

Evaluating the Effectiveness of Attack-Agnostic Features for Morphing Attack Detection

Laurent Colbois^{1,2} and Sébastien Marcel^{1,2}

¹ Idiap Research Institute, Switzerland

² Université de Lausanne, Switzerland

{laurent.colbois, sebastien.marcel}@idiap.ch

Abstract

Morphing attacks have diversified significantly over the past years, with new methods based on generative adversarial networks (GANs) and diffusion models posing substantial threats to face recognition systems. Recent research has demonstrated the effectiveness of features extracted from large vision models pretrained on bonafide data only (attack-agnostic features) for detecting deep generative images. Building on this, we investigate the potential of these image representations for morphing attack detection (MAD). We develop supervised detectors by training a simple binary linear SVM on the extracted features and one-class detectors by modeling the distribution of bonafide features with a Gaussian Mixture Model (GMM). Our method is evaluated across a comprehensive set of attacks and various scenarios, including generalization to unseen attacks, different source datasets, and print-scan data. Our results indicate that attack-agnostic features can effectively detect morphing attacks, outperforming traditional supervised and one-class detectors from the literature in most scenarios. Additionally, we provide insights into the strengths and limitations of each considered representation and discuss potential future research directions to further enhance the robustness and generalizability of our approach.

1. Introduction

Morphing attacks pose a significant threat to face recognition (FR) systems. These attacks involve creating a composite passport image that merges facial features from two distinct source identities. This manipulated image is then submitted to governmental services for passport applications, a process still allowed in several European countries where applicants can provide their own photographs. In successful morphing attacks, both contributing individuals can then authenticate against the altered image, enabling them to share a single passport. This undermines the security and effectiveness of automated border control (ABC) systems.

Historically, morphing attacks were primarily generated

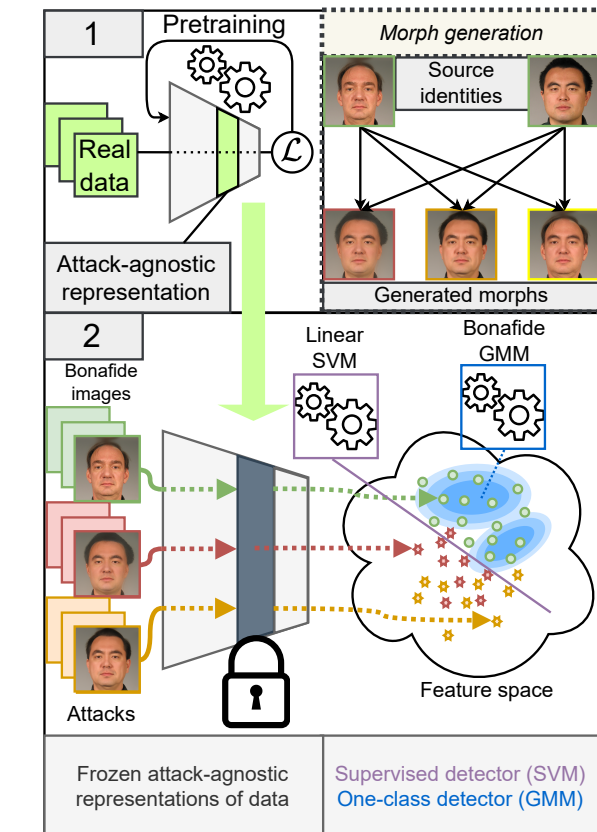


Figure 1. We tackle the problem of MAD using pretrained attack-agnostic extractors. Morph generation: we generate morphs using a variety of algorithms (landmark-based, GAN-based, and diffusion-based). Stage 1: the attack-agnostic extractor is a large vision model trained on real images for a pretext task. We reuse it to summarize any image by extracting an internal representation as the feature vector. Stage 2: features are extracted for bonafide images and face morphs. We train a supervised morphing attack detector as a linear SVM on top of this features space. We train a one-class detector by modeling the distribution of bonafide features with a GMM, then using the likelihood of incoming samples as the discriminative score.

using straightforward image processing algorithms. These methods, typically relying on facial landmark warping,

alignment, and pixel averaging of the source images, are known as **landmark-based** morphing techniques [9]. The vulnerability of existing FR systems to such attacks has been well documented, prompting extensive research into morphing attack detection (MAD) methods.

More recently, advances in generative artificial intelligence have introduced fundamentally different morphing algorithms, particularly those based on GANs [23, 20, 25] and on diffusion models [5]. Resulting morphs are referred to as **deep morphs**, as they leverage deep neural networks for their creation. Initially less effective than landmark-based morphs, deep morphs have rapidly improved, and now also pose a significant concern as they reach real-world applicability.

