FIQA-FAS: Face Image Quality Assessment Based Face Anti-Spoofing

Ya-Chi Liang[1] Min-Xuan Qiu[1] Shang-Hong Lai[1]
[1]National Tsing Hua University, Taiwan
judy22prince@gmail.com maisiechiu@gapp.nthu.edu.tw lai@cs.nthu.edu.tw

Abstract

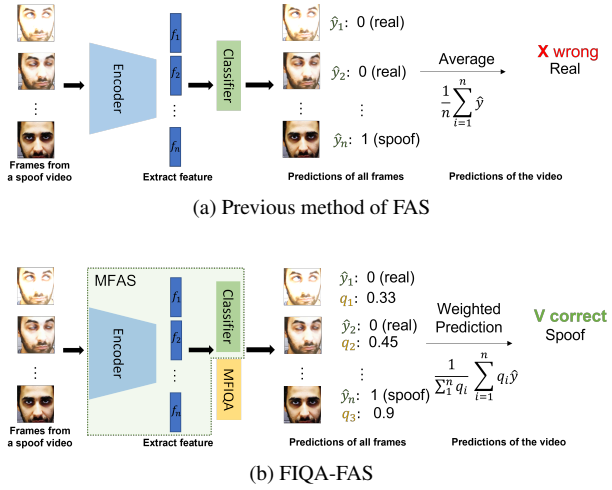
Face anti-spoofing (FAS) is to protect facial recognition systems against presentation attacks. However, recent research on FAS often neglects real-world conditions, such as changing illumination, varying angles of face, and motion blur within a video. These conditions lead to inconsistent feature quality across face images, where low-quality features can cause the model to learn unreliable information during training. Moreover, frames with low feature quality within videos result in inaccurate decisions. To address this issue, we propose the Face Image Quality Assessment Based Face Anti-Spoofing System (FIQA-FAS), which integrates a face image quality assessment module with a face anti-spoofing module. FIQA-FAS assesses the feature quality extracted from each face image and uses the quality score to compute a weighted prediction for deciding if the face in a video is live or spoof. We demonstrate the effectiveness of FIQA-FAS through experiments on the SIW and SIW-M datasets. To further demonstrate our model's capabilities, we introduce a novel simulated scenario that mimics the real world, where our model outperforms other SOTA.

1. Introduction

As technology advances rapidly, facial recognition has become a ubiquitous tool across various applications, offering unprecedented convenience and accuracy for tasks such as smartphone unlocking, self-service border control, and mobile payments. However, this convenience comes with significant security challenges. Facial recognition systems are susceptible to a variety of Presentation Attacks (PAs), including replay, print, and 3D mask attacks, posing serious threats to the integrity of these systems. In response, Presentation Attack Detection, also known as Face Anti-Spoofing (FAS), has been developed to effectively distinguish between real and spoof faces, thereby increasing the overall security of these systems.

In real-world scenarios, some frames within a video may be affected by intense movement or changes in lighting conditions, resulting in blurriness or abnormal illumination in some frames, thus affecting their feature quality. Frames with low feature quality lack crucial identity information, potentially degrading the performance of facial recognition. To mitigate this issue, many recent facial recognition works [13, 16] incorporate Face Image Quality Assessment (FIQA). These studies assess the feature quality of each frame, prioritizing those with more identity information. The strategy guarantees the stable and dependable performance of facial recognition.

However, integration of FIQA with FAS is rare, failing to account for the feature quality across video frames. Fig. 1a demonstrates that frames of low feature quality are often incorrectly predicted due to the absence of spoof cues, re-



(a) Previous method of FAS

(b) FIQA-FAS

Figure 1. Comparison of previous FAS methods with ours for inference. (a) Previous methods use an encoder to encode the features of all frames, classify each as real or spoofed, and average the results. This equal treatment of low- and high-quality frames can lead to errors in assessing video authenticity. (b) FIQA-FAS obtains features from the M_{FAS} encoder and passes them in parallel to a classifier and M_{FIQA}, generating real or spoofed decisions $\hat{y}_i$ and feature quality scores q_i for each frame. By weighting the prediction with q_i, frames with more significant spoof cues exert greater influence, enhancing the accuracy in distinguishing real from spoofed videos.

sulting in diminished prediction accuracy overall.

To address this issue, we propose an innovative Face Image Quality Assessment Based Face Anti-Spoofing(FIQA-FAS) that integrates the face image quality assessment module (M_{FIQA}) with the face anti-spoofing module (M_{FAS}). During training, M_{FIQA} outputs the quality scores to ensure that frames with varying feature qualities contribute differently to the loss function, as illustrated in Fig. 2. During inference, higher feature quality frames with clear spoofing cues are prioritized; conversely, lower quality frames likely to lead to incorrect predictions have their influence minimized, as illustrated in Fig. 1b. Moreover, our method does not rely on human-defined quality labels and utilizes very simple network architecture. To sum up, FIQA-FAS efficiently generates quality scores that reflect spoof cues and significantly enhances the reliability of detecting spoofing attempts, even in dynamically changing real-world scenarios.

The FAS benchmark datasets only capture diverse environments and lighting conditions across different videos and fail to collect the rapid changes that can occur within a single video. It does not reflect real-world scenarios. Therefore, we introduce a simulated scenario into the SIW [17] and SIW-M [18] datasets to better mimic real-world conditions. Additionally, we compare the currently available state-of-the-art Face anti-spoofing methods [6, 11, 28] under this scenario. Our approach achieves superior results compared to these methods.

