Deep Convolutional Pooling Transformer for Deepfake Detection

TIANYI WANG, The University of Hong Kong
HARRY CHENG, Shandong University
KAM PUI CHOW, The University of Hong Kong
LIQIANG NIE, Harbin Institute of Technology (Shenzhen)

Recently, Deepfake has drawn considerable public attention due to security and privacy concerns in social media digital forensics. As the wildly spreading Deepfake videos on the Internet become more realistic, traditional detection techniques have failed in distinguishing between real and fake. Most existing deep learning methods mainly focus on local features and relations within the face image using convolutional neural networks as a backbone. However, local features and relations are insufficient for model training to learn enough general information for Deepfake detection. Therefore, the existing Deepfake detection methods have reached a bottleneck to further improve the detection performance. To address this issue, we propose a deep convolutional Transformer to incorporate the decisive image features both locally and globally. Specifically, we apply convolutional pooling and re-attention to enrich the extracted features and enhance efficacy. Moreover, we employ the barely discussed image keyframes in model training for performance improvement and visualize the feature quantity gap between the key and normal image frames caused by video compression. We finally illustrate the transferability with extensive experiments on several Deepfake benchmark datasets. The proposed solution consistently outperforms several state-of-the-art baselines on both within- and cross-dataset experiments.

CCS Concepts: • Security and privacy → Human and societal aspects of security and privacy; • Applied computing → Investigation techniques; System forensics;

Additional Key Words and Phrases: Deepfake detection, image keyframes, transformer

ACM Reference format:
Tianyi Wang, Harry Cheng, Kam Pui Chow, and Liqiang Nie. 2023. Deep Convolutional Pooling Transformer for Deepfake Detection. ACM Trans. Multimedia Comput. Commun. Appl. 19, 6, Article 179 (May 2023), 20 pages. https://doi.org/10.1145/3588574

1 INTRODUCTION

The fast-developing facial manipulation technique, Deepfake (examples are shown in Figure 1), has achieved remarkable success with various face synthesis algorithms in recent years [11, 25, 32, 38]. While the high-quality Deepfake videos can facilitate human lives in multiple domains such as
Fig. 1. Examples of Deepfake. The source identity is swapped onto the target, maintaining the facial expression and action of the target person.

Educational media and digital communications [58], they can be easily misused by criminals for malicious purposes and cause potential threats in societal, political, and business realms [28, 37, 52]. Recognized as the most severe artificial intelligence threat in 2020 [17], Deepfake has been vastly adopted to generate fake celebrity porn videos and fake politician speeches [18]. Famous actress Emma Watson and the well-known singer Ariana Grande are representative victims of Deepfake porn videos [24]. Moreover, the wildly spreading fake Barack Obama insulting Donald Trump video has reflected its severe social impact [5]. The number of Deepfake videos is growing at an unexpected rate such that it has almost doubled within only nine months in 2019 [51]. Considering the free access to the Deepfake technique with open-source implementations and packaged Deepfake applications circulating on the Internet, anyone can become a victim of the Deepfake technique. Consequently, it is unrealistic to distinguish real and fake videos manually due to the vast amount of existing Deepfake videos and their increasing authenticity. Therefore, protecting human lives against Deepfake attacks with the help of automatic Deepfake detection solutions is highly desired to preserve the social order.

Deepfake detection is normally framed as a binary classification task by the existing approaches. While the classic forgery detection methods [42, 57] are insufficient to identify hyper-realistic Deepfake materials, deep learning models are frequently utilized. Specifically, the image frames are extracted from the candidate video and fed into the designed deep neural networks to determine the authenticity. Early approaches mainly focus on the local features in the input image to look for abnormal feature patches and resort to the pre-trained convolutional neural network (CNN) backbones followed by a fully connected layer to conduct the classification. In particular, the convolutional kernel studies information within each receptive field while ignoring the correlations between positions that are far away. Meanwhile, the goal of Deepfake detection has shifted from within-dataset to cross-dataset test on unseen Deepfake videos for generalization purposes. As the CNNs have shown poor generalization ability in feature learning, recent algorithms have
gradually introduced the idea of the attention mechanism using CNNs to enlarge the local image feature areas for both within- and cross-dataset performance enhancement. However, recent synthetic work has gradually fulfilled the research gap and local artifacts are now difficult to be discovered. As finding consistency and relation between any two positions within an image can be meaningful such that the synthetic eyes and mouth may not look consistent with respect to the facial expressions, those current solutions are still limited in learning global features for further improving the Deepfake detection ability. On the other hand, the low resolution of the randomly extracted image frames has limited the potential ability of the detection models as they naturally contain restricted image features rather than high-resolution ones.

To handle the aforementioned issues, we present a keyframe-based deep convolutional Transformer model, which jointly leverages convolutional pooling and re-attention approaches to study the decisive image features and relations both locally and globally. In detail, a stack of CNNs is adopted to extract local features from the input face images, and an enhanced deep Transformer with convolutional pooling and re-attention follows to enrich the global feature learning and to analyze the corresponding relations between image feature patches. Specifically, the convolutional pooling refines dominant features with dimension reduction, and the re-attention technique maintains the diversity within attention maps in a deep model. Moreover, since videos are commonly compressed for storage saving, unavoidable information loss occurs when the normal image frames are reconstructed and randomly extracted from the videos. On the contrary, the keyframes within a video are the only image frames that carry complete frame information with high resolution and do not suffer from information loss during the image frame reconstruction process. To the best of our knowledge, the importance of the keyframes has been untapped in the existing Deepfake detection approaches. In order to bridge this gap, in this study, we extract all keyframes from the videos for both training and testing processes, and thus emphasize the power of keyframes in boosting Deepfake detection performance. We conduct extensive experiments on several commonly-used Deepfake datasets. The experimental results prove that our approach outperforms several baselines in both within- and cross-dataset evaluations. We further visualize the gap in the extracted features between normal and key image frames to verify the robustness of the latter.

The contributions of this work are threefold:

— We introduce the scheme of learning image features both locally and globally using the proposed deep convolutional Transformer model via jointly integrating convolutional pooling and re-attention strategies, which well explores the global features within a face image yet ignored by the existing methods.

— Significantly distinguished from the existing Deepfake detection solutions that seldomly consider the information loss in image extraction from the compressed videos, we are the first on extracting the keyframes and exhibiting their importance both statistically and visually in experiments, to the best of our knowledge.

— Besides the considerable achievement of within-dataset performance, we verify the robustness of the proposed model in extensive cross-dataset experiments, which outperforms the existing state-of-the-art baselines on several popular Deepfake datasets.

2 RELATED WORK

2.1 Deepfake Generation

The term Deepfake is firstly introduced by the Reddit user ‘deepfakes (DF)’ when the open-source implementation was released simultaneously in 2017. The leading architecture in the existing Deepfake generation work is an autoencoder for facial identity swap. Specifically, a shared encoder learns the identity-independent features from the input faces of different
identities, and two individual decoders each generates synthesized faces with the desired identity. When performing face-swapping, a target face is passed into the autoencoder. The shared encoder performs facial feature extraction and converts the input target face to a context vector. Thenceforth, the corresponding decoder analyzes the context vector and generates a look maintaining the facial expression of the target face while having the identity of the desired source face.

