Multi-Stream CNN with Attention for Single-Image Face Spoofing Detection

Rafał Klinowski[0009−0007−5055−5679] and
Mirosław Kordos[0000−0002−2031−7561]

University of Bielsko-Biała,
Department of Computer Science and Automatics,
43-309 Bielsko-Biała ul. Willowa 2, Poland
rklinowski@student.ubb.edu.pl, mkordos@ubb.edu.pl

Abstract. We present a study of a face spoofing detection system based on a multi-stream CNN architecture that operates on a single RGB image taken with a simple camera, such as the front-facing camera of a smartphone. The network extracts and processes features through four parallel streams: a local texture convolution with a Convolutional Block Attention Module, a color analysis with large-kernel convolutions, a context analysis with dilated convolutions, and frequency analysis with the Fast Fourier Transform to identify high-frequency artifacts. The proposed architecture achieved very high accuracy (for example, 99.28% accuracy, and an AUC value of 99.94% on the CelebA-Spoof dataset, and 97.13% accuracy and 99.18% AUC for the CASIA-FAS dataset).

Keywords: Face Anti-Spoofing · Convolutional Neural Network · Multi-Stream Network · Attention Mechanism

1 Introduction

Face spoofing detection determines whether a face presented to a camera is an authentic face of a real person who is in front of the camera, or is displayed on an electronic device or in another way (spoof) to cheat the system.

A single image contains many different features, such as texture details, color distribution, context of objects, or its representation in the frequency domain. These features can be extracted and processed using different models. Alternatively, the different models can be incorporated into a single multi-modal model.

Such multi-modal models allow for deep feature extraction, which is especially relevant for face anti-spoofing due to the variety of spoofing methods used. It is important that the models generalize well by learning patterns that allow them to effectively detect spoofing attempts of different types in variable conditions and environments. In particular, the models can be implemented by various deep neural networks. Compared to an ensemble of several models that work in parallel, the use of a single multi-stream network can be beneficial due to its ability to synthesize diverse feature representations easier optimization of parameters of the entire model. The simultaneous processing of the input by several streams

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_43

allows the network to capture complex relationships between these features and classify the images based on a broader feature set, as opposed to the use of several independent methods and processing only their final predictions, which in turn further help the network generalize and perform well on unseen data.

This paper proposes a multi-stream neural network architecture for classifying faces as authentic or spoof. Unlike methods proposed by other authors, our method detects spoofing attempts based on features extracted from only a single RGB image taken with a simple camera, such as a USB camera or a front-facing camera of a smartphone. Though the use of a simple camera leads to less information contained in the image, it is also an important research topic to allow face anti-spoofing systems to work without the requirement of specialized hardware, such as infrared or 3D cameras.

2 Literature Review

Face anti-spoofing research has been approached from many directions. One of them is the use of an ensemble of models (deep learning methods and other algorithms) that each focus on different qualities of the image, as presented in [7]. Although this approach is fast, efficient, uses only a single RGB image for classification, and shows high accuracy, the downside is that it is vulnerable to variability in environmental conditions, such as lighting and background. Additionally, the use of an ensemble does not allow the method to fully capture relationships between different features, as each model makes a classification based on a narrower feature set, and the final decision is made based only on these single partial classifications and not on their interactions.

For that reason some research on face anti-spoofing focuses on developing architectures that process several images taken simultaneously (multi-modal networks). Specifically, works like the paper by H. Xue, J. Ma, X. Guo [15] focus on developing multi-modal networks that process: an RGB image, an IR image, a depth image, and a thermal image, all at once, to produce the classification result. Similar research has been conducted by G. Chen et al. [4], who propose a multi-stream network trained on three modalities: RGB, infrared, and depth. Both papers propose solutions that achieve high accuracy, though they require very specific hardware to capture several types of images simultaneously, which is a limiting factor for the use of such systems. In our case, we do not have multiple modalities available in the source photos.

