

Keyframe and GAN-Based Data Augmentation for Face Anti-Spoofing

Jarred Orfao^a and Dustin van der Haar^b

Academy of Computer Science and Software Engineering, University of Johannesburg,
Kingsway Avenue and University Rd, Auckland Park, South Africa

Keywords: Face Anti-Spoofing, Generative Data Augmentation, Keyframe Selection.

Abstract: As technology improves, criminals, find new ways to gain unauthorised access. Accordingly, face spoofing has become more prevalent in face recognition systems, requiring adequate presentation attack detection. Traditional face anti-spoofing methods used human-engineered features, and due to their limited representation capacity, these features created a gap which deep learning has filled in recent years. However, these deep learning methods still need further improvements, especially in the wild settings. In this work, we use generative models as a data augmentation strategy to improve the face anti-spoofing performance of a vision transformer. Moreover, we propose an unsupervised keyframe selection process to generate better candidate samples for more efficient training. Experiments show that our augmentation approaches improve the baseline performance of the CASIA-FASD and achieve state-of-the-art performance on the Spoof in the Wild database for protocols 2 and 3.

1 INTRODUCTION

Face recognition is a physical biometric modality that has historically struggled with user acceptance due to its non-contact nature (Jain et al., 2007). In recent years, developments in technology and COVID-19 protocols have enabled facial recognition to become a more common method of user authentication in public and private settings (Bischoff, 2021), including workplaces, trains, and airports, using devices such as computers and mobile phones. Although facial recognition has taken massive strides from its inception, each component has inherent vulnerabilities. The biometric sensor is the first component of an authentication system that a user interacts with, making it the easiest to access and least expensive to attack. Unlike other components, the system has no control over the input that the sensor is exposed to, thereby making it susceptible to presentation attacks, such as face spoofing.


A face spoofing attack is when an attacker presents a two-dimensional medium, such as a photo or video, or a three-dimensional medium, such as a mask, of an enrolled user (commonly referred to as a *victim*), to the biometric sensor to gain illegal access (Daniel and Anitha, 2018). As an authenticated face is often the only hurdle to accessing physical and


digital assets, it is imperative that face spoofing is detected.

In this study, we use deep learning (a Vision Transformer) to detect two-dimensional face spoofing attacks and generative models to enhance the results further. To our knowledge, this paper is the first work using Generative Adversarial Networks (GANs) as a data augmentation strategy for face anti-spoofing. We summarise the main contributions of this paper as follows:

1. We show the effectiveness of GANs as a data augmentation strategy for face anti-spoofing compared to traditional augmentation approaches.
2. We propose an unsupervised keyframe selection process for more effective candidate generation and investigate the relationship between variability and image fidelity and its role in artefact detection.
3. We explore when data augmentation should be performed, the optimal data augmentation percentage and the number of frames to consider for face anti-spoofing.
4. We benchmark our approach against the current state-of-the-art face anti-spoofing approaches on public datasets.

The rest of the paper is structured accordingly: We begin with a discussion of face anti-spoofing and similar work in Section 2, followed by an explanation of

^a  <https://orcid.org/0000-0003-1430-0488>

^b  <https://orcid.org/0000-0002-5632-1220>

the proposed method in Section 3. In Section 4, we describe the experiment setup, followed by the analysis of the results in Section 5 and a conclusion in Section 6.

2 RELATED WORK

Face anti-spoofing has been an active research topic for more than 15 years, with publications increasing yearly (Yu et al., 2021). Despite researchers' efforts, face recognition systems are still vulnerable to simple, non-intrusive attack vectors. These attack vectors prey on the biometric sensor and are categorised accordingly (Hernandez-Ortega et al., 2021):

1. A *photo attack* is when an attacker presents a printed image of a *victim* to the biometric sensor.
2. A *warped-photo attack* is an extension of a photo attack, implemented by manipulating the printed image to simulate facial motion.
3. A *cut-photo attack* is an extension of a photo attack, implemented by blinking behind eye holes removed from the printed image.
4. A *video replay attack* is when an attacker uses a device to replay a video containing a *victim's* face to the biometric sensor.
5. A *3D-mask attack* is when an attacker wears a 3D mask, replicating a *victim's* facial features, in front of the biometric sensor.
6. A *DeepFake attack* occurs when an attacker uses deep learning methods to replace a person's face in a video with a victim's.

Thankfully, face anti-spoofing methods have advanced significantly in recent years. Traditionally, researchers achieved face anti-spoofing by using human vitality cues and handcrafted features. The vitality cues include eye blink detection (Li, 2008), face movement analysis (Bao et al., 2009) and gaze tracking (Ali et al., 2013). However, these approaches are susceptible to cut-photo and video-replay attacks, making them unreliable. In contrast, handcrafted features such as Local Binary Patterns (LBP) (Boulkenafet et al., 2016), Speeded-Up Robust Features (SURF) (Boulkenafet et al., 2017) and Shearlets (van der Haar, 2019) have extracted effective spoof patterns in real-time with minimal resources. However, handcrafted features require a hands-on approach from feature engineers to select the essential features from images, which becomes more difficult as the number of classification classes increase (Walsh et al., 2019). Moreover, each feature

requires handling multiple parameters, which all need fine-tuning.

In contrast, deep learning approaches discover descriptive patterns independently with minimal human intervention. Convolutional Neural Network (CNN) architectures have been successful with face anti-spoofing but require a large amount of data to train models sufficiently and are prone to overfitting. To avoid overfitting a dataset during training, researchers use regularisation methods, such as dropout, especially when training a model with no prior knowledge (Ur Rehman et al., 2017). Another strategy that has achieved success is using a pre-trained model and fine-tuning selected layers (Nagpal and Dubey, 2019), (George and Marcel, 2021). This approach allows a model to apply the features learned from a large dataset to a similar task with a smaller dataset. Some researchers have been successful in training CNNs with auxiliary information. Using this approach, (Liu et al., 2018) achieved competitive results by fusing the depth map of the last frame with the corresponding remote photoplethysmography signal acquired over a sequence of frames to determine the final spoof score. However, their approach requires multiple frames, which limits its applicability.

