ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240548

Unified Video and Image Representation for Boosted Video Face Forgery Detection

Haotian Liu [a], **Chenhui Pan** [b], **Yang Liu** [a], **Guoying Zhao** [a] **and Xiaobai Li** [b, a,*]

[a]Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland
[b]State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou, China

Abstract. Face forgery detection is crucial in preserving the security and integrity of facial data amidst the rapid developments in face manipulation techniques and deep generative models. Existing methods for video face forgery detection typically assume that all frames in a forged video are manipulated, while identifying partially forged videos with only a subset of altered frames is still a challenge to be solved. To address this issue, we propose a novel framework, i.e., the UVIF, that utilizes additional annotated images to provide fine-grained supervision for detecting partial forgeries in videos. The UVIF integrates a unified encoder and a multi-task learning paradigm to model both facial videos and images for boosted video face forgery detection. A 2D backbone with temporal fusion modules is employed for the unified encoder. A pseudo labeling process is also designed for facial video frames to bridge the representation of individual video frames and static images. Extensive experiments on benchmark datasets demonstrate the effectiveness of our framework, outperforming state-of-the-art methods in detecting partially forged videos while introducing no additional computational overhead. Our code is available at https://github.com/haotianll/UVIF.

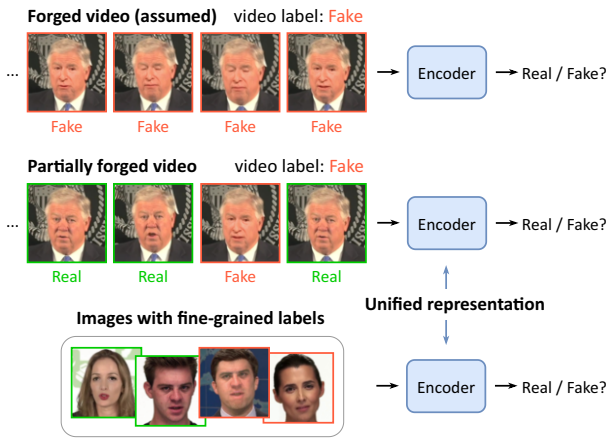

Figure 1. Previous video forgery detection studies assumed all frames of forged videos are fake, while in reality some videos might be partially forged containing a subset of fake frames. This poses challenges for video face forgery detection in realistic scenarios. To address this issue, this paper proposes utilizing additional images to provide fine-grained supervision for detecting partial forgeries in videos, by employing unified representations for facial videos and images.

1 Introduction

Face forgery detection aims to distinguish between authentic and fabricated faces [25]. This task is essential in preventing malicious uses of face manipulation techniques and AI-generated content (AIGC) [18, 41], thereby upholding the security and integrity of facial data.

Current research on video forgery detection is primarily categorized into image-based and video-based methods. Image-based methods [1, 42, 56, 40, 4, 37] utilize static facial images as inputs and perform image classification to verify their authenticity. When dealing with a video clip, these methods extract individual facial frames and assign classification labels to each frame. In contrast, video-based methods [19, 23, 27, 57, 49, 55] process facial video clips directly with 3D backbones and complete a video classification task. Given the critical role of temporal inconsistency between video frames in forgery analysis, video-based methods [27, 57, 49, 55] tend to achieve higher accuracy compared to image-based methods.

A significant limitation of current studies on face forgery detection is the assumption [42] that all frames in a fake video are manipulated, underpinning the preprocessing pipelines, model designs, and training strategies commonly employed. However, such an assumption does not hold across all forgery detection tasks in realistic scenarios, that some fake videos may only contain a portion of manipulated frames. This discrepancy poses challenges for both image-based and video-based detection methods previously proposed. For image-based methods, there are no fine-grained labels for each video frame, making them impractical to process partial forgery videos for training. For video-based methods, tailored strategies need to be designed in the model pipelines to manage the uncertainty presented by a mix of authentic and forged frames. Due to the lack of fine-grained supervision for each video frame, the feature extraction process in video-based methods will also be impacted.

Some previous work [32] tried to use multiple instance learning (MIL) [24] to tackle partially forged videos, but the performance of this approach is limited due to lack of fine-grained supervision. Specifically, MIL uses a set of labeled bags containing many instances for training. For a binary classification, a positive bag can contain both positive and negative instances, while a negative bag contains only negative ones. The MIL resembles the partial video forgery detection task, i.e., a forgery video can be viewed as a positive bag. However, current MIL methods [32, 45, 44, 54, 35] only focus on group instances based on feature similarity, without access to instance-level labels during training, limiting their effectiveness in

* Corresponding Author. Email: xiaobai.li@zju.edu.cn.

detecting partially forged videos.

Intuitively, detecting forged frames is essential in verifying the authenticity of facial videos. The knowledge of detecting forged images can also contribute to this, as they have a similar representation learning process, i.e., extracting discriminate features related to forgery clues. Facial image instances with fine-grained labels are readily available in existing face forgery detection datasets. Therefore, we propose incorporating annotated facial images to improve the detection accuracy of partially forged videos, as illustrated in Figure 1.

In light of this motivation, we propose a novel framework named UVIF, i.e., Unified Video and Image representation for Forgery detection in facial videos. The UVIF framework can simultaneously process both video and image inputs with one single model. This integration is facilitated by a 2D backbone combined with temporal fusion designs. It follows a multi-task learning optimization paradigm, which encompasses both video and image face forgery detection. Here, image forgery detection serves as an auxiliary task that introduces fine-grained supervision to enhance video forgery detection. As a result, our UVIF framework can significantly improve the model's representation for partial video face forgery detection.

The contributions of this paper include:

1) We propose a novel method, the UVIF, that utilizes additional annotated images to provide fine-grained supervision for detecting partial forgeries in videos.

2) In UVIF, a unified encoder and a multi-task learning paradigm are integrated to model both facial videos and images for boosted video face forgery detection. A 2D backbone with a temporal fusion module is employed for the unified encoder.

