Acoustic or Pattern? Speech Spoofing Countermeasure based on Image Pre-training Models

Jingze Lu
lujingze@hccl.ioa.ac.cn
Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, CAS.
University of Chinese Academy of Sciences
Beijing, China

Zhuo Li
lizhuo@hccl.ioa.ac.cn
Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, CAS.
University of Chinese Academy of Sciences
Beijing, China

Yuxiang Zhang
zhangyuxiang@hccl.ioa.ac.cn
Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, CAS.
University of Chinese Academy of Sciences
Beijing, China

Wenchao Wang
wangwenchao@hccl.ioa.ac.cn
Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, CAS.
Beijing, China

Pengyuan Zhang
zhangpenyuan@hccl.ioa.ac.cn
Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, CAS.
University of Chinese Academy of Sciences
Beijing, China

ABSTRACT

Traditional speech spoofing countermeasures (CM) typically contain a frontend which extracts two-dimensional feature from the waveform, and a Convolutional Neural Network (CNN) based backend classifier. This pipeline is similiar to an image classification task, in some degree. Pre-training is a widely used paradigm in many fields. Self-supervised pre-trained frontends such as Wav2Vec 2.0 have shown superior improvement in the speech spoofing detection task. However, these pre-trained models are only trained by bonafide utterances. Moreover, acoustic pre-trained frontends can also be used in the text-to-speech (TTS) and voice conversion (VC) task, which reveals that commonalities of speech are learnt by them, rather than discriminative information between real and fake data. The speech spoofing detection task and the image classification task share the same pipeline. Based on the hypothesis that CNNs follow the same pattern in capturing artefacts in these two tasks, we apply an image pre-trained CNN model to detect spoofed utterances, counterintuitively. To supplement the model with potentially missing acoustic features, we concatenate Jitter and Shimmer features to the output embedding. Our proposed CM achieves top-level performance on the ASVspoof 2019 dataset.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

DDAM '22, October 14, 2022, Lisboa, Portugal

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9496-3/22/10…$15.00
https://doi.org/10.1145/3552466.3556524

CCS CONCEPTS

• **Security and privacy** → *Human and societal aspects of security and privacy*; • **Information systems** → *Multimedia information systems.*

KEYWORDS

anti-spoofing, audio deepfakes, image pre-training, Wav2Vec2

ACM Reference Format:
Jingze Lu, Zhuo Li, Yuxiang Zhang, Wenchao Wang, and Pengyuan Zhang. 2022. Acoustic or Pattern? Speech Spoofing Countermeasure based on Image Pre-training Models. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia (DDAM '22), October 14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3552466.3556524

1 INTRODUCTION

The automatic speaker verification (ASV) [34] system plays an important role in data security and passing certification in recent years. However, just like other biometric systems, the ASV system is not completely secure in the face of illegal means or spoofing attacks [58] ,such as voice imitating and replay attack. Meanwhile, as the rapid development of deep learning in the field of speech acoustics, text-to-speech (TTS) [36] and voice conversion (VC) [23] technologies are able to used to generate spoofed utterances which are natural enough to confuse people. Therefore, there is an urgent need to develop a sufficiently reliable spoof speech detection system to protect ASV systems as well as human beings.

Traditional speech spoofing countermeasures typically contain a frontend which extracts the feature from an input utterance, and a backend that output classification score. However, existing handcrafted or Deep Neural Network (DNN) based features perform poorly in the face of multi-domain conditions or low signal noise ratio (SNR) environments [8, 59, 61]. "Pre-training and fine-tuning"

has become a common paradigm in many fields. To achieve a robust frontend, some researchers turn their eyes on self-supervised pre-trained speech model [29, 30, 43, 55]. Using Wav2vec 2.0 [3] or HuBERT [16] as the frontend, their countermeasures (CMs) show superior performance on the ASVspoof datasets [47, 59] and the ADD2022 Challenge [61].

Despite the acoustic pre-trained frontend has been proven fit for many tasks, it is still necessary to explore whether it is optimal in this task from more perspectives. For one reason, such pre-trained model learns from large-scale dataset which only contain bonafide utterances. Intuitively, this kind of training strategies makes it difficult for them to learn discriminative representations to distinguish spoof utterances from bonafide ones. In addition, the acoustic pre-trained models can also boost the performance of TTS and VC tasks [4, 33, 37, 49], which means the embeddings extracted by these frontends contain more commonalities between bonafide and spoof utterances.

These potential issues motivate us to consider another way, which is adpoting image pre-trained backends to the CMs. The mainstream CMs include a hand-crafted feature, such as Short Time Fourier Transform (STFT) or Constant-Q cepstrum coefficients (CQCC) [46], and a Convolutional Neural Network (CNN) based backend. Artefacts of the fake utterances sometimes could be directly seen from these features output by frontends. For people and CNN classifiers, the process of distinguishing such image-like features is similar to the process of distinguishing pictures. In this perspective, we assume that the significant features which learned by CNN-based classifiers of the CMs are patterns, contours and textures of the image-like features. In other words, CNN has similar learning pattern in the image classification task and speech spoofing detection task. Therefore, it is possible that CNN models pre-trained on the large-scale image classifition datasets such as ImageNet [9] have better ability to capture such patterns and artefacts than classifiers without pre-training. The above assumptions, as well as the potential issues of acoustic pre-trained frontends, motivate we to introduce the image-pretrained CNN-based model to the speech spoofing CMs.