Recently, researchers have achieved state-of-the-art performance by using a hybrid approach of handcrafted features with deep learning. Wu et al. (2021) created a DropBlock layer, which randomly discards a part of the feature map to learn location-independent features. Furthermore, their method acts as a data augmentation strategy because the blocked regions can represent occlusions, thus increasing the training samples and reducing the risk of overfitting. Similarly, inspired by how LBPs describe local relations, Yu et al. (2020) created a Central Difference Convolution layer. This layer has the same sampling step as a traditional convolutional layer but prefers to aggregate the centre-oriented gradient of the sampled values during the aggregation step. In doing so, they obtain intensity-level semantic information and gradient-level detailed messages, which they prove are essential for face anti-spoofing.

The above analysis shows that hybrid approaches reap the benefits of traditional and deep learning methods. Furthermore, approaches such as Wu et al. (2021) also act as a data augmentation strategy. To fairly evaluate the effectiveness of our data augmentation strategy, we will fine-tune a vision transformer similar to George and Marcel (2021), which we will discuss in the next section.

3 PROPOSED METHOD

This paper proposes a transfer learning approach to face anti-spoofing using a pre-trained Vision Transformer (ViT) model. Furthermore, we use generative data augmentation to optimise this model by synthesising candidate samples using StyleGAN3 models. In the following sections, we will discuss the different stages of the training pipeline, illustrated in Figure 1.

3.1 Preprocessing

We preprocessed each video to avoid background and dataset bias. First, we employed the MTCNN algorithm (Zhang et al., 2016) for face detection. Next, we rotated the detected region (to align the eye centres horizontally), scaled the region (to minimise the background), and cropped the region to produce a square patch containing the subject’s eyebrows and mouth. Although it is possible to scale the detected region to remove the background altogether, the crop patch does not have eyebrows or a mouth. Since these facial features are essential for portraying emotion, we decided to include slightly more background to ensure they are both present in the cropped patch.

3.2 Data Augmentation

We followed a generative data augmentation approach by synthesising new training images rather than applying an affine transformation to existing images. We trained a StyleGAN3 model for each attack vector and used these models to generate new candidate samples. We also performed data augmentation using traditional methods to compare it against our generative approach.

$$N_I = \left(\frac{N_S \times P}{100\% - P} \right) \div N_A \quad (1)$$

In equation (1), N_I is the number of images generated for each attack vector; N_S is the number of samples present in the training protocol; N_A is the number of attack vectors present in the training protocol, and P is the desired data augmentation percentage. Using equation 1, we calculated the number of images necessary to achieve the desired data augmentation percentage for each attack vector.

3.3 Model Training

Vision transformers have received much attention in recent years (Han et al., 2022). Initially, researchers used transformers for natural language pro-

cessing (Vaswani et al., 2017), but their success in this field caught the attention of computer vision researchers. Kolesnikov et al. (2021) harnessed the power of transformers for computer vision tasks by making minor alterations. Instead of providing tokens to a transformer as input, they divided an image into patches and used the patch embeddings. In doing so, they created what is now known as a Vision Transformer (ViT). Vision transformers have achieved remarkable results in image classification tasks (Krishnan and Krishnan, 2021).

We followed a similar approach to George and Marcel (2021), who achieved state-of-the-art results in face anti-spoofing using their ViT. Moreover, we regard face anti-spoofing as a binary classification problem in which a sample is either *bona fide* or a *spoof*. For clarity, we define a *bona fide sample* as a genuine sample directly acquired from an individual. Furthermore, we define a *spoof sample* as a fabricated sample of an individual captured indirectly from a presented medium. Lastly, we employ a static face anti-spoofing approach by analysing each frame independently.

4 EXPERIMENT SETUP

4.1 Datasets

There are a variety of publicly available datasets for face anti-spoofing. To evaluate the effectiveness of our approach, we used the CASIA Face Anti-Spoofing Database (CASIA-FASD) and Spoof in the Wild (SiW) Database.

CASIA-FASD (Zhang et al., 2012). In 2012, fifty subjects participated in producing 600 videos captured in natural scenes with no artificial environment unification. These researchers recorded each subject normally (N) and created cut-photo (C), warped-photo (W) and video-replay (R) attacks using a low (1), medium (2) and high-resolution (HR) camera. For clarity, we denote A as the attack vector and B as the resolution in $A.B$. For example, W_{HR} corresponds to a warped-photo attack sample captured with a high-resolution camera.

This dataset contains seven protocols for training and evaluating a model’s performance. Protocols 1 to 3 correspond to training and testing using only low, medium, and high-resolution videos. Similarly, protocols 4 to 6 correspond to training and testing using only warped, cut-photo and video-replay attack videos. Lastly, protocol 7 uses all the videos for training and testing.

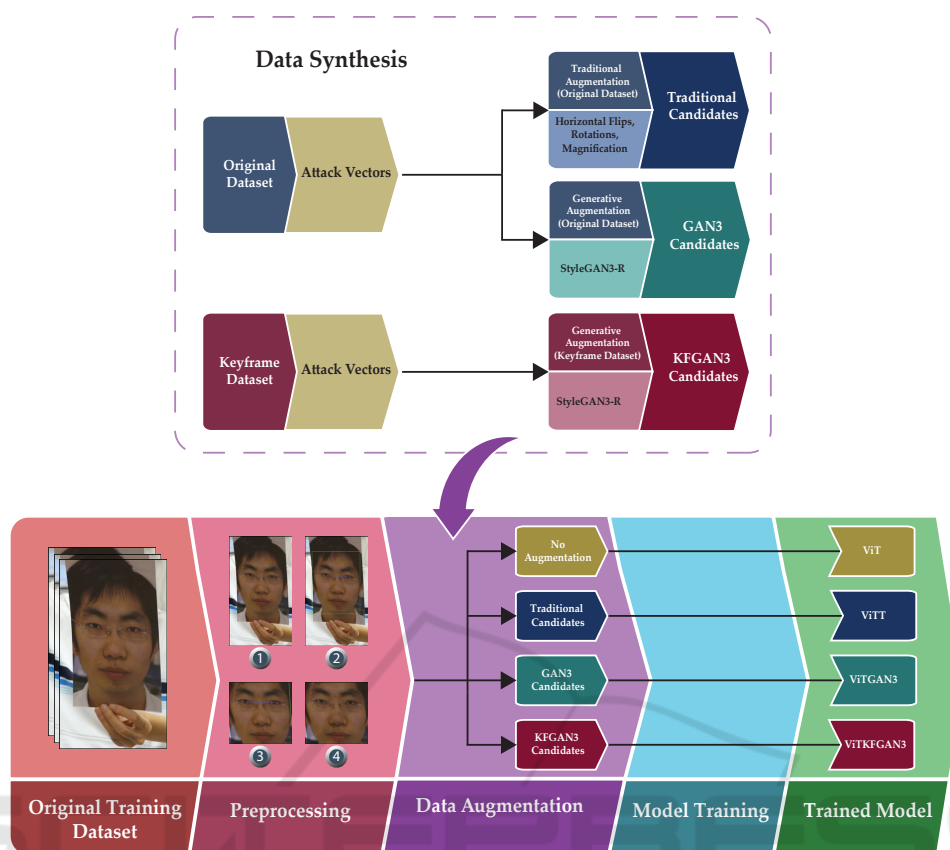


Figure 1: The stages of the proposed training pipeline.