Detection of deep morphs has often been approached by adapting existing methods designed for landmark-based morphs, such as incorporating deep morphs into the training datasets for data-driven detectors. We propose to consider the reverse perspective: treating MAD as a *deepfake* detection problem, specifically focusing on detecting deep synthetic images. With respect to typical deepfake detection, one must then address two additional challenges: keeping the ability to handle the fundamentally different nature of landmark-based morphs, and ensuring robustness against print-scan post-processing—a degradation not typically considered in deepfake detection literature.

Recent advancements in deepfake detection have demonstrated the unexpected effectiveness of using internal features from large vision models trained exclusively on real data. These features, which are then **attack-agnostic**, can be used in conjunction with simple downstream classifiers to perform detection. Notably, features extracted using pre-trained CLIP models, originally trained for image-caption alignment, have shown promise in previous studies [4, 14].

This study focuses on evaluating the applicability of attack-agnostic features for MAD. Specifically:

- We develop and evaluate MAD systems using simple probe classifiers trained on attack-agnostic feature representations.
- We develop and evaluate MAD systems based on one-class modeling of the bona fide class, and detecting morphs as out-of-distribution samples, an approach which is enabled by the use of attack-agnostic representations.
- We compare our methodology against traditional supervised convolutional neural network (CNN) training, through extensive experiments involving three different datasets and five types of morphing attacks spanning three categories: landmark-based, GAN-based, and diffusion-based. Our evaluation includes a variety of scenarios, focusing on the generalization capability

ties across different families of attacks, across source datasets, and across domains (digital to print-scan).

Source code for regenerating the morphs and reproduce the results is released publicly.¹

2. Related work

Morphing attack detection systems can be broadly categorized into single MAD and differential MAD. Single MAD aims to assess the authenticity of a single image, such as a registered passport picture, while differential MAD also exploits probe information, such as the live-captured image of the passport holder at the Automated Border Control (ABC) gate. We focus here on single MAD which our work is concerned with.

MAD systems can be categorized into those using handcrafted features and those using deep features [21]. Handcrafted features typically rely on texture cues (e.g. Local Binary Patterns (LBPs)) or image forensic cues (e.g. frequency content, Photo Response Non-Uniformity (PRNU)). Deep features, on the other hand, are learned in a data-driven manner by training a neural network (usually a CNN) on examples of bonafide and morphed images.

A significant portion of research has focused on landmark-based morphs, with somewhat more limited attention given to GAN-based morphs and almost none to the more recent diffusion-based morphs, such as those introduced in [5]. Common benchmark datasets, such as the NIST FATE MORPH [1] and the SOTAMD dataset [19], include only a single type of deep morph (GAN-based) or none at all. Similarly, the largest available dataset, SMDD [6], based on synthetic identities, includes only a single landmark-based morphing attack. In practice, handcrafted features developed for landmark-based MAD are not particularly effective for deep morphs, as demonstrated in [22]. The effectiveness of deep features is strongly dependent on the training data, and generalization from a training dataset containing only landmark-based morphs to one containing deep morphs is not guaranteed, as observed in [5].

Notable exceptions include two works that approach MAD as an anomaly detection problem. Both design an image-reconstruction network that aims to degrade then reconstruct bonafide input images. This process is done through an autoencoder in [8], and by a noise-denoise process in [10] using diffusion models. They then observe that the reconstruction error differs between bonafide images and morphs, although it is *lower* for morphs in [8] but *higher* in [10]. The reconstruction error is thus discriminative for detection purposes. One main advantage of such approaches is that they are one-class, relying only on bonafide data and not on specific attacks in the training set, making

¹https://gitlab.idiap.ch/bob/bob.paper.ijcb2024_agnostic_features_mad

them less prone to bias towards a specific family of morphing methods. However, evaluation on diffusion morphs, for example, is not provided in these works.

Finally, [3] demonstrates that generic pre-existing GAN-image detectors are quite effective out-of-the-box for detecting GAN-based morphs in the digital domain. This suggests potential in leveraging methodologies from deep synthetic image detection research and applying them to MAD. The two main additional challenges are handling landmark-based morphs, which are of a different nature, and dealing with the print-scan domain. Recent progress in synthetic image detection, as shown in [14] and [4], indicates that internal representations from existing large vision models, pretrained on auxiliary tasks and real data only (hence, **attack-agnostic**), can be surprisingly effective for synthetic image detection by training a simple downstream classifier on top of extracted features. Similarly to one-class approaches, this method is less prone to overspecialize for a family of morphing algorithms, given that the selected representations are based only on bonafide data.

The core goal of our work is to carefully examine the applicability of attack-agnostic features in the context of MAD, particularly with the inclusion of landmark-based morphs and print-scan data.

3. Methodology

3.1. Morph Datasets

We create morphs using three distinct source datasets: the FRLI dataset [2], the FRGC dataset [17], and the FFHQ dataset [11]. The FRLI and FRGC datasets are extensively utilized in prior research on face morphing due to their constrained facial images (frontal pose, neutral expression) with consistent backgrounds and illumination. These characteristics render them suitable for morph generation. In contrast, the FFHQ dataset, collected from Flickr, exhibits greater diversity. We hypothesize that employing a more diverse source dataset is advantageous for our research objectives, particularly in studying cross-source dataset generalization and one-class modeling of the bonafide class.

