Evaluate The Efficacy of Attention Mechanisms in Face Anti-Spoofing With an Explainable AI Approach

1st Thinh Nguyen Le Quang
Software Engineering
FPT University
Can Tho, Vietnam
nlqthinh.work@gmail.com

2nd Dat Vo Minh
Software Engineering
FPT University
Can Tho, Vietnam
vominhdat312@gmail.com

3rd Anh Dinh The
Software Engineering
FPT University
Can Tho, Vietnam
theanh290702@gmail.com

4th Phuc Phan Hong
Software Engineering
FPT University
Can Tho, Vietnam
phucphan1421@gmail.com

5th Hoang Ngoc Tran
Software Engineering
FPT University
Can Tho, Vietnam
hoang2531992@gmail.com

***Abstract*—In this study, we propose the eXplainable AI (XAI) to evaluate the influence of Channel Attention and Spatial Attention in Convolutional Neural Networks (CNN) models on the CelebA-Spoof dataset. Specifically, we implemented four models: CNN, CNN with Channel Attention, CNN with Spatial Attention, and CNN combining both Channel Attention and Spatial Attention. The experimental results show a significant improvement in accuracy, achieving 0.9875, 0.9900, 0.9905, and 0.9960 for each model respectively. To better understand how and why these models work, we use eXplainable AI (XAI) interpretation methods including Grad-CAM, SHAP, and LIME. Additionally, our study not only enhances understanding of attention techniques' impact on classification accuracy but also contributes to elucidating the model's decision-making process, thereby fostering the development of more accurate and interpretable image classification models.**

***Index Terms*—Explainable AI, Face Anti-Spoofing, Convolutional Neural Networks, Channel Attention, Spatial Attention.**

I. Introduction

In recent years, Deep Learning has made great strides in the field of computer vision, especially in tasks such as facial recognition and distinguishing between spoof and live images. One of the most important factors contributing to this success is the development of Convolutional Neural Networks [1], CNNs designed to effectively recognize image features. However, despite its ability to learn strong feature representations, traditional CNNs may ignore some important spatial and channel information when processing images. This leads to the research and application of attention techniques such as Channel Attention and Spatial Attention to improve the model's recognition and classification ability.

Channel Attention focuses on enhancing the distinction between feature channels, while Spatial Attention enhances the ability to perceive the spatial location of important features in the image. Combining these methods can notably enhance CNN performance, especially in tasks like facial spoofing detection, where accurate feature recognition is crucial. Transparency and trust in AI models are important, driving the need for interpretability. The idea we use eXplainable AI (XAI) to evaluate many models is similar to the article [2] but we use more techniques such as Grad-CAM, SHAP, and LIME to shed light on model decision-making, aiding research and development efforts.

II. Related work

The paper [3] introduces a generalized face anti-spoofing framework that combines depth estimation, face analysis, and live/fake classification, improving fake face detection and generalization to unseen domains using meta-learning. However, it requires large amounts of high-quality data, which may not be practical. In [4], lightweight network models are shown to achieve comparable accuracy to larger models with lower computational costs, suitable for resource-constrained applications but raising questions about accuracy versus efficiency. Papers [5] and [6] explore network depth and lightweight architectures for tamper recognition, with [5] highlighting the benefits of deeper networks and [6] introducing FeatherNet, a high-performance, low-complexity architecture. Despite advancements, previous studies often neglect model decision interpretation, essential for trust and transparency, and the combined impact of Channel and Spatial Attention techniques on spoof detection remains underexplored. Our paper addresses this by evaluating the influence of these attention mechanisms on classification performance using the CelebA-Spoof dataset. We demonstrate significant accuracy improvements and employ eXplainable AI (XAI) methods such as Grad-CAM, SHAP, and LIME to explain model decisions, enhancing transparency and reliability in facial spoofing detection systems.

