Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 33 No. 1s (2026)

Advanced Liveness Detection Using Vision Transformers and the DINO Framework

T. Sai Lalith Prasad
Assistant Professor Artificial Intelligence and Data Science
Vignan Institute of Technology and Science

Patvadi Keerthi
Artificial Intelligence and Data Science
Vignan Institute of Technology and Science Hyderabad, India

Pasupula Srivardhan
Artificial Intelligence and Data Science
Vignan Institute of Technology and Science Hyderabad, India

Vanga Ajay
Artificial Intelligence and Data Science
Vignan Institute of Technology and Science Hyderabad, India
Hyderabad, India lalithresearch15@gmail.comkeerthipatvadi@gmail.com
srivardhanpaspula3535@gmail.com vangaajay72@gmail.com

Article History:
Received: 04-02-2026
Revised: 20-03-2026
Accepted: 10-04-2026

Abstract:
Face recognition systems have been increasing used in banking, mobile authentication and secure access control - however, they are prone to presentation attacks such as printed photos, replay videos and spoof masks. This project proposes a strong face anti-spoofing or liveness detection using Vision Transformer (ViT) and DINO based self-supervised learning. Unlike typical CNN-based approaches relying considerably on local texture clues and large labeled data, in this approach the approach takes global facial patterns into account using fully attention mechanisms of transformer and helps to create such discriminative representations learning from augmented facial images. The system consists of face preprocessing, feature ex- traction and binary classification to determine whether the input is live or spoof. It is tested on standard biometric anti-spoofing measures such as Accuracy, APCER, BPCER and ACER. The proposed method attempts to enhance the generalization to the unseen spoofing attacks while preserving the practicality for deployment. This work introduces the potential of transformer based self-supervised models to secure and reliable biometric authentication systems.

Index Terms—Face Anti-Spoofing, Liveness Detection, Vision Transformer (ViT), DINO, Self-Supervised Learning, Biometric Authentication

I. INTRODUCTION

Face recognition systems have been extraordinarily success- ful in recent years with the advancement in deep learning and vast amounts of training data. The introduction of Vision Transformer (ViT) proved that the self attention mechanisms can be used effectively to model the long-range dependencies within images by representing them as a sequence of patches rather than just using convolutional operations [1]. This paradigm shift demonstrated the benefits of global contextual modeling to be superior to classical and convolution-based architectures in different vision tasks.

Taking transformer architectures as base, self-supervised learning methods have further improved representation learn- ing capabilities. In particular, Caron et al. has proposed DINO, a self-distillation framework for Vision Transformers that allows learning powerful visual features without explicit labels [2]. Such approaches go a long way to reduce the reliance on large datasets with annotations, as well as enhance feature generalization.

Despite these developments, face recognition systems are still susceptible to presentation attacks, including printed pho- tographs, replayed videos, and three-dimensional masks. A very thorough survey on the topic of face anti-spoofing by Yu et. al. emphasized that deep learning based face anti-spoofing techniques still suffer from cross-dataset generalization and unseen attack cases [3]. Recent multimodal approaches, such as polarized image translation techniques, have tried to in- crease the robustness of the approach by utilizing comple- mentary sensing modalities [4].

To further deal with the generalization problem, one-class learning schemes have been proposed to focus mainly on the modeling of bona fide samples and the less dependence on diverse spoof examples [5]. Additionally, multimodal data fusion frameworks have shown better performances using RGB, depth, and other biometric cues [6]. Earlier CNN-based approaches, such as patch-based methods and methods that rely on depth, ensured the establishment of strong approaches to spoof detection through fine-grained spatial features [7].

Although these approaches have given promising results, many existing systems are based on convolutional networks which mainly collect local texture information. Motivated by the success of transformer architectures and self-supervised learning, this work proposes a Vision Transformer-based face anti-spoofing framework that is enhanced with the self-supervised representation learning. The purpose of the pro- posed method is to utilize global contextual modeling and identify subtle spoofing artifacts to enhance the robustness against various presentation attacks.

II. RELATED WORK

Deep learning has given a great boost to the field of face anti-spoofing (FAS). Liu et al. [8] proposed the auxiliary su- pervision strategies that exploited the depth and physiological signals to improve the spoof detection performance beyond the simple binary classification. Their work had shown that auxiliary tasks enhance feature discriminability.

