Context De-confounded Emotion Recognition

Dingkang Yang[1,2] Zhaoyu Chen[1] Yuzheng Wang[1] Shunli Wang[1] Mingcheng Li[1] Siao Liu[1]
Xiao Zhao[1] Shuai Huang[1] Zhiyan Dong[1] Peng Zhai[1] Lihua Zhang[1,2,3,4§]
[1]Academy for Engineering and Technology, Fudan University [2]Institute of Meta-Medical, IPASS
[3]Jilin Provincial Key Laboratory of Intelligence Science and Engineering, Changchun, China
[4]Engineering Research Center of AI and Robotics, Ministry of Education, Shanghai, China
{dkyang20,lihuazhang}@fudan.edu.cn

Abstract

Context-Aware Emotion Recognition (CAER) is a crucial and challenging task that aims to perceive the emotional states of the target person with contextual information. Recent approaches invariably focus on designing sophisticated architectures or mechanisms to extract seemingly meaningful representations from subjects and contexts. However, a long-overlooked issue is that a context bias in existing datasets leads to a significantly unbalanced distribution of emotional states among different context scenarios. Concretely, the harmful bias is a confounder that misleads existing models to learn spurious correlations based on conventional likelihood estimation, significantly limiting the models' performance. To tackle the issue, this paper provides a causality-based perspective to disentangle the models from the impact of such bias, and formulate the causalities among variables in the CAER task via a tailored causal graph. Then, we propose a Contextual Causal Intervention Module (CCIM) based on the backdoor adjustment to de-confound the confounder and exploit the true causal effect for model training. CCIM is plug-in and model-agnostic, which improves diverse state-of-the-art approaches by considerable margins. Extensive experiments on three benchmark datasets demonstrate the effectiveness of our CCIM and the significance of causal insight.

1. Introduction

As an essential technology for understanding human intentions, emotion recognition has attracted significant attention in various fields such as human-computer interaction [1], medical monitoring [28], and education [40]. Previous works have focused on extracting multimodal emotion cues from human subjects, including facial expressions [9, 10, 49], acoustic behaviors [2, 50, 52], and body postures [25, 53], benefiting from advances in deep learning algorithms [6, 7, 21, 26, 27, 43, 44, 46, 47, 54, 55, 59]. Despite the impressive improvements achieved by subject-centered approaches, their performance is limited by natural and unconstrained environments. Several examples in Figure 1 (left) show typical situations on a visual level. Instead of well-designed visual contents, multimodal representations of subjects in wild-collected images are usually indistinguishable (*e.g.*, ambiguous faces or gestures), which forces us to exploit complementary factors around the subject that potentially reflect emotions.

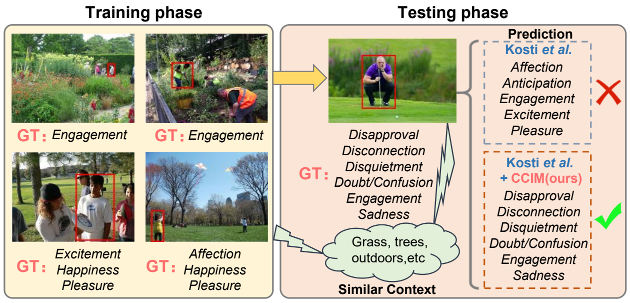
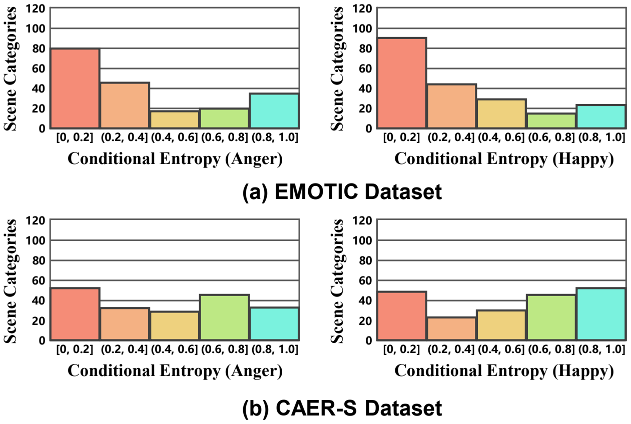

Training phase
GT: Engagement
GT: Engagement
GT: Excitement Happiness Pleasure
GT: Affection Happiness Pleasure
Testing phase
GT: Disapproval Disconnection Disquietment Doubt/Confusion Engagement Sadness
Grass, trees, outdoors,etc
Similar Context
Prediction
Kosti et al.
Affection Anticipation Engagement Excitement Pleasure
✗
Kosti et al. + CCIM(ours)
Disapproval Disconnection Disquietment Doubt/Confusion Engagement Sadness
✓

Figure 1. Illustration of the context bias in the CAER task. GT means the ground truth. Most images contain similar contexts in the training data with positive emotion categories. In this case, the model learns the spurious correlation between specific contexts and emotion categories and gives wrong results. Thanks to CCIM, the simple baseline [19] achieves more accurate predictions.

Inspired by psychological study [3], recent works [19, 22, 23, 29, 56] have suggested that contextual information contributes to effective emotion cues for Context-Aware Emotion Recognition (CAER). The contexts are considered to include the place category, the place attributes, the objects, or the actions of others around the subject [20]. The majority of such research typically follows a common pipeline: (1) Obtaining the unimodal/multimodal representations of the recognized subject; (2) Building diverse contexts and extracting emotion-related representations; (3) Designing

[§]Corresponding Author.



(a) EMOTIC Dataset

(b) CAER-S Dataset

Figure 2. We show a toy experiment on the EMOTIC [20] and CAER-S [22] datasets for scene categories of angry and happy emotions. More scene categories with normalized zero-conditional entropy reveal a strong presence of the context bias.

fusion strategies to combine these features for emotion label predictions. Although existing methods have improved modestly through complex module stacking [12,23,51] and tricks [16, 29], they invariably suffer from a context bias of the datasets, which has long been overlooked. Recalling the process of generating CAER datasets, different annotators were asked to label each image according to what they subjectively thought people in the images with diverse contexts were feeling [20]. This protocol makes the preference of annotators inevitably affect the distribution of emotion categories across contexts, thereby leading to the context bias. Figure 1 illustrates how such bias confounds the predictions. Intrigued, most of the images in training data contain vegetated scenes with positive emotion categories, while negative emotions in similar contexts are almost nonexistent. Therefore, the baseline [19] is potentially misled into learning the spurious dependencies between context-specific features and label semantics. When given test images with similar contexts but negative emotion categories, the model inevitably infers the wrong emotional states.

