IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 May 2025, accepted 10 June 2025, date of publication 18 June 2025, date of current version 27 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3580950

RESEARCH ARTICLE

3DDGD: 3D Deepfake Generation and Detection Using 3D Face Meshes

HICHEM FELOUAT[1,2], HUY H. NGUYEN[1,2], (Member, IEEE), JUNICHI YAMAGISHI[1,3], (Senior Member, IEEE), AND ISAO ECHIZEN[1,2,4], (Senior Member, IEEE)

[1]Informatics Program, The Graduate University for Advanced Studies, SOKENDAI, Kanagawa 240-0115, Japan
[2]Information and Society Research Division, National Institute of Informatics (NII), Tokyo 101-8430, Japan
[3]Digital Content and Media Sciences Research Division, National Institute of Informatics (NII), Tokyo 101-8430, Japan
[4]Department of Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8656, Japan

Corresponding author: Hichem Felouat (hichemfel@nii.ac.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) under KAKENHI Grants JP21H04907 and JP24H00732; by the Japan Science and Technology Agency (JST) under CREST Grants JPMJCR18A6 and JPMJCR20D3, including the AIP Challenge Program; by JST AIP Acceleration Grant JPMJCR24U3; and by JST K Program Grant JPMJKP24C2, Japan.

ABSTRACT 3D face technology is revolutionizing various fields by providing superior security and realism compared with 2D methods. In biometric authentication, 3D facial features serve as unique, hard-to-forge identifiers, improving accuracy in facial recognition for border control and criminal identification. Additionally, 3D avatars enhance virtual interactions. In this study, we aimed to strengthen 3D facial biometric systems against deepfakes. Key contributions include proving the superior protection of 3D faces over 2D ones, creating a dataset of real and fake 3D faces, and developing advanced models for accurate 3D deepfake detection. We evaluated our models for generalization to other datasets and stability when changing training data. Our experiments used the mesh multi-layer perceptron model for deepfake detection along with self-attention mechanisms and the newly introduced TabTransformer model. Results indicate that 3D face meshes greatly improve security by distinguishing real faces from deepfakes. Future work will focus on enhancing detection tools and integrating geometric features with facial textures for more accurate 3D deepfake detection. The dataset and models are publicly available on GitHub, excluding licensed elements: https://github.com/hichemfelouat/3DDGD

INDEX TERMS 3D deepfake detection, 3D deepfake generation, 3D face reconstruction, 3D biometric systems.

I. INTRODUCTION

The use of 3D facial recognition has grown rapidly in various fields as it offers several benefits over traditional 2D methods. One key application is biometric authentication, where 3D facial features provide a unique and difficult-to-copy identifier, ensuring secure access to devices and sensitive information. Additionally, 3D face scans enhance the accuracy and reliability of facial recognition, improving security measures for border control and criminal identification. In the metaverse and virtual reality, expressive 3D avatars add a human element to digital interactions, making social engagement in virtual spaces more natural and engaging.

There are several reasons for the preference for 3D face models over 2D ones. Unlike 2D models that typically reflect age, race, and skin color, 3D models reflect the geometric details of facial features, capturing every bump and curve with precision. This high level of detail enables systems using 3D models to detect even the smallest facial features, similar to how someone closely examines a sculpture. As a result, systems using 3D models often outperform the best

The associate editor coordinating the review of this manuscript and approving it for publication was Mohamad Afendee Mohamed.

2D facial recognition systems, making them much harder to trick [1], [2]. While getting a 2D face image of a person is relatively easy through social media and public records, obtaining a 3D face image is more difficult. It requires specialized scanners and software, which can be difficult to access and make forgery more expensive, adding an extra layer of security.

In this study, we aimed to strengthen 3D facial biometric authentication systems against the rising threat of 3D deepfakes. This paper focuses on protecting facial recognition and remote identity-proofing systems from 3D deepfake attacks by making three key contributions. First, it shows that 3D face images offer better protection than 2D ones because of their complexity, which makes them much harder for attackers to generate. Second, it presents a comprehensive dataset of both real and fake 3D face images. Lastly, it introduces an advanced model that enables accurate detection of deepfakes in 3D faces. Through these efforts, the study aims to enhance the security and reliability of 3D facial recognition systems in our increasingly digital world.

II. RELATED WORK

This section reviews the existing literature on 3D deepfakes. We cover techniques for 3D deepfake generation, methods for deepfake detection, key 3D face datasets, the role of deep learning in 3D face modeling, and 3D face mesh-based deep-learning approaches.

A. 3D DEEPFAKE GENERATION

3D deepfake generation involves creating realistic 3D representations of human faces using advanced computer vision and machine learning techniques. These deepfakes can be categorized into several types on the basis of the underlying processes involved:

Entire 3D Face Generation: This process involves creating a 3D face model from various inputs. The generation can be driven by inputting specific prompts that detail the desired features [3], by inputting random noises [4], or by directly manipulating features and expressions to achieve a certain appearance [5].

3D Face Reconstruction: This process converts a 2D face image into a 3D model. By using a single 2D image or multi-view 2D images, a corresponding 3D representation is reconstructed, enabling a more lifelike depiction of the individual [6], [7], [8], [9].

3D Face Animation: In this process, sound is synchronized with the 3D facial image, enabling the lips, head, and facial muscles to move naturally in response to spoken words. This technique often creates lifelike video content in which the 3D facial image mimics real human speech and expressions [10], [11].

3D Face Manipulation: This process involves altering the 3D face model by changing its facial expressions, apparent age, gender, or other attributes. Manipulation enables customization of the face to achieve specific desired outcomes, such as simulating different emotions or altering physical characteristics [12].

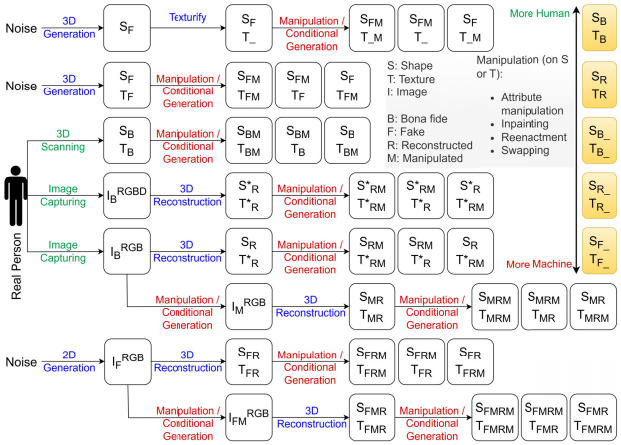


FIGURE 1. Methods for generating 3D deepfakes.

Figure 1 illustrates the main methods for generating 3D deepfakes. One approach involves generating a 3D face from random noise using a deep-learning model, resulting in a face shape or a textured face. This generated face can be manipulated through attribute manipulation, reenactment, or face swapping. Another approach starts with a direct 3D scan of a face, which can then be manipulated. Alternatively, an RGB or RGBD image can be used. In this case, 3D face reconstruction models are used to generate a 3D face using various manipulations. The RGB image can also be modified before applying the reconstruction model. Lastly, a 2D fake image can be generated from random noise, followed by the application of a 3D face reconstruction model, which may also be modified beforehand. Once the 3D face is generated through any of these methods, it can be further manipulated to achieve the desired outcome.

B. 3D DEEPFAKE DETECTION

Deepfake detection has become an increasingly important area of research due to the rapid advancements in generative artificial intelligence and the proliferation of synthetic media. While most research has focused on 2D deepfakes [13], the emergence of 3D face deepfakes introduces new challenges due to the added complexity of spatial information and the ability to manipulate facial geometry more realistically.

Detecting 3D deepfakes is similar to detecting 2D deepfakes; both involve binary classification. The key differences are the features used and the network architecture used for classification. While there has been extensive research on detecting 2D deepfakes, there is less work on detecting 3D deepfakes, as most facial recognition systems, remote identity-proofing systems, and biometric identity cards still rely on 2D images. Nevertheless, there have been several notable attempts in this area. For example, Chen et al. [14] devised a technique for detecting 3D face mask attacks by combining information from the face's texture and shape.

