ADVERFACIAL: PRIVACY-PRESERVING UNIVERSAL ADVERSARIAL PERTURBATION AGAINST FACIAL MICRO-EXPRESSION LEAKAGES

†*Yin-Yin Low,* †*Angeline Tanvy,* †*Raphaël C.-W. Phan,* ‡*Xiaojun Chang*

†School of IT, Monash University, Malaysia campus
‡ReLER Lab, University of Technology Sydney, Australia

ABSTRACT

Privacy safeguards are crucial, notably now with increased virtual conferencing usage during the Covid pandemic. In contrast to conventional facial expressions that are visually obvious to humans, micro-expressions are involuntary and transient facial expressions, commonly manifested involuntarily when we aim to withhold our emotions. Advanced micro-expression recognition techniques exist that can reveal the genuine emotions that people attempt to conceal, thus threatening individual emotional privacy, as fundamental human rights would dictate that one should have a choice of what emotion is being shown or not shown. We propose the novel universal adversarial perturbation-based approach - AdverFacial - for privacy concealment against automated micro-expression analysis via deep learning techniques. We derive the optimal strategy to achieve micro-expression misclassification with a high success rate, low perceptibility and cross neural network transferability. We perform experiments on two popular datasets with state-of-the-art microexpression spotting and recognition models and demonstrate our approach's effectiveness in emotional concealment.

***Index Terms*—** Micro-expression recognition, withholdment, emotional privacy, universal adversarial patterns

1. INTRODUCTION

Facial expression recognition is an active area of research for various application domains including human-computer interaction, security and health. In contrast to normal facial expressions that may be posed or convincingly acted out by talented humans, facial micro-expressions generally represent the actual emotion of a person, as it is a spontaneous reaction expressed involuntarily through the face [1, 2, 3]. In recent years, deep learning based approaches have gained popularity and now widely implemented in solving various computer vision problems including Image Classification [4].

In general, deep learning techniques have shown to outperform hand-crafted techniques in most of these problems. Moreoever, a few convolutional neural network based approaches have been proposed for automated micro-expression recognition [5, 6], achieving considerable accuracies.

Deep neural networks, while powerful in learning from complicated visual data, are vulnerable to small noise i.e. adversarial perturbations [7]. Szegedy et al. [8] showed that adversarial perturbations imperceptible to humans can cause misclassification in DNN-based image classifiers; while Mopuri et al. [9] proposed "image-agnostic" perturbations.

In this paper, we focus on safeguarding individual emotional privacy against automated micro-expression spotting and recognition tasks. Micro-expressions are emotions that humans intentionally suppress. Thus, humans will feel their privacy is violated if automatic algorithms, such as those increasingly deployed in social media platforms, can recognize their micro-expressions through the shared videos. Essentially, this notion of privacy is in terms of ensuring micro-expression mis-classifications, and thus withholding the true emotion currently felt by the human.

There have been some research work in this direction, namely: adversarial image perturbation (AIP) [10], generative adversarial network based algorithm [11] and mapping distortion based protection (MDP) [12]. Specifically, AIP introduces a general game theoretical framework for the user-recogniser dynamics that involve the current state-of-the-art AIP and person recognition techniques. Meanwhile, MDP protects the privacy by modifying the original dataset with its corresponding label. In contrast to our approach which focuses on videos, these methods are for protecting image privacy.

Research in attacking video-based classifiers has emerged over the years. [13] were the first to propose white-box attack on video action recognition focusing on networks with a CNN+RNN architecture to compute sparse adversarial perturbations. Essentially, they searched for the perturbation over the video space. Note that the duration of a micro-expression is usually only 1/25 to 1/5 of a second, such attacks may not fit with the frame rate in micro-expression settings. [14] proposed a method that adds a universal adversarial framing on the border of the image. On the other hand, [15] selects a specific subset of frames to perturb. Meanwhile, [16] relied on a GAN-like model to generate, in interaction with a dis-

This research is funded in part by the Partenariat Hubert Curien (PHC)-Hibiscus MyPAIR funding under the project GAN Games.

criminator, the universal perturbation by up-sampling from a random latent noise vector, and this is subsequently added to the video clip. More recently, [17] considered searching for perturbations based on intermediate layers of the neural network, and adding the perturbations directly to the video. [18] introduced a flickering temporal perturbation against action classification tasks. However, such attacks may overly perturb the videos without learning the spatio-temporal features (across temporal frames), which are commonly extracted by DNN models for micro-expression recognition. Furthermore, they require artificially inducing light sources to the frames.

