2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

GestSpoof: Gesture Based Spatio-Temporal Representation Learning For Robust Fingerprint Presentation Attack Detection.

Bhavin Jawade, Shreeram Subramanya, Atharv Dabhade, Srirangaraj Setlur, Venu Govindaraju
{bjawade, sgudemar, atharvda, setlur, govind}@buffalo.edu
University At Buffalo

***Abstract*— Fingerprint spoof attacks represent one of the most prevalent forms of biometric presentation attacks. While significant progress has been made in framing fingerprint spoof detection as a general image classification problem, limited attention has been given to treating it as a temporal learning problem. The distinctions in the elastic properties between authentic and synthetically created counterfeit fingerprints can be more accurately captured under motion-induced gestures during acquisition. In this study, we introduce a novel method for detecting fake fingerprints by deliberately introducing distortions through sliding and twisting motions during acquisition. As widely used spoof datasets such as those from LivDet 2009 to 2021 or MSU FPAD lack the temporal information essential for this investigation, we assembled a new dataset focused on distortion-based fake and real fingerprints, encompassing various types of spoof materials and diverse distortions. This gesture-equipped dataset comprises more than 3680 videos gathered from 184 unique fingers. Additionally, we present a novel spatial-temporal multi-modal network for detecting fingerprint spoofs using intentional-distortion. Our proposed approach yields significantly improved results compared to traditional static classification-based methods for spoof detection, across various metrics and for both known and unknown (generalization) scenarios, thereby highlighting the substantial impact that introducing gestures can have on enhancing fingerprint spoof detection. The dataset can be downloaded from here: https://www.buffalo.edu/cubs/research/datasets/gestspoof-dataset.html**

I. Introduction

Fingerprint recognition systems are one of the oldest and most widely used person identification systems due to the uniqueness, and permanence of fingerprints as well as their ease of enrollment, and authentication. Over the years, there have been significant contributions to improving fingerprint verification performance [11], [17], [16], but improving fingerprint presentation attack detection continues to be a significant challenge. [24] highlights several vulnerabilities faced by automated fingerprint recognition systems. A fingerprint consists of a flowing pattern of ridges and valleys and its properties can be mimicked by fabricating finger-like artifacts (*e.g.* gummy fingers) to generate fake fingerprints. Failure to detect these fake fingerprints has been demonstrated as a key limitation of existing real-world biometric systems. Recently, in July 2022, [4] a group of seven individuals cloned 2,500 fingerprints on butter paper using a polymer sheet and controlled heating with a specific chemical to deceive biometric systems for accessing bank accounts via the Aadhar Enabled Payment System (AEPS). In April 2019, a Galaxy S10 user successfully spoofed their smartphone's ultrasonic in-display fingerprint sensor using a 3D printed fingerprint. All these attacks underscore the susceptibility of real-world systems to fingerprint presentation vulnerabilities.

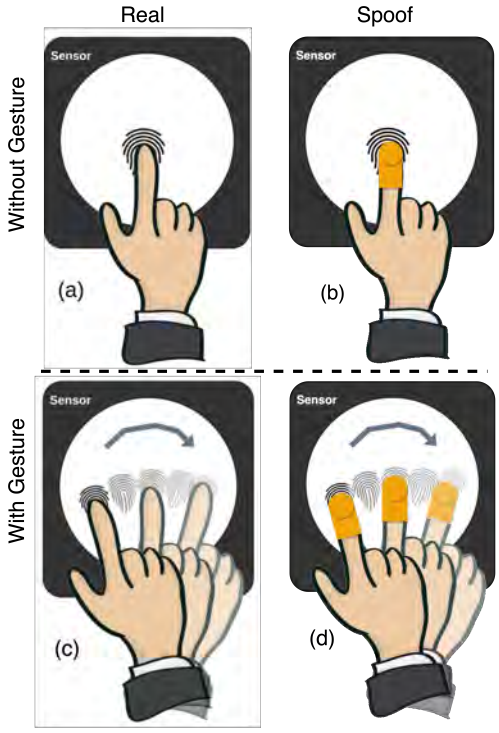

Fig. 1: (a) and (b) show existing fingerprint spoof detection approaches based static image classification. (c) and (d) show our proposed approach that utilizes the intentional distortion induced in the form of gestures to amplify elastic differences.

Fingerprints can be physically replicated using a variety of molding and casting techniques. Typically, a negative of a finger is employed to fabricate molds, creating finger-like artifacts aimed at identity subversion. Common materials such as gelatin, wood glue, and silicone are used to produce these molds, while casts like body double, Aljasafe, or dental mold serve as casting agents. A more intricate spoofing method involves capturing an image of a finger and 3D-printing high-quality replicas to deceive identification systems.

Various studies have implemented hardware and software-based solutions to improve fingerprint presentation attack detection. Between 2009 and 2019, the International Fin-

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

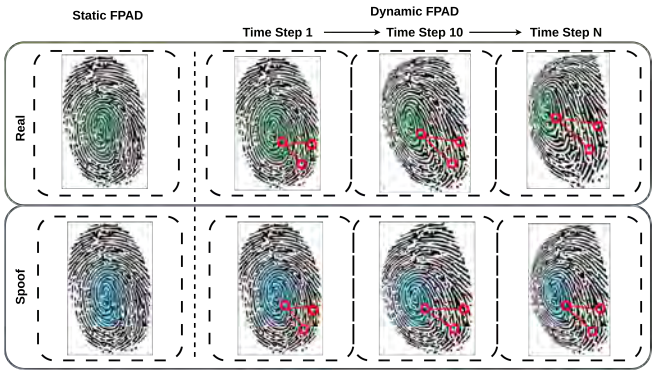

Fig. 2: An illustration to demonstrate the role of intentional motion in amplifying elastic properties. On the left we show two static fingerprints - real and spoof. On the right, we show how the fingerprint distorts during the motion. This distortion is different for spoof and real since they have different coefficient of friction. Regions in green and blue highlight the relatively displacement of central region of the fingerprint with respect to outer region caused by the friction during motion. It can also be observed that the relative position of minutia (red) changes during motion.

