A Comparative Study on Face-Spoofing Detection Algorithms

Kartik Agrawal Harishankar M Riya Ann Easow Kaustubh Gupta

Abstract

With the widespread adoption of face recognition technology into our daily routines, the need for a robust defense against face spoofing attacks becomes imperative. While a number of face spoof detection techniques have been proposed, many of them fail to generalize to unseen scenarios. In response to this, we propose a comparative study that contrasts classical methods, leveraging techniques such as Local Binary Patterns (LBP) and Image Quality Assessment (IQA), with cutting-edge deep learning methodologies like Single Side Domain Generalization (SSDG).

We analyze the efficacy of these methods through qualitative and quantitative analysis, shedding light on their strengths and limitations. Through our work, we highlight the difficulties of achieving cross-dataset generalization. We experiment with diverse pre-processing strategies to improve the state of the art[1].

1. Introduction

In recent years, the integration of face recognition technology into various aspects of daily life, including smartphone authentication and access control, has become commonplace. This popularity of face recognition systems has sparked rising interest among researchers in studying the vulnerabilities associated with these systems against different types of attacks [21]. Among the different threats analyzed, face spoofing attacks are of interest to the biometric community. In these attacks, the intruder employs synthetically produced artifacts (e.g., prints, video, face masks) to fraudulently access the biometric system. Conventional digital security measures, such as encryption and digital signatures, often prove inadequate in defending against these sophisticated attacks. This shows the need to develop efficient protection measures against these attacks.

Some differences may be challenging even for human observers to discern visually. 3D objects possess distinct optical characteristics absent in synthetically produced objects. Thus, these distinctions become apparent when images are translated into a suitable feature space. Extensive

[1]Our implementation is provided here

(a) (b)

Figure 1. An example is shown (a) an image captured from a real face and (b) a print image used for a spoof attack.[24]

works on algorithms for spoofing detection have explored various methodologies ranging from traditional methods (e.g. SVM based classifier with texture based features) to complex deep neural networks for feature extraction and generalization across domains (different environments) of usage.

The traditional methods can be broadly categorized based on the cues used in face spoof detection into the following groups : (i) motion-based methods, (ii) texture-based methods, (iii) methods based on image quality analysis, and (iv) methods based on other cues(voice, 3D depth, etc.). The deep learning-based models are better than the traditional methods in learning the discriminating features between the real and spoof attack attempts. However, generalization across different datasets still poses a challenge.

In our work, we emphasize traditional methods based on texture and image quality analysis as well as deep learning models, which give prominence to domain generalization.

2. Literature Review

The following section reviews key studies that form the theoretical framework for our proposed methodologies:

2.1. Traditional Methods

The traditional methods aim to detect spoofing attempts using manually designed features based on the various aspects of the input image. The local texture-based information of the input image is one class of features which can be used to distinguish between the real and fake images. Tan *et*

al. [24] considers the Lambertian reflectance to discriminate between real and spoof images. Boulkenafet *et al*. [5] has proposed the use of speed-up robust features from different colour spaces which are encoded using Fischer vectors to detect spoof attacks. The use of multi-scale local binary patterns (LBP) was exploited by Määttä *et al*. [16] to encode the micro-texture patterns, construct an enhanced feature histogram and use them to discriminate between authentic and spoofed images.

J. Galbally *et al*. [9] in their work make use of image quality assessment(IQA) for liveness detection motivated by the assumption that fake images captured in an attack attempt will have different quality than a real image acquired in the normal operation scenario. Following this "quality-difference" assumption, they consider a feature space of 25 complementary image quality measures(IQMs), which is then combined with simple classifiers to detect real and fake access attempts.

Wen *et al*. [29] has proposed a method based on image distortion analysis (IDA) to classify between real and fake cases. Similar to [9], the proposed methodology focuses on the use of face image quality differences (between real and spoofed images) and related features based on specular reflection, blurriness, chromatic moments and colour diversity of the image

2.2. Deep Learning Methods

With the development of deep learning, Yang *et al*. [31] pioneered the use of CNNs to extract features better for face anti-spoofing. In his experiments, he demonstrated that, combined with some data preprocessing, the face anti-spoofing performance increases drastically. Building upon this, Atoum *et al*. [1] introduced a two-stream CNN-based approach capable of extracting both local features and depth information. Additionally, Song *et al*. .[23] proposed three discriminative representations for face presentation attack detection.

Deep Learning methods based on temporal cues have also evolved. Xu *et al*. [30] combined CNNs with Long Short-Term Memory (LSTM) networks to simultaneously capture temporal relations and local features. Liu *et al*. [15] designed a CNN-RNN model to estimate face depth and remote Photoplethysmography(rPPG) signals, enhancing the discrimination between real and fake faces by incorporating auxiliary supervision.

However, it was identified that while the deep learning methods achieved good results on the same dataset, they struggled to generalize to other datasets.

2.3. Domain Generalization

Yunpei Jia *et al*. [12] proposed a novel adaptation of domain generalization techniques for face anti-spoofing to address this challenge. The core concept involves learning a generalized feature space wherein the feature distribution of real faces is compact, while that of fake faces is dispersed across domains but compact within each domain.

Yunpei Jia *et al*. have extended their research in [13]. They have proposed a unified unsupervised and semi-supervised domain adaptation network to further reduce the discrepancy between the source and target domains.

3. Methodology

Through our work, we plan to implement and perform a comparative study between the traditional methods and a few deep learning methods. Following this, we try to improve these methods by exploring various ways, such as feature concatenation and image preprocessing.

3.1. Multiscale Local Binary Patterns (LBP)

The LBP texture analysis (as in [16]) involves applying the LBP operator to the given image to transform it into a new space where the real and spoof images have significant difference on the basis of texture. The LBP operator labels the image pixels by thresholding the neighbouring pixels with the centering value and considering the resulting bit pattern as a binary number.