In light of the promising synthetic performance of Generative Adversarial Networks (GAN) [16] in domains such as face recognition [55, 63, 65], face synthesis [66], and human parsing [64], recent Deepfake research has exploited GAN for better synthesis authenticity. For example, FaceShifter [32] and SimSwap [6] both propose specific modules to preserve facial attributes that are hard to handle and maintain the fidelity for arbitrary identities. HifiFace [56] and MegaFS [68] achieve high-resolution identity swap on 512 and 1,024 resolutions for arbitrary identities, respectively.

2.2 Deepfake Detection
Since the threat of Deepfake firstly attracted public attention, most existing Deepfake detection methods have utilized CNN-based architectures as the main backbones along with other fine-tuning steps. An early approach [2] proposes optical flow analysis with pre-trained VGG16 [45] and ResNet50 [19] CNN backbones and achieves preliminary within-dataset test performance on the FaceForensics++ (FF++) dataset [43]. Bonettini et al., [4] studied the combination of different pre-trained CNN models starting from EfficientNetB4 [47] as the backbone for Deepfake detection. The DFT-MF [22] approach focuses on open mouths with teeth and uses the standard CNN model to detect Deepfake by isolating, analyzing, and verifying lip and mouth movements. Later, Rossler et al. [43] employed the pre-trained well-designed Xception [7] network and achieved the state-of-the-art detection performance over all other existing approaches at the time on the FF++ dataset. However, since the CNN architecture lacks generalization ability and mainly focuses on local features, even Xception is restricted in learning the global features for further performance improvements. Therefore, the most recent solutions have introduced the attention mechanism using a convolutional layer with the kernel size of one and claim to enlarge local feature areas and the corresponding relations. The SRM [36] network constructs two streams of Xception backbones that focus on RGB frames and high-frequency frames, respectively, and studies the cross-modal relations between the two streams. The MAT model [62] adopts EfficientNetB4 as the backbone to detect Deepfake and introduces the convolutional attention layer to study features within different local parts. Recently, unlike previous work that directly utilizes cross-entropy loss for real and fake classification, Sun et al. [46] proposed Dual Contrastive Learning (DCL) to study positive and negative paired data for Deepfake detection with better transferability. It is also worth noting that a Transformer based approach, CViT [59], has been proposed with source code published and has shown promising results, and several latest models [8, 9, 26] are further designed following similar ideas but devoting better detection performance.

2.3 Vision Transformers
The success of Transformer [53] in Natural Language Processing (NLP) in 2017 has attracted researchers to explore its potential ability in the computer vision domain. Since the first vision Transformer (ViT) [13] achieves relatively reasonable performance on downstream vision tasks, various innovative ViT architectures have been attempted. An early representative novel Swin Transformer [35] devises hierarchical architecture with shifted windows to make it compatible with a broad range of vision tasks. Later, a Contextual Transformer [34] is proposed to learn contextual information among input keys that have been ignored by the classic self-attention. Yao
et al. [61] conducted Wavelet Vision Transformer (Wave-ViT) following the wavelet theory, which specifically solves the inevitable information dropping caused by over-aggressive down-sampling when dealing with high computational costs. Recently, a Dual Vision Transformer (Dual-ViT) [60] further enhances the computational efficiency of self-attention with the state-of-the-art vision task performance by incorporating a critical semantic pathway that compresses token vectors into global semantics. Downstream tasks that satisfactory performance has been achieved by Transformer architectures include image recognition, object detection, and semantic segmentation. However, the promising idea of Transformer and self-attention has been barely discussed and attempted in the Deepfake detection domain. Therefore, in this study, we fulfill the research gap by presenting a deep convolutional Transformer model.

3 METHODOLOGY

3.1 Framework

In this section, we briefly introduce the framework and workflow of the proposed approach. As shown in Figure 2, the model mainly consists of the following parts: convolutional feature extraction, depth-wise separable convolutional projection, and pooling Transformer. To be more specific, the face images are fed to the convolutional feature extraction module after being extracted and cropped from the candidate Deepfake image frames. The convolutional feature extraction module studies the local features within each face image and acquires the representative information. We then perform the depth-wise separable convolution upon the learned local feature information and transmit the results to the pooling Transformer. The pooling Transformer analyzes the global image features, and the attention map diversity is preserved by the re-attention mechanism. Particularly, the re-attention mechanism is applied to each attention map within the pooling Transformer module by unique learnable transformation matrices for each layer. Thereafter, the ultimately learned feature information is reshaped with multilayer perceptron and performed Deepfake detection.

In the following parts, we first introduce the idea of utilizing image keyframes within a commonly seen compressed video. We then illustrate the convolutional local feature extraction module upon the input images. After that, we demonstrate the Transformer part for global features and relation analysis. In particular, we present the idea of convolutional pooling and re-attention within an enhanced deep Transformer.
3.2 Keyframe Extraction

In real-life scenarios, video compression is usually operated to maintain as few complete image frames as possible for storage saving. The commonly used compression formats include H.264, MPEG-4, and JPEG [40]. In general, three types of image frames, I-frame, P-frame, and B-frame, are derived within a video under H.264 compression. A 10-frame H.264 compressed video clip example showing the dependencies among the frames is exhibited in Figure 3. I-frame, also known as keyframe or intra-frame, is the only type of image frame that carries complete image information with the largest sizes and resolutions within a compressed video. The keyframe is generally at the beginning or end of an action [48]. In between two keyframes are the P-frames and B-frames. P-frame, namely the predicted picture, holds only the variations between the current image frame and the closest preceding I-frame or P-frame in a video, occupying much less disk space than a keyframe. B-frame, the bidirectional predicted picture, is encoded from an interpolation of the closest succeeding and preceding image frames, either I-frame or P-frame, requiring the least amount of disk space [3, 54]. Other compression formats such as MPEG-4 and JPEG follow the same rule as H.264 except for the absence of B-frame. In Deepfake detection, the image frame extraction process suffers information loss of the P-frames and B-frames during image frame reconstruction in video decoding. Considering that the hyper-realistic Deepfake videos contain few fake cues that can be easily discovered, it is significant to collect as many keyframes from the video when detecting a fake face.