In summary, this paper makes the following contributions:

1. The proposed FIQA-FAS method is the first work that combines face image quality assessment with face anti-spoofing, prioritizing high-quality frames for enhanced anti-spoofing accuracy.

2. FIQA-FAS outperforms SOTA methods in experiments on simulated real-world scenarios, demonstrating its superiority in different environmental conditions.

2. Related Work

2.1. Face Anti-Spoofing

Several earlier studies [2, 3, 5, 9, 22] have utilized texture analysis techniques to distinguish live faces from spoofs. These studies relied on manually created descriptors for identifying spoof textures, employing techniques such as LBP, HOG, SURF, and SIFT. Later, some methods [12, 24] based on mixing hand-crafted and deep learning were proposed to achieve more reliable performance. However, these hand-crafted descriptors are not robust enough to be used for different scenarios, When the light source or background changes, the performance of such methods is likely to degrade significantly. Moreover, they are highly dependent on specific expert knowledge.

More recently, numerous learning-based methods [6, 15, 19, 28] have been proposed to detect presentation attacks, achieving remarkable performance. They use CNN encoders to extract features with spoof cues and treat it as a binary classification task, labeling "0" and "1" for real and spoof faces. [6] proposed extracting features from face images using an auxiliary classifier to aid the encoder in focusing on spoof cue extraction. Wang *et al*. [28] proposed disentangled representation learning to learn the spoof-related feature through two-stage training. However, these FAS methods extract features without considering the reliability of these features. When a video contains blurriness, occlusion, or poor lighting, the spoof cues in the features of some frames are poor. These features are not ideal for making a good prediction and may lead the model to learn useless information.

While [7, 8] consider image quality, they utilize it as the criterion for distinguishing between live and spoof faces, rather than assigning lower importance to low-quality images.

Considering the above factors, we believe that the FAS system needs to address the issue of low-quality features, which are unreliable for face anti-spoofing. Thus, we propose FIQA-FAS to assess the quality of features and determine whether the frame's prediction contains sufficient spoof cues to make a correct prediction.

2.2. Face Image Quality Assessment

Unconstrained face recognition has always been challenging due to the presence of low-quality face images. Traditional Face Image Quality Assessment(FIQA) methods, such as the ISO/IEC 19794-5 and ICAO 9303 standards, define the quality standards for facial images under conditions of occlusion, blur, etc. In recent years, many methods have focused on learning-based research, such as FaceQNet [10] and Best-Rowden [1]. These approaches generate quality measures through network regression trained from human-labeled data. However, these quality labels are error-prone because they use these hand-crafted labels as ground truth to train the network. Furthermore, this approach is not entirely objective, as humans may not accurately assess the features most important for recognition systems. Considering only the similarity between two pictures can also lead to errors due to differences in age or attire of the same person. As a result, these approaches may overlook some factors that affect quality scores. From an efficiency standpoint, the method based on manual labeling can be very time-consuming.

Since the above methods are not only time-consuming and inefficient but also lack objectivity, in recent years, many learning-based FIQA methods have been developed

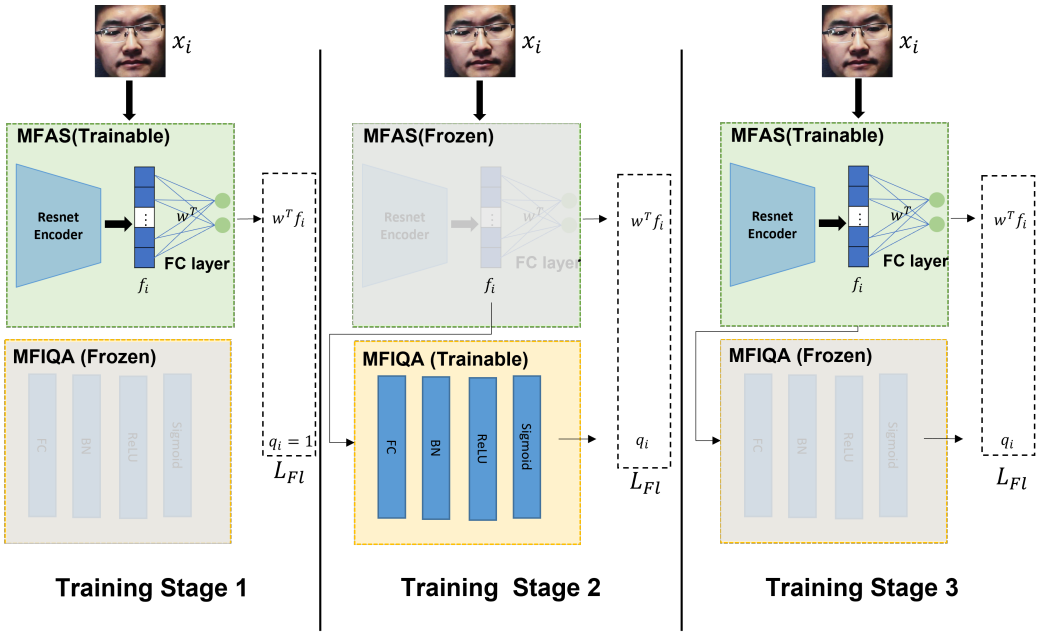


Figure 2. Explanation of the FIQA-FAS training pipeline. During Stage 1, M_{FAS} is trainable while M_{FIQA} is frozen, with all q_i set to 1. In Stage 2, M_{FIQA} becomes trainable, and M_{FAS} is frozen, employing the weights from Stage 1 training. The f_i extracted by the M_{FAS} encoder is input into M_{FIQA} to obtain the feature quality q_i. In Stage 3, M_{FAS} becomes trainable again, using M_{FIQA} weights trained in Stage 2, which remain frozen. The quality score q_i during this stage is derived from the output of M_{FIQA} for each f_i.