Our study makes a valuable contribution in protecting emotional privacy against automatic leakages because our primary contribution is beyond leveraging the field of adversarial attacks, which has so far been mostly applied to image classification tasks [7, 8, 9]. In contrast to the current adversarial attacks on video action recognition tasks [13, 14, 15], our proposed AdverFacial framework is the first-known to adopt dynamic imaging [19], which summarizes the subtle and involuntary movements of the micro-expression image sequences based on intermediate CNN layers into a single dynamic imaging frame, in universal attacks. This is the first-known video-based universal perturbation technique focusing on spatio-temporal features for the challenging subtle micro-expressions problem, where variations are infinitesimal and occur within fractions of a second.

AdverFacial safeguards emotion privacy against automatic emotional micro-expression recognition through adversarial perturbations with a comparatively high fooling rate in the automatic micro-expression classifier and cross-models transferability while keeping the perturbations imperceptible to human eyes. In this way, we can still protect the emotional privacy even when the facial videos or datasets are leaked.

2. METHODOLOGY

2.1. Preliminaries

Micro-expression Recognition Task: A micro-expression classifier $F_\theta(X) = y$ accepts an input $X = [x_1, x_2, \ldots, x_T] \in \mathbb{R}^{T\times W\times H\times C}$ comprising T consecutive frames each of dimensions H, W and C notably the height, width and number of color channels for each frame. The classifier produces an output $y \in \mathbb{R}^K$ which can be treated as the probability distribution over the micro-expression output domain, where K is the number of classes. The classifier F implicitly depends on some parameters θ that are fixed during the privacy protection process. Note that adversarial video is denote as $\hat{X} = X + \delta$ where the video perturbation $\delta \in \mathbb{R}^{T\times W\times H\times C}$ and each individual adversarial frame by $\hat{x}_i = x_i + \delta$. $\hat{X}$ is adversarial such that $F_\theta(\hat{X}) \neq F_\theta(X)$, while minimizing the distance between $\hat{X}$ and X to be negligibly small under the selected visual quality metric to ensure the visual perceptibility is maintained despite the perturbation.

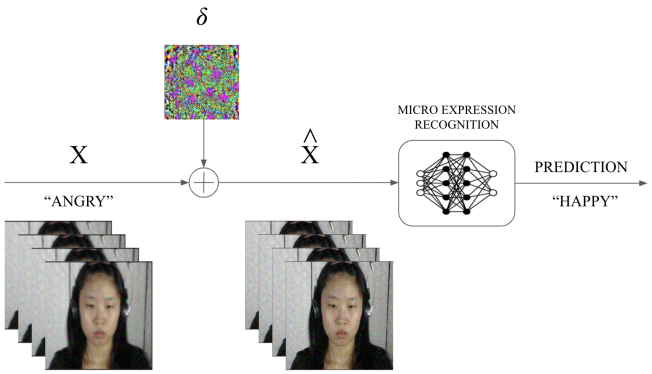

Fig. 1. AdverFacial Universal Perturbation framework.

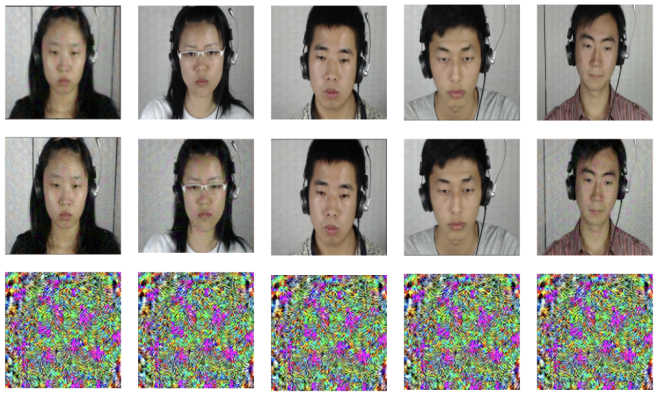

Fig. 2. Demos of crafted AdverFacial adversarial examples for CAS(ME)^2 dataset. Rows respectively show the original videos, adversarial examples and adversarial noise.

