electronics

MDPI

Article

A Transformer-Based DeepFake-Detection Method for Facial Organs

Ziyu Xue [1,2,*], Qingtong Liu [2], Haichao Shi [3], Ruoyu Zou [1] and Xiuhua Jiang [1,4]

[1] School of Information and Communication Engineering, Communication University of China, Beijing 100024, China
[2] Academy of Broadcasting Science, NRTA, Beijing 100866, China
[3] Institute of Information Engineering, CAS, Beijing 100093, China
[4] Peng Cheng Laboratory, Shenzhen 518055, China
* Correspondence: xzy_88@126.com or xueziyu@abs.ac.cn

Abstract: Nowadays, deepfake detection on subtle-expression manipulation, facial-detail modification, and smeared images has become a research hotspot. Existing deepfake-detection methods on the whole face are coarse-grained, where the details are missing due to the negligible manipulated size of the image. To address the problems, we propose to build a transformer model for a deepfake-detection method by organ, to obtain the deepfake features. We reduce the detection weight of defaced or unclear organs to prioritize the detection of clear and intact organs. Meanwhile, to simulate the real-world environment, we build a Facial Organ Forgery Detection Test Dataset (FOFDTD), which includes the images of mask face, sunglasses face, and undecorated face collected from the network. Experimental results on four benchmarks, i.e., FF++, DFD, DFDC-P, Celeb-DF, and for FOFDTD datasets, demonstrated the effectiveness of our proposed method.

Keywords: generated face; image-forensics detection; generative adversarial network

Citation: Xue, Z.; Liu, Q.; Shi, H.; Zou, R.; Jiang, X. A Transformer-Based DeepFake-Detection Method for Facial Organs. *Electronics* **2022**, *11*, 4143. https://doi.org/10.3390/electronics11244143

Academic Editor: Byung Cheol Song

Received: 16 October 2022
Accepted: 9 December 2022
Published: 12 December 2022

1. Introduction

Nowadays, the deepfake methods have been applied to various scenarios and have achieved great success, such as actor face-swapping and virtual hosts. However, the incidents of using face-swapping methods on celebrities have confused the public and disrupted the social order, which has also brought harmful effects at the same time. It has become increasingly important to explore the deepfake-detection methods.

Generally, the deepfake methods can be divided into two categories regarding the type of synthetic object: complete-facial synthesis and facial-organ synthesis. Identifying the complete-facial synthesis is the mainstream method. Earlier methods [1,2] detect the forgery regions by combining physical methods such as symbiotic matrix and spectral consistency, with deep learning methods. Meanwhile, some other methods [3,4] put main emphasis on CNN models to extract synthetic trace-features and detect the superimposed noise generated during the synthesis process [5–7]. Recently, Ni et al. [8] proposed a detection method based on consistent representation learning by capturing the different representations with different augmentations and calculating the distance for the different representations. Wang et al. [9] proposed to combine the color domain and frequency domain, using a frequency-domain filter-based multi-scale transformer. Yuan et al. [10] achieved deepfake discrimination by classifying fake methods using multi-class forgery-classification tasks. With the improvement of the effect of deepfake approaches, fewer and fewer traces of forgery are detected, especially for the subtle operations on expressions and organs. The identification methods in complete-facial synthesis have reflected some shortcomings, and need to be addressed.

With the development of deepfake methods, deepfake detection on partial areas has become a hot topic. Existing approaches [11–13] have utilized deep learning models to ex-

tract physiological features, which focus on the detection of unusual physiological features. Meanwhile, it has been of vital importance for research to detect facial organs consistently. Matern et al. [14] detect a fake face by the light reflections and color inconsistencies in specific areas such as eyes and teeth. Nirkin et al. [15] employ contextual association to detect inconsistencies in organs of the face after facial recognition. In recent years, with the rapid development of the attention mechanism and visual transformer, some methods [16,17] integrate the attention mechanism to make the discriminator focus more on the manipulated partial-areas. The visual-semantic transformer [18] divides the person's face into organs, and builds an attention map for the parsing map to judge the faked region by blocks.

Although some promising methods have focused on facial-region deepfake detection, most of the existing methods still detect forgeries region-by-region without considering organ-level detection. In particular, as the forgery methods tend to diversify, the forged people may wear masks, sunglasses, or other accessories, which will cause the partial face to be occluded. The features extracted from the original methods are further reduced, leading to poor detection performance. At the same time, to avoid being detected, some of the forged media are blocked with cropping, masking, down-sampling, and other operations, which will also reduce the accuracy of existing deepfake-detection methods.

To address the above problems, we propose a novel deepfake-detection framework based on the transformer to detect organs and whole faces. Meanwhile, we build a test dataset that simulates the real-world scenarios to verify the effectiveness of our method. Our contribution can be summarized as follows:

1. We propose a novel deepfake-detection method based on transformer architecture, which can identify facial-detail editing and is robust to synthetic facial-recognition methods dealing with occluded masks or sunglasses.
2. The method focuses on organ-based forgery detection, which trains different transformers for different organs. Each transformer can work independently and flexible. At the same time, the weight of obscured and stained organs are reduced automatically.
3. We propose a deepfake-detection dataset, namely Facial Organ Fake Detection Test Dataset (FOFDTD). It is consisting of 750 authentic images, 750 GAN-generated images, and 900 forgery images made by humans, including masks, sunglasses, and undecorated faces. All the authentic images in the FOFDTD are collected from the Internet, and are mainly used for deepfake detection in real-world.

