Towards Unbiased Visual Emotion Recognition via Causal Intervention

Yuedong Chen
yuedong.chen@monash.edu
Monash University
Melbourne, Australia

Xu Yang
101013120@seu.edu.cn
Southeast University
Nanjing, China

Tat-Jen Cham
astjcham@ntu.edu.sg
Nanyang Technological University
Singapore, Singapore

Jianfei Cai
jianfei.cai@monash.edu
Monash University
Melbourne, Australia

ABSTRACT

Although much progress has been made in visual emotion recognition, researchers have realized that modern deep networks tend to exploit dataset characteristics to learn spurious statistical associations between the input and the target. Such dataset characteristics are usually treated as dataset bias, which damages the robustness and generalization performance of these recognition systems. In this work, we scrutinize this problem from the perspective of causal inference, where such dataset characteristic is termed as a confounder which misleads the system to learn the spurious correlation. To alleviate the negative effects brought by the dataset bias, we propose a novel Interventional Emotion Recognition Network (IERN) to achieve the backdoor adjustment, which is one fundamental deconfounding technique in causal inference. Specifically, IERN starts by disentangling the dataset-related context feature from the actual emotion feature, where the former forms the confounder. The emotion feature will then be forced to see each confounder stratum equally before being fed into the classifier. A series of designed tests validate the efficacy of IERN, and experiments on three emotion benchmarks demonstrate that IERN outperforms state-of-the-art approaches for unbiased visual emotion recognition.

CCS CONCEPTS

• Information systems → Multimedia information systems; • Human-centered computing → Human computer interaction (HCI).

KEYWORDS

Causal intervention, backdoor adjustment, facial expression recognition, image emotion recognition, dataset bias

1 INTRODUCTION

Visual emotion recognition, including facial expression recognition (FER) [7, 20, 23, 50] and image emotion recognition (IER) [29, 52, 53], has attracted great attention for decades, playing a vital role in lots of daily scenes such as mental health care and driver drowsiness.
detection. While deep neural networks (DNNs) show promising performances on several existing benchmarks [10, 21], it has been recently observed that DNNs may “cheat” by relying on unintended cognitive level [58] dataset characteristics, e.g., scene contexts [29] and human attributes [45], instead of the actual affective level [58] causal relationships between the input image and the output emotion label, harming their effectiveness in practice. Intuitively, such dataset characteristics are leveraged by DNNs as the shortcut [14] to learn spurious statistical associations between variables.

DNNs may not appear to be affected by dataset characteristics if the training and test set data are from the same distribution [37], under which most emotion benchmarks are collected and constructed. However, a robust practical emotion recognition system is expected to have consistent generalization even for data from different distributions [37]. Here we design a toy experiment to showcase how DNNs behave on data from both the same distribution and different distributions, as depicted in Fig. 1. From a facial expression dataset, we rendered additional images by separately applying blur and noise degradation. They were next partitioned into two datasets (blue dashed vs orange solid borders) by selecting subsets of expressions and degradation labels. A VGG-11 [38] was then trained (randomly initialized and without data augmentation) on samples drawn from the blue dataset. Such a trained model performed well on samples drawn from the same blue dataset (data from the same distribution), shown in the blue confusion matrix of Fig. 1, but is impaired dramatically when tested on samples from the orange dataset (data from a different distribution). Instead of recognizing expressions based on relevant expression features, the model has been biased towards recognizing the degradation effects, which are easier to detect.

Unlike existing methods in emotion recognition that address the dataset bias issue via solutions such as re-sampling and building larger datasets, we choose to leverage causal inference [32, 33, 37] to tackle it. In causal inference, entities or attributes that mislead DNNs to learn spurious correlations between input and output are termed as confounders [32, 33, 35], which refers to the dataset characteristics in our emotion recognition context. For example, in the above toy experiment, the model has been misled to confound blur with disgust and sadness, and fails to recognize anger when blur co-occurs. An intuitive solution for the toy experiment is to add noise or blur to all input images (see the bottom left of Fig. 1), whose underlying principle is to deconfound the emotion recognition systems from the confounder / dataset characteristics. This solution is also known as causal intervention [32].

To deal with more sophisticated confounders, i.e., unknown dataset characteristics and image scenes, we analyze the problem with Pearl’s structure causal model (SCM) [32], and propose a novel framework we call Interventional Emotion Recognition Network (IERN), to do the intervention by embedding the backdoor adjustment theorem [32]. Although backdoor adjustment has been incorporated in some recent deep learning models for other vision tasks [12, 40, 51], all of these only approximate the intervention via memory-query operations, rather than applying real intervention as is done in our IERN.

Specifically, our IERN starts by disentangling the dataset-related context feature from the actual emotion feature, where the former represents the confounder. Following application of the backdoor adjustment theorem, the emotion feature will then be forced to see each confounder stratum equally before being fed into the classifier. We showcase how IERN realizes the backdoor adjustment theorem effectively on the facial expression recognition task via a mixed-dataset configuration, where emotion-independent dataset characteristics are treated as confounder, with dataset name being the label. On the image emotion recognition task, IERN outperforms the state-of-the-art approach by a significant margin under the challenging cross-dataset setting, where image scenes are treated as confounders. The contributions are threefold.

- We are the first to tackle dataset bias in visual emotion recognition from a causality perspective.
- We propose a novel trainable framework, named Interventional Emotion Recognition Network (IERN), to realize the backdoor adjustment theorem.
- With rigorous experiments done for both the mixed-dataset and cross-dataset on existing benchmarks, we show how IERN effectively alleviates negative effects raised by dataset bias and outperforms state-of-the-art approaches.