3) A pseudo labeling process is designed for facial video frames to bridge the representation of video frames and static images.

4) Extensive experiments demonstrate that UVIF significantly outperforms SOTA methods in detecting partially forged videos. Further ablation tests show the efficiency of the proposed approach, that a small set of added images e.g., 20k, are sufficient to achieve a significant performance boost across videos with various forgery ratios, e.g., as low as 10%.

2 Related Work

2.1 Face Forgery Detection

Face forgery detection aims to detect forged or synthesized faces in images and videos, which is very important for the authenticity of visual information that we see every day. Some methods [1, 42] directly used convolution neural networks (CNNs) and performed a binary classification task for image face forgery detection. Subsequent approaches further exploit spatial forgery patterns of facial images, including local textures [56], frequency domain [40, 29, 37], and inconsistency information [31, 11, 46, 4]. It is worth noting that some image-based methods have been applied to video face forgery detection. They converted facial videos to individual image frames and generated classification labels for each frame during training. However, these methods assumed that all frames in a facial video are the same kind, i.e., all as real, or all as fake, making them unsuitable to process partially forged videos.

As temporal inconsistency between video frames is also crucial, many video face forgery detection methods [19, 23, 27, 57, 49, 55] have been proposed leveraging the inconsistency information for video face forgery detection. They took the temporal dimension into consideration and processed facial video inputs directly. Early studies focused on mining temporal clues by using prior knowledge, such as eye blinks [33], lip motions [21], and biological signals [7]. Some methods [19, 20, 23, 49] directly processed facial video clips with a 3D CNN or recurrent neural network (RNN), which achieved better accuracy than image-based methods. With the advance of vision transformers [12, 2], recent works [27, 57, 55] also proposed using transformer structures to enhance the representation of inter-frame relations during feature extraction. Despite their swift progress, the task of partially forged video detection still needs to be further explored.

2.2 Multiple Instance Learning

Multiple instance learning (MIL) is a form of weakly supervised learning with broad applications in medical imaging and video analysis [5]. In MIL, a bag of instances is annotated with a single bag-level label, while the exact label for each instance is unavailable. For a binary classification, negative bags only contain negative instances, while positive bags can contain both positive and negative instances. Early studies [15, 52] of MIL focused on extracting and aggregating features of each instance in a bag via deep neural networks and pooling operations. Ilse et al. [24] first proposed utilizing the attention mechanism to increase the weights of key instances and thus enhance the representation of bag features. Recently, some MIL approaches [28, 44, 54, 35] have also been developed based on attention mechanisms and contrastive learning. Li et al. [32] proposed a sharp MIL, i.e., S-MIL, for face forgery detection to handle the problem of partially manipulated faces in videos. It treated facial frames and videos as instances and bags in MIL, respectively, and designed a sharp loss emphasizing hard instances to address the partially forged videos. Nevertheless, the performance of MIL methods is limited due to the lack of instance-level labels during training. In this paper, we propose to address the partial forgery video detection task from a new perspective, i.e., by incorporating additional image instances with annotations to compensate for the absence of fine-grained supervision information.

2.3 Unified Architecture Design

The unified architecture design [13, 16] has gained significant attention recently. It can process input data of different modalities or perform multi-task learning with a single model. As video and image data are highly related in structure, i.e., video can be viewed as a sequence of images, some methods [34, 3, 2, 39, 38] have been developed to introduce temporal modeling to CNN [22] or transformer [12] backbones originally designed for 2D images for video data processing. Moreover, other methods [13, 39] investigated using image data and fine-grained annotations to assist video tasks by joint training of video and image data. With the increasing popularity of transformers for visual tasks, some latest methods [16, 17, 50] used a shared transformer backbone to encode or align data of multiple modalities, such as image, video, audio, and text, into a unified feature space. Due to the good generalization and scaling capability of transformers, these methods can learn excellent feature representation from diverse training data across modalities. The unified architecture design has been demonstrated promising results in other visual tasks, which inspires us to apply it for detecting forged facial videos integrated with annotated images.

3 Methodology

An overview of the proposed UVIF method is illustrated in Figure 2, which integrates a unified encoder and a multi-task learning

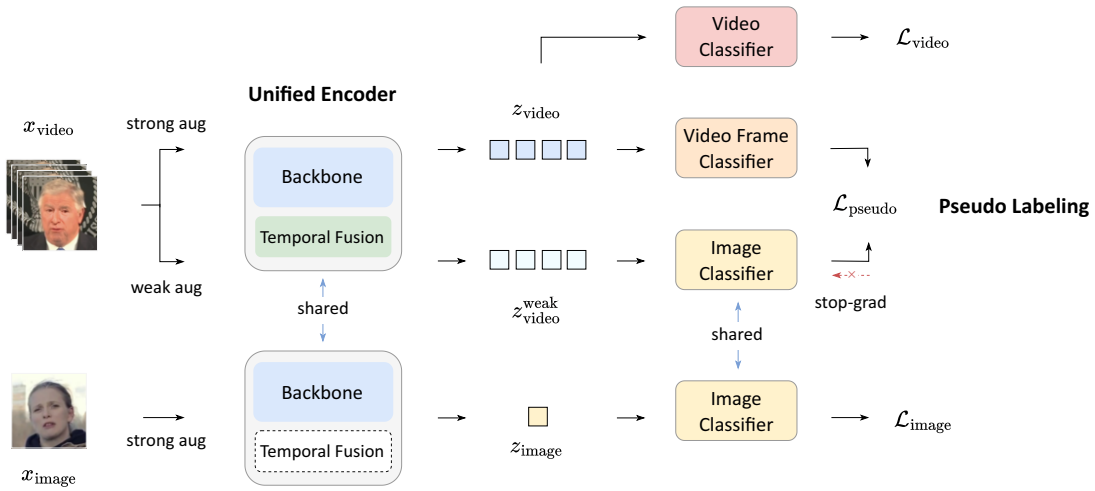

Figure 2. The UVIF framework comprises three primary components: 1) A unified encoder extracts features from both facial videos and images, and its temporal fusion modules are only applied to video clips. 2) A multi-task learning paradigm is adopted for the training of the video and image classifiers. 3) An auxiliary pseudo labeling process to bridge the representations of video frames and images. Both weak and strong augmentations are applied to each video input to create two views for pseudo labeling. The UVIF utilizes only the unified encoder and video classifier for video forgery detection during testing.