However, intuitively, CNN backend over-learnt on the image classifiction dataset may have difficulty in learning acoustic features. Therefore, we conduct experiments to complement the embeddings of the models with additional acoustic features. By analyzing the spectrum of the utterances through GradCAM [35], an explainable visualization method, we find that the CNN-based CMs pay attention to fundamental frequency (F0) and its adjacent harmonics. This phenomenon may indicate that F0, as one of the vital feature using in the process of TTS and VC, has implicit information in distinguishing bonafide and spoof utterances. Thus, we extract Jitter and Shimmer features, which are closely related to F0, and concatenate them to the hidden embeddings.

In conclusion, our main contributions in this work include:

- Treat the speech spoofing detection as an image-like feature classification task, we investigate whether image pre-training is effective for CMs. To the best of our knowledge, this is the first work that apply image pre-trained model in spoofed speech detection task.
- Only using image-pretrained model may reduce the ability of classifiers to detect artefacts from acoustic features. To complement the model with acoustic features, we analyze the significance of Jitter and Shimmer features and concatenate them to the hidden embedding.
- A CM combine image pre-training and acoustic features is proposed, which shows superior performance on widely-used dataset.

The rest of the paper is organized as follows: Section 2 introduces some related works. Section 3 describes the methodology and the pipeline of our proposed paradigm. Section 4 introduces the experimental parameter settings. While in section 5, the results and analysis of systems are presented. The paper ends in section 6 with conclusions.

2 RELATED WORKS

2.1 Speech Spoof Detection

In response to the threat of a variety of speech spoofing attacks, the Automatic Speaker Verification Spoofing and Countermeasures (ASVspoof) Challenges were successfully held every two years since 2015 [47, 59]. Based on these challenges, [51] subtracts the log spectrogram of the vocoderfiltered audio from the one of the original audio to highlight the replay channel information. FIR filter is adopted as data augment strategy in [48]. Besides, researchers have conducted a great variety of studies on the three tasks of logical access (LA), physical access (PA) and DeepFake (DF) [11, 19, 42, 63].

The first Audio Deep synthesis Detection (ADD 2022) [61] challenge involves some challenging attacking situations ignored by the ASVspoof Challenges, including low-quality fake audio detection (LF) and partially fake audio detection (PF). In this challenge, self-supervised pre-training method has shown dominance in multiple tasks [29, 30]. Besides, [60] introduces neural stitching technology to reduce overfitting. For the PF task, inspired by the extraction-based question answering strategy, [56] proposes a self-attention-based countermeasure to discover fake span.

Several other studies related on speech spoof detection are also reported previously. [54] investigates CM training using active learning (AL) to select the training data. SHapley Additive exPlanations (SHAP) are applied in [14] to provide explanation for the behaviour of different spoofing detection models. [53] augments the CMs with confidence estimators to achieve a CM which has the capablity for opting for abstention. Mutli-task learning strategy is used to alleviate cross-domain mismatch in [31]. [26] first investigates the vulnerability of CMs under the adversarial attacks with the fast gradient sign method (FGSM) and the projected gradient descent (PGD) method. And in [57], adversarial training, a proactive defense method, is introduced to mitigate the vulnerability of CMs against adversarial attacks.

In conclusion, at present, speech spoofing detection task is a hot issue that needs to be solved urgently. The research community has made efforts to increase the robustness and the interpretability of the CMs, and decrease the vulnerability of them.

2.2 Class Activation Maps

The conception of class activation maps (CAM) is first introduced in [64]. CAM strategy is designed to visualize the discriminant basis

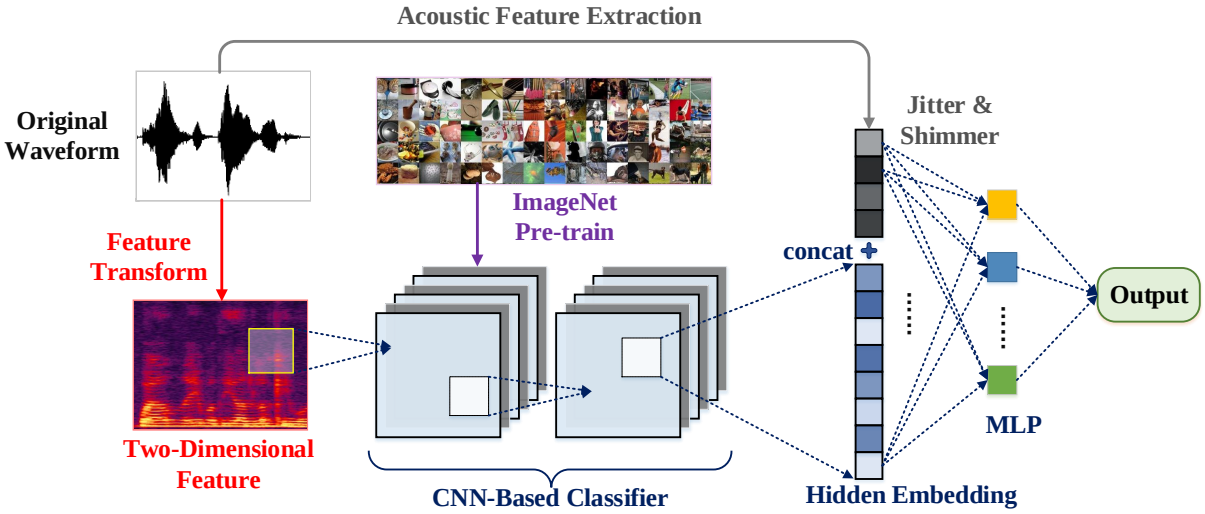

Figure 1: The pipeline of our proposed countermeasure.