For the FRLI and FRGC datasets, we select identity pairs for morph creation following previous research works [13] and [25], respectively. This results in 1,140 pairings for FRLI and 2,521 pairings for FRGC. For the FFHQ dataset, we initially select 10,000 images from the original dataset of 70,000 images, focusing on those with the most frontal poses, which are then randomly paired to form 5,000 morphing pairs. While this process might yield some unrealistic morphs (e.g., morphs between different genders), it allows the creation of a large set of samples containing the relevant attack artifacts and showcasing high diversity. Hence, this set remains valuable for training MAD systems.

Using consistent pairings, we generate morphs from

Table 1. Number of samples in each dataset and split. We indicate the number of attack samples *per morphing algorithm*, i.e. the total number of attack samples used in experiments should be obtained by multiplying the provided value by the number of considered morphing algorithms.

Src. dataset	# bonafide		# per attack	
	Train	Test	Train	Test
FRGC	9228	2304	2014	507
FRLI	-	204	-	1140
FFHQ	8000	2000	4000	1000

Table 2. Available image sets. Attacks are grouped into higher level families indicated in the first row. Most attacks are available in digital format (◦), some of them have their test set in print-scan format as well (●). The FRGC-MIPGAN attack is used only for testing purpose in the print-scan domain (★).

	LB		GAN		Diff		
	LB-Complete[12]	LB-Combined[13]	SG2-W[20]	SG2-W+[23]	MorDIFF[5]	MIPGAN[25]	Bonafide
FRGC	◦	●	◦	◦	◦	★	●
FRLI	◦	◦	◦	◦	◦		◦
FFHQ	◦	◦	◦	◦	◦		◦

these source datasets employing five different attack algorithms. These include two landmark-based algorithms (LB-Complete [12] and LB-Combined [13]), two GAN-based algorithms (SG2-W [11] and SG2-W+ [23]), and one diffusion-based algorithm (MorDIFF [5]). Examples of the generated morphs are presented in Figure 2.

For the bonafide sets, we use original images from the source datasets. For FRLI, we use the only available 204 frontal images, some of which have also been used as sources for morphing. Due to this low amount of bonafide images, we restrict the usage of FRLI to test purposes. For FRGC and FFHQ, we select bonafide images containing identities never used for morphing, with 11,532 and 10,000 images, respectively. We split both the bonafide sets and attack sets into training and test sets using an 80-20 ratio, ensuring that identities are disjoint between the training and test sets for bonafide images, and that pairs of identities are disjoint between the training and test sets for the attacks. The exact number of samples in each dataset is detailed in Table 1.

Additionally, we create a "real-world" test dataset by printing and scanning a subset of images. Specifically, the

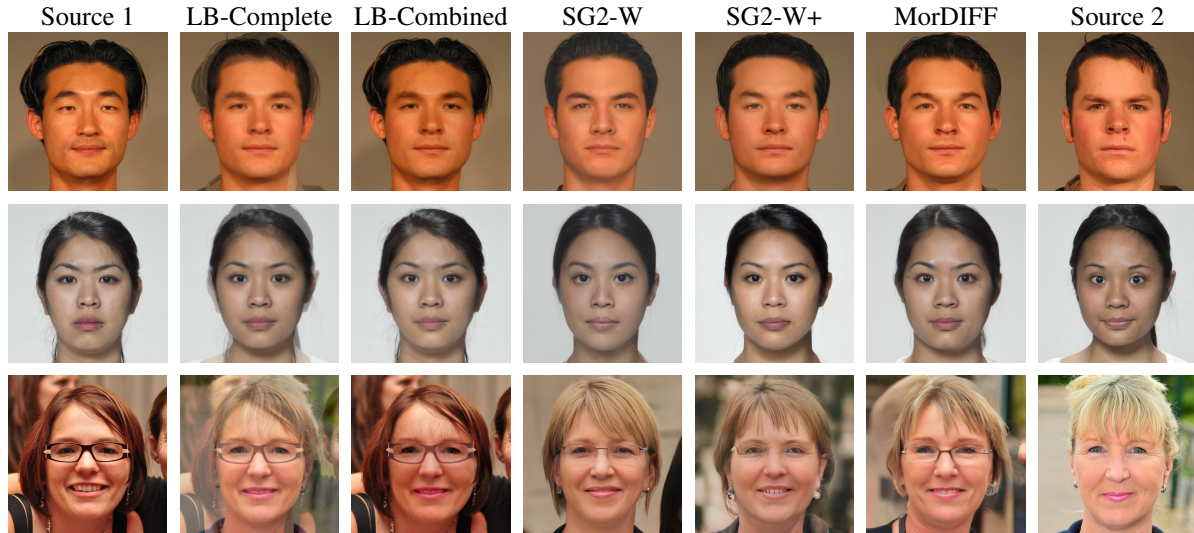


Figure 2. Examples of generated morphs using as source dataset respectively FRGC (first row), FRL (second row) and FFHQ (third row). The first and last column show the two real sources for which a morph must be created, and other columns show the results using each considered morphing algorithm.