An extension of the LBP operator which looks for uniform patterns in the image is used here. A local binary pattern is called uniform if the binary pattern contains at most two bit wise transitions from 0 to 1 or vice versa when the bit pattern is traversed circularly. There is a separate label for each uniform pattern, and all the non-uniform patterns are given a single label. The histogram of the transformed image is utilized as a discriminatory feature for classification between authentic and spoofed images. Fig.2 shows the real and printed faces in the transformed domain.

Figure 2. Examples of two images (a real face (left), face print (right)) in the original space and the corresponding LBP images using basic LBP as a feature space

3.2. Image Quality Assessment (IQA)

We perform IQA of the input sample using a wide range of general image quality measures (IQMs) [9], which exploit complementary image quality properties. The feature vectors consists of 23 Image Quality Measures(IQMs) based on [9] and are used to classify between real and spoof images.

Full-reference(FR) IQA methods rely on the availability of a clean reference image to estimate the quality of the

test sample. However, in the problem of fake detection, we do not have access to such a reference image and only the test sample is known. Hence, to overcome this limitation, the same strategy already successful for image manipulation detection [4] and steganalysis [3] is implemented here.

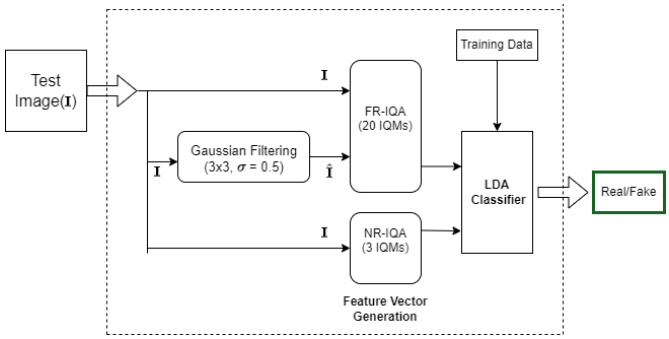


Figure 3. General Diagram showing the implementation of Image Quality Assessment(IQA) in the present work. [9].

The input image I is filtered with a low-pass Gaussian kernel to generate a smoothed version $\hat{I}$. The quality between both images (I and $\hat{I}$) is computed according to the corresponding full-reference IQA metric. The fact that loss of quality produced by Gaussian filtering differs between real and spoof images is used as a classification criteria.

No-reference(NR) IQA methods try to tackle the challenging problem of assessing the quality of images, in the absence of a reference. It is based on the principle that, in general, the human visual system does not require a reference sample to determine the quality level of an image. In our work, the NR-IQA methods estimate the quality of the test image based on some pre-trained statistical models.

3.3. Deep Learning Methods

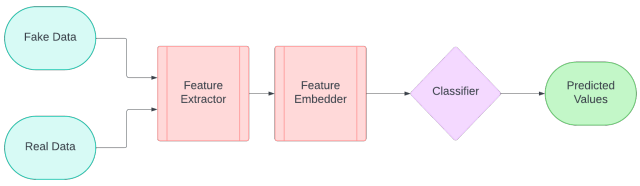


Figure 4. Baseline Model

We plan to implement a Baseline classifier to predict between the real and the fake faces. To do this, we use a pre-trained Feature Extractor, then pass it to a Feature Embedder, and finally use a Artificial Neural Network classifier with L2-Normalization to classify the images as real or fake.

Figure 5. Feature Space
Circular blobs represent the fake data while the cross represents real data. Clearly, the real data is compact while the fake data is dispersed.[12]

3.4. Single Side Domain Generalization (SSDG)

We implement the Single-Side Domain Generalization model, training it on two datasets and subsequently evaluating its performance on a distinct dataset. The method employs Single-Side Adversarial Training. The model has a domain discriminator, which discriminates between the features of real faces and fake faces, and a feature generator, which competes with the domain discriminator to make the real faces from the different datasets compact. Furthermore, an asymmetric triplet loss is utilized to effectively separate fake faces from different domains and aggregate real faces from various domains. The approach incorporates feature and weight normalization techniques to enhance the generalization capability further.

We plan to compare the performance of the Baseline model with our SSDG model, and create inferences from the two models.

4. Datasets

In our project, we work with 3 different datasets:

1. **NUAA ImposterDB** [25]: It has 5105 Real Images, and 7509 Spoof Images from 15 different subjects. The type of spoofing attacks used in this dataset was only print type (flat or wrapped over the face).
2. **LCC-FASD [26] + CASIA [33]** : The Large Crowdcollected Facial Anti Spoofing Dataset (LCC-FASD) contains 1942 real images and 16885 spoof images of over 243 different identities and uses a wide range of spoofing and recording devices.
   The CASIA dataset provides us with 12 video clips from 50 subjects, 3 genuine and 9 spoof. From the video clips, we collect the individual frames to put into our dataset. In total, we get 99 real images and 463 spoof images from 50 different subjects. This dataset uses print-type

spoofing attacks as well as using tablets.
The combined version of LCC-FASD and CASIA databases is available on Kaggle.

3. **Idiap Replay Attack [7]**: The Replay-Attack Database comprises 1300 video clips depicting both photo and video-attack attempts against 50 distinct clients captured under various lighting conditions. Different devices were used to generate the attack: print, mobile (phone), high definition (tablet), photo and video. The dataset is divided into 3 subsets: training set(360 videos), development(360 videos), and test set(480 videos). 25 frames per video were taken in our project.

