MADPromptS: Unlocking Zero-Shot Morphing Attack Detection with Multiple Prompt Aggregation

Eduarda Caldeira
Fraunhofer IGD
Darmstadt, Germany
eduarda.caldeira@igd.fraunhofer.de

Fadi Boutros
Fraunhofer IGD
Darmstadt, Germany
fadi.boutros@igd.fraunhofer.de

Naser Damer
Fraunhofer IGD and Department of Computer Science, TU Darmstadt
Darmstadt, Germany
naser.damer@igd.fraunhofer.de

Abstract

Face Morphing Attack Detection (MAD) is a critical challenge in face recognition security, where attackers can fool systems by interpolating the identity information of two or more individuals into a single face image, resulting in samples that can be verified as belonging to multiple identities by face recognition systems. While multimodal foundation models (FMs) like CLIP offer strong zero-shot capabilities by jointly modeling images and text, most prior works on FMs for biometric recognition have relied on fine-tuning for specific downstream tasks, neglecting their potential for direct, generalizable deployment. This work explores a pure zero-shot approach to MAD by leveraging CLIP without any additional training or fine-tuning, focusing instead on the design and aggregation of multiple textual prompts per class. By aggregating the embeddings of diverse prompts, we better align the model's internal representations with the MAD task, capturing richer and more varied cues indicative of bona-fide or attack samples. Our results show that prompt aggregation substantially improves zero-shot detection performance, demonstrating the effectiveness of exploiting foundation models' built-in multimodal knowledge through efficient prompt engineering. The code is publicly released: https://github.com/EduardaCaldeira/MADPromptS

CCS Concepts

• **Computing methodologies** → **Biometrics**; **Computer vision**.

Keywords

Computer Vision, Foundation Models, Morphing Attack Detection

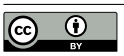
ACM Reference Format:
Eduarda Caldeira, Fadi Boutros, and Naser Damer. 2025. MADPromptS: Unlocking Zero-Shot Morphing Attack Detection with Multiple Prompt Aggregation. In *Proceedings of the 1st International Workshop & Challenge on Subtle Visual Computing (SVC '25), October 27–28, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3728425.3759909

1 Introduction

In recent years, the research community's strong focus on deep learning (DL) techniques enabled the high-paced development of high-performing systems in different domains, including face recognition (FR) [2, 15]. Despite the unquestionable benefits associated with this evolution, the same scientific advances capable of enhancing biometric systems can also be maliciously deployed to attack them [11, 18], raising concerns about their secure deployment. Some of these malicious samples are created by interpolating identity information of two or more individuals in a single face image, resulting in samples that can be verified as belonging to multiple identities by FR systems and by humans. These attacks are known as morphing attacks (MA) due to their inherent property of incorporating defining features from multiple identities, contributing to impaired FR functionality when left undetected [4, 10, 46]. The inability to efficiently detect these samples is particularly problematic in high-security applications, as they potentiate crimes such as identity theft. To mitigate such risks, various morphing attack detection (MAD) systems have been proposed in the recent years [4, 5, 12, 16, 23, 30, 39]. These algorithms aim at distinguishing MAs from authentic (bonafide) samples before they are fed to FR systems, removing malicious samples from the recognition framework at an early stage and preventing them from being considered for face verification.

Foundation models (FMs) are large-scale networks that can be trained with unlabeled data, following a self-supervised learning paradigm [27, 31, 36]. This allows FMs to be trained in massive and diverse datasets, resulting in trained models that can efficiently generalize to a wide variety of tasks [1]. Due to this property, FMs can be directly deployed for classification of samples belonging to categories that were not necessarily analyzed during their training stage (zero-shot learning), which makes them powerful tools in fields that address several tasks, such as natural language processing [3] and computer vision [27, 31, 36, 40]. While FMs have shown significant zero-shot capacity across several downstream tasks, they achieve less optimal performance when applied to domain-specific settings [40], for which an adaption to the downstream task is usually performed [6, 7, 21]. Despite allowing for a beneficial balance between pre-trained FMs in-built information and the acquisition of domain-specific knowledge [5, 33], this adaption process requires model fine-tuning, resulting in decreased computational efficiency when compared with zero-shot learning, which does not require any MAD training data. Prompt engineering has recently gained attention as an effective strategy to boost the zero-shot performance of foundation models without any additional computational burden [36]. By feeding multimodal FMs with carefully designed textual prompts describing each class, it is possible to better align the input representations with the built-in knowledge of the FM and, consequently, take better advantage of the FMs zero-shot capacity.

In this work, we explore the potential of prompt engineering for zero-shot learning MAD. In particular, we analyze how the



SVC '25, Dublin, Ireland

ACM ISBN 979-8-4007-1837-3/2025/10
https://doi.org/10.1145/3728425.3759909

utilization of multiple prompts to describe each possible output class impacts zero-shot performance, highlighting the benefits of a careful selection of the prompt sets used during inference. The obtained results show that while using a single prompt per class provides a general description of the desired label, using multiple prompts allows the incorporation of more specific characteristics, shifting the attention of the FM to a broader spectrum of details. While some sets of prompts do not contribute to increase zero-shot performance, disjoint sets that contribute positively to the FM's performance present complementary properties, with their joint utilization resulting in further MAD performance improvements. Hence, this study highlights the benefits that can arise from efficient prompt engineering strategies, providing important insights regarding the importance of correctly exploiting textual prompts to the user's advantage by leveraging FM's zero-shot learning ability to a more complete extent.

2 Related Work

This section presents an overview of recent works proposing MAD solutions, followed by a discussion of recent advances in foundation models and their applications within the biometric domain.

2.1 Morphing Attack Detection

Interpolating the identities of two or more face images in one image, such that it can be verified as belonging to those identities [4] is a major risk to many processes involving automatic or manual identity verification [19, 41]. Detecting such attacks became a major challenge given the realistic appearance and the ease of creation of such morphing samples, motivating the research community towards the development of more accurate and generalizable MAD solutions [4, 12, 16, 23, 30, 39, 45]. From an operational point of view, MAD solutions can be categorized as single image MAD strategies [4, 12, 16, 23, 26, 30, 39, 45], which base their decision solely on the inspected image, and differential MAD solutions [9], which consider a live captured image along with the inspected sample. The latter strategies generally show higher accuracy in detecting morphing attacks, since they have access to additional information to make a prediction [9, 37, 38]. However, their applicability is limited in several use cases, as it requires performing a live capture under operator supervision. Hence, several studies have developed single-image MAD systems [4, 12, 16, 23, 26, 30, 39, 45], which can be applied without the need to perform a live capture, allowing MAD in a wider range of real-world applications (e.g. analyzing standalone documents).