bonafide test samples from FRGC, morph test samples created using FRGC with the LB-Combined and MorDIFF algorithms, and an additional set of FRGC morphs created using another unseen algorithm, MIPGAN [25]. This simulates a challenging scenario where we must generalize from the digital to the print-scan domain, and towards unseen attacks. The morphs are printed at a size of 35mm x 35mm then rescanned at a resolution of 300 DPI, using a *Kyocera TASKalfa 2554ci* (laser printer + scanner). As preprocessing, all images are cropped to 256x256 pixels while ensuring consistent landmark alignment.

Available image sets are summarized in Table 2. For the experiments, we regroup attacks into higher level families, respectively landmark-based (LB), GAN-based (GAN), and diffusion-based (Diff).

3.2. Evaluation Scenarios

We aim to evaluate the performance of MAD systems in various settings, with a focus on assessing generalization capability across unseen attack families (LB, GAN, or Diff), unseen source datasets, and different domains (digital to print-scan). Additionally, we seek to evaluate the performance of one-class detectors trained solely on bonafide data. The following evaluation scenarios are considered:

1. **Baseline** : the detector is trained and tested on digital bonafide and morph samples from the same source dataset (FRGC or FFHQ), with all families of attacks seen during training.
2. **Generalization to unseen attacks** : unlike the baseline, the detector is trained using only a single family

of attacks (LB, GAN, or Diff.) and tested on the other two families.

3. **Generalization to different source datasets** : unlike the baseline, the detector is tested on bonafide and morph samples from an unseen source dataset, specifically FRL.
4. **Generalization to print-scan data** : unlike the baseline, the detector is tested on print-scanned bonafide and morph samples. This scenario is evaluated only using FRGC, for which print-scanned data is available.
5. **One-Class Detection** : the detector is trained solely on bonafide samples and then tested on all attacks. For this setting, we restrict ourselves to a single source dataset and to the digital domain.

The first four scenarios involve training the detector in a supervised manner as a binary classifier. In the last scenario, one-class detectors are achieved by modeling the statistical distribution of the features of bonafide samples, then using the likelihood score of incoming samples under the learned distribution as the discriminative score. For both types of systems, performance is evaluated by reporting the Detection Equal Error Rate (D-EER) on the respective test sets.

3.3. Models

We consider two types of detection models. The first type, which is the focus of our study, involves training a simple downstream classifier on top of pretrained features extracted from an attack-agnostic vision model, i.e., a network trained solely on bonafide data for some auxiliary task

(cf. Figure 1). The second type is used for comparison purposes, and consist in fully training a convolutional neural network directly on image samples, either as a binary classifier (in the supervised setting) or as an autoencoder (in the one-class setting).

3.3.1 Probed Attack-Agnostic Models

We consider the following attack-agnostic feature extractors:

- **RN50-IN** [16] : this baseline extractor is a ResNet50 network trained for image classification on ImageNet. We use the output of the penultimate layer before the image classification layer as the feature representation of images.
- **DINOv2** [15] : this extractor is trained in a self-supervised manner with the goal of learning general image representations. It has demonstrated effectiveness for a broad variety of downstream classification tasks and serves as a more sophisticated baseline compared to RN50-IN. We specifically use the ‘giant’ variant, and use the learned general representation as feature vector.
- **CLIP** [18] : this vision-language model is trained to represent matched image-caption pairs jointly in the same feature space. Despite being trained for a seemingly unrelated task, previous research [14, 4] has shown that CLIP-extracted features showcase strong discriminative power to differentiate between bonafide and synthetic images. We use the L/14 variant as suggested by [14], and use the output of the vision encoder as feature vector.
- **AIM** [7] : this extractor is pretrained for auto-regressive image modeling, which involves decomposing images into ordered sequences of patches and predicting subsequent patches using only the context of previous patches. This auto-regressive objective is theoretically equivalent to learning the true underlying image distribution. Trained on a massive dataset of 12.8 billion images, AIM has the potential to approximate the distribution of “natural” images. Given that deep synthetic images typically exhibit salient statistical differences from bonafide ones [4], we hypothesize that they might lie outside of the distribution learned by AIM. We use the 600M variant, and use the pool-averaged output of the trunk as the feature representation.
- **DNADet** [24] : this extractor is originally designed to improve the accuracy of source attribution for GAN-generated images. It is pretrained using real images

for a task of patchwise contrastive learning of image transformations, where images undergo various degradations (e.g., blurring, JPEG compression) and are decomposed into patches. The model learns to represent patches subject to the same degradations close to each other, and patches subject to different degradations far apart, and additionally has to classify incoming patches based on their applied degradation. Given the already demonstrated efficacy of this pretraining in learning salient features to differentiate various GAN models, DNADet is a strong candidate for synthetic image detection, particularly as its pretraining data includes face images, making it also content-specific for our case. We use the output of the penultimate layer, right before the fully connected layer used for the classification, as the feature representation.

