electronics

MDPI

Article

Quantitative Study of Swin Transformer and Loss Function Combinations for Face Anti-Spoofing

Liang Yu Gong *,†,‡ and Xue Jun Li *,†,‡

Department of Electrical and Electronic Engineering, Auckland University of Technology, Auckland 1010, New Zealand

* Correspondence: liangyu.gong@autuni.ac.nz (L.Y.G.); xuejun.li@aut.ac.nz (X.J.L.)

† Current address: 6 Saint Paul Street, Auckland 1010, New Zealand.

‡ These authors contributed equally to this work.

Abstract: Face anti-spoofing (FAS) has always been a hidden danger in network security, especially with the widespread application of facial recognition systems. However, some current FAS methods are not effective at detecting different forgery types and are prone to overfitting, which means they cannot effectively process unseen spoof types. Different loss functions significantly impact the classification effect based on the same feature extraction without considering the quality of the feature extraction. Therefore, it is necessary to find a loss function or a combination of different loss functions for spoofing detection tasks. This paper mainly aims to compare the effects of different loss functions or loss function combinations. We selected the Swin Transformer as the backbone of our training model to extract facial features to ensure the accuracy of the ablation experiment. For the application of loss functions, we adopted four classical loss functions: cross-entropy loss (CE loss), semi-hard triplet loss, L1 loss and focal loss. Finally, this work proposed combinations of Swin Transformers and different loss functions (pairs) to test through in-dataset experiments with some common FAS datasets (CelebA-Spoofing, CASIA-MFSD, Replay attack and OULU-NPU). We conclude that using a single loss function cannot produce the best results for the FAS task, and the best accuracy is obtained when applying triplet loss, cross-entropy loss and Smooth L1 loss as a loss combination.

Keywords: loss functions; Swin Transformer; pixel-wise loss; face anti-spoofing detection

Academic Editors: Dah-Jye Lee, Taehyeon Kim and KyungTaek Lee

Received: 25 November 2024
Revised: 6 January 2025
Accepted: 21 January 2025
Published: 23 January 2025

Citation: Gong, L.Y.; Li, X.J. Quantitative Study of Swin Transformer and Loss Function Combinations for Face Anti-Spoofing. *Electronics* **2025**, *14*, 448. https://doi.org/10.3390/electronics14030448

1. Introduction

Face anti-spoofing (FAS) has always been a difficult task that has attracted much attention in cybersecurity. Especially with the widespread application of facial recognition, many online financial systems require facial information (Face ID) to unlock accounts. In the beginning, most work utilised traditional methods, such as Local Binary Pattern [1], to extract hand-crafted features in face anti-spoofing tasks. This method could efficiently detect the images' texture information; however, it is not commonly used for detecting different spoof types. Since CNNs show relatively good performance in image classification work, some researchers have selected ResNet [2] or Xception [3] as the backbone to design spoofing detectors. In the process of developing face-spoofing detectors, it is inevitable that the most efficient backbone with a loss function will be found to test the classification effect of anti-spoofing. With the continuous development of VIT, Swin Transformer has been used as an excellent feature extractor in many image recognition fields. However, spoof (real) facial information will become more and more abundant as feature extractors continue to evolve. Therefore, finding a suitable loss function combination is a basic task for

the convergence of the model and improvement of the recognition effect. For the proposed method, we first chose the Swin Transformer as the human face representation extractor and performed an ablation test to examine different loss function combinations. Specifically, we utilised cross-entropy loss (CE loss) as one standard loss function, and then we separately combined it with triplet loss [4], Smooth L1 loss and focal loss. The performance of different loss function combinations was tested, and we analysed the different loss functions' advantages and disadvantages. The second objective of this article is to build an Auto-Encoder architecture to generate the corresponding real clue maps and spoof clue maps, which are prepared for calculating the pixel-wise loss. The entire architecture contains two stages: the first stage is a Swin Transformer-based Encoder to extract facial features, and then two separated ResNet Decoders are designed to reconvert clue maps. These clue maps are used to calculate the pixel-wise loss. The loss functions are combined and grouped with the final loss for model training. Finally, we applied this architecture to train and validate it using common deepfake datasets to observe the model's distinguishing ability. There are three main contributions of this work: (1) Firstly, to test the performance of pixel-wise loss, such as Smooth L1 loss, we propose an Auto-Encoder structure to generate live and spoof clue maps. (2) Since we obtained the corresponding generated clue maps, we tested different single loss functions' performance, including CE loss, semi-hard triplet loss, pixel-wise loss and focal loss, in face-spoofing tasks. (3) Finally, this research also tested the in-dataset results for loss function combinations and found the loss function combination with the best performance. Our paper focuses on AI applications, and our work could help future researchers select loss functions and loss function combinations, which would enhance their efficiency in research.

