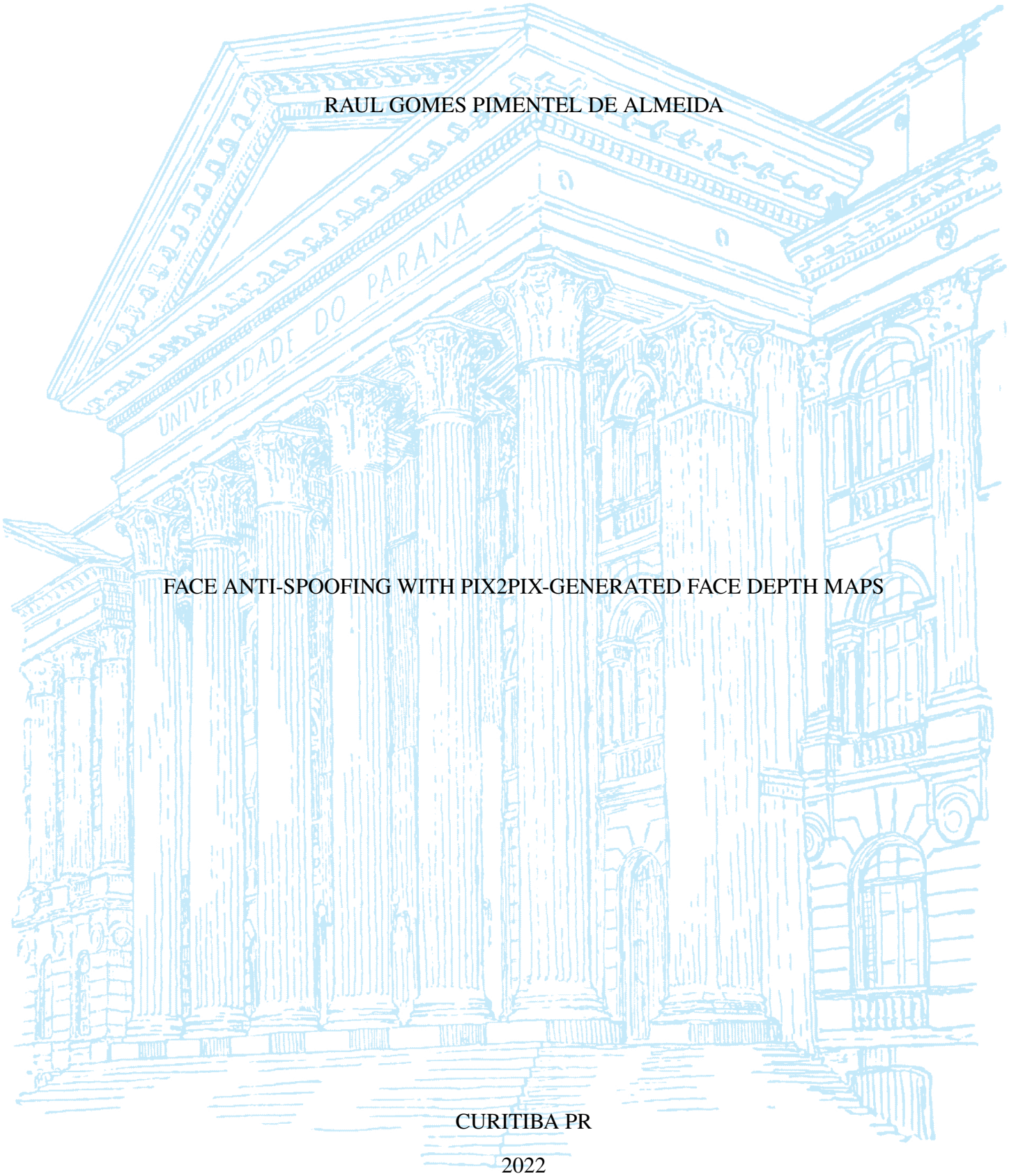
UNIVERSIDADE FEDERAL DO PARANÁ

RAUL GOMES PIMENTEL DE ALMEIDA



FACE ANTI-SPOOFING WITH PIX2PIX-GENERATED FACE DEPTH MAPS

CURITIBA PR

2022

RAUL GOMES PIMENTEL DE ALMEIDA

FACE ANTI-SPOOFING WITH PIX2PIX-GENERATED FACE DEPTH MAPS

Trabalho apresentado como requisito parcial à conclusão do Curso de Bacharelado em Ciência da Computação, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: David Menotti Gomes.

CURITIBA PR

2022

Universidade Federal do Paraná
Setor de Ciências Exatas
Curso de Ciência da Computação

Ata de Apresentação de Trabalho de Graduação II

Título do Trabalho: FACE ANTI-SPOOFING WITH PIX2PIX-GENERATED FACE DEPTH MAPS

Autor(es):
GRR 20182544 None: RAUL GOMES PIMENTEL DE ALMEIDA

GRR _____ Nome: _____

GRR _____ Nome: _____

Apresentação: Data: 16 / 09 / 2022 Hora: 13:00 Local: https://meet.google.com/drj-wyfj-bzd

Orientador: DAVID MENOTTI GOMES _____

Membro 1: GUSTAVO FÜHR _____

Membro 2: ANDRÉ RICARDO ABED GRÉGIO _____

(nome) (assinatura)

AVALIAÇÃO – Produto escrito		ORIENTADOR	MEMBRO 1	MEMBRO 2	MÉDIA
Conteúdo	(00-40)				
Referência Bibliográfica	(00-10)				
Formato	(00-05)				
AVALIAÇÃO – Apresentação Oral					
Domínio do Assunto	(00-15)				
Desenvolvimento do Assunto	(00-05)				
Técnica de Apresentação	(00-03)				
Uso do Tempo	(00-02)				
AVALIAÇÃO – Desenvolvimento					
Nota do Orientador	(00-20)		**********	**********	
NOTA FINAL		**********	**********	**********	**80**

Pesos indicados são sugestões.
Conforme decisão do colegiado do curso de Ciência da Computação, a entrega dos documentos comprobatório de trabalho de graduação 2 deve deve respeitar os seguintes procedimentos: Orientador deve abrir um processo no Sistema Eletrônico de Informações (SEI – UFPR); Selecionar o tipo: Graduação: Trabalho Conclusão de Curso; informar os interessados: nome do aluno e o nome do orientador; anexar esta ata escaneada e a versão final do pdf da monografia do aluno.; Tramita o processo para CCOMP (Coordenação Ciência da Computação).

Ao registrador que primeiro guardou um caractere do texto fonte dedico como saudosa lembrança este texto compilado

AGRADECIMENTOS

Minha família me ensinou o início do que eu sei. Meus amigos, um pouco do resto. A última parte aprendi de uma mistura confusa de acaso e intenção. Sem qualquer um destes três, não teria aprendido nada. Obrigado!

Agradeço ao meu orientador por uma orientação completa e organizada.

À minha mãe por me levar à escola no primeiro dia de aula da minha vida, e em boa parte dos que vieram depois.

Ao meu pai, por me ensinar infinitésimos e logaritmos de uma maneira intuitiva antes que eu soubesse dizer o que é uma fração.

À minha irmã, por insistir em nossa amizade.

À minha namorada, por apoiar minhas decisões nessa jornada.

A todas as minhas amizades, por fazerem de minha vida uma história na qual quero participar.

Conforme me aproximo da obtenção de um bacharelado, os rostos que vejo pelos corredores da universidade são cada vez mais estranhos. O corpo estudantil que eu conheci quando entrei aqui não é mais o mesmo e, de alguma forma, permanece Um. Eu agradeço acima de tudo à inevitável passagem do tempo, que sempre me obrigou à impermanência.

RESUMO

Face spoofing consiste em simular características biométricas faciais de uma pessoa de maneira a personificá-la em um sistema de reconhecimento facial (por exemplo, em aplicações de pagamento digital e mídia social). Detecção de vivacidade facial ou *face anti-spoofing* é o problema de reconhecer ataques como estes. O problema de detecção de vivacidade facial e sua fundamentação teórica são apresentados, seguidos de um estudo de trabalhos relacionados. Em seguida uma abordagem baseada no uso de profundidade estimada pela rede Pix2Pix para a detecção de vivacidade facial em imagem é apresentada, implementada e finalmente testada em protocolos relevantes para a avaliação em comparação com o estado da arte.

Palavras-chave: Detecção de vivacidade facial. Face Anti-Spoofing. Detecção de Ataque de Apresentação.

ABSTRACT

Face spoofing consists of simulating a person's facial biometric traits in order to impersonate them in a face recognition system (for example, in digital payment and social media applications). Face liveness detection or face anti-spoofing is the problem of recognizing such attacks. The problem of face liveness detection and its theoretical foundation are presented, followed by a study of related work. Afterwards an approach based on using Pix2Pix-estimated depth information for image face liveness detection is presented, implemented and finally tested on relevant protocols for evaluation in comparison with the state of the art.

Keywords: Face Liveness Detection. Face Anti-Spoofing. Presentation Attack Detection.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ACRÔNIMOS

APCER	Attack Presentation Classification Error Rate
BPCER	Bonafide Presentation Classification Error Rate
FAR	False Acceptance Rate
FRR	False Rejection Rate
HTER	Half Total Error Rate
EER	Equal Error Rate
AUC	Area Under the Curve
FAS	Face Anti-Spoofing
PAD	Presentation Attack Detection
GAN	Generative Adversarial Network
cGAN	Conditional Generative Adversarial Network
ResNet	Residual Network
ReLU	Rectified Linear Unit
BN	Batch Normalization
RGB	Red, Green and Blue
SVM	Support Vector Machine
IDA	Image Distortion Analysis
CNN	Convolutional Neural Network
PCA	Principal Component Analysis
rPPG	REmote Photoplethysmography
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory Network
SWIR	Shortwave Infrared
SGD	Stochastic Gradient Descent

SUMÁRIO

1 INTRODUCTION

In recent years, the use and capacity of mobile phones have grown so as to make these devices an important part of everyday life. This growth was accompanied by a shift in interactions; many activites that could only be done in a computer are now made available in the user's pocket. Important applications followed this tendency, including banking (from internet banking to mobile applications) and documentation (from government-issued physical documents to verified applications). This was only possible due to accurate verification capability, both in hardware and software, with algorithms such as those of face recognition.

There are, however, strategies for misleading these verification systems, which become easier as the verification becomes decentralized (i.e., the user is not required to be in a physical installation of a company or institution to have their identity verified). For face recognition, a range of spoofing strategies have been developed and are increasingly honed for improved attacks of impersonation: for an example, a malicious invader A could pretend to be some other person B by placing a photo of B in front of their phone's camera in the face verification step of an application's authentication process - the algorithm would recognize the face and let the attacker in. Naturally, access to privileged information and resources should be reserved to privileged users only, and the impact of vulnerabilities to spoofing attacks such as this implies in a fragile structure for very essential information exchanges in modern days.

Attacks are the most varied, exploiting weaknesses even in the model's learning process (such as bias and limited representation). In this context, it is fundamental that the task of face anti-spoofing (FAS), also named presentation attack detection (PAD) and liveness detection (here, a face's image is said to be *live* if it is not a spoof), is further researched, so as to enhance current models' performance.