SiW (Liu et al., 2018). In 2018, 165 subjects participated in producing 4 478 videos, each with a different distance, pose, illumination and expression. This dataset contains high-resolution normal (N), photo (P) and video-replay (R) attack videos. The researchers captured the normal videos using two lighting variations: no lighting variation (NLV) and different lighting variation (ELV). To create the photo attacks, they captured low-resolution (LR) and high-resolution (HR) images of each subject, which they then printed on glossy and matte paper. Lastly, the video-replay attacks utilised a Samsung Galaxy S8 (SGS8), an iPhone 7 Plus (IP7P), iPad Pro 2017 (IPP2017) and an ASUS MB168B laptop screen (ASUS) to display the bona fide videos.

This dataset contains three protocols to evaluate a model’s performance. Protocol 1 evaluates the generalisation capabilities by only training on the first 60 frames of the training set videos with mainly frontal view faces and testing on all the frames of the test set videos. Since we cannot guarantee that a generated image will be within the first 60 frames, we will not perform this protocol. Protocol 2 evaluates the generalisation capability on cross-mediums of the

same spoof type. This protocol follows a leave-one-out (LOO) strategy, repeated for all mediums: using three replay-attack mediums for training and leaving the remaining medium for testing. Lastly, protocol 3 evaluates a model’s performance on unknown presentation attacks. This protocol is similar to protocol 2, using attack vectors rather than spoof mediums. For clarity in later sections, we denote LOO_X as the group left out for training and used exclusively for testing, where X is a video-replay spoof medium (ASUS, IP7P, IPP2017 or SGS8) in protocol two and an attack vector (P or R) in protocol three.

Since protocols 2 and 3 utilise various training combinations, it is essential to use the mean and standard deviation of the combinations when reporting metrics. For comparability with other work, we used the videos of subjects 90 for training and 75 for testing. Figures 2 and 3 display a sample frame from each video captured for subjects 75 and 90, respectively. We display the sample type (*ST*) and medium (*M*) for each sample in both figures.

Although CASIA-FASD is an older and smaller dataset compared to other face anti-spoofing datasets, we selected it because it contains cut-photo attacks

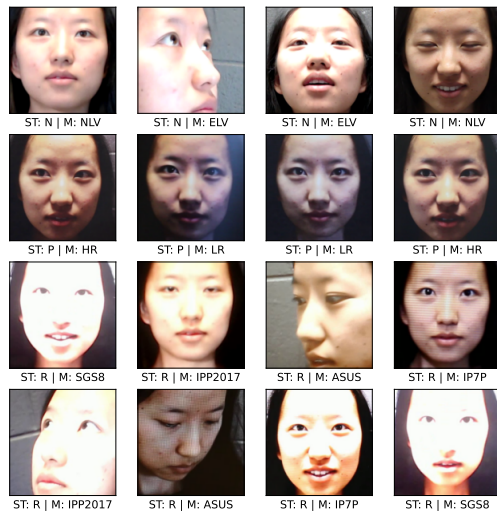


Figure 2: The middle frame of each video for subject 75 from the SiW database, where ST and M correspond to the sample type and medium.

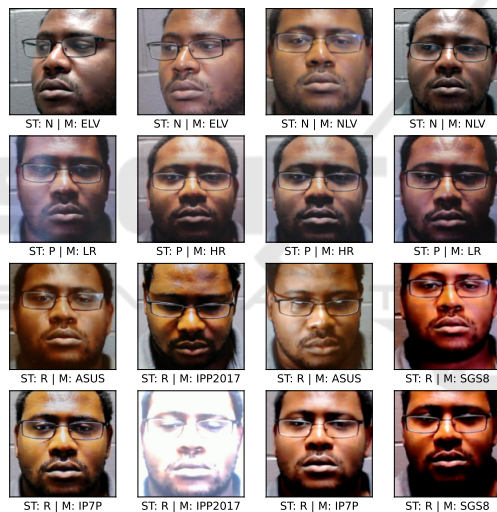


Figure 3: The middle frame of each video for subject 90 from the SiW database, where ST and M correspond to the sample type and medium.

and low-resolution videos. Furthermore, we selected SiW for its variation in the subjects' distance, pose, illumination and expression.

4.2 Dataset Augmentation

A traditional approach to data augmentation involves applying random transformations to an image, such as varying brightness, rotating, or cropping (Pérez-Cabo et al., 2019). However, some of these transformations could adversely affect the sample's label due to the nature of the problem environment (Shorten and Khoshgoftaar, 2019). For instance, in some

video-replay attack samples, the LCD screen backlight makes them appear brighter than the corresponding bona fide sample, as shown in Figures 2 and 3. Thus, altering the brightness of the spoof samples may lead to an unclear decision boundary due to label contradictions. Hence, when using the traditional data augmentation approach, we only use random horizontal flips, rotations (within 15°), and magnifications (within 20%).

In contrast to traditional data augmentation, GANs can create new samples that match the characteristics of an image domain while maintaining the label given to the original samples. We selected the StyleGAN3 architecture for the generative data augmentation approach. StyleGAN3 improved the image synthesis by linking details to depicted object surfaces rather than absolute coordinates (Karras et al., 2021). Additionally, it inherits the enhanced training stabilisation from StyleGAN2 (Karras et al., 2020), enabling it to achieve good results on smaller datasets. Thus, StyleGAN3 can synthesise high-fidelity images with limited data, making it state-of-the-art in image generation.