Earlier, Chingovska et al. [9] demonstrated effectiveness of Local Binary Patterns (LBP)

in detection of spoof attack through the capturing of micro-texture features between real and fake faces. Although effective, handcrafted descriptors had trouble with generalization effect.

To standardize the evaluation procedures in the biometric presentation attack detection, ISO / IEC 30107-3 [10] defined metrics such as APCER and BPCER providing a unified benchmarking framework for anti-spoofing systems.

With the advent of transformer architectures, the approach known as data-efficient image transformers (DeiT) was pro- posed by Touvron et al. [11], showing that attention-based models have good capability of performing competitively than before with limited data for labeling. Similarly, He et al.

[12] proposed Masked Autoencoders (MAE) which allowed scalable self-supervised learning of vision transformers.

In order to solve domain generalization problems in face anti-spoofing, Wang et al. [13] proposed a multi-domain learning framework to make it more robust across datasets. Yu et al. [14] further improved CNN-based methods by proposed Central Difference Convolutional Networks (CDCN) which focused on gradient-based texture cues for spoof detection. Liu et al. [15] introduced a zero-shot face anti-spoofing approach based on deep tree learning, which focuses on unseen types of attacks.

Beyond supervised approaches, self-supervised representa- tion learning approaches have come into the focus. He et al. [16] proposed Momentum Contrast, MoCo, a contrastive learning framework for unsupervised visual representation learning. Grill et al. [17] later proposed BYOL which removed the negative sample pairs requirement and still achieved a good representation quality.

Although these works enhanced the performance of spoof detection significantly, combining transformer-based global modeling and robust face anti-spoofing is an open research direction.

III. PROPOSED METHODOLOGY

A. *Problem Definition*

Let $x \in \mathrm{R}^{H \times W \times 3}$ denote an input facial image and $y \in \{0, 1\}$ denote its corresponding label, where 1 represents a bona fide (live) sample and 0 represents a spoof attack. The objective is to learn a mapping function

$$f_\theta : x \rightarrow y \tag{1}$$

that remains robust under domain variations such as illu- mination changes, device differences, and unseen presentation attacks.

Unlike conventional convolutional neural networks that fo- cus primarily on local texture cues, the proposed approach leverages global contextual modeling using transformer-based architecture combined with self-supervised representation learning.

B. *Vision Transformer Backbone*

Given a preprocessed and aligned facial image, the image is divided into non-overlapping patches of size $P \times P$. The total number of patches is

$$N = \frac{H \times W}{P^2}. \quad (2)$$

Each patch x_i is flattened and linearly projected into a D- dimensional embedding space:

$$z_i = x_i E, \quad (3)$$

where $E \in \mathrm{R}^{(P^2 C) \times D}$ represents the projection matrix.
A learnable classification token z_{cls} is prepended to the patch sequence and positional embeddings are added:

$$z_0 = [z_{cls}; z_1; \ldots; z_N] + E_{pos}. \quad (4)$$

The transformer encoder consists of stacked multi-head self- attention (MHSA) and feed-forward layers. The self-attention mechanism is computed as

$$\text{Attention}(Q, K, V) = \text{softmax}\frac{QK^T}{\sqrt{d_k}}V, \quad (5)$$

where Q, K, and V denote query, key, and value matrices respectively, and d_k is the scaling factor.
This formulation enables global interaction among all fa- cial patches, allowing the model to capture distributed spoof artifacts. As shown in Fig. 1, the proposed model combines transformer-based global feature modeling and self-supervised distillation to enhance the performance of liveness detection.

C. *Self-Supervised Distillation*

To improve feature generalization, a self-supervised dis- tillation strategy is employed. The framework consists of a student network f_s and a teacher network f_t with identical architectures.
The teacher parameters are updated using exponential mov- ing average:

$$\theta_t \leftarrow \lambda\theta_t + (1 - \lambda)\theta_s, \quad (6)$$

where λ is the momentum coefficient.
Given two augmented views of the same input, the student and teacher produce probability distributions p_s and p_t. The self-supervised objective is defined as:

$$\mathbf{L}_{SSL} = -\sum_i p_t(i) \log p_s(i). \quad (7)$$

This consistency-based learning encourages invariant and discriminative feature

representations without heavy reliance on labeled spoof data.