2. Related Work

2.1. Face Anti-Spoofing

Because facial recognition is convenient with high accuracy, facial ID has been widely applied in checking in and mobile payments [5]. Therefore, current systems are vulnerable to Spoofing Attacks, such as paper masks and screen replay attacks. Most banking systems use an interactive detection system for face anti-spoofing, requiring the tester to act accordingly to determine authenticity, and the silent liveness detection system supervises the sample to a certain extent. We reviewed several representative FAS methods, and there are two main methods for detecting spoof images in the FAS field. The first one utilises a convolutional neural network (CNN) for classification, and the second one uses hybrid learning methods. CNNs have always had great advantages in classification tasks, and because of the maturity of some classification models, applying the CNN architecture as a face feature extractor is the first choice for most researchers. For instance, the researchers in [6] developed a fine-grained network that encompasses diverse spoofing methods. Their research demonstrates that multi-class supervision is more effective than binary classification and is capable of representing certain detailed paper attack and replay attack information. However, most face anti-spoofing datasets have an asymmetric distribution of real and spoofing classes; therefore, some work [7] proved that it is useful to apply focal loss or pixel-wise loss to supervise model learning. On the other hand, the hybrid learning method combines hand-crafted features with machine learning- or deep learning-extracted features. The main innovation is in designing different feature fusion methods or combining multiple data modalities [8]. Khamari [9] utilises Local Binary Pattern (LBP) and Weber descriptors to extract hand-crafted features and then embeds them into CNN-extracted features to obtain richer human facial information, specifically focusing more on the edge information. In order to filter out the uncorrelated deep learning representations, some proposed methods [10] use hand-crafted features, which is also an effective hybrid learning

approach. Our proposed method can be considered a CNN classification method, but we utilise an Auto-Encoder to reconvert the "clue" maps and apply semi-hard triplet loss and pixel-wise loss functions to train the model so that they will increase the model's generalisation ability. In particular, it is more effective in distinguishing some hard samples and can achieve better evaluation performance.

2.2. Data Augmentations

Data augmentation is a technique that transforms, modifies, or combines existing data to generate more samples. It can increase the robustness of the training model and prevent overfitting in computer vision. In this work, there are five data augmentation methods applied to change the input's pixel values and positions. Base transformation is the most common and performs pixel normalisation and then resizes to specific tensor sizes. Random Erasing [11] means randomly erasing a portion of an image during training and replacing this area with some padding zero values, which has been widely used in image classification and object detection. Random Crop entails randomly cropping a subregion of a facial image, and the subregion's position is random. After randomly cropping an area, the image is resized to the corresponding height and width. DFDC Selim contains Gaussian noise, Gaussian blur and random shift. Random Augmentation involves randomly selecting one of the data augmentations, for example, Random Erasing or Random Resize Crop. As compared to the design of traditional data augmentations for image classification, we have largely increased the randomness of input data and minimised the occurrence of overfitting.

2.3. Swin Transformer

Convolutional neural networks (CNNs) are extensively utilised in image classification tasks. But, Vision Transformers with self-attention mechanisms [12] have opened up a new realm for classification in the computer vision domain. They separate images into embedding patches and integrate their relationships. After obtaining the patch embeddings, these embeddings apply linear projection to calculate the corresponding Query (Q), Key (K) and Values (V). Then, the attention score is calculated using Equation (1).

$$Attention(Q, K, V) = SoftMax(QK^T/\sqrt{d} + B)V \quad (1)$$

where Q, K and V are Query, Key and Value, respectively; d is the dimension of Key; $SoftMax$ represents the softmax function; and B represents bias.

Swin Transformer [13] has been growing in strength among various Vision Transformer (ViT) models. Swin Transformer employs shifted windows and cyclic shifts to acquire the interaction information between separate patches. This is different from the way that other ViTs compute global self-attention. It can largely address the drawbacks of some common ViTs that lack overlapping image patches. In addition, patch merging is effective in constructing hierarchical representations by adjusting the resolution and channels of features at each stage. This Transformer architecture reduces the computational complexity, and it demonstrates that it can serve as a general backbone for different recognition tasks. This proposed model architecture currently achieves SOTA accuracy with suitable Floating-Point Operations and parameters.

