Fourier Transform and Attention Guided Deep Neural Network for Face Anti-Spoofing in Medical Applications

Zhanseri Ikram
Al-Farabi Kazakh National University, Almaty, Kazakhstan
Narxoz University, Almaty, Kazakhstan

***Abstract*—Face recognition systems have become prevalent in mobile devices and security applications, increasing the demand for robust face presentation attack detection. Early efforts based on handcrafted features struggled to cope with variations in illumination, pose, and attack modalities, prompting a transition toward deep learning solutions capable of extracting subtle discriminative cues. A novel architecture built upon an EfficientNet-V2 backbone, combined with a Shuffle Attention module and Fourier heads, was developed to capture both spatial and frequency domain characteristics. A dual-path approach processes each input face image through conventional convolutional blocks and a 2D Discrete Fourier Transform path, with dedicated Fourier heads reconstructing frequency maps that reveal minute discrepancies between genuine and spoofed presentations. Experimental evaluation on the Oulu-NPU dataset demonstrates strong performance across four protocols, including robust detection under varying environmental conditions, low error rates with novel attack types, and consistent results across different sensor inputs. Metrics such as APCER, BPCER, and ACER validate the method's ability to distinguish between live and fake faces reliably. The outcomes suggest that combining spatial and frequency cues addresses limitations observed in earlier approaches, offering valuable insights for deployment in security-sensitive applications and setting a strong foundation for future research in face anti-spoofing.**

***Keywords*—Liveness detection; face anti-spoofing; deep learning; CNN; frequency domain**

I. Introduction

Face recognition systems have grown in popularity in various applications such as mobile devices and security systems. Early research in face liveness detection relied on handcrafted features such as local binary patterns and histograms of oriented gradients to capture subtle texture and motion cues [1-2], using simple image analysis and rule-based methods to separate real faces from spoof attempts. However, they often struggled with changes in illumination and pose and could not adapt well to different attack scenarios [3-4].

The field later shifted to the use of deep learning methods. Convolutional Neural Networks brought a new way to extract features from face images and proved better at detecting minute differences between live faces and fakes [5-7]. Recent studies have also focused on integrating attention mechanisms and frequency domain analysis into CNNs to improve face anti-spoofing [8-9], showing improvements in accuracy and robustness compared to traditional approaches.

In this work, EfficientNet-V2-S was used as the backbone network. EfficientNet-V2-S is known for its speed and accuracy in image processing tasks [10]. The network works by processing an input face image through several convolutional blocks that extract features at different scales. Shuffle Attention [11] introduces a novel mechanism for processing deep convolutional features by partitioning input channels and applying concurrent spatial operations. Grouping feature maps allows parallel extraction of spatial and channel information, followed by a channel shuffle that redistributes features effectively. Extensive experiments on standard networks such as ResNet-50 indicate significant performance improvements in classification and detection tasks while maintaining low computational overhead. Researchers have integrated the module into multiple vision tasks, yielding promising accuracy and efficiency results. Recent empirical evaluations validate the approach as a valuable alternative to traditional attention methods in diverse computer vision applications. Further studies corroborate findings.

Alongside the backbone and attention module, Fourier heads are added to the architecture. They reconstruct Fourier maps from the features extracted by the backbone. The Fourier maps capture frequency domain information that often holds clues to subtle differences between live faces and spoof attempts [12-14]. The Fourier heads use a series of deconvolution and dilated convolution layers to build these maps. By including frequency information, the network can detect differences that may not be clear in the spatial domain alone.

The network is built in steps that allow it to process the image, compute attention weights, and reconstruct Fourier maps. The outputs from the different stages are then combined to produce a final prediction. Such a design aims to capture both spatial and frequency cues that are important for distinguishing genuine faces from presentation attacks. Each component of the network plays a role in addressing the weaknesses of earlier methods and in improving the overall detection accuracy.

Face anti-spoofing remains a subject of great interest due to the increasing use of face recognition in security-sensitive applications. As presentation attacks grow more complex, researchers must continue to search for better ways to detect both obvious and subtle forms of spoofing. The need for robust and efficient systems is clear given the evolving methods of attack and the widespread use of biometric systems. Current study adds to this growing body of work by combining modern

network architectures, attention mechanisms, and Fourier domain analysis to address these challenges. Ongoing research in this area is essential to keep up with the rapid progress in attack techniques and to ensure the safety and reliability of face recognition systems [15].

II. Related Works

Face anti-spoofing has attracted significant research interest owing to the security implications of deploying face recognition systems. Over time, methods evolved from reliance on handcrafted features like local binary patterns and histograms of oriented gradients to modern deep learning frameworks capable of automatically extracting discriminative representations. Early techniques concentrated on identifying subtle texture and motion cues to differentiate between live and spoof presentations, though challenges with varying illumination and pose limited their practical applications.

A. Face Anti-Spoofing: Definitions and Impact

Face anti-spoofing deals with the task of telling apart live faces from fake ones produced by photographs, video replays, or masks. The goal is to protect systems that use face recognition by stopping attempts to bypass security. Research in this area has shown that proper detection is vital for preventing unauthorized access and misuse in various applications [16].

B. Traditional Methods

Earlier work in face anti-spoofing focused on hand-designed features and rule-based procedures. Texture features extracted using methods such as local binary patterns and histogram of oriented gradients were combined with classifiers like support vector machines. Thus, such systems worked by identifying subtle texture differences between real faces and printed or replayed images. While these methods were simple and fast, their performance dropped in conditions with varying lighting, pose, or image quality [17].

C. Machine Learning Methods

As the field advanced, researchers introduced machine learning methods that improved detection rates. Classifiers were built on top of handcrafted or shallow learned features extracted from facial images. Several studies combined texture information with motion cues, attempting to capture both spatial and temporal variations in the input. Fusion of multiple cues improved performance under diverse conditions, although the reliance on manually designed features still limited the overall accuracy [18].