2 RELATED WORK

Visual Emotion Recognition. The majority of emotion recognition works belong to Facial Expression Recognition (FER), where the input are human facial images with simple and limited backgrounds [16, 43, 57]. Most methods focus on enhancing performance by making use of related auxiliary priors, e.g., facial landmarks [19, 55, 56], local regions [9, 17, 48], action units [7, 8, 24], optical flows [39], etc. The assumption of these methods is that the training data is fairly distributed, which may be wrong and could harm the generalization performance of trained models. Recently, the bias problem has received much attention, and most approaches [13, 22, 28, 36] mainly aim to intuitively disentangle the bias features, so as to build more robust emotion features.

Others belong to Image Emotion Recognition (IER), which aims to recognize the emotions invoked from images that contain a variety of objects rather than recognizing emotions of human faces. Approaches along this line mainly deal with social media photos [44, 47], manually generated images [18], artistic photographs [1, 58], web crawled natural images [29, 52, 53], etc. The most relevant work is [29], which identified the image scene as the bias factor, and managed to alleviate the bias issue by building a new larger and more balanced dataset using web data.

Different from all approaches that handle dataset bias in visual emotion recognition via heuristic solutions, our approach is the first to offer a fundamental understanding of the bias issue via causal inference, and propose a new trainable unbiased framework based on causal intervention, by employing the feature disentanglement technique and the center loss function [46].

Causal inference. Causality [32, 35, 37] can help pursue the causal effect between two observed variables, rather than depend only on their correlation. Many works have been proposed to explore how causality can be leveraged in machine learning [4, 5, 27, 30]. Recently, causal inference has been introduced into the deep learning framework for several computer vision tasks, including image captioning [51], image classification [6, 25, 40], semantic segmentation [12], and few-shot learning [54].
Our work differs from causality-based deep learning methods in two aspects. Firstly, ours is the first to address the biased visual emotion recognition problem. Secondly and more importantly, we propose a novel solution by learning confounder features and conducting real intervention, while previous methods had chosen to model confounders using a predefined dictionary and approximate intervention via memory-query operations by adopting the Normalized Weighted Geometric Mean (NWGM) [3, 49].

3 METHODOLOGY

Given input image X, visual emotion recognition aims to solve the problem \( P(Y|X) \), where Y is the emotion label. In this section, we will detail how confounder D undermines the objective of \( P(Y|X) \) (Section 3.1) to raise dataset bias, how such a bias effect can be removed (Section 3.2), and what our unbiased solution is (Section 3.3).

3.1 SCM for Emotion Recognition

Structural causal models (SCM) [32, 37] are designed to analyze the causal relationships. Fig. 2 shows a SCM in our context constructed among input X, emotion Y and confounder D. The direction of an edge denotes only the causal relationship, while information can flow bidirectionally, e.g., \( D \rightarrow X \) means that D is the cause and X is the effect, while information can still flow from X to D.

\[ X \rightarrow Y \]

It is the intended causal relationship that the network is supposed to learn, which is to recognize emotion Y based on input image X. For simplicity, we mix the use of symbol X, denoting either input images or image features.

\[ D \rightarrow X, \; D \rightarrow Y \]

Confounder D is the undesirable context feature, e.g., *unknown dataset characteristics, image scenes*. Without disentanglement, confounder are normally embedded in the input features, and thus we have \( D \rightarrow X \). Similarly, confounder also affects networks in predicting emotion, so we have \( D \rightarrow Y \).

\[ X \leftarrow D \rightarrow Y \] (backdoor path)

The existence of confounder D enables a backdoor path between X and Y, by which networks can be biased to build up spurious correlation between X and Y, leading to degradation of its generalization performance. For example, as shown in Fig. 1, instead of identifying the correct *anger* image features to recognize anger, i.e. \( X \rightarrow Y \), networks may learn that *noise is anger*, i.e. \( D \rightarrow X \), and should be predicted as *anger*, i.e. \( X \leftarrow D \rightarrow Y \).

The backdoor path can be explained with the law of total probability. Specifically, \( P(Y|X) \) can be decomposed into

\[ P(Y|X) = \sum_d P(Y|X, D = d)P(D = d|X) \]

where \( d \) is one of the strata (a.k.a. “level”) of confounder D, e.g., \( d_{\text{blur}}, d_{\text{noise}} \). Assuming that spurious correlation leads to \( P(d_{\text{noise}}|X_{\text{anger}}) \approx 1 \), then \( P(Y|X_{\text{anger}}, d_{\text{noise}}) \) will become \( P(Y|X_{\text{anger}}, d_{\text{noise}}, \text{)} \), resulting in bias towards \( d_{\text{noise}} \).

3.2 Intervention with Backdoor Adjustment

Once the confounder is identified, a straightforward solution to alleviate the dataset bias is to build a larger dataset to ensure that emotion is balanced among all confounder strata, similar to what has been done in [29]. However, such solutions might not be practical due to three main reasons. Firstly, it is time-consuming and costly to collect and label a large amount of data. Secondly, if the confounder is too complex, it is hard to construct a balanced new dataset. Thirdly, there may be ethical or practical issues in collecting certain types of data, e.g., forcing a child to cry in a park in order to get negative-emotional park images.