3. Proposed Methods

We followed the spoofing detector pipeline and designed our model architecture with four parts: data pre-processing, feature extraction, Decoder design and loss calculation. The whole architecture (see Figure 1) contains two stages. Firstly, we make use of a facial detector to locate and crop the facial areas. Subsequently, we randomly employ data aug-

mentations to acquire data views. These data views are then fed into the feature extractor (Swin Transformer) for extraction, and this constitutes the output of Stage 1. Once the extracted features are collected, we use two separated Decoders to restore the relative clue maps, since we regard the ideal of the generated clue maps tending to be all-zero and all-one maps. We utilise Smooth L1 loss for model training. Finally, the output of Stage 1 directly goes through a Multi-Layer Perception (MLP) for binary classification after applying L2 generalisation to prevent overfitting. This research work first compared different single loss functions' performance, and then we selected CE loss and combined different loss functions to test their performance. Finally, the most effective loss combinations for FAS tasks were selected based on in-dataset experiments.

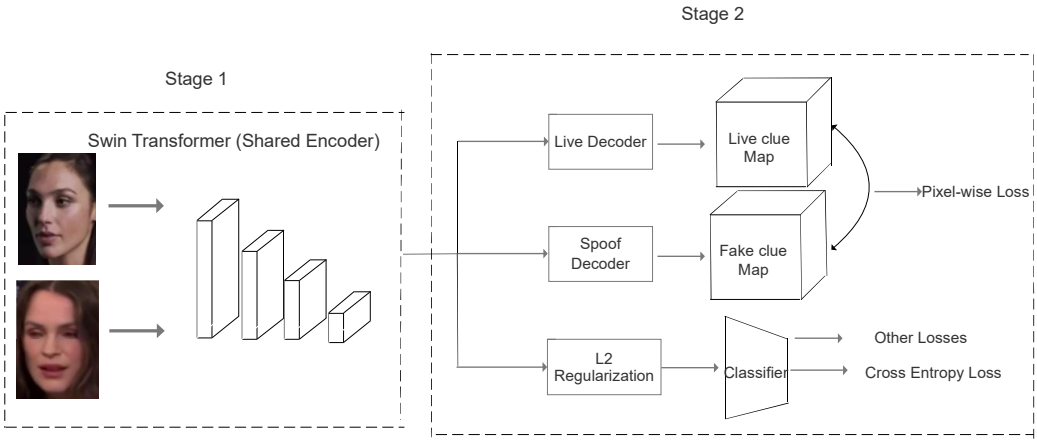

Figure 1. An explanation of the proposed model architecture.

3.1. Pre-Processing Work

Since most of the fake datasets are in the form of videos, we take image frames from B videos and extract M frames from each video on average; meanwhile, the total collected frames can be calculated. Then, we utilise a facial detector [14] to locate human face positions. Once the human faces are located, we apply face alignment and crop areas and resize them to specific widths and heights for model training. In addition, we enlarge the size of facial images by 10% to involve facial background information. Eventually, we obtain a total of $M \times N$ frames for the feature extractor and resize them all to $3 \times 224 \times 224$. Finally, five random data augmentations are applied, and the input view sets are of the size $[B, M, 3, 224, 224]$. These augmented tensors serve as the final input for our designed model, as depicted in Figure 2.

Figure 2. Our proposed augmented data for one single facial image.

3.2. Model Architecture

The entire model architecture utilises an Auto-Encoder because the Encoder part can compress high-dimensional data into lower dimensions and then reduce the computational complexity of the data. In addition, the Auto-Encoder is more flexible in the selection of

Encoder and Decoder models. To verify the classification effect of pixel-wise loss on face anti-spoofing, we extended the original classifier based on the Swin Transformer extractor and designed a Decoder module to restore the extracted features into the corresponding clue maps. We specify that the ideal real-face maps are all-zero matrices, and conversely, the spoof-face maps are all-one matrices. The whole architecture is divided into two stages. Stage 1 is a Swin Transformer-based Encoder. Firstly, the facial images are split into patches by convolutional operations, and we regard them as the "tokens". After applying 12 basic Swin Transformer blocks, the output features have a size of $[49, 768]$. These extracted features are stored in the latent space for adding different loss functions in Stage 2, such as pixel-wise loss and triplet loss.