More intrigued, a toy experiment is performed to verify the strong bias in CAER datasets. This test aims to observe how well emotions correlate with contexts (*e.g.*, scene categories). Specifically, we employ the ResNet-152 [15] pre-trained on Places365 [58] to predict scene categories from images with three common emotion categories (*i.e.*, "anger", "happy", and "fear") across two datasets. The top 200 most frequent scenes from each emotion category are selected, and the normalized conditional entropy of each scene category across the positive and negative set of a specific emotion is computed [30]. While analyzing correlations between scene contexts and emotion categories in Figure 2 (*e.g.*, "anger" and "happy"), we find that more scene categories with the zero conditional entropy are most likely to suggest the significant context bias in the datasets, as it shows the presence of these scenes only in the positive or negative set of emotions. Concretely, for the EMOTIC dataset [20], about 40% of scene categories for anger have zero conditional entropy while about 45% of categories for happy (*i.e.*, happiness) have zero conditional entropy. As an intuitive example, most party-related scene contexts are present in the samples with the happy category and almost non-existent in the negative categories. *These observations confirm the severe context bias in CAER datasets, leading to distribution gaps in emotion categories across contexts and uneven visual representations.*

Motivated by the above observation, we attempt to embrace causal inference [31] to reveal the culprit that poisons the CAER models, rather than focusing on beating them. As a revolutionary scientific paradigm that facilitates models toward unbiased prediction, the most important challenge in applying classical causal inference to the modern CAER task is how to reasonably depict true causal effects and identify the task-specific dataset bias. To this end, this paper attempts to address the challenge and rescue the bias-ridden models by drawing on human instincts, *i.e.*, looking for the causality behind any association. Specifically, we present a causality-based bias mitigation strategy. We first formulate the procedure of the CAER task via a proposed causal graph. In this case, the harmful **context bias** in datasets is essentially an unintended **confounder** that misleads the models to learn the spurious correlation between similar contexts and specific emotion semantics. From Figure 3, we disentangle the causalities among the input images $\boldsymbol{X}$, subject features $\boldsymbol{S}$, context features $\boldsymbol{C}$, confounder $\boldsymbol{Z}$, and predictions $\boldsymbol{Y}$. Then, we propose a simple yet effective Contextual Causal Intervention Module (CCIM) to achieve context-deconfounded training and use the *do*-calculus $P(\boldsymbol{Y}|do(\boldsymbol{X}))$ to calculate the true causal effect, which is fundamentally different from the conventional likelihood $P(\boldsymbol{Y}|\boldsymbol{X})$. CCIM is plug-in and model-agnostic, with the backdoor adjustment [14] to de-confound the confounder and eliminate the impact of the context bias. We comprehensively evaluate the effectiveness and superiority of CCIM on three standard and biased CAER datasets. Numerous experiments and analyses demonstrate that CCIM can significantly and consistently improve existing baselines, achieving a new state-of-the-art (SOTA).

The main contributions can be summarized as follows:

- To our best knowledge, we are the first to investigate the adverse context bias of the datasets in the CAER task from the causal inference perspective and identify that such bias is a confounder, which misleads the models to learn the spurious correlation.

- We propose CCIM, a plug-in contextual causal intervention module, which could be inserted into most CAER models to remove the side effect caused by the confounder and facilitate a fair contribution of diverse contexts to emotion understanding.

- Extensive experiments on three standard CAER datasets show that the proposed CCIM can facilitate existing models to achieve unbiased predictions.

2. Related Work

Context-Aware Emotion Recognition. As a promising task, Context-Aware Emotion Recognition (CAER) not only draws on human subject-centered approaches [4, 49, 52] to perceive emotion via the face or body, but also considers the emotion cues provided by background contexts in a joint and boosting manner. Existing CAER models invariably extract multiple representations from these two sources and then perform feature fusion to make the final prediction [12,22–24,29,37,51,56]. For instance, Kosti *et al.* [19] establish the EMOTIC dataset and propose a baseline Convolutional Neural Network (CNN) model that combines the body region and the whole image as the context. Hoang *et al.* [16] propose an extra reasoning module to exploit the images, categories, and bounding boxes of adjacent objects in background contexts to achieve visual relationship detection. For a deep exploration of scene context, Li *et al.* [23] present a body-object attention module to estimate the contributions of background objects and a body-part attention module to recalibrate the channel-wise body feature responses. Although the aforementioned approaches achieve impressive improvements by exploring diverse contextual information, they all neglect the limitation on model performance caused by the context bias of the datasets. Instead of focusing on beating the latest SOTA, we identify the bias as a harmful confounder from a causal inference perspective and significantly improve the existing models with the proposed CCIM.

Causal Inference. Causal inference is an analytical tool that aims to infer the dynamics of events under changing conditions (*e.g.*, different treatments or external interventions) [31], which has been extensively studied in economics, statistics, and psychology [11, 41]. Without loss of generality, causal inference follows two main ways: structured causal model [32] and potential outcome framework [38], which assist in revealing the causality rather than the superficial association among variables. Benefiting from the great potential of the causal tool to provide unbiased estimation solutions, it has been gradually applied to various computer tasks, such as computer vision [5, 34, 39, 42, 45] and natural language processing [17, 35, 57]. Inspired by visual commonsense learning [45], to our best knowledge, this is the first investigation of the confounding effect through causal inference in the CAER task while exploiting causal intervention to interpret and address the confounding bias from contexts.

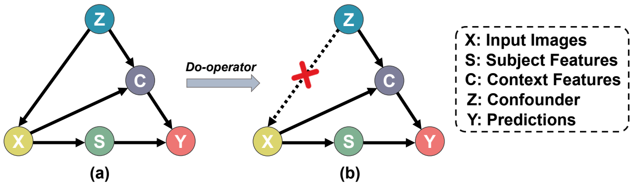


Figure 3. Illustration of our CAER causal graph. (a) The conventional likelihood $P(\boldsymbol{Y}|\boldsymbol{X})$. (b) The causal intervention $P(\boldsymbol{Y}|do(\boldsymbol{X}))$.

3. Methodology

3.1. Causal View at CAER Task

Firstly, we formulate a tailored causal graph to summarize the CAER framework. In particular, we follow the same graphical notation in structured causal model [32] due to its intuitiveness and interpretability. It is a directed acyclic graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$ that can be paired with data to produce quantitative causal estimates. The nodes $\mathcal{N}$ denote variables and the links $\mathcal{E}$ denote direct causal effects. As shown in Figure 3, there are five variables involved in the CAER causal graph, which are input images $\boldsymbol{X}$, subject features $\boldsymbol{S}$, context features $\boldsymbol{C}$, confounder $\boldsymbol{Z}$, and predictions $\boldsymbol{Y}$. Note that our causal graph is applicable to a variety of CAER methods, since it is highly general, imposing no constraints on the detailed implementations. The details of the causal relationships are described below.

$\boldsymbol{Z} \rightarrow \boldsymbol{X}$**.** Different subjects are recorded in various contexts to produce images $\boldsymbol{X}$. On the one hand, the annotators make subjective and biased guesses about subjects' emotional states and give their annotations [18, 20], *e.g.*, subjects are usually blindly assigned positive emotions in vegetation-covered contexts. On the other hand, the data nature leads to an unbalanced representation of emotions in the real world [13]. That is, it is much easier to collect positive emotions in contexts of comfortable atmospheres than negative ones. The context bias caused by the above situations is treated as the harmful confounder $\boldsymbol{Z}$ to establish spurious connections between similar contexts and specific emotion semantics. For the input images $\boldsymbol{X}$, $\boldsymbol{Z}$ determines the biased content that is recorded, *i.e.*, $\boldsymbol{Z} \rightarrow \boldsymbol{X}$.