TABLE I
EVALUATION METRICS COMPARISON

Model	Acc.	APCER	BPCER	ACER
EfficientNet-B2	88.2	22.5	1.0	11.75
EfficientNet-B2 (NS)	92.4	9.8	0.7	5.25
MobileViT	97.0	5.5	0.4	2.95
ViT-DINO (Proposed)	**99.8**	**1.6**	**0.1**	**0.8**

False Accepts

$$APCER = \text{Total Attack Samples'} \quad (11)$$

False Rejects

$$BPCER = \text{Total Genuine Samples'} \quad (12)$$

$$ACER = \frac{APCER + BPCER}{}$$

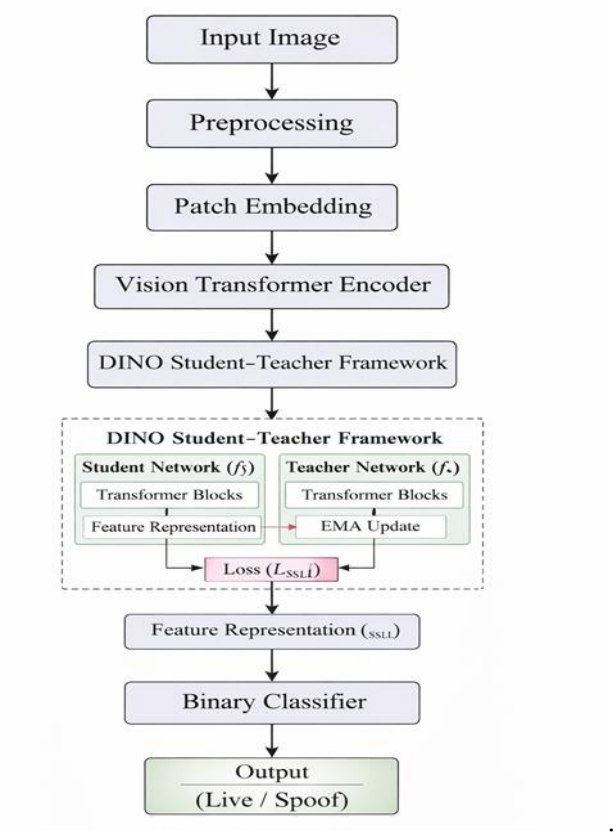

Fig. 1. The overall system Architecture

D. *Supervised Liveness Classification*

The final representation corresponding to the classification token is passed through a fully connected layer:

$$\hat{y} = \sigma(Wz_{cls} + b), \quad (8)$$

where W and b are learnable parameters and σ denotes the sigmoid activation function.
Binary cross-entropy loss is used for supervised optimiza- tion:

$$L_{sup} = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})]. \quad (9)$$

The total training objective combines supervised and self-supervised losses:

$$L_{total} = L_{sup} + \alpha L_{SSL}, \quad (10)$$

where α controls the contribution of the self-supervised term.

E. *Evaluation Metrics*

The proposed model is evaluated using ISO/IEC 30107-3 compliant metrics:These metrics ensure standardized benchmarking of presen- tation attack detection performance.

IV. Results and Discussion

A. *Quantitative Performance Appraisal*

The quantitative performance comparison of EfficientNet- B2, EfficientNet-B2 with Noisy Student training, MobileViT and the proposed ViT-DINO framework is shown in Table 1. The classification accuracy of the proposed model is 99.8%, which is improved significantly compared to EfficientNet- B2(88.2%), EfficientNet-B2(Noisy Student)(92.4%) and Mo- bileViT(97.0%). This large improvement shows the power of transformer-based global feature modeling in combination with self-supervised distillation.

In addition to overall accuracy, security-related evaluation metrics also attest the robustness of the proposed approach. The Attack Presentation Classification Error Rate (APCER) is lowered to 1.6%, as compared to 22.5% for EfficientNet- B2 and 5.5% for MobileViT. This dramatic reduction is an indication of superior capability in the detection of spoof attacks and the minimization of false acceptances. Simulta- neously, the Bona Fide Presentation Classification Error Rate (BPCER) is also reduced down to 0.1%, which guarantees that real users are only mis-classified rarely. The resulting Average Classification Error Rate (ACER) of 0.8% is a balanced and very reliable result from both attack and bona fide samples. The decrease in ACER from 11.75% (EfficientNet-B2) to 0.8% demonstrates the tremendous progress made by the proposed framework. As shown in Fig. 2, the proposed ViT-DINO model consistently outperforms baseline models in overall classification accuracy.