For supervised modeling, we train a downstream linear probe on top of the extracted features, specifically a binary linear Support Vector Machine (SVM), preceded by a Principal Component Analysis (PCA) decomposition achieving 99% of explained variance. This initial projection mitigates the challenges posed by the high dimensionality of certain feature spaces.

For one-class modeling, we fit the distribution of bonafide features using a GMM, also preceded by PCA decomposition achieving 99% of explained variance. The log-likelihood of incoming samples under this statistical model is then used to distinguish between bonafide samples, which are expected to have high log-likelihood values, and attacks, which are expected to have low log-likelihood values. To determine the optimal number of components for the GMM (ranging from 1 to 256), as well as the type of covariance matrix (diagonal or spherical), we perform 4-fold cross-validation on the training set. The validation set includes attack samples, and the D-EER on the validation set is used as the selection criterion.

3.3.2 Reference MAD Models

For the supervised detection setting, we use as comparative reference the MixFaceNet architecture, which has been employed in prior work as a backbone for training MAD systems. The model is a CNN trained as a binary classifier directly on image examples. We reproduce the backbone setup and training process as described in [6], and reuse their provided code.² For the one-class setting, we compare our models to the SPL-MAD model from [8]. The SPL-MAD model is trained as a convolutional autoencoder on the Casia-WebFace dataset (bonafide face images). At inference time, the authors observe that the reconstruction er-

²<https://github.com/naserdamer/SMDD-Synthetic-Face-Morphing-Attack-Detection-Development-dataset>

Table 3. **Baseline.** D-EER (%) on the test split when all attacks are seen at training time. **Bold** values indicate setups where probed attack-agnostic models perform better than the MixFaceNet MAD reference. Underlined values are the best performing models.

Src dataset	FRGC			FFHQ		
	LB	GAN	DIFF	LB	GAN	DIFF
Model						
AIM	<u>0.00</u>	<u>0.00</u>	<u>0.00</u>	0.20	0.05	0.05
CLIP	<u>0.00</u>	<u>0.00</u>	0.13	1.45	0.40	1.65
DNADet	<u>0.00</u>	0.04	<u>0.00</u>	5.70	5.75	6.10
DINOv2	<u>0.00</u>	<u>0.00</u>	<u>0.00</u>	5.50	2.45	3.25
RN50-IN	0.04	0.17	<u>0.00</u>	9.70	7.25	6.35
MixFaceNet *	<u>0.00</u>	<u>0.00</u>	<u>0.00</u>	6.70	5.30	7.05

ror is smaller for morphs than for bonafide images, and thus can be used as a discriminative score for detection, despite using only bonafide data at training time. We use the code and pretrained model provided by the authors³.

4. Results

Baseline performance Table 3 shows results for the baseline scenario where both training and testing data come from the same source dataset in the digital domain, and all attacks are known during training. For attacks from a constrained dataset (FRGC), all methods perform well, achieving nearly perfect separation between bonafide and attack samples regardless of the attack family. However, with a more diverse dataset (FFHQ), performance declines for most methods, and differences become more evident. Here, our linear probes generally outperform the MixFaceNet detectors, with AIM and CLIP achieving the best results across all attack families.

Generalization to unseen attacks Table 4 presents results where only one attack family is known during training. For FRGC attacks, AIM features perform best except when only diffusion attacks are known; in this case, MixFaceNet is superior for diffusion to landmark-based generalization, and CLIP is best for diffusion to GAN generalization. The DNADet probe shows comparable generalization from diffusion to both landmark-based and GAN attacks, though it performs slightly worse than MixFaceNet overall.

For FFHQ attacks, CLIP probes consistently outperform AIM probes and MixFaceNet across all generalization scenarios. Our linear probing approach also frequently surpasses MixFaceNet.

Generalization to difference source datasets Table 5 reports results when the source dataset differs between train-

ing and testing, focusing on the FRLL dataset. The experiment highlights in particular the importance of source dataset diversity for effective generalization. Indeed, detectors trained on FFHQ attacks generally perform better on FRLL attacks than those trained on FRGC. This trend holds for both our linear probes and the MixFaceNet detector, with DINOv2 being a notable exception. Overall, linear probes typically outperform MixFaceNet. When trained on FFHQ attacks, AIM and DNADet probes achieve perfect separation between FRLL morphs and bonafide samples but perform poorly when trained on FRGC attacks. In this latter case, CLIP probes provide the most balanced performance for generalization to unseen datasets across all attack types.

