# STIDNet: Identity-Aware Face Forgery Detection with Spatiotemporal Knowledge Distillation

Mingqi Fang, Lingyun Yu*, Hongtao Xie, Qingfeng Tan*,
Zhiyuan Tan, Amir Hussain, Zezheng Wang, Jiahong Li, and Zhihong Tian

***Abstract*—The impressive development of facial manipulation techniques has raised severe public concerns. identity-aware methods, especially suitable for protecting celebrities, are seen as one of promising face forgery detection approaches with additional reference video. However, without in-depth observation of fake video's characteristics, most existing identity-aware algorithms are just naive imitation of face verification model and fail to exploit discriminative information. In this paper, we argue that it is necessary to take both spatial and temporal perspectives into consideration for adequate inconsistency clues and propose a novel forgery detector named *SpatioTemporal IDentity Network* (STIDNet). To effectively capture heterogeneous spatiotemporal information in a unified formulation, our STIDNet is following a knowledge distillation architecture that the student identity extractor receives supervision from a *Spatial Information Encoder* (SIE) and a *Temporal Information Encoder* (TIE) through multi-teacher training. Specifically, a regional sensitive identity modeling paradigm is proposed in SIE by introducing facial blending augmentation but with uniform identity label, thus encourage model to focus on spatial discriminative region like outer face. Meanwhile, considering the strong temporal correlation between audio and talking face video, our TIE is devised in a cross-modal pattern that the audio information is introduced to supervise model exploiting temporal personalized movements. Benefit from knowledge transfer from SIE and TIE, STIDNet is able to capture individual's essential spatiotemporal identity attributes and sensitive to even subtle identity deviation caused by manipulation. Extensive experiments indicate the superiority of our STIDNet compared with previous works. Moreover, we also demonstrate STIDNet is more suitable for real world implementation in terms of model complexity and reference set size.**

***Index Terms*—Face Forgery Detection, Knowledge Distillation, Video Forensics, Deep Learning.**

## I. Introduction

With the development of deep learning and generative adversarial networks [1]–[15], recent advancements in face manipulation techniques enable the creation of incredibly realistic fake videos (so called deepfakes) [3], [4], [16]–[23]. Potential indiscriminate usage of this technology arouses public's concern of protecting personal portraits from manipulation, especially for famous people like politicians, celebrities and corporate leaders. On the one hand, it is easier to collect a large quantity of their video material from internet. On the other hand, malicious spread of their fake videos may bring serious consequences, such as slandering reputation [24] or guiding public opinion [25].

Previous researches devoted to detecting forgery content are mostly based on anomaly-aware: they represent forgery detection as a binary classification task and try to obtain forensic clues by exploiting generation artifacts [26]–[30] or perceiving fake video's unnatural phenomenon, e.g. discordant headpose [31] or abnormal eye blinking pattern [32]. Because of simple binary classification strategy and limited manipulated samples during training, their performances drop dramatically when encountering unseen manipulation methods. Although recent methods try to improve generalization by locating blending boundary [33], [34], tracking lip movements [35], [36] and realizing feature disentanglement [37], [38], these anomaly-aware attempts can only mitigate but arduous to solve the generalization problem.

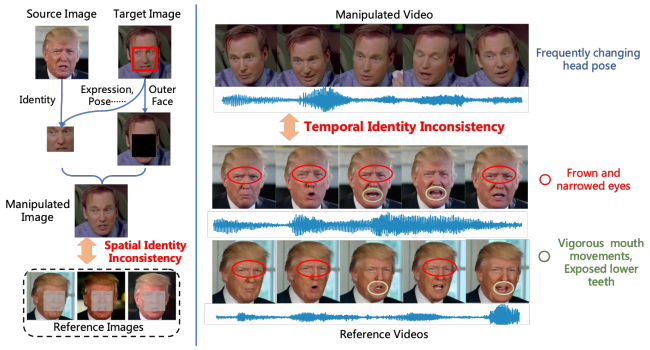



Fig. 1. **Fake video's typical spatiotemporal inconsistencies.** For spatial inconsistency (left), the outer face of manipulated image presents to be inconsistent with that of reference image in identity information. For temporal inconsistency (right), the manipulated video gets a large number of differences with pristine videos in temporal identity characteristics. For instance, the manipulated video changes head pose frequently while this individual in reference videos is used to maintaining a stable head pose during talking. Besides, some personalized movements are not reflected in manipulated videos, such as frown, narrowed eyes, vigorous mouth movements and exposed lower teeth.

Inspired by observation of inevitable identity deviation during manipulation [39]–[42], it is feasible to turn this binary classification problem to an identity verification task, i.e., distinguish whether the person from suspect video is the same one in reference videos. This identity-aware idea pays attention to obtain characteristic identity information instead of manipulated content to improve generalization ability. However, present researches on identity-aware detection are still preliminary. For instance, they just naively imitate the face verification algorithms [39] or even directly use pretrained face verification model [41] to extract identity information. Obviously, this attempts lack observation of manipulated content's characteristic and result in poor discriminative ability for forgery detection. To obtain discriminative identity information, ID-Reveal [40] tries to capture temporal identity

*Lingyun Yu and Qingfeng Tan are the corresponding authors.

information by modeling 3D facial sequence in an adversarial training manner. However, predicting 3D parameters frame by frame leads to high compute consumption. Besides, overly focusing on temporal information while ignoring spatial clues causes poor detection performance. After this, a recent method ICT-Ref [42] is devoted to detecting identity inconsistency of manipulated face and gets remarkable performance. But it extremely relies on a huge reference set, making it impossible to implement in real world scenario.

In fact, identity inconsistency between pristine and manipulated videos extensively exists in both spatial and temporal perspectives. Figure 1 (left) demonstrates the typical process of fake content generation. The inner face of target image is manipulated with identity transferred from source image, and then is blended with target image's outer face. For spatial aspect, this mixture operation results in an identity deviation between manipulated face and source face, mainly reflected in the outer region that provided by target. Meanwhile, for temporal aspect, the temporal identity characteristics, such as head movement pattern, facial expression habit and specific pronunciation movements like raising eyebrow or pursing lips, is followed by target face instead of source face, leading to a temporal inconsistency between manipulated video and reference video with source identity. As shown in Figure 1 (right), these fake videos may look extremely realistic individually, but we can still tell the temporal characteristic difference compared with the real sample from the same person. Moreover, considering audio's temporal properties and its strong correlation with facial motion, such as lip movements and head pose pattern, it will be easier to distinguish temporal identity difference with the help of audio track. Based on the above observation, we argue that it is significant to simultaneously focus on spatial and temporal information for a good identity-aware face forgery detection model.

In this paper, to effectively capture heterogeneous spatiotemporal information in a unified formulation, we propose a multi-teacher multi-modal knowledge distillation architecture, termed *SpatioTemporal IDentity Network* (STIDNet). The STIDNet consists of two exceptional teacher networks (i.e., the spatial teacher network and temporal teacher network) and a student identity extractor. While distinguishing input video's identity label, student network simultaneously receives transferred knowledge from two teacher networks for obtaining spatiotemporal discriminative identity representation. Specifically, for the spatial teacher, a novel *Spatial Information Encoder* (SIE) is devised to exploit regional sensitive identity information. During the training of SIE, we adopt the facial blending augmentation by randomly changing inner face content while maintaining the uniform identity label, encouraging model to capture outer face forensic clues. Meanwhile, for the temporal teacher, although existing methods try to encode temporal identity information by analyzing action units [43] or behavior habits [39], they are all based on single modality and handcrafted features, leading to insufficient mining of temporal clues. Instead of focusing on single modality, we develop a new temporal modeling paradigm named *Temporal Information Encoder* (TIE), which introduces the audio information for exploiting temporal personalized movements in a cross-model training manner. Notably, the training of STIDNet is totally under a generic audio-visual dataset (such as VoxCeleb2 [44]) without manipulation, thus model is not limited to existing manipulation methods and the generalization ability is guaranteed. Besides, benefit from knowledge distillation framework, our method is able to get competitive performance with a much lighter student backbone during identification, improving the real world implementation flexibility.

In experiments, we quantitatively verify the effectiveness of our STIDNet. With spatiotemporal knowledge distillation from SIE and TIE, our method outperforms existing methods on various datasets and exhibits excellent robustness to common video corruptions. Besides, through the comparison of model complexity and reference set size, our STIDNet is more suitable for application in the real world.

Our contributions are as follows:

- We propose the STIDNet, a novel face forgery detection model that focusing on spatiotemporal identity inconsistencies of manipulated videos with knowledge distillation.
- A novel spatial information encoder is devised to guide network to exploit spatial discriminative identity information, and a temporal information encoder is recommended to supervise network comprehending temporal information through cross-modal learning.
- We achieve state-of-the-art performance in adequate generalization and robustness experiments with a light backbone.

