Face Morphing Attack Detection with Denoising Diffusion Probabilistic Models

Marija Ivanovska, Vitomir Struc
Faculty of Electrical Engineering, University of Ljubljana, Trzaska cesta 25, Ljubljana, Slovenia

***Abstract*—Morphed face images have recently become a growing concern for existing face verification systems, as they are relatively easy to generate and can be used to impersonate someone's identity for various malicious purposes. Efficient Morphing Attack Detection (MAD) that generalizes well across different morphing techniques is, therefore, of paramount importance. Existing MAD techniques predominantly rely on discriminative models that learn from examples of bona fide and morphed images and, as a result, often exhibit sub-optimal generalization performance when confronted with unknown types of morphing attacks. To address this problem, we propose a novel, diffusion–based MAD method in this paper that learns only from the characteristics of bona fide images. Various forms of morphing attacks are then detected by our model as out-of-distribution samples. We perform rigorous experiments over four different datasets (CASIA-WebFace, FRLL-Morphs, FERET-Morphs and FRGC-Morphs) and compare the proposed solution to both discriminatively-trained and once-class MAD models. The experimental results show that our MAD model achieves highly competitive results on all considered datasets.**

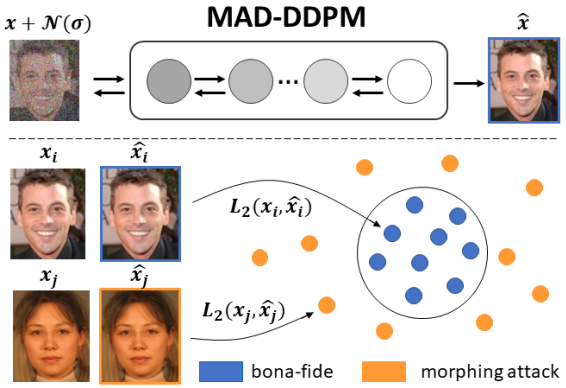

Figure 1. **Illustration of MAD-DDPM.** MAD-DDPM is a (reconstruction-based) one-class face morphing attack detection (MAD) model that uses a probabilistic (denoising) diffusion process to learn the distribution of bone-fide samples. At run-time, face morphs are detected based on the produced reconstruction error. Unlike the majority of competing MAD techniques, MAD-DDPM requires no attack examples during training.

I. Introduction

Automatic face recognition systems (FRSs) are widely used to verify a person's identity by matching the face image of an individual to the data enrolled in the system's database. While such systems are today widely deployed and highly accurate [13], they are known to be prone to certain types of attacks with manipulated data, such as morphing attacks [12], [15], [36]. Because face morphs are created by blending/morphing the facial appearances of at least two different people, a single morphed image can be utilized to falsely authenticate all individuals, whose face has been used during the morph-generation process.

With recent advancements in generative models and the availability of open-source morphing techniques, the generation of highly realistic, high–quality morphed face images has become an almost effortless process. The successful detection of *face morphing attacks* is, hence, crucial for the prevention of illegal activities [7]. While significant progress has been achieved in morphing attack detection (MAD), the majority of existing solutions learn to detect morphed faces discriminatively, i.e., by analyzing and learning the differences between bona fide and morphed samples. Such techniques have been shown to be very accurate when evaluated on morphing techniques seen during training, but often fail to detect morphs created by unknown morphing attacks. Moreover, when evaluated on data from unknown sources, their accuracy is usually adversely affected by domain shifts.

To address the generalization capabilities of MAD models, some researchers explored the use of the one-class models [4], [11], [16], where only bona fide images are used in the training phase. Such models, are generally expected to generalize better to unseen morphs and are also at the heart of this work. Specifically, we propose in this paper a novel one-class MAD technique (MAD-DDPM) that exploits Denoising Diffusion Probabilistic Models (DDPMs) for the detection task. We evaluate the model in comprehensive experiments over multiple datasets and in comparison to both discriminatively-trained and one-class MAD competitors with promising results.

II. Related work

In this section, we discuss background information and related work on morphing-attack-detection (MAD) and diffusion models to provide the necessary context for our research. For a more comprehensive coverage of these topics, the reader is referred to some of the excellent surveys available in the literature, e.g., [3], [37].

A. Morphing attack detection

Existing morphing attack detection (MAD) models can in general be grouped into single-image (S-MAD) and differential (D-MAD) models. The first category of models examines facial morphs independently one from the other, while the latter are comparing manipulated samples to a reference. D-MADs are generally very accurate in closed-group problems, while S-MADs are predominantly used to detect attacks without prior knowledge of the subjects'

identities. We limit the literature review in this section to S-MADs only, as they are most closely related to our work.