We chose the alias-free rotation equivariant architecture (StyleGAN3-R) and trained each model using the associated training configuration for 2 GPUs with a 256 by 256-pixel resolution for 5000 *king*, where *king* is the number of thousand images from the training set. We selected the model with the lowest Fréchet Inception Distance (FID) (Heusel et al., 2017) to generate the candidate samples. We used ordered seeds from 1 to N_f and a truncation ψ value of 1 (for maximum variation).

For clarification, we synthesised images for each attack vector separately, with each image labelled as a spoof. Although it is possible to generate images using bona fide samples, labelling them as such could introduce a DeepFake attack vulnerability. Figure 4 illustrates the synthesised images for each SiW attack vector using each data augmentation approach.

4.3 Face Anti-Spoofing

Training a ViT from beginning to end is very resource-intensive, requiring adequate hardware and a large dataset. Fortunately, researchers developed a technique to alleviate this burden known as transfer learning (Li and Lee, 2021). Transfer learning enables us to use the knowledge learned by a model for one task and apply it to another. Following (George and Marcel, 2021), we used a pre-trained ImageNet ViT-B/32 model and fine-tuned the output to meet our needs. We resized the input images to 224 by 224 pixels and replaced the last layer with a two-

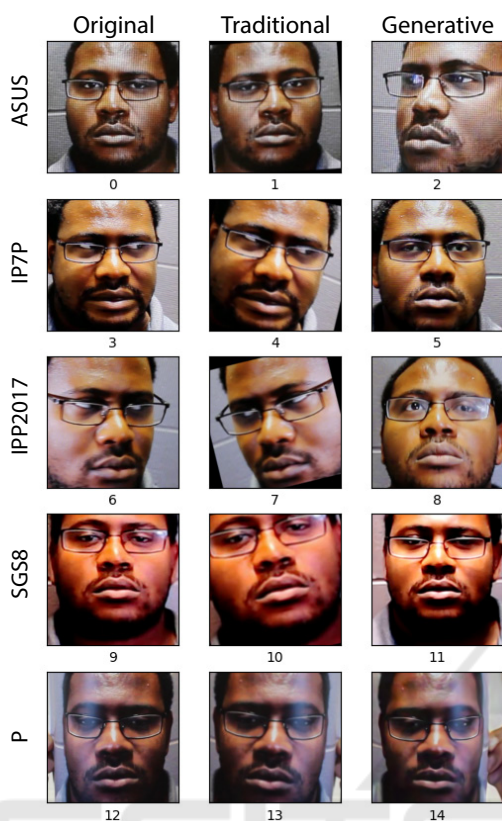


Figure 4: The synthesised images for each SiW attack vector. The columns and rows correspond to the sample sets and attack vectors.

node dense layer with a SoftMax activation. We froze all the layers before the final layer and trained the model using binary cross-entropy loss, optimised with Adam (Kingma and Ba, 2014) at a learning rate of $1e-4$. Using early stopping (with a patience value of 15), we trained each model for 70 epochs and restored the version with the lowest validation loss before testing.

We determined the optimal data augmentation percentage by performing a hyperparameter search (Liaw et al., 2018) for the following augmentation percentages: 5, 10, 20, and 30. Using a stratified 3-fold cross-validation (Rodríguez and Lozano, 2007) approach, we split the training dataset into 80% for training and 20% for validation, each with a batch size of 32. We repeated each trial twice and averaged the following ISO/IEC₃₀₁₀₇₋₃ (ISO/IEC, 2017) metrics to compare our approach with similar work: Bonafide Presentation Classification Error Rate (BPCER), Attack Presentation Classification Error Rate (APCER) and Average Classification Error Rate (ACER). BPCER is the proportion of bona fide samples incorrectly classified as an attack. Similarly, APCER is the proportion of attack samples incor-

rectly classified as bona fide. Finally, ACER is the mean of APCER and BPCER. Additionally, we report the Equal Error Rate (ERR) for comparisons with older face anti-spoofing approaches, which in the context of biometric anti-spoofing, is the point where the APCER and BPCER are equal (Ben Mabrouk and Zagrouba, 2018).

4.4 Keyframe Selection

Hardly any movement in a one-second video can result in 24 near-duplicate frames. We investigated these near-duplicate frames' effect on the ViT training. To do this, we employed the following three-stage unsupervised keyframe selection process.

Stage 1: Feature Extraction. We extracted features from the preprocessed frames using a ResNet-50 backbone, pre-trained on the VGGFace2 dataset. This large facial recognition dataset contains 9131 subjects of various ages, ethnicities and professions in various poses and illumination (Cao et al., 2018), making it suitable for our task.

Stage 2: Feature Clustering. We clustered the extracted features using Lloyd's K-Means clustering algorithm. To do so, we used the Facebook AI Similarity Search (FAISS) (Johnson et al., 2019) library, which harnesses the power of GPUs and is currently the fastest implementation of Lloyd's K-Means clustering algorithm. We conducted a hyperparameter search to find the K that maximises the Silhouette Score (Shahapure and Nicholas, 2020) to determine the optimal K.

Stage 3: Keyframe Selection. For each category, we calculated the mean optimal K. For CASIA-FASD, we used the attack vectors as the categories; for Spoof in the Wild, we used the medium names associated with each session. We again clustered the extracted features; however, we used the corresponding categorical mean optimal K to obtain the cluster centroids. Finally, we used vector quantisation to determine the features closest to these centroids and selected the corresponding images as the keyframes. Figures 5 and 6 illustrate the final result of the keyframe selection process.

Table 1 shows the ablation study for selecting the optimal K for CASIA-FASD. Looking at this table, we can see that the standard deviation is greater than the mean, implying that the optimal Ks are moderately dispersed. The dispersion occurs between the 75th and 100th percentiles values. If we look at the

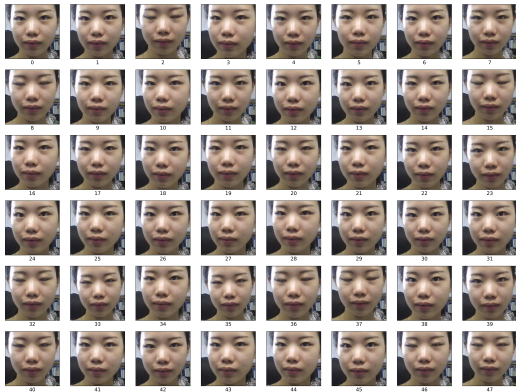


Figure 5: The original frames for subject 18, video 1 in the CASIA-FASD test release.