State-of-the-art FAS models often exploit secondary domain information, such as face depth maps, due to their discriminative advantage over image-domain representations. Face depth maps consist of 3D masks representing a person's face shape. While for bonafide images they maintain this aspect, for spoof images (since there is no face depth) the depth map is empty (a black image). Figure 1.1 exemplifies face depth in the CASIA-FASD (Zhang et al., 2012) dataset for samples of both classes.

Figure 1.1 illustrates two images: one bonafide (real, genuine) and one spoof.

Figura 1.1: Examples of bonafide (left) and attack (right) images from the CASIA-FASD dataset (Zhang et al., 2012)

1.1 MOTIVATION

Any system that uses face recognition is subject to face spoofing attacks and requires a face anti-spoofing strategy for assurance of its proper functionality and security. Examples of such systems are digital payment and social media applications.

At the same time, as prevention methods are improved, attack strategies are also further developed. This adversarial scenario creates the need for continuous development of models that can effectively detects spoofing attempts.

This work's motivation is to, in its approach and explorations, improve the general understanding of how security can be increased in the field of presentation attack detection.

1.2 CHALLENGES

In all its importance this task remains a difficult one, particularly due to variety in imaging conditions (variations in light, contrast and color) and attack types among the train and test phases (or train phase and real-world use). The principal aim of research in this field is to improve model generalization capabilities with respect to this variety.

For this reason most evaluation protocols rely on cross-dataset scenarios or divide the data in contrasting subsets - one example to this is the OULU-NPU dataset (Boulkenafet et al., 2017), which proposes four intra-dataset protocols for challenging imaging variety among the train and test sets.

1.3 FACE FORGERY

Another problem that has gotten growing attention over recent years is the one of face forgery detection, which consists of recognizing that an image has been digitally manipulated (e.g., to put someone's face in it). Face forgeries are often used for the same purposes as face spoofings, and therefore it could be expected of this work to be at least partially concerned with this additional problem. Face forgery detection is, however, not as important a task as face anti-spoofing because the way forgeries are presented often share characteristics with spoofs (for example, showing a forgery to a camera by holding a screen that displays it - in this case, a face anti-spoofing algorithm should detect this presentation as a display attack) or exploit other security weaknessess of a system, such as injecting the forgery as if it was coming directly from the acquisition device - then it can be argued that it would not be a matter of face anti-spoofing, but of the hardware or software system's integrity.

All of that depends, of course, on how a face recognition system operates. If the face image is captured in a controlled environment, such as a mobile phone app, then forgeries would fall into one of the two described cases; if, however, it is uploaded by a user in a browser application, a forgery is simply another uploaded image and therefore forgery detection should be used to prevent these attacks. Face forgery does not fit the scope of this work.

1.4 PROPOSED APPROACH

A common path for improving a model's performance in FAS is to provide it with additional information about the input image, i.e., to use both the RGB input and its correspondent in another domain where discrimination is easier. One example of such a domain is face depth, where spoofs are modeled to have none and bonafides are modeled to have a deep face mask. Figure 1.2 illustrates how genuine and spoof images differentiate in the depth domain.

Previous works report good results when using face depth information for presentation attack detection (Atoum et al., 2017; Liu et al., 2018; Shao et al., 2019; Wang et al., 2019, 2020c; Zhang et al., 2020; Zheng et al., 2021; Wang et al., 2021b), as explored in Chapter 3.

Figura 1.2: Examples face depth maps of bonafide (left) and attack (right) images from the CASIA-FASD dataset (Zhang et al., 2012). The depth of most attack types will be completely empty because the image consists of a plain surface and not an actual face.

The approach proposed in this work is to use a a deep learning-based strategy for detecting face liveness in images by processing information in the image (RGB) and depth domains, with the particular condition of the depth domain information being obtained with the Pix2Pix network (Isola et al., 2017), which has previously shown high effectiveness in qualitative evaluation of image domain translation performance in many instances of the problem (this network is further explained in Chapter 2).

1.5 OBJECTIVES

This work aims to answer four main questions regarding the proposed approach. The context for each question is further developed in the next chapters, and Chapters 5 and 6 discuss conclusions.

The first question is whether and how auxiliary depth information enhances a classifier's performance, i.e., if and how much a model's prediction and generalization capabilities are improved once depth information is provided to it.

The second relates to how an input's RGB image and corresponding depth map should be given to input as the classifier network, a procedure that this work refers to as the *fusion method* (see Chapter 4). The two considered procedures are concatenation and blending.

Since the proposed approach is classifier model-agnostic, it is easily modifiable to use different backbones for classification. The effectiveness of the recent ConvNeXt family (Liu et al., 2022) is evaluated for this task both with and without auxiliary depth information.

The final question is that of the effectiveness of the Pix2Pix network for face depth estimation in the context of face anti-spoofing, considering that the depth maps are very different between spoofs and genuine images. This is the main consideration in this work.

1.6 CONTRIBUTIONS

This work successfully explores all the posed questions and provides answers for them with the support of experiments on different intra- and cross-dataset protocols from the CASIA-FASD (Zhang et al., 2012) and Replay-Attack (Chingovska et al., 2012) datasets. Both the performed experiments and the subsequent discussions are available in Chapter 5. This allows for reinforced understanding of auxiliary domain information effects on model performance in two aspects: prediction performance and generalization capability.

1.7 OUTLINE

The chapters in this work are organized as follows. Chapter 2 describes fundamental concepts and metrics related to liveness detection. Chapter 3 presents a study of related works. Chapter 4 presents this work's proposed approach, which is evaluated as described and presented in Chapter 5. Limitations and possibilites for future work are finally discussed in Chapter 6.

2 THEORETICAL BACKGROUND

Relevant concepts and metrics related to the field of face liveness detection and networks used in Chapters 4 and 5 are now presented. A foundational understanding of machine learning is expected from the reader.

2.1 CONCEPTS

A **presentation** is the capture of a face through an adequate acquisition device (i.e., a picture of the user's face). It can be an attack or a bonafide (genuine presentation).

The task of liveness detection is framed as a binary classification problem where attacks and bonafides belong to the positive and negative classes, respectively.

2.2 METRICS

The **Attack Presentation Classification Error Rate (APCER)** represents the proportion of attack presentations incorrectly classified as bonafide, that is, the false negative (FN) rate among false negatives and true positives (TP), i.e.,

$$APCER = \frac{FN}{FN + TP}. \tag{2.1}$$

Similarly, the **Bonafide Presentation Classification Error Rate** represents the proportion of bonafide presentations incorrectly classified as attacks, that is, the false positive (FP) rate among false positives and true negatives (TN), i.e.,

$$BPCER = \frac{FP}{FP + TN}. \tag{2.2}$$

The **Average Classification Error Rate** is the average of the $APCER$ and $BPCER$ metrics, i.e.,

$$ACER = \frac{APCER + BPCER}{2}. \tag{2.3}$$

The **Half Total Error Rate** is the average between the false acceptance rate (FAR, $\frac{FN}{FN+TP}$) and false rejection rate (FRR, $\frac{FP}{FP+TN}$), i.e.,

$$HTER = \frac{FAR + FRR}{2}. \tag{2.4}$$

2.3 NETWORKS

This section presents network architectures that are applied to this work's approach in later chapters. The first, Pix2Pix, will be used for depth estimation (i.e., translation from the image domain to the depth domain), while the rest (ResNet, ConvNeXt and FeatEmbedder) is used for liveness classification.

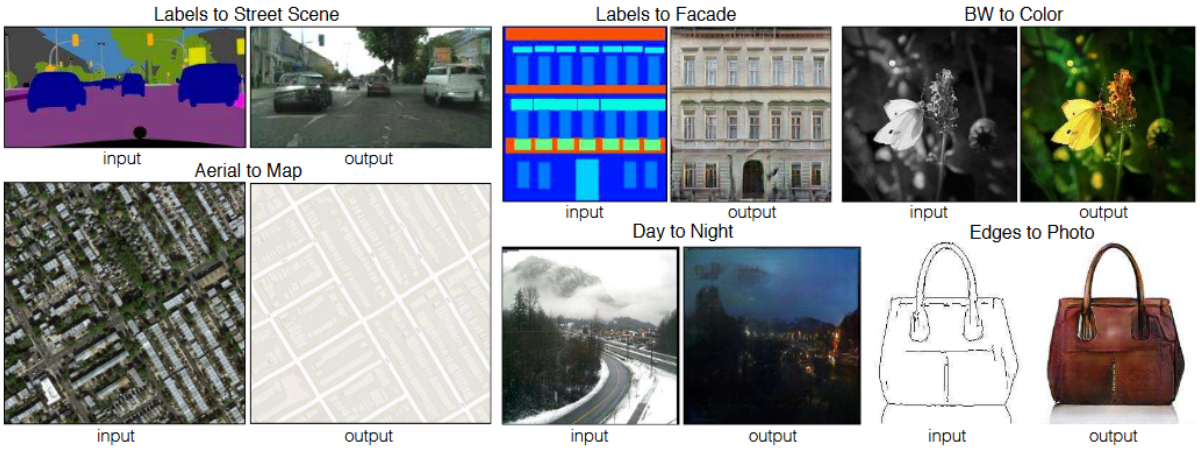
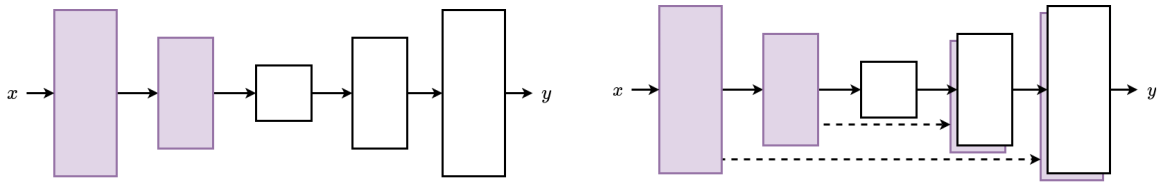


Figura 2.1: Image-to-image translation examples. Source: Isola et al. (2017).

2.3.1 Pix2Pix

The Pix2Pix network, introduced by Isola et al. (2017), is a conditional generative adversarial network (GAN) where the generator is an encoder-decoder model with skip connections between symmetrical layers (a U-Net) and the discriminator considers only one patch of the image at a time (the discriminator is run convolutionally across the image and the final output is an average of all its results). Figure 2.1 shows instances of the image-to-image translation problem where this network has shown good results. Figure 2.2 illustrates how the U-Net generator network compares to classic generator encoder-decoder networks: the encoder-decoder networks have no concatenation connections between layers.