In Stage 2, we use ResNet as the Decoder because its skip connection can prevent gradient explosion and vanishing from happening during training procedures. Before feeding low-dimensional features to the Decoder, we reshape the features to $[12, 56, 56]$ for generating the corresponding clue maps, and this stage contains two separately trained Decoders with different parameters. Since the Decoders are based on the ResNet architecture, we use 1 convolution block with 13 residual blocks. The module components of the designed architectures are shown in Table 1.

Table 1. The components of the Decoder.

Decoder Layers	Output Feature Size	Components
Decoder Conv1	$64 \times 28 \times 28$	2×2, stride = 2
Decoder Basic Block1	$128 \times 28 \times 28$	$[3 \times 3, 3 \times 3] \times 3$
Decoder Basic Block2	$256 \times 14 \times 14$	$[3 \times 3, 3 \times 3] \times 4$
Decoder Basic Block3	$1024 \times 7 \times 7$	$[3 \times 3, 3 \times 3] \times 6$
Reshape	224×224	None

3.3. Loss Functions

The cross-entropy function L_{CE} (as shown in Equation (2)), also known as Kullback–Leibler divergence (KL divergence), is a traditional measurement of the difference between two probability distributions. When two probability differences have a large distribution, the KL divergence also becomes larger. Therefore, it is very suitable for logistic regression tasks to use the CE loss function to compute the difference between the distribution of predicted values and the distribution of labels. At the same time, the CE loss gradient update is more efficient than other pixel-wise losses; the gradient of cross-entropy loss with the softmax activation function for binary classification is shown in Equation (3). Compared with the MSE loss function, the CE loss function's output value is more stable; for example, MSE always outputs a lower loss function in these outliers, which slows down model updating progress. Therefore, the cross-entropy function is more biased towards calculating the probability distribution distance of the data, while the MSE is more biased towards calculating the "numerical distance".

$$L_{CE} = -[y \log(p) + (1 - y) \log(1 - p)] \tag{2}$$

$$\frac{\partial L}{\partial z} = p - y \tag{3}$$

where p is the predicted value, y represents the Ground-Truth (GT) Label, and z represents the output of the model. The Ground-Truth Label is considered the absolute true result and is the standard reference point for training and evaluating model performance.

However, the CE loss function is very sensitive to outliers (or noise). Especially when the prediction probability of the model is very close to 0 or 1, the loss value can increase rapidly, resulting in a gradient explosion that affects the training process. This is not benefi-

cial for face anti-spoofing tasks, because detecting spoofing samples is always regarded as anomaly detection. In addition, falsified face datasets usually have unbalanced data categories, so using only the CE loss function in this case cannot achieve the optimal effect. CE loss, especially binary cross-entropy loss, has fixed weights for each class prediction and does not take into account unbalanced data. The trained model will sometimes be biased towards the class with large amounts of data. Therefore, focal loss is an improved cross-entropy loss that is widely applied in this imbalanced-category situation. In focal loss, α is the balanced parameter to adjust the impact of positive and negative samples. γ is the focusing parameter to control the loss scaling of samples. To prevent imbalanced data, we fuse the balanced parameter α with the traditional focal loss, which is shown in Equation (4).

$$L_{FL} = -[\alpha y(1-p)^{\gamma} log(p) + (1-\alpha)(1-y)p^{\gamma} log(1-p)] \tag{4}$$

where p is the predicted value, and y denotes the Ground-Truth Label; α represents the balanced parameter and γ is the focusing parameter, which are both adjustable hyperparameters.

Semi-hard triplet loss is another loss function in classification, and the main purpose of applying this function is to shorten the same-class samples' distances and enlarge the different-class samples' distances. One of the most typical applications of this function is Person Re-identification because this function is good for determining whether the samples match or not. Each triplet loss has three main components: anchor (a), positive (p) and negative (n), where the anchor and positive belong to the same class. We aim to use an online method to create positives and semi-hard negatives in one mini-batch. To distinguish samples more easily, we stipulate that the Euclidean distance between (a) and (p) is shorter than the distance between (a) and (n). Meanwhile, positive exemplars with a margin are farther away from (a) than from (n). Put another way, this process makes the negative embeddings limit the margin area to separate the slight differences between real and fake samples. The embedding distance comparison is demonstrated in Equation (5), and the semi-hard triplet loss function is presented in Equation (6). Semi-hard triplet loss is more beneficial for classifying samples by computing the distances between them, and the margin can make sure that the distance between negative samples and anchor points is large enough. However, the challenges of applying this loss are negative sample selection and triplet generation, which is complex and requires large computing resources. In addition, it is susceptible to overfitting, especially on small-scale datasets.