D. Deep Learning Methods

Deep learning has introduced new possibilities in various domains, ranging from computer vision to medicine [19-21]. One of these fields involves addressing security issues such as face anti-spoofing, where models automatically learn meaningful features directly from raw data. Convolutional neural networks have been applied to capture intricate spatial details that differentiate live faces from spoof attacks. Some works introduced additional supervision or auxiliary tasks to guide the network toward discriminative patterns. Recent studies have also brought in attention mechanisms that help the network focus on regions with strong cues and methods that analyze frequency domain information to detect hidden patterns. Many deep learning approaches have reached state-of-the-art results on standard datasets such as CASIA-FASD and Replay-Attack, proving their advantage over earlier methods [22-23].

E. Challenges

Despite progress, several challenges remain. One difficulty is the wide range of attack types, from simple printed photos to complex 3D masks. Another is the gap between controlled experimental conditions and real-world scenarios. Variations in camera quality, lighting, and environmental conditions can lead to degraded performance when a model is deployed outside its training setting. Moreover, models sometimes struggle with unexpected spoofing methods not seen during training, which calls for solutions that can generalize well across different domains [24].

Another persistent challenge is the limited availability of large-scale, diverse datasets that accurately represent real-world spoofing variations. Many face anti-spoofing datasets are collected under controlled conditions, resulting in a lack of generalization when applied to heterogeneous environments such as hospitals or telemedicine platforms. Medical applications introduce unique challenges, including the presence of masks, occlusions, and patient motion artifacts, which complicate spoofing detection. Moreover, privacy constraints often limit the amount of publicly available medical facial data for model training, further exacerbating overfitting and bias issues [25-26]. To mitigate these problems, researchers have begun exploring domain adaptation, transfer learning, and synthetic data generation using generative adversarial networks (GANs) to expand dataset diversity and improve model robustness [27].

Another significant challenge is the computational complexity and real-time inference requirements for deployment in practical medical environments. Deep neural networks with attention and frequency-domain components, such as Fourier-based modules, demand high processing power, which can hinder integration into low-latency telemedicine systems or edge devices. Lightweight architectures, model pruning, and knowledge distillation are being investigated to reduce inference time while maintaining accuracy [28]. Furthermore, there is growing attention to explainability and transparency in model decisions, especially in healthcare contexts, where interpretability is critical for clinical trust and regulatory approval [29]. Future advancements will likely focus on integrating efficient model compression, interpretable attention visualization, and adaptive frequency analysis to enhance both performance and reliability in medical face anti-spoofing systems.

F. Research Gaps and Opportunities

Current work has made notable strides, yet there are several open areas for investigation. One gap lies in fully exploiting the frequency domain, as few methods integrate frequency information with spatial features for a more robust detection scheme [30]. Further studies are needed to understand how attention mechanisms can be combined with alternative representations to better capture subtle spoof cues. In addition, there is an opportunity to improve the generalization of anti-spoofing systems by addressing the domain shift between training and real-world environments [31]. Future research

should consider the integration of multiple modalities and richer representations to build systems that are both accurate and robust against evolving spoof attacks.

III. Materials and Methods

A. Proposed Model

The method illustrated in Fig. 1 begins by detecting a face using the Multitask Cascaded Convolutional Network [32]. The bounding box output by MTCNN is cropped and resized to 512 × 512 pixels. Each face image then follows two main paths. The first branch is a common path, where various stochastic augmentations (e.g., random horizontal flips, color jitter) were applied. The augmented image is forwarded through a pretrained EfficientNetV2-S backbone, which consists of eight blocks of modified convolutional layers. EfficientNetV2-S - because it offers a strong accuracy-latency trade-off for 512×512 inputs while maintaining a moderate parameter count, which is important for real-time mobile or access-control deployments. Its compound scaling and fused-MBConv layers provide large effective receptive fields and robust texture sensitivity without incurring the inference cost of deeper backbones (e.g., ResNet-101/152) or the capacity limits of lighter models (e.g., MobileNet variants). The second path is Frequency. The same 512×512 image go through a 2D Discrete Fourier Transform to produce a frequency-domain representation.