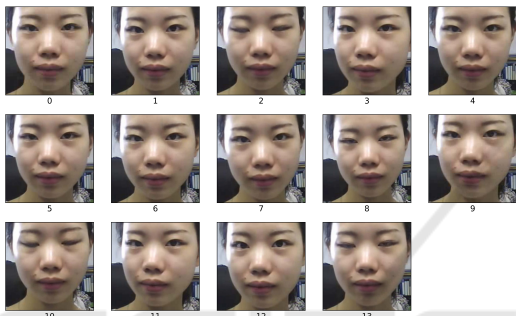


Figure 6: The keyframes for subject 18, video 1 in the CASIA-FASD test release.

75th percentile, the values are close to the mean except for the medium resolution warped-photo attack category (W_2). Thus, using the categorical-mean optimal K is better than the individual-video optimal K due to outlier videos. Figure 7 depicts the Keyframe reduction effect on CASIA-FASD. The unsupervised keyframe reduction process reduced the 110 882 frames in the original dataset to 7850 keyframes, resulting in a 1412% frame reduction.

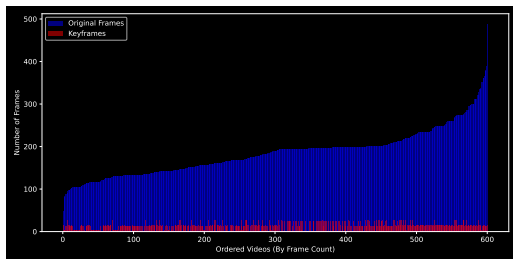


Figure 7: The number of original frames (blue) vs the number of keyframes (red) for each video in the CASIA-FASD ordered by video frame count.

To avoid confusion, we introduce notations to distinguish the Vision Transformer models using their corresponding data augmentation approach. We begin by representing the baseline model, trained on

Table 1: The mean, std. deviation and five-number-summary for the optimal K of each attack vector in CASIA-FASD.

Attack Vector	Mean	Five-Number-Summary				
		Min	25 th	50 th	75 th	Max
N_1	14.2 ± 15	2	2	5	23.3	42
N_2	9.6 ± 16.8	2	2	2.5	4.3	54
N_HR	12.3 ± 20.2	2	2	3	5.3	68
C_1	2.2 ± 0.2	2	2	2	2	3
C_2	2.2 ± 0.5	2	2	2	2	4
C_HR	3.1 ± 4.2	2	2	2	2	21
W_1	32.8 ± 21.4	2	20.3	31.5	49.5	73
W_2	17.4 ± 30.9	2	2	3	4.5	95
W_HR	20.7 ± 28.5	2	2.8	4.5	39.8	101
R_1	14.1 ± 20.8	2	2	2.5	18.3	63
R_2	10.8 ± 13	2	2	2	21.3	40
R_HR	18.2 ± 29.5	2	2	2.5	14.5	85

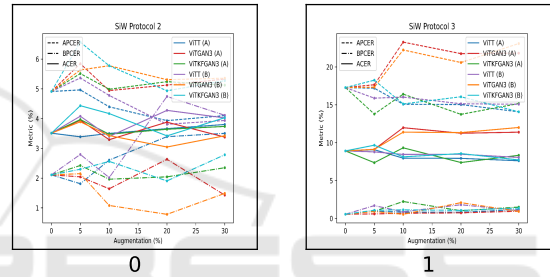


Figure 8: The average metrics of the traditional (ViTT) and generative (ViTGAN3 and ViTKFGAN3) data augmentation models performed before (B) and after (A) the validation split for SiW protocols 2 and 3, using data augmentation percentages: 5, 10, 20 and 30.

the original dataset with no data augmentation as ViT. Next, we represent a ViT model optimised using the traditional data augmentation approach as ViTT. Since our generative data augmentation approach uses StyleGAN3-R models, we represent a ViT model optimised with this approach as ViTGAN3. We introduce KFGAN3 as a keyframe generative data augmentation approach, in which we train the StyleGAN3 models using only the keyframes. Thus, we represent a ViT model optimised with the keyframe generative data augmentation approach as ViTKFGAN3.

5 RESULTS

We conducted an ablation study using SiW protocols 1 and 2 to determine when to perform data augmentation, the optimal data augmentation percentage, and the optimal number of frames to detect face spoofing.

Table 2: The performance of the baseline (ViT), traditional data augmentation (ViTT) and generative data augmentation (ViTGAN3, ViTKFGAN3) models for SiW protocols 2 and 3, in terms of ACER (%) for the image-based and video-based classification approaches, using a window size of 5, 7, 10, and 15.

Model	Aug. (%)	Protocol 2					Protocol 3				
		Image	5	7	10	15	Image	5	7	10	15
ViT (Baseline)	0	3.51	0	0	0	0	8.92	2.83	1.49	1.19	0
	5	4.08	0.78	0.26	0.26	0	8.79	2.08	2.08	1.19	0.3
ViTT (Before)	10	3.4	0	0	0	0	8.45	2.9	2.9	2.01	0.3
	20	4.28	0.26	0.26	0.26	0.26	8.48	2.98	2.68	1.49	0
	30	4.02	0.26	0.26	0	0	8.06	2.98	2.98	2.38	0.3
	5	3.39	0	0	0	0	9.08	3.13	2.53	1.93	1.34
ViTT (After)	10	3.5	0	0	0	0	7.96	1.79	1.79	1.49	1.04
	20	3.66	0	0	0	0	7.94	3.27	2.68	1.19	0.3
	30	3.8	0.26	0.26	0.26	0	7.62	2.68	2.68	1.79	0.3
ViTGAN3 (Before)	5	3.89	0	0	0	0	9.12	2.68	2.38	2.08	1.79
	10	3.43	0	0	0	0	11.4	5.21	5.21	4.02	4.32
	20	3.04	0.52	0.52	0	0	11.34	3.57	3.57	2.68	2.68
	30	3.42	0	0	0	0	12.01	5.51	5.51	4.61	5.21
ViTGAN3 (After)	5	3.94	1.04	1.04	1.04	0.52	9.13	2.38	2.08	1.79	0.89
	10	3.29	0.26	0.26	0.26	0.26	11.98	5.36	5.36	4.46	4.46
	20	3.88	0.78	0.26	0.52	0.26	11.25	5.95	5.65	4.76	3.87
	30	3.38	0	0	0	0	11.39	8.78	8.78	8.18	7.59
ViTKFGAN3 (Before)	5	4.43	1.3	1.3	0.78	0.52	9.67	3.87	3.87	3.27	2.68
	10	4.17	0.52	0.52	0.52	0.52	8.13	3.13	3.13	1.93	1.04
	20	3.42	0	0	0	0	8.57	2.08	2.83	2.83	2.23
	30	4.03	0	0	0	0	7.73	2.08	1.49	0.595	0
ViTKFGAN3 (After)	5	3.97	0.26	0.52	0	0	7.37	1.93	2.53	1.34	1.93
	10	3.47	0	0	0	0	9.33	2.9	2.31	2.01	2.23
	20	3.64	0	0	0	0	7.38	1.64	1.64	1.64	1.04
	30	3.74	0	0	0	0	8.34	1.79	1.19	0.595	0