Regardless of the face morphing technique used, the generated morphs typically contain image irregularities, such noise, pixel discontinuities, distortions, spectrum discrepancies, and similar artifacts. With early MADs, such irregularities were often detected using hand-crafted techniques utilizing photo-response non-uniformity (PRNU) noise [35], reflection analysis [38] or texture-based descriptors, such as LBP [24], LPQ [25] or SURF [20]. Although these methods yielded promising results, their generalization capabilities were shown to be limited [4].

More recent MADs take advantage of the capabilities of data–driven, deep-learning algorithms [15]. Raghavendra *et al.* [29] were amongst the first to propose transfer learning, with pretrained deep models for this task. In their work, attacks were detected with a simple, fully-connected binary classifier, fed with fused VGG19 and AlexNet features, pretrained on ImageNet. Wandzik *et al.* [40], on the other hand, achieved high detection accuracy with features extracted with general-purpose face recognition systems (FRSs), fed to an SVM. Ramachandra *et al.* [30] utilized Inception models in a similar manner, while Damer *et al.* [7] argued that pixel-wise supervision, where each pixel is classified as a bona fide or a morphing attack, is superior, when used in addition to the binary, image-level objective. Recently, MixFaceNet [2] by Boutros *et al.* achieved state-of-the-art results in different detection tasks, including face morphing detection [5]. This model represents a highly efficient architecture that captures different levels of attack cues through differently-sized convolutional kernels.

Different from the supervised techniques discussed above, some authors have advocated the use of *one-class learning* models trained on bone-fide samples only to improve the generalization capabilities of the MAD techniques. Damer *et al.* [4], for example, were among the first to achieve significant performance generalization on unseen attacks with two different one–class methods, i.e. a one-class support vector machine (OCSVM) and an isolation forest (ISF). Similar generalization capabilities were later demonstrated in [16], where Ibsen *et al.* explored the use of a Gaussian Mixture Model (GMM), Variational Autoencoder (VAE) and Single-Objective Generative Adversarial Active Learning (SO-GAAL) in addition to an OCSVM. In a recent study, Fang *et al.* [11] proposed a one–class convolutional autoencoder, enhanced with a self-paced learning (SPL) algorithm. Here, the authors found that morphing attacks are easier to reconstruct in comparison to non-manipulated samples. The MAD-DDPM model, proposed in this paper, also falls into the group of one-class learning models, but relies on a probabilistic diffusion process to learn the distribution of bona-fide face images.

B. Diffusion models

Denoising Diffusion Probabilistic Models (DDPMs) have recently been found to be exceptionally powerful models for various computer-vision tasks [3], [21], [32]. DDPMs, first introduced by Ho *et al.* [14], were shown to be able to generate high–quality images sampled from pure Gaussian noise. These methods learn to gradually add noise to training samples and to perform denoising, by iteratively maximizing the data likelihood. Although early models have shown impressive generative capabilities, their sampling techniques are time-consuming and often affect the image quality of the generated samples.

Shortly after the initial release of DDPMs, Nichol *et al.* [23] proposed an improved optimization criterion that significantly sped up the noise removal, while maintaining the quality of the generated data. Song *et al.* [39] proposed their own solution for faster sampling and easier deployment of the diffusion process. Dhariwal *et al.* [10] built on these findings and showed that DDPMs can outperform GANs on image synthesis. In a recent study, Karras *et al.* [17] explored different approaches for image generation with diffusion and provided guidelines related to the architectural design and the optimization strategy of DDPMs. Rombach *et al.* [31] successfully reduced the complexity of the diffusion models, by implementing the diffusion process in the latent space of a pretrained autoencoder with minimal degradation in image quality.

Although DDPMs were primarily developed for the generation of new data, they have also been adapted to one-class-learning algorithms. Wolleb *et al.* [41], for example, trained an image-to-image diffusion model, that learned to reconstruct medical images of healthy subjects through the iterative denoising process. A similar technique, proposed by Wyatt *et al.* [42], used Simplex noise, instead of the common Gaussian noise. In contract, Pinaya *et al.* [28] detected anomalies by utilizing diffusion models in the latent space, where the noise reversal is much more efficient.

III. Methodology

A considerable cross-section of existing MAD techniques uses discriminatively trained models for the detection of facial morphs and, as a result, often struggles with the generalization to unseen morphing attacks. In this section, we propose a novel MAD model, MAD-DDPM, that is trained with bona-fide samples only and is, therefore, expected to generalize better to various (unknown, unseen) types of morphing attacks.