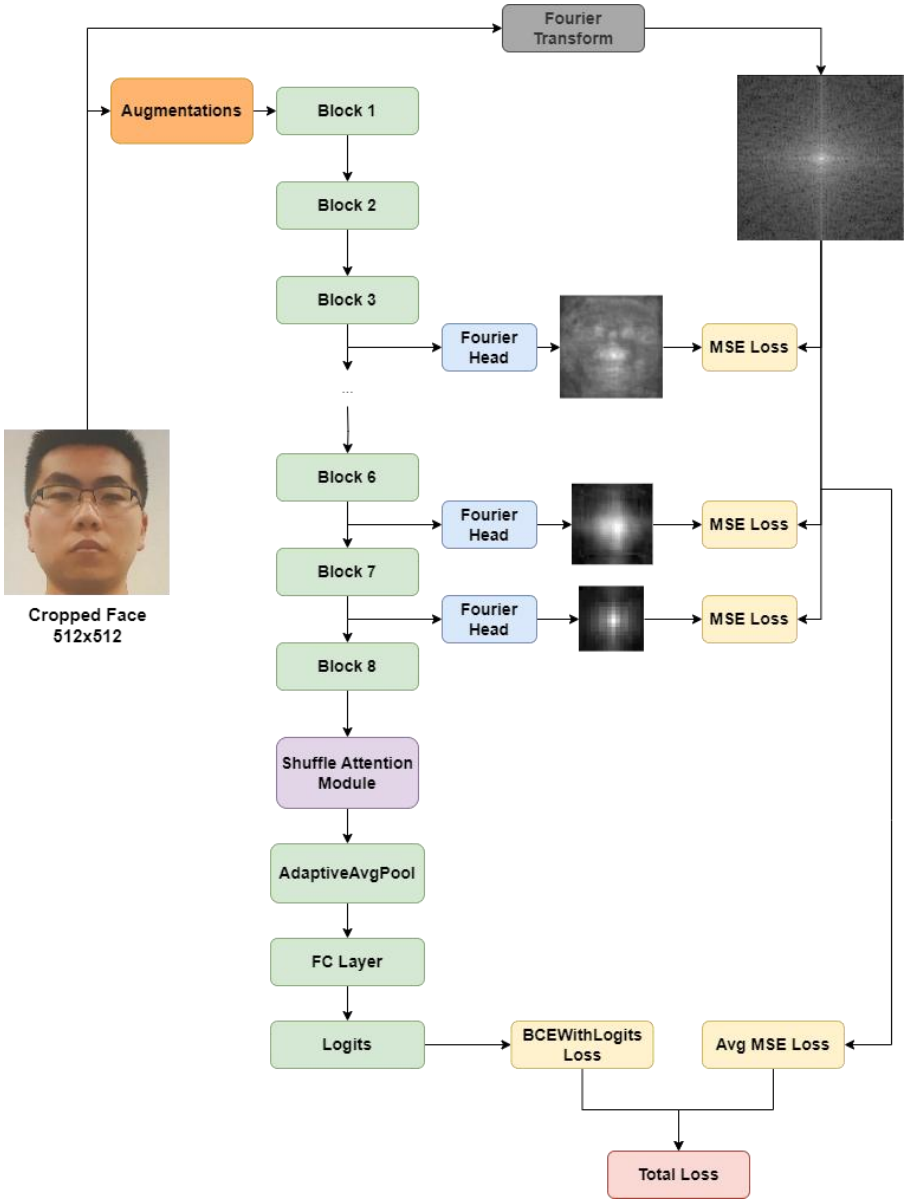


Fig. 1. Architecture of the proposed model.

After Blocks 3, 6, and 7 of the EfficientNetV2-S backbone in Fig. 2, intermediate feature maps X_3, X_6 and X_7, are each passed to a custom "Fourier Head" to reconstruct the ground-truth frequency map F. Each head contains six sequential convolutional layers (with kernel sizes of 3 × 3 or 1 × 1) interspersed with batch normalization and ReLU activation. After the third convolution, a dropout layer with probability $p = 0.5$ is inserted to reduce overfitting. The output of the fifth

convolutional layer is a single-channel feature map $\widehat{F_k} \in \mathbb{R}^{WxH}$ (where k = 3,6,7 indicates the block index), intended to approximate F. The Mean Squared Error loss for each head k is computed as Eq. (1):

$$\mathcal{L}_{MSE,k} = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}(F_{x,y} - \hat{F}_{k,x,y})^2 \quad (1)$$

After the final block, Block 8, of the EfficientNetV2-S backbone, the resulting feature map $X_8 \in \mathbb{R}^{1280x16x16}$ is processed by a Shuffle Attention module. Shuffle Attention rearranges spatial and channel dimensions for efficient information exchange without additional squeeze-and-excitation blocks. It was decided to choose Shuffle Attention over squeeze-and-excitation or CBAM because SA jointly captures channel and spatial dependencies with minimal parameter overhead and no extra bottlenecks, which is advantageous at the 16×16, 1280-channel stage of EfficientNetV2-S. The output of Shuffle Attention is a refined feature map $X_{att} \in \mathbb{R}^{1280x16x16}$.

Following the Shuffle Attention block, the refined features X_{att} pass through an Adaptive Average Pooling layer, resulting in a single vector $\boldsymbol{z}$. A fully connected layer transforms $\boldsymbol{z}$ into logits $p \in \mathbb{R}^2$ (or a single logit for binary tasks). The classification loss is the binary cross-entropy (BCE) with logits, as in Eq. (2):

$$BCELoss = -\frac{1}{M}\sum_{m=1}^{M}(y_m * \log(\sigma(p_m)) + (1 - y_m) * log(1 - \sigma(p_m))) \quad (2)$$

where, $y_m \in \{0,1\}$ is the label for sample m, p_m is the predicted logit, σ (·) is the logistic sigmoid function, and M is the number of samples in a batch. The total loss $\mathcal{L}_{total}$ combines BCE and the average MSE from the three Fourier heads, as in Eq. (3):

$$\mathcal{L}_{total} = \mathcal{L}_{BCE} + \frac{1}{3}\sum_{k\in\{3,6,7\}}\mathcal{L}_{MSE,k} \quad (3)$$

Thus, by enforcing frequency consistency, the model is encouraged to learn both spatial and frequency-based characteristics pertinent to face spoofing detection.

B. Dataset

All experiments are conducted on the Oulu-NPU [33] dataset illustrated in Fig. 3, which contains 4,950 recorded videos. From each video, 8 frames at indices {10,30,50,70,90,100,120,140} were extracted, totaling approximately 39,000 images. These videos contain:

- Six smartphone cameras (Samsung Galaxy S6 edge, HTC Desire EYE, MEIZU X5, ASUS Zenfone Selfie, Sony XPERIA C5 Ultra Dual, OPPO N3).
- Three recording sessions with varying illumination and backgrounds.
- Two display types and two printer types.
- 55 participants.
- Four standard protocols for training and testing under different environmental and cross-device conditions.

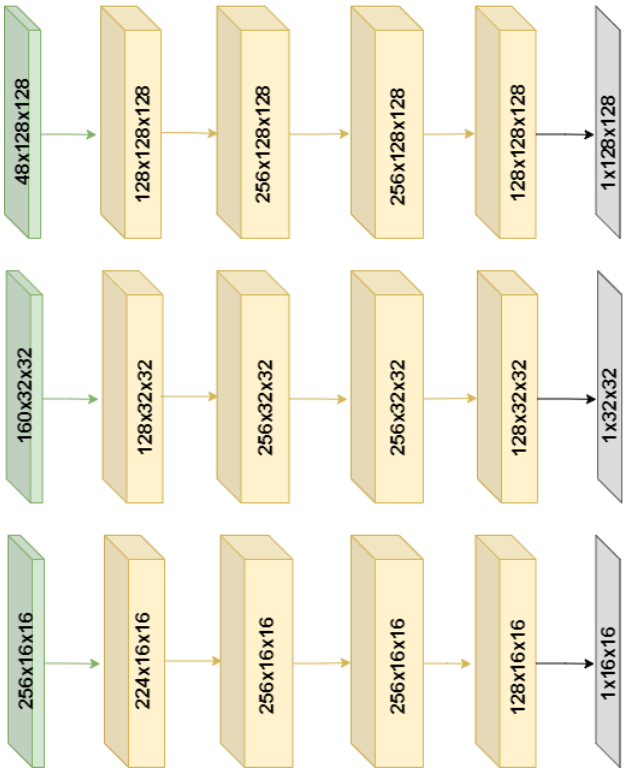


Fig. 2. Custom Fourier heads. From top to bottom: after 3rd block, after 6th block and after 7th block.