gerprint Liveness Detection Competition (LivDet) introduced numerous datasets. While hardware-based approaches ([24]) capture liveness based on temperature, conductivity of the material, and spectrography properties, software-based solutions ([1], [23]) extract morphological, physiological, and texture-based features from a fingerprint image or sequence of images that are trained on various algorithms such as SVMs and CNNs to distinguish the liveness characteristics. Chugh et.al [8] employed a jointly learned CNN-LSTM model, capturing the spatio-temporal dynamics from a 10-frame color sequence shot at 8fps during finger presentation to the sensor.

Though there are several datasets proposing spoof detection methods, there has been limited work done in proposing fingerprint PAD as a temporal learning problem. Current temporal/dynamic FPAD datasets [8], [18], [2] do not incorporate intentional distortion and are only capable of capturing minute changes in ridge-valley patterns, introduced due to physiological factors, such as variations in perspiration levels and small texture changes due to flattening of skin under pressure. Although these datasets and associated algorithms capture some material properties, they might not be significant enough to differentiate spoofs at scale. In this work, we present a *proof-of-concept*, to study potential of incorporating intentional gesture (refer figure 1) as a means to capture significant and amplified elastic differences for fingerprint presentation attack detection. The effectiveness of our approach stems from the fact that different materials have different coefficient of friction against the capture surface which leads to different distortions (hereby referred to as spatial-temporal properties) while in motion. These friction induced elastic difference (refer figure 2) can only be captured when a significant translation/twisting gesture is performed as in our approach and are not captured in existing dynamic FPAD datasets since they do not encompass significant motion for friction to play a role.

The key contributions of this work are summarized below:

1) We propose a novel approach towards fingerprint presentation attack detection based on gestures acquired through intentional distortions (motion) during acquisition.
2) As part of the study, we release a gesture based spoof and real fingerprint dataset - "GestSpoof" with different kinds of spoof materials and motions.
3) We also present baseline static image based evaluation results that only utilize spatial features, along with dynamic video based evaluation results that utilize both spatial and temporal features to distinguish real and fake fingerprint.
4) We also propose a novel multi-modal architecture that utilizes spatio-temporal ridge features along with spatio-temporal minutiae features, and show that proposed architecture out-performs the baseline methods.
5) Through this study we demonstrate that additional temporal information captured through the intentional-distortion induced during acquisition, aids in improving presentation attack detection substantially with respect to just performing static fingerprint acquisition.

II. Related Works

Over the past few years, fingerprint-based biometric recognition systems have achieved a high degree of accuracy. However, due to their widespread use, there's growing concern about their vulnerability to recognizing fake fingerprints. These finger-like artifacts are created using various materials to replicate the ridge-valley structure of a live finger. Each material introduces varying degrees of distortion in an attempt to mimic real fingerprints.

Existing spoof datasets: In recent years, there have been significant efforts to improve Fingerprint Presentation Attack Detection (FPAD), leading to the release of a few publicly available datasets. One of the early endeavors to establish such a spoofing database was the [12] ATVS-FFp DB. This dataset includes data from 17 users, representing a total of 68 unique identities. For every genuine finger in this dataset, two counterfeit fingers were produced.

The International Fingerprint Liveness Detection Competition (LivDet), has been held since 2009, to address the problem of presentation attack detection and assess the performance by benchmarking various SOTA fingerprint presentation attack detection (FPAD) algorithms. Throughout its seven editions, the competition has generated over twenty-five datasets using fourteen different scanners. Numerous materials have been tested with various mold types to fabricate fingerprint replicas, with the intent of producing successful spoofs to mimic the ridge valley patterns. The inaugural edition featured three optical sensors, and the fingerprints were fabricated using gelatin, silicone, and play-doh as spoof materials through a consensual method. [30] LivDet-11 introduced four datasets, with fingerprints produced through both consensual and non-consensual methods. Each iteration

Dataset	# Fingers	# Real	# Fake	Temporal	Gesture	Materials Used
LivDet 2009 [25]	254	5500	5500	✗	✗	Gelatine, Silicone and Play-Doh
LivDet 2011 [30]	200	3000	3000	✗	✗	Gelatine, Silgum, Ecoflex ...
LivDet 2013 [13]	225	8000	8000	✗	✗	Gelatine, Modasil, Ecoflex ...
LivDet 2015 [26]	100	4500	5948	✗	✗	Body Double, EcoFlex, Wood Glue ...
LivDet 2017 [27]	150	8099	9685	✗	✗	Body Double, Liquid Ecoflex , Body Double ...
LivDet 2019 [28]	-	6029	6936	✗	✗	Gelatine, Wood Glue, Latex ...
LivDet 2021 [5]	66	10700	11740	✗	✗	GLS20, Body Double, Mix 1 ...
PB Spoof-Kit [7]	-	1000	900	✗	✗	Crayola, Wood glue, 2D print ...
MSU-FPAD [7]	-	9000	10500	✗	✗	2D Print-Matte Paper, 2D Print (Transparency) ...
ATVS-FFp [12]	68	816	816	✗	✗	Silicone, Play-Doh
Tsinghua [18]	60	300	470	✓	✗	Silicone
BSL [2]	90	900	400	✓	✗	Silicone, gelatin, latex, wood glue
T. Chugh et.al [8]	685	26650	32910	✓	✗	Ecoflex, Crayola Model Magic, Dragon Skin ...
GestSpoof	184	920 (videos) 132466 (frames)	2760 (videos) 478194 (frames)	✓	✓	Body Double, EcoFlex, Gelatine