## II. Related Works

### A. Face Forgery Detection

**Anomaly-Aware approaches.** Most of the existing face forgery detection methods can be regarded as anomaly-aware fundamentally [26]–[33], [35]–[38], [45]–[47]. Earlier attempts pay attention to handcrafted features such as head pose [31], eye blinking [32] or face warping artifacts [26]. Rössler et al. [27] demonstrate that simply training a Xception network [1] can achieve remarkable intra-dataset performance. Recently, some works focus on exploiting the frequency domain [30] or tracking dynamic inconsistency [48] to mine forensic clues. However, despite their impressive intra-dataset detection precision, these models' performances always have great decline when encountering unseen manipulation methods.

To improve generalization performance, several researches attempt to refine anomaly-aware approaches by designing complex optimization losses or introducing auxiliary tasks. For example, based on the observation of universal inconsistent lip movements in fake videos, a series of lip-related auxiliary tasks have been utilized for capturing general manipulation clues [35], [36]. Besides, Zheng et al. [46] reduce the spatial kernel sizes of convolutions to 1 to enhance temporal representations for detecting unseen manipulation methods. Moreover, facial feature disentanglement is also introduced for generalization improvement [37], [38]. Nevertheless, these attempts under binary supervision can't get rid of manipulated training samples and are hard to solve generalization problem

radically. In this paper, Instead of anomaly-aware idea, we turn to identity-aware detection and represent forgery detection as a verification task, thus the manipulated content isn't involved in training and the generalization performance is guaranteed.

**Identity-Aware approaches.** Recently, the conception of identity-aware detection has been proposed to extract someone's personal traits that usually been modified during manipulation. Present attempts are mostly naive imitation of face verification models and ignore analyzing spatiotemporal generation characteristics of fake video [39], [41]. In addition to above preliminary attempts, Cozzolin et al. [40] introduce 3D facial parameters and adversarial learning for temporal identity extraction, but it ignores rich spatial information and results in poor performance. Besides, a recent method ICT-Ref [42] tries to exploit identity inconsistency of manipulated face with transformer architecture. However, its performance heavily relies on a huge reference set with various identities, which is difficult for real world implementation. In this paper, we propose a STIDNet to extract identity feature from spatiotemporal perspectives simultaneously. Besides, knowledge transfer with two experienced teacher networks also prompts our STIDNet to exploit comprehensive information and reduce the reliance on reference data in real world scenario.

### B. Knowledge Distillation

Recent research on knowledge distillation [49]–[59] are mainly devoted to realizing the model compression and knowledge transfer. Typical knowledge distillation methods generally follow a teacher-student learning strategy, where the student models with lightweight architecture or limited information are optimized to align with teacher model semantically. Specifically, various distillation strategies are studied for efficient knowledge transfer. Romero et al. [52] regard the middle-layer output from teacher model as a kind of essential hint information, and optimize several guided layers in student model to fit this specific feature distribution. Moreover, instead of learning individual feature representation as the transferred knowledge, Park et al. [58] point out that the structured relation between features generally convey more essential information, and propose a relational knowledge distillation strategy for modeling the relation between features. In addition, ICKD [59] is proposed to model the inter-channel correlation for knowledge distillation. The diversity and homology in feature space are exploited respectively for effective teacher-student alignment.

Except for the single-modal attempts, multi-modal supervision signals are also utilized for promoting the knowledge distillation performance. For example, Aytar et al. [55] employ a visual teacher model to guide the student to learn richer sound representation. Moreover, the multi-modal information is introduced in CCL [56] for promoting the video classification performance. And an experienced visual teacher is proven to be beneficial for depth estimation in [57].

However, existing knowledge distillation methods are not specially designed for face forgery detection, where tracing subtle forensics clues becomes the main issue. These direct teacher-student alignment strategies universally will force the student model to fit large teacher arduously, thus make subtle forensics clues drown in complex facial information. On the contrary, in our STIDNet, only the student identity extractor is utilized to process the whole facial video individually. Afterward, two experienced teacher models are designed to give fine-grained supervisions from spatial and temporal perspectives respectively, which further purifies the student model representation and promotes the subtle forensics clues tracing.

### C. Spatiotemporal Modeling

Spatiotemporal modeling is widely utilized in various video-based tasks [60]–[64] for spatiotemporal dependence analysis and temporal variation perception. For example, in the field of few-shot action recognition, a recent research STRM [60] is proposed for effective spatiotemporal relation modeling. Both the spatial local patch features and temporal global frame features are aggregated for action representation enhancement. Besides, to improve the pedestrian trajectory prediction performance, a graph-based spatiotemporal transformer is introduced in [61] for trajectory modeling. The spatial relation between pedestrians are adopted to encourage the temporal trajectory prediction. Moreover, a spatiotemporal curriculum dropout strategy [62] is proposed to discard the difficult spatiotemporal graph node at the beginning of training, which encourages model to learn the robust spatiotemporal dependence step by step.

Furthermore, facial videos generally contain abundant expressions and movements information. Thus, it is of great benefit to introduce spatiotemporal modeling into various face-related tasks. For instance, Xia et al. [63] propose a multi-branch spatiotemporal network for facial expression recognition, where both the appearance and optical flow are considered for a comprehensive spatiotemporal representation. Besides, to improve the face anti-spoofing performance, Wang et al. [64] propose a spatiotemporal propagation module to process both spatial depth information and temporal movement feature simultaneously.

However, existing attempts merely learn spatiotemporal representation by the intrinsic modeling of input video, which are hard to introduce external priors for effective forgery detection. For spatial aspect, without the guidance of regional sensitive information, existing methods generally fail to exploit discriminative local region. Meanwhile, for temporal aspect, normal methods neglect the strong temporal correlation between talking face video and audio information. Lack of audio supervision generally results in poor temporal identity representation ability. Different from this, in the STIDNet, two well-designed teachers are proposed to provide essential spatiotemporal prior for forgery detection. An innovative spatial information encoder is devised to guide model to trace the spatial local identity inconsistency, and a temporal information encoder is recommended to help trace temporal identity clues through cross-modal learning.

## III. Proposed Method

To simultaneously capture spatiotemporal forensic clues for effective forgery detection, we propose the STIDNet in

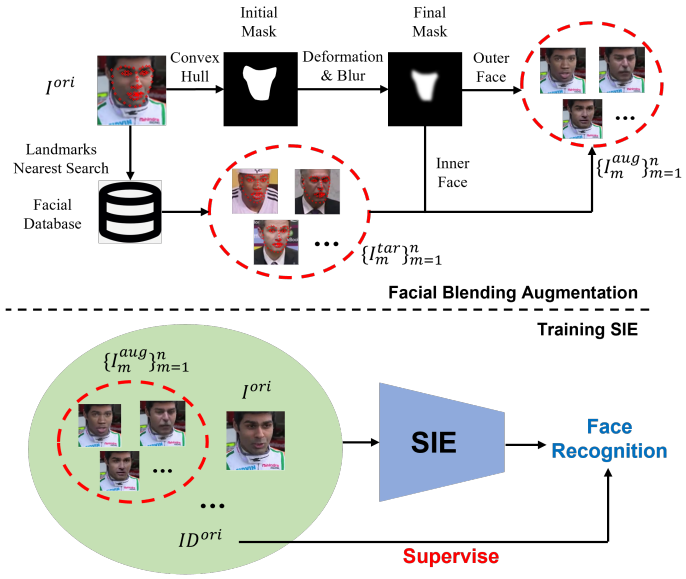



Fig. 2. **Preparation of the SIE.** We augment the original image by swapping other individual's inner face. Then all images are uniformly labeled with the original identity, accordingly a specific face recognition model is trained as our SIE.


Fig. 3. **Illustration of facial blending augmentation.** The augmented faces are various in inner face region, while they share the same outer face identity.

a multi-teacher multi-modal knowledge distillation manner. An overview of our STIDNet is illustrated in Figure 4, which mainly consists a student identity extractor and two teacher networks, i.e., the *Spatial Information Encoder* (SIE) and *Temporal Information Encoder* (TIE). Firstly, in Section III-A, we introduce the preparation of two teacher networks for comprehensive spatiotemporal forensics clues exploiting. Afterward in Section III-B, with two experienced teacher networks, we describe how to utilize these effective spatiotemporal supervisions and train the STIDNet through knowledge distillation. Finally, in Section III-C, we demonstrate the inference phase of STIDNet for identity-aware face forgery detection.

### *A. Spatial and Temporal Information Encoders*

In this section, we introduce the preparation of the *Spatial Information Encoder* (SIE) and *Temporal Information Encoder* (TIE) as our spatial and temporal teacher models respectively.

**Spatial Information Encoder.** To encode the spatial identity information, a direct attempt is adopting the off-the-shelf face recognition model [65], [66] for feature extraction. Unfortunately, existing face recognition algorithms are more inclined to characterize the inner face information, and ignore the outer face region because of its less discriminative in recognition task. On the contrary, as we analyzed above, outer face usually plays a more important role in face forgery detection. Therefore, it is necessary to retrain a face recognition model that more sensitive to outer face identity.