C. Experimental Setup

All models are trained and evaluated on a workstation with an NVIDIA RTX 4090 GPU (24 GB VRAM), 64 GB of CPU memory, and 24 CPU cores. The main hyperparameters are: batch size 16, Adam optimizer with weight_decay = 5×10^{-5}, number of epochs 30, ReduceLROnPlateau scheduler with patience = 3, reduce factor = 0.5. The metric used to trigger reduction in the learning rate is APCER, with an initial learning rate: 1×10^{-4}. Preliminary experiments showed that higher learning rates as 10^{-2} or 10^{-3} led to poor convergence, even when the Shuffle Attention block was removed. Consequently, a smaller learning rate with a pre-trained backbone yielded more stable results, suggesting that retaining initialization from EfficientNetV2-S is essential for this task. Each training epoch processes every sample in the dataset once. Early stopping is triggered if no improvement is observed in the validation metrics for more than five epochs, and the best model snapshot based on the validation set is preserved for final testing.

Fig. 3. Dataset samples. From left to right: live face, printed attack and replay attack.

Several transformations are applied to modify image properties while preserving essential features. Horizontal flipping reverses the image along the vertical axis, helping the model generalize to different orientations. Adjustments in contrast, gamma, and brightness alter pixel intensity distributions, simulating variations in lighting conditions. Elastic transformations, grid distortions, and optical distortions introduce non-linear deformations, making the model more resilient to geometric variations. Cutout randomly masks portions of the image, encouraging the model to focus on essential patterns rather than specific regions. Shift, scale, and rotation transformations modify the position and size of objects, aiding in generalization across different viewpoints.

IV. Results

A. Evaluation Metrics

In our evaluation of biometric authentication systems, we use three key metrics: Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER), and Average Classification Error Rate (ACER).

APCER [Eq. (4)] measures the frequency of falsely accepting an attack attempt. A higher APCER indicates a greater likelihood of incorrectly classifying an attack as a genuine attempt. BPCER [Eq. (5)] represents the proportion of falsely rejected genuine presentations. A high BPCER means that genuine users are more frequently denied access. ACER [Eq. (6)] provides an overall error measurement by averaging APCER and BPCER. Lower ACER values indicate better system performance in distinguishing between genuine and attack presentations.

$$APCER = \frac{Number\ of\ False\ Accepts}{Total\ Number\ of\ Attack\ Presentations} \quad (4)$$

$$BPCER = \frac{Number\ of\ False\ Rejects}{Total\ Number\ of\ Genuine\ Presentations} \quad (5)$$

$$ACER = \frac{APCER+BPCER}{2} \quad (6)$$

B. Experimental Results

1) State-of-the-art comparison: The experimental outcomes were obtained from four distinct protocols designed to evaluate face presentation attack detection under various challenging conditions, as illustrated in Table I.

Protocol I focuses on generalization across environmental variations. Data recorded in three sessions with differing illumination and background settings were partitioned into training, development, and evaluation sets. The proposed method achieved zero errors, with APCER, BPCER, and ACER all measuring 0%. An analysis of the error metrics confirms an exceptional ability to differentiate between authentic and attack samples under fluctuating lighting and background conditions.

Protocol II examines the impact of novel attack artifacts generated by previously unseen print and video-replay attacks. A new print attack and a video-replay attack were incorporated into the test set to assess vulnerability to unconventional spoofing techniques. The approach recorded an ACER of 1.6%, with APCER at 0.4% and BPCER at 2.8%. A detailed evaluation of the figures reveals a robust performance in minimizing false acceptance errors, demonstrating a strong capacity to counteract artifacts introduced by diverse display and printing sources.

Protocol III investigates sensor interoperability using a Leave One Camera Out setup. Video recordings from five smartphones contributed to training and tuning, while recordings from a sixth device provided the testing material. An ACER of $0.8 \pm 0.6\%$ was obtained, with both APCER and BPCER reflecting similarly low error margins. Consistency in performance across various camera sensors confirms a reliable level of operation among different acquisition devices. A careful review of the stability of these metrics confirms the method's applicability in scenarios involving heterogeneous imaging sensors.

Protocol IV presents the most challenging scenario by combining environmental variations, attack types, and sensor diversity. Under these complex conditions, the method achieved an ACER of $1.9 \pm 2.0\%$, outperforming competing approaches that exhibited higher error rates. Error metrics for both APCER and BPCER were noticeably reduced under the simultaneous influence of multiple factors.

2) Contribution of the modules and ablation: To prove that the combination of additional modules works, the different stages by adding modules progressively were tested. In Table II, the EfficientNet-V2-S model achieves APCER of 0.8%, BPCER of 3.8%, and ACER of 2.4%. Introducing a Fourier Heads at the 3rd, 6th, and 7th blocks results in lower error rates, reducing APCER to 0.7%, BPCER to 3.2%, and ACER to 2.1%, suggesting an improvement in attack detection.

Further refinement by integrating the Shuffle Attention module into those blocks leads to the best performance, with APCER dropping to 0.4%, BPCER to 2.8%, and ACER to 1.6%. The consistent reduction in error rates indicates that the combination of Fourier Heads and SA Module contributes to more effective feature extraction, strengthening the model's ability to distinguish genuine faces from spoofing attempts.

3) Dynamic thresholding: Dynamic thresholding is a technique used in biometric security systems to adaptively determine decision boundaries based on varying conditions. Unlike fixed thresholding, which applies a single static threshold across all samples, dynamic thresholding adjusts the classification boundary according to contextual factors such as environmental conditions, device variability, or individual user characteristics. This approach enhances the robustness of biometric systems by mitigating the impact of variations in data distributions and reducing misclassification errors. In presentation attack detection, dynamic thresholding is particularly effective in balancing security and usability, as it allows the system to refine its decision-making process based on real-time inputs.