Data Augmentation Before or After the Validation Split. As illustrated in Figure 8, the generative data augmentation approach (GAN3) performed better for both protocols and achieved its lowest ACER before the validation split. In contrast, the traditional data augmentation approach performed better and achieved its lowest ACER after the split. Interestingly, the keyframe generative data augmentation (KFGAN3) performed better before the split for protocol two and after the split for protocol 3. Tables 5 and 6 confirm these results on the CASIA-FASD, except for the ViTGAN3 ACER, which performed slightly better after the split.

The Optimal Data Augmentation Percentage (DAP). As shown in Figure 8, the GAN3 approach works best with a DAP of 20% and 0% for protocols 2 and 3, respectively. Similarly, the KFGAN3 approach works best with a DAP of 20% and 5% for protocols 2 and 3, respectively. Lastly, the traditional approach works best with a DAP of 5% and 30% for protocols 2 and 3, respectively. Tables 5 and 6 confirm that beyond 30%, we encounter diminishing returns.

Image-Based Classification Versus Video-Based Classification. One advantage of using a static analysis approach to face anti-spoofing is that it can be incorporated into image-based and video-based recognition systems. Accordingly, we investigated the effectiveness of our approach in each scenario by using image-based and video-based classification. We treated each video frame as a separate presentation attempt for image-based classification and classified them independently. For video-based classification, we treated each video as a separate presentation attempt by aggregating the predictions from the first 'n' frames and using the majority vote as the final label. When n is even, the vote favours the spoof label. The study results are shown in Table 2.

Starting with the image-based classification method, we found that the GAN3 approach performed the best in SiW protocol 2 but the worst in protocol 3. In contrast, the KFGAN3 approach performed the best in protocol 3 but the worst in protocol 2. Upon investigation, we observed that the FID of the KFGAN3 models was much larger than the GAN3 models, as shown in Table 4. Since FID is a measure of similarity between the generated and original samples, there is more variability in the training set using KFGAN3

Table 3: The performance of the models trained with the traditional (ViTT) and generative data augmentation (ViTGAN3 and ViTKFGAN3) approaches, compared to similar work models on protocols 2 and 3 of SiW. We denote ‘-B’ and ‘-A’ as our models trained with data augmentation before and after the validation split, respectively. As a reminder, we did not perform protocol one since we could not guarantee that the generated image would appear within the first 60 frames.

Model	Classification Approach	Metric (%)	Protocol	
			2	3
Auxiliary (Liu et al., 2018)	Video-based	APCER	0.57 ± 0.69	8.31 ± 3.81
		BPCER	0.57 ± 0.69	8.31 ± 3.80
		ACER	0.57 ± 0.69	8.31 ± 3.81
CDCN (Yu et al., 2020)	Image-based	APCER	0.00 ± 0.00	1.67 ± 0.11
		BPCER	0.13 ± 0.09	1.76 ± 0.12
		ACER	0.06 ± 0.04	1.71 ± 0.11
CDCN++ (Yu et al., 2020)	Image-based	APCER	0.00 ± 0.00	1.97 ± 0.33
		BPCER	0.09 ± 0.1	1.77 ± 0.1
		ACER	0.04 ± 0.05	1.90 ± 0.15
FAS-SGTD (Wang et al., 2020)	Video-based	APCER	0.00 ± 0.00	2.63 ± 3.72
		BPCER	0.04 ± 0.08	2.92 ± 3.42
		ACER	0.02 ± 0.04	2.78 ± 3.57
FasTCO (Xu et al., 2021)	Video-based	APCER	0.02 ± 0.02	2.73 ± 0.91
		BPCER	0.00 ± 0.00	1.28 ± 0.21
		ACER	0.01 ± 0.01	2.00 ± 0.56
PatchNet (Wang et al., 2022)	Image-based	APCER	0.00 ± 0.00	3.06 ± 1.10
		BPCER	0.00 ± 0.00	1.83 ± 0.83
		ACER	0.00 ± 0.00	2.45 ± 0.45
ViTGAN3-B (20%)	Image-based	APCER	5.31 ± 5.86	20.59 ± 18.63
		BPCER	0.78 ± 0.78	2.08 ± 4.49
		ACER	3.04 ± 2.91	11.34 ± 8.57
ViTKFGAN3-A (5%)	Image-based	APCER	5.51 ± 5.88	13.79 ± 11.62
		BPCER	2.42 ± 2.77	0.95 ± 0.84
		ACER	3.97 ± 2.62	7.37 ± 5.45
ViTKFGAN3-A (30%) Window size of 10	Video-based	APCER	0.00 ± 0.00	1.19 ± 2.78
		BPCER	0.00 ± 0.00	0.00 ± 0.00
		ACER	0.00 ± 0.00	0.595 ± 1.39
ViTKFGAN3-A (30%) Window size of 15	Video-based	APCER	0.00 ± 0.00	0.00 ± 0.00
		BPCER	0.00 ± 0.00	0.00 ± 0.00
		ACER	0.00 ± 0.00	0.00 ± 0.00

Table 4: The FID for the StyleGAN3 models trained using the original frames (GAN3) and keyframes (KFGAN3) for each attack vector in SiW.

Model	Attack Vector FID				
	ASUS	IP7P	IPP2017	SGS8	P
GAN3	25.33	18.18	34.76	35.14	26.47
KFGAN3	54.66	57.91	99.41	82.98	66.38

than GAN3-generated candidates. Furthermore, performing the data augmentation after the split can increase the variability of the training set. The high variability seems beneficial for unseen presentation attacks (protocol 3), and the low variability appears beneficial for unseen spoof mediums of the same type (protocol 2).