3.3 Convolutional Local Feature Extraction

We extract the local features from the input keyframe face images using a stack of CNNs. Specifically, the convolutional local feature extraction module contains 17 convolutional layers, each with a kernel size of three, and the stride and padding values are set to one. In particular, each convolutional layer is followed by batch normalization and a GELU activation function for feature normalization and non-linearity, respectively. The stack of CNNs is distributed into five, each followed by a max-pooling layer with a kernel size of two and stride value of two to concentrate on the dominant features in the input image. Except for the first group of CNNs that raises the number of input image channels from 3 to 32, the first convolutional layer of each group doubles the number of input channels. In addition, each max-pooling layer reduces the intermediate image feature dimensions by half. As a result, the convolutional local feature extraction module extracts a local feature with dimension $512 \times 7 \times 7$ by inputting an image with dimension $3 \times 224 \times 224$. The extracted local features are then fed into the Transformer part for global feature learning and relation analysis.
3.4 Deep Transformer with Convolutional Pooling and Re-attention

The Transformer architecture [53] is firstly introduced with the innovative multi-head self-attention mechanism, which learns global features and relations with low costs. In this section, inspired by the scheme of pooling ViT [21], we boost its learning ability with depth-wise separable convolutional projection and deep re-attention to analyze the global features and relations within the input image.

3.4.1 Convolutional Pooling Transformer. The success of the Transformer architecture in the NLP domain is attributed to the autoencoder design and the multi-head self-attention mechanism. While the auto-encoder structure plays the role to fit the most classic NLP tasks, multi-head self-attention is the one that learns global features at a low cost. In specific, a Transformer includes various Transformer blocks, each containing a multi-head self-attention section followed by residual normalization and feedforward layers. In each Transformer block, three learnable linear projections $W_Q, W_K, \text{and } W_V$ are performed on the input $X$ to generate Query ($Q$), Key ($K$), and Value ($V$) matrices by

$$Q, K, V = XW_Q, XW_K, XW_V.$$  \hspace{1cm} (1)

Then, the $Q$, $K$, and $V$ matrices are arranged for self-attention computation by

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right) V,$$  \hspace{1cm} (2)

where $d_K$ is the dimension of each entry vector in $K$.

In the existing efforts, a convolutional implementation of the attention mechanism with a kernel size of one fails to study the relation between any two feature patches within an entire feature map. Meanwhile, the functionality is similar to a fully connected layer except for the modification of the number of channels. On the other hand, a CNN with a kernel size that is the same as the feature map dimension considers all feature patches and their relations by the convolutional computation. However, such an operation downscales the feature map to the dimension of one, which significantly discards determinant features and deteriorates the model performance. In our proposed approach, we apply a pooling Transformer to fit feature dimensions for image analysis. The self-attention mechanism makes the distance between any two locations in the feature map one and achieves global feature relation learning at a low cost. A multi-head self-attention is an ensemble of self-attentions, each with a unique group of learnable linear projections for $Q$, $K$, and $V$ computation, offering multiple perspectives of global feature learning. In particular, for a squared input image fed to the convolutional local feature extraction module, a feature map maintaining the original relative position information with dimension $w \times w \times c$ is derived, where $w$ is the width of the extracted feature map and $c$ is the number of channels. The feature map can be regarded as $w^2$ feature patches in $c$ different channels of perspectives. The $Q$, $K$, and $V$ matrices are then generated and flattened to the shape of $w^2 \times c$. We append the special learnable CLS token [53] for classification purposes and split the channels regarding $h$ attention heads, which devotes the result with dimension $(w^2 + 1) \times he$, where $c = he$ and each of the $w^2 + 1$ rows represents the embedded feature information of a feature patch. In the self-attention calculation for each attention head, the multiplication of $Q$ and $K$ projected from the input $X$ devotes the following matrix,

$$QK = \begin{bmatrix} q_1K, q_2K, \ldots, q_{w^2+1}K \end{bmatrix},$$  \hspace{1cm} (3)

where $K = [k_1, k_2, \ldots, k_{w^2+1}]$. The entries on the $QK$ matrix diagonal represent the self-relations for each feature patch, and the rest entries represent the relative relations between any two different feature patches. Omitting the constant computations and the softmax function, a
Fig. 4. Framework of the Transformer with convolutional pooling. The three phases are partitioned by two convolutional pooling layers, resulting in doubling the number of channels and reducing half of the feature map dimensions.

multiplication with $V$ further enhances the relation learning within the input image and derives the following matrix with dimension $(w^2 + 1) \times e$,

$$QKV = \begin{bmatrix} q_1KV, & q_2KV, & \ldots, & q_{(w^2+1)KV} \end{bmatrix}, \quad (4)$$

where $V = [v_1, v_2, \ldots, v_e]$ and each entry of $V$ is the entire feature map at one of the $e$ feature channels. As a result, relations between any two feature patch positions within an input image are studied and applied to assist Deepfake detection.

Moreover, since the self-attention computation breaks the relative positional layout of the feature patches, a positional embedding is introduced to maintain meaningful relations between feature patches, and we apply sinusoidal position encoding to each feature patch after convolutional local feature extraction by

$$PE(p, 2i) = \sin \left( \frac{p}{\lambda \frac{d_{model}}{2i}} \right), \quad (5)$$

and

$$PE(p, 2i + 1) = \cos \left( \frac{p}{\lambda \frac{d_{model}}{2i + 1}} \right), \quad (6)$$

where $\lambda$ is a constant controlling the sinusoidal cycle speed, $p$ is the position of the feature patch, $d_{model}$ is the feature patch dimension, and $i$ is the real-time running dimension in range $(0, d_{model})$.

The role of pooling is to adjust the receptive field, and the interaction within feature patches increases as the number of feature patches decreases after the pooling operation. In detail, the Transformer blocks are phased into three, with a convolutional pooling between every two consecutive phases. The convolutional pooling (as shown in Figure 4) is implemented as a convolutional layer with non-linearity and takes charge of refining dominant features. In particular, each convolutional pooling layer doubles the number of feature channels and reduces the feature dimension by half.
Considering the advantages of CNNs over linear projections on image analysis, we further introduce the depth-wise separable convolutions into the generation process of $Q$, $K$, and $V$ matrices to enrich image feature learning for multi-head self-attention computation. In detail, we replace the three fully connected layers for the generation of $Q$, $K$, and $V$, each with a depth-wise separable convolution followed by batch normalization when projecting the input features in each Transformer block. To maintain stable model architecture, the separable convolution is set to only perform projections while maintaining the feature dimensions as being extracted from the local feature extraction module. The workflow of the separable convolutional projection is shown in Figure 5. The depth-wise separable convolution is to avoid the possible overfitting and time-consuming problem due to the large number of parameters in a standard CNN \cite{7}. Specifically, the depth-wise separable convolution divides a standard CNN into a depth-wise convolution and a point-wise convolution. The former is a spatial convolution that only modifies the feature map dimension with the channels unchanged, while the latter only changes the number of channels. By adopting the depth-wise separable convolution for matrix projections, the proposed model can analyze richer image features.