By fusing these dynamic features, the proposed technique enhances the accuracy of anti-spoofing measures in biometric systems. Zhu et al. [15], [16] created a method for detecting face forgeries by decomposing a face image into five components (3D geometry, common texture, identity texture, ambient light, and direct light). This method identifies inconsistencies in the 3D geometry, which helps distinguish between real and manipulated facial images. Singh and Ramachandra [17] explored the creation of 3D face morphing attacks, their effect on biometric systems, and methods to detect them. They examined how these attacks can be generated, evaluated the vulnerabilities they exploit, and presented detection strategies for countering them. This paper focuses on 3D deepfake detection using 3D face meshes. In Subsection II-E, we review various methods for mesh classification that can be adapted or serve as inspiration for 3D deepfake detection.

C. 3D FACE DATASETS

Over the past decade, the field of 3D face reconstruction has seen significant advancements, largely driven by the availability of comprehensive 3D face datasets. These datasets are essential for training and evaluating deep-learning models for tasks such as 3D face recognition, expression analysis, and facial geometry reconstruction. Existing 3D face datasets can be categorized into two groups in accordance with the devices used for acquiring the 3D face models. First, depth sensors or scanners [18], [19], [20], [21], [22], [23] provide direct 3D information although their spatial resolution is limited. Second, sparse multi-view camera systems offer higher resolution, although they can suffer from unstable or inaccurate reconstruction [4], [24], [25]. Table 1 presents a quantitative comparison of available 3D datasets, including such factors as the number of subjects, expressions, and resolution.

TABLE 1. Quantitative comparison of 3D face datasets.

	Release Year	Number of Subjects	Expressions	Resolution
BU-3DFE [18]	2006	100	25	Low
BU-4DFE [19]	2008	101	6	Low
Bosphorus [21]	2008	105	35	Medium
BJUT-3D [20]	2009	500	1–3	High
FaceWarehouse [22]	2013	150	20	Low
4DFAB [23]	2018	180	6	High
D3DFACS [24]	2011	10	38	Medium
BP4D-Spontanous [25]	2014	41	27	Medium
FaceScape [4]	2020	938	20	Very High

D. 3D FACE DEEP LEARNING

Geometric deep learning for 3D face analysis fundamentally depends on the representation of the 3D face shape. Several representation formats have been explored for this purpose, including multi-view images, voxels, point clouds, and triangle meshes. Each representation captures distinct aspects of the 3D facial geometry for the development of advanced deep-learning models. Further information can be found in the comprehensive surveys [26], [27], [28].

With the multi-view image format, multiple images are taken from different angles around an object instead of a single image. These images collectively contain more information about the object's 3D structure. This process involves rendering a 3D shape into multiple 2D views, enabling the use of a convolutional neural network (CNN) to process the resulting images directly [29], [30]. However, when dealing with complex shapes, especially those with highly curved surfaces, the limited number of views often results in not capturing critical 3D information.

With the voxel-based format, the 3D space is divided into a grid of volumetric elements (voxels), each representing a value at a point in space, similar to 3D pixels. The voxel representation facilitates the straightforward application of a 3D CNN to volumetric data by defining a 3D convolutional operator. This approach has been widely adopted for various tasks such as classification [31], semantic segmentation [32], and alignment [33]. However, the high computational and memory demands associated with volumetric data substantially limit the scalability and efficiency of this approach.

The point cloud format provides a straightforward yet dense representation of 3D shapes by capturing spatial coordinates in a simple format, making this approach increasingly popular for deep-learning applications involving 3D shapes. Methods such as PointNet [34] and PointNet++ [35] directly process raw point clouds, while PointCNN [36] leverages geometric features through convolutional operations on the points. The more recently introduced point cloud transformer framework [37] uses transformer-based attention mechanisms to enhance the processing of point clouds. Despite these advancements, point clouds have a limited ability to fully capture detailed geometric features.

With the triangle mesh format, vertices connected by edges are used to form a continuous surface. By capturing both the geometric and topological properties of the surface, triangle meshes provide richer and more comprehensive shape representations, enabling detailed and accurate visualizations.

E. 3D FACE MESH-BASED DEEP LEARNING

The field of 3D face mesh-based deep learning has advanced dramatically in recent years with the introduction of new methods using the basic structure of mesh data. A mesh typically consists of three main components: vertices, edges, and faces. They are fundamental in various deep-learning techniques for analyzing 3D face meshes. Mesh-based deep-learning networks are usually designed with two key aspects in mind: how are the geometric features to be extracted from each mesh component (vertex, edge, face) and the need for the network architecture to include key modules, such as attention mechanisms, convolutional layers, and pooling operations, that are specifically adapted for mesh data.

Recent advancements in this field have introduced various models that leverage these mesh components in innovative ways. Notable examples include the MeshWalker model [38], which processes mesh data by ''walking'' over the vertices in sequence, enabling the model to learn patterns on the basis of the order in which the vertices are visited. The MeshCNN model [39] uses a traditional CNN to work directly on mesh edges, enabling the extraction of features from the mesh's geometry. The PD-MeshNet model [40] combines point cloud data with mesh structures, leveraging the relationship between points and the mesh to improve 3D shape analysis. The SubdivNet model [41] uses subdivision techniques to refine mesh resolution, progressively enhancing detail and enabling more precise feature extraction. The HodgeNet model [42] uses mathematical concepts from Hodge theory to capture the topological and geometric properties of meshes for better feature representation. The DNF-Net model [43] introduces a dual representation of mesh normals (both vertex and face normals) to capture finer details of the mesh surface. The DiffusionNet model [44] uses a diffusion process to spread information across the mesh, capturing both local and global features for a more comprehensive analysis. The Laplacian2Mesh model [45] uses the Laplacian operator, a mathematical tool often used in mesh processing, to learn the mesh features. It helps smooth and refine the mesh while preserving important geometric details. The mesh multi-layer perceptron model (MLP) model [46] applies the MLP architecture to mesh data by directly processing the vertices and faces. It focuses on learning mesh features without the need for complex convolutional or graph-based operations.

In addition to these mesh-based models, recent transformer-based models have also shown strong performance in 3D face mesh analysis. These models leverage the attention mechanism to process the mesh data more effectively. The Graphormer model [47] applies the attention mechanism to graph-structured data, including meshes, effectively capturing complex relationships between mesh components. The Katam 3D model [48] uses a transformer architecture tailored for mesh data and learns intricate geometric patterns through the attention mechanism. The Transformesh model [49] focuses attention on crucial parts of the mesh, enhancing the accuracy of 3D face analysis. The MeshFormers model [50] combines the strengths of transformers with mesh processing techniques. It uses the attention mechanism to dynamically weigh the importance of different mesh regions, leading to more precise feature extraction and analysis.

These models represent recent advancements in 3D mesh-based deep-learning and offer various approaches to effectively processing and analyzing complex 3D face mesh structures.

III. METHODOLOGY

This section outlines the methodology used in this research. We begin by detailing the assumptions made during our study. Next, we describe the 3DDGD dataset created for this study and used in our experiments. We then discuss the input features used in our model, and finally, we provide an overview of the classification network architecture used.

A. ASSUMPTIONS

In this study, we assumed that the recent and rapid advancements in generative artificial intelligence have made generating 2D facial images a highly accurate and sophisticated process. Such progress poses a significant threat to systems that rely on facial recognition or remote identity verification as it enables malicious actors to create realistic deepfake images with relative ease. Furthermore, we assume that obtaining a 3D facial image of a target person is more complex than acquiring a 2D image. While 2D images can be easily obtained from social media, online searches, and public records, obtaining a 3D face scan typically requires specialized hardware and software, like the multi-view system used to capture the detailed facial images in the FaceScape dataset, as illustrated in Figure 2 [4]. The complexity, expense, and need for sophisticated technology make 3D face images far less accessible and therefore more secure for facial recognition and remote identity verification systems.