The experimental results show obvious performance dif- ferences between the convolution-based architecture and transformer-based architecture. EfficientNet-B2, with a feature extraction limited to convolution, has a high APCER of 22.5%, thus making it vulnerable to

spoof attacks. Although Noisy Student training strategy enhances performance by decreasing ACER to a value of 5.25, it is not as good as transformer-based approaches.

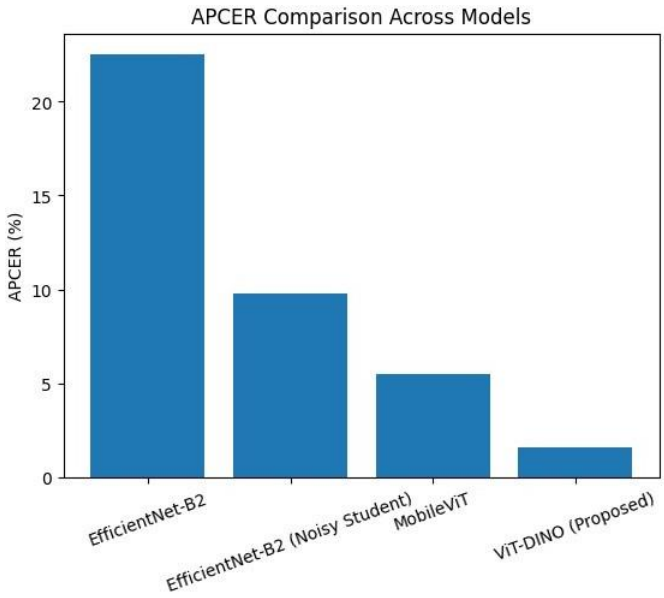

Fig. 2. Accuracy comparision across models

MobileViT with its integration of light transformer com- ponents are resulting in better accuracy and low error rate than CNN-based models. However, still, its hybrid archi- tecture restricts the use of global attention mechanisms. In contrast, the proposed ViT-DINO framework uses a fully transformer-based encoder that is able to model long-range dependencies among all facial patches. This way, the system can detect inconsistencies in reflectance which are distributed and subtle spoof artifacts that are easily missed by localized convolutional filters. The consistent decrease in APCER in all comparisons establishes the benefit of global contextual modeling to presentation attack detection. Fig. 3 illustrates the substantial reduction in attack presentation errors achieved by the proposed model.

B. Effects of Self Supervised Representation Learning

A major part of the performance increases brought are due to using DINO-based self-supervised distillation. The student- teacher learning paradigm imposes feature consistency across different augmented views of the same input that encourages invariant and discriminative representations. This limits the perception of variations in the environment such as illumi- nation changes and device characteristics.

The very low BPCER of 0.1% shows that the model is effec- tive in preserving the real facial features and in suppressing the spurious artefacts in the spoofed faces. The dramatic decrease of APCER is another proof of the better attack generalization capability. The integrated supervised liveness classification and self-supervised feature alignment of the proposed framework helps to provide its superior robustness without focusing on individual dataset-

specific spoof patterns.

