Every Frame You Take: The Video-based Morphing Attack Detection Challenge

Nicolò Di Domenico[1,^], Lorenzo Pellegrini[1,^], Simone Maurizio La Cava[3,^], Guido Borghi[2,^], Fabio Notaro[1,^], Nicolò Penserini[1,^], Nicolas Tazzieri[1,^], Annalisa Franco[1,^], Matteo Ferrara[1,^], Gian Luca Marcialis[3,^], Ananya Dhulipala[4,†], Mohammad Javad Shamani[5,†], Asif Hussain Khan[6,†], Yannick Dayer[6,†], Hatef Otroshi Shahreza[6,†], Alain Komaty[6,†], Sébastien Marcel[6,†], Anant Gupta[7,†], Gourav Gupta[7,†], Leon Todorov[8,†], Nikola Maric[9,†], Marija Ivanovska[8,†]

[1]University of Bologna, Italy [2]University of Modena and Reggio Emilia, Italy
[3]University of Cagliari, Italy [4]Eli Lilly and Company, India
[5]St. Bonaventure University, United States [6]Idiap Research Institute, Switzerland
[7]Arogyapandit Private Limited, India [8]University of Ljubljana, Slovenia
[9]Jožef Stefan Institute, Slovenia

[^]Organizers [†]Participants

{nicolo.didomenico, l.pellegrini}@unibo.it, simonem.lac@unica.it

Abstract

Face morphing attacks represent a critical threat to Face Recognition systems deployed in operational scenarios such as automated border control gates and remote identity verification platforms. Although most Morphing Attack Detection (MAD) research focuses on static images, real-world operational biometric systems typically acquire short video sequences that provide additional temporal and quality-related information. To address this gap, "Every Frame You Take: The Video-based Morphing Attack Detection (V-MAD) Challenge", organized in conjunction with the IEEE/IAPR International Joint Conference on Biometrics 2026, promotes research on Video-based MAD under realistic conditions. The competition introduces a novel benchmark composed of real and synthetic acquisition datasets and evaluates submitted methods under two complementary settings: a simple scenario aligned with the training distribution and a significantly more challenging scenario designed to assess robustness against unseen morphing conditions and acquisition variability. A total of 25 systems were submitted and evaluated on sequestered datasets using standard and custom MAD metrics. This paper summarizes the challenge organization, datasets, evaluation protocol, participating methods, and final results, highlighting emerging trends and future directions for robust Video-based MAD systems.

1. Introduction

Face morphing attacks [22] have emerged as one of the most critical threats against modern Face Recognition systems (FRs) deployed in operational scenarios such as Automated Border Control (ABC) gates. By blending the facial characteristics of two different individuals into a single synthetic identity image, morphing attacks can allow multiple subjects to successfully authenticate against the same identity document while maintaining visual plausibility to both human operators and automated biometric systems [47].

Over the last decade, the research community has proposed a large variety of Morphing Attack Detection (MAD) approaches to counter this threat. Generally, existing methods are divided into Single-image MAD (S-MAD), where only the ICAO-compliant [31, 32] document image is analyzed, and Differential MAD (D-MAD), where the document image is compared against a trusted live capture [46]. Recent advances in Deep Learning and feature fusion analysis have significantly improved detection performance under controlled conditions [9, 27, 48]. Generally, most existing MAD benchmarks and algorithms are designed only for single- and double-image analysis.

The use of a single or a pair of images in input only partially reflects real operational biometric systems, such as ABC gates. Indeed, in practical deployments, FRs typically acquire short video streams rather than isolated frames [13]. The availability of multiple frames introduces both new vulnerabilities and new opportunities.

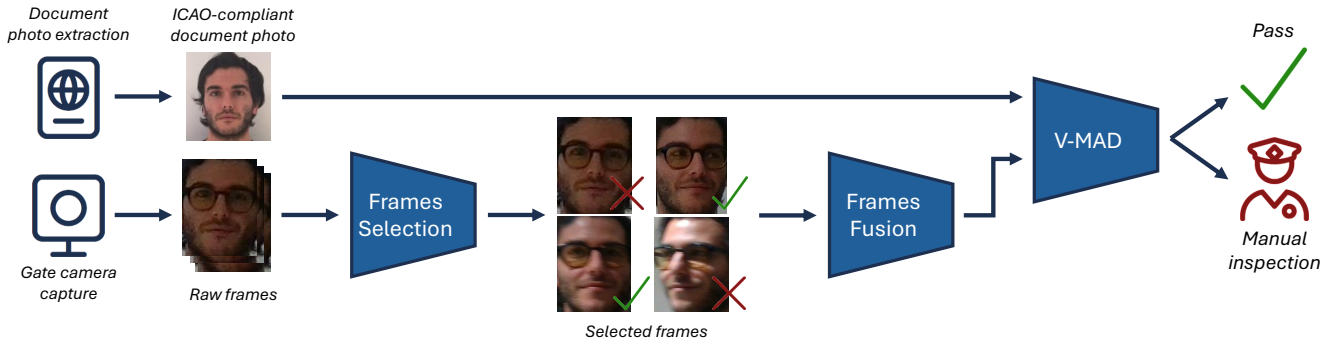

Figure 1. Visual summarization of the V-MAD task. The input consists of an ICAO-compliant image and a short video sequence. Raw frames can be selected (*Frame Selection*) to reduce the impact of several elements (such as non-frontal poses or uneven illumination); in addition, features extracted from single frames can be merged (*Frame Fusion*) for a more robust result of the final V-MAD model.

On one hand, repeated acquisition attempts may increase the probability of a successful morphing attack. On the other hand, video sequences provide richer temporal and quality-related information that can potentially improve robustness of MAD systems.

To address this gap between research benchmarks and operational deployments, Borghi *et al.* [7] introduced the task of Video-based Morphing Attack Detection (V-MAD), demonstrating that combining information across multiple probe frames can substantially improve MAD performance compared to traditional frame-level approaches. Their work established the first systematic formulation of V-MAD, investigating score-level fusion strategies, face image quality assessment, and machine learning-based fusion methods in realistic border-control scenarios.

1.1. Video-based Morphing Attack Detection

As mentioned, in real operational scenarios, Face Recognition systems typically acquire short video streams rather than isolated facial images [13]. Consequently, traditional D-MAD approaches may not fully exploit the information naturally available during the verification process.

The availability of multiple probe frames fundamentally changes the landscape of morphing attacks, with important implications for MAD systems from both adversarial and defensive perspectives.

For attackers, the sequential nature of video acquisition offers a clear advantage: unlike single-image verification, a video stream effectively provides repeated comparison attempts across all frames. In many automated border control and identity verification systems, a single successful frame-level match may be sufficient to satisfy the verification threshold and grant unauthorized access [23, 26]. As a result, even a morphed image of moderate quality may yield a high similarity score in at least a subset of frames, bypassing detection if the system relies on independent frame-level decisions rather than aggregating evidence over time.