$\boldsymbol{Z} \rightarrow \boldsymbol{C} \rightarrow \boldsymbol{Y}$**.** $\boldsymbol{C}$ represents the total context representation obtained by contextual feature extractors. $\boldsymbol{C}$ may come from the aggregation of diverse context features based on different methods. The causal path $\boldsymbol{Z} \rightarrow \boldsymbol{C}$ represents the detrimental $\boldsymbol{Z}$ confounding the model to learn unreliable emotion-related context semantics of $\boldsymbol{C}$. In this case, the impure $\boldsymbol{C}$ further affects the predictions $\boldsymbol{Y}$ of the emotion labels and can be reflected via the link $\boldsymbol{C} \rightarrow \boldsymbol{Y}$. Although $\boldsymbol{Z}$ potentially provides priors from the training data to better estimation when the subjects' features are ambiguous, it misleads the model to capture spurious "*context-emotion*" mapping during training, resulting in biased predictions.

$\boldsymbol{X} \rightarrow \boldsymbol{C} \rightarrow \boldsymbol{Y}$ & $\boldsymbol{X} \rightarrow \boldsymbol{S} \rightarrow \boldsymbol{Y}$. $\boldsymbol{S}$ represents the total subject representation obtained by subject feature extractors. Depending on distinct methods, $\boldsymbol{S}$ may come from the face, the body, or the integration of their features. In CAER causal graph, we can see that the desired effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ follows from two causal paths: $\boldsymbol{X} \rightarrow \boldsymbol{C} \rightarrow \boldsymbol{Y}$ and $\boldsymbol{X} \rightarrow \boldsymbol{S} \rightarrow \boldsymbol{Y}$. These two causal paths reflect that the CAER model estimates $\boldsymbol{Y}$ based on the context features $\boldsymbol{C}$ and subject features $\boldsymbol{S}$ extracted from the input images $\boldsymbol{X}$. In practice, $\boldsymbol{C}$ and $\boldsymbol{S}$ are usually integrated to make the final prediction jointly, *e.g*., feature concatenation [29].

According to the causal theory [31], the confounder $\boldsymbol{Z}$ is the common cause of the input images $\boldsymbol{X}$ and corresponding predictions $\boldsymbol{Y}$. The positive effects of context and subject features providing valuable semantics follow the causal paths $\boldsymbol{X} \rightarrow \boldsymbol{C}/\boldsymbol{S} \rightarrow \boldsymbol{Y}$, which we aim to achieve. Unfortunately, the confounder $\boldsymbol{Z}$ causes the negative effect of misleading the model to focus on spurious correlations instead of pure causal relationships. This adverse effect follows the backdoor causal path $\boldsymbol{X} \leftarrow \boldsymbol{Z} \rightarrow \boldsymbol{C} \rightarrow \boldsymbol{Y}$.

3.2. Causal Intervention via Backdoor Adjustment

In Figure 3(a), existing CAER methods rely on the likelihood $P(\boldsymbol{Y}|\boldsymbol{X})$. This process is formulated by Bayes rule:

$$P(\boldsymbol{Y}|\boldsymbol{X}) = \sum_{\boldsymbol{z}} P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{S} = f_s(\boldsymbol{X}), \boldsymbol{C} = f_c(\boldsymbol{X}, \boldsymbol{z}))P(\boldsymbol{z}|\boldsymbol{X}), \tag{1}$$

where $f_s(\cdot)$ and $f_c(\cdot)$ are two generalized encoding functions that obtain the total $\boldsymbol{S}$ and $\boldsymbol{C}$, respectively. The confounder $\boldsymbol{Z}$ introduces the observational bias via $P(\boldsymbol{z}|\boldsymbol{X})$. To address the confounding effect brought by $\boldsymbol{Z}$ and make the model rely on pure $\boldsymbol{X}$ to estimate $\boldsymbol{Y}$, an intuitive idea is to intervene $\boldsymbol{X}$ and force each context semantics to contribute to the emotion prediction fairly. The process can be viewed as conducting a randomized controlled experiment by collecting images of subjects with any emotion in any context. However, this intervention is impossible due to the infinite number of images that combine various subjects and contexts in the real world. To solve this, we stratify $\boldsymbol{Z}$ based on the backdoor adjustment [31] to achieve causal intervention $P(\boldsymbol{Y}|do(\boldsymbol{X}))$ and block the backdoor path between $\boldsymbol{X}$ and $\boldsymbol{Y}$, where *do*-calculus is an effective approximation for the imaginative intervention [14]. Specifically, we seek the effect of stratified contexts and then estimate the average causal effect by computing a weighted average based on the proportion of samples containing different context prototypes in the training data. In Figure 3(b), the causal path from $\boldsymbol{Z}$ to $\boldsymbol{X}$ is cut-off, and the model will approximate causal intervention $P(\boldsymbol{Y}|do(\boldsymbol{X}))$ rather than spurious association $P(\boldsymbol{Y}|\boldsymbol{X})$. By applying the Bayes rule on the new graph, Eq. (1) with the intervention is formulated as:

$$P(\boldsymbol{Y}|do(\boldsymbol{X})) = \sum_{\boldsymbol{z}} P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{S} = f_s(\boldsymbol{X}), \boldsymbol{C} = f_c(\boldsymbol{X}, \boldsymbol{z}))P(\boldsymbol{z}). \tag{2}$$

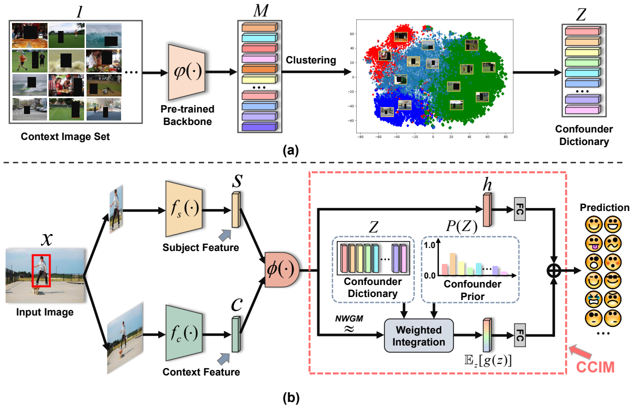


Figure 4. (a) The generation process of the confounder dictionary $\boldsymbol{Z}$. (b) A general pipeline for the context-deconfounded training. The red dotted box shows the core component that achieves the powerful approximation to causal intervention: our CCIM.

As $\boldsymbol{z}$ is no longer affected by $\boldsymbol{X}$, the intervention intentionally forces $\boldsymbol{X}$ to incorporate every $\boldsymbol{z}$ fairly into the predictions of $\boldsymbol{Y}$, subject to the proportion of each $\boldsymbol{z}$ in the whole.

3.3. Context-Deconfounded Training with CCIM

To implement the theoretical and imaginative intervention in Eq. (2), we propose a Contextual Causal Intervention Module (CCIM) to achieve the context-deconfounded training for the models. From a general pipeline of the CAER task illustrated in Figure 4(b), CCIM is inserted in a plug-in manner after the original integrated feature of existing methods. Then, the output of CCIM performs predictions after passing the final task-specific classifier. The implementation of CCIM is described below.