Generalization to print-scan data Table 6 shows results when detectors trained on digital data are evaluated on print-scan data. This scenario is challenging because artifacts left by deep morph generators on generated samples are likely degraded during the print-scan process. Most detectors, which achieved perfect separation in the baseline protocol, show significant performance drops on print-scan data (notably LB-PS and Diff-PS, even though they contain attacks whose digital counterpart has been seen during training). The MIPGAN-PS attacks are particularly challenging due to being totally unseen during training. Nevertheless, our linear probes still generally outperform MixFaceNet, with DINOv2 features being the most effective, followed by CLIP.

One-class detector Finally, Table 7 presents the performance of one-class detectors trained only on bonafide data. It is important to note that the comparison to SPL-MAD is not entirely fair, as SPL-MAD is trained on Casia-Webface data, while our detectors are specifically tuned to the considered source dataset, providing an advantage. Nonetheless, for FRGC attacks, AIM and DNADet probes show quite strong performance, and significantly better than SPL-MAD. DNADet probes in particular lead to an impressive D-EER of under 1% for all considered families of attacks, even though the detector is never exposed to any attack for its development. For FFHQ attacks however, the overall detection performance is unsatisfactory, with CLIP features proving to be the most effective in this scenario.

4.1. Discussion

The results demonstrate that the considered attack-agnostic feature representations are highly effective for morphing attack detection. Training simple probes on these features consistently outperforms a CNN detector trained end-to-end on image samples across all generalization scenarios. They also lead to improved performance over an out-of-the-box one-class detector from the recent literature. However, *which* representation is the most effective

³<https://github.com/meilfang/SPL-MAD>

Table 4. **Unseen attacks generalization.** D-EER (%) on the test split when a single family of attacks is seen at training time. **Bold** values indicate setups where probed attack-agnostic models perform better than the MixFaceNet MAD reference. Underlined values are the best performing models.

	Train attacks	LB			GAN			Diff		
		Test attacks	LB	GAN	Diff	LB	GAN	Diff	LB	GAN
Src. dataset	Model									
	AIM	<u>0.00</u>	0.00	0.00	0.22	<u>0.00</u>	0.39	33.81	8.68	<u>0.00</u>
	CLIP	<u>0.00</u>	0.00	1.22	5.21	<u>0.00</u>	5.03	4.34	0.22	<u>0.00</u>
FRGC	DNADet	<u>0.00</u>	2.91	0.00	9.77	<u>0.00</u>	0.65	1.39	1.09	<u>0.00</u>
	DINOv2	<u>0.00</u>	0.69	1.13	10.81	<u>0.00</u>	5.90	7.34	2.78	<u>0.00</u>
	RN50-IN	0.09	6.03	0.00	11.50	0.09	0.74	2.86	4.86	<u>0.00</u>
	MixFaceNet *	<u>0.00</u>	0.48	0.13	2.73	<u>0.00</u>	1.87	<u>0.95</u>	0.61	<u>0.00</u>
	AIM	0.00	11.20	19.60	12.90	0.00	11.90	27.90	13.00	0.00
FFHQ	CLIP	1.25	0.90	8.30	5.70	0.00	7.90	7.75	1.20	0.30
	DNADet	2.30	17.05	27.20	16.15	1.95	41.45	25.30	26.35	0.90
	DINOv2	5.10	8.25	12.20	20.75	0.30	17.85	19.05	10.15	0.55
	RN50-IN	7.75	33.00	17.65	33.90	2.45	36.60	21.90	26.95	2.25
	MixFaceNet *	5.00	18.10	33.75	24.05	0.85	34.55	26.15	26.15	2.40

Table 5. **Source dataset generalization.** D-EER (%) on FRLL bona fide & morph images when all attacks based on a *different* source dataset are seen at training time. **Bold** values indicate setups where probed attack-agnostic models perform better than the MixFaceNet MAD reference. Underlined values are the best performing models.

Train src. dataset	FRGC			FFHQ		
	LB	GAN	DIFF	LB	GAN	DIFF
Model						
AIM	<u>1.47</u>	23.53	11.76	0.00	0.00	0.00
CLIP	6.86	<u>4.90</u>	7.84	3.43	0.49	0.98
DNADet	10.29	35.78	42.65	0.00	0.00	0.00
DINOv2	9.80	8.33	<u>3.92</u>	15.20	13.24	5.88
RN50-IN	13.24	29.41	19.61	1.96	38.73	0.98
MixFaceNet *	12.75	28.92	20.10	2.94	11.76	1.47

is scenario-dependent.

The key outcomes can be summarized as follows:

- **DNADet features** are particularly effective for one-class modeling in the digital domain and when targeting a single passport standard. The DNADet one-class detector achieves a D-EER under 1% for all attack families on FRGC attacks. However, these features exhibit poor performance in print-scan generalization. This limitation is likely due to DNADet’s pretraining task of contrastive learning of image transformations, which may result in a different representation manifold for print-scan images compared to digital ones. Incorporating print-scan data into the bonafide training set may resolve this issue, which we plan to explore in future work.

