Counterfactual Thinking Driven Emotion Regulation for Image Sentiment Recognition

Xinyue Zhang[1,2] , **Zhaoxia Wang**[4] , **Hailing Wang**[2,3] and **Guitao Cao**[1,2,3]*
[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@stu.ecnu.edu.cn,
gtcao@sei.ecnu.edu.cn

Abstract

Image sentiment recognition (ISR) facilitates the practical application of affective computing on rapidly growing social platforms. Nowadays, region-based ISR methods that use affective regions to guide emotion prediction have gained significant attention. However, existing methods lack a causality-based mechanism to guide affective region generation and effective tools to quantitatively evaluate their quality. Inspired by the psychological theory of Emotion Regulation, we propose a counterfactual thinking driven emotion regulation network (CTERNet), which simulates the Emotion Regulation Theory by modeling the entire process of ISR based on human causality-driven mechanisms. Specifically, we first use multi-scale perception for feature extraction to simulate the stage of situation selection. Next, we combine situation modification, attentional deployment, and cognitive change into a counterfactual thinking based cognitive reappraisal module, which learns both affective regions (factual) and other potential affective regions (counterfactual). In the response modulation stage, we compare the factual and counterfactual outcomes to encourage the network to discover the most emotionally representative regions, thereby quantifying the quality of affective regions for ISR tasks. Experimental results demonstrate that our method outperforms or matches the state-of-the-art approaches, proving its effectiveness in addressing the key challenges of region-based ISR.

1 Introduction

Emotion is a state that individuals constantly evoke and experience [Cui *et al.*, 2023]. With the continuous development of social communication channels, vast amounts of visual information, particularly images, now flood social media platforms. Image Sentiment Recognition (ISR) focuses on detecting and classifying the emotions conveyed by images through visual analysis. On a personal level, ISR can play a crucial role in identifying early signs of mental health issues, such as depression and anxiety, enabling timely intervention [She *et al.*, 2019]. On a commercial level, ISR can analyze users' emotional states and their fluctuations on social media, helping platforms enhance user experience, identify social trends, and track emerging hot topics [Wang *et al.*, 2020]. As a result, ISR has found widespread applications across diverse fields, including education [Tan *et al.*, 2023] and opinion mining [Li *et al.*, 2019].

Traditional ISR methods rely on handcrafted features to identify the emotions conveyed by images. With the development of the ISR field, these methods can be broadly categorized into image-level and region-based approaches. Image-level methods employ convolutional neural networks to analyze an entire image and determine its predominant emotion [You *et al.*, 2015]. In contrast, region-based methods focus on specific affective regions within an image, which are critical for evoking emotional responses [Zhang *et al.*, 2022b; She *et al.*, 2019; Zhang *et al.*, 2024]. As a result, region-based ISR has emerged as a primary area of research interest. However, these methods face two significant challenges:

First, current ISR methods rely on correlation-based learning [You *et al.*, 2016; Zhang *et al.*, 2023], focusing on statistical associations rather than causal relationships. This reliance often results in overfitting to dataset biases or spurious correlations, ultimately limiting the model's ability to accurately identify regions that genuinely evoke emotions [Rao *et al.*, 2021]. To address these limitations, some researchers have attempted to model the innate causality mechanisms of humans [Zhang *et al.*, 2024; Yang *et al.*, 2023a]. These causality mechanisms enable humans to naturally mitigate emotional content biases, allowing fair emotional judgments despite visual biases [Yang *et al.*, 2024]. For example, flowers are often perceived as positive imagery [You *et al.*, 2016], yet this perception is not solely based on their visual features but also influenced by causal reasoning rooted in personal experiences and cultural context. Despite cultural differences, humans consistently interpret emotions in visual content by leveraging these experiences and contexts,

[*]Corresponding Author

resonating with familiar symbols and associations through well-developed causal reasoning processes [You *et al.*, 2015; Sun *et al.*, 2022]. Secondly, the lack of effective tools to evaluate affective regions often causes models to focus on irrelevant attributes. For instance, if "joy" samples frequently include a table, the model might associate the table with joy. Region-based methods address this using additional annotations (e.g., bounding boxes) [Yang *et al.*, 2018b], yet this approach is labor-intensive and difficult to scale.

Therefore, we integrate the innate counterfactual thinking ability of humans to adjust predictions more effectively. This approach mimics the natural human emotion judgment process, driven by causality, within an ISR model. According to Gross's Emotion Regulation Theory, emotion regulation involves processes that influence the occurrence, experience, and expression of emotions [Gross, 2015]. These processes are categorized into five stages: situation selection, situation modification, attentional deployment, cognitive change, and response modulation [Gross, 2002]. Inspired by this, we build a **C**ounterfactual **T**hinking driven **E**motion **R**egulation **Net**work (CTERNet) for the ISR task, which aims to simulate the Emotion Regulation process of humans by re-evaluating the scenarios of emotional stimuli to change the meaning of the outcomes. The CTERNet is employed with a response-focused strategy [Gross, 2002] in Emotion Regulation, where the adjustment occurs after the emotion has already been generated and the emotional response has been activated. The response-focused strategy aligns well with how humans experience emotions when observing an image. For instance, as shown in Figure 1(a), when people notice the rose (Area 1) in a picture and interpret the emotion conveyed as "joy", they might then reflect due to the nature of cognitive load and attention [Sweller, 2020]: *What if I do not notice the Area 1 in the image*? That is, after obtaining a prediction, humans will reflect through counterfactual thinking [Rao *et al.*, 2021], which can re-examine the problem to ensure that no important information has been overlooked and to obtain the correct emotion prediction.

