

Towards Generalizable Morph Attack Detection with Consistency Regularization

Anonymous IJCB 2023 submission

Abstract

Though recent studies have made significant progress in morph attack detection by virtue of deep neural networks, they often fail to generalize well to unseen morph attacks. With numerous morph attacks emerging frequently, generalizable morph attack detection has gained significant attention. This paper focuses on enhancing the generalization capability of morph attack detection from the perspective of consistency regularization. Consistency regularization operates under the premise that generalizable morph attack detection should output consistent predictions irrespective of the possible variations that may occur in the input space. In this work, to reach this objective, two simple yet effective morph-wise augmentations are proposed to explore a wide space of realistic morph transformations in our consistency regularization. Then, the model is regularized to learn consistently at the logit as well as embedding levels across a wide range of morph-wise augmented images. The proposed consistency regularization aligns the abstraction in the hidden layers of our model across the morph attack images which are generated from diverse domains in the wild. Experimental results corroborate the idea and demonstrate the superior generalization and robustness performance of our proposed method compared to the state-of-the-art studies.

1. Introduction

In face recognition systems, the face is tightly tied to an individual's identity. By disrupting this particular link, morph attacks may pose a potential hazard [39, 36, 33, 5]. As a result, morph attack detection plays an important role in face recognition systems [6]. Morph attacks take place when a single morphed image can be used to prove the existence of two or more different people. Morphed images are crafted by interpolating facial landmarks or the latent representations between two or more individuals.

Despite the success of previous morph attack detection studies in recent years [25, 6, 32], their performance diminishes significantly on unseen morph attacks in real-world scenarios. A morph detection model should be general-



Figure 1. Illustration of the proposed SM augmentation for an identity sample in the Twin dataset [27]. The face images in the green and red bounding boxes correspond to the input bona fide and generated self-morphed images, respectively.

izable to out-of-sample distributions, independent of the training distribution. Previous studies in morph detection [34, 37] mainly overlook the domain shift issue in the real world and benchmark their methods on intra-domain scenarios. That is, the same methods are generally adopted in both training and test phases to create morph images. Moreover, generalization performance to post-processing operations such as print-scan conversion or JPEG compression are not taken into account in these works. As such, they obtain satisfactory performance on limited test sets, but struggle to retain its performance beyond their training regime. Using unlabeled target data, the domain adaptation (DA) methodology can be employed to mitigate the domain shift challenge [22]. However, in practical settings, we are not provided with the unlabeled target domain. To address this challenge in a more general setting, domain generalization methodology has attracted much attention recently [47]. It aims to learn the domain invariant features without accessing target domain data.

Recently, a few studies have been conducted to enhance the generalization of morph attack detection. For example, Damer et al. [7] embrace the observation that fusing several detectors which are trained on different types of morph attacks generalize better to new morph attacks. Also, in [6], the pixel-wise supervision is incorporated into the binary classification to empower the morph detector with the generalization capability. However, these works [7, 6] typically rely on the diverse morph images generated from different morph attacks, which can be challenging to provide. Moreover, their generalization ability is restricted to certain situations far from real-world scenarios.

In this paper, we approach the generic morph attack detection differently by learning consistent feature representations. Our approach relies on the consistency regularization concept [2, 35, 11], that a generic model should predict consistent results regardless of the plausible variations that the input images may undergo. Various factors may induce these variations, including brightness, lighting, image style, camera sensors, and the type of morphing attacks. Thus, a model with higher consistency under plausible common image corruptions such as noise, blur, contrast and compression corruptions is expected to generalize better to new unexplored domains [11, 50, 1]. Bearing these insights in mind, the consistent predictions are imposed on our model across a wide range of morph-wise augmented images. Building upon this concept, we can minimize the soft output class distributions between different morph attack variations. However, this minimization would not necessarily align the abstraction in the hidden layers of our model across morph attack images generated from different domains. To overcome this issue, we regularize our model to ensure consistent learning at both the logit and feature representation levels. For this objective, several regularization branches are first integrated into the intermediate layers of our model and the embedding levels are computed at these branches. Then, we jointly regularize feature representation as well as final soft output class distribution. To learn the domain-shared feature representation, adversarial feature learning [23] is adopted among the morph-wise augmented images at different feature representation levels. In this respect, a feature extractor competes with a domain discriminator to learn a domain-shared feature representation and the domain discriminator determines whether the input images come from the intact morph images or the augmented ones. An overview of our proposed architecture is illustrated in Figure 3.

To encourage our model to learn generic representation for morph attack detection, we incorporate cross-domain morph attacks into our consistency regularization. Note that blindly exposing a model to random image transformations does not necessarily enhance the cross-domain performance. Rather, it could hurt the inter-domain performance

of our morph attack detection. With this consideration in mind, we propose two morph-wise augmentations, namely Inter-domain Style Mixup (ISM), and Self-morphing (SM) augmentations, to explore a wide space of realistic morph transformations in our consistency regularization. The ISM augmentation employs the photo-realistic style transfer [42] to synthesize unseen morph attacks with new styles, while keeping the content of the input morph images unchanged. Also, the SM augmentation synthesizes morph attacks with minimal visual artifacts using several instances of the same identity. The motivation of the proposed SM augmentation stems from the fact that in realistic morph attack scenarios, the visible morphing artifact are further post-processed and eliminated carefully. This process results in hardly recognizable but valuable morph images, which still contain imperceptible morphing artifacts. Our major contributions in this paper are

- We regularize morph attack detection model to predict consistent results regardless of potential variations caused by diverse morph attacks, image quality, and environmental situations.
- We propose two morph-wise augmentations to explore a wide space of realistic morph attack transformations in our consistency regularization.
- We carry out extensive evaluations on several datasets to validate the generalization capability of our morph attack detection.