$$d(a,p) < d(a,n) < d(a,p) + margin \tag{5}$$

$$L_{tri} = \frac{1}{n}\sum_{i=0}^{n}[\| f_a - f_n \|_2^2 - \| f_a - f_p \|_2^2 - margin] \tag{6}$$

where f_a represents the anchor feature, and a is the anchor sample; f_p denotes the positive embedding, and p is the positive sample with the same class as the anchor; n is the negative sample, and f_n represents the negative embedding; and $margin$ represents the enforced distance between positive and negative samples.

To measure the numerical difference between the generated clue map and the idealised clue map, the proposed method utilises Smooth L1 loss in the Decoder. This function combines L1 loss and L2 loss; the loss is shown in Equation (6). When the error between generated maps and idealised maps is small, we apply L2 loss, which helps to optimise the gradient descent during the early stage of training. When the error is large, the gradient of L1 loss no longer increases, which reduces the impact of large errors on the model. Smooth L1 loss takes advantage of both L1 and L2 losses: it modifies the non-smoothing at zero

points and thus is useful for processing outliers. In particular, as illustrated in Equation (7), we use spoof clue maps to compute Smooth L1 loss.

$$L_{pixel} = \begin{cases} 0.5(M_1 - M_S)^2 & if|M_1 - M_S| < 1 \\ |M_1 - M_S| - 0.5 & otherwise \end{cases} \quad (7)$$

where M_1 denotes all-one maps; M_S is the generated Spoof Embeddings.

Similarly, we use live clue maps to compute Smooth L1 loss, which can be written as in Equation (8).

$$L_{pixel} = \begin{cases} 0.5(M_L - M_0)^2 & if|M_L - M_0| < 1 \\ |M_L - M_0| - 0.5 & otherwise \end{cases} \quad (8)$$

where M_0 represents all-zero maps; M_L denotes the generated Live Embeddings.

4. Experiments

4.1. Datasets

Four face-spoofing datasets were used in this experiment: CelebA-Spoof [15], OULU-NPU [16], CASIA-MFSD [17] and Replay attack. Due to the large amount of data in CelebA-Spoofing, we chose it as the training set, and the other two datasets were used for cross-dataset testing.

CelebA-Spoofing contains 625,337 images, and the live images amount to more than 202,599, which originated from 10,117 subjects. This dataset provides various illumination conditions and 43 attributes with rich annotations. The spoofing types comprise paper print, paper cut, 3D masks and replay attacks, and these spoofing methods are the most common in the real world. The OULU dataset has 4950 real videos and fake videos with different shooting cameras and backgrounds. There are four protocols for evaluating different unseen environmental conditions for PAD attacks and printer attacks. CASIA-MFSD is a small-scale spoofing dataset. It encompasses 600 video clips involving 50 subjects for both the training and testing stages. For every video folder, 12 video clips are included, among which 3 of them belong to the liveness classes and were shot by different cameras. As a result, this dataset is unbalanced. However, it is quite suitable for cross-dataset experiments to demonstrate the robustness of our model and loss function. Replay attack is a dataset specifically designed for testing the performance of replay attack FAS tasks. This dataset contains facial video data from different cameras and at different angles. It is widely used to study and develop replay attack detection systems, and it contains about 1300 videos covering 50 subjects. The video length is typically 10–15 s, and both real facial videos and forged (replay) videos are included. In this experiment, we utilised four main evaluation metrics to present our proposed method's performance: accuracy, Attack Presentation Classification Error Rate (APCER), Bona fide Presentation Classification Error Rate (BPCER) and Average Classification Error Rate (ACER). The last three metrics are error rates, which means the lower their values, the better the performance of the proposed method.