We therefore turn to a more elegant solution of applying the intervention [32] by blocking the backdoor path in the SCM (Fig. 2 Right). Specifically, instead of learning the correlation \( P(Y|X) \), we set our target as \( P(Y|do(X)) \), where \( do(\cdot) \) is the do-operator (a.k.a. do-calculus), denoting the experimental intervention. Since both D and X are accessible in our tasks, we can solve \( P(Y|do(X)) \) by applying the backdoor adjustment theorem [32]:

\[ P(Y|do(X)) = \sum_d P(Y|X, D = d)P(D = d) \]

Intuitively, the backdoor adjustment removes the risk of spurious correlation arising from \( P(D|X) \) via a two-step strategy. It starts by estimating the causal effect in each stratum of the confounder, followed by summing the effects of all strata, weighted by their known prevalence in the population. By doing so, it ensures that the contribution of the causal effect of an unobserved stratum is no different from that of an observed one. In this way, the overall causal effect would not be biased towards the observed strata. For example, with the help of the backdoor adjustment, each stratum of the confounder, including *noise, blur and normal*, will contribute equally to *anger*, forcing the model to learn to recognize *anger* from the intended affective features.

3.3 IERN

To overcome dataset bias in visual emotion recognition, we propose our Interventional Emotion Recognition Network (IERN) by embedding the backdoor adjustment. Compared with heuristic debiasing methods like building larger datasets [29], IERN is designed based on well-studied deconfounding theory. Also, unlike NWGM approximation-based backdoor adjustment models [12, 51] that absorb the probability sum into the network, IERN applies the real intervention \( P(Y|do(X)) \). Fig. 3 gives an overview of the IERN, which is composed of four parts: Backbone, Feature Disentanglement, Confounder Builder and Classifier. The backbone refers to an existing image feature extractor, e.g., VGG [38] or ResNet [15], the
Figure 3: Overview of Interventional Emotion Recognition Network (IERN). $f_b, g_e, d_e, g_c, d_c, g_r$ and $f_c$ refer to backbone, emotion generator, emotion discriminator, context generator, context discriminator, reconstruction network and classifier, respectively. In training, Feature Disentanglement splits features into emotion and context features. Confounder Builder groups context features and gets their centers as confounder features. Each emotion feature is combined with all strata of the confounder individually, and forwarded to the Classifier. Only the top branch will be leveraged in testing, using the learned confounders.

The disentanglement part separates emotion and context features, the confounder builder does the causal intervention, and the classifier predicts the emotion label. Formally, $f_b, g_e, d_e, g_c, d_c, g_r$ and $f_c$ refer to backbone, emotion generator, emotion discriminator, context generator, context discriminator, reconstruction network and classifier, respectively.

**Feature Disentanglement.** As mentioned in Section 3.1, without proper handling, context features are mixed with emotion features. Thus, IERN starts by applying feature disentanglement to separate emotion and context features. Specifically, to obtain the clean emotion feature, the disentanglement is conducted by a dual-discriminator structure, consisting of an emotion discriminator $d_e$ and a context discriminator $d_c$, where $d_e$ is to ensure the extracted feature contains emotional information while $d_c$ is to ensure that the extracted feature does not contain any context information.

This procedure is implemented by alternating the optimization between the feature generators and the discriminators with the following training objective:

$$
\mathcal{L}_c = \min_{d_e} l_{CE}(d_e(g_e(f_b(x))), y_e) + \min_{g_e} l_{MSE}(d_e(g_e(f_b(x))), \frac{1}{N_e}),
$$

where $l_{CE}$ and $l_{MSE}$ denote the cross-entropy loss and the mean squared error loss, respectively; $d_e^\theta$ and $g_e^\theta$ are the trainable parameters of $d_e$ and $g_e$, respectively; and $g_e(f_b(x))$ is the intended clean emotion feature, obtained by feeding the input image $x$ first into the backbone $f_b$ and then into the emotion feature generator $g_e$. Also, $y_e$ is the ground-truth emotion label for $x$, and $N_e$ is the number of levels in the confounder. Intuitively, in Eq. (3), $d_e$ is optimized to predict the ground truth emotion label given the generated emotion feature $g_e(f_b(x))$ as input, so as to ensure that $g_e(f_b(x))$ contains the correct emotion information. Here, $g_e$ is optimized to encourage the softmax output of $d_c$ to be equally distributed among all confounder levels, so as to ensure that $g_c(f_b(x))$ contains no context information.

To obtain clean context features, we adopt a similar optimization approach with the following objective function:

$$
\mathcal{L}_c = \min_{d_e^\theta} l_{CE}(d_e^\theta(g_e^\theta(f_b(x))), y_c) + \min_{g_e^\theta} l_{MSE}(d_e^\theta(g_e^\theta(f_b(x))), \frac{1}{N_c}),
$$

where $d_e^\theta$ and $g_e^\theta$ are the trainable parameters of the context discriminator $d_e$ and the context generator $g_e$, respectively; $y_c$ is the confounder label, and $N_c$ is the number of emotion classes.

To ensure that the separated features fall within reasonable domains, IERN should be capable of reconstructing the base feature $f_b(x)$, given the emotion features and the context features as input. Thus, a feature reconstruction loss function is added as

$$
\mathcal{L}_r = \min_{g_r^\theta} l_{MSE}(g_r^\theta(g_c(f_b(x))), g_r^\theta(f_b(x))).
$$

where $g_r^\theta$ is the trainable parameters of the feature reconstruction network $g_r$.

**Cofounder Builder.** The purpose of the confounder builder is to combine each emotion feature with different context features so as to avoid the bias towards the observed context strata. To limit the diversity of the context features, we propose to use the center of all context features within each specific stratum as a cofounder feature. This is reasonable since context features are usually similar within the same stratum while different across different strata. Essentially, a confounder feature represents the general concept of a specific stratum. So we adopt the center loss [46] to learn cofounder features with the objective function as

$$
\mathcal{L}_{CB} = \min_{g_c, g_e} l_{MSE}(g_c(f_b(x)), G_{Ce}).
$$
where \( C_{y_e} \) denotes the learned confounder feature of the \( y_e \)-th stratum of the confounder, and \( C = \{ C_1, C_2, ..., C_{N_C} \} \), with all being learnable parameters.

