Surveillance Face Anti-spoofing

Hao Fang, Ajian Liu, Jun Wan, *Senior Member, IEEE*, Sergio Escalera, *Senior Member, IEEE*
Chenxu Zhao, Xu Zhang, Stan Z. Li, *Fellow , IEEE*, Zhen Lei, *Senior Member, IEEE*

***Abstract*—Face Anti-spoofing (FAS) is essential to secure face recognition systems from various physical attacks. However, recent research generally focuses on short-distance applications (*i.e.*, phone unlocking) while lacking consideration of long-distance scenes (*i.e.*, surveillance security checks). In order to promote relevant research and fill this gap in the community, we collect a large-scale *Su*rveillance *Hi*gh-*Fi*delity *Mask* (SuHiFiMask) dataset captured under 40 surveillance scenes, which has 101 subjects from different age groups with 232 3D attacks (high-fidelity masks), 200 2D attacks (posters, portraits, and screens), and 2 adversarial attacks. In this scene, low image resolution and noise interference are new challenges faced in surveillance FAS. Together with the SuHiFiMask dataset, we propose a Contrastive Quality-Invariance Learning (CQIL) network to alleviate the performance degradation caused by image quality from three aspects: (1) An Image Quality Variable module (IQV) is introduced to recover image information associated with discrimination by combining the super-resolution network. (2) Using generated sample pairs to simulate quality variance distributions to help contrastive learning strategies obtain robust feature representation under quality variation. (3) A Separate Quality Network (SQN) is designed to learn discriminative features independent of image quality. Finally, a large number of experiments verify the quality of the SuHiFiMask dataset and the superiority of the proposed CQIL.**

***Index Terms*—Face anti-spoofing, Dataset, Surveillance scenes.**

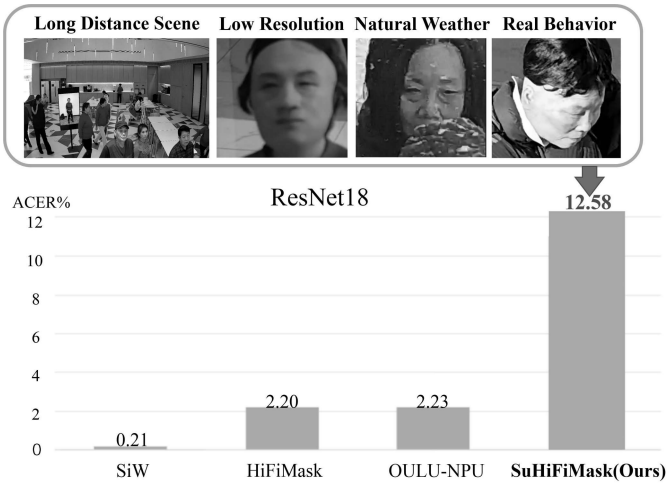

Fig. 1. Performance comparisons on the SiW, HiFiMask, OULU-NPU, and proposed SuHiFiMask dataset using the same ResNet18 network. It shows significant performance degradation under surveillance FAS.

I. Introduction

FACE Presentation Attack Detection (PAD) technology is a crucial step to enhance the security of face recognition systems and plays an increasingly important role in resisting malicious attacks, such as print-attack [1], replay-attack [2], or face-mask [3]. Although current works [4]–[13] have achieved satisfactory performance in short-distance applications, such as phone unlocking, face payment, and access authentication, they are still sensitive to face quality and fail in long-distance applications, which hinders the expansion of FAS to surveillance scenarios.

With the popularity of remote cameras and the improvement of surveillance networks, the development of smart cities has put forward higher requirements for traditional visual technologies in surveillance. Benefited from the release of face recognition datasets [14]–[16] in the surveillance scene and driven by related algorithms [17]–[19], the face recognition system has gradually got rid of the constraint of verification distance, and can use the surveillance camera to complete real-time capture, self-service access control, and self-service supermarket payment. However, the FAS community is still stuck in the protection of the face recognition system under short-distance conditions, and cannot serve for the detection of spoofing faces under a long-distance natural behavior. We analyze two reasons that hinder the development of PAD technologies: (1) **Lack of a dataset that can truly simulate the attack in surveillance.** The existing FAS datasets, whether 2D print or replay attacks [4], [20], [21], or 3D mask attacks [3], [22]–[25], require the subjects to face the acquisition device under distance constraints. However, diversified surveillance scenes, rich spoofing types, and natural human behavior are important assessment factors for the surveillance FAS dataset collection. (2) **Low-quality faces in the surveillance scenarios cannot meet the requirements of fine-grained feature-based FAS tasks.** The existing FAS algorithms, whether based on color-texture feature learning [26]–[29], face depth structure fitting [4], [6], or remote photoplethysmography (rPPG)-based detection [30]–[32], require high-quality image details to ensure high performances. As illustrated in Fig. 1, the resolution of faces under long-distance surveillance is small and contains noise from motion blur, occlusion, bad weather, and other bad factors. These are new challenges for algorithm

Corresponding author: Jun Wan (e-mail: jun.wan@ia.ac.cn).
Hao Fang, Ajian Liu, Jun Wan and Zhen Lei are with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation Chinese Academy of Sciences (CASIA) and School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing, China (e-mail: {fanghao2021, ajian.liu, jun.wan, zhen.lei}@ia.ac.cn).
Xu Zhang is with Beijing Normal University School of Artificial Intelligence, Beijing, China (e-mail: xuzhang0908@mail.bnu.edu.cn).
Chenxu Zhao is with the SailYond Technology, Beijing, China (e-mail: zhaochenxu@sailyond.com).
Sergio Escalera is with the Universitat de Barcelona (UB), Barcelona, Computer Vision Center (CVC), and Aalborg University (AAU) (e-mail: sergio@maia.ub.es).
Stan Z. Li is with Westlake University, Hangzhou, China (e-mail: stan.zq.li@westlake.edu.cn).

design in surveillance FAS.

In order to fill the gap in surveillance scenes of the FAS community, we target to solve two challenging problems analyzed above from two aspects: data collection and algorithm design. In Tab. I, we collect a large-scale FAS dataset based on surveillance scenes, namely SuHiFiMask. It has the following advantages: (a) ***Rich surveillance scenes.*** It includes 40 real surveillance scenes, such as movie theaters, security gates, and parking lots, which cover most face recognition scenes as much as possible. (b) ***Realistic distribution of human faces and natural behavior.*** It involves 101 participants of different ages, and genders distribution participating in the data collection. These subjects perform natural behaviors in daily life. (c) ***Rich spoofing Attacks.*** It has 232 high-fidelity masks (*i.e.* resin, plaster, silicone, headgear, head mold), 200 2D attacks (*i.e.* posters, portraits, and screens), and 2 adversarial attacks. (d) ***Realistic lighting and diverse weather.*** We collect data under real outdoor scenes with different weather (*i.e.*, sunny, snowy day) and light (*i.e.*, day and night).

For the algorithm design, the Contrastive Quality-Invariance Learning network (CQIL) is proposed in Fig. 5, which includes an Image Quality Variable (IQV) module and a two-stream framework consisting of a contrastive learning branch and a Separate Quality Network (SQN) branch. The IQV module is used to recover discriminative information related to FAS in the picture by super-resolution and deliver quality differences in contrast to the contrastive learning network backbone and SQN branch. The contrastive learning backbone [38] contains the online network and the target network. The online network continuously fits the target network during training, learning to approximate the same class with different quality distributions in the shared potential space. The SQN consists of a Quality-Invariance backbone network (CQI) (composed of a central differential convolution operator [7]), a quality discriminator for separating quality, and the main classifier. CQI can effectively extract fine-grained features under environmental changes. The sample pairs generated by IQV are fed into CQI through adversarial learning, which allows CQI to focus on encoding features related to liveness while separating out the interference caused by quality. The main contributions of this paper are summarized below:

- To the best of our knowledge, this is the first work to extend FAS to real surveillance scenes rather than mimicking low-resolution images and surveillance environments. We promote the development of this scenario through data collection and algorithm design.
- We collect a large-scale surveillance FAS dataset, SuHiFiMask, including 101 participants of different ages, 232 masks and 200 2D attacks. A total of $10,195$ videos were collected by 7 mainstream cameras in 40 real scenes.
- We propose a novel Contrastive Quality-Invariance Learning (CQIL) network to enhance the detection of face attacks in surveillance. Among them, an Image Quality Variable (IQV) module is designed to recover the FAS information in images and construct sample pairs to simulate face quality differences in realistic surveillance. A contrastive learning branch to obtain features robust to quality changes. And a Separate Quality Network (SQN) branch based on adversarial learning is introduced to further guide the model to learn quality-independent liveness features.
- Extensive experiments are conducted on the SuHiFiMask and three other public datasets to demonstrate the challenges of the SuHiFiMask and the effectiveness of the proposed method.

II. Related Work

In this section, we review the current FAS works in constrained environments and some preliminary attempts in the surveillance scenes.

FAS under constrained Environments.