paradigm to model facial videos and images for boosted video face forgery detection. In the following parts, we first detail the problem formulation and our motivation for detecting partial forgery videos (Section 3.1). Then, we introduce how to achieve unified modeling of both facial videos and images within a single model (Section 3.2). Additionally, to bridge the representation of video frames and static images, we design a novel pseudo labeling process for video frames (Section 3.3). Finally, the detailed architecture of the proposed UVIF framework is described (Section 3.4).

3.1 Problem Formulation

This paper concerns video face forgery detection, focusing on differentiating whether a video instance contains fake faces, as a binary classification task. Let X represent a facial video clip and Y represent the binary classification label of the entire video. In which, $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ consists of a sequence of video frames, and T is the number of frames. $Y \in \{0, 1\}$ denotes whether the video is real or fake, respectively. The goal of the video face forgery detection task is to train a model that can accurately predict the binary label Y for a given video input X.

For real video samples, the faces in every frame are genuine. Conversely, if faces in one or more frames of a video are manipulated, the entire video is labeled as fake. Most previous research only considered the ideal prerequisite that every frame of a fake video is fake, which does not hold in realistic scenarios. This paper aims to address the more complex issue of partial face forgery detection, where a fake video may contain only a portion of fake frames.

Specifically, assuming that there are binary frame labels $Y_{\text{frame}} = \{y_1, y_2, ..., y_T\}$ for each video frame, where $y_i \in \{0, 1\}$, for $i = 1, 2, ..., T$, we can describe this premise as:

$$Y = \begin{cases} 1, & \exists y_i = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

This formulation shares a similar setting as multiple instance learning (MIL) [24]. We can take video frames $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ as a bag of instances in MIL, and Y is the classification label of the bag. During training, only the label Y is used while the individual labels Y_{frame} of each frame are not available.

The description above indicates that detecting fake frames is essential in verifying the authenticity of facial videos. This requires the classification model to extract discriminate features related to fake frames. Nevertheless, the model only uses video-level labels during training, resulting in a lack of fine-grained supervision for each video frame. One direct approach is annotating each frame in a video with classification labels, but it is time-consuming and impractical for long and complex facial videos.

In this paper, we introduce fine-grained supervision to partial video forgery detection with an additional set of facial images, which is more efficient and feasible. For an image face forgery dataset, let $X_{\text{image}} = \{\mathbf{x}'_1, \mathbf{x}'_2, ..., \mathbf{x}'_K\}$ denote a set of K facial images, and $Y_{\text{image}} = \{y'_1, y'_2, ..., y'_K\}$ represent their corresponding binary classification labels. Compared to the video label Y, the image label Y_{image} is more fine-grained and precise, and such annotated facial images are readily available in face forgery detection datasets. As a static image can be viewed as a frame in a facial video, the annotated pair of X_{image} and Y_{image} can be regarded as a substitute for the unavailable pair X and Y_{frame} of video frames. Therefore, our research focuses on how to utilize a set of labeled facial images to provide fine-grained supervision for the detection of partially forged videos.

3.2 Unified Video and Image Modeling

To effectively utilize a set of facial images in boosting partial face forgeries in videos, a unified model is required that can process both video and image inputs and perform classification tasks for both modalities during training.

Inspired by the popular unified architecture designs [34, 2, 39], we propose to use a 2D backbone with temporal fusion modules to achieve unified video and image modeling. We choose this architecture for two reasons: First, although typical 3D-based backbones are the prior choice for video classification tasks, they encounter incompatible or redundant problems when processing image data due to complicated temporal fusion operations or pipelines. Second, using two unshared backbones for video and image data will result in distinct feature spaces, and implicit interactions between video and image features will be significantly constrained during training. Therefore, we decide to use a 2D backbone to implement the unified en-

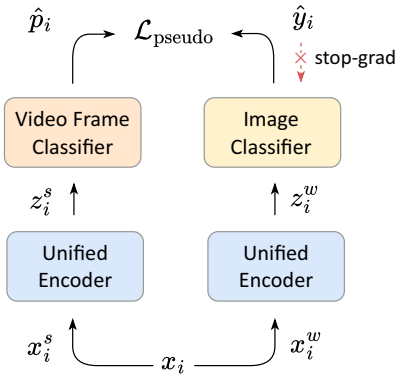

Figure 3. Illustration of the pseudo labeling process. For each video frame x_i, the image classifier uses weak augmented view x_i^w to generate the pseudo label $\hat{y}_i$, and the video frame classifier uses strong augmented x_i^s to predict the probability $\hat{p}_i$.

coder for facial video and image inputs. When dealing with video inputs, temporal fusion modules are applied to compensate for the lack of temporal interaction in 2D backbones. The detailed architecture is introduced in Section 3.4.

During training, the framework is optimized following the multi-task learning paradigm. A training batch is constructed by randomly sampling a set of facial videos and images. Two unshared classifiers are employed to make predictions of video and image inputs. The classification losses $\mathcal{L}_{\text{video}}$ and $\mathcal{L}_{\text{image}}$ for video and image are computed as follows:

$$\mathcal{L}_{\text{video}} = \mathcal{L}_{\text{CE}}(Y, p), \tag{2}$$
$$\mathcal{L}_{\text{image}} = \mathcal{L}_{\text{CE}}(y', p'), \tag{3}$$

where p and p' are the predicted probabilities of a facial video or image, Y and y' are corresponding ground truth labels, and $\mathcal{L}_{\text{CE}}$ is the standard cross entropy loss in classification tasks. $\mathcal{L}_{\text{video}}$ and $\mathcal{L}_{\text{image}}$ are further averaged according to the number of videos and images in the training batch.