**Classifier.** As mentioned in Section 3.2, the backdoor adjustment aims to weigh the causal effect of each stratum with its prevalence in the population, i.e., \( P(D = d) \), instead of \( P(D = d|X) \). To ensure fairness, our deconfounded classifier is set such that for any emotion feature, it is present in all confounder strata equally, i.e., \( P(D = d) = 1/N_C \), via the following training objective:

\[
\mathcal{L}_{\text{Cls}} = \min_{f^0, g^0, f^g} \sum_{i=1}^{N_C} f_e(g_e(f_b(x), C_i)), y_e. \tag{7}
\]

where \( f^0 \) denotes the trainable parameters of the classification network \( f_e \). Given a separated emotion feature \( g_e(f_b(x)) \), IERN reuses the reconstruction network \( g_r \) to combine it with each learned confounder feature \( C_i \), so as to ensure that for any emotion, the classifier can be challenged with that emotion existing in every stratum of the confounder equally, thus avoiding bias. In other words, \( P(Y|X, D = d) \) is obtained by \( f_e(g_r(g_e(f_b(x), C_i))) \), while by forwarding \( N_C \) times and taking the average, we have \( P(Y|do(X)) \), where \( P(D = d) = 1/N_C \).

**Training Phase.** In general, IERN is trained with the weighted combination of the above loss functions. The overall objective is

\[
\mathcal{L}_{do} = \lambda_1 \mathcal{L}_e + \lambda_2 \mathcal{L}_{CB} + \lambda_3 \mathcal{L}_{\text{Cls}} \tag{8}
\]

where \( \lambda_1, \lambda_2 \) and \( \lambda_3 \) are hyper-parameters. IERN is end-to-end trainable, and different components need to be updated sequentially within one training iteration. The detailed training procedure is given in Algorithm 1.

**Algorithm 1 IERN Training Procedure**

1. for number of training iterations do
2. Forward \( f_b, g_e, g_r, d_e, d_r, g_r \)
3. Freeze \( d^e_\theta \) and \( d^g_\theta \), backward \( d_e \) using 1st term of \( \mathcal{L}_e \) and \( d_r \) using 1st term of \( \mathcal{L}_r \)
4. Freeze \( d^e_\theta \) and \( d^g_\theta \), backward \( g_e \) using 2nd term of \( \mathcal{L}_e \) and \( \mathcal{L}_r \), backward \( g_r \) using 2nd term of \( \mathcal{L}_e \) and \( \mathcal{L}_r \) and \( \mathcal{L}_{CB} \), backward \( g_r \) using \( \mathcal{L}_r \), backward \( C \) using \( \mathcal{L}_{CB} \)
5. for \( i = 1 \rightarrow N_C \) do
6. Forward \( f_e(g_r(g_e(f_b(x)), C_i)) \)
7. Backward \( f_e, g_e \) and \( f_b \) using \( \mathcal{L}_{\text{Cls}} \)

**Testing Phase.** Given a test image, IERN will predict its emotion by going through only the top branch (see Fig. 3), which is

\[
y_e = \arg \max \sigma \left( \frac{1}{N_C} \sum_{i=1}^{N_C} f_e(g_e(f_b(x), C_i)) \right). \tag{9}
\]

where \( \sigma \) denotes the softmax function, and \( C_i \) is the confounder feature learned in the training phase. Following the backdoor adjustment theorem, to ensure the prediction of emotion is not biased towards specific strata at test time, we need to combine the emotion feature with each confounder stratum feature and take an average.

### 4 EXPERIMENTS

We conducted experiments on both facial expression recognition (FER) (Section 4.1) and image emotion recognition (IER) (Section 4.2). For the former, we designed a challenging yet practical out-of-distribution (o.o.d.) test configuration through mixing different benchmarks, motivated by the fact that in the real world, images provided to practical recognition systems are taken by different people with different preferences, cameras and geographical locations, and this distribution cannot be controlled [37]. And we conduct ablation study and compare to related SOTA, including DACL [13] and NWGM-based methods [12], using the introduced configurations. As for the latter, IERN is compared, under the challenging cross-dataset setting, to Curriculum Learning [29], which is the SOTA for unbiased image emotion recognition.

**Implementation details.** To maintain a fair comparison, ResNet-50 [15] was chosen as the backbone \( f_b \) following the setting of [29]. While \( g_e \) and \( g_r \) are constructed by using the residual block [15] as building blocks, \( d_e \) and \( d_r \) are composed of convolution layers, and \( f_e \) is composed of fully connected layers. Note that other backbones or building blocks can also be adopted, as long as they follow the framework depicted in Fig. 3. The hyper-parameters \( \lambda_1 \) and \( \lambda_3 \) in Eq. (6) were set as default to \( \lambda_1 = \lambda_3 = 1 \), while \( \lambda_2 = 5 \times 10^{-4} \) were set mainly to balance the loss value magnitudes. We initialized \( f_b \) with ImageNet [11] pretrained weights, while other components were randomly initialized. An Adam optimizer was used with a learning rate of \( 2 \times 10^{-4} \), decayed using the standard warm-up strategy. All models were trained till converged before testing. Specifically, in the mix-dataset experiments (Section 4.1), IERN were trained for 80 epochs, while in the cross-dataset experiments (Section 4.2), IERN were trained for 140 epochs. All test sets remain untouched during training. We implemented IERN with PyTorch [31] and trained it using 1-3 TITAN X GPUs.