The remainder of this paper is structured as follows. Section 2 provides a literature review of recent research in morph detection, domain generalization, and state-of-the-art data augmentation methods. Section 3 describes our methodology in detail, including our proposed morph-wise data augmentation and the proposed consistency regularization. We provide comprehensive evaluations in section 4 to assess the impact of synthetic morphs and proposed consistency regularization on the generalization performance. In this section, we also evaluate the generalization and robustness performance of our proposed morph detector compared with the state-of-the-art studies. Finally, Section 5 concludes this paper.

2. Related Work

2.1. Morph Detection

Morph attack detection studies [7, 8, 33, 32, 25, 6, 38] can be categorized into single and differential morph detection. Single morph detection attempts to distinguish the morphed image from the bona fide one. Differential morph detection, on the other hand, compares the potential morphed image with a second reliable image of the probe such

as a live capture to make its prediction. Recently, deep learning models have been widely used for morph detection. With the advent of deep learning, several morph detection studies have been carried out in recent years. Soleymani et al. [37] train a disentangling network that produces disentangled representations for landmarks and facial appearance. They generate triplets of images, whereby each intermediate image takes the landmarks from one image and the appearance from the other image. To improve unknown re-digitized morph attacks detection, authors in [6] adjust pixel-wise supervision in the training to capture more informative morphing artifacts. Besides, they make a morphing dataset accessible to the public, which comprises digital and re-digitized morph attacks as well as bona fide images. Using a convolutional neural network, a de-morphing-based method is suggested in [28] to unravel the chip image and identify morphing presentation assaults in actual automated border control systems. Damer *et al.* [4] create bona fide face images of non-existing people and develop a synthetic-based morph attack detection testset with StyleGAN2-ADA [20], whereby the legal and ethical difficulties associated with biometric data use can be mitigated. Equipped with SMDD dataset in [4], Ivanovska et al. [17] train Xception and HRNet networks to demonstrate the potential of synthetic morph data and justifies its importance for morph detection models across three limited morph datasets.

2.2. Domain Generalization

In supervised learning, the training data is assumed to be from the same distribution as the test data. However, in most real-world scenarios with out-of-distribution (OOD) data, this assumption could be violated, and consequently, these algorithms suffer significant performance drops on an OOD data [23, 50]. Domain generalization is intended to learn domain-invariant representations that are generalizable to an unseen domain based on labeled source domains [47]. A number of studies in this area have been made with respect to data augmentation, domain alignment, ensemble learning, and self-Supervised learning. Regarding data augmentation, the studies [43, 16, 40, 41] simulate domain shift through transferring the styles of source domain with external styles to learn domain-invariant representations. In the domain alignment category, some researchers [29, 3, 48] work on normalization operations to eliminate information that aggravate domain shift issue. Although domain generalization has achieved impressive results in image classification, object detection and semantic segmentation, little attention has been paid to the generalization capabilities of morph attack detection. In addition, the existing studies [7, 6] do not learn the invariant representations to the application-specific textural distortion. Therefore, we encourage our model to learn domain-invariant morphing attack feature representation which is found beneficial to mit-

igate domain shift challenge.

2.3. Data Augmentation

Consistency-based methods rely on generating diverse yet reasonable augmentation of the input data. Data augmentation is one of the most effective solutions to improve model performance generalization without incurring computational cost in the inference time. Expanding the diversity of the training data with data augmentation can be regarded as a useful regularizer to mitigate overfitting issue [18, 10, 44]. It can also enhance the robustness of deep neural networks against input distribution shifts. Conventional data augmentations include simple label-invariant image transformations such as flipping, translation, jittering, and random cropping. As an example, CutOut randomly removes a square region in the input samples [10]. Recently, different studies on data augmentation have proposed to synthesize mixed samples and employ a sequence of image transformations. For instance, Mixup is the seminal study that linearly interpolates between two or more input samples to synthesize new samples [44]. Another group of studies such as style randomization [18] utilizes neural style transfer [19] to modify the distribution of low-level features such as color and texture information in the training samples [49, 18]. Based on the observations outlined in [14], they leverage Stylized ImageNet to mitigate the texture bias in deep neural networks and improve the generalization performance against distribution shifts.

3. Methodology

3.1. Problem Definition

In the context of domain generalization, it is assumed that the data from a source domain D_S is accessible to train our model. The ultimate objective is to train a model capable of performing as well as possible on data from unseen domains D_T , without requiring additional model updates based on target domain D_T . With no prior knowledge on D_T , we regularize our model to learn semantic consistency between several different landmark-based and GAN-based morphing attacks. These attacks include Print and Scan [45], StyleGAN2 [36], WebMorph [36], OpenCV [36], and FaceMorpher [36] attacks. By doing so, we enforce our model to learn consistent representations in respect to different morphing attack artifacts rather than the domain-specific features relevant to the identity information. Moreover, considering diversity and realism, we propose two morph-wise augmentations to synthesize novel morph domains and enrich the training source domain in the consistency regularization. In what follows, the proposed morph-wise augmentations are first explained. Then, the consistent regularisation learning is addressed in more detail.



