MADation: Face Morphing Attack Detection with Foundation Models

Eduarda Caldeira[1], Guray Ozgur[1], Tahar Chettaoui[1], Marija Ivanovska[2],
Peter Peer[2], Fadi Boutros[1], Vitomir Struc[2], Naser Damer[1,3]

[1] Fraunhofer IGD, Germany, [2] University of Ljubljana, Slovenia
[3] TU Darmstadt, Germany
Email: maria.eduarda.loureiro.caldeira@igd.fraunhofer.de

Abstract

Despite the considerable performance improvements of face recognition algorithms in recent years, the same scientific advances responsible for this progress can also be used to create efficient ways to attack them, posing a threat to their secure deployment. Morphing attack detection (MAD) systems aim to detect a specific type of threat, morphing attacks, at an early stage, preventing them from being considered for verification in critical processes. Foundation models (FM) learn from extensive amounts of unlabelled data, achieving remarkable zero-shot generalization to unseen domains. Although this generalization capacity might be weak when dealing with domain-specific downstream tasks such as MAD, FMs can easily adapt to these settings while retaining the built-in knowledge acquired during pretraining. In this work, we recognize the potential of FMs to perform well in the MAD task when properly adapted to its specificities. To this end, we adapt FM CLIP architectures with LoRA weights while simultaneously training a classification header. The proposed framework, MADation surpasses our alternative FM and transformer-based frameworks and constitutes the first adaption of FMs to the MAD task. MADation presents competitive results with current MAD solutions in the literature and even surpasses them in several evaluation scenarios. To encourage reproducibility and facilitate further research in MAD, we publicly release the implementation of MADation at https://github.com/gurayozgur/MADation.

1. Introduction

The high focus of the research community on the study of deep learning techniques in recent years has led to the development of high-performing systems in several fields, including face recognition (FR) [5, 20]. However, the same scientific advances used to improve the recognition power of FR systems can also be used to create efficient ways to attack them [15, 23], posing a threat to their secure deployment. Morphing attacks (MA) constitute an example of such threats, as their generation process aims at incorporating features from more than one identity, resulting in a sample that can be verified by multiple people by the same FR system. When left undetected, these attacks can lead to several dangerous situations [7, 13, 56], such as allowing multiple people to pass border control with the same passport or letting a criminal travel under the identity of another person [36]. To address this problem, several morphing attack detection (MAD) systems have been proposed [7,17,21,27,38,46]. MAD algorithms aim at distinguishing unaltered images (bona-fide samples) from MAs to identify malicious samples at an early stage and prevent them from being considered for verification in critical processes.

Foundation models (FM) are large-scale networks that can be trained with self-supervised learning, which allows them to learn from unlabelled data. The fact that no labelling is required for FMs' training samples highly simplifies the data acquisition task, allowing FMs to be trained in massive and diverse datasets. This training paradigm results in models that can efficiently generalize to a wide variety of assignments [4], making them particularly beneficial for fields that address several tasks, such as natural language processing (NLP) [6] and computer vision (CV) [34, 40, 45, 47]. Despite the recent attention given to FMs, their adaption to perform biometrics tasks is still very limited. While very recent works have used FMs to generate synthetic face images [41], perform iris segmentation [22] and FR [11], the utility of FMs for most biometrics fields is still highly under-explored. This literature gap should be carefully addressed, especially taking into account that biometrics tasks such as MAD may strongly benefit from FMs' high generalization power, provided their efficient adaption to the domain specificities of the downstream task [11], and high generalizability when it comes to sub-domains [45], i.e. different morphing mechanisms in the MAD case.

This work explores the potential of using FMs as the basis for the downstream MAD task. Given the domain-specific nature of the MAD task and knowing that FMs often underperform in specialized settings [53], we propose to adapt the pre-trained FM CLIP [45] to MAD with low-rank adaption (LoRA), while simultaneously training a header to perform classification. This allows the FM to better align its feature space with the specificities of the downstream MAD task and still take advantage of the built-in knowledge acquired during pre-training. This adaption paradigm corresponds to our proposed framework, MADation. We further evaluate whether MADation properly takes advantage of FMs' properties by comparing it with alternative FM and transformer-based methods. We start by assessing the importance of adapting CLIP to the downstream MAD task by evaluating its zero-shot performance on this task (TI). Then, to verify whether LoRA adaption improves MAD performance, we use the FM as a frozen feature extractor. In this scenario, the FM's feature space is not aligned with the specificities of MAD and only the classification layer is trained (FE). Finally, to ensure that MADation's performance is not only deriving from its architecture but is also dependent on the FM's built-in knowledge acquired during pre-training, we assess the importance of the FM's pre-trained weights by comparing MADation with models following the same architectures trained from scratch (ViT-FS). The developed experiments highlight the efficiency of MADation compared with the remaining methods, reducing the average EER by 16.93 pp. and 8.10 pp. compared to ViT-FS and FE, respectively, for CLIP ViT-L. Furthermore, MADation revealed competitive performance levels compared with recent MAD solutions, highlighting FM's potential in domain-specific tasks such as MAD.

2. Related Work

MAD: MAs are face images that result from the fusion of identity information belonging to two or more identities, allowing them to be simultaneously verifiable as belonging to all of them. An example of an MA can be found in Figure 1. Both human observers and FR systems are vulnerable to MAs [24, 49], leading to dangerous situations, such as multiple people being able to pass border control with the same passport [7]. To address the threat posed by MAs, several studies have proposed MAD systems [7, 17, 21, 27, 38, 46]. These systems can address the MAD task from two different perspectives, depending on the MAD operational scenario. Differential MAD solutions [12] are fed two samples simultaneously: a live capture of the individual claiming that the investigated image represents their identity and the investigated image itself. Although this approach is useful in scenarios such as border control, the fact that two images need to be compared to perform the detection limits its applicability in several scenarios [12], e.g. analysing standalone documents. Hence, several studies have developed single-image MAD systems [7, 17, 21, 27, 31, 38, 46, 55], which can detect whether the investigated image is a morph based only on its characteristics. Ramachandra *et al.* [46] introduced a handcrafted-feature-based approach that extracts textural features across multiple scales and classifies them using collaborative representation. [17] deviated from the common binary classification of the whole investigated image by learning to classify each of its pixels (or pixel blocks) as bona-fide or MA. Fang *et al.* [21] proposed an unsupervised approach that used self-paced learning to assign smaller weights to suspicious samples, which generally correspond to MAs, allowing for training a robust autoencoder for anomaly detection even when the training data is polluted with MAs. Neto *et al.* [38] determined whether the analyzed sample contained two independent identities by separating its identity information into two orthogonal latent vectors. [7] trained an autoencoder on bona-fide samples to distil identity knowledge to a MAD system, following distinct distillation techniques for bona-fide and MAs. [55] performed MAD with ViT architectures, showing promising results. [31] developed a self-supervised diffusion model that reconstructs bona-fide images from noisy inputs. As the model is trained on bona-fide samples alone, it leads to higher error rates when fed MAs, which can be identified through anomaly detection. [27] promoted the SYN-MAD 2022 competition on MAD based on synthetic training data, presenting a comprehensive analysis of the results of seven submitted approaches. This work also focuses on single-image MAD due to its wider utility in real-world scenarios, whether they offer a live probe or not.

