ATTENTION-BASED DUAL-STREAM VISION TRANSFORMER FOR RADAR GAIT RECOGNITION

Shiliang Chen[⋆] *Wentao He*[⋆] *Jianfeng Ren*[⋆] *Xudong Jiang*[†]

[⋆] The School of Computer Science, University of Nottingham Ningbo China
[†] School of Electrical & Electronic Engineering, Nanyang Technological University

ABSTRACT

Radar gait recognition is robust to light variations and less infringement on privacy. Previous studies often utilize either spectrograms or cadence velocity diagrams. While the former shows the time-frequency patterns, the latter encodes the repetitive frequency patterns. In this work, a dual-stream network with attention-based fusion is proposed to fully aggregate the discriminant information from these two representations. Both streams are analyzed through the Vision Transformer, which well captures the gait characteristics embedded in these representations. The proposed method is validated on a large benchmark dataset for radar gait recognition, showing that it significantly outperforms state-of-the-art solutions.

***Index Terms*—** Radar gait recognition, Spectrogram, Cadence velocity diagram, Vision transformer, Attention-based fusion

1. INTRODUCTION

Human gait recognition has become increasingly attractive in biometric applications such as public safety monitoring, health screening and human-computer interaction [1]. Traditional gait recognition methods [2, 3] often require videos captured from a side view and heavily depend on lighting conditions. Taking pictures of people may cause privacy issues. In contrast, radar can capture the micro-Doppler signatures (mDS) of front-view gait features from a moving target [4], revealing human dynamics and recognizing human motions and gestures [4–6]. More importantly, radar has less invasion of privacy and can work robustly in a variety of real-world situations such as dim conditions. In this paper, the problem of front-view human identification using radar mDS is studied.

Many mDS representations have been developed, *e.g.*, spectrograms [7–9], cadence velocity diagrams (CVD) [10] and cepstrograms [11]. The spectrogram is a time-varying representation of mDS in the time-frequency domain [4], which has been applied to human activity recognition [6, 12]. The time-varying gait information such as the swinging speed of arms and legs can be well encoded in the spectrogram. However, the gait spectrograms of different people have very little differences, while the same person may have different gait characteristics, which makes the human identification using radar gait features particularly challenging. The CVD is another mDS representation by taking Fast Fourier Transform of the spectrogram along the time axis [10]. The CVD is less studied for gait recognition [13]. It provides a useful measurement of repetition patterns for different velocities of body parts, which is a supplement to the spectrogram.

Traditional classifiers such as Naïve Bayes [14, 15], support vector machines [16] and k-nearest-neighbour classifiers [17] have been used to classify mDS, while deep convolutional neural networks (DCNNs) often produce better performance. DCNNs have been used on spectrograms for radar gesture recognition and action classification [4, 6, 18], while other representations such as CVD are not fully explored but could provide complementary information to spectrogram. It is hence advantageous to use both spectrogram and CVD for classification. In addition, those DCNNs directly adopted from image recognition tasks often ignore the unique physical nature of the radar signal compared to optical images.

In this work, an Attention-based Dual-Stream Vision Transformer (ADS-ViT) is proposed to recognize people through radar gaits. As spectrograms often focus on the short-time-varying nature of mDS only, the CVD is introduced to capture the information on how often different frequencies repeat across a long duration. The proposed dual-stream network extracts features from these two representations simultaneously. Both streams are designed based on the Vision Transformer (ViT) to deeply exploit the gait characteristics embedded in patches of spectrogram and CVD. Furthermore, an attention-based fusion mechanism is designed to optimally combine features from two streams. The proposed framework is validated on a large dataset for radar gait recognition and significantly outperforms state-of-the-art models.

Our contributions are two-fold: 1) The proposed ADS-ViT effectively extracts and fuses features from spectrogram and CVD for radar gait recognition. 2) Both streams of ADS-ViT utilize the patch-processing ability of ViT to effectively capture the gait characteristics embedded in patches corresponding to frequency bands of spectrogram and CVD.

This work was supported in part by the National Natural Science Foundation of China under Grant 72071116, and in part by the Ningbo Municipal Bureau of Science and Technology under Grant 2019B10026.