Moreover, 3D face models offer enhanced security by focusing on the geometric details of facial features, such as the bumps and curves that define an individual's unique facial structure, rather than the attributes like age, race, or skin color often used in 2D models. The depth and precision of 3D facial data enable capturing even the tiniest details. This level of detail makes it much harder for attackers to create convincing fakes or bypass 3D-based recognition systems. 3D face recognition models have been shown to outperform state-of-the-art 2D models, offering a robust defense against attempts to deceive facial recognition systems [1], [51].

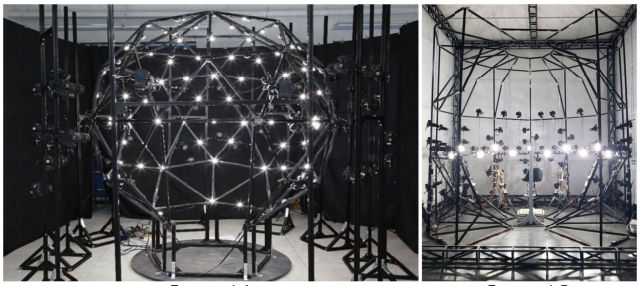


FIGURE 2. Photographs of the multi-view system used to reconstruct the high-quality, detailed 3D faces in the FaceScape dataset showing two frameworks. The system consists of 68 DSLR cameras, 30 of which capture 8K images focusing on the front side, while the remaining cameras capture 4K images of the side views [4].

B. 3DDGD DATASET

The 3DDGD dataset was developed to serve as a public 3D deepfake dataset to enhance the protection of 3D face recognition and remote identity verification systems against 3D deepfake attacks. In addition, it offers protocols for evaluating the performance of 3D deepfake detection models.

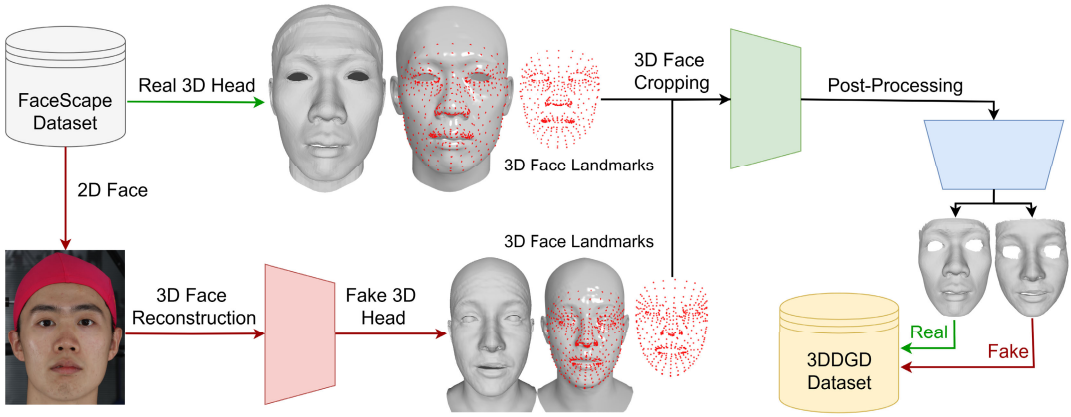


FIGURE 3. Flowchart illustrating the creation of the 3DDGD dataset. Real 3D head models were taken from the FaceScape dataset, while fake models were generated from 2D images in the same dataset using 3D face reconstruction methods. 3D facial landmarks were used to crop the faces, and post-processing was used to enhance the resulting models. Notably, the 3D deepfakes in our dataset were created by reconstructing 3D faces from 2D images.

The dataset is diverse in terms of age and gender, with mostly Asian identities, and includes 20 different facial expressions. Unlike existing deepfake detection datasets, which primarily focus on 2D deepfakes, the 3DDGD dataset specifically targets 3D deepfake scenarios. The subsequent sections detail the dataset creation process, as illustrated in Figure 3.

1) REAL 3D MODELS

The FaceScape dataset was selected as the source of real 3D face models due to its high-quality 3D facial geometry and diverse expressions [4]. Each of the 938 subjects (ranging in age from 16 to 70; with the majority Asian) in the FaceScape dataset exhibited 20 distinct expressions. The dataset was created using a multi-view system to reconstruct detailed 3D face models, Figure 2. The dataset provides 3D models of the entire head; therefore, we needed to crop the face on the basis of 3D facial landmarks, as illustrated in Figure 3.

2) FAKE 3D MODELS

We used two 3D face reconstruction methods to generate synthetic 3D face models. Frontal face images from the FaceScape dataset were selected due to their richer facial details compared with side views, as illustrated in Figure 3. We chose the reconstruction methods based on two main criteria: they must ensure accessibility and transparency and create detailed and accurate facial geometries. The methods we considered are shown in Figure 4 (from [9]), which compares different 3D face reconstruction methods. We selected DECA [6] and EMOCA [7] as the most suitable methods in accordance with these criteria. Since these methods generate complete 3D head models, we cropped the faces from the reconstructed models to focus solely on the facial region.

3) 3D FACE CROPPING

To extract the 3D face from the 3D head model, we used 3D facial landmarks,[1] assigning each point a specific index

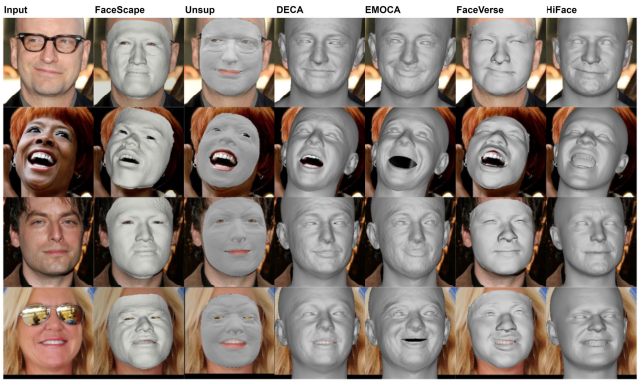


FIGURE 4. Detailed shape reconstruction comparison among different methods. Displayed from left to right: FaceScape [4], Unsup [52], DECA [6], EMOCA [7], FaceVerse [53], and HiFace [9].

(Figure 5 a). These indexed points enabled us to identify the points surrounding the face and define a bounding box, from which we extracted all relevant vertices corresponding to the face. The eyes are hollow in the real 3D models, whereas in the fake models, the eyes are filled. To maintain consistency between the two sets, we removed the eyes from the fake models to resemble the real ones. However, this eye removal process introduced several distortions (Figure 5 b and c). To ensure similar conditions in both sets, we applied the same removal procedure to the real models, which involved defining a polygon around the eye region and removing its vertices.

4) POSTPROCESSING

After 3D face cropping and eye area hollowing, a post-processing stage was necessary to enhance the quality of the resulting models. This involved basic operations like removing small, unconnected, or irrelevant parts that may appear after the eye area hollowing, especially in the fake models, as shown in Figure 5 (b and c). Additionally, it was important to refine the facial borders to reduce artifacts

[1] https://github.com/cse15-sip-interns/3d_face_landmark_identification.git

introduced during the cropping process. To achieve this, we represented the face as a graph (Figure 5 c) and used a heuristic method to remove vertices with degrees less than 4 ($d(vertex) < 4$). Before this, we determined the maximum and minimum bounding boxes for the face and eyes and used the heuristic method within these bounds to reduce the search space and processing time.