2. Related Work

The early forgery-detection methods mainly focused on the whole image. Fridrich et al. [19] proposed a novel strategy to detect the forgery images. The method begins by assembling a model of noise components into a union of several different sub-models, formed from the joint distribution of adjacent samples obtained using linear and nonlinear high-pass filters that quantify image-noise residuals. Cozzolino et al. [20] proposed the local descriptors, which can be regarded as a CNN, to detect the image-noise residual, find the forgery regions, and label them. Tan et al. [21] proposed a scaling method that uniformly scales all depth-, width-, and resolution-dimensions, using a simple yet highly effective compound coefficient.

With the development of image-synthesis technology, face-oriented forgery has become an important research direction. Compared with the earlier detection methods, face-oriented forgery detection has a specific dependence on physiological features. Meanwhile, traditional image-forensics techniques are usually not well suited to deepfake detection, due to the compression that strongly degrades the data and the post-processing operations which confuse detection or image artifacts. Afchar et al. [22] presented a method to automatically and efficiently detect face forgery. The model has a low number of layer networks, and focuses on the subtle features of the image. Güera et al. [23] proposed a temporal-aware pipeline to detect deepfake media, using a CNN to extract the features in the frame level and using an RNN to classify them. Chollet [24] proposed a deep convolutional-neural-network

architecture illuminated by inception, where inception modules have been replaced with depth-wise separable convolutions. Experiments have proved that the model explicitly affects the identification of forged faces.

At the same time, research on frequency detection or inconsistency is also the focus of research on facial forgery. Qian et al. [25] utilized frequency-aware decomposed-image components and local frequency statistics, in face-forgery detection Luo et al. [26] proposed utilizing high-frequency features for detection by combining models such as a multi-scale high-frequency feature-extraction module and a residual-guided spatial-attention module. Ni [8] proposed consistent representation learning (CORE), which constrains the consistency of different representations. The method based on the different representations is first captured with different augmentations, and then the cosine distance of the representations is regularized, to enhance consistency. Although face-based detection performs well in most deep-forgery-detection scenarios, the detection granularity of those methods is relatively coarse. There is still room for improvement when detecting fine-grained forgery methods.

Detecting partial regions is a fine-grained method aiming to detect a forged face. Matern et al. [14] used several characteristic artifacts, such as eyebrow color and geometric analysis, to detect face forgery. Chen et al. [27] proposed DefakeHop, which uses the successive subspace learning (SSL) principle and the channel-wise Saab transform to extract features, and the feature-distillation model to reduce the spatial dimension. In addition, dividing the synthetic image into the partial area for detection, is a novel method. Chen et al. [28] proposed a framework to build correlation between the partial regions, to avoid the overfitting problem caused by global supervision. These methods still have certain limitations, and the detection effect for some situations needs improvement, such as occluded and low-resolution images.

Recently, multiple-attention mechanisms and multiple models have become essential methods for solving deepfake discrimination. Zhao et al. [29] proposed a multi-attentional deepfake-detection method to detect the subtle and partial features in real and fake images. The technique had three key components: multiple spatial-attention-heads, a textural-feature-enhancement block, and an aggregate module. Wang et al. [9] proposed a multi-modal, multi-scale transformer, to detect deepfake images. The model can detect image patches of different sizes, to find the local inconsistencies at different spatial-levels.

The above methods can identify the deepfake content within limits. However, the above methods still need to detect the critical organs of the face, which leads to the detection effect of the process not being robust in low-resolution images and organ-occlusion-images detection. In particular, post-processing methods or image defilements are added to some deepfake images to blind the deep-forgery-discrimination model, which makes the existing detection-methods based on the whole face, more difficult.

Based on this, we employ the transformer to perform local-organ detection and combine the full-face features for analysis. Finally, the method employs a classifier to synthesize the identification results. The experiment shows that our method can adapt well to face occlusion, image defacement, or low resolution.

3. Proposed Method

3.1. Overview of the Framework

We propose a novel deepfake-detection architecture. Firstly, we extracted the critical organs, which include the eyes, nose, mouth, eyebrows, and ears, and we built the transformer encoder and calculated the feature vector by organ. Meanwhile, we also built a transformer encoder for the whole face, to supplement the discrimination. After that, we combined the feature vectors of each organ and the whole face, to form a feature-vector group. Finally, we classified the vector group. Moreover, the feature weights were set to 0 when the organ was stained, as shown in Figure 1. The input image is shown in $I \in P^{H \times W \times 3}$, where H, W represents the height and width, respectively.