Figure 2. Illustration of the proposed ISM augmentation for an identity sample in the Twin dataset [27]. The face images in the red bounding box correspond to the input face image and the others indicate the augmented ones with the same class label.

3.2. Morph-wise Augmentation

Self-morphing Augmentation. The key idea in SM Augmentation is that hardly detectable morph attacks in reality with imperceptible morphing artifacts could enforce our model to learn more generalizable representations. As such, the SM Augmentation lies in the guidance of such high-quality morph attacks with minimal visible morphing artifacts, which are synthesized by two look-alike subjects. To promote the effectiveness of morph attack images, different instances of the same identity are employed in our morphing attacks. Formally, let x_i represents an instance i , which belongs to the identity x . First, x_i is randomly augmented with different image transformations, including Color, Gaussian noise, Blurring, Contrast, Brightness, Shear, Translate, and Compression operations. Then, the output self-morphed image would be calculated as follows:

$$x_{ij}^{SM} = \Psi(\hat{x}_i, x_j), \quad (1)$$

where $\hat{x}_i, x_j$ represent the augmented and pristine instances of identity x , and x_{ij}^{SM} is the output self-morph image. Also, Ψ denote the adopted morphing attacks, which include the StyleGAN, OpenCV, and FaceMorpher approaches. The SM Augmentation enriches the diversity of morph attacks so that our model can explore a large space of morphing artifacts and would not overfit to visible morphing artifacts. Figure 1 represents an example of SM augmentation.

Inter-domain Style Mixup Augmentation. ISM Augmentation aims to cover unconstrained scenarios that may occur in the real-world morph attacks in our consistency regularization. Equipped with the photorealistic style transfer method WCT2 [42], ISM Augmentation manipulates the low-level style information in the source domain without compromising the high-level semantic information. This operation is formulated as follows:

$$x_{ij}^{ISM} = \Omega(x_i^{co}, x_j^{st}), \quad (2)$$

where Ω is the ISM transformation, the x_i^{co} , x_j^{st} , and x_{ij}^{ISM} indicates the content image, the style image, and the output augmented image with a shared ground-truth label. Recall that the generated morph (or bona fide) images contain the content of the source morph (or bona fide) images and the style of the target morph (or bona fide) images. As such, the output generated synthetic images share identity information with the source images and low-level features with the target images. In a quest to find a dataset from a range of possibilities that can be used as the target domain, we opt for the FFHQ dataset [21]. An example of ISM augmentation is shown in Figure 2.

3.3. Consistency Regularization

The baseline model is built from a feature extraction module with several levels (expressed as $K = \kappa_1 \circ \kappa_2 \circ \dots \circ \kappa_K$) and a linear classifier c_f . Each level of the model is followed by a feature alignment module γ_i and an auxiliary shallow classifier c_i . The feature alignment module γ_i balances the feature dimension among different depths of the model so that the semantic abstractions would be regularizing throughout the network. Using the auxiliary classifier α_i and baseline model, the outputs at different levels can be obtained as below:

$$\alpha_1(x) = c_1 \circ \gamma_1 \circ \kappa_1(x), \quad (3)$$

$$\alpha_i(x) = c_i \circ \gamma_i \circ \kappa_i(x) \dots \circ \kappa_1(x), \quad (4)$$

$$\text{Baseline}(x) = c_f \circ \kappa_f(x) \circ \kappa_{f-1}(x) \circ \dots \circ \kappa_1(x). \quad (5)$$

While the logit outputs contain limited probabilistic information over classes, the feature representations at deep and shallow levels of the network encode richer information, which respectively capture the category-level semantic information and boundary content in the input images. Thus, an intuitive solution to setting up a powerful classifier is to incorporate a hierarchy of feature representations at several levels as $F_{cat} = \text{Cat}([F_1, \dots, F_N])$. To do so, the feature representation F_i is extracted at different levels by means of the auxiliary classifiers c_i as $F_i = \gamma_i \circ \kappa_i(x) \dots \circ \kappa_1(x)$, where $i \in 1, \dots, N-1$, and $N-1$ is the number of levels in the baseline model. Afterwards, a linear classifier c_{cat} operates on top of the concatenated features as given by:

$$\alpha_{cat}(x) = c_{cat} \circ \text{PWConv}(\text{Cat}([F_1, \dots, F_N])), \quad (6)$$