From a defensive standpoint, videos provide richer and more diverse evidence than a single image, enabling more robust detection. The temporal dimension introduces natural variability in face dynamics, camera behavior, and environmental conditions: consecutive frames may exhibit meaningful variations in head pose, expression, illumination, depth-of-field, motion blur, and overall quality. This variability stems from both physical and technical factors. The subject may shift position, blink, rotate the head, or change expression, while camera processes such as autofocus, auto-exposure, and compression further alter frame appearance. Consequently, some frames in a sequence may show a frontal, well-lit, sharp, and unobstructed face, whereas others may suffer from motion blur, shadows, off-axis gaze, or sensor and compression noise.

The main advantage of V-MAD over classical D-MAD is that the decision no longer depends on a single probe acquisition. Instead, the system can exploit complementary evidence from multiple frames, thereby reducing the impact of the aforementioned conditions and other acquisition artifacts [36] (see Fig. 1). Indeed, the original V-MAD study showed that even simple score-level fusion strategies can improve detection performance with respect to selecting a single frame [7]. Moreover, quality-aware strategies can further support the selection or weighting of more reliable frames by incorporating Face Image Quality Assessment (FIQA) scores into the aggregation process [7, 49].

In the literature, only limited previous works investigated multiple acquisitions or multi-camera settings for morphing attack detection [51, 52]. However, these works do not fully address the document-to-video formulation considered in V-MAD, where a single document image is compared against a live sequence acquired during an operational verification attempt. Therefore, V-MAD represents a natural evolution of D-MAD toward more realistic biometric deployment scenarios, opening the way to robust frame aggregation, quality-aware decision strategies, temporal consistency analysis, and video-oriented deep learning solutions.

Building upon this framework, the "*Every Frame You*

Take: The Video-based Morphing Attack Detection (V-MAD) Challenge"[1] was organized in conjunction with IEEE/IAPR International Joint Conference on Biometrics (IJCB) 2026 with the goal of fostering research on sequence-based morphing attack detection under realistic operational conditions. The competition introduces a novel benchmark specifically designed for video-based MAD, including both real and synthetic acquisition scenarios characterized by pose variations, illumination changes, occlusions, and heterogeneous acquisition conditions. Unlike previous MAD evaluations, the challenge explicitly frames morphing attack detection as a sequence-level biometric security problem, encouraging participants to investigate video-oriented deep learning strategies.

Summarizing, this paper presents an overview of the competition, including the adopted datasets, evaluation protocol, participating solutions, and results. Furthermore, we discuss the main methodological trends emerging from the challenge and highlight future research directions toward robust and operationally realistic V-MAD systems.

2. Related Work

2.1. Face Morphing Attacks

Face morphing consists of generating a synthetic facial image by blending the biometric characteristics of two or more different identities into a single image. Within the context of electronic Machine-Readable Travel Documents (eMRTDs), morphing attacks represent a severe security vulnerability because a morphed document image may successfully match multiple individuals while remaining visually plausible [22]. Previous studies demonstrated that morphing attacks can deceive both human inspectors and commercial face recognition systems, especially in unconstrained operational scenarios [45, 47].

The growing accessibility of generative artificial intelligence techniques, including GANs [28] and Diffusion Models [44], has further increased the realism and diversity of morphed images. In addition, modern retouching and restoration techniques [8, 17] can effectively suppress visible artifacts that are typically exploited by traditional detection algorithms [11]. These developments have considerably increased the difficulty of the task and motivated the need for more robust and realistic detection paradigms.

2.2. Synthetic Data for Morphing Attack Detection

The growing demand for large-scale and privacy-compliant biometric datasets has progressively increased interest in the use of synthetic facial data for both face recognition and presentation attack detection research [10, 20].

Indeed, synthetic data generation offers several advantages, including reduced privacy concerns, improved demographic balancing, controllable acquisition conditions, and the possibility of generating large amounts of annotated data at limited cost [14, 18].

Recent studies investigated the effectiveness of synthetic facial images for improving the robustness and generalization capabilities of biometric systems. For instance, the SDFR [50] and FRCSyn-onGoing [41] initiatives highlighted the potential of synthetic data and their combination with real data to improve FRs performance across heterogeneous operational conditions. Similarly, synthetic identities and AI-generated facial images have recently been explored within the morphing attack detection domain to alleviate the limited availability of operational morphing datasets and facilitate reproducible benchmarking [12, 30].

At the same time, the increasing realism of generative models introduces additional challenges for biometric security. Modern synthetic face generation approaches based on GANs and Diffusion Models can produce highly realistic identities and morphing attacks that substantially reduce the effectiveness of traditional artifact-based MAD methods [28, 44]. Consequently, evaluating MAD systems under synthetic acquisition conditions is becoming increasingly important to assess their robustness.

Motivated by these considerations, the V-MAD Challenge includes not only real operational acquisition scenarios, but also a fully synthetic benchmark, specifically designed to investigate the role of synthetic identities, simulated acquisition variability, and AI-generated content in video-based morphing attack detection.

3. The Challenge

3.1. Datasets

At the time of writing, no public datasets for the V-MAD task are available in the literature. Then, the first step of the competition is to acquire and create a common dataset.

A dataset suitable for V-MAD requires a specific structure than databases commonly used for traditional MAD: indeed, for each subject, at least one ICAO-compliant [32, 31] frontal photo must be included, usable as a document-like reference image, and one or more video sequences captured under realistic conditions, such as to simulate the behavior of FRs in an operational environment. Additional desirable properties include high-resolution frames with variability in pose, camera distance, illumination, and occlusions, together with several identities.

Based on these criteria, several public or semi-public datasets, including ChokePoint [55], PaSC [5], MBGC [43], YouTube Faces [25], IJB-A [35], and UMD Faces [3], are not fully suitable for V-MAD. ChokePoint, for instance, includes a few subjects and no ICAO-compliant frontal pho-

[1] Official challenge website: https://sites.google.com/view/vmad26competition

tos, while PaSC lacks ICAO-compliant images and has a non-immediate access procedure. Hence, two proprietary datasets, named VAMONOT-R (real) and VAMONOT-S (synthetic), have been introduced for the competition.

3.1.1 VAMONOT-R

The dataset is designed to reproduce in a controlled yet realistic way the scenario in which a document image is compared with a live sequence acquired at a security checkpoint. For each subject, a high-resolution frontal photo is collected along with video sequences recorded in two environments. Frontal images are captured with a Nokia Lumia 930 at 3024×5376 pixels, whereas videos are captured with an Intel RealSense D435i, providing RGB and depth data at 1280×720 pixels and 30 FPS. On average, videos are 14.3 ± 3.5 s long.