Foundation Models: FMs contain many trainable parameters, enabling FM to learn from large and diverse datasets. This intensive training leads to high adaptability, which is particularly useful in areas that deal with a wide range of tasks, such as CV [11, 22, 34, 40, 41, 45]. DINOv2 networks [40] is a series of self-supervised pre-trained visual models able to generate universal features that can be used to perform both image-level and pixel-level visual tasks. The Segment Anything Model (SAM) [34] can perform image segmentation on a wide range of domains. Contrastive Language-Image Pretraining (CLIP) [45] is a multimodal FM trained to interpret visual and textual inputs simultaneously, effectively learning the correlation between images and their textual description.

Although vision FMs' prove to be generalizable for several downstream tasks, they achieve less optimal performance when applied to some specific settings [53], which require adapting FMs to the desired downstream task [9, 10, 25]. AdaptFormer [9] used two identical MLP branches instead of the MLP block in the transformer encoder. One of them replicates the original network to help maintain its properties, while the other allows for task-specific finetuning. ViT-Adapter [10] led to state-of-the-art (SOTA) COCO results for plain ViT networks, by combining the reconstruction of fine-grained multi-scale features with the introduction of image-related inductive biases into the FM.

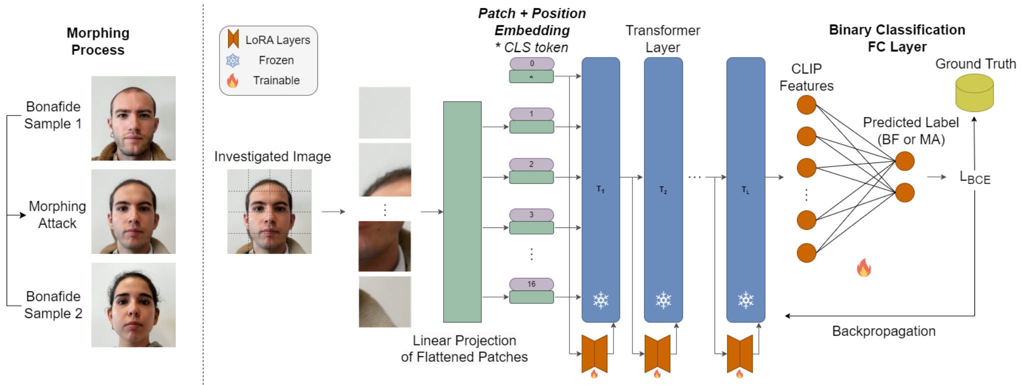

Figure 1. Morphing attack generation and MADation's pipeline. The left side of the figure depicts a morphing sample and the two bona-fide identities that were morphed to generate it, using [13]. Keep in mind that attackers commonly choose to morph faces with similar features for higher success [16]. The right side represents MADation's pipeline, consisting of an adapted FM followed by a binary fully connected classification layer. The embedding space of the FM is adapted by fine-tuning the LoRA parameters and the classification layer is simultaneously trained to produce the MAD predictions. Better visualized in colour.

Hu *et al.* [25] inserted trainable rank decomposition matrices in the pre-trained FM, allowing for low-dimensional reparametrization. During the adaption process, the FM's pre-trained weights are kept frozen and only these matrices' weights are updated, allowing for an effective adaption with minimal computational cost. Using these matrices, also known as Low-Rank Adaptation (LoRA) layers, to adapt FMs has led to performance improvements in a broad range of tasks, such as capsule endoscopy diagnosis [54], plant phenotyping [8] and FR [11]. In this work, we select LoRA to adapt CLIP to the MAD task due to its promising results reported in the literature, namely in biometrics [11].

Despite the growing attention given to FMs by the community, their application in biometrics is still limited to a few recent works [11, 22, 41]. In [22], SAM [34] was fine-tuned to iris segmentation. [41] used FMs to generate identity-specific facial images. A recent work [11] used LoRA [25] to fine-tune DINOv2 [40] and CLIP [45] FMs to the FR task under several data availability settings. The experiments showed that FMs can be efficiently used in FR, especially in low data availability scenarios, where the proposed technique surpassed models trained from scratch.

In this work, we contribute to the list of recent advancements in FMs' application to biometrics by proposing a MAD solution based on FMs, MADation. In particular, we recognize the FMs' generalization potential and their utility to tasks such as MAD when properly adapted to their domain-specific characteristics [11]. To this end, MADation incorporates LoRA layers in the analysed FM while simultaneously training an extra classification layer.

3. Methodology

3.1. Preliminary on CLIP

CLIP [45] is a multimodal FM that interprets visual and textual inputs simultaneously. CLIP was trained in a large dataset where each image is paired with a textual description, allowing it to learn the relationship between these two modalities. When a pair is fed to CLIP, its components are processed by two distinct encoders that are simultaneously trained using a contrastive learning approach that evaluates the cosine similarity between the features extracted for image and text, maximizing or minimizing it for positive and negative pairs, respectively. This allows CLIP to effectively learn the correlation between images and textual descriptions, resulting in a model generalizable across distinct tasks [45] with very competitive zero-shot learning results.

In this work, we use CLIP due to its capacity to generalize well to domain-specific downstream tasks when properly adapted [11]. We assess CLIP's abilities as a MAD model by evaluating its zero-shot learning performance (Section 3.3). To this end, CLIP is fed with image-text pairs where the text input corresponds to the possible image labels ('face image morphing attack' and 'bona-fide presentation' based on the ISO/IEC 20059 standard [29]). For the remaining approaches, only the image encoder is used and CLIP works as a feature extractor on top of which a classification layer is added, following the recent successful utilization of FMs in downstream tasks such as FR [11] and image segmentation [34]. These approaches include our proposed framework, MADation (Section 3.2), and two additional frameworks designed to assess the importance of the FM's built-in knowledge and of adapting its feature space to the downstream domain (Section 3.3). When applicable (TI, FE, and MADation), CLIP was initialized with the pre-trained weights made publicly available in [45][1].

3.2. MADation

In this work, we take advantage of the high generalization capacity of FMs and assess their usefulness in the downstream MAD task. Although we acknowledge the benefits that can arise from directly deploying FMs to MAD, we also recognize the limited ability of these networks to perform well in domain-specific tasks such as MAD without

[1] https://github.com/OpenAI/CLIP

any adaption. Thus, the proposed framework, MADation, adapts an FM by using LoRA layers to shift its pre-trained feature space in a direction that facilitates the MAD task while simultaneously training a classification layer. A visual representation of MADation is depicted in Figure 1.