3.4.2 Deep Re-attention. A deep model with a mass of Transformer blocks is necessary for complex image tasks. However, a deep ViT model empirically suffers from the attention collapse issue, such that the attention maps tend to be overly similar, and obvious performance decline is observed as the model goes deeper \cite{67}. In other words, the multi-head self-attention mechanism begins to
taper off in feature learning efficacy as the model further proceeds. Therefore, in this study, we construct a deep pooling Transformer with a total of 24 Transformer blocks in three phases and employ the re-attention technique [67] with a learnable transformation matrix \( \Theta \) to retain attention map diversity. Specifically, observing that the attention maps from different attention heads are not affected by the attention collapse, the re-attention technique resorts to cross-head communication by

\[
\text{Re-Attention}(Q, K, V) = \Theta^T \text{softmax} \left( \frac{QK^T}{\sqrt{d_K}} \right) V,
\]

the multiplication of the learnable parameter \( \Theta \) and the self-attention map along the head dimension. The attention map diversity is improved with the unique learnable transformation matrix \( \Theta \) for the multi-head self-attention within each Transformer block. Consequently, richer features and relations can be learned and boost the Deepfake detection performance.

3.4.3 Binary Classification. Following the convention for binary classification, we take the learnable CLS token from the output after the last Transformer block and pass it through a fully connected layer followed by the softmax layer for Deepfake detection prediction on the input face image frame. The model is tuned with the cross entropy loss function by

\[
L_{CE} = -\sum_{i=1}^{2} t_i \log p_i,
\]

where \( t_i \) is the groundtruth value and \( p_i \) is the softmax prediction for class \( i \).

4 EXPERIMENTS
In this section, we first detail the implementation of the model training process. Then, we describe the selected Deepfake datasets for training and testing. Thereafter, we conduct experiments to evaluate the model performance on the FF++ dataset and several other Deepfake datasets. Finally, we propose several ablation studies to evaluate the importance of keyframes, the performance of each model component, and the effectiveness of different model depths.

4.1 Implementation Details
We used the DLIB library [29] to detect faces for all image frames and saved the aligned facial images with a size of 224 × 224. The depth of the pooling Transformer is set to 24 of three phases with 8, 8, and 8 Transformer blocks and 4, 8, and 16 attention heads, respectively. The sinusoidal cycle constant \( \lambda \) is set to 10,000. Adam optimizer with a learning rate of \( 1 \times 10^{-4} \) and weight decay of \( 1 \times 10^{-4} \) is used for optimization. The models are trained on the Tesla V100 GPU with batch size 32. All models are trained and tested at frame level upon the extracted facial images.

4.2 Datasets
FF++ [43] is currently the most widely adopted dataset in the existing Deepfake detection studies. It includes 1,000 original pristine videos collected from YouTube and four subsets each containing 1,000 fake videos synthesized from the 1,000 original real ones using a unique facial manipulation technique. The four facial manipulations are known as FaceSwap (FS), DF, Face2Face (F2F) [50], and NeuralTextures (NT) [49]. In general, the FF++ dataset provides an official split with a ratio of 720:140:140 for training, validation, and testing videos. Three qualities have been released, namely, Raw, HQ (c23), and LQ (c40), where the latter two are compressed using the H.264 codec with different compression levels. We adopted the HQ (c23) version dataset, which is similar to the real-world Deepfake, and followed the official split ratio to execute image frame extraction and face extraction using the DLIB library.
As discussed in Section 3.2 and following the early approach for keyframe extraction [27], we first derived all keyframe faces from the videos using FFmpeg [15] and augmented the real keyframe faces for a balanced keyframe dataset. Due to the limited amount, it is insufficient to train the model to converge with keyframes only. Therefore, we further randomly extracted sufficient normal frames from each video. Specifically, to maintain a balanced dataset between the real and fake faces, we extracted three times more normal frame faces from each real video than the fake one. In this study, the proposed model is trained on the FF++ training set and evaluated through the within-dataset test on the FF++ testing set with a mixture of keyframe and normal frame faces. Meanwhile, we collected various popular Deepfake benchmark datasets for extensive experiments on cross-dataset evaluation, namely, Deepfake Detection Challenge (DFDC) [12], Celeb-DF [33], and DeeperForensics-1.0 (DF-1.0) [23]. A summary of the attending datasets is listed in Table 1.

DFDC [12] is one of the largest public Deepfake datasets [20, 31, 69] by far with 128,154 videos, which includes synthesized videos using eight different facial manipulation methods as listed in Table 1. However, its large quantity has caused storage issues in data preprocessing such that the existing approaches have been seldomly attempted on the whole dataset. In this study, we acquired the entire DFDC dataset and randomly shuffled 10% of the whole dataset for cross-dataset evaluation.

Celeb-DF [33] is one of the most challenging Deepfake detection datasets because of the high quality and high difficulty of its official testing set with 518 videos. It has failed most of the existing baselines at a time, even in the within-dataset performance on the testing set. We considered the Celeb-DF official testing dataset with 518 videos in the extensive experiments and evaluated the robustness of our approach against the existing baseline models. It is worth noting that the official testing set is imbalanced with a ratio of 178:340 for real and fake.

DF-1.0 [23] is the first large dataset that has been manually added deliberate distortions and perturbations to the clean Deepfake videos. The additional noise aims to imitate real-world scenarios and has brought new challenges to the Deepfake detection task. The dataset contains a total of 60,000 videos with a ratio of 5:1 for real and fake. The fake videos are generated based on the real source videos, and seven types and five levels of distortions and perturbations are added accordingly. We followed the official split scheme and performed Deepfake detection on the 1,000 testing set videos with mixed levels of random distortions and perturbations added. To keep a balanced testing dataset, we randomly sampled the same amount of real videos in the experiment.

Since the above three datasets are rather recently published with high-quality videos, we successfully extracted sufficient numbers of keyframes to establish each testing set for cross-evaluation. Unless an imbalanced official testing set is given (Celeb-DF), we constructed balanced testing sets for all experiments. Besides, all comparative experiments are conducted on the same datasets utilized by our proposed model for complete fairness.

---

1Random augmentations include random rotation, transposition, horizontal flipping, vertical flipping, and so on.
Table 2. Quantitative Frame-level ACC and AUC Performance Comparison on Each Testing Set after Training on the FF++ Dataset

| Method       | Test Dataset | FF++ | DFDC | Celeb-DF | DF-1.0 |
|--------------|--------------|------|------|----------|--------|
|              | ACC   | AUC  | ACC  | AUC  | ACC  | AUC  | ACC  | AUC  | ACC  | AUC  |
| MesoNet [1]  | 61.03% | 58.13% | 50.02% | 50.16% | 36.73% | 50.01% | 50.05% | 50.21% |
| Capsule [39] | 76.40% | 83.44% | 51.30% | 56.16% | 61.96% | 59.93% | 59.29% | 61.46% |
| FFD [10]     | 82.29% | 82.48% | 59.44% | 59.47% | 46.19% | 55.86% | 53.69% | 53.81% |
| CViT [59]    | 83.05% | 91.08% | 60.76% | 67.43% | 53.26% | 63.60% | 54.97% | 58.52% |
| MAT [62]     | 87.50% | 94.85% | 63.16% | 69.56% | 44.78% | 57.20% | 56.90% | 61.72% |
| SRM [36]     | 88.17% | 94.93% | 59.93% | 64.80% | 52.95% | 60.90% | 55.83% | 62.54% |
| Xception [7] | 90.08% | 96.51% | 58.77% | 66.95% | 54.24% | 65.86% | 54.76% | 67.03% |
| Our Approach | 92.11% | 97.66% | 65.76% | 73.68% | 63.27% | 72.43% | 62.46% | 78.19% |

The best performance is marked as bold.

