

# Post-Hoc Adversarial Stickers Against Micro-Expression Leakage

Pei-Sze Tan\*, Sailaja Rajanala, Yee-Fan Tan, Arghya Pal, Chun-Ling Tan, Raphaël C.-W. Phan, Huey-Fang Ong  
*CyPhi ( $\Psi\Phi$ ) AI Lab, School of Information Technology, Monash University, Malaysia campus*  
 Email: \*tan.peisze@monash.edu

**Abstract**—Securing micro-expressions against leakage is crucial for privacy, as these subtle facial movements convey genuine emotions and are inherently personal. This study aims to protect micro-expression data from potential adversarial attacks, ensuring the preservation of individuals’ privacy and preventing unauthorized access or misuse of sensitive emotional information. Unlike traditional methods, which often require training and extensive access to models, this research introduces a novel post-hoc method that does not require additional training. We focus on physical adversarial attacks in micro-expression recognition, involving intentional manipulation of visual cues to deceive recognition systems and protect individual emotional privacy. Our approach leverages a causal discovery algorithm to identify causal relationships between facial parts, enabling rapid identification of the optimal locations for adversarial patches in frames with triggered micro-expressions. This method exhibits a more consistent attack success rate than randomly placed adversarial stickers, demonstrating effective generalization across different emotions, stickers, and models. Particularly relevant in scenarios with restricted access to the model, our technique requires only a single interaction during the attack process, highlighting its efficiency and minimal need for querying the target model. The proposed method effectively balances privacy protection with high generalization capability, setting a new standard for defending against adversarial threats in micro-expression recognition. The code is available at <https://github.com/noobasuna/au-sticker>.

**Index Terms**—Adversarial sticker, causality, micro-expression

## I. INTRODUCTION

**Adversarial attacks** play a crucial role in preserving privacy, particularly in areas like face recognition and natural language processing. These attacks introduce subtle perturbations that deceive machine learning models without significantly altering the original content, ensuring that sensitive information remains protected. For example, Zhang et al.[1] propose an Adversarial Privacy-preserving Filter (APF) that generates adversarial noise to mislead face recognition algorithms while maintaining the quality of images. Similarly, Zhao et al.[2] employ a collaborative framework where adversarial gradients are transferred between the user’s device and the cloud, enhancing the system’s ability to confuse face recognition algorithms. Shamshad et al. [3] extend this approach by using adversarial latent codes to protect facial privacy. Their method leverages user-defined makeup prompts in generative models to deceive face recognition systems, achieving high privacy preservation without visually altering the person’s identity. However, while much work has focused on facial recognition, few efforts have explored physical adversarial

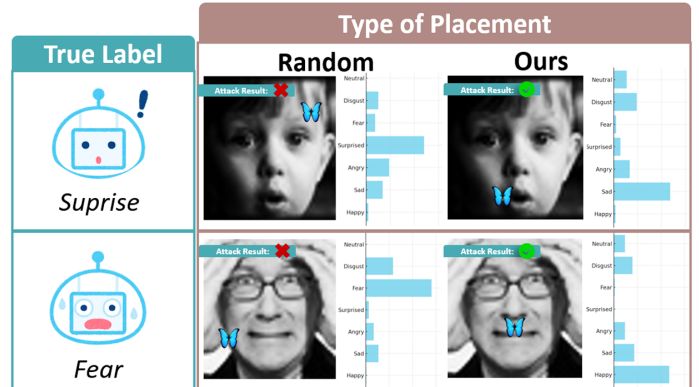


Fig. 1: Adversarial examples on the FER2013 dataset, comparing random placement with our sticker method. The goal is to predict the optimal sticker position that shifts the probability mass from the true expression to neighboring classes, thereby confusing the model and preserving privacy by lowering the likelihood of correct emotion identification.

attacks, particularly in the context of emotion recognition. Low et al. [4] address this gap by introducing AdverFacial, a framework that applies universal adversarial perturbations to deceive automated micro-expression recognition systems.

**Physical adversarial attacks** offer a compelling solution for protecting emotional privacy, particularly in video-based applications like virtual meetings and surveillance. Unlike digital attacks, which can be complex and computationally intensive, physical adversarial attacks are faster, less complex, and easily extendable to different datasets with examples given in Fig. I. They work by introducing physical perturbations, such as stickers, to obscure facial regions, thus preventing recognition systems from detecting micro-expressions.