To address the above issue, in this paper we propose a novel regional sensitive *Spatial Information Encoder* (SIE) with effective facial blending augmentation, which can provide essential outer face identity information for forgery detection. The detail of face blending augmentation and SIE training process is illustrated in Figure 2. For each original facial image, we firstly augment it to obtain abundant facial images with various inner regions, then we utilize these faces to train a SIE that is sensitive to the outer face identity.

Firstly, for the facial blending augmentation, we propose to randomly swap the inner identity of input facial image, accordingly restrict model's attention to inner region and guide it to focus on outer face identity. Specifically, inspired by [33], we propose to apply the augmentation with a facial landmark matching strategy. As illustrated in Figure 2 (top), given a pristine original image $I^{ori}$, firstly we extract its facial landmarks with the pretrained landmark detector [67]. Then we apply the *Landmarks Nearest Search* in a large-scale facial database to search the *top n* similar images with various identities. a target set $\{I_m^{tar}\}_{m=1}^n$ is constructed with these $n$ images. Meanwhile, the landmark convex hull of the original image $I^{ori}$ is computed as an initial facial mask. To promise the flexibility of the mask region among various manipulation techniques, we further adopt random shape deformation and Gaussian blur to initial mask to obtain the final mask. Finally, we separately compute the outer face region of original image $I^{ori}$ and inner face regions of $n$ target images $\{I_m^{tar}\}_{m=1}^n$ according to this facial mask, and blend them together one by one to get the $n$ augmented images $\{I_m^{aug}\}_{m=1}^n$. Figure 3 gives an example of augmented faces, it can be observed that the augmented faces are various in inner face region, while they share the same outer face identity as original image $I^{ori}$.

With the help of face blending augmentation, we can extend one original image $I^{ori}$ to total $n+1$ facial images, that share the consistent outer face identity but with various inner face identities. Then we utilize these images to train a regional sensitive SIE. As shown in Figure 2 (bottom), during the training of SIE, we uniformly categorize the both original and augmented images into the identity label of the original image $ID^{ori}$. Then we train the SIE under a face recognition protocol, i.e., distinguishing the identity label of input images. Therefore, to effectively identify the input images, our SIE is more inclined to focus on outer face region and exploit outer face identity information for more discriminative face forgery detection. Notably, we uniformly sample $k$ images with various backgrounds as the original images for each $ID^{ori}$ during

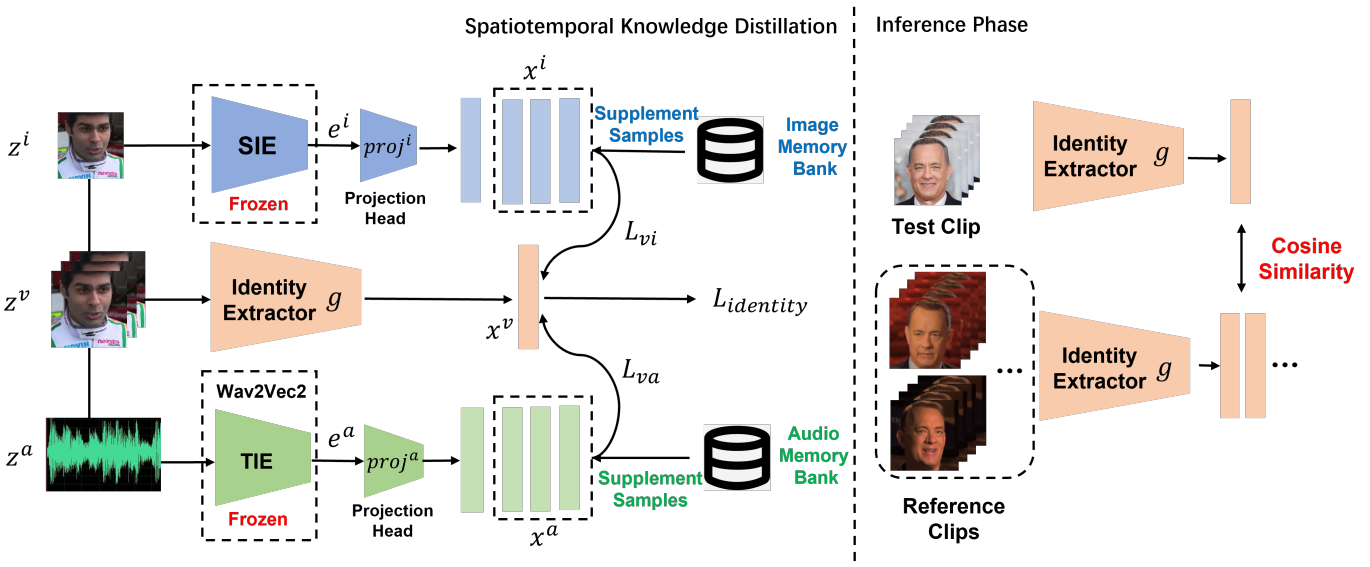



Fig. 4. **Overview of the STIDNet**. In spatiotemporal knowledge distillation, we simultaneously distill the knowledge from SIE and TIE to supervise student identity extractor distinguishing the identity information. In inference phase, we distinguish the authenticity by comparing similarity between test clip and pristine reference clips.

implementation. That is, for each identity, total $k \times (n+1)$ images can be obtained after facial blending augmentation, which effectively promises the background diversity of input images during the training of SIE.

**Temporal Information Encoder.** For the temporal identity representation, previous works usually delve into single-modal features and make less progress [39], [40]. Actually, the audio data encodes rich context information temporally and appears in strong correlation with visual content in talking face video [36], [47], [55], [56]. Thus, we propose to apply audio information in TIE and transfer this temporal supervision signal in a cross-modal manner.

To effectively manage audio information, we employ a pretrained Wav2Vec2 [68] model as the baseline of our TIE. The Wav2Vev2 model is trained under a self-supervise learning manner and exploits the semantic context representation encoded from input raw audio. It achieves impressive performances in various downstream tasks after finetuning. In our STIDNet, we directly employ the original Wav2Vec2 model to obtain temporal supervision signal.

### *B. Spatiotemporal Knowledge Distillation*

After obtaining the experienced SIE and TIE as the spatiotemporal teachers, in this section we will further describe how to utilize these effective spatiotemporal supervisions and train the STIDNet through knowledge distillation. The overall spatiotemporal knowledge distillation framework is illustrated in Figure 4. Firstly we give the formulation of our STIDNet, afterward the loss functions for optimization are introduced in detail.

**Formulation.** We assume access to a large talking face dataset $D$ without any manipulation. A sample $z \in D$ consists of a video clip $z^v \in \mathbb{R}^{T_v \times H \times W \times 3}$ and its corresponding one-dimensional raw audio $z^a \in \mathbb{R}^{T_a}$, where $T_v$, $H$ and $W$ respectively represent the number of frames, image height and width, $T_a$ is the length of input raw audio. In addition, we randomly sample a single frame image $z^i \in \mathbb{R}^{H \times W \times 3}$ from $z^v$. The identity label of $z^v$, $z^a$ and $z^i$ is denoted as $y$.

As the knowledge distillation process illustrated in Figure 4 (left), our multi-teacher knowledge distillation architecture consists two pretrained teacher networks $SIE$, $TIE$ and a student identity extractor $g$, where $g$ is a lightweight video backbone for video identity extraction. While the student $g$ trying to distinguish the input video clip's identity label, supervision signals from $SIE$ and $TIE$ can help the student to focus on spatiotemporal discriminative forensic clues. Specifically, for the single image $z^i$ and raw audio $z^a$, two frozen teachers respectively obtain embeddings $e^i = SIE(z^i)$ and $e^a = TIE(z^a)$, which are then passed through two projection heads $proj^i$ and $proj^a$ to project embeddings to the common space, $x^i = proj^i(e^i)$ and $x^a = proj^a(e^a)$. Besides, for the video input $z^v$, our student $g$ extracts its feature $x^v = g(z^v)$.

With the obtained image, audio and video features $x^i$, $x^a$ and $x^v$, on the one hand, the video feature $x^v$ is used to distinguish input video's identity label in an identity recognition training manner. On the other hand, the image feature $x^i$ and audio feature $x^a$ simultaneously supervise the student $g$ capturing abundant spatiotemporal information through contrastive learning, i.e., pulling together the positive pairs and pushing away negative pairs.

Meanwhile, various researches [69]–[71] point that the number of negative samples per batch plays a very important role in contrastive learning. Therefore, we propose to establish two memory banks for $x^i$ and $x^a$ to provide abundant negative samples during optimization. These memory banks are constantly updating and we randomly sample a set of features for training every batch. Notably, consider that the $SIE$ and $TIE$ are frozen and only parameters from $proj^i$ and $proj^a$ are updating during training, concerns about the hysteresis of memory banks are almost negligible.