4.3 Performance Evaluation

We first evaluated our approach on the FF++ testing set after training and then examined the model transferability on DFDC, Celeb-DF, and DF-1.0 testing sets. We also performed comparative tests against the state-of-the-art Deepfake detection approaches to justify the performance of our model. Both the proposed model and the comparative models are trained on the FF++ dataset.

For fairness, we only adopted the state-of-the-art baselines with publicly available source code and trained and tested each comparative method following the corresponding default optimal parameter settings on the same datasets we used for our model evaluation. After training, our proposed model achieves a 92.11% accuracy (ACC) and a 97.66% area under the receiver operating characteristic (ROC) curve (AUC) score on the FF++ testing set. As Table 2 columns two and three shown, although some comparative approaches have derived competitive performance, our proposed model has outperformed all baseline models in the comparative experiment on FF++.

We further conducted extensive experiments on cross-dataset tests for DFDC, Celeb-DF, and DF-1.0. As Table 2 columns four to nine shown, our proposed Deepfake detection approach outperforms all state-of-the-art baselines in the comparative tests for the cross-dataset experiment. It can be concluded that the CNN-based approaches such as MesoNet [1], Capsule [39], and FFD [10] mainly focus on the local features within the face images, which lack global information for further enhancement and thus have poor transferability when cross-evaluated on DFDC, Celeb-DF, and DF-1.0. It can also be observed that the Transformer based CViT [59] approach directly adopts the ViT [13] where the utilized vanilla Transformer is unstable on image analysis tasks. Besides, the recent state-of-the-art models MAT [62] and SRM [36] have shown relatively competitive abilities for both within-dataset and cross-dataset evaluations compared to the early approaches. Nevertheless, the Xception [7] method which utilized the XceptionNet CNN backbone has derived surprisingly promising performance with even higher accuracies and AUC scores than the most recent state-of-the-art models. This might be due to the robustness of the perfect XceptionNet architecture design.

As a result, although the manipulation techniques in the testing sets are unseen during the training process for the cross-dataset evaluation, our proposed method is the only one that achieves over 60% accuracy and 70% AUC scores for all of DFDC, Celeb-DF, and DF-1.0 comparing to other approaches. According to the AUC metric, we observed that the high-quality Celeb-DF dataset was the most challenging one for most approaches. It is worth noting that our model also performs stably on the DF-1.0 dataset with random levels of distortions and perturbations, which means the model is robust against additional artificial noise.
4.4 Keyframe Analyses

In this study, we emphasized the significant role that the keyframes play in video compression, and we claimed that our model studies richer features and relations from the keyframes than from the normal frames. In particular, we conducted ablation studies setting the variation in the attendance of keyframe faces during the training and testing process. Firstly, having the model architecture fixed, we enforced two training sessions with one adopting normal image frames and keyframes, and the other merely normal image frames. Models trained in the two sessions are evaluated on the same group of testing sets as introduced in section 4.3. As shown in Table 3, the model session trained with keyframes performs remarkably better than the one trained with only normal image frames. Specifically, the attendance of keyframes has boosted the model performance by 6.22%, 4.96%, 8.15%, and 10.19% when tested on FF++, DFDC, Celeb-DF, and DF-1.0, respectively. The results have proved the superiority of keyframes over normal frames in feature richness.

Moreover, we employed the gradient-based visualization approach, the Grad-CAM [44], for feature heatmap visualization upon our proposed model. We visualized the facial features extracted by the proposed model from randomly selected pairs of keyframes and normal frames. We then displayed four pairs of keyframes and normal frames for each facial manipulation technique in the FF++ dataset along with the heatmaps of their extracted features (Figure 6), where each pair includes a keyframe and a normal frame face derived from the same video. Specifically, every two rows exhibit pairs of facial images and feature heatmaps, and every two columns illustrate the pairs of keyframes and normal frames for each manipulation technique. For a feature heatmap, the hotter (colored with deep red) an area is, the more features are extracted by the model. It can be easily observed that although faces from the same video contain ‘hot’ features at similar locations, the keyframe faces are generally ‘hotter’ with larger red area than the normal frames. This means that the keyframe faces have relatively more decisive features extracted by the proposed model, which helps derive better detection performance. Consequently, the displayed features from any pair of faces are sufficient to conclude the advantage of keyframes over normal frames in Deepfake detection.

To imitate the possible real-life cases, we further probed the importance of keyframes by training and testing with different combinations of keyframes and normal frames. In detail, we trained the proposed model with four different settings, namely, keyframes (K), keyframes with data augmentation (K_{aug}), normal frames (N), and the combination of key and normal frame faces (K + N). The results are presented in Table 3(b), showing that the model trained with keyframes and data augmentation performs best in terms of ACC and AUC.

### Table 3. Experimental Results and Statistical Information Regarding the Keyframe Settings

| Test Data | Trained on FF++ | |
|-----------|----------------|-----|
|           | w/o keyframes  | w/keyframes |
| FF++      | ACC 83.74%     | 92.11% |
|           | AUC 91.44%     | 97.66% |
| DFDC      | ACC 61.41%     | 65.76% |
|           | AUC 68.72%     | 73.68% |
| Celeb-DF  | ACC 52.72%     | 63.27% |
|           | AUC 64.28%     | 72.43% |
| DF-1.0    | ACC 55.94%     | 62.46% |
|           | AUC 68.00%     | 78.19% |

4.4 Keyframe Analyses

In this study, we emphasized the significant role that the keyframes play in video compression, and we claimed that our model studies richer features and relations from the keyframes than from the normal frames. In particular, we conducted ablation studies setting the variation in the attendance of keyframe faces during the training and testing process. Firstly, having the model architecture fixed, we enforced two training sessions with one adopting normal image frames and keyframes, and the other merely normal image frames. Models trained in the two sessions are evaluated on the same group of testing sets as introduced in section 4.3. As shown in Table 3, the model session trained with keyframes performs remarkably better than the one trained with only normal image frames. Specifically, the attendance of keyframes has boosted the model performance by 6.22%, 4.96%, 8.15%, and 10.19% when tested on FF++, DFDC, Celeb-DF, and DF-1.0, respectively. The results have proved the superiority of keyframes over normal frames in feature richness.