4.2. Implementation Details

In this section, we apply random data augmentation methods, which include basic transformation, Random Erasing, Random Crop and DFDC Selium, for our proposed method. All images are resampled and positioned within the facial areas by a facial detector [14] with an NMS value of 0.5 in the pre-processing step. Moreover, the bounding boxes of these images are enlarged by a factor of 1.2. Then, we resize them to $3 \times 224 \times 224$. Then, we make use of the Adam optimiser, and its learning rate is 0.0001 and weight decay

is 5×10^{-4} specifically. A dropout rate of 0.2 is applied to the MLP to prevent overfitting. The semi-hard triplet loss margin is set to 0.5, and the alpha (α) and gamma (γ) parameters are set to 0.75 and 2, respectively, in binary cross-entropy focal loss (focal loss). In addition, the pre-trained model parameter "swin tiny patch4 window7 224.pth" is chosen for the Encoder, the batch size is set to 32, and the total training epoch is set to 30.

4.3. Ablation Test for Single Loss Function

In this section, we use cross-entropy loss, semi-hard triplet loss, binary cross-entropy focal loss and Smooth L1 loss to compare different loss functions' performance in distinguishing in-dataset data. The feature extractor is Swin Transformer, and the training dataset is CelebA-Spoofing because if the training data amount is low, Swin Transformer cannot learn useful information with limited epochs, and it cannot perform relatively well on small-scale datasets. The best accuracy, APCER, BPCER and ACER are summarised in Table 2. This ablation test shows that the most effective loss function for face anti-spoofing is focal loss, which can reach 86.73% accuracy, 4.365% APCER, 12.703% BPCER and 8.534% ACER, because weighting different categories with different sample proportions increase the effectiveness of learning the data features of the category with less data in the case of an imbalanced data distribution. Cross-entropy loss, as a standard and common classification loss function, shows the second-best performance. Only triplet loss does not perform well in distinguishing spoofed faces, and the BPCER also increases by approximately 5%. The clustering function of triplet loss is not obvious, and the recognition function decreases significantly for both the live and spoof categories. Smooth L1 loss reaches an accuracy of 77.16% but obtains the lowest APCER results. This means that this loss function can effectively learn spoof clue features, but it cannot effectively distinguish real faces. The main reason that the BPCER greatly increases is that the number of spoof images used for training is much larger than the real-face number, leading to a higher bona fide error rate. Triplet loss and Smooth L1 loss rely heavily on calculating the similarity (Euclidean distance) between different kinds of features, which leads to model overfitting and decreases accuracy in the validating set.

Table 2. An ablation test for the single loss function. The backbones are all Swin Transformers but utilise different loss functions to supervise model learning and update parameters.

Loss Functions	Accuracy (%)	APCER (%)	BPCER (%)	ACER (%)
CE loss	84.35	6.444	15.672	11.058
Semi-hard triplet loss	71.44	47.665	20.710	34.188
Focal loss	86.73	4.365	12.703	8.534
Smooth L1 loss	77.16	2.780	31.528	17.153

4.4. Ablation Test for Loss Combination

There were three loss function combinations used in this ablation test, which were triplet loss with CE loss, CE loss with semi-hard triplet loss, and Smooth L1 loss and focal loss with semi-hard triplet loss and Smooth L1 loss. Specifically, the Encoder part's architectures were entirely the same in the ablation experiment. Semi-hard triplet loss was utilised for representation extractions of the Swin Transformer Encoder in the second testing method, and the third method applied Smooth L1 loss to the "Clue Maps" (Decoder part). We also conducted an in-dataset experiment and adopted CelebA-Spoofing as the training set. Subsequently, this work calculated the relative evaluation metrics separately within 30 epochs, and the evaluation results are presented in Table 3.

Table 3. In-dataset validation results for four loss function combinations.

Loss Combination	Epoch	Accuracy (%)	APCER (%)	BPCER (%)	ACER(%)
CE loss	18	84.35	6.444	15.672	11.508
CE + triplet	22	89.75	8.401	12.494	10.448
CE + triplet + Smooth L1	17	93.52	2.184	8.773	5.479
Focal + triplet + Smooth L1	6	88.86	1.324	15.391	8.357