III. Proposed Method

A. Network architecture

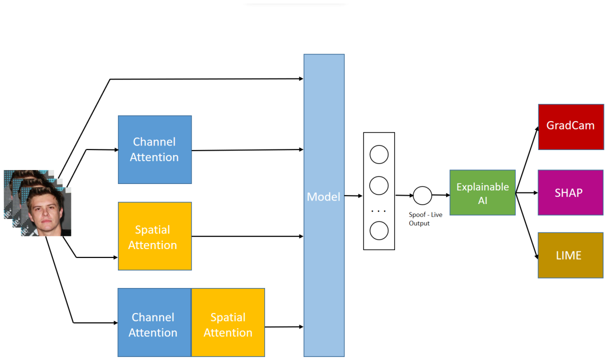

Fig. 1. The overall architecture for spoof images classification.

The input will be an image of size 64x64x3 which will be inspected according to the workflow as shown in Figure 1, we will experiment with no attention, channel attention, spatial attention, then combined both that attached to the top of the model. The model in the figure is the best model with the highest accuracy chosen for experimentation. Before selecting the best model, we trained various pre-trained models and self-tuned models. Finally, we will then use explainable AI to explain each model's predictions.

Channel attention mechanisms focus on ”what” is important in the input feature maps by assigning different weights to different channels. The process involves global average pooling and global max pooling to generate context descriptors. These descriptors are then passed through convolutional layers and activation functions to produce a channel attention map. This map emphasizes the most informative channels, allowing the model to focus on more relevant features.

Spatial attention mechanisms focus on ”where” is important in the input feature maps by highlighting significant spatial locations. This is achieved by applying average pooling and max pooling along the channel axis, followed by concatenation of these descriptors. The combined descriptor is then passed through a convolutional layer to generate a spatial attention map. This map identifies key spatial regions in the input images, enabling the model to concentrate on these areas.

Combining channel and spatial attention mechanisms allows the model to leverage both ”what” and ”where” information. This combined approach aims to enhance the model's ability to focus on the most relevant features and spatial locations, leading to improved classification performance.

B. Evaluation metrics

For evaluating our model on a balanced dataset, we use four key metrics:

- **Accuracy** (A): Defined as $A = \frac{TP+TN}{N}$, where TP, TN, and N represent the number of true positives, true negatives, and the total number of samples, respectively.
- **Precision** (P): Given by $P = \frac{TP}{TP+FP}$, with FP indicating false positives, measures the accuracy of positive predictions.
- **Recall** (R): Calculated as $R = \frac{TP}{TP+FN}$, where FN stands for false negatives, evaluates the model's ability to identify all actual positives.
- **F1-Score** ($F1$): The harmonic mean of Precision and Recall, $F1 = 2 \times \frac{P \cdot R}{P+R}$, providing a balance between them.

C. XAI techniques

Our **Grad-CAM** implementation generates a heatmap based on the gradients of the target class (either ”spoof” or ”live”) with respect to the output feature maps of the last convolutional layer. This process is mathematically modeled as follows:

Given an image input X, the Grad-CAM heatmap for a class c (identified by prediction index) from the last convolutional layer is computed as:

$$\text{Grad-CAM}(X, c) = \text{ReLU}\left(\sum_k \alpha_k^c \cdot A^k\right) \tag{1}$$

where:

- A^k are the activations of the last convolutional layer for the input X.
- α_k^c are the channel-wise gradient weights, calculated by global average pooling the gradients of Y^c (the output for class c) with respect to A^k.
- The ReLU function is applied to focus on features that have a positive influence on the class of interest.

In the **LIME** approach we've used, the main idea is to approximate the model locally by perturbing the input image around its vicinity and observing the changes in predictions. The critical part of LIME, as implemented, involves creating superpixel-based perturbations and fitting a simple model to these perturbations to infer the importance of each superpixel.