**Loss functions.** As we mentioned above, our video identity extractor $g$ has two main optimization targets: distinguishing input video clip's identity and receiving the supervision from teacher networks $SIE$ and $TIE$. Following, we will explain in detail separately.

Firstly, for the identity classification task, to enhance the intra-class compactness and inter-class discrepancy of identity features, the cosine-based loss [65] is introduced for optimization rather than widely used cross-entropy loss.

$$L_{identity} = -log\frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s\cos(\theta_{y_i}+m)} + \sum_{j=1,j\neq y_i}^{N} e^{s\cos\theta_j}} \quad (1)$$

Where the $\theta_j$ is the angle between $W_j$ and $x_i^v$, $W_j \in \mathbb{R}^d$ denotes the *j*-th column of the last face recognition fully-connected layer weight $W \in \mathbb{R}^{d\times N}$, $x_i^v \in \mathbb{R}^d$ is the *i*-th sample of the input batch and its identity label is $y_i$. $s$ is a scale hyperparameter and $m$ is the additive angular margin. We set $s = 30$ and $m = 0.2$ by default.

Besides, for the knowledge distillation loss functions, given the image feature $x^i$ and audio feature $x^a$, we propose to respectively distill the spatiotemporal information to student video network $g$ by contrastive learning. Formally, The contrastive loss $L_{vi}$ (based on InfoNCE [72]) between video and image features can be derived as following.

$$L_{vi} = -log\frac{e^{\phi(x^v,x_+^i)/\tau_i}}{\sum_{k=1}^{B+M} e^{\phi(x^v,x_k^i)/\tau_i}} \quad (2)$$

Where $\phi$ denotes computing cosine similarity, $\tau_i$ is the temperature. $x_+^i$ is the paired positive sample for $x^v$ in whole batch's image embeddings $\{x_k^i\}_{k=1}^{B+M}$. Here, $B$ is the batch size and $M$ is the number of additional sampled image features from image memory bank. The video-audio distillation loss between $x^v$ and $x^a$ has the same form with $L_{vi}$ and is denoted as $L_{va}$. Moreover, the distillation temperature for video-audio paired is denoted as $\tau_a$. Specifically, $L_{va}$ can be derived as following.

$$L_{va} = -log\frac{e^{\phi(x^v,x_+^a)/\tau_a}}{\sum_{k=1}^{B+M} e^{\phi(x^v,x_k^a)/\tau_a}} \quad (3)$$

Finally, the objective is given by

$$L = L_{identity} + \lambda_{image}L_{vi} + \lambda_{audio}L_{va} \quad (4)$$

where $\lambda_{image}$ and $\lambda_{audio}$ are the scaling factors.

### C. Inference Phase

Figure 4 (right) illustrates the inference phase of STIDNet. Only the student identity extractor $g$ is adopt during inference. Given a suspect video and a set of reference videos from the same individual . We firstly embed both of them with the identity extractor $g$. Then we compute the similarity of suspect embedding with each reference embeddings respectively. Finally, the maximum similarity is determined as similarity score. A lower score indicates that the suspect video is more likely to be fake. On the contrary, an input video with high score can be seen as the real one.

## IV. Experiments

In this section, firstly we describe the detailed introduction about experimental setup in Section IV-A. Afterwards, in Section IV-B, we provide the detection performance comparison with state-of-the-art methods. Then in Section IV-C, the robustness evaluation is conducted in our STIDNet. Finally, Section IV-D reports the ablation experiment and visualization results to analyze the STIDNet.

### A. Experimental setup

**Dataset.** We use the **VoxCeleb2** [44] dataset for training, which contains 6,112 identities with more than 150,000 speech videos. For forgery datasets, we choose the following widely used benchmarks for test. (1) **DeepFake Detection** (DFD) [23] contains 363 real and 3068 manipulated videos released by Google. (2) **preview DeepFake Detection Challenge** (pDFDC) [22] includes realistic facial manipulation videos that are subjected to strong perturbations. (3) **FaceForensics++** (FF++) [27] consists of 1,000 real videos and 4,000 fake videos generated by four facial manipulation methods: Deepfakes [17], Face2Face [19], FaceSwap [18] and NeuralTextures [20]. (4) **FaceShifter** (FSh) [3] contains 1,000 fake videos generated from real videos in FF++. (5) **Celeb-Deepfake** [21] is a challenging dataset for forgery detection. It includes an original version CD1 with 408 real videos and 795 manipulated videos. Its extending version CD2 contains 590 real videos and 5639 fake videos. For the selection of reference information, we follow the provision in [42] and randomly sample 10 video clips of each protected individual as reference set.

**Evaluation metric.** To solve the imbalance of real and fake samples in forgery datasets, the widely applied AUC (area under the Receiver Operating Characteristic curve) [33], [35]–[40], [42], [45]–[47] is used as evaluation metric in experiments.

**Implementation details.** Firstly, for the data preprocessing and augmentation, we extract facial images from video with pretrained face detector [67] frame by frame, and then resize them to $112 \times 112$. The *Horizontal Flipping* with probability 0.5 is applied as image augmentation. Moreover, each video clip $z^v$ consists of 30 such facial frames sampled with 10 FPS sampling rate. The audio input $z^a$ is sampled with 16kHz from synchronized audio track. And the image input $z^i$ is a single frame that is randomly sampled from video input $z^v$. Furthermore, the original image $I^{ori}$ that used for training SIE follows the same preprocessing and augmentation strategy as $z^i$.

Besides, for the network structure, a 26-layer configuration Channel-Separated Convolutional Network (CSN-26) [2] is adopted as the video identity extractor $g$, and the $SIE$ backbone is a ResNet-18 [73]. In addition, we utilize the off-the-shelf Wav2Vec2 model [68] with a combination of CNN and transformer architecture as the $TIE$. A single fully-connection (FC) layer is employed as the projection heads $proj^i$ and $proj^a$ respectively. Moreover, the teacher networks $SIE$ and $TIE$ are kept frozen during training. During inference, only the student identity extractor $g$ is utilized.

TABLE I
CROSS-DATASET GENERALIZATION. AUC SCORES (%) ON UNSEEN DATASETS. HERE WE MARK THE DATASET-REF METHODS WITH GRAY DUE TO ITS EXCESSIVELY HUGE REFERENCE SET AND LESS PRACTICABILITY. BEST RESULTS **EXCEPT DATASET-REF METHODS** ARE DENOTED IN **BOLD**.

| Method | | | DFD | pDFDC | FF++ | FSh | CD1 | CD2 | Avg |
|---|---|---|---|---|---|---|---|---|---|
| Anomaly-Aware | | Multi-task [29] | 65.2 | - | - | - | 72.3 | 61.1 | 66.2 |
| | | MesoInc4 [28] | 59.1 | 74.0 | 63.4 | - | 42.3 | 53.6 | 58.5 |
| | | Xception [27] | 95.6 | 79.0 | - | - | 75.0 | 77.8 | 81.9 |
| | | DSP-FWA [26] | 91.0 | - | 81.9 | - | 78.5 | 81.4 | 83.2 |
| | | FFD [45] | 76.6 | 69.0 | 92.3 | - | 74.2 | 77.8 | 78.0 |
| | | Face X-ray [33] | 94.1 | - | **98.5** | - | 74.8 | 75.4 | 85.7 |
| | | LipForensics [35] | - | - | - | 97.1 | - | 82.4 | 89.8 |
| | | FTCN [46] | - | - | - | 98.8 | - | 86.9 | 92.9 |
| | | RealForensics [47] | - | - | - | 99.7 | - | 86.9 | 93.3 |
| | | Zhao et al. [36] | - | - | - | 97.8 | - | 84.2 | 91.0 |
| | | ICT [42] | 84.1 | - | 90.2 | - | 81.4 | 85.7 | 85.4 |
| Identity-Aware | Dataset-Ref | ICT-Ref [42] | 93.2 | - | 98.6 | - | 96.4 | 94.4 | 95.7 |
| | | A&B [39] | 77.0 | 60.0 | - | - | - | 56.0 | 64.3 |
| | Individual-Ref | ID-Reveal [40] | 96.0 | 91.0 | 89.5 | - | - | 84.0 | 90.1 |
| | | DFR [41] | 92.5 | 95.8 | 81.9 | 98.6 | 88.3 | 88.2 | 90.9 |
| | | **STIDNet(Ours)** | **96.1** | **99.4** | 95.8 | **99.9** | **92.1** | **91.4** | **95.8** |

Moreover, for the hyperparameter and optimization settings, we set the knowledge distillation temperatures $\tau_i$ and $\tau_a$ in Formula 2 and Formula 3 to 0.1, 0.5 respectively. And the loss function hyperparameters $\lambda_{image}$ and $\lambda_{audio}$ in Formula 4 are set to 1 and 0.5. Meanwhile, our STIDNet is trained by SGD optimizer with learning rate $1 \times 10^{-1}$ and weight decay $5 \times 10^{-4}$. We train for 50 epochs with the batch size of B=32, and the learning rate is divided by 10 every 10 epochs. Furthermore, we additionally sample M=512 samples from image and audio memory banks respectively every mini-batch during training. The maximum capacity of memory bank is 16,384 and the earlier incoming sample is preferentially discarded once the capacity is exceeded. Besides, for preparation of SIE, we extract $k = 40$ facial images every identity as the original images. For each original image, we retrieve the *top n=10* target images for facial blending augmentation, i.e., total $40 \times (10 + 1) = 440$ images are utilized for training of SIE per identity.