Fine-Tuning with LoRA: As mentioned in Section 2, the FMs' drop in performance when facing domain-specific scenarios such as MAD can be addressed with ViT adapters [9, 10, 25]. In this work, we select LoRA [25] to adapt the selected FM due to its promising results reported in the literature, namely in biometrics [11]. LoRA employs a low-dimensional reparametrization strategy, demonstrating effectiveness comparable to training the full parameter space [1] while greatly reducing the number of parameters that need to be updated. In this method, the pre-trained weights of the FM, $W_0 \in \mathbb{R}^{d\times k}$, are kept unchanged, and trainable rank-decomposition matrices are added within each layer of the transformer, enabling effective adaptation with minimal computational overhead. The low-rank decomposition introduced by LoRA updates W_0 as follows:

$$W_0 + \Delta W = W_0 + \gamma_r BA, \tag{1}$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are the trainable rank-decomposition matrices, with the rank $r << min(d,k)$, and γ_r is a scaling factor. Originally, γ_r was defined as $\frac{\alpha}{r}$, but this formulation often causes gradient collapse as the rank r increases, resulting in a lack of performance gains despite the use of additional trainable parameters for fine-tuning [32]. To address this issue, rank-stabilized LoRA (rsLoRA) [32] modifies the scaling factor to $\frac{\alpha}{\sqrt{r}}$, preventing gradient collapse and enabling better performance at higher ranks. For this reason, we opt to employ rsLoRA to fine-tune CLIP, setting $\gamma_r = \frac{\alpha}{\sqrt{r}}$. Since γ_r is a constant and the original CLIP weights are kept frozen, only the matrices A and B are updated. Once the adaption process is complete, the final model weights are calculated as $W = W_0 + \gamma_r BA$. As no extra parameters are added, the original model's computational efficiency during inference is maintained.

The FM's image encoder contains alternating multi-headed self-attention (MSA) layers and multilayer perceptron (MLP) blocks. Layer normalization and residual connections are applied before and after each block, respectively. For simplicity and parameter efficiency, LoRA is only applied on the MSA weights, leaving the MLP unaltered [25]. Although LoRA can be applied to the query, key, value and output (q, k, v and o, respectively) projection matrices in the MSA, we only adapt q and v matrices following the results obtained in the original study where LoRA was proposed [25] and recent work using FMs in FR [11]. The MSA layers run h parallel heads, each with a unique set of q, k and v. The LoRA layers in each head function independently and have distinct weights. When an embedding x is fed to the MSA, the q, k and v projection layers in head i (Q_i, K_i and V_i, respectively) are calculated as follows:

$$\begin{aligned} Q_i &= W_i^q x + \gamma_r B_i^q A_i^q x, \\ K_i &= W_i^k x, \\ V_i &= W_i^v x + \gamma_r B_i^v A_i^v x, \end{aligned} \tag{2}$$

where W_i^q, W_i^k and W_i^v are the frozen projection layers for q, k and v, respectively, and A_i^q, B_i^q, A_i^v and B_i^v correspond to the trainable LoRA layers. Q_i, K_i and V_i are then used to compute the attention score of head i, using the dimension of the key vectors, d_k, as a scaling factor:

$$Attention(Q_i, K_i, V_i) = Softmax\Big(\frac{Q_i K_i^T}{\sqrt{d_k}}\Big)V_i. \tag{3}$$

The MSA layer's output is determined by concatenating all heads' attention scores along the feature dimension and feeding the resultant vector to the projection layer O:

$$Multihead(Q, V, K) = Concat(head_1, ..., head_k)W^0. \tag{4}$$

The final output of the MSA is processed by the frozen MLP, completing the execution of a ViT block. The output of block l is then processed by block $l+1$, which consists of a new MSA adapted with LoRA followed by a frozen MLP.

Classification: Using LoRA to fine-tune CLIP allows it to produce a final embedding space adapted to the downstream MAD task. In this scenario, the FM works as a feature extractor on top of which classification can be performed. Hence, an additional fully connected layer with two output neurons followed by a softmax layer is added on top of the FM, resulting in the complete detector architecture. This layer is trained along with fine-tuning the LoRA parameters, using the binary cross-entropy loss:

$$L_{BCE} = -(y\, log(\tilde{y}) + (1-y)\, log(1-\tilde{y})), \tag{5}$$

where y and $\tilde{y}$ represent the sample's ground truth and predicted labels, respectively. After training, and during the feedforward MAD process, all MADation's weights are kept frozen. The detection score is obtained from the output of the binary classification layer, with the highest output score defining the model prediction for each sample.

3.3. MAD Baselines

To provide a comprehensive analysis of the use of FMs to perform MAD and prove the effectiveness of MADation when compared with alternative baseline solutions, we considered three alternative FM or transformer-based approaches for MAD.

Text-Image (TI) MAD: FMs like CLIP have demonstrated exceptional zero-shot learning performance across several downstream tasks, including action recognition in videos, sentiment analysis and car model classification [45]. Taking this into account, we evaluate the selected FM zero-shot learning performance on MAD, by simultaneously using its text and image encoders. To this end, TI processes image-text pairs where the textual input specifies the two possible classification labels. The predicted label is determined by analysing the similarity score between the image embedding and the input text embeddings. However, the potential of this technique should not be overestimated based on its success in general downstream domains such as the ones specified above given the domain-specific nature of MAD. The fact that no adaption is performed to the specific requirements of the downstream MAD task makes TI prone to underperform when compared with methods that consider MAD's specific characteristics, such as MADation.

ViT Trained from Scratch (ViT-FS) MAD: The remarkable performance of FMs is closely related to their underlying architectures, which are often based on ViT networks [2]. These architectures have demonstrated promising performance in tasks such as MAD [55] and presentation attack detection (PAD) [26]. In this work, we also examine whether ViT networks can effectively perform MAD, allowing us to assess their contribution to the proposed FM-based methodology, MADation. Specifically, we train from scratch the same ViT-B and ViT-L used to evaluate MADation, using only the selected MAD training datasets. These transformers' parameters are randomly initialized, meaning that ViT-FS does not benefit from prior knowledge acquired during massive training and thus cannot be considered an FM. This approach enables a direct comparison between FM-based methods, such as MADation, and visual transformers, allowing us to assess how valuable the in-built knowledge of FMs is for downstream tasks such as MAD.

Feature Extractor (FE) MAD: To determine the importance of the network adaption allowed by LoRA in MADation, we develop an experiment where the FM is not fine-tuned. In this scenario, the FM works as a frozen feature extractor on top of which a fully connected layer is trained to perform classification, using the binary cross-entropy loss (Equation 5). This experiment allows us to assess the suitability of the FM's original feature space to discriminate between MAD classes while quantitatively measuring the improvements introduced by adapting the FM's weights to the domain-specific downstream MAD task with LoRA.

4. Experimental Setup

Datasets: The Synthetic Morphing Attack Detection Development (SMDD) dataset [14] was selected as the training dataset of the proposed models. SMDD [14] is a synthetic-based MAD dataset, consisting of 25k bona-fide images generated using the StyleGAN2-ADA framework [33, 52] and 15k morphing attacks created from the bona-fide samples using the OpenCV morphing technique [37]. The SMDD dataset was chosen for training as it ensures privacy by avoiding the use of real face images and because it has shown remarkable success in MAD solution development [7, 13, 38] and public competitions [27]. The benchmarking datasets proposed in these works, MAD22 [27] and its extension MorDIFF [13], were also chosen to ensure the results' comparability and a data domain identity disjoint from the SMDD training data. The evaluation benchmarks are based on the Face Research Lab London (FRLL) dataset [19] and thus contain the same 204 bona-fide images. Additionally, the same image pairs were used to create the morphing attacks in MAD22 and MorDIFF. The MAD22 dataset includes morphed images from five different approaches: three image-level techniques (FaceMorpher, OpenCV [37], and Webmorph) and two GAN-based representation-level methods (MIPGAN I and II [56]). The morphing samples of the MorDIFF dataset were generated with a diffusion autoencoder [42]. In addition to MAD22 [27], the FRLL-Morphs dataset [48] is utilized for evaluation so that a comparison is possible with methods that are not evaluated on MAD22. The FRLL-Morphs dataset [48], similarly derived from the Face Research London Lab dataset [19], serves as a benchmark for MAD evaluation. The dataset comprises 204 genuine samples and over 1,000 morphed faces per technique, generated using five distinct morphing methods: Style-GAN2 [33, 52], WebMorph [18], AMSL [39], FaceMorpher [43], and OpenCV [37].