Face spoofing (*e.g.*, presentation attacks) is the typical physical attack to deceive the face recognition systems, where attackers present faces from spoof mediums, such as a photograph, screen, or mask, instead of a living human. According to the spoof mediums, we can roughly classify the existing attacks into 2D [4], [20], [21] and 3D attacks [3], [25]. Replay-Attack [2] and CASIA-FASD [1] are early FAS datasets, commonly used as benchmark for domain generalization evaluation. The spoof medium of the former is an electronic screen, while the latter introduces additional paper mediums based on different resolutions. With the advancement of acquisition equipment in mobile phones, there are also some high-resolution datasets recorded by replaying face video with a smartphone, such as Replay-Mobile [39], OULU-NPU [20], and SiW [4]. CelebA-Spoof [40] introduces rich attribute annotation information, which can be used as an auxiliary task to improve the generalization of the model in various attacks. Recently, with the cost reduction of multi-spectral sensors and the popularity of use scenes, some new sensors have been introduced to provide more possibilities for FAS methods. Holger *et al.* [23] use multi-spectral short wave infrared (SWIR) imaging to ensure the authenticity of a face even in the presence of partial disguises and masks. Zhang *et al.* [21] collect a CASIA-SURF dataset with 3 modalities (*i.e.*, RGB, Depth and NIR) using Intel RealSense SR300 camera, and propose a multi-modal multi-scale fusion method for FAS. Similarly, Liu *et al.* [41] introduce a CASIA-SURF CeFA dataset, covering 3 ethnicities, $1,607$ subjects, and $23,538$ videos with 1280×720 resolution. As attack techniques are constantly upgraded, some new types of attacks have emerged, *e.g.*, face mask [3], [25], [29]. Nesli *et al.* [3] provide a 3DMAD which is recorded using the Microsoft Kinect sensor and consists of Depth and RGB modalities with 3D masks. George *et al.* [29] introduce a WMCA database with four channels, *e.g.*, color, depth, near-infrared, and thermal, for face PAD which contains a wide variety of 2D and 3D presentation attacks, and propose MC-CNN method aiming to detect sophisticated attacks with multiple channels information. Heusch *et al.* [42] collect an HQ-WMCA database, which can be viewed as an extension of the WMCA [29] database via adding a new sensor acting in the shortwave infrared (SWIR) spectrum. A large-scale High-Fidelity Mask dataset, namely CASIA-SURF HiFiMask (briefly HiFiMask) was collected by

TABLE I
COMPARISON OF PUBLIC FACE ANTI-SPOOFING DATASETS. * NOTES THAT SUHIFIMASK IS FOCUSED ON SURVEILLANCE SCENES WHERE BOTH REAL PEOPLE AND FAKE ATTACKS APPEAR AT THE SAME TIME. THEREFORE, THE NUMBER OF REAL VIDEOS AND THE NUMBER OF FAKE VIDEOS ARE BOTH 10195.

Dataset, Year	#sub.	Distance (Long/Short)	Materials	Scenes	Light, Weather	Attacks	Devices	#Videos (#Live/#Fake)
3DMAD [33], 2013	17	Short	Paper, Resin	Constrained scenes	Adjustment	2D/2.5D image	Kinect	255(170/85)
3DFS-DB [22], 2016	26	Short	Plastic	Office	Adjustment	2D/2.5D image, 3D Mask	Kinect, Carmine 1.09	520(260/260)
BRSU [23], 2016	137	Short	Silicone, Plastic, Resin, Latex	Disguise, Counterfeiting	Adjustment	2D image	SWIR, Color	141(0/141)
MARsV2 [34], 2016	12	Short	ThatsMyFace, REAL-F	Office	Six directions of light	3D Mask	Logitech C920, Industrial Cam, EOS M3, Nexu 4, IPhone 6, Samsung S7, Sony Tablet S	1008 (504/504)
SMAD [35], 2017	Online	Short	Silicone	-	Varying light	2D image, 3D Mask	Varying Cam	130(65/65)
MLFP [36], 2017	10	Short	Latex, Paper	Indoor, Outdoor	Daylight	2D image	Visible, Near infrared, Thermal	1350 (150/1200)
ERPA [37], 2017	5	Short	Resin, Silicone	Indoor	Room light	3D Mask	Xenic Gobi, Thermal Cam	86
WMCA [29], 2019	72	Short	Plastic, Silicone, Paper	Indoor	Office/LED/Day light	2D image, 3D Mask	Intel RealSense SR 300, Seek Thermal, Compact PRO	1670 (347/1332)
3DMask [26], 2020	48	Short	Plaster	Indoor, Outdoor	Six directions of light	2D image, 3D Mask	Apple, Huawei, Samsung	1152 (288/864)
HiFiMask [25], 2021	75	Short	Transparent, Plaster, Resin	White, Green, Tricolor, Sunshine, Shadow, Motion	Six directions of light	2D image, 3D Mask	IPhone 11, IPhone X, MI10, P40, S20, Vivo, HJIM	54,600 (13,650/40,950)
SuHiFiMask (ours), 2022	101	**Long**	Resin, Plaster, Silicone, Paper	Security check lane, Theater, Parking lot [1]	Day/Night light, Sunny/Windy/ Cloudy/Snowy day	2D image, Video replay, 3D Mask	Surveillance cameras[2]	10,195* (10,195/10,195)

[1] 40 real surveillance environments, including indoor as well as outdoor. Please see Fig.2 in *Appendix* for more details.
[2] dahua: DH-IPC-HFW4843M, DH-P80A1-SA; HIKVISION: DS-2CD3T87WD-L, DS-2CD3T86FWDV2-I3S; TP-LINK: TL-IPC586FP, TL-IPC586HP; ZHONGDUN: ZD5920-Gi4N (Brand name: Camera model) .

Liu *et al.* [25]. Specifically, it consists of a total amount of $54,600$ videos which are recorded from 75 subjects with 7 kinds of sensors. Although the resolution and fidelity of these datasets are increasing high (*i.e.*, resolution from 320×240 [2] to $1,920 \times 1,080$ [4], and spoofing types from print [1] to mask [25]), they are all oriented to FAS in a close constrained environment, ignoring the application requirements of remote surveillance scenes.

The essence of FAS is a defensive measure for face recognition systems and has been studied for over a decade. Early works were mainly based on color texture [2], [43] and motion analysis [44]. The former is based on the consideration that the fake face is different from the live face in texture details, such as color distortions, and specular highlights, due to the intervention of spoofing mediums. However, these algorithms are not accurate enough because of the use of handcrafted features, such as LBP [2], HoG [45], and SURF [46]. The latter analyzes the attack samples as static or non-rigid motion compared with live faces from the perspective of motion. Unfortunately, these methods become vulnerable if someone presents a replay attack or a print attack with cut eye/mouth regions. Instead of using pre-defined features such as LBP and HOG, CNN-based methods [47], [48] design a unified framework of feature extraction and classification in an end-to-end manner. However, they treat FAS as a binary classification task, and will highly depend on the liveness-unrelated cues, such as color distortion, shape deformation, or background information. Intuitively, the live faces in any scene have consistent face-like geometry. Inspired by this, some works [4], [7], [49] leverage the physical-based depth information instead of binary classification loss as supervision, which are more faithful attack clues in any domain. Another works [8], [9], [50]–[52] treat FAS as a feature disentangled representation learning. Although these CNN-based methods achieve near-perfect performance under known attack clues, they still show poor generalization in the face of unknown attacks. To solve this problem, there are also some methods [53]–[56] that focus on improving the generalization of FAS in unknown domains. Examples are MADDG [5], SSDG [57], are SSAN [58], which aim to learn a generalized feature space via adversarial training and triplet loss strategies. In the case of FMeta [59], MT-FAS [60], D^{2}AM [11], and SDA [61], they aim to find the generalized feature directions via meta-learning strategies.

FAS in Surveillance Environments. The task of face recognition in surveillance has been widely concerned by researchers, including data collection and algorithm design. SCface [16] was the first face recognition dataset released to simulate research in surveillance scenes, which contains $4,160$ still images captured by five different quality cameras. The QMUL-Survface dataset [14] further complements the low-resolution face recognition dataset by collecting $463,507$ face images from $15,573$ different identities in the real world using surveillance cameras. Then, IJB-C [15] aims to improve the representation of the global population by adding a list of names containing specific occupations such as artists, public speakers, and journalists from different countries to the surveillance scenario. In addition, based on these datasets, face recognition algorithms for surveillance scenes have been in full swing. Li *et al.* [18] introduce the adversarial generative networks and fully convolutional architectures to recognize ground-resolution faces in supervised discriminative learning. Considering the incompleteness of these datasets, Zhong *et al.* [19] propose a sigmoid-constrained hypersphere loss (SFace) to reduce the intra-class distance of high-quality samples while preventing over-fitting label noise. Kim *et al.* [17] propose an adaptive marginal function to adjust the importance of different samples by emphasizing the role of clean samples in classification.

In the FAS community, Chen *et al.* [62] explore the face anti-spoofing in surveillance scenes for the first time and proposed a dataset and benchmark. As for the dataset, they release the GREAT-FASD-S, which is first collected by two

multi-modal cameras, and then processed into low-quality images. And for the method, they propose the DAM-SE module to select the most informative channels and recover the image with the nearest neighbor interpolation algorithm. Aravena *et al.* [63] demonstrates that discarding a suitable percentage of low-quality samples can effectively improve the performance of the PAD algorithm. However, the nearest neighbor interpolation algorithm can not recover the original information by filling pixels with low-resolution images, and the method of directly discarding low-quality samples does not directly face the challenge of FAS in surveillance scenes.

Fig. 2. An overview of some characteristics of SuHiFiMask. From top to bottom: forms of attack, diverse weather, and surveillance scenes.