Our **SHAP** implementation seeks to explain individual predictions by computing the contribution of each feature (pixel in images) to the prediction. The SHAP values are calculated using a background dataset to approximate how the presence or absence of features in the input affects the model's output. SHAP values explain the output of model f for an input x by computing the contribution of each feature:

$$\phi(f, x) = f(x) - E[f(x)] \tag{2}$$

where:

- $E[f(x)]$ is the expected value of the model output over the background dataset.
- $\phi(f, x)$ represents the SHAP values, indicating the deviation of the model's prediction for x from the expected prediction, attributed to the features present in x.

IV. Experiments

In this study, we are using the GPU P100 implemented in Keras on Tensorflow 2.0 which is available on Kaggle. The P100 GPU is a powerful graphics processing unit, equipped with 16 GB of GPU memory, which provides significant computational capacity essential for our research. Additionally, we employed early stopping with a patience of 10 epochs to prevent overfitting during training. We utilized the Adam optimizer with a learning rate schedule, which exponentially decays from an initial value of 0.0001 over 100,000 decay steps with a decay rate of 0.96 and a staircase decay mode. The model was trained for 50 epochs using a sparse categorical cross-entropy loss function and evaluated on a separate validation set.

A. Dataset

CelebA-Spoof [6] is a dataset consisting of 625,537 pictures of 10,177 subjects that was used for our experiments, designed to evaluate anti-face spoofing techniques. It includes a variety of facial images with different lighting conditions, poses, and spoofs, providing a comprehensive environment to try out. We took 200,000 images and divided them into 100,000 live images and 100,000 spoof images, of which 180,000 images were used for training, 18,000 images were used for validation, and 2000 images were used for testing of our models.

B. Experiments and Results

Our experiment evaluates several pre-trained models followed by our custom-tuned CNN to evaluate their performance in detecting spoofing attempts. We have train and test with the following pre-trained models: Xception, ResNet50, VGG16, DenseNet, EfficiencyNetB7, and MobileNetV3 with the following results shown in Table I.

TABLE I
Pre-trained models training and test results

	Accuracy	Precision	Recall	F1-Score
Xception	0.81100	0.81312	0.81156	0.81084
Resnet50	0.97200	0.97197	0.97205	0.97200
VGG16	0.95400	0.95399	0.95407	0.95400
Densenet121	0.92350	0.92347	0.92351	0.92348
EfficientNetB7	0.87250	0.87247	0.87252	0.87248
MobileNetV3	0.95850	0.95854	0.95862	0.95850
Custom-tuned CNN	**0.98750**	**0.98763**	**0.98767**	**0.98750**

Custom fine-tuned CNN model is a model that we fine-tune and design ourselves. In the custom fine-tuned CNN model, the input is normalized by the Lambda layer to ensure pixel values are clustered around 0, improving convergence during training. This is followed by a series of Conv2D layers with a small number of filters (10 and then 3) are applied, followed by LeakyReLU activation to introduce nonlinearity. The model has several convolutional blocks, each with a Conv2D layer, followed by BatchNormalization. The number of filters in these layers gradually increases (16, 32, 64, 96, 128, and 192), allowing the network to learn a hierarchy of increasingly complex features. MaxPooling layers are interleaved between convolution blocks to reduce the spatial dimension of the feature map, thereby reducing the number of parameters and computational cost. After the final Conv2D block with 256 and then 128 filters, the model uses a Flatten layer to convert the feature maps to 1D vectors, followed by a Dropout layer to minimize overfitting. The classification is performed by Dense Layer with softmax activation function to give the probability for two classes. Based on the experimental results shown in Table II, our custom-tuned Convolutional Neural Network (CNN) has proven to be highly effective and specifically designed for the task of anti-face spoofing.