5. Preprocessing

Initially, we use HAAR cascade classifier to detect the location of the face in the image and crop it out. This way the final classification results are ensured to be unbiased and not dependent on contextual-specific artifacts such as unwanted changes in the background, different sizes of the heads etc [9]. Following this, the image is normalized. For the classical methods, we have resized the image to 64x64, and for the deep learning method, we have resized it to 256x256 following the implementations in the respective papers.

Further, for the SSDG model, we have applied Random Horizontal Flip which randomly flips images horizontally with a probability of 0.5. This increases the diversity of the training dataset.

6. Experimental Setup

6.1. Multiscale Local Binary Patterns (LBP)

The following notation is used for the LBP operator: $LBP^{u2}_{P,R}$. The subscript represents the pixel neighborhood of radius R with P sampling points. The superscript u2 stands for using uniform patterns and labeling all remaining patterns with a single label. The detailed description and implementation of the LBP operator can be found in [20])

We apply the $LBP^{u2}_{8,1}$ operator on the normalized input face image and divide the resulting LBP face image into 3 x 3 overlapping regions (with an overlapping size of 96 pixels). The local 10-bin histogram of each of the regions is computed and collected into a 90-bin histogram. Then, we compute two other histograms from the whole face image using $LBP^{u2}_{8,2}$ and $LBP^{u2}_{16,1}$ operators, yielding two more 10-bin histograms and appended to the previous 90-bin histogram. Hence, the length of the final enhanced feature histogram vector, which is used as the feature vector of each image, is 110.

Using the LBP feature space and feature vector generated for each image in the training set, we try to fit a Linear Discriminant Analysis (LDA) classifier model.

6.2. Image Quality Assessment

The IQMs used to generate the feature vector are formally described in Table 1 (1).

6.2.1 Full-Reference IQMs

The FR-IQMs have been classified into the different categories according to the image property assessed:

1. **Pixel Difference Measures**: (MSE, PSNR, SNR, SC, MD, AD, NAE, RAMD, LMSE) These features compute the distortion between two images based on their pixel-wise differences.
   (a) In the RAMD entry in Table 1, max_r is defined as the r -highest pixel difference between two images. In the present implementation, R = 10.
   (b) In the LMSE entry in Table 1, $h(I_{i,j}) = I_{i+1,j} + I_{i-1,j} + I_{i,j+1} + I_{i,j-1} - 4I_{i,j}$.
2. **Correlation-based measures**: (NXC, MAS, MAMS) A variant of correlation-based approach is obtained by considering the statistics of the angles between the pixel vectors of the original and distorted images.
   (a) In the MAS and MAMS entries in Table 1, $\alpha_{i,j}$ denotes the angle between two vectors, defined as $\alpha_{i,j} = \frac{2}{\pi} \arccos \frac{<I_{i,j}, \hat{I}_{i,j}>}{||I_{i,j}||.||\hat{I}_{i,j}||}$ where $< I_{i,j}, \hat{I}_{i,j} >$ denotes the scalar product and $\frac{2}{\pi}$ is a normalizing factor.
3. **Edge-based measures**: (TCD, TED) The structural distortion of an image is closely connected to its edge deterioration (edges and corners being highly informative elements of an image); hence we analyse two measures of edge quality: TED and TCD.
   (a) In the TED entry, I_E and $\hat{I}_E$ refer to the edge map created after applying the Sobel operator (gradient-analysis) to the images.
   (b) For TCD, The Harris corner detector [10] is used to to compute the number of corners N_{cr} and $\hat{N}_{cr}$ found in I and $\hat{I}$.
4. **Spectral Distance Measures**: (SME, SPE) The analysis of frequency domain (based on Fourier Transform) can provide significant distinguishing factors between original and distorted image.
   (a) In Table 1, F and $\hat{F}$ are the respective Fourier transforms of I and $\hat{I}$), and $arg(F)$ denotes phase
5. **Gradient Based Measures**: (GME, GPE) Distortions present in the image can sometimes be projected in its gradient map and hence are valuable for analysing quality.
   (a) For GME & GPE, G and $\hat{G}$ are the gradient maps of I and $\hat{I}$ defined as $G = G_x + iG_y$, where $G_x$and G_y are the gradients along the x and y directions
6. **Structural Similarity Measures**: Structural Similarity Index Measure (SSIM) assesses the likeness between an

Table 1. List of the 23 Image Quality Measures(IQMs) implemented for Image Quality Assessment(section 6.2) [9]. All the measures were directly taken or adapted from the references mentioned. $\mathbf{I}$ denotes the input test image of size (MxN) and $\hat{\mathbf{I}}$ denotes the smoothed version of the input image after filtering it with a Gaussian Kernel($\sigma = 0.5$ and $size = 3 \times 3$). For other undefined functions and variables specifications, refer to section 6.2 for the corresponding description.