that are trained without explicit quality score labels. Examples include PFE [25], ADRL [23], SER-FIQ [27], SDD-FIQA [20], EQFace [16], and AdaFace [13]. [20] proposed a novel learning-based method, employing a recognition model to collect the intra-class and inter-class similarity distribution. They used the Wasserstein Distance [21] between the two distributions to denote the quality pseudo-label. [13] proposed an adaptive margin loss function that assigns varying importance to images based on their quality, assuming that the relative importance of easy or hard samples should correspond to image quality. This approach emphasizes easy samples when image quality is poor to avoid unidentifiable images. The paper [16], most closely related to our method and a significant inspiration for our network design, proposed a simple explicit quality network for face recognition, providing a quantitative and clear quality value when extracted by a feature vector. This study claims to be the first to implement these two functions in a network.

These studies state that the quality of a face image indicates its reliability for recognition performance. They obtain good recognition results by integrating image quality with face recognition. While some methods provide a clear face quality score and achieve good performance, they are specifically designed for face recognition systems to evaluate image quality. Furthermore, these methods have not yet been applied in Face Anti-Spoofing.

3. Proposed Method

Since there are varying spoof cues across frames in the real world, not every frame's feature and predictions are reliable. Face Image Quality Assessment Based Face Anti-Spoofing (FIQA-FAS) addresses this issue by incorporating two modules: the Face Image Quality Assessment Module (M_{FIQA}) and the Face Anti-Spoofing Module (M_{FAS}). M_{FAS} is designed to distinguish whether each frame is real or a spoof, while M_{FIQA} evaluates the quality scores of features, q_i, extracted by M_{FAS} from each frame. q_i serves as a weighted factor in both the loss function and the weighted prediction mechanism we designed. By prioritizing higher-quality frames, FIQA-FAS focuses more on spoof-related features, achieving robust performance in distinguishing live videos from spoofed.

Sec. 3.1 details the architecture of the M_{FAS}. Sec. 3.2 explores the architecture of the M_{FIQA}. The utilized loss function is discussed in Sec. 3.3. Sec. 3.4 provides a thorough overview of the entire training pipeline. This section introduces a three-stage training process that ensures the quality score output from M_{FIQA} accurately reflects the reliability of the features. Finally, Sec. 3.5 explains the weighted prediction mechanism we have designed and details how these two modules work during inference to accurately identify videos as live or spoofed.

3.1. Module of Face Anti-Spoofing

The Module of Face Anti-Spoofing(M_{FAS}) primarily focuses on extracting features from the input frames for predicting real or spoof by utilizing a ResNet101-based architecture. Let x_i denote the i-th input frame, where i ranges from 1 to n, with n being the total number of images. The ResNet encoder extracts feature $f_i \in \mathbb{R}^{512}$ for frame x_i.

During training, the feature f_i is simultaneously directed to the Fully Connected (FC) layer, undergoing multiplication with the FC layer's weights $w \in \mathbb{R}^{512\times 2}$, and to the M_{FIQA} module to determine the feature quality score, q_i, which ranges from 0 to 1, as illustrated in Fig. 2.

During inference, the feature f_i is directed to the FC layer of M_{FAS} for real or spoofed prediction, $\hat{y}_i \in (0, 1)$, indicating whether x_i is real or spoofed. Additionally, f_i is fed to the M_{FIQA} module to obtain the feature quality score, q_i, as shown in Fig. 1b.

3.2. Module of Face Image Quality Assessment

Frames with high-quality features, capable of achieving more accurate predictions and containing an abundance of spoof cues, significantly enhance the model's reliability. Conversely, features from low-quality frames often incorporate irrelevant data into the model, impairing its spoof detection capabilities. To address this issue, the Face Image Quality Assessment Module (M_{FIQA}) aims to evaluate each feature vector f_i extracted by M_{FAS}, as mentioned in Sec. 3.1, assigning a specific quality score q_i.

The process begins with f_i being forwarded to a Fully Connected (FC) layer, immediately followed by a Batch Normalization (BN) layer, and then activated using the ReLU function. It then proceeds to a second FC layer and a sigmoid layer, ensuring the quality value is normalized between 0 and 1, thus obtaining the quality score q_i for f_i. This entire process can be referenced in Fig. 2.

This quality score, q_i, is then incorporated into the loss function with the feature vector f_i, as specified in Eq. (3), and also merged with the prediction $\hat{y}_i$ for weighted score prediction during inference, as detailed in Eq. (5).

M_{FIQA} enables the model to prioritize high-quality frames throughout the training and inference stages, ensuring a focus on the most informative frames. Moreover, the module's architecture is simple yet significantly aids in making the face anti-spoofing decision more robust.