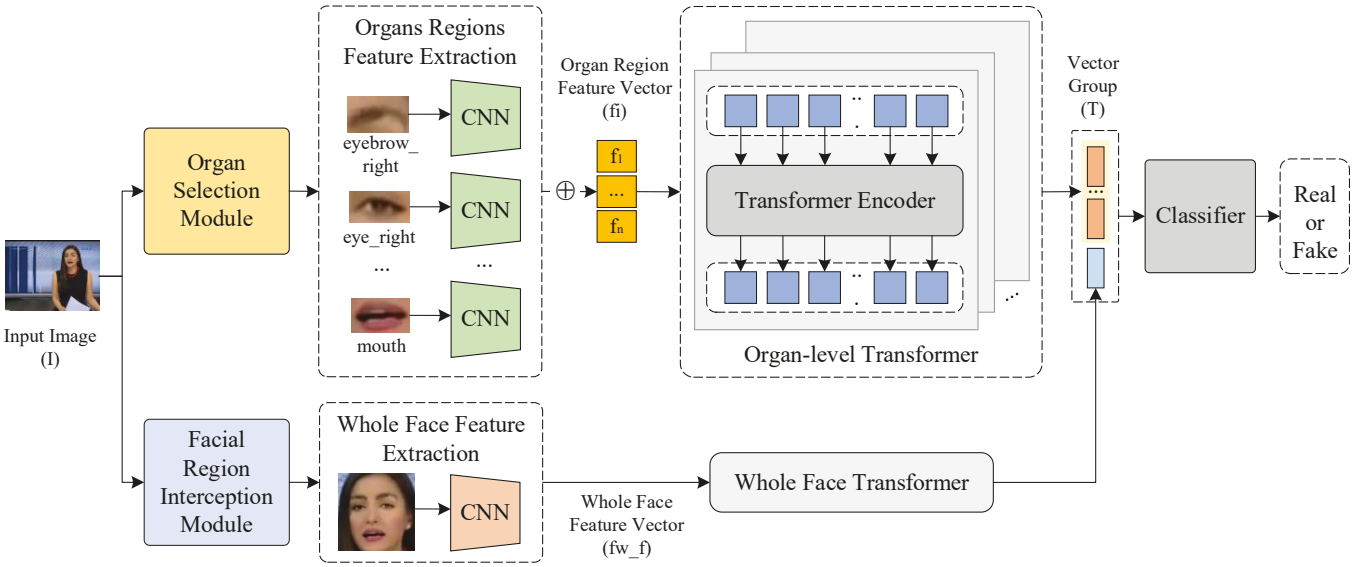


Figure 1. Our proposed method framework includes an organ-selection module, facial-region interception module, organ-level transformer, and classifier. The organ-selection module is mainly used to select clear organs and then utilizes the extractor to obtain features organ by organ. The facial-region interception module mainly frames the facial area and then uses the feature extractor to obtain the features of the whole face. Finally, we use a classifier to classify the results.

3.2. Organ Selection and Feature Extraction

We employed the Dlib [30] to obtain 68 coordinate points in the facial RoI to extract critical organs such as the eyes, nose, mouth, eyebrows, and ears. Each part of the region can be written as $\left(\tau_i^{H_i \times W_i \times 3}, \partial_i\right)$, where H_i, W_i are the width and height of the region τ_i, respectively, and 3 is the number of color channels. ∂_i is the weight occupied by the region τ_i, and we set the weight of occluded or unclear organs to 0, to ensure their features did not affect the final results. We utilized the multiple CNNs to extract the feature map $f_i \in \tau_i^{(H_i/4)\times(W_i/4)\times C}$ by organ, and the extracted features were used as input for the organ-level transformer. Moreover, the fusion module excludes the part where the weight is 0.

3.3. Organ-Level Transformer

We set up a multi-group transformer model for different organs, and each organ corresponded to a transformer. We set the feature map, f_i, of each organ, τ_i, as the input, which was partitioned into spatial blocks of various sizes. The self-attention of the spatial blocks was calculated using different headers. Following the method 9, we extracted the block with the shape $r_h \times r_h \times C$ from f_i. We reshaped it to a 1-dimensional vector of h-heads. After that, the flattened vector was embedded into the sequence, using the fully connected layer to form $Q_i^h \in \tau_i^{(H_i/4r_h)\times(W_i/4r_h)\times(r_h \times r_h \times C)}$. In addition, the key embeddings, K_i^h, and the value embeddings, V_i^h, had the same operation. The attention matrix corresponding to organ τ_i can be calculated using Equation (1).

$$A_i^h = softmax\left(\frac{Q_i^h\left(K_i^h\right)^T}{r_h \times r_h \times C}\right)V_i^h \tag{1}$$

Then A_i^h was reduced to the resolution of the original space, and the features of different heads were stitched together, to obtain the output $T_i \in \tau_i^{(H_i/4)\times(W_i/4)\times(r_h \times r_h \times C)}$ through the 2D residual block.

After each organ calculation, all the vectors were reorganized into a vector group, T, as shown in Equation (2).

$$T = (T_1, T_2, \ldots, T_n) \tag{2}$$

where n represents the number of valid organs.

3.4. Whole-Face Transformer and Classifier Network

Following the method 9, we built a transformer based on the whole face, which received the facial features, W_{w_f}, extracted from the feature model. After the feature extraction by the transformer, we sent them to the classifier network.

The classifier consists of a linear layer and a softmax normalization layer, which can map the probability, p, to represent the vector group, T, into scalars. The two-dimension output of the softmax layer was used as the final probability 8, and the output probability was used to distinguish the deepfake faces.

3.5. Loss Functions

We utilized the standard cross-entropy loss as the classification loss, as shown in Equation (3).

$$l(p) = ylogp + (1 - y)log(1 - p) \tag{3}$$

where y denotes the ground-truth and p is the predicted result for an organ. For a single image with N organs, the classification loss can be calculated as follows:

$$l\left(p_1^N\right) = \sum_{n=1}^{N} (l(p_n)) \tag{4}$$

4. Facial-Organ Forgery-Detection Test Dataset

We proposed a test dataset FOFDTD to simulate an occluded-face forgery in the real world. FOFDTD has 750 authentic images, 750 GAN-generated images, and 900 forgery images made by humans. Moreover, it consists of masks, sunglasses, and undecorated faces. The actual images were collected from the Internet (www.baidu.com, accessed on 30 August 2022), and we used StarGAN [31] to make the fake images. Meanwhile, the artificial images were used to ensure authenticity, as shown in Figure 2. We can download FOFDTD at www.github.com/ZiyuXue/FOFDTD (accessed on 30 August 2022).

4.1. GAN-Generated Image

GAN-generated images use StarGAN to edit and modify the character attributes in the entire image and the images without adding manual operations. The forgery methods mainly include appearance changes such as darker skin color, lighter skin color, and hair-color change, and emotional changes such as anger, joy, and smile, as shown in Figure 3.