TABLE I: Comparison against existing datasets (Statistics of existing datasets from [19])

of the competition has experimented with different types of spoof materials and sensors to produce and enroll spoof fingerprints, aiming to test the efficacy of PAD algorithms. The 2021 edition of LivDet involved two scanners, GreenBit and Dermalog, creating two datasets using each sensor. [23] proposed a method combining temporal features from perspiration and morphology for liveness detection. They optimized multiple features through a feature selection technique for specific sensors, employing conventional classification methods. The [7] MSU-FPAD dataset was introduced using two distinct fingerprint readers, Guardian 200 and Lumidigm Venus 302, and comprised 9,000 live samples alongside 10,500 fake samples. Meanwhile, the [7]Precise Biometrics Spoof-Kit (PBSKD) utilized 10 spoof materials for fabrication, amassing 1,000 live samples and 900 fake samples. Chugh et.al [8] developed a dataset comprising 26,650 live frames from 685 individuals, employing three different types of spoof materials, all captured without any motion. Table I shows a detailed comparison of existing datasets against our proposed GestSpoof dataset. As it can be observed that even though a few recent datasets incorporated temporal features, GestSpoof is the first dataset which includes intentional motion or gesture.

Static Approaches For Spoof Detection: [7] tackled the challenge of creating generalized algorithms for spoof fingerprint detection. Utilizing the public domain LivDet datasets from 2011, 2013, and 2015, they extracted local patches based on minutiae location and orientation, subsequently training MobilNet CNN models aligned with these patches. [15] introduced an automated spoof detector that leverages an ensemble of One-class SVMs. This was designed to distinguish between live and fake fingerprints on the LivDet 2011 dataset, specifically to address class imbalances. Such imbalances often occur due to the presence of fewer fake samples during training in comparison to live samples. Moreover, this method aims to overcome the challenge of identifying previously unseen fake fingerprints, an issue frequently found in conventional machine-learning classifiers. [9] showcased a GAN-based strategy, intending to produce high-quality plain fingerprints, both live and fake. This approach emulates a genuine fingerprint database, addressing the shortage of publicly available data. T. Chugh [8] suggested a style-transfer-based wrapper. Its purpose is to boost the generalization performance of spoof detectors against novel fabricated fake fingerprints, unseen during training. This enhancement was achieved by incorporating synthetic live fingerprints and crafted spoof samples, complemented by the LivDet 2017 dataset. Table I compares existing datasets against "GestSpoof".

Dynamic Approaches For Spoof Detection: Limited research [29], [22], has explored the integration of temporal information for spoof detection. [1] employed the dynamics of imaging on a touch-based fingerprint reader, leveraging characteristics such as perspiration and skin distortion to distinguish between live and fake fingerprints. In a bid to enhance liveness detection, [23] proposed an algorithm that utilizes a blend of dynamic perspiration features—sourced from analyzing multiple frames of the same finger—and static morphology-based features. These static features are extracted either from a single finger impression or the differences between finger impressions, drawing from the LivDet dataset. [2] presented an approach that hinges on skin elastic properties. This method extracts skin distortion data to discern between live and fake fingerprints. For this study, a database was compiled, comprising the thumbs and forefingers of 45 volunteers, as well as 40 fabricated fake fingers spanning 10 image sequences. The skin distortion data is culled from static features like temperature, impedance, and spectroscopy. Additionally, dynamic attributes are employed to determine a distortion code, calculated via optical flow. This code, coupled with a temporal-distortion map, is instrumental in assessing fingerprint liveness.

In prior studies, there have been efforts to integrate temporal data by observing perspiration and skin distortions. Notably, the videos used in these works often display static frames with minimal motion. In contrast to these, our proposal introduces a unique approach involving deliberate motion-based gestures like sliding, dragging, and twisting. By doing so, we aim to enhance the detectable elastic distortions within the skin, providing more effective spatio-

temporal features for refined spoof detection.

III. Gesture Based Dataset

GestSpoof stands apart from existing datasets in its unique utilization of intentional distortion as a key element in its data collection process. Participants were tasked with deliberately introducing distortions through a series of gestures, including twisting and sliding motions in various directions. Due to the friction between the sensor surface and the fingertip of an individual, the unique behavior of skin elasticity is accentuated. Thus our approach to capture intentional distortion serves as a powerful magnifier, as it amplifies the elastic properties of the skin and the spoof materials. Consequently, GestSpoof provides this additional modality of temporal information, enabling a more nuanced and robust detection of spoofed fingerprints.

A. Collection Methodology

GestSpoof dataset consists of 184 unique fingers from 23 participants. For all 184 subjects, we fabricated high-quality fingerprint spoofs using three distinct spoof and cast material combinations. These include: "Body Double" spoofs created using Alja Safe [1] molds, "Ecoflex 50" spoofs created using Alja Safe molds, and "Gelatin Spoofs" created using Body Double molds. These spoof materials have been widely referenced in literature [7] [28] [27] [26] [13] [30] [25] since they are easy to use, can capture impressions, and are challenging to detect. We also utilized different molding materials since moisture retention within the mold material also determines the hardness of the spoof, thereby affecting the elastic properties. We create spoof fingerprints using silicone polymers EcoFlex and Body Double, along with protein-based Gelatine. Body Double in our study was utilizied as both a casting material and a spoof material. On the other hand, Alja Safe was only used to create casts. To address curing inhibition issues between EcoFlex and Body Double (which typically bond together and are difficult to separate), we use Alja-safe as cast, which is a a crystalline silica-free seaweed-based powder. These materials have different hardness properties based on their Shore score levels, resulting in different elastic behaviours when dragged or moved on the sensor surface. Body Double being a fast set polymer captures prints quickly but lacks skin-like movement, while EcoFlex is stretchable and behaves skin-like under motion. Gelatin, derived from animal tissue, replicates human skin's chemical properties, resulting in similar distortion patterns. Overall as part of the study we created 552 different fingerprint spoofs for 184 unique fingers.