Image Pre-Processing: Before being used as an input to the FM, each sample is cropped following [14] and then resized to 224 × 224 pixels to comply with the image resolution originally used to train CLIP [45]. During training, all samples are also subject to data augmentation using random horizontal flipping, following [38, 50]. Due to the success achieved by tokenization in FMs' NLP applications, Alexey *et al.* [2] proposed preprocessing images into tokens before feeding them to the FM's image encoder. All the experiments mentioned in this document follow this tokenization process, represented in Figure 1. Each input sample is divided into non-overleaping regions which are fed to a linear projection layer [2]. The resultant patch embeddings are appended to a learnable embedding, the class (CLS) token [2], which constitutes an image representation that helps classify the input into predefined categories. Additional position embeddings are considered to preserve the spatial order of the original sample's patches. The final embedding vector is then fed to the image encoder [2].

Model Architecture: CLIP [45] released four different models with two architectures: base and large. CLIP's base architecture contains 86M parameters and has 2 variants with different patch sizes: 16 and 32. The large version of CLIP has 0.3 billion parameters and includes a variant pre-trained at a higher resolution of 336 pixels for one additional epoch to boost performance [51]. Following the results achieved by recent work that used CLIP in FR [11], we decide to consider one version of each architecture, namely CLIP's base version with a patch size of 16 and CLIP's large version trained without high-resolution images, which we refer to as ViT-B and ViT-L, respectively. These architectures were used in all the settings described in Sections 3.2 and 3.3, namely TI, ViT-FS, FE and MADation.

Implementation Details: All models were trained for 40 epochs using the AdamW optimizer [35] with momentum of 0.9 and weight decay of 0.05 [11]. For MADation, we used ViT-B and ViT-L architectures with model and header learning rates of 1e-5 and 1e-4, respectively. The LoRA parameters were set to $r = 2$, $\alpha = \{4, 8\}$ and $dropout = \{0.4, 0.2\}$ for ViT-B and ViT-L respectively. For FE, we used a header learning rate of 1e-2 for both architectures. ViT-FS models were trained with learning rates of 1e-5/5e-5 (model/header) for ViT-B and 1e-4/1e-4 for ViT-L. The batch size was set to 256 for all settings.

Evaluation Metrics: The MAD evaluation metrics were selected to ensure compliance with the ISO/IEC 30107-3 [28] standard and enable consistent benchmarking and comparability with previous studies [13, 27]. Performance is reported using the Bona-fide Presentation Classification Error Rate (BPCER), the Attack Presentation Classification Error Rate (APCER), and the detection Equal Error Rate (EER). The BPCER quantifies the proportion of bona-fide images misclassified as attack samples, while the APCER measures the proportion of attack images misclassified as bona-fide samples. The detection EER is the error rate at the operating point where the BPCER and APCER are equal, offering a concise measure of the system's overall performance balance. 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%.

5. Results and Discussion

Zero-Shot MAD (TI): As described in Section 3.3, the FM's zero-shot performance was evaluated by simultaneously pairing images with textual prompts describing the two possible classification labels, 'face image morphing attack' and 'bona-fide presentation' [29]. Table 1 displays TI results for both ViT-B and ViT-L. The results obtained with ViT-B reveal the performance limitations of TI, as it performs close to random in 3 out of the 6 evaluated datasets and results in high EER values for all of them. Although a similar tendency is verified for ViT-L in some evaluation datasets, this network performs significantly better than ViT-B and even achieves competitive results with recent MAD solutions from the literature (Table 2) in MIPGAN I and MIPGAN II. The higher zero-shot MAD capacity of ViT-L is justified by its higher number of parameters, which allow it to learn a wider spectrum of features during its pretraining stage and thus perform better for a wider variety of tasks, as demonstrated in [45]. Nonetheless, the global performance of both networks in the TI scenario is still far from satisfactory in comparison to other options described later in this section, highlighting the limitations of FM's in domain-specific scenarios such as MAD. These limitations can be largely overcome by adapting the FM to the downstream MAD task, as will be later shown in this section.

Table 1. Evaluation results for CLIP ViT-B and ViT-L for four sets of experiments: TI, ViT-FS, FE and MADation. The best and second-best results achieved for each metric in each test dataset are highlighted in bold and underlined, respectively.