Recent single-image MAD works have explored diverse methods to detect morphing attacks, ranging from handcrafted features [39] to advanced deep learning techniques [5, 26, 45]. Ramachandra *et al.* [39] proposed to extract multi-scale textural descriptors and classify them using collaborative representation. Another work [12] suggested that each pixel (or block of pixels) should be individually classified as bona-fide or morphing attack, shifting away from the common global classification towards a more localized detection. Unsupervised [16] and self-supervised [26] approaches have also been proposed to tackle the MAD task. In [16], the authors trained a robust autoencoder for anomaly detection using an unsupervised self-paced learning approach. This approach identifies suspicious training samples and assigns them less importance during training, despite the datasets being polluted with morphed samples. [26] trained a self-supervised diffusion model to reconstruct bona-fide samples. While authentic samples can be easily reconstructed by this model, its ability to reconstruct morphed samples is significantly lower which results in higher reconstruction errors, allowing for detecting these malicious samples. [23] promoted the SYN-MAD 2022 competition on MAD based on synthetic training data, presenting a comprehensive analysis of the results of seven submitted approaches. In [4, 30], morphing attacks were detected through the identification of independent identity information in each analyzed sample. Neto *et al.* [30] used orthogonal vectors to identify the presence of more than one identity in the input samples. [4] used knowledge distillation to transfer information from an autoencoder trained on bona-fide samples, following distinct distillation techniques for bona-fide images (single identity distillation) and morphing attacks (double identity distillation). A vision transformer architecture for MAD was presented in [45], showing promising results. Very recently, MADation [5] used LoRa layers to fine-tune a foundation model to the downstream MAD task, highlighting the potential of FM in domain-specific tasks such as MAD. The work proposed in this document also focuses on single-image MAD due to its wider utility in real-world scenarios where a live probe might not be available.

2.2 Foundation Models

Foundation models are large-scale networks that can be trained with unlabeled data, following a self-supervised learning paradigm. This allows FMs to be trained on massive and diverse datasets, resulting in trained models that can efficiently generalize to a wide variety of tasks [1]. The DINOv2 family of networks [31] comprises self-supervised visual models capable of producing universal features that can be applied to both image-level and pixel-level tasks. The Segment Anything Model (SAM) [27] demonstrates strong generalization capabilities, enabling zero-shot image segmentation across a wide range of domains. Contrastive Language-Image Pre-training (CLIP) [36] is a multimodal FM constituted by a visual encoder and a text encoder. This allows CLIP to consider visual and textual information simultaneously during its training process and effectively learn the correlation between images and their textual description.

Recent advances in FMs have led the scientific community to explore their applicability to a wide range of downstream tasks, including in biometrics [1, 42]. One of the first works that applied foundation models to the face recognition task [8] concluded that using LoRA layers [21] to fine-tune FMs to the FR downstream task consistently outperforms training those models from scratch in low data availability scenarios. MADation [5] highlighted the adaptability of CLIP [36] to the domain-specific MAD task by fine-tuning its image encoder with LoRA layers. Similarly, FoundPAD [33] leverages LoRA-adapted CLIP and a binary classifier for face presentation attack detection (PAD), achieving superior performance across cross-dataset evaluations compared to state-of-the-art (SOTA) PAD systems. These results highlight the high adaptability of FMs to unseen downstream tasks even in domain-specific scenarios such as MAD and PAD. Arc2Face [34] adopts foundation models to extract

identity embeddings from face images, using these as conditions for a diffusion model to generate identity-specific synthetic faces. CLIB-FIQA [32] used CLIP to perform face image quality assessment by aligning the visual features of input face images with textual descriptions of image quality factors such as pose and expression.

Apart from visual encoders, recent developments in large language models (LLMs) and, in particular, multimodal versions of GPT-4, have opened new research directions in the biometrics field. Hassanpour et al. [20] assessed GPT-4's performance in tasks such as FR, gender classification, and age estimation. DeAndres-Tame et al. [13] conducted a thorough evaluation of GPT-4 for face verification and soft-biometric attribute estimation, providing a complementary explainability analysis of the model's decisions. Despite the absence of fine-tuning to biometric tasks, GPT-4 managed to achieve 94% verification accuracy on the LFW dataset [22] and 96.3% gender classification accuracy on MAAD-Face [43]. Furthermore, GPT-4's capacity to match human faces was proved to be on par with average human performance [28]. Farmanifard and Ross [17] explored GPT-4's capabilities in iris recognition under zero-shot settings.

Recent works have explored the use of FM for biometric tasks, including MAD [5], primarily by fine-tuning these models on task-specific datasets. While fine-tuning FM can yield strong performance on detecting MAs [5], this approach often overlooks one of the key strengths of FMs, which is their ability to perform zero-shot predictions on previously unseen scenarios. By relying heavily on fine-tuning [5] or training models from scratch [26, 30, 45], these approaches may limit the generalization capacity of the models, which is especially critical in MAD applications where new morphing techniques continually emerge and the ability to detect unseen attacks without extensive retraining or fine-tuning is essential. This paper unleashes the power of multimodal FMs under a zero-shot prediction scenario, namely CLIP, for MAD by leveraging multiple prompts, aiming to better capture diverse cues of bona-fide and attack samples, and thus, improve the MAD performance.

3 Zero-shot MAD using CLIP and multiple textual prompts aggregation

This work leverages the zero-shot learning capability of the multimodal (text–image) foundation model CLIP for MAD by employing multiple carefully designed textual prompts per class (attack or bona-fide). By averaging these prompts, we aim to better align CLIP's aggregated text embeddings with the image embeddings and, thus, capture more diverse cues without any need for fine-tuning. In this section, we first present preliminaries on CLIP, followed by a detailed description of the single- and multiple-prompt strategies for MAD.

3.1 Preliminary on CLIP

CLIP [36] is a multimodal FM constituted by a visual encoder and a text encoder. CLIP was trained using a massive dataset where each image is paired with a textual description that may accurately describe it (positive pair) or not (negative pair), utilizing a contrastive learning paradigm where the cosine similarity between the features extracted for image and text is maximized/minimized for positive/negative pairs. This allows CLIP to learn the relationship between visual and textual inputs and simultaneously interpret them during inference, resulting in a model generalizable across distinct tasks [36] with very competitive zero-shot learning results.

3.2 MADPrompts

In this work, we explore the zero-shot learning capacity of multimodal FMs in the MAD task when single and multiple prompts are used to describe each possible classification label. Using CLIP to classify images that are not present during the CLIP training process might result in suboptimal performance, particularly in domain-specific settings such as MAD [5] or face recognition [8]. However, works that highlight this phenomenon [5, 8, 33] only considered a single yet simple text prompt and do not provide extensive zero-shot evaluation using multiple prompts. Furthermore, they sometimes failed to adhere to the input image specifications that are expected to work optimally for CLIP. Hence, we propose to perform such an evaluation to reach a more comprehensive understanding of FMs' zero-shot capability and raise awareness towards the importance of efficient prompt engineering.