3.3. Loss Function

The following Eq. (1) presents the general softmax loss. Let $f_i \in \mathbb{R}^{512}$ denote the feature vector extracted by M_{FAS} for the input frame x_i, with the ground truth y_i belonging to class j, where $j \in (0, 1)$ represents real (0) or spoof (1). The weights of the Fully Connected layer in M_{FAS}, denoted by $w \in \mathbb{R}^{512\times 2}$, match the 512-dimensional feature vector f_i and the two face anti-spoofing classes. $w_j \in \mathbb{R}^{512}$ denotes the weight vector associated with class j. The variable n represents the number of samples in this batch.

$$L_{sf} = \sum_{i=1}^{n} -log\frac{e^{w_{y_i}^T f_i}}{\sum_{j=0}^{1} e^{w_j^T f_i}} \tag{1}$$

To prioritize higher-quality features, Eq. (2) incorporates the predicted quality score q_i as a confidence weight into the softmax loss, inspired by EQ-face [16]. q_i, ranging from 0 to 1, denotes the quality score of the extracted feature f_i, and is used as a confidence weight. A higher q_i results in more loss, encouraging the model to focus on features with higher quality scores.

$$L_{sf\_weighted} = \sum_{i=1}^{n} -log\frac{q_i \cdot e^{w_{y_i}^T f_i}}{\sum_{j=0}^{1} q_i \cdot e^{w_j^T f_i}} \tag{2}$$

To address the issue of the imbalanced number of samples between real and spoof videos in existing face anti-spoofing datasets, we integrate $L_{sf_{weighted}}$ with focal loss to further enhance the model's learning capability. See Eq. (3)

$$L_{FL} = \sum_{i=1}^{n} -(1 - \frac{q_i \cdot e^{w_{y_i}^T f_i}}{\sum_{j=0}^{1} q_i \cdot e^{w_j^T f_i}})^{\gamma} log\frac{q_i \cdot e^{w_{y_i}^T f_i}}{\sum_{j=0}^{1} q_i \cdot e^{w_j^T f_i}} \tag{3}$$

3.4. Training Pipeline

To ensure that q_i accurately reflects the quality of features in face images, we have developed a training pipeline, drawing inspiration from [16, 26]. This pipeline is depicted in Fig. 2 and involves a three-stage training process for the models M_{FAS} and M_{FIQA}

Stage 1: Train M_{FAS} and Freeze M_{FIQA}. In the first stage, the primary objective is to enhance the capability of M_{FAS} in extracting features for classifying images as either real or spoofed. At this point, the weights of M_{FIQA} are frozen, and q_i in Eq. (3) is set to 1 for every face image. This configuration allows the whole module to concentrate on processing and estimating the features of real and fake images.

Stage 2: Train M_{FIQA} and Freeze M_{FAS}. In this stage, the M_{FAS} module is frozen while M_{FIQA} becomes trainable. Additionally, we set q_i to the output of M_{FIQA}. Samples that result in a lower value in Eq. (1) are considered more informative regarding spoof features, as they contribute to more accurate predictions. This suggests that such samples should be given more attention by the model during training and deserve a higher q_i for their reliability. Since M_{FIQA} aims to minimize the loss L_{FL}, as described in Eq. (3), it assigns higher q_i to samples with a lower value in Eq. (1) and conversely, lower q_i to those with a higher value. Therefore, in this second stage, we successfully prioritize

higher-quality features by making them more influential in the model's learning process, while reducing the impact of lower-quality features. This stage empowers M_{FIQA} with the capability to accurately assign quality scores to features extracted by M_{FAS}, which can reflect spoof cues.

Stage 3: Train M_{FAS} and Freeze M_{FIQA}. In the final stage, we freeze M_{FIQA} and train M_{FAS}. Since M_{FIQA} assigns a relative quality value q_i to each feature, the contribution to the loss of a feature is based on the quality score generated by M_{FIQA}. This means the model can focus more on samples with high-quality features.

The Reason for the Three-Stage Training. We believe it's essential to clarify why we adopt a three-stage training pipeline. If M_{FAS} and M_{FIQA} were trained simultaneously from the start, the values of q_i would progressively diminish to decrease the loss during training, eventually nearing 0 for every feature. Consequently, q_i could not accurately reflect feature quality.

Hence, the initial stage is designed to fix q_i at 1 and to train the M_{FAS} network independently. This stage enables the network to make reasonably accurate predictions for both live and spoofed face images. In the second stage, face images that are more prone to misclassification suggest they contain features that could lead to inaccurate prediction, indicating their quality is comparatively lower. Therefore, the features of these images should not be heavily weighted. Conversely, an image likely to be predicted correctly is of better quality. M_{FIQA} will learn the suitable quality score for each feature. Finally, once q_i is capable of differentiating between face images of varying qualities, we freeze the parameters of M_{FIQA} and utilize the q_i assessed by M_{FIQA} as the weight for features f_i from M_{FAS}. This strategy ensures that the model pays more attention to high-quality features during training.