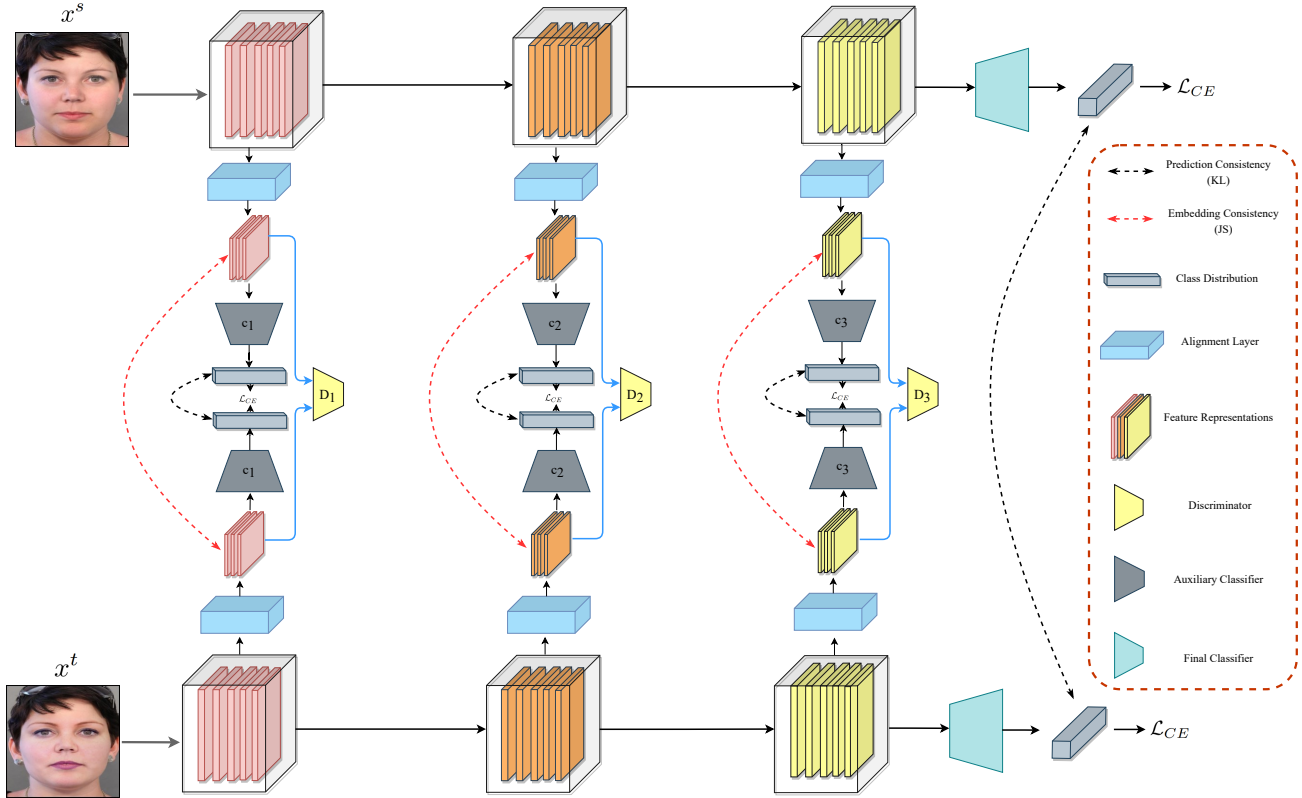


Figure 3. Illustration of the proposed architecture.

where PWConv denotes the point-wise convolution layer. Then, on a mixed set of raw and the morph-wise augmented images, the auxiliary classifiers ($c_{i \in 1, \dots, N-1, cat}$) and baseline backbone are trained using the standard cross-entropy (CE) loss function given as:

$$L_{cls} = \sum_{i=1}^{N+1} \mathcal{L}_{CE}(\sigma(\alpha_i(x); \tau), y), \quad (7)$$

where x and y are the input sample and its groundtruth and L_{CE} indicates the CE loss function. Also, $\alpha_N(x)$ and $\alpha_{N+1}(x)$ denote the final prediction of the baseline model and c_{cat} is the classifier, respectively. Here, $\sigma(\alpha(x); \tau)$ denotes the Softmax operation with temperature τ . An increase in $\tau > 1$ leads to a softer probability distribution. The operation would be a normal Softmax if $\tau = 1$.

Prediction-level Consistency Regularization To encourage the model to yield the same output distribution, we begin our consistency regularization with matching the class posterior distributions between predictions of the auxiliary classifiers $\alpha_i(x)$ for the source D_S and augmented target domain D_T . Ideally, for even an unlabeled example, a robust model should produce consistent predictions no matter how it has been deformed and distorted. To achieve this,

the Kullback–Leibler (KL) divergence minimization is applied as below:

$$L_{label} = \sum_{i=1}^{N+1} D_{KL}(\sigma(\alpha_i(x^s); \tau), \sigma(\alpha_i(x^t); \tau)), \quad (8)$$

where $x_s \in D_S$, $x_t \in D_T$, and temperature scaling τ produces the soft output probability. Note that in the training process of morph class, D_T includes the morph samples that are either augmented by the ISM and SM augmentations or are generated by a morph attack different from x_s . This regularizes a consistent posterior distribution whereby the classes with near-zero probabilities would not be simply discarded.

Embedding-level Consistency Regularization We argue that a highly generalized model should behave consistently in feature representation space regardless of the styles and domains of the input images. Such representations encode the beneficial contents relevant to image intensity and spatial correlation. To fully meet this requirement, the feature representation $F_{i \in 1, \dots, N-1}$ at different levels of the backbone model is extracted separately. Afterwards, the feature representation F_i are matched between the source D_S and generated ISM target domains D_T by the

Table 1. Cross-morph evaluations of the proposed method with the state-of-the-art studies on FRGC datasets. The results are in terms of APCER1 (@BPCER=1%), APCER5 (@BPCER=5%), APCER (@BPCER10=10%), EER, and AUC metrics. GAN refers to the StyleGAN2 [36].