Regarding the optimal frame window, we found no need for data augmentation for protocol 2 when using frame sizes 5, 7, 10 and 15. Moreover, we found that the baseline, traditional, and KFGAN3 approaches achieved state-of-the-art performance across

both protocols using a frame size of 15. Despite this remarkable performance, we found that the KFGAN3 approach achieves comparable performance with less processing time, using a window size of 10. As the window size increases, the image-based classification trend emerges with reduced error values. As for the CASIA-FASD, we found that the KFGAN3 approach performed the best for image and video-based classification when using a window size of 7. Even though the other approaches achieved the same ACER and EER values using the same window size, the KFGAN3 approach achieved the lowest values simultaneously. Thus, the optimal window size for video-based classification lies between the first 7 (CASIA-FASD) and 15 frames (SiW). From this study, we suspect that the first few video frames can effectively reveal whether a presentation is genuine or a spoof.

Tables 3 and 7 benchmark our approach against similar research. Although our independent frame approach improved the baseline model’s performance, more was needed to compete with the state-of-the-

Table 5: The performance of the baseline (ViT), traditional data augmentation (ViTT) and generative data augmentation (ViTGAN3, ViTKFGAN3) models for CASIA-FASD protocol 7 in terms of EER (%) for the image-based and video-based (window size of 7) classification approaches.

Model	Aug. (%)	EER (%)	
		Image	Video
ViT (Baseline)	0	1.75	2.01
	5	1.96	2.18
ViTT (Before)	10	2	1.83
	20	2.12	2.36
	30	2.18	2.36
ViTT (After)	5	1.89	1.65
	10	1.96	1.82
	20	2.13	2
	30	2.21	1.82
ViTGAN3 (Before)	5	1.77	1.65
	10	1.72	1.47
	20	1.75	1.71
	30	1.81	1.29
ViTGAN3 (After)	5	1.72	1.53
	10	1.8	1.83
	20	1.83	1.65
	30	1.87	1.83
ViTKFGAN3 (Before)	5	1.71	1.83
	10	1.71	1.29
	20	1.73	1.47
	30	1.85	1.65
ViTKFGAN3 (After)	5	1.78	1.83
	10	1.82	1.83
	20	1.86	2.01
	30	1.9	2.01

art image-based classification methods. However, our video-based classification approach achieved top performance across both SiW protocols. Although the baseline and traditional approaches obtained the same results as the KFGAN3 approach using a window size of 15, we decided the KFGAN3 approach was the best when considering its performance on CASIA-FASD.

Upon investigating our suspicion of the first few frames' effectiveness, we found that (Xu et al., 2021) also encountered this phenomenon and hypothesised large motions and illumination changes to be the cause. Unlike (Xu et al., 2021), who developed a module to estimate uncertainty for a sequence of frames, we found improved performance using a majority vote on the first fifteen frames.

6 CONCLUSION

In this work, we optimised the performance of a vision transformer using generative and traditional data augmentation approaches. We trained a StyleGAN3 model for each attack vector and used these models to generate candidate samples. We extended this ap-

Table 6: The performance of the baseline (ViT), traditional data augmentation (ViTT) and generative data augmentation (ViTGAN3, ViTKFGAN3) models for CASIA-FASD protocol 7 in terms of ACER (%) for the image-based and video-based (window size of 7) classification approaches.

Model	Aug. (%)	ACER (%)	
		Image	Video
ViT (Baseline)	0	1.42	1.42
	5	1.39	1.45
ViTT (Before)	10	1.38	1.27
	20	1.38	1.39
	30	1.41	1.45
	5	1.37	1.11
ViTT (After)	10	1.39	1.17
	20	1.41	1.2
	30	1.42	1.2
	5	1.37	1.2
ViTGAN3 (Before)	10	1.36	1.11
	20	1.41	1.36
	30	1.42	1.14
	5	1.36	1.11
ViTGAN3 (After)	10	1.35	1.27
	20	1.35	1.23
	30	1.36	1.3
	5	1.42	1.3
ViTKFGAN3 (Before)	10	1.36	1.11
	20	1.34	1.14
	30	1.38	1.23
	5	1.36	1.27
ViTKFGAN3 (After)	10	1.36	1.2
	20	1.35	1.36
	30	1.34	1.3

Table 7: The performance of the models trained with the traditional (ViTT) and generative data augmentation (ViTGAN3 and ViTKFGAN3) compared to similar work models using CASIA-FASD protocol 7. We denote '-B' and '-A' as our models trained with data augmentation before and after the validation split. Moreover, 'I' and 'V' correspond to image-based and video-based classification approaches, with 'W-N' denoting the first N frames.

Model	Approach	EER (%)
DoG (Zhang et al., 2012)	I	17.00
LBP (Boulkenafet et al., 2016)	I	3.20
Deep LBP (Li et al., 2017a)	I	2.30
Hybrid CNN (Li et al., 2017b)	I	2.2
Attention CNN (Chen et al., 2020)	V	3.145
Dropblock (Wu et al., 2021)	I	1.12
ViTKFGAN3-B (10%)	I	1.71
ViTKFGAN3-B W-7 (10%)	V	1.29

proach by training the StyleGAN3 models using only keyframes rather than all the training samples. We implemented the traditional data augmentation approach using random rotations, horizontal flips and magnifications.

We found that the samples generated using keyframes increased the variability among the train-

ing and validation sets, whereas the samples generated using all the frames increased the similarity. We found high variability beneficial for unknown presentation attack detection and high similarity beneficial for unknown presentation attacks of the same kind.

We explored face anti-spoofing performance using image-based and video-based classification methods. We found the first few frames more effective for detecting face spoofing attacks than using each frame independently. The keyframe data augmentation approach using the first 15 frames achieved the top performance for Spoof in the Wild protocols 2 and 3.

The SiW and CASIA-FASD results proved keyframe data augmentation to be the most effective approach. Furthermore, we suspect augmenting training sets with generated spoof images can make deep learning models more robust against DeepFake attacks. We will investigate this in future work, along with more advanced GAN image generation techniques.

ACKNOWLEDGEMENTS

The support and resources from the South African Lengau cluster at the Centre for High-Performance Computing (CHPC) are gratefully acknowledged.