3.2.1 Single-Prompt Inference. When a multimodal FM is used to infer the label of a sample x_i in a zero-shot learning setting, two parallel processing steps are required, one for each encoder. The visual encoder, E_v, processes x_i, producing an image embedding $e_i = E_v(x_i)$. The text encoder, E_t, processes a list of text prompts, each describing one of the possible labels of x_i, y_i. For a binary task such as MAD, where $y_i \in \{0, 1\}$, E_t is fed two individual text prompts describing bona-fide (p_{BF}) and morphing attack (p_{MA}) samples. This results in two text embeddings, $e_{BF} = E_t(p_{BF})$ and $e_{MA} = E_t(p_{MA})$, that represent the two possible labels in a feature space with the same dimensionality as e_i. It is important to note that all the considered embeddings are normalized, thus laying on a unit hypersphere with the same dimensionality. As explained in Section 3.1, the contrastive learning paradigm of multimodal FMs like CLIP results in a strong correlation between the feature space of the visual and text embeddings. Hence, the zero-shot learning prediction for x_i, $\hat{y}_i$, can be determined by selecting the label whose embedding is more similar to e_i:

$$\hat{y}_i = \begin{cases} 0, & \phi(e_i, e_{BF}) > \phi(e_i, e_{MA}) \\ 1, & \text{otherwise} \end{cases}, \tag{1}$$

where $\phi(.)$ is the cosine similarity function.

3.2.2 Prompts Aggregation. Although CLIP can achieve remarkable zero-shot evaluation results in several tasks using a single text prompt to define each class, this model has been shown to perform better when the information of several text prompts defining each possible class is combined into a single text embedding [36]. The proposed approach combines multiple context prompts on the text feature space by averaging their contributions before comparing the final feature vector with the visual embedding, increasing CLIP's zero-shot performance on e.g., ImageNet by 3.5 percentage points [36]. These results suggest that customizing the prompts used to perform zero-shot classification can largely contribute to its success.

Taking this into consideration, we further adapt CLIP's zero-shot evaluation to include multiple text prompts per class. In this scenario, p_{BF} and p_{MA} are substituted by sets of prompts where each

Table 1: Lists with the three sets of prompts representing characteristics linked to face images used in this work: identity, presentation and appearance. For each possible label, the "{}" field is substituted by its corresponding ISO/IEC 20059 compliant definition ("face image morphing attack" and "bona-fide presentation").

Identity	Presentation	Appearance
male {}.	frontal {}.	bearded {}.
female {}.	profile {}.	moustached {}.
young {}.	tilted {}.	smiling {}.
elderly {}.	rotated {}.	frowning {}.
child {}.	upward {}.	eyeglasses {}.
adult {}.	downward {}.	sunglasses {}.
asian {}.	sideways {}.	wrinkled {}.
black {}.	leftward {}.	balding {}.
white {}.	rightward {}.	occluded {}.
latino {}.	angled {}.	scarred {}.
middle eastern {}.	inclined {}.	pierced {}.
indian {}.	declined {}.	tanned {}.
blonde {}.	oblique {}.	pale {}.
brunette {}.	twisted {}.	makeup {}.
redhead {}.	turned {}.	freckled {}.
tall {}.	slanted {}.	chubby-cheeked {}.
short {}.	offcenter {}.	sweaty {}.
thin {}.	misaligned {}.	dirty {}.
obese {}.	skewed {}.	blinking {}.
teen {}.	asymmetric {}.	tearful {}.

entry represents a possible bona-fide or morphing attack description, respectively. Let these sets be defined as $P_{BF} = \{p_{BF_1}, p_{BF_2}, ..., p_{BF_N}\}$ and $P_{MA} = \{p_{MA_1}, p_{MA_2}, ..., p_{MA_N}\}$, respectively. Each of the entries of P_{BF} and P_{MA} is individually fed to the text encoder, generating its embedding in the textual feature space. The final embedding representations used for each label (e_{BF} and e_{MA}) can then be obtained by averaging the contributions of all the embeddings belonging to the corresponding set:

$$e_{BF} = \frac{1}{N}\sum_{j=1}^{N} E_t\left(p_{BF_j}\right), \quad p_{BF_j} \in P_{BF} \tag{2}$$

$$e_{MA} = \frac{1}{N}\sum_{j=1}^{N} E_t\left(p_{MA_j}\right), \quad p_{MA_j} \in P_{MA} \tag{3}$$

Both text embeddings are normalized before being compared with e_i to predict the input sample's class (Equation 1) to ensure that the embeddings being compared lay on top of a unit hypersphere of the same dimensionality. Note that the usage of multiple prompts per class does not imply additional computational costs when compared with the single text prompt scenario, as the final textual embedding representations e_{BF} and e_{MA} can be pre-computed once and utilized during inference to match with the image embedding.

4 Experimental Setup

This section presents the experimental setups followed in the paper.

4.1 Model Architecture

CLIP [36] released four different models with two architectures: base and large. CLIP base architecture has 86M parameters and is available in 2 variants with different patch sizes. CLIP lage contains 0.3 billion parameters and also includes two variants, one of which is pre-trained at a higher resolution for one additional epoch [44]. The zero-shot MAD performance of these architectures has been assessed by a recent study [5] that revealed CLIP large architecture's superiority by a significant margin of 12.73 percentage points in terms of average EER across MAD22 [23] and its extension MorDIFF [10]. The higher zero-shot MAD capacity of the large architecture ViT-L is justified by its higher number of parameters, which allows it to learn a more complete set of features and thus perform better in a wider variety of tasks without access to extra knowledge (zero-shot learning) [5]. Taking these results into consideration, we selected CLIP large trained without high-resolution images as the FM architecture used in this work. This architecture is from now on referred to as ViT-L.

4.2 Text Prompts

In this work, we evaluate CLIP's zero-shot learning performance in the MAD task when single and multiple prompts are used to describe each classification label. When a single prompt is used per class, we follow the textual descriptions proposed in [5], to provide directly comparable results while complying with the ISO/IEC 20059 standard [25]. However, differently from [5], we propose to add a dot to each of the suggested textual prompts to comply with the settings followed during CLIP's training process [36]. Hence, the single prompt per class scenario utilizes two possible prompts to describe the analyzed samples: "face image morphing attack." and "bona-fide presentation.". The multiple prompt scenario requires a more careful design of the textual inputs, as it delves into more detailed image specificities instead of focusing solely on its possible labels. To provide a thorough investigation of the usefulness of different image attributes in boosting CLIP's zero-shot performance, we considered three attribute lists representing characteristics linked to face images: identity (ID), presentation (Pr) and appearance (Ap). The list of prompts for these three categories are listed in Table 1. For each possible label, the {} field is substituted by the correspondent ISO/IEC 20059 compliant definition, as described above. For morphing attack samples, for example, this results in the following ID prompt list: ["male face image morphing attack.", "female face image morphing attack.", ..., "teen face image morphing attack."]. Note that all prompts used in the multiple prompt scenario also include a dot in the end, following CLIP's training settings [36]. The different lists of prompts are also combined to verify their grouped contribution to zero-shot evaluation, resulting in four extra evaluation settings: ID+Pr, ID+Ap, Pr+Ap and All.

4.3 Image Pre-Processing

Before being evaluated by the FM, each sample was cropped following [10] and then resized to 224 × 224 pixels and normalized following the same setting used during CLIP's training (μ =[0.48145466, 0.4578275, 0.40821073], σ =[0.26862954, 0.26130258, 0.27577711]) [36]. These pre-processing steps ensure that the images fed to CLIP comply with the image resolution and normalization settings originally used to train this FM [36].