### *B. Comparison with State-of-the-art Methods*

In this section, we compare our approach with state-of-the-art methods from both anomaly-aware and identity-aware categories. Consider that the generalization ability is the most important property for detection model, our comparison is following a cross-dataset protocol [33], [35], [36], [39], [40], [42], [46], [47], i.e., all the methods are tested in the manipulation dataset that unseen during training.

**Comparison with anomaly-aware methods.** The anomaly-aware methods include Multi-task [29], MesoInc4 [28], Xception [27], DSP-FWA [26], FFD [45], Face X-ray [33], LipForensics [35], FTCN [46], RealForensics [47], Zhao et al. [36], and ICT [42]. As shown in Table I, our STIDNet gets the best detection performance in five forgery datasets, suggesting that our model performs well when exposed to unseen manipulation methods.

Comparison with other state-of-the-art anomaly-aware methods, our STIDNet exhibits distinguished advantage in several challenging datasets. Specifically, for the advanced method FTCN [46] devoted to exploiting the temporal incoherence clues, STIDNet outperforms it by 4.5% on the challenging CD2 dataset. Moreover, for another typical method Face X-ray [33] that focuses on blending boundary detection, STIDNet also outperforms it by 2.0% and 16% on the DFD and CD2 datasets respectively. We analyze that, general anomaly-aware method merely perceive the inconsistency clues like temporal incoherence and blending boundary from a single video. However, for several challenging datasets like DFD and CD2, the advanced manipulation techniques have eliminated most of the inconsistency clues, and existing anomaly-aware methods are hard to exploit enough forensics information merely relying on the single video. On the contrary, the additional reference data in identity-aware methods provides kind of strong prior knowledge for effective forgery detection. In our STIDNet, we propose to comprehensively exploit this essential identity prior from both spatial and temporal perspectives. The effective multi-teacher knowledge distillation architecture enables extraction of discriminative identity representation and promotes the detection performance.

**Compared with identity-aware methods.** For identity-aware methods, it is further divided into two secondary categories Dataset-Ref and Individual-Ref according to the size of reference set. The Dataset-Ref methods need to build an extremely huge reference dataset including various individuals and retrieve the paired identity during test. Meanwhile, a more flexible and practical solution is the Individual-Ref method. It only requires the reference data from single individual and executes the verification task during test. In our comparison, the ICT-Ref [42] and A&B [39] belong to Dataset-Ref methods, while the Individual-Ref methods include ID-Reveal [40], DFR [41] and our STIDNet. Notably, The Dataset-Ref methods consume a lot of resources and are impractical for

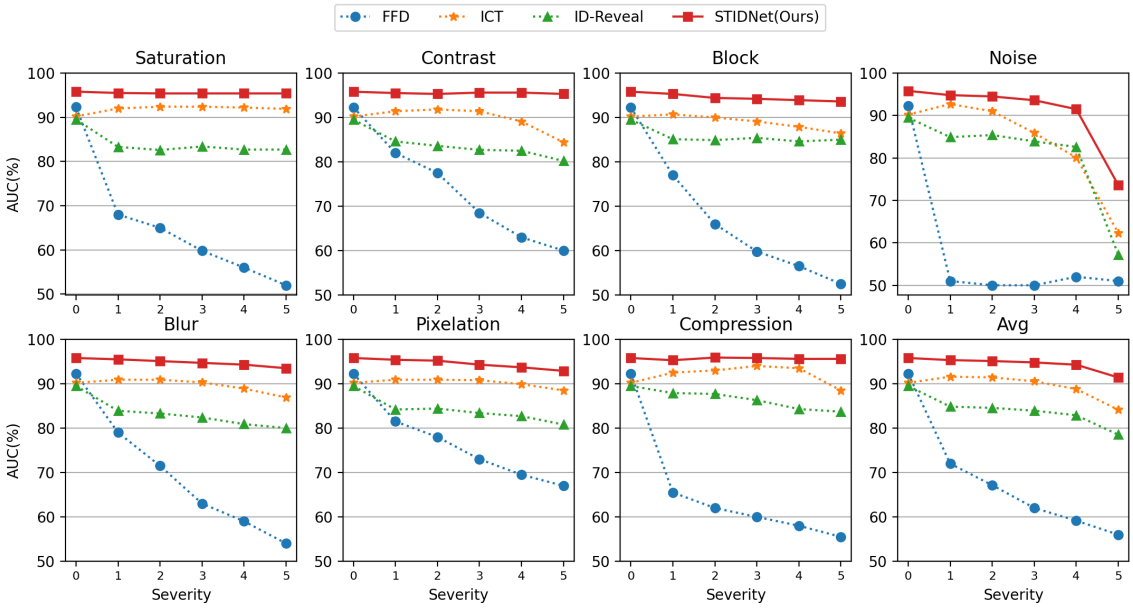



Fig. 5. **Robustness to various unseen corruptions**. Model's detection AUC to various severity of seven common video corruptions. "Avg" denotes the average AUC across all corruptions. Our STIDNet exhibits strong robustness while get remarkable detection AUC.

real world implementation. On the one hand, these methods require a large quantity of reference data to ensure detection performance, which is dozens or even hundreds of times compared with the demand of Individual-Ref methods. On the other hand, it is impossible to establish such large reference dataset in real world implementation for Dataset-Ref methods.

Table I presents the comparison of our STIDNet and other identity-aware methods, where the Dataset-Ref methods are marked with gray because of its excessively huge reference set and less practicability. It can be observed that our STIDNet achieves remarkable performance among all datasets. Specifically, for another typical identity-aware method DFR, STIDNet outperforms it on all six datasets, especially on two challenging dataset DFD and CD2 by 3.6% and 3.2% respectively. Meanwhile, our method also gets competitive performance compared with Dataset-Ref method ICT-Ref in the case of a much smaller reference set, which further proves the effectiveness and practicability of our method with limited reference data. In fact, compared with other methods that merely exploit the identity information from restricted perspective, our STIDNet proposes to comprehensively capture the spatiotemporal identity characteristics for effective identity-aware forgery detection. In spatial perspective, an innovate SIE is proposed to guide model to focus on essential outer face identity. Meanwhile, in temporal aspect, an audio teacher TIE is adopted for personalized temporal identity characteristic modeling. This effective multi-teacher knowledge distillation architecture enables the comprehensive identity characteristics capturing in our STIDNet.

However, we also notice that our STIDNet gets an unsatisfied performance on the FF++ dataset, which is lower than the state-of-the-art method by 2.7% on detection AUC. This disadvantage is mainly due to the misclassification of face reenactment manipulation like NeuralTextures [20]. Specifically, STIDNet achieves 84.7% performance on the NeuralTextures videos in FF++. Theoretically, identity-aware methods are more suitable for detecting the fake videos that the identity information is manipulated, i.e., the face swapping videos. For face reenactment videos, the facial identity is intentionally preserved unchanged, and it is hard to exploit enough identity inconsistency clues. In fact, detecting face reenactment videos is the general shortcoming for all identity-aware methods. However, benefit from the focus on temporal identity characteristics, our identity-aware STIDNet still makes certain progress in detecting face reenactment videos, which reflects in relatively high performance on FF++ compared with other identity-aware methods.