2.2. AdverFacial Generation for Emotional Privacy

As shown in Figure 1, we propose a micro-expression privacy protection framework called AdverFacial. In more detail, AdverFacial generates the privacy-preserving videos through the infusion of adversarial frames with universal perturbation vectors. As discussed above, hiding the ground-truth micro-expression classification results between video frames and its prediction label is critical for privacy protection. Therefore, instead of storing the original dataset directly, we suggest applying the universal perturbation to video frames prior to being streamed or shared on social media. Our proposed AdverFacial framework seeks a universal perturbation δ such that $||\delta||_p \leq \epsilon$ for some negligible visual quality threshold ϵ, while fooling automated micro-expression recognition of videos X. The algorithm starts with $\delta = 0$ (no perturbation) and iteratively updates the minimal perturbation $\triangle\delta_i$ that sends the current perturbed frame $x_i + \delta$ to the decision boundary of the classifier. These iterative updates continue until the termination conditions are satisfied, when the empirical "fooling rate" $Err(\cdot)$ on the perturbed data set $\hat{X}$ exceeds the target threshold $1 - \zeta$. Algorithm 1 shows our algorithm pseudocode.

Spatio-temporal features identify the micro-level informa-

Algorithm 1 Computation of AdverFacial Universal Perturbation

Input Set of X of input Dynamic Imaging preprocessed videos, classifier $F_\theta(\cdot)$, desired ℓ_p norm of perturbation ϵ and desired accuracy on perturbed samples ζ

Output Universal Perturbation δ

Initialize $\delta \leftarrow 0$
while $Err(\hat{X}) \leq 1 - \zeta$ **do**
  for all $x_i \in X$ **do**
    if $F_\theta(x_i + \delta) = F_\theta(x_i)$ **then**
      Compute the minimimal perturbation:
      $\triangle\delta_i \leftarrow arg\min_r ||r||_2$ s.t. $F_\theta(x_i + \delta + r) \neq F_\theta(x_i)$
      Update the perturbation:
      $\delta \leftarrow project(\delta + \triangle\delta_i)$
    end if
  end for
end while

tion that changes along the temporal dimension. Based on these spatio-temporal features, the micro-expression classifier distinguishes between the emotion classes. Note that the i^{th} perturbation δ_i corresponds to the i^{th} frame x_i of the video, which can be represented by three scalars. Thus, $\delta = [\delta_1, \delta_2, \ldots, \delta_T] \in \mathbb{R}^{T\times1\times1\times3}$ has in total $3T$ parameters to optimize in the spatio-temporal domains. To generate an adversarial perturbation, the objective function is formulated as:

$$arg\min_{\delta} \lambda D(\delta) + \frac{1}{N}\sum_{n=1}^{N} \ell(F_\theta(X_n + \delta), t_n) \quad (1)$$

$$\text{s.t. } \hat{x}_i \in [V_{min}, V_{max}]^{H\times W\times C} \text{ and } ||\delta||_p \leq \epsilon \quad (2)$$

where N is the total number of training videos, X_n is the n^{th} video, $F_\theta(X_n + \delta)$ is the classifier output (probability distribution), t_n is the original micro-expression label and ϵ denotes the maximum allowed perturbation strength. Note that the first term in Equation (1) is a regularization term, while the second term is the adversarial classification loss which will discussed later in this paper. The parameter λ weights the relative importance of being adversarial and also the regularization terms. The function $D(\cdot)$ determines the regularization term that allows us to achieve better imperceptibility for the human observer. The first constraint in Equation (2) guarantees that after applying the adversarial perturbation, the perturbed video will be clipped between valid values: V_{min}, V_{max} which respectively represent the minimum and maximum allowed pixel intensity.

2.3. Dynamic Imaging

In our proposed framework, dynamic imaging [19] transforms the video sequences into a frame instance by conserving the spatio-temporal information. As discussed, micro expressions are rapid and of short duration in nature. Thereby, they will appear only in a few frames of a video. To extract these momentary changes from the video, we adopt the dynamic imaging technique to generate the universal perturbation. These dynamic images are subsequently processed by the proposed AdverFacial framework for further training.

2.4. Regularization Terms

In contrast to Universal Adversarial Perturbation [7], we quantify the distortion introduced by the perturbation δ with $D(\cdot)$ in the spatio-temporal domain. This metric is constrained such that the perturbation δ is imperceptible to the human observer. In contrast to previous related works [13, 15], in our case the focus is on the unnoticability to the human observer. In order to achieve the most imperceptible perturbation, we implement a regularization term to control the different aspects of human perception. To simplify the definition of our regularization terms and metrics, we define the following notations for video perturbation of $X = [x_1, x_2, \ldots, x_T] \in \mathbb{R}^{T\times H\times W\times C}$.