4.4 Datasets

To evaluate the zero-shot learning capacity of CLIP and ensure consistent benchmarking and comparison with earlier research [5, 10, 23], MAD22 [23] and its extension MorDIFF [10] were selected as evaluation datasets. These benchmarks are based on the Face Research Lab London (FRLL) dataset [14]. They use the same set of 204 bona-fide images and use distinct morphing techniques to create morphing attacks from the same pairs of bona-fide samples. MAD22 includes five sets of morphed samples; two of them were generated by GAN-based representation-level methods (MIPGAN I and II [46]) while the remaining three derive from image-level techniques (FaceMorpher, OpenCV [29], and Webmorph). MorDIFF's morphing samples were generated with a diffusion autoencoder [35].

4.5 Evaluation Metrics

The metrics used to perform MAD evaluation in this study were chosen to allow for consistent benchmarking [5, 10, 23] while ensuring conformity with the ISO/IEC 30107-3 [24] standard. These metrics include the Bona-fide Presentation Classification Error Rate (BPCER) and the Attack Presentation Classification Error Rate (APCER), which measure the proportion of bona-fide images misclassified as attack samples and the proportion of attacks misclassified as bona-fide samples, respectively. To cover different operational points and present comparative results, we report both the APCER at fixed BPCER values and the BPCER at fixed APCER values, evaluated at values of 1%, 10%, and 20%. We further report the detection Equal Error Rate (EER), which provides a succinct indicator of the overall performance balance of the system as it corresponds to the error rate at the operating point where the BPCER and APCER are equal.

4.6 Explainability

To provide a comprehensive analysis of the performed experiments, we analyze the relation between different text embeddings and the input images, supporting our quantitative results with visual explanations. To that end, the similarity score between e_i and a text embedding (obtained by passing single or multiple text prompts to CLIP's text encoder) is backpropagated through the image encoder. This results in a text-conditioned image heatmap that highlights the image regions that are more responsible for the similarity to the analyzed textual embedding (e_{BF} or e_{MA}). We start by comparing how the samples from each dataset are activated by single textual prompts with bona-fide or morphing descriptions, as it is expected that they generate distinct heatmaps. We further analyze the impact that using different sets of multiple prompts has on the activations, to verify whether distinct descriptions contribute differently to the final activations.

5 Results and Discussion

This section presents a detailed analysis of the results obtained in this work. We start by assessing the importance of complying to the training settings followed during the multimodal FM training process to achieve optimal zero-shot performance. Then, we explore the potential of using multiple textual prompts per label to increase zero-shot learning performance. Finally, we provide an explainability evaluation that showcases CLIP's focus shift when analyzing samples with distinct labels, highlighting its capacity to distinguish bona-fide samples from morphing attacks without fine-tuning to the MAD task.

5.1 The Power of Compliance

Table 2 presents CLIP's zero-shot learning performance when using a single prompt to represent each class. The two first sections present the results of MADation (TI) [5] and our implementation of TI (TI w/o Dot), which used "face image morphing attack" and "bona-fide presentation" as input text prompts, based on the ISO/IEC 20059 standard. While MADation (TI) normalizes the input samples with a mean and standard deviation of 0.5 on all dimensions, our modified version, TI w/o Dot, defines the normalization hyperparameters as those used during CLIP training (μ =[0.48145466, 0.4578275, 0.40821073], σ =[0.26862954, 0.26130258, 0.27577711]). The last section, TI-Dot, follows the same sample normalization as TI w/o Dot but further complies with CLIP's training settings by adding the dot character (".") in the end of both text prompts.

The analysis of the results shows that TI w/o Dot surpassed MADation's TI [5] implementation across all evaluated benchmarks and metrics. Although the normalization setting used by TI (μ =[0.5, 0.5, 0.5], σ =[0.5, 0.5, 0.5]) is commonly followed in MAD systems training and evaluation, this strategy is suboptimal when evaluating CLIP's zero-shot performance, as this FM learned to classify images normalized with a distinct distribution. These findings highlight the importance of correctly adapting the input images' pre-processing depending on the model used to classify them.

When comparing TI w/o Dot and TI-Dot, the superiority of the latter method is showcased by its decreased average error across 6 of the 7 analyzed metrics. In particular, adding a single dot character (".") to the input text prompts fed to CLIP reduced the MAD EER by 1.46 percentage points. These results support the previously withdrawn conclusions, highlighting the importance of correctly following the settings used to train FM's when using them for zero-shot evaluation on domain-specific tasks such as MAD. Taking this into consideration, all the remaining experiments presented in this paper follow the same normalization settings as TI-Dot and use input text prompts that include the dot character.

5.2 Multiple Prompt Aggregation

Table 3 presents CLIP zero-shot learning performance when multiple text prompts' contributions are averaged for each class (morphing attacks and bona-fide). These results are directly compared with the scenario where a single text prompt per label is used (TI-Dot) to quantify the impact of using multiple text prompts per class. It can be seen that from 6 out of the 7 multiple prompt settings surpassed TI-Dot in terms of average EER. In particular, the Pr+Ap setting achieved the best overall performance, reducing the average EER achieved with a single prompt per class by the considerable margin of 2.71 percentage points.

While it is important to determine which prompt setting leads to better zero-shot performance, it is also relevant to understand how the different sets of prompts contribute to improving the performance achieved with a single text prompt per class. An initial assessment of each individual set of prompts (identity, presentation, appearance) can be done by comparing the performance of settings ID, Pr, and Ap, respectively, with the baseline, TI-Dot. While

Table 2: Evaluation results for CLIP ViT-L using different normalization settings and prompt design structures. The best result achieved for each metric in each test dataset is highlighted in bold.

Method	Test data	EER (%)	APCER (%) @ BPCER (%)			BPCER (%) @ APCER (%)		
			1.00	10.00	20.00	1.00	10.00	20.00
TI [5]	FaceMorph	44.60	98.40	79.70	63.60	99.02	87.25	76.96
	MIPGAN_I	18.90	71.80	32.20	17.80	69.61	33.82	18.14
	MIPGAN_II	12.80	56.70	17.00	8.90	59.31	17.16	8.33
	OpenCV	35.47	96.24	77.54	63.11	96.08	73.53	55.39
	WebMorph	25.20	94.80	52.00	30.20	87.75	50.98	32.35
	MorDIFF	42.60	97.80	79.60	69.50	97.06	83.33	68.63
	Average	*29.93*	*85.96*	*56.34*	*42.19*	*84.81*	*57.68*	*43.30*
	Worst	*44.60*	*98.40*	*79.70*	*69.50*	*99.02*	*87.25*	*76.96*
TI w/o Dot (ours)	FaceMorph	**17.70**	**49.50**	**22.90**	17.30	**83.82**	35.78	**13.73**
	MIPGAN_I	6.90	**24.30**	5.70	3.40	35.78	4.41	1.96
	MIPGAN_II	**3.40**	**10.51**	**1.10**	0.70	16.18	**1.47**	**0.49**
	OpenCV	16.26	**59.25**	**20.73**	14.63	78.92	28.43	**10.29**
	WebMorph	**17.60**	**62.40**	**24.00**	**16.40**	**67.65**	**28.43**	**14.71**
	MorDIFF	32.90	**93.90**	62.30	49.70	94.12	65.69	48.04
	Average	15.79	**49.98**	22.79	17.02	62.75	27.37	14.87
	Worst	32.90	**93.90**	62.30	49.70	94.12	65.69	48.04
TI-Dot (ours)	FaceMorph	18.10	61.20	25.10	**16.40**	88.73	**34.31**	17.65
	MIPGAN_I	**5.40**	26.50	**4.50**	**1.60**	**24.02**	**3.43**	**1.47**
	MIPGAN_II	3.50	13.41	1.20	**0.20**	**10.78**	**1.47**	**0.49**
	OpenCV	**16.06**	66.97	21.14	**10.98**	**67.16**	**21.57**	10.78
	WebMorph	18.40	75.60	30.60	17.40	77.45	32.35	19.12
	MorDIFF	**24.50**	94.70	**52.20**	**30.10**	**91.18**	**51.96**	**29.90**
	Average	**14.33**	56.40	**22.46**	**12.78**	**59.89**	**24.18**	**13.24**
	Worst	**24.50**	94.70	**52.20**	**30.10**	**91.18**	**51.96**	**29.90**