III. SuHiFiMask

In order to fill the gap in the face anti-spoofing dataset of surveillance scenes and promote the research of related algorithms, we collected the SuHiFiMask dataset that has the following advantages over existing datasets:

Advantage 1: To the best of our knowledge, SuHiFiMask is the first dataset collected based on real surveillance scenes, rather than the low-quality datasets obtained by manual degradation, such as GREAT-FASD-S [62]. Compared to previous PAD datasets in controlled environments, the one we present inevitably introduces low-resolution face, pedestrian occlusion, changeable posture, motion blur, and other challenging situations, which greatly increases the challenge of FAS tasks. In addition, as shown in the third column of the Tab. I, we define the dataset with the distance between the camera and the subject less than one meter as the short distance dataset, while the dataset with the distance between the camera and the subject greater than three meters is defined as the long-distance type. **Advantage 2:** SuHiFiMask considers the most comprehensive attack types, each of which contains diverse spoofing methods. As shown in Tab. I, 2D image, video replay and 3D mask all appear in SuHiFiMask to evaluate the algorithm's perception of changes for paper color, screen moire and face structure in surveillance scenes. Different from the attack type under classical more constrained environments, as shown in Fig. 2, we introduce paper posters, humanoid stand-ups in 2D image, and headgear, head mold in the 3D mask to minimize the spoofing trace in the surveillance scenes. In order to effectively prevent criminals from hiding their identities through local occlusion during security inspection, we introduce two most effective adversarial attacks (ADV), instead of simply masking the face with paper classes [29] and partial paper [64]. **Advantage 3:**

We designed 40 common real-world surveillance scenes, including daily life scenes (*e.g.*, cafes, cinemas, and theaters) and security check scenes (*e.g.*, security check lanes and parking lots) for deploying face recognition systems. In fact, the rich natural behaviors in different surveillance scenes greatly increase the difficulty of PAD due to pedestrian occlusion and non-frontal views. **Advantage 4:** We collect data in four types of weather (*e.g.*, Sunny, Windy, Cloudy and Snowy days) and natural lighting (*e.g.*, Day and Night lights) to fully simulate the complex and changeable surveillance scenes. Different weather and light bring diverse image style information and image artifacts, which will put forward higher requirements for the generalization of PAD technology.

Based on the above acquisition advantages, our SuHiFiMask contains $10,195$ videos from 101 subjects of different age groups, which are collected by 7 mainstream surveillance cameras and see Fig.1 in *Appendix* for more details. In particular, as shown in the second and third rows of Fig. 2, SuHiFiMask is focused on surveillance scenes, and both real and fake attacks appear at the same time.

As shown in the Fig. 3, the existing FAS dataset of a video contains only one real person or one type of attack. The subject faces the camera and remains stationary during the shooting to ensure the clarity of the collected data. In contrast, videos based on a surveillance scene contain multiple real people and multiple types of attacks. Subjects are not required to face the camera and move randomly in the scene while filming. This leads to low-resolution of face images, pedestrian occlusions, non-frontal poses, and other disturbances that affect the stability and generalization of the algorithm. Thus, the surveillance scene-based FAS dataset poses a greater challenge than the existing FAS dataset.

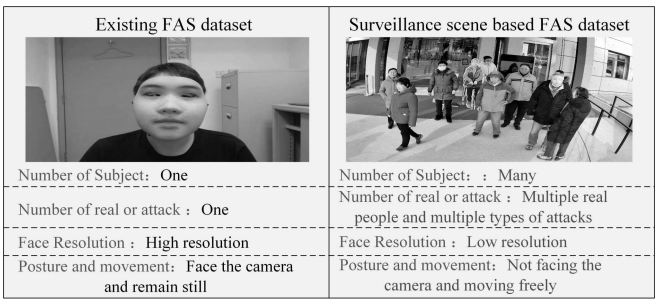

Fig. 3. Comparison of the existing FAS dataset and the surveillance scene-based dataset. The left image is from MARsV2 and the right image is from the proposed SuHiFiMask.

A. Acquisition Details of SuHiFiMask.

Scenes and props. In order to cover real surveillance environments as much as possible, we carefully selected and rented 40 real-world scenarios that include daily places, such as cafes, yoga studios, and movie theaters, as well as security checkpoints, such as security lanes, parking lots, and entrance/exit gates. We provide 232 masks as the candidate

TABLE II
STATISTICAL INFORMATION FOR EACH PROTOCOL OF THE PROPOSED SUHIFIMASK DATASET. NOTE THAT 1, 2, 3 IN THE FOURTH COLUMN MEAN RESIN, SILICONE, AND PLASTER. 4 REPRESENTS HEADGEAR AND HEAD MOLD.

Pro.	Subset	#Subject	Mask	Quality score	#Live	#Mask	#Other attack	#All
1	Train	40	1&2&3&4	[0, 1]	118,520	60,715	22,333	201,568
	Dev	10	1&2&3&4	[0, 1]	23,304	11,856	5663	40,823
	Test	51	1&2&3&4	[0, 1]	69,878	42,569	19,743	132,190
2.1	Train	101	1&2&3	[0, 1]	100,990	40,454	0	141,444
	Dev	101	1&2&3	[0, 1]	20,521	19,608	0	40,129
	Test	101	4	[0, 1]	42,539	21,199	0	63,738
2.2	Train	101	1&2&4	[0, 1]	78,961	36,829	0	115,790
	Dev	101	1&2&4	[0, 1]	20,505	19,052	0	39,557
	Test	101	3	[0, 1]	42,521	28,366	0	70,887
2.3	Train	101	1&3&4	[0, 1]	77,952	28,994	0	106,946
	Dev	101	1&3&4	[0, 1]	20,594	17,714	0	38,308
	Test	101	2	[0, 1]	42,498	44,104	0	86,602
2.4	Train	101	2&3&4	[0, 1]	79,102	29,087	0	108,189
	Dev	101	2&3&4	[0, 1]	20,627	18,068	0	38,695
	Test	101	1	[0, 1]	42,513	42,887	0	85,400
3	Train	101	1&2&3&4	[0.4, 1]	64,276	35,898	58,889	159,063
	Dev	101	1&2&3&4	[0.3, 0.4)	37,990	24,031	27,255	89,276
	Test	101	1&2&3&4	[0, 0.3)	84,368	43,820	36,369	164,557

pool for selection according to the scene requirements. Among them, some high-fidelity plaster and resin masks are from HiFiMask [25], and silicone material headgear and head mold masks are new additions to reduce the forgery traces exposed in the monitoring perspective, where the numbers of plaster masks, resin masks, silicone masks, headgear, and head mold were 93, 93, 23, 10, and 13, respectively. In addition to mask attacks, we printed 2D images of 50 subject in the form of humanoid upright cards and posters and provided video attacks by displaying images on a movable TV. In particular, in order to effectively prevent criminals from hiding their identity information during security checks in surveillance scenes, we crafted adversarial mask [65] and adversarial hat [66] that can induce face recognition systems to categorize the registered identity as unknown identity, aiming to increase the challenge to algorithm stability.

Data collection and processing rules. To ensure the quality and challenge of data, we implemented the following criteria before each shot: a) Device adjustment. We adjusted the positions and angles of each camera to ensure that the entire scene is captured. b) Sample balance. We arranged a consistent number of live and fake subjects to ensure sample balance in SuHiFiMask. c) Static 2D attacks. We deployed posters and humanoid upright-card and electronic screen-based photos with the same identity as the subjects at random locations.

We also considered the following criteria during each shot: a) We designed specific movement routes for each subject to ensure adequate pedestrian occlusion, versatile posture and comprehensive perspective. b) We requested subjects in different scenes to perform scene-related behaviors, such as eating and chatting in daily scenes, and self-service check-in security check scenes.

After data collection, we performed the following pre-processing: a) Face detection. We used RetinaFace [67] to detect the face in each frame of the original video, and discarded those frames where the face could not be detected. b) Face tracking. We use face similarity to track the position of faces in consecutive frames and name each of the different face tracking boxes. c) Video sampling. We sample each video in 10-frame intervals and store the cropped face image in the corresponding face tracking box folder. d) Dataset naming rules. We named the folder of this video according to the following rule: $Group\_Scene\_Camera\_Epoch\_Time$.

Ethical and legal considerations. Since we collected our filming scenes from real-world environments, we have a responsibility to maintain the public environment and protect pedestrian safety. We commissioned two companies to legally authorize the scenes for data collection. SuHiFiMask is a dataset consisting of videos taken from subjects of different age groups, and although this is not a subject explicitly modeled for human behavior, the relevant challenge factors are related to humans. Based on the consideration of the protection of human rights and legal interests, our collection process follows a strictly ethical procedure. We commission a data acquisition company to develop strict standards and obtain the signature authorization of all human subjects. The collected images and videos will be used to develop, train and optimize face anti-spoofing technologies to the extent permitted by Chinese laws. The dataset is balanced in terms of gender and age, that is, there is no hazard in terms of ethics.

B. Evaluation protocol and Statistics.

We define three protocols for SuHiFiMask to fully evaluate the performance in surveillance environments: Protocol 1-ID, Protocol 2-Mask, and Protocol 3-quality.

Protocol 1-ID. Protocol 1 aims to evaluate the comprehensive performance of the algorithm being migrated to long-distance surveillance scenes. Compared with the classical constrained environment datasets, protocol 1 includes various unique factors in surveillance scenes, such as low resolution, pedestrian occlusion, changeable posture, motion blur, and other complex weather, which pose greater challenges to algorithm design. As shown in Tab. II, we divide the training set, development set, and testing set according to the identity information, including 40, 10, and 50 subjects, respectively.