A. Overview of MAD-DDPM

A high-level overview of the proposed MAD-DDPM model is presented in Figure 2. The model follows the (self-supervised, one-class) reconstruction-based framework to anomaly detection [1], [45], where a generative model is learned to reconstruct so-called *normal* training data, i.e., bona-fide face images, frome noisy inputs. Because *anomalies* (face morphs in our case) deviate from the distribution of the normal samples, they are expected to generate (comparably) larger reconstruction errors. Consequently, these errors can be used to determine whether the given data sample is normal (bona-fide) or anomalous (morph), as illustrated on the left of Figure 2.

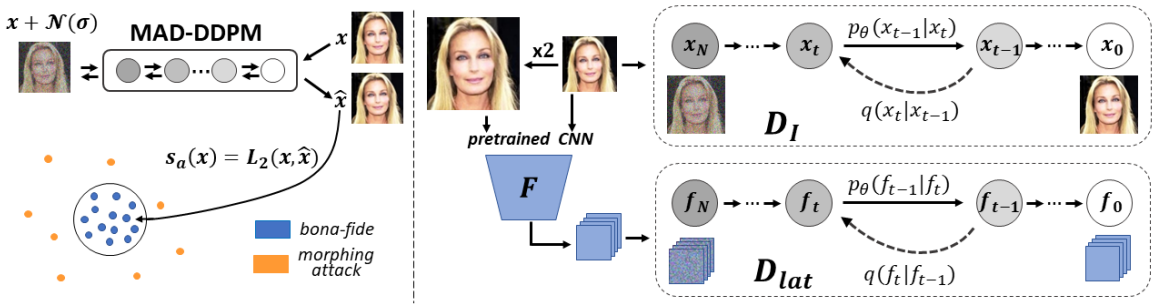

Figure 2. **High-level overview of the proposed MAD-DDPM model.** MAD-DDPM is a one-class learning model that uses a reconstruction-based measure to determine whether the input images are bona fide or face morphs (shown on the left). At the core of the technique is a two-branch reconstruction procedure that uses denoising diffusion probabilistic models (DDPMs) learned over only bona-fide samples as the basis for the detection tasks (shown on the right). Here, the first branch models the distribution on bona-fide samples directly in the pixel-space (for low-level artifact detection), while the second captures the distribution of higher-level features extracted with a pretreind CNN F.

While different generative models have been used in the literature for reconstruction-based anomaly detection (e.g., autoencoders, GANs, etc.), they were often observed to generalize too well beyond the training data, leading to comparable reconstructions for both normal and anomalous data. For the MAD-DDPM we, therefore, design a powerful reconstruction procedure that: (i) results in larger differences in the reconstructions of bona-fide and morphed images than competing (one-class) MAD solutions, and (ii) consequently results in better performance. The procedure is based on Denoising Diffusion Probabilistic Models (DDPMs) and the following considerations:

Complementary data representation: The reconstruction task is learned over two data representation, i.e., (i) the pixel space, where the goal is to model image-level (bona-fide) facial characteristics and to facilitate the detection of low-level image artifacts, and (ii) a feature space that captures higher-level semantic cues of the training data, enabling the detection of potentially more abstract data irregularities.

Compact distribution modelling: To ensure the generative models do not generalize too well beyond the data used for learning, efficient modeling techniques are needed that result in compact distributions of the training data. In MAD-DDPM, we model the data distribution of the bona-fide samples using a probabilistic denoising diffusion process across two data representations, which allows us to efficiently capture the characteristics of the bona-fide samples in a compact manner. This leads to highly competitive MAD performance, as demonstrated in Section V.

In the following sections, we present the theoretical background behind DDPMs, discuss the design of the MAD-DDPM reconstruction procedure, and elaborate on the detection-score computation step.

B. Denoising Diffusion Probabilistic Models (DDPMs)

DDPMs are likelihood–based generative methods, that learn to model a given data distribution $p_{data}(\mathbf{x})$ with standard deviation σ_{data} by employing a two–stage approach [14]. In the first stage, a forward diffusion process is applied to the data $\mathbf{x}_0 \quad p_{data}(\mathbf{x})$, by gradually corrupting the sample x_0 with Gaussian noise $N(0, \sigma^2\mathbf{I})$. The noising technique results in a noisy sample x_N and represents a non–homogeneous Markov chain:

$$q(\mathbf{x}_t j\mathbf{x}_t \quad _1) = N(\mathbf{x}_t j\mathbf{x}_t \quad _1 \sqrt{1 \quad \beta_t}, \beta_t\mathbf{I}) \tag{1}$$

where t is the time step from a predefined time sequence $ft_0, t_1, ...t_N g$, while $\beta_t = \sigma_t^2$ defines the amount of noise added at each step and its value is determined by a variance schedule. Following the recommendations from [17], we implement a linear variance schedule, found to work best in terms of sampling speed and generated data quality. The forward process defined with Eq. (1) enables fast sampling of $\mathbf{x}_t$ at any time step t:

$$q(\mathbf{x}_t j\mathbf{x}_0) = N(\mathbf{x}_t; p\overline{\bar{\alpha}_t}\mathbf{x}_0, (1 \quad \bar{\alpha}_t)\mathbf{I})$$

where $\bar{\alpha}_t = 1 \quad \beta_t$ and $\bar{\alpha}_t = \Pi_{i=1}^{t}\alpha_i$. In the second stage, a generative model parametrized by θ performs sequential denoising of $\mathbf{x}_N$ according to:

$$p_\theta(\mathbf{x}_t \quad _1 j\mathbf{x}_t, \mathbf{x}_0) = N(\mathbf{x}_t \quad _1 j\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t\mathbf{I}) \tag{2}$$

where $t = t_N, t_N \quad _1, ...t_0$, $\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) = \frac{p\overline{\bar{\alpha}_{t-1}}\beta_t}{1 \quad \bar{\alpha}_t}\mathbf{x}_0 + \frac{p\overline{\bar{\alpha}_t}(1 \quad \bar{\alpha}_{t-1})}{1 \quad \bar{\alpha}_t}x_t$ and $\tilde{\beta}_t = \frac{1 \quad \bar{\alpha}_{t-1}}{1 \quad \bar{\alpha}_t}\beta_t$. The mean function $\tilde{\mu}_t$ is optimized by an aproximator $D_\theta(\mathbf{x}, \sigma)$, trained to minimize the expected L_2 denoising error:

$$L = \mathbb{E}_{\mathbf{x} \quad p_{data}}\mathbb{E}_{\mathbf{n} \quad N(0,\sigma^2\mathbf{I})} jjD(\mathbf{x} + \mathbf{n}; \sigma) \quad \mathbf{x}jj_2^2 \tag{3}$$

where $D_\theta(\mathbf{x}, \sigma)$ is a neural network. For MAD-DDPM, an unconditional U-Net [33] architecture, originally proposed in [14], is selected for the implementation of this network. For efficiency reasons, we leverage the recently published DPM-Solver [19], a dedicated high-order solver for diffusion ordinary differential equations (ODEs).

C. Reconstruction and Detection-Score Computation

MAD-DDPM uses a two-branch reconstruction procedure to model the distribution of the bona fide samples, as shown in Figure 2. The first DDPM branch, D_I, is modeling the distribution of bona fide face images in the pixel space. The second DDPM branch, D_{lat}, operates in

the feature space of a pretrained convolutional network F, that extracts high-level image representations over two different scales. Here, the calculated feature maps are concatenated before feeding them to the dedicated diffusion model. Each DDPM branch of the model is learned independently of the other to reduce the computational effort and reduce cross-talk and interactions between low-level image characteristics and higher-level semantic cues.

During run-time, the probability of an image $\mathbf{x}_n$ to be a morphing attack is quantified using the score s_a, calculated by summing up the reconstruction errors of the two diffusion branches, i.e.:

$$s_a(\mathbf{x}_n) = \mathsf{D}_I(\mathbf{x}_n + \mathbf{n}_I; \sigma_I) + \mathsf{D}_{lat}(F(\mathbf{x}_n) + \mathbf{n}_F; \sigma_F) \quad (4)$$

Because our main goal is to detect face morphing artifacts, MAD-DDPM performs the iterative noising with a relatively low σ_{max}, which leads to moderately noised samples. In contrast to existing generative DDPMs, our model is, therefore, unable to generate new samples directly from noise. Instead, it is conditioned on the noisy input $\mathbf{x}_n + \mathbf{n}_I$ and aims to recover information that has been obscured during the forward noising process.

IV. Experimental Setup

A. Datasets

We primarily use four publicly available datasets for the experiments: CASIA-WebFace [43], FERET-Morphs, FRLL-Morphs and FRGC-Morphs [34]. Images from all datasets are first preprocessed by RetinaFace [9] to localize the facial areas. Next, these areas are cropped with a margin equal to 5% of the bounding box height. With this strategy, we ensure, that the cropped images include pixels surrounding the face area, as this is where a considerable amount of morphing artifacts is typically located. Finally, the cropped images are resized to 224 224 pixels and fed to the MAD model. The training of the model is performed in a one–class learning manner, with bona fide images only. In the testing phase, we use three different datasets consisting of both, bona fide and morphed images.

Training data. The MAD models are trained on CASIA-WebFace [43], a large-scale dataset used commonly for face verification and identification tasks. The dataset consists of 494.414 face images of 10.575 unique subjects, collected from the internet. The dataset was designed to include a wide variety of face poses and expressions, captured under different illumination settings and with different image resolutions.