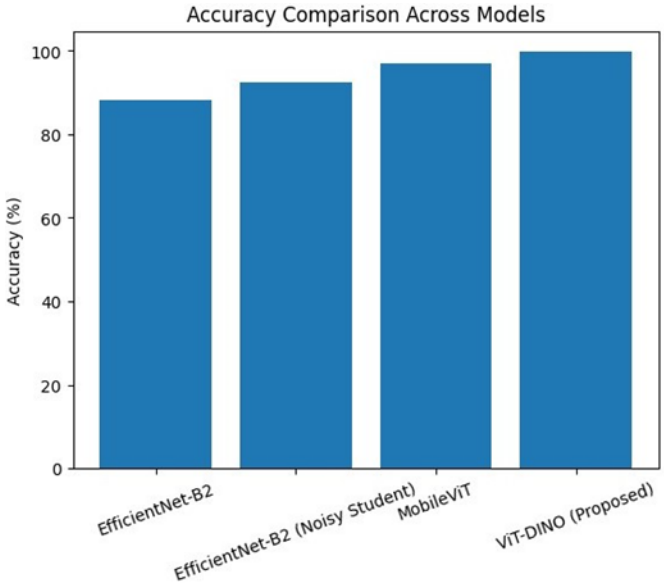

Fig. 3. APCER Models across comparision

C. *Real-Time Deployment Testings*

To assess whether this model was practical, the proposed model was tested using real time deployment conditions. As shown in Fig. 7, the system is able to detect real face with high confidence scores, which demonstrates the stability of liveness detection in the live capture scenario. In contrast, successful spoof attempt detection has been presented in Fig. 8 and Fig. 9 in the form of mobile device displays. The system detects the spoofed face accurately and gives a ”Fake Alert” notification with high confidence about spoofing.

These real time experiments validate that the model has a high level of reliability outside evaluation with the controlled dataset. The consistent detection of both the live and spoof samples in the dynamic scenarios indicate the good general- ization ability and practical robustness. As the results in Fig. 4 show, the proposed model can accurately detect a bona fide face in real-time at a high confidence score, which shows stable liveness prediction under the condition of live capture. Fig. 5 shows successful detection of a spoof attack presented through a mobile device display where the system successfully classifies the sample to be fake and returns a spoof alert. As shown in Fig. 6, the system discriminates between live and spoof presentations at the same time, and can successfully detect the true face as well as mark the spoof sample as false.

D. *Discussion*

The experimental results verify that Vision Transformers and self-supervised distillation combination indeed leads to a significant improvement in anti-spoofing performance on faces. The transformer backbone helps to obtain global atten- tion across the facial regions and enhance the ability to detect the distributed artifacts, and the self-supervised learning mech-

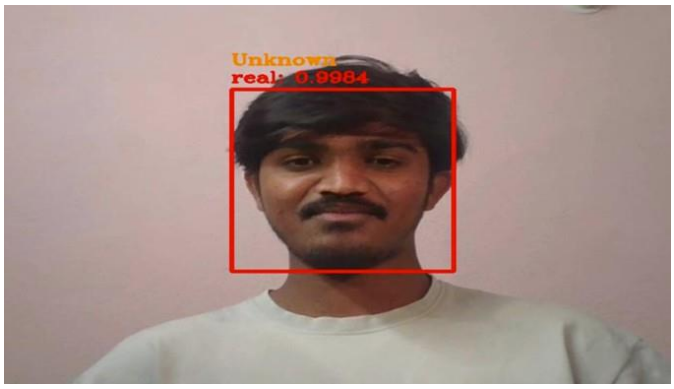

Fig. 4. Real time liveness detection of a genuine face based on the proposed ViT-DINO framework.

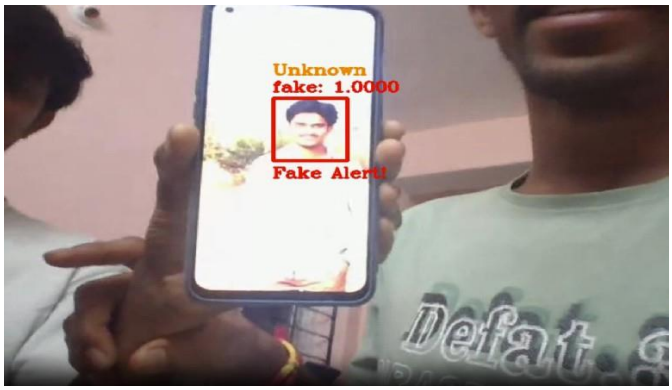

Fig. 5. Real-time detection of a spoof attack that is presented through a mobile display with fake alert notification

anism helps to boost the feature invariance and robustness. The considerable decrease in ACER, with good accuracy-to-nearly- perfect accuracy, shows that the proposed framework provides a good balance between security and usability.

Although the increased computational resources required by transformer-based architectures in comparison with lightweight CNN models, the achieved performance gains make them worth deploying in high security biometric sys- tems. Future optimization strategies can focus on model com- pression and efficient models deployment to further improve the practical scalability.