Specifically, we first utilize a multi-scale perception network to extract features from images, simulating various situation selections. Next, we integrate counterfactual thinking to combine situation modification, attentional deployment, and cognitive change into a counterfactual-thinking-based cognitive reappraisal module (C^{2}RM). To clarify the causal relationships between affective regions and emotional predictions within the counterfactual framework, we build a structural causal graph In C^{2}RM, we simulate human emotional arousal under different situations by learning affective regions (factual) and exploring alternative potential regions (counterfactual). During the response modulation stage, emotional predictions are obtained by comparing the impacts of factual and counterfactual regions. This approach enables context-aware predictions and evaluates the quality of identified affective regions. To focus on emotionally representative regions and minimize biased sentimental cues, we apply the total effect (TE) [Pearl, 2014; Yang *et al.*, 2024].

The main contributions are summarized as follows:

- We propose CTERNet, a novel method based on Emotion Regulation Theory, for ISR through total effect analysis. This method equips machines with the ability to compare factual and counterfactual outcomes based on causality.
- We construct counterfactual thinking based cognitive reappraisal module (C^{2}RM) as the core component of CTERNet, simulating human emotional arousal across various situations.
- We utilize structural causal graphs to reformulate CTERNet, revealing the ISR process from image input to prediction, identifying irrelevant attributes, and intuitively providing counterfactual thinking interventions.

2 Related Work

2.1 Region-based Image Sentiment Recognition

Region-based ISR methods aim to identify specific areas within an image that strongly convey emotions [Yang *et al.*, 2018a]. Existing approaches mainly follow two strategies. The first is precise annotation-based methods (e.g., bounding boxes and segmentation masks) [Zhang *et al.*, 2022b], which require additional prior knowledge, such as generating proposals to locate bounding boxes. However, these methods are time-consuming and labor-intensive compared to image-level annotations [Zhang *et al.*, 2023]. Additionally, they often retain proposals focused on foreground objects after processing through a localization regressor, potentially leading to information loss during emotional analysis [She *et al.*, 2019]. To address these issues, a mainstream weakly-supervised strategy has emerged, offering soft proposals for evoking emotions. This strategy identifies attention-grabbing regions related to human visual attention, leveraging visual saliency in weakly-supervised ISR [You *et al.*, 2017; Fan *et al.*, 2017; He *et al.*, 2019; Zhang *et al.*, 2024].

2.2 Counterfactual in Emotion Regulation

In psychology, Emotion Regulation involves managing emotions through planned efforts. Despite varying definitions [Thompson, 1994; Feldner *et al.*, 2003; Gross, 2015; Cludius *et al.*, 2020], Gross's process model is widely accepted [Gross, 2002; Gross and Feldman Barrett, 2011], distinguishing between antecedent-focused and response-focused strategies. Cognitive reappraisal, a form of cognitive change, alters perceptions of emotional events by hypothesizing different outcomes [McRae *et al.*, 2012; Theodorou *et al.*, 2023], essentially using counterfactual thinking [Rye *et al.*, 2008; Sirois *et al.*, 2010; Parikh *et al.*, 2022]. This involves evaluating past events to predict, reason, and attribute causality. Some studies suggest a causal relationship between emotion and counterfactual thinking during cognitive reappraisal, aiding adaptation to social and environmental needs.

3 Methodology

Emotion Regulation Theory, as illustrated in Figure 1(a), delves into the traditional understanding of emotions or sentiments, suggesting that our emotional experiences are part of a broader causal chain [Zhang *et al.*, 2024; Coëgnarts and Kravanja, 2016; Gross, 2015], which lay the foundation for