Moreover, we employed the gradient-based visualization approach, the Grad-CAM [44], for feature heatmap visualization upon our proposed model. We visualized the facial features extracted by the proposed model from randomly selected pairs of keyframes and normal frames. We then displayed four pairs of keyframes and normal frames for each facial manipulation technique in the FF++ dataset along with the heatmaps of their extracted features (Figure 6), where each pair includes a keyframe and a normal frame face derived from the same video. Specifically, every two rows exhibit pairs of facial images and feature heatmaps, and every two columns illustrate the pairs of keyframes and normal frames for each manipulation technique. For a feature heatmap, the hotter (colored with deep red) an area is, the more features are extracted by the model. It can be easily observed that although faces from the same video contain ‘hot’ features at similar locations, the keyframe faces are generally ‘hotter’ with larger red area than the normal frames. This means that the keyframe faces have relatively more decisive features extracted by the proposed model, which helps derive better detection performance. Consequently, the displayed features from any pair of faces are sufficient to conclude the advantage of keyframes over normal frames in Deepfake detection.

To imitate the possible real-life cases, we further probed the importance of keyframes by training and testing with different combinations of keyframes and normal frames. In detail, we trained the proposed model with four different settings, namely, keyframes (K), keyframes with data augmentation (K_{aug}), normal frames (N), and the combination of key and normal frame faces (K + N). The results are presented in Table 3(b), showing that the model trained with keyframes and data augmentation performs best in terms of ACC and AUC.
The reason for having the keyframes with data augmentation setting is the lack of sufficient FF++ keyframe data for training the model. Statistics of the four training data settings are listed in Table 3, and we kept the datasets always roughly balanced for real and fake. Models trained with the above four settings are each evaluated on FF++ with the settings keyframes (K), normal frames (N), and the combination of key and normal frames (K + N). The results are exhibited in Table 4.

It can be observed that training with sole keyframes derives unsatisfactory performance on all testing sets due to the issue of insufficient amount of training data. On the contrary, the model trained with an augmented keyframe dataset achieves significantly better performance, and the highest accuracy and AUC score are derived on the keyframe testing set as expected. As for the model trained with solely normal frames, although evaluating on the normal frame testing set outperforms that on the keyframe testing set, the overall performance is relatively weak due to the poor feature richness issue. Expectedly, training with the combination of key and normal frames has devoted the best performance on all three testing sets.

In real life, it is unpredictable and uncontrollable whether a candidate Deepfake image is a key or normal frame. With the performance reported in Table 4, it can be concluded that our proposed model after training with the combination of key and normal frames can achieve satisfactory detection performance on both key and normal frame Deepfake candidate faces. To be more specific, training with the combination of key and normal frames maintains robust model
Table 4. Performance Evaluation using Different Training and Testing Settings with the FF++ Dataset

| Tested on FF++ | Trained on FF++ | K | Kaug | N | K + N |
|---------------|----------------|----|------|---|------|
|               |                | ACC | AUC | ACC | AUC | ACC | AUC | ACC | AUC |
| K             | 82.03%         | 90.20% | 89.41% | 96.11% | 82.48% | 90.34% | 92.09% | 97.64% |
| N             | 72.66%         | 82.01% | 83.54% | 93.87% | 84.18% | 91.83% | 92.12% | 97.69% |
| K + N         | 75.09%         | 84.23% | 85.06% | 94.35% | 83.74% | 91.44% | 92.11% | 97.66% |

K: keyframe faces solely; Kaug: keyframe faces with data augmentation; N: normal frame faces solely; K + N: the combination of key and normal frame faces.

Table 5. Ablation Study on Module Cumulation Performance

| Method                  | Test Dataset | FF++ |        |        |        |        |        |
|-------------------------|--------------|------|--------|--------|--------|--------|--------|
|                         |              | ACC  | AUC    | ACC    | AUC    | ACC    | AUC    |
| Vanilla Transformer     |              | 83.21% | 91.60% | 62.62% | 66.90% | 52.71% | 62.18% |
| + Pooling               |              | 87.38% | 94.63% | 62.73% | 69.36% | 54.28% | 63.90% |
| + Conv. Projection      |              | 91.45% | 97.36% | 62.92% | 70.34% | 60.29% | 68.26% |
| + Re-attention          |              | 92.11% | 97.66% | 65.76% | 73.68% | 63.27% | 72.43% |

The components pooling, convolutional (conv.) projection, and re-attention are cumulatively added starting from a vanilla Transformer.

detection performance, in other words, the optimal performance in this study, regardless of the types of testing image frames.

4.5 Model Component Analysis

To evaluate the contribution of each module, we conducted an ablation study on different components of the proposed model. In specific, having the local feature extraction module fixed, every other module is cumulatively added starting from a vanilla Transformer. As exhibited in Table 5, pooling, convolutional projection, and re-attention modules are added step by step, and each module consistently boosts the detection performance as expected.

It can be observed that the pooling module and the convolutional projection idea improve the performance evenly with around 3% with respect to the AUC score when tested on FF++. While for the cross-dataset evaluation, the pooling module is more helpful on DFDC than on Celeb-DF, and the convolutional projection is more effective on Celeb-DF than on DFDC. Meanwhile, the two modules both boost the performance on DF-1.0 by around 4% of AUC scores. Besides, re-attention is more instrumental in cross-dataset performance by adding diversity to the attention maps. In specific, it only gains less than 1% AUC score improvement on FF++, which is relatively trivial compared to the performance of cross-dataset evaluation. It is also worth noting that each of the three components has brought significant performance enhancement when tested on DF-1.0. This informs that the components are still powerful and helpful with reliable robustness even when the candidate face image contains artificial noise.

4.6 Model Depth Analysis

In this study, the model depth of the pooling Transformer is set to 24 with three phases having the pooling layers in between every two consecutive phases. In this section, we conducted an ablation
study on the effectiveness of various model depths. Specifically, we introduced three experiments with different model depths and evaluated the performance accordingly. The three experiments are conducted upon the model with depths [4, 4, 6], [8, 8, 8], and [10, 10, 12] for the three phases and having the total depths of 14, 24, and 32 from shallow to deep, respectively.

As Table 6 demonstrated, it can be concluded that deeper detection models do not always derive better performance than shallow ones. To take a closer look, the detection performance on FF++ consistently increases as the model depth becomes larger. However, the rate of increase is relatively trivial with around a 1% AUC score gap between the adjacent experiments on FF++. On the contrary, a model with depth 24 achieves the best cross-dataset performance with 5.8%, 7.51%, and 11.28% advantages on DFDC, Celeb-DF, and DF-1.0, respectively, over the shallow model with depth 14 concerning the AUC score. Meanwhile, it can be observed that the model starts to overfit on FF++ with performance damping by 2.38%, 1.55%, and 5.18% AUC scores on DFDC, Celeb-DF, and DF-1.0, respectively, as the model depth goes deeper from 24 to 32. Thus, model depth 24 is the optimal setting in this study.