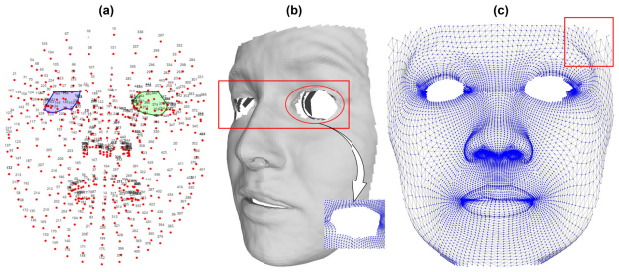


FIGURE 5. Illustration of basic operations in 3D face cropping and resulting artifacts. (a) Plotting of 3D facial landmarks to identify key points surrounding the face and eyes. (b) Visualization of disconnected face regions and artifacts following eye hollowing process. (c) Representation of 3D face as a mesh to highlight outlier vertices around facial and eye boundaries.

5) DATASET STATISTICS

The FaceScape dataset primarily consists of individuals of Asian descent. Figure 6 illustrates the age and gender distribution for both the real and fake datasets; more details are in Table 2. The real dataset is larger and includes more subjects compared with the fake dataset, which includes only 359 of the 938 subjects. This discrepancy is because the FaceScape dataset does not provide 2D images for all subjects. Additionally, the 3D fake faces were reconstructed using two different models.

FIGURE 6. Histogram illustrating distribution of subjects' age and gender across real and fake datasets, providing a comparative view of the demographic characteristics in each dataset.

C. INPUT FEATURES

Given a connected triangular mesh as input, where the mesh is defined by the pair $(\mathbf{V}, \mathbf{F})$, with $\mathbf{V} = \{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_n}\}$ representing the set of vertex positions in $\mathbb{R}^3$ and $\mathbf{F}$ defining the connectivity through triplets of vertices for triangular meshes, we extract various geometric features to construct a synthesized shape descriptor. In geometric deep learning, intrinsic and extrinsic features are crucial as they characterize the shape from different perspectives [39], [41], [44], [45], [46]. Intrinsic features capture shape variations while extrinsic features encode coordinate-dependent properties, Figure 7 [45]. The combination of these features forms a comprehensive geometric descriptor. Research indicates that hand-crafted shape descriptors significantly enhance network performance [39], [45], [46]. Therefore, integrating intrinsic shape descriptors and extrinsic shape descriptors is essential for effective geometric feature representation.

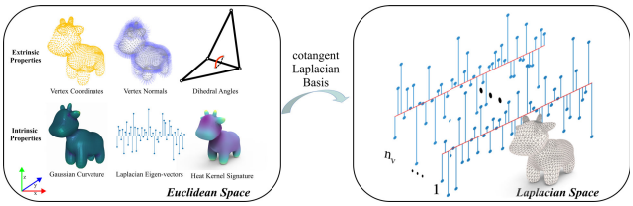


FIGURE 7. Our approach involves computing the mesh's intrinsic and extrinsic geometric features and transforming them into spectral signals for network input [45].

Our method for extracting features from 3D face meshes leverages the Laplacian-based approach from Dong et al. [45]. We construct a detailed feature set for each vertex in the mesh, combining both extrinsic (shape-related) and intrinsic (geometry-related) properties, Figure 7 [45]. Below, we outline the process step-by-step to clarify key concepts.

1) STEP 1: EXTRACTION OF GEOMETRIC FEATURES

Each 3D face mesh consists of a set of n vertices, where each vertex is described by a 39-dimensional feature vector. These features are categorized as follows:

- **Extrinsic Features (9D):** These features describe the absolute geometric properties of the vertex, including:
  1) 3D Vertex Coordinates (3D): (x, y, z) position of the vertex.
  2) Vertex Normals (3D): Normal vector (n_x, n_y, n_z) encoding local surface orientation.
  3) Dihedral Angles (3D): Angles between adjacent faces, capturing local sharpness.
- **Intrinsic Features (30D):** These describe the mesh's inherent geometric properties, independent of its global positioning:
  1) Gaussian Curvature (1D): Measures local shape at a vertex.
  2) Heat Kernel Signature (9D): Encodes multi-scale geometric information.

3) Eigenvectors of Cotangent Weight Matrix (20D): The 20 smallest non-zero eigenvectors provide a spectral representation of the mesh.

Thus, for each vertex, we construct a 39-dimensional feature vector, denoted as:

$$\mathbf{G} \in \mathbb{R}^{n \times 39} \tag{1}$$

where n represents the total number of vertices in the mesh.

2) STEP 2: TRANSFORMATION TO LAPLACIAN SPECTRAL DOMAIN

To improve computational efficiency and standardize the feature size across meshes, we transform $\mathbf{G}$ into the Laplacian spectral domain. This step reduces the dimensionality from n (which varies with mesh complexity) to a fixed k, producing a new feature matrix $\tilde{\mathbf{G}} \in \mathbb{R}^{k \times 39}$. The transformation follows this equation:

$$\tilde{\mathbf{G}}_{\mathbf{k}\times\mathbf{39}} = \Phi_{\mathbf{k}}^{\mathbf{T}} \mathbf{G}_{\mathbf{n}\times\mathbf{39}} \tag{2}$$

where:

- $\Phi_{\mathbf{k}} \in \mathbb{R}^{n \times k}$ are the first k eigenvectors of the Laplacian matrix.
- The hyperparameter k controls the spectral dimensionality, typically chosen from {16, 64, 154}.
- The transformation ensures dimensionality reduction, making $\tilde{\mathbf{G}}$ independent of the original mesh resolution.

3) STEP 3: STANDARDIZATION FOR EFFICIENT PROCESSING

The transformed feature matrix $\tilde{\mathbf{G}}$ has a consistent shape of $\mathbb{R}^{k \times 39}$, regardless of the complexity of the input mesh. This standardization:

- Facilitates uniform processing in subsequent deep learning models.
- Reduces computational overhead while retaining rich geometric information.
- Allows for seamless integration of additional task-specific features via simple concatenation.

D. CLASSIFICATION NETWORK

The proposed model uses the TabTransformer pipeline (Figure 8) to classify 3D face meshes as real or fake. The input to the classification network is feature matrix $\tilde{\mathbf{G}}$, which encapsulates the necessary features of the 3D face meshes. Given the subtle differences between real and fake 3D face models, as illustrated in Figure 9, we used the self-attention mechanism [54] within TabTransformer to capture fine-grained details crucial for accurate classification.

The TabTransformer pipeline was chosen due to its proven effectiveness in handling large feature sets and its successful application in previous research [55], [56], [57]. The network starts with an embedding layer that linearly transforms the input features. This is followed by multiple encoding layers, each consisting of a multi-head attention mechanism and a feed-forward network designed to enhance the model's ability to focus on different aspects of the input features.

Finally, the output layer, which uses a sigmoid activation function, predicts the probability of the mesh being real or fake. The binary cross-entropy loss function is used to optimize the model during training, ensuring precise differentiation between the two classes.

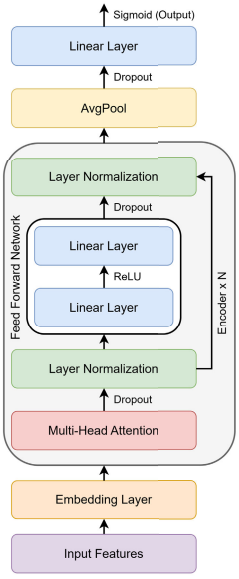


FIGURE 8. Our TabTransformer pipeline for mesh classification tasks.

IV. EXPERIMENTS

In this section, we describe the experiments conducted to assess our dataset's and classifier's effectiveness. We focus on two primary research questions:

RQ1: How well do models trained on our dataset perform when tested on different datasets?

RQ2: How does changing the training dataset affect the stability of the models?

We aimed to evaluate the models' generalization capabilities and robustness by addressing these questions, providing valuable insights into their performance in real-world applications.