The acquisition setup introduces controlled variability through two environments and three capture configurations. The first environment is an entrance with mixed natural and artificial lighting, while the second relies only on artificial lighting and presents more challenging conditions. In each environment, subjects are recorded while looking directly at the camera, freely moving the head, or moving with partial facial occlusions caused by accessories such as a hat, scarf, or both. Subjects were able to wear prescription glasses in non-occluded sequences. This protocol introduces variations in distance, pose, illumination, background, and facial quality. An example is reported in Figure 2.

The final dataset includes 65 subjects, each with an ICAO-compliant frontal photo, two non-ICAO frontal photos, and six video sequences, obtained by combining the two environments and the three pose/occlusion configurations. Videos are decomposed into RGB and depth frames, for a total of about 330k frames, with an average of 800 frames per video and approximately 5k images per subject. Although the competition protocol considers only ICAO-compliant frontal images and RGB video frames, depth data and non-ICAO photos are retained in the dataset to support future studies on distance-aware V-MAD and quality-based frame selection. Finally, 325 morphed images are generated from ICAO-compliant photos by selecting subject pairs based on the cosine similarity of their facial embeddings, thereby obtaining realistic and biometrically plausible morphing attacks. The dataset is partially released[2].

3.1.2 VAMONOT-S

The dataset is the synthetic counterpart of VAMONOT-R and is built by sourcing 55 identities from the ONOT [18] synthetic dataset of ICAO-compliant facial images. The use of synthetic data is particularly advantageous, as it reduces the constraints related to real-world data privacy while explicitly controlling identity, acquisition conditions, and sample variability. In fact, recent generative models are now capable of producing highly realistic facial images, making synthetic datasets a valuable resource for evaluating biometric systems.

Figure 2. Example of VAMONOT-R data acquired for a given subject, including an ICAO-compliant image and frames extracted from video sequences under different conditions.

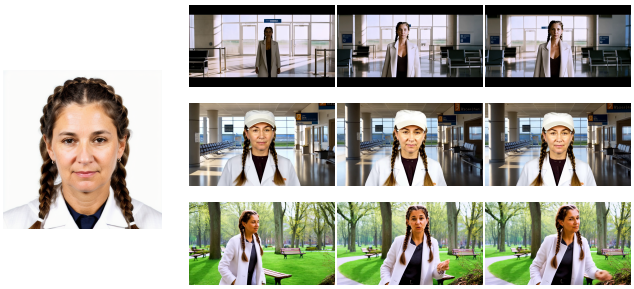

Figure 3. Example of VAMONOT-S synthetic data acquired for a given subject. The dataset includes an ICAO-compliant image and frames extracted from video sequences under different conditions.

Synthetic video sequences are generated using tools based on Hailuo AI[3] and the Minimax S2V-01 model. The videos are designed to simulate realistic verification conditions, including both a scenario similar to an airport security checkpoint, where the subject walks towards a fixed camera while maintaining a frontal view, and less constrained scenarios characterized by natural head and body movements (Fig. 3). In this way, VAMONOT-S provides a fully synthetic benchmark for V-MAD, complementary to the real-world acquisition of VAMONOT-R but maintaining its structure based on the comparison between a document image and a live sequence. This dataset is publicly released[4].

3.2. Challenge Setup

To encourage participation, we publicly release a split of both datasets for training and keep the rest sequestered for testing. In both VAMONOT-S and VAMONOT-R, subject identities are split into strictly disjoint subsets, ensuring that no morphed image is generated by combining one identity from the training set with another from the test set. Specifically, we release half of VAMONOT-S and five selected identities from VAMONOT-R, while the remaining identi-

[2] https://miatbiolab.csr.unibo.it/vamonot-r

[3] https://hailuoai.video

[4] https://miatbiolab.csr.unibo.it/vamonot-s

ties are reserved for evaluation.

To further reduce the development effort for participating teams, given the relatively short submission window, we provide multiple representations for each video. In particular, we release the original frames, as well as face-cropped frames obtained using MediaPipe [40], and also OFIQ [42] quality metrics computed for each frame. Moreover, to limit the number of frames to be processed and reduce inference time, videos are temporally downsampled before release, retaining one frame every four frames for VAMONOT-R and one frame every two frames for VAMONOT-S. This way, each video contains roughly 120 to 150 frames.

The challenge is organized into two scenarios (*Simple* and *Hard*) of increasing difficulty, defined according to the morphing algorithms used, the type of acquisition environment, and the level of face occlusion and head poses. The same submitted method is evaluated in both scenarios with an independent leaderboard.

In the *Simple* scenario, morphed images are generated using the algorithm described in [24]. Since this is a landmark-based morphing algorithm, some images may contain visible artifacts, mainly due to slight inaccuracies in the landmark detection stage. Sequences are acquired in a scenario with mostly uniform lighting, and all subjects present a fully visible face without any garment. In half of the sequences, the head pose is controlled and mainly frontal, while in the other half, slow rotations are present. The released training subsets of both datasets contain only morphs generated with this morphing algorithm, allowing participants to iterate on and tune their solutions.

Conversely, in the *Hard* scenario, morphed images are generated using two algorithms, namely those described in [4] and [19], which are landmark-based and diffusion-based morphing algorithms, respectively. These algorithms are not included in the released training data and are therefore unseen by the participating teams. The sequences are acquired in more challenging conditions, *i.e.*, in an environment with low artificial light where subjects present different garments and head poses. Since models are evaluated on this scenario only after the submission window closes, teams cannot iteratively adapt their methods to these morphing algorithms. This setting is designed to assess the cross-algorithm robustness and generalization capability of the submitted solutions.

The number of bona fide (BF) and morphing (M) attack attempts that are used for both training and evaluation is reported in Table 1, split by both scenario and dataset.

Split	Scenario	Dataset	# BF	# M
Training	Simple	VAMONOT-S	87	327
		VAMONOT-R	20	68
Testing	Simple	VAMONOT-S	78	294
		VAMONOT-R	231	863
	Hard	VAMONOT-S	165	1233
		VAMONOT-R	356	2709

Table 1. Number of bona fide and morphing attack attempts used for evaluation, split by both scenario and dataset.

3.3. Metrics

The competition measures the performance of the submitted methods by leveraging common metrics for the MAD task. In particular, we report the Bona fide Sample Classification Error Rate (BSCER), which quantifies the percentage of genuine samples that are falsely classified as morphed. Conversely, the Morphing Attack Classification Error Rate (MACER) denotes the fraction of morphed images that are incorrectly identified as genuine. In the MAD literature, BSCER is usually evaluated at predefined MACER operating points. Following this convention, we report $\text{BSCER}_{0.1}$ ($\text{B}_{0.1}$), $\text{BSCER}_{0.05}$ ($\text{B}_{0.05}$), and $\text{BSCER}_{0.01}$ ($\text{B}_{0.01}$), representing the minimum BSCER achievable while constraining MACER to be at or below 10%, 5%, and 1%, respectively. Additionally, to obtain a compact indicator of overall performance, we report the Equal Error Rate (EER), defined as the operating point at which the BSCER and MACER curves intersect.