	Method	APCER1%	APCER5%	APCER10%	EER	AUC
MIPGAN	ConvNext [24]	17.40	3.07	1.20	16.33	99.17
	Inception [17]	61.98	36.68	23.82	17.26	91.12
	Residual [33]	-	-	-	6.67	-
	GRL	00.00	00.00	00.00	4.28	99.99
StyleGAN	ConvNext [24]	44.60	14.52	2.80	7.65	97.57
	Inception [17]	50.60	32.39	25.56	17.26	94.89
	GRL	00.00	00.00	00.00	00.00	100.00
OpenCV	ConvNext [24]	60.68	29.66	12.65	11.50	95.27
	Inception [17]	00.00	00.00	00.00	00.00	100.00
	GRL	00.00	00.00	00.00	00.00	100.00

Jensen-Shannon Divergence (JSD). We also integrate the discriminator $Dsc_{i \in 1, \dots, N}$ (fed by the feature representation F_i) into our regularization framework to classify samples in the source D_S domain from the generated target one D_T . These operations can be summarized as follows:

$$L_{emb} = \sum_{i=1}^N D_{JS}(F_i(x^s), F_i(x^t)) + \eta \log(Dsc_i(F_i(x^s))) + \eta \log(1 - Dsc_i(F_i(x^t))), \quad (9)$$

where $Dsc_{i \in 1, \dots, N}$ is trained with the CE loss function, and D_{JS} refers to the JSD loss function. Compared with simple KL-divergence and Mean Square Error (MSE), JSD regularizes higher degree of consistency for feature representations and encourages more flexible optimization. Also, η is the weight parameter that adjusts the importance of JSD regularization compared to CE optimization. Minimization in the Equation 9 realizes an adversarial process wherein the JSD regularization attempt to fool the discriminator $D_{i \in 1, \dots, N}$ by learning indistinguishable feature representations.

Overall Loss Finally, the overall objective function in our training optimization can be summarized as follows:

$$L_{total} = L_{cls} + \mu L_{label} + \delta L_{emb}, \quad (10)$$

where δ , and μ indicate the weighting parameters that balance the impact of different loss functions. In the inference step, the discriminators $D_{i \in 1, \dots, N}$ and the auxiliary classifiers are detached and removed from the backbone model, thereby incurring no extra computational overhead compared to the baseline model.

4. Experiments

This section provides the explanation about the test and train datasets, the implementation details and evaluation protocols. Also, a comparative evaluation is performed to demonstrate that the proposed method with generalizable representation learning (called GRL) significantly outperforms its competitors in respect to the generalization and accuracy performance.

4.1. Evaluation Settings

To fully assess the generalization capability of our morph attack detection, the experimental evaluations are carried out in two settings. In the first setting, we study the generalization performance of our method from one morph attack to the unseen attacks. More specifically, the bona fide images remain unchanged, yet the domain discrepancy exists in the morph attack and morph artifacts. In this setting, the FRGC dataset [31] is adopted as the training data and the morph attacks are generated via the FaceMorpher method [36]. Also, the target domain belongs to the FRGC morph faces which are created by the other morph attacks such as StyleGAN2, MIPGAN, and OpenCV approaches. When employed in an identification document, images may undergo various post-processing operations such as JPEG compression, resizing, and print-scan transformations, leading to the new types of artifacts in addition to the morph ones. In this regard, to benchmark the robustness of our method against these artifacts, we craft the printed-and-scanned MIPGAN (MIPGAN-PS) and JPEG compressed (MIPGAN-JPG) test sets using the FRGC images.

In the second setting, to perform morph detection on a cross-domain dataset, we employ the Twins morph dataset [27] as the training set. The Twins dataset is composed of 9,052 bona fide and 12,991 morphed images. To generate high-quality morphs in this dataset, identical twin pairs are selected as the contributing subjects in the landmark-based and Generative Adversarial Network (GAN)-based morphing methodologies. The FaceMorpher library [36] and the pre-trained StyleGAN2 model [20] are utilized as the former and latter methodologies, respectively. To corroborate the effectiveness of morph detection method over morph images with different distributions compared to the training set, FRGC [31], AMSL [26], FERET [30], VISAPP17 [26], and FRLL [9, 36] datasets are targeted in our experimental evaluations. To fully study the generalization capability of our morph detection model, we also benchmark our model against a wide range of landmark-based and GAN-based morphing attacks. These attacks consists of Print and Scan [45], StyleGAN2 [36], WebMorph [36], OpenCV [36], and FaceMorpher [36] attacks. At last, we also conduct ablation experiments to validate the importance of SM and ISM augmentations in our proposed consistency regularization.

Table 2. Cross-domain comparison of the proposed GRL with the state-of-the-art studies. The evaluations are in terms of the EER metric.