Micro-expressions, which are brief and involuntary facial movements lasting only a fraction of a second [5], pose a unique challenge for recognition systems. State-of-the-art techniques, including convolutional neural networks (CNNs)[6] and graph neural networks (GNNs)[7], have been developed to detect these fleeting expressions. However, physical adversarial attacks, such as location-aware stickers, present an effective way to disrupt these systems by targeting key facial regions.

Our work leverages physical perturbations to exploit causal relationships identified through graph-based methods for micro-expression recognition. By targeting the most influential nodes in the causal graph, those with the highest outdegree, which are likely common causes for downstream nodes, we

can effectively occlude or manipulate these critical areas. Altering or intervening on these key causal nodes disrupts the processes that lead to the manifestation of emotions, thereby challenging the recognition of emotional states and enhancing privacy protection. This approach is more efficient than training numerous adversarial examples and provides a practical solution for real-world applications where quick and effective privacy protection is needed [4][8][9].

To address these challenges, we proposed this post-hoc method with the contributions of:

- **Introduction of Physical Attacks for Emotion Concealment:** We utilize adversarial stickers to physically obscure facial expressions, protecting emotional privacy.
- **Post-Hoc Causal Relationship Analysis:** We apply conditional independence tests to identify causal relationships among facial regions and determine optimal sticker placement.
- **Cross-domain Effectiveness and One-time query:** Our method achieves high attack success rates across various datasets, models, and adversarial stickers with minimal computational resources, demonstrating its efficiency in real-world scenarios.

## II. METHODOLOGY

Our aim is to identify weaknesses in the classifier’s design, particularly its vulnerability to sticker-based adversarial attacks. We focus on this without employing any adversarial training processes or optimizing the patch location.

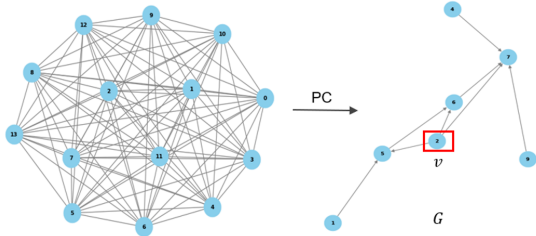


Fig. 2: Illustration of a fully connected undirected graph during the skeleton discovery process using the Peter-Clark (PC) algorithm. Each node represents a Region of Interest (ROI). The resulting causal graph, denoted as  $G$ , highlights the vertex with the most outgoing edges,  $v$ , in a red box.

**Causal Discovery-based Location Optimization:** To determine the most effective location for placing an adversarial sticker, we start by exploring the causal relationships within facial Action Units (AUs). These AUs represent various combinations of facial muscle movements that correlate with specific emotions. This exploration involves identifying 14 key Regions of Interest (ROIs) on the face, based on the method proposed in [10]. Each ROI corresponds to specific AUs, which are muscle movements that commonly occur during micro-expressions [11].

Formally, given the dataset as  $\mathcal{D} = \{(x_i, \{AU_i\})\}_{i=1}^N$ , where  $x_i$  represents facial features in pixel space, and  $AU_i$  corresponds to the set of activated AUs in  $x_i$ . From each  $x_i$ , we extract 14 ROIs, each containing one or more activated AUs. This process allows us to map an image to its respective

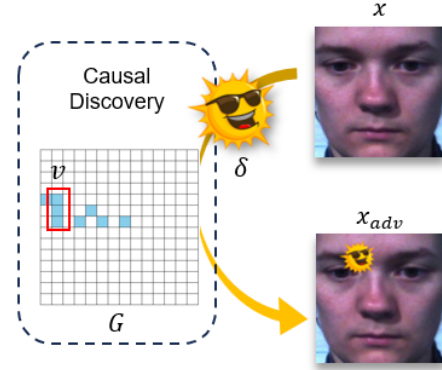


Fig. 3: The **upper eyebrow region**,  $v$ , is found to be the most influential node (parent) in the causal graph  $G$  due to its highest **outdegree**, indicating numerous outgoing causal connections. Placing the **sticker** here effectively fools the model. Best view in color.

active AUs. We use a deep learning model  $\mathcal{M}(\phi)$  to transform raw facial features  $x_i$  into a feature vector  $f_i \in \mathbb{R}^d$ , where  $x_i \rightarrow f_i$  and  $d$  denotes the dimensionality of the feature space. We need to find a causal relationship graph  $G = (V, E)$  between the facial ROIs.