of the CNN-based classifiers, by computing a weighted sum of the feature maps of the last convolutional layer and upsampling it to the size of the input image. Gradient-weighted Class Activation Mapping (Grad-CAM) [35] is improved upon CAM technology. Weighting the activations by the average gradient, Grad-CAM can produce visual explanations for CNN-based models for almost any structure. In recent years, several new CAMs have appeared in the field of computer vision. In the speech spoof detection task, using CAM to analyze the artefacts of the two dimensional hand-crafted feature has been reported in several studies [6, 28].

2.3 Pre-trained Method

"Pre-training and fine-tuning" paradigm has the ability to dig information from large-scale datasets, and has become a common paradigm in the field of CV and NLP in recent years [15]. A large number of downstream tasks achieve breakthrough by fine-tuning networks which were pre-trained for ImageNet classification in the field of CV. Meanwhile, in the field of NLP, state-of-the-art of many downstream tasks is refreshed along with the appearance of pre-trained models like BERT [10]. Similarly, pre-trained models have also achieved significant improvement in the field of Acoustics. Models such as Wav2Vec 2.0 [3] and XLS-R [2] have been popular in Automatic Speech Recognition (ASR) task and speaker verification task. Researchers have also investigated using these unsupervised or self-supervised models to detect fake utterances [29, 30, 43, 55]. However, such unsupervised pre-trained models have never seen spoof utterances from their pre-training dataset. In addition, such pre-trained model also fit for TTS and VC, which reveals that more commonalities between bonafide and spoof utterances are learned by these models, rather than difference. Thus, as an investigation to introducing pre-training paradigm to the speech spoofing detection task, we turn our eyes on image pre-training.

3 METHODOLOGY

In this section, we will introduce our proposed countermeasure equipped with image pre-trained method and hand-crafted features. Figure 1 shows the pipeline of the proposed CM.

3.1 Image Pre-trained Method

Self-supervised pre-training paradigm has been already widely used in tasks in the field of acoustics like ASR, TTS etc. Several studies have also reported excellent performance using pre-train method in detecting spoof speech. However, intuitively, the embeddings learned from the large-scale unsupervised datasets **do not** contain representations which can distinguish spoof utterances from bonafide ones. For one reason, the large-scale dataset on which pre-trained models trained contains natural speech only. Moreover, these pre-trained models can also boost the performance of TTS and VC tasks, which means that embeddings extracted by these frontends contain more commonalities between voices, rather than difference between bonafide and spoof utterances.

Mainstream speech spoofing countermeasures contain a two dimensional feature extractor and a CNN-based classifier. In this perspective, compared to other acoustics tasks, spoofed speech detection task is more similar to an image classification task. The CNN of these two tasks have similar learning pattern. CNN of both of them try to learn and capture textures and patterns from the feature-maps, and classify them correctly. Based on these analyses, compared to traditional acoustics pre-trained models, we try to apply image pre-trained models.

ImageNet [9] is an influential large-scale labeled dataset, with a huge amount of data and categories. Many downstream tasks in the CV field benifit from pre-trained models based on it. In this work, we apply a ResNet34 model pre-trained on ImageNet to the speech spoofing detection task.

Both of them learn and capture textures and patterns from the image-like features.

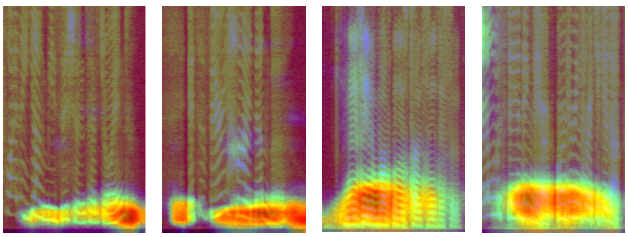

Figure 2: Grad-CAM heatmaps of speech samples from the ASVspoof dataset. The heatmaps are visualized based on STFT feature and CNN classifier.

3.2 Acoustic Features

The image pre-trained CNN-based model only has the prior information of image patterns. Intuitively, compared to pre-trained models widely used in the field of acoustics, pre-trained model based on ImageNet is lack of the characteristics of acoustic features. To verify this hypothesis, we try to supplement the image-based pre-train model with acoustic features.

Short time Fourier Transform (STFT) is a commonly used feature in a large amount of acoustic tasks. Compared to the original waveform, this kind of transform only lose phase information. The STFT feature represents the energy distribution of of a speech utterance in the time-frequency domain. The F0, formant, harmonics and other acoustic features of speech can all be represented in the STFT feature. We use Grad-CAM, a visualization method, on the STFT feature to find out which acoustic features are the CNN-based classifier really focus on. Figure 2 show the heatmaps of a SE-ResNet based classifier on several example utterances from the ASVspoof 2019 LA dataset. We randomly sampled several bonafide and spoof utterances. Among them, the spoofed utterances are generated by different spoofing attack strategies. It could be observed from the heatmaps that the classifiers mainly focus on F0 and its adjacent harmonics.