TABLE I. EXPERIMENTAL RESULTS AND COMPARISON WITH DIFFERENT STATE-OF-THE-ART METHODS ON OULU-NPU DATASET

Protocol	Method	APCER (%)	BPCER (%)	ACER (%)
1	STASN [35]	1.2	2.5	1.9
	Auxiliary [36]	1.6	1.6	1.6
	Disentangle [37]	1.7	0.8	1.3
	STDN [38]	0.8	1.3	1.1
	CDCN [39]	0.4	1.7	1
	DC-CDN [40]	0.5	0.3	0.4
	NAS-FAS [41]	0.4	0	0.2
	PatchNet [42]	0	0	0
	TransFas [43]	0.8	0	0.4
	Ours	0	0	0
2	STASN [35]	4.2	0.3	2.2
	Auxiliary [36]	2.7	2.7	2.7
	Disentangle [37]	1.1	3.6	2.4
	STDN [38]	2.3	1.6	1.9
	CDCN [39]	1.5	1.4	1.5
	DC-CDN [40]	0.7	1.9	1.3
	NAS-FAS [41]	1.5	0.8	1.2
	PatchNet [42]	1.1	1.2	1.2
	TransFas [43]	1.5	0.5	1
	Ours	0.4	2.8	1.6
3	STASN [35]	4.7±3.9	0.9±1.2	2.8±1.6
	Auxiliary [36]	2.7±1.3	3.1±1.7	2.9±1.5
	Disentangle [37]	2.8±2.2	1.7±2.6	2.2±2.2
	STDN [38]	1.6±1.6	4.0±5.4	2.8±3.3
	CDCN [39]	2.4±1.3	2.2±2.0	2.3±1.4
	DC-CDN [40]	2.2±2.8	1.6±2.1	1.9±1.1
	NAS-FAS [41]	2.1±1.3	1.4±1.1	1.7±0.6
	PatchNet [42]	1.8±1.5	0.6±1.2	1.2±1.3
	TransFas [43]	0.6±0.7	1.1±2.5	0.9±1.1
	Ours	0.8±0.5	0.8±0.8	0.8±0.6
4	STASN [35]	6.7±10.6	8.3±8.4	7.5±4.7
	Auxiliary [36]	9.3 ± 5.6	10.4 ± 6.0	9.5 ± 6.0
	Disentangle [37]	5.4±2.9	3.3±6.0	4.4±3.0
	STDN [38]	2.3±3.6	5.2±5.4	3.8±4.2
	CDCN [39]	4.6±4.6	9.2±8.0	6.9±2.9
	DC-CDN [40]	5.4±3.3	2.5±4.2	4.0±3.1
	NAS-FAS [41]	4.2±5.3	1.7±2.6	2.9±2.8
	PatchNet [42]	2.5±3.8	3.3±3.7	2.9±3.0
	TransFas [43]	2.1±2.2	3.8±3.5	2.9±2.4
	Ours	2.5±2.5	1.2±2.1	1.9±2.0

TABLE II. INDIVIDUAL CONTRIBUTION OF EACH MODULE TESTED ON PROTOCOL II

Method	APCER (%)	BPCER (%)	ACER (%)
EfficientNet-V2-S	0.8	3.8	2.4
EfficientNet-V2-S w/ 3rd, 6th, 7th blocks Fourier Heads	0.7	3.2	2.1
EfficientNet-V2-S+SA Module w/ 3rd, 6th, 7th blocks Fourier Heads	**0.4**	**2.8**	**1.6**

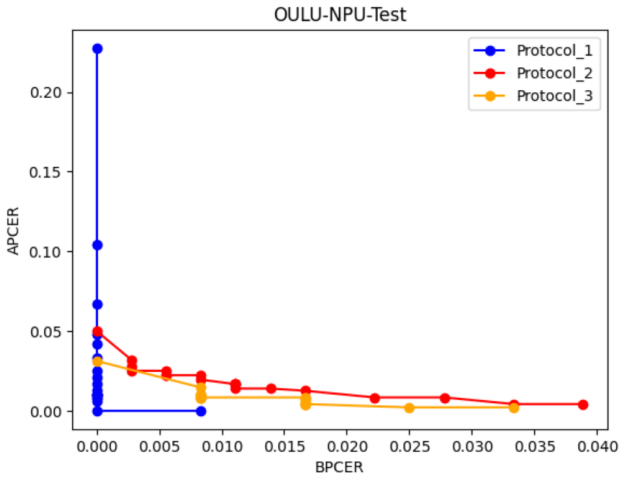


Fig. 4. Dynamic thresholding for 3 protocols.

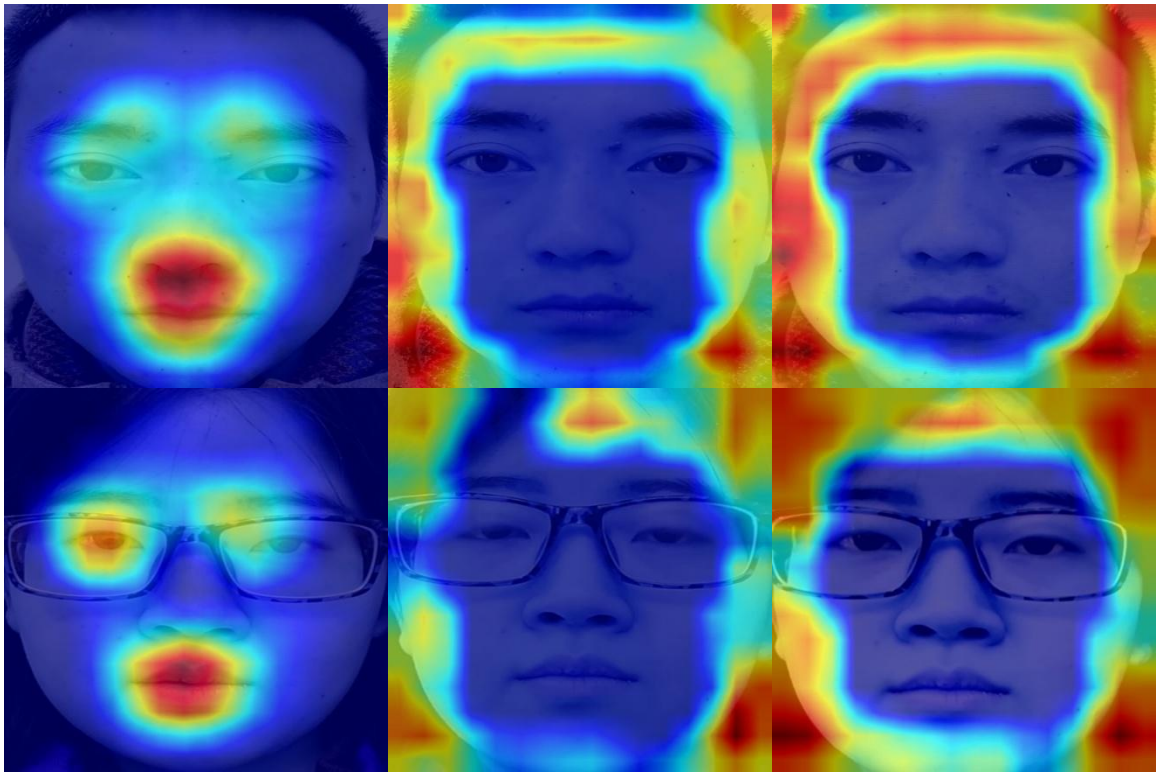
Fig. 4 illustrates the performance of three different protocols under dynamic thresholding on the OULU-NPU-Test dataset. The *x*-axis represents the BPCER, which indicates the proportion of genuine samples misclassified as attacks. The *y*-axis corresponds to the APCER, reflecting the proportion of attack presentations misclassified as genuine. Lower values on both axes indicate a more effective system, as it minimizes both types of errors.