To achieve this, we apply the PC algorithm [12] to the facial regions adjacency graph, identifying causal relationships between the ROIs based on observational data. The PC algorithm differentiates causation from correlation, offering insights into the mechanisms behind emotional expressions. By using  $f_j$  as input and performing the Chi-square test of independence, we derive a causal graph. For a more detailed mathematical explanation, please refer to [12].

Let  $G = (V, E)$  be the resulting causal graph from the PC algorithm, where each node  $v_i \in V$  represents a Region of Interest (ROI) in the facial feature space, and each directed edge  $(v_i, v_j) \in E$  indicates a causal relationship between two ROIs. In this context,  $v_i \rightarrow v_j$ , signifying that the activation or movement in ROI  $v_i$  influences the behavior of ROI  $v_j$ . The outdegree  $d(v_i)$  of a node  $v_i$  corresponds to the number of outgoing edges, representing how much influence that particular ROI exerts on others. We define  $v^*$  as the node in  $V$  with the highest outdegree, which corresponds to the most influential ROI, potentially serving as a causal parent to multiple downstream ROIs.

$$v^* = v_i \in V : \forall v_j \in V, d(v_i) \geq d(v_j) \quad (1)$$

Equation 1 ensures that  $v^*$  is the node with the maximum outdegree in  $G$ . Fig.2 illustrates this process.

**Placement of Adversarial Patch:** The goal, as shown in Fig. 3, is to create an adversarial image  $x_{adv}$  that causes misclassification [9] of the original image,  $x$  using an adversarial perturbation  $\delta$ :

$$x_{adv} = (1 - \mathbf{A}) \odot x + \mathbf{A} \odot \delta \quad (2)$$

By placing the sticker, we effectively perform a do-calculus intervention, denoted as  $\text{do}(v = \delta)$ , on the most significant

causal node, thereby intervening on the cause  $\rightarrow$  effect link. In this case, the cause is a facial part, and by occluding or blocking it, we disrupt the resulting effect, which is the expression of emotion. This approach aligns with concepts from Pearl’s theory of causality [13], where interventions on causal relationships are modeled. Here, the binary matrix  $\mathbf{A}^{W \times H}$  serves as a mask for the patch area, containing the information about the patch location. The variables  $W$  and  $H$  represent the width and height of the image, respectively. The adversarial patch  $\delta$  is placed at the coordinates corresponding to the identified node  $\mathbf{v}$ , as determined above. Instead of generating dynamic adversarial patches, our method uses fixed pattern stickers as perturbations. These stickers, as shown in Table II, are positioned at the most influential points on the face identified by our causal analysis. The patch size is consistent throughout the experiments. The current sticker size is determined by ablation studies, optimizing for the most effective attack results while ensuring it minimally impacts the overall facial features, maintaining a natural appearance on the human face. When pasting it on an action units area, we did not cover the whole region but only placed it centrally within the AU region, aligning the perturbation with the key causal points.

### III. EXPERIMENTAL DETAILS

#### A. Datasets and Patch Pattern

The experiments are conducted with CASMEII [14], SAMM [15], and SMIC [16] datasets. These datasets are all equipped with fixed cameras and lightning to minimize noise disturbance in the environment. Video stimuli were used to induce the micro-expressions from the participants. These samples were classified into three classes: positive, negative, and surprise. We conducted experiments on a diverse variety of stickers. A few sticker examples (Table II(a) and (b)) from [9] and added some stickers we found that are more likely to be pasted on human faces during events or situations (Table II(c) and (d)). While emoji stickers (Table II(e) and (h)) are tested on the target model to use them as easy-to-reproduce and non-threatening stickers. These stickers could be substituted with tattoos and body painting in real-life situations.

TABLE I: Total queries in baselines that learn an optimal patch or location vs. our method that uses causal inference to determine the most impactful location to place the patch.

	Number of Queries
[9], [17], [18]	Number of Iterations $\times$ Population Size
Ours	Number of Images

#### B. Target Model and Evaluation Settings

The model we are attacking is the state-of-the-art micro-expressions recognition model, Muscle motion-guided network (MMNet) [19] and Spatio-Temporal Convolutional Neural Networks (STCNN) [6]. Microexpression recognition datasets have a small sample size. MMNet employs a dual-branch paradigm, utilizing a continuous attention block to extract muscle motion patterns and integrating a position calibration module to generate position embeddings. The

STCNN incorporates complete spatial information through the utilization of a 3D-CNN. To avoid overfitting, the evaluation protocol used in our experiments is leave-one-subject-out (LOSO) validation on single-database settings. The type of physical attack we add to the testing inputs are the adversarial stickers, which could be referred to [9]. Meanwhile, we are using the 2D pasting method to add the sticker patch on the subjects’ faces. In this work, we are only testing this attack in the digital world.