#	Type	Acronym	Name	Ref.	Description
1	FR	MSE	Mean Squared Error	[2]	$MSE(\mathbf{I},\hat{\mathbf{I}}) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\mathbf{I}_{i,j} - \hat{\mathbf{I}}_{i,j}\right)^2$
2	FR	PSNR	Peak Signal to Noise Ratio	[11]	$PSNR(\mathbf{I},\hat{\mathbf{I}}) = 10\log(\frac{\max(\mathbf{I}^2)}{\text{MSE}(\mathbf{I},\hat{\mathbf{I}})})$
3	FR	SNR	Signal to Noise Ratio	[32]	$SNR(\mathbf{I},\hat{\mathbf{I}}) = 10\log(\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(\mathbf{I}_{i,j})^2}{N\cdot M\cdot MSE(\mathbf{I},\hat{\mathbf{I}})})$
4	FR	SC	Structural Content	[8]	$SC(\mathbf{I},\hat{\mathbf{I}}) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(\mathbf{I}_{i,j})^2}{\sum_{i=1}^{N}\sum_{j=1}^{M}(\hat{\mathbf{I}}_{i,j})^2}$
5	FR	MD	Maximum Difference	[8]	$MD(\mathbf{I},\hat{\mathbf{I}}) = \max\left\lvert\mathbf{I}_{i,j} - \hat{\mathbf{I}}_{i,j}\right\rvert$
6	FR	AD	Average Difference	[8]	$AD(\mathbf{I},\hat{\mathbf{I}}) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\mathbf{I}_{i,j} - \hat{\mathbf{I}}_{i,j}\right)$
7	FR	NAE	Normalized Absolute Error	[8]	$NAE(\mathbf{I},\hat{\mathbf{I}}) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\lvert\mathbf{I}_{i,j} - \hat{\mathbf{I}}_{i,j}\rvert}{\sum_{i=1}^{N}\sum_{j=1}^{M}\lvert\mathbf{I}_{i,j}\rvert}$
8	FR	RAMD	R-Averaged MD	[2]	$RAMD(\mathbf{I},\hat{\mathbf{I}},R) = \frac{1}{R}\sum_{r=1}^{R}\max_r\left\lvert\mathbf{I}_{i,j} - \hat{\mathbf{I}}_{i,j}\right\rvert$
9	FR	LMSE	Laplacian MSE	[8]	$LMSE(\mathbf{I},\hat{\mathbf{I}}) = \frac{\sum_{i=1}^{N-1}\sum_{j=2}^{M-1}(h(\mathbf{I}_{i,j}) - h(\hat{\mathbf{I}}_{i,j}))^2}{\sum_{i=1}^{N-1}\sum_{j=2}^{M-1}h(\mathbf{I}_{i,j})^2}$
10	FR	NXC	Normalized Cross-Correlation	[8]	$NXC(\mathbf{I},\hat{\mathbf{I}}) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(\mathbf{I}_{i,j}\cdot\hat{\mathbf{I}}_{i,j})}{\sum_{i=1}^{N}\sum_{j=1}^{M}(\mathbf{I}_{i,j})^2}$
11	FR	MAS	Mean Angle Similarity	[2]	$MAS(\mathbf{I},\hat{\mathbf{I}}) = 1 - \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(\alpha_{i,j})$
12	FR	MAMS	Mean Angle Magnitude Similarity	[2]	$MAMS(\mathbf{I},\hat{\mathbf{I}}) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(1 - [1-\alpha_{i,j}]\,[1 - \frac{\lVert\mathbf{I}_{i,j} - \hat{\mathbf{I}}_{i,j}\rVert}{255}])$
13	FR	TED	Total Edge Difference	[17]	$TED(\mathbf{I},\hat{\mathbf{I}}) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left\lvert\mathbf{I}_{\mathbf{E}i,j} - \hat{\mathbf{I}}_{\mathbf{E}i,j}\right\rvert$
14	FR	TCD	Total Corner Difference	[17]	$TCD(\mathbf{I},\hat{\mathbf{I}}) = \frac{\lvert N_{cr} - \hat{N}_{cr}\rvert}{\max(N_{cr},\hat{N}_{cr})}$
15	FR	SME	Spectral Magnitude Error	[19]	$SME(\mathbf{I},\hat{\mathbf{I}}) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(\lvert\mathbf{F}_{i,j}\rvert - \lvert\hat{\mathbf{F}}_{i,j}\rvert)^2$
16	FR	SPE	Spectral Phase Error	[19]	$SPE(\mathbf{I},\hat{\mathbf{I}}) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\lvert\arg(\mathbf{F}_{i,j}) - \arg(\hat{\mathbf{F}}_{i,j})\rvert^2$
17	FR	GME	Gradient Magnitude Error	[14]	$GME(\mathbf{I},\hat{\mathbf{I}}) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(\lvert\mathbf{G}_{i,j}\rvert - \lvert\hat{\mathbf{G}}_{i,j}\rvert)^2$
18	FR	GPE	Gradient Phase Error	[14]	$GPE(\mathbf{I},\hat{\mathbf{I}}) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\lvert\arg(\mathbf{G}_{i,j}) - \arg(\hat{\mathbf{G}}_{i,j})\rvert^2$
19	FR	SSIM	Structural Similarity Index	[28]	Refer to [28]
20	FR	VIF	Visual Information Fidelity	[22]	Refer to [22]
21	NR	JQI	JPEG Quality Index	[27]	Refer to [27]
22	NR	HLFI	High-Low Frequency Index	[34]	$SME(\mathbf{I}) = \frac{\sum_{i=1}^{i_l}\sum_{j=1}^{j_l}\lvert\mathbf{F}_{i,j}\rvert - \sum_{i=i_h+1}^{N}\sum_{j=j_h+1}^{M}\lvert\mathbf{F}_{i,j}\rvert}{\sum_{i=1}^{N}\sum_{j=1}^{M}\lvert\mathbf{F}_{i,j}\rvert}$
23	NR	NIQE	Naturalness Image Quality Estimator	[18]	Refer to [18]

original image, represented as I, and its approximation, denoted as $\hat{I}$. It operates by gauging a similarity metric that closely aligns with the subjective qualitative perception of humans.

7. **Information Theoretic Measures**: (VIF) VIF between I and $\hat{I}$ can be quantified using the mutual information between them. The VIF metric measures the quality fidelity as the ratio between the total information (measured in terms of entropy) ideally extracted by the brain from the whole distorted image and the total information conveyed within the complete reference image.