The paper by N. Li et al. [9] presents a convolutional neural network based on the Dual-path Adaptive Channel Attention module, which filters the features of the input facial images to extract key information. It also presents a feature constraints method based on Inner Similarity Estimation, which enhances intra-class consistency by reducing the distance between samples and their class center, and thus improving class separability. However, the authors note that the performance of their method may be affected by the use of low-quality images, such as those produced by a simple RGB camera.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_43

X. Shu et al. [11] propose a dual-stream architecture operating on just RGB images, where one stream processes the image directly, and the other stream converts to grayscale to extract face reflection features. The proposed architecture also includes a paralleled CBAM (PCBAM) attention block. The experiments performed by the authors show promising results, although the authors note that many existing methods, including the one proposed by them, are susceptible to lighting conditions.

3 Proposed Method

Our proposed method uses an architecture consisting of four parallel streams:

- Local Texture with Attention Mechanism: a CNN architecture with three convolutional blocks and an Attention Block for detecting local patterns;
- Color Analysis: a stream for detecting global color distribution and anomalies by using larger convolution kernels;
- Context Analysis: a stream with dilated convolutions for broader, global context analysis;
- Frequency Analysis: a network stream that operates on images processed by the Fast Fourier Transform [10], detecting high-frequency artifacts characteristic of spoofing attempts. This stream learns the frequency domain features based on a representation obtained after applying the FFT.

All four streams operate on RGB images of size (256×256). As a standard practice, the images are normalized to ImageNet normalization values of $mean = [0.485, 0.456, 0.406]$ and $std = [0.229, 0.224, 0.225]$, to ensure that all channel information is equally relevant and to set an accurate baseline [2].

In the proposed method, only the first stream uses the attention mechanism; it amplifies texture analysis by helping the network focus on relevant details rather than general features. Applying the Attention mechanism to the Color and Context Analysis streams is unnecessary, as they focus on global feature distribution rather than local details.

The frequency analysis stream requires additional preprocessing, which includes a conversion to grayscale and the use of the Fast Fourier Transform to convert the data to the frequency domain. This preprocessing is performed internally within the network.

We evaluated two attention block architectures:

- The simpler architecture proposed by J. Hu et al. in "Squeeze and Excitation Networks" [5];
- The more complex architecture of the Convolutional Block Attention Module, proposed by S. Woo et al. [14]

The results showed that the network performs marginally better with the latter attention block architecture, which is why it was used in the finalized

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_43

multi-stream network. Due to space constraints within this paper, for detailed descriptions of these attention block architectures, we refer to the cited works.

The total number of trainable parameters for the proposed network architecture is $1,954,347$, which ensured short training times of 22.8 minutes per epoch for the largest CelebA-Spoof dataset using our software and hardware specified at the end of the next subsection. This also allowed for easy iteration of the architecture and parameters. A diagram of this network is shown in Figure 1.

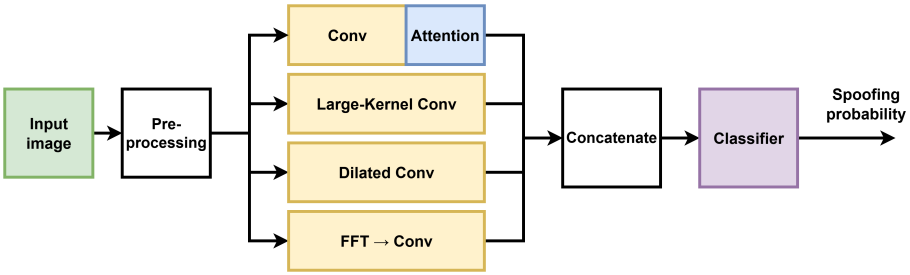

Fig. 1: Diagram showing the proposed multi-stream network architecture. Once processed, the feature maps are flattened and concatenated, and then passed to the classifier network.