3.5. Weighted Score Prediction

In previous face anti-spoofing work, researchers typically averaged all frame predictions within a video to obtain the video prediction by Eq. (4). X represents the set of input frames from a video, with n denoting the total number of frames contained within X.

$$Predict(X) = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i \tag{4}$$

However, as shown in Fig. 1a, previous methods cannot prioritize more reliable frames to make more precise predictions for a video, potentially yielding incorrect results due to the impact of low-quality feature predictions. To fully leverage the module, as illustrated in Fig. 1b, FIQA-FAS integrates the quality scores into the prediction process. It uses the following equation to determine whether an entire video is real or spoof:

$$Weighted\_Predict(X) = \frac{\sum_{i=1}^{n} q_i\hat{y}_i}{\sum_{i=1}^{n} q_i} \tag{5}$$

In Eq. (5), x_i represents the i-th frame within X. The prediction for frame x_i, derived from M_{FAS}, is denoted by $\hat{y}_i$, where $\hat{y}_i \in (0, 1)$ indicates whether x_i is real or spoofed. The quality score, q_i, evaluated by M_{FIQA} for the feature f_i extracted from x_i by M_{FAS} and ranging from 0 to 1, acts as a weighting factor. We predict whether the video is real or spoofed by taking a weighted average of the predictions for all frames, as described in Eq. (5).

4. Experiments

In this section, we evaluate our method's performance across SIW [17], SIW-M [18], and simulated scenarios.

4.1. Implementation Detail

Both training and testing data will be cropped by the face detector Dlib [14] before being fed into our system. These cropped images will then be uniformly resized to 112 $\times$ 112. In the optimizer settings, we selected SGD to optimize our model and set the initial learning rate to 0.05. The weight decay is set at 5e-4, and the minimum learning rate value is 1e-5. Our training batch size is set to 256. We then set the upper limit of the training epochs to 200 for each stage. Our method is implemented in PyTorch, and the γ in Eq. (3) is set to 2.

4.2. Databases and Simulated Scenarios

SIW [17]: The SiW dataset contains live and spoof videos, featuring 165 subjects. Each subject has 8 live videos and 20 spoof videos, totaling 4,620 videos. The live videos were collected over several sessions, during which subjects may move backward and forward, change the yaw angle of their heads within a range of -90 to 90 degrees, make different facial expressions, or be recorded with additional lighting variation within a video. This setup closely mimics real-world conditions and is suitable for our testing. Protocol 1 of the SiW dataset aims to evaluate the generalization of presentation attack (PA) detection methods under various face poses and expressions. Protocols 2 and 3 are designed to assess the generalization capability across media of the same spoof type and the performance against unknown PAs, respectively.

SiW-M [17]: The SiW-M dataset comprises 660 live videos from 493 subjects and 968 videos depicting 13 types of spoofing attacks from 700 subjects. Similar to SiW, SiW-M exhibits great diversity in environmental conditions such as pose, lighting, and expression within videos, and it also demonstrates excellent diversity in attack types and the

Method	protocol 1	low blurriness	middle blurriness	high blurriness	low lightness w/ noise	high lightness w/ noise
LGSC [6]	0.002	0.008	0.065	0.117	0.063	0.057
Dual-stage [28]	**0.000**	0.051	0.076	0.192	0.075	0.123
SSDG [11]	**0.000**	0.008	0.036	0.052	**0.054**	0.091
Ours-without WP	0.012	0.043	0.057	0.073	0.085	0.056
Ours-with WP	0.003	**0.004**	**0.009**	**0.017**	**0.054**	**0.045**

Table 1. The comparison to the SOTA on the SiW under protocal 1 and 5 simulated scenarios with ACER.

Method	custom protocol	low blurriness	middle blurriness	high blurriness	low lightness w/ noise	high lightness w/ noise
LGSC [6]	0.048	0.073	0.071	0.057	**0.056**	0.077
Dual-stage [28]	0.062	0.059	0.083	0.150	0.088	0.073
Ours-without WP	0.075	0.069	0.071	0.071	0.086	0.099
Ours-with WP	**0.044**	**0.052**	**0.054**	**0.056**	0.082	**0.070**

Table 2. The comparison to the SOTA on the SiW-M under custom protocol and 5 simulated scenarios with ACER.

number of subjects. The dataset defines leave-one-out testing protocols, which entail training the model with 12 types of spoofing attacks and 80% of the live videos and testing on the remaining attack type plus 20% of the live videos. However, our goal is to evaluate our method's performance under various poses, lighting conditions, and expressions within a video. Therefore, we customized a SiW-M protocol for our method by dividing the entire SiW-M database into new training and testing datasets in an 8:2 ratio that contains all spoofed attacks. In our custom protocol, subjects that appear in the training set do not appear in the testing set.

Simulated Scenarios of SIW and SiW-M To assess the effectiveness of FIQA-FAS under real-world conditions, we randomly selected portions of frames from every video and applied various image processing techniques to both SIW Protocol 1 and the custom SIW-M protocol. These techniques include applying three levels of blurriness, two levels of lightness, and additional noise. This simulated scenario aims to replicate real-world challenges, such as abrupt lighting changes, head pose alterations, or motion blur from moving subjects within a single video. Following is the detail of how we performed image processing.

For blurriness, we utilized the Gaussian Blur function from OpenCV to create images with low, medium, and high levels of blur, mimicking the effect of low-resolution input images. To adjust lightness, we applied Gamma Correction, also known as Power Law Transform, simulating conditions 1.5 times darker and brighter to represent both low and high lightness. Furthermore, since real-life images often suffer from degradation due to noise from imaging devices and external environmental factors during transmission, we introduced Gaussian noise to our lightness-adjusted data. We set the noise's standard deviation to 0.1 and the mean value to 0. Samples from the simulated scenarios can be viewed in Fig. 3.