Confounder Dictionary. Since the number of contexts is large in the real world and there is no ground-truth contextual information in the training set, we approximate it as a stratified confounder dictionary $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N]$, where N is a hyperparameter representing the size, and each $\boldsymbol{z}_i \in \mathbb{R}^d$ represents a context prototype. As shown in Figure 4(a), we first mask the target subject in each training image based on the subject's bounding box to generate the context image set I. Subsequently, the image set I is fed to the pre-trained backbone network $\varphi(\cdot)$ to obtain the context feature set $\boldsymbol{M} = \{\boldsymbol{m}_k \in \mathbb{R}^d\}_{k=1}^{N_m}$, where N_m is the number of training samples. To compute context prototypes, we use the K-Means++ with principle component analysis to learn $\boldsymbol{Z}$ so that each $\boldsymbol{z}_i$ represents a form of context cluster. Each cluster $\boldsymbol{z}_i$ is set to the average feature of each cluster in the K-Means++, *i.e.*, $\boldsymbol{z}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} \boldsymbol{m}_j^i$, where N_i is the number of context features in the i-th cluster.

Instantiation of the Proposed CCIM. Since the calculation of $P(\boldsymbol{Y}|do(\boldsymbol{X}))$ requires multiple forward passes of all $\boldsymbol{z}$, the computational overhead is expensive. To reduce the computational cost, we apply the Normalized Weighted Geometric Mean (NWGM) [48] to approximate the above

Category	EMOT-Net [19]	EMOT-Net + **CCIM**	GCN-CNN [56]	GCN-CNN + **CCIM**	CAER-Net [22]	CAER-Net + **CCIM**	RRLA [23]	VRD [16]	EmotiCon [29]	EmotiCon + **CCIM**
Affection	26.47	34.87	47.52	36.18	22.36	23.08	37.93	44.48	38.55	40.77
Anger	11.24	13.05	11.27	12.53	12.88	12.99	13.73	30.71	14.69	15.48
Annoyance	15.26	18.04	12.33	13.73	14.42	15.28	20.87	26.47	24.68	24.47
Anticipation	57.31	94.19	63.2	92.32	52.85	90.03	61.08	59.89	60.73	95.15
Aversion	7.44	13.41	6.81	15.41	3.26	12.96	9.61	12.43	11.33	19.38
Confidence	80.33	74.9	74.83	75.01	72.68	73.24	80.08	79.24	68.12	75.81
Disapproval	16.14	19.87	12.64	14.45	15.37	16.38	21.54	24.54	18.55	23.65
Disconnection	20.64	27.72	23.17	30.52	22.01	23.39	28.32	34.24	28.73	31.93
Disquietment	19.57	19.12	17.66	20.85	10.84	18.1	22.57	24.23	22.14	26.84
Doubt/Confusion	31.88	19.35	19.67	20.43	26.07	17.66	33.5	25.42	38.43	34.28
Embarrassment	3.05	6.23	1.58	9.21	1.88	5.86	4.16	4.26	10.31	16.73
Engagement	86.69	88.93	87.31	96.88	73.71	70.04	88.12	88.71	86.23	97.41
Esteem	17.86	21.69	12.05	22.72	15.38	16.67	20.5	17.99	25.75	27.44
Excitement	78.05	73.81	72.68	73.21	70.42	71.08	80.11	74.21	80.75	81.59
Fatigue	8.87	9.96	12.93	12.66	6.29	9.73	17.51	22.62	19.35	15.53
Fear	15.7	9.04	6.15	10.31	7.47	6.61	15.56	13.92	16.99	15.37
Happiness	58.92	78.09	72.9	75.64	53.73	62.34	76.01	83.02	80.45	83.55
Pain	9.46	14.71	8.22	15.36	8.16	9.43	14.56	16.68	14.68	17.76
Peace	22.35	22.79	30.68	23.88	19.55	20.21	26.76	28.91	35.72	38.94
Pleasure	46.72	46.59	48.37	45.52	34.12	35.37	55.64	55.47	67.31	64.57
Sadness	18.69	17.47	23.9	22.08	17.75	13.24	30.8	42.87	40.26	45.63
Sensitivity	9.05	7.91	4.74	8.02	6.94	4.74	9.59	15.89	13.94	17.04
Suffering	17.67	15.35	23.71	18.45	14.85	11.89	30.7	46.23	48.05	21.52
Surprise	22.38	13.12	8.44	13.93	17.46	11.7	17.92	16.27	19.6	26.81
Sympathy	15.23	32.6	19.45	33.95	14.89	28.59	15.26	15.37	16.74	47.6
Yearning	9.22	10.08	9.86	11.58	4.84	8.61	10.11	10.04	15.08	12.25
mAP	$27.93^{\dagger}$	$\mathbf{30.88}^{\dagger}$ (↑ 2.95)	$28.16^{\dagger}$	$\mathbf{31.72}^{\dagger}$ (↑ 3.56)	$23.85^{\dagger}$	$\mathbf{26.51}^{\dagger}$ (↑ 2.66)	32.41^{*}	35.16^{*}	$35.28^{\dagger}$	$\mathbf{39.13}^{\dagger}$ (↑ 3.85)

Table 1. Average precision (%) of different methods for each emotion category on the EMOTIC dataset. $*$: results from the original reports. $\dagger$: results from implementation. The footnotes $*$ and $\dagger$ of Tables 2 and 3 follow the same interpretation.

expectation at the feature level as:

$$P(\boldsymbol{Y}|do(\boldsymbol{X})) \approx P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{S}=f_s(\boldsymbol{X}), \boldsymbol{C}=\sum_{\boldsymbol{z}} f_c(\boldsymbol{X}, \boldsymbol{z})P(\boldsymbol{z})). \tag{3}$$

Inspired by [45], we parameterize a network model to approximate the above conditional probability of Eq. (3) as follows:

$$P(\boldsymbol{Y}|do(\boldsymbol{X})) = \boldsymbol{W}_h\boldsymbol{h} + \boldsymbol{W}_g\mathbb{E}_{\boldsymbol{z}}[g(\boldsymbol{z})], \tag{4}$$

where $\boldsymbol{W}_h \in \mathbb{R}^{d_m \times d_h}$ and $\boldsymbol{W}_g \in \mathbb{R}^{d_m \times d}$ are the learnable parameters, and $\boldsymbol{h} = \phi(\boldsymbol{s}, \boldsymbol{c}) \in \mathbb{R}^{d_h \times 1}$. $\phi(\cdot)$ is a fusion strategy (*e.g.*, concatenation) that integrates $\boldsymbol{s}$ and $\boldsymbol{c}$ into the joint representation $\boldsymbol{h}$. Note that the above approximation is reasonable, because the effect on $\boldsymbol{Y}$ comes from $\boldsymbol{S}$, $\boldsymbol{C}$, and the confounder $\boldsymbol{Z}$. Immediately, we approximate $\mathbb{E}_{\boldsymbol{z}}[g(\boldsymbol{z})]$ as a weighted integration of all context prototypes:

$$\mathbb{E}_{\boldsymbol{z}}[g(\boldsymbol{z})] = \sum_{i=1}^{N} \lambda_i \boldsymbol{z}_i P(\boldsymbol{z}_i), \tag{5}$$

where λ_i is a weight coefficient that measures the importance of each $\boldsymbol{z}_i$ after interacting with the origin feature $\boldsymbol{h}$, and $P(\boldsymbol{z}_i) = \frac{N_i}{N_m}$. In practice, we provide two implementations of λ_i: dot product attention and additive attention:

$$\text{Dot Product}: \lambda_i = softmax(\frac{(\boldsymbol{W}_q\boldsymbol{h})^T(\boldsymbol{W}_k\boldsymbol{z}_i)}{\sqrt{d}}), \tag{6}$$

$$\text{Additive}: \lambda_i = softmax(\boldsymbol{W}_t^T \cdot Tanh(\boldsymbol{W}_q\boldsymbol{h} + \boldsymbol{W}_k\boldsymbol{z}_i)), \tag{7}$$

where $\boldsymbol{W}_t \in \mathbb{R}^{d_n \times 1}$, $\boldsymbol{W}_q \in \mathbb{R}^{d_n \times d_h}$, and $\boldsymbol{W}_k \in \mathbb{R}^{d_n \times d}$ are mapping matrices.