This ablation experiment reveals that the supervised learning FAS method does not have the best performance. When adding the semi-hard triplet loss to the Encoder, the APCER increases, indicating that the model makes more incorrect predictions for negative samples. However, the BPCER drops by approximately 3%. Therefore, the combination of CE loss and semi-hard triplet loss in the Encoder can enhance the validation accuracy, especially for the live-sample sets. In the end, we utilise CE loss, which is applied to the auxiliary classifier. Meanwhile, Smooth L1 loss and semi-hard triplet loss are both applied to the generated "Clue Maps". As a result, all the error rates decrease significantly, and the validation accuracy increases substantially. Thus, it is feasible to reconvert the facial representations extracted from the Encoder to the clue maps of the original input size. Moreover, the unsupervised losses applied to the Decoder can improve the evaluation performance for both spoof- and live-sample sets. In addition, we chose the best-performing loss function in the single-loss ablation test, focal loss, to combine with semi-hard triplet loss and Smooth L1 loss to test the in-dataset relative evaluations; this combination reached the lowest APCER, which means it can effectively distinguish most spoofing types, but the BPCER increased substantially in the experiment. Since the evaluation performance of focal loss is largely determined by two of its own adjustable hyperparameters, the hyperparameters set in the commonly used binary classification tasks cannot solve the problem of FAS to the greatest extent.

4.5. Cross-Dataset Study

Since the combination of CE loss, semi-hard triplet loss and Smooth L1 loss showed the highest accuracy in the in-dataset experiment, we utilised this model to test cross-dataset performance with some benchmark models. In this cross-dataset experiment, we used two protocols, and the first one entailed using CASIA-MFSD as a training dataset and testing on Replay attacks [18]. Then, to better prove our model's generalisation ability, we chose Replay attacks as the training set and CASIA-MFSD as the testing set as the second protocol. The performance results are shown in Tables 4 and 5 accordingly. In the first protocol, our method obtains the lowest ACER, which reaches 23.8%. However, the Replay attacks dataset does not have other spoofing types, but CASIA-MFSD contains paper attacks and different replay attacks. It shows the second-best performance among the listed benchmark methods. Even though some of the methods fuse different data modalities or features, our work shows competitive results. This is because Swin Transformer can use the correlation between patch embeddings, which means it can extract more information, and pixel-wise loss (Smooth L1 loss) is useful for measuring the outline and colour of the images. Thus, this loss function is efficient for classifying spoofing images because different spoof types have obvious colour differences.

Based on this proposed methodology, we also applied our model to FaceForensic++, which is a common deepfake dataset. FaceForensics++ contains multiple forgery methods, including deepfakes, neural texture, Face2Face and Faceswap. In particular, neural texture samples have undergone by professional post-processing, and it is hard to distinguish fakes by the human eye. It only reached 62.5% in the in-dataset experiment. The main reason is that deepfake data were generated by multiple forgery methods. Because of the colour

blending and other post-processing work on generated fake media, the deepfake frames look more realistic and do not present obvious clues that the model could use to extract the relative features and distinguish them. Thus, this combination of loss functions only performs well in the FAS task.

Table 4. The experimental results when trained on CASIA-MFSD and tested on Replay attacks.

Models	ACER
Motion	50.2%
CNN [19]	48.5%
LBP [9]	47.0%
Auxiliary [20]	27.6%
FaceDS [21]	28.5%
This work	23.8%

Table 5. The inter-dataset experiment when trained on Replay attacks and tested on CASIA-MFSD.

Models	ACER
Motion	47.9%
CNN [19]	45.5%
LBP [9]	39.6%
Auxiliary [20]	28.4%
FaceDS [21]	41.1%
This work	33.1%

5. Conclusions and Future Work

In this paper, we used a Swin Transformer-based feature extractor and designed a clue map generator for computing Smooth L1 loss. In the ablation test, we verified the performance of a single loss function and a combination of different loss functions in face anti-spoofing. Among them, focal loss had the strongest distinguishing ability for spoofing and live faces in the test of the single loss function. CE loss with semi-hard triplet loss and Smooth L1 loss presented the best performance in face anti-spoofing, which reached 93.52% accuracy in the in-dataset experiments and 23.8% and 33.1% ACER in cross-dataset experiments. In addition, we also applied our model architecture to deepfake datasets but could not reach a new SOTA performance when compared with other benchmark models. In real life, spoofing detection is mainly used to unlock financial accounts using facial information, and most bank systems use interactive detection methods to determine whether the tester is a real face through specific instructions (such as nodding, closing eyes, etc.). Our method is a silent detection method and can be used as an auxiliary classifier in practical applications to increase the accuracy of detection.