### 4.1 Mix-Dataset Experiments for FER

**Dataset settings.** The facial expression experiments were conducted on three lab-controlled benchmarks, including CK+ [26], MMI [42], and Oulu-CASIA [57]. In all the datasets, we only selected those sequences that contain frontal faces and are labeled with six basic expression labels, i.e., anger, disgust, fear, happiness, sadness and surprise, resulting in CK+ with 327 sequences of 118 subjects, MMI with 208 sequences of 32 subjects, and Oulu-CASIA with 480 sequences of 80 objects. As a general approach [21], three peak frames were extracted from each sequence for further experiments.

To ascertain the effectiveness of IERN in dealing with dataset bias, practical o.o.d. tests were designed by mixing the three datasets in a three-fold cross-validation setting. With reference to Tab. 1: in

|       | AN | DI | FE | HA | SA | SU |
|-------|----|----|----|----|----|----|
| CK+   | 1  | 2  | 3  | 1  | 2  | 3  |
| MMI   | 3  | 1  | 2  | 3  | 1  | 2  |
| Oulu-CASIA | 2  | 3  | 1  | 2  | 3  | 1  |
was measured in terms of classification accuracy, averaged across were included in the test set, while the other images were used to This is reasonable since dataset bias in practice can be very complex vary across different people on CK+ and MMI. The performance was included in the test set, while the other images were used to construct the training set. Likewise, the second fold comprises a test set of the ‘2’ image sets, and so on. Here we chose to apply a sequence-independent setting rather than a person-independent one, the reason is that using the latter will make the data very unbalanced across different folds, since the numbers of emotions vary across different people on CK+ and MMI. The performance was measured in terms of classification accuracy, averaged across all three folds (% is omitted in all result tables). We point out that under such a setting, the dataset label is the confounder label, i.e., $g_c$ were grouped and separated, while confounder features were found near the center of each context feature distribution as expected. Therefore, by using the Confounder Builder, the feature of each stratum of the confounder is correctly modeled, allowing the following do intervention to be applied correctly.

Confounder features are correctly modeled. We plotted the extracted context features $g_c(f_b(x))$ and confounder center features $C$ using t-SNE. As seen in the left part of Fig. 4, context features were grouped and separated, while confounder features were found near the center of each context feature distribution as expected. Therefore, by using the Confounder Builder, the feature of each stratum of the confounder is correctly modeled, allowing the following do intervention to be applied correctly.

Dataset is the confounder. The right part of Fig. 4 depicts the accuracy confusion matrix for $d_c$, given the extracted context feature $g_c(f_b(x))$ as input. We can see that the network achieved over 90% accuracy in predicting the confounder / dataset label, which is consistent with the finding of the Name That Dataset Game conducted in both [41] and [29]. Compared with the emotion prediction results shown in Tab. 2, it is clear that the network is better at learning to recognize the cognitive level features (dataset characteristics) than to recognize the affective level features (image emotion). Without intervention, the network would thus be biased by the more easily learned dataset features.

IERN vs Baseline in detail. To get a closer view of how IERN outperforms Baseline in dealing with dataset bias, the accuracy confusion matrices of both models on the third fold setting are shown in Fig. 5. Compared to IERN, the results of Baseline were biased by the training data. For example, fear images in the test

Table 2: FER accuracy for the comparisons using the designed o.o.d. setting. The best results are highlighted in bold

|                | Anger | Disgust | Fear  | Happiness | Sadness | Surprise | Average |
|----------------|-------|---------|-------|-----------|---------|----------|---------|
| Baseline       | 10.13 | 30.60   | 15.29 | 49.56     | 10.00   | 77.12    | 35.57   |
| Disentanglement| 1.27  | 27.68   | 18.80 | 67.19     | 15.95   | 80.07    | 38.95   |
| DACL [13]      | 5.27  | 30.68   | 29.82 | 47.99     | 11.90   | 83.50    | 38.01   |
| IERN (Ours)    | 10.13 | 28.07   | 23.56 | 81.85     | 29.05   | 79.09    | 45.50   |

Ablation study of IERN. We compared IERN to two variants: 1) Baseline, which is the backbone of IERN, i.e., ResNet-50. 2) Disentanglement, which directly feeds the output of $g_c$ into $f_c$, bypassing the Confounder Builder. The results are shown in Tab. 2. 

Compared to Baseline, Disentanglement saw a 3.38% improvement in terms of average accuracy. This is because Disentanglement can distill the expected emotion feature whilst dispensing the undesirable confounding feature, so that the classifier can be better learned. This finding has also been explored and verified recently by an existing approach that dealt with biased facial attribute classification [2]. IERN improved on Disentanglement by 6.55%. From a causality point of view, IERN invokes causal intervention, so that it can address the bias better as previously discussed. From a deep learning point of view, IERN can be seen as applying feature-level data augmentation. It combines emotion features with all kinds of confounder features to provide coverage even for combinations missing from training data. So it encounters greater data diversity and is thus more robust than pure disentanglement.

We also evaluated the related SOTA, DACL [13], on the designed setting by training and testing with the released codes. As shown in Tab. 2, IERN outperformed DACL with a healthy margin. Besides, the performance of DACL is close to Disentanglement, since they share similar underlying principles as mentioned in Section 2.

Figure 4: Left: t-SNE maps of context and confounder features (best viewed in high-resolution color). Right: confusion matrix of the confounder prediction.