4.2. Artificial Image

Artificial images use PhotoShop and other software to modify the local details of the face and retain the background or irrelevant details. For example, in the group "eyebrow-color change", we only adjusted the eyebrow color, and did not process the background and other facial details. The manual methods included darker skin color, lighter skin color, thinner face, eyebrow-color change, eye-size change, and lip-color change, as shown in Figure 4.

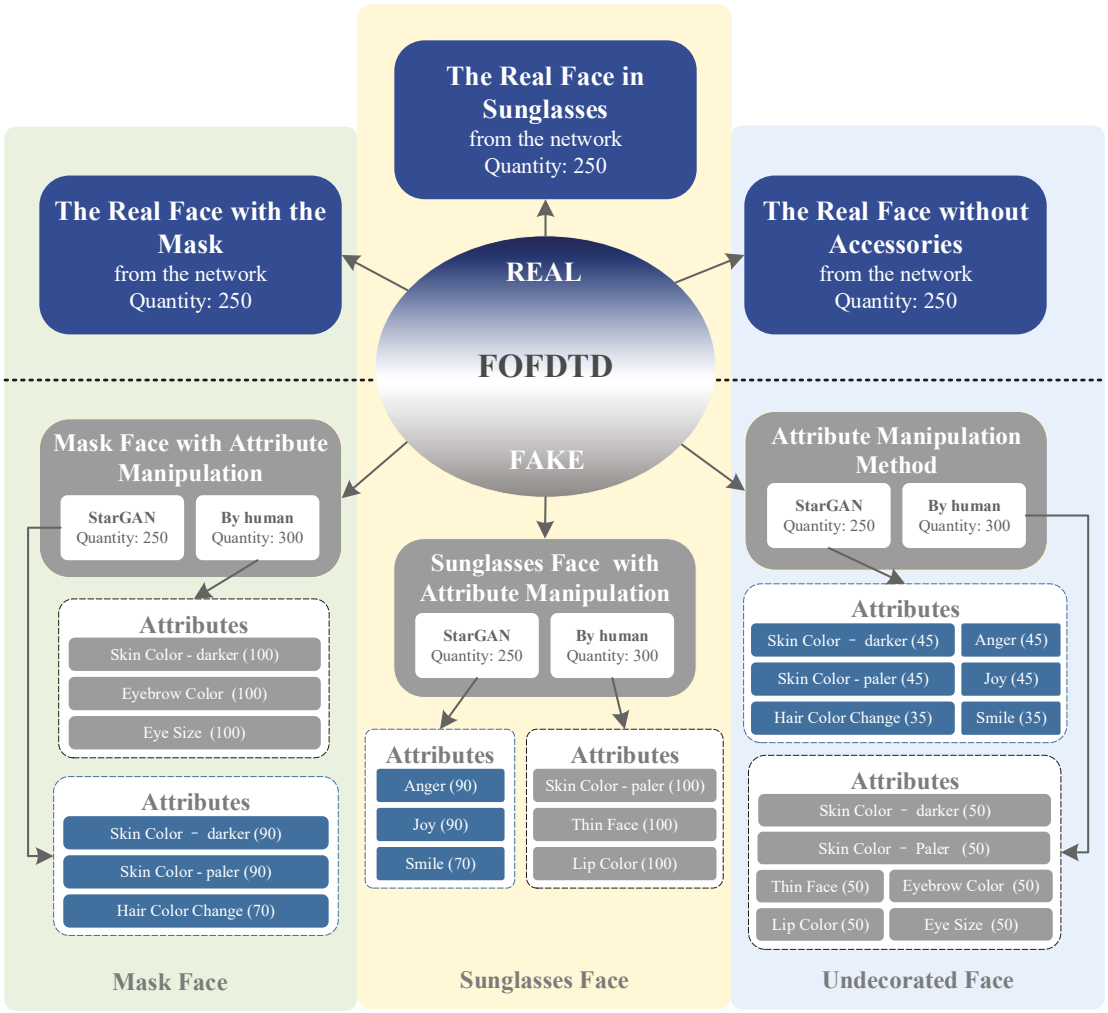


Figure 2. Overview of Focal-Organ Fake-Detection Test Dataset (FOFDTD).

Figure 3. The example of GAN-generated images in FOFDTD.