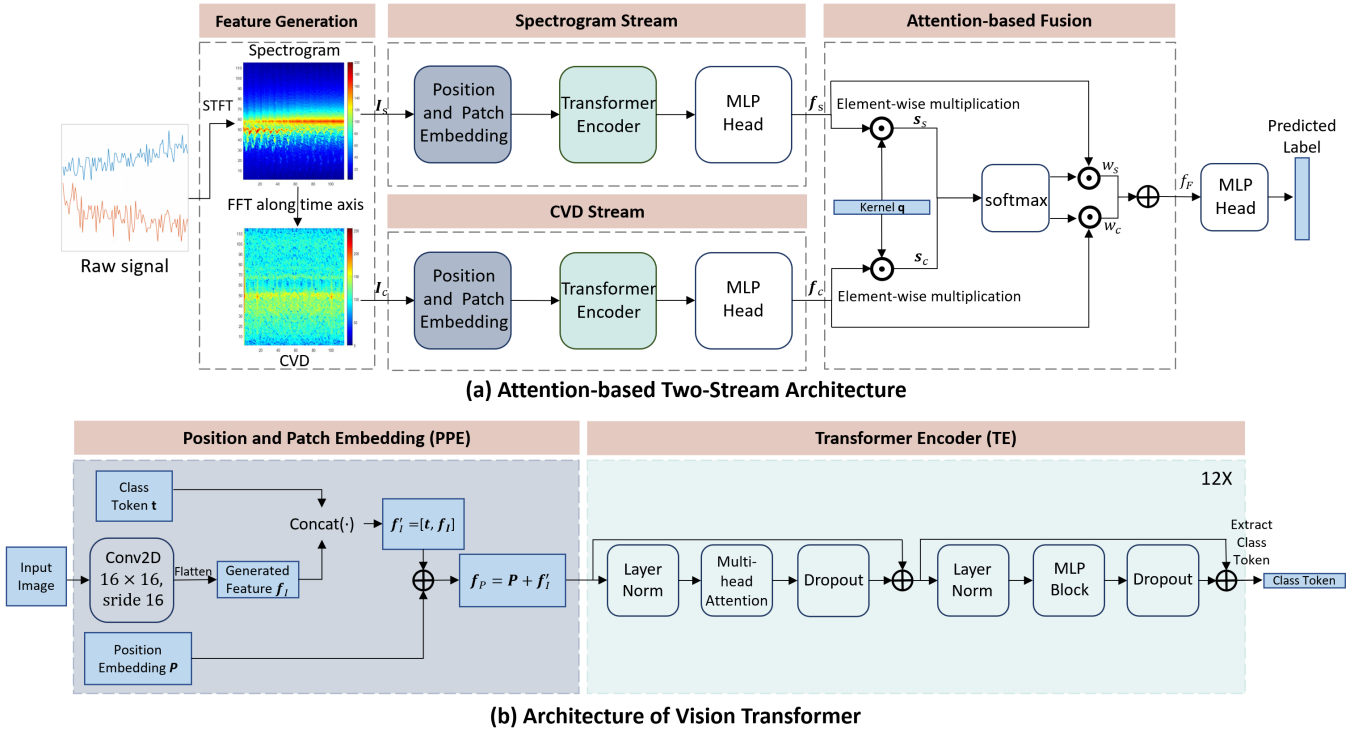

Fig. 1. Overview of the proposed method shown in (a) and pipeline of ViT shown in (b).

2. PROPOSED ATTENTION-BASED DUAL-STREAM VISION TRANSFORMER

2.1. Overview of Proposed Method

The proposed ADS-ViT for radar gait recognition is shown in Fig. 1(a). To better exploit the discriminant information embedded in the radar signal, two initial feature representations, spectrogram [8, 9] and CVD [10] are jointly utilized in the proposed framework. Two subnetworks, spectrogram stream and CVD stream, are used to extract features from each of these two representations, respectively. For spectrograms and CVDs, different image patches correspond to different frequency bands. To robustly extract low-level features from these frequency bands, a network based on the Vision Transformer (ViT) [19] is designed, where the local features are encoded through the Position and Patch Embedding mechanism of ViT. ViT could also capture the high-level semantic information across different frequency bands by utilizing the self-attention mechanism in the Transformer Encoder of ViT. An attention-based fusion mechanism is proposed to find the most discriminant features, and fuse the features of two streams into one feature map. Finally, a fully-connected layer is used to predict the person's identity given the radar signal.