For future work, there are additional loss functions, such as consistency loss, and loss function combinations that could be tested to obtain better evaluation metrics in the FAS task. In addition, the current research only used CelebA-Spoofing, OULU-NPU, CASIA-MFSD and Replay attacks. Part of our future work will consider using SiW [22] to enhance the model evaluation and increase the model's generalisation abilities.

Author Contributions: Conceptualisation, L.Y.G. and X.J.L.; methodology, L.Y.G. and X.J.L.; software, L.Y.G.; validation, L.Y.G.; formal analysis, L.Y.G.; investigation, L.Y.G.; resources, L.Y.G. and X.J.L.; data curation, L.Y.G. and X.J.L.; writing—original draft preparation, L.Y.G. and X.J.L.; writing—review and editing, L.Y.G. and X.J.L.; visualisation, L.Y.G. and X.J.L.; supervision, X.J.L.; project administration, X.J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
2. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
3. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807. [CrossRef]
4. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A unified embedding for face recognition and clustering. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 815–823. [CrossRef]
5. Yu, Z.; Qin, Y.; Li, X.; Zhao, C.; Lei, Z.; Zhao, G. Deep Learning for Face Anti-Spoofing: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 5609–5631. [CrossRef]
6. Xu, X.; Xiong, Y.; Xia, W. On Improving Temporal Consistency for Online Face Liveness Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seattle, WA, USA, 13–19 June 2020.
7. Liu, A.; Wan, J.; Escalera, S.; Escalante, H.J.; Tan, Z.; Yuan, Q.; Wang, K.; Lin, C.; Guo, G.; Guyon, I.; et al. Multi-Modal Face Anti-Spoofing Attack Detection Challenge at CVPR2019. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019; pp. 1601–1610. [CrossRef]
8. Liu, A.; Tan, Z.; Wan, J.; Liang, Y.; Lei, Z.; Guo, G.; Li, S.Z. Face Anti-Spoofing via Adversarial Cross-Modality Translation. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 2759–2772. [CrossRef]
9. Khammari, M. Robust face anti-spoofing using CNN with LBP and WLD. *IET Image Process.* **2019**, *13*, 1880–1884. [CrossRef]
10. Li, L.; Feng, X.; Boulkenafet, Z.; Xia, Z.; Li, M.; Hadid, A. An original face anti-spoofing approach using partial convolutional neural network. In Proceedings of the 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA), Oulu, Finland, 12–15 December 2016; pp. 1–6. [CrossRef]
11. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random Erasing Data Augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13001–13008. [CrossRef]
12. Vaswani, A. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 6000–6010.
13. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002.
14. Guo, J.; Deng, J.; Lattas, A.; Zafeiriou, S. Sample and Computation Redistribution for Efficient Face Detection. *arXiv* **2021**, arXiv:2105.04714.
15. Zhang, Y.; Yin, Z.-f.; Li, Y.; Yin, G.; Yan, J.; Shao, J.; Liu, Z. CelebA-Spoof: Large-Scale Face Anti-Spoofing Dataset with Rich Annotations. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.
16. Boulkenafet, Z.; Komulainen, J.; Li, L.; Feng, X.; Hadid, A. OULU-NPU: A Mobile Face Presentation Attack Database with Real-World Variations. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 612–618. [CrossRef]
17. Zhang, Z.; Yan, J.; Liu, S.; Lei, Z.; Yi, D.; Li, S.Z. A face antispoofing database with diverse attacks. In Proceedings of the 2012 5th IAPR International Conference on Biometrics (ICB), New Delhi, India, 29 March–1 April 2012; pp. 26–31. [CrossRef]
18. Chingovska, I.; Anjos, A.; Marcel, S. On the effectiveness of local binary patterns in face anti-spoofing. In Proceedings of the 2012 BIOSIG—Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 6–7 September 2012; pp. 1–7.
19. Lin, C.; Liao, Z.; Zhou, P.; Hu, J.; Ni, B. Live face verification with multiple instantialized local homographic parameterization. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018.

20. Liu, Y.; Jourabloo, A.; Liu, X. Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 389–398.. [CrossRef]
21. Jourabloo, A.; Liu, Y.; Liu, X. Face De-spoofing: Anti-spoofing via Noise Modeling. In Proceedings of the Computer Vision—ECCV 2018: 15th European Conference, Proceedings, Part XIII, Munich, Germany, 8–14 September 2018.
22. Spoof in Wild. Available online: https://cvlab.cse.msu.edu/siw-spoof-in-the-wild-database.html (accessed on 10 November 2024).