CAUSAL PLACEMENT OF ADVERSARIAL STICKERS AGAINST EMOTION LEAKAGE

*Pei-Sze Tan**, *Sailaja Rajanala, Arghya Pal, Chun-Ling Tan, Raphaël C.-W. Phan, Huey-Fang Ong*

CyPhi ($\Psi\Phi$) AI Research Lab, School of IT, Monash University, Malaysia campus

ABSTRACT

Securing micro-expressions against leakage is paramount for privacy as these subtle facial movements convey genuine emotions, making them inherently personal. Protecting micro-expression data from potential adversarial attacks ensures the preservation of individuals' privacy, preventing unauthorized access and misuse of sensitive emotional information. This study focuses on physical adversarial attacks in micro-expression recognition, involving intentional manipulation of visual cues to deceive automatic recognition systems that would otherwise violate individual emotional privacy. Exploiting subtle micro-expressions and pasting patches to cover genuine emotions, this research introduces a novel approach, AU-Sticker. Leveraging the Peter-Clark algorithm, our methodology rapidly identifies optimal locations for adversarial patches in frames with triggered micro-expressions. Integrating advanced computer vision techniques with causal reasoning based on action units, AU-Sticker enhances the security against micro-expression leakage and exhibits a more consistent attack success rate than randomly placed adversarial stickers. Quantifying based on query numbers involves assessing the number of attempts needed to learn an optimal adversarial patch. This is particularly relevant in scenarios where attackers may encounter restricted access to query the model. Our method is more effective and has a good generalization with a lower query number, involving just a single interaction with the target model during the adversarial attack process.

Index Terms— Micro-expression recognition, facial action units, physical attacks, adversarial sticker, causality

1. INTRODUCTION

Facial tattoos and stickers have emerged as fashionable aesthetic choices. In various contexts, individuals adopt facial tattoos for accessories or religious reasons, while social media filters apply cartoon stickers to videos or images, reflecting the prevalence of such enhancements in real-life situations.

Interestingly, these actually pose a challenge in micro-expression recognition due to their transient nature. Notably, micro-expressions are hidden emotions that people show involuntarily. These fleeting facial expressions, lasting only a fraction of a second [1], often differ from conventional recognition methodologies.

State-of-the-art approaches in micro-expression recognition typically leverage deep learning techniques, utilizing convolutional neural networks (CNNs) [2] and graph neural networks [3] to capture and analyze subtle facial movements. While these methods have shown promising results in controlled environments, they often fall short when faced with adversarial attacks. The difficulty in micro-expression recognition is further compounded by adversarial stickers, which can be strategically placed on the face to deceive recognition systems. Furthermore, the challenge of placing adversarial stickers in the expression recognition model is vital for ensuring the optimal resilience to expression leakage in real-world scenarios. Current research lacks a comprehensive solution that can effectively locate adversarial stickers in the context of facial expression recognition.

Adversarial patches were designed to study the security, privacy and robustness aspects of the deep learning models. Understanding the weaknesses of machine learning models will foster the development of more robust systems addressing these vulnerabilities. Adversarial stickers proposed in [4] represents a physically plausible and inconspicuous attack technique utilizing readily available stickers found in everyday life. The real-world implications of successful adversarial attacks include potential security breaches in applications relying on accurate emotion recognition, emphasizing the importance of ongoing research in securing these systems [5]. Low et al. showed a novel universal adversarial perturbation approach in micro-expression analysis [6] and multimodal emotion recognition task [7] with low perceptibility and high adversarial transferability. These methods above are non-physical attacks that humankind could not see with bare eyes, which are imperceptible and ideal for stealthy digital manipulation. In our work, we considered the actuality and diversity of human society by applying the attacks by visible patches.

In this paper, we propose a novel method named AU-Sticker, which locates adversarial stickers with causal knowledge of facial action units (AUs) [8] in micro-expression detection tasks to address the challenges above. Our approach combines advanced computer vision techniques with causal reasoning to accurately identify the optimal location of pasting adversarial stickers on human face in one iteration without

*Corresponding Author

any additional optimization or training.

Our contribution is using the physical attack (i.e., adversarial stickers) to help hide the genuine emotions of a subject effectively by applying the conditional independence tests method to find the causal relation of the facial action units while considering the causal sufficiency. We demonstrate an advanced framework capable of effectively deceiving both macro and micro-expression recognition systems using physical adversarial attacks. This framework boasts a higher success rate in attacks while requiring fewer queries for optimal results.