Figure 5: Confusion matrices of accuracy for both Baseline (left) and IERN (right) on the third fold setting.
Table 3: Facial expression recognition accuracy for the comparison with other methods using the modified setting (moving 10% of images from the original test set into the training set). The best results for each category are highlighted in bold

|                | Anger | Disgust | Fear | Happiness | Sadness | Surprise | Average |
|----------------|-------|---------|------|-----------|---------|----------|---------|
| Baseline       | 24.48 | 46.10   | 49.44| 80.04     | 46.83   | 89.31    | 58.55   |
| Re-sampling    | 23.78 | 48.49   | 57.78| 89.73     | 43.92   | 90.22    | 61.59   |
| NWGM-based [12]| 30.77 | 59.74   | 52.78| 81.59     | 45.24   | 95.29    | 63.63   |
| IERN (Ours)    | 59.44 | 59.52   | 60.56| 83.92     | 62.43   | 93.66    | 71.71   |

Figure 6: Two random samples of anger from each dataset. See the mouth region, while it is consistently opened within MMI, such manifestation is inconsistent in other two datasets.

set are from CK+ (see Tab. 1 number 3). Baseline predicted over 50% images as anger, disgust, happiness or sadness, which are all related to CK+ in the training set, while IERN alleviated the bias and improved the performance by near 20%.

Limitation of using dataset as the confounder. We noticed that the accuracy results of anger for both Baseline and IERN are 0 as shown in Fig. 5. The reason is that the manifestation of anger in MMI is inconsistent with that in CK+ and Oulu-CASIA, as depicted in Fig. 6, where the mouth region in MMI (consistently opened within MMI) has a very different shape compared to the other two datasets. Since in the third fold, IERN was trained with anger images from only CK+ and Oulu-CASIA, it failed to recognize anger images in MMI. IERN was rather intended to solve biases caused by dataset characteristics that are consistently shared across the whole dataset, e.g., images in MMI have a blue background, CK+ contains mainly gray images, images in Oulu-CASIA are blurry.

The manifestation differences also make the Disentanglement perform much worse than Baseline and IERN in the anger expression as shown in Tab. 2. On the first and second folds, anger faces in the training set are mixed with images from MMI and another dataset. Under such settings, Feature Disentanglement failed to disentangle emotion and context features due to the inconsistent manifestation, resulting in wrong emotion features. Unlike Disentanglement that predicted emotion labels purely based on the extracted emotion features, our IERN still combined emotion with context features in the following steps before applying classification, allowing it to overcome the flaws to some extent.

Re-sampling is essentially intervention. The most widely adopted method in handling dataset bias is to balance the training data through re-sampling. However, in our o.o.d. test setting, the training data does not contain any image that shares the same attribute as those in the test set, which makes it impossible to directly use re-sampling.

In fact, re-sampling can be considered as a vanilla form of intervention, since it also aims to solve the bias through balancing the data across all strata of the confounder. For better comparison between IERN and re-sampling, we next modified the setting by moving 10% of images (sequence-independent) from the test set into the training set for each attribute in each fold, so that re-sampling can be applied. The performances are reported in Tab. 3.

Compared to Baseline, re-sampling achieved a higher performance with 3.04% improvement. The reason is that re-sampling is approximately intervention, so it can help alleviate the bias and improve the robustness of the model. However, IERN significantly outperformed re-sampling by 10.12%. This is because by modeling unobserved data using confounder features, IERN can improve not only the balance but also the diversity of training data, while re-sampling does not introduce new data.

Comparison to NWGM-based method. We also compared IERN to the NWGM-based method [12], given that both methods are based on backdoor adjustment. The confounder features in NWGM are built as a predefined dictionary, by grouping images in the training set according to confounder labels and averaging the extracted features of each group [12, 51].

Using the same modified setting as the re-sampling experiment, the results of the NWGM method are shown in Tab. 3. Although both re-sampling and NWGM are approximate interventions, NWGM outperformed re-sampling by 2.04%, suggesting that it is better to model confounders explicitly in the network. However, IERN significantly outperformed NWGM by 8.08%. This is because IERN conducts real intervention by modeling confounder features as trainable parameters, while the pre-computed confounders in NWGM might not be sufficiently representative.

4.2 Cross-Dataset Experiments for IER

Dataset settings. Experiments related to image emotion were conducted on three in-the-wild benchmarks, including Deep Sentiment [52], Emotion-6 [29] and WEEmo [29]. All datasets are based on images collected from the internet, e.g., Flickr, Google. Deep Sentiment contains 1269 images and Emotion-6 has 8350 images, both of which were manually labeled by several human subjects. WEEmo has around 268,000 images, which were automatically labelled by query keywords. Following [29], we report results based on the binary emotion labels, i.e., positive and negative.