4 Experimental Evaluation

4.1 Setup

To test the proposed multi-stream architecture, several datasets were used, ranging from a few hundred images to over 600,000. We used the following datasets:

- CelebA-Spoof with $N = 546,702$ images (183,964 images of authentic faces and 362,738 images of spoof faces). It was the largest of the datasets used, and for that reason it was chosen as the baseline for model tuning [17];
- Large Crowdcollected Facial Anti-Spoofing Dataset with $N = 16,284$ images (1,942 images of authentic faces and 14,342 images of spoof faces) [12];
- Spoofing Dataset with $N = 2412$ images (1,261 images of authentic faces and 1,151 images of spoof faces) [3];
- A Small Dataset with $N = 477$ images (243 images of authentic faces and 234 images of spoof faces), partly collected by the authors and partly adopted from the previously mentioned datasets;
- The CASIA Face Anti-Spoofing with $N = 4,063$ images (995 images of authentic faces and 3,068 images of spoof faces) [8];

The CASIA Dataset was split into training and test sets by the original authors to prevent data leakage, and the split was retained. For all other datasets, a 5-fold stratified cross-validation was performed.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_43

Our code was written using the PyTorch 2.2.2 with Python 3.12 and used a constant random seed value of 42. All experiments were performed on computers equipped with Windows Server 2019, 64 GB RAM, Intel Core Ultra 9 285K, NVIDIA RTX 5080 GPUs with CUDA 12.8. The source code and the datasets used in this paper are available at `https://github.com/Stukeley/MultiStreamCNN`.

4.2 Results

During the development of the four stream network, an ablation study was performed. First, each individual stream was tested. The results are in Table 1.

Table 1: The results obtained for architectures consisting of a single stream, on the CelebA-Spoof dataset.

	Local Texture (no Attention)	**Local Texture (Attention)**	**Color Analysis**	**Context Analysis**	**Frequency Analysis**
Acc.(%)	98.23 ± 0.99	98.82 ± 0.19	96.08 ± 0.48	98.16 ± 0.34	71.91 ± 11.42
Prec.(%)	99.04 ± 0.49	98.77 ± 0.53	96.30 ± 1.78	98.04 ± 0.79	80.90 ± 9.63
Recall(%)	98.30 ± 1.86	99.47 ± 0.37	97.93 ± 1.83	99.21 ± 0.36	81.80 ± 28.59
F1(%)	98.66 ± 0.77	99.12 ± 0.14	97.08 ± 0.35	98.62 ± 0.25	75.95 ± 18.16

Then, starting from the Local Texture stream, which had the best performance, subsequent streams were added in the order of their performance, which created the following configurations:

- Two streams: Local Texture and Context Analysis;
- Three streams: Local Texture, Context, and Color Analysis;
- Four streams: Local Texture, Context, Color, and Frequency Analysis.

The results of this analysis are shown in Table 2. The same configurations were tested with a soft voting classifier, the results of which are in Table 3. The analysis proves that the introduction of additional network streams increases the network's performance.

Table 2: The results obtained for different numbers of proposed neural network streams on the CelebA-Spoof dataset.

	Texture and Context	**Texture, Context and Color**	**Texture, Context, Color, and Frequency**
Accuracy (%)	98.94 ± 0.14	99.21 ± 0.06	99.37 ± 0.07
Precision (%)	98.82 ± 0.32	99.25 ± 0.24	99.34 ± 0.20
Recall (%)	99.60 ± 0.15	99.57 ± 0.19	99.72 ± 0.15
F1 (%)	99.21 ± 0.10	99.41 ± 0.04	99.53 ± 0.05

After initial evaluation, the architecture was finalized - networks were made deeper by adding more convolution blocks, and more complex by increasing the number of channels in each layer to further improve performance, which increased

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_43

Table 3: The results obtained with a voting classifier for different numbers of proposed neural network streams on the CelebA-Spoof dataset.