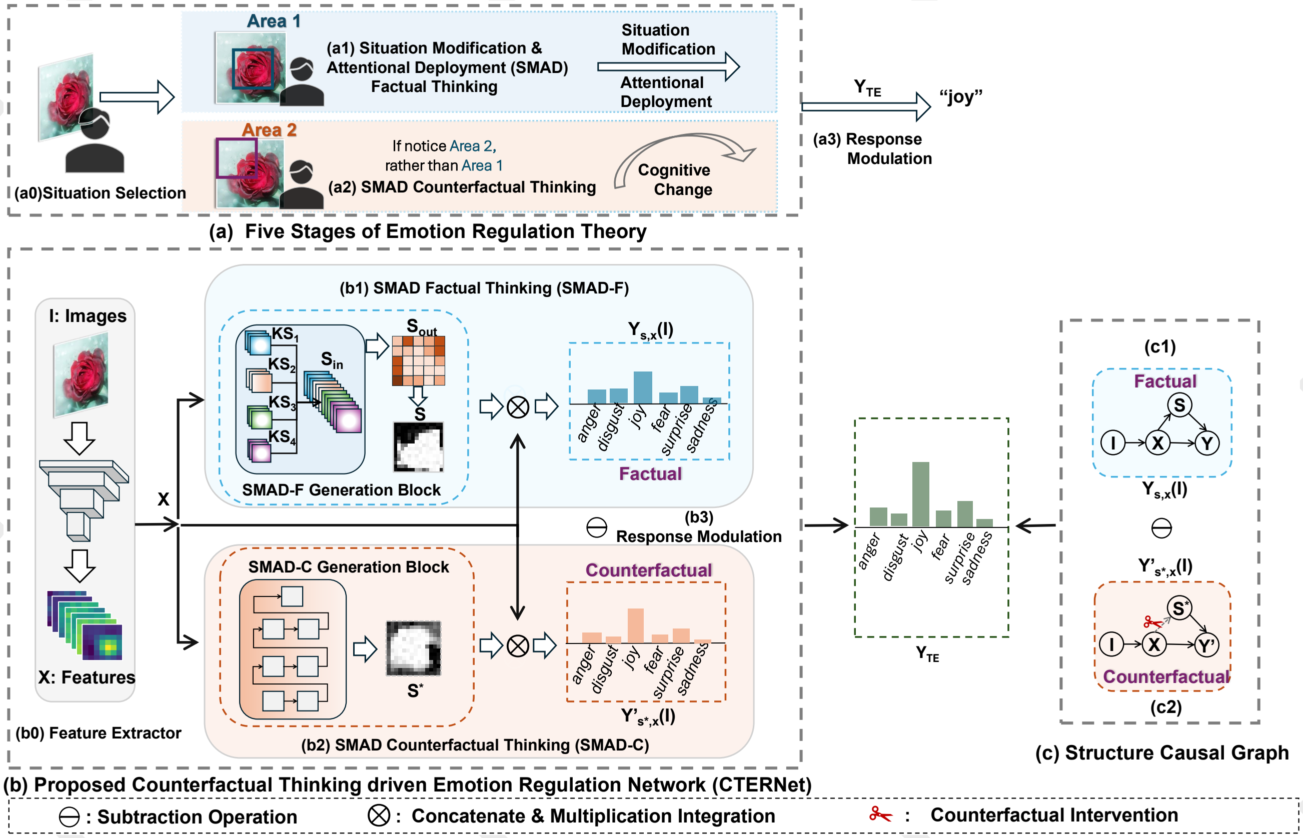
Area 1
(a1) Situation Modification & Attentional Deployment (SMAD) Factual Thinking
Situation Modification
Attentional Deployment
Area 2
If notice Area 2, rather than Area 1
(a2) SMAD Counterfactual Thinking
Cognitive Change
(a0)Situation Selection
Y_{TE}
"joy"
(a3) Response Modulation
(a) Five Stages of Emotion Regulation Theory
I: Images
X: Features
(b0) Feature Extractor
X
(b1) SMAD Factual Thinking (SMAD-F)
KS_1
KS_2
KS_3
KS_4
S_{in}
S_{out}
S
SMAD-F Generation Block
$Y_{s,x}(I)$
anger
disgust
joy
fear
surprise
sadness
Factual
(b3) Response Modulation
SMAD-C Generation Block
S^*
Counterfactual
$Y'_{s^*,x}(I)$
(b2) SMAD Counterfactual Thinking (SMAD-C)
(b) Proposed Counterfactual Thinking driven Emotion Regulation Network (CTERNet)
Y_{TE}
(c1)
Factual
S
I
X
Y
$Y_{s,x}(I)$
⊖
$Y'_{s^*,x}(I)$
S'
Y'
Counterfactual
(c2)
(c) Structure Causal Graph
⊖ : Subtraction Operation
⊗ : Concatenate & Multiplication Integration
✂ : Counterfactual Intervention

Figure 1: The Framework of the proposed CTERNet. (a) Five Stages of Emotion Regulation Theory: These stages include situation selection, situation modification, attentional deployment, cognitive change, and response modulation. (b) The Proposed CTERNet: The Proposed CTERNet simulates the four stages of Emotion Regulation Theory. The final output of the CTERNet system, Y_{TE}, can be obtained through response modulation (b3), which involves adding the factual prediction (from SMAD-F) and its counterfactual prediction (from SMAD-C). (c) Structural Causal Graph: We built this Structural Causal Graph, which corresponds to the proposed CTERNet (Figure 1 (b)).

the proposed method in this proposal. According to Emotion Regulation [Gross, 2002; Pearl and Mackenzie, 2018; Rao *et al.*, 2021; Pearl, 2022], as shown in Figure 1(a), there are **five stages** including ***situation selection*** (Figure 1(a0)), ***situation modification*** and ***attentional deployment*** (Figure 1(a1)), ***cognitive change*** (Figure 1(a2)), and ***response modulation*** (Figure 1(a3)). To better simulate the Emotion Regulation process (illustrated in Figure 1(a)) within Emotion Regulation Theory, we propose a novel counterfactual thinking driven emotion regulation network (CTERNet) for identifying emotional or sentiment categories, such as positive, negative, and their respective subcategories. The outputs (binary or multiclass) depend on the corresponding datasets available and the requirements.

The proposed CTERNet simulates the five stages of Emotion Regulation Theory. **Feature extractor**, as shown in block (b0) in Figure 1(b) simulates ***situation selection*** as shown in Figure 1(a0). The pairs: **SMAD-F** (Figure 1 b1)) and **SMAD-C** (Figure 1(b2)) are the core components of the CTERNet, which simulate various emotional arousal processes, such as ***SMAD factual thinking*** (a1) and ***SMAD counterfactual thinking*** (a2) in Emotion Regulation Theory. The effects of **SMAD-F** (factual thinking), depicted in block (b1), are compared with those of **SMAD-C** (counterfactual thinking), illustrated in block (Figure 1(b2)), before conducting ***response modulation*** (Figure 1(a3)). The final output of the CTERNet, Y_{TE}, can be obtained through ***response modulation*** (a3), which involves adding the factual prediction (from **SMAD-F**) and its counterfactual prediction (from **SMAD-C**).