	Methods	FRLL AMSL	FRLL Webmorpher	FRLL OpenCV	FRLL StyleGAN	FRLL FaceMorpher	FERET OpenCV	FERET StyleGAN	FERET FaceMorpher	FRGC OpenCV	FRGC StyleGAN	FRGC FaceMorpher	FRGC MIPGAN	VISAPP17	LMA-DRD
EER	SPL-MAD [12]	12.09	15.72	5.78	12.92	4.67	30.21	28.95	25.76	19.54	15.57	18.42	-	-	29.54
	MixFacenet [17]	15.18	12.35	4.39	8.99	3.87	-	-	-	-	-	-	-	-	23.72
	Inception [17]	10.79	9.86	5.38	11.37	3.17	-	-	-	-	-	-	-	-	19.01
	PW-MAD [17]	15.18	16.65	2.42	16.64	2.20	-	-	-	-	-	-	-	-	20.39
	Hamza [15]	-	-	-	-	-	13.5	-	11.5	-	-	-	-	-	-
	Quality [13]	7.91	7.13	5.41	7.04	3.60	12.29	13.99	10.80	24.48	14.32	24.17	-	-	25.09
	OrthoMAD [25]	14.80	15.23	0.73	6.54	0.98	-	-	-	-	-	-	-	-	-
	Residuals (LMA) [32]	-	-	-	-	-	-	-	-	-	0.17	-	13.92	-	-
	Mutual [37]	3.11	-	-	-	-	-	-	-	-	-	-	-	4.69	-
	Scale-Space Gradients [33]	-	-	-	-	-	-	-	0.98	-	-	-	6.67	-	-
	GRL	1.53	5.23	1.05	5.54	1.79	6.80	8.39	7.75	0.06	0.24	0.12	0.40	0.00	13.09

Table 3. Robustness evaluation of the proposed method against post-processing artifacts compared with the state-of-the-art studies on FRGC MIPGAN dataset. The results are in terms of APCER1 (@BPCER=1%), APCER5 (@BPCER=5%), APCER10 (@BPCER=10%), EER, and AUC metrics. PS and JPG refer to the MIPGAN-PS and MIPGAN-JPG test sets.

	Method	APCER1%	APCER5%	APCER10%	EER	AUC
PS	ConvNext [24]	95.71	75.90	60.50	31.19	76.11
	Inception [17]	86.88	71.48	54.48	27.40	77.73
	Residual [33]	-	-	-	9.63	-
	GRL	67.60	19.41	9.63	10.84	95.28
JPG-112	ConvNext-112 [24]	95.71	75.90	60.50	31.19	94.73
	Inception-112 [17]	76.97	49.13	37.35	21.48	86.65
	GRL-112	39.75	6.024	1.74	5.67	98.08
JPG-64	ConvNext-64 [24]	61.98	36.68	23.82	17.77	91.36
	Inception-64 [17]	91.83	75.23	57.69	27.65	78.21
	GRL-64	37.75	16.46	9.10	9.50	96.63
	GRL-32	65.59	42.83	27.84	16.33	90.75

Evaluation Metrics. To gauge the performance of our morph attack detection, the Attack Presentation Classification Error Rate (APCER) is adopted. This metric computes the ratio of morph attacks which are incorrectly classified as bona fide. Also, to gain a comprehensive performance of the morph attack detection, the Area-Under-the-Curve (AUC) and Detection Equal Error Rate (D-EER) are computed. It is worth noting that the D-EER reports the classification error where APCER is equal to BPCER.

Implementation Detail. To pre-process the training data, the MTCNN model [46] is utilized to detect and align face images. Then, the captured faces are re-scaled to 512×512 resolution. The ConvNext network [24] is also selected as the backbone model. In order to train our backbone model,

the Stochastic Gradient Descent (SGD) with momentum 0.9 is employed. The initial learning rate, batch size, and the total number of epochs are set to $1e-4$, 64, and 50, respectively. The hyperparameters used in Equations 7, 9, and 10 are set to $\tau = 0.1$, $\eta = 0.1$, $\mu = 0.05$, and $\delta = 0.1$.

4.2. Results on Unseen Morph Attacks

In this evaluation setting, we measure how well the learned representations in the domain-specific morph artifacts may transfer to other types of morph artifacts. From Table 1, we can observe that the proposed method achieves much higher generalization performance under all evaluation metrics over the baseline model (ConvNext) and other morph detection models when tested on the MIPGAN, StyleGAN2, and OpenCV morph attacks. For instance, compared to the baseline model, our approach improves the APCER1% on MIPGAN, StyleGAN2 and OpenCV morph attacks from 17.4%, 44.60%, and 60.68% to 0%. The reason behind such a pronounced generalization is that the proposed consistency regularising enforces the model to learn domain-agnostic feature representations. These results demonstrate the effectiveness of the proposed GRL to greatly benefit the out-of-distribution generalization of the morph attack detection.

4.3. Results on Unseen Post-processing Artifacts

The objective of this experiment is to benchmark how robust the GRL is against new types of artifacts induced by post-processing operations. The vulnerability assessments of our method against Print/Scan and JPEG compression operations are presented in Table 3. The FRGC dataset is the target test set wherein the morph attacks are generated by the MIPGAN approach. It is evident from the results in this analysis that the proposed consistency regularization equipped with the morph-wise augmentations can po-

tentially gain considerable robustness against unseen post-processing artifacts in morph attack generation compared with vanilla morph attack detection studies. In Print/Scan and JPEG compressed morph images, our morph detection model significantly outperforms the other studies. It is worth highlighting that the performance of the proposed GRL against compressed images with resolution 32×32 surpasses the baseline model against compressed images with resolution 128×128 .