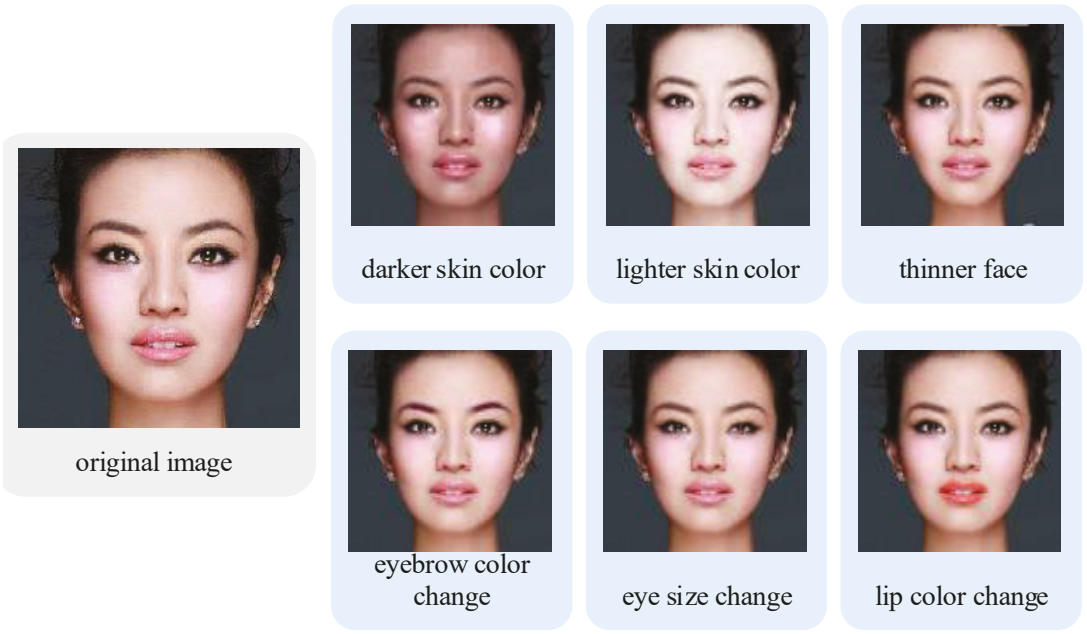


Figure 4. The example of artificial images in FOFDTD.

5. Experimental Results

5.1. Implementation Details

Dataset: We selected five mainstream datasets, including FaceForensics++ (FF++) [32], Celeb-DF [33], deepfakeDetection (DFD) [34], and DFDC Preview (DFDC-P) [35]. FF++ is widely used for deepfake facial-detection. It includes 1000 real videos and 4000 fake videos. In the real videos, there are 720 videos for training and 280 for validation and testing. Celeb-DF contains 590 real videos obtained from the Youtube website and 5939 fake videos, which have a more realistic faking effect compared with earlier methods. The DFD dataset contains 363 real and 3068 fake videos, with FF++ as the base data. DFDC-P contains 1131 real and 4119 fake videos, all of which are low quality and modified in terms of gender, age, etc. Meanwhile, we also examined our model in FOFDTD.

Implementation Protocol: Our model was trained using TensorFlow on one NVIDIA RTX 3090 (24G) GPU. The accuracy (ACC) and area under the curve (AUC) were used as the evaluation criteria.

5.2. Ablation Study

We conducted ablation experiments for different organ-combinations, to verify the effectiveness of our method. We selected FF++ to train and test our method, and the results are shown in Table 1. The selected combinations included four groups: "eyebrows + eyes + nose", "nose + mouth + ears", "eyes + nose + mouth", and "eyes + nose + mouth + eyebrows + ears".

Table 1. Results of key-organ ablation experiments.

Combination	Simulated Scene	ACC	AUC
eyebrows + eyes + nose	Face with mask	86.73	83.24
nose + mouth + ears	Face with sunglasses	86.52	84.13
eyes + nose + mouth	Cover the eyebrows	95.43	93.64
eyes + nose + mouth + eyebrows + ears	Undecorated face	99.67	99.93

We also mapped the experimental group to the corresponding scene. The group "eyebrows + eyes + nose" represented the face with a mask in the real world, the group "nose + mouth + ears" represented the face with sunglasses in the real world, and the group

"eyes + nose + mouth + eyebrows + ears" represented the undecorated face, as shown in Table 1.

We reduced the weight of the covered organs to 0. For example, in the group "eyebrows + eyes + nose", we reduced the weight of the mouth to 0. Table 1 shows that our method is acceptable when detecting missing- organs forged faces. Our proposed method had a better detection effect on "undecorated face." We also found that the "eyes + nose + mouth" group showed better detection results. The artifacts on the eyes, nose, and mouth were more distinct in the detection process than those on the eyebrow.

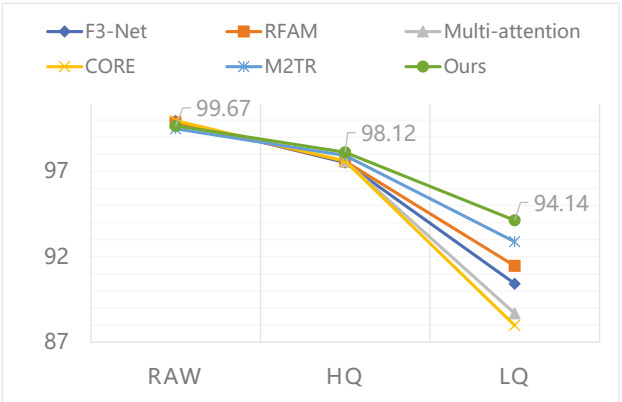
5.3. Comparison with Other Methods

We evaluated the detection results at the frame level using the FF++ dataset, which includes RAW, LQ, and HQ. Table 2 shows our method's ACC and AUC results and state-of-the-art on the FF++ dataset. The best experimental group in the ablation experiment was selected for comparison with other methods. It can be seen that our method had a superior detection performance on the HQ and LQ datasets. Our model can reduce the weight of unclear organs after zoom, to guarantee the overall effect of the method. From the table, our method improved by 1.25% in ACC and 1.12% in AUC, the results from M2TR 9 in the LQ dataset. Moreover, our method showed a better result for the HQ datasets.

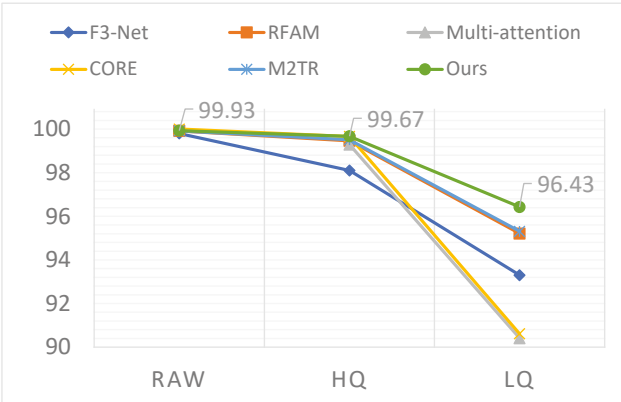
Table 2. Detection results of Acc and AUC at frame level for our method and the state-of-the-art (FF++).