To establish the final ranking on the leaderboard based on a single metric, the competition relies on an aggregate measure referred to as the Weighted Average Error across Datasets (WAED) [6], which combines the main performance indicators across both datasets into a single scalar value, providing a more balanced view of overall system quality than any single metric alone.

We define WAED as:

$$\text{WAED} = \sum_{E \in \mathcal{E}} \sum_{D \in \mathcal{D}} w_D w_E E(D) \tag{1}$$

where $E(D)$ is the value of the error indicator E measured on the dataset D; w_E is a weighting factor assigned to each error indicator to focus on the most relevant ones for a real-world scenario (*e.g.* $\text{B}_{0.01}$). The weights considered for the WAED metric computation are chosen arbitrarily, by assigning the majority of the weight to the closest real-world operating point (*i.e.* $\text{B}_{0.01}$, $w_E = 0.4$), followed by the EER ($w_E = 0.3$), and finally the other two chosen operating points (*i.e.* $\text{B}_{0.05}$, $w_E = 0.2$ and $\text{B}_{0.1}$, $w_E = 0.1$); w_D is a dataset weight that attempts to capture the dataset complexity, quantified in our experiments by the similarity of the morphed images to the two contributing subjects. In particular, we arbitrarily assign a weight of $w_D = 0.33$ to VAMONOT-S, which is generally simpler and partially available for training, whereas VAMONOT-R is given a weight of $w_D = 0.67$, being the case more challenging and, more importantly, closer to real operating conditions.

4. Evaluation

4.1. Description of Submissions

A total of 25 systems were submitted to the challenge by a heterogeneous set of participants, including university research groups, industrial teams, and independent researchers. This diversity reflects the growing interest surrounding morphing attack detection as an increasingly relevant problem at the intersection of biometrics, multimedia forensics, and trustworthy AI.

Given the large number of submissions and the strong methodological overlap observed among several approaches, the following discussion focuses primarily on the top-performing methods across the two challenge scenarios (see Tab. 2). In particular, we summarize the main architectural choices and processing strategies adopted by the most competitive solutions, highlighting both their common design principles and their distinguishing characteristics.

MAD	Affiliation	Country
thisisananya	Eli Lilly and Company	India
Idiap-BSP	Idiap Research Institute	Switzerland
ArogyaPandit	Arogyapandit Private Limited	India
Dragons	University of Ljubljana	Slovenia
SBU-ML	St. Bonaventure University	USA
TCDML	Trinity College Dublin	Ireland
GHSYS	University College Cork	Ireland
AADE, smartALC, TCVG, FactIT, KLsys, IDECVG	Independent researcher	

Table 2. Top-performing MAD systems across the two challenge scenarios and related affiliations.

thisisananya (#1 Simple, #8 Hard). The approach proposed by team thisisananya uses a weighted ensemble of EfficientNet-B0 [53] based document-video attention models. Each model uses the document image as the reference input and learns to combine information from multiple cropped video frames in relation to that document. For each attempt, the video is represented with up to eight frames cropped on the face: if more frames are available, eight are sampled uniformly across the sequence; if fewer are available, the remaining ones are padded with zero tensors. When no cropped frames are available, the system returns a score of 0.5, assuming there is no reliable video input for a confident decision. During training, OFIQ-based [1, 42] frame quality filtering was applied before sampling. Frames with unified quality scores below 40 were discarded to reduce the influence of blur, poor illumination, pose variation, and unreliable detections. This reduced noisy training inputs and encouraged the model to learn identity-related morphing cues rather than artifacts from poor-quality crops, while the submitted inference pipeline remained simpler by using the provided cropped frames directly. The architecture uses a shared EfficientNet-B0 encoder for both document images and video frames, producing 1280-dimensional features in a common representation space. Most of the encoder is frozen, while the final four EfficientNet feature blocks are fine-tuned. The fusion module that follows the encoders is an 8-head multi-head attention layer [54]. The document feature is used as the query, and the video-frame features are used as keys and values. Instead of treating all sampled frames equally, the model can assign greater importance to frame features that are most informative for the document image. This direction was chosen because morphing cues may be subtle and may not appear consistently across all frames. The attended 1280-dimensional representation is finally passed to a lightweight binary classifier. Models were trained using subject-disjoint cross-validation to support generalization to unseen identities. The final ensemble includes five checkpoints that minimize cross-entropy and four that minimize focal loss [37], the latter being used to improve sensitivity to difficult morphing cases. At inference, each checkpoint is evaluated with five deterministic test-time augmentations, producing 45 morph-class probabilities per attempt. The final score is a weighted average, assigning weight 1.0 to cross-entropy checkpoints and 2.0 to focal loss checkpoints.

Idiap-BSP (#1 Simple, #11 Hard). Team Idiap-BSP proposes a dual-branch architecture to detect whether an ID document has been morphed by comparing it with a live, trusted video of the presenting subject. Given an ID document image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and $\{\mathbf{F}_t\}_{t=1}^{T}$ frames from the video, process both modalities through independent but weight-shared Vision Encoders. Each branch employs an InternVL3-1B [57] vision encoder (0.3B parameters) to extract fixed-dimensional embeddings. The ID document is encoded once to produce $\mathcal{E}_{\text{doc}} \in \mathbb{R}^{1024}$, while each video frame is encoded independently. Frame-level embeddings are aggregated (mean pooling) to obtain $\mathcal{E}_{\text{frames}} \in \mathbb{R}^{1024}$. The two embeddings are then concatenated and passed to a multilayer perceptron (MLP) that outputs a binary prediction, *i.e.*, bona fide or morphed. The core idea is to treat the live video as a trusted identity reference. Using a shared pretrained vision backbone across both branches enforces a unified embedding space, enabling direct comparison between document and frame embeddings. Aggregating independently encoded frames preserves temporal diversity while avoiding explicit sequence modeling, improving robustness to pose, expression, and illumination variations. Freezing the encoder reduces overfitting while leveraging strong pretrained visual priors. The vision encoder is kept frozen, and only the MLP classifier head is trained. Their system was trained for 20 epochs, using a batch size of 16 and a learning rate of 5×10^{-4}.