6.2.2 No-Reference IQMs

Based on the images that were used to train the NR-IQM models and on a priori knowledge, the NR-IQMs used in this work are broadly divided into the following categories:

1. **Distortion-specific approaches:** (JQI, HLFI) Based on previously acquired knowledge about the quality loss caused by a specific distortion. JQI evaluates the quality in images affected by the block artifacts due to low bit-rates compression algorithms like JPEG. HLFI considers local gradients as a blind metric to detect blur & noise and computes the difference between the power in the lower and upper frequencies of the Fourier Spectrum.
   (a) In Table 1 for HLFI, i_l, i_h, j_l, j_h are the indices corresponding to the lower and upper frequency thresholds respectively. For our work, $i_l = i_h = 0.15N$ & $j_l = j_h = 0.15M$
2. **Natural Scene Statistic approach:** (NIQE) Here, natural scene distortion-free images are used to train the initial model. The rationale behind this relies on the hypothesis that undistorted images of the natural world contain certain *regular* properties which can help to evaluate the perceptual quality of an image. The NIQE is a completely blind IQM based on the construction of statistical features related to a multi variate Gaussian natural scene model.

Using these IQMs, a 23 length feature vector was constructed for each image in the training set, which was then used to fit the LDA model.

6.3. Deep Learning Methods

6.3.1 Loss Function

For our baseline classifier, we have selected binary cross-entropy loss as we are dealing with only the classification problem and not domain generalization.

However, for our Single Side Domain Generalization model, we have used three different loss functions:

1. *Asymmetric Triplet Loss*: As mentioned earlier, to separate the fake faces and to aggregate the real faces from different domains, we use the Asymmetric Triplet Loss.

$$\min_{G} \mathcal{L}_{\text{AsTrip}}(G) = \sum_{x_i^\alpha, x_i^p, x_i^n} \big(\|f(x_i^\alpha) - f(x_i^p)\|_2^2 - \|f(x_i^\alpha) - f(x_i^p)\|_2^2 + \alpha \big) \tag{1}$$

   where the labels of anchor x_i^α and positive example x_i^p are the same, while those of x_i^α and negative example x_i^n are different. The α is the predefined margin.
2. *Cross Entropy Loss* ($\mathcal{L}_{\text{Cls}}$): To make sure the classification happens smoothly, we use the Cross Entropy Loss to train the Classifier.
3. *Adversarial Loss*: As all the real faces are collected from taking pictures of real people, the distribution of all these images should be compact. Therefore, we try to learn a generalized feature space using Single Side Adversarial Learning.
   Let X_r and X_f denote the real and fake faces in the N source domains, $D = D_1, D_2, ..., D_N$. Let G_r and G_f be the feature extractors (collectively denoted by G), which extract the features Z_r and Z_f from X_r and X_f, respectively. The domain generator D uses Z_r to know which source domain the input features come from. So, the loss function required to train the generalized feature space can be given by -

$$\min_{D} \max_{G} \mathcal{L}_{\text{Ada}}(G, D) = -\mathbb{E}_{x,y \sim X_r, Y_D} \left[\sum_{n=1}^{N} \mathbb{I}_{[n=y]} \log D(G(x)) \right]$$

   where Y_D represents the set of domain labels.

We combine all three of these losses together to get the final loss function for our SSDG model:

$$\mathcal{L}_{\text{SSDG}} = \mathcal{L}_{\text{Cls}} + \lambda_1 \mathcal{L}_{\text{Ada}} + \lambda_2 \mathcal{L}_{\text{AsTrip}} \tag{2}$$

where λ_1 and λ_2 are the balanced parameters.

6.3.2 Implementation Details

We conducted extensive experimentation, systematically employing one dataset for testing while utilizing the remaining two for training. This approach was applied across all possible combinations, and the models were trained and tested on each for 200 epochs. For optimization, we used the SGD optimizer with momentum 0.9 and weight decay 5e-4. The hyperparameter α is set to 1. We adjust the learning rate by 0.1 after 100 epochs.

We use the architecture based on ResNet-18 proposed in [12], which replaces the last average pooling layer with a global pooling layer (GAP). A fully connected layer (FC) with 512 hidden units acts as the bottleneck layer. The resulting architecture employs a linear model for the anti-spoofing classifier.

The feature generator, embedder and classifier is same for both our baseline classifier and the SSDG model. However, in addition, the SSDG model has a domain discriminator with two FC layers.

7. Evaluation Metrics

We use three main metrics to evaluate our model:

1. *Half Total Error Rate*:
   HTER provides a single measure of performance, balancing the rates of falsely accepting impostors and falsely rejecting genuine users.

$$\text{HTER} = \frac{\text{FAR} + \text{FRR}}{2} \tag{3}$$

2. *AUC score and ROC curve*:
   The AUC score represents the area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate (sensitivity) against the false positive rate (specificity) for different threshold values. An AUC score of 1 represents perfect classification and a score of 0.5 represents random guessing.
3. *Accuracy*:
   Accuracy is calculated as the ratio of the number of correctly predicted instances to the total number of instances in the dataset. It provides an intuitive measure of the model's overall correctness in making predictions.
4. *t-SNE*:
   t-distributed Stochastic Neighbor Embedding, or t-SNE in short, is a non-linear visualization technique [6]. It is calculated by checking the similarity between two points, which is just the conditional probability that point A would have point B as its neighbor.