Jitter feature is quantified as the cycle-to-cycle variations of F0. In [32], it has been reported that Jitter feature is effective for distinguishing spoofing attack strategies. Based on these studies and observations, to supplement the image pre-trained based model with discriminative acoustic information, we concatenate Jitter feaure to the hidden embedding. Besides, Shimmer feature, which is quantified as the cycle-to-cycle variations of waveform amplitude, is usually used in conjunction with Jitter feature. By capturing instant-to-instant perturbations of the utterances, these two features are also reported effective to distinguish bonafide and spoof speech in [12].

These Jitter and Shimmer features are extracted by applying the Praat voice analysis tool [1]. Nine types of Jitter and Shimmer features we used are listed below [44, 45]:

- **Jitter (local, absolute):** Represents the average absolute difference between two consecutive periods, expressed as:

$$jitta = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}| \tag{1}$$

where T_i are the extracted F0 period lengths and N is the number of extracted F0 periods.

- **Jitter (local):** Represents the average absolute difference between two consecutive periods, divided by the average period, expressed as:

$$jitt = \frac{jitta}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100 \tag{2}$$

- **Jitter (rap):** Defined as the Relative Average Perturbation, the average absolute difference between a period and the average of it and its two neighbours, divided by the average period.
- **Jitter (ppq5):** Represents the ratio of disturbance within five periods, the average absolute difference between a period and the average containing its four nearest neighbor periods.
- **Shimmer (local):** Represents the average absolute difference between the amplitudes of two consecutive periods, divided by the average amplitude, expressed as:

$$shim = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100 \tag{3}$$

where A_i are the amplitude and N is the number of extracted F0 periods.

- **Shimmer (local, dB):** Represents the average absolute difference of the base 10 logarithm of the difference between two consecutive periods, expressed as:

$$shdB = \frac{1}{N-1}\sum_{i=1}^{N-1}|20 \times \log(A_{i+1}/A_i)| \tag{4}$$

- **Shimmer (apq3):** Represents the quotient of amplitude disturbance within three periods, in other word, the average absolute difference between the amplitude of a period and the mean amplitudes of its two neighbors, divided by the average amplitude.
- **Shimmer (apq5):** Represents the ratio of perturbation amplitude of five periods.
- **Shimmer (apq11):** Represents the ratio of perturbation amplitude of eleven periods

These features have different distributions between natural speech and synthesized speech, which indicate discriminative information, as shown in Figure 3. It could be observed from the figure that Jitter and Shimmer features of fake utterances have more outliers.

3.3 Image Pre-training Based Countermeasure

Figure 1 shows the whole pipeline of our image pre-training based CM. The original waveform is firstly transformed into a two dimensional feature. Next, the feature is sent to the CNN-based classifier, which is pre-trained on the ImageNet [9]. Meanwhile, to supplement the classifier with acoustic information, Jitter and Shimmer features are extracted from the original waveform and concatenated to the hidden embedding output by the classifier. Finally, the concat embedding is sent to a multilayer perceptron (MLP) and achieve the classification score.

4 EXPERIMENTAL SETUP

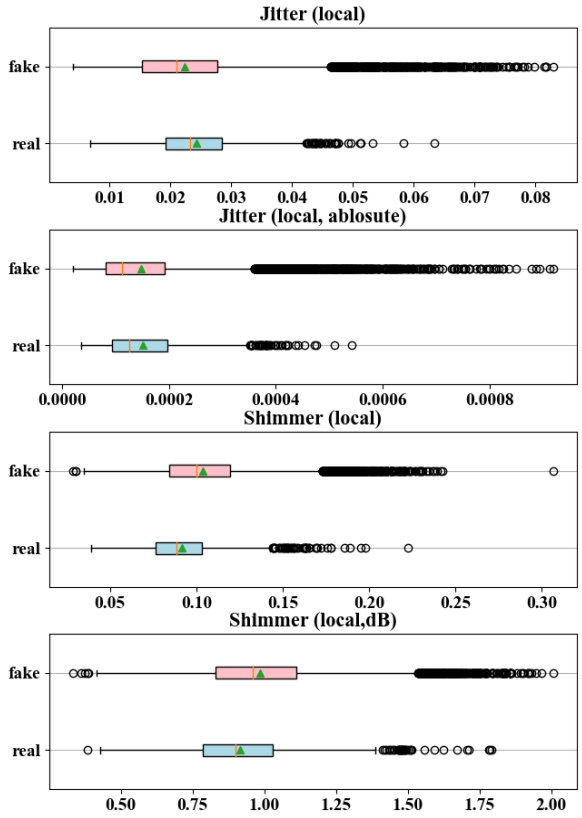

Figure 3: The values of several Jitter and Shimmer static features calculated on the ASVspoof 2019 dataset, shown as boxplot. The green triangles are the means of the data and the hollow circles show the outliers.