Testing data. Testing is done on three common morphing datasets proposed by Sarkar *et al.* in [34], i.e. FRLL-Morphs, FERET-Morphs and FRGC-Morphs. All morphed face images were created by merging bona fide samples from their respective source datasets, i.e. FRLL [8], [22], FERET [27] and FRGC [26]. For the generation of the landmark–based morphs, the authors used OpenCV and FaceMorpher, while deep–learning–based morphs were generated with StyleGAN2. Additionally, image samples

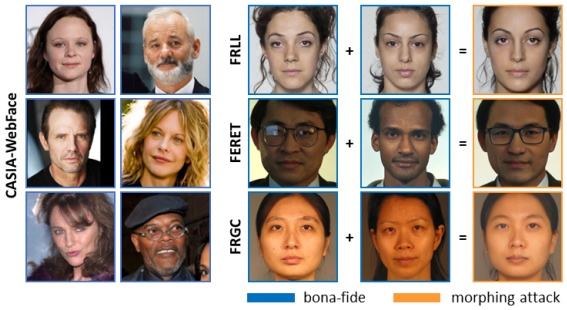

Figure 3. **Sample images from the datasets used for the evaluation.** The figure shows bona fide training images from CASIA-WebFace [43] (left) and bona fide and morphed examples from FERET-Morphs, FRLL-Morphs and FRGC-Morphs [34] (right).

Table I
NUMBER OF BONA FIDE IMAGES (BF) AND MORPHING ATTACKS (MA) IN EACH TEST DATASET. THE MORPHS ARE GENERATED BY 5 MORPHING METHODS, I.E. OPENCV (OCV), FACEMORPHER (FM), STYLEGAN (SG), AMSL, WEBMORPH (WM).

Dataset	Image size	BF	OCV	FM	SG	AMSL	WM
FRLL-M	1350 × 1350	204	1221	1222	1222	2175	1221
FERET-M	512 × 768	1.413	529	529	529	/	/
FRGC-M	227 × 277	3.167	964	964	964	/	/

from FRLL, have been used as a source for morph generation with two more morphing methods, AMSL [22] and Webmorph.The characteristics of the datasets are given in Table I and a few examples are presented in Figure 3.

Training data for supervised MADs. MAD-DDPM is also evaluated against selected discriminatively trained MAD methods, learned on morphs from 3 datasets not used for our evaluations, i.e. LMA-DRD [7], MorGAN [6] and SMDD [5]. LMA-DRD morphs are generated with OpenCV. Digital morphs are labeled with D, while re–digitalized (printed then scanned) morphs are labeled with PS. MorGAN also consists of two types of moprhs: LMA, generated with OpenCV and deep learning-based morphs, generated with a GAN model. SMDD, on the other hand, contains synthetically generated data, where both, bona fide and attack samples are created with StyleGAN2.

B. Evaluation metrics

The model evaluation follows the testing protocol proposed in [11]. Based on morphing attack scores, we first calculate the proportion of attack samples misclassified as bona fide, i.e. the Attack Presentation Classification Error Rate (APCER). We also calculate the proportion of bona fide samples misclassified as attacks, i.e. the Bona fide Presentation Classification Error Rate (BPCER). The overall detection accuracy is then reported in terms of the Equal Error Rate (EER), where APCER equals BPCER.

C. Implementation details

The input to MAD-DDPM consists of RGB images and the corresponding feature maps extracted with a WideResNet50 [44] (model F), pretrained on ImageNet. The feature extraction is performed on two different scales, to better capture differently sized patterns. First, images

Table II
COMPARISON OF MAD-DDPM AND THE CURRENT SOTA ONE-CLASS SPL-MAD APPROACH IN TERMS OF EER (%).

Dataset	Morphing methods	SPL-MAD [11]	MAD-DDPM (Ours)
FRLL-M	OpenCV	3.63	**3.55**
	FaceMorpher	**2.98**	4.04
	StyleGAN2	15.14	**10.96**
	WebMorph	**12.29**	14.49
	AMSL	**11.22**	11.67
FERET-M	OpenCV	32.13	**30.81**
	FaceMorpher	27.69	**25.14**
	StyleGAN2	32.57	**23.25**
FRGC-M	OpenCV	36.11	**27.17**
	FaceMorpher	23.99	**23.23**
	StyleGAN2	36.79	**11.41**
Average performance		21.32	**16.88**

of size 224 224 are fed to the WideResnet to calculate feature maps of size 1024 14 14. Next, each RGB image is resized and split into 4 non–overlapping patches, that are passed through WideResNet, to get 4 additional feature maps. The DDPM branch, labeled as D_I (Figure 2), is then optimized on raw RGB images with $\sigma_{max} = 8$, while D_{lat} is trained on the concatenated feature maps, with $\sigma_{max} = 2$. The σ_{max} values were determined based on preliminary experiments. The DDPMs in the proposed model are optimized with AdamW [18]. The learning rate is set to 0.0001, β_1 and β_2 to 0.95 and 0.999, respectively, while the weight decay is set to 0.001.