In this way, the model is supervised by both video and image inputs. Apart from video face forgery detection, it also learns to extract discriminative features to identify the authenticity of facial images. As a result, the model's representation is improved by incorporating the fine-grained labels from the image set for training.

3.3 Bridging Video and Image Representation

According to the problem formulation, the unified modeling process actually takes the annotated image set $X_{\text{image}} = \{\mathbf{x}'_1, \mathbf{x}'_2, ..., \mathbf{x}'_K\}$ and $Y_{\text{image}} = \{y'_1, y'_2, ..., y'_K\}$ as substitutes of original video frames $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$, while the frame-level labels $Y_{\text{frame}} = \{y_1, y_2, ..., y_T\}$ are still not available for training. In other words, the model does not receive any direct supervision for each video frame in X during training. Therefore, we further introduce an auxiliary pseudo labeling process for facial videos. It generates pseudo labels $\hat{Y}_{\text{frame}} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_T\}$ for each video frame based on the image classifier, aiming to bridge the representation of individual video frames and images.

As shown in Figure 3, the pseudo labeling process involves two augmented views $\mathbf{x}_i^{\text{w}}$ and $\mathbf{x}_i^{\text{s}}$ of each video frame $\mathbf{x}_i$, for $i = 1, 2, ..., T$. In which, $\mathbf{x}_i^{\text{w}}$ is only applied with weak augmentations, including resizing, cropping, and horizontal flipping, while $\mathbf{x}_i^{\text{s}}$ is a view with additional strong augmentation, such as image compression and color perturbations. The two views $\mathbf{x}_i^{\text{w}}$ and $\mathbf{x}_i^{\text{s}}$ are fed into the unified encoder with shared parameters to obtain two output feature vectors $\mathbf{z}_i^{\text{w}}$ and $\mathbf{z}_i^{\text{s}}$, respectively. The image classifier is applied to $\mathbf{z}_i^{\text{w}}$ from the view with weak augmentation to get the soft pseudo label $\hat{y}_i$. Then, another video frame classifier is introduced, and it takes $\mathbf{z}_i^{\text{s}}$ as input to predict the probabilities $\hat{p}_i$ for each video frame. Note that the video frame classifier has the same structure as the image classifier but with unshared parameters.

Based on the outputs $\hat{y}_i$ and $\hat{p}_i$ of two views, we define the pseudo labeling loss for each video frame as:

$$\mathcal{L}_{\text{pseudo}} = \mathcal{L}_{\text{CE}}(\text{stopgrad}(\hat{y}_i), \hat{p}_i), \tag{4}$$

where $\mathcal{L}_{\text{CE}}$ denotes the cross entropy, and $\mathcal{L}_{\text{pseudo}}$ is averaged over all the video frames. A stop-gradient operation $\text{stopgrad}(\cdot)$ is applied to $\hat{y}_i$ to avoid the collapse of training.

The pseudo labeling process is similar to the semi-supervised learning frameworks [48, 6]. It generates reliable pseudo labels from a weak augmented view and provides supervision to a strong view input during training. As the image classifier is applied for the generation of pseudo labels, the pseudo labeling loss $\mathcal{L}_{\text{pseudo}}$ can guide the unified encoder to minimize the gap between features of video frames and static images. Besides, as the video frame classifier is dropped during testing, this pseudo labeling process only affects the representation learning of the unified encoder.

3.4 The Architecture

The UVIF framework comprises three primary components, i.e., a unified encoder, video and image classifiers, and an auxiliary pseudo labeling process. It adopts a multi-task learning paradigm for optimization, as depicted in Figure 2.

3.4.1 Unified Encoder

The unified encoder processes two sets of inputs in a batch, i.e., images and video clips. We define an input video clip $x_{\text{video}} \in \mathbb{R}^{T\times 3\times H_0\times W_0}$ and an input image $x_{\text{image}} \in \mathbb{R}^{3\times H_0\times W_0}$. In which, T denotes the sampled frames of a video clip during training, H_0 and W_0 are the input height and width of video frames or images, and both x_{video} and x_{image} are applied with strong data augmentation. The unified encoder then generates a video feature map $f_{\text{video}} \in \mathbb{R}^{T\times C\times H\times W}$ and an image feature map $f_{\text{image}} \in \mathbb{R}^{C\times H\times W}$, respectively. Here, C is the channels of the feature map, while H and W are the height and width of the feature map.

The unified encoder is implemented by using typical 2D CNN [22, 53] or transformer [12] backbones. For the design of temporal fusion modules, we utilize the Temporal Shift Module (TSM) [34] for CNN backbones [22, 53], and we use 3D positional encoding [2, 51] and temporal attention operations [2, 51] for transformer backbones including ViT [12].

3.4.2 Video and Image Classifiers

Two unshared classifiers are applied to accomplish video and image face forgery detection tasks. Specifically, the video classifier first performs global average pooling (GAP) over $f_{\text{video}} \in \mathbb{R}^{T\times C\times H\times W}$ to obtain feature vectors $z_{\text{video}} \in \mathbb{R}^{T\times C}$ for each video frame, and the image classifier generates $z_{\text{image}} \in \mathbb{R}^{C}$ in a similar way. Then, two vanilla multilayer perceptrons (MLPs) are used to get the predicted probabilities for each video or image. Note that the predicted probability of a video is averaged over the predictions of all video frames following [34].

Table 1. Comparison with the state-of-the-art methods on the ForgeryNet [23] validation set. The results marked with † are cited from the original ForgeryNet paper [23], where the code and models are not publicly available. All the remaining results are reimplemented using the same protocol for a fair comparison. The #params denotes the number of parameters, and the FLOPs are measured under the spatial size 224 × 224. Bold indicates the best results.