4.1 Datasets and Metrics

To investigate the robustness of the proposed countermeasure, experiments are conducted on the series of ASVspoof datasets, which is a group of influential datasets in this field. The ASVspoof 2019 logical access (LA) dataset is based on speech derived from the VCTK base corpus [50]. Fake utterances in the dataset are generated by 17 different TTS and VC systems. The ASVspoof 2021 (LA) dataset use the same attack strategies as the 19LA dataset, while its utterances are transmitted over different various systems including voice-over-IP (VoIP) and a public switched telephone network (PSTN). The deepfake (DF) track of 21 dataset collects about 600K utterances processed with various lossy codecs which are used typically for media storage. Data in the DF track also contain out-of-distribution data, such as utterances in other languages and data from the other datasets.

The equal error rate (EER) is used as the evaluation metric in this work. EER is defined as the point where the false acceptance rate (FAR) and the false rejection rate (FRR) are equal. The other metric is minimum normalized tandem detection cost function (min-tDCF) [22].

4.2 Front-End and Model Architecture

For the image pre-trained method, we choose STFT as the input feature. STFT feature is a two dimensional vector, which represent the energy distribution in the time-frequency domain of an utterance. The STFT is extracted with window length 1024, hop length 160 and a 1024-point FFT.

We choose SENet pre-trained on the ImageNet [9] as the CNN-based classifier. SENet is an improved version of the ResNet , with squeeze-and-excitation (SE) [17] block. A Global Average Pooling layer is connected after the SENet block, which calculate the mean value of each channels. The loss function used in this binary classification task is the angular margin based softmax loss (A-softmax) [27].

To supplement the image pre-trained model with acoustic information, we extract nine kinds of Jitter and Shimmer features from original waveform, and concatenate them with the 128-dimensional embeddings output by the global average pooling layer. The embeddings are then sent to a multilayer perceptron (MLP) and achieve the classification score.

Meanwhile, as a control experiment, we also build a pipeline based on Wav2Vec 2.0 frontend. The pre-trained model used is Wav2Vec2-large-xlsr-53. The dimension of features output by the last layer of Wav2Vec frontend are first reduced from 1024 to 128 through a fully connected layer. The embedding features are then obtained by mean pooling in the time domain. Embedding features are fed into a fully connected layer for binary classification.

4.3 Details of Systems Implementation

For the CNN-based CM, the layers of the CNN backend are initialized with filters pretrained on ImageNet [9]. During the warm-up period, we fix the weights and biases of these initialized layers, only update those of the MLPs layers. After the first five epochs, we train all convolutional layers and MLP layers jointly. The loss function is minimized using Adam [21] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$ and weigh decay 10^{-4}. Warm-up steps of the loss function are set to 1000, after which, the learning rate decreases proportionally to the inverse square root of the step number. All models are trained with 40 epochs, in which the model with the lowest loss on the development dataset was selected as the final model.

For the Wav2Vec 2.0 based CM, the weights of the Wav2Vec 2.0 frontend are updated during the training period. The optimizer is Adam with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-8}$ and weight decay 10^{-4}. The learning rate is initialized to 1×10^{-6}. The stepLR is used as scheduler with a step size of 10 epochs and a coefficient of 0.5. The model with the smallest dev set EER is selected as the final model for evaluation.

5 RESULT AND ANALYSIS

Table 1 shows the ablation results of our proposed image pre-trained paradigm. The results show that compared to the original CNN based countermeasure, on all the three datasets, the image pre-trained model combined with Jitter and Shimmer features achieve boosted performance. Among them, the result evaluated on the ASVspoof 2019 LA dataset is significantly reduced by 60.27%. The result on the 21 LA and 21 DF achieve 30.14% and 4.01% relative

Table 1: Results of ablation experiments of our proposed method. EER(%) is the eval metric of all results. *Prog* and *Eval* in the table are two evaluation sets of the ASVspoof Challenge. (1: Only STFT feature and ResNet34; 2: Apply image pre-train; 3: Add Jitter and Shimmer features to the embeddings.)

Model	2019 LA	2021 LA		2021 DF	
		prog	eval	prog	eval
1	2.19	13.19	14.40	11.93	28.21
1 + 2	1.42	12.23	12.36	9.31	26.94
1 + 3	2.06	12.95	15.45	8.72	24.25
1 + 2 + 3	**0.87**	**9.50**	**10.06**	**8.15**	**27.08**

Table 2: Comparison of the image pre-training with Wav2Vec frontend. EER(%) is the eval metric of all results.

Model	19 LA	21 LA		21 DF	
		prog	eval	prog	eval
Img-pre + J&S	**0.87**	**9.50**	**10.06**	**8.15**	**27.08**
Wav2Vec 2.0	1.78	8.77	6.8	1.66	8.7

EER reduction, respectively. The boosted performance prove our hypothesis, which is the image pre-trained model can enhance the ability of the CNN classifier to detect pattern and artefacts. Meanwhile, the ablation experiments prove that adding acoustic features to the embedding output by CNN based classifiers can improve the performance of CMs. Moreover, only applying image pre-training or concatenating the Jitter and Shimmer features is also effective, to a certain extent.

Table 2 shows the comparison results between our proposed image pre-training method and a CM based on Wav2Vec 2.0 frontend on the three datasets. It could be observed from the table that our proposed method achieves better performance on the 19 dataset. It achieve a competitive while slightly worse performance compared to the Wav2Vec based CM on the 21 LA dataset. And the results on the DF dataset are far inferior to the Wav2Vec. The reason for this phenomenon is that the utterances in the 21LA and 21DF dataset are processed by various codec algorithms, which may fade the textures and artefacts learnt by backend. Moreover, the 21DF dataset contains out-of-distribution utterances from the other datasets and in other language. The Wav2Vec frontend has seen large amount of speech data so it is more robust to various kinds of speech utterances. Therefore, the Wav2Vec frontend is better at handling out-of-distribution data in the DF dataset. By contrast, the image pre-trained model are more capable of capturing artefacts and patterns of in-distribution data, which is more similar to the training data. It is worth noting that, the Wav2Vec is a frontend, while the image pre-trained model is a backend. Combining these two frontend and backend may merge their strengths, which will be investigated in our future work.