A. DATASET PROCESS

We created a second dataset using a different source, resolution, and data collection method, specifically a structured light system. This new dataset was derived from the BU-3DFE dataset [18] using the same method shown in Figure 3. One challenge we encountered was the variation in the number of vertices in the 3D face meshes, which hindered feature extraction. We needed a standardized and normalized approach for all 3D models to address this. Initially, we tried standardizing the number of vertices in each 3D face mesh, but this approach led to significant defects. We then standardized the number of faces to 15,000 using the Simplify Quadratic Decimation method [59]. This method is commonly used in computer graphics and 3D modeling and reduces the complexity of a 3D mesh (faces and vertices)

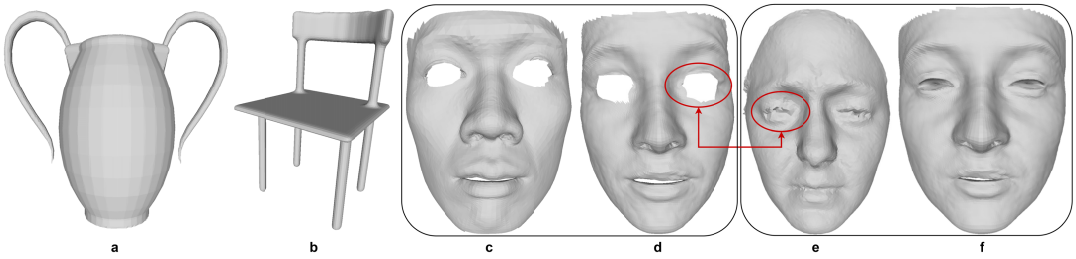


FIGURE 9. This figure illustrates the topological differences between various 3D objects represented as 3D meshes as well as the differences between real and fake 3D faces. Images (a) and (b), sourced from the COSEG dataset [58], demonstrate significant topological differences. In contrast, images (c) and (d) show real and fake 3D faces from the first dataset (FaceScape), while images (e) and (f) show real and fake 3D faces from the second dataset (BU-3DFE). Notably, there is a substantial topological similarity between the real and fake faces in each dataset. Note that the eye area was not cropped in the second dataset.

while preserving its overall shape and essential features. Additionally, the 3D face meshes had varying dimensions for the 3D coordinates of the vertices. Therefore, we standardized the coordinate space to the range $[-0.5, 0.5]$. Both datasets were split into a training set (80%) and a test set (20%). Comprehensive statistics for both datasets are provided in Table 2.

TABLE 2. Comprehensive statistics of datasets used in this study.

		Source	No. of Expressions	No. of Identities	No. of Samples
Dataset 1	Real	FaceScape	20	846	16888
	Fake	DECA	20	359	7143
		EMOCA	20	250	4958
					28989
Dataset 2	Real	BU-3DFE	25	100	2500
	Fake	DECA	25	100	2500
					5000

B. 3D DEEPFAKE DETECTION AND RESULTS

In our 3D deepfake detection experiments, we adapted a model initially designed for 3D mesh classification known for its high accuracy in that domain. The results are presented in Tables 3 and 4.

We modified the mesh-MLP model[2] [46], originally developed for multi-class 3D mesh classification, to perform binary classification. This involved adjusting its output layer and incorporating sigmoid activation and binary cross-entropy loss functions. Building on this foundation, we developed two variants: the mesh-MLP-SA model, which integrates a self-attention mechanism, and the mesh-MLP-MHA model, which incorporates a multi-head self-attention mechanism. These adaptations enabled us to leverage the models' strengths in 3D mesh representation while tailoring them to the specific needs of deepfake detection. Additionally, we introduced our novel TabTransformer model, designed specifically for this study, with its architecture detailed in Figure 8. This ensemble of models, each with unique enhancements, forms a preliminary toolkit for addressing the challenges of 3D deepfake detection.

[2]https://github.com/QiujieDong/TaskDrivenNet2Mesh.git

TABLE 3. Results of 3D deepfake detectors trained on Dataset 1 and evaluated on Dataset 1 and Dataset 2, reported as F1 scores (%). Green highlights the highest scores, while red indicates the second-highest.

Models	Tested on Dataset 1				Tested on Dataset 2			
	G0	G1	G2	All	G0	G1	G2	All
Mesh_MLP	95.39	95.66	95.61	**96.30**	63.36	60.72	62.13	61.78
Mesh_MLP_SA	**97.61**	**98.30**	**98.38**	95.57	71.59	71.79	**72.50**	60.62
Mesh_MLP_MHA	85.16	96.50	95.41	**97.87**	**72.25**	**79.77**	**75.55**	**73.32**
TabTransformer	**98.17**	**98.01**	**97.98**	93.87	**71.66**	**72.26**	71.71	**83.18**

TABLE 4. Results of 3D deepfake detectors trained on Dataset 2 and evaluated on Dataset 2 and Dataset 1, reported as F1 scores (%). Green highlights the highest scores, while red indicates the second-highest.

Models	Tested on Dataset 2				Tested on Dataset 1			
	G0	G1	G2	All	G0	G1	G2	All
Mesh_MLP	**95.25**	**96.39**	**95.27**	**96.57**	61.56	**66.06**	61.58	**62.43**
Mesh_MLP_SA	**95.57**	**95.66**	**95.69**	**94.91**	61.54	61.01	**62.50**	61.83
Mesh_MLP_MHA	93.48	92.38	92.51	94.79	**61.77**	**64.17**	59.45	62.08
TabTransformer	93.41	93.12	93.02	94.47	**62.91**	63.98	**62.58**	**63.29**

To further enhance detection performance and generalization, we employed an Ensemble-TabTransformer approach, in which three TabTransformer models were independently trained with varied initializations. The ensemble aggregated predictions using soft-voting, allowing it to leverage complementary strengths of the individual models. As shown in Table 5, this method consistently outperformed single-model baselines, demonstrating improved robustness and reliability in 3D deepfake detection tasks.

C. ADVERSARIAL ATTACK

Evaluating the robustness of 3D deepfake detectors in adversarial settings is essential to ensure their reliability in real-world scenarios. In this experiment, we specifically investigate adversarial attacks at the semantic level targeting the proposed 3D deepfake detection models, rather than a traditional 3D face recognition system. The goal is to assess

TABLE 5. Performance of Ensemble-TabTransformer for 3D deepfake detection trained on Dataset 1 and tested on Dataset 1 and Dataset 2 (F1 score in %).

	Tested on Dataset 1				Tested on Dataset 2			
Model	G0	G1	G2	All	G0	G1	G2	All
Ensemble TabTransformer	97.48	95.74	97.04	97.18	81.34	83.98	88.46	80.87

whether subtle manipulations of 3D face meshes can deceive the detector into misclassifying deepfakes as real samples or vice versa.

Unlike traditional image-based attacks, 3D mesh attacks are challenging due to the complex geometric structure of mesh surfaces. These surfaces contain semantic and non-semantic features, such as vertex connections and curvature, which make adversarial perturbations difficult to craft. Recent studies highlight that effective adversarial meshes must be imperceptible and geometrically realistic [60], [61] [62], [63].

Our approach introduces perturbations in the spectral domain of the mesh inspired by the SAGA [60] method. We generate adversarial examples by adding controlled spectral noise to the clean 3D face mesh eigencomponents, followed by mesh reconstruction. This method allows for precise geometric changes while maintaining the natural appearance of the mesh.

For evaluation, we used the test set of Dataset 1 to generate adversarial examples, as spectral transformations and reconstructions are computationally intensive. Despite the added complexity, our models showed strong robustness. Although accuracy slightly decreased under attack conditions, the models remained effective, demonstrating resilience to adversarial manipulation as shown in Table 6.

These results emphasize the importance of evaluating adversarial robustness for 3D-based deepfake detectors and biometric authentication systems. Understanding how 3D models respond to spectral attacks helps to develop more secure and adaptable architectures.

TABLE 6. Robustness evaluation of models against adversarial attack, trained and tested on Dataset 1. Results are reported as F1 scores (%).