ArogyaPandit (#3 Simple). The method submitted by team ArogyaPandit uses a quality-aware dual-branch late-fusion strategy. They treat V-MAD as a document-to-video comparison problem rather than independent frame classification, combining document artifact evidence with identity-consistency evidence from the live sequence. More in detail, in the artifact branch, the document image is processed by an EfficientNet-B0 [53] binary classifier with sigmoid output. The artifact detector was initialized from ImageNet-pretrained weights [15] and trained to separate bona fide and morphed face images using VAMONOT labelled samples together with compressed ChiMo [6] morph samples. Training used 224×224 resizing, ImageNet normalization, random horizontal flipping, brightness/contrast augmentation, binary cross-entropy loss, and AdamW optimization. In the identity branch, they use ArcFace [16] to extract L2-normalized 512-dimensional embeddings from the document image and detected video faces. Frames are filtered using face detection confidence and OFIQ-derived [42] quality indicators, including unified quality, sharpness, yaw, and pitch. Frames with yaw or pitch above 25 degrees, unified quality below 15, or detector confidence below 0.8 are discarded. Up to five accepted frame embeddings are weighted by a composite score combining OFIQ values, detector confidence, and a logistic pose penalty, then aggregated into a quality-weighted centroid representing the live subject. Finally, identity evidence is computed as the cosine distance between the document embedding and the live-video centroid. A raw score is formed as 80% document artifact probability and 20% identity distance, then calibrated with a temperature-scaled sigmoid.

Dragons (#4 Simple). Team Dragons's submission combines complementary S-MAD and D-MAD evidence through a calibrated additive fusion rule. The document face is first detected and aligned using the SCRFD-10G keypoint detector [29]. The resulting crop is processed by SelfMAD++ [33], a self-supervised foundation-model-based S-MAD classifier trained to recognize generic synthetic morphing artifacts, producing a morph posterior sMAD $\in [0, 1]$. In parallel, we compute a face-recognition-based D-MAD cue between the document image and the bona-fide video. The document crop is embedded with LVFace-L [56], pretrained using large-scale face-recognition data [2], to obtain an identity descriptor e_{doc}. For the video, 30 frames are uniformly sampled and processed by the same SCRFD + LVFace-L pipeline. The frame-level embeddings are ranked by face-detection confidence, and the top five are averaged and L2-normalized to obtain e_{video}. The identity discrepancy is measured as $d_{FR} = 1 - \cos(e_{doc}, e_{video})$. The final score is obtained as $s_{final} = \text{clip}(\text{sMAD} + W \cdot \text{clip}(\frac{d_{FR} - \text{PLO}}{\text{PHI} - \text{PLO}}, 0, 1) - 0.5, 0, 1)$, where PLO is the 10th percentile of bona-fide d_{FR} scores and PHI is the mean morph distance.

This asymmetric calibration makes the fusion conservative when decreasing the morph posterior, requiring very strong bona fide identity agreement, while allowing larger identity discrepancies to increase the final morph score. The fusion weight W and calibration thresholds were tuned on the training split from both datasets.

SBU-ML (#2 Hard), all remaining discussed teams. The methods proposed by SBU-ML and all other teams propose two-stream classification frameworks that jointly analyze the document image and a selection of eight frames from the live video sequence, leveraging an EVA-02-based [21] Vision Transformer pipeline. All systems process one static image together with eight video frames sampled at regular intervals throughout the video: sequences shorter than eight frames are zero-padded, tensors are resized to 224×224 pixels and normalized with ImageNet [15] mean and variance, and training samples are augmented with random horizontal flips and brightness-contrast variations. Architecturally, all teams use a two-stream EVA-02 design in which both streams share the pretrained transformer backbone. In all cases, the image stream produces a compact spatial representation, while the video stream generates frame-level embeddings that are averaged over time. The image and video representations are then concatenated and passed to a simple classification head consisting of a 512-unit linear layer, batch normalization, ReLU activation, dropout regularization, and a final linear layer. Team SBU-ML avoids data leakage by splitting the dataset at the subject level, ensuring that samples from the same individual appear exclusively in the training or validation subset. The same team also uses automatic mixed-precision and gradient scaling to improve efficiency and numerical stability, performs a hyperparameter search, trains with AdamW [39], applies cosine annealing [38], and uses early stopping with best-weight retention. Training details are fully available only for the SBU-ML team; for the remaining teams, these optimization and evaluation details are unknown.

4.2. Results

4.2.1 Simple Scenario

Results obtained on the *Simple* scenario are reported in Table 3. Overall, the competition revealed remarkably strong performance across most participating methods, with several approaches achieving near-perfect detection capability. In particular, the top-ranked systems obtained WAED values close to zero (*e.g.*, ≤ 0.1 for the four top-performing systems), demonstrating that the proposed benchmark can be effectively addressed when acquisition conditions and morph generation processes remain relatively aligned with the training distribution.

To facilitate comparison, following [7], we include an

#	Team Name	WAED	EER_S	EER_R
1	thisisananya	0.00%	0.00%	0.00%
1	Idiap-BSP	0.00%	0.00%	0.00%
3	ArogyaPandit	0.08%	0.00%	0.39%
4	Dragons	0.77%	1.15%	0.84%
5	AADE	0.80%	0.00%	1.68%
-	*Baseline*	*65.1%*	*26.9%*	*28.1%*

Table 3. Results on the *Simple* scenario for methods that obtained WAED $\leq 1\%$ for VAMONOT-R (EER_R) and VAMONOT-S (EER_S.)

adapted D-MAD system [48] as a baseline, where the final sequence score is determined as the average score across all document-frame pairs.

A common trend among the best-performing submissions is the adoption of dual-branch architectures jointly processing the document image and the associated video sequence. Most approaches relied on pretrained visual backbones combined with lightweight temporal aggregation strategies, typically based on frame-level feature averaging or attention-based fusion mechanisms. Interestingly, explicit temporal modeling through recurrent or transformer-based sequence analysis was less prominent than expected, suggesting that robust frame-level representations and effective aggregation strategies may already provide sufficient discriminative information under moderate operational variability.

Another recurring design choice concerns the use of quality-aware frame selection or filtering procedures. More in detail, only team ArogyaPandit relied on pre-computed frame-level OFIQ metrics (in particular, unified quality score, sharpness, and head pose pitch and yaw) to compute a quality score to weigh the single frames' contributions; all other teams that introduced frame-level weighing either relied on learned attention mechanisms or confidence-based frame ranking. This trend further confirms the importance of acquisition quality in operational V-MAD scenarios.

Furthermore, we observe that for the *Simple* scenario there is no clear winner in terms of face preprocessing: indeed, teams thisisananya and AADE rely on pre-cropped frames obtained via MediaPipe [40], rather than implementing their own face detection pipeline; conversely, teams ArogyaPandit and Idiap-BSP process raw frames directly, without any face detection step run beforehand, while team Dragons relies on face detection and landmark extraction with dlib [34].

The relatively homogeneous high performance observed in this scenario also suggests that current deep learning approaches are already capable of effectively exploiting the complementary information provided by document images and video sequences when the evaluation conditions remain sufficiently close to the training distribution. Under these circumstances, the main differences between methods appear to be more related to architectural details than to fundamentally different modeling assumptions.