Table 3 shows the comparison of our proposed image pre-trained model with other recently proposed state-of-the-art systems on the ASVspoof 2019 dataset. Our system is only slightly worse than AASIST system [20] with RawWave on the EER metric. It is worth mentioning that our proposed CM is improved from a 2.19 EER baseline. Our proposed method is a kind of paradigm. Similarly, it is also possible to get a performance improvement using our method for other better performed CNN based CM. Overall, our approach makes a mediocre CM achieve top-level performance on the ASVspoof 2019 dataset.

To make the system more robust to the multi-domain conditions, we have also tried to use data augment strategy on our proposed CM. RawBoost [40] method is an effective data augment method for multi-domain conditions. Adding this kind of data augmentation

Table 3: Comparison with recently proposed state-of-the-art systems on the ASVspoof2019 LA dataset. The results are reported using Min-tDCF and EER(%), which are both the smaller the better. All the results listed below are based on single model, without any kind of score-level ensemble.

Model	2019 LA	
	min-tDCF	EER(%)
Img-pre + J&S(proposed)	**0.027**	**0.87**
AASIST [20]	0.028	0.83
RAWGAT-ST [38]	0.034	1.06
STFT(Low)+ResNet [63]	0.037	1.14
Wav2Vec 2.0[55]	0.100	1.28
Res-TSSDNet [18]	0.048	1.64
Raw PC-DARTS [13]	0.052	1.77
CQT+MCG-Res2Net [25]	0.052	1.78
LFCC+LCNN+LSTM [52]	0.052	1.92
STFT+ResNet+OCsoftmax [62]	0.059	2.19
LFCC+GMM [41]	0.090	3.50
HuBERT [55]	0.157	3.55
Siamese CNN [24]	0.093	3.79
GAT-S [39]	0.091	4.48
RawNet2 [39]	0.155	5.54

Table 4: Results of our proposed method with Rawboost. EER(%) is the eval metric of all results.(1: Only STFT feature and ResNet34; 2: Apply image pre-train; 3. Add Jitter and Shimmer features to the embeddings.) Compared to CMs with acoustic pre-trained frontends and some CNN based CMs.

Model	2021 LA	2021 DF
1	14.40	28.21
1+2+3+RawBoost(proposed)	**7.71**	**19.11**
Wav2Vec 2.0	6.8	8.7
Wav2Vec 2.0 [55]	9.66	4.75
Wav2Vec 2.0 [30]	7.20	5.68
HuBERT [55]	9.55	13.07
GMM+LCNN(Ensemble) [7]	3.62	18.30
ECAPA-TDNN(Ensemble) [5]	5.46	20.33

can boost the robustness of CMs against multiple codec environments. Table 4 shows the result of our CM which adopts Rawboost on the ASVspoof 2021 LA and DF datasets, as well as some CMs with acoustic pre-trained frondend and some CNN based CMs. The results shows that after adding the data augment strategy, EER of our proposed relatively reduce 46.46% and 32.26% on the ASVspoof 2021 LA and DF dataset, respectively. After adding data augment strategy in the pipeline, the proposed CM achieve competitive results compared to the CM with acoustic pre-trained frontend on the 21 LA dataset. Our results are worse than other CNN based model on the 21 LA dataset. However, our training paradigm achieve great improvement compared to the original training strategy without it. It is possible to get a similar performance improvement using our training paradigm for other better performed CNN based CM. On the DF dataset, the performance is still much worse than acoustic pre-trained frontends. However, on the DF dataset, other CNN based CMs also cannot achieve similar results to them. Same as the analysis above, an acoustic pre-trained frontend has unique advantages when dealing with out-of-distribution data, like utterances of other languages and from other datasets. Some types of acoustic characteristics **can not** be easily captured as a pattern by CNN-based classifier, without pre-trained on diverse large-scale speech datasets. Therefore, combining the acoustic pre-trained frontend and image pre-trained backend and merge their advantages might be a reasonable hypothesis. We will conduct research on it in our future work.

6 CONCLUSION

In this work, based on the hypothesis that CNN based classifiers treat speech spoofing detection task as a image classification task and capture patterns and artefacts from the two dimensional acoustic features, we introduce image pre-training paradigm to the speech spoofing CM. Models pre-trained by ImageNet are adopted and achieve better performance. To supplement the model with potentially missing acoustic features, we concatenate Jitter and Shimmer features to the output embedding. Although our proposed CM achieve top performing results on the in-distribution data, results also prove that it is not good at dealing with out-of-distribution data compared to acoustic pre-trained frontend such as Wav2Vec 2.0. However, our proposed paradigm is to use a pre-trained backend, while Wav2Vec 2.0 is a frontend. Combining them may merge their advantages and achieve good performance on both in-distribution data and out-of-distribution data. Our future work will focus on this possibility.