**Attack Success Rate (ASR)** measures the effectiveness of an adversarial perturbation in causing misclassification or manipulation of the model’s output. A high attack success rate indicates the potency of the crafted adversarial example in deceiving the model.

#### C. Results and Discussion

The effectiveness of our method, which strategically places adversarial stickers based on learned causal knowledge, is demonstrated in Table II by comparing it the sticker placement strategy in [8]. We compare with only one method as existing placing adversarial stickers approaches are limited to single-image analysis and do not suit multi-frame micro-expressions. Also, our post-hoc method uniquely protects without revisiting the model, which other methods cannot do. However, in the current study, we focus on MER, using a dataset-independent causal graph that delivers consistent and comparable improvements across datasets. It is very likely that the observed differences between databases can be attributed to biases, which could require a debiasing strategy such as [20].

Causal sticker placement consistently outperforms random placement, exploiting the model’s vulnerabilities more effectively. Our approach, incorporating causal knowledge, achieves a higher success rate in tricking micro-expression recognition models compared to other methods [9] and [17], which focus on face verification model augmentation.

Table I compares the number of iterations required to identify the correct ROI (1 in our case) without location optimization or distribution updating. Our method demonstrates high attack performance even with a lower query number, highlighting its robustness and efficiency for potential real-world deployment.

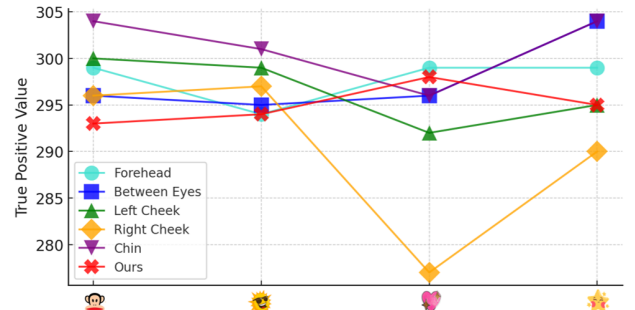










Fig. 4: The true positive count (TP  $\downarrow$ ) of each sticker and each position on the face with MMNet are recorded. The lower the TP, the better. Our method’s ability to perform well across various sticker patterns showcases its generalization capabilities, reinforcing its efficacy in diverse scenarios.

TABLE II: Comparative studies of our method of causally placing the adversarial sticker vs placement employed in [8] with Attack Success Rate (ASR  $\uparrow$ ) as the evaluation metric. The value in **bold** shows higher in ASR. *On average there is a 12.13% increase in ASR for MMNet and a 7.44% increase in ASR for STCNN.*