ID falls behind TI-Dot in 6 out of the 7 averaged metrics, Pr and Ap managed to surpass the single text prompt approach in all 7 metrics, with Pr performing better than Ap in 5 of them. These results suggest that while both presentation and appearance attribute information positively contribute to increasing the performance of the FM zero-shot performance, the incorporation of id-related information is not beneficial. This conclusion is also supported by the fact that ID+Pr shows increased performance in comparison to ID while falling behind Pr in all 7 averaged metrics. Similar conclusions can also be withdrawn when comparing ID+Ap with ID and Ap or All with Pr+Ap. The Pr+Ap setting, on the other hand, manages to surpass Pr and Ap in most of the considered evaluation metrics, suggesting that the presentation and appearance attributes provide information presenting complementary benefits to the FM zero-shot capacity. Overall, Pr+Ap proved to be the best performing approach, surpassing all remaining strategies in 3 out of the 7 average metrics, including the EER. The best average value for the 4 remaining metrics was achieved by either Pr or Ap, supporting the individual contribution of these two settings towards increased zero-shot performance.

While this study does not aim to achieve SOTA performance but to provide a comprehensive analysis of zero-shot performance when using different text input prompts and raise awareness towards the importance of efficient prompt engineering and prompt aggregation, the comparison between our MADPromptS strategy and MADation [5] derives naturally from the previous comparison established with the zero-shot learning results presented in this work (Table 2). It is interesting to notice that Pr+Ap surpasses the fine-tuned CLIP model proposed in [5] by 0.32 percentage points in terms of average EER, without requiring any extra fine-tuning or adaption to the downstream MAD task. The complete assessment of the results provided in this section reveals the importance of selecting appropriate text prompts when using FMs in the zero-shot evaluation setting, highlighting the benefits that can arise

Table 3: Evaluation results for CLIP ViT-L using a single prompt per class (TI-Dot) and multiple prompts per class. The sets of prompts used for multiple text prompt zero-shot evaluation provide more detailed descriptions than TI-Dot, highlighting identity (ID), presentation (Pr) or appearance (Ap) characteristics, as well as mixtures of these categories (ID+Pr, ID+Ap, Pr+Ap and All). The best result achieved for each metric in each test dataset is highlighted in bold.

Method	Test data	EER (%)	APCER (%) @ BPCER (%)			BPCER (%) @ APCER (%)		
			1.00	10.00	20.00	1.00	10.00	20.00
TI-Dot	FaceMorph	18.10	61.20	25.10	16.40	88.73	34.31	17.65
	MIPGAN_I	5.40	26.50	4.50	1.60	24.02	3.43	1.47
	MIPGAN_II	3.50	13.41	1.20	0.20	10.78	1.47	**0.49**
	OpenCV	16.06	66.97	21.14	10.98	67.16	21.57	10.78
	WebMorph	**18.40**	75.60	**30.60**	**17.40**	**77.45**	**32.35**	**19.12**
	MorDIFF	24.50	94.70	52.20	30.10	91.18	51.96	29.90
	Average	14.33	56.40	22.46	12.78	59.89	24.18	13.24
	Worst	24.50	94.70	52.20	30.10	91.18	51.96	29.90
ID	FaceMorph	19.90	62.80	31.60	19.90	84.80	38.24	20.10
	MIPGAN_I	7.00	21.30	5.60	2.70	25.98	5.39	1.47
	MIPGAN_II	5.01	13.61	2.10	0.80	19.12	3.43	0.49
	OpenCV	16.06	55.69	21.65	11.69	67.65	22.06	12.75
	WebMorph	22.00	**67.20**	36.20	24.40	85.78	38.24	24.51
	MorDIFF	21.40	82.30	43.20	27.00	86.27	37.75	24.02
	Average	15.23	50.48	23.39	14.43	61.60	24.19	13.89
	Worst	**22.00**	82.30	43.20	27.00	86.27	**38.24**	**24.51**
Pr	FaceMorph	14.00	**53.60**	**17.40**	10.60	71.57	21.08	**8.33**
	MIPGAN_I	5.40	18.90	3.10	0.80	18.63	**2.94**	0.98
	MIPGAN_II	3.40	9.91	**0.90**	0.20	**8.33**	0.98	**0.49**
	OpenCV	13.21	57.62	**15.85**	10.06	**56.37**	20.10	**7.35**
	WebMorph	25.60	81.60	43.60	30.20	83.82	47.06	33.33
	MorDIFF	14.70	73.90	20.40	9.70	72.55	19.12	**10.29**
	Average	12.72	49.26	**16.88**	10.26	51.88	18.55	**10.13**
	Worst	25.60	81.60	43.60	30.20	**83.82**	47.06	33.33
Ap	FaceMorph	15.50	62.70	23.70	**9.10**	**66.67**	**18.63**	11.76
	MIPGAN_I	6.40	19.00	**2.50**	0.60	14.71	3.43	0.98
	MIPGAN_II	**3.00**	12.21	1.00	**0.10**	9.80	1.47	**0.49**
	OpenCV	13.52	62.30	20.33	**8.13**	57.84	17.65	9.80
	WebMorph	24.60	76.80	46.60	26.80	92.16	57.35	31.86
	MorDIFF	15.00	78.90	30.20	**8.20**	**60.78**	**17.65**	12.75
	Average	13.00	51.99	20.72	**8.82**	**50.33**	19.36	11.27
	Worst	24.60	78.90	46.60	26.80	92.16	57.35	31.86
ID+Pr	FaceMorph	16.20	58.70	23.30	13.90	79.90	27.45	10.78
	MIPGAN_I	5.90	19.20	3.90	1.90	23.04	3.43	0.98
	MIPGAN_II	3.20	11.31	1.30	0.20	10.78	1.96	**0.49**
	OpenCV	13.62	56.61	19.21	10.57	62.25	21.08	9.31
	WebMorph	23.40	75.60	41.40	27.80	86.76	45.59	27.45
	MorDIFF	18.40	79.00	**15.80**	31.80	81.86	27.45	15.69
	Average	13.45	50.07	17.49	14.36	57.43	21.16	10.78
	Worst	23.40	79.00	41.40	31.80	86.76	45.59	27.45
ID+Ap	FaceMorph	16.10	60.40	25.60	14.30	77.45	26.96	12.75
	MIPGAN_I	6.50	17.70	3.90	0.80	16.67	4.41	0.98
	MIPGAN_II	4.60	10.71	1.10	0.20	11.76	1.96	**0.49**
	OpenCV	13.85	55.59	19.51	10.06	63.24	19.61	9.31
	WebMorph	22.20	71.20	41.00	26.00	88.73	48.53	27.94
	MorDIFF	17.70	79.10	35.00	16.20	78.43	27.94	15.20
	Average	13.49	49.12	21.02	11.26	56.05	21.57	11.11
	Worst	22.20	79.10	**41.00**	**26.00**	88.73	48.53	27.94
Pr+Ap	FaceMorph	**12.90**	55.40	19.80	9.50	70.59	19.61	9.80
	MIPGAN_I	**4.50**	**16.70**	3.20	**0.50**	**13.73**	3.43	**0.49**
	MIPGAN_II	3.60	**9.81**	**0.90**	**0.10**	9.80	**0.49**	**0.49**
	OpenCV	12.80	55.79	18.70	9.04	56.86	**16.67**	7.84
	WebMorph	23.60	77.60	47.00	28.80	87.75	51.96	31.86
	MorDIFF	**12.30**	**73.40**	26.80	8.60	69.61	**17.65**	11.27
	Average	**11.62**	**48.12**	19.40	9.42	51.39	**18.30**	10.29
	Worst	23.60	**77.60**	47.00	28.80	87.75	51.96	31.86
All	FaceMorph	14.90	58.60	21.40	11.80	75.49	25.00	10.78
	MIPGAN_I	5.40	17.10	3.80	0.60	17.16	3.43	**0.49**
	MIPGAN_II	3.80	10.21	10.00	0.30	10.78	0.98	**0.49**
	OpenCV	**12.30**	**55.18**	18.50	9.65	62.25	19.61	8.82
	WebMorph	22.80	74.80	42.60	27.20	85.78	47.06	28.92
	MorDIFF	15.50	77.60	30.40	12.00	76.47	24.02	13.24
	Average	12.45	48.92	21.12	10.26	54.66	20.02	10.46
	Worst	22.80	**77.60**	42.60	27.20	85.78	47.06	28.92