Table 6. **Print-scan generalization.** D-EER (%) on test split when all digital attacks are seen at training time, but test attacks are in the print-scan domain. **Bold** values indicate setups where probed attack-agnostic models perform better than the MixFaceNet MAD reference. Underlined values are the best performing models.

Src dataset	FRGC		
	LB-PS	MIPGAN-PS	DIFF-PS
Test on			
Model			
AIM	4.77	32.47	30.12
CLIP	<u>3.99</u>	15.02	14.97
DNADet	16.19	60.50	56.55
DINOv2	8.85	<u>5.60</u>	<u>7.51</u>
RN50-IN	20.57	35.24	26.78
MixFaceNet *	22.05	50.22	32.34

- **AIM features** excel for generalizing to unseen attacks but show inconsistencies in other generalization scenarios. While AIM features behave overall similarly to DNADet features, their more irregular performance across different attack families may limit their practicality in real-world applications.
- **DINOv2 features** are particularly suitable for print-scan generalization. In scenarios where we assume a limited known set of possible attacks (i.e., all attacks can be seen during training), these features are valuable when generating actual print-scan data for train-

Table 7. **One-class model.** D-EER (%) on the test split when only bona fide sample are seen at training time. We compare to the SPL-MAD model from [8]. **Bold** values indicate setups where probed attack-agnostic models perform better than the SPL-MAD reference. Underlined values are the best performing true one-class models.

Src. dataset	FRGC			FFHQ		
	LB	GAN	DIFF	LB	GAN	DIFF
Model						
AIM	6.08	0.39	0.00	34.40	56.10	<u>7.20</u>
CLIP	23.87	1.52	20.92	14.50	4.75	27.70
DNADet	0.87	0.82	0.48	27.10	29.10	32.80
DINOv2	35.72	32.86	30.16	35.80	48.90	34.00
RN50-IN	51.56	43.23	18.75	46.75	61.25	46.10
SPL-MAD *	16.28	11.02	20.23	28.15	14.10	34.20

ing is impractical or too time-consuming. Future work should in particular verify if this print-scan generalization performance holds across a wider variety of physical devices.

- **CLIP features**, even though they are rarely the best, consistently perform well across all generalization scenarios, making them interesting for scenarios where multiple generalization challenges are simultaneous. By enabling robust generalization to unseen attacks, strong source dataset generalization, and decent print-scan generalization, they become a strong candidate for training detectors in a supervised way on a small set of attacks. In the one-class setting, CLIP features, while less effective than DNADet on FRGC attacks, are the most effective for FFHQ attacks. Coupled with their strong source dataset generalization capability, this fact makes them potentially well-suited for developing more general-purpose one-class MAD systems that target *multiple* passport standards.

5. Conclusion

Our work highlighted the superior effectiveness of training simple probes on top of attack-agnostic features in morphing attack detection (MAD) compared to traditional supervised CNN training (MixFaceNet) and a one-class detector from the literature (SPL-MAD).

In particular, DNADet features led to remarkable performance in one-class detection scenarios, achieving a D-EER of less than 1% for all attack families on the FRGC dataset. This underscores its efficacy in detecting morphs without prior exposure to attack samples. However, this performance was limited to the digital domain, with DNADet showing low efficacy for generalization to the print-scan domain. This indicates the need to explore whether the inclusion of bonafide print-scan data in the training set of the

DNADet one-class model might enable similar performance in the print-scan domain.

Conversely, DINOv2 excelled in print-scan generalization, making it a promising candidate for contexts where generating large enough print-scan data for training is impractical.

Finally, CLIP, while not always the top performer, consistently delivered solid results across all generalization scenarios. This highlights its potential for developing more versatile MAD systems capable of handling various types of generalization.

Future work will focus on several key areas to further enhance the robustness and generalizability of our proposed approach. First, a more systematic evaluation of one-class detection performance is necessary, particularly to ensure fairer comparisons with existing methods, notably by making sure equivalent bonafide sets are seen at training time. Second, an evaluation of the one-class performance of DNADet in the print-scan domain is needed, likely requiring the inclusion of bonafide print-scan data in the training set. Third, there is potential in specializing attack-agnostic extractors by continuing pretraining using content-specific data, such as bonafide face images. In this work, only DNADet had been pretrained on face data. Lastly, the print-scan generalization capabilities of DINOv2 should be evaluated using additional print-scan devices to verify its effectiveness across a broader range of physical conditions.

In conclusion, the study validates the effectiveness of attack-agnostic representations for MAD, with DNADet and CLIP feature representations standing out in one-class and generalist performances, respectively, and DINOv2 in print-scan generalization. The outlined future work aims to address current limitations and further optimize these models for practical deployment in diverse real-world scenarios.