8. Experimental Results

8.1. Classical Methods

8.1.1 Training on individual datasets

Measure	Dataset		
	NUAA	Replay	LCC
Accuracy	97.70	91.94	86.67
AUC	0.99	0.72	0.92
HTER	0.02	0.44	0.18

Table 2. Evaluation results for a LBP based classifier on various datasets

Measure	Dataset		
	NUAA	Replay	LCC
Accuracy	94.25	91.98	91.56
AUC	0.98	0.77	0.97
HTER	0.06	0.44	0.08

Table 3. Evaluation results for a IQA based classifier on various datasets

8.1.2 Training on multiple datasets

	Model trained on								
	L and N			N and R			R and L		
	HTER	AUC	ACC	HTER	AUC	ACC	HTER	AUC	ACC
L	0.48	0.66	89.83	0.48	0.61	70.60	0.47	0.66	88.53
N	0.06	0.98	92.98	0.043	0.99	95.62	0.31	0.74	60.91
R	0.53	0.5	65.52	0.34	0.78	73.52	0.25	0.84	79.00

Table 4. Evaluation results of LBP based Classifier for training on multiple datasets. L: LCC, N: NUAA , R: Replay Attack

	Model trained on								
	L and N			N and R			R and L		
	HTER	AUC	ACC	HTER	AUC	ACC	HTER	AUC	ACC
L	0.48	0.66	85.69	0.49	0.58	69.43	0.46	0.724	88.99
N	0.09	0.96	89.53	0.07	0.98	93.22	0.41	0.37	60.95
R	0.21	0.72	79.08	0.30	0.80	77.52	0.680	0.96	88.70

Table 5. Evaluation results of IQA based Classifier for training on multiple datasets. L: LCC, N: NUAA , R: Replay Attack

LBP based Classifier: We can see that the LBP based classifier gave good results while training and testing on the same dataset. However, when training on multiple datasets (taken two at a time) ,the quality of results in terms of the evaluation metrics have significantly gone down.

IQA based Classifier: Similar to the LBP based classifer,the IQA based model gives good performance w.r.t. individual datasets training whereas performance drops drastically when trained on multiple datasets.

Thus we can conclude from the experimental results the classifiers for spoof detection obtained for classical methods give good performance with respect to specific scenarios and fail to generalize when employed in an environment where it was not trained on.

8.2. Baseline Classifier

	Model trained on								
	L and N			N and R			R and L		
	HTER	AUC	ACC	HTER	AUC	ACC	HTER	AUC	ACC
L	0.12	0.94	91.19	0.28	0.76	65.00	0.25	0.80	74.33
N	0.00	1.00	100	0.00	1.00	100	0.26	0.79	73.33
R	0.30	0.86	70.00	0.00	1.00	100	0.34	0.89	66.00

Table 6. Evaluation results of the baseline classifier on different datasets.
L: LCC-FASD [26], N: NUAA [25], R: Replay Attack [7]

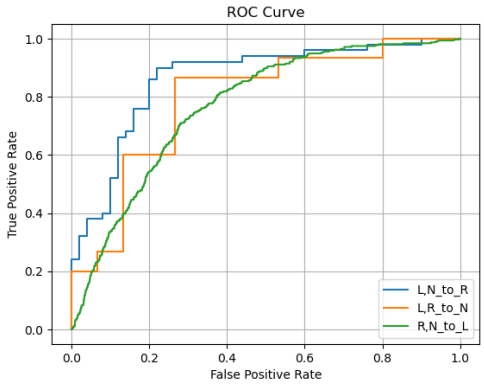


Figure 6. ROC curves for Baseline Model tested on a different domain than the training domains.

8.3. Domain Generalization

	Model trained on								
	L and N			N and R			R and L		
	HTER	AUC	ACC	HTER	AUC	ACC	HTER	AUC	ACC
L	0.09	0.98	98.07	0.18	0.89	82.11	0.02	0.99	99.68
N	0.22	0.99	83.33	0.00	1.00	100	0.30	0.98	70.00
R	0.07	0.98	93.00	0.00	1.00	100	0.00	1.00	100

Table 7. Evaluation results of SSDG on different datasets.
L: LCC-FASD [26], N: NUAA [25], R: Replay Attack [7]

Compared to the baseline classifier, the SSDG model performed better except when the model was trained on Replay-Attack and LCC-FASD. In this particular case, the baseline classifier marginally outperformed the SSDG model in terms of accuracy on the NUAA dataset. However, it's important to note that the baseline model's accuracy on its respective training datasets was much lower than that of the SSDG model. This suggests that the SSDG model demonstrates significant improvements over the baseline model in generalizing to diverse domains.

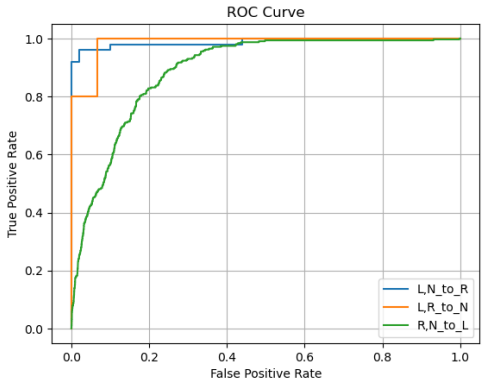


Figure 7. ROC curves for Domain Generalization tested on a different domain than the training domains

8.4. Domain Generalization with Pre-processing

In addition to the above, we experiment with various pre-processing methods.

1. *Histogram Equalization*: It enhances the contrast of the images to improve the feature generation of the images.
2. *Local Binary Patterns*: We pass Local Binary Patterns (LBP) applied on the HSV channels of the images to our model [1]. It helps the model detect the patterns in the images and provides texture information as well.

The model was trained on LCC-FASD and Replay-Attack and tested on NUAA.

Pre-processing	HTER	AUC	ACC
Histogram Equalization	0.86	0.43	56.67
LBP	0.45	0.82	55.00

Table 8. Evaluation results of SSDG with various pre-processing.

The pre-processing methods exhibit a decrement in performance compared to models without pre-processing, implying ineffectiveness in enhancing model generalization across diverse datasets. However, notable proficiency is observed when evaluating these models on the same datasets used for training. This suggests limitations in the generalization capacity of the employed pre-processing techniques.

8.5. t-SNE plots

In this section, we plot the results of applying t-SNE on the feature vectors generated on the training dataset.