With reference to [18], we apply the perceptibility regularization term that forces the adversarial perturbation to be as minimal over the three color channels of each frame. By having no temporal constraint, which relates to the 'perceptibility' of the adversarial pattern:

$$D(\delta) = \frac{1}{3T}||\delta||_2^2 \quad (3)$$

where $||\cdot||$ is defined in Equation (4) with $p = 2$ as:

$$||X||_p = (\sum_{i_1=1}^{T}\cdots\sum_{i_4=1}^{C}|x_{i_1\ldots i_4}|^p)^{\frac{1}{p}} \quad (4)$$

where $i_1, i_2, \ldots i_4$ refer to the dimensions of the micro-expression video sequences.

2.5. Evaluation Metric

To quantitatively examine the performance of AdverFacial in the micro-expression privacy protection task through misclassification, we consider state-of-the-art evaluation metrics of adversarial attacks as follows:

White-box Fooling Rate: is the percentage of adversarial videos that are successfully misclassified. It is computed based on the reduction in accuracy due to adversarial attack and causing false predictions as per the baseline [7].

Adversarial Transferability: attacking the target model using the adversarial videos which are generated on the source model as suggested in [7, 8].

Perceptibility: Mean Absolute Perturbation per pixel

$$perceptibility_i(\delta) = \frac{1}{3T}||\delta||_1 \quad (5)$$

where $||\cdot||_1$ is defined as per Equation (4) with $p = 1$. The perceptibility values in this paper will be presented as per-

cents from the full applicable values of the video span, e.g.,

$$perceptibility(\delta) = \frac{perceptibility_i(\delta)}{V_{max} - V_{min}} * 100$$

3. EXPERIMENTS

3.1. Experimental Setup

Dataset: To validate the performance of our approach, we conducted experiments on benchmarked spontaneous micro expression datasets: SMIC [1], CASME II [2] , and CAS(ME)^2 [3]. These datasets were designed for the problem of detecting and recognizing spontaneous micro-expressions. These three corpora have the following characteristics:

- The SMIC dataset has 164 spontaneous microexpressions from 16 subjects. The participants undergo high emotional arousal through highly emotional clips and suppress their facial expressions in an interrogation room setting with a punishment threat.
- The CASME II dataset has 256 micro-expressions from 26 subjects. It has higher video quality and image size compared with SMIC.
- The CAS(ME)^2 dataset, which contains 206 videos, is the latest version of the CASME series of datasets on facial micro-expressions.

Micro-expression Recognition Model: We apply the state-of-the-art CNN-based micro-expression models as our baseline targeted model for privacy protection: Lateral Accretive Hybrid Network (LEARNet) [5] and Spatiotemporal Convolutional Neural Networks (STCNN) [6]. Our experiments follow the white-box setting, which assumes knowledge of the targeted model, its parameter values and architecture.

Parameters: In this section, we generalize the attack to cause misclassification to all videos from a specific class with a single universal adversarial perturbation δ. The parameter configurations for the micro-expression dataset are $T = 96, H = 224, W = 224, C = 3$, and $V_{min} = -1, V_{max} = 1$. The perturbation size ϵ is 0.031 and the maximum number of iterations is 1,000. Results are reported for $p = 2, \epsilon = 2000$. We have randomly partitioned the available benchmarked datasets with a ratio of 80:20 respectively. Furthermore, the training set is divided into training and validation set with a 70:30 ratio respectively. We developed the single universal δ by solving the optimization problem in Equation (1), where $\{X_n\}_{n=1}^{N}$ is the training set defined as the entire evaluation-split of the micro-expression dataset. Once the universal δ was computed, we evaluated its fooling ratio, transferability and perceptibility on a random sub-sample from the test-split.

3.2. Privacy Protection

We perform novel adversarial attacks on video-based micro-expression recognition to confuse the automatic emotion classifier so that classification accuracy is reduced, thereby guarding emotional privacy against automated emotional recognition while keeping the perturbations imperceptible to human eyes.

Table 1. White-box Fooling Rate(%) performance against different architectures with various datasets.

Attack	CASME-II	CAS(ME)^2	SMIC
AF [14]	20.08	31.65	18.68
AdverFacial	65.49	68.02	62.12

Table 2. Adversarial Transferability performance against different architectures with the same datasets, CAS(ME)^2.