Acknowledgments

This work was supported by the Swiss Center for Biometrics Research & Testing and the Idiap Research Institute.

References

- [1] Face Analysis Technology Evaluation (FATE) MORPH. https://pages.nist.gov/frvt/html/frvt_morph.html. 2
- [2] Face Research Lab London Set, May 2017. 3
- [3] L. Colbois and S. Marcel. On the detection of morphing attacks generated by GANs. In *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, Sept. 2022. 3
- [4] D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, and L. Verdoliva. Raising the Bar of AI-generated Image Detection with CLIP, Nov. 2023. 2, 3, 5

- [5] N. Damer, M. Fang, P. Siebke, J. N. Kolf, M. Huber, and F. 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, Apr. 2023. [2](#), [3](#)
- [6] N. Damer, C. A. F. López, M. Fang, N. Spiller, M. V. Pham, and F. Boutros. Privacy-friendly Synthetic Data for the Development of Face Morphing Attack Detectors. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1605–1616, June 2022. [2](#), [5](#)
- [7] A. El-Nouby, M. Klein, S. Zhai, M. A. Bautista, A. Toshev, V. Shankar, J. M. Susskind, and A. Joulin. Scalable Pre-training of Large Autoregressive Image Models, Jan. 2024. [5](#)
- [8] M. Fang, F. Boutros, and N. Damer. Unsupervised Face Morphing Attack Detection via Self-paced Anomaly Detection, Aug. 2022. [2](#), [5](#), [8](#)
- [9] M. Ferrara, A. Franco, and D. Maltoni. The magic passport. In *IEEE International Joint Conference on Biometrics*, pages 1–7, Sept. 2014. [2](#)
- [10] M. Ivanovska and V. Štruc. Face Morphing Attack Detection with Denoising Diffusion Probabilistic Models. *2023 11th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, Apr. 2023. [2](#)
- [11] T. Karras, S. Laine, and T. Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, June 2019. [3](#)
- [12] A. Makrushin, T. Neubert, and J. Dittmann. Automatic Generation and Detection of Visually Faultless Facial Morphs: In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 39–50, Porto, Portugal, 2017. SCITEPRESS - Science and Technology Publications. [3](#)
- [13] T. Neubert, A. Makrushin, M. Hildebrandt, C. Krätzer, and J. Dittmann. Extended StirTrace benchmarking of biometric and forensic qualities of morphed face images. *IET Biom.*, 2018. [3](#)
- [14] U. Ojha, Y. Li, and Y. J. Lee. Towards Universal Fake Image Detectors that Generalize Across Generative Models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24480–24489, Vancouver, BC, Canada, June 2023. IEEE. [2](#), [3](#), [5](#)
- [15] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, Apr. 2023. [5](#)
- [16] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. [5](#)
- [17] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 947–954 vol. 1, June 2005. [3](#)
- [18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, July 2021. [5](#)
- [19] K. Raja, M. Ferrara, A. Franco, L. Spreeuwens, I. Batskos, F. De Wit, M. Gomez-Barrero, U. Scherhag, D. Fischer, S. K. Venkatesh, J. M. Singh, G. Li, L. Bergeron, S. Isadskiy, R. Ramachandra, C. Rathgeb, D. Frings, U. Seidel, F. Knopjes, R. Veldhuis, D. Maltoni, and C. Busch. Morphing Attack Detection-Database, Evaluation Platform, and Benchmarking. *IEEE Transactions on Information Forensics and Security*, 16:4336–4351, 2021. [2](#)
- [20] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel. Are GAN-based morphs threatening face recognition? In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2959–2963, May 2022. [2](#), [3](#)
- [21] U. Scherhag, C. Rathgeb, and C. Busch. Face Morphing Attack Detection Methods. In C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, and C. Busch, editors, *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*, pages 331–349. Springer International Publishing, Cham, 2022. [2](#)
- [22] J. E. Tapia and C. Busch. Face Feature Visualisation of Single Morphing Attack Detection. *2023 11th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, Apr. 2023. [2](#)
- [23] S. Venkatesh, H. Zhang, R. Ramachandra, K. Raja, N. Damer, and C. Busch. Can GAN Generated Morphs Threaten Face Recognition Systems Equally as Landmark Based Morphs? - Vulnerability and Detection. In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, Apr. 2020. [2](#), [3](#)
- [24] T. Yang, Z. Huang, J. Cao, L. Li, and X. Li. Deepfake Network Architecture Attribution. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4):4662–4670, June 2022. [5](#)
- [25] H. Zhang, S. Venkatesh, R. Ramachandra, K. Raja, N. Damer, and C. Busch. MIPGAN—Generating Strong and High Quality Morphing Attacks Using Identity Prior Driven GAN. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(3):365–383, July 2021. [2](#), [3](#), [4](#)