We built this Structural Causal Graph [Pearl and others, 2000; Pearl, 2014], which corresponds to the proposed CTERNet (Figure 1(c)). By introducing such a structural causal graph, we aim to simplify the proposed method to enable the novice to understand the relationship between different variables, I, X, S, S^*, $Y_{s,x}(I)$ and $Y'_{s^*,x}(I)$. I represent input images, X represents the extracted features, S represents pseudo sentiment maps obtained by module **SMAD-F** (Figure 1(b1)) (also illustrated as (Figure 1(c1)), S^* represents pseudo sentiment maps obtained by module **SMAD-C** (Figure 1(b2)) (also illustrated as Figure 1(c2)), $Y_{s,x}(I)$ represents the final emotion response prediction output by module **SMAD-F** (Figure 1(b1)) (also illustrated as Figure 1(c1)), and $Y'_{s^*,x}$ represents the final emotion response prediction output by module **SMAD-C** (Figure 1(b2)) (also illustrated as Figure 1(c2)). The link $I \to X \to S$ represents the extraction of features X from I through the corresponding encoder in the ISR model, followed by the generation of pseudo sentiment maps S. The link $I \rightarrow (X, S) \rightarrow Y$ signifies the final emotion classification prediction obtained by integrating features X and sentiment regions S, capturing the causal effect of X on the model's prediction Y.

3.1 Feature Extractor (b0)

Feature extractor (Figure 1(b0)) simulates the ***situation selection*** (Figure 1(a0)) of the Emotion Regulation Theory. By processing and analyzing the visual information, individuals can assess the emotional impact of the current context and make appropriate choices [Gross and Feldman Barrett, 2011; Gross, 2015]. In Emotion Regulation Theory, ***situation selection*** (Figure 1(a0)) refers to the process of regulating emotions by choosing to engage with or avoid certain emotional situations (such as Area 1, Area 2, in Figure 1(a)). Therefore, $I \rightarrow X$ as a feature extraction process (Figure 1(b0)) naturally aligns with the ***situation selection*** stage in Emotion Regulation Theory. By selecting specific emotion scenarios (i.e., avoiding certain scenarios) [Gross and John, 2003], humans can achieve the goal of emotional regulation. To simulate the different emotional scenarios directed by various visual inputs, we utilize Res2Net-101 [Gao *et al.*, 2019] as the backbone for the $I \rightarrow X$ process.

3.2 Situation Modification and Attentional Deployment Factual Thinking (SMAD-F) (b1)

After selecting a scenario through ***situation selection*** (Figure 1(a0)), the corresponding module for the proposed CTERNet is **Feature Extractor** shown in Figure 1(b0). This scenario is then modified to obtain and modify the emotional impact, which is known as ***situation modification and attentional deployment*** (SMAD) factual thinking (Figure 1(a1)). For example, suppose we choose the emotional scenario in Area 1 from Figure 1(a). In that case, we can generate different situations (e.g., different features including low-level features and high-level features) based on the Area 1. The corresponding module for the proposed CTERNet is **SMAD-F** as shown in Figure 1(b1). As shown in block (b1) of Figure 1, this preliminary control of emotional scenarios is simulated using different multi-scale convolutional filters, similar to the way situation modification work in the human emotion perception process. The process can be represented by the following equation.

$$S_{in} = \sum_{g=1}^{G} Conv(KS_g \times KS_g, X_g) \tag{1}$$

where $g \in G$ represents dividing X into g groups based on the number of channels. For each group, we use convolutional filters with a kernel size (*KS*), $KS = 2g+1$ for emotion perception. Following the practice in [Zhang *et al.*, 2022a], we set g to 4.

The situations in ***situation modification*** have different aspects [Gross, 2002], therefore, we concatenate the outputs from different emotional scenario perceptions along the channel dimension as the input of ***attentional deployment***. Through ***attentional deployment***, we can select the specific situation aspect to focus on. To mimic the ***attentional deployment*** stage, we reshape the feature map S_{in} to a size of $C \times (W \times H)$. We then multiply the reshaped feature map by the transpose of S_{in}, and after normalization, we obtain a feature map S_{out} of size $C \times C$. Next, S_{out} is multiplied by the transpose of S_{in}, and the result is reshaped back to $C \times W \times H$ with a learnable parameter φ, which is shown in Equation below.

$$S_{out} = \sum_{j=1}^{C} (\varphi \sum_{i=1}^{C} (S_{ij} \times (S_{in})_i) + (S_{in})_j) \tag{2}$$

where S_{ij} represents the dependency between the i-th and j-th channels, defined as $\frac{exp((S_{in})_i \cdot (S_{in})_j)}{\sum_{i=1}^{C} exp((S_{in})_i \cdot (S_{in})_j)}$. This process simulates the operation of ***attentional deployment***, helping CTERNet focus on the most discriminative and important regions of the image.

3.3 Situation Modification and Attentional Deployment Counterfactual Thinking (SMAD-C) (b2)

SMAD-C component (in Figure 1(b2)), also known as ***cognitive change*** as shown in Figure 1(a2), in Emotion Regulation Theory, involves selecting potential interpretations of the significance of emotional events. Based on the counterfactual thinking strategy in Emotion Regulation Theory [Parikh *et al.*, 2022], we consider: *if the model sees other potential affective regions (e.g., Area 2), what would the prediction be?* By learning from both the affective regions identified by the **SMAD-F** module (factual) (see Figure 1(b1)) and other potential affective regions identified by the **SMAD-C** module (counterfactual) (see Figure 1(b2)), we enhance the effectiveness of the ISR, which consider the **SMAD-F** only rather than the pairs (both **SMAD-F** and **SMAD-C**). This approach helps extract more robust emotional features. The causal relationships in the proposed method can be expressed as following:

$$Y_{s,x}(I) = Y(S = s, X = x \mid I) \tag{3}$$

$Y_{s,x}(I)$ contains confusing emotional guidance information. By assuming different variables, counterfactual interventions can be achieved [Pearl and Mackenzie, 2018; Pearl, 2022]. By introducing a structural causal graph, we can directly manipulate the values of several variables to analyze causal relationships and observe their effects [VanderWeele, 2015; Rao *et al.*, 2021]. In the ISR tasks, we conduct counterfactual intervention $do(S = s^*)$ as potential conditions [VanderWeele, 2015] for the affective region (s^*) (counterfactual) to replace the affective region map (S) (factual), as illustrated in the potential generation for **SMAD-C** in block (b3) of Figure 1(b), and in the block (c2) of Figure 1(c). In practice, we use random attention allocation, uniform attention allocation, and reverse attention allocation as counterfactuals [Rao *et al.*, 2021]. Subsequently, ***response modulation*** affects human emotional perception outcomes after the emotional response tendencies have already been elicited. To simulate this stage, we leverage the cross-spatial pooling strategy, dividing the input feature map's channels into multiple categories, which is shown as Equation (4). For each classification category cl,

Methods	FI	Emotion6
SentiBank [Borth *et al.*, 2013]	49.23	35.24
DeepSentiBank [Chen *et al.*, 2014]	51.52	42.53
MAP [He *et al.*, 2019]	68.13	60.47
WSCNet [She *et al.*, 2019]	70.07	58.25
MSRCA [Zhang *et al.*, 2022b]	72.60	55.60
Yang *et al.* [Yang *et al.*, 2023b]	71.13	**60.54**
DCNet [Zhang *et al.*, 2023]	71.65	59.60
CausVSR [Zhang *et al.*, 2024]	**72.57**	59.82
CTERNet	**72.71**	**60.49**

Table 1: Accuracy comparison on multi-class datasets. The highest accuracy model is indicated in bold black, and the second-best is indicated in bold blue.

the feature maps $S_{\text{out},i} \quad (i = 1, 2, \ldots, l)$ have all pixels in each channel averaged by the global average pooling operation (G_{GAP}), thereby achieving dimensionality reduction and information aggregation of the input feature map to produce the final output $S = f_{cs}(S_{out})$.

$$f_{cs} = \sum_{cl=1}^{CL}\left(\frac{1}{l}\sum_{i=1}^{l} G_{GAP}(S_{\text{out},i})\right)\left(\frac{1}{l}\sum_{i=1}^{l} f_{cl,i}\right) \tag{4}$$

As shown in the structural causal graph, we combine the pseudo sentiment maps S and X for the final prediction P:

$$P = f_{fully}(G_{\text{GAP}}(t(S, X))) \tag{5}$$

where $t(\cdot)$ represents the concatenate operation, and f_{fully} represents the fully connected layer, used to calculate the prediction scores for different emotional categories. The effect of learned Emotion Regulation on actual emotion prediction can be represented by the addition or sum of the factual thinking prediction, Y (from **SMAD-F**) and its counterfactual prediction Y' (from **SMAD-C**).

3.4 Response Modulation (b3)

The total effect (TE) [Pearl, 2014; Yang *et al.*, 2024], Y_{TE}, which is the **final emotion response** of the CTERNet can be obtained using the following equation:

$$Y_{TE} = \mathbb{E}[Y(S = s, X = x|I) - Y(S = s^{*}, X = x|I)] \tag{6}$$

We utilize Y_{TE} as a supervision signal to guide the generation of affective regions. In combination with the original losses in region-based ISR, the total loss function can be expressed as shown in Equation (7).

$$\mathcal{L}_{\text{Total}} = \mathcal{L}(S, Y_{label}) + \mathcal{L}(Y_{TE}, Y_{label}) \tag{7}$$

where $\mathcal{L}(\cdot)$ represents cross-entropy loss, and Y_{label} denotes the ground truth. The loss $\mathcal{L}(S, Y_{label})$ calculates the difference between the generated affective regions S and the true emotion labels.

4 Experiments

4.1 Datasets and the Evaluation Metric

The experiments were conducted on datasets of various scales, including Flickr and Instagram (FI) [You *et al.*, 2016], Emotion6 [Peng *et al.*, 2016], ArtPhoto [Machajdik and Hanbury, 2010; Yang *et al.*, 2018b], and Twitter II [Borth *et al.*, 2013; Zhang *et al.*, 2024].

Like all other ISR works, we utilize classification accuracy for evaluation.

4.2 Implementation Details

In terms of model architecture, we utilized a Res2Net-101 [Gao *et al.*, 2019] pre-trained on the ImageNet dataset, implemented using the PyTorch framework, to parameterize the feature extraction network. The output of the last linear layer was replaced to generate the task-specific number of neurons for ISR prediction. For dataset processing, we set the input image size to 448×448. Data augmentation on the training set included random cropping, random horizontal flipping, and image normalization. For the test set, we applied center cropping and image normalization. We used the SGD optimizer with a learning rate of 0.0001 and a momentum of 0.9. The learning rate was decayed by a factor of 0.1 every 10 epochs. All experiments were implemented on NVIDIA Geforce RTX 2080 Ti GPUs.

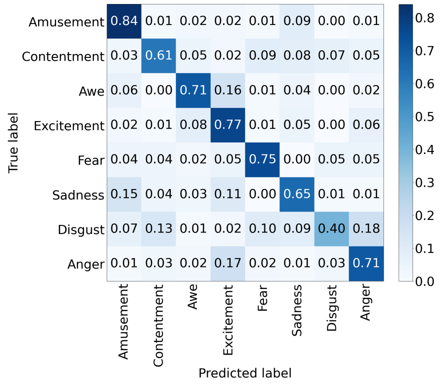

Figure 2: Confusion matrix for the FI dataset with eight emotions.