(a) Encoder-Decoder network

(b) U-Net network

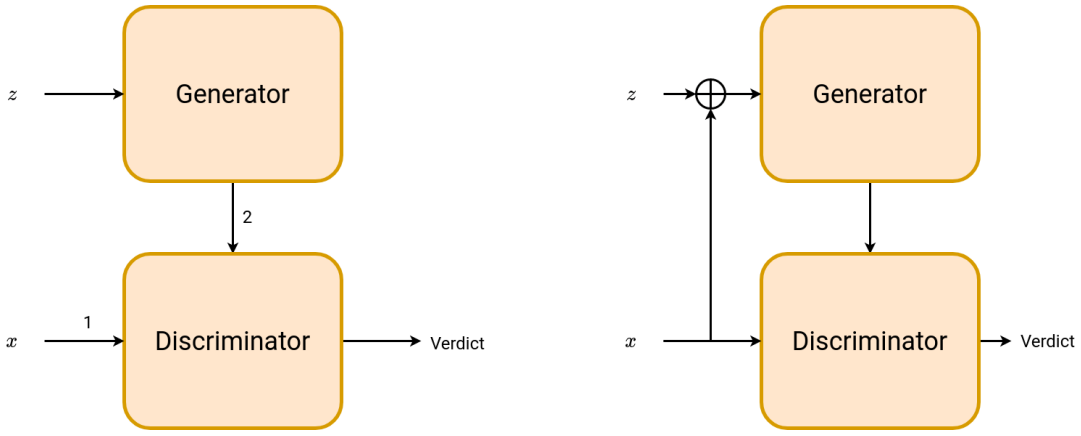
Figura 2.2: Possibilities for generator networks. Pix2Pix uses U-Net. Based on Figure 3 of Isola et al. (2017).

Figure 2.3 illustrates the main difference between conditional and unconditional GANs. While in a traditional (unconditional) GAN the generator only receives as input a noise vector z and the discriminator receives as input either the generator's output $G(z)$ or a real sample x, in a conditional GAN both the generator and the discriminator networks have access to a conditioning input that serves as control to the desired output, that is, the generator receives as input both z and x and the discriminator receives as input both $G(z,x)$ and x.

2.3.2 Backbone Architectures

For backbone classifier networks, two families and one individual model have been chosen. The two families are ResNet and ConvNeXt; the individual model, FeatEmbedder, is the same classifier network as in a similar work (Wang et al., 2021b) which is described in Chapter 3.

Residual Networks, or ResNets (He et al., 2015), compose a classic architecture for deep learning in computer vision. Their main contribution is the introduction of learning residual functions which reference layer inputs instead of learning unreferenced functions. This intuitively

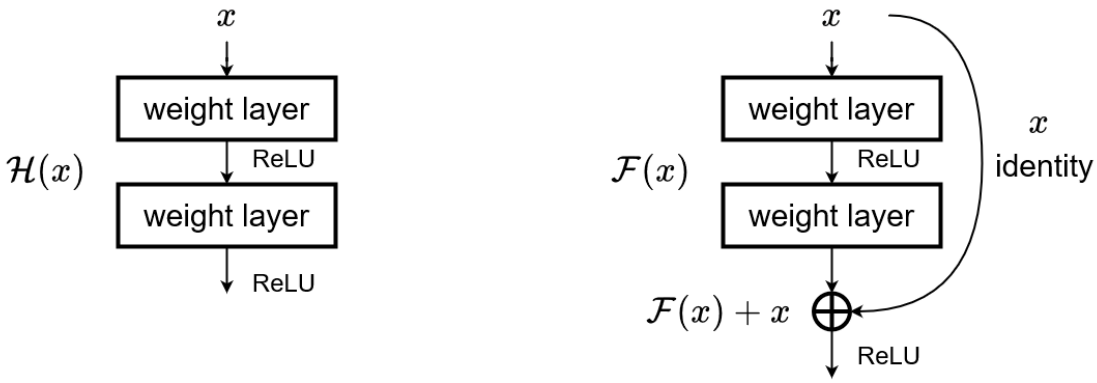


(a) GAN architecture

(b) Conditional GAN (cGAN) architecture

Figura 2.3: Comparison of GANs and cGANS

allows network layers to form modularized understanding of images and has been shown to improve optimization and accuracy. Figure 2.4 demonstrates how a residual block works in comparison to non-residual blocks: the residual connection allows the block to learn in an independent manner.



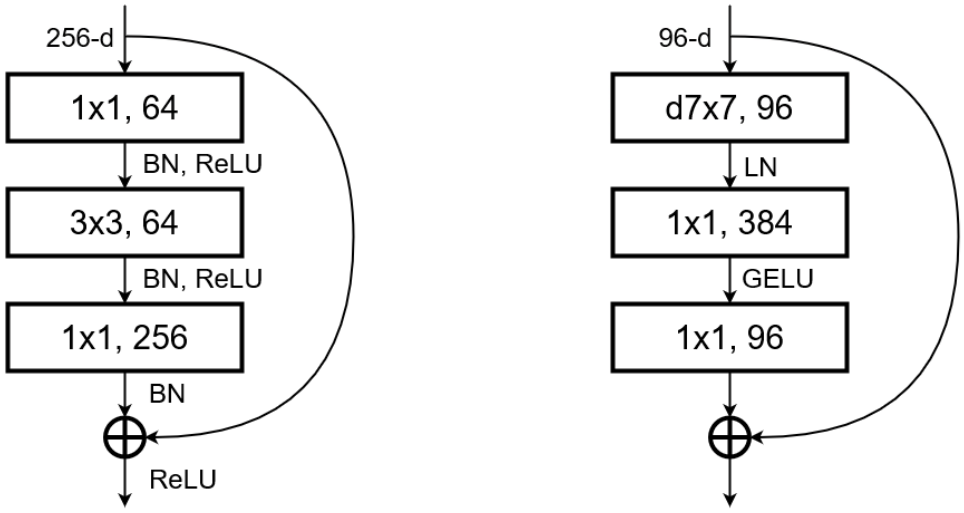
(a) Block with no residual connections

(b) Residual block

Figura 2.4: ResNet block comparison. Based on Figure 2 of He et al. (2015).

Over the years since ResNets, Transformer networks - which come from the field of Natural Language Processing, based on attention function as opposed to convolutions - have shown great performance, often superior to convolutional networks, in a range of computer vision tasks. This success is mostly attributed to the transformer aspect of these networks, i.e., attention functions are considered better suited for these tasks. Liu et al. (2022) argue, however, that this success is due to other design decisions - many that could be applied to convolutional neural networks as well. With that in mind, the authors of this paper develop ConvNeXt, a family of purely convolutional networks based on ResNets that follow the design of Swin Transformers. This covers different characteristics of the network, from the training procedure to kernel sizes. These detail interventions render the authors state-of-the-art results in different computer vision tasks. Figure 2.5 compares ResNet blocks to ConvNeXt blocks.

Wang et al. (2021b) use as a classifier network the model to be in this work denominated as FeatEmbedder. The authors do not discuss the reasoning behind the model choice, and it is



(a) Block with no residual connections

(b) Residual block

Figura 2.5: Comparison between ResNet and ConvNeXt blocks. Based on Figure 4 of Liu et al. (2022).

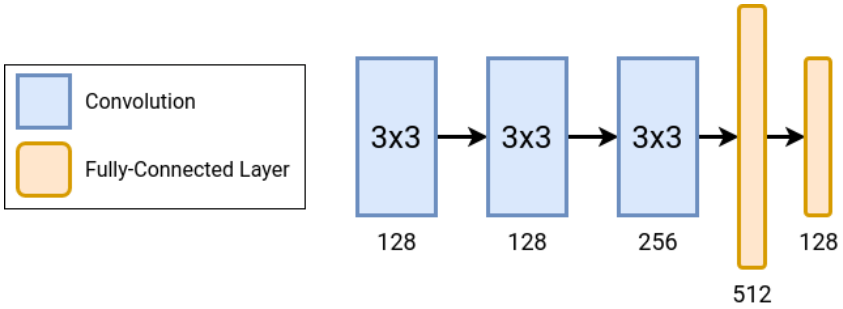


Figura 2.6: FeatEmbedder architecture. Number below layer indicate input resolution, and all layers are followed by batch normalization and ReLU steps. For more details, please refer to the official implementation (Wang et al., 2020b).

made available in their public GitHub repository (Wang et al., 2020b). Figure 2.6 illustrates its architecture.

3 RELATED WORK

This chapter explains all the datasets and methods proposed in studied works. Current challenges in the field are discussed, as are the difficulties with datasets.

3.1 DATASETS

With the growth in the field of face liveness detection research, there have been many different approaches to dataset collection and what matters in a dataset. For some works, for example, multiple image modes are important, as many methods take advantage of information external to what an RGB image can entail. Currently, there is a focus on sample abundance and variability in acquisition conditions (such as subject movement, camera quality and environmental light), subject characteristics (gender, race, etc) and attack methods. Table 3.1 summarizes the studied datasets.

Name	Year	Citation	Samples	Subjects	Attack types	Additional description
NUAA	2010	Tan et al. (2010)	5105 real, 7509 fake	15	1	
PRINT-ATTACK	2011	Anjos e Marcel (2011)	200 real, 200 fake	50	1	
CASIA	2012	Zhang et al. (2012)	150 real, 450 fake	50	3	
Replay-Attack	2012	Chingovska et al. (2012)	200 real, 1000 fake	50	3	Different recording conditions
MSU-MFSD	2015	Wen et al. (2015)	110 real, 330 fake	55	3	
MSU-USSA	2016	Patel et al. (2016)	1140 real, 9120 fake	1140	2	Multiple devices for replay spoofing;
MLFP	2017	Agarwal et al. (2017)	150 real, 1200 fake	10	2	Visible, near infra-red and thermal modes for each sample
Oulu-NPU	2017	Boulkenafet et al. (2017)	990 real, 3960 fake	55	4	Varied environments
SiW	2018	Liu et al. (2018)	1320 real, 3300 fake	165	6	Varying subject movements, camera angles and facial expressions
ROSE-Youtu	2018	Li et al. (2018)	3350 in total	20	3	
SiW-M	2019	Liu et al. (2019c)	660 real, 968 fake	493	13	Different scenarios (varying movement, light, camera quality, distance to camera)
HQ-WMCA	2020	Mostaani et al. (2020)	555 real, 2349 fake	51	10	Focus on varied attacks and a multi-modal character
DMAD	2020	Wang et al. (2020c)	900 real, 1800 fake	300	6	

Tabela 3.1: Studied datasets' main characteristics.

This work's evaluation protocols rely on the CASIA-FASD and Replay-Attack datasets. For the next section, all acronyms for datasets and challenges mentioned in the results of methods that have been studied are described in Table 3.2.

Acronym	**Dataset or Challenge**	**Contemplated by this study**
E	NenuLD	No
H	In-house	Does not apply
L	ChaLearn Face Anti-Spoofing Challenge	Yes
R	ROSE-Youtu	Yes
A	SMAD	No
C	CASIA-FASD	Yes
C_M	CASIA-MFSD	No
C_S	CASIA-SURF	No
D	DMAD	Yes
I	Replay-Attack	Yes
M	MSU-MFSD	Yes
N	NUAA	Yes
S	SiW	Yes
S_M	SiW-M	Yes
W	HQ-WMCA	Yes
Y	The Extended Yale Face Database B	No

Tabela 3.2: Datasets and Challenges mentioned in studied methods' results.

3.2 METHODS FOR FACE ANTI-SPOOFING

Tan et al. (2010) implement a lambertian model with two different strategies for obtaining latent features and two extensions to a sparse logistic regression model that make it faster and more accurate.

Chingovska et al. (2012), alongside presenting a dataset that remains relevant to this day, study the effectiveness of using texture features based on Local Binary Patterns and their variations on classification. The reported results present no consistency in regards to types of attacks or cross-database scenarios.