Sticker	Region	MMNet [19]			STCNN [6]		
		CASMEII	SAMM	SMIC	CASMEII	SAMM	SMIC
(a) 	[8]	2.56 $\pm$ 0.64	31.58 $\pm$ 0.75	63.06 $\pm$ 1.91	40.38 $\pm$ 2.56	31.58 $\pm$ 7.52	64.33 $\pm$ 4.46
	Ours	<b>3.85</b>	<b>32.33</b>	<b>66.24</b>	<b>42.31</b>	<b>37.59</b>	<b>69.43</b>
(b) 	[8]	3.85 $\pm$ 0.64	31.57 $\pm$ 1.50	62.42 $\pm$ 1.27	41.67 $\pm$ 3.21	35.34 $\pm$ 3.76	61.15 $\pm$ 2.55
	Ours	<b>5.77</b>	<b>32.33</b>	<b>63.69</b>	<b>41.03</b>	<b>36.09</b>	<b>74.52</b>
(c) 	[8]	2.56 $\pm$ 0.64	30.83 $\pm$ 2.56	62.42 $\pm$ 0.64	40.38 $\pm$ 1.28	33.08 $\pm$ 0.75	59.87 $\pm$ 3.82
	Ours	<b>9.62</b>	<b>33.02</b>	<b>63.69</b>	<b>42.31</b>	<b>33.08</b>	<b>61.78</b>
(d) 	[8]	2.56 $\pm$ 0.64	32.33 $\pm$ 1.50	62.42 $\pm$ 0.64	42.95 $\pm$ 3.21	<b>35.34 <math>\pm</math> 0.75</b>	63.69 $\pm$ 7.01
	Ours	<b>3.21</b>	<b>32.33</b>	<b>65.61</b>	<b>41.03</b>	32.33	<b>64.97</b>
(e) 	[8]	3.20 $\pm$ 1.28	<b>43.59 <math>\pm</math> 1.50</b>	63.06 $\pm$ 1.91	34.59 $\pm$ 10.92	33.83 $\pm$ 3.76	57.96 $\pm$ 3.82
	Ours	<b>3.85</b>	41.67	<b>64.33</b>	<b>37.82</b>	<b>31.58</b>	<b>64.97</b>
(f) 	[8]	3.21 $\pm$ 1.92	41.67 $\pm$ 0.75	62.42 $\pm$ 0.64	40.38 $\pm$ 5.79	36.84 $\pm$ 0.75	58.60 $\pm$ 0.64
	Ours	<b>3.21</b>	<b>41.67</b>	<b>63.06</b>	<b>44.23</b>	<b>36.09</b>	<b>61.78</b>
(g) 	[8]	2.56 $\pm$ 0.64	41.67 $\pm$ 0.75	63.69 $\pm$ 1.27	41.67 $\pm$ 10.1	<b>33.08 <math>\pm</math> 0.75</b>	61.15 $\pm$ 5.10
	Ours	<b>3.85</b>	<b>41.67</b>	<b>63.69</b>	<b>40.38</b>	31.58	<b>61.78</b>
(h) 	[8]	3.21 $\pm$ 0.64	41.67 $\pm$ 0.75	63.01 $\pm$ 1.27	37.82 $\pm$ 0.98	31.58 $\pm$ 5.26	61.15 $\pm$ 4.46
	Ours	<b>4.49</b>	<b>42.31</b>	<b>64.33</b>	<b>45.51</b>	<b>31.58</b>	<b>64.97</b>

By emphasizing the comparable success rates in these non-traditional scenarios, we highlight the adaptability of our method. From Fig 4, this comparison is a compelling illustration of the consistency of true positive value across different stickers. By carefully articulating the causal knowledge behind our approach and its consistent success across unconventional scenarios, we strengthen the argument that our method not only stands out in specific contexts but also excels in generalization, making it suitable for a wide range of sticker patterns.

#### D. Patch Size Ablation Studies.

The most effective patch size was determined through ablation studies. We provide in Table III to demonstrate how different patch sizes affect effectiveness. The patch size is consistently set in the experiments as  $S = \left(\frac{kW}{10}, \frac{kH}{10}\right)$ , where  $k$  is a configurable parameter, with  $k = 2$  being the default setting. We only include  $k$  as 1 and 3 as larger patches cover too much of the face, making them less effective or impractical for real-life applications.

k	MMNet			STCNN		
	CASMEII	SAMM	SMIC	CASMEII	SAMM	SMIC
1	2.56	33.02	62.42	37.82	35.34	59.87
3	2.56	32.33	63.69	40.38	40.62	64.33
2	<b>3.85</b>	<b>32.33</b>	<b>66.24</b>	<b>42.31</b>	<b>37.59</b>	<b>69.43</b>

TABLE III: Attack success rate for different patch sizes by using the sticker (a) in Table II. The default setting results are in **bold**.

#### E. Causal Analysis in Different Facial Regions

Causal analysis in different facial regions identifies the most influential areas for expressions, enhancing recognition accuracy and reducing bias. It helps models focus on relevant regions, improves interpretability, and ensures reliable, fair analysis across diverse populations. The differences in influential ROIs between CASMEII and SAMM are due to distinct

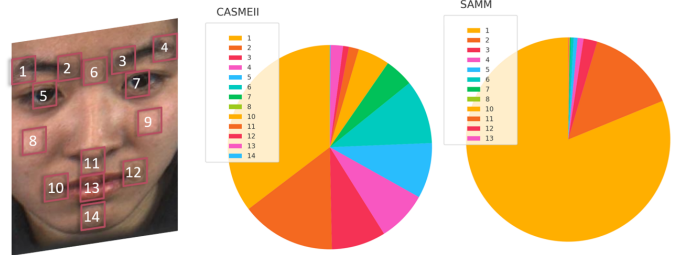


Fig. 5: The pie charts illustrate the distribution of influential Regions of Interest (ROIs) of each image in the CASMEII and SAMM datasets. The facial images on the left show the corresponding facial regions to paste the stickers.

causal relationships in each dataset shown in Fig. 5. CASMEII captures a broader range of ROIs, reflecting complex interactions of facial movements from diverse expressions and subjects. In contrast, SAMM's fewer impactful ROIs suggest more localized influences, likely due to a narrower focus on specific or uniform expressions. These variations underscore the importance of context-specific analysis in understanding micro-expressions. Our method could capture these subtle expressions and generalize them across different datasets and subjects which also reflected in our results.