#	Team Name	WAED	EER_S	EER_R
1	TCDML	2.58%	1.80%	3.36%
2	SBU-ML	2.66%	3.60%	1.67%
3	TCVG	3.95%	1.88%	3.36%
4	GHSYS	7.34%	3.64%	4.00%
5	smartALC	8.67%	3.60%	5.44%
6	FactIT	9.47%	3.68%	5.04%
7	KLsys	9.57%	3.60%	7.88%
8	thisisananya	11.4%	9.09%	2.30%
9	IDECVG	13.5%	1.30%	9.27%
10	AADE	15.3%	1.80%	8.42%
11	Idiap-BSP	15.3%	3.89%	2.23%
-	*Baseline*	*61.3%*	*31.5%*	*23.9%*

Table 4. Results on the *Hard* scenario for methods that obtained WAED $\leq 16\%$.

4.2.2 Hard Scenario

Results on the *Hard* scenario, reported in Table 4, reveal a substantially more challenging evaluation setting. Compared to the previous scenario, all methods experience a visible performance degradation, although with markedly different robustness levels. The best-performing method, TCDML, achieves a WAED of 2.58%, followed closely by SBU-ML and TCVG, indicating that robust generalization under unseen conditions remains feasible but significantly more difficult.

One of the most interesting observations emerging from this scenario is the strong methodological convergence among the top-performing solutions. While the *Simple* scenario exhibited a relatively heterogeneous set of approaches, the highest-ranked methods in the *Hard* scenario shared the same underlying design principles, and progressively converged toward highly similar architectural paradigms based on dual-stream document-video processing, pretrained visual backbones, and lightweight feature aggregation strategies. Furthermore, the vast majority of the best-performing teams in this scenario relied almost exclusively on pre-cropped frames using MediaPipe [40], without leveraging the quality metrics provided by OFIQ [42].

This behavior is itself an important outcome of the competition. The increased complexity of the *Hard* scenario, characterized by previously unseen morphing conditions and stronger acquisition variability, appears to significantly reduce the effectiveness of highly specialized or dataset-dependent optimization strategies. As a consequence, only a relatively narrow family of robust and generalizable pipelines consistently maintained competitive per-

#	Team Name	WAED$_S$	WAED$_H$	ΔWAED
1	TCDML	3.45%	2.58%	-0.86%
2	KLsys	9.19%	9.57%	+0.38%
3	SBU-ML	1.98%	2.66%	+0.68%
4	TCVG	3.25%	3.95%	+0.70%
5	smartALC	6.03%	8.67%	+2.64%
6	GHSYS	3.05%	7.34%	+4.30%
7	FactIT	4.90%	9.47%	+4.57%
8	IDECVG	6.93%	13.5%	+6.52%
9	thisisananya	0.00%	11.4%	+11.4%
10	AADE	0.80%	15.3%	+14.5%
11	Idiap-BSP	0.00%	15.3%	+15.3%
12	Dragons	0.77%	24.2%	+23.4%
13	ArogyaPandit	0.08%	71.8%	+71.7%

Table 5. Absolute difference in WAED obtained by comparing Simple (WAED$_S$) and Hard (WAED$_H$) scenarios.

formance across heterogeneous conditions.

Another relevant observation concerns the increased importance of generalization-oriented training procedures. Methods that rely on stronger regularization, subject-disjoint partitioning, pretrained large-scale visual encoders, and conservative aggregation strategies generally exhibited greater stability under the *Hard* scenario. Conversely, several approaches that achieved near-perfect performance in the *Simple* track experienced substantially greater degradation when evaluated under previously unseen conditions.

At the same time, the relatively limited performance gap among the best-performing submissions indicates that future improvements in realistic V-MAD scenarios may increasingly depend on advances in robustness-oriented training, synthetic data utilization, domain generalization strategies, and temporal consistency modeling rather than on entirely new architectural designs.

4.3. Cross-Scenario Robustness Analysis

Table 5 reports the absolute performance variation between the *Simple* and *Hard* scenarios. The results highlight substantial differences in cross-scenario robustness among the submitted methods.

Among the most stable approaches, TCDML achieved the best overall robustness, showing even a slight performance improvement in the *Hard* scenario. Similarly, SBU-ML, TCVG, and KLsys exhibited only limited performance degradation, suggesting that their pipelines successfully captured more generalizable representations rather than overfitting to the specific properties of the original benchmark. Conversely, several methods that achieved extremely low error rates in the *Simple* scenario suffered considerably larger degradations in the Hard setting. In particular, thisisananya, Idiap-BSP, and ArogyaPandit obtained near-perfect results under the original protocol but experienced substantial increases in WAED when evaluated on the harder benchmark. This behavior suggests that excellent performance under controlled conditions does not necessarily translate into operational robustness under realistic distribution shifts.

Specifically, the *Simple*-to-*Hard* degradation can be interpreted in light of the different nature of the two scenarios. In the Simple setting, the morphing algorithm used at test time is represented in the released training data, and the acquisition conditions are comparatively controlled. As a result, near-perfect performance may partly reflect the ability to exploit cues associated with known morphing artifacts and dataset-specific acquisition regularities.

Conversely, the *Hard* scenario introduces a stronger distribution shift, including unseen landmark-based and diffusion-based morphing algorithms, low artificial illumination, stronger pose variability, and garments or partial occlusions. Therefore, methods relying on algorithm-specific artifacts or overly tuned decision boundaries may suffer substantial degradation. This explains why some systems that achieve near-perfect *Simple* results show reduced robustness in the *Hard* setting, and reinforces the need for cross-algorithm and cross-condition evaluation protocols in operational V-MAD. These findings further reinforce the central role of generalization in operational V-MAD. The *Hard* scenario appears to significantly reduce the benefits of dataset-specific optimization strategies while favoring architectures and training procedures capable of maintaining stable performance across heterogeneous acquisition conditions and unseen morph generation processes.

Overall, the challenge demonstrates that video-based MAD constitutes a highly promising direction for operational morphing attack detection. At the same time, the results clearly highlight that robust cross-condition generalization remains one of the primary open challenges for future V-MAD research.

4.4. Computational Analysis

Figure 4 reports the average inference runtime per attempt measured across datasets and tasks. Submissions were run on a server with an Intel Xeon W5-3423 24-core CPU, 64 GB of RAM, and an Nvidia RTX A4000 GPU with 16 GB of VRAM. Overall, the majority of submissions exhibit relatively low computational requirements, with most approaches operating below 0.25 s per attempt. This suggests that many V-MAD pipelines are already compatible with near-real-time deployment constraints typically encountered in operational biometric verification systems.

Among the fastest approaches, GHSYS, SBU-ML, and FactIT achieve inference times of $0.13 - 0.14$ s per attempt while maintaining competitive detection performance, particularly in the *Hard* scenario. This highlights the effectiveness of lightweight aggregation strategies combined with efficient pretrained visual backbones.