More specifically, the proposed ADS-ViT network can be formulated as a quadruplet $\mathcal{Q} = (\mathcal{V}_s, \mathcal{V}_c, \mathcal{F}, \mathcal{C})$, where $\mathcal{V}_s$ and $\mathcal{V}_c$ are the ViT networks for spectrogram and CVD respectively, as shown in Fig. 1(b), $\mathcal{F}$ is an attention-based fusion network and $\mathcal{C}$ is the classifier at the end. Denote the generated spectrogram and CVD as $\boldsymbol{I}_s$ and $\boldsymbol{I}_c$, respectively. The features $\boldsymbol{f}_s$ from the spectrogram stream and $\boldsymbol{f}_c$ from the CVD stream are generated as:

$$\boldsymbol{f}_s = \mathcal{V}_s(\boldsymbol{I}_s), \tag{1}$$

$$\boldsymbol{f}_c = \mathcal{V}_c(\boldsymbol{I}_c), \tag{2}$$

These two sets of features are fused by the attention-based fusion network $\mathcal{F}$ to generate the features $\boldsymbol{f}_F$ as:

$$\boldsymbol{f}_F = \mathcal{F}(\boldsymbol{f}_s, \boldsymbol{f}_c). \tag{3}$$

Finally, the label $\hat{l}$ is predicted by the classifier $\mathcal{C}$ as:

$$\hat{l} = \mathcal{C}(\boldsymbol{f}_F). \tag{4}$$

2.2. Time-frequency Representations

Spectrograms [8, 9] and CVDs [10] are jointly utilized for robust radar gait recognition in this paper.

Spectrogram: Tahmoush and Silvious [13] modeled Doppler of each body part of a walking human as sinusoidal modulation in the spectrogram. This model reveals the velocities of each body part and determines the necessary human gait characteristics. The extracted features are the mean Doppler velocity and the size of torsos in the spectrogram. The mDS of a person is found relatively consistent and different walking

people exhibit discriminative characteristics in the spectrogram [13]. The radar spectrogram can well manifest the time-varying characteristics of a person's gait. A sample spectrogram is shown in Fig. 1(a).

Cadence velocity diagram: After deriving the spectrogram, the CVD [10] is obtained by taking the Fourier transform of the spectrogram along the time axis. The derived CVD, as shown in Fig. 1(a), is a matrix with rows representing Doppler frequencies and columns representing cadence frequencies, which measures how frequently different frequencies appear in the signal over the observation duration. It encodes useful information of the repetition of velocities, which would be the key feature to identify a walking person.

2.3. Vision Transformer for Spectrogram and CVD

Two network branches based on the Vision Transformer are designed to extract features from spectrograms and CVDs, respectively. The ViT network consists of two parts: Position and Patch Embedding (PPE) and Transformer Encoder (TE), as shown in Fig. 1(b). The original size of spectrogram/CVD is 115×115. To fit the pretrained ViT network, the input spectrogram/CVD is converted to a 3-channel image $\boldsymbol{I} \in \mathcal{R}^{224\times224\times3}$, by replicating the resized grayscale image for three channels. In the PPE stage, $\boldsymbol{I}$ is convolved into the feature map of size $14 \times 14 \times 768$ via a 16×16 kernel with a stride of 16, and flattened to the features $\boldsymbol{f}_I \in \mathcal{R}^{196\times768}$. Similarly as in [19], the learnable classification token $\boldsymbol{t} \in \mathcal{R}^{1\times768}$ is prepended to $\boldsymbol{f}_I$ as:

$$\boldsymbol{f}'_I = [\boldsymbol{t}, \boldsymbol{f}_I]. \tag{5}$$

The position information is embedded by adding a learnable position matrix $\boldsymbol{P} \in \mathcal{R}^{197\times768}$ for retaining positional knowledge of each patch, similarly as in [19], which results in the feature map $\boldsymbol{f}_P \in \mathcal{R}^{197\times768}$ with position embedding to the Transformer Encoder as:

$$\boldsymbol{f}_P = \boldsymbol{P} + \boldsymbol{f}'_I. \tag{6}$$

The TE module consists of 12 repeating blocks, where each contains a multi-head self-attention layer. This module could learn the global relations between patches. The learned knowledge is represented in the classification tokens for all patches and used as the output of the ViT network.