We followed the same configurations as [29], which mainly used 80% of images for training and the rest for testing. According to the analysis given in [29], the image scene is the cause of the bias, and is thus the confounder. Places365-CNNs [59] was used to detect scene categories, which contains over 300 classes. To improve error tolerance and reduce computation, we further developed a two-step
Table 4: Cross-dataset validation for image emotion recognition. Values in cells are the results of Curriculum Learning [29], and values inside brackets are the improvements of IERN. “Self Test” means the self-validation (data from the same distribution) experiments, and “Mean Others” means the average of cross-datasets (data from different distributions) validations.

| Train on:    | Deep Sentiment | Emotion-6 | WEBEmo | Self Test | Mean Others |
|--------------|----------------|-----------|--------|-----------|-------------|
| Deep Sentiment | 78.74 [+ 3.01] | 49.76 [+ 8.18] | 47.79 [+ 3.41] | 78.74 [+ 3.01] | 48.78 [+ 5.79] |
| Emotion-6     | 54.33 [+ 5.19] | 77.72 [+ 1.85] | 64.30 [+ 0.99] | 77.72 [+ 1.85] | 59.32 [+ 3.09] |
| WEBEmo        | 68.50 [+ 2.13] | 78.38 [+ 0.53] | 81.41 [+ 0.31] | 81.41 [+ 0.31] | 73.44 [+ 1.33] |

Figure 7: Emotion distribution within each scene stratum for both training (left) and test sets (right). SEN, EMO and WEB refer to Deep Sentiment, Emotion-6 and WEBEmo dataset, respectively. P denotes Positive emotion and N denotes Negative emotion (best viewed in high-resolution color).

We noticed that IERN had a lower improvement when trained on WEBEmo compared to training on the other two datasets, which is mainly because WEBEmo is already well-balanced. WEBEmo was proposed in [29] as a solution to address the bias problem, with emotion data fairly distributed across scenes. To verify, we plot the emotion distribution within each scene stratum for both training and test sets of all datasets. As shown in Fig. 7, compared with Deep Sentiment and Emotion-6, WEBEmo has better-balanced emotion distributions in many scenes.

We also found that IERN gained less improvement on the self-validation experiments than on cross-dataset ones, especially on Deep Sentiment and Emotion-6. The reason is that self-validation (data from the same distribution) might not be affected by bias [37]. As shown in Fig. 7, within the same dataset, the training and test set have similar bias in the distribution. Thus, although Curriculum Learning might learn biased features during training, it may still achieve good performance on a test set that has the same biases.

Conversely, in the cases of cross-datasets (data from different distributions) validation, this is no longer true. In particular for many scenes, the biases in Deep Sentiment and Emotion-6 are in opposite directions, e.g., in beauty_salon, the emotion of the former tends to be positive, while for the latter it tends to be negative. So without proper handling, Curriculum Learning may suffer from opposing biases when trained on one dataset and tested on another, while IERN can remove the bias in training via re-balancing, thus achieving higher performance.

5 CONCLUSION

We have studied the dataset bias problem in visual emotion recognition from a causality perspective. We have proposed a novel end-to-end trainable framework, IERN, to model the backdoor adjustment theorem of causal intervention in order to achieve unbiased emotion recognition. Extensive experiments on two visual emotion recognition tasks have shown the effectiveness of IERN and its individual components. Future works include applying our solution to other vision tasks and considering more complex confounders, e.g., use gender / ethnicity as confounders, automatically learn the number of confounders by embedding clustering algorithms.

ACKNOWLEDGMENTS

This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from Singapore Telecommunications Limited (Singtel), through Singtel Cognitive and Artificial Intelligence Lab for Enterprises (SCALE@NTU). This research is also partially supported by FIT Start-up Grant.
REFERENCES

[1] Xavier Alameda-Pineda, Elisa Ricci, Yan Yan, and Niecu Sebe. 2016. Recognising emotions from abstract paintings using non-linear matrix completion. In IEEE Conf. Comput. Vis. Pattern Recog. IEEE, 5240–5248.

[2] Mehdin Avi, Andrew Zimmerman, and Christoffer Nellaker. 2018. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In Eur. Conf. Comput. Vis. 0–0.

[3] Peter Baldi and Peter Sadowski. 2014. The dropout learning algorithm. Artificial Intelligence 210 (2014), 78–122.

[4] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. 2020. A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms. In Int. Conf. Learn. Represent. https://openreview.net/forum?id=rKyWlWgFpS

[5] Michel Besserve, Arash Mehrjou, Rémy Sun, and Bernhard Schölkopf. 2020. Counterfactuals uncover the modular structure of deep generative models. In Int. Conf. Learn. Represent. https://openreview.net/forum?id=SJhDPxgEKh

[6] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. 2015. Visual causal feature learning. In Conf. Uncertainty Artificial Intell. AUAI Press, Arlington, Virginia, USA, 181–190.

[14] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. 2015. Facial expression recognition: History, trends, and affect-related applications. IEEE Trans. Pattern Anal. Mach. Intell. 38, 5 (2016), 1548–1568.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In IEEE Conf. Comput. Vis. Pattern Recog. IEEE Computer Society, 248–255.

[16] Zhang Dong, Zhang Hanwang, Tan Jinghui, Hua Xiongsheng, and Sun Qianru. 2020. Causal Intervention for Weakly Supervised Semantic Segmentation. In Adv. Neural Inform. Process. Syst.

[17] Amir Hossein Farzaneh and Xiaojun Qi. 2021. Facial expression recognition. In IEEE Conf. Comput. Vis. Image Process. IEEE, 1–4.

[18] Brendan Jou, Subhabrata Bhattacharya, and Shih-Fu Chang. 2014. Predicting viewer perceived emotions in animated GIFs. In ACM Int. Conf. Multimedia. 213–216.

[19] Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. 2020. A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms. In Int. Conf. Learn. Represent. https://openreview.net/forum?id=rKyWlWgFpS

[20] Judea Pearl and Dana Mackenzie. 2018. The Book of Why: The New Science of Cause and Effect. Hachette UK.

[21] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In Conference on Empirical Methods in Natural Language Processing 1532–1543.

[22] Xiuyi Jia, Xiang Zheng, Weiwei Li, Changqing Zhang, and Zechao Li. 2019. Facial motion prior networks for facial expression recognition. In IEEE Conf. Comput. Vis. Pattern Recog. 13014–13023.