Protocol I, represented in blue, demonstrates a steep decline in APCER as BPCER increases. At very low BPCER values, APCER remains considerably high, indicating that the system is prone to false acceptance when a strict threshold is applied. This pattern suggests that the protocol may be highly sensitive to variations in presentation attacks but suffers from instability when distinguishing bona fide samples. The rapid decrease in APCER as BPCER increases further suggests that a slight relaxation of the threshold significantly improves attack detection, albeit at the cost of increased false rejection.

Protocol II, visualized in red, follows a more stable trajectory, exhibiting a gradual reduction in APCER as BPCER increases. Compared to Protocol I, this approach achieves a more balanced trade-off between false acceptance and false rejection, indicating that it may be more reliable in practical deployment scenarios. The smoother trendline suggests that Protocol II is less sensitive to minor variations in input data, making it a more consistent option for presentation attack detection.

Protocol III, shown in orange, consistently maintains the lowest APCER across the observed BPCER range. The results indicate that this protocol outperforms the other two in minimizing attack presentation errors while maintaining a relatively low bona fide classification error. The stability of the curve suggests that Protocol III achieves a well-calibrated balance between security and accessibility, making it the most effective among the three.

4) Grad-Cam results: Grad-Cam [34] results in Fig. 5 provide an insightful interpretation of the deep learning model's decision-making process for detecting presentation attacks. Heatmaps indicate which facial regions contribute most to classification, revealing distinct patterns across live, print, and video-replay samples.



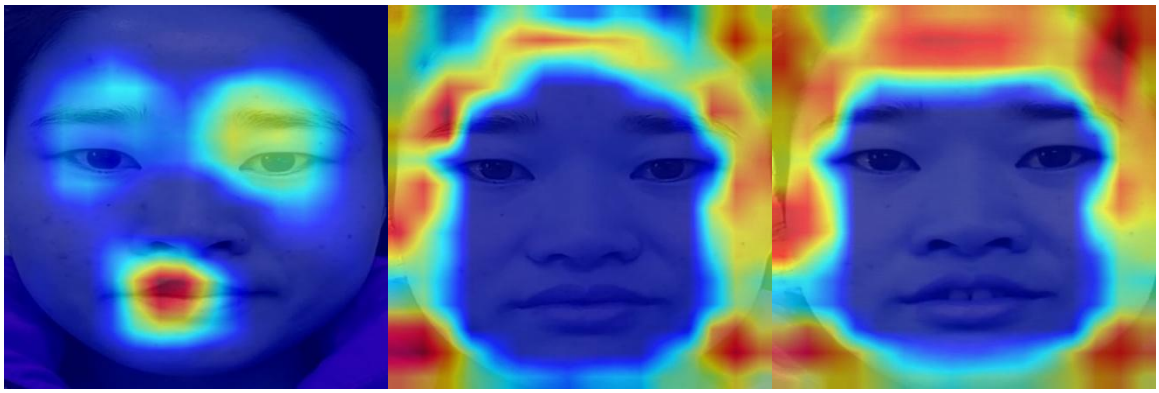


Fig. 5. Grad-Cam results. From left to right columns: live, print, and video-replay.

In live face images, attention is primarily focused around the mouth and nose, where color, texture, and depth cues play a significant role in distinguishing genuine faces. The uniform blue background suggests that the model effectively ignores non-relevant areas, reinforcing its reliance on intrinsic facial details.

For print and video-replay attacks, activations shift towards facial boundaries, where red and yellow highlights suggest the model detects inconsistencies in edges and lighting artifacts. The presence of strong activations around the periphery indicates an effort to identify unnatural textures, which are often present in spoofing attempts.

Images with glasses introduce additional complexity, with attention distributed around the frames and nose bridge. Higher activations in print and video attacks suggest reflections and distortions are key factors in classification.

Variations in heatmap intensity between attack types confirm the model's ability to differentiate genuine and spoofed faces. Attention shifts in predictable ways, aligning with expected signs of presentation attacks, demonstrating the model's capacity to detect spoofing artifacts with reliability.

Overall, the experimental results indicate several advantages of the proposed method. The approach demonstrates flawless performance in the face of environmental variations, maintains a low false acceptance rate against novel attack artifacts, and exhibits consistent error rates across different sensor inputs. When confronted with a combination of challenges, it delivers competitive performance with reduced error metrics, suggesting promising applicability for real-world face presentation attack detection systems.

V. Discussion

The experimental results indicate a promising advancement in face presentation attack detection. Findings confirm that spatial and frequency domain information combined into a unified architecture can significantly improve detection accuracy. The adoption of EfficientNet-V2-S as the backbone, combined with a Shuffle Attention module, has provided a robust framework capable of extracting discriminative features even under adverse conditions. Integration of Fourier heads to reconstruct frequency maps introduces a novel perspective that captures subtle differences between live and spoofed faces, as evidenced by the low error rates across various protocols.