REFERENCES

[1] 2022. Praat software website. http://www.fon.hum.uva.nl/praat/.

[2] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296* (2021).

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems* 33 (2020), 12449–12460.

[4] Li-Wei Chen and Alexander Rudnicky. 2022. Fine-grained style control in Transformer-based Text-to-speech Synthesis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7907–7911.

[5] Xinhui Chen, You Zhang, Ge Zhu, and Zhiyao Duan. 2021. UR channel-robust synthetic speech detection system for ASVspoof 2021. *arXiv preprint arXiv:2107.12018* (2021).

[6] Xingliang Cheng, Mingxing Xu, and Thomas Fang Zheng. 2019. Replay detection using CQT-based modified group delay feature and ResNeWt network in ASVspoof 2019. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 540–545.

[7] Rohan Kumar Das. 2021. Known-unknown data augmentation strategies for detection of logical access, physical access and speech deepfake attacks: ASVspoof 2021. In *Proc. ASVspoof2021 Workshop*.

[8] Rohan Kumar Das, Jichen Yang, and Haizhou Li. 2020. Assessing the scope of generalized countermeasures for anti-spoofing. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6589–6593.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[11] Roberto Font, Juan M Espín, and María José Cano. 2017. Experimental analysis of features for replay attack detection-results on the ASVspoof 2017 Challenge.. In *Interspeech*. 7–11.

[12] Yang Gao, Jiachen Lian, Bhiksha Raj, and Rita Singh. 2021. Detection and evaluation of human and machine generated speech in spoofing attacks on automatic speaker verification systems. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 544–551.

[13] Wanying Ge, Jose Patino, Massimiliano Todisco, and Nicholas Evans. 2021. Raw differentiable architecture search for speech deepfake and spoofing detection. *arXiv preprint arXiv:2107.12212* (2021).

[14] Wanying Ge, Jose Patino, Massimiliano Todisco, and Nicholas Evans. 2022. Explaining deep learning models for spoofing and deepfake detection with SHapley Additive exPlanations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6387–6391.

[15] Kaiming He, Ross Girshick, and Piotr Dollár. 2019. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4918–4927.

[16] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460.

[17] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[18] Guang Hua, Andrew Beng Jin Teoh, and Haijian Zhang. 2021. Towards end-to-end synthetic speech detection. *IEEE Signal Processing Letters* 28 (2021), 1265–1269.

[19] Sally Jones et al. 2011. Speech is silver, silence is golden: The cultural importance of silence in Japan. *The ANU Undergraduate Research Journal* 3 (2011), 17–27.

[20] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. 2022. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6367–6371.

[21] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[22] Tomi Kinnunen et al. 2020. Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2195–2210.

[23] Tomi Kinnunen, Zhi-Zheng Wu, Kong Aik Lee, Filip Sedlak, Eng Siong Chng, and Haizhou Li. 2012. Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4401–4404.

[24] Zhenchun Lei, Yingen Yang, Changhong Liu, and Jihua Ye. 2020. Siamese Convolutional Neural Network Using Gaussian Probability Feature for Spoofing Speech Detection.. In *INTERSPEECH*. 1116–1120.

[25] Xu Li, Xixin Wu, Hui Lu, Xunying Liu, and Helen Meng. 2021. Channel-wise gated res2net: Towards robust detection of synthetic speech attacks. *arXiv preprint arXiv:2107.08803* (2021).

[26] Songxiang Liu, Haibin Wu, Hung-yi Lee, and Helen Meng. 2019. Adversarial attacks on spoofing countermeasures of automatic speaker verification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 312–319.

[27] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 212–220.

[28] Jingze Lu, Yuxiang Zhang, Wenchao Wang, and Pengyuan Zhang. 2022. Robust Cross-SubBand Countermeasure Against Replay Attacks. In *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*. 126–132. https://doi.org/10.21437/Odyssey.2022-18

[29] Zhiqiang Lv, Shanshan Zhang, Kai Tang, and Pengfei Hu. 2022. Fake Audio Detection Based On Unsupervised Pretraining Models. In *ICASSP 2022-2022 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 9231–9235.

[30] Juan M Martín-Doñas and Aitor Álvarez. 2022. The Vicomtech Audio Deepfake Detection System Based on Wav2vec2 for the 2022 ADD Challenge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 9241–9245.

[31] Yichuan Mo and Shilin Wang. 2022. Multi-Task Learning Improves Synthetic Speech Detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6392–6396.

[32] Nicolas M Müller, Franziska Dieckmann, and Jennifer Williams. 2022. Attacker Attribution of Audio Deepfakes. *arXiv preprint arXiv:2203.15563* (2022).

[33] Sravya Popuri, Peng-Jen Chen, Changhan Wang, Juan Pino, Yossi Adi, Jiatao Gu, Wei-Ning Hsu, and Ann Lee. 2022. Enhanced Direct Speech-to-Speech Translation Using Self-supervised Pre-training and Data Augmentation. *arXiv preprint arXiv:2204.02967* (2022).

[34] Douglas A Reynolds. 1995. Speaker identification and verification using Gaussian mixture speaker models. *Speech communication* 17, 1-2 (1995), 91–108.

[35] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2016. Grad-CAM: Why did you say that? *arXiv preprint arXiv:1611.07450* (2016).