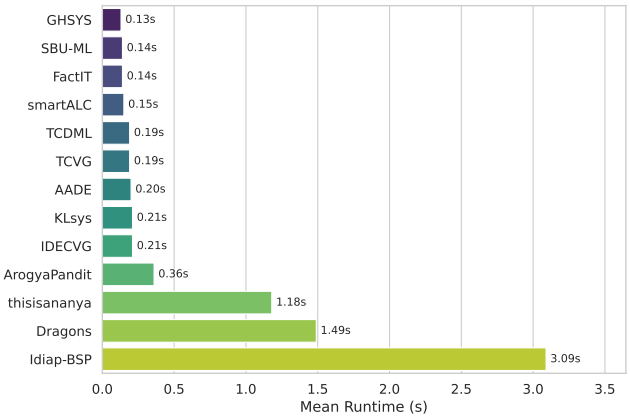

Figure 4. Mean inference runtime per attempt across datasets and tasks measured in seconds.

In contrast, methods that rely on heavier architectures or more computationally demanding processing stages exhibit substantially larger runtimes. In particular, Idiap-BSP requires more than 3 seconds per attempt due to the use of the InternVL3-based multimodal encoder, while Dragons and thisisananya also present noticeably higher inference costs. Although these approaches still achieve good performance, the results highlight the practical trade-off between computational complexity and operational deployability.

Interestingly, the six best-performing methods on the *Hard* scenario maintain a favorable balance between robustness and efficiency, requiring only ≤ 0.2 seconds per attempt. More generally, the competition results suggest that robust V-MAD performance does not necessarily require extremely large architectures or expensive temporal modeling strategies. Instead, carefully designed lightweight aggregation mechanisms combined with strong pretrained representations may already provide an effective trade-off between computational efficiency and detection robustness.

Overall, the runtime analysis further confirms the practical viability of video-based morphing attack detection for real operational scenarios, while simultaneously highlighting the importance of balancing robustness, efficiency, and scalability in future V-MAD research.

5. Conclusions and Future Works

The IJCB 2026 V-MAD Challenge introduced the first competition specifically dedicated to Video-based Morphing Attack Detection under operationally realistic conditions. The proposed benchmark included two complementary evaluation settings: a *Simple* scenario, where acquisition and morph generation conditions remain relatively aligned with the training distribution, and a significantly more challenging *Hard* scenario designed to evaluate robustness under stronger acquisition variability and previously unseen morphing conditions.

Overall, the competition results demonstrate that leveraging multiple video frames can substantially improve robustness compared to traditional frame-level MAD approaches, especially when combined with quality-aware frame selection and pretrained visual representations. At the same time, the *Hard* scenario clearly highlighted that robust generalization across heterogeneous operational conditions remains a major open challenge for current V-MAD systems. One of the most interesting outcomes of the competition is the strong methodological convergence observed among the top-performing solutions in the *Hard* scenario. Despite the diversity of implementation choices, the best-performing methods progressively converged toward a relatively consistent family of dual-stream document-video architectures that rely on lightweight temporal aggregation and robust pretrained visual backbones. This suggests that the V-MAD task is beginning to identify a set of effective architectural priors for operational morphing attack detection. More generally, the challenge highlights how the increasing realism of AI-generated facial content is rapidly transforming the detection of morphing attacks into a central research topic at the intersection of biometrics, multimedia forensics, trustworthy AI, and operational security. The strong participation and the quality of the submitted methods further confirm the relevance and urgency of developing robust detection systems capable of operating under realistic deployment conditions.

Future work will concentrate on expanding both the challenge framework and the underlying datasets. On the data side, we plan to increase the number of subjects and the volume of video sequences, acquiring them under more diverse conditions and quality levels, addressing the current limitation posed by the relatively small size of the two proposed datasets and the resulting constraints on possible performance analyses of participating solutions. In parallel, upcoming editions of the challenge will examine more complex operational scenarios, including stronger domain shifts, cross-device acquisition, temporal consistency analysis, and the integration of multimodal information such as depth data. We will also design more advanced evaluation protocols, with a focus on robustness, demographic fairness, explainability, computational efficiency, and long-term operational reliability in real-world scenarios.

Acknowledgments

This project received funding from the European Union's Horizon Europe research and innovation program under Grant Agreement No.101121280. This text reflects only the author's views, and the commission is not liable for any use that may be made of the information contained therein. We also acknowledge the support of the European funds from the Emilia-Romagna Region under the Fse+ 2021-2027 program for Lorenzo Pellegrini.

References

[1] Iso/iec 29794-5:2025: Information technology – biometric sample quality – part 5: Face image data, 2025.

[2] X. An, X. Zhu, Y. Gao, Y. Xiao, Y. Zhao, Z. Feng, L. Wu, B. Qin, M. Zhang, D. Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1445–1449, 2021.

[3] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa. Umdfaces: An annotated face dataset for training deep networks. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, page 464–473. IEEE Press, 2017.

[4] I. Batskos and L. Spreeuwers. Improving fully automated landmark-based face morphing. In *2024 12th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2024.

[5] J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–8, 2013.

[6] G. Borghi, N. Di Domenico, A. Franco, M. Ferrara, and D. Maltoni. Revelio: A modular and effective framework for reproducible training and evaluation of morphing attack detectors. *IEEE Access*, 11:120419–120437, 2023.

[7] G. Borghi, A. Franco, N. Di Domenico, M. Ferrara, and D. Maltoni. V-mad: Video-based morphing attack detection in operational scenarios. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2024.

[8] G. Borghi, A. Franco, G. Graffieti, and D. Maltoni. Automated artifact retouching in morphed images with attention maps. *IEEE Access*, 9:136561–136579, 2021.

[9] G. Borghi, E. Pancisi, M. Ferrara, and D. Maltoni. A double siamese framework for differential morphing attack detection. *Sensors*, 21(10):3466, 2021.

[10] F. Boutros, V. Struc, J. Fierrez, and N. Damer. Synthetic data for face recognition: Current state and future prospects. *Image and Vision Computing*, 135:104688, 2023.

[11] S. Concas, S. M. La Cava, A. Panzino, E. Masala, G. Orrú, and G. L. Marcialis. Deceptive beauty: Evaluating the impact of beauty filters on deepfake and morphing attack detection. In *2025 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*, pages 61–66. IEEE, 2025.

[12] N. Damer, C. A. F. López, M. Fang, N. Spiller, M. V. Pham, and F. Boutros. Privacy-friendly synthetic data for the development of face morphing attack detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1606–1617, 2022.

[13] J. S. Del Rio, D. Moctezuma, C. Conde, I. M. de Diego, and E. Cabello. Automated border control e-gates and facial recognition systems. *computers & security*, 62:49–72, 2016.

[14] R. Delussu, L. Putzu, and G. Fumera. Synthetic data for video surveillance applications of computer vision: A review. *International Journal of Computer Vision*, 132(10):4473–4509, 2024.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[16] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[17] N. Di Domenico, G. Borghi, A. Franco, and D. Maltoni. Face restoration for morphed images retouching. In *2024 12th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2024.