Methods	RAW		HQ		LQ	
	ACC	AUC	ACC	AUC	ACC	AUC
Steg.Features [19]	97.63	-	70.97	-	55.98	-
LD-CNN [20]	98.57	-	78.45	-	58.69	-
MesoNet [22]	95.23	-	83.10	-	70.47	-
F^3-Net [25]	99.95	99.80	97.52	98.10	90.43	93.30
RFAM [28]	99.87	99.92	97.59	99.46	91.47	95.21
Multi-attention [29]	-	-	97.60	99.29	88.69	90.41
CORE [8]	99.97	100.00	97.61	99.66	87.99	90.61
M2TR [9]	99.50	99.92	97.93	99.51	92.89	95.31
Ours	99.67	99.93	98.12	99.67	94.14	96.43

Figure 5 shows a line chart, in which our model has better adaptability for HQ and LQ data. Our method performs well in the LQ dataset, mainly because the method reduces the weight of obscured or low-resolution organs, and ensures the detection of clear organs and faces. This also proves the proposed method has relatively little impact on the change in segmentation ratio, and can be adapted to detect the scaled synthetic images.



(a) The line chart of ACC in RAW, HQ and LQ dataset (%).



(b) The line chart of AUC in RAW, HQ and LQ dataset (%).

Figure 5. The line chart of ACC and AUC in the RAW, HQ, and LQ dataset.

5.4. Cross-Dataset Evaluation

Mainstream Dataset. We also evaluated our approach on the cross-dataset, as shown in Table 3. We trained our approach in FF++ and tested in DFD, DFDC-P, and Celeb-DF. In Table 3, Xception, Local-relation, HFF, and CORE methods were trained using real and faked images in FF++. Note that LSC only used real images for training. The experimental results are referred to in 8.

Table 3. The cross-dataset evaluation results for DFD, DFDC-P, and Celeb-DF datasets (AUC%).

Methods	DFD	DFDC-P	Celeb-DF
Xception [24]	87.86	-	73.04
Local-relation [28]	89.24	76.53	78.26
HFF [26]	91.90	-	79.40
LSC [36]	-	74.37	81.80
CORE [8]	93.74	75.74	79.45
Ours	94.32	75.93	82.43

As shown in Table 3, our model performed better on DFD and Celeb-DF datasets. Meanwhile, the local-relation method's AUC was 0.6% higher than our method for DFDC-P. That is because the DFD and Celeb-DF datasets have some lower-resolution data. Our method focuses on organ detection and does not rely on high resolution, which is more suitable for such detection tasks. In this way, low-resolution faces did not affect subsequent results. Meanwhile, our method was relatively more stable in the cross-dataset experiment than the others.

FOFDTD. We evaluated the performance effects of the baseline and advanced methods [9,21] on FOFDTD. We set up two group experiments on GAN-generated and artificial datasets, as shown in Tables 4 and 5. Moreover, the model we used was consistent with the above experiments.

Table 4. Cross-dataset evaluation results on FOFDTD (GAN-generated) dataset (ACC% and AUC%).

Methods	All		Mask Face		Sunglasses		Undecorated	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Efficientnet-B4 [21]	49.1	78.5	48.7	68.5	48.8	89.4	49.7	83.2
M2TR [9]	61.7	63.8	57.6	58.8	66.5	70.6	61.0	65.4
Ours	64.2	66.7	60.5	60.2	68.9	72.3	63.3	67.4

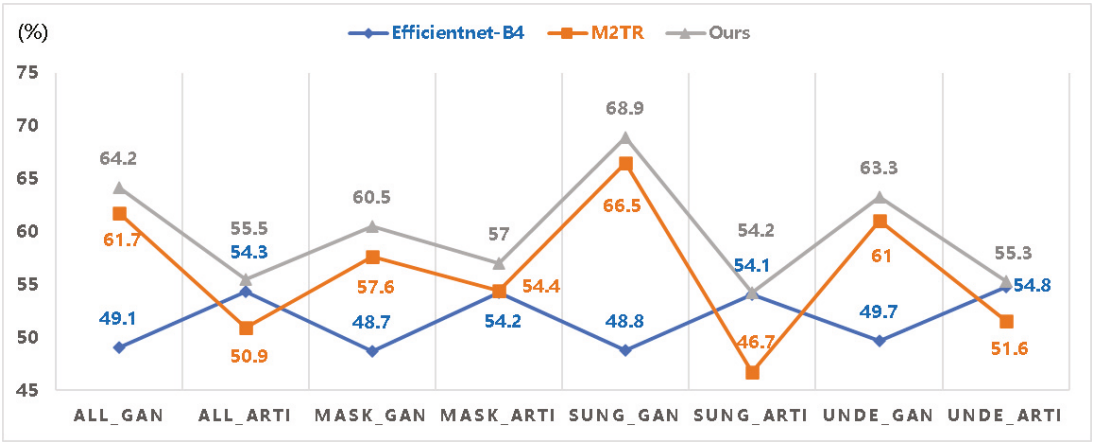
Table 5. Cross-dataset evaluation results on FOFDTD (artificial) dataset (ACC% and AUC%).