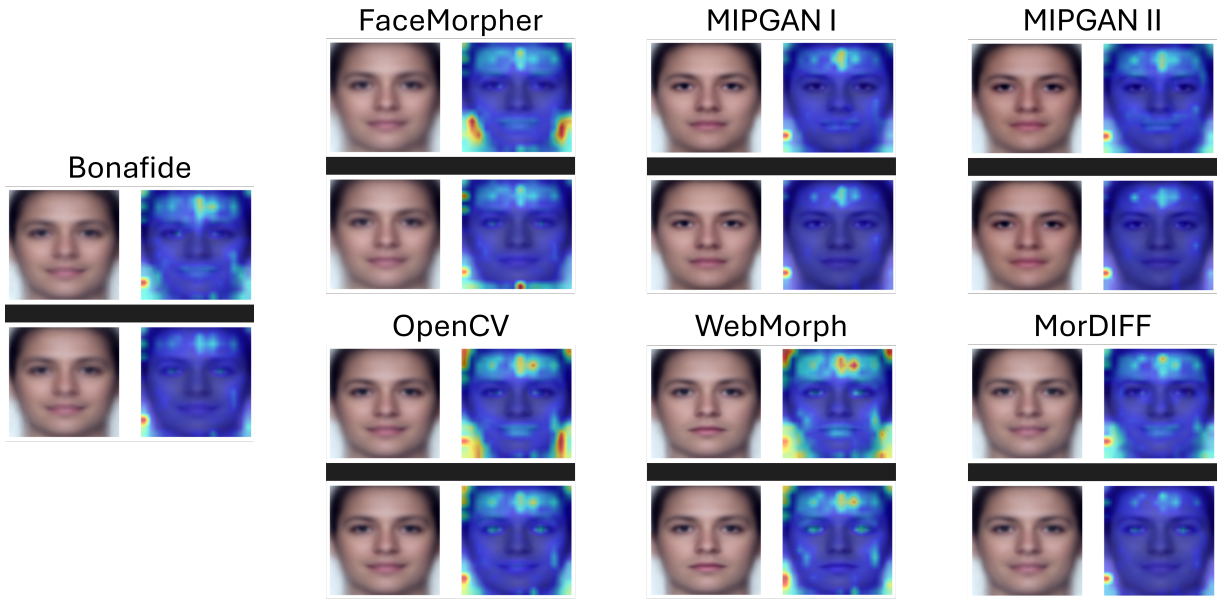


Figure 1: CLIP's average activation heatmaps for different MAD datasets when using a single text prompt to describe each possible class (TI-Dot). Each block represents one dataset, displaying an average representation of all its samples as well as the correspondent average activation heatmap. For each dataset, the first and second line highlight the heatmap associated with the morphing attack and the bona-fide prompt, respectively. Since all the evaluation datasets considered in this work share the same set of bona-fide samples, the average bona-fide activation maps are highlighted separately (first column). The average input sample and heatmap associated with each dataset (second to foruth columns) correspond only to its morphing attack samples.