Performance under controlled environmental variations demonstrated strong discrimination between genuine and attack samples. Minimal errors in protocols addressing different illumination and background conditions suggest that the network effectively adapts to changes in lighting and scene composition. Robust outcomes in scenarios involving previously unseen print and video-replay attacks further support the method's aptitude for recognizing unconventional spoof artifacts. Results reveal a capacity for mitigating false acceptance errors, which remains critical in maintaining the integrity of biometric systems.

Interoperability across different sensor types has been confirmed through experiments using diverse smartphone cameras. Consistent performance in sensor-specific testing scenarios reinforces the network's reliability, even when presented with imaging data acquired from heterogeneous sources. Outcomes recorded in the most challenging scenario, where multiple factors were simultaneously present, indicate competitive error metrics that rival or surpass existing methods. Low Average Classification Error Rates across protocols suggest that the design effectively balances the detection of genuine presentations with the rejection of spoof attacks.

Observations imply that the fusion of spatial features and frequency representations addresses limitations inherent in earlier approaches based solely on handcrafted or spatial features. The network's ability to capture minute texture and motion cues, while simultaneously processing global frequency information, contributes to its overall robustness. A careful analysis of the error metrics shows potential avenues for further exploration, particularly regarding the model's performance in real-world environments where attack types may be more diverse.

Limitations related to domain shifts between training and deployment environments persist, inviting additional research to refine generalization capabilities. Future investigations might consider the incorporation of alternative modalities or additional attention mechanisms to further improve detection rates under evolving spoof conditions. Overall, the findings present compelling evidence that integrating modern network architectures with both spatial and frequency domain analysis provides a viable pathway toward more reliable and secure face presentation attack detection systems.

VI. Conclusion

This work demonstrates that explicitly coupling spatial features with frequency-aware supervision yields robust face presentation attack detection across environmental, attack-type, and sensor variations. An EfficientNetV2-S backbone provides an effective accuracy-latency balance for deployment; Fourier heads guide the network to preserve high-frequency spoof cues that are often attenuated in spatial pipelines; and, Shuffle Attention efficiently enhances discriminative regions without heavy computation. Across Oulu-NPU's four protocols, the system attains 0% ACER in Protocol I, 1.6% in Protocol II, 0.8 ± 0.6% in Protocol III, and 1.9 ± 2.0% in Protocol IV, indicating strong generalization to novel artifacts and unseen sensors. Beyond raw numbers, the ablations and interpretability results clarify why the method works: frequency supervision reduces bona-fide rejections under photometric variation, while attention improves rejection of attack-specific structures, together producing consistent gains.

For security-sensitive biometric applications, reducing APCER directly lowers the risk of unauthorized access, and lowering BPCER preserves user experience. The proposed design advances practical anti-spoofing by improving both, without sacrificing throughput or model size, making it suitable for edge and mobile deployments. Performance under severe domain shift can still degrade. Future work will: i) incorporate temporal cues (e.g., rPPG or micro-motion) alongside frequency supervision, ii) explore self-supervised or domain-generalization objectives to handle unseen attacks, and iii) calibrate dynamic thresholds using cohort/device context to further stabilize APCER-BPCER trade-offs in the field.

Acknowledgments

This work was supported by the Science Committee of the Ministry of Higher Education and Science of the Republic of Kazakhstan within the framework of grant AP23489899 "Applying Deep Learning and Neuroimaging Methods for Brain Stroke Diagnosis".

References

[1] Patel, D., Shah, H., & Kumar, P. (2018). Liveness detection: A review of traditional and deep learning methods. Journal of Visual Communication and Image Representation, 54, 234–245.

[2] Kumar, P., Patel, D., & Shah, H. (2017). Anti-spoofing techniques for biometric systems. IEEE Transactions on Information Forensics and Security, 12(8), 1870–1884.

[3] Li, S., Liu, Z., & Zhang, H. (2019). A survey on face anti-spoofing methods. Pattern Recognition Letters, 115, 61–72.

[4] Anjos, A., Menotti, D., & Batista, G. (2020). Deep learning methods for face anti-spoofing. IEEE Transactions on Biometrics, Behavior, and Identity Science.

[5] Wang, L., Yu, G., & Li, X. (2021). Fourier analysis in face anti-spoofing. IEEE Access, 9, 102354–102363.

[6] Altayeva, A., Omarov, B., Jeong, H. C., & Im Cho, Y. (2016). Multi-step face recognition for improving face detection and recognition rate. Far East Journal of Electronics and Communications, 16(3), 471.

[7] Yang, Y., Chen, Z., & Liu, Z. (2020). Attention mechanisms in convolutional neural networks: A review. IEEE Signal Processing Magazine, 37(3), 75–85.

[8] Zhang, Y., Zhou, K., & Lin, J. (2021). A novel Fourier head for face anti-spoofing. In Proceedings of the International Conference on Pattern Recognition.

[9] Tan, M., & Le, Q. V. (2021). EfficientNetV2: Smaller models and faster training. arXiv preprint arXiv:2104.00298.

[10] Zhang, X., Yang, J., & Wang, L. (2020). Shuffle attention for neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[11] Zhang, Q.-L., & Yang, Y.-B. (2021). SA-Net: Shuffle attention for deep convolutional neural networks. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2235-2239. https://doi.org/10.1109/ICASSP39728.2021.9414568

[12] Feng, W., & Xie, X. (2019). Frequency domain analysis for robust face anti-spoofing. IEEE Transactions on Information Forensics and Security, 14(10), 2802–2813.

[13] Omarov, B., Baikuvekov, M., Sultan, D., Mukazhanov, N., Suleimenova, M., & Zhekambayeva, M. (2024). Ensemble Approach Combining Deep Residual Networks and BiGRU with Attention Mechanism for Classification of Heart Arrhythmias. Computers, Materials & Continua, 80(1).

[14] Guo, Y., Zhao, G., & Zhang, R. (2020). Exploring frequency cues in face anti-spoofing with deep networks. IEEE Access, 8, 75678–75689.

[15] Cheng, W., & Li, X. (2022). Continuous advances in face anti-spoofing:A review of recent methods. IEEE Transactions on Biometrics, Behavior, and Identity Science, 4(1), 1–14.