Unlike real-world images in which an object could be positioned anywhere in an image but interpreted the same, patches of spectrogram or CVD generated from the radar signal have unique physical meanings when positioned at different locations in the image. Each patch in spectrogram or CVD contains information in different frequency bands at different time instances (or different cadence velocities for CVD). Even patches similar in appearance could be physically different based on their frequency bands and temporal positions.

The commonly-used convolutional neural networks such as AlexNet [18], VGG [4] and ResNet [4] often use the same convolutional kernel across the whole image, while the proposed ADS-ViT treats patches differently, *i.e.*, it splits the spectrogram/CVD into patches and embeds their positions so that the information embedded in different frequency bands could be properly aligned and effectively extracted. This partially justifies the superior performance of the proposed ADS-ViT over other models. The multi-head self-attention mechanism in the Transformer Encoder could further extract the reserved discriminant information from each patch and process this information globally to encode the global relations among patches.

2.4. Attention-based Fusion

Inspired by Chen *et al.* [20], an attention-based fusion architecture shown in Fig. 1(a) is developed to fuse the complementary features extracted from both spectrogram and CVD. The target of the attention-based feature-level fusion is to find a set of weights $\{\boldsymbol{w}_i \in \mathcal{R}^{1\times768}\}$ for the features $\{\boldsymbol{f}_i \in \mathcal{R}^{1\times768}\}$ to obtain an aggregated feature $\boldsymbol{f}_a \in \mathcal{R}^{1\times768}$:

$$\boldsymbol{f}_a = \sum_{i=1}^{n} \boldsymbol{w}_i \odot \boldsymbol{f}_i, \tag{7}$$

where $\odot$ denotes element-wise multiplication, n is the number of features and $n = 2$ in this paper. In the attention-based fusion model, the kernel $\boldsymbol{q} \in \mathcal{R}^{1\times768}$ is required to be trained. The process begins with the element-wise multiplication between the feature vector $\boldsymbol{f}_i$ and the kernel $\boldsymbol{q}$:

$$\boldsymbol{s}_i = \boldsymbol{q} \odot \boldsymbol{f}_i, \tag{8}$$

where $\boldsymbol{s}_i \in \mathcal{R}^{1\times768}$ are the confidence scores for feature representations. A softmax function is then applied to $\boldsymbol{s}_i$ to assure that the derived weights $\sum_i \boldsymbol{w}_i = \boldsymbol{1} \in \mathcal{R}^{1\times768}$:

$$\boldsymbol{w}_i = e^{\boldsymbol{s}_i} \oslash \sum_j e^{\boldsymbol{s}_j}, \tag{9}$$

where $\oslash$ denotes element-wise division. The fused feature vector is finally generated using Eq. (7). This attention-based fusion could well highlight the most discriminant features by training the kernel function $\boldsymbol{q}$.

3. EXPERIMENTAL RESULTS

3.1. Dataset

There is no publicly available dataset for radar gait recognition, and hence a K-MC1 radar transceiver and an ST200 evaluation system launched by RFbeam Microwave GmbH are employed to collect a radar gait dataset by the authors. There are two data collection sessions at least two weeks apart. In each session, each volunteer walks 10 sequences. (Some only participate in the first session.) The dataset consists of 1670 walking sequences from 98 volunteers. For each volunteer,

50% of sequences are randomly selected as the training set and the rest are used for testing. In each sequence, a volunteer walks away from the radar along a corridor about 40 meters long, turns around and walks back towards the radar, lasting about 30 seconds. The sampling rate of the original signal is 125k. After a decimation of 64, the sampling rate is about 1.95k. The spectrogram is built using 128 sample points with an overlapping ratio of 90%. After removing the cluster and non-informative high-frequency components, the spectrum has 115 data points. To enrich the dataset, each sequence is cut into multiple frames of 115 data points, with a stride of 10. As a result, a total number of 45,768 frames of size 115×115 are generated, in which 22,894 are used for training and 22,874 for testing.