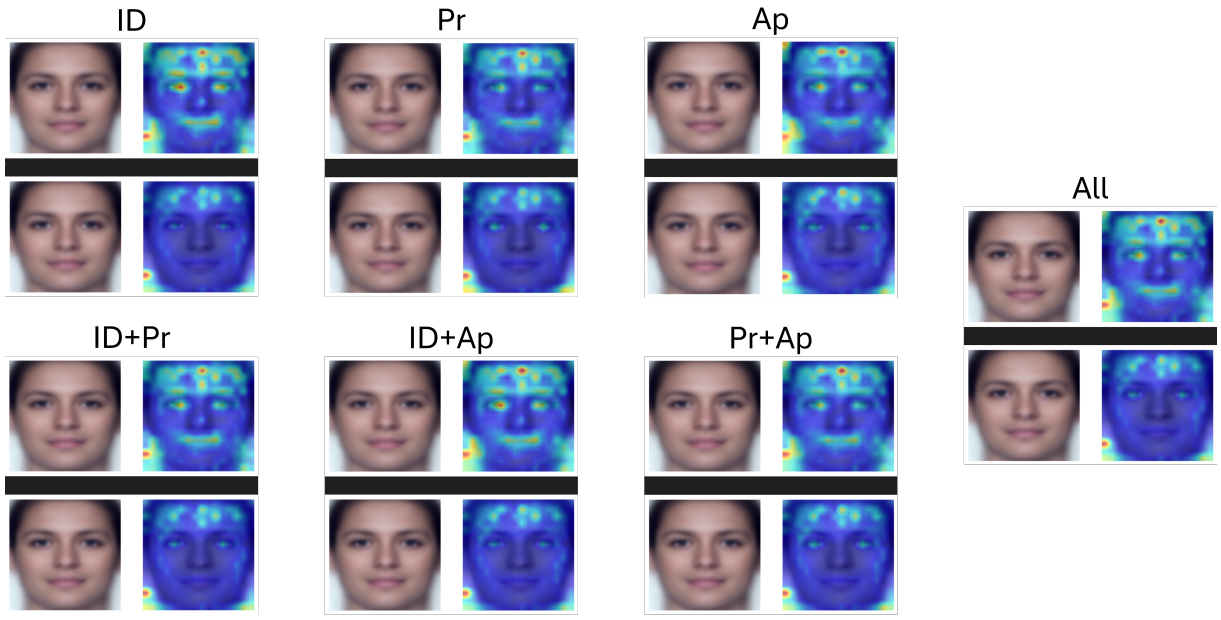


Figure 2: CLIP's average activation heatmaps for the morphing samples on the MorDIFF dataset when using multiple text prompts to describe each possible class (ID, Pr, Ap, ID+Pr, ID+Ap, Pr+Ap, All). For each setting, the first and second line highlight the heatmap associated with the morphing attack and the bona-fide sets of prompts, respectively.

from efficient prompt engineer. Hence, this work paves the way towards more efficient prompt engineering and aggregation, providing important insights regarding the importance of correctly exploiting textual prompts to the user's advantage by leveraging FM's zero-shot learning ability to a fuller extent.

5.3 Explainability

Figure 1 shows CLIP's average activation heatmaps of different datasets when using a single text prompt to describe each possible class (TI-Dot). For each dataset, the first and second row highlight the heatmap associated with the morphing attack and the bona-fide prompt, respectively. All the evaluation datasets considered in this work share the same set of bona-fide samples. Hence, the activation maps for the bona-fide samples are highlighted separately, and the average input sample and heatmap associated with each dataset corresponds only to its morphing attack samples. It should be noted that all these datasets use the same pairs of bona-fide samples to create the morphing, which justifies the high similarity between the average input samples shown for each scenario. It is possible to observe that the heatmaps obtained when the morphing prompt is encoded by the text encoder significantly differ from the ones associated with bona-fide text prompts. In particular, the usage of morphing prompts makes CLIP focus its attention in detailed areas of the face, such as the mouth and jawline. These areas are particularly prone to contain artifacts in morphing samples [10, 46], highlighting CLIP's capacity to focus on important characteristics that might indicate the presence of a malicious sample without any fine-tuning (zero-shot learning). It is also worth notice that the heatmaps for presented morphing prompts generally show more activation when analyzing morphing samples in comparison with bona-fide images. These results complement the previous analysis regarding the effectiveness of the proposed method, providing a visual and easily interpretable explanation of the obtained results.

Similar heatmaps were also plotted for the different multiple prompt scenarios analyzed in this work (Figure 2). This plot follows the same presentation logic as Figure 1 and focuses on the MorDIFF dataset, for which our ID+Pr multiple prompt strategy surpassed the baseline (TI-Dot) by a larger margin. It can be seen that using different sets of prompts to describe the input samples results in different activation patterns, supporting their distinct influence in CLIP's performance. In particular, it is interesting to observe that combining different sets of prompts in the proposed multiple text prompt zero-shot approach results in activation maps that combine characteristics from all the incorporated sets. As an example, when analyzing the morphing heatmaps of ID, Pr and Ap it is clear that ID results in a stronger activation in the eye region, followed by Ap and finally by Pr. This tendency is kept when analysing joint contributions, with ID+Ap showcasing the strongest activation in the eye region, followed by ID+Pr and, finally, by Pr+Ap. The complementary nature of these activations is also in line with the conclusions withdrawn in the previous section revealing that sets of prompts that contribute to increase the baseline (TI-Dot) performance can generally be combined to boost CLIP's zero-shot MAD capacity even more (Pr+Ap vs Pr and Ap) while combining a set that negatively impacts CLIP performance with other prompt collections reduces the performance achieved with a single benefitial prompt set (ID+Pr vs Pr, for example).

Overall, it is possible to conclude that using single or multiple prompts per class results in significantly different CLIP activations (Figures 1 and 2), and that CLIP follows distinct attention patterns when using different multiple prompt sets. These conclusions are in line with the results displayed in Table 3, highlighting the importance of effective prompt engineering to take the best possible advantage of FM's zero-shot learning capacity.

6 Conclusion

This work offers a comprehensive analysis of the use of multimodal FMs for the critical task of MAD under a zero-shot setting. We demonstrate that careful alignment with the model's original training conditions, including appropriate textual prompt design, can significantly enhance zero-shot performance without any fine-tuning. Beyond this, we introduce and evaluate a strategy of aggregating multiple carefully designed textual prompts per class, enabling the model to capture more diverse and discriminative cues relevant to distinguishing bona-fide from attack samples. While using a single prompt per class provides only a general description of the target label, employing multiple prompts incorporates more specific characteristics, shifting the model's attention to a broader range of details. Our experiments show that disjoint sets of prompts exhibit complementary capabilities when combined. Moreover, leveraging prompt aggregation not only enhances zero-shot MAD performance but can even surpass fine-tuned models, highlighting the untapped potential of well-designed textual prompts as a simple yet effective alternative to task-specific fine-tuning. These findings emphasize the importance of prompt design for zero-shot FM predictions, paving the way for more generalizable and scalable MAD solutions in biometric security applications.

Acknowledgments

This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[1] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the Opportunities and Risks of Foundation Models. *CoRR* abs/2108.07258 (2021).

[2] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. 2022. Elasticface: Elastic margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1578–1587.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya

Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 159, 25 pages.

[4] Eduarda Caldeira, Pedro C Neto, Tiago Gonçalves, Naser Damer, Ana F Sequeira, and Jaime S Cardoso. 2023. Unveiling the two-faced truth: Disentangling morphed identities for face morphing detection. In *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE, 955–959.

[5] Eduarda Caldeira, Guray Ozgur, Tahar Chettaoui, Marija Ivanovska, Peter Peer, Fadi Boutros, Vitomir Struc, and Naser Damer. 2025. MADation: Face Morphing Attack Detection with Foundation Models. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV) Workshops*. 1650–1660.

[6] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. 2022. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems* 35 (2022), 16664–16678.

[7] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. 2022. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534* (2022).

[8] Tahar Chettaoui, Naser Damer, and Fadi Boutros. 2024. FRoundation: Are Foundation Models Ready for Face Recognition? *arXiv preprint arXiv:2410.23831* (2024).

[9] Naser Damer, Viola Boller, Yaza Wainakh, Fadi Boutros, Philipp Terhörst, Andreas Braun, and Arjan Kuijper. 2019. Detecting face morphing attacks by analyzing the directed distances of facial landmarks shifts. In *Pattern Recognition: 40th German Conference, GCPR 2018, Stuttgart, Germany, October 9-12, 2018, Proceedings 40*. Springer, 518–534.