#### IV. CONCLUSION

We introduce a novel post-hoc approach that uses causal discovery to strategically place adversarial stickers to attack the micro-expression recognition model, enhancing the security of micro-expression recognition without additional training. This novel approach promises to strengthen the security of micro-expression analysis across various applications and consistently outperforms random placements, generalizing well across emoticons, stickers, and models while requiring minimal queries. This efficient, one-time query approach sets a new benchmark for defending against adversarial threats in micro-expression analysis.



## REFERENCES

- [1] Jiaming Zhang, Jitao Sang, Xian Zhao, Xiaowen Huang, Yanfeng Sun, and Yongli Hu, "Adversarial privacy-preserving filter," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1423–1431.
- [2] Xian Zhao, Jiaming Zhang, and Xiaowen Huang, "Apf: An adversarial privacy-preserving filter to protect portrait information," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2813–2815.
- [3] Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar, "Clip2protect: Protecting facial privacy using text-guided makeup via adversarial latent search," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20595–20605.
- [4] Yin-Yin Low, Angeline Tanvy, Raphaël C.-W. Phan, and Xiaojun Chang, "Adverfacial: Privacy-preserving universal adversarial perturbation against facial micro-expression leakages," in *ICASSP*, 2022, pp. 2754–2758.
- [5] Hong-Xia Xie, Ling Lo, Hong-Han Shuai, and Wen-Huang Cheng, "Au-assisted graph attention convolutional network for micro-expression recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2871–2880.
- [6] Reddy Sai Prasanna Teja, Karri Surya Teja, Shiv Ram Dubey, and Snehasis Mukherjee, "Spontaneous facial micro-expression recognition using 3d spatiotemporal convolutional neural networks," *International Joint Conference on Neural Networks*, 2019.
- [7] Shu-Min Leong, Fuad Noman, Raphaël C.-W. Phan, Vishnu Monn Baskaran, and Chee-Ming Ting, "Graphex: Facial action unit graph for micro-expression classification," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 3296–3300.
- [8] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer, "Adversarial patch," *arXiv preprint arXiv:1712.09665*, 2017.
- [9] Xingxing Wei, Ying Guo, and Jie Yu, "Adversarial sticker: A stealthy attack method in the physical world," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [10] Walied Merghani and Moi Hoon Yap, "Adaptive mask for region-based facial micro-expression recognition," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 2020, pp. 765–770.
- [11] Paul Ekman and Erika L. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*, Oxford University Press, 04 2005.
- [12] Peter Spirtes, Clark N Glymour, and Richard Scheines, *Causation, prediction, and search*, MIT press, 2000.
- [13] Judea Pearl, "The do-calculus revisited," *arXiv preprint arXiv:1210.4852*, 2012.
- [14] Karen L Schmidt and Jeffrey F Cohn, "Dynamics of facial expression: Normative characteristics and individual differences," in *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001*. IEEE, 2001, pp. 547–550.
- [15] Adrian K Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap, "Samm: A spontaneous micro-facial movement dataset," *IEEE transactions on affective computing*, vol. 9, no. 1, pp. 116–129, 2016.
- [16] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg)*. IEEE, 2013, pp. 1–6.
- [17] Xingxing Wei, Shouwei Ruan, Yinpeng Dong, and Hang Su, "Distributional modeling for location-aware adversarial patches," 2023.
- [18] Xingxing Wei, Yao Huang, Yitong Sun, and Jie Yu, "Unified adversarial patch for cross-modal attacks in the physical world," 2023.
- [19] Hanting Li, Mingzhe Sui, Zhaoqing Zhu, and Feng Zhao, "Mmnet: Muscle motion-guided network for micro-expression recognition," *arXiv preprint arXiv:2201.05297*, 2022.
- [20] Pei-Sze Tan, Sailaja Rajanala, Arghya Pal, Shu-Min Leong, Raphaël C-W Phan, and Huey Fang Ong, "Causally uncovering bias in video micro-expression recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 5790–5794.