	Texture and Context	**Texture, Context and Color**	**Texture, Context, Color, and Frequency**
Accuracy (%)	98.29 ± 0.56	98.45 ± 0.34	97.98 ± 0.57
Precision (%)	97.72 ± 0.90	98.40 ± 0.72	97.21 ± 0.89
Recall (%)	99.76 ± 0.90	99.29 ± 0.62	99.83 ± 0.10
F1 (%)	98.73 ± 0.41	98.84 ± 0.25	98.50 ± 0.42

the total number of parameters to the previously specified $1,954,347$. Next, a baseline was established based on the results obtained on the CelebA-Spoof dataset. Hyperparameters used for training were tuned for the final network architecture, using the Optuna framework [1]. The parameters were obtained through repeated evaluations of the network's performance on the CelebA-Spoof dataset, each time using the same training and test split of the data in a 3-fold stratified cross-validation over 5 training epochs.

After parameter tuning, the following values were used: A learning rate of $2.6e^{-4}$, a weight decay of $1.8e^{-3}$, the AdamW optimizer, a train split of 80%, a batch size of 32, 20 epochs, a fully-connected layer size of the classifier of 384, a dropout rate of 0.43, and the threshold that separates faces classified as authentic from those classified as spoof of 0.3.

Table 4: Performance metrics across evaluated datasets, part 1.

	Small Dataset	**Spoofing Dataset**	**Large Crowdcollected**
Accuracy (%)	95.18 ± 1.94	98.76 ± 0.42	99.20 ± 0.27
Precision (%)	92.74 ± 1.97	98.45 ± 0.69	99.27 ± 0.43
Recall (%)	97.87 ± 2.33	98.96 ± 0.21	99.83 ± 0.16
F1 (%)	95.23 ± 1.92	98.70 ± 0.43	99.55 ± 0.15
AUC (%)	99.14 ± 0.90	99.86 ± 0.11	99.81 ± 0.16
APCER (%)	7.41 ± 1.00	1.43 ± 1.07	5.65 ± 0.99
NPCER (%)	2.14 ± 2.48	1.04 ± 0.72	0.17 ± 0.09
ACER (%)	4.77 ± 1.68	1.24 ± 0.62	2.91 ± 0.46

Table 5: Performance metrics across evaluated datasets, part 2. (*) denotes evaluation without the use of cross-validation.

	CelebA-Spoof	**CASIA (*)**
Accuracy (%)	99.28 ± 0.04	97.13
Precision (%)	99.07 ± 0.12	98.72
Recall (%)	99.85 ± 0.08	97.47
F1 (%)	99.46 ± 0.03	98.09
AUC (%)	99.94 ± 0.01	99.18
APCER (%)	1.85 ± 0.55	3.92
NPCER (%)	0.15 ± 0.08	2.54
ACER (%)	1.00 ± 0.23	3.23

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_43

Then, the performance of the network was evaluated on all aforementioned datasets based on the following metrics: accuracy, precision, recall, F1 score, and Area Under the Curve. All metrics assume "spoof" as the positive value for the classification (1), and "authentic" as the negative value (0). The results are in Tables 4 and 5. A comparison to the methods proposed by other authors in [6] [16] [17] [13] can be found in Table 6.

Table 6: Comparison of our proposed method with methods proposed by other authors for an intra-dataset evaluation on the CelebA-Spoof dataset.

	MobileNetV2 +PTA	**BASN**	**AENet$_{c,s,g}$**	**ChinaTelecom (CVPR 2023)**	**DPM**	**Ours**
APCER (%)	1.21	4.0	2.29	1.30	0.84	1.85 ± 0.55
NPCER (%)	3.05	1.1	0.96	1.91	0.82	0.15 ± 0.08
ACER (%)	2.13	2.6	1.63	1.60	0.83	1.00 ± 0.23

5 Conclusions

Based on the results obtained, we can see that the proposed method achieves very high accuracy, AUC, and ACER values, especially in comparison to other methods. These scores indicate that our multi-stream network performs well compared to the solutions proposed by other authors. By simultaneously processing four different qualities of the image - local texture, global color distribution, broader context, and high-frequency artifacts - the network acts as an ensemble, and therefore yields better results compared to a single-stream network. Additionally, the use of four streams within a single network makes it more efficient and allows it to classify based on features obtained from the streams, and not only on partial probabilities returned by methods within an ensemble.