### C. Robustness to Common Corruptions

In addition to generalization ability, another significant indicator for face forgery detector is the anti-corruption robustness [30], [35], [36], [38], [42], [47]. Facial videos spread on social media are usually subjected to common corruptions and detector may fail to capture adequate forensic clues due to video degradation. To demonstrate the robustness of our model, we follow [35], [42], [47], [74] to assess robustness to various perturbations. The FFD [45], ICT [42], ID-Reveal [40] are selected for comparison and all the models are tested on FF++ samples that were exposed to various unseen corruptions. The perturbations are proposed in [74] and include seven different operations, i.e., Color Saturation Change, Color Contrast Change, Local Block-Wise Distortion, White Gaussian Noise in Color Components, Gaussian Blur, Pixelation and Video Compression. Each perturbation have five intensity levels. Figure 5 demonstrates the model's detection AUC to various severity of seven video corruptions. It is evident

TABLE II
**FRAMEWORK ABLATION.** ANALYZING THE EFFECT OF SIE AND TIE. AUC SCORES (%) ON CD2 AND DFD ARE PRESENTED.

| Method | CD2 | DFD |
|---|---|---|
| only CSN | 84.5 | 92.8 |
| CSN + SIE | 91.0 | 95.5 |
| CSN + TIE | 85.7 | 93.4 |
| **STIDNet(Ours)** | **91.4** | **96.1** |

TABLE III
**ABLATION ABOUT DESIGN OF SIE.** ANALYZING THE DESIGN OF SIE. WE HAVE FOUR DIFFERENT EXPERIMENT SETTINGS AND THE AUC SCORES (%) ON CD2 AND DFD ARE PRESENTED.

| Method | CD2 | DFD |
|---|---|---|
| STIDNet w/o SIE | 85.7 | 93.4 |
| FR | 88.8 | 94.4 |
| Masked Face | 89.9 | 95.0 |
| **STIDNet(Ours)** | **91.4** | **96.1** |

that while achieving remarkable detection precision, the line representing our STIDNet is almost horizontal against most corruptions in different severity level. The Gaussian Noise at level 5 is an exception that the video content is greatly destroyed and other methods also get a significant decline here. For the experiment result illustrated in Figure 5, we analyze that the detection robustness problem is mainly due to the gradually erased forensics clues by common corruptions. For the compared methods like FFD, although these methods are able to maintain satisfied detection performance on the clean videos, it will be hard for them to exploit enough forensics clues in the degradation videos, and results in poor robustness. On the contrary, in our STIDNet, we propose to adequately exploit the identity information from both spatial and temporal aspects. This comprehensive perspective enables our model to capture enough forensics clues even in a degradation scenario. Moreover, the focus on semantic-level identity information also promotes the robustness of our STIDNet.

### *D. Analysis of STIDNet*

**Framework ablation.** As shown in Table II, we ablate the importance of two teacher networks SIE and TIE by testing the detection performance on CD2 and DFD dataset. It can be observed that knowledge transfer from both SIE and TIE are beneficial to face forgery detection. A CSN network [2] with SIE get 6.5% and 2.7% AUC increase than simply training a CSN model. At the same time, the supervision signal from TIE also promotes the model to capture temporal clues and the *CSN + TIE* performs better than *only CSN* by 1.2% and 0.6%. Finally, our STIDNet with multi-teacher supervision gets the best performance, suggesting that both spatial and temporal information are significant for capturing identity inconsistency.

**Design of SIE.** In our framework, we use the SIE as spatial teacher and train it with facial blending augmentation. Here we discuss the effectiveness of introducing augmentation and have four different experiment settings: (1) *STIDNet w/o SIE*: remove the SIE branch. (2) *FR*: using a pretrained face recognition network [65] as SIE. (3) *Masked Face*: retrain a face recognition network with input image that inner face region is masked. Specifically, we directly utilize the facial mask in facial blending augmentation (Section III-A) to mask the inner face region, which promises the comparison fairness with our STIDNet. (4) *STIDNet (Ours)*: train the SIE with facial blending augmentation. The detection AUC scores on CD2 and DFD are presented in Table III. As expected, training SIE with facial blending augmentation, i.e., our STIDNet, gets the best performance compared with other three settings. For *FR*, it can be observed that the addition of a simple face recognition model can increase model's performance by 3.1% and 1.0%, we analyze that the practiced teacher model effectively supervise the backbone to focus on spatial identity information. Besides, the *Masked Face* model gets a better performance than *FR* because of its tendency to pay attention to outer face region. Finally, our STIDNet with facial blending augmentation get higher performance than *Masked Face* by 1.5% and 1.1%. In our opinion, although both two methods are aimed to guide model to focus on outer face region, it is evident that the blending measure is smoother than the direct mask operation, thus guarantees the completeness of input facial image. Consider that SIE works as teacher network for providing supervision signal, training the SIE with blending data is more suitable for knowledge transfer.

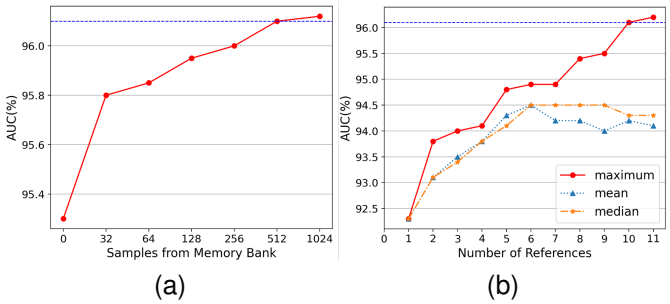



Fig. 6. (a) **Effect of memory bank.** AUC scores(%) as a function of the number of samples from memory bank, the negative sample supplement from memory banks promotes model's performance. (b) **Effect of reference set.** AUC scores(%) as a function of the number of references. The maximum strategy gets the best performance compared with mean and median strategy.

**The effect of memory bank.** In STIDNet, we introduce the image and audio memory banks to provide abundant negative samples for knowledge distillation. Figure 6a shows the effectiveness of supplementary samples from memory bank in DFD dataset. It can be seen that the design of memory bank improves model's discriminating ability and results in higher AUC score, which proves that supplement of negative samples effectively prompt the knowledge transfer. Meanwhile, model's detection performance increases with the number of samples and tend to be stationary when it comes to 512. As a result, we choose the 512 as sampling number for balance between performance and computational cost.

**The effect of reference set.** Figure 6b illustrates the detection AUC of STIDNet varies with size of reference set. Obviously, a larger reference set can provide more information to forgery

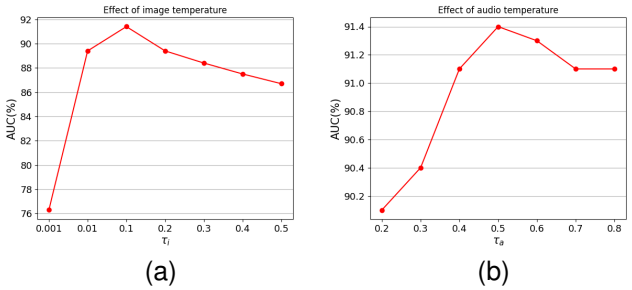



Fig. 7. **Effect of temperature hyperparameters.** Effect of image temperature hyperparemeter $\tau_i$ (a) and audio temperature hyperparameter $\tau_a$ (b).

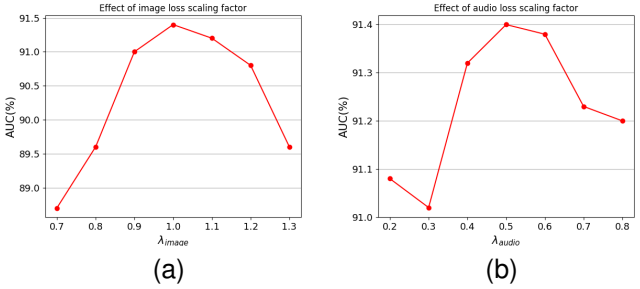



Fig. 8. **Effect of loss function scaling factors.** Effect of the loss function scaling factors $\lambda_{image}$ (a) and $\lambda_{audio}$ (b).

detection. Same as the provision in [42], in our above experiments we uniformly assign 10 video clips to each identity for a fair comparison. Besides, It can be observed from Figure 6b that our STIDNet can still get 92.3% AUC even with 1 reference video clip, which means our model doesn't rely on a carefully selected reference set and shows great robustness.

Moreover, in our strategy, we calculate the maximum cosine similarity among whole reference set as the final similarity score. Figure 6b also provides the AUC of calculating mean and median scores, while the maximum strategy gets the best result. We analyze that the reference set generally contains various facial videos with different facial expressions, emotions and backgrounds, which reflects a degree of diversity. Meanwhile, the identity-aware forgery detection is actually a verification task, and the main target of our STIDNet is searching for the *"most similar"* sample among various videos as the reference data. Therefore, it is more suitable to take the "maximum cosine similarity strategy" with the increase of reference set. Besides, other identity-aware methods such as [40], [42] also adopt this relatively loose measurement strategy for better performance.

**Analysis of hyperparameters.** To further analyze the effect of hyperparameters in STIDNet, in Figure 7 and Figure 8 we detailedly evaluate the model detection performance on CD2 dataset with various hyperparameter settings, including the knowledge distillation temperatures $\tau_i$, $\tau_a$ and the loss function scaling factors $\lambda_i$ and $\lambda_a$.