### 5 CONCLUSION

As the real-life Deepfake videos circulating on the Internet become more realistic and difficult to be identified, Deepfake detection models are facing considerable challenges, especially the transferability to the unseen Deepfake contents. In this article, we present a novel deep convolutional Transformer using convolutional pooling and re-attention techniques for facial feature learning both locally and globally in Deepfake detection. We also demonstrate the significance of the barely discussed image keyframes in image feature learning as normal image frames suffer from information loss with common video compression. Compared with the existing Deepfake detection baselines, our approach achieves state-of-the-art performance in both within- and cross-dataset experiments with robust transferability. Furthermore, the visualization of the learned facial features has shown richer information obtained from the keyframes over the normal frames. Therefore, thoroughly exploring the image keyframes and studying the decisive features and relations both locally and globally can facilitate further performance enhancement in Deepfake detection.

Despite satisfactory experiment performance on Deepfake detection tasks, we are aware of the potential limitations. While the keyframes are of high image quality and contain more dominant features, they may be utilized in Deepfake generation. If the opponents intentionally raise the authenticity of Deepfake upon the keyframes, our proposed solution is highly likely to be fooled.

### ACKNOWLEDGMENTS

We would like to thank Yaning Zhang for her help with running and monitoring some of the model-tuning experiments on the remote device.
Deep Convolutional Pooling Transformer for Deepfake Detection

REFERENCES

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. MesoNet: A compact facial video forgery detection network. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (Dec 2018). DOI: https://doi.org/10.1109/wifs.2018.8630761

[2] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. 2019. Deepfake video detection through optical flow based CNN. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop. IEEE, 1205–1207.

[3] Apple. 2021. MPEG-2 Reference Information. Retrieved August 29, 2021 from https://tinyurl.com/m7cet9mc.

[4] Nicolò Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. 2021. Video face manipulation detection through ensemble of CNNs. In Proceedings of the 2020 25th International Conference on Pattern Recognition. IEEE, 5012–5019.

[5] BuzzFeedVideo. 2018. You Won’t Believe What Obama Says In This Video! Retrieved August 29, 2021 from https://www.youtube.com/watch?v=cQ54GDm1eL0.

[6] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. 2020. SimSwap: An efficient framework for high fidelity face swapping. In Proceedings of the MM’20: The 28th ACM International Conference on Multimedia.

[7] Francois Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 1251–1258.

[8] Davide Alessandro Cocconi, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. 2022. Combining efficientnet and vision transformers for video deepfake detection. In Proceedings of the Image Analysis and Processing –ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part III. Springer-Verlag, Berlin, 219–229. DOI: https://doi.org/10.1007/978-3-031-06433-3_19

[9] Davide Alessandro Cocconi, Giorgos Kordopatis Zilos, Giuseppe Amato, Roberto Caldelli, Fabrizio Falchi, Symeon Papadopoulos, and Claudio Gennaro. 2022. MINTIME: Multi-Identity Size-Invariant Video Deepfake Detection. (2022). Retrieved from https://doi.org/10.48550/ARXIV.2211.10996. Accessed February 20, 2022.

[10] Hao Dang, Fung Liu, Joel Shroouwer, Xiaoming Liu, and Anil K. Jain. 2020. On the detection of digital face manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[11] deepfakes. 2019. FACESWAP. Retrieved August 29, 2021 from https://faceswap.dev/.

[12] Brian Dollhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The DeepFake Detection Challenge (DFDC) Dataset. arXiv: cs.CV/2006.07397. Retrieved from https://arxiv.org/abs/2006.07397. Accessed August 29, 2021.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations.

[14] Tanima Dutta and Hari Prabhat Gupta. 2017. An efficient framework for compressed domain watermarking in P frames of high-efficiency video coding (HEVC)–encoded video. ACM Transactions on Multimedia Computing, Communications, and Applications 13, 1, Article 12 (jan 2017), 24 pages. DOI: https://doi.org/10.1145/3002178

[15] FFmpeg. 2021. FFmpeg. Retrieved August 29, 2021 from https://www.ffmpeg.org/.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (Eds.), Vol. 27. Curran Associates, Inc.

[17] Mark Greaves. 2021. Deepfakes’ ranked as most serious AI crime threat. Retrieved November 16, 2021 from https://tinyurl.com/hlfmrat4.

[18] Haya R. Hasan and Khaled Salah. 2019. Combating deepfake videos using blockchain and smart contracts. IEEE Access 7 (2019), 41596–41606. https://ieeexplore.ieee.org/document/8668407.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 770–778.

[20] Yinan He, Bei Gan, Siyu Chen, Yichun Yin, Guojun Yin, Lu Sheng, Jing Shao, and Ziwei Liu. 2021. ForgeryNet: A versatile benchmark for comprehensive forgery analysis. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4358–4367. https://doi.org/10.1109/CVPR4637.2021.00434

[21] Byeongho Heo, Sangdoo Yun, Donghyoon Han, Sanghyuk Chun, Junseuk Choe, and Seong Joon Oh. 2021. Rethinking spatial dimensions of vision transformers. In Proceedings of the International Conference on Computer Vision.

[22] Mousa Tayseer Jafar, Mohammad Ababneh, Mohammad Al-Zoube, and Ammar Elhassan. 2020. Forensics and analysis of deepfake videos. In Proceedings of the 2020 11th International Conference on Information and Communication Systems. IEEE, 053–058.
[23] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. 2020. DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2889–2898.

[24] Leo Kelion. 2018. Deepfake porn videos deleted from internet by Gfycat. Retrieved August 29, 2021 from https://www.bbc.com/news/technology-42905185.

[25] Ira Kemelmacher-Shlizerman. 2016. Transfiguring portraits. ACM Trans. Graph. 35, 4, Article 94 (July 2016), 8 pages. DOI: https://doi.org/10.1145/2997824.2929871

[26] Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. 2022. Hybrid transformer network for deepfake detection. In Proceedings of the 19th International Conference on Content-Based Multimedia Indexing. Association for Computing Machinery, New York, 8–14. https://doi.org/10.1145/3549555.3549588

[27] Naghmeh Khodabakhshi and Mohamed Hefeeda. 2013. Spider: A system for finding 3D video copies. ACM Trans. Multimedia Comput. Commun. Appl. 9, 1, Article 7 (Feb 2013), 20 pages. DOI: https://doi.org/10.1145/2422956.2422963

[28] Jan Kietzmann, Linda W. Lee, Ian P. McCarthy, and Tim C. Kietzmann. 2020. Deepfakes: Trick or treat? Business Horizons 63, 2 (2020), 135–146.

[29] Davis King. 2021. dlib 19.22.1. Retrieved August 29, 2021 from https://pypi.org/project/dlib/.

[30] Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, Yoshua Bengio and Yann LeCun (Eds.).

[31] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. 2021. KoDF: A large-scale Korean deepfake detection dataset. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 10744–10753.

[32] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 2019. FaceShifter: Towards high fidelity and occlusion aware face swapping. arXiv:1912.13457 (2019). Retrieved from https://arxiv.org/abs/1912.13457. Accessed August 29, 2021.

[33] Yuan Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-DF: A large-scale face forgery detection dataset. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 2889–2898.