A single threshold value of the network output to discern between authentic and spoof faces performs a binary classification, and discards the probability and context information. Considering also these values to improve the prediction will be the subject of our further research.

Our future research will also focus on a closer evaluation of the network's streams, especially we plan to perform further work on the network's streams' architecture. We also plan to evaluate the proposed architecture across different datasets to ensure that the network generalizes well and that it learns the specific patterns that indicate a spoofing attempt.

Acknowledgment

This research was funded by The National Centre for Research and Development of Poland (NCBiR), project nr POIR.01.01.01-00-1144/19.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_43

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD. p. 2623–2631. KDD '19, Association for Computing Machinery, New York, NY, USA (2019)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
3. Face: Spoofing dataset (sep 2024), https://universe.roboflow.com/face-hgc6e/spoofing-wqkzq, visited on 2025-02-18
4. Geng Chen, Wuyuan Xie, D.L.Y.L.M.W.: mmfas: Multimodal face anti-spoofing using multi-level alignment and switch-attention fusion. In: The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25). pp. 58–66 (2025)
5. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks (2019), https://arxiv.org/abs/1709.01507
6. Khabarlak, K.: Post-train adaptive mobilenet for fast anti-spoofing (2022), https://arxiv.org/abs/2207.13410
7. Klinowski, R., Kordos, M.: Face spoofing detection with stacking ensembles in work time registration system. Applied Sciences **15**(15) (2025), https://www.mdpi.com/2076-3417/15/15/8402
8. Li, A., Tan, Z., Li, X., Wan, J., Escalera, S., Guo, G., Li, S.Z.: Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing (2020), https://arxiv.org/abs/2003.05136
9. N Li, Z Weng, F.L.Z.L.W.W.: Dual-path adaptive channel attention network based on feature constraints for face anti-spoofing. IEEE Access **13**, 22855–67 (2025), https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10855450
10. Park, C.S.: 2d discrete fourier transform on sliding windows. IEEE Transactions on Image Processing **24**(3), 901–907 (2015)
11. Shu, X., Li, X., Zuo, X., Xu, D., Shi, J.: Face spoofing detection based on multi-scale color inversion dual-stream convolutional neural network. Expert Systems with Applications **224**, 119988 (2023), https://www.sciencedirect.com/science/article/pii/S0957417423004906
12. Timoshenko, D., Simonchik, K., Shutov, V., Zhelezneva, P., Grishkin, V.: Large crowdcollected facial anti-spoofing dataset. In: 2019 Computer Science and Information Technologies (CSIT). pp. 123–126 (2019)
13. Wang, D., Guo, J., Shao, Q., He, H., Chen, Z., Xiao, C., Liu, A., Escalera, S., Escalante, H.J., Lei, Z., Wan, J., Deng, J.: Wild face anti-spoofing challenge 2023: Benchmark and results (2023), https://arxiv.org/abs/2304.05753
14. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module (2018), https://arxiv.org/abs/1807.06521
15. Xue, H., Ma, J., Guo, X.: A hierarchical multi-modal cross-attention model for face anti-spoofing. Journal of Visual Communication and Image Representation **97**, 103969 (2023), https://www.sciencedirect.com/science/article/pii/S1047320323002195
16. Zhang, Y., Wu, Y., Yin, Z., Shao, J., Liu, Z.: Robust face anti-spoofing with dual probabilistic modeling. Pattern Recognition **167**, 111700 (2025), https://www.sciencedirect.com/science/article/pii/S0031320325003607
17. Zhang, Y., Yin, Z., Li, Y., Yin, G., Yan, J., Shao, J., Liu, Z.: Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In: European Conference on Computer Vision (ECCV) (2020), https://arxiv.org/abs/2007.12342

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_43