Firstly, for the temperatures $\tau_i$, $\tau_a$ in Formula 2 and Formula 3, in Figure 7a, we set $\tau_i$ to various values between 0.001 to 0.5 and fix $\tau_a$ to be 0.5 for evaluation. Besides, in Figure 7b

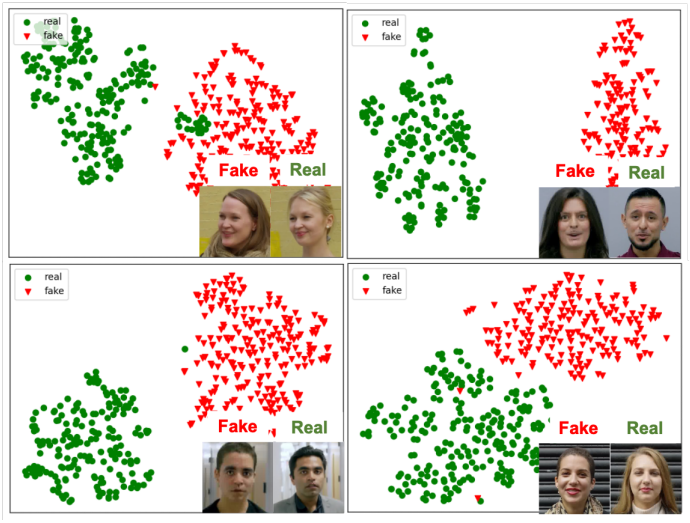



Fig. 9. **t-SNE visualization**. Four identities in DFD dataset are randomly sampled for visualization, there is clear boundary between real and fake samples.

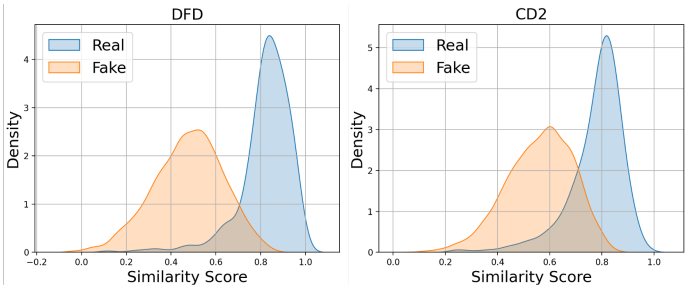



Fig. 10. **Distribution of similarity scores**. The probability density distribution of similarity score in DFD and CD2 datasets. It can be observed that the boundary between real and fake is stably maintained around 0.7 among different datasets.

we test the model performance on multiple $\tau_a$ from 0.2 to 0.8 with $\tau_i = 0.1$. It can be observed that STIDNet achieves the best performance with $\tau_i = 0.1$ and $\tau_a = 0.5$ respectively. We analyze that, it is harder for model to transfer the knowledge from other modalities than the inter-modal learning. Therefore, a smaller video-image temperature $\tau_i$ and a larger video-audio temperature $\tau_a$ is more beneficial for discriminative knowledge distillation. Moreover, the video-image temperature $\tau_i$ can not be too small for better model convergence. Thus, the setting of $\tau_i = 0.1, \tau_a = 0.5$ achieves the best performance, and we implement it in the optimization of STIDNet.

Moreover, Figure 8 illustrates the effect of scaling factors $\lambda_{image}$ and $\lambda_{audio}$ in Formula 4. Specifically, Figure 8a analyzes the effect of $\lambda_{image}$ with a fixed $\lambda_{audio} = 0.5$. Similarly, Figure 8b shows the model performance with a various $\lambda_{audio}$ from 0.2 to 0.8, and we accordingly set the $\lambda_{image} = 1$. It can be observed that the setting of $\lambda_{image} = 1$ and $\lambda_{audio} = 0.5$ achieves the best performance respectively, which also indicates that it is beneficial to assign a relative larger weight for spatial perspective knowledge distillation. Therefore, we uniformly set $\lambda_{image} = 1$ and $\lambda_{audio} = 0.5$ during implementation.

**Visualization.** To intuitively show the forgery discriminating

ability of STIDNet, we randomly sample four identities from DFD dataset and adopt t-SNE [75] to visualize their identity embeddings respectively. As is shown in Figure 9, Although our STIDNet is trained under a totally generic dataset without any manipulation, it still exhibits excellent discriminating ability when managing manipulated faces. For the same identity, boundary between real and fake faces is clear and the distribution appears good intra-class compactness.

To further illustrate feature distribution of real and fake faces, Figure 10 presents the probability density distribution of similarity score in DFD and CD2 datasets. We can observe that the real and fake samples tend to appear as bimodal distribution, while real faces get a higher similarity score with the reference set. Moreover, the similarity score boundary between real and fake samples is stably maintained around 0.7 among various datasets. It allows us to simply set the decision threshold as 0.7 in real world scenario implementation, while other methods like ID-Reveal [40] needs to test on a set of data for determining threshold each time.

## V. Conclusion

In this paper, we detailedly discuss the generation characteristic of manipulated videos, and point out it is essential to take both spatial and temporal perspectives into consideration for effective identity-aware face forgery detection. Correspondingly, a novel **SpatioTemporal IDentity Network** (STIDNet) is proposed to comprehensively capture spatiotemporal identity forensics clues with multi-teacher knowledge distillation. Extensive experiments demonstrate the superiority and robustness of our method. In comparison with other methods, it is surprising to find that our STIDNet can effectively utilize the reference video as a kind of strong identity prior, and accordingly promote the performance of forgery detection. Moreover, two experienced teachers also provide the equally essential spatiotemporal supervision in STIDNet, thus promising the adequate forensics clues collection in generalization and robustness evaluation. In the future, we will explore how to take full advantage of the pretrained identity extraction models, and further exploit more abundant and discriminative identity characteristics for identity-aware face forgery detection.

**Acknowledgements.** This work is supported by the National Key Research and Development Program of China (No.2023QY0204), the National Nature Science Foundation of China (U23B2028, 62232006, U2336202, 62032006, 62102127), the UK Engineering and Physical Sciences Research Council (EPSRC) Grants Ref.EP/M026981/1, EP/T021063/1, EP/T024917/.

## References

[1] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[2] D. Tran, H. Wang, L. Torresani, and M. Feiszli, "Video classification with channel-separated convolutional networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5552–5561.

[3] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5074–5083.

[4] R. Chen, X. Chen, B. Ni, and Y. Ge, "Simswap: An efficient framework for high fidelity face swapping," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2003–2011.

[5] Y. Ren, Y. Xiao, Y. Zhou, Z. Zhang, and Z. Tian, "Cskg4apt: A cybersecurity knowledge graph for advanced persistent threat organization attribution," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[6] J. Qiu, L. Du, D. Zhang, S. Su, and Z. Tian, "Nei-tte: intelligent traffic time estimation based on fine-grained time derivation of road segments for smart city," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 4, pp. 2659–2666, 2019.

[7] C. Lu, W. Zheng, H. Lian, Y. Zong, C. Tang, S. Li, and Y. Zhao, "Speech emotion recognition via an attentive time–frequency neural network," *IEEE Transactions on Computational Social Systems*, 2022.

[8] A. Dash, A. Chakraborty, S. Ghosh, A. Mukherjee, and K. P. Gummadi, "Fairir: Mitigating exposure bias from related item recommendations in two-sided platforms," *IEEE Transactions on Computational Social Systems*, 2022.

[9] M. Rattaphun, W.-C. Fang, and C.-Y. Chiu, "Attention on global-local representation spaces in recommender systems," *IEEE Transactions on Computational Social Systems*, 2021.

[10] W.-H. Cheng, S. Song, C.-Y. Chen, S. C. Hidayati, and J. Liu, "Fashion meets computer vision: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 4, pp. 1–41, 2021.

[11] H.-J. Chen, K.-M. Hui, S.-Y. Wang, L.-W. Tsao, H.-H. Shuai, and W.-H. Cheng, "Beautyglow: On-demand makeup transfer framework with reversible generative network," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2019, pp. 10 042–10 050.

[12] S. C. Hidayati, T. W. Goh, J.-S. G. Chan, C.-C. Hsu, J. See, L.-K. Wong, K.-L. Hua, Y. Tsao, and W.-H. Cheng, "Dress with style: Learning style from joint deep embedding of clothing styles and body shapes," *IEEE Transactions on Multimedia*, vol. 23, pp. 365–377, 2020.

[13] Y.-H. Kuo, H.-T. Lin, W.-H. Cheng, Y.-H. Yang, and W. H. Hsu, "Unsupervised auxiliary visual words discovery for large-scale image object retrieval," in *CVPR 2011*. IEEE, 2011, pp. 905–912.

[14] C.-H. Chang, M.-C. Hu, W.-H. Cheng, and Y.-Y. Chuang, "Rectangling stereographic projection for wide-angle image visualization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2824–2831.

[15] C. H. Sio, Y.-J. Ma, H.-H. Shuai, J.-C. Chen, and W.-H. Cheng, "S2siamfc: Self-supervised fully convolutional siamese network for visual tracking," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1948–1957.

[16] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–14, 2018.

[17] "Deepfakes," https://github.com/deepfakes/faceswap.

[18] "Faceswap," https://github.com/MarekKowalski/FaceSwap.

[19] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387–2395.

[20] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–12, 2019.

[21] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3207–3216.