2. METHODOLOGY

Our objective is to uncover vulnerabilities inherent the classifier's design, specifically focusing on susceptibility to sticker-based adversarial attacks without any adversarial training process and patch location optimization. Fig.1 illustrates an overview of our framework.

2.1. Causal Relation Discovery between Facial Action Unit

To determine the optimal location for a sticker, we initiate the process by investigating the causal relationships within facial action units (AUs). This exploration involves identifying 14 crucial Regions of Interest (ROIs) on the face, following the methodology proposed in [9]. Each ROI corresponds to specific facial action units (AUs) [8] commonly triggered during micro-expressions. The mentioned AUs represent various movements of facial muscles. Specific combinations of these facial muscle movements correspond to the expression of particular emotions.

We represent the dataset as $\mathcal{D} = \{(x_i, \mathrm{AU}_i)\}_{i=1}^{N}$, where x_i denotes facial features in pixel space, and AU_i is the corresponding set of activated action units (AUs). From x_i, we extract 14 Regions of Interest (ROIs), each comprising one or a few activated AUs. This enables us to map an image to its respective active action units. Subsequently, we employ a deep learning model $\mathcal{M}(\phi)$, denoted as $\mathcal{M}(\phi)$:$x_i \rightarrow f_i$, where $f_i \in R^d$, to transform raw facial features into a suitable format for causal inference. Here, $\mathcal{M}(\phi)$ is the deep learning model, and d represents the dimensionality of the transformed feature space.

By passing the action unit adjacency graph to the Peter-Clark (PC) algorithm [10], we identify causal relationships between AUs from our observational data. By distinguishing causation from correlation, it helps us understand the mechanisms behind emotional expressions. By applying PC algorithm with f_j as input, we will have a causal graph after performing the Chi-square test of independence. Refer to [10] for more detailed mathematical elaboration.

Let G be the resulting causal graph from PC algorithm, where nodes represent variables (each ROI feature representations) and edges indicate causal relationships. Let V represent the set of all nodes corresponding to ROIs in G, where v_i denotes a specific node in V, and $d(v_i)$ is the outdegree of node v_i. We define v as the node in V with the highest outdegree, representing the node that exerts the greatest influence on other nodes in the network:

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

Here, Equation 1 ensures that v is the node with the maximum outdegree in the G. Fig.2 shows an example of this process.

2.2. Placement of Adversarial Patch

From Fig. 3, the goal is to create an adversarial image x_{adv} that causes misclassification.

$$x_{adv} = (1 - (A, \theta)) \odot x + (A, \theta) \odot \delta \qquad (2)$$

The image x is combined with an adversarial patch δ resulting in the final perturbed image x_{adv}. Here, the binary matrix $A^{W \times H}$, serves as a mask for the patch area i.e., it contains the information of the patch location. W is the width and H is the height of the image. Our method refrains from generating the adversarial patch δ; instead and instead chooses fixed pattern stickers as perturbations as illustrated in Table 1. The sticker δ is positioned at the coordinates corresponding to the identified node v_i as determined above. The patch size is consistently set in the experiments, specifically as $S = \left(\frac{2W}{10}, \frac{2H}{10}\right)$.

3. EXPERIMENTAL DETAILS

3.1. Datasets and Patch Pattern

The experiments are conducted with CASMEII [13], SAMM [14], and SMIC [15] 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. We utilized 26 subjects with 250 video samples from the CASMEII dataset, 159 videos from 29 participants in the SAMM dataset, and 157 videos from the SMIC dataset. 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 1(a) and (b)) from [4] and added some stickers we found that are more likely to be pasted on human faces during events or situations (Table 1(c) and (d)). These stickers could be substituted with tattoos and body painting in real-life situations.

3.2. Attack Model and Evaluation Settings

The model we are attacking is the state-of-the-art micro-expressions recognition model, Muscle motion-guided network (MMNet) [11] and Spatio-Temporal Convolutional