	Tested on Dataset 1			
Models	G0	G1	G2	All
Mesh_MLP_MHA	84.14	86.49	93.59	92.29
TabTransformer	87.88	85.75	92.98	89.76
Ensemble TabTransformer	95.97	94.78	95.83	96.73

D. COMPUTATIONAL COSTS AND DEPLOYMENT

Understanding the computational cost of deepfake detection is crucial to enable real-world deployment. In this section, we evaluate the inference efficiency of our proposed 3D deepfake detection models, providing a detailed comparison with existing 2D-based approaches. We also discuss the practical implications of these findings for deployment in real-world scenarios.

1) TRAINING AND INFERENCE COSTS

Our 3D-based deepfake detection framework comprises three primary computational stages:

1) **Data Preparation:** Loading the raw inputs, cropping of the 3D facial region, and normalizing the resulting mesh representations.
2) **Feature Extraction:** Computing both extrinsic and intrinsic geometric features, including Laplacian-based spectral transformations.
3) **Inference:** Performing classification using deep neural networks.

To evaluate the computational efficiency of our pipeline, we benchmark each stage using NVIDIA Tesla T4 and A100 GPUs. A detailed breakdown of time consumption for each stage is shown in Figure 10. All 3D face meshes are normalized to a consistent structure, containing approximately 15,000 triangular faces and 7,500-8,000 vertices, depending on mesh connectivity. This preprocessing step ensures structural uniformity, which is essential for consistent and reliable feature extraction.

Among the three stages, feature extraction contributes the most to the overall computational cost, with intrinsic feature computation being particularly demanding. This involves calculating the eigenvectors of the cotangent weight matrix and computing the heat-kernel signature, both computationally intensive. Data preparation also adds to the overhead, though to a lesser extent.

Training time depends primarily on the size of the data set and the computational cost of intrinsic feature extraction. However, it remains practical for both research and deployment scenarios, offering a favorable trade-off between detection accuracy and computational efficiency.

Our findings indicate that the inference stage is relatively lightweight compared to pre-processing and feature extraction. Future work could explore optimization strategies such as accelerating spectral feature computation, parallelizing data preprocessing, and applying model compression techniques such as quantization to further reduce training and inference latency.

2) COMPARISON WITH 2D DEEPFAKE DETECTION

Deploying our 3D framework in practical settings requires balancing its computational demands with its enhanced detection capabilities. Table 7 evaluates our approach against 2D deepfake detection across four criteria: data availability, feature variety, inference time, and adversarial robustness using a rating scale (−: Weak, +: Good, *: Very Good).

a: EXISTING DATASETS

2D deepfake detection benefits from a wealth of image and video datasets, such as FaceForensics++ (FF++) [64],

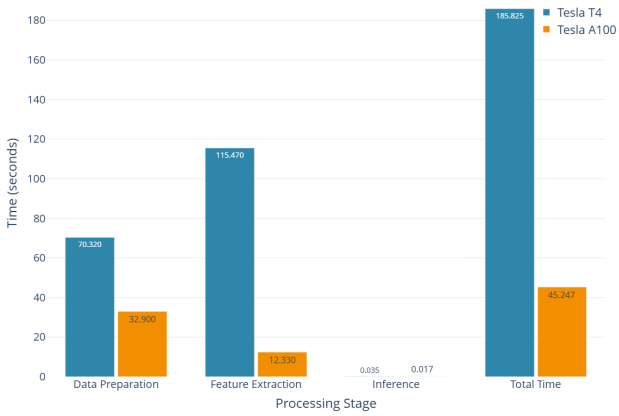


FIGURE 10. Comparison of average computational costs (in seconds) for one example on Tesla T4 and Tesla A100 GPUs across data preparation, feature extraction, inference, and total processing time.

TABLE 7. Comparison of 3D deepfake detection with 2D deepfake detection. Ratings: –: Weak, +: Good, *: Very Good.

Criterion	2D Detection	3D Detection
Data Availability	*	-
Feature Variety	+	*
Inference Time	*	+
Adversarial Robustness	- +	+ *

and is therefore rated as *Very Good* in terms of dataset availability. In contrast, 3D deepfake detection faces a significant limitation due to the scarcity of publicly available 3D face mesh datasets (see Table 1), leading to a *Weak* rating in this aspect.

b: FEATURE VARIETY

The proposed 3D deepfake detection framework is rated as *Very Good* because it effectively utilizes both extrinsic and intrinsic geometric features, capturing detailed spatial structures [39], [45], [50]. Furthermore, it enables the fusion of geometric and textural features [17], surpassing the feature extraction capabilities of 2D methods, which primarily rely on pixel-based analysis and are rated as *Good*.

c: INFERENCE TIME

2D detection methods achieve *Very Good* performance in terms of inference speed due to their optimized image-based processing pipelines. In contrast, our 3D approach is computationally more intensive but remains efficient, as shown in Figure 10, earning a *Good* rating.

d: ADVERSARIAL ROBUSTNESS

The proposed 3D detection framework demonstrates strong resilience against 3D-aware manipulations, with a robustness rating of *Good - Very Good*, Table 6, however, further investigation is required. In comparison, 2D deepfake detection methods, rated as *Weak - Good*, exhibit vulnerabilities to advanced deepfake techniques that exploit 2D representations, making them more susceptible to sophisticated attacks [13].

e: REAL-WORLD DEPLOYMENT

The proposed 3D deepfake detection system is well-suited for high-security applications, such as biometric verification and identity authentication, where robustness and security take precedence over computational speed. In contrast, 2D detection methods excel in large-scale, real-time scenarios, such as video forensics and social media monitoring, benefiting from their extensive dataset availability and lower computational overhead. This trade-off highlights the complementary nature of our framework, positioning it as a specialized solution for security-critical environments. Future advancements in hardware acceleration and algorithmic optimizations are expected to enhance the efficiency of 3D deepfake detection, making real-time deployment feasible even on resource-constrained devices.

E. DISCUSSION

Detecting deepfakes using 3D face meshes presents a unique challenge compared with classifying other 3D objects represented by meshes. Unlike the clearly evident topological differences found among various 3D objects (such as chairs, vases, and aliens in the COSEG dataset [58]), the distinctions between real and fake 3D faces are often subtle and difficult to detect, as illustrated in Figure 9. This subtlety necessitates using advanced techniques like self-attention to capture the nuanced details that distinguish genuine faces from deepfakes.

Our findings, as presented in Tables 3 and 4, demonstrate that incorporating an attention mechanism into the mesh-MLP model (resulting in Mesh-MLP-SA) enhanced overall performance. Furthermore, adding a multi-head attention mechanism (mesh-MLP-MHA) improved the model's ability to generalize to unseen data although this comes at the cost of slightly weaker performance on the training dataset. Notably, all models exhibited stability across different datasets, highlighting the need for an architecture that can both generalize well to unseen data and maintain stability across various training datasets. This led to our development of the TabTransformer architecture (Figure 8), which showed promising results on tabular data containing numerous features. Although we did not crop the eye area in the dataset derived from the BU-3DFE dataset (Figure 9), the models used in this study achieved an acceptable generalization rate on unseen data. Furthermore, they demonstrated stability when the training dataset was changed. These results address research questions **RQ1** and **RQ2**.

Dong et al. showed that, despite the usefulness of convolutions, a simple architecture relying only on MLPs can effectively handle mesh classification and semantic segmentation [46]. Our findings corroborate this; however, given the topological similarities between real and fake 3D faces, we found it necessary to introduce attention

mechanisms to capture and differentiate the fine details in the features.

Our preliminary results are promising and suggest potential for further development. The use of 3D face meshes has the potential to greatly enhance security in systems that rely on facial images for authentication or verification purposes.

F. LIMITATIONS AND FUTURE WORK

One of the major challenges in this field is collecting a sufficiently large and diverse dataset, encompassing various races, ages, and genders, to effectively train deep-learning models. Scanning 3D faces requires significant resources, including specialized tools and volunteer participation. Since most features and expressions critical for identifying a person are concentrated on the face, it is essential to crop or segment the 3D face. However, current tools and models lack the capability to efficiently handle all types of data or existing datasets.