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets. Our experiments are conducted on three standard datasets for the CAER task, namely EMOTIC [20], CAER-S [22], and GroupWalk [29] datasets. **EMOTIC** contains 23,571 images of 34,320 annotated subjects in uncontrolled environments. The annotation of these images contains the bounding boxes of the target subjects' body regions and 26 discrete emotion categories. The standard partitioning of the dataset is 70% training set, 10% validation set, and 20% testing test. **CAER-S** includes 70k static images extracted from video clips of 79 TV shows to predict emotional states. These images are randomly split into training (70%), validation (10%), and testing (20%) images. These images are annotated with 7 emotion categories: Anger, Disgust, Fear, Happy, Sad, Surprise, and Neutral. **GroupWalk** consists of 45 videos that were captured using stationary cameras in 8 real-world settings. The annotations consist of the following discrete labels: Angry, Happy, Neutral, and Sad. The dataset is split into 85% training set and 15% testing set.

Evaluation Metrics. Following [19, 29], we utilize the mean Average Precision (mAP) to evaluate the results on the EMOTIC and GroupWalk. For the CAER-S, the standard classification accuracy is used for evaluation.

4.2. Model Zoo

Limited by the fact that most methods are not open source, we select four representative models to evaluate the effectiveness of CCIM, which have different network structures and contextual exploration mechanisms. **EMOT-Net** [19] is a baseline Convolutional Neural Net-

Category	EMOT-Net [19]	EMOT-Net + **CCIM**	GNN-CNN [56]	GNN-CNN + **CCIM**	CAER-Net [22]	CAER-Net + **CCIM**	EmotiCon [29]	EmotiCon + **CCIM**
Angry	57.65	62.41	51.92	54.07	45.18	50.43	68.85	75.93
Happy	71.32	75.68	63.37	70.25	56.59	60.71	72.31	79.15
Neutral	43.1	41.03	40.26	39.49	39.32	37.84	50.34	48.66
Sad	61.24	63.84	58.15	61.85	52.96	54.06	70.8	73.48
mAP	58.33†	**60.74**† (↑ 2.41)	53.43†	**56.42**† (↑ 2.99)	48.51†	**50.76** † (↑ 2.25)	65.58†	**69.31** † (↑ 3.73)

Table 2. Average precision (%) of different methods for each emotion category on the GroupWalk dataset.

Methods	**Accuracy** (%)	**Methods**	**Accuracy** (%)
CAER-Net [22]	73.47†	EmotiCon [29]	88.65†
CAER-Net + **CCIM**	**74.81**† (↑ 1.34)	EmotiCon + **CCIM**	**91.17**† (↑ 2.52)
EMOT-Net [19]	74.51†	SIB-Net [24]	74.56*
EMOT-Net + **CCIM**	**75.82**† (↑ 1.31)	GRERN [12]	81.31*
GNN-CNN [56]	77.21†	RRLA [23]	84.82*
GNN-CNN + **CCIM**	**78.66**† (↑ 1.45)	VRD [16]	90.49*

Table 3. Emotion classification accuracy (%) of different methods on the CAER-S dataset.

work (CNN) model with two branches. Its distinct branches capture foreground body features and background contextual information, respectively. **GCN-CNN** [56] utilizes different context elements to construct an affective graph and infer the affective relationship according to the Graph Convolutional Network (GCN). **CAER-Net** [22] is a two-stream CNN model following an adaptive fusion module to reason emotions. The method focuses on the context of the entire image after hiding the face and the emotion cues provided by the facial region. **EmotiCon** [29] introduces three context-aware streams. Besides the subject-centered multimodal extraction branch, they propose to use visual attention and depth maps to learn the scene and socio-dynamic contexts separately. For EMOT-Net, we re-implement the model following the available code. Meanwhile, we reproduce the results on the three datasets based on the details reported in the SOTA methods above (*i.e.*, GCN-CNN, CAER-Net, and EmotiCon).

4.3. Implementation Details

Confounder Setup. Firstly, except for the annotated EMOTIC, we utilize the pre-trained Faster R-CNN [36] to detect the bounding box of the target subject for each training sample on both CAER-S and GroupWalk. After that, the context images are generated by masking the target subjects on the training samples based on the bounding boxes. Then, we use the ResNet-152 [15] pre-trained on Places365 [58] dataset to extract the context feature set $\boldsymbol{M}$. Each context feature $\boldsymbol{m}$ is extracted from the last pooling layer, and the hidden dimension d is 2048. The rich scene context semantics in Places365 facilitate obtaining better context prototypes from the pre-trained backbone. In the EMOTIC, CAER-S, and GroupWalk, the default size N (*i.e.*, the number of clusters) of $\boldsymbol{Z}$ is 256, 128, and 256, respectively.

Training Details. The CCIM and reproducible methods are implemented through PyTorch platform [33]. All models are trained on four Nvidia Tesla V100 GPUs. For a fair comparison, the training settings (*e.g.*, loss function, batch size, learning rate strategy, etc) of these models are consistent with the details reported in their original papers. For the implementation of our CCIM, the hidden dimensions d_m and d_n are set to 128 and 256, respectively. The output dimension d_h of the joint feature $\boldsymbol{h}$ in the different methods is 256 (EMOT-Net), 1024 (GCN-CNN), 128 (CAER-Net), and 78 (EmotiCon).

4.4. Comparison with State-of-the-art Methods

We comprehensively compare the CCIM-based models with recent SOTA methods, including RRLA [23], VRD [16], SIB-Net [24], and GRERN [12]. The default setting uses the dot product attention of Eq. (6).

Results on the EMOTIC Dataset. In Table 1, we observe that CCIM significantly improves existing models and achieves the new SOTA. Specifically, the CCIM-based EMOT-Net, GCN-CNN, CAER-Net and EmotiCon improve the mAP scores by 2.95%, 3.56%, 2.66%, and 3.85%, respectively, outperforming the vanilla methods by large margins. In this case, these CCIM-based methods achieve competitive or better performance than the recent models RRLA and VRD. We also find that CCIM greatly improves the AP scores for some categories heavily persecuted by the confounder. For instance, CCIM helps raise the results of "Anticipation" and "Sympathy" in these CAER methods by 29%$\sim$37% and 14%$\sim$29%, respectively. Due to the adverse bias effect, the performance of most models is usually poor on infrequent categories, such as "Aversion" (AP scores of about 3%$\sim$12%) and "Embarrassment" (AP scores of about 1%$\sim$10%). Thanks to CCIM, the AP scores in these two categories are achieved at about 12%$\sim$19% and 5%$\sim$16%.

Results on the GroupWalk Dataset. As shown in Table 2, our CCIM effectively improves the performance of EMOT-Net, GCN-CNN, CAER-Net, and EmotiCon on the GroupWalk dataset. The mAP scores for these models are increased by 2.41%, 2.99%, 2.25%, and 3.73%, respectively.