MAD-DDPM is implemented in Python 3.8 with PyTorch 1.9 and CUDA 11.7. Experiments were run on a single NVIDIA GeForce RTX 3090, where MAD-DDPM required around $0.6s$ to perform the MAD procedure for a single image on average. Additional implementation details can be found in the publicly released source code (made available after the reviewing procedure).

V. RESULTS

Comparison to One-Class Competitors. We first compare MAD-DDPM to the current state-of-the-art (SOTA) one-class SPL-MAD approach [11]. The results in Table II show that MAD-DDPM achieves very competitive results on FRLL-Morphs. In the detection of StyleGAN morphing attacks, it outperforms SPL-MAD by over 5% in terms of EER, while producing comparable results on the remaining morphs. On the other two datasets, MAD-DDPM consistently outperforms SPL-MAD across all types of morphing attacks. Overall, MAD-DDPM achieves an average EER of 16.88%, outperforming the current one-class SOTA method by a margin of over 4%.

Comparison to Discriminative MAD models. Similarly to [11], we also compare MAD-DDPM to SOTA discriminative MAD techniques in Table III, i.e., MixFaceNet [2], PW-MAD [7] and Inception [30]. The discriminative models are learned in a two-class setting, where a different set of morphing attacks is chosen in each training session. Although the best EER in individual categories of morphing attacks is achieved by the discriminative MADs, none of the trained discriminative models shows consistently superior results across different datasets and morphing attack types. Moreover, the average morphing attack detection performance is by far the highest for MAD-DDPM with an average EER of 16.88%. Based on these results, we conclude that our one-class MAD-DDPM approach demonstrates strong generalization capabilities.

Ablation study. MAD-DDPM is trained on three different data sources, i.e. RGB images (I) and feature maps from two different image scales (S1 and S2). The contribution of each data source is investigated in an ablation study, where we train three independent DDPMs, one for each data source. A separate DDPM, is trained with concatenated CNN features of both scales. As can be seen from Table IV, among all ablated models, the highest detection accuracy is achieved by the DDPM trained on RGB images. We hypothesize, that due to the nature of DDPMs, such approach efficiently detects high-frequency components representing image artifacts induced by the morphing techniques. Conversely, the extracted features encode high-level semantics that are comparably less informative (yet still important) for the morphing detection task. They do however consistently boost the detection performance in all test datasets. The complete MAD-DDPM model outperforms all ablated models with an average EER of 16.88%.

VI. CONCLUSION

We presented a one-class model for morphing attack detection (MAD) that relies on denoising diffusion probabilistic models (DDPM). In comprehensive experiments, the model was shown to result in highly competitive performance on multiple datasets. As part of our future work, we plan to incorporate additional proxy task into the proposed model to further improve results.

REFERENCES

[1] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon. GANomaly: Semi-supervised Anomaly Detection via Adversarial Training". In *ACCV*, 2019.

[2] F. Boutros, N. Damer, M. Fang, F. Kirchbuchner, and A. Kuijper. MixFaceNets: Extremely Efficient Face Recognition Networks. In *IEEE IJCB*, 2021.

[3] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah. Diffusion Models in Vision: A Survey. *arXiv:2209.04747*, 2022.

[4] N. Damer, J. H. Grebe, S. Zienert, F. Kirchbuchner, and A. Kuijper. On the Generalization of Detecting Face Morphing Attacks as Anomalies: Novelty vs. Outlier Detection. In *IEEE BTAS*, 2019.

[5] 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. *IEEE CVPRW*, 2022.

[6] N. Damer, A. M. Saladié, A. Braun, and A. Kuijper. MorGAN: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Generative Adversarial Network. In *IEEE BTAS*, 2018.

[7] N. Damer, N. Spiller, M. Fang, F. Boutros, F. Kirchbuchner, and A. Kuijper. PW-MAD: Pixel-Wise Supervision for Generalized Face Morphing Attack Detection. In *Springer AVC*, 2021.

[8] L. DeBruine and B. Jones. Face Research Lab London Set, 2017.

[9] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In *IEEE CVPR*, 2020.

[10] P. Dhariwal and A. Nichol. Diffusion Models Beat GANs on Image Synthesis. In *NeurIPS*, 2021.

[11] M. Fang, F. Boutros, and N. Damer. Unsupervised Face Morphing Attack Detection via Self-paced Anomaly Detection. In *IJCB*, 2022.

Table III
COMPARISON OF MAD-DDPM AND DISCRIMINATIVE SOTA MADS IN TERMS OF EER (%).