Frames within the same videos, both processed and unprocessed, underwent prediction by Eq. (5) when determining the videos' authenticity and we evaluated our performance using Average Classification Error Rate(ACER). It's noteworthy that our goal is to assess whether we can handle significant variations in different frames within the same video, leading to varied feature quality for each frame. Hence, our testing focused solely on SIW Protocol 1, the custom protocol for SIW-M, and the Simulated Scenario for both SIW and SIW-M. Although the Replay Attacks dataset [4] is consistent with our scenario, current research, in particular intra-dataset testing on Replay Attacks, has achieved an ACER of 0.0. We have achieved the same level of performance, so we will not discuss this result in the following sections. Additionally, since we introduce new scenarios for face anti-spoofing, which involve distinct training and testing data, the comparison methods in our paper are limited to studies that have released their training and inference code. All the results in the following section will be presented by video-level ACER.

4.3. Experiments Results

Experiments on SiW. We conducted comparisons between our method and other state-of-the-art (SOTA) methods on the SIW dataset. Under Protocol 1, our method performed comparably to the SOTA methods. However, in the simulated scenarios described in Sec. 4.2, our method significantly outperformed these methods. The results demonstrate that FIQA-FAS can effectively handle real-world scenarios that may present sudden changes in illumination, motion blur, and noise within a video. The results are presented in Tab. 1. **Experiments on SiW-M.** We compared our method with state-of-the-art (SOTA) methods using the custom protocol to determine whether incorporating quality information offers an improvement. As illustrated in Tab. 2, our method outperformed these methods. Moreover, in simulated scenarios that included variations in blurriness and lightness, as well as the introduction of Gaussian noise, our method outperformed the others in most scenarios, particularly demonstrating the lowest error rate across different

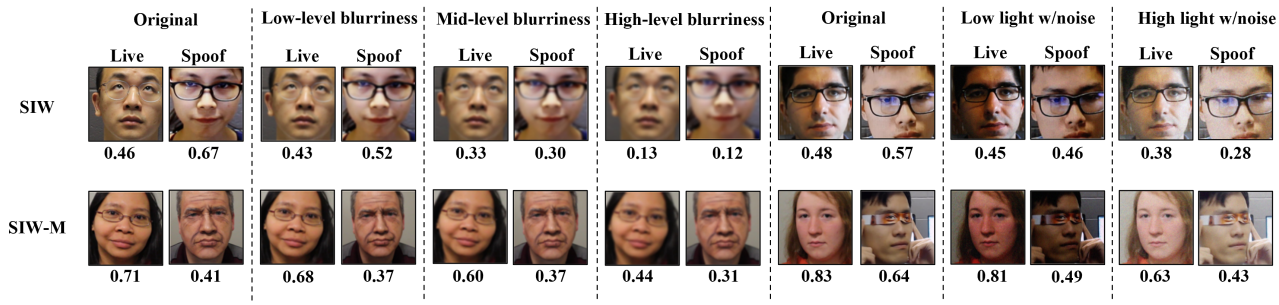


Figure 3. Display the original live and spoofed images, along with three blur simulations and two lightness adjustments. The top row samples are from SIW, and the bottom row is from SiW-M. Under each image, display the quality scores, qi, generated by M_{FIQA}

levels of blurriness. The results demonstrate that FIQA-FAS is more suitable for real-world scenarios than other methods.

5. Ablation Study

5.1. Effectiveness of Training Strategy

Tab. 3 shows our ablation study with and without certain components in our method and with different backbones. The ablation study was performed on the original SiW Protocol 1 and the custom SiW-M protocol. If the model is trained without the quality module or without using focal loss, relying solely on general cross-entropy, we observe a considerable degradation in accuracy. This indicates that the quality assessment module and the application of quality scores to the loss function are crucial. In addition, ResNet101 improves performance, which is why it was chosen as our backbone. We also visualize the feature distribution using t-SNE for the settings mentioned above, as shown in Fig. 4. The figure illustrates the feature distribution from our ablation studies on the SIW in the upper row and SIW-M in the bottom row. Each point in the figure represents a frame: red points for real data and purple points for fake data. The visualization includes results from our full method, variations with the backbone changed to ResNet50 or ResNet152, and our method trained without the M_{FIQA} and without employing focal loss. Our method demonstrates a less overlapping region, indicating a clearer decision boundary.

5.2. Effectiveness of Weighted Prediction in Face Anti-Spoofing

In Tab. 1 and Tab. 2, the results of 'Ours-without WP' are based on Eq. (4), while 'Ours-with WP' is based on Eq. (5). These results demonstrate that using quality scores as weighted factors for each frame's prediction to decide the authenticity of videos performs better than simply averaging all frame predictions. To demonstrate the advantages of incorporating image quality into our testing process, we tested our method by retaining only the top 75%, 50%, and 25% of high-quality test frames on SIW. As shown in Tab. 4, the results indicate that retaining the top 25% of frames yields the best performance. This suggests that lower-quality data may not contribute effectively to better discrimination. We can emphasize the importance of prioritizing higher-quality frames for improved performance.

Settings	SiW	SiW-M
Ours-trained without MFIQA	0.016	0.084
Ours-trained without focal loss	0.004	0.113
Ours-trained on resnet50	0.004	0.069
Ours-trained on resnet152	0.010	0.052
Ours-trained on resnet101	**0.003**	**0.044**

Table 3. Ablation study of our training strategy.