4.4. Results on Unseen Datasets

As discussed previously, in the second setting, we explore the generalization capability of the proposed GRL to a wide range of unseen bona fide images, the domain of which are far away from our training set. In addition, since the distribution of morph images is also of great importance, unseen morph attacks are also integrated into the test data as well. As reported in Table 2, our GRL remarkably outperforms the state-of-the-art methods. In our evaluation, the best results of other studies are reported for comparison evaluations. For instance, in OrthoMAD, the best performance with the SMDD training set are reported. This is also the case for other studies. An in-depth analysis in Table 2 reveals that the superiority of GRL over the state-of-the-art studies such as Quality [13] and SPL-MAD [12] is more noticeable on FRGC dataset than the others. The results validate that the state-of-the-art studies significantly lag behind the proposed GRL in both out-of-distribution and in-distribution morph attack detection. Note that on some test sets such as FRLL, the proposed GRL achieves better results on in-distribution morph attacks compared to the out-of-distribution morph attacks; however, this gap cannot be observed in other test sets such as FRGC. In short, taking these results into account, we can substantiate the generalization capability of GRL in out-of-distribution morph attack detection while retaining its high in-distribution performance for morph attack detection.

4.5. Ablation Study

To determine the contribution of SM, and ISM augmentations and also the embedding- and prediction-level consistency regularizations in the proposed morph detection, we eliminate them from the proposed GRL and follow the second setting of our training as mentioned in subsection 4.1. Then, the trained degraded versions of GRL are assessed individually on the MIPGAN test set. The results in Table 4 verify that each one of the proposed components, namely SM, and ISM augmentations, L_{label} and L_{emb} consistency regularizations, plays an important role in the generalization and robustness performance of the proposed morph detection. Interestingly, we can observe that the SM augmentation yields higher gains in AUC and EER metrics compared to the ISM augmentation. This is due to the fact that

Table 4. Ablation evaluations of the proposed method on MIPGAN test set. The results are in terms of EER, and AUC metrics.

Metric	Baseline	+ SM	+ ISM	+ L_{label}	+ L_{emb}	GRL
EER	11.24	4.21	9.47	6.74	1.79	0.4
AUC	96.03	99.18	97.45	97.91	99.50	99.95

the ISM augmentation can synthesize high-quality morphs images with minimal visual artifacts using a wide range of instances of the same identity, which are imperative for morph detection. In addition, it is observed that the proposed L_{emb} consistency regularization leads to more significant improvements in AUC and EER metrics in comparison to the L_{label} consistency regularization. In summary, these improvements prove the contribution of GRL in morph detection.

5. Conclusions

In this paper, we present a morph attack detection with strong generalization ability to different morph attacks. To make our detector generalize better to unseen face morph attacks, we propose the ISM and SM morph-wise augmentations to explore a wide space of realistic morph attack artifacts in our consistency regularization. The ISM augmentation synthesizes unseen morph attacks with new styles, whilst preserving the content of the input morph images. Moreover, the SM augmentation generates realistic morph attacks with imperceptible visual morph artifacts. To improve the generalization performance of our detector against unseen face morph attacks, we encourage our model to predict consistent output regardless of the input variations simulated for different domains. To this end, we regularize our model to learn consistently at the logit and feature representation levels. Experimental results on several datasets demonstrate the generalization ability of our proposed model while keeping high in-domain performance. Also, our proposed consistency regularization improves the robustness of our model against post-processing artifacts.

References

- [1] A. Abuduweili, X. Li, H. Shi, C.-Z. Xu, and D. Dou. Adaptive consistency regularization for semi-supervised transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6923–6932, 2021. 2
- [2] P. Bachman, O. Alsharif, and D. Precup. Learning with pseudo-ensembles. In *Advances in Neural Information Processing Systems*, volume 27, 2014. 2
- [3] S. Choi, S. Jung, H. Yun, J. T. Kim, S. Kim, and J. Choo. RobustNet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11580–11590, 2021. 3