3.2. Experimental Settings

AlexNet has been used as the backbone of the Deep Convolutional Neural Network on spectrograms of the radar signal to identify people [18]. In [4], VGG16 [21] and ResNet18 [22] have been used for radar gait recognition. These three approaches are state-of-the-art methods for radar target recognition. They are implemented and evaluated on our benchmark dataset. As the proposed ADS-ViT utilizes both spectrogram and CVD, AlexNet [18], VGG16 [4] and ResNet18 [4] are applied on the CVD for comparison as well.

The size of frames is 115 initially and resized to 224 to fit the networks. Stochastic Gradient Descent is used as the optimizer of all models with a momentum of 0.9 and a weight decay of 5×10^{-5}. The cross-entropy loss function is used. For learning parameters, a linear learning rate warmup and decay strategy is used with a learning rate scheduler, and the initial learning rate is set to 0.01 for ResNet, 0.001 for AlexNet and 0.0001 for VGG following the parameter settings in [4] and [18] for a fair comparison. The batch size is 128 for all models. The proposed ADS-ViT initialize ViT networks with pretrained weights for both spectrogram and CVD streams. Models are trained for 500 epochs to guarantee convergence, and the optimal performance on the test set is reported.

3.3. Comparisons to State-of-the-art Approaches

The comparisons to state-of-the-art models are summarized in Table 1. Most existing models are applied on spectrograms, *e.g.*, AlexNet [18], VGG16 [4] and ResNet18 [4]. As the proposed method utilizes both spectrogram and CVD, these three CNN models are applied on the CVD as well. In addition, the proposed attention-based dual-stream architecture is also applied on ResNet18 to verify its effectiveness.

Table 1 shows that the proposed ADS-ViT significantly outperforms all compared methods. The proposed attention-based two-stream architecture significantly improves the classification accuracy from 85.56% to 87.08% when applied on the ResNet18. The proposed ADS-ViT further improves the classification accuracy to 91.02%. The achieved significant performance gain is attributed to the proposed ADS-ViT in two folds. First, by dividing the spectrogram/CVD into patches, the proposed model extracts the local features from different frequency bands using the Position and Patch Embedding mechanism. Secondly, the proposed model extracts the global features across frequency bands using the self-attention mechanism embedded in the transformer encoder.

Table 1. Comparisons to state-of-the-art AlexNet [18], VGG16 [4] and ResNet18 [4]. These models are implemented and evaluated on both spectrograms and CVDs on our benchmark dataset for comparison.

Method	Accuracy
DCNN-AlexNet on Spectrogram [18]	71.56%
VGG16 on Spectrogram [4]	69.24%
ResNet18 on Spectrogram [4]	85.56%
DCNN-AlexNet on CVD	72.44%
VGG16 on CVD	79.83%
ResNet18 on CVD	80.73%
Attention-Based Dual-Stream ResNet18	87.08%
Proposed ADS-ViT	**91.02%**

Second, while the spectrogram encodes the short-duration information using the short-time Fourier transform, the long-duration repeating frequency patterns are captured by the CVD. By utilizing both spectrogram and CVD, the short-duration repetitive patterns in the spectrogram and the long-duration repetitive frequency patterns in the CVD are captured by the two streams of the proposed ADS-ViT, respectively. The proposed attention-based fusion scheme well integrates the discriminant information extracted from spectrogram and CVD, which leads to the superior performance of the proposed ADS-ViT.

4. CONCLUSION

The proposed Attention-based Dual-Stream ViT well solves the problem of radar gait recognition. The radar signal can be represented as a spectrogram or a cadence velocity diagram. The complementary nature of these two representations motivates us to utilize both representations for better classification performance. We propose to use the Vision Transformer to split the spectrogram and CVD into patches and embed their positions so that the information embedded in different frequency bands could be effectively extracted and properly aligned. This enables the model to exploit the deep physical knowledge of the radar signal. The proposed attention-based fusion scheme integrates the discriminant features from the two representations. The proposed model is compared with state-of-the-art models on our benchmark dataset. The experimental results demonstrate that the proposed model significantly outperforms all compared methods.

5. REFERENCES

[1] Zhenyuan Zhang, Zengshan Tian, and Mu Zhou, "Latern: Dynamic continuous hand gesture recognition using FMCW radar sensor," *IEEE Sensors J.*, vol. 18, no. 8, pp. 3278–3289, 2018.

