment recognition. In *26th European Conference on Artificial Intelligence*, pages 3050 – 3057, 2023.

[Zhao *et al.*, 2014] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 47–56, 2014.

[Zhao *et al.*, 2016] Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, Wenlong Xie, Xiaolei Jiang, and Tat-Seng Chua. Predicting personalized emotion perceptions of social images. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1385–1394, 2016.

[Zhou *et al.*, 2016] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[Zhu *et al.*, 2017] Yi Zhu, Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Soft proposal networks for weakly supervised object localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1841–1850, 2017.