Method	Backbone	#params (M)	FLOPs (G)	Acc	AUC
TSM [34] †	ResNet-50	24	132	88.04	93.05
SlowFast [14] †	3D ResNet-50	34	51	**88.78**	93.88
TSM [34]	ResNet-50	24	132	80.89	88.66
TSM [34]	ResNet-101	43	251	81.48	88.08
SlowOnly [14]	3D ResNet-50	32	168	79.61	86.71
SlowFast [14]	3D ResNet-50	34	51	83.20	90.99
SlowFast [14]	3D ResNet-101	62	97	83.42	91.25
STIL [19]	SCNet-50	23	151	81.18	87.61
FTCN [57]	3D ResNet-50	57	68	74.40	80.08
TimeSformer [3]	ViT-B	86	281	78.11	86.60
Swin [36]	VideoSwin-T	28	88	80.57	88.17
Swin [36]	VideoSwin-S	50	166	82.38	89.96
VideoMAEv2 [51]	ViT-S	22	57	78.23	85.83
UniFormer [30]	UniFormer-S	21	110	82.59	89.16
MIL [24]	ResNet-50	24	132	81.41	88.28
S-MIL [32]	ResNet-50	24	132	81.38	88.33
DSMIL [28]	ResNet-50	24	132	81.55	88.44
UVIF (**Ours**)	ResNet-50	24	132	85.32	93.45
UVIF (**Ours**)	ResNet-101	43	251	86.57	**94.42**

3.4.3 Pseudo Labeling Process

Apart from the input video x_{video} with strong augmentation and its corresponding feature vector z_{video}, the pseudo labeling process involves another weak augmented video $x_{\text{video}}^{\text{weak}} \in \mathbb{R}^{T\times 3\times H_0\times W_0}$, and it is also fed into the unified encoder to get video feature map $f_{\text{video}}^{\text{weak}} \in \mathbb{R}^{T\times C\times H\times W}$. After global average pooling, the image classifier uses $z_{\text{video}}^{\text{weak}} \in \mathbb{R}^{T\times C}$ to generate the pseudo labels for each video frame, while another video frame classifier is introduced that uses z_{video} to get the predicted probabilities for each video frame.

3.4.4 Optimization

The overall framework is optimized end-to-end within a multi-task learning paradigm, employing the following loss function:

$$\mathcal{L} = \mathcal{L}_{\text{video}} + \mathcal{L}_{\text{image}} + \mathcal{L}_{\text{pseudo}} \,, \tag{5}$$

where $\mathcal{L}_{\text{video}}$ and $\mathcal{L}_{\text{image}}$ are the classification losses for the video and image classifiers, respectively, and $\mathcal{L}_{\text{pseudo}}$ is the pseudo labeling loss for each video frame. Note that the proposed method incurs no additional computational overhead during testing as only the video classifier is applied.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets and Evaluation Metrics

We conduct experiments on publicly available face forgery detection datasets ForgeryNet [23] and DFDC (preview) [10].

ForgeryNet is a large-scale face forgery detection dataset with over 220k facial video flips and 2.9m static images based on over 5k subjects. It contains 15 facial manipulation approaches and over 36 mix-perturbations, and the majority of forged videos in ForgeryNet are partially manipulated, which makes it very challenging for face forgery detection. For forgery video classification, we follow [23] and take about 140k videos as the training set and 18k as the validation set. As for the image set, ForgeryNet actually includes over 2.3m training images and 150k validation images, but we only train our method with a randomly selected subset of 100k images (less than 5%) from the entire training set if not specified.

DFDC is the preview dataset of the Deepfake Detection Challenge [10], which includes 1131 real facial video clips and 4113 forged ones from two unknown synthesis methods. DFDC also features many partially forged video clips that contain both actual and manipulated frames. We follow the original dataset partition in [10] and use 4464 videos for training and 780 for testing.

For evaluation metrics, we follow [23, 10] and adopt standard video-level Accuracy (Acc) and Area under the ROC curve (AUC) to evaluate the performance of video face forgery detection.

4.1.2 Implementation Details

We use MMEngine [8] to implement our method. As a typical case of our method, we adopt ResNet [22] equipped with TSM [34] for the unified encoder and utilize pre-trained weights on ImageNet-1k for initialization.

For data processing, we adopt some settings introduced in [23]. We use RetinaFace [9] to extract facial regions from each video frame or image. We enlarge the detected bounding boxes with a factor of 1.3 to obtain the cropped faces for training and evaluation. We randomly sample 32 frames with temporal stride 4 from each video clip during training and use a center clip of 32 frames for evaluation. We resize the resolution of each video frame or image to 224 × 224.

For data augmentation, we follow [23] and use weak and strong augmentations during training. Specifically, weak augmentation only uses geometric transforms, including random resizing of range $[1, 8/7]$, random cropping, and random horizontal flipping. In contrast, strong augmentation also applies a few more perturbations, such as image compression, random color distortions, Gaussian blur, CLAHE, and channel shuffle. We use strong augmentation as the default training setting of video frames or images for all the baselines and compared methods. Besides, our method also leverages both weak and strong augmentations to construct two different views of pseudo labeling process.

The batch size of video clips is set to 16 for video classification baselines during training. For each iteration, our method also randomly selects 128 images from the training image set. Note that our method randomly selects videos and images without requiring them

Table 2. Results on the DFDC [23] testing set. Results marked with † are cited from original papers. Bold indicates the best results.