Method		Test data	EER (%)	APCER (%) @ BPCER (%)			BPCER (%) @ APCER (%)		
				1.00	10.00	20.00	1.00	10.00	20.00
ViT-B	TI	FaceMorph	51.50	98.40	88.20	81.40	99.51	93.63	85.29
		MIPGAN_I	36.40	99.80	81.10	65.30	86.76	55.88	46.57
		MIPGAN_II	33.40	99.60	76.00	55.30	80.39	49.02	43.63
		OpenCV	47.15	99.90	83.74	74.90	98.04	81.37	70.10
		WebMorph	35.60	98.20	70.20	57.20	86.76	61.27	48.53
		MorDIFF	51.90	100.00	92.60	86.70	99.02	92.65	85.29
		Average	*42.66*	*99.32*	*81.97*	*70.13*	*91.75*	*72.30*	*63.24*
		Worst	*51.90*	*100.00*	*92.60*	*86.70*	*99.51*	*93.63*	*85.29*
	ViT-FS	FaceMorph	5.38	8.77	2.49	0.90	20.98	0.49	**0.00**
		MIPGAN_I	32.87	85.66	61.35	47.41	100.00	49.02	49.02
		MIPGAN_II	27.19	94.92	64.94	44.42	100.00	57.84	30.88
		OpenCV	16.30	50.40	26.42	14.27	100.00	56.31	34.47
		WebMorph	22.80	83.60	58.00	44.40	100.00	52.94	32.35
		MorDIFF	28.14	84.73	52.00	35.93	100.00	56.31	34.37
		Average	*22.13*	*68.01*	*44.20*	*31.22*	*86.83*	*40.68*	*26.41*
		Worst	*32.87*	*94.92*	*64.94*	*47.41*	*100.00*	*57.84*	*49.02*
	FE	FaceMorph	2.89	4.89	1.30	0.20	11.22	0.49	0.49
		MIPGAN_I	26.00	83.27	55.68	36.06	77.94	50.98	32.84
		MIPGAN_II	34.26	91.43	74.70	57.27	84.80	65.20	51.96
		OpenCV	14.88	39.98	20.34	9.21	61.27	18.63	10.78
		WebMorph	32.80	91.40	71.40	49.80	84.80	66.18	52.94
		MorDIFF	17.86	50.90	27.05	13.77	59.22	24.27	12.62
		Average	*21.45*	*60.31*	*41.74*	*27.72*	*63.21*	*37.62*	*26.94*
		Worst	*34.26*	*91.43*	*74.70*	*57.27*	*84.80*	*66.18*	*52.94*
	MADation (ours)	FaceMorph	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
		MIPGAN_I	33.37	82.97	55.18	43.92	94.12	72.55	52.94
		MIPGAN_II	22.21	79.98	34.66	24.30	84.80	47.55	26.47
		OpenCV	3.85	11.64	1.82	1.11	23.53	0.98	**0.00**
		WebMorph	**10.80**	60.00	**11.40**	**5.00**	**51.47**	**11.76**	**4.41**
		MorDIFF	**1.10**	**1.60**	**0.00**	**0.00**	**1.94**	**0.00**	**0.00**
		Average	***11.89***	*39.36*	*17.18*	*12.39*	***42.64***	*22.14*	*13.97*
		Worst	*33.37*	*82.97*	*55.18*	*43.92*	*94.12*	*72.55*	*52.94*
ViT-L	TI	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*
	ViT-FS	FaceMorph	22.63	75.17	38.68	24.93	88.29	40.98	24.88
		MIPGAN_I	23.80	79.08	42.93	25.50	91.18	46.57	28.43
		MIPGAN_II	21.81	80.28	36.65	25.40	91.67	25.00	40.69
		OpenCV	30.47	84.72	59.92	44.23	94.12	60.29	42.16
		WebMorph	33.60	91.60	59.80	48.60	100.00	75.49	52.45
		MorDIFF	40.92	94.51	77.94	67.86	100.00	81.55	67.96
		Average	*28.87*	*84.23*	*52.65*	*39.42*	*94.21*	*57.59*	*40.15*
		Worst	*40.92*	*94.51*	*77.94*	*67.86*	*100.00*	*81.55*	*67.96*
	FE	FaceMorph	9.77	44.17	9.77	4.09	35.12	10.24	5.37
		MIPGAN_I	23.51	88.84	55.28	31.37	71.57	40.69	27.45
		MIPGAN_II	21.81	82.37	45.42	25.10	69.61	32.84	23.53
		OpenCV	15.89	55.77	25.40	10.83	48.53	22.06	12.75
		WebMorph	26.40	86.60	56.80	37.80	68.63	41.67	29.90
		MorDIFF	22.85	87.03	50.70	29.14	67.48	35.92	24.27
		Average	*20.04*	*74.13*	*40.56*	*23.06*	*60.16*	*30.57*	*20.54*
		Worst	*26.40*	*88.84*	*56.80*	*37.80*	***71.57***	*41.67*	*29.90*
	MADation (ours)	FaceMorph	0.40	0.40	**0.00**	**0.00**	0.49	**0.00**	**0.00**
		MIPGAN_I	20.32	**55.88**	**29.08**	20.32	79.41	35.78	**15.69**
		MIPGAN_II	**9.06**	**19.42**	**9.06**	**5.58**	100.00	**5.39**	**0.98**
		OpenCV	**2.23**	**3.74**	**1.32**	**0.71**	**15.69**	**0.00**	**0.00**
		WebMorph	20.40	**47.60**	20.40	20.40	82.35	37.25	13.24
		MorDIFF	19.26	48.40	24.45	19.26	84.47	34.95	15.53
		Average	*11.94*	***29.24***	***14.05***	***11.04***	*60.40*	***18.90***	***7.57***
		Worst	***20.40***	***55.88***	***29.08***	*20.40*	*100.00*	***37.25***	***15.69***

Baselines toward MADation (ViT-FS and FE): We further explore two alternative approaches, ViT-FS and FE. ViT-FS is trained from scratch, making it possible to assess the potential of the underlying ViT architectures for the MAD task without relying on the built-in knowledge of FMs. FE makes use of this knowledge by using CLIP as a frozen feature extractor on top of which a binary fully connected layer is trained to perform classification, allowing us to assess the suitability of the FM's original feature space to discriminate between MAD classes. The results achieved by both approaches are presented in Table 1. When the ViT-B architecture is considered, FE outperforms ViT-FS in 4 of the 6 considered benchmarking datasets, with an average EER difference of 0.68 pp.. For ViT-L, FE largely surpasses ViT-FS performance, improving the average EER by 8.83 pp.. Similar tendencies can be

Table 2. Results comparison between MADation and previous MAD solutions. All methods are trained on SMDD [14], and evaluated on MAD22 [27] and its extension MorDIFF [13]. Specific values unavailable in the original papers are marked with "-". The best and second-best results achieved for each metric in each test dataset are highlighted in bold and underlined, respectively.

