DFGC-VRA: DeepFake Game Competition on Visual Realism Assessment

Organizers

Bo Peng[1], Xianyun Sun[2], Caiyong Wang[2], Wei Wang[1], Jing Dong[1], Zhenan Sun[1]*
and Competition Teams
Rongyu Zhang[3], Heng Cong[3], Lingzhi Fu[3], Hao Wang[3], Yusheng Zhang[3], Hanyuan Zhang[4], Xin Zhang[4], Boyuan Liu[4], Hefei Ling[4], Luka Dragar[5], Borut Batagelj[5], Peter Peer[5], Vitomir Štruc[6], Xinghui Zhou[7], Kunlin Liu[7], Weitao Feng[7], Weiming Zhang[7], Haitao Wang[8], Wenxiu Diao[9]

[1]Center for Research on Intelligent Perception and Computing (