GestSpoof was collected by recording fingerprint acquisition videos for various kinds of intentional motions. Our dataset consists of 5 different kinds of motions including horizontal sliding, vertical sliding, two diagonal sliding and twisting (refer figure 3). Overall our dataset consists of 3680 videos out of which 920 are live finger videos and 2760 are fake finger videos. All fingerprint videos were acquired

[1]Alja Safe - https://sculpturesupply.com/products/alja-safe

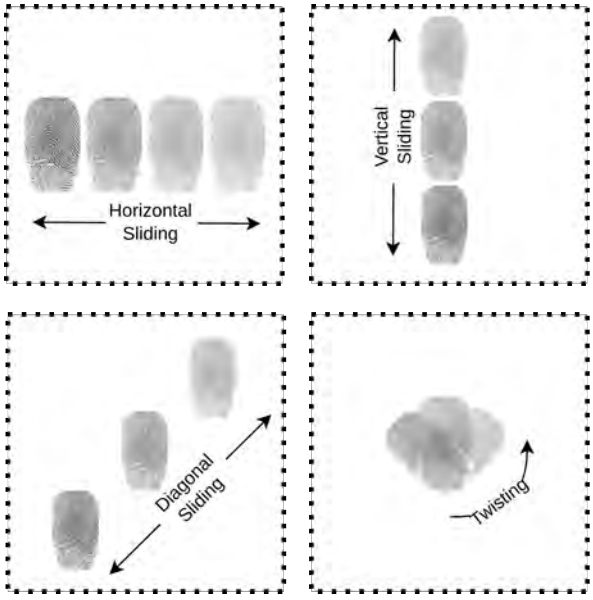

Fig. 3: Different types of gestures incorporated in GestSpoof.

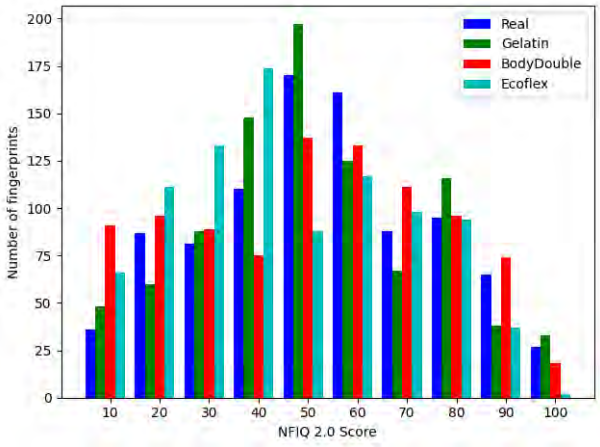

Fig. 4: NFIQ 2.0 score distribution of grayscale and enhanced real and spoof fingerprints in GestSpoof.

using a *Futronic FS64 EBTS* Flatbed sensor at frame rate of 30 FPS. The video duration vary between 5-20 seconds, amounting to approximately 926,591 frames in the entire dataset.

We have divided the dataset into disjoint testing and training sets. The training set features 136 unique fingers, encompassing 2,720 videos, while the testing set includes 48 unique fingers with 960 videos. Figure 4 shows the NFIQ 2.0 score distribution of real, gelatin, bodydouble and ecoflex fingerprints for the highest NFIQ score frame. NFIQ 2.0 scores range from 0 - 100. In the figure 4 we presented scores binned in bin size of 10.

IV. Proposed Method

A. Architecture

In this section, we describe the baseline spatio-temporal video classification approach for our dataset. Given a real

Property	Count
Number of Subjects	23
Number of Unique Fingers	184
Number of Videos Real	920
Number of Videos Fake	2760
Number of Spoof Materials	3
Number of Intentional Motions	5
Number of Frames	610660

TABLE II: Summary of dataset statistics.

Fig. 5: Sample frames from GestSpoof dataset for different types of spoof materials

or fake video V, we extract the frames from each video. Let $V = [F^1, F^2, F^3...]$ be the frames in a video. In each frame, the fingerprint is present at a certain location. To get fine-grained temporal information about elastic changes in the skin and to ignore large spatial motions occurring in a frame, we first detect the fingerprint region. We start by binarizing each frame using adaptive gaussian thresholding followed by a morphological opening operation to remove noise from the binarized frame. We then detect contours in the binarized frames and select the largest contour as the fingerprint region.

By extracting the fingerprint from a larger frame and stacking the frames together to form a video, we are able to suppress large fingerprint motions within the capture region and only focus on fine-grained differences (distortions) of the ridges and valleys across the frames. We refer to these cropped videos as camera-tracking fingerprint videos V_c since they represent the scenario when there is no relative motion between the acquisition camera and the fingerprint. Fingerprint extracted from different frames can be of different dimensions therefore before stacking them together we pad the frames to be of the same dimension (H, W).