Method		Test data	EER (%)	APCER (%) @ BPCER (%)			BPCER (%) @ APCER (%)		
				1.00	10.00	20.00	1.00	10.00	20.00
MixFaceNet-MAD [13, 14]		FaceMorph	4.60	5.50	3.60	2.90	-	-	-
		MIPGAN_I	16.70	75.80	22.20	14.50	-	-	-
		MIPGAN_II	20.62	81.58	32.03	20.62	-	-	-
		OpenCV	8.33	36.38	6.50	3.76	-	-	-
		WebMorph	18.20	74.00	24.00	17.60	-	-	-
		MorDIFF	8.50	33.40	7.40	4.10	-	-	-
Inception-MAD [13, 46]		FaceMorph	**0.00**	1.70	**0.00**	**0.00**	-	-	-
		MIPGAN_I	10.90	50.90	**13.70**	5.70	-	-	-
		MIPGAN_II	16.22	82.48	25.83	11.41	-	-	-
		OpenCV	7.52	28.66	5.49	3.05	-	-	-
		WebMorph	18.00	85.20	27.40	13.40	-	-	-
		MorDIFF	5.30	17.20	3.50	2.50	-	-	-
MorphHRNet [27, 30]		FaceMorph	5.90	31.20	4.30	2.40	48.04	1.96	1.47
		MIPGAN_I	15.30	89.80	21.90	13.00	75.98	24.02	11.27
		MIPGAN_II	10.41	84.18	11.01	6.11	61.27	11.27	2.94
		OpenCV	5.69	66.97	3.76	1.63	33.82	1.96	1.47
		WebMorph	9.80	90.02	11.20	4.20	56.86	10.78	3.92
		MorDIFF	-	-	-	-	-	-	-
Con-Text Net A [27]		FaceMorph	**0.00**	99.90	**0.00**	**0.00**	100.00	**0.00**	**0.00**
		MIPGAN_I	12.30	**41.90**	14.10	8.10	**59.31**	**16.18**	**6.37**
		MIPGAN_II	12.91	43.44	14.51	8.61	59.31	19.61	5.88
		OpenCV	17.48	70.93	26.52	15.75	74.02	32.84	16.18
		WebMorph	26.20	89.20	45.60	31.00	93.14	48.53	31.86
		MorDIFF	-	-	-	-	-	-	-
E-CBAM@VCMI [27]		FaceMorph	41.20	100.00	92.80	62.80	100.00	95.10	82.35
		MIPGAN_I	32.50	99.90	84.90	60.20	78.92	53.43	40.69
		MIPGAN_II	25.93	99.60	64.66	37.34	**56.86**	30.39	30.39
		OpenCV	27.54	98.58	48.68	33.03	76.47	50.49	38.24
		WebMorph	30.60	99.00	86.80	46.80	69.12	47.06	38.24
		MorDIFF	-	-	-	-	-	-	-
Con-Text Net B [27]		FaceMorph	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
		MIPGAN_I	30.30	71.50	53.00	39.80	87.75	60.29	35.59
		MIPGAN_II	29.43	67.67	51.25	39.14	91.18	61.76	47.06
		OpenCV	22.66	57.83	34.45	23.68	82.35	43.14	23.53
		WebMorph	31.40	81.00	59.60	43.80	94.61	55.39	43.80
		MorDIFF	-	-	-	-	-	-	-
Xception [27, 30]		FaceMorph	0.60	0.50	**0.00**	**0.00**	1.47	1.47	1.47
		MIPGAN_I	36.90	97.90	80.40	57.40	86.27	57.35	49.02
		MIPGAN_II	44.54	99.50	92.49	77.08	90.20	67.65	56.86
		OpenCV	7.32	21.75	6.61	2.54	35.29	4.90	1.47
		WebMorph	14.60	49.40	23.00	10.80	53.92	21.57	11.76
		MorDIFF	-	-	-	-	-	-	-
D-FW-MixFaceNet [44]		FaceMorph	0.10	-	-	**0.00**	-	-	-
		MIPGAN_I	**6.70**	-	-	**1.20**	-	-	-
		MIPGAN_II	6.61	-	-	**1.00**	-	-	-
		OpenCV	13.72	-	-	9.04	-	-	-
		WebMorph	10.80	-	-	7.40	-	-	-
		MorDIFF	-	-	-	-	-	-	-
D-FW-CDCN [44]		FaceMorph	**0.00**			44.10			
		MIPGAN_I	11.90	-	-	3.80	-	-	-
		MIPGAN_II	14.11	-	-	8.51	-	-	-
		OpenCV	**0.30**	-	-	**0.00**	-	-	-
		WebMorph	**0.00**	-	-	64.00	-	-	-
		MorDIFF	-	-	-	-	-	-	-
MADation (ours)	ViT-B	FaceMorph	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
		MIPGAN_I	33.37	82.97	55.18	43.92	94.12	72.55	52.94
		MIPGAN_II	22.21	79.98	34.66	24.30	84.80	47.55	26.47
		OpenCV	3.85	**11.64**	**1.82**	1.11	**23.53**	**0.98**	**0.00**
		WebMorph	10.80	60.00	11.40	5.00	51.47	**11.76**	4.41
		MorDIFF	**1.10**	**1.60**	**0.00**	**0.00**	**1.94**	**0.00**	**0.00**
	ViT-L	FaceMorph	0.40	0.40	**0.00**	**0.00**	0.49	**0.00**	**0.00**
		MIPGAN_I	20.32	55.88	29.08	20.32	79.41	35.78	15.69
		MIPGAN_II	9.06	**19.42**	**9.06**	5.58	100.00	**5.39**	**0.98**
		OpenCV	19.26	48.40	24.45	19.26	84.47	34.95	15.53
		WebMorph	2.23	**3.74**	**1.32**	**0.71**	**15.69**	**0.00**	**0.00**
		MorDIFF	20.40	47.60	20.40	20.40	82.35	37.25	13.24

observed for metrics such as APCER@BPCER=20% and APCER@BPCER=10%. The superiority of FE when compared to ViT-FS possibly derives from the large number of trainable parameters of ViT-B and ViT-L. These large-scale networks require large amounts of training data to properly learn the considered task without overfitting, which might undermine ViT-FS' capacity to learn the MAD task given the reduced size of the SMDD dataset. On the other hand, FE benefits from the FM's previous knowledge, which was acquired during a pre-training phase with a massive amount of training data, which justifies its superior performance. Furthermore, the fact that this performance difference is higher when using ViT-L (8.83 pp. vs 0.68 pp. difference on average EER when using ViT-L and ViT-B, respectively) mostly derives from ViT-FS decreased performance in this scenario, which further reinforces that the limited capacity of ViT-FS to learn the MAD task is strongly correlated with the large number of trainable parameters of the considered networks. Hence, it can be concluded that taking advantage of the built-in knowledge of the FM results in better performance than training the network from scratch with a reduced amount of training data. Nonetheless, FE's MAD performance still has the potential for further enhancement, making it worth exploring whether adapting the FM to the downstream MAD task results in increased performance.

MADation: As previously discussed, FMs can generalize to a wide variety of downstream tasks but can show limited capacity when handling domain-specific tasks such as MAD. The previously analysed FE results highlight this characteristic, since the EER values still leave space for improvement, revealing that the FM's feature space is most likely significantly misaligned for the MAD task. Although it is also possible that the selected classification network has saturated its capacity given the FM's feature space and should thus be deeper, we argue that the problem is most likely arising from the feature space misalignment which can be corrected through FM's adaption, as highlighted in previous studies [8, 11, 54]. Furthermore, this type of adaption allows the network to have more flexibility without giving up on the knowledge acquired during the FM's pre-training phase, which might constitute a good trade-off between the properties of FE and ViT-FS. Hence, we propose to adapt CLIP to the MAD task with LoRA layers, resulting in our proposed approach, MADation. We evaluate MADation using the same benchmarks as FE and ViT-FS, to allow for a fair comparison with these approaches, as shown in Table 1. It can be seen that MADation achieves the best and/or second-best performance levels for most of the evaluated metrics and benchmarks. In particular, ViT-B is the best-performing method on average in 2 out of the 7 evaluated metrics, and the second-best-performing method on the remaining 5. For these 5 metrics, ViT-L presents the best overall performance. The analysis of the average EER also reveals MADation's superiority when compared with the remaining approaches, as it surpasses ViT-FS and FE by 10.24 pp. and 9.56 pp., respectively, for ViT-B and by 16.93 pp. and 8.10 pp., respectively, for ViT-L. MADation's improvements when compared with FE prove the importance of performing a correct FM adaption to downstream domain-specific tasks such as MAD. Simultaneously, MADation's superiority regarding ViT-FS shows that the network adaption provided by LoRA does not prevent the final network from benefitting from the knowledge acquired during pre-training. Hence, it is possible to conclude that MADation reaches an efficient trade-off between preserving the FM's built-in knowledge and fine-tuning it to the downstream task, resulting in improved MAD performance.

Comparison with the recent MAD approaches: To

Table 3. Results comparison between MADation and previous MAD solutions. All methods are trained on SMDD [14], and evaluated on FRLL-Morphs [48]. Specific values unavailable in the original papers are marked with "-". The best and second-best results achieved for each metric in each test dataset are highlighted in bold and underlined, respectively.