REFERENCES

- Ali, A., Deravi, F., and Hoque, S. (2013). Directional sensitivity of gaze-collinearity features in liveness detection. In *2013 Fourth International Conference on Emerging Security Technologies*, pages 8–11.
- Bao, W., Li, H., Li, N., and Jiang, W. (2009). A liveness detection method for face recognition based on optical flow field. In *2009 International Conference on Image Analysis and Signal Processing*, pages 233–236.
- Ben Mabrouk, A. and Zagrouba, E. (2018). Abnormal behavior recognition for intelligent video surveillance systems: A review. *Expert Systems with Applications*, 91:480–491.
- Bischoff, P. (2021). Facial recognition technology (frt): 100 countries analyzed - comparitech.
- Boulkenafet, Z., Komulainen, J., and Hadid, A. (2016). Face spoofing detection using colour texture analysis. *IEEE Transactions on Information Forensics and Security*, 11(8):1818–1830.
- Boulkenafet, Z., Komulainen, J., and Hadid, A. (2017). Face antispoofing using speeded-up robust features and fisher vector encoding. *IEEE Signal Processing Letters*, 24(2):141–145.
- Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74.
- Chen, H., Hu, G., Lei, Z., Chen, Y., Robertson, N. M., and Li, S. Z. (2020). Attention-based two-stream convolutional networks for face spoofing detection. *IEEE Transactions on Information Forensics and Security*, 15:578–593.
- Daniel, N. and Anitha, A. (2018). A study on recent trends in face spoofing detection techniques. In *2018 3rd International Conference on Inventive Computation Technologies (ICICT)*, pages 583–586.
- George, A. and Marcel, S. (2021). On the effectiveness of vision transformers for zero-shot face anti-spoofing. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8.
- Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., and Tao, D. (2022). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.
- Hernandez-Ortega, J., Fierrez, J., Morales, A., and Galbally, J. (2021). Introduction to presentation attack detection in face biometrics and recent advances.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. ISO/IEC (2017). *International Organization for Standardization. Information technology - Biometric presentation attack detection - Part 3: Testing and reporting*. Technical Report ISO/IEC FDIS 30107-3:2017(E), Geneva, CH.
- Jain, A. K., Flynn, P., and Ross, A. A. (2007). *Handbook of biometrics*. Springer Science & Business Media.
- Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.
- Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. (2020). Training generative adversarial networks with limited data. In *Proc. NeurIPS*.
- Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., and Aila, T. (2021). Alias-free generative adversarial networks. In *Proc. NeurIPS*.
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization.
- Kolesnikov, A., Dosovitskiy, A., Weissenborn, D., Heigold, G., Uszkoreit, J., Beyer, L., Minderer, M., Dehghani, M., Houlsby, N., Gelly, S., Unterthiner, T., and Zhai, X. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.
- Krishnan, K. S. and Krishnan, K. S. (2021). Vision transformer based covid-19 detection using chest x-rays. In *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*, pages 644–648.
- Li, J.-W. (2008). Eye blink detection based on multiple gabor response waves. In *2008 International Conference on Machine Learning and Cybernetics*, volume 5, pages 2852–2856.
- Li, L., Feng, X., Jiang, X., Xia, Z., and Hadid, A. (2017a). Face anti-spoofing via deep local binary patterns. In

- 2017 IEEE International Conference on Image Processing (ICIP), pages 101–105.
- Li, L., Xia, Z., Li, L., Jiang, X., Feng, X., and Roli, F. (2017b). Face anti-spoofing via hybrid convolutional neural network. In *2017 International Conference on the Frontiers and Advances in Data Science (FADS)*, pages 120–124.
- Li, T.-W. and Lee, G.-C. (2021). Performance analysis of fine-tune transferred deep learning. In *2021 IEEE 3rd Eurasia Conference on IOT, Communication and Engineering (ECICE)*, pages 315–319.
- Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J. E., and Stoica, I. (2018). Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.
- Liu, Y., Jourabloo, A., and Liu, X. (2018). Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 389–398.
- Nagpal, C. and Dubey, S. R. (2019). A performance evaluation of convolutional neural networks for face anti spoofing. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.
- Pérez-Cabo, D., Jiménez-Cabello, D., Costa-Pazo, A., and López-Sastre, R. J. (2019). Deep anomaly detection for generalized face anti-spoofing. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1591–1600.
- Rodríguez, J. and Lozano, J. (2007). Repeated stratified k-fold cross-validation on supervised classification with naive bayes classifier: An empirical analysis.
- Shahapure, K. R. and Nicholas, C. (2020). Cluster quality analysis using silhouette score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 747–748.
- Shorten, C. and Khoshgoftaar, T. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6.
- Ur Rehman, Y. A., Po, L. M., and Liu, M. (2017). Deep learning for face anti-spoofing: An end-to-end approach. In *2017 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pages 195–200.
- van der Haar, D. T. (2019). Face antispoofing using shears: An empirical study. *SAIEE Africa Research Journal*, 110(2):94–103.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Walsh, J., O’ Mahony, N., Campbell, S., Carvalho, A., Krpalkova, L., Velasco-Hernandez, G., Harapanahalli, S., and Riordan, D. (2019). Deep learning vs. traditional computer vision.
- Wang, C.-Y., Lu, Y.-D., Yang, S.-T., and Lai, S.-H. (2022). Patchnet: A simple face anti-spoofing framework via fine-grained patch recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20249–20258.
- Wang, Z., Yu, Z., Zhao, C., Zhu, X., Qin, Y., Zhou, Q., Zhou, F., and Lei, Z. (2020). Deep spatial gradient and temporal depth learning for face anti-spoofing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5041–5050.
- Wu, G., Zhou, Z., and Guo, Z. (2021). A robust method with dropblock for face anti-spoofing. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.
- Xu, X., Xiong, Y., and Xia, W. (2021). On improving temporal consistency for online face liveness detection system. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 824–833.
- Yu, Z., Qin, Y., Li, X., Zhao, C., Lei, Z., and Zhao, G. (2021). Deep learning for face anti-spoofing: A survey.
- Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., Zhou, F., and Zhao, G. (2020). Searching central difference convolutional networks for face anti-spoofing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5294–5304.
- Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.
- Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., and Li, S. Z. (2012). A face antispoofing database with diverse attacks. In *2012 5th IAPR International Conference on Biometrics (ICB)*, pages 26–31.