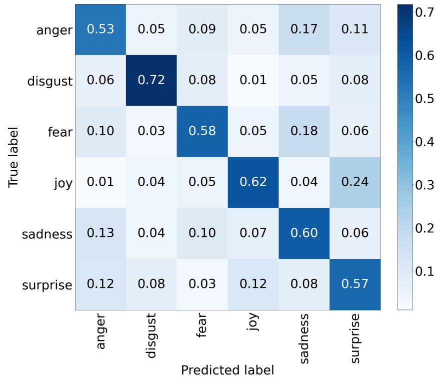

Figure 3: Confusion matrix for the Emotion6 dataset with six emotions.

Methods	FI	Emotion6	ArtPhoto	Twitter II
SentiBank [Borth *et al.*, 2013]	56.47	-	67.74	65.93
DeepSentiBank [Chen *et al.*, 2014]	64.39	-	68.73	70.23
PCNN [You *et al.*, 2015]	75.34	-	70.84	77.68
Zhu *et al.* [Zhu *et al.*, 2017]	84.26	-	75.5	-
Panda *et al.* [Panda *et al.*, 2018]	84.81	77.72	-	-
Yang *et al.* [Yang *et al.*, 2018b]	86.35	81.26	74.8	80.48
WSCNet [She *et al.*, 2019]	-	-	-	81.35
MSRCA [Zhang *et al.*, 2022b]	87.40	83.00	-	-
DCNet [Zhang *et al.*, 2023]	90.58	83.26	**79.13**	82.50
CausVSR [Zhang *et al.*, 2024]	**90.93**	**83.30**	78.98	**82.86**
CTERNet	**91.01**	**83.67**	**79.27**	**83.33**

Table 2: Accuracy comparison on binary-class datasets. The highest accuracy model is indicated in bold black, and the second-best is indicated in bold blue.

Vanilla Model	SMAD-F	SMAD-C	FI	Emotion6	ArtPhoto	TwitterII
✓			71.45	58.95	76.72	82.07
✓	✓		72.31	60.06	78.84	82.96
✓	✓	✓	**72.71**	**60.49**	**79.27**	**83.33**

Table 3: Impacts of the CTERNet Framework Structure.

4.3 Comparisons on Multi-class Datasets

We compared the CTERNet with classic and state-of-the-art ISR models on multi-class datasets, as shown in Table 1. WSCNet is a classic weakly-supervised ISR method that has inspired much work in the field [She *et al.*, 2019]. Yang's model uses a semantic embedding space to explore image-emotion relationships [Yang *et al.*, 2023b]. MSRCA introduces a multi-level sentiment region correlation analysis module and leverages a Transformer encoder for rich emotion interaction [Zhang *et al.*, 2022b]. CausVSR models ISR based on Emotion Regulation Theory, using a front-door adjustment to reduce contextual confounding [Zhang *et al.*, 2024]. Compared to these models, CTERNet achieved the highest accuracy on the FI dataset (72.71%), outperforming MSRCA and CausVSR by 0.11% and 0.14%, respectively. On the Emotion6 dataset, it performed well (60.49%), slightly behind Yang's model (60.54%), which employs multi-task learning and a fusion strategy to better utilize limited samples in smaller datasets. In contrast, CTERNet emphasizes causal relationships through Emotion Regulation strategies and counterfactual thinking, making it more effective on complex, larger-scale datasets by leveraging more information for causal analysis.

Furthermore, we present the confusion matrices for the FI and Emotion6 multi-class datasets in Figure 2 and Figure 3 to analyze the performance of CTERNet. The matrices show that the model performs well in recognizing most emotions on both datasets. In the FI dataset, significant misclassifications occur between similar emotions, such as Anger and Disgust, and Excitement and Awe, with Disgust being the hardest to recognize. This may be due to the greater diversity in how Disgust is expressed across contexts, making it more challenging for the model. In the Emotion6 dataset, overall performance is strong, but Anger shows lower accuracy, with misclassifications primarily between Anger and Sadness. This could be due to visual similarities, as training samples for both emotions often feature black-and-white images with dark tones.

4.4 Comparisons on Binary-class Datasets

We also conducted experiments focused on sentiment polarity analysis. Besides the ArtPhoto and Twitter II datasets, which inherently include sentiment polarity classification, we converted the original labels of Emotion6 and FI into two polarities: positive and negative. As shown in Table 2, CTERNet demonstrated excellent performance across all polarity datasets, showcasing its robust capabilities in polarity sentiment classification tasks. In comparison, while other methods performed well on certain datasets, none surpassed the overall performance of CTERNet.

4.5 Ablation Studies

By systematically conducting ablation studies on components of the CTERNet model that simulate Emotion Regulation Theory, we observe the impact of each module on the overall performance. These studies are conducted on the multi-class datasets and the binary classification datasets.

Impacts of the CTERNet Framework Structure. Table 3 presents the ablation study results for each regulation process within the CTERNet framework: i) The significant performance improvement when incorporating the **SMAD-F** module into the vanilla model indicates that our simulation of situation modification and attention allocation provides valuable emotional semantics, effectively aiding CTERNet in continuously identifying and adjusting affective regions. ii) Adding the **SMAD-C** module, based on counterfactual thinking, further enhances the results. Comparing facts with counterfactual assumptions helps reduce bias in the affective region gen-

Potential Generation Used in SMAD-C	FI	Emotion6	ArtPhoto	TwitterII
Vanilla Model	72.31	60.06	78.84	82.96
Uniform Attention Allocation	72.58	60.24	79.12	83.20
Reverse Attention Allocation	72.36	60.11	78.88	82.84
Random Attention Allocation	**72.71**	**60.49**	**79.27**	**83.33**

Table 4: Impacts of Leveraging Different Potential Generation in SMAD-C.