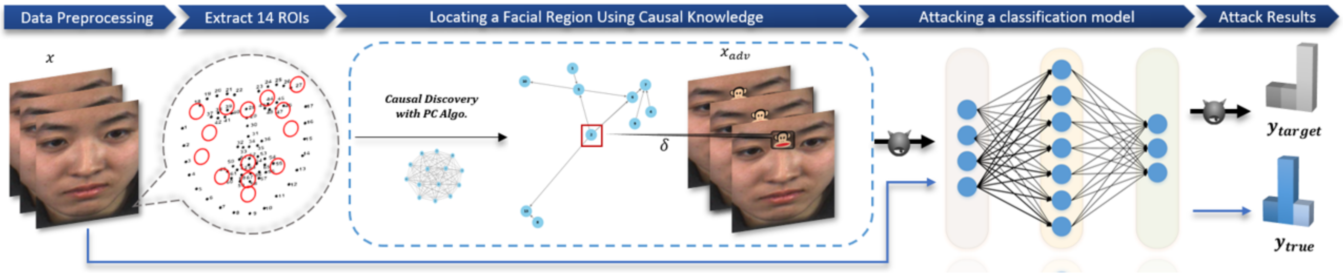

Data Preprocessing
Extract 14 ROIs
Locating a Facial Region Using Causal Knowledge
Attacking a classification model
Attack Results
x
Causal Discovery with PC Algo.
x_{adv}
δ
y_{target}
y_{true}

Fig. 1. The physical attack involves placing adversarial stickers to target micro-expression recognition models. Initially, we extract causal relations between facial muscles, specifically known as action units. Subsequently, the sticker is positioned on the most influential action unit, often the one acting as the cause to all other significant muscles in the face. We have named this proposed method as AU-Sticker.

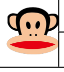
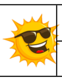
Sticker	Region	MMNet [11]			STCNN [2]		
		CASMEII	SAMM	SMIC	CASMEII	SAMM	SMIC
(a) 	[12]	2.56 ± 0.64	31.58 ± 0.75	63.06 ± 1.91	40.38 ± 2.56	31.58 ± 7.52	64.33 ± 4.46
	Ours	**3.85**	**32.33**	**66.24**	**42.31**	**37.59**	**69.43**
(b) 	[12]	3.85 ± 0.64	31.57 ± 1.50	62.42 ± 1.27	**41.67 ± 3.21**	35.34 ± 3.76	61.15 ± 2.55
	Ours	**5.77**	**32.33**	**63.69**	41.03	**36.09**	**74.52**
(c)	[12]	2.56 ± 0.64	30.83 ± 2.56	62.42 ± 0.64	40.38 ± 1.28	33.08 ± 0.75	59.87 ± 3.82
	Ours	**9.62**	**33.02**	**63.69**	**42.31**	**33.08**	**61.78**
(d)	[12]	2.56 ± 0.64	32.33 ± 1.50	62.42 ± 0.64	**42.95 ± 3.21**	**35.34 ± 0.75**	63.69 ± 7.01
	Ours	**3.21**	**32.33**	**65.61**	41.03	32.33	**64.97**

Table 1. Comparative studies of our method of causally placing the adversarial sticker vs placement employed in [12] with Attack Success Rate (ASR ↑) as the evaluation metric. The value in **bold** shows higher ASR.

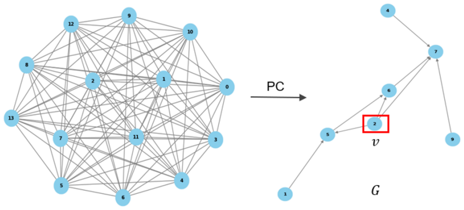


Fig. 2. Consider a fully connected undirected graph undergoing skeleton discovery using the PC algorithm, each node represents an ROI. The outcome is a causal graph, denoted as G, wherein the vertex with the most outgoing edges v is highlighted in a red box.

	Number of Queries
[4, 16, 17]	Number of Iterations × Population Size
AU-Sticker	Population Size

Table 2. 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.

Neural Networks (STCNN) [2]. 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 [4]. 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. The evaluation metric is the Attack Success Rate (ASR). 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.

3.3. Results and Discussion

The effectiveness of our method, which strategically places adversarial stickers based on learned causal knowledge, is demonstrated in Table 1 by comparing it the sticker placement

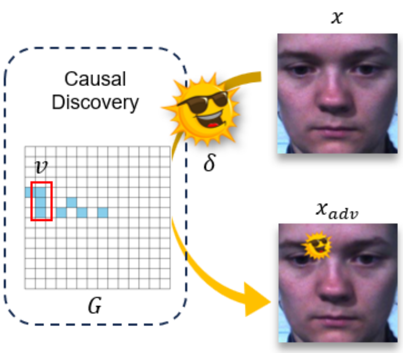


Fig. 3. We propose a novel method that enhances the effectiveness of location-aware patches by discovering the causal relations among facial regions. Here, v is recorded as the most influential part by count, which also happens to be the parent node in all the relations in the final directed causal graph G from the PC algorithm.

strategy in [12]. 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 [4] and [16], which focus on face verification model augmentation.