Method	Acc	AUC
TSM-Res50 [34]	81.08	90.67
TSM-Res101 [34]	82.11	91.07
SlowFast-Res50 [14]	83.53	92.10
SlowFast-Res101 [14]	83.14	92.43
Xception-avg [42] †	84.58	-
STIL [19]	86.23	93.16
VideoMAEv2-ViT-S [51]	83.01	90.52
VideoSwin-T [36]	79.54	89.04
Uniformer-S [30]	85.07	94.02
S-MIL [32] †	83.78	-
S-MIL-T [32] †	85.11	-
UVIF-Res50 (**Ours**)	83.40	93.54
UVIF-Res101 (**Ours**)	**87.00**	**94.95**

Table 3. Effectiveness of the primary components of UVIF with ResNet-50, i.e., $\mathcal{L}_{\text{image}}$ in unified modeling, temporal fusion modules for videos, and $\mathcal{L}_{\text{pseudo}}$ in pseudo labeling. Bold indicates the best results.

$\mathcal{L}_{\text{image}}$	Temporal	$\mathcal{L}_{\text{pseudo}}$	Acc	AUC
-	-	-	68.24	75.92
✓	-	-	76.17	85.10
-	✓	-	80.89	88.66
✓	✓	-	84.41	92.85
✓	✓	✓	**85.32**	**93.45**

to contain corresponding subjects.

The proposed framework is trained end-to-end with an SGD optimizer on two NVIDIA Tesla V100 GPUs. The models are trained for 100k iteration on ForgeryNet [23] and 20k on DFDC [10] to reach converging. The base learning rate is set to 0.01. A linear warm-up schedule of 10^{-3} is used for the first 2k iterations, and then a one-cycle [47] decay schedule is applied. The weight decay is 10^{-4} and the SGD momentum is 0.9. We also utilize the same hyper-parameter setting for other compared methods unless specified.

4.2 Comparison to State of the Art

We evaluate the performance of our proposed methods on the ForgeryNet [23] dataset, and compare them with a range of previous state-of-the-art methods, including CNN-based and transformer-based methods for video face forgery detection, as well as typical multiple instance learning methods, as detailed in Table 1. Most comparison methods are initialized with pre-trained weights on Kinetics-400 [26] instead of ImageNet-1k [43] for better performance, except [34, 19]. The number of frames for [3, 51] is reduced to 16 due to the GPU memory limit. The TSM [34] methods can be viewed as the baselines for the UVIF, as they have the equivalent architecture during evaluation. The comparison results show that our proposed UVIF methods outperform previous methods by a large margin. The best model, UVIF-ResNet-101, achieves 86.57% accuracy and 94.42% AUC, surpassing SlowFast [14] by +3.15% and +3.17%, respectively. Compared to SlowFast, our UVIF has fewer parameters but more FLOPs due to the differences in architecture. Slowfast leverages 3D convolution and temporal pooling operations for feature extraction, while our UVIF is based on 2D backbones equipped with temporal fusion modules. Besides, MIL methods [24, 32, 28] are also implemented based on the architecture TSM-ResNet-50, and we take each video clip input as a bag in MIL. The results in Table 1 indicate that although MIL methods can improve the AUC to some extent by grouping similar video frames, their performance benefits are still limited. In contrast, our UVIF methods can significantly improve detection performance by utilizing fine-grained image annotations.

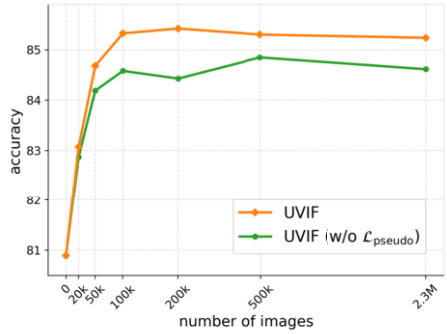

Figure 4. Ablation on the number of images for training. Results of two UVIF settings, i.e., with or without pseudo labeling process, are illustrated. All results are based on ResNet-50.

We also compare our proposed UVIF with existing methods on the DFDC [10] dataset, as presented in Table 2. Here, we use the selected image set of ForgeryNet to train our method. All methods extract video frames within the whole video clip for evaluation following the protocol of [10, 32, 19]. The results suggest that our UVIF can significantly enhance the detection accuracy compared to the TSM [34] baseline, e.g., +2.32% accuracy for ResNet-50. Our method also demonstrates competitive performance compared to the state-of-the-art methods on the DFDC dataset.

4.3 Ablation Study

4.3.1 Effectiveness of the Core Components

We first perform ablation experiments to analyze the components of the UVIF framework. Specifically, our UVIF framework contains three primary components: supervision $\mathcal{L}_{\text{image}}$ of images in unified modeling, temporal fusion modules for video inputs, and the loss $\mathcal{L}_{\text{pseudo}}$ in pseudo labeling process. The ablation results are shown in Table 3. The first row denotes training a vanilla ResNet-50 model on the videos from ForgeryNet [23] training set. Applying temporal fusion modules in row 3 can bring significant performance gains (+12.65% accuracy and +12.74% AUC), indicating the importance of temporal information in face forgery detection. Additionally, utilizing images together with videos for training (row 4) has led to a promising accuracy of 84.41%, highlighting the benefits of fine-grained annotations of images for video forgery detection. Interestingly, row 2 also shows that training images can still improve the performance of a model even without temporal fusion. This further reflects the effectiveness of annotated images, as the features of images and individual video frames are very similar. Finally, adding the pseudo labeling process during training in row 5 produces the best results for UVIF, indicating its efficacy in bridging the representations of images and video frames during training.

4.3.2 Number of Training Images

We then test the performance of UVIF using different numbers of training images, as illustrated in Figure 4. We randomly sample $20k, 50k, 100k, 200k, 500k$, and $2.3m$ (all) training images from ForgeryNet [23], and perform experiments under two UVIF settings, i.e., with or without pseudo labeling process. The results show that the accuracy of UVIF methods significantly improves when the training images increase from 0 to $100k$ and reaches saturation at $100k$. Thus, $100k$ is sufficient for the current video sample set, and adding more images, i.e., even up to $2.3m$ won't make further help.