Wen et al. (2015) propose to approach detectyion with an ensemble of two Support Vector Machine (SVM) classifiers trained for different attacks and fed an Image Distortion Analysis (IDA) feature vector extracted from specular reflection, blurriness, chromatic moment and color intensity information. The authors list as future work the development of features tailored to specific use cases so as to facilitate the task.

Yang et al. (2015) also work on tailored information, introducing a strategy based on identifying each individual in a system to use spoofing classifier specialized on the individuals. This approach does not escalate.

Patel et al. (2016) build on top of the IDA-based strategy by using different intensity channels, image regions and feature descriptors for extracting information on surface reflection, moire patterns, color distortion and shape deformation. One proposed idea for future work is to combine this method with movement cues, such as eye blinking, to improve detection performance.

Li et al. (2016) extract features for detection with a CNN. The extracted features go through a Principal Component Analysis (PCA) process in order to prevent overfitting and reduce dimensionality, and are then finally fed into an SVM for classification. There could be room for improvement in using a secondary neural network in the intermediary process instead of applying the traditional PCA method.

Boulkenafet et al. (2016) recognize that using luminance information to decide liveness is a relevant strategy and then propose to use color textures instead. Some common approaches for improved performance are not implemented and listed as future work, such as normalizing the input to the face's bounding box. Furthermore, the authors incentive the study of whether different color aspects prove useful in detecting different types of attacks.

Killioğlu et al. (2017) use eye (pupil) movement detection for determining whether a sample is real or not. This approach is shown to perform worse on glass-wearing subjects, and would not work against certain present-day masked attacks.

Atoum et al. (2017) introduce the first approach studied that uses depth estimation for detecting liveness. Their method consists of a two-stream CNN, where one stream extracts local features while the second obtains holistic depth (i.e., checks whether the image has a face-like depth pattern). This approach achieved state-of-the-art results and the authors suggested as future work a deeper study of possible stream fusion strategies.

Chan et al. (2018) propose using as input to the model two images: one taken with a flash (weaker than a comercial camera flash) and one without. This would make it easier to handle low light and noise scenarios and would enhance differences between attackers and legitimate users. The proposed method uses descriptors to capture texture and structure from both images. The approach of taking two pictures in different conditions, however, relies on specific hardware and would not apply to many scenarios where face liveness detection is relevant.

Ito et al. (2017) present a case study of face liveness detection, showcasing an approach that consists of classifying with an SVM from CNN-generated features. A possible enhancement to the method would be to further improve the CNN architecture to skip the SVM.

Li et al. (2018) propose an unsupervised domain adaptation scheme to learn the classifier for the target domain where an embedding function maps the source and target domains to a space where distribution similarity can be measured and optimized. The authors list as future work the application of this strategy to zero-shot scenarios.

Liu et al. (2018) reach state-of-the-art results with a model that consists of two parts, namely a CNN for estimating face depth (with pixel-wise supervision) and an RNN for estimating rPPG (remote photoplethysmography) signals (sequence-wise supervision). These two parts are fused for liveness detection.

Singh e Arora (2018) consider the average intensity between indicators of eye blink sequence and lip and chin movement, classifying liveness by comparing the calculated average to threshold values. A possible direction for improvement would be to feed a neural network with the intensity sequences instead of calculating average values.

Jourabloo et al. (2018) introduce a strategy of de-spoofing for liveness classification. More specifically, the presented model uses a CNN for dividing an input image into live content and spoof noise (which is modeled via the loss function to be zero in the case of live samples). Achieved results are on level with state of the art, though this model remains vulnerable to low-resolution images.

Luo et al. (2018) argue that single-scale schemes (i.e., those where the input image is cropped to the face bounding box before being fed to the model) exclude possibly valuable information present in the background, and propose instead to use multiple bounding boxes for cropping, forming a sequence of different scales from a single image which can then be used with LSTMs for feature generation and consequent classification. Possibilities for future work include a learnable decision of scales for sequence generation, which could enhance background information capture.

Liu et al. (2019c) approach the problem with a deep tree network, partitioning samples into semantic subgroups in an unsupervised manner. This enable the decision making process to be specialized, as the classification is done based only on similar attacks. Reported results are on par with state of the art.

Chen et al. (2020) argue that many existing methods are hindered by illumination variation, and so propose an illumination-invariant method based on a two-stresm convolutional neural network which works on two complementary spaces: the original imaging space (RGB), with detailed texture at the cost of high light sensitivity, and the illumination-invariant multi-scale retnex space (MSR), with lower light sensitivity and face information. The MSR images offer discriminative information for face spoofing detection, and the two network streams are fused with an attention-based method. Reported results achieve state-of-the-art levels.

Liu et al. (2019b) take advantage of a Microsoft Kinect device as input for a liveness classification network, which does not fit into this work's scope but demonstrates a clever use of the technology.

Liu et al. (2019a) report the state of the art through the lens of a face liveness detection challenge (built around the CASIA-SURF dataset) and its results. The first Observation is that the problem remains challenging due to lack of generalization in the following aspects: for intra-dataset testing, there is often still a performance gap between testing and validation, and for cross-dataset scenarios, existing methods often rely too heavily on known data and may be found fragile once confronted with unknown acquisition devices, attack methods and spoofing mediums. Another issue is that the ubiquitous use of softmax loss might lead models to value arbitrary cues which are not actually indicators of spoofing - when such cues disappear during testing, models fail to generalize. The authors also argue that supervision should be designed from essential differences between live and spoof faces, such as rPPG signals (as they can reflect human physiological signs), depth imaging, light reflection and (in the case of videos) inter-frame variations. Finally, the authors warn readers about less recent models' incapability to handle new 3D masks in particular.

Yang et al. (2019) use LSTM networks for classification and argue that capturing a large dataset is an easier task in the case of face spoofing than often regarded, obtaining data from a video social network. The legal aspect of this collection is not discussed.

Shen et al. (2019) present a multi-stream, bag-of-local-features-based CNN trained on the CASIA-SURF dataset that renders good results on the ChaLearn challenge.

Shao et al. (2019) propose to learn a feature space shared between domains through multi-adversarial discriminative domain generalization with auxiliary depth. The adversarial scheme consists of a generator producing domain-shared features and multiple domain discriminators, so that learned features are domain-wise indistinguishable. Reported results are state-of-the-art.

Wang et al. (2019) present a multi-modal approach where four branches (namely RGB, depth, infrared and a concatenation of all three) are fused with a spatial and channel attention module. Reported results are competitive with the state of the art.

Koshy e Mahmood (2019) combine texture analysis with convolutional neural networks, where a nonlinear-diffused image is fed into a CNN. This work shows that the smoothness of a diffused image can be an important factor in determining an image's liveness.

Li et al. (2019) observe that, when it comes to replay attacks, motion blur analysis is an useful aspect for classification, as blur width and intensity variation are different in fake and real input samples. The classification strategy the authors present is essentially to extract blur intensity and with features with a convolutional neural network and a local similar pattern method, respectively, and then fuse those features for detection. Reported results are competitive with the state of the art, and suggested improvements include exploring better feature fusion strategies.

Wang et al. (2020c) seek to mix depth and movement in pattern recognition, with a residual spatial gradient block for detecting discriminative details, a spatio-temporal propagation module for encoding spatio-temporal information, and a contrastive depth loss for improved depth-supervised attack detection generality. This work also demonstrate the effectiveness of using depth maps in this task. Achieved results are state-of-the-art.

Wang et al. (2020a) present a strategy based on two steps, namely a disentangled representation learning one and multi-domain learning one (which starts off with the first step's output as its input). This strategy renders the authors state-of-the-art results.

Garg et al. (2020) propose the use of a deep belief network for liveness classification. The results are not, however, presented on the most relevant datasets of the time of publication.

Heusch et al. (2020) explore the use of SWIR (shortwave infrared) as input to recent CNN-based models, where a difficulty for generalization is found. The authors suggest two possible reasons: the usage of different wavelengths in the SWIR between datasets and the difference in image quality between datasets. Something to note is that only two datasets were used, as SWIR input is not as widespread a capture modality as others.

Zhang et al. (2020) introduce an adversarial scheme with auxiliary texture and depth supervision where the learning process is enhanced by mixing samples' disentangled liveness and content features. Reported results are competitive with the state of the art.

Deb e Jain (2021) use a self-supervised regional fully convolutional network trained to learn local discriminative cues from the input's facial regions. Reported results are comparable to the state of the art. The authors suggest that this approach is severly hindered by small amounts of data and low-resolution data.

Yu et al. (2020) apply a central difference convolutional network to the problem, improving it with a neural architecture search and multiscale attention fusion module. The buse of central difference convolutions render the authors results competitive with the state of the art.

Jia et al. (2020) develop an end-to-end single-side domain generalization framework where the learned generalized feature space is such that real samples are grouped regardless of

their domain and spoof samples are grouped by domain (and distant from the real region). This is enabled by the employment of an asymmetric triplet mining strategy. Reported results are state-of-the-art, and the authors suggest as future work the use of asymmetric design for dividing fake faces according to attack types rather than databases.

Liu et al. (2021b) argue that current work does not generalize well when there is a variety of presentation attacks because they cannot extract features well enough and thus propose a multi-modality data-based two-stage cascade framework that can selectively fuse low- and high-level features from different modalities to improve feature representation. This idea renders the authors state-of-the-art results.

Zheng et al. (2021) use a two-stream spatial-temporal network to explore both depth and multi-scale information with a temporal shift module, which can extract temporal information without additional calculations by separating data movement and calculation in convolutions, and a depth information estimation network, which contains an attention module based on estimated depth and fusions multi-scale features from both streams. Reported results are competitive with the state of the art.

Liu et al. (2021a) recognize that many modalities usually help classification yet are not very common as input, and so introduce an RGB-based strategy that can be extended with other modalities during testing by translating their content.

Chen et al. (2021) seek to generalize accross different acquistion devices with a two-branch scheme. The first branch extracts camera-invariant spoofing features from a high-frequency domain and the second branch extracts both low- and high-frequency features from an enhanced image. Reported results are competitive with the state of the art.

Wang et al. (2021b) generate depth maps in an adversarial scheme to then feed a classifier network with the source RGB input and intermediate features from the adversarial scheme's generator network. Reported results are competitive with the state of the art. This work also demonstrates how well-separated depth maps are when compared to source images.

Purnapatra et al. (2021) report on the Face Liveness Detection Competition, 2021 edition. This competition was open to both academia and industry and focused on generalizability of models. General performance of competitors was worse than in other competitions, which is attributed by the authors to increased complexity (i.e., more attack types and instruments) and the absence of a training dataset (competitors used whatever training resources were available).

Sanghvi et al. (2021) employ three subnetworks to detect different types of attacks (print, replay and mask attacks). This approach allows the reporting of type of attack detected without computational overhead. Possibilites for future work include exploring different architectures for each subnetwork and applying the strategy to other biometric attack detection tasks.

George e Marcel (2021) study the effectiveness of vision transformers for zero-shot face anti-spoofing, fine-tuning a source vision transformer for transfer learning. This very simple intervention allows for results that are competitive with the state of the art.