Protocol 2-Mask. Protocol 2 evaluates the generalization of the algorithm for the 'unseen' 3D facial mask type. The diversity and unpredictability of mask materials are important characteristics of spoofing means and are easily interfered with

by other liveness-unrelated factors. Thus, the generalization to mask materials is an important evaluation index. In this work, we divide protocol 2 into four sub-protocols by using the 'leave-one-type-out testing' method, in which one unknown 3D mask material is divided into the testing set for each sub-protocol. As shown in Tab. II, '1', '2', '3', and '4' in the fourth column indicate that the 3D mask material is headgear/head mold, resin, silicone, and plaster, respectively.

Fig. 4. Some samples from protocol 3. The number below the image represents the quality score of the faces. Samples with different score intervals are proportionally categorized as the training set, development set, and testing set.

Protocol 3-Quality. Protocol 3 evaluates the robustness of the algorithm to image quality degradation. Variable quality and disturbances are factors that affect the stability of the algorithm. Therefore, the robustness of the algorithm to quality degradation is an important metric to be evaluated. In this work, as shown in Fig. 4, we use the SER-FIQ [68] algorithm to calculate the image quality score which ranges from 0 to 1. As shown in the fifth column of Tab. II, we assign images with scores [0.4, 1] as the training set, scores [0.3, 0.4) as the development set, and scores [0, 0.3) as the testing set.

IV. Methodology

In this section, we present a Contrastive Quality-Invariance Learning (CQIL) network for FAS tasks based on long-distance surveillance scenes. As shown in Fig. 5, CQIL contains an Image Quality Variable module (IQV) and a dual-stream framework with a contrastive learning branch and a Separate Quality Network (SQN) branch. IQV processes low-quality images into high-quality images by super-resolution and sends them to the contrastive learning branch and the SQN branch. The contrastive learning branch trains the network by using high-quality and low-quality images as input to the online network and the target network, respectively. The SQN branch makes the features extracted by the encoder independent of quality by adversarial learning. In addition, CQI uses high-quality images after super-resolution as input to extract richer discriminative features.

Image Quality Variable Module (IQV). In contrast to the classical constrained environment, the difficulty of the FAS task based on surveillance scenes is the low resolution and variable quality of the images, which leads to insufficient information contained in the images and severely interferes with the extraction of robust features. To solve this problem, a possible solution is to increase the resolution of the image and extract robust invariant features. Inspired by CSRI [69], we introduce the Image Quality Variable (IQV) module to improve the image resolution of SuHiFiMask and recover information relevant to the FAS task. In addition, IQV tags the images processed by the SR network with label 0 and the original images with label 1. Then IQV sends them to the contrastive learning branch and the SQN branch. Since SuHiFiMask is the first unconstrained PAD dataset, there is no high-quality image as ground truth to optimize the super-resolution network. Thus, we use the existing high-definition PAD dataset to train the super-resolution network. As shown in Fig. 5, this process can be expressed as follows: 1) We degrade the high-fidelity dataset OULU-NPU [20] into a low-quality dataset using pre-processing methods such as interpolation and gaussian blurring. 2) We feed degraded low-resolution images into an SR network and use its original data for supervision to train the SR network. 3) We use the SR network with shared parameters to process SuHiFiMask's images into high-quality images. Unlike the standalone super-resolution tasks, we combine the SR tasks with the FAS tasks by integrating the IQV module into the framework with the following two advantages below:

- Training the FAS network with SR network-boosted resolution images can improve the performance of the FAS network.
- The improved performance of other networks in CQIL can better guide the SR network to recover information related to the FAS task in the image.

Finally, MSE loss is used to constrain the super-resolution network:

$$\mathcal{L}_{mse} = \frac{1}{\mathrm{n}} \sum \left(\hat{\mathrm{y}}_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}}\right)^2 \tag{1}$$

where n represents the number of pixels in the image, $\hat{\mathrm{y}}_{\mathrm{i}}$, y_{i} denote the pixel value of the image after super-resolution and the pixel value of ground truth respectively.

Contrastive Learning Branch. To improve the robustness of FAS networks in a quality-variant surveillance environment, we propose a branch based on contrastive learning. Inspired by the BYOL [38], this branch obtains robustness to quality variations by fitting the distribution of potential features for different quality pictures. Specifically, during the training process, due to the constraints of Eq. 2 and Eq. 3, the online network will gradually fit the target network by closing the same class in the potential feature space for pairs of images of different quality, which makes it to obtain a powerful

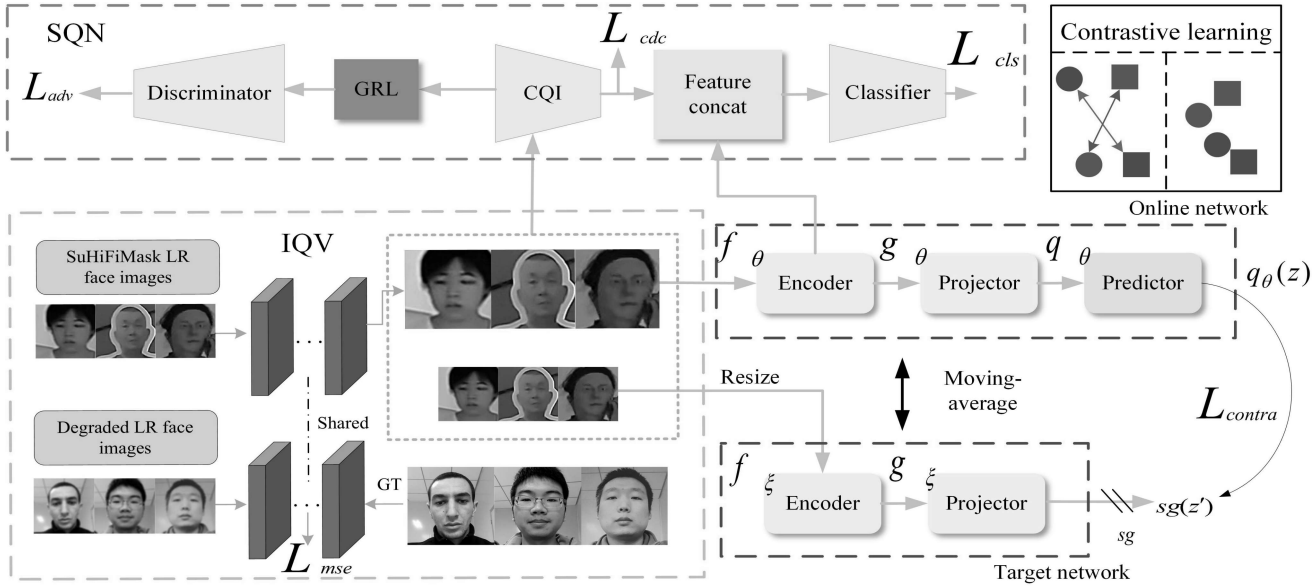

Fig. 5. Contrastive Quality-Invariance Learning (CQIL) network. IQV recovers information from the images and constructs sample pairs of different qualities. Sample pairs are sent to the contrastive learning branch and the SQN branch. The contrastive learning branch consists of an online network (encoder, projector, predictor) and a target network (encoder, projector). Above the image the SQN branch is shown, which contains the discriminator, CQI, GRL, and the main classifier.

feature representation while ignoring the negative impact from different quality distributions.

$$\mathcal{L}_{\theta,\xi} \triangleq \left\|\overline{q_\theta}\left(z_\theta\right) - \overline{z}'_\xi\right\|_2^2 = 2 - 2 \cdot \frac{\left\langle q_\theta\left(z_\theta\right), z'_\xi \right\rangle}{\left\|q_\theta\left(z_\theta\right)\right\|_2 \cdot \left\|z'_\xi\right\|_2} \tag{2}$$

$$\mathcal{L}_{contra} = \mathcal{L}_{\theta,\xi} + \widetilde{\mathcal{L}}_{\theta,\xi} \tag{3}$$

where $q_\theta\left(z_\theta\right)$ is the prediction of the online network output and z'_ξ is the projection of the target network output, then we use ℓ_2-normalize to turn $q_\theta\left(z_\theta\right)$ and z'_ξ into $\overline{q_\theta}\left(z_\theta\right)$ and $\overline{z}'_\xi$. In addition, $\widetilde{\mathcal{L}}_{\theta,\xi}$ is the result of $\mathcal{L}_{\theta,\xi}$ symmetrization.

As shown in Fig. 5, image pairs of different quality generated by IQV are sent to the online and target networks. The online network is composed of an encoder network (Interchangeable backbone networks), a projector (Projection of extracted features into the latent space), and a predictor (with the same multi-layer perceptron structure). Similarly, the target network has an encoder and a projector with different weights from the online network. Unlike the weight update of the online network, the parameters of the target network are not updated in gradient descent [38], and the process can be expressed as follows:

$$\xi \leftarrow \tau\xi + (1-\tau)\theta \tag{4}$$

The parameters ξ and θ represent the parameters to be updated for the target network and the online network, respectively. The parameters θ of the online network are updated by the optimization of the loss function, the parameters ξ of the target network are updated by perceiving an exponential moving-average [70] of the online parameters and we perform the moving-average after each step by target decay rate τ.