Results on the CAER-S Dataset. The accuracy of different methods on the CAER-S dataset is reported in Table 3. The performance of EMOT-Net, GCN-CNN, and CAER-Net is consistently increased by CCIM, making each context prototype contribute fairly to the emotion classification results. These models are improved by 1.31%, 1.45%, and 1.34%, respectively. Moreover, the CCIM-based EmotiCon

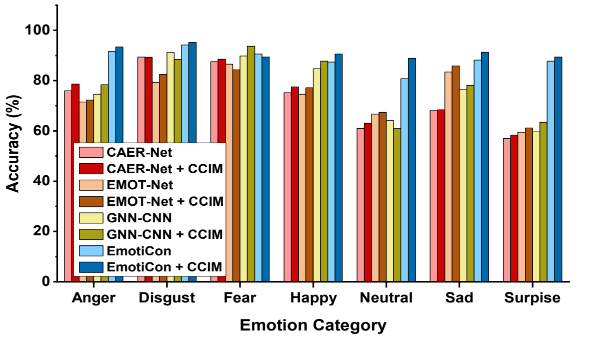


Figure 5. Emotion classification accuracy (%) for each category of different methods on the CAER-S dataset.

achieves a significant gain of 2.52% and outperforms all SOTA methods with an accuracy of 91.17%.

Discussion from the Causal Perspective. **(i)** Compared to the CAER-S (average gain of 1.66% across models), the performance improvements on the EMOTIC (average gain of 3.26%) and GroupWalk (average gain of 2.85%) are more significant. The potential reason is that the samples in these two datasets come from uncontrolled real-world scenarios that contain various context prototypes, such as rich scene information and agent interaction. In this case, CCIM can more effectively eliminate spurious correlations caused by the adequately extracted confounder and provide sufficient gains. **(ii)** Furthermore, CCIM can provide better gains for fine-grained methods of modeling context semantics. For instance, EmotiCon (average gain of 3.37% across datasets) with two contextual feature streams significantly outperforms EMOT-Net (average gain of 2.22%) with only one stream. We argue that the essence of fine-grained modeling is the potential context stratification within the sample from the perspective of backdoor adjustment. Fortunately, CCIM can better refine this stratification effect and make the models focus on contextual causal intervention across samples to measure the true causal effect. **(iii)** According to Tables 1 and 2, and Figure 5, while the causal intervention brings gains for most emotions across datasets, the performance of some categories shows slight improvements or even deteriorations. A reasonable explanation is that the few samples and insignificant confounding effects of these categories result in over-intervention. However, the minor sacrifice is tolerable compared to the overall superiority of our CCIM.

4.5. Ablation Studies

We conduct thorough ablation studies in Table 4 to evaluate the implementation of the causal intervention. To explore the effectiveness of CCIM when combining methods that model context semantics at different granularities, we choose the baseline EMOT-Net and SOTA EmotiCon.

Rationality of Confounder Dictionary $\boldsymbol{Z}$. We first provide a random dictionary with the same size to replace the tailored confounder dictionary $\boldsymbol{Z}$, which is initialized by randomization rather than the average context features. Experimental results (3, 4) show that the random dictionary would significantly hurt the performance, proving the validity of our context prototypes. Moreover, we use the ResNet-152 pre-trained on ImageNet [8] to replace the default settings (1, 2) for extracting context features. The decreased results (5, 6) suggest that context prototypes based on scene semantics are more conducive to approximating the confounder than those based on object semantics. It is reasonable as scenes usually include objects, *e.g.*, in Figure 1, "grass" is the child of the confounder "vegetated scenes".

ID	Setting	EMOTIC	CAER-S	GroupWalk
		mAP (%)	Accuracy (%)	mAP (%)
(1)	EMOT-Net + **CCIM**	**30.88**	**75.82**	**60.74**
(2)	EmotiCon + **CCIM**	**39.13**	**91.17**	**69.31**
(3)	(1) w/ Random $\boldsymbol{Z}$	26.56	73.36	57.45
(4)	(2) w/ Random $\boldsymbol{Z}$	35.12	87.34	65.62
(5)	(1) w/ ImageNet Pre-training	28.72	74.75	58.96
(6)	(2) w/ ImageNet Pre-training	37.48	90.46	68.28
(7)	(1) w/ ResNet-50	29.53	75.34	59.92
(8)	(2) w/ ResNet-50	38.86	90.41	68.85
(9)	(1) w/ VGG-16	28.78	74.95	59.47
(10)	(2) w/ VGG-16	37.93	89.82	68.11
(11)	(1) w/ Additive Attention	30.79	75.64	**60.85**
(12)	(2) w/ Additive Attention	**39.16**	91.08	69.26
(13)	(1) w/o λ_i	30.05	75.21	59.83
(14)	(2) w/o λ_i	38.53	89.67	68.75
(15)	(1) w/o $P(\boldsymbol{z}_i)$	30.63	75.59	59.94
(16)	(2) w/o $P(\boldsymbol{z}_i)$	39.05	90.06	69.15
(17)	(1) w/o Masking Strategy	29.86	74.84	59.22
(18)	(2) w/o Masking Strategy	38.06	90.57	67.79

Table 4. Ablation study results on all three datasets. w/ and w/o are short for with and without, respectively.

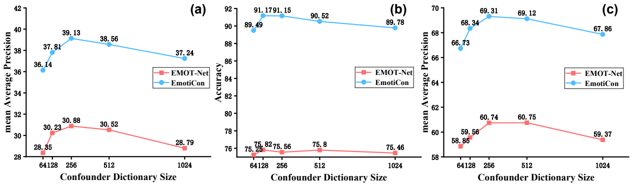


Figure 6. Ablation study results for the size N of the confounder dictionary $\boldsymbol{Z}$ on three datasets. (a), (b), and (c) from the EMOTIC, CAER-S, and GroupWalk datasets, respectively.

Robustness of Pre-trained Backbones. The experiments (7, 8, 9, 10) in Table 4 show that the gain from CCIM increases as more advanced pre-trained backbone networks are used, which indicates that our CCIM is not dependent on a well-chosen pre-trained backbone $\varphi(\cdot)$.

Effectiveness of Components of $\mathbb{E}_{\boldsymbol{z}}[g(\boldsymbol{z})]$. First, we report the results in experiments (11, 12) using the additive attention weight λ_i in Eq. (7). The competitive performance demonstrates that both attention paradigms are meaningful and usable. Furthermore, we evaluate the effectiveness of the weighted integration by separately removing the weights λ_i and prior probabilities $P(\boldsymbol{z}_i)$ in the $\mathbb{E}_{\boldsymbol{z}}[g(\boldsymbol{z})]$. The decreased results (13, 14, 15, 16) suggest that depict-