For every video V_c we also extract minutiae points for each frame ${F_c}^i$ using minDTCT minutiae detector. Detected minutiae points are plotted on a white image with the same dimensions as the cropped frame with red circle depicting the minutiae location and a small line depicting the minutiae direction. By stacking minutiae frames together in the temporal dimension, we create a minutiae relative motion video denoted by $M_c = [m_c^1, m_c^2, m_c^3...]$. Similar to V_c, we pad each minutiae frame m_c^i to be of the same dimension (H, W).

We adopt a video transformers based approach for our spatio-temporal baseline. Given a video V_c, we pass each frame F_c^i through a feature extractor θ_e to compute the frame level features of dimension (B, D), where B is the batch size and D is the feature vector length. We stack the frame level features temporally to get a video level feature vector of dimension (B, T, D), where T is the temporal dimension or the number of frames. Next, we pass the temporally stacked feature map through a transformer encoder block θ_t. The feature maps are appended with temporal positional encoding. This transformer block performs temporal self-attention across frame level features. We denoted the resultant video representation as $x_v \in \mathbb{R}^{B,T,D}$. We perform the same feature extraction for the minutiae videos, to a get minutiae video representation of dimension $x_m \in \mathbb{R}^{B,T,D}$.

To perform a temporal feature fusion of minutiae and fingerprint frame representations we use multi-headed cross-attention between two feature maps. We pass the concatenated fingerprint and minutiae feature vectors through the cross-attention block denoted by θ_c. The output dimension of cross-attended feature vector is $(B, 2 \text{ x } T, D)$.

To compute the fused embedding, we take the mean in the temporal dimension to get one fused vector of dimension (B, D). We use a dense fully connected layer to get logits for two binary classes (Spoof and Real).

B. Optimization

Given the high class imbalance in the training set, with spoof fingerprints having nearly three times the number of samples of real fingerprints, we optimize focal loss instead of the traditional cross-entropy loss.

$$\text{FL}(p_t) = -(1 - p_t)^{\gamma} \cdot \log(p_t)$$

where: $p_t p$ is the predicted probability of the true class. γ is a tunable focusing parameter. When $\gamma = 0$, it becomes the standard cross-entropy loss.

Focal loss applies a $(1 - p_t)^{\gamma}$ factor to the conventional cross-entropy loss, in order to focus learning on hard misclassified samples where p_t is the probability of the class. Here γ is the focusing parameter that adjusts the rate at which easy samples are down weighted. We also apply a weighted reduction to compute the average loss for the two classes.

V. Experiments

In this section, we will describe the experimental setting and results obtained for the proposed method.

A. Implementation Details

In our spatio-temporal architecture we utilize 30 frames per video for both fingerprints and minutiae plots. The frame dimensions H, W are (224, 244). The feature extractor θ_e that we use in the proposed approach is an ImageNet

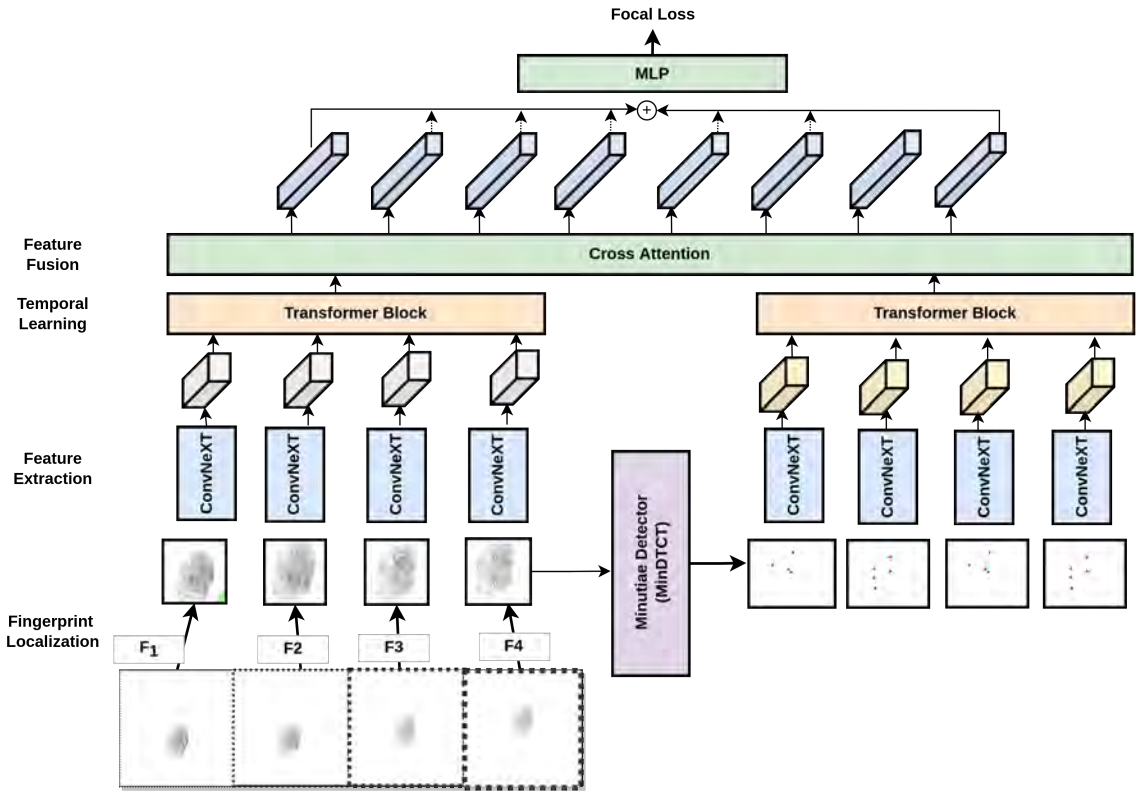

Fig. 6: Architecture of our spatio-temporal network

pretrained ConvNext model. The self-attention transformer block θ_t consists of 2 encoder layers with 2 heads. The multi-head cross-attention θ_c block consists of 1 encoder layer with 4 heads. The batch size B utilized during training is 6 and the dimension of the intermediate feature vector D is 768. The dense fully connected layer consists of 2 linear transformations with ReLU non-linearity and dropout of 0.2.