Table 2 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.

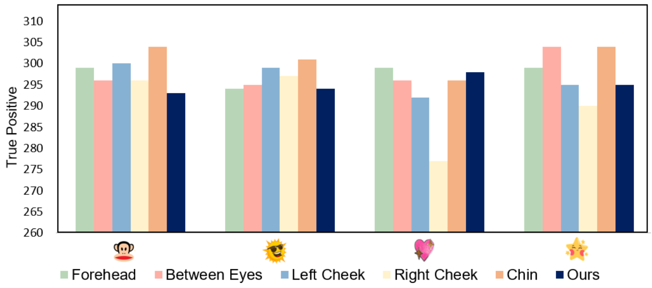


Fig. 4. The true positive count (TP ↓) of each sticker and each position on the face with MMNet are recorded. The ability of our method to perform well across various sticker patterns showcases its generalization capabilities, reinforcing its efficacy in diverse scenarios.

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 broader applicability of our proposed solution beyond conventional contexts. 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 applications.

Note that in this work, we only focus on the stickers' location and apply them with a fixed sticker size. When pasting it on an action units area, we did not cover the whole region but only placed it in the middle of the AU region. The consideration of patch pattern variation is still in our exploration.

3.4. Applying Physical Attacks on Macro Expression Recognition Model

Besides the micro-expression detection task, we also performed the adversarial sticker in Table 1(a) attack by placing them with our proposed position localization method on typical images of facial emotion, which are fairly noticeable on the facial expression recognition (FER) model.

	Clean	Random	**AU-Sticker (Ours)**
ACC	0.58	0.51	**0.50**

Table 3. The accuracy (ACC ↓) of classification results on FER model.

With the experiment on datasets, FER2013 [18] in CNN VGG-16 classification model [19], we computed the instances of successful testing cases when we pasted the stickers with our proposed method prove that placing the adversarial stickers performed well in the macro expression images, where placing stickers in both random and causal ways could attack the model efficiently with evidence of lower true positives than the clean testing datasets. The slight 1% drop in accuracy with causal attacks compared to random ones can be explained by the small size of the input images. These small images make it easier for random changes to accidentally affect crucial facial features, leading to a subtle but noticeable impact on accuracy.

4. CONCLUSION

In this work, we present an approach named AU-Sticker, designed to fool the micro-expression recognition classification model by applying a causal discovery algorithm to locate the adversarial stickers. This novel approach promises to strengthen the security of micro-expression analysis across various applications and demonstrates an increased attack success rate compared to randomly placed adversarial stickers. Moreover, the versatility of this framework extends to macro-expression recognition models, showcasing its applicability in broader contexts.

5. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay, "A survey on adversarial attacks and defences," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 1, pp. 25–45, 2021.

[6] 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.

[7] Yin-Yin Low, Raphaël C.-W. Phan, Arghya Pal, and Xiaojun Chang, "Usurp: Universal single-source adversarial perturbations on multimodal emotion recognition," in *2023 IEEE International Conference on Image Processing (ICIP)*, 2023, pp. 2150–2154.

[8] 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.

[9] 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.

[10] Peter Spirtes, Clark N Glymour, and Richard Scheines, *Causation, prediction, and search*, MIT press, 2000.

[11] Hanting Li, Mingzhe Sui, Zhaoqing Zhu, and Feng Zhao, "Mmnet: Muscle motion-guided network for micro-expression recognition," *arXiv preprint arXiv:2201.05297*, 2022.

[12] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer, "Adversarial patch," *arXiv preprint arXiv:1712.09665*, 2017.

[13] 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.

[14] 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.

[15] 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.

[16] Xingxing Wei, Shouwei Ruan, Yinpeng Dong, and Hang Su, "Distributional modeling for location-aware adversarial patches," 2023.

[17] Xingxing Wei, Yao Huang, Yitong Sun, and Jie Yu, "Unified adversarial patch for cross-modal attacks in the physical world," 2023.

[18] Ian J. Goodfellow, Dumitru Erhan, and et al. Carrier, Pierre Luc, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing*, Berlin, Heidelberg, 2013, pp. 117–124, Springer Berlin Heidelberg.

[19] Yann LeCun, Yoshua Bengio, et al., "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, pp. 1995, 1995.