The growing prevalence of models that animate 3D faces or modify their expressions and features necessitates the development of new tools for 3D face cropping and segmentation. Such tools will be crucial for detecting general deepfakes of 3D faces or specific alterations, such as lip movements during speech. In addition to working on the development of such tools, we will also explore integrating geometric features and facial textures to develop more accurate and efficient methods for 3D deepfake detection.

V. CONCLUSION

In this study, we aimed to achieve three key objectives: enhancing security by using 3D faces for authentication, creating a comprehensive 3D face dataset, and evaluating deep-learning models for detecting 3D deepfakes. We developed a large, diverse dataset of real and fake face models generated through 3D reconstruction from 2D images along with a smaller dataset to test model generalization and stability. Our evaluation of these models showed that they can be used to effectively detect 3D deepfakes while generalizing well to unseen data and remaining stable with different training conditions. These findings indicate that using 3D faces can significantly improve the security of facial recognition and remote identity verification systems.

REFERENCES

[1] S. Z. Gilani and A. Mian, "Learning from millions of 3D scans for large-scale 3D face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1896–1905.

[2] K. W. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition," *Comput. Vis. Image Understand.*, vol. 101, no. 1, pp. 1–15, Jan. 2006.

[3] M. Wu, H. Zhu, L. Huang, Y. Zhuang, Y. Lu, and X. Cao, "High-fidelity 3D face generation from natural language descriptions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 4521–4530.

[4] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao, "FaceScape: A large-scale high quality 3D face dataset and detailed riggable 3D face prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 598–607.

[5] F. Taherkhani, A. Rai, Q. Gao, S. Srivastava, X. Chen, F. de la Torre, S. Song, A. Prakash, and D. Kim, "Controllable 3D generative adversarial face model via disentangling shape and appearance," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 826–836.

[6] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3D face model from in-the-wild images," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–13, Aug. 2021.

[7] R. Daněček, M. J. Black, and T. Bolkart, "EMOCA: Emotion driven monocular face capture and animation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20279–20290.

[8] B. Lei, J. Ren, M. Feng, M. Cui, and X. Xie, "A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 394–403.

[9] Z. Chai, T. Zhang, T. He, X. Tan, T. Baltrusaitis, H. Wu, R. Li, S. Zhao, C. Yuan, and J. Bian, "HiFace: High-fidelity 3D face reconstruction by learning static and dynamic details," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 9053–9064.

[10] B. Thambiraja, I. Habibie, S. Aliakbarian, D. Cosker, C. Theobalt, and J. Thies, "Imitator: Personalized speech-driven 3D facial animation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 20564–20574.

[11] F. Nocentini, T. Besnier, C. Ferrari, S. Arguillere, S. Berretti, and M. Daoudi, "ScanTalk: 3D talking heads from unregistered scans," 2024, *arXiv:2403.10942*.

[12] J. Ling, Z. Wang, M. Lu, Q. Wang, C. Qian, and F. Xu, "Structure-aware editable morphable model for 3D facial detail animation and manipulation," in *Proc. Eur. Conf. Comput. Vis.*, Jan. 2022, pp. 249–267.

[13] H. Felouat, H. H. Nguyen, T.-N. Le, J. Yamagishi, and I. Echizen, "EKYC-DF: A large-scale deepfake dataset for developing and evaluating eKYC systems," *IEEE Access*, vol. 12, pp. 30876–30892, 2024.

[14] S. Chen, W. Li, H. Yang, D. Huang, and Y. Wang, "3D face mask anti-spoofing via deep fusion of dynamic texture and shape clues," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 314–321.

[15] X. Zhu, H. Wang, H. Fei, Z. Lei, and S. Z. Li, "Face forgery detection by 3D decomposition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2928–2938.

[16] X. Zhu, H. Fei, B. Zhang, T. Zhang, X. Zhang, S. Z. Li, and Z. Lei, "Face forgery detection by 3D decomposition and composition search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8342–8357, Jul. 2023.

[17] J. M. Singh and R. Ramachandra, "3-D face morphing attacks: Generation, vulnerability and detection," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 6, no. 1, pp. 103–117, Jan. 2024.

[18] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit. (FGR06)*, Apr. 2006, pp. 211–216.

[19] L. Yin, X. Chen, Y. Sun, T. Worm, and M. J. Reale, "A high-resolution 3D dynamic facial expression database," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Sep. 2008, pp. 1–6.

[20] B. Yin, S. Yan-feng, C. Wang, and G. Yun, "BJUT-3D large scale 3D face database and information processing," *J. Comput. Res. Develop.*, vol. 46, no. 6, p. 1009, Jun. 2009.

[21] A. Savran, N. Alyüz, H. Dibeklioglu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3D face analysis," in *Proc. 1st Eur. Workshop Biometrics Identity Manag.* Cham, Switzerland: Springer, Jan. 2008, pp. 47–56.

[22] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "FaceWarehouse: A 3D facial expression database for visual computing," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 3, pp. 413–425, Mar. 2014.

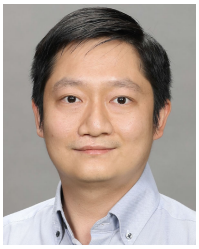
[23] S. Cheng, I. Kotsia, M. Pantic, and S. Zafeiriou, "4DFAB: A large scale 4D database for facial expression analysis and biometric applications," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5117–5126.

[24] D. Cosker, E. Krumhuber, and A. Hilton, "A FACS valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2296–2303.

[25] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "BP4D-spontaneous: A high-resolution spontaneous 3D dynamic facial expression database," *Image Vis. Comput.*, vol. 32, no. 10, pp. 692–706, Oct. 2014.