B. Evaluation Metrics

We present performance of our method using four evaluation metrics, F1 Score, Accuracy, True accept rate at a given False accept rate (TAR@FAR) and Attack Presentation Classification Error Rate at a given Bona Fide Presentation Classification Error Rate (APCER@BPCER). Given the class imbalance between the real and the spoof classes, F1 score is a better comparison criteria then accuracy. For PAD systems, APCER@BPCER is a common evaluation metric. APCER represents the error rate when an attack presentation is incorrectly classified as a bona fide presentation, while BPCER denotes the error rate when a bona fide presentation is incorrectly classified as an attack presentation. These metrics are computed as:

$$\text{APCER} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

$$\text{BPCER} = \frac{\text{FN}}{\text{FN} + \text{TP}}$$

We present the TAR at FAR of 1% and 0.2% and APCER at BPCER of 1% and 0.2% as has been the practice in this domain.

C. Static Image Based Baselines

To evaluate static image based approaches on our proposed dataset, we explore several architectures, starting with ResNet's [14], a widely recognized deep residual network. All ResNet based models benchmarked in this study were pre-trained on ImageNet and then finetuned on GestSpoof. We considered three varinats: ResNet-18, ResNet-34 and ResNet-50, with the major distinguishing factor being the number of parameters, which increases as the depth of the network grows. Following that we examined Vision Transformer (ViT) [10]. Unlike conventional convolutional architectures, ViT leverages the transformer mechanism, allowing for global attention across the entire image, potentially capturing long-range dependencies better. We then explored Swin Transformer V2 [20], This newer iteration of the transformer-based model introduces a hierarchical structure and shifted windows to further enhance both local and global attention capabilities. Lastly, we evaluated Cross ViT[6], a hybrid architecture that integrates the strengths of both ResNet and ViT. By fusing the spatial hierarchies of ResNet with global attention mechanism of ViT.

D. Dynamic Video Based Baselines

For the spatio-temporal evaluation on our proposed dataset, we present the baseline results with TimeSformer [3] base

Method	F1 Score	Accuracy	TAR@FAR=1%	TAR@FAR=0.2%	APCER@BPCER=1%	APCER@BPCER=0.2%
Static Image Based Approaches (10 Frame)						
ResNet 18 [14]	66.49%	64.41%	49.37%	48.54%	75.83%	87.50%
ResNet 34 [14]	66.69%	64.62%	49.23%	48.12%	74.16%	86.25%
ResNet 50 [14]	65.78%	63.68%	49.51%	48.26%	84.58%	94.16%
ViT Base [10]	66.30%	64.09%	47.98%	47.15%	67.08%	76.25%
CrossViT [6]	65.35%	63.16%	48.68%	47.01%	82.91%	95.00%
SwinV2 [20]	65.83%	63.89%	51.17%	48.95%	67.08%	71.25%
Static Image Based Approaches (1 Frame)						
ResNet 18 [14]	71.24%	69.30%	45.90%	45.21%	98.54%	99.58%
ResNet 34 [14]	71.47%	69.51%	48.12%	28.98%	98.12%	99.58%
ResNet 50 [14]	72.24%	70.34%	47.15%	47.01%	93.75%	95.83%
ViT Base [10]	73.39%	71.69%	42.09%	34.11%	89.16%	94.58%
CrossViT [6]	71.38%	69.40%	47.98%	23.09%	100.00%	98.75%
SwinV2 [20]	69.78%	67.63%	48.68%	48.26%	80.83%	86.66%
Dynamic Video Based Approaches						
TimesFormer	75.44%	77.12%	49.88%	48.54%	72.48%	81.91%
VideoModel	83.84%	85.10%	50.83%	49.44%	61.48%	72.91%

TABLE III: Overall Comparison of Spoof Detection Performance of Static Image Methods against Dynamic Spatio-Temporal Methods

Method	F1 Score	Accuracy	TAR@FAR=1%	TAR@FAR=0.2%	APCER@BPCER=1%	APCER@BPCER=0.2%
Static Image Based Approaches (10 Frame)						
Bodydouble						
ResNet 50 [14]	70.35%	71.46%	50.00%	50.00%	80.42%	90.83%
ViT Base [10]	69.67%	70.62%	50.00%	47.92%	61.67%	73.75%
Ecoflex						
ResNet 50 [14]	72.21%	73.39%	49.79%	48.13%	66.25%	66.25%
ViT Base [10]	74,42%	75.05%	48.54%	45.64%	64.17%	68.75%
Gelatin						
ResNet 50 [14]	71.55%	72.71%	49.17%	47.50%	85.41%	88.75%
ViT Base [10]	70.70%	71.45%	47.50%	45.41%	76.67%	77.50%
Static Image Based Approaches (1 Frame)						
Bodydouble						
ResNet 50 [14]	73.91%	73.96%	50.42%	47.50%	93.75%	96.67%
ViT Base [10]	72.87%	72.92%	47.50%	40.00%	89.58%	92.08%
Ecoflex						
ResNet 50 [14]	73.60%	73.60%	45.64%	36.51%	90.83%	96.67%
ViT Base [10]	71.05%	71.10%	39.83%	12.44%	90.00%	90.00%
Gelatin						
ResNet 50 [14]	67.29%	67.29%	47.92%	47.92%	95.41%	98.12%
ViT Base [10]	70.20%	70.20%	38.75%	32.08%	84.16%	96.25%
Dynamic Video Based Approaches						
Bodydouble						
VideoModel	77.90%	77.90%	57.25%	49.44%	62.76%	71.93%
Ecoflex						
VideoModel	75.11 %	74.86 %	54.38 %	50.00 %	69.38%	78.45%
Gelatin						
VideoModel	77.81	75.79	56.38	49.44	64.25%	74.21 %