In Fig.8a 8b 8c, we see the t-SNE plots for the LBP feature vectors generated over different datasets. We observe that the feature vectors corresponding to real and spoof images of the LCC dataset have a very high overlap making the classification difficult. In case of Replay-attack also, the feature vectors overlap each together but clearly they form clusters whereas in NUAA they are clearly distinguishable.

In Fig.8d 8e 8f, we see the t-SNE plots for the the IQA feature vectors generated over different datasets. We observe that similar to the LBP case, the feature vectors corresponding to the real and spoof images of the LCC dataset have a very high overlap making the classification difficult. Incase of NUAA & Replay attack the feature vectors have significantly lower overlap.

For the Deep Learning methods, we present the results obtained when we used NUAA and Replay-Attack for training and LCC-FASD for testing. Our analysis is based on 75 images from both real and fake image categories across each dataset.

In Fig.8g, utilizing solely the feature generator with pre-trained weights, the distinction between real and fake images is subtle. However, post-training the model without domain generalization (Fig.8h), a discernible boundary emerges between the two. Following training with the SSDG model, there is a clear boundary between the real and fake images (Fig.8i). As expected, real images tend to cluster together, while fake images scatter across disparate points within the feature space, justifying its superior performance.

9. Future Work

From the results & observations, we observe that the feature space generated using texture (LBP) and IQA based attributes, when applied to classical machine learning methodologies fail to perform well on a new scenario. Hence, there is a need to explore techniques which could generalise well to new environments. Devising new feature spaces or combining different feature spaces together (like LBP measures and IQA measures) are two possible directions of solution to the above scenario.

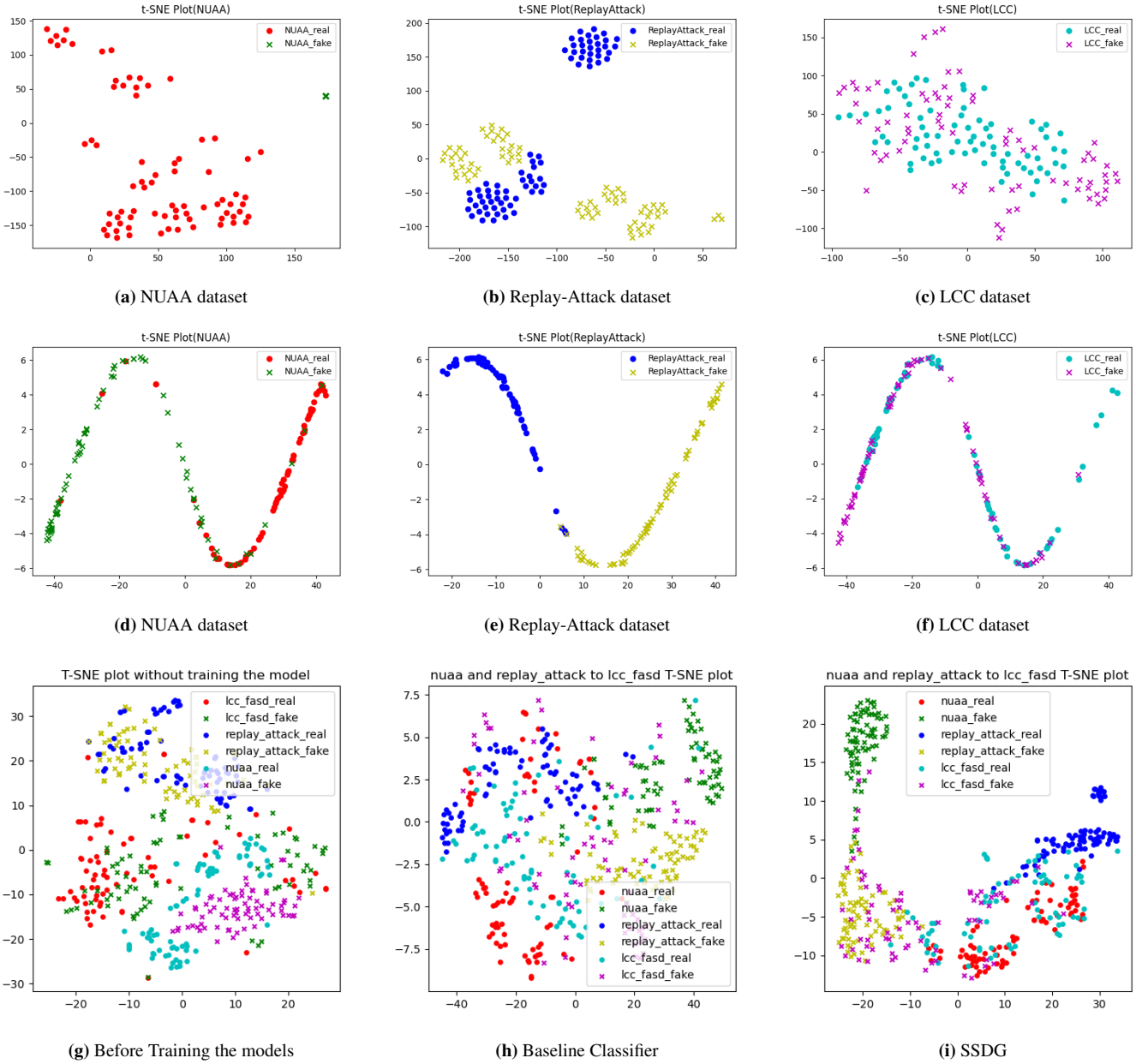
For the domain generalization methods, our current focus has predominantly been on the RGB color space. However, alternative color spaces may offer advantages in face anti-spoofing compared to RGB. The HSV and YC_bC_r color spaces, for instance, segregate illuminance and chrominance information, thereby giving additional discriminative cues for learning [1]. In our work, we have briefly touched upon using LBP on the HSV color space as well. Experiments on solely the HSV and YC_bC_r space can be tried out as well.