[2] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognit.*, vol. 98, pp. 107069, 2020.

[3] Cheng Zhang, Wu Liu, Huadong Ma, and Huiyuan Fu, "Siamese neural network based gait recognition for human identification," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.* IEEE, 2016, pp. 2832–2836.

[4] Pia Addabbo, Mario Luca Bernardi, Filippo Biondi, Marta Cimitile, Carmine Clemente, and Danilo Orlando, "Temporal convolutional neural networks for radar micro-Doppler based gait recognition," *Sensors*, vol. 21, no. 2, pp. 381, 2021.

[5] Xueru Bai, Ye Hui, Li Wang, and Feng Zhou, "Radar-based human gait recognition using dual-channel deep convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9767–9778, 2019.

[6] Hoang Thanh Le, Son Lam Phung, Abdesselam Bouzerdoum, and Fok Hing Chi Tivive, "Human motion classification with micro-Doppler radar and Bayesian-optimized convolutional neural networks," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.* IEEE, 2018, pp. 2961–2965.

[7] Jianfeng Ren, Xudong Jiang, Junsong Yuan, and Nadia Magnenat-Thalmann, "Sound-event classification using robust texture features for robot hearing," *IEEE Trans Multimedia*, vol. 19, no. 3, pp. 447–458, 2017.

[8] Jianfeng Ren and Xudong Jiang, "Regularized 2-D complex-log spectral analysis and subspace reliability analysis of micro-Doppler signature for UAV detection," *Pattern Recognit.*, vol. 69, pp. 225–237, 2017.

[9] Jianfeng Ren and Xudong Jiang, "A three-step classification framework to handle complex data distribution for radar UAV detection," *Pattern Recognit.*, vol. 111, pp. 107709, 2021.

[10] Svante Björklund, Tommy Johansson, and Henrik Petersson, "Evaluation of a micro-Doppler classification method on mm-wave data," in *Proc. 2012 IEEE Radar Conf.* IEEE, 2012, pp. 0934–0939.

[11] RIA Harmanny, JJM De Wit, and G Prémel Cabic, "Radar micro-Doppler feature extraction using the spectrogram and the cepstrogram," in *Proc. 2014 Eur. Radar Conf.* IEEE, 2014, pp. 165–168.

[12] Youngwook Kim and Taesup Moon, "Human detection and activity classification based on micro-Doppler signatures using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 8–12, 2015.

[13] Ann-Kathrin Seifert, Moeness G Amin, and Abdelhak M Zoubir, "New analysis of radar micro-Doppler gait signatures for rehabilitation and assisted living," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.* IEEE, 2017, pp. 4004–4008.

[14] Francesco Fioranelli, Matthew Ritchie, and Hugh Griffiths, "Centroid features for classification of armed/unarmed multiple personnel using multistatic human micro-Doppler," *IET Radar Sonar Navig.*, vol. 10, no. 9, pp. 1702–1710, 2016.

[15] Shihe Wang, Jianfeng Ren, and Ruibin Bai, "A regularized attribute weighting framework for naive bayes," *IEEE Access*, vol. 8, pp. 225639–225649, 2020.

[16] Rezaul K Begg, Marimuthu Palaniswami, and Brendan Owen, "Support vector machines for automated gait classification," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 5, pp. 828–838, 2005.

[17] WL van Rossum, L Anitori, P van Dorp, JJM de Wit, and RIA Harmanny, "Classification of human gaits using interrupted radar measurements," in *Proc. 2017 IEEE Radar Conf.* IEEE, 2017, pp. 0514–0519.

[18] Peibei Cao, Weijie Xia, Ming Ye, Jutong Zhang, and Jianjiang Zhou, "Radar-ID: human identification based on radar micro-Doppler signatures using deep convolutional neural networks," *IET Radar Sonar Navig.*, vol. 12, no. 7, pp. 729–734, 2018.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. 8th Int. Conf. Learn. Represent.*, 2020.

[20] Haonan Chen, Guosheng Hu, Zhen Lei, Yaowu Chen, Neil M Robertson, and Stan Z Li, "Attention-based two-stream convolutional networks for face spoofing detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 578–593, 2019.

[21] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.