MAD type		**Discriminatively trained**															**One-class**
MAD model		**MixFaceNet** [2]					**PW-MAD** [7]					**Inception** [30]					**MAD-DDPM (Ours)**
Test data \ Train data		D	PS	LMA	GAN	SMDD	D	PS	LMA	GAN	SMDD	D	PS	LMA	GAN	SMDD	
FRLL-M	OpenCV	8.82	13.22	8.91	17.66	4.39	17.33	15.69	13.96	45.59	**2.42**	13.72	10.76	6.86	55.89	5.38	3.55
	FaceMorpher	7.80	10.97	7.34	15.65	3.87	13.88	15.14	10.92	44.57	**2.20**	16.62	15.81	6.32	66.14	3.17	4.04
	StyleGAN2	20.07	15.29	13.41	23.51	**8.89**	29.97	27.64	18.11	48.53	16.64	37.24	19.58	20.56	55.03	11.37	10.96
	WebMorph	25.97	29.04	20.61	30.39	12.35	33.78	28.51	35.75	52.43	16.65	57.38	58.32	30.88	77.42	**9.86**	14.49
	AMSL	24.53	27.59	19.24	30.03	15.18	36.25	32.95	34.38	48.52	15.18	49.02	61.44	**9.80**	86.49	10.79	11.67
FERET-M	OpenCV	28.12	32.19	31.57	33.86	31.74	37.27	45.29	34.27	43.11	39.93	**6.39**	7.23	42.12	13.62	59.32	30.81
	FaceMorpher	22.57	29.48	27.90	31.81	23.69	35.16	44.30	28.24	40.40	29.41	**5.17**	6.91	36.53	18.36	46.94	25.14
	StyleGAN2	29.57	29.02	35.46	39.41	39.85	44.25	45.30	29.70	42.47	47.20	9.03	**7.12**	35.29	15.09	60.05	23.25
FRGC-M	OpenCV	23.81	25.04	31.62	21.11	20.67	57.06	48.60	29.74	53.55	26.45	34.32	**13.65**	36.17	59.66	19.63	27.17
	FaceMorpher	22.83	23.54	29.38	19.98	18.10	56.00	50.70	30.49	51.61	23.40	34.96	19.71	35.10	56.91	**16.06**	23.23
	StyleGAN2	32.71	28.68	21.70	21.95	11.62	37.38	38.42	16.43	26.62	14.32	41.14	25.85	36.19	47.03	15.26	**11.41**
Average performance		22.43	24.01	22.47	26.03	17.30	36.21	35.69	25.64	45.22	21.25	27.73	22.40	26.89	50.15	23.44	**16.88**

*D: LMA-DRD (D), PS: LMA-DRD (PS), LMA: MorGAN (LMA), GAN: MorGAN (GAN)

Table IV
RESULTS OF THE ABLATION STUDY.

Dataset	**Morphs**	**Data source**				
		I	**S1**	**S2**	**S1 + S2**	**I + S1 + S2**
FRLL-M	OpenCV	6.55	32.02	30.88	24.65	**3.55**
	FaceMorpher	4.21	26.63	21.91	20.18	**4.04**
	StyleGAN2	12.11	26.60	20.70	17.51	**10.96**
	WebMorph	14.82	32.92	41.85	38.74	**14.49**
	AMSL	12.02	34.53	40.97	34.89	**11.67**
FERET-M	OpenCV	31.38	41.78	40.07	37.61	**30.81**
	FaceMorpher	25.52	35.73	32.70	31.56	**25.14**
	StyleGAN2	34.59	32.70	23.44	**23.15**	23.25
FRGC-M	OpenCV	28.42	28.53	28.01	**25.00**	27.17
	FaceMorpher	24.69	24.38	25.00	22.71	**23.23**
	StyleGAN2	12.34	24.69	15.35	13.09	**11.41**
Average performance		18.79	30.96	29.17	26.28	**16.88**

[12] M. Ferrara, A. Franco, and D. Maltoni. *On the Effects of Image Alterations on Face Recognition Accuracy*. 2016.

[13] K. Grm, V. Štruc, A. Artiges, M. Caron, and H. K. Ekenel. Strengths and Weaknesses of Deep Learning Models for Face Recognition against Image Degradations. *IET Biometrics*, 2018.

[14] J. Ho, A. Jain, and P. Abbeel. Denoising Diffusion Probabilistic Models. In *NIPS*, 2020.

[15] M. Huber, F. Boutros, A. T. Luu, K. Raja, R. Ramachandra, N. Damer, P. C. Neto, T. Gonçalves, A. F. Sequeira, J. S. Cardoso, J. Tremoço, M. Lourenço, S. Serra, E. Cermeño, M. Ivanovska, B. Batagelj, A. Kronovšek, P. Peer, and V. Štruc. SYN-MAD 2022: Competition on Face Morphing Attack Detection Based on Privacy-aware Synthetic Training Data. In *IJCB*, 2022.