- [4] N. Damer, C. A. F. López, M. Fang, N. Spiller, M. V. Pham, and F. Boutros. Privacy-friendly synthetic data for the development of face morphing attack detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022. 3
- [5] N. Damer, A. M. Saladie, A. Braun, and A. Kuijper. MorGAN: Recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network. In *International Conference on Biometrics Theory, Applications and Systems*, pages 1–10, 2018. 1
- [6] 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 *International Symposium on Visual Computing*, pages 291–304, 2021. 1, 2, 3
- [7] N. Damer, S. Zienert, Y. Wainakh, A. M. Saladie, F. Kirchbuchner, and A. Kuijper. A multi-detector solution towards an accurate and generalized detection of face morphing attacks. In *International Conference on Information Fusion*, pages 1–8, 2019. 2, 3
- [8] L. Debiase, N. Damer, A. M. Saladié, C. Rathgeb, U. Scherhag, C. Busch, F. Kirchbuchner, and A. Uhl. On the detection of GAN-based face morphs using established morph detectors. In *International Conference on Image Analysis and Processing*, pages 345–356, 2019. 2
- [9] L. DeBruine and B. Jones. Face research lab london set, May 2017. 6
- [10] T. DeVries and G. W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 3
- [11] Y. Fan, A. Kukleva, D. Dai, and B. Schiele. Revisiting consistency regularization for semi-supervised learning. *International Journal of Computer Vision*, pages 1–18, 2022. 2
- [12] M. Fang, F. Boutros, and N. Damer. Unsupervised face morphing attack detection via self-paced anomaly detection. In *IEEE International Joint Conference on Biometrics*, pages 1–8, 2022. 7, 8
- [13] B. Fu and N. Damer. Face morphing attacks and face image quality: The effect of morphing and the unsupervised attack detection by quality. *IET Biometrics*, 11(5):359–382, 2022. 7, 8
- [14] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. 3
- [15] M. Hamza, S. Tehsin, H. Karamti, and N. S. Alghamdi. Generation and detection of face morphing attacks. *IEEE Access*, 10:72557–72576, 2022. 7
- [16] J. Huang, D. Guan, A. Xiao, and S. Lu. FSDR: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6891–6902, 2021. 3
- [17] M. Ivanovska, A. Kronovšek, P. Peer, V. Štruc, and B. Batagelj. Face morphing attack detection using privacy-aware training data. *arXiv preprint arXiv:2207.00899*, 2022. 3, 6, 7
- [18] P. T. Jackson, A. A. Abarghouei, S. Bonner, T. P. Breckon, and B. Obara. Style augmentation: data augmentation via style randomization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, volume 6, pages 10–11, 2019. 3
- [19] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711, 2016. 3
- [20] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems*, pages 12104–12114, 2020. 3, 6
- [21] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2019. 4
- [22] W. M. Kouw and M. Loog. A review of domain adaptation without target labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):766–785, 2019. 1
- [23] H. Li, S. J. Pan, S. Wang, and A. C. Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018. 2, 3
- [24] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 6, 7
- [25] P. C. Neto, T. Gonçalves, M. Huber, N. Damer, A. F. Sequeira, and J. S. Cardoso. OrthoMAD: Morphing attack detection through orthogonal identity disentanglement. In *International Conference of the Biometrics Special Interest Group*, pages 1–5, 2022. 1, 2, 7
- [26] 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*, 7(4):325–332, 2018. 6
- [27] K. O’Haire, S. Soleymani, J. Dawson, and N. M. Nasrabadi. Ultimate morph database generation using identical twins. In *International Joint Conference on Biometrics*, 2022. 1, 4, 6
- [28] D. Ortega-Delcampo, C. Conde, D. Palacios-Alonso, and E. Cabello. Border control morphing attack detection with a convolutional neural network de-morphing approach. *IEEE Access*, 8:92301–92313, 2020. 3
- [29] X. Pan, P. Luo, J. Shi, and X. Tang. Two at once: Enhancing learning and generalization capacities via IBN-Net. In *Proceedings of the European Conference on Computer Vision*, pages 464–479, 2018. 3
- [30] P. Phillips, H. Wechsler, J. R. Huang, and P. J. Rauss. The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16:295–306, 1998. 6
- [31] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 1, pages 947–954, 2005. 6

- [32] K. Raja, G. Gupta, S. Venkatesh, R. Ramachandra, and C. Busch. Towards generalized morphing attack detection by learning residuals. *Image and Vision Computing*, 126:104535, 2022. 1, 2, 7
- [33] R. Ramachandra and G. Li. Residual colour scale-space gradients for reference-based face morphing attack detection. In *International Conference on Information Fusion*, pages 1–8, 2022. 1, 2, 6, 7
- [34] R. Ramachandra, K. Raja, and C. Busch. Algorithmic fairness in face morphing attack detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 410–418, 2022. 1
- [35] M. Sajjadi, M. Javanmardi, and T. Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, 2016. 2
- [36] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel. Vulnerability analysis of face morphing attacks from landmarks and generative adversarial networks. *arXiv preprint*, Oct. 2020. 1, 3, 6
- [37] S. Soleymani, A. Dabouei, F. Taherkhani, J. Dawson, and N. M. Nasrabadi. Mutual information maximization on disentangled representations for differential morph detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1731–1741, 2021. 1, 3, 7
- [38] S. Venkatesh, R. Ramachandra, K. Raja, L. Spreeuwiers, R. Veldhuis, and C. Busch. Morphed face detection based on deep color residual noise. In *International Conference on Image Processing Theory, Tools and Applications*, pages 1–6, 2019. 2
- [39] S. Venkatesh, R. Ramachandra, K. Raja, L. Spreeuwiers, R. Veldhuis, and C. Busch. Detecting morphed face attacks using residual noise from deep multi-scale context aggregation network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 280–289, 2020. 1
- [40] Y. Wang, L. Qi, Y. Shi, and Y. Gao. Feature-based style randomization for domain generalization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5495–5509, 2022. 3
- [41] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, and Q. Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021. 3
- [42] J. Yoo, Y. Uh, S. Chun, B. Kang, and J.-W. Ha. Photorealistic style transfer via wavelet transforms. In *International Conference on Computer Vision*, pages 9036–9045, 2019. 2, 4
- [43] X. Yue, Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, and B. Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2100–2110, 2019. 3
- [44] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 3
- [45] 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, 2021. 3, 6
- [46] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 7
- [47] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022. 1, 3
- [48] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2020. 3
- [49] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30:8008–8018, 2021. 3
- [50] R. Zhu and S. Li. Crossmatch: Cross-classifier consistency regularization for open-set single domain generalization. In *International Conference on Learning Representations*, 2022. 2, 3