[26] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: Going beyond Euclidean data," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 18–42, Jul. 2017.
[27] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," 2021, *arXiv:2104.13478*.
[28] Y.-P. Xiao, Y.-K. Lai, F.-L. Zhang, C. Li, and L. Gao, "A survey on deep geometry learning: From a representation perspective," *Comput. Vis. Media*, vol. 6, no. 2, pp. 113–133, Jun. 2020.
[29] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 945–953.
[30] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas, "Volumetric and multi-view CNNs for object classification on 3D data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5648–5656.
[31] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D ShapeNets: A deep representation for volumetric shapes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1912–1920.
[32] L. Tchapmi, C. Choy, I. Armeni, J. Gwak, and S. Savarese, "SEGCloud: Semantic segmentation of 3D point clouds," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 537–547.
[33] R. Hanocka, N. Fish, Z. Wang, R. Giryes, S. Fleishman, and D. Cohen-Or, "ALIGNet: Partial-shape agnostic alignment via unsupervised learning," *ACM Trans. Graph.*, vol. 38, no. 1, pp. 1–14, Feb. 2019.
[34] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 77–85.
[35] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2017, pp. 5099–5108.
[36] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, "PointCNN: Convolution on X-transformed points," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2018, pp. 828–838.
[37] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "PCT: Point cloud transformer," *Comput. Vis. Media*, vol. 7, no. 2, pp. 187–199, Jun. 2021.
[38] A. Lahav and A. Tal, "MeshWalker: Deep mesh understanding by random walks," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 1–13, Dec. 2020.
[39] R. Hanocka, A. Hertz, N. Fish, R. Giryes, S. Fleishman, and D. Cohen-Or, "MeshCNN: A network with an edge," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–12, Aug. 2019.
[40] F. Milano, A. Loquercio, A. Rosinol, D. Scaramuzza, and L. Carlone, "Primal-dual mesh convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2020, pp. 952–963.
[41] S.-M. Hu, Z.-N. Liu, M.-H. Guo, J.-X. Cai, J. Huang, T.-J. Mu, and R. R. Martin, "Subdivision-based mesh convolution networks," *ACM Trans. Graph.*, vol. 41, no. 3, pp. 1–16, Jun. 2022.
[42] D. Smirnov and J. Solomon, "HodgeNet: Learning spectral geometry on triangle meshes," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–11, Aug. 2021.
[43] X. Li, R. Li, L. Zhu, C.-W. Fu, and P.-A. Heng, "DNF-net: A deep normal filtering network for mesh denoising," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 10, pp. 4060–4072, Oct. 2021.
[44] N. Sharp, S. Attaiki, K. Crane, and M. Ovsjanikov, "DiffusionNet: Discretization agnostic learning on surfaces," *ACM Trans. Graph.*, vol. 41, no. 3, pp. 1–16, Jun. 2022.
[45] Q. Dong, Z. Wang, M. Li, J. Gao, S. Chen, Z. Shu, S. Xin, C. Tu, and W. Wang, "Laplacian2Mesh: Laplacian-based mesh understanding," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 7, pp. 4349–4361, Jul. 2024.
[46] Q. Dong, X. Gong, R. Xu, Z. Wang, J. Gao, S. Chen, S. Xin, C. Tu, and W. Wang, "A task-driven network for mesh classification and semantic part segmentation," *Comput. Aided Geometric Design*, vol. 111, Jun. 2024, Art. no. 102304.
[47] C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen, and T.-Y. Liu, "Do transformers really perform badly for graph representation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, Dec. 2021, pp. 28877–28888.
[48] H. Katam, "3D mesh segmentation using transformer based graph operations," Ph.D. dissertation, Fakultät für Informatik, Technische Universität München, Munich, Germany, 2021.
[49] I. Sarasua, S. Pölsterl, and C. Wachinger, "TransforMesh: A transformer network for longitudinal modeling of anatomical meshes," in *Proc. 12th Int. Workshop Mach. Learn. Med. Imaging*, Strasbourg, France. Cham, Switzerland: Springer, Sep. 2021, pp. 209–218.
[50] H.-Y. Peng, M.-H. Guo, Z.-N. Liu, Y.-L. Yang, and T.-J. Mu, "Meshformers: Transformer-based networks for mesh understanding," *SSRN Electron. J.*, Jan. 2022. [Online]. Available: https://ssrn.com/abstract=4313526
[51] Y. Jing, X. Lu, and S. Gao, "3D face recognition: A comprehensive survey in 2022," *Comput. Vis. Media*, vol. 9, no. 4, pp. 657–685, Dec. 2023.
[52] Y. Chen, F. Wu, Z. Wang, Y. Song, Y. Ling, and L. Bao, "Self-supervised learning of detailed 3D face reconstruction," *IEEE Trans. Image Process.*, vol. 29, pp. 8696–8705, 2020.
[53] L. Wang, Z. Chen, T. Yu, C. Ma, L. Li, and Y. Liu, "FaceVerse: A fine-grained and detail-controllable 3D face morphable model from a hybrid dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20301–20310.
[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.
[55] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "TabTransformer: Tabular data modeling using contextual embeddings," 2020, *arXiv:2012.06678*.
[56] Z. Wang and J. Sun, "TransTab: Learning transferable tabular transformers across tables," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2022, pp. 2902–2915.
[57] S. Luetto, F. Garuti, E. Sangineto, L. Forni, and R. Cucchiara, "One transformer for all time series: Representing and training with time-dependent heterogeneous tabular data," 2023, *arXiv:2302.06375*.
[58] Y. Wang, S. Asafi, O. van Kaick, H. Zhang, D. Cohen-Or, and B. Chen, "Active co-analysis of a set of shapes," *ACM Trans. Graph.*, vol. 31, no. 6, pp. 1–10, Nov. 2012.
[59] G. Zhou, S. Yuan, and S. Luo, "Mesh simplification algorithm based on the quadratic error metric and triangle collapse," *IEEE Access*, vol. 8, pp. 196341–196350, 2020.
[60] T. Stolik, I. Lang, and S. Avidan, "SAGA: Spectral adversarial geometric attack on 3D meshes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4261–4271.
[61] Y. Huang, Y. Dong, S. Ruan, X. Yang, H. Su, and X. Wei, "Towards transferable targeted 3D adversarial attack in the physical world," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 24512–24522.
[62] J. Zhang, L. Chen, B. Liu, B. Ouyang, Q. Xie, J. Zhu, W. Li, and Y. Meng, "3D adversarial attacks beyond point cloud," *Inf. Sci.*, vol. 633, pp. 491–503, Jul. 2023.
[63] C. Xiao, D. Yang, B. Li, J. Deng, and M. Liu, "MeshAdv: Adversarial meshes for visual recognition," 2018, *arXiv:1810.05206*.
[64] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–11.

HICHEM FELOUAT received the B.S. and M.S. degrees in computer science from the University of Jijel, Algeria, in 2013 and 2015, respectively, and the Ph.D. degree in computer science from the University of Blida, Algeria, in 2021. He is currently pursuing the Ph.D. degree with the Echizen Laboratory, National Institute of Informatics, Tokyo, Japan. His research interests include security and privacy in biometrics and machine learning.



HUY H. NGUYEN (Member, IEEE) received the Ph.D. degree in computer science from The Graduate University for Advanced Studies, SOKENDAI, Japan, in 2022. He is currently a Project Assistant Professor with the Echizen Laboratory, National Institute of Informatics, Tokyo, Japan. His research interests include security and privacy in biometrics and machine learning.

JUNICHI YAMAGISHI (Senior Member, IEEE) received the Ph.D. degree from Tokyo Institute of Technology (Tokyo Tech), Tokyo, Japan, in 2006. From 2007 to 2013, he was a Research Fellow with the Centre for Speech Technology Research, The University of Edinburgh, U.K. He became an Associate Professor with the National Institute of Informatics, Japan, in 2013, where he is currently a Professor. His research interests include speech processing, machine learning, signal processing, biometrics, digital media cloning, and media forensics. He served as a co-organizer for the bi-annual ASVspoof challenge and the bi-annual voice conversion challenge. He also served as a member for the IEEE Speech and Language Technical Committee from 2013 to 2019, an Associate Editor for IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING from 2014 to 2017, and the Chairperson for ISCA SynSIG from 2017 to 2021. He is currently a Principal Investigator on the JST-CREST and ANR supported VoicePersona Project and a Senior Area Editor for IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.

ISAO ECHIZEN (Senior Member, IEEE) received the B.S., M.S., and D.E. degrees from Tokyo Institute of Technology, Japan, in 1995, 1997, and 2003, respectively. In 1997, he joined Hitachi Ltd., where he was a Research Engineer with the Systems Development Laboratory, until 2007. He was a Visiting Professor with the University of Freiburg, Germany, and the University of Halle-Wittenberg, Germany. He is currently the Director and a Professor with the Information and Society Research Division, National Institute of Informatics (NII), the Director of the Global Research Center for Synthetic Media, NII, a Professor with the Department of Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo, and a Professor with the Graduate Institute for Advanced Studies, The Graduate University for Advanced Studies, SOKENDAI, Japan. He is also engaged in research on multimedia security and multimedia forensics. He is the Research Director of the CREST FakeMedia Project, Japan Science and Technology Agency (JST). He was a member of the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society. He is an IEICE Fellow, the Japanese Representative on IFIP TC11 (Security and Privacy Protection in Information Processing Systems), a Member-at-Large of the Board-of-Governors of APSIPA, and an Editorial Board Member of IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, the *EURASIP Journal on Image and Video Processing*, and the *Journal of Information Security and Applications* (Elsevier). He received the Best Paper Award from IEICE in 2023, the Best Paper Award from IPSJ in 2005 and 2014, the IPSJ Nagao Special Researcher Award in 2011, the DOCOMO Mobile Science Award in 2014, the Information Security Cultural Award in 2016, and the IEEE Workshop on Information Forensics and Security Best Paper Award in 2017.