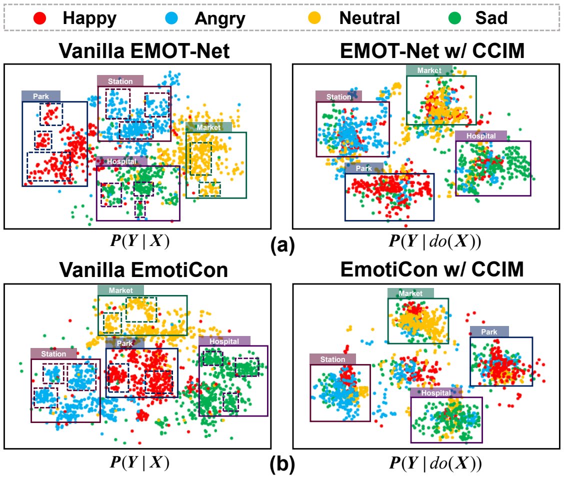
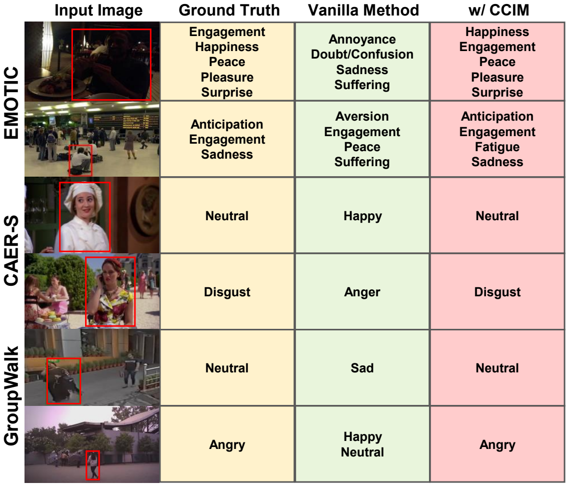


Figure 7. Visualization results of vanilla and CCIM-based EMOT-Net and EmotiCon models on the GroupWalk dataset.



	Input Image	Ground Truth	Vanilla Method	w/ CCIM
EMOTIC		Engagement Happiness Peace Pleasure Surprise	Annoyance Doubt/Confusion Sadness Suffering	Happiness Engagement Peace Pleasure Surprise
		Anticipation Engagement Sadness	Aversion Engagement Peace Suffering	Anticipation Engagement Fatigue Sadness
CAER-S		Neutral	Happy	Neutral
		Disgust	Anger	Disgust
GroupWalk		Neutral	Sad	Neutral
		Angry	Happy Neutral	Angry

Figure 8. Qualitative results of the vanilla and CCIM-based EMOT-Net on three datasets.

ing the importance and proportion of each confounder is indispensable for achieving effective causal intervention.

Effect of Confounder Size. To justify the size N of the confounder $\boldsymbol{Z}$, we set N to 64, 128, 256, 512, and 1024 on all datasets separately to perform experiments. The results in Figure 6 show that selecting the suitable size N for a dataset containing varying degrees of the harmful bias can well help the models perform de-confounded training.

Necessity of Masking Strategy. The masking strategy aims to mask the *recognized subject* to learn prototype representations using pure background contexts. Note that other subjects are considered as background to provide the socio-dynamic context. The gain degradation from the experiments (17, 18) is observed when the target subject regions are not masked. It is unavoidable because the target subject-based feature attributes would impair the context-based confounder dictionary $\boldsymbol{Z}$ along the undesirable link $\boldsymbol{S} \rightarrow \boldsymbol{Z}$, affecting causal intervention performance.

4.6. Qualitative Results

Difference Between $P(\boldsymbol{Y}|\boldsymbol{X})$ and $P(\boldsymbol{Y}|do(\boldsymbol{X}))$. To visually show the difference between the models approximate $P(\boldsymbol{Y}|\boldsymbol{X})$ and $P(\boldsymbol{Y}|do(\boldsymbol{X}))$, we visualize the distribution of context features learned by EMOT-Net and EmotiCon in testing samples on the GroupWalk. These sample images contain four real-world contexts, *i.e*., park, market, hospital, and station. Figure 7 shows the following observations. In vanilla models, features with the same emotion categories usually cluster within similar context clusters (*e.g*., context features of the hospital with the sad category are closer), implying that biased models rely on context-specific semantics to infer emotions lopsidedly. Conversely, in the CCIM-based models, context-specific features form clusters containing diverse emotion categories. The phenomenon suggests that the causal intervention promotes models to fairly incorporate each context prototype semantics when predicting emotions, alleviating the effect of harmful context bias.

Case Study of Causal Intervention. In Figure 8, we select two representative examples from each dataset to show the performance of the model before and after the intervention. For instance, in the first row, the vanilla baseline is misled to predict entirely wrong results because the subjects in the dim scenes are mostly annotated with negative emotions. Thanks to causal intervention, CCIM corrects the bias in the model's prediction. Furthermore, in the fifth row, CCIM disentangles the spurious correlation between the context ("hospital entrance") and the emotion semantics ("sad"), improving the model's performance.

5. Conclusion

This paper proposes a causal debiasing strategy to reduce the harmful bias of uneven distribution of emotional states across diverse contexts in the CAER task. Concretely, we disentangle the causalities among variables via a tailored causal graph and present a Contextual Causal Intervention Module (CCIM) to remove the adverse effect caused by the context bias as a confounder. Numerous experiments prove that CCIM can consistently improve existing models and promote them to a new SOTA. The model-agnostic and plug-in CCIM undoubtedly has excellent superiority over complex module stacking in previous approaches.

Acknowledgements

This work is supported in part by the National Key R&D Program of China (2021ZD0113502, 2021ZD0113503), in part by the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0103), and in part by the China Postdoctoral Science Foundation under Grant (BX20220071, 2022M720769).

References

[1] Abeer Ali Alnuaim, Mohammed Zakariah, Aseel Alhadlaq, Chitra Shashidhar, Wesam Atef Hatamleh, Hussam Tarazi, Prashant Kumar Shukla, and Rajnish Ratna. Human-computer interaction with detection of speaker emotions using convolution neural networks. *Computational Intelligence and Neuroscience*, 2022, 2022. 1

[2] Souha Ayadi and Zied Lachiri. Deep neural network for visual emotion recognition based on resnet50 using song-speech characteristics. In *2022 5th International Conference on Advanced Systems and Emergent Technologies (IC_ASET)*, pages 363–368, 2022. 1

[3] Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290, 2011. 1

[4] Uttaran Bhattacharya, Trisha Mittal, Rohan Chandra, Tanmay Randhavane, Aniket Bera, and Dinesh Manocha. Step: Spatial temporal graph convolutional networks for emotion perception from gaits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1342–1350, 2020. 3

[5] Yingjie Chen, Diqi Chen, Tao Wang, Yizhou Wang, and Yun Liang. Causal intervention for subject-deconfounded facial action unit recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 374–382, 2022. 3

[6] Zhaoyu Chen, Bo Li, Shuang Wu, Jianghe Xu, Shouhong Ding, and Wenqiang Zhang. Shape matters: deformable patch attack. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–548, 2022. 1

[7] Zhaoyu Chen, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Wenqiang Zhang. Towards practical certifiable patch defense with vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15148–15158, 2022. 1

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 7

[9] Yangtao Du, Dingkang Yang, Peng Zhai, Mingchen Li, and Lihua Zhang. Learning associative representation for facial expression recognition. In *IEEE International Conference on Image Processing (ICIP)*, pages 889–893, 2021. 1

[10] Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2402–2411, 2021. 1

[11] E Michael Foster. Causal inference and developmental psychology. *Developmental psychology*, 46(6):1454, 2010. 3

[12] Qinquan Gao, Hanxin Zeng, Gen Li, and Tong Tong. Graph reasoning-based emotion recognition network. *IEEE Access*, 9:6488–6497, 2021. 2, 3, 6