[23] Yuanqing Yang, Keqiang Huang, and Hanwang Zhang. 2021. Feature decomposition and reconstruction learning for effective facial expression recognition. In IEEE Conf. Comput. Vis. Pattern Recog. 7660–7669.

[24] Mengyi Liu, Shaoxin Li, Shiguang Shan, and Xilin Chen. 2013. Au-aware deep feature learning. In Adv. Neural Inform. Process. Syst. 8026–8037.

[25] Villa, Olexa Bilaniuk, Anirudh Goyal, and Yoshua Bengio. 2021. Towards Causal Representation Learning. arXiv preprint arXiv:2102.11071 (2021).

[26] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. 2015. Facial expression recognition: History, trends, and affect-related applications. IEEE Trans. Pattern Anal. Mach. Intell. 38, 5 (2016), 1548–1568.

[27] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. Int. Conf. Learn. Represent. (2015).

[28] Ning Sun, Qi Li, Ruizhi Huang, Jinxin Liu, and Guang Han. 2019. Deep spatial-temporal feature fusion for facial expression recognition in static images. Pattern Recogn. Letters 119 (2019), 49–61.

[29] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. 2020. Long-Tailed Classification by keeping the Good and Removing the Bad Momentum Causal Effect. In Adv. Neural Inform. Process. Syst.

[30] Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In IEEE Conf. Comput. Vis. Pattern Recog. IEEE, 1521–1528.

[31] Judea Pearl and Dana Mackenzie. 2018. Learning independent causal mechanisms. In PMLR, 4036–4044.

[32] Judea Pearl et al. 2000. Elements of causal inference. 1532–1543.

[33] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2015. Elements of causal inference. The MIT Press.

[34] Delian Ruan, Yan Yan, Shengli Lai, Zhenhua Chai, Chunhua Shen, and Hanzi Wang. 2021. Feature decomposition and reconstruction learning for effective facial expression recognition. In IEEE Conf. Comput. Vis. Pattern Recog. 7660–7669.

[35] Xiaohui Tang, Jianqiang Huang, and Hanwang Zhang. 2020. Long-Tailed Classification by keeping the Good and Removing the Bad Momentum Causal Effect. In Adv. Neural Inform. Process. Syst.

[36] Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In IEEE Conf. Comput. Vis. Pattern Recog. IEEE, 1521–1528.

[37] Judea Pearl and Dana Mackenzie. 2018. Learning independent causal mechanisms. In PMLR, 4036–4044.

[38] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. Int. Conf. Learn. Represent. (2015).

[39] Ning Sun, Qi Li, Ruizhi Huang, Jinxin Liu, and Guang Han. 2019. Deep spatial-temporal feature fusion for facial expression recognition in static images. Pattern Recogn. Letters 119 (2019), 49–61.

[40] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. 2020. Long-Tailed Classification by keeping the Good and Removing the Bad Momentum Causal Effect. In Adv. Neural Inform. Process. Syst.

[41] Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In IEEE Conf. Comput. Vis. Pattern Recog. IEEE, 1521–1528.

[42] Judea Pearl and Dana Mackenzie. 2018. Learning independent causal mechanisms. In PMLR, 4036–4044.

[43] Judea Pearl and Dana Mackenzie. 2018. Learning independent causal mechanisms. In PMLR, 4036–4044.

[44] Judea Pearl and Dana Mackenzie. 2018. Learning independent causal mechanisms. In PMLR, 4036–4044.

[45] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Marcelo Magnavita, and Alexander Torralba. 2021. Joint Discriminative and Mutual Adaptation Networks for Facial Expression Recognition. In Int. Conf. Mach. Learn. 108355.

[46] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. 2020. Long-Tailed Classification by keeping the Good and Removing the Bad Momentum Causal Effect. In Adv. Neural Inform. Process. Syst.

[47] Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In IEEE Conf. Comput. Vis. Pattern Recog. IEEE, 1521–1528.

[48] Judea Pearl and Dana Mackenzie. 2018. Learning independent causal mechanisms. In PMLR, 4036–4044.

[49] Judea Pearl and Dana Mackenzie. 2018. Learning independent causal mechanisms. In PMLR, 4036–4044.
[51] Xu Yang, Hanwang Zhang, and Jianfei Cai. 2021. Deconfounded image captioning: A causal retrospect. *IEEE Trans. Pattern Anal. Mach. Intell.* (2021).

[52] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*. 381–388.

[53] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Building a large scale dataset for image emotion recognition: the fine print and the benchmark. In *AAAI*. 308–314.

[54] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. 2020. Interventional Few-Shot Learning. In *Adv. Neural Inform. Process. Syst.*

[55] Kaihao Zhang, Yongzhou Huang, Yong Du, and Liang Wang. 2017. Facial expression recognition based on deep evolulational spatial-temporal networks. *IEEE Trans. Image Process*. 26, 9 (2017), 4193–4203.

[56] Tong Zhang, Wenming Zheng, Zhen Cui, Yuan Zong, Jingwei Yan, and Keyu Yan. 2016. A deep neural network-driven feature learning method for multi-view facial expression recognition. *IEEE Trans. Multimedia* 18, 12 (2016), 2528–2536.

[57] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z. Li, and Matti Pietikäinen. 2011. Facial expression recognition from near-infrared videos. *Image and Vision Computing* 29, 9 (2011), 607–619.

[58] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. 2014. Exploring principles-of-art features for image emotion recognition. In *ACM Int. Conf. Multimedia*. 47–56.

[59] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* (2017).