[18] N. Di Domenico, G. Borghi, A. Franco, and D. Maltoni. Onot: a high-quality icao-compliant synthetic mugshot dataset, 2024.

[19] N. Di Domenico, A. Franco, M. Ferrara, and D. Maltoni. Arc2morph: Identity-preserving facial morphing with arc2face. In *2024 IEEE 18th international conference on automatic face and gesture recognition (FG)*. IEEE, 2026.

[20] M. Fang, M. Huber, J. Fierrez, R. Ramachandra, N. Damer, A. Alkhaddour, M. Kasantcev, V. Pryadchenko, Z. Yang, H. Huangfu, et al. Synfacepad 2023: Competition on face presentation attack detection based on privacy-aware synthetic training data. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11. IEEE, 2023.

[21] Y. Fang, Q. Sun, X. Wang, T. Huang, X. Wang, and Y. Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024.

[22] M. Ferrara, A. Franco, and D. Maltoni. The magic passport. In *2014 IEEE International Joint Conference on Biometrics*, pages 1–7. IEEE, 2014.

[23] M. Ferrara, A. Franco, D. Maltoni, and C. Busch. Morphing attack potential. In *2022 International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2022.

[24] M. Ferrara, A. Franco, D. Maltoni, et al. Decoupling texture blending and shape warping in face morphing. In *BIOSIG*, pages 25–33, 2019.

[25] C. Ferrari, S. Berretti, and A. Del Bimbo. Extended youtube faces: a dataset for heterogeneous open-set face identification. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3408–3413, 2018.

[26] Frontex. Automated biometric border crossing systems based on electronic passports and facial recognition: Rapid and smartgate. Technical report, European Agency for the Management of Operational Cooperation at the External Borders of the Member States of the European Union, 2010.

[27] D. Ghiani, J. D. R. Chivata, S. Lilliu, S. M. La Cava, M. Micheletto, G. Orrú, F. Lama, and G. L. Marcialis. Fragile watermarking for image certification using deep steganographic embedding. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2025.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[29] J. Guo, J. Deng, A. Lattas, and S. Zafeiriou. Sample and computation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714*, 2021.

[30] M. Huber, F. Boutros, A. T. Luu, K. Raja, R. Ramachandra, N. Damer, P. C. Neto, T. Gonçalves, A. F. Sequeira, J. S. Cardoso, et al. Syn-mad 2022: Competition on face morphing attack detection based on privacy-aware synthetic training data. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2022.

[31] International Standards Organization. ISO/IEC 19794-5 — Information technology — Biometric data interchange formats — Part 5: Face image data. Standard, International Organization for Standardization, 2011.

[32] International Standards Organization. ISO/IEC 39794-5 — Information technology — Extensible biometric data interchange formats — Part 5: Face image data. Standard, International Organization for Standardization, 2019.

[33] M. Ivanovska, L. Todorov, P. Peer, et al. Selfmad++: Self-supervised foundation model with local feature enhancement for generalized morphing attack detection. *Information fusion*, page 103921, 2025.

[34] D. E. King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.

[35] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1939, 2015.

[36] S. M. La Cava, G. Orrù, M. Drahansky, G. L. Marcialis, and F. Roli. 3d face reconstruction: the road to forensics. *ACM Computing Surveys*, 56(3):1–38, 2023.

[37] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, Oct 2017.

[38] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[39] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[40] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, et al. Mediapipe: A framework for perceiving and processing reality. In *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)*, volume 2019, page 2. Long Beach, CA, 2019.

[41] P. Melzi, R. Tolosana, R. Vera-Rodriguez, M. Kim, C. Rathgeb, X. Liu, I. DeAndres-Tame, A. Morales, J. Fierrez, J. Ortega-Garcia, et al. Frcsyn-ongoing: Benchmarking and comprehensive evaluation of real and synthetic data to improve face recognition systems. *Information Fusion*, 107:102322, 2024.

[42] J. Merkle, C. Rathgeb, B. Herdeanu, B. Tams, D. Lou, A. Dörsch, M. Schaubert, J. Dehen, L. Chen, X. Yin, D. Huang, A. Stratmann, M. Ginzler, M. Grimmer, and C. Busch. Open source face image quality (ofiq): Implementation and evaluation of algorithms. Technical report, German Federal Office for Information Security, 2024.

[43] P. J. Phillips, P. Flynn, W. Scruggs, A. O'Toole, D. Bolme, K. Bowyer, B. Draper, G. Givens, Y. Lui, H. Sahibzada, J. Scallan, and S. Weimer. Overview of the multiple biometrics grand challenge. pages 705–714, 06 2009.

[44] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[45] S. L. Sanna, A. Panzino, S. M. La Cava, S. Concas, L. Regano, D. Maiorca, G. L. Marcialis, and G. Giacinto. Pixel perfect or perfectly fake? exploring the robustness of digital forensic tools against facial deepfakes and morphed images. In *2025 33rd European Signal Processing Conference (EUSIPCO)*, pages 1357–1361. IEEE, 2025.

[46] U. Scherhag, C. Rathgeb, and C. Busch. Face morphing attack detection methods. In *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*, pages 331–349. Springer, 2022.

[47] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch. Face recognition systems under morphing attacks: A survey. *IEEE Access*, 7:23012–23026, 2019.

[48] U. Scherhag, C. Rathgeb, J. Merkle, and C. Busch. Deep face representations for differential morphing attack detection. *IEEE transactions on information forensics and security*, 15:3625–3639, 2020.

[49] T. Schlett, C. Rathgeb, O. Henniger, J. Galbally, J. Fierrez, and C. Busch. Face image quality assessment: A literature survey. *ACM Computing Surveys (CSUR)*, 54(10s):1–49, 2022.

[50] H. O. Shahreza, C. Ecabert, A. George, A. Unnervik, S. Marcel, N. Di Domenico, G. Borghi, D. Maltoni, F. Boutros, J. Vogel, et al. Sdfr: Synthetic data for face recognition competition. In *2024 IEEE 18th international conference on automatic face and gesture recognition (FG)*, pages 1–9. IEEE, 2024.

[51] J. M. Singh, R. Ramachandra, K. B. Raja, and C. Busch. Robust morph-detection at automated border control gate using deep decomposed 3d shape & diffuse reflectance. In *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 106–112. IEEE, 2019.

[52] J. M. Singh, S. Venkatesh, and R. Ramachandra. Robust face morphing attack detection using fusion of multiple features and classification techniques. In *2023 26th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2023.

[53] M. Tan, Q. E. Le, et al. Rethinking model scaling for convolutional neural networks. In *Proceedings of the International conference on machine learning, Long Beach, CA, USA*, volume 15, 2019.

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[55] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. Lovell. Chokepoint dataset, June 2017.

[56] J. You, S. Li, Y. Sun, J. Wei, M. Guo, C. Feng, and J. Ran. Lvface: Progressive cluster optimization for large vision models in face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11840–11849, 2025.

[57] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.