Separate Quality Network (SQN). For FAS data in surveillance scenes, which contains many variations (*e.g.*, environment, light, weather), we need operators that are more robust to variations to describe the required fine-grained information. Inspired by central differential convolution (CDC) [7], we use CDC to form a quality-independent backbone network (CQI) in the second branch, exploiting its powerful representation ability to extract fine-grained features under environmental variations. In addition, we use cross-entropy loss as a supervision of CQI, so that this network can capture the cues related to liveness more robustly.

The sample pairs generated by the IQV module have the following characteristics: 1) Both the super-resolution network processed images and the original images contain the object of the face (live or attack) in the center of their images, so even samples with very different quality share the same semantic feature space. 2) Although the quality of each image is different, they all contain discriminative information. Therefore, we make the discriminative features extracted by CQI independent of quality by adversarial learning. Specifically, we use the adversarial loss to optimize the backbone network CQI. And the gradient reversal layer (GRL) [71] allows the parameters of the quality discriminator to be optimized in the reverse direction. This process can be formulated as follows:

$$\begin{aligned} &\min_D \max_C \mathcal{L}_{adv}(C, D) = \\ &- E_{(x.y)\sim(X,Y_Q)} \sum_{i=1}^{N} 1[i=y] \log D(C(x)) \end{aligned} \tag{5}$$

where Y_Q is the set of quality labels, N is the number of images of different quality, C stands for the CQI network backbone where we extracted the liveness-related information, and D represents the quality discriminator. Finally, we concatenate the features extracted by CQI with those extracted by the contrastive learning branch and input them to the classifier for classification.

Algorithm 1 Contrastive Quality-Invariance Learning (CQIL)

Input: image set X, label set Y, HD image set H.

Initialize: encoder f_θ, projector g_θ, predictor q_θ
Initialize: encoder f'_ξ, projector g'_ξ
Initialize: network n of IQV, encoder c of CQI
while not end of training **do**
  sample batch $\mathcal{A} \leftarrow \{x_i \sim X\}_{i=1}^{N}$, $\mathcal{B} \leftarrow \{y_i \sim Y\}_{i=1}^{N}$
  sample batch $\mathcal{H} \leftarrow \{h_i \sim H\}_{i=1}^{N}$
  for $x_i \in \mathcal{A}$ **do**
    $l_i \leftarrow h_i$ ▷ Degradation into lq images
    compute $\mathcal{L}_{mse}$, see Eq. 1
    $x_{i1} \leftarrow n(x_i), x_{i2} \leftarrow x_i$ ▷ Generate image pairs
    $y_{i1}, y_{i2} \leftarrow x_{i1}, x_{i2}$ ▷ Generate quality labels
    $z_1 \leftarrow g_\theta\left(f_\theta\left(x_{i1}\right)\right)$, $z_2 \leftarrow g_\theta\left(f_\theta\left(x_{i2}\right)\right)$
    $z'_1 \leftarrow g_\theta\left(f_\theta\left(x_{i2}\right)\right)$, $z_2\prime \leftarrow g_\theta\left(f_\theta\left(x_{i1}\right)\right)$
    compute $\mathcal{L}_{contra}$, see Eq. 2, Eq. 3
    $\mathcal{X} \leftarrow [x_{i1}, x_{i2}]$, $\mathcal{Y} \leftarrow [y_{i1}, y_{i2}]$, $\mathcal{Z} \leftarrow c(\mathcal{X})$
    compute $\mathcal{L}_{adv}$, by Eq. 5
    $z_3 \leftarrow c(x_i)$, $z_4 \leftarrow f_\theta(x_i)$
    compute $\mathcal{L}_{cls}$, $\mathcal{L}_{cdc}$, $\mathcal{L}_{total}$, by Eq. 6
  end for
end while
$\Delta\theta = backward\left(\mathcal{L}_{total}\right)$
$\theta \leftarrow \theta - learningrate \cdot \Delta\theta$
update ξ by Eq. 4
update network n, encoder c

Overall Loss. As mentioned, CQI is used to extract quality-independent discriminative features, and these features are concatenated with the robust features extracted from the contrastive learning branch and fed to the main classifier. Therefore, the cross-entropy loss $\mathcal{L}_{cdc}$ and $\mathcal{L}_{cls}$ is well constrained for both CQI and the main classifier. In summary, the overall loss function $\mathcal{L}_{total}$ for stable and reliable training can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{total} = \lambda_1 \cdot \mathcal{L}_{cls} + \lambda_2 \cdot \mathcal{L}_{contrast} + \lambda_3 \cdot \mathcal{L}_{adv} \\ + \lambda_4 \cdot \mathcal{L}_{cdc} + \lambda_5 \cdot \mathcal{L}_{mse}\end{aligned} \tag{6}$$

whereλ_1, λ_2, λ_3, λ_4 and λ_5 are five hyper-parameters to balance the proportion of the different loss functions.

V. Experiments

A. Experiments Settings

Dataset and Protocols. In experiments, a total of five datasets were used: OULU-NPU [20], CASIA-MFSD [1], RepalyAttack [2], MARsV2 [34] and the SuHiFiMask dataset. First, we conducted ablation experiments on three protocols of the proposed SuHiFiMask to demonstrate the effectiveness of each component of the proposed CQIL. Second, we present the respective baselines for the different protocols for the proposed dataset. Finally, we design several different cross-testing experiments to demonstrate the importance of the proposed dataset and the effectiveness of the method.

Training Setting. Our proposed method is implemented with Pytorch. In the training stage, models are trained with Adam optimizer and the initial learning rate is $2e-4$. The batch size is set to 6 for CQIL. The epoch of the intra-testing is set to 10, and the lr decreases by 0.2 times per epoch. The epoch of the inter-testing is 300, and lr decreases by 0.2 times per 50 epochs. λ_1, λ_2, λ_3, λ_4 and λ_5 are set to 2, 1.5, 0.5, 1.5, 0.5 respectively.

Performance Metrics and Implementation Details. We accept the Attack Presentation Classification Error Rate (APCER), Bonafide Presentation Classification Error Rate (BPCER), and ACER [72] as the evaluation metrics in our experiments. The ACER on each testing set is determined by the threshold value of the performance on the development set. In cross-testing experiments, we use Half Total Error Rate (HTER) [73] and Area Under Curve (AUC) as evaluation metrics. We use the ResNet18 [48], ViT [74], and the CDCN [7] network as the backbone, and report their results in experiments.

B. Ablation Study.

Here we conduct ablation experiments to verify the contribution of each module of the proposed CQIL on the three protocols of the SuHiFiMask dataset.

TABLE III
The ablation study of different Components. The evaluation metric is ACER (%).

Method	Prot.1	Prot.2	Prot.3
ResNet18	12.58	16.55±51.71	17.64
CQIL-Model-1	11.97	16.01±50.23	17.45
CQIL-Model-2	11.75	15.67±48.12	16.54
CQIL-Model-3	10.90	15.14±46.66	16.13
CQIL-Model-4	**10.69**	**14.90±45.92**	**15.98**

Advantage of the proposed architecture. We compare four architectures with ResNet18 to demonstrate the advantages of each module of the proposed method. The CQIL-model-1 is a contrastive learning network with ResNet18 as its backbone. Since the training of the contrastive learning network requires the output of the IQV module, we use images processed by cubic interpolation and nearest-neighbor interpolation to mimic samples of different quality to eliminate the impact of the IQV module on performance. In addition, we additionally supervise the training of the online encoder using cross-entropy loss. In the testing phase, we use the features extracted by the online encoder for classification. As shown in Tab. III, CQIL-model-1 has a significant improvement in performance on all three protocols compared to ResNet18, which demonstrates that the contrastive learning branch using quality change as a contrast improves the robustness of the network in surveillance scenes.

Advantage of SQN branch. Our proposed SQN branch takes sample pairs of different qualities generated by the IQV module as input and lets the discriminative features extracted by the encoder CQI be independent of the quality by adversarial learning. CQIL-model-2 extends the SQN branch on the basis of CQIL-model-1. In the testing phase, we concatenate the features extracted by the contrastive learning branch with the features extracted from CQI in the SQN

branch for classification. As shown in Tab. III, the performance of CQIL-model-2 is significantly improved on all three protocols, and the performance improvement is especially obvious in protocol 3, which verifies that SQN trained with samples of different quality have the ability to extract discriminative features independent of quality.

Advantage of IQV module. CQIL-model-3 extends the complete IQV module based on CQIL-model-2 but uses low-quality original images to train the CQI encoder. CQIL-model-4 extends CQIL-model-3 by training CQI encoders using high-quality images generated by SR networks. The improved performance of CQIL-model-3 in Tab. III demonstrates that the sample pairs constructed by IQV more closely match the quality variation in the surveillance scene and IQV can effectively improve the performance of the SQN branch and contrastive learning branch. The improved performance of CQIL-Model-4 further validates the two advantages of IQV modules: 1) The SR network processed images can be used for CQI encoder training, thus improving the performance of the FAS task. 2) The performance-improved FAS network can better guide the SR network to recover the discriminative information of the images.

C. Intra-Testing.

Here, we conduct experiments on three different protocols of SuHiFiMask, showing that SuHiFiMask poses a challenge to existing FAS studies while also testing the performance of our proposed CQIL method in different data distributions.

TABLE IV
THE RESULTS OF INTRA-TESTING ON THREE PROTOCOLS OF SUHIFIMASK.