TABLE IV: Segregated performance on each spoof type. Above we present results after training on all spoof and evaluating for each spoof type separately.

pretrained on K400 dataset and finetuned on GestSpoof. We also evaluate Time distributed Convnext [21] (Small) [2] finetuned on GestSpoof training set (referred to as VideoModel (Ridge)). Finally, we present results with our proposed architecture with cross attention based fusion of Ridge and Minutiae spatio-temporal features.

[2]Pretrained Models from Video Transformers Pytorch: https://github.com/fcakyon/video-transformers

E. Results

In table III we present overall performance of static image-based and dynamic video-based approaches for spoof detection. For static image based methods we present single frame results along with 10 frame score fusion results. This is performed to ensure that improvement in performance using intentional motion is not merely because of use of more frames. It can be observed from table III, as expected

Method	F1 Score	TAR@FAR=1%	APCER@BPCER=1%
Static Image Based Approaches (10 Frame)			
Hold out spoof - Bodydouble			
ResNet 50 [14]	66.24%	49.93%	86.25%
ViT Base [10]	73.68%	42.99%	73.75%
Hold out spoof - Ecoflex			
ResNet 50 [14]	65.22%	50.21%	77.50%
ViT Base [10]	61.63%	46.88%	**56.67%**
Hold out spoof - Gelatin			
ResNet 50 [14]	64.88%	49.79%	88.75%
ViT Base [10]	67.94%	46.19%	70.83%
Static Image Based Approaches (1 Frame)			
Hold out spoof - Bodydouble			
ResNet 50 [14]	71.93%	47.57%	90.83%
ViT Base [10]	73.19%	25.24%	96.67%
Hold out spoof - Ecoflex			
ResNet 50 [14]	70.50%	46.88%	97.50%
ViT Base [10]	66.07%	43.34%	79.17%
Hold out spoof - Gelatin			
ResNet 50 [14]	71.18%	47.16%	86.67%
ViT Base [10]	73.17%	43.41%	86.67%
Dynamic Video Based Approaches			
Hold out spoof - Bodydouble			
VideoModel	**79.23%**	**54.29%**	**61.50%**
Hold out spoof - Ecoflex			
VideoModel	**74.67%**	**51.46%**	71.23%
Hold out spoof - Gelatin			
VideoModel	**78.85%**	**52.78%**	**67.11%**

TABLE V: Hold out Performance - Here we show the performance of the models on unseen (unknown) spoof types, where the model is trained on two spoof types and evaluated on all three.

Ablation Setting		F1 Score	TAR@FAR	APCER@BPCER
Ridge	Min			
✓		79.34%	49.76%	72.80%
✓	✓	83.34%	50.83%	61.48%

TABLE VI: Abalation Study Table

10 frame score aggregated results for static image-based baselines are significantly better than the single frame results. Our proposed video-based approach that utilizes both ridge and minutiae spatio-temporal information achieves 83.84 % F1 score which is nearly 10 % higher than the best image-based static method, substantiating the contribution of temporal features for spoof detection. Though these results demonstrate substantial growth when incorporating temporal features, there is high potential for improvement for future spoof detection works that utilize Gestspoof dataset.

VI. Discussion

Segregated Analysis: In table IV we present segregated results for each spoof type. In this case, real and spoof samples per evaluation are equal (920). Among, segregated results, it can be observed that Ecoflex has lower F1 score than Body Double and Gelatin with dynamic approaches, which shows that ecoflex mimics the motion properties of real-finger relatively more closely than other materials used in the study.

Holdout-set Analysis: In table V we present holdout results, where we evaluate performance of methods on Gestspoof when one spoof type is unknown or unseen. In this case, we train the method on two spoof types and evaluate it on the third spoof type. It can be observed that even in the holdout setting where one spoof type is unseen, our proposed video-based model performs significantly better than static image-based models, emphasizing on the role intentional distortion can play in improving FPAD.

Abalation Study: In table VI we present abalation experiments to study contribution of different branches of the network to the overall performance. As it can be observed ridge features alone give 79.34% in F1 Score and 72.80% in APCER@BPCER=1%. Incorporation of minutia features improve the F1 score by 4% and decrease the APCER error rate by 10%. This demonstrates that minutia drift features can significantly incorporate discriminative fine-grained temporal information for spoof detection over just ridge features.

VII. Conclusions

In this work, we have proposed a novel approach to fingerprint spoof detection based on gesture induced elastic distortions. We collect and release first of its kind motion based fingerprint spoof dataset "GestSpoof" which contains spoofs created using different types of materials along with videos captured under different types of motions. We also present baseline static image only and dynamic video-based results for spoof detection on our dataset along with a novel spatio-temporal approach that combines the ridge features with minutiae features. Future works, can utilize GestSpoof to benchmark novel approaches for fingerprint presentation attack detection using more sophisticated and fine-grained features to capture the spatio-temporal differences between the motions of real and spoof fingers. This gesture augmented fingerprint spoof detection approach can be integrated to existing smartphones without hardware changes thereby improving their security and robustness against fingerprint presentation attacks.

VIII. Acknowledgement

This work was conducted at the Center for Unified Biometrics and Sensors (CUBS) at the University at Buffalo and was supported by the Center for Identification Technology Research (CITeR) and the National Science Foundation through grant #1822190.