Method		Test data	EER (%)	BPCER (%) @ APCER (%)		
				1.00	10.00	20.00
OrthoMAD [38]		FRLL-Style-GAN2	6.54	13.74	-	3.76
		FRLL-WebMorph	15.23	70.92	-	9.50
		FRLL-OpenCV	**0.73**	**0.73**	-	0.32
		FRLL-AMSL	14.80	65.05	-	10.89
		FRLL-FaceMorpher	0.98	2.37	-	0.08
IDistill [7]		FRLL-Style-GAN2	**1.96**	**8.51**	-	**0.08**
		FRLL-WebMorph	4.01	14.41	-	0.33
		FRLL-OpenCV	2.46	6.14	-	0.16
		FRLL-AMSL	4.00	21.10	-	2.85
		FRLL-FaceMorpher	2.05	4.26	-	0.16
MixFaceNet [14]		FRLL-Style-GAN2	8.99	42.16	**8.82**	4.41
		FRLL-WebMorph	12.35	80.39	15.20	7.84
		FRLL-OpenCV	4.39	26.47	1.96	1.47
		FRLL-AMSL	15.18	49.51	21.08	11.76
		FRLL-FaceMorpher	3.87	23.53	0.49	0.49
PW-MAD [14]		FRLL-Style-GAN2	16.64	80.39	25.98	13.24
		FRLL-WebMorph	16.65	80.39	24.02	13.24
		FRLL-OpenCV	2.42	22.06	0.49	0.49
		FRLL-AMSL	15.18	96.57	24.02	5.88
		FRLL-FaceMorpher	2.20	26.47	0.49	**0.00**
Inception [14]		FRLL-Style-GAN2	11.37	72.06	13.73	6.86
		FRLL-WebMorph	9.86	53.92	9.80	2.94
		FRLL-OpenCV	5.38	38.73	1.96	0.98
		FRLL-AMSL	10.79	72.06	12.75	4.90
		FRLL-FaceMorpher	3.17	30.39	0.49	0.49
WB-Avcivas [3]		FRLL-Style-GAN2	14.87	-	-	-
		FRLL-WebMorph	19.32	-	-	-
		FRLL-OpenCV	7.91	-	-	-
		FRLL-AMSL	18.23	-	-	-
		FRLL-FaceMorpher	17.11	-	-	-
MADation (ours)	ViT-B	FRLL-Style-GAN2	17.21	54.85	26.69	13.10
		FRLL-WebMorph	**3.42**	**5.88**	**0.49**	**0.00**
		FRLL-OpenCV	2.97	4.41	0.49	0.49
		FRLL-AMSL	**3.85**	**12.07**	**2.89**	**2.41**
		FRLL-FaceMorpher	1.35	1.47	**0.00**	**0.00**
	ViT-L	FRLL-Style-GAN2	24.96	94.17	49.03	22.33
		FRLL-WebMorph	4.07	6.86	1.47	1.47
		FRLL-OpenCV	0.99	0.98	**0.00**	**0.00**
		FRLL-AMSL	7.26	21.26	10.63	5.80
		FRLL-FaceMorpher	**0.74**	**0.98**	0.98	0.98

further extend our study and verify if MADation shows competitive performance with recent MAD approaches, we compare our proposed framework with several MAD architectures previously proposed in the literature. The comparison considered all the reported results in the literature that complied with the SYN-MAD 2022 competition [27] by training on SMDD [14] and testing on the MAD22 [27] (and its derivatives [13]) as well as the works trained on SMDD and tested on the FRLL-Morphs [48] evaluation benchmark. This might have missed comparisons to some published MAD techniques (that did not follow this protocol or are not publicly available) but provides a wide comparison with many of them and relies on a public and clear benchmark. In Table 2, we start by comparing the MAD techniques [13, 14, 27, 30, 44, 46] with public results on MAD22 [27] and its extension MorDIFF [13], including some of the submitted solutions to the SYN-MAD 2022 competition [27]. MADation presents the best and/or second-best performance in 23/21 (ViT-B/ViT-L) out of the 42 evaluated scenarios. Note that no other approach in Table 2 evaluated BPCER at a fixed APCER for MorDIFF. In particular, ViT-B presents remarkable results in FaceMorph and OpenCV, while ViT-L consistently outperforms the remaining techniques in FaceMorph, MIPGAN II and WebMorph. Although the D-FW [44] approach presents very competitive performance levels for all available metrics, it should be kept in mind that these models use a multi-task learning framework that incorporates 3D facial information. This leads to a significant increase in computational demands and justifies the achieved performance levels. Furthermore, this information could also be included in the proposed MADation approach and would likely result in increased performance. Nonetheless, applications such as these fall out of the scope of the current work, which focuses on providing a computationally inexpensive solution to the MAD task through the usage of LoRA layers to adapt a pre-trained FM. We further extend our study to the well-known FRLL-Morphs dataset [48] in Table 3, to present a more comprehensive comparison with previous works [3, 7, 14, 38]. It can be seen that MADation scores first and/or second place in 12/9 (ViT-B/ViT-L) out of the 20 evaluated scenarios. In particular, ViT-B achieves the lowest EER for FRLL-WebMorph and FRLL-AMSL, reducing the previously best EER values by 0.59 pp. and 0.15 pp., respectively. Overall, MADation presents competitive performances with a wide set of recent MAD solutions, highlighting the importance of exploring FM's potential in biometrics tasks such as MAD.

6. Conclusion

This work presents MADation, the first approach that takes advantage of the generalization capabilities of FMs to perform MAD. To ensure that the pre-trained FM can align its feature space with the domain specificities of MAD, we adapt CLIP with LoRA layers while simultaneously training a classification layer to perform MAD. Through extensive benchmarking on several datasets and comparison with other transformer and FM-based approaches, we show that MADation can take advantage of the knowledge acquired during CLIP's pre-training with massive amounts of data while constituting a flexible approach that efficiently aligns the produced feature space with MAD specificities, resulting in increased performance. Furthermore, MADation showed competitive performance with recent MAD solutions in several evaluation benchmarks, demonstrating the potential of FM's in domain-specific tasks such as MAD provided their correct adaption to the downstream task, even when little training data is available.

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, and the Slovenian ARIS research programs Metrology and Biometric Systems (P2-0250) and Computer Vision (P2-0214).

References

[1] Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020. 4

[2] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020. 5

[3] Ismail Avcibas. Morphed face detection with wavelet-based co-occurrence matrices. *IEEE Signal Process. Lett.*, 31:1344–1348, 2024. 8

[4] 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. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. 1

[5] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1578–1587, 2022. 1

[6] 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. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. 1

[7] Eduarda Caldeira, Pedro C Neto, Tiago Gonçalves, Naser Damer, Ana F Sequeira, and Jaime S Cardoso. Unveiling the two-faced truth: Disentangling morphed identities for face morphing detection. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pages 955–959. IEEE, 2023. 1, 2, 5, 8

[8] Feng Chen, Mario Valerio Giuffrida, and Sotirios A Tsaftaris. Adapting vision foundation models for plant phenotyping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 604–613, 2023. 3, 7

[9] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022. 2, 4

[10] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 2, 4

[11] Tahar Chettaoui, Naser Damer, and Fadi Boutros. Froundation: Are foundation models ready for face recognition? *arXiv preprint arXiv:2410.23831*, 2024. 1, 2, 3, 4, 5, 6, 7

[12] Naser Damer, Viola Boller, Yaza Wainakh, Fadi Boutros, Philipp Terhörst, Andreas Braun, and Arjan Kuijper. 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*, pages 518–534. Springer, 2019. 2