Prot.	Method	APCER%	BPCER%	ACER%
1	ResNet18	13.59	11.57	12.58
	ViT	13.45	9.89	11.67
	CDCN	20.46	18.95	20.41
	CQIL (ours)	**11.09**	**10.29**	**10.69**
2	ResNet18	20.46±184.60	12.74±1.43	16.60±51.05
	ViT	19.56±181.71	12.25±0.42	15.89±45.01
	CDCN	24.88±55.77	24.44±12.51	24.66±16.46
	CQIL (ours)	**18.83±169.37**	**10.88±0.34**	**14.86±46.04**
3	ResNet18	21.04	13.64	17.64
	ViT	19.61	13.95	16.78
	CDCN	28.70	25.89	27.30
	CQIL (ours)	**19.14**	**12.82**	**15.98**

Experiments on Protocol 1-ID. In protocol 1, the data distribution is similar for different sets. The training set, development set, and testing set contain all attack types, and also contain data for all quality scores. The protocol is appropriate to evaluate the performance of the FAS algorithm in long-distance surveillance scenes. As shown in Tab. IV, the proposed CQIL ranks first for three performance metrics (11.09%, 10.29%, 10.69%, respectively) compared to the generic network backbone ResNet, ViT, and the FAS task network CDCN with robust feature representation on the Protocol 1, showing that the proposed method performs well in the FAS task based on surveillance scene with low resolution and many interferences.

Experiments on Protocol 2-Mask. We verify the algorithm's ability to discriminate between different types of masks by protocol 2. As shown in Tab. V, our proposed CQIL achieved good results except for the APCER on protocol 2.1 and protocol 2.2 which was not the highest performance, which proves that our method can extract discriminative features in low-quality mask images. It is worth mentioning that the testing set of protocol 2.1 is composed of headgear and head mold. These two types of masks are very similar to the human head structure, so the algorithm can no longer use features such as mask contours as a basis for prediction. Thus, the performance of CQIL on protocol 2.1 demonstrates the importance of CQI encoders that can extract fine-grained features.

TABLE V
THE RESULTS OF INTRA-TESTING ON FOUR SUB-PROTOCOLS OF SUHIFIMASK PROTOCOL 2.

Prot.	Method	APCER%	BPCER%	ACER%
2.1	ResNet18	43.33	13.96	28.65
	ViT	42.54	12.05	27.29
	CDCN	**35.47**	20.69	28.08
	CQIL (Ours)	41.18	**11.81**	**26.49**
2.2	ResNet18	**8.50**	11.58	10.04
	ViT	8.56	11.40	9.98
	CDCN	14.72	21.41	18.06
	CQIL (Ours)	8.88	**10.26**	**9.57**
2.3	ResNet18	17.52	11.51	14.52
	ViT	13.70	12.33	13.02
	CDCN	26.62	29.20	27.91
	CQIL (Ours)	**13.61**	**10.92**	**12.27**
2.4	ResNet18	12.48	13.90	13.19
	ViT	13.33	13.20	13.26
	CDCN	22.72	26.47	24.59
	CQIL (Ours)	**11.64**	**10.54**	**11.09**

TABLE VI
THE RESULTS OF CROSS-DATASET TESTING FOR CASIA-MFSD, REPLAY-ATTACK AND SUHIFIMASK. THE EVALUATION METRIC IS HTER(%).

Method	Train	CASIA-MFSD		ReplayAttack	
	Test	Replay-Attack	**SuHiFi-Mask (Ours)**	CASIA-MFSD	**SuHiFi-Mask(Ours)**
ResNet18		36.3	**44.5**	**50.9**	42.1
ViT		34.9	**42.8**	44.8	**45.9**
CDCN		15.6	**45.9**	32.6	**41.4**
AUX.(Depth)		27.6	**43.8**	28.4	**39.6**

TABLE VII
CROSS-TESTING RESULTS ON THE MARSV2 DEGRADED WITH DIFFERENT SIZE OF THE GAUSSIAN KERNEL WHEN TRAINED ON THE PROPOSED SUHIFIMASK.

Method	Train	**SuHiFiMask (ours)**					
	Test	MARsV2		MARsV2-3×3		MARsV2-5×5	
	Metric	HTER(%)↓	AUC(%)↑	HTER (%)↓	AUC (%)↑	HTER(%)↓	AUC (%)↑
ResNet18		27.2	79.6	29.5	79.2	32.5	**75.5**
CDCN		37.6	66.8	41.6	61.7	51.2	52.7
AUX.(Depth)		26.8	79.4	41.1	63.7	48.7	54.5
CQIL (ours)		**21.8**	**87.5**	**26.2**	**81.4**	**30.9**	74.0

Experiments on Protocol 3-Quality. Protocol 3 evaluates the stability of the algorithm to image quality degradation. Since the training, development, and testing sets of this protocol differ only in quality, the algorithm is needed to learn a general feature extraction method on data with different quality

distributions. As shown in Tab. IV, our algorithm ranks first on protocol 3 (APCER, BPCER, and ACER are 19.14%, 12.82%, 15.98%, respectively), which proves that our algorithm is effective in extracting discriminative features independent of quality.

D. Inter-Testing.

To evaluate the difficulty of surveillance-based FAS tasks and the effectiveness of CQIL working on low-quality datasets, we design a number of cross-testing experiments.

Cross-dataset. To further evaluate the difficulty of the long-distance PAD task based on surveillance scenes, we design two cross-dataset experiments. (1) We train the model on the CASIA-MFSD dataset and perform the cross-test evaluation on the proposed SuHiFiMask and ReplayAttack datasets. (2) We train the model on the ReplayAttack dataset and evaluate it on the SuHiFiMask and CASIA-MFSD datasets for cross-testing. As shown in Tab. VI, the performance of the model tested on the proposed SuHiFiMask is significantly degraded relative to the performance testing on ReplayAttack or CASIA-MFSD. For example, the HTER (%) of the CDCN trained on CASIA-MFSD was increased by 30.3% for the test on the proposed dataset compared to the test on ReplayAttack. This shows the performance of existing algorithms degrades significantly when they encounter negative factors such as low resolution, motion blur, and occlusion. In particular, CQIL is a PAD method based on low-quality data and requires low-resolution images as input. So the generality of the method will be evaluated in the next subsection.

Cross-quality. To demonstrate the generality of our method to low-quality datasets, we design a series of experiments across the quality. We train the different methods on the proposed SuHiFiMask and test them on MARsV2 after the degradation of gaussian kernels of different sizes. Specifically, since no existing work has provided available low-quality PAD datasets, we simulate low-quality datasets with different degrees of degradation by means of a gaussian kernel to verify the generality of different methods on low-quality datasets. Fig. 6 shows several samples of MARsV2 after treatment with gaussian kernels of different sizes. As shown in Tab. VII, our CQIL achieves good performance on the MARsV2 dataset at all degradation degrees. This demonstrates that our method can encode quality-independent discriminative features. However, there is a domain gap between the manually degraded low-quality dataset and the dataset based on the surveillance scenes. This results in CQIL not being able to take full advantage of encoders trained on low-quality data in real surveillance scenes.

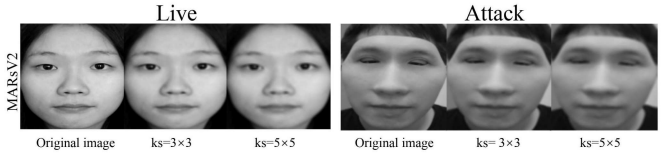

Fig. 6. MARsV2 with different sizes of the gaussian kernel processing.

E. Visualization Analysis

In this section, we further visualize the difficulties that low-quality data poses to FAS work and the performance of CQIL in surveillance scenes. First, we compare the features learned by ResNet18 on protocol 1 of HiFiMask, a dataset for the constrained environment, and on protocol 1 of SuHiFiMask, a proposed surveillance scene-based dataset. As shown in Fig 7, the performance of the algorithm degrades significantly on SuHiFiMask, which indicates that the low-quality data in the surveillance scenes add difficulties to the FAS work. Next, we compare the features learned by CQIL and ResNet18 on protocol 3 of the proposed SuHiFiMask. Compared with ResNet18, the proposed CQIL is able to better distinguish between real faces and attacks, which demonstrates the better discriminative representation capacity of the proposed CQIL in surveillance scenes.

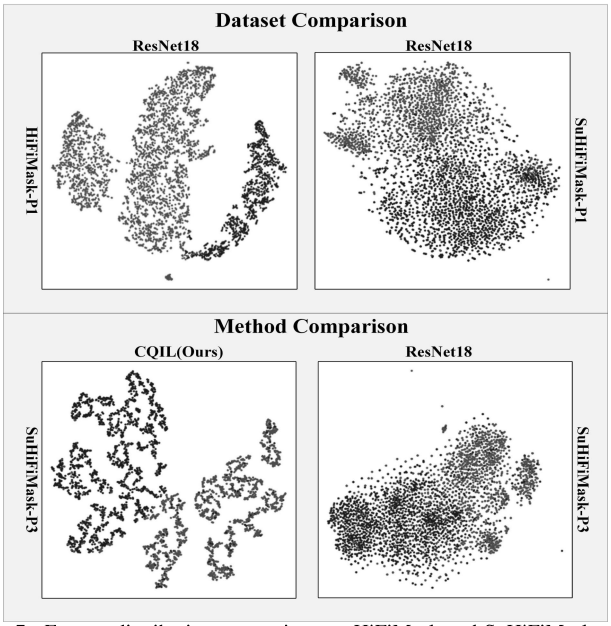

Fig. 7. Feature distribution comparison on HiFiMask and SuHiFiMask using t-SNE [75]. The points with different colors denote features from different classes (blue: real faces; red: attack samples).