Source Model	STCNN	LEARNet
STCNN [6]	69.50	55.20
LEARNet [5]	62.10	68.02

Due to the feature representation in a small spatio-temporal window for micro-expression datasets, it might be unfair to compare AdverFacial with [13, 16]. Note that our experiments follows the white-box setting; therefore, [15, 17] which are black box techniques are not relevant in this paper. The results of the privacy protection of micro-expression prediction labels are shown in Table 1. Under the white-box setting, all attacks are generated on the target model and used to attack this model. As compared with Adversarial Framing (AF) [14], the AF attack can only reduce the fooling rate results slightly in all cases. In contrast, the AdverFacial approach can effectively improve the fooling rate results in attacking different video frame lengths for all the three datasets.

3.3. Adversarial Transferability

We evaluate the transferability of the universal perturbations across models on the same micro-expression dataset. It is unfair to compare AdverFacial with AF on transferability since the latter is not designed towards that goal. Table 2 shows the results of the transferability. Result shows that the AdverFacial successfully generates transferable adversarial perturbations in privacy protection. The high effectiveness of the attack applied across models indicates that our attack is transferable between these different models.

4. CONCLUSION

In this paper, we propose the AdverFacial framework which adopts the universal adversarial perturbation on videos in order to effectively safeguard the emotional privacy against automated micro-expression recognition tasks. This is achieved by perturbing the videos without visually affecting the perceptual quality. Thus, we show that it is possible to protect emotional privacy even when one's facial videos are publicly available, while keeping such perturbations imperceptible to human eyes.

5. REFERENCES

[1] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikainen, "A spontaneous micro-expression database: Inducement, collection and baseline," *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013.

[2] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu, "Casme ii: An improved spontaneous micro-expression database and the baseline evaluation," *PLoS ONE*, vol. 9, no. 1, 2014.

[3] Fangbing Qu, Su-Jing Wang, Wen-Jing Yan, and Xiaolan Fu, "Cas(me)2: A database of spontaneous macro-expressions and micro-expressions," *Human-Computer Interaction. Novel User Experiences Lecture Notes in Computer Science*, p. 48–59, 2016.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, 2012, NIPS'12.

[5] Monu Verma, Santosh Kumar Vipparthi, Girdhari Singh, and Subrahmanyam Murala, "Learnet: Dynamic imaging network for micro expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 1618–1627, 2020.

[6] Zhaoqiang Xia, Xiaopeng Hong, Xingyu Gao, Xiaoyi Feng, and Guoying Zhao, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 626–640, 2020.

[7] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[8] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations*, 2014.

[9] Konda Reddy Mopuri, Utkarsh Ojha, Utsav Garg, and R. Venkatesh Babu, "Nag: Network for adversary generation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[10] Seong Joon Oh, Mario Fritz, and Bernt Schiele, "Adversarial image perturbation for privacy protection a game theory perspective," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[11] Yuwen He, Chunhong Zhang, Xinning Zhu, and Yang Ji, "Generative adversarial network based image privacy protection algorithm," *International Conference on Graphics and Image Processing (ICGIP 2018)*, 2019.

[12] Yiming Li, Peidong Liu, Yong Jiang, and Shu-Tao Xia, "Visual privacy protection via mapping distortion," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[13] Xingxing Wei, Jun Zhu, and Hang Su, "Sparse adversarial perturbations for videos," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, 2021, pp. 8973–8980.

[14] Michał Zajac, Konrad Zołna, Negar Rostamzadeh, and Pedro O. Pinheiro, "Adversarial framing for image and video classification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 10077–10078, Jul. 2019.

[15] Zhipeng Wei, Jingjing Chen, Xingxing Wei, Linxi Jiang, Tat-Seng Chua, Fengfeng Zhou, and Yu-Gang Jiang, "Heuristic black-box adversarial attacks on video recognition models," in *The AAAI Conference on Artificial Intelligence (AAAI) 2020, New York, NY, USA, February 7-12, 2020*. 2020, pp. 12338–12345, AAAI Press.

[16] Shasha Li, Ajaya Neupane, Sujoy Paul, Chengyu Song, Srikanth V. Krishnamurthy, Amit K. Roy-Chowdhury, and Ananthram Swami, "Stealthy adversarial perturbations against real-time video classification systems," in *Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. 2019, The Internet Society.

[17] Shangyu Xie, Han Wang, Yu Kong, and Yuan Hong, "Universal 3-dimensional perturbations for black-box attacks on video recognition systems," *CoRR*, vol. abs/2107.04284, 2021.

[18] Roi Pony, Itay Naeh, and Shie Mannor, "Over-the-air adversarial flickering attacks against video recognition networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 515–524.

[19] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould, "Dynamic image networks for action recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3034–3042.