[22] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (dfdc) preview dataset," *arXiv preprint arXiv:1910.08854*, 2019.

[23] "Deepfake detection dataset," https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html.

[24] L. Floridi, "Artificial intelligence, deepfakes and a future of ectypes," *Philosophy & Technology*, vol. 31, no. 3, pp. 317–321, 2018.

[25] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[26] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," *arXiv preprint arXiv:1811.00656*, 2018.

[27] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.

[28] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2018, pp. 1–7.

[29] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," in *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2019, pp. 1–8.

[30] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *European conference on computer vision*. Springer, 2020, pp. 86–103.

[31] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8261–8265.

[32] T. Jung, S. Kim, and K. Kim, "Deepvision: Deepfakes detection using human eye blinking pattern," *IEEE Access*, vol. 8, pp. 83 144–83 154, 2020.

[33] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face x-ray for more general face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5001–5010.

[34] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 720–18 729.

[35] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5039–5049.

[36] H. Zhao, W. Zhou, D. Chen, W. Zhang, and N. Yu, "Self-supervised transformer for deepfake detection," *arXiv preprint arXiv:2203.01265*, 2022.

[37] J. Liang, H. Shi, and W. Deng, "Exploring disentangled content information for face forgery detection," *arXiv preprint arXiv:2207.09202*, 2022.

[38] S. Dong, J. Wang, J. Liang, H. Fan, and R. Ji, "Explaining deepfake detection by analysing image matching," in *European Conference on Computer Vision*. Springer, 2022, pp. 18–35.

[39] S. Agarwal, H. Farid, T. El-Gaaly, and S.-N. Lim, "Detecting deepfake videos from appearance and behavior," in *2020 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2020, pp. 1–6.

[40] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, and L. Verdoliva, "Id-reveal: Identity-aware deepfake video detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 108–15 117.

[41] S. Ramachandran, A. V. Nadimpalli, and A. Rattani, "An experimental evaluation on deepfake detection using deep face recognition," in *2021 International Carnahan Conference on Security Technology (ICCST)*. IEEE, 2021, pp. 1–6.

[42] X. Dong, J. Bao, D. Chen, T. Zhang, W. Zhang, N. Yu, D. Chen, F. Wen, and B. Guo, "Protecting celebrities from deepfake with identity consistency transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9468–9478.

[43] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes." in *CVPR workshops*, vol. 1, 2019, p. 38.

[44] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[45] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, 2020, pp. 5781–5790.

[46] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 044–15 054.

[47] A. Haliassos, R. Mira, S. Petridis, and M. Pantic, "Leveraging real talking faces via self-supervision for robust forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 950–14 962.

[48] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, and L. Ma, "Delving into the local: Dynamic inconsistency learning for deepfake video detection." AAAI, 2022.

[49] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[50] H. Li, "Exploring knowledge distillation of deep neural networks for efficient hardware solutions," *University Of Stanford: CS230 course report*, 2018.

[51] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers." in *Interspeech*, 2017, pp. 3697–3701.

[52] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[53] B. Heo, M. Lee, S. Yun, and J. Y. Choi, "Knowledge distillation with adversarial samples supporting decision boundary," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3771–3778.

[54] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[55] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," *Advances in neural information processing systems*, vol. 29, 2016.

[56] Y. Chen, Y. Xian, A. Koepke, Y. Shan, and Z. Akata, "Distilling audio-visual knowledge by compositional contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7016–7025.

[57] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2827–2836.

[58] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3967–3976.

[59] L. Liu, Q. Huang, S. Lin, H. Xie, B. Wang, X. Chang, and X. Liang, "Exploring inter-channel correlation for diversity-preserved knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8271–8280.

[60] A. Thatipelli, S. Narayan, S. Khan, R. M. Anwer, F. S. Khan, and B. Ghanem, "Spatio-temporal relation modeling for few-shot action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 958–19 967.

[61] C. Yu, X. Ma, J. Ren, H. Zhao, and S. Yi, "Spatio-temporal graph transformer networks for pedestrian trajectory prediction," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 2020, pp. 507–523.

[62] H. Wang, J. Chen, T. Pan, Z. Fan, X. Song, R. Jiang, L. Zhang, Y. Xie, Z. Wang, and B. Zhang, "Easy begun is half done: spatial-temporal graph modeling with st-curriculum dropout," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 4, 2023, pp. 4668–4675.

[63] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 626–640, 2019.

[64] 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.

[65] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[66] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274.

[67] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct 2016.

[68] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[69] 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.
[70] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742.
[71] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
[72] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
[73] 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.
[74] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2889–2898.
[75] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

**Mingqi Fang** received the B.E. degree in electronics and information engineering at Northwestern Polytechnical University, Xi'an, China, in 2021. He is currently pursuing the M.S. degree at School of Information Science and Technology, University of Science and Technology of China, Hefei, China. His research interests include face analysis and face forgery detection.

**Lingyun Yu** received her Ph.D. degree in Control Science and Engineering from University of Science and Technology of China, in 2020. She is currently an Associate Professor with the School of Information Science and Technology, University of Science and Technology of China, Hefei, China. Her research interests cover talking head generation, face forgery detection, multi-modal learning, and video synthesis.

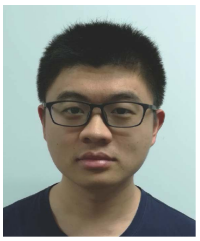
**Hongtao Xie** received the Ph.D. degree in computer application technology from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2012. He is currently a Professor with the School of Information Science and Technology, University of Science and Technology of China, Hefei, China. His research interests include multimedia content analysis and retrieval, deep learning, and computer vision.

**Qingfeng Tan** received the Ph.D. degree in information security from the University of Chinese Academy of Sciences, Beijing, China, in 2017. He is currently an Associate Professor with the Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou, China. His current research interests include anonymous communication and privacy protection.

**Zhiyuan Tan** is a associate professor at the School of Computing at the Edinburgh Napier University (ENU), the United Kingdom. His current research interests include cybersecurity, machine learning, data analytics, virtualisation, and cyberphysical system. Dr Tan is now working on the following research topics: (1) Adversarial machine learning for Anomaly/Malware detection; (2) Virtualisation security based on non-parametric behaviour modelling; (3) Knowledge transfer (Transfer Machine Learning) in cyber-security problems; and (4) IoT Security with focuses on Cloud and Edge computing security issues.

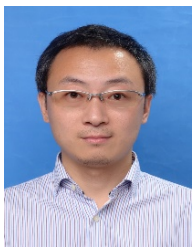
**Amir Hussain** received his B.Eng (highest 1st Class Honours with distinction) and Ph.D degrees, from the University of Strathclyde, Glasgow, U.K., in 1992 and 1997, respectively.

Following postdoctoral and academic positions at the Universities of West of Scotland (EPSRC postdoctoral fellow: 1996-98), Dundee (Research Lecturer: 1998-2000) and Stirling (Lecturer: 2000-4; Senior Lecturer: 2004-8; Reader: 2008-12; Professor: 2012-18) respectively, he joined Edinburgh Napier University (ENU) in 2018 as a Professor in the School of Computing. He is founding Director of the Centre for AI and Robotics (CAIR) and Head of the Data Science and Cyber Analytics (DSCA) Research Group (managing over 20 academics and research staff). He is also founding Head of the Cognitive Big Data Analytics (CogBiD) Research Lab, and co-Lead of the Centre for Cardio-Vascular Health (with the School of Health and Social Care).



**Zezheng Wang** received the BS and MS degrees from Tianjin University, Tianjin, China, in 2015 and 2018, respectively. He is currently an artificial intelligence engineer with Xiaohongshu Inc., Beijing, China. His current research interests include, machine learning, face analysis, face anti-spoofing, and multi-modal analysis.

**Jiahong Li** received the BS degrees from Huazhong University of Science and Technology, Wuhan, China, in 2010 and MS degrees from University of Chinese Academy of Sciences, Beijing, China, in 2013. He is currently an artificial intelligence research and development engineer with Beijing Kwai Technology Co., Ltd, Beijing, China. His current research interests include computer vision, multimedia, and face anti-spoofing.



**Zhihong Tian** is currently a Professor, and Dean, with the Cyberspace Institute of Advanced Technology, Guangzhou University, Guangdong Province, China. Guangdong Province Universities and Colleges Pearl River Scholar (Distinguished Professor). He is also a part-time Professor at Carlton University, Ottawa, Canada. Previously, he served in different academic and administrative positions at the Harbin Institute of Technology. He has authored over 200 journal and conference papers in these areas. His research interests include computer networks and cyberspace security. His research has been supported in part by the National Natural Science Foundation of China, National Key research and Development Plan of China, National High-tech R&D Program of China (863 Program), and National Basic Research Program of China (973 Program). He also served as a member, Chair, and General Chair of a number of international conferences. He is a Senior Member of the China Computer Federation, and a Member of IEEE.