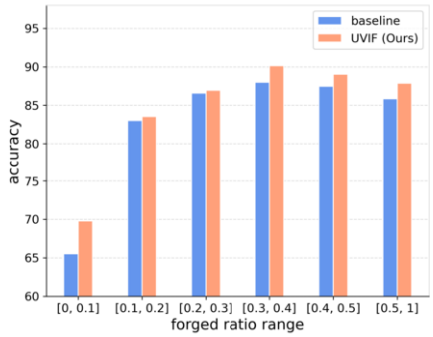

Figure 5. Accuracy achieved on videos with different forged ratios from the ForgeryNet [23] validation set. The baseline denotes ResNet-50 equipped with TSM [34].

Table 4. Comparing different backbones for the unified encoder. The baseline is a backbone with temporal fusion designs.

Backbone	baseline		UVIF (**Ours**)	
	Acc	AUC	Acc	AUC
ViT-S [12]	78.23	85.83	78.38	86.85
ConvNeXt-T [53]	81.56	88.43	84.94	93.35
ResNet-18 [22]	81.17	88.24	83.98	92.18
ResNet-50 [22]	80.89	88.66	85.32	93.45
ResNet-101 [22]	81.48	88.08	86.57	94.42

$100k$ images with annotations are easily available from open-access datasets, which brings a significant level of video forgery detection boost. Even with a small portion of added images, e.g., $20k$, the UVIF is able to achieve apparent performance improvements compared to the baseline (83.06% v.s. 80.89% accuracy). This indicates that the UVIF model gradually learns from the supervision information through training on annotated images, rather than relying on a large number of images to improve its representation. Meanwhile, the comparison of the green curve and the orange curve shows that 1) the pseudo labeling process is always effective using various numbers of training images, and 2) the performance improvement is larger when the added images are $100k$ and more. This shows that the model requires a certain number of images to train a reliable image classifier for pseudo label generation.

4.3.3 *Performance on Videos with Different Forged Ratios*

Figure 5 illustrates the accuracy achieved on videos with different forged ratios from the ForgeryNet [23] validation set. We utilize the annotations for temporal forgery localization [23] task to compute the ratio of forged frames of each partial forged video, and then divide them into six groups based on their forged ratios: [0, 0.1], [0.1, 0.2], [0.2, 0.3], [0.3, 0.4], [0.4, 0.5], and [0.5, 1]. The sample distribution of each group is 0.07, 0.19, 0.22, 0.22, 0.15, and 0.15, respectively. The results suggest that the UVIF can consistently enhance the baseline methods in all forged ratio groups. Notably, at the lowest forged ratio group [0, 0.1], although baseline accuracy is the lowest, the advantage of our method is significant (+4.29% accuracy). This shows how our approach leverages annotated images to learn distinctive representations that differentiate between real and fake video frames, therefore improving the accuracy of detecting partial forgeries in videos.

4.3.4 *Different Backbones*

We also conduct ablation experiments on the backbones for the unified encoder, as presented in Table 4. The backbones that we choose include ViT [12], ConvNeXt [53], and ResNet [22]. The results suggest that our proposed UVIF can consistently improve the forgery detection accuracy of different backbones, especially for CNN-based ones. As the capacity of the ResNet backbones increases, the accuracy of the baselines remains almost the same, while our methods can result in more accuracy improvements. This indicates that our UVIF effectively enhances the model's representation for face forgery detection. Besides, our method has a minor improvement for the ViT backbone. One possible reason is that the ViT down-samples the images greatly at the start of feature extraction, resulting in the model not learning discriminative feature information of facial forgery from the annotated images.

5 Conclusion

In this paper, we present UVIF, an end-to-end multi-task learning framework for video face forgery detection. Our method establishes a unified representation of facial videos and images by processing them together within a single model. By utilizing the fine-grained annotations from the image set, the UVIF framework can bring significant performance gains for detecting partial forgeries in videos. In the future, we will study to build extended multi-task learning frameworks for facial video and image data to expand their applicability to other forgery detection tasks beyond classification.

Acknowledgements

This work was supported by the Research Council of Finland (former Academy of Finland) ICT 2023 project TrustFace (grant 345948), Academy Professor project EmotionAI (grants 336116, 345122, 359854), the University of Oulu & Research Council of Finland Profi 7 (grant 352788), the Infotech Oulu, the Finnish Cultural Foundation for North Ostrobothnia Regional Fund under Grant 60231712, and in part by the Instrumentarium Science Foundation under Grant 240016. As well, the authors wish to acknowledge CSC-IT Center for Science, Finland, for computational resources.

References

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, pages 1–7. IEEE, 2018.

[2] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid. Vivit: A video vision transformer. In *Int. Conf. Comput. Vis. (ICCV)*, pages 6836–6846, 2021.

[3] G. Bertasius, H. Wang, and L. Torresani. Is space-time attention all you need for video understanding? In *Int. Conf. Mach. Learn. (ICML)*, volume 2, page 4, 2021.

[4] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang. End-to-end reconstruction-classification learning for face forgery detection. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 4113–4122, 2022.

[5] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition (PR)*, 77:329–353, 2018.

[6] B. Chen, J. Jiang, X. Wang, P. Wan, J. Wang, and M. Long. Debiased self-training for semi-supervised learning. *Adv. Neural Inform. Process. Syst. (NIPS)*, 35:32424–32437, 2022.

[7] U. A. Ciftci, I. Demir, and L. Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 2020.

[8] M. Contributors. MMEngine: Openmmlab foundational library for training deep learning models. 2022.

[9] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 5203–5212, 2020.

[10] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

[11] 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 *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 9468–9478, 2022.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *Int. Conf. Learn. Represent. (ICLR)*, 2021.

[13] H. Duan, Y. Zhao, Y. Xiong, W. Liu, and D. Lin. Omni-sourced webly-supervised learning for video recognition. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 670–688. Springer, 2020.

[14] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In *Int. Conf. Comput. Vis. (ICCV)*, pages 6202–6211, 2019.