[16] M. Ibsen, L. J. Gonzalez-Soler, C. Rathgeb, P. Drozdowski, M. Gomez-Barrero, and C. Busch. Differential Anomaly Detection for Facial Images. In *IEEE WIFS*, 2021.

[17] T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the Design Space of Diffusion-Based Generative Models. In *NIPS*, 2022.

[18] I. Loshchilov and F. Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019.

[19] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. *CoRR*, abs/2206.00927, 2022.

[20] A. Makrushin, C. Kraetzer, J. Dittmann, C. Seibold, A. Hilsmann, and P. Eisert. Dempster-Shafer Theory for Fusing Face Morphing Detectors. In *EUSIPCO*, 2019.

[21] N. G. Nair and V. M. Patel. T2V-DDPM: Thermal to Visible Face Translation using Denoising Diffusion Probabilistic Models, 2022.

[22] T. Neubert, A. Makrushin, M. Hildebrandt, C. Kraetzer, and J. Dittmann. Extended StirTrace benchmarking of biometric and forensic qualities of morphed face images. *IET Biometrics*, 2018.

[23] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021.

[24] T. Ojala, M. Pietikäinen, and D. Harwood. A Comparative Study of Texture Measures With Classification Based on Featured Distributions. *Pattern Recognition*, 1996.

[25] V. Ojansivu and J. Heikkilä. Blur Insensitive Texture Classification Using Local Phase Quantization. In *Img. and Sig. Processing*, 2008.

[26] 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 *CVPR*, volume 1, pages 947–954 vol. 1, 2005.

[27] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss. The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295–306, 1998.

[28] W. H. L. Pinaya, M. S. Graham, R. Gray, P. F. da Costa, P.-D. Tudosiu, P. Wright, Y. H. Mah, A. D. MacKinnon, J. T. Teo, R. Jager, D. Werring, G. Rees, P. Nachev, S. Ourselin, and M. J. Cardoso. Fast Unsupervised Brain Anomaly Detection and Segmentation with Diffusion Models. In *MICCAI*, 2022.

[29] R. Raghavendra, K. B. Raja, S. Venkatesh, and C. Busch. Transferable Deep-CNN Features for Detecting Digital and Print-Scanned Morphed Face Images. In *IEEE CVPRW*, 2017.

[30] R. Ramachandra, S. Venkatesh, K. Raja, and C. Busch. Detecting Face Morphing Attacks with Collaborative Representation of Steerable Features. In *CVIP*, 2020.

[31] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *CVPR*, 2022.

[32] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi. Palette: Image-to-Image Diffusion Models. In *ACM SIGGRAPH*, 2022.

[33] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma. PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications. In *ICLR*, 2017.

[34] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel. Vulnerability Analysis of Face Morphing Attacks from Landmarks and Generative Adversarial Networks. 2020.

[35] U. Scherhag, L. Debiasi, C. Rathgeb, C. Busch, and A. Uhl. Detection of Face Morphing Attacks Based on PRNU Analysis. *IEEE TBBIS*, 2019.

[36] U. Scherhag, R. Raghavendra, K. B. Raja, M. Gomez-Barrero, C. Rathgeb, and C. Busch. On the vulnerability of face recognition systems towards morphed face attacks. In *IWBF*, 2017.

[37] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch. Face Recognition Systems Under Morphing Attacks: A Survey. *IEEE Access*, 2019.

[38] C. Seibold, A. Hilsmann, and P. Eisert. Reflection Analysis for Face Morphing Attack Detection. In *EUSIPCO*, 2018.

[39] J. Song, C. Meng, and S. Ermon. Denoising Diffusion Implicit Models. In *ICLR*, 2021.

[40] L. Wandzik, G. Kaeding, and R. V. Garcia. Morphing Detection Using a General-Purpose Face Recognition System. In *EUSIPCO*, 2018.

[41] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin. Diffusion Models for Medical Anomaly Detection. In *MICCAI*, 2022.

[42] J. Wyatt, A. Leach, S. M. Schmon, and C. G. Willcocks. AnoDDPM: Anomaly Detection with Denoising Diffusion Probabilistic Models using Simplex Noise. In *CVPRW*, 2022.

[43] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning Face Representation from Scratch. *CoRR*, abs/1411.7923, 2014.

[44] S. Zagoruyko and N. Komodakis. Wide Residual Networks. *CoRR*, abs/1605.07146, 2016.

[45] V. Zavrtanik, M. Kristan, and D. Skočaj. Reconstruction by Inpainting for Visual Anomaly Detection. *Patt. Rec.*, 2021.