Method	SiW-prorocol1	SiW with low blurriness	SiW with middle blurriness	SiW with high blurriness
Ours-75% frames	**0.0**	0.003	0.005	0.015
Ours-50% frames	**0.0**	0.003	0.003	0.007
Ours-25% frames	**0.0**	**0.002**	**0.002**	**0.005**

Table 4. Evaluation of our method by retaining top 75%, top 50%, and top 25% high-quality test frames on the SIW dataset during inference.

5.3. Image Quality Scores

Quality scores were collected from the original SiW and SiW-M datasets, as well as from these datasets under the simulated scenarios mentioned in Sec. 4.2. The distribution of these quality scores is visualized in Fig. 5. These figures reveal that, under various scenarios, image processing affecting lightness results in a greater number of low-quality frames and fewer high-quality frames compared to processing affecting blur. This finding matches with our experimental results in Tab. 1 and Tab. 2, which show that blurriness does not impact the ACER as significantly as lightness adjustments do, especially when compared to the ACER without image processing (SIW Protocol 1 and the custom protocol of SIW-M).

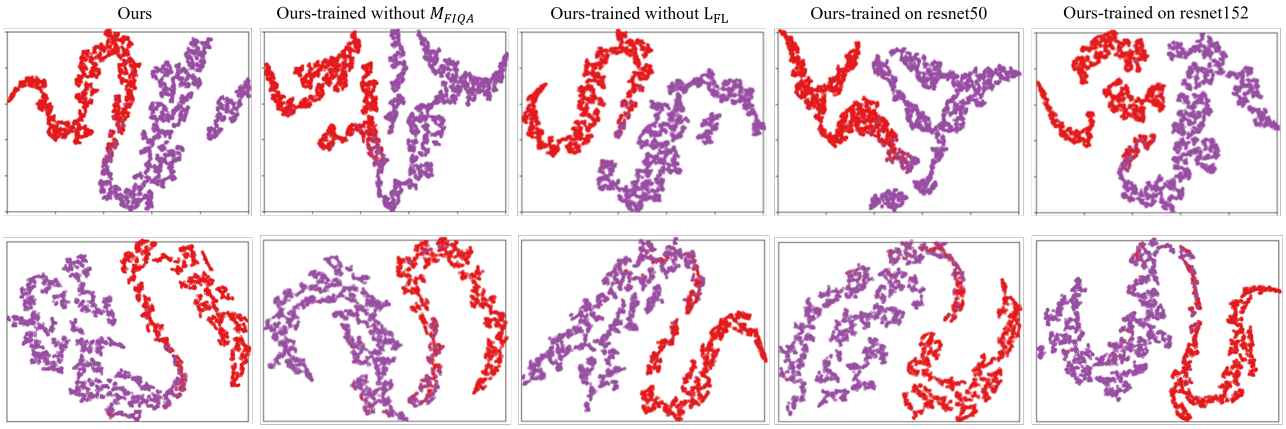
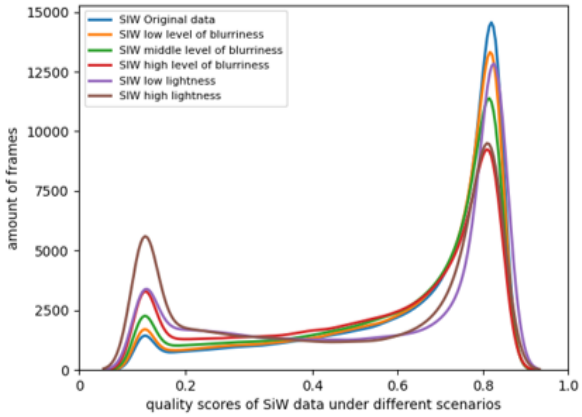
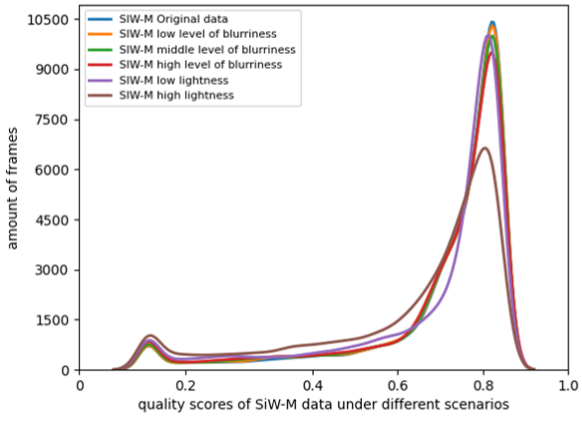


Figure 4. The figure displays feature distribution via t-SNE from ablation studies on SIW (upper row) and SIW-M (bottom row). In the visualization, red points indicate real data, and purple points represent fake data. It shows outcomes for our complete method, adjustments with ResNet50 or ResNet152 backbones, and versions omitting the M_{FIQA} or focal loss.



(a) SiW



(b) SiW-M

Figure 5. Feature quality distribution of image in SiW and SIW-M as well as with simulation scenarios, estimated by M_{FIQA}. Simulated data has lower image quality than the original.

In addition, in Fig. 3, we present the practical quality scores generated by M_{FIQA} for face images within the SIW and SiW-M datasets. Images affected by poor lightness and greater blurriness receive lower quality scores.