[15] J. Feng and Z.-H. Zhou. Deep miml network. In *Proc. AAAI Conf. Artif. Intell. (AAAI)*, volume 31, 2017.

[16] R. Girdhar, M. Singh, N. Ravi, L. Van Der Maaten, A. Joulin, and I. Misra. Omnivore: A single model for many visual modalities. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 16102–16112, 2022.

[17] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. Imagebind: One embedding space to bind them all. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 15180–15190, 2023.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, 2020.

[19] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, F. Huang, and L. Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *ACM Int. Conf. Multimedia (ACMMM)*, pages 3473–3481, 2021.

[20] D. Güera and E. J. Delp. Deepfake video detection using recurrent neural networks. In *IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, pages 1–6. IEEE, 2018.

[21] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 5039–5049, 2021.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 770–778, 2016.

[23] Y. He, B. Gan, S. Chen, Y. Zhou, G. Yin, L. Song, L. Sheng, J. Shao, and Z. Liu. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 4360–4369, 2021.

[24] M. Ilse, J. Tomczak, and M. Welling. Attention-based deep multiple instance learning. In *Int. Conf. Mach. Learn. (ICML)*, pages 2127–2136. PMLR, 2018.

[25] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, and Y. Liu. Countering malicious deepfakes: Survey, battleground, and horizon. *Int. J. Comput. Vis. (IJCV)*, 130(7):1678–1734, 2022.

[26] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[27] S. A. Khan and H. Dai. Video transformer for deepfake detection with incremental learning. In *ACM Int. Conf. Multimedia (ACMMM)*, pages 1821–1828, 2021.

[28] B. Li, Y. Li, and K. W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 14318–14328, 2021.

[29] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 6458–6467, 2021.

[30] K. Li, Y. Wang, G. Peng, G. Song, Y. Liu, H. Li, and Y. Qiao. Uniformer: Unified transformer for efficient spatial-temporal representation learning. In *Int. Conf. Learn. Represent. (ICLR)*, 2022.

[31] 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 *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 5001–5010, 2020.

[32] X. Li, Y. Lang, Y. Chen, X. Mao, Y. He, S. Wang, H. Xue, and Q. Lu. Sharp multiple instance learning for deepfake video detection. In *ACM Int. Conf. Multimedia (ACMMM)*, pages 1864–1872, 2020.

[33] Y. Li, M.-C. Chang, and S. Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, pages 1–7, 2018.

[34] J. Lin, C. Gan, and S. Han. Tsm: Temporal shift module for efficient video understanding. In *Int. Conf. Comput. Vis. (ICCV)*, pages 7083–7093, 2019.

[35] K. Liu, W. Zhu, Y. Shen, S. Liu, N. Razavian, K. J. Geras, and C. Fernandez-Granda. Multiple instance learning via iterative self-paced supervised contrastive learning. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3355–3365, 2023.

[36] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu. Video swin transformer. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3202–3211, 2022.

[37] C. Miao, Z. Tan, Q. Chu, H. Liu, H. Hu, and N. Yu. F2trans: High-frequency fine-grained transformer for face forgery detection. *IEEE Trans. Inf. Forensics Secur. (TIFS)*, 18:1039–1051, 2023.

[38] J. Park, J. Lee, and K. Sohn. Dual-path adaptation from image to video transformers. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 2203–2213, 2023.

[39] A. Piergiovanni, W. Kuo, and A. Angelova. Rethinking video vits: Sparse video tubes for joint image and video learning. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 2214–2224, 2023.

[40] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 86–103, 2020.

[41] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 10684–10695, 2022.

[42] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Int. Conf. Comput. Vis. (ICCV)*, pages 1–11, 2019.

[43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis. (IJCV)*, 115(3): 211–252, 2015.

[44] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Adv. Neural Inform. Process. Syst. (NIPS)*, 34: 2136–2147, 2021.

[45] X. Shi, F. Xing, Y. Xie, Z. Zhang, L. Cui, and L. Yang. Loss-based attention for deep multiple instance learning. In *Proc. AAAI Conf. Artif. Intell. (AAAI)*, volume 34, pages 5742–5749, 2020.

[46] K. Shiohara and T. Yamasaki. Detecting deepfakes with self-blended images. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 18720–18729, 2022.

[47] L. N. Smith and N. Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artif. Intell. Mach. Learn. Multi-Domain Oper. Appl.*, volume 11006, pages 369–386. SPIE, 2019.

[48] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Adv. Neural Inform. Process. Syst. (NIPS)*, 30, 2017.

[49] H. Wang, Z. Liu, and S. Wang. Exploiting complementary dynamic incoherence for deepfake video detection. *IEEE Trans. Circuit Syst. Video Technol. (TCSVT)*, 2023.

[50] J. Wang, Y. Ge, R. Yan, Y. Ge, K. Q. Lin, S. Tsutsui, X. Lin, G. Cai, J. Wu, Y. Shan, et al. All in one: Exploring unified video-language pre-training. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 6598–6608, 2023.

[51] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 14549–14560, 2023.

[52] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu. Revisiting multiple instance neural networks. *Pattern Recognition (PR)*, 74:15–24, 2018.

[53] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 16133–16142, 2023.

[54] H. Zhang, Y. Meng, Y. Zhao, Y. Qiao, X. Yang, S. E. Coupland, and Y. Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 18802–18812, 2022.

[55] C. Zhao, C. Wang, G. Hu, H. Chen, C. Liu, and J. Tang. Istvt: Interpretable spatial-temporal video transformer for deepfake detection. *IEEE Trans. Inf. Forensics Secur. (TIFS)*, 18:1335–1348, 2023.

[56] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu. Multi-attentional deepfake detection. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 2185–2194, 2021.

[57] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen. Exploring temporal coherence for more general video face forgery detection. In *Int. Conf. Comput. Vis. (ICCV)*, pages 15044–15054, 2021.