Methods	All		Mask Face		Sunglasses		Undecorated	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Efficientnet-B4 [21]	54.3	50.0	54.2	50.6	54.1	53.8	54.8	47.9
M2TR [9]	50.9	51.7	54.4	52.7	46.7	49.4	51.6	53.8
Ours	55.5	55.0	57.0	55.5	54.2	53.4	55.3	56.2

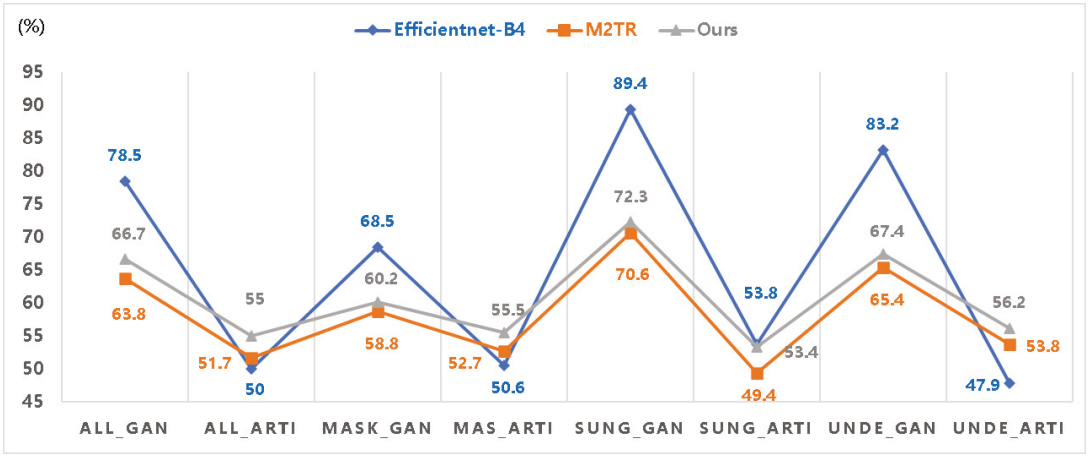
Table 4 shows that our method had the best ACC score compared with other methods. The average ACC is 15.1% higher than EfficientNet-B4 and 2.5% higher than M2TR, but EfficientNet-B4 performs well for AUC.

We synthesized Tables 4 and 5, to form Figure 6. Figure 6a is the ACC result and Figure 6b is the AUC result. At the same time, we put similar groups next to each other. For example, the first two columns in Figure 6a represent the overall ACC generated by GAN and artificially, to facilitate comparison between the similar groups. It should be noted that during the detection process, a suitable method's ACC and AUC curves should

be relatively gentle, and the severe fluctuation is not conducive for application to the actual situation. Figure 6 shows that, compared with others, our method was more stable and had better robustness for deepfake detection in different forgery types.



(**a**)



(**b**)

Figure 6. The line chart of the comparison in ACC and AUC. (**a**) The line chart of the comparison between GAN-generated and artificial datasets in ACC (%). (**b**) The line chart of the comparison between GAN-generated and artificial datasets in AUC (%).

5.5. Complexity Measure

We compared the time complexity with methods [9,21] as shown in Table 6. The experiment showed that our approach and M2TR [9] had more time complexity than the basic method [21] based on CNN, which is a common problem with transformer-based methods.

Table 6. Time-complexity comparison table with the state-of-the-art.

Methods	Input Size	Params	Flops
Efficientnet-B4 [21]	320 × 320	0.129 M	0.1391 G
M2TR [9]	320 × 320	0.345 M	1.5217 G
Ours	320 × 320	1.221 M	3.8573 G

Compared with M2TR, the time complexity of our method was higher, mainly because our model has at least one organ-level transformer. Although our approach ignores some unclear organs, organ-level transformer detection still consumes a lot of computation.

To optimize efficiency, we will choose the more critical organ-level transformers for the following work step.

6. Conclusions

In this paper, we propose a transformer-based deepfake-detection method for facial organs, which can effectively distinguish deepfake media. Our method is robust to detect subtle expression-manipulation, partial detail-modification, and stained deepfake images. We also build transformers at organ level, to obtain the features. The accuracy was increased by reducing the weights of organs that were stained, defaced, and of low-quality. A whole-face transformer was also used to assist in the detection of partial information. Moreover, we built a test dataset to simulate the realistic scenarios for facial-organ deepfake discrimination, named FOFDTD. The dataset consists of the mask face, sunglasses face, and undecorated face. To verify the effectiveness of our method, we evaluated our method with the FF++, DFD, DFDC-P, Celeb-DF, and FOFDTD datasets. The results demonstrated that our method are superior to the state-of-the-art methods.