[13] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 3

[14] Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 2, 4

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2, 6

[16] Manh-Hung Hoang, Soo-Hyung Kim, Hyung-Jeong Yang, and Guee-Sang Lee. Context-aware emotion recognition based on visual relationship detection. *IEEE Access*, 9:90465–90474, 2021. 2, 3, 5, 6

[17] William Huang, Haokun Liu, and Samuel R Bowman. Counterfactually-augmented snli training data does not yield better generalization than unaugmented data. *arXiv preprint arXiv:2010.04762*, 2020. 3

[18] Bryan D Jones. Bounded rationality. *Annual review of political science*, 2(1):297–321, 1999. 3

[19] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *Proceedings of the IEEE/CVF Conference on computer Vision and Pattern Recognition (CVPR)*, pages 1667–1675, 2017. 1, 2, 3, 5, 6

[20] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2755–2766, 2019. 1, 2, 3, 5

[21] Haopeng Kuang, Dingkang Yang, Shunli Wang, Xiaoying Wang, and Lihua Zhang. Towards simultaneous segmentation of liver tumors and intrahepatic vessels via cross-attention mechanism. *arXiv preprint arXiv:2302.09785*, 2023. 1

[22] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10143–10152, 2019. 1, 2, 3, 5, 6

[23] Weixin Li, Xuan Dong, and Yunhong Wang. Human emotion recognition with relational region-level analysis. *IEEE Transactions on Affective Computing*, 2021. 1, 2, 3, 5, 6

[24] Xinpeng Li, Xiaojiang Peng, and Changxing Ding. Sequential interactive biased network for context-aware emotion recognition. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–6, 2021. 3, 6

[25] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10631–10642, 2021. 1

[26] Yang Liu, Jing Liu, Mengyang Zhao, Dingkang Yang, Xiaoguang Zhu, and Liang Song. Learning appearance-motion normality for video anomaly detection. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2022. 1

[27] Yang Liu, Dingkang Yang, Yan Wang, Jing Liu, and Liang Song. Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models. *arXiv preprint arXiv:2302.05087*, 2023. 1

[28] Weizhi Meng, Yong Cai, Laurence T Yang, and Wei-Yang Chiu. Hybrid emotion-aware monitoring system based on brainwaves for internet of medical things. *IEEE Internet of Things Journal*, 8(21):16014–16022, 2021. 1

[29] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14234–14243, 2020. 1, 2, 3, 4, 5, 6

[30] 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. 2

[31] Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009. 2, 3, 4

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

[33] Automatic Differentiation In Pytorch. Pytorch, 2018. 6

[34] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10860–10869, 2020. 3

[35] Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5434–5445, 2021. 3

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems (NIPS)*, 28, 2015. 6

[37] Shulan Ruan, Kun Zhang, Yijun Wang, Hanqing Tao, Weidong He, Guangyi Lv, and Enhong Chen. Context-aware generation-based net for multi-label visual emotion recognition. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2020. 3

[38] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005. 3

[39] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3716–3725, 2020. 3

[40] Dahiru Tanko, Sengul Dogan, Fahrettin Burak Demir, Mehmet Baygin, Sakir Engin Sahin, and Turker Tuncer. Shoelace pattern-based speech emotion recognition of the lecturers in distance education: Shoepat23. *Applied Acoustics*, 190:108637, 2022. 1

[41] Hal R Varian. Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences*, 113(27):7310–7315, 2016. 3

[42] Shunli Wang, Shuaibing Wang, Bo Jiao, Dingkang Yang, Liuzhen Su, Peng Zhai, Chixiao Chen, and Lihua Zhang. Caspacenet: Counterfactual analysis for 6d pose estimation in space. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10627–10634, 2022. 3

[43] Shunli Wang, Dingkang Yang, Peng Zhai, Chixiao Chen, and Lihua Zhang. Tsa-net: Tube self-attention network for action quality assessment. In *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, pages 4902–4910, 2021. 1

[44] Shunli Wang, Dingkang Yang, Peng Zhai, Qing Yu, Tao Suo, Zhan Sun, Ka Li, and Lihua Zhang. A survey of video-based action quality assessment. In *International Conference on Networking Systems of AI (INSAI)*, pages 1–9, 2021. 1

[45] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10760–10770, 2020. 3, 5

[46] Yuzheng Wang, Zhaoyu Chen, Dingkang Yang, Yang Liu, Siao Liu, Wenqiang Zhang, and Lizhe Qi. Adversarial contrastive distillation with adaptive denoising. *arXiv preprint arXiv:2302.08764*, 2023. 1

[47] Yuzheng Wang, Zuhao Ge, Zhaoyu Chen, Xian Liu, Chuangjia Ma, Yunquan Sun, and Lizhe Qi. Explicit and implicit knowledge distillation via unlabeled data. *arXiv preprint arXiv:2302.08771*, 2023. 1

[48] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, pages 2048–2057, 2015. 4

[49] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, page 1642–1651, 2022. 1, 3

[50] Dingkang Yang, Shuai Huang, Yang Liu, and Lihua Zhang. Contextual and cross-modal interaction for multi-modal speech emotion recognition. *IEEE Signal Processing Letters*, 29:2093–2097, 2022. 1

[51] Dingkang Yang, Shuai Huang, Shunli Wang, Yang Liu, Peng Zhai, Liuzhen Su, Mingcheng Li, and Lihua Zhang. Emotion recognition for multiple context awareness. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 13697, pages 144–162, 2022. 2, 3

[52] Dingkang Yang, Haopeng Kuang, Shuai Huang, and Lihua Zhang. Learning modality-specific and -agnostic representations for asynchronous multimodal language sequences. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, page 1708–1717, 2022. 1, 3

[53] Dingkang Yang, Yang Liu, Can Huang, Mingcheng Li, Xiao Zhao, Yuzheng Wang, Kun Yang, Yan Wang, Peng Zhai, and Lihua Zhang. Target and source modality co-reinforcement for emotion understanding from asynchronous multimodal sequences. *Knowledge-Based Systems*, 265:110370, 2023. 1

[54] Kun Yang, Jing Liu, Dingkang Yang, Hanqi Wang, Peng Sun, Yanni Zhang, Yan Liu, and Liang Song. A novel ef-

ficient multi-view traffic-related object detection framework. *arXiv preprint arXiv:2302.11810*, 2023. 1

[55] Peng Zhai, Jie Luo, Zhiyan Dong, Lihua Zhang, Shunli Wang, and Dingkang Yang. Robust adversarial reinforcement learning with dissipation inequation constraint. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5431–5439, 2022. 1

[56] Minghui Zhang, Yumeng Liang, and Huadong Ma. Context-aware affective graph reasoning for emotion recognition. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 151–156. IEEE, 2019. 1, 3, 5, 6

[57] Wenkai Zhang, Hongyu Lin, Xianpei Han, and Le Sun. Debiasing distantly supervised named entity recognition via causal intervention. *arXiv preprint arXiv:2106.09233*, 2021. 3

[58] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017. 2, 6

[59] Ruichao Zhu, Jiafu Wang, Tianshuo Qiu, Dingkang Yang, Bo Feng, Zuntian Chu, Tonghao Liu, Yajuan Han, Hongya Chen, and Shaobo Qu. Direct field-to-pattern monolithic design of holographic metasurface via residual encoder-decoder convolutional neural network. *Opto-Electronic Advances*, pages 220148–1, 2023. 1