[16] Boulkenafet, I., Komulainen, J., & Hadid, A. (2016). Face spoofing detection from single images using spatio-temporal features. IEEE Transactions on Information Forensics and Security, 11(12), 2718–2732.

[17] Chingovska, I., Anjos, A., & Marcel, S. (2012). On the effectiveness of local binary patterns in face anti-spoofing. In 2012 IEEE International Joint Conference on Biometrics (IJCB) (pp. 1–7). IEEE.

[18] Patel, V. M., & Chellappa, R. (2013). Face anti-spoofing using support vector machines with handcrafted features. In Proceedings of the IEEE Workshop on Biometrics (pp. 1–6).

[19] Al Noman, M. A., Zhai, L., Almukhtar, F. H., Rahaman, M. F., Omarov, B., Ray, S., ... & Wang, C. (2023). A computer vision-based lane detection technique using gradient threshold and hue-lightness-saturation value for an autonomous vehicle. International Journal of Electrical and Computer Engineering, 13(1), 347.

[20] Akshay Dilip Lahe, Guddi Singh, "A Survey on Security Threats to Machine Learning Systems at Different Stages of its Pipeline", International Journal of Information Technology and Computer Science(IJITCS), Vol.15, No.2, pp.23-34, 2023. DOI:10.5815/ijitcs.2023.02.03

[21] Omarov, B., Tursynova, A., & Uzak, M. (2023). Deep learning enhanced internet of medical things to analyze brain computed tomography images of stroke patients. International Journal of Advanced Computer Science and Applications, 14(8).

[22] Atoum, Y., Liu, S., & Maybank, S. (2017). Face spoofing detection using patch-based convolutional neural networks. In 2017 International Joint Conference on Biometrics (IJCB) (pp. 1–8). IEEE.

[23] Li, J., Liao, R., & Liu, M. (2018). Deep texture features for face anti-spoofing. IEEE Transactions on Image Processing, 27(9), 4489–4501.

[24] Yang, Y., Liu, Z., & Chen, X. (2018). Challenges in face anti-spoofing: A review. IEEE Transactions on Biometrics, Behavior, and Identity Science, 1(2), 107–119.

[25] Omarov, B., Batyrbekov, A., Dalbekova, K., Abdulkarimova, G., Berkimbaeva, S., Kenzhegulova, S., ... & Omarov, B. (2020, December). Electronic stethoscope for heartbeat abnormality detection. In International Conference on Smart Computing and Communication (pp. 248-258). Cham: Springer International Publishing.

[26] Bala Dhandayuthapani V., "Enhancing Jakarta Faces Web App with AI Data-Driven Python Data Analysis and Visualization", International Journal of Information Technology and Computer Science(IJITCS), Vol.16, No.5, pp.36-51, 2024. DOI:10.5815/ijitcs.2024.05.03

[27] Lin, J. D., Lin, H. H., Dy, J., Chen, J. C., Tanveer, M., Razzak, I., & Hua, K. L. (2021). Lightweight face anti-spoofing network for telehealth applications. IEEE Journal of Biomedical and Health Informatics, 26(5), 1987-1996.

[28] Ifeoluwani Jenyo, Elizabeth A. Amusan, Justice O. Emuoyibofarhe, "A Trust Management System for the Nigerian Cyber-health Community",

International Journal of Information Technology and Computer Science(IJITCS), Vol.15, No.1, pp.9-20, 2023. DOI:10.5815/ijitcs.2023.01.02

[29] Adeniran, A. A., Onebunne, A. P., & William, P. (2024). Explainable AI (XAI) in healthcare: Enhancing trust and transparency in critical decision-making. World Journal of Advanced Research and Reviews, 23(3), 2647-2658.

[30] Liu, M., & Liu, X. (2020). Future directions in face anti-spoofing research. IEEE Signal Processing Magazine, 37(5), 60–70.

[31] Wang, Y., Zhang, Z., & Li, W. (2021). Integrating frequency and spatial cues for robust face anti-spoofing. IEEE Transactions on Multimedia, 23, 1572–1583

[32] Zhang, K., Zhang, Z., & Li, Z. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. In 2016 IEEE Signal Processing Letters (Vol. 23, pp. 1499-1503). IEEE. https://doi.org/10.1109/LSP.2016.2603342

[33] Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., & Hadid, A. (2017). OULU-NPU: A mobile face presentation attack database with real-world variations. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017).

[34] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision, 128(4), 1187–1200. https://doi.org/10.1007/s11263-019-01228-7

[35] Yang, X., Luo, W., Bao, L., Gao, Y., Gong, D., Zheng, S., Li, Z., Liu, W.: Face anti-spoofing: Model matters, so does data. In: proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3507–3516 (2019)

[36] Liu, Y., Jourabloo, A., Liu, X.: Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 389–398 (2018)

[37] Ke-Yue Zhang, Taiping Yao, Jian Zhang, Ying Tai, Shouhong Ding, Jilin Li, Feiyue Huang, Haichuan Song, and Lizhuang Ma. Face anti-spoofing via disentangled representation learning. In ECCV, 2020. 1, 5, 6

[38] Y. Liu, J. Stehouwer, and X. Liu, “On disentangling spoof trace for generic face anti-spoofing,” in Proc. ECCV, 2020, pp. 406–422.

[39] Z. Yu et al., “Searching central difference convolutional networks for face anti-spoofing,” in Proc. CVPR, 2020, pp. 5295–5305.

[40] Z. Yu, Y. Qin, H. Zhao, X. Li, and G. Zhao, “Dual-cross central difference network for face anti-spoofing,” in Proc. IJCAI, 2021, pp. 1281–1287.

[41] Z. Yu, J. Wan, Y. Qin, X. Li, S. Z. Li, and G. Zhao, “NAS-FAS: Static-dynamic central difference network search for face anti-spoofing,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 9, pp. 3005–3023, Sep. 2021.

[42] C.-Y. Wang, Y.-D. Lu, S.-T. Yang, and S.-H. Lai, “PatchNet: A simple face anti-spoofing framework via fine-grained patch recognition,” in Proc. CVPR, 2022, pp. 20281–20290.

[43] Z. Wang, Q. Wang, W. Deng, and G. Guo, "Face Anti-Spoofing Using Transformers With Relation-Aware Mechanism," IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 4, no. 3, July 2022, p. 439.