These findings affirm the practical meaning of our quality scores, supporting the concept that frames compromised by poor lightness and blurriness can lead to a higher error rate. Consequently, such frames should not be given priority during training and inference.

6. Conclusions

Given the real world's frequent abrupt changes in lighting and motion blur, leading to videos with frames of varying feature quality and potential for incorrect predictions, we introduced FIQA-FAS. This pioneering system integrates face image quality assessment with face anti-spoofing, prioritizing high-quality frames to enhance video-based face anti-spoofing accuracy. Through clever design of training and inference mechanisms, our quality assessment module achieves robustness in detecting face spoofing, despite its simple architecture. Our method has proven to be more robust than state-of-the-art methods through experiments with realistic scenarios on the SIW and SIW-M datasets. This demonstrates our superior performance and feasibility for real-world scenarios. In future work, we aim to explore our quality module's adaptability by applying it to a range of state-of-the-art FAS anti-spoofing techniques. By incorporating the module directly before the fully connected layer to derive quality scores, we anticipate broadening its application. Additionally, we plan to conduct extensive testing across various public datasets using widely accepted cross-dataset protocols. This will help us assess the generalization capabilities of our model in more diverse scenarios.

References

[1] Lacey Best-Rowden and Anil K. Jain. Automatic face image quality prediction, 2017. 2

[2] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face anti-spoofing based on color texture analysis. In *IEEE International Conference on Image Processing (ICIP), 2015*, pages 2636–2640. IEEE, 2015. 2

[3] Z. Boulkenafet, J. Komulainen, and A. Hadid. Face spoofing detection using colour texture analysis. *IEEE Transactions on Information Forensics and Security*, 11(8):1818–1830, Aug 2016. 2

[4] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*, pages 1–7, 2012. 6

[5] Tiago de Freitas Pereira, Jukka Komulainen, André Anjos, José Mario De Martino, Abdenour Hadid, Matti Pietikäinen, and Sébastien Marcel. Face liveness detection using dynamic texture. *EURASIP Journal on Image and video Processing*, 2014(1):2, 2014. 2

[6] Haocheng Feng, Zhibin Hong, Haixiao Yue, Yang Chen, Keyao Wang, Junyu Han, Jingtuo Liu, and Errui Ding. Learning generalized spoof cues for face anti-spoofing, 2020. 2, 6

[7] Emna Fourati, Wael Elloumi, and Aladine Chetouani. Face anti-spoofing with image quality assessment. In *2017 2nd International Conference on Bio-engineering for Smart Technologies (BioSMART)*, pages 1–4. IEEE, 2017. 2

[8] Javier Galbally and Sébastien Marcel. Face anti-spoofing based on general image quality assessment. In *2014 22nd international conference on pattern recognition*, pages 1173–1178. IEEE, 2014. 2

[9] Diego Gragnaniello, Giovanni Poggi, Carlo Sansone, and Luisa Verdoliva. An investigation of local descriptors for biometric spoofing detection. *IEEE transactions on information forensics and security*, 10(4):849–863, 2015. 2

[10] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. Faceqnet: Quality assessment for face recognition based on deep learning, 2019. 2

[11] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 6

[12] Mohammed Khammari. Robust face anti-spoofing using cnn with lbp and wld. *IET Image Processing*, 13(11):1880–1884, 2019. 2

[13] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 3

[14] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009. 5

[15] Lei Li, Xiaoyi Feng, Zinelabidine Boulkenafet, Zhaoqiang Xia, Mingming Li, and Abdenour Hadid. An original face anti-spoofing approach using partial convolutional neural network. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6, 2016. 2

[16] Rushuai Liu and Weijun Tan. Eqface: A simple explicit quality network for face recognition, 2021. 1, 3, 4

[17] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision, 2018. 2, 5

[18] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing, 2019. 2, 5

[19] Yaojie Liu, Joel Stehouwer, and Xiaoming Liu. On disentangling spoof traces for generic face anti-spoofing. In *In Proceeding of European Conference on Computer Vision (ECCV 2020)*, Virtual, 2020. 2

[20] Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Liujuan Cao, and Yuan-Gen Wang. Sdd-fiqa: Unsupervised face image quality assessment with similarity distribution distance. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7670–7679, 2021. 3

[21] Victor M Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6:405–431, 2019. 3

[22] Keyurkumar Patel, Hu Han, and Anil K Jain. Secure face unlock: Spoof detection on smartphones. *IEEE transactions on information forensics and security*, 11(10):2268–2283, 2016. 2

[23] Yongming Rao, Jiwen Lu, and Jie Zhou. Attention-aware deep reinforcement learning for video face recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3951–3960, 2017. 3

[24] Yasar Abbas Ur Rehman, Lai-Man Po, and Jukka Komulainen. Enhancing deep discriminative feature maps via perturbation for face presentation attack detection. *Image and Vision Computing*, 94:103858, 2020. 2

[25] Yichun Shi and Anil K. Jain. Probabilistic face embeddings, 2019. 3

[26] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6902–6911, 2019. 4

[27] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness, 2020. 3

[28] Yu-Chun Wang, Chien-Yi Wang, and Shang-Hong Lai. Disentangled representation with dual-stage feature learning for face anti-spoofing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1955–1964, January 2022. 2, 6