Quan et al. (2021) present a semi-supervised scheme that requires few training samples (about 50, as the authors describe it). During training, the model progessively adopts unlabeled data with reliable pseudo labels, exploiting the temporal consistency in videos to make this easier (i.e., it is easier to determine the liveness of a frame when the liveness of a neighbour frame is already known). Reported results are competitive with the state of the art.

Wang et al. (2021a) apply an unsupervised model adaptor scheme to adapt the trained model to new domains, seeking an improvement in domain adaptation and generalization. This renders competitive results.

Wang et al. (2022) split the complete image representation into content and style representations, which are obtained in an adversarial scheme. Furthermore, the style features are

refined through contrastive learning, which allows for clusterizing classes regardless of domain. Reported results are highly competitive with the state of the art.

Table 3.3 presents an overview of results for the studied methods.

Tabela 3.3: Related works results summary. Datasets are mentioned by their acronyms as presented in Table 3.2, and cross-dataset results are indicated by * superscripts. Numbers after acronyms indicate protocols.

Year	Work	Dataset	Results
2012	Chingovska et al. (2012)	I	13.87% HTER
		N	13.17% HTER
		C	18.21% HTER
2015	Wen et al. (2015)	I	7.41% HTER
		C	13.3% EER
		M	8.58% EER
	Yang et al. (2015)	C	4.95% EER
		I	2.51% HTER
2016	Patel et al. (2016)	*M	9.27% HTER
		*I	3.5% HTER
		*C	2% HTER
	Li et al. (2016)	I	2.9% EER, 6.1% HTER
		C	4.5% EER
	Boulkenafet et al. (2016)	I	0.4% EER, 2.8% HTER
		C	2.1% EER
		M	4.9% EER
		*I	30.3% Average HTER
		*C	37.7% Average HTER
		*M	33.9% Average HTER
2017	Killioğlu et al. (2017)	Y	89.7% Accuracy
	Atoum et al. (2017)	I	0.1% EER, 0.72% HTER
		C	2.67% EER, 0% HTER
		M	0.35% EER, 0.21% HTER
	Chan et al. (2018)	H	1.17% Average HTER
	Ito et al. (2017)	I	2.4% EER
2018	Li et al. (2018)	I	1.4% HTER
		C	5.5% EER
		M	5.8% EER
		R	8% EER
	Liu et al. (2018)	O1	1.6% APCER, 1.6% BPCER, 1.6% ACER
		O2	2.7% APCER, 2.7% BPCER, 2.7% ACER
		O3	2.7% APCER, 3.1% BPCER, 2.9% ACER
		O4	9.3% APCER, 10.4% BPCER, 9.5% ACER
		S1	3.58% APCER, 3.58% BPCER, 3.58% ACER

Continued on next page

Tabela 3.3 – continued from previous page

Year	Work	Dataset	Results
2018		S2	0.57% APCER, 0.57% BPCER, 0.57% ACER
		S3	8.31% APCER, 8.31% BPCER, 8.31% ACER
		*I	27.6%
		*C_M	28.4%
	Luo et al. (2018)	C	2.4% EER
		I	0.02% EER, 0.39% HTER
	Jourabloo et al. (2018)	I	28.5% HTER
		C_M	41.1% HTER
		O1	1.2% APCER, 1.7% BPCER, 1.5% ACER
		O2	4.2% APCER, 4.4% BPCER, 4.3% ACER
		O3	4% APCER, 3.8% BPCER, 3.6% ACER
		O4	5.1% APCER, 6.1% BPCER, 5.6% ACER
	Singh e Arora (2018)	H	99.41% accuracy
2019	Liu et al. (2019c)	C, M, I	95.9% AUC
		S_M	17.1% APCER, 16.6% BPCER, 16.8% ACER, 16.1% EER
	Liu et al. (2019b)	E	99.8% Accuracy
	Chen et al. (2020)	*C	33.4% HTER
		*I	30% HTER
		O1	5.1% APCER, 6.7% BPCER, 5.9% ACER
		O2	7.6% APCER, 2.2% BPCER, 4.9% ACER
		O3	3.9% APCER, 7.3% BPCER, 5.6% ACER
		O4	11.3% APCER, 9.7% BPCER, 9.8% ACER
	Yang et al. (2019)	S1	1% ACER
		S2	0.28% ACER
		S3	12.10% ACER
		O1	1.2% APCER, 2.5% BPCER, 1.9% ACER
		O2	4.2% APCER, 0.3% BPCER, 2.2% ACER
		O3	4.7% APCER, 0.9% BPCER, 2.8% ACER
		O4	6.7% APCER, 8.3% BPCER, 7.5% ACER
		*I	18.7% HTER
		*C_M	25% HTER
	Shen et al. (2019)	L	99.8% TPR@FPR=10e-4
	Shao et al. (2019)	*M	17.69% HTER, 88.06% AUC
		*C	24.5% HTER, 84.51% AUC
		*I	22.19% HTER, 84.99% AUC
		*O	27.98% HTER, 80.02% AUC
	Wang et al. (2019)	C_S	0.2% APCER, 0.3% NPCER, 0.2% ACER
	Koshy e Mahmood (2019)	N	100% Accuracy
	Li et al. (2019)	I	0% APCER, 0% BPCER
		O	5.3% APCER, 4.7% BPCER

Continued on next page

Tabela 3.3 – continued from previous page

Year	Work	Dataset	Results
2020	Wang et al. (2020c)	O1	2% APCER, 0% BPCER, 1% ACER
		O2	2.5% APCER, 1.3% BPCER, 1.9% ACER
		O3	3.2% APCER, 2.2% BPCER, 2.7% ACER
		O4	6.7% APCER, 3.3% BPCER, 5% ACER
		S1	0.4% ACER
		S2	0.02% ACER
		S3	2.78% ACER
		D	4.55% ACER
		I	17% HTER
		C_M	22.8% HTER
	Wang et al. (2020a)	*M	90.1% AUC, 17.02% HTER
		*C	87.43% AUC, 19.68% HTER
		*I	86.72% AUC, 20.87% HTER
		*O	81.47% AUC, 25.02% HTER
	Garg et al. (2020)	N	99% Accuracy
	Zhang et al. (2020)	O1	1.7% APCER, 0.8% BPCER, 1.3% ACER
		O2	1.1% APCER, 3.6% BPCER, 2.4% ACER
		O3	2.8% APCER, 1.7% BPCER, 2.2% ACER
		O4	5.4% APCER, 3.3% BPCER, 4.4% ACER
		S1	0.28% ACER
		S2	0.1% ACER
		S3	5.59% ACER
		*C	30.3% HTER
		*I	22.4% HTER
	Deb e Jain (2021)	*I	19.9% HTER
		*C	41.9% HTER
		O1	1.5% APCER, 7.7% BPCER, 4.6% ACER
		O2	3.1% APCER, 3.7% BPCER, 3.4% ACER
		O3	2.9% APCER, 2.7% BPCER, 2.8% ACER
		O4	8.3% APCER, 13.3% BPCER, 10.8% ACER
	Yu et al. (2020)	O1	0.4% APCER, 0% BPCER, 0.2% ACER
		O2	1.8% APCER, 0.8% BPCER, 1.3% ACER
		O3	1.7% APCER, 2% BPCER, 1.8% ACER
		O4	4.2% APCER, 5.8% BPCER, 5% ACER
		S1	0.12% ACER
		S2	0.04% ACER
		S3	1.9% ACER
		*I	6.5% HTER
		*C_M	29.8% HTER
	Jia et al. (2020)	*C	10.44% HTER, 95.94% AUC
		*O	15.61% HTER, 91.54% AUC
		*M	7.38% HTER, 97.17% AUC
		*I	11.71% HTER, 96.59% AUC
2021	Liu et al. (2021b)	C_S	0.296% ACER

Continued on next page

Tabela 3.3 – continued from previous page

Year	Work	Dataset	Results
2021	Zheng et al. (2021)	N	99.5% Accuracy, 0.995% AUC, 0.1% EER
		S1	0% ACER
		S2	0% ACER
		S3	4.35% ACER
		O1	1.4% APCER, 1.8% BPCER, 1.6% ACER
		O2	2.6% APCER, 0.8% BPCER, 1.7% ACER
		O3	2% APCER, 3.9% BPCER, 2.8% ACER
		O4	4.2% APCER, 4.6% BPCER, 4.4% ACER
		*C_M	27.4% HTER
		*I	28.1% HTER
	Liu et al. (2021a)	C_S	2.4% APCER, 1.7% BPCER, 2.0% ACER
	Chen et al. (2021)	*I	21.3% HTER
		*M	14.8% HTER
		*C	32.3% HTER
	Wang et al. (2021b)	C	1.34% EER
		I	0.06% EER, 0.02% HTER
		O1	0.78% APCER, 1.06% BPCER, 0.92% ACER
		O2	3.84% APCER, 2.11% BPCER, 2.88% ACER
		O3	1.9% APCER, 3.8% BPCER, 2.8% ACER
		O4	4.0% APCER, 3.0% BPCER, 3.5% ACER
		*M	19.4% HTER, 86.87% AUC
		*C	22.03% HTER, 87.71% AUC
		*I	21.43% HTER, 88.81% AUC
		*O	18.26% HTER, 89.4% AUC
		*S_M, A (mixed)	36.93% APCER, 12.51% BPCER, 24.72% ACER
	George e Marcel (2021)	*S_M	14.7% HTER
		*W	12.7% HTER
	Quan et al. (2021)	I	0.28% EER
		C	0.53% EER
		M	0.18% EER
		O1	0.6% APCER, 0% BPCER, 0.4% ACER
		O2	1.7% APCER, 0.6% BPCER, 1.2% ACER
		O3	1.5% APCER, 2.2% BPCER, 1.7% ACER
		O4	5.2% APCER, 4.6% BPCER, 4.8% ACER
		*M	7.82% HTER, 97.67% AUC
		*C	4.01% HTER, 98.96% AUC
		*I	10.36% HTER, 97.16% AUC
		*O	14.23% HTER, 93.66% AUC
	Wang et al. (2021a)	*M	15.4% HTER, 91.8% AUC
		*C	24.5% HTER, 84.4% AUC
		*I	16.4% HTER, 92% AUC
		*O	23.1% HTER, 84.3% AUC

Continued on next page

Tabela 3.3 – continued from previous page

Year	Work	Dataset	Results
2022	Wang et al. (2022)	*M	6.67% HTER, 98.75% AUC
		*C	10% HTER, 96.67% AUC
		*I	8.88% HTER, 96.79% AUC
		*O	13.72% HTER, 93.63% AUC

3.3 KEY PROBLEMS

From this understanding of the state of the art, one main problem can be identified: generalization capability in classification models. This applies to multiple aspects of the classification, such as acquisition device, attack types and environmental conditions.

Another important issue for the scope of this work is that many methods exploit the abundance of information in videos, with models that extract cross-frame features. Since this work handles single-image liveness detection, its approach must face the difficulties that come with having even less input resources.

3.4 CONCLUDING REMARKS

The state of the art in liveness detection was studied and discussed, including the most relevant dataset. This work will build upon the idea of using estimated depth for enhanced classification, in particular using the network architectures presented in Chapter 2 for depth estimation and liveness classification.

4 PROPOSED SOLUTION FOR LIVENESS DETECTION

This chapter explains the proposed approach for liveness detection, a model based on two neural networks - one for obtaining depth and one for classification. The general architecture is demonstrated and its details are further discussed.