V. CONCLUSION

This paper introduced a transformer-based anti-spoofing method for the face image that combines Vision Transformer backbone and DINO-based self-supervised distillation for robust face liveness detection. The proposed approach takes advantage of global self-attention to model the distributed spoof artifacts with improved feature invariance by student- teacher representation learning. Experimental results show that the ViT-DINO model is much better than conventional CNN- based and hybrid models, achieving an accuracy of 99.8% and a

low ACER of 0.8%. Furthermore, real-time experiments at deployment verify the effectiveness of the model in the

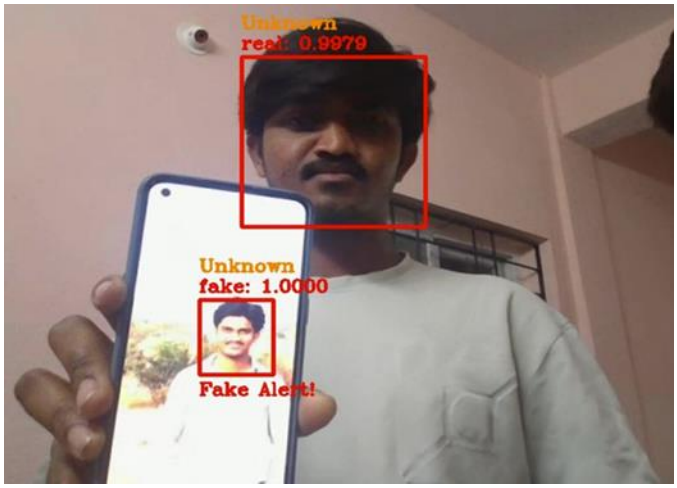

Fig. 6. Real-time identification of both live and spoof faces in one evaluation scenario

practical conditions for (accurately) distinguishing bona fide from spoof presentations. Overall, the results show that the combination of the transformer architectures and the self-supervised learning approach offers a very reliable and scal- able approach to presentation attack detection in modern biometric authentication systems.

REFERENCES

[1] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representa- tions (ICLR)*, 2021.

[2] M. Caron et al., "Emerging properties in self-supervised vision trans- formers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 9650– 9660, 2021.

[3] Z. Yu, Y. Qin, X. Li, C. Zhao, Z. Lei, and G. Zhao, "Deep learning for face anti-spoofing: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 3997–4024, Nov. 2021.

[4] Y. Tian, Y. Huang, K. Zhang, Y. Liu, and Z. Sun, "Polarized image trans- lation from nonpolarized cameras for multimodal face anti-spoofing," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 1324–1337, 2023.

[5] S. Lim, Y. Gwak, W. Kim, J.-H. Roh, and S. Cho, "One-class learning method based on live correlation loss for face anti-spoofing," *IEEE Access*, vol. 8, pp. 151239–151248, 2020.

[6] W. Liu, H. Wang, Y. Zhao, and S. Wang, "Data-fusion-based two-stage cascade framework for multimodality face anti-spoofing," *IEEE Trans. Cogn. Dev. Syst.*, vol. 14, no. 3, pp. 742–754, 2022.

[7] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and

depth-based CNNs," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, pp. 319–328, 2017.

[8] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti- spoofing: Binary or auxiliary supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 389–398, 2018.

[9] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Proc. IEEE Int. Conf. Biometrics*, pp. 1–7, 2012.

[10] ISO/IEC JTC 1/SC 37, "ISO/IEC 30107-3: Information technology — Biometric presentation attack detection — Part 3: Testing and reporting," International Organization for Standardization, 2017.

[11] H. Touvron et al., "Training data-efficient image transformers & distil- lation through attention," in *Proc. Int. Conf. Machine Learning (ICML)*, pp. 10347–10357, 2021.

[12] K. He, X. Chen, S. Xie, Y. Li, P. Dolla´r, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 16000–16009, 2022.

[13] Z. Wang, C. Zhao, Y. Qin, Q. Zhou, Z. Lei, and S. Z. Li, "Multi- domain learning for face anti-spoofing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 919–928, 2019.

[14] Z. Yu, X. Li, J. Shi, Z. Xia, and G. Zhao, "Searching central difference convolutional networks for face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 5295–5305, 2020.

[15] Y. Liu, J. Stehouwer, and X. Liu, "Deep tree learning for zero-shot face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4680–4689, 2019.

[16] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 9729–9738, 2020.

[17] J.-B. Grill et al., "Bootstrap your own latent: A new approach to self- supervised learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.