VI. Conclusion

In this paper, we release the first large-scale FAS dataset based on surveillance scenes, SuHiFiMask, with three challenging protocols. We hope that this will fill the gap in FAS research in long-distance surveillance scenes. In addition, we propose a Contrastive Quality-Invariance Learning (CQIL) network to recover image information using super-resolution and enhance the robustness of the algorithm to quality variations by fitting the quality variance distribution. Finally, we conduct comprehensive experiments on SuHiFiMask and three other datasets to verify the importance of the datasets for the FAS task and the effectiveness of the proposed method.

References

[1] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face anti-spoofing database with diverse attacks," in *2012 5th IAPR international conference on Biometrics (ICB)*. IEEE, 2012, pp. 26–31.

[2] I. Chingovska, A. Anjos, and S. 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)*. IEEE, 2012, pp. 1–7.

[3] E. Nesli and S. Marcel, "Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect," in *IEEE 6th International Conference on Biometrics: Theory, Applications and Systems (BTAS'13)*, 2013, pp. 1–8.

[4] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 389–398.

[5] R. Shao, X. Lan, J. Li, and P. C. Yuen, "Multi-adversarial discriminative deep domain generalization for face presentation attack detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 023–10 031.

[6] A. George and S. Marcel, "Deep pixel-wise binary supervision for face presentation attack detection," in *2019 International Conference on Biometrics (ICB)*. IEEE, 2019, pp. 1–8.

[7] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, "Searching central difference convolutional networks for face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5295–5305.

[8] K.-Y. Zhang, T. Yao, J. Zhang, Y. Tai, S. Ding, J. Li, F. Huang, H. Song, and L. Ma, "Face anti-spoofing via disentangled representation learning," in *European Conference on Computer Vision*. Springer, 2020, pp. 641–657.

[9] Y. Liu, J. Stehouwer, and X. Liu, "On disentangling spoof trace for generic face anti-spoofing," in *European Conference on Computer Vision*. Springer, 2020, pp. 406–422.

[10] B. Yang, J. Zhang, Z. Yin, and J. Shao, "Few-shot domain expansion for face anti-spoofing," *arXiv preprint arXiv:2106.14162*, 2021.

[11] Z. Chen, T. Yao, K. Sheng, S. Ding, Y. Tai, J. Li, F. Huang, and X. Jin, "Generalizable representation learning for mixture domain face anti-spoofing," *arXiv preprint arXiv:2105.02453*, 2021.

[12] A. Liu, Z. Tan, J. Wan, Y. Liang, Z. Lei, G. Guo, and S. Z. Li, "Face anti-spoofing via adversarial cross-modality translation," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2759–2772, 2021.

[13] X. Li, J. Wan, Y. Jin, A. Liu, G. Guo, and S. Z. Li, "3dpc-net: 3d point cloud network for face anti-spoofing," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2020, pp. 1–8.

[14] Z. Cheng, X. Zhu, and S. Gong, "Surveillance face recognition challenge," *arXiv preprint arXiv:1804.09691*, 2018.

[15] H. Nada, V. A. Sindagi, H. Zhang, and V. M. Patel, "Pushing the limits of unconstrained face detection: a challenge dataset and baseline results," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2018, pp. 1–10.

[16] M. Grgic, K. Delac, and S. Grgic, "Scface–surveillance cameras face database," *Multimedia tools and applications*, vol. 51, no. 3, pp. 863–879, 2011.

[17] M. Kim, A. K. Jain, and X. Liu, "Adaface: Quality adaptive margin for face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 18 750–18 759.

[18] P. Li, L. Prieto, D. Mery, and P. J. Flynn, "On low-resolution face recognition in the wild: Comparisons and new techniques," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 8, pp. 2000–2012, 2019.

[19] Y. Zhong, W. Deng, J. Hu, D. Zhao, X. Li, and D. Wen, "Sface: Sigmoid-constrained hypersphere loss for robust face recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 2587–2598, 2021.

[20] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "Oulu-npu: A mobile face presentation attack database with real-world variations," in *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*. IEEE, 2017, pp. 612–618.

[21] S. Zhang, A. Liu, J. Wan, Y. Liang, G. Guo, S. Escalera, H. J. Escalante, and S. Z. Li, "Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 2, pp. 182–193, 2020.

[22] J. Galbally and R. Satta, "Three-dimensional and two-and-a-half-dimensional face recognition spoofing using three-dimensional printed models," *IET Biometrics*, 2016.

[23] H. Steiner, A. Kolb, and N. Jung, "Reliable face anti-spoofing using multispectral swir imaging," in *2016 international conference on biometrics (ICB)*. IEEE, 2016, pp. 1–8.

[24] S. Liu, B. Yang, P. C. Yuen, and G. Zhao, "A 3d mask face anti-spoofing database with real world variations," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 100–106.

[25] A. Liu, C. Zhao, Z. Yu, J. Wan, A. Su, X. Liu, Z. Tan, S. Escalera, J. Xing, Y. Liang *et al.*, "Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 2497–2507, 2022.

[26] Z. Yu, J. Wan, Y. Qin, X. Li, S. Z. Li, and G. Zhao, "Nas-fas: Static-dynamic central difference network search for face anti-spoofing," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 9, pp. 3005–3023, 2020.

[27] S. Jia, G. Guo, and Z. Xu, "A survey on 3d mask presentation attack detection and countermeasures," *Pattern recognition*, vol. 98, p. 107032, 2020.

[28] S. Jia, X. Li, C. Hu, G. Guo, and Z. Xu, "3d face anti-spoofing with factorized bilinear coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 4031–4045, 2020.

[29] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel, "Biometric face presentation attack detection with multi-channel convolutional neural network," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 42–55, 2019.

[30] S.-Q. Liu, X. Lan, and P. C. Yuen, "Remote photoplethysmography correspondence feature for 3d mask face presentation attack detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 558–573.

[31] B. Lin, X. Li, Z. Yu, and G. Zhao, "Face liveness detection by rppg features and contextual patch-based cnn," in *Proceedings of the 2019 3rd international conference on biometric engineering and applications*, 2019, pp. 61–68.

[32] S.-Q. Liu, X. Lan, and P. C. Yuen, "Multi-channel remote photoplethysmography correspondence feature for 3d mask face presentation attack detection," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2683–2696, 2021.

[33] N. Erdogmus and S. Marcel, "Spoofing 2d face recognition systems with 3d masks," in *2013 International Conference of the BIOSIG Special Interest Group (BIOSIG)*. IEEE, 2013, pp. 1–8.

[34] S. Liu, P. C. Yuen, S. Zhang, and G. Zhao, "3d mask face anti-spoofing with remote photoplethysmography," in *European Conference on Computer Vision*. Springer, 2016, pp. 85–100.

[35] I. Manjani, S. Tariyal, M. Vatsa, R. Singh, and A. Majumdar, "Detecting silicone mask-based presentation attack via deep dictionary learning," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1713–1723, 2017.

[36] A. Agarwal, D. Yadav, N. Kohli, R. Singh, M. Vatsa, and A. Noore, "Face presentation attack with latex masks in multispectral videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 81–89.

[37] S. Bhattacharjee and S. Marcel, "What you can't see can help you-extended-range imaging for 3d-mask presentation attack detection," in *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2017, pp. 1–7.

[38] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[39] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel, "The replay-mobile face presentation-attack database," in *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2016, pp. 1–7.

[40] Y. Zhang, Z. Yin, Y. Li, G. Yin, J. Yan, J. Shao, and Z. Liu, "Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations," in *European Conference on Computer Vision*. Springer, 2020, pp. 70–85.

[41] A. Liu, Z. Tan, J. Wan, S. Escalera, G. Guo, and S. Z. Li, "Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1179–1187.

[42] G. Heusch, A. George, D. Geissbühler, Z. Mostaani, and S. Marcel, "Deep models and shortwave infrared information to detect face presentation attacks," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 4, pp. 399–409, 2020.

[43] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face antispoofing using speeded-up robust features and fisher vector encoding," *IEEE Signal Processing Letters*, vol. 24, no. 2, pp. 141–145, 2016.

[44] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," in *2007 IEEE 11th international conference on computer vision*. IEEE, 2007, pp. 1–8.

[45] W. R. Schwartz, A. Rocha, and H. Pedrini, "Face spoofing detection through partial least squares and low-level descriptors," in *2011 International Joint Conference on Biometrics (IJCB)*. IEEE, 2011, pp. 1–8.

[46] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face spoofing detection using colour texture analysis," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 8, pp. 1818–1830, 2016.

[47] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," *arXiv preprint arXiv:1408.5601*, 2014.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[49] Z. Wang, Z. Yu, C. Zhao, X. Zhu, Y. Qin, Q. Zhou, F. Zhou, and Z. Lei, "Deep spatial gradient and temporal depth learning for face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5042–5051.

[50] A. Jourabloo, Y. Liu, and X. Liu, "Face de-spoofing: Anti-spoofing via noise modeling," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 290–306.

[51] A. Liu, J. Wan, N. Jiang, H. Wang, and Y. Liang, "Disentangling facial pose and appearance information for face anti-spoofing," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 4537–4543.