TABLE II
Attention model training and test results

	Accuracy	Precision	Recall	F1-Score
Custom-tuned CNN	0.98750	0.98763	0.98767	0.98750
Custom-tuned CNN + Channel Attention	0.99000	0.98999	0.99001	0.99000
Custom-tuned CNN + Spatial Attention	0.99050	0.99047	0.99056	0.99050
Custom-tuned CNN + Channel Attention + Spatial Attention	**0.99600**	**0.99601**	**0.99598**	**0.99600**

Our custom-tuned CNN model outperforms pre-trained models with an accuracy of 0.98750. Given its outstanding performance, it was chosen for further experiment with attention mechanisms. We then incorporated the spatial and channel attention modules into our custom-tuned CNN model. Integrating these attention mechanisms aims to enhance the model's focus on relevant features to detect tampering. The accuracy results are as Table II.

Adding channel attention improved accuracy slightly, while spatial attention had a more pronounced effect, suggesting that focusing on specific spatial regions of the input image is very important to the mission. The synergy of combining both channel and spatial attention is evident, delivering outstanding precision of 0.99600 and consistent improvements in precision, recall, and F1 score. This integration demonstrates the effectiveness of the attention mechanism in enhancing the model's ability to focus on discriminatory features to combat face spoofing, leading to more accurate classification.

C. Explainable AI

The Grad-CAM shown in Figure 2 visualization a heatmap overlay on the original image, where the regions that contribute the most to the model's output have been highlighted, these regions correspond to the regions in the image that the model image is considered important for its prediction. In the case of different attack scenarios—print attack, playback/video, and 3D masking, Grad-CAM results generally cluster around facial features, all four models showing no difference The Grad-CAM display results are too large, but somehow we know better what model is important to rely on to make predictions.

LIME shown in Figure 3 visualization provides clear division of influential areas by highlighting superpixels. These highlighted areas are areas where perturbation will have a significant impact on the model's predictions. In the output

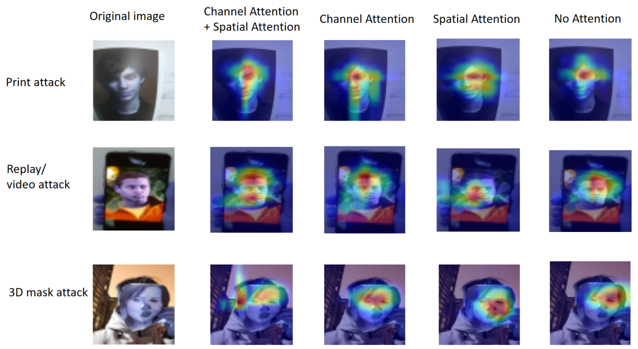

Fig. 2. GradCam illustration for spoof images

of LIME, we observe that not only facial features but also background elements and patterns surrounding the subject are considered. In the model that combines channel attention and spatial attention, the output retains many important features and seems more accurate. For example, the replay attack image in the image in Figure 2, the model focuses on the phone to make predictions more than other models.

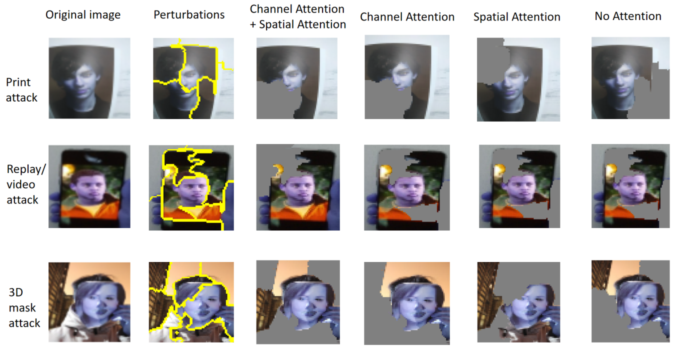

Fig. 3. LIME illustration for spoof images

SHAP shown in Figure 4 visualization represents pixel-level contributions to the model's predictions by displaying them in grid format on the original image. Red and blue squares represent positive and negative SHAP values, respectively, suggesting which pixels push the model's output toward or away from a particular layer. These figures suggest that certain image patches have more influence on the model's decisions, potentially indicating areas with or without spoofing artifacts. Models with a combination of channel attention and spatial attention display red squares on important features more than other models.