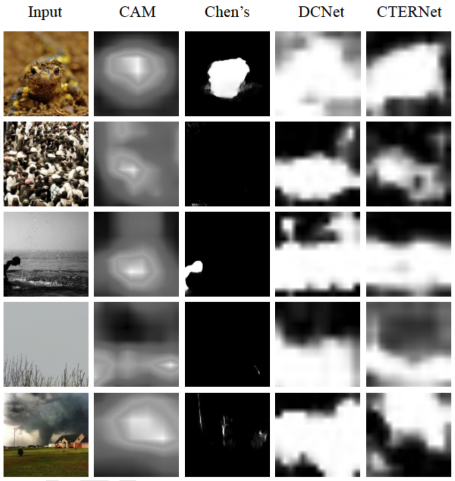

Figure 4: Comparison of visualization of affective regions on Emotion6 dataset. The 1st column shows the original input images, the 2nd column displays the class activation maps generated using [Zhou *et al.*, 2016], the 3rd column presents the saliency maps produced by [Chen *et al.*, 2020], the 4th column illustrates the affective region maps generated by DCNet [Zhang *et al.*, 2023], and the 5th column depicts the results of CTERNet.

eration process. Additionally, this demonstrates that the total causal effect is essential for improving de-biasing gains.

Impacts of the Potential Generation in SMAD-C. We applied three different strategies [Rao *et al.*, 2021], including random attention allocation, uniform attention allocation, and reverse attention allocation, to the attention allocation phase of Emotion Regulation as counterfactual scenarios. The results, shown in Table 4, indicate that random attention allocation outperforms the other strategies. However, the performance of reverse attention allocation is suboptimal, even falling below the vanilla model's performance on the ArtPhoto and TwitterII datasets. We analyzed that the random attention allocation strategy can provide diverse attention patterns in different contexts, capturing more emotional information and contextual variations. This diversity aligns better with natural human attention, helping the model to more comprehensively understand and regulate emotions. In contrast, the reverse attention allocation strategy may disrupt the organization and expression of emotional features, leading to reduced focus on key emotional characteristics and thereby weakening the model's ability to regulate emotions.

4.6 Visualization

We conducted a visual analysis on Emotion6 using different affective region generation methods. CAM model [Zhou *et al.*, 2016] can highlight the regions of the image that the model focuses on when making classification decisions. Similarly, saliency object detection method of Chen et al. [Chen *et al.*, 2020] can identify the most prominent areas in an image, which are the parts most likely to attract human attention first. However, the ISR task differs from ordinary image classification tasks. ISR not only focuses on objects in the image but also involves more abstract and complex features that guide its decisions. As shown in Figure 4, CTERNet exhibits more precise targeting when the image content is simple and features a single main subject (e.g., the first and third rows). In the first row, while all tested models detect relevant affective regions, CTERNet considers both the foreground toad and the background environment. Similarly, in the third row, CTERNet captures both the person and the seaside environment, unlike the saliency-based method, which focuses only on the foreground object. Additionally, CTERNet avoids the interference in the upper affective region generated by DCNet, providing more accurate delineation of affective regions. For more complex images with multiple targets or lower contrast (e.g., the second, fourth, and fifth rows), the saliency-based method struggles to identify significant regions accurately. In contrast, CTERNet not only detects clear affective regions but also identifies finer details compared to other models. This highlights CTERNet's superior ability to handle complex scenarios, making it highly effective for ISR tasks.

5 Conclusion

This paper simulates the Emotion Regulation Theory by modeling the entire process of image sentiment recognition based on human causality-driven mechanisms. Specifically, we focus on counterfactual thinking as a crucial strategy in cognitive change. By leveraging counterfactual causal relationships, we jointly model the stages of situation modification, attentional deployment, and cognitive change in Emotion Regulation Theory. This approach allows us to learn both affective regions (factual) and other potential affective regions (counterfactual), simulating humans' diverse emotional arousal processes in different situations. During the response modulation phase, we obtain emotional prediction results based on interpreting various emotional scenarios, comparing the effects of factual and counterfactual information on the final emotional prediction. This process also quantifies the quality of the affective regions. Extensive experiments on public image sentiment recognition datasets demonstrate the performance of the proposed model.

Acknowledgments

This research was supported by National Natural Science Foundation of China under Grant Nos. 61871186, 61771322.

References

[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.

[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.

[Cludius *et al.*, 2020] Barbara Cludius, Douglas Mennin, and Thomas Ehring. Emotion regulation as a transdiagnostic process. *Emotion*, 20(1):37, 2020.

[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.

[Cui *et al.*, 2023] Jingfeng Cui, Zhaoxia Wang, Seng-Beng Ho, and Erik Cambria. Survey on sentiment analysis: evolution of research methods and topics. *Artificial Intelligence Review*, 56(8):8469–8510, 2023.

[Fan *et al.*, 2017] Shaojing Fan, Ming Jiang, Zhiqi Shen, Bryan L Koenig, Mohan S Kankanhalli, and Qi Zhao. The role of visual attention in sentiment prediction. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 217–225, 2017.