[13] Naser Damer, Meiling Fang, Patrick Siebke, Jan Niklas Kolf, Marco Huber, and Fadi Boutros. Mordiff: Recognition vulnerability and attack detectability of face morphing attacks created by diffusion autoencoders. In *2023 11th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2023. 1, 3, 5, 6, 7, 8

[14] Naser Damer, César Augusto Fontanillo López, Meiling Fang, Noémie Spiller, Minh Vu Pham, and Fadi Boutros. Privacy-friendly synthetic data for the development of face morphing attack detectors. In *CVPR Workshops*, pages 1605–1616. IEEE, 2022. 5, 7, 8

[15] Naser Damer, Alexandra Mosegui Saladie, Andreas Braun, and Arjan Kuijper. 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)*, pages 1–10. IEEE, 2018. 1

[16] Naser Damer, Alexandra Mosegui Saladie, Steffen Zienert, Yaza Wainakh, Philipp Terhörst, Florian Kirchbuchner, and Arjan Kuijper. To detect or not to detect: The right faces to morph. In *2019 International Conference on Biometrics, ICB 2019, Crete, Greece, June 4-7, 2019*, pages 1–8. IEEE, 2019. 3

[17] Naser Damer, Noémie Spiller, Meiling Fang, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. 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*, pages 291–304. Springer, 2021. 1, 2

[18] Lisa DeBruine. debruine/webmorph: Beta release 2. *Zenodo https://doi. org/10*, 5281, 2018. 5

[19] Lisa DeBruine and Benedict Jones. Face Research Lab London Set. 5 2017. 5

[20] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos 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. 1

[21] Meiling Fang, Fadi Boutros, and Naser Damer. Unsupervised face morphing attack detection via self-paced anomaly detection. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11. IEEE, 2022. 1, 2

[22] Parisa Farmanifard and Arun Ross. Iris-sam: Iris segmentation using a foundational model. *arXiv preprint arXiv:2402.06497*, 2024. 1, 2, 3

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

[24] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. On the effects of image alterations on face recognition accuracy. *Face recognition across the imaging spectrum*, pages 195–222, 2016. 2

[25] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 3, 4

[26] Hsin-Ping Huang, Deqing Sun, Yaojie Liu, Wen-Sheng Chu, Taihong Xiao, Jinwei Yuan, Hartwig Adam, and Ming-Hsuan Yang. Adaptive transformers for robust few-shot cross-domain face anti-spoofing. In *ECCV (13)*, volume 13673 of *Lecture Notes in Computer Science*, pages 37–54. Springer, 2022. 5

[27] 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. 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. 1, 2, 5, 6, 7, 8

[28] International Organization for Standardization. ISO/IEC DIS 30107-3:2016: Information Technology – Biometric presentation attack detection – P. 3: Testing and reporting, 2017. 6

[29] International Organization for Standardization. ISO/IEC DIS 20059: Information technology — Methodologies to evaluate the resistance of biometric recognition systems to morphing attacks, 2023. 3, 6

[30] Marija Ivanovska, Andrej Kronovsek, Peter Peer, Vitomir Struc, and Borut Batagelj. Face morphing attack detection using privacy-aware training data. *CoRR*, abs/2207.00899, 2022. 7, 8

[31] Marija Ivanovska and Vitomir Struc. 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*, pages 1–6. IEEE, 2023. 2

[32] Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with lora. *arXiv preprint arXiv:2312.03732*, 2023. 4

[33] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020. 5

[34] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1, 2, 3

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[36] Andrey Makrushin, Tom Neubert, and Jana Dittmann. Automatic generation and detection of visually faultless facial morphs. In *VISIGRAPP (6: VISAPP)*, pages 39–50. SciTePress, 2017. 1

[37] Satya Mallick. Face morph using opencv — c++ / python. *LearnOpenCV*, 1(1), 2016. 5

[38] Pedro C Neto, Tiago Gonçalves, Marco Huber, Naser Damer, Ana F Sequeira, and Jaime S Cardoso. Orthomad: Morphing attack detection through orthogonal identity disentanglement. In *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2022. 1, 2, 5, 8

[39] Tom Neubert, Andrey Makrushin, Mario Hildebrandt, Christian Kraetzer, and Jana Dittmann. Extended stirtrace benchmarking of biometric and forensic qualities of morphed face images. *IET Biometrics*, 7(4):325–332, 2018. 5

[40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2, 3

[41] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model for id-consistent human faces. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1, 2, 3

[42] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *CVPR*. IEEE, 2022. 5

[43] Alyssa Quek. Facemorpher, 2019. 5

[44] Harsh Rachalwar, Meiling Fang, Naser Damer, and Abhijit Das. Depth-guided robust face morphing attack detection. In *IJCB*, pages 1–9. IEEE, 2023. 7, 8

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4, 5, 6

[46] Raghavendra Ramachandra, Sushma Venkatesh, Kiran Raja, and Christoph Busch. 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*, pages 255–265. Springer, 2019. 1, 2, 7, 8

[47] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1

[48] Eklavya Sarkar, Pavel Korshunov, Laurent Colbois, and Sébastien Marcel. Vulnerability analysis of face morphing

attacks from landmarks and generative adversarial networks. *CoRR*, abs/2012.05344, 2020. 5, 8

[49] Ulrich Scherhag, Ramachandra Raghavendra, Kiran B Raja, Marta Gomez-Barrero, Christian Rathgeb, and Christoph Busch. On the vulnerability of face recognition systems towards morphed face attacks. In *2017 5th international workshop on biometrics and forensics (IWBF)*, pages 1–6. IEEE, 2017. 2

[50] Clemens Seibold, Wojciech Samek, Anna Hilsmann, and Peter Eisert. Accurate and robust neural networks for face morphing attack detection. *J. Inf. Secur. Appl.*, 53:102526, 2020. 5

[51] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. *Advances in neural information processing systems*, 32, 2019. 5

[52] Sushma Venkatesh, Haoyu Zhang, Raghavendra Ramachandra, Kiran B. Raja, Naser Damer, and Christoph Busch. Can GAN generated morphs threaten face recognition systems equally as landmark based morphs? - vulnerability and detection. In *IWBF*, pages 1–6. IEEE, 2020. 5

[53] An Wang, Mobarakol Islam, Mengya Xu, Yang Zhang, and Hongliang Ren. SAM meets robotic surgery: An empirical study on generalization, robustness and adaptation. *CoRR*, abs/2308.07156, 2023. 2

[54] Bowen Zhang, Ying Chen, Long Bai, Yan Zhao, Yuxiang Sun, Yixuan Yuan, Jianhua Zhang, and Hongliang Ren. Learning to adapt foundation model dinov2 for capsule endoscopy diagnosis. *arXiv preprint arXiv:2406.10508*, 2024. 3, 7

[55] Haoyu Zhang, Raghavendra Ramachandra, Kiran Raja, and Christoph Busch. 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*, pages 1510–1518, June 2024. 2, 5

[56] Haoyu Zhang, Sushma Venkatesh, Raghavendra Ramachandra, Kiran Bylappa Raja, Naser Damer, and Christoph Busch. MIPGAN - generating strong and high quality morphing attacks using identity prior driven GAN. *IEEE Trans. Biom. Behav. Identity Sci.*, 3(3):365–383, 2021. 1, 5