[36] Vadim Shchemelinin and Konstantin Simonchik. 2013. Examining vulnerability of voice verification systems to spoofing attacks by means of a TTS system. In *International Conference on Speech and Computer*. Springer, 132–137.

[37] Hubert Siuzdak, Piotr Dura, Pol van Rijn, and Nori Jacoby. 2022. WavThruVec: Latent speech representation as intermediate features for neural speech synthesis. *arXiv preprint arXiv:2203.16930* (2022).

[38] Hemlata Tak, Jee-weon Jung, Jose Patino, Madhu Kamble, Massimiliano Todisco, and Nicholas Evans. 2021. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. *arXiv preprint arXiv:2107.12710* (2021).

[39] Hemlata Tak, Jee-weon Jung, Jose Patino, Massimiliano Todisco, and Nicholas Evans. 2021. Graph attention networks for anti-spoofing. *arXiv preprint arXiv:2104.03654* (2021).

[40] Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans. 2022. Rawboost: A Raw Data Boosting and Augmentation Method Applied to Automatic Speaker Verification Anti-Spoofing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6382–6386.

[41] Hemlata Tak, Jose Patino, Andreas Nautsch, Nicholas Evans, and Massimiliano Todisco. 2020. Spoofing attack detection using the non-linear fusion of sub-band classifiers. *arXiv preprint arXiv:2005.10393* (2020).

[42] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. 2021. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6369–6373.

[43] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee weon Jung, Junichi Yamagishi, and Nicholas Evans. 2022. Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation. In *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*. 112–119. https://doi.org/10.21437/Odyssey.2022-16

[44] João Paulo Teixeira and André Gonçalves. 2016. Algorithm for jitter and shimmer measurement in pathologic voices. *Procedia Computer Science* 100 (2016), 271–279.

[45] João Paulo Teixeira, Carla Oliveira, and Carla Lopes. 2013. Vocal acoustic analysis–jitter, shimmer and hnr parameters. *Procedia Technology* 9 (2013), 1112–1122.

[46] Massimiliano Todisco, Héctor Delgado, and Nicholas Evans. 2017. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language* 45 (2017), 516–535.

[47] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. 2019. ASVspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441* (2019).

[48] Anton Tomilov, Aleksei Svishchev, Marina Volkova, Artem Chirkovskiy, Alexander Kondratev, and Galina Lavrentyeva. 2021. STC antispoofing systems for the ASVspoof2021 challenge. In *Proc. ASVspoof 2021 Workshop*. 61–67.

[49] Benjamin van Niekerk, Marc-André Carbonneau, Julian Zaïdi, Matthew Baas, Hugo Seuté, and Herman Kamper. 2022. A comparison of discrete and soft speech units for improved voice conversion. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6562–6566.

[50] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. 2017. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)* (2017).

[51] Xingming Wang, Xiaoyi Qin, Tinglong Zhu, Chao Wang, Shilei Zhang, and Ming Li. 2021. The DKU-CMRI System for the ASVspoof 2021 Challenge: Vocoder based Replay Channel Response Estimation. *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge* (2021), 16–21.

[52] Xin Wang and Junich Yamagishi. 2021. A comparative study on recent neural spoofing countermeasures for synthetic speech detection. *arXiv preprint arXiv:2103.11326* (2021).

[53] Xin Wang and Junichi Yamagishi. 2022. Estimating the confidence of speech spoofing countermeasure. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6372–6376.

[54] Xin Wang and Junich Yamagishi. 2022. Investigating Active-learning-based Training Data Selection for Speech Spoofing Countermeasure. *arXiv preprint arXiv:2203.14553* (2022).

[55] Xin Wang and Junichi Yamagishi. 2022. Investigating Self-Supervised Front Ends for Speech Spoofing Countermeasures. In *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*. 100–106. https://doi.org/10.21437/Odyssey.2022-14

[56] Haibin Wu, Heng-Cheng Kuo, Naijun Zheng, Kuo-Hsuan Hung, Hung-Yi Lee, Yu Tsao, Hsin-Min Wang, and Helen Meng. 2022. Partially Fake Audio Detection by Self-Attention-Based Fake Span Discovery. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 9236–9240.

[57] Haibin Wu, Songxiang Liu, Helen Meng, and Hung-yi Lee. 2020. Defense against adversarial attacks on spoofing countermeasures of asv. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6564–6568.

[58] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. 2015. Spoofing and countermeasures for speaker verification: A survey. *speech communication* 66 (2015), 130–153.

[59] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al. 2021. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537* (2021).

[60] Rui Yan, Cheng Wen, Shuran Zhou, Tingwei Guo, Wei Zou, and Xiangang Li. 2022. Audio Deepfake Detection System with Neural Stitching for ADD 2022. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 9226–9230.

[61] Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, et al. 2022. Add 2022: the first audio deep synthesis detection challenge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 9216–9220.

[62] You Zhang, Fei Jiang, and Zhiyao Duan. 2021. One-class learning towards synthetic voice spoofing detection. *IEEE Signal Processing Letters* 28 (2021), 937–941.

[63] Yuxiang Zhang, Wenchao Wang, and Pengyuan Zhang. 2021. The Effect of Silence and Dual-Band Fusion in Anti-Spoofing System. In *Proc. Interspeech 2021*. 4279–4283. https://doi.org/10.21437/Interspeech.2021-1281

[64] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929.