[34] Zhe Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision.

[35] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. 2021. Generalizing face forgery detection with high-frequency features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 16317–16326.

[36] Kirsti Melville. 2019. The insidious rise of deepfake porn videos — and one woman who won’t be silenced. Retrieved November 16, 2021 from https://www.abc.net.au/news/2019-08-30/11437774.

[37] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. 2018. RSGAN: Face swapping and editing using face and hair representation in latent spaces. In Proceedings of the ACM SIGGRAPH 2018 Posters (SIGGRAPH’18). Association for Computing Machinery, New York, Article 69, 2 pages. DOI: https://doi.org/10.1145/3230744.3230818

[38] Huy H. Nguyen, Junchi Yamagishi, and Isao Echizen. 2019. Use of a Capsule Network to Detect Fake Images and Videos. (2019). arXiv:cs.CV/1910.12467. Retrieved from https://arxiv.org/abs/1910.12467. Accessed August 29, 2021.

[39] Thomas Norman. 2014. 11 - information technology systems infrastructure. In Computer Systems Design (Second Edition). Butterworth-Heinemann, Boston, 203–249.

[40] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr. Dpfks, Carl ShiftFacen, Luis R. P., Jian Jiang, Sheng Zhang, Pingyu Wu, Bo Zhou, and Weiming Zhang. 2021. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. arXiv:cs.CV/2005.05535. Retrieved from https://arxiv.org/abs/2005.05535. Accessed August 29, 2021.

[41] Hareesh Ravi, A. V. Subramanyam, and Sabu Emmanuel. 2016. Forensic analysis of linear and nonlinear image filtering using quantization noise. ACM Trans. Multimedia Comput. Commun. Appl. 12, 3, Article 39 (March 2016), 23 pages. DOI: https://doi.org/10.1145/2857069

[42] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. 2019. FaceForensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 1–11.

[43] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision. 618–626.

[44] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations (ICLR). IEEE.
Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research), Vol. 97. PMLR, 6156–6164.

TechBytes. 2021. Retrieved August 29, 2021 from https://tinyurl.com/8tjsj9th9.

Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred neural rendering: Image synthesis using neural textures. ACM Trans. Graph. 38, 4, Article 66 (July 2019), 12 pages.

Justus Thies, Michael Zollhöfer, Marc Stamina, Christian Theobalt, and Matthias Nießner. 2016. Face2Face: Real-time face capture and reenactment of RGB videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2387–2395.

Rob Toews. 2020. Deepfakes Are Going To Wreak Havoc On Society. We Are Not Prepared. Retrieved August 29, 2021 from https://tinyurl.com/58mpuac7.

Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. 2020. Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion 64 (2020), 131–148. https://www.sciencedirect.com/science/article/abs/pii/S1566253520303110.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc.

Krishna Rao Vijayanagar. 2020. I, P, and B-frames – Differences and Use Cases Made Easy. Retrieved August 29, 2021 from https://tinyurl.com/syvfx27thk.

Qingzhong Wang, Pengfei Zhang, Haoyi Xiong, and Jian Zhao. 2021. Face. evolVe: A high-performance face recognition library. arXiv:2107.08621. Retrieved from https://arxiv.org/abs/2107.08621. Accessed February 20, 2022.

Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. 2021. HiFiFace: 3D shape and semantic prior guided high fidelity face swapping. In Proceedings of the 13th International Joint Conference on Artificial Intelligence, IJCAI-21, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 1136–1142. DOI: https://doi.org/10.24963/ijcai.2021/157

Longyin Wen, Honggang Qi, and Siwei Lyu. 2018. Contrast enhancement estimation for digital image forensics. ACM Trans. Multimedia Comput. Commun. Appl. 14, 2, Article 49 (may 2018), 21 pages. DOI: https://doi.org/10.1145/3183518

Mika Westerlund. 2019. The emergence of deepfake technology: A review. Technology Innovation Management Review 9, 11 (11/2019 2019), 40–53.

Deressa Wodajo and Solomon Atnafu. 2021. Deepfake Video Detection Using Convolutional Vision Transformer. (2021). arXiv:2102.11126. Retrieved from https://arxiv.org/abs/2102.11126. Accessed February 20, 2022.

Ting Yao, Yehao Li, Yingwei Pan, Yu Wang, and Tao Mei. 2022. Dual vision transformer. arXiv:2207.04976 (2022). Retrieved from https://arxiv.org/abs/2207.04976. Accessed February 20, 2022.

Ting Yao, Yingwei Pan, Yehao Li, Chong-Wah Ngo, and Tao Mei. 2022. Wave-ViT: Unifying wavelet and transformers for visual representation learning. In Proceedings of the European Conference on Computer Vision.

Hangqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. 2021. Multi-attentional deepfake detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2185–2194.

Jian Zhao, Yu Cheng, Yan Xu, Lin Xiong, Jiashu Li, Fang Zhao, Karlekar Jayashree, Sugiri Pranata, Zhengmei Shen, Junliang Xing, et al. 2018. Towards pose invariant face recognition in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2207–2216.

Jian Zhao, Jiashu Li, Yu Cheng, Terence Sim, Shuicheng Yan, and Jiashi Feng. 2018. Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. In Proceedings of the 26th ACM International Conference on Multimedia. Association for Computing Machinery, New York, 792–800. DOI: https://doi.org/10.1145/3240508.3240509

Jian Zhao, Lin Xiong, Yu Cheng, Yi Cheng, Jiashu Li, Li Zhou, Yan Xu, Jayashree Karlekar, Sugiri Pranata, Zhengmei Shen, Junliang Xing, Shuicheng Yan, and Jiashi Feng. 2018. 3D-aided deep pose-invariant face recognition. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI-18. International Joint Conferences on Artificial Intelligence Organization, 1184–1190. DOI: https://doi.org/10.24963/ijcai.2018/165

Jian Zhao, Lin Xiong, Panasonc Karlekar Jayashree, Jiashu Li, Fang Zhao, Zhecan Wang, Panasonc Sugiri Pranata, Panasonc Zhengmei Shen, Shuicheng Yan, and Jiashi Feng. 2017. Dual-agent GANs for photorealistic and identity preserving profile face synthesis. In Proceedings of the Advances in Neural Information Processing Systems, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc.
[67] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Qibin Hou, and Jiashi Feng. 2021. DeepViT: Towards Deeper Vision Transformer. arXiv:2103.11886. Retrieved from https://arxiv.org/abs/2103.11886. Accessed February 20, 2022.

[68] Yuhao Zhu, Qi Li, Jian Wang, Chengzhong Xu, and Zhenan Sun. 2021. One shot face swapping on megapixels. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4834–4844.

[69] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. 2020. WildDeepfake: A challenging real-world dataset for deepfake detection. In Proceedings of the 28th ACM International Conference on Multimedia. 2382–2390. DOI: https://doi.org/10.1145/3394171.3413769

Received 16 September 2022; revised 29 January 2023; accepted 15 March 2023