[Feldner *et al.*, 2003] Matthew T Feldner, Michael J Zvolensky, Georg H Eifert, and Adam P Spira. Emotional avoidance: An experimental test of individual differences and response suppression using biological challenge. *Behaviour research and therapy*, 41(4):403–411, 2003.

[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.

[Gross and Feldman Barrett, 2011] James J Gross and Lisa Feldman Barrett. Emotion generation and emotion regulation: One or two depends on your point of view. *Emotion review*, 3(1):8–16, 2011.

[Gross and John, 2003] James J Gross and Oliver P John. Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *Journal of personality and social psychology*, 85(2):348, 2003.

[Gross, 2002] James J Gross. Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology*, 39(3):281–291, 2002.

[Gross, 2015] James J Gross. Emotion regulation: Current status and future prospects. *Psychological inquiry*, 26(1):1–26, 2015.

[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.

[Li *et al.*, 2019] Zuhe Li, Yangyu Fan, Bin Jiang, Tao Lei, and Weihua Liu. A survey on sentiment analysis and opinion mining for social multimedia. *Multimedia Tools and Applications*, 78:6939–6967, 2019.

[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.

[McRae *et al.*, 2012] Kateri McRae, Bethany Ciesielski, and James J Gross. Unpacking cognitive reappraisal: goals, tactics, and outcomes. *Emotion*, 12(2):250, 2012.

[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.

[Parikh *et al.*, 2022] Natasha Parikh, Felipe De Brigard, and Kevin S LaBar. The efficacy of downward counterfactual thinking for regulating emotional memories in anxious individuals. *Frontiers in psychology*, 12:712066, 2022.

[Pearl and Mackenzie, 2018] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

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

[Pearl, 2014] Judea Pearl. Interpretation and identification of causal mediation. *Psychological methods*, 19(4):459, 2014.

[Pearl, 2022] Judea Pearl. Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl*, pages 373–392. 2022.

[Peng *et al.*, 2016] Kuan-Chuan Peng, Amir Sadovnik, Andrew Gallagher, and Tsuhan Chen. Where do emotions come from? predicting the emotion stimuli map. In *2016 IEEE international conference on image processing (ICIP)*, pages 614–618. IEEE, 2016.

[Rao *et al.*, 2021] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1025–1034, 2021.

[Rye *et al.*, 2008] Mark S Rye, Melissa B Cahoon, Rahan S Ali, and Tarika Daftary. Development and validation of the counterfactual thinking for negative events scale. *Journal of personality assessment*, 90(3):261–269, 2008.

[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.

[Sirois *et al.*, 2010] Fuschia M Sirois, Jennifer Monforton, and Melissa Simpson. "if only i had done better": Perfectionism and the functionality of counterfactual thinking. *Personality and social psychology bulletin*, 36(12):1675–1692, 2010.

[Sun *et al.*, 2022] Teng Sun, Wenjie Wang, Liqaing Jing, Yiran Cui, Xuemeng Song, and Liqiang Nie. Counterfactual reasoning for out-of-distribution multimodal sentiment analysis. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 15–23, 2022.

[Sweller, 2020] John Sweller. Cognitive load theory and educational technology. *Educational technology research and development*, 68(1):1–16, 2020.

[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.

[Theodorou *et al.*, 2023] Annalisa Theodorou, Giuseppina Spano, Gregory N Bratman, Kevin Monneron, Giovanni Sanesi, Giuseppe Carrus, Claudio Imperatori, and Angelo Panno. Emotion regulation and virtual nature: cognitive reappraisal as an individual-level moderator for impacts on subjective vitality. *Scientific Reports*, 13(1):5028, 2023.

[Thompson, 1994] Ross A Thompson. Emotion regulation: A theme in search of definition. *Monographs of the society for research in child development*, pages 25–52, 1994.

[VanderWeele, 2015] Tyler VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015.

[Wang *et al.*, 2020] Zhaoxia Wang, Seng-Beng Ho, and Erik Cambria. A review of emotion sensing: categorization models and algorithms. *Multimedia Tools and Applications*, 79:35553–35582, 2020.

[Yang *et al.*, 2018a] Jufeng Yang, Dongyu She, Yu-Kun Lai, and Ming-Hsuan Yang. Retrieving and classifying affective images via deep metric learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[Yang *et al.*, 2018b] 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.*, 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.

[Yang *et al.*, 2024] Dingkang Yang, Kun Yang, Mingcheng Li, Shunli Wang, Shuaibing Wang, and Lihua Zhang. Robust emotion recognition in context debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12447–12457, 2024.

[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 AAAI conference on Artificial Intelligence*, volume 29, 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.

[You *et al.*, 2017] Quanzeng You, Hailin Jin, and Jiebo Luo. Visual sentiment analysis by attending on local image regions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[Zhang *et al.*, 2022a] Hu Zhang, Keke Zu, Jian Lu, Yuru Zou, and Deyu Meng. Epsanet: An efficient pyramid squeeze attention block on convolutional neural network. In *Proceedings of the asian conference on computer vision*, pages 1161–1177, 2022.

[Zhang *et al.*, 2022b] 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.

[Zhang *et al.*, 2024] Xinyue Zhang, Zhaoxia Wang, Hailing Wang, Jing Xiang, Chunwei Wu, and Guitao Cao. Causvsr: Causality inspired visual sentiment recognition. IJCAI, 2024.

[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] Xinge Zhu, Liang Li, Weigang Zhang, Tianrong Rao, Min Xu, Qingming Huang, and Dong Xu. Dependency exploitation: A unified cnn-rnn approach for visual emotion recognition. In *IJCAI*, pages 3595–3601, 2017.