Further, other asymmetric designs can be explored, such as dividing the data based on the attack types rather than the databases. [12]

References

[1] Yousef Atoum, Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Face anti-spoofing using patch and depth-based cnns. 2017. 2, 8

[2] Ismail Avcibas, Bulent Sankur, and K. Sayood. Statistical



(a) NUAA dataset **(b)** Replay-Attack dataset **(c)** LCC dataset

(d) NUAA dataset **(e)** Replay-Attack dataset **(f)** LCC dataset

(g) Before Training the models **(h)** Baseline Classifier **(i)** SSDG

Figure 8. (a)-(c) : t-SNE visualisation plots of LBP feature vector on various datasets
(d)-(f) : t-SNE visualisation plots of IQA feature vector on various datasets
(g)-(i) : t-SNE plots of Deep Learning Methods

evaluation of quality measures. j electron imaging. *J. Electronic Imaging*, 11:206–223, 2002. 5

[3] I. Avcibas, N. Memon, and B. Sankur. Steganalysis using image quality metrics. *IEEE Transactions on Image Processing*, 12(2):221–229, 2003. 3

[4] Sevinc Bayram, Ismail Avcibas, Bulent Sankur, and Nasir Memon. Image manipulation detection. *J. Electronic Imaging*, 15:041102, 2006. 3

[5] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face antispoofing using speeded-up robust features and fisher vector encoding. *IEEE Signal Processing Letters*, 24(2):141–145, 2017. 2

[6] T. Tony Cai and Rong Ma. Theoretical foundations of t-sne for visualizing high-dimensional clustered data. 2021. 6

[7] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face antispoofing. In *International Conference of Biometrics Special Interest Group (BIOSIG)*, pages 1–7, 2012. 4, 7

[8] A.M. Eskicioglu and P.S. Fisher. Image quality measures and their performance. *IEEE Transactions on Communications*, 43(12):2959–2965, 1995. 5

[9] Javier Galbally, Sébastien Marcel, and Julian Fierrez. Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. *IEEE Transactions on Image Processing*, 23(2):710–724, 2014. 2, 3, 4, 5

[10] Chris Harris, Mike Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, pages 10–5244. Citeseer, 1988. 4

[11] Q. Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics Letters*, 44:800 – 801, 2008. 5

[12] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. *arXiv preprint arXiv:2004.14043*, 2020. 2, 3, 6, 8

[13] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Unified unsupervised and semi-supervised domain adaptation network for cross-scenario face anti-spoofing. *Pattern Recognition*, 115:107888, 2021. 2

[14] Anmin Liu, Weisi Lin, and Manish Narwaria. Image quality assessment based on gradient similarity. *IEEE Transactions on Image Processing*, 21(4):1500–1512, 2012. 5

[15] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 389–398, 2018. 2

[16] Jukka Määttä, Abdenour Hadid, and Matti Pietikäinen. Face spoofing detection from single images using micro-texture analysis. In *2011 international joint conference on Biometrics (IJCB)*, pages 1–7. IEEE, 2011. 2

[17] Maria G. Martini, Chaminda T.E.R. Hewage, and Barbara Villarini. Image quality assessment based on edge preservation. *Signal Processing: Image Communication*, 27(8):875–882, 2012. Special issue on: pervasive mobilemultimedia. 5

[18] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 5

[19] Norman B. Nill and Brian Bouzas. Objective image quality measure derived from digital image power spectra. *Optical Engineering*, 31(4):813 – 825, 1992. 5

[20] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002. 4

[21] S. Prabhakar, S. Pankanti, and A.K. Jain. Biometric recognition: security and privacy concerns. *IEEE Security Privacy*, 1(2):33–42, 2003. 1

[22] H.R. Sheikh and A.C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006. 5

[23] Xiao Song, Xu Zhao, Liangji Fang, and Tianwei Lin. Discriminative representation combinations for accurate face spoofing detection. *Pattern Recognition*, 85:220–231, 2019. 2

[24] Xiaoyang Tan, Yi Li, Jun Liu, and Lin Jiang. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI 11*, pages 504–517. Springer, 2010. 1, 2

[25] Xiaoyang Tan, yi Liu, Jun Liu, and Lin Jiang. Face liveness detection from a single image with sparse low rank bilinear discriminative model. 6316:504–517, 2010. 3, 7

[26] Denis Timoshenko, Konstantin Simonchik, Vitaly Shutov, Polina Zhelezneva, and Grishkin Valery. Large crowdcollected facial anti-spoofing dataset. pages 123–126, 2019. 3, 7

[27] Zhou Wang, H.R. Sheikh, and A.C. Bovik. No-reference perceptual quality assessment of jpeg compressed images. In *Proceedings. International Conference on Image Processing*, pages I–I, 2002. 5

[28] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 5

[29] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, 2015. 2

[30] Zhenqi Xu, Shan Li, and Weihong Deng. Learning temporal features using lstm-cnn architecture for face anti-spoofing. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 141–145, 2015. 2

[31] Jianwei Yang, Zhen Lei, and Stan Z. Li. Learn convolutional neural network for face anti-spoofing, 2014. 2

[32] S.s Yao, Weisi Lin, Ee Ong, and Zhongkang Lu. Contrast signal-to-noise ratio for image quality assessment. pages 397–400, 2005. 5

[33] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Li. A face antispoofing database with diverse attacks. *Proceedings - 2012 5th IAPR International Conference on Biometrics, ICB 2012*, pages 26–31, 2012. 3

[34] Xiang Zhu and Peyman Milanfar. A no-reference sharpness metric sensitive to blur and noise. In *2009 International Workshop on Quality of Multimedia Experience*, pages 64–69, 2009. 5