[52] J. Stehouwer, A. Jourabloo, Y. Liu, and X. Liu, "Noise modeling, synthesis and classification for generic object anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7294–7303.

[53] G. Wang, H. Han, S. Shan, and X. Chen, "Cross-domain face presentation attack detection via multi-domain disentangled representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6678–6687.

[54] S. Liu, K.-Y. Zhang, T. Yao, M. Bi, S. Ding, J. Li, F. Huang, and L. Ma, "Adaptive normalized representation learning for generalizable face anti-spoofing," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1469–1477.

[55] S. Liu, K.-Y. Zhang, T. Yao, K. Sheng, S. Ding, Y. Tai, J. Li, Y. Xie, and L. Ma, "Dual reweighting domain generalization for face presentation attack detection," *arXiv preprint arXiv:2106.16128*, 2021.

[56] H.-P. Huang, D. Sun, Y. Liu, W.-S. Chu, T. Xiao, J. Yuan, H. Adam, and M.-H. Yang, "Adaptive transformers for robust few-shot cross-domain face anti-spoofing," *arXiv preprint arXiv:2203.12175*, 2022.

[57] Y. Jia, J. Zhang, S. Shan, and X. Chen, "Single-side domain generalization for face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8484–8493.

[58] Z. Wang, Z. Wang, Z. Yu, W. Deng, J. Li, T. Gao, and Z. Wang, "Domain generalization via shuffled style assembly for face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4123–4133.

[59] R. Shao, X. Lan, and P. C. Yuen, "Regularized fine-grained meta face anti-spoofing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 974–11 981.

[60] Y. Qin, Z. Yu, L. Yan, Z. Wang, C. Zhao, and Z. Lei, "Meta-teacher for face anti-spoofing," *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[61] J. Wang, J. Zhang, Y. Bian, Y. Cai, C. Wang, and S. Pu, "Self-domain adaptation for face anti-spoofing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 2746–2754.

[62] X. Chen, S. Xu, Q. Ji, and S. Cao, "A dataset and benchmark towards multi-modal face anti-spoofing under surveillance scenarios," *IEEE Access*, vol. 9, pp. 28 140–28 155, 2021.

[63] C. Aravena, D. Pasmino, J. E. Tapia, and C. Busch, "Impact of face image quality estimation on presentation attack detection," *arXiv preprint arXiv:2209.15489*, 2022.

[64] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, "Deep tree learning for zero-shot face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4680–4689.

[65] A. Zolfi, S. Avidan, Y. Elovici, and A. Shabtai, "Adversarial mask: Real-world adversarial attack against face recognition models," *arXiv preprint arXiv:2111.10759*, 2021.

[66] S. Komkov and A. Petiushko, "Advhat: Real-world adversarial attack on arcface face id system," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 819–826.

[67] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5203–5212.

[68] P. Terhorst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "Serfiq: Unsupervised estimation of face image quality based on stochastic embedding robustness," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5651–5660.

[69] Z. Cheng, X. Zhu, and S. Gong, "Low-resolution face recognition," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 605–621.

[70] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[71] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[72] international organization for standardization, "Iso/iec jtc 1/sc 37 biometrics: Information technology biometric presentation attack detection part 1: Framework." in *https://www.iso.org/obp/ui/iso*, 2016.

[73] S. Bengio and J. Mariéthoz, "A statistical significance test for person authentication," in *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop*, no. CONF, 2004.

[74] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[75] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

Appendix

A. Sample of faces

As shown in the Fig. 8, we have listed some samples of pre-processed face images. The figure contains six sections, which are listed as close-up samples of real people, resin masks, silicone masks, plaster masks, headgear, head molds and other forms of attacks. Each column in the first five sections of the figure represents a mainstream surveillance camera, where C1 to C7 represents DS-2CD3T87WD-L, DS-2CD3T86FWDV2-I3S, TL-IPC586HP, TL-IPC586FP, DH-IPC-HFW4843M, DH-P80A1-SA, and ZD5920-Gi4N cameras. Each row in the first five sections of the figure represents a weather or shooting time, with samples taken on sunny days, cloudy days, windy days, snowy days, and nights, respectively. The sixth section of the figure lists head molds and other forms of attack, from left to right, in each column are adversarial masks, adversarial hats, replay attacks in electronic screens, posters, cardboards, and head molds.

B. Sample of scenes

As shown in the Fig. 9, we have listed all 40 scenes included in the SuHiFiMask, which include daily life scenes (*e.g.*, cafes, cinemas, and theaters) and security check scenes (*e.g.*, security check lanes and parking lots) for deploying face recognition systems. On the left side of the figure is the number of each scene in the row, which is the basis for naming the videos in the dataset. In addition, we need to increase the relevance of the data content and surveillance scenes by asking the subjects do scene-related behaviors in the scenes, such as asking the subjects sit around a coffee table and drink coffee in the coffee shop scenes. It is worth mentioning that some scenes in real life are vulnerable to attack in both day and night, so we identify the day and night of this scene as two different scenes, such as parking lot (day) and parking lot (night).

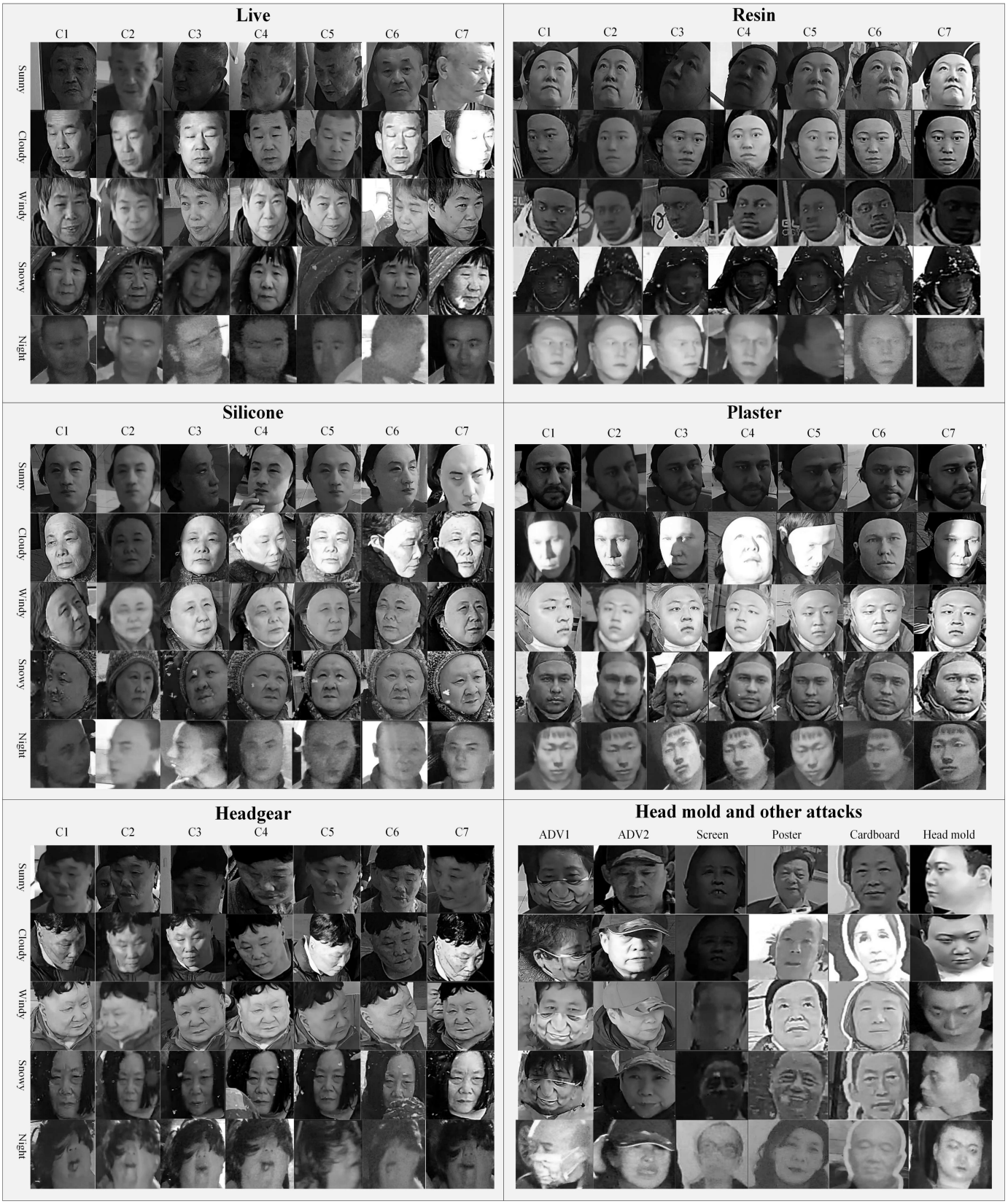

Fig. 8. We show some samples taken on snowy, cloudy, windy, sunny and night time days. Among them, C1 to C7 represents the surveillance cameras with DS-2CD3T87WD-L, DS-2CD3T86FWDV2-I3S, TL-IPC586HP, TL-IPC586FP, DH-IPC-HFW4843M, DH-P80A1-SA, and ZD5920-Gi4N, respectively.

Fig. 9. We show the 40 surveillance scenes included in SuHiFiMask, S1-S40 on the left side of the figure are the numbers of each scene.