References

[1] A. Abhyankar and S. Schuckers. Fingerprint liveness detection using local ridge frequencies and multiresolution texture analysis techniques. In *2006 International Conference on Image Processing*, pages 321–324, 2006.

[2] A. Antonelli, R. Cappelli, D. Maio, and D. Maltoni. A new approach to fake finger detection based on skin distortion. In D. Zhang and A. K. Jain, editors, *Advances in Biometrics*, pages 221–228, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[3] G. Bertasius, H. Wang, and L. Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[4] M. Buddi. Gang clones fingerprints to hack a/cs through aadhaar-based pay. *The Times of India*, June 17 2022.

[5] R. Casula, M. Micheletto, G. Orrù, R. Delussu, S. Concas, A. Panzino, and G. L. Marcialis. Livdet 2021 fingerprint liveness detection competition - into the unknown. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–6, 2021.

[6] C.-F. R. Chen, Q. Fan, and R. Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021.

[7] T. Chugh, K. Cao, and A. K. Jain. Fingerprint spoof buster: Use of minutiae-centered patches. *IEEE Transactions on Information Forensics and Security*, 13(9):2190–2202, 2018.

[8] T. Chugh and A. K. Jain. Fingerprint spoof detection: Temporal analysis of image sequence. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2020.

[9] T. Chugh and A. K. Jain. Fingerprint spoof detector generalization. *IEEE Transactions on Information Forensics and Security*, 16:42–55, 2021.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] J. J. Engelsma, K. Cao, and A. K. Jain. Learning a fixed-length fingerprint representation. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1981–1997, 2019.

[12] J. Galbally, J. Fierrez, F. Alonso-Fernandez, and M. Martinez-Diaz. Evaluation of direct attacks to fingerprint verification systems. *Telecommunication Systems*, 47:243–254, 2011.

[13] L. Ghiani, D. Yambay, V. Mura, S. Tocco, G. L. Marcialis, F. Roli, and S. Schuckcrs. Livdet 2013 fingerprint liveness detection competition 2013. In *2013 International Conference on Biometrics (ICB)*, pages 1–6, 2013.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] I. ISO. Iec 30107-1: 2016 information technology—biometric presentation attack detection—part 1. *Information Technology Task Force (ITTF): Geneva, Switzerland*, 2016.

[16] B. Jawade, A. Agarwal, S. Setlur, and N. Ratha. Multi loss fusion for matching smartphone captured contactless finger images. In *2021 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2021.

[17] B. Jawade, D. D. Mohan, S. Setlur, N. Ratha, and V. Govindaraju. Ridgebase: A cross-sensor multi-finger contactless fingerprint dataset. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9, 2022.

[18] J. Jia, L. Cai, K. Zhang, and D. Chen. A new approach to fake finger detection based on skin elasticity analysis. In S.-W. Lee and S. Z. Li, editors, *Advances in Biometrics*, pages 309–318, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[19] H. Li and R. Ramachandra. Deep learning based fingerprint presentation attack detection: A comprehensive survey. *arXiv preprint arXiv:2305.17522*, 2023.

[20] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo. Swin transformer v2: Scaling up capacity and resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11999–12009, 2022.

[21] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[22] E. Marasco and A. Ross. A survey on antispoofing schemes for fingerprint recognition systems. *ACM Comput. Surv.*, 47(2), nov 2014.

[23] E. Marasco and C. Sansone. Combining perspiration- and morphology-based static features for fingerprint liveness detection. *Pattern Recognition Letters*, 33(9):1148–1156, 2012.

[24] S. Marcel, J. Fierrez, and N. Evans, editors. *Handbook of Biometric Anti-Spoofing*. Advances in Computer Vision and Pattern Recognition. Springer Singapore, 3 edition.

[25] G. L. Marcialis, A. Lewicke, B. Tan, P. Coli, D. Grimberg, A. Congiu, A. Tidu, F. Roli, and S. Schuckers. First international fingerprint liveness detection competition—livdet 2009. In P. Foggia, C. Sansone, and M. Vento, editors, *Image Analysis and Processing – ICIAP 2009*, pages 12–23, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[26] V. Mura, L. Ghiani, G. L. Marcialis, F. Roli, D. A. Yambay, and S. A. Schuckers. Livdet 2015 fingerprint liveness detection competition 2015. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6, 2015.

[27] V. Mura, G. Orrù, R. Casula, A. Sibiriu, G. Loi, P. Tuveri, L. Ghiani, and G. L. Marcialis. Livdet 2017 fingerprint liveness detection competition 2017. In *2018 international conference on biometrics (ICB)*, pages 297–302. IEEE, 2018.

[28] G. Orrù, R. Casula, P. Tuveri, C. Bazzoni, G. Dessalvi, M. Micheletto, L. Ghiani, and G. L. Marcialis. Livdet in action - fingerprint liveness detection competition 2019. In *2019 International Conference on Biometrics (ICB)*, pages 1–6, 2019.

[29] R. Plesh, K. Bahmani, G. Jang, D. Yambay, K. Brownlee, T. Swyka, P. Johnson, A. Ross, and S. Schuckers. Fingerprint presentation attack detection utilizing time-series, color fingerprint captures. In *2019 International Conference on Biometrics (ICB)*, pages 1–8, 2019.

[30] D. Yambay, L. Ghiani, P. Denti, G. L. Marcialis, F. Roli, and S. Schuckers. Livdet 2011 — fingerprint liveness detection competition 2011. In *2012 5th IAPR International Conference on Biometrics (ICB)*, pages 208–215, 2012.