V. Conclusion

In this research, we presented the integration of attention mechanisms into a custom-tuned convolutional neural network (CNN) for the anti-face spoofing task. Our experimental results demonstrate the effectiveness of channel and spatial attention

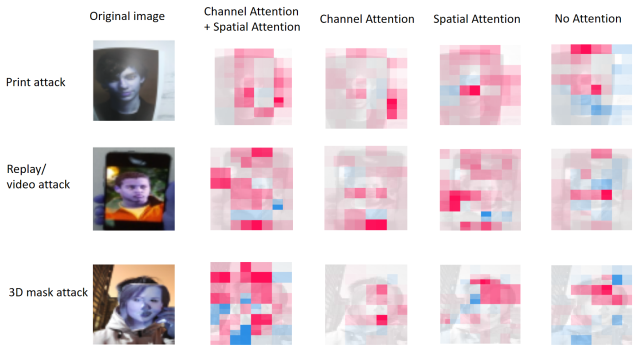

Fig. 4. SHAP illustration for spoof images

in enhancing the model's discrimination ability. Channel attention effectively fine-tunes feature representation on a channel-by-channel basis, selectively emphasizing informative features, while spatial attention provides the model with the ability to focus into prominent spatial regions in the input image illustrated with XAI techniques that contribute to clarifying the model's predictions.

References

[1] Nagpal, Chaitanya, Dubey, Shiv Ram. (2019). A Performance Evaluation of Convolutional Neural Networks for Face Anti Spoofing. 1-8. 10.1109/IJCNN.2019.8852422.

[2] L. -D. Quach, K. N. Quoc, A. N. Quynh, H. T. Ngoc and N. Thai-Nghe, ”Tomato Health Monitoring System: Tomato Classification, Detection, and Counting System Based on YOLOv8 Model With Explainable MobileNet Models Using Grad-CAM++,” in IEEE Access, vol. 12, pp. 9719-9737, 2024, doi: 10.1109/ACCESS.2024.3351805. keywords: Predictive models;Computational modeling;Data models;YOLO;Solid modeling;Physiology;Diseases;Explainable AI;Smart agriculture;Farming;Object detection;Crops;Explainable AI;interpretability;XAI;MobileNet models;smart farming;Grad-CAM++;fruits object detection,

[3] Chuang, Chu-Chun, Chien-Yi Wang, and Shang-Hong Lai. “Generalized Face Anti-Spoofing via Multi-Task Learning and One-Side Meta Triplet Loss.” arXiv, November 29, 2022. http://arxiv.org/abs/2211.15955.

[4] Y. Martínez-Díaz, H. Méndez-Vázquez, L. S. Luevano and M. Gonzalez-Mendoza, ”Exploring the Effectiveness of Lightweight Architectures for Face Anti-Spoofing,” 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 2023, pp. 6392-6402, doi: 10.1109/CVPRW59228.2023.00680. keywords: Performance evaluation;Computer vision;Analytical models;Computational modeling;Computer architecture;Robustness;Task analysis,

[5] Simonyan, Karen, Zisserman, Andrew. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition.

[6] P. Zhang et al., ”FeatherNets: Convolutional Neural Networks as Light as Feather for Face Anti-Spoofing,” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 2019, pp. 1574-1583, doi: 10.1109/CVPRW.2019.00199. keywords: Face;Computer architecture;Task analysis;Feature extraction;Network architecture;Machine learning;Support vector machines,

[7] Zhang, Yuanhan, Zhenfei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. “CelebA-Spoof: Large-Scale Face Anti-Spoofing Dataset with Rich Annotations.” arXiv, August 1, 2020. http://arxiv.org/abs/2007.12342.