Author Contributions: Z.X. designed the study. Z.X. and Q.L. performed the experiments and analyzed the data. Z.X. wrote the paper. H.S. and X.J. guided the research and reviewed the manuscript. R.Z. labeled the test set. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Fiscal Expenditure Program of China under Grant 130016000000200003.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Nataraj, L.; Mohammed, T.M.; Manjunath, B.S.; Chandrasekaran, S.; Flenner, A.; Bappy, J.H.; Roy-Chowdhury, A.K. Detecting GAN generated fake images using co-occurrence matrices. *Electron. Imaging* **2019**, *2019*, 532-1–532-7. [CrossRef]
2. Barni, M.; Kallas, K.; Nowroozi, E.; Tondi, B. CNN detection of GAN-generated face images based on cross-band co-occurrences analysis. In Proceedings of the 2020 IEEE International Workshop on Information Forensics and Security (WIFS), New York, NY, USA, 6–11 December 2020; pp. 1–6.
3. Mi, Z.; Jiang, X.; Sun, T.; Xu, K. Gan-generated image detection with self-attention mechanism against gan generator defect. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 969–981. [CrossRef]
4. Wang, S.Y.; Wang, O.; Zhang, R.; Owens, A.; Efros, A.A. CNN-generated images are surprisingly easy to spot... for now. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8695–8704.
5. Hu, J.; Liao, X.; Wang, W.; Qin, Z. Detecting Compressed Deepfake Videos in Social Networks Using Frame-Temporality Two-Stream Convolutional Network. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 1089–1102. [CrossRef]
6. Chen, B.; Liu, X.; Zheng, Y.; Zhao, G.; Shi, Y.Q. A robust GAN-generated face detection method based on dual-color spaces and an improved Xception. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 3527–3538. [CrossRef]
7. He, Y.; Yu, N.; Keuper, M.; Fritz, M. Beyond the spectrum: Detecting deepfakes via re-synthesis. *IJCAI* **2021**, 2534–2541.
8. Ni, Y.; Meng, D.; Yu, C.; Quan, C.; Ren, D.; Zhao, Y. CORE: Consistent Representation Learning for Face Forgery Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12–21.
9. Wang, J.; Wu, Z.; Ouyang, W.; Han, X.; Chen, J.; Jiang, Y.G.; Li, S.N. M2tr: Multi-modal multi-scale transformers for deepfake detection. In Proceedings of the 2022 International Conference on Multimedia Retrieval, Newark, NJ, USA, 27–30 June 2022; pp. 615–623.
10. Yuan, Y.; Fu, X.; Wang, G.; Li, Q.; Li, X. Forgery-Domain-Supervised Deepfake Detection with Non-Negative Constraint. *IEEE Signal Process. Lett.* **2022**. [CrossRef]
11. Ciftci, U.A.; Demir, I.; Yin, L. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**. [CrossRef] [PubMed]
12. Agarwal, S.; Farid, H.; El-Gaaly, T.; Lim, S.N. Detecting deep-fake videos from appearance and behavior. In Proceedings of the 2020 IEEE International Workshop on Information Forensics and Security (WIFS), New York, NY, USA, 6–11 December 2020; pp. 1–6.

13. Hu, S.; Li, Y.; Lyu, S. Exposing GAN-generated Faces Using Inconsistent Corneal Specular Highlights. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 2500–2504.
14. Matern, F.; Riess, C.; Stamminger, M. Exploiting visual artifacts to expose deepfakes and face manipulations. In Proceedings of the 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa Village, HI, USA, 9–11 January 2019; pp. 83–92.
15. Nirkin, Y.; Wolf, L.; Keller, Y.; Hassner, T. DeepFake detection based on discrepancies between faces and their context. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 6111–6121. [CrossRef] [PubMed]
16. Dang, H.; Liu, F.; Stehouwer, J.; Liu, X.; Jain, A.K. On the detection of digital face manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5781–5790.
17. Fernando, T.; Fookes, C.; Denman, S.; Sridharan, S. Detection of fake and fraudulent faces via neural memory networks. *IEEE Trans. Inf. Forensics Secur.* **2020**, *16*, 1973–1988. [CrossRef]
18. Xu, Y.; Jia, G.; Huang, H.; Duan, J.; He, R. Visual-Semantic Transformer for Face Forgery Detection. In Proceedings of the 2021 IEEE International Joint Conference on Biometrics (IJCB), Shenzhen, China, 4–7 August 2021; pp. 1–7.
19. Fridrich, J.; Kodovsky, J. Rich models for steganalysis of digital images. *IEEE Trans. Inf. Forensics Secur.* **2012**, *7*, 868–882. [CrossRef]
20. Cozzolino, D.; Poggi, G.; Verdoliva, L. Recasting residual-based local descriptors as convolutional neural networks: An application to image forgery detection. In Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, Philadelphia, PA, USA, 20–22 June 2017; pp. 159–164.
21. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 6105–6114.
22. Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. Mesonet: A compact facial video forgery detection network. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 11–13 December 2018; pp. 1–7.
23. Güera, D.; Delp, E.J. Deepfake video detection using recurrent neural networks. In Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 27–30 November 2018; pp. 1–6.
24. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
25. Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; Shao, J. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 86–103.
26. Luo, Y.; Zhang, Y.; Yan, J.; Liu, W. Generalizing face forgery detection with high-frequency features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 16317–16326.
27. Chen, H.S.; Rouhsedaghat, M.; Ghani, H.; Hu, S.; You, S.; Kuo CC, J. Defakehop: A light-weight high-performance deepfake detector. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; pp. 1–6.
28. Chen, S.; Yao, T.; Chen, Y.; Ding, S.; Li, J.; Ji, R. Local relation learning for face forgery detection. *AAAI Conf. Artif. Intell.* **2021**, *35*, 1081–1088. [CrossRef]
29. Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; Yu, N. Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2185–2194.
30. Dlib. Dlib C++ Library [EB/OL]. Available online: http://dlib.net/ (accessed on 14 September 2022).
31. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8789–8797.
32. Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 17 October–2 November 2019; pp. 1–11.
33. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-df: A large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3207–3216.
34. Dolhansky, B.; Howes, R.; Pflaum, B.; Baram, N.; Ferrer, C.C. The deepfake detection challenge (dfdc) preview dataset. *arXiv* **2019**, arXiv:1910.08854.
35. Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; Ferrer, C.C. The Deepfake Detection Challenge (DFDC) Dataset. *arXiv* **2020**, arXiv:2006.07397.
36. Zhao, T.; Xu, X.; Xu, M.; Ding, H.; Xiong, Y.; Xia, W. Learning self-consistency for deepfake detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 15023–15033.