[10] Naser Damer, Meiling Fang, Patrick Siebke, Jan Niklas Kolf, Marco Huber, and Fadi Boutros. 2023. Mordiff: Recognition vulnerability and attack detectability of face morphing attacks created by diffusion autoencoders. In *2023 11th International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 1–6.

[11] Naser Damer, Alexandra Mosegui Saladie, Andreas Braun, and Arjan Kuijper. 2018. Morgan: Recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network. In *2018 IEEE 9th international conference on biometrics theory, applications and systems (BTAS)*. IEEE, 1–10.

[12] Naser Damer, Noémie Spiller, Meiling Fang, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. 2021. Pw-mad: Pixel-wise supervision for generalized face morphing attack detection. In *Advances in Visual Computing: 16th International Symposium, ISVC 2021, Virtual Event, October 4-6, 2021, proceedings, Part I*. Springer, 291–304.

[13] Ivan DeAndres-Tame, Ruben Tolosana, Rubén Vera-Rodríguez, Aythami Morales, Julian Fiérrez, and Javier Ortega-Garcia. 2024. How Good Is ChatGPT at Face Biometrics? A First Look Into Recognition, Soft Biometrics, and Explainability. *IEEE Access* 12 (2024), 34390–34401.

[14] Lisa DeBruine and Benedict Jones. 2017. Face Research Lab London Set. (5 2017). doi:10.6084/m9.figshare.5047666.v5

[15] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.

[16] Meiling Fang, Fadi Boutros, and Naser Damer. 2022. Unsupervised face morphing attack detection via self-paced anomaly detection. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 1–11.

[17] Parisa Farmanifard and Arun Ross. 2024. ChatGPT Meets Iris Biometrics. In *IJCB*. IEEE, 1–10.

[18] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. 2014. The magic passport. In *IJCB*. IEEE, 1–7.

[19] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. 2016. On the effects of image alterations on face recognition accuracy. *Face recognition across the imaging spectrum* (2016), 195–222.

[20] Ahmad Hassanpour, Yasamin Kowsari, Hatef Otroshi-Shahreza, Bian Yang, and Sébastien Marcel. 2024. Chatgpt and Biometrics: an Assessment of Face Recognition, Gender Detection, and Age Estimation Capabilities. In *ICIP*. IEEE, 3224–3229.

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[22] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 2008. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.

[23] Marco Huber, Fadi Boutros, Anh Thi Luu, Kiran Raja, Raghavendra Ramachandra, Naser Damer, Pedro C Neto, Tiago Gonçalves, Ana F Sequeira, Jaime S Cardoso, et al. 2022. 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)*. IEEE, 1–10.

[24] International Organization for Standardization. 2017. ISO/IEC DIS 30107-3:2016: Information Technology – Biometric presentation attack detection – P. 3: Testing and reporting.

[25] International Organization for Standardization. 2023. ISO/IEC DIS 20059: Information technology — Methodologies to evaluate the resistance of biometric recognition systems to morphing attacks.

[26] Marija Ivanovska and Vitomir Struc. 2023. Face Morphing Attack Detection with Denoising Diffusion Probabilistic Models. In *11th International Workshop on Biometrics and Forensics, IWBF 2023, Barcelona, Spain, April 19-20, 2023*. IEEE, 1–6. doi:10.1109/IWBF57495.2023.10156877

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.

[28] Robin SS Kramer. 2025. Face to face: Comparing ChatGPT with human performance on face matching. *Perception* 54, 1 (2025), 65–68.

[29] Satya Mallick. 2016. Face Morph Using OpenCV — C++ / Python. *LearnOpenCV* 1, 1 (2016). https://learnopencv.com/face-morph-using-opencv-cpp-python/

[30] Pedro C Neto, Tiago Gonçalves, Marco Huber, Naser Damer, Ana F Sequeira, and Jaime S Cardoso. 2022. Orthomad: Morphing attack detection through orthogonal identity disentanglement. In *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 1–5.

[31] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

[32] Fu-Zhao Ou, Chongyi Li, Shiqi Wang, and Sam Kwong. 2024. CLIB-FIQA: Face Image Quality Assessment with Confidence Calibration. In *CVPR*. IEEE, 1694–1704.

[33] Guray Ozgur, Eduarda Caldeira, Tahar Chettaoui, Fadi Boutros, Raghavendra Ramachandra, and Naser Damer. 2025. FoundPAD: Foundation Models Reloaded for Face Presentation Attack Detection. In *WACV (Workshops)*. IEEE, 697–707.

[34] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. 2024. Arc2Face: A Foundation Model for ID-Consistent Human Faces. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[35] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. 2022. Diffusion Autoencoders: Toward a Meaningful and Decodable Representation. In *CVPR*. IEEE.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[37] Raghavendra Ramachandra, Sushma Venkatesh, Naser Damer, Narayan Vetrekar, and Rajendra S. Gad. 2024. Multispectral Imaging for Differential Face Morphing Attack Detection: A Preliminary Study. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*. IEEE, 6173–6181. doi:10.1109/WACV57701.2024.00607

[38] Raghavendra Ramachandra, Sushma Venkatesh, and Guoqiang Li. 2025. PoolAtnRes: Towards Generalisable Differential Morphing Attack Detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2025, Tucson, AZ, USA, February 26 - March 6, 2025*. IEEE, 9312–9321. doi:10.1109/WACV61041.2025.00902

[39] Raghavendra Ramachandra, Sushma Venkatesh, Kiran Raja, and Christoph Busch. 2019. Detecting face morphing attacks with collaborative representation of steerable features. In *Proceedings of 3rd International Conference on Computer Vision and Image Processing: CVIP 2018, Volume 1*. Springer, 255–265.

[40] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714* (2024).

[41] Ulrich Scherhag, Ramachandra Raghavendra, Kiran B Raja, Marta Gomez-Barrero, Christian Rathgeb, and Christoph Busch. 2017. On the vulnerability of face recognition systems towards morphed face attacks. In *2017 5th international workshop on biometrics and forensics (IWBF)*. IEEE, 1–6.

[42] Hatef Otroshi Shahreza and Sébastien Marcel. 2025. Foundation models and biometrics: A survey and outlook. *Authorea Preprints* (2025).

[43] Philipp Terhörst, Daniel Fährmann, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. 2021. MAAD-Face: A Massively Annotated Attribute Dataset for Face Images. *IEEE Trans. Inf. Forensics Secur.* 16 (2021), 3942–3957.

[44] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. 2019. Fixing the train-test resolution discrepancy. *Advances in neural information processing systems* 32 (2019).

[45] Haoyu Zhang, Raghavendra Ramachandra, Kiran Raja, and Christoph Busch. 2024. Generalized Single-Image-Based Morphing Attack Detection Using Deep Representations from Vision Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 1510–1518.

[46] Haoyu Zhang, Sushma Venkatesh, Raghavendra Ramachandra, Kiran Bylappa Raja, Naser Damer, and Christoph Busch. 2021. MIPGAN - Generating Strong and High Quality Morphing Attacks Using Identity Prior Driven GAN. *IEEE Trans. Biom. Behav. Identity Sci.* 3, 3 (2021), 365–383. doi:10.1109/TBIOM.2021.3072349