4.1 PROPOSED ARCHITECTURE

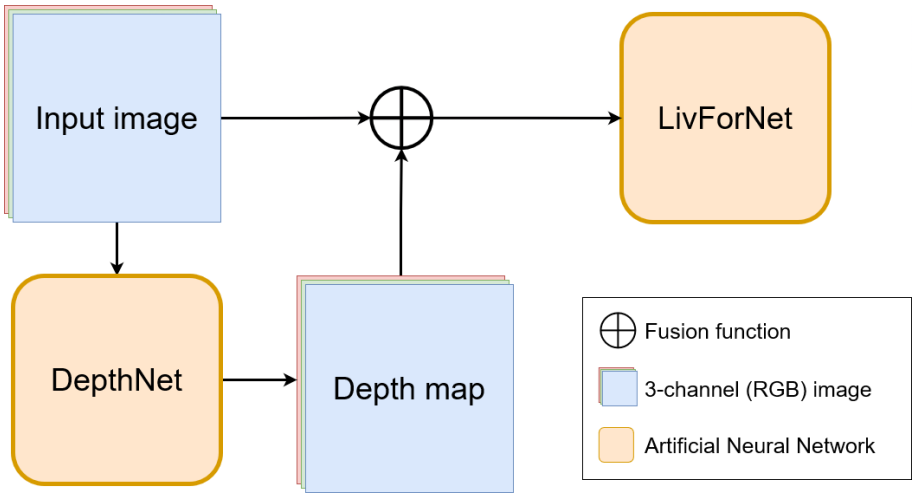


Figura 4.1: Proposed network architecture

Previous work has already used depth for enhanced liveness detection (Atoum et al., 2017; Liu et al., 2018; Shao et al., 2019; Wang et al., 2019, 2020c; Zhang et al., 2020; Zheng et al., 2021; Wang et al., 2021b), as explored in Chapter 3. The usual approach is to obtain the depth map for a given input in an adversarial scheme, and this work maintains this idea. The main difference in this proposal is using the Pix2Pix network for depth map estimation, as its aptitude for image domain translation may be useful in obtaining depth.

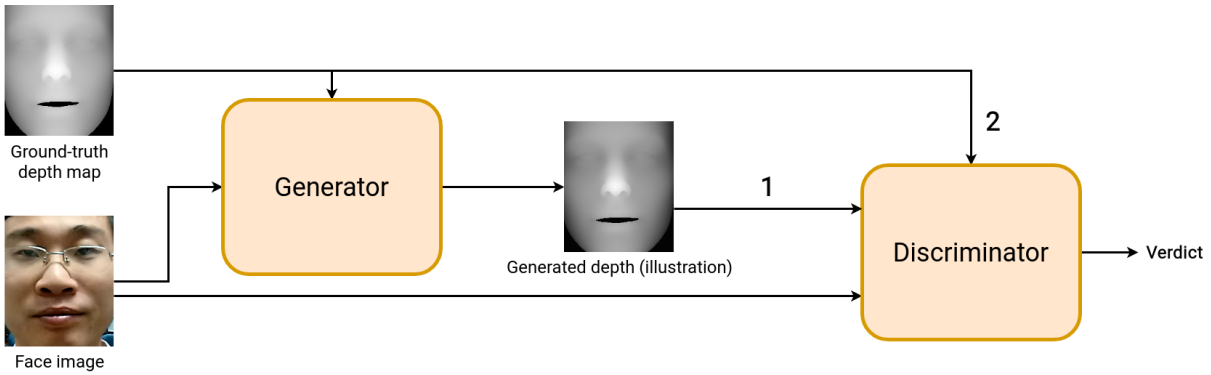


Figura 4.2: Conditional adversarial scheme for learning to generate depth maps. The discriminator first observes a pair with a generated depth map (1) and then one with a ground-truth depth (2).

Figure 4.2 illustrates the adversarial scheme for learning to generate face depth. From a ground-truth depth map D and an RGB face image I, a generator network G is trained to output

a realistic face depth map $G(I, D)$. At the same time, the discriminator network D is trained to determine whether an input (x, y), corresponding to a face image and a depth map, is real or not - that is, if y is a generated depth map. In this approach, the adversarial scheme is the Pix2Pix network (Isola et al., 2017).

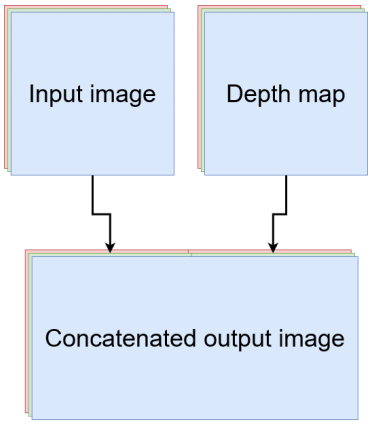


Figura 4.3: Illustration of the concatenation function.

Figure 4.1 illustrates the network's architecture. From a given input, the Pix2Pix network obtains its depth map, which is fused to the original image and fed as input to the classifier network (more on the fusion method below). Do notice this scheme is not end-to-end differentiable: the Pix2Pix network is completely trained before the backbone classifier.

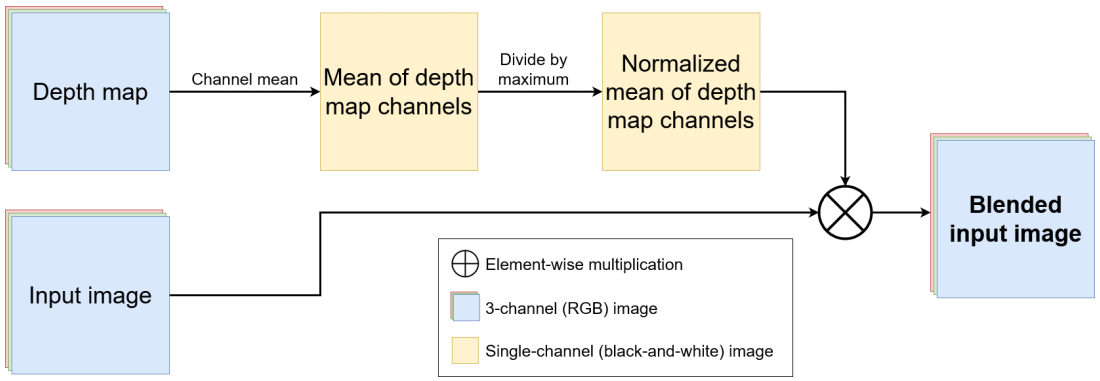


Figura 4.4: Illustration of the blending function.

The effectiveness of seven different backbone classifier networks is studied: FeatEmbedder, the same as in (Wang et al., 2021b); three sizes of ConvNeXt and three sizes of the classic ResNet. As the ConvNeXt family has recently shown superior performance in diverse computer vision tasks, the ResNet results are expected to serve as baseline values improved by the use of ConvNeXt, and that FeatEmbedder can be an approximation of how Pix2Pix compares to the adversarial scheme implemented in (Wang et al., 2021b).

$$\left(\frac{\overline{D}}{max_{\{max_{\overline{D}}, 1\}}} \right) \odot I \tag{4.1}$$

Two possibilities for fusion methods are explored, namely concatenation and blending. The concatenation function is illustrated in Figure 4.3; it is performed with respect to the channels axis (i.e., the depth map is appended horizontally to the original image). From two images of dimensions $C \times H \times W$ (channels, height and width, respectively), the output of the concatenation

function will be an image of dimensions $C \times H \times 2W$. The blending function, illustrated in Figure 4.4 and described by Equation 4.1, involves two intermediary depth pre-processing steps: from an input image I and its corresponding depth map D, the mean of all three depth map channels $\overline{D}$ (i.e., the depth map's luminance) is obtained and normalized by its maximum (or not normalized, in the particular case where the maximum equals zero). Finally, the Hadamard product between the normalized $\overline{D}$ and I is computed.

4.2 CONCLUDING REMARKS

This work's proposed approach has been presented in detail, in different levels of abstraction. The adversarial scheme is not further discussed, and remains as here presented in the experiments in Chapter 5. Two possible fusion methods were presented, namely blending and concatenation, and the choice of a definite (or "best") method for final experimentation is left for Chapter 5, where the decision is made after adequate empirical evaluation.

5 EXPERIMENTS

This chapter is split into two major sections: Section 5.1 describes the methodology applied for evaluating the proposed architecture, while Section 5.2 presents the obtained results.

5.1 METHODOLOGY

5.1.1 Databases

Two databases are used for training and evaluation, namely CASIA-FASD and Replay-Attack. Both databases include one set for training and one for validation. For model validation, results are presented in intra-dataset protocols, i.e., the model is trained and validated in disjoint subsets of the same dataset. For model testing, results are presented in both intra- and cross-dataset protocols.

Table 5.1 describes these protocols and their respective acronyms. HTER values are reported for all protocols.

Acronym	**Train Dataset**	**Validation Dataset**
C	CASIA-FASD	CASIA-FASD
R	Replay-Attack	Replay-Attack
CR	CASIA-FASD	Replay-Attack
RC	Replay-Attack	CASIA-FASD

Tabela 5.1: Evaluation protocols to be used in this work.

Since the chosen databases are composed of videos and this work contemplates *image* liveness, an additional frame extraction step is necessary. For each video sample, five equidistant frames are obtained to compose a resulting image dataset. Each frame has the same label as the source video and is a separate sample in the resulting dataset, which means two frames from the same source video will be treated as completely different image samples (i.e., their obvious similarities are ignored).

Two extra steps are also taken: face cropping and depth map ground truth generation. In the face cropping step, all samples are cropped to the face region. In the depth map ground truth generation step, all samples are mapped to a depth map - if the sample is a spoof, its depth map is just a black map; if it is real, the depth map is generated with the 3DDFA_V2 network (Guo et al., 2020). Both steps use code from the official 3DDFA_V2 GitHub repository (Guo et al., 2018), and the whole process of frame extraction is illustrated in Figure 5.1

5.1.2 Networks and Training Settings

The backbone models used for classification are FeatEmbedder, ResNet (18, 50 and 152) and ConvNeXt (tiny, base and large). All seven networks are described in Chapter 2. Pix2Pix is used for depth estimation.

Both the ResNet and ConvNeXt families have available PyTorch implementations with pretrained weights, which are used. This means all backbones go through a finetuning process on the train/test protocols, except for FeatEmbedder (which has no pretrained weights). The FeatEmbedder official implementation, made available by the authors (Wang et al., 2021b), is used.

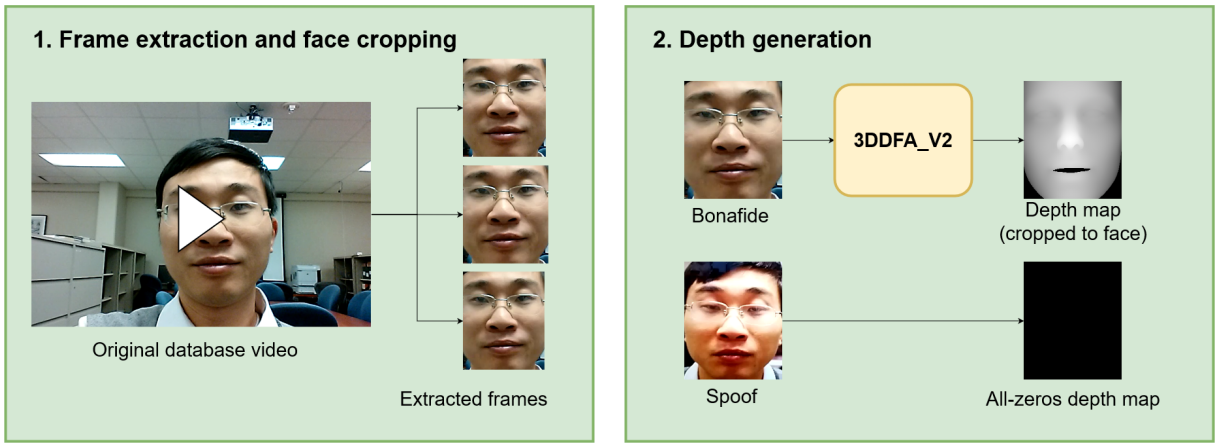


Figura 5.1: Processes for video frame extraction (1) and face depth generation (2). Samples belong to the MSU-MFSD (Wen et al., 2015) dataset.

For depth estimation, the Pix2Pix used network has no preset weights, and the used implementation follows the official one made available by the authors. Training takes 500 epochs, with a batch size of 16.

For all backbone classifier networks a Cross-Entropy loss function with label smoothing (0.1) is used. For the ConvNeXt models, the AdamW optimizer is used, with a learning rate of 0.004 and a 0.05 weight decay factor. For ResNet models, the Stochastic Gradient Descent (SGD) optimizer is used, with a learning rate of 0.1, a weight decay of 0.0001 and 0.9 momentum. Finally, for the FeatEmbedder network, the Adam optimizer is used, with a weight decay of 0.01. The training stage takes 600 epochs, with a batch size of 8.

All input images are resized to a resolution of 256 by 256, and no data augmentation is employed. Finally, all methods are implemented with the Python PyTorch (Paszke et al., 2019) library, so any parameters not detailed in this work follow the defaults from PyTorch functionality.

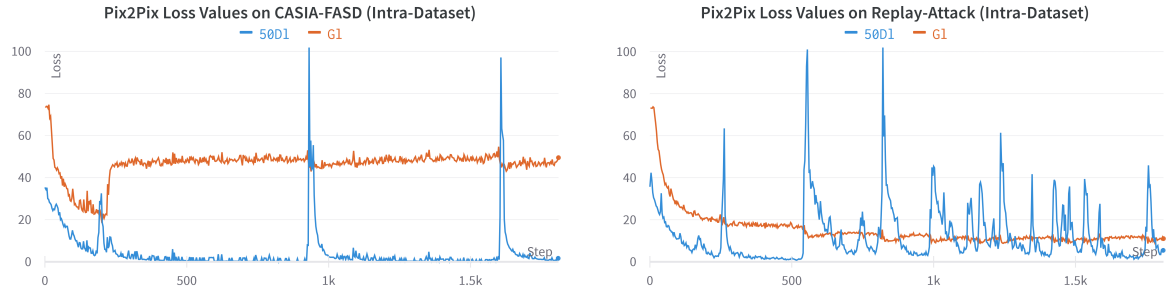
5.2 RESULTS

Table 5.2 compares HTER values for each fusion method across all networks in intra-dataset protocols. Since the ground-truth depth represents an upper bound of depth map quality, this allows for the analysis of two important questions: how the two fusion methods compare to each other and how well Pix2Pix-generated depth compares to the ground truth.

Model	**C**				**R**			
	GT		**Gen**		**GT**		**Gen**	
	Bl	**Cat**	**Bl**	**Cat**	**Bl**	**Cat**	**Bl**	**Cat**
FeatEmbedder	3%	7.556%	24.944%	49.444%	0%	14.375%	31.238%	50%
ResNet 18	3%	3.111%	50%	50%	3.375%	0%	26.237%	26.725%
ResNet 50	50%	3.37%	50.074%	50%	0%	0%	27.925%	26.013%
ResNet 152	4.889%	7.667%	50%	50%	1.25%	0%	29.738%	30.487%
ConvNeXt Tiny	50%	50%	50%	50%	50%	50%	50%	50%
ConvNeXt Base	3%	50%	50%	50%	50%	50%	50%	50%
ConvNeXt Large	3%	50%	50%	50%	50%	0%	50%	50%

Tabela 5.2: HTER values for ground-truth (GT) and generated (Gen) depth with both fusion methods (Bl: Blending, Cat: Concatenation) across all backbone classifiers on intra-dataset protocols

From Table 5.2, Table 5.3 is constructed, listing the fusion method with the lowest HTER value for each protocol considering ground-truth depth, generated depth or both. On Replay-Attack with ground-truth depth, there is a tie between both methods because an error rate of 0% is obtained. With CASIA-FASD, there is also little difference between ground-truth results (3% to 3.111%), which indicates both fusion methods are adequate or at least that they are not as strong determiners of performance as the input depth map's quality. For all subsequent analysis, the blending function is chosen for fusing depth due to its higher frequency in Table 5.3. Furthermore, the issue with generated depth map quality can be observed in Figure 5.2. The Pix2Pix networks could not find stability in convergence.



(a) Convergence graph for CASIA-FASD.

(b) Convergence graph for Replay-Attack.

Figura 5.2: Pix2Pix model convergence for both datasets on intra-dataset protocols. *Dl* is the discriminator loss (scaled for easier observation) and *Gl* is the generator loss, both as described in (Isola et al., 2017).

	Ground Truth	**Generated**	**Both**
C	Blending	Blending	Blending
R	-	Concatenation	Concatenation

Tabela 5.3: Fusion method with the lowest HTER value for each protocol with ground-truth depth, generated depth and both.

For model testing, Tables 5.4 and 5.5 compare HTER values for the chosen fusion method across all networks for ground-truth and generated depth against using no depth information in intra- and cross-dataset protocols, respectively. Note that all intra-dataset data is present in Table 5.2 except for the non-depth columns (repeated columns are added for readability). Again, since the ground-truth depth maps are an upper bound of depth map quality, this allows an understanding of how useful the depth information is for these classifiers.

5.3 CONCLUDING REMARKS

With the results from Section 5.2, an analysis of the main questions in this work is performed.

The first one is how depth information enhances a classifier's performance for the task of face liveness detection. From Tables 5.4 and 5.5 it can be observed that in most variations of models and protocols the depth-fed backbones perform better (even when this depth is not ground truth). This indicates that depth does, in fact, contribute to the classification capability of a model. This has been explored in related work (Zheng et al., 2021), and the presented results are consistent with findings of other authors.

The second question comes from Chapter 4. Two methods were presented for fusing depth maps with face images; which one performs best? Table 5.2 show the HTER values

Model	C			R		
	N	GT	Gen	N	GT	Gen
FeatEmbedder	47.185%	3%	24.944%	50%	0%	31.238%
ResNet 18	50%	3%	50%	12.85%	3.375%	26.237%
ResNet 50	47.889%	50%	50.074%	11.85%	0%	27.925%
ResNet 152	49.407%	4.889%	50%	17.825%	1.25%	29.738%
ConvNeXt Tiny	50%	50%	50%	50%	50%	50%
ConvNeXt Base	50%	3%	50%	50%	50%	50%
ConvNeXt Large	50%	3%	50%	50%	50%	50%

Tabela 5.4: HTER values for backbones with no depth (N), ground-truth depth (GT) and generated depth (Gen) on intra-dataset protocols.

Model	CR			RC		
	N	GT	Gen	N	GT	Gen
FeatEmbedder	58.625%	10%	50.25%	47.444%	33.111%	43.593%
ResNet 18	50.575%	0%	50%	54.296%	3%	43.63%
ResNet 50	52.525%	50%	50%	54.407%	3%	43.63%
ResNet 152	50%	50%	50%	57.185%	4.222%	43.667%
ConvNeXt Tiny	50%	0%	50%	50%	3%	50%
ConvNeXt Base	50%	50%	50%	50%	3%	50%
ConvNeXt Large	50%	0%	50%	50%	3%	50%

Tabela 5.5: HTER values for backbones with no depth (N), ground-truth depth (GT) and generated depth (Gen) on cross-dataset protocols.

for all variations of protocols and networks with both ground-truth and generated depth maps. Additionally, Table 5.3 summarizes the best fusion method for each combination of depth type and protocol. As every result but one for ground-truth depth are actually draws, and the one exception consists of a 3.7% difference, it is reasonable to reach the conclusion that the fusion method is not as strong a factor in model performance as the depth map quality. For the generated depth maps, results are balanced (half the results are in favor of blending fusion, and the other half are in favor of concatenation). When considering both types of depth maps, results are also balanced, since this result will be biased towards the less-effective type (generated, as opposed to ground-truth). It is understood that further experiments should be performed to better understand the relationship between fusion methods and model performance.

Another matter was how well the ConvNeXt family would perform in the task of face anti-spoofing. With no information beyond the RGB input, results are far from the state of the art, but depth information may make these models stronger competitors (Tables 5.4 and 5.5). However, this improvement is not consistent across evaluation protocols.

Finally, Pix2Pix is discussed as a cGAN for depth estimation. Isola et al. (2017) aimed to develop a framework that would require no hand-engineering of loss functions for different image-to-image translation tasks. The developed adversarial scheme has shown good results in different instances of the image-to-image translation problem, with even positive community engagement. For depth estimation in the scenario of face spoofing, however, where very similar images may have very different face depth maps, Pix2Pix could not perform well enough. This

can be observed in the difference between errors from ground-truth to generated depth in Tables 5.4 and 5.5.

6 CONCLUSION

The trajectory of the state of the art in face liveness detection has been thoroughly studied and summarized in this work. Major datasets have also been listed, and the current challenges in the field were discussed. From an understanding of these challenges a novel approach was proposed, namely using the Pix2Pix network for estimating depth that would then be used as auxiliary information in determining the presence of spoofs in face images. Presented experiments validate the value of depth in this task, and the Pix2Pix network reached insufficient results when trained in an isolated manner.

This work opens many possibilites for future development. First, it could be valuable to study the viability of including Pix2Pix in an end-to-end differentiable training scheme with the classifier network, so as to improve depth map generation quality. Following this idea, the Pix2Pix generator and discriminator networks could show good performance when inserted in an architecture such as the one presented by Wang et al. (2021b), where (1) there's end-to-end model integration and (2) it is not the generator's output that is fed as input to the classifier backbone, but an inner layer.

Another path is to replace Pix2Pix with 3DDFA_V2 (Guo et al., 2020), finetuning it to generate empty depth maps for spoof images. This has two advantages over the Pix2Pix approach: first, the 3DDFA_V2 has been pretrained for adequate 3DMM parameter generation, allowing it to achieve the state of the art (which is why it has been chosen for ground-truth depth map creation in the present work). Finetuning weights poses a clear advantage over training from randomly initialized ones (i.e., what was done with Pix2Pix). Second, since the 3DDFA_V2 network generates 3DMM parameters - orders of magnitude fewer than the images generated by Pix2Pix -, it has a better tendency towards convergence (Guo et al., 2020).

REFERÊNCIAS

Agarwal, A., Yadav, D., Kohli, N., Singh, R., Vatsa, M. e Noore, A. (2017). Face presentation attack with latex masks in multispectral videos. Em *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Anjos, A. e Marcel, S. (2011). Counter-measures to photo attacks in face recognition: A public database and a baseline. Em *2011 International Joint Conference on Biometrics (IJCB)*, páginas 1–7.

Atoum, Y., Liu, Y., Jourabloo, A. e Liu, X. (2017). Face anti-spoofing using patch and depth-based cnns. Em *2017 IEEE International Joint Conference on Biometrics (IJCB)*, páginas 319–328.

Boulkenafet, Z., Komulainen, J. e Hadid, A. (2016). Face spoofing detection using colour texture analysis. *IEEE Transactions on Information Forensics and Security*, 11(8):1818–1830.

Boulkenafet, Z., Komulainen, J., Li, L., Feng, X. e Hadid, A. (2017). Oulu-npu: A mobile face presentation attack database with real-world variations. Em *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, páginas 612–618.

Chan, P. P. K., Liu, W., Chen, D., Yeung, D. S., Zhang, F., Wang, X. e Hsu, C.-C. (2018). Face liveness detection using a flash against 2d spoofing attack. *IEEE Transactions on Information Forensics and Security*, 13(2):521–534.

Chen, B., Yang, W., Li, H., Wang, S. e Kwong, S. (2021). Camera invariant feature learning for generalized face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 16:2477–2492.

Chen, H., Hu, G., Lei, Z., Chen, Y., Robertson, N. M. e Li, S. Z. (2020). Attention-based two-stream convolutional networks for face spoofing detection. *IEEE Transactions on Information Forensics and Security*, 15:578–593.

Chingovska, I., Anjos, A. e Marcel, S. (2012). On the effectiveness of local binary patterns in face anti-spoofing. Em *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*, páginas 1–7.

Deb, D. e Jain, A. K. (2021). Look locally infer globally: A generalizable face anti-spoofing approach. *IEEE Transactions on Information Forensics and Security*, 16:1143–1157.

Garg, S., Mittal, S., Kumar, P. e Anant Athavale, V. (2020). Debnet: Multilayer deep network for liveness detection in face recognition system. Em *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)*, páginas 1136–1141.

George, A. e Marcel, S. (2021). On the effectiveness of vision transformers for zero-shot face anti-spoofing. Em *2021 IEEE International Joint Conference on Biometrics (IJCB)*, páginas 1–8.

Guo, J., Zhu, X. e Lei, Z. (2018). 3ddfa. `https://github.com/cleardusk/3DDFA`.

Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z. e Li, S. Z. (2020). Towards fast, accurate and stable 3d dense face alignment. Em *Proceedings of the European Conference on Computer Vision (ECCV)*.

He, K., Zhang, X., Ren, S. e Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Heusch, G., George, A., Geissbühler, D., Mostaani, Z. e Marcel, S. (2020). Deep models and shortwave infrared information to detect face presentation attacks. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(4):399–409.

Isola, P., Zhu, J.-Y., Zhou, T. e Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. Em *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ito, K., Okano, T. e Aoki, T. (2017). Recent advances in biometrie security: A case study of liveness detection in face recognition. Em *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, páginas 220–227.

Jia, Y., Zhang, J., Shan, S. e Chen, X. (2020). Single-side domain generalization for face anti-spoofing. Em *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jourabloo, A., Liu, Y. e Liu, X. (2018). Face de-spoofing: Anti-spoofing via noise modeling. Em *ECCV*.

Killioğlu, M., Taşkiran, M. e Kahraman, N. (2017). Anti-spoofing in face recognition with liveness detection using pupil tracking. Em *2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, páginas 000087–000092.

Koshy, R. e Mahmood, A. (2019). Optimizing deep cnn architectures for face liveness detection. *Entropy*, 21(4).

Li, H., Li, W., Cao, H., Wang, S., Huang, F. e Kot, A. C. (2018). Unsupervised domain adaptation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(7):1794–1809.

Li, L., Feng, X., Boulkenafet, Z., Xia, Z., Li, M. e Hadid, A. (2016). An original face anti-spoofing approach using partial convolutional neural network. Em *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, páginas 1–6.

Li, L., Xia, Z., Hadid, A., Jiang, X., Zhang, H. e Feng, X. (2019). Replayed video attack detection based on motion blur analysis. *IEEE Transactions on Information Forensics and Security*, 14(9):2246–2261.

Liu, A., Tan, Z., Wan, J., Liang, Y., Lei, Z., Guo, G. e Li, S. Z. (2021a). Face anti-spoofing via adversarial cross-modality translation. *IEEE Transactions on Information Forensics and Security*, 16:2759–2772.

Liu, A., Wan, J., Escalera, S., Jair Escalante, H., Tan, Z., Yuan, Q., Wang, K., Lin, C., Guo, G., Guyon, I. e Li, S. Z. (2019a). Multi-modal face anti-spoofing attack detection challenge at cvpr2019. Em *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Liu, S., Song, Y., Zhang, M., Zhao, J., Yang, S. e Hou, K. (2019b). An identity authentication method combining liveness detection and face recognition. *Sensors*, 19(21).

Liu, W., Wei, X., Lei, T., Wang, X., Meng, H. e Nandi, A. K. (2021b). Data fusion based two-stage cascade framework for multi-modality face anti-spoofing. *IEEE Transactions on Cognitive and Developmental Systems*, páginas 1–1.

Liu, Y., Jourabloo, A. e Liu, X. (2018). Learning deep models for face anti-spoofing: Binary or auxiliary supervision. Em *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, Y., Stehouwer, J., Jourabloo, A. e Liu, X. (2019c). Deep tree learning for zero-shot face anti-spoofing. Em *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 4675–4684.

Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T. e Xie, S. (2022). A convnet for the 2020s. *CoRR*, abs/2201.03545.

Luo, S., Kan, M., Wu, S., Chen, X. e Shan, S. (2018). Face anti-spoofing with multi-scale information. Em *2018 24th International Conference on Pattern Recognition (ICPR)*, páginas 3402–3407.

Mostaani, Z., George, A., Heusch, G., Geissbühler, D. e Marcel, S. (2020). The high-quality wide multi-channel attack (HQ-WMCA) database. *CoRR*, abs/2009.09703.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. e Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. Em Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. e Garnett, R., editores, *Advances in Neural Information Processing Systems 32*, páginas 8024–8035. Curran Associates, Inc.

Patel, K., Han, H. e Jain, A. K. (2016). Secure face unlock: Spoof detection on smartphones. *IEEE Transactions on Information Forensics and Security*, 11(10):2268–2283.

Purnapatra, S., Smalt, N., Bahmani, K., Das, P., Yambay, D., Mohammadi, A., George, A., Bourlai, T., Marcel, S., Schuckers, S., Fang, M., Damer, N., Boutros, F., Kuijper, A., Kantarci, A., Demir, B., Yildiz, Z., Ghafoory, Z., Dertli, H., Ekenel, H. K., Vu, S., Christophides, V., Dashuang, L., Guanghao, Z., Zhanlong, H., Junfu, L., Yufeng, J., Liu, S., Huang, S., Kuei, S., Singh, J. M. e Ramachandra, R. (2021). Face liveness detection competition (livdet-face) - 2021. Em *2021 IEEE International Joint Conference on Biometrics (IJCB)*, páginas 1–10.

Quan, R., Wu, Y., Yu, X. e Yang, Y. (2021). Progressive transfer learning for face anti-spoofing. *IEEE Transactions on Image Processing*, 30:3946–3955.

Sanghvi, N., Singh, S. K., Agarwal, A., Vatsa, M. e Singh, R. (2021). Mixnet for generalized face presentation attack detection. Em *2020 25th International Conference on Pattern Recognition (ICPR)*, páginas 5511–5518.

Shao, R., Lan, X., Li, J. e Yuen, P. C. (2019). Multi-adversarial discriminative deep domain generalization for face presentation attack detection. Em *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shen, T., Huang, Y. e Tong, Z. (2019). Facebagnet: Bag-of-local-features model for multi-modal face anti-spoofing. Em *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, páginas 1611–1616.

Singh, M. e Arora, A. S. (2018). A novel face liveness detection algorithm with multiple liveness indicators. *Wireless Personal Communications*, 100(4):1677–1687.

Tan, X., Li, Y., Liu, J. e Jiang, L. (2010). Face liveness detection from a single image with sparse low rank bilinear discriminative model. Em Daniilidis, K., Maragos, P. e Paragios, N., editores, *Computer Vision – ECCV 2010*, páginas 504–517, Berlin, Heidelberg. Springer Berlin Heidelberg.

Wang, G., Han, H., Shan, S. e Chen, X. (2020a). Cross-domain face presentation attack detection via multi-domain disentangled representation learning. Em *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 6677–6686.

Wang, G., Lan, C., Han, H., Shan, S. e Chen, X. (2019). Multi-modal face presentation attack detection via spatial and channel attentions. Em *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, páginas 1584–1590.

Wang, J., Zhang, J., Bian, Y., Cai, Y., Wang, C. e Pu, S. (2021a). Self-domain adaptation for face anti-spoofing. *CoRR*, abs/2102.12129.

Wang, Y., Song, X., Xu, T., Feng, Z. e Wu, X.-J. (2020b). Dfa. `https://github.com/Elroborn/DFA`.

Wang, Y., Song, X., Xu, T., Feng, Z. e Wu, X.-J. (2021b). From rgb to depth: Domain transfer network for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 16:4280–4290.

Wang, Z., Wang, Z., Yu, Z., Deng, W., Li, J., Gao, T. e Wang, Z. (2022). Domain generalization via shuffled style assembly for face anti-spoofing.

Wang, Z., Yu, Z., Zhao, C., Zhu, X., Qin, Y., Zhou, Q., Zhou, F. e Lei, Z. (2020c). Deep spatial gradient and temporal depth learning for face anti-spoofing. Em *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 5041–5050.

Wen, D., Han, H. e Jain, A. K. (2015). Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761.

Yang, J., Lei, Z., Yi, D. e Li, S. Z. (2015). Person-specific face antispoofing with subject domain adaptation. *IEEE Transactions on Information Forensics and Security*, 10(4):797–809.

Yang, X., Luo, W., Bao, L., Gao, Y., Gong, D., Zheng, S., Li, Z. e Liu, W. (2019). Face anti-spoofing: Model matters, so does data. Em *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 3502–3511.

Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., Zhou, F. e Zhao, G. (2020). Searching central difference convolutional networks for face anti-spoofing. Em *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 5294–5304.

Zhang, K.-Y., Yao, T., Zhang, J., Tai, Y., Ding, S., Li, J., Huang, F., Song, H. e Ma, L. (2020). Face anti-spoofing via disentangled representation learning. Em *European Conference on Computer Vision*, páginas 641–657. Springer.

Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D. e Li, S. Z. (2012). A face antispoofing database with diverse attacks. Em *2012 5th IAPR International Conference on Biometrics (ICB)*, páginas 26–31.

Zheng, W., Yue, M., Zhao, S. e Liu, S. (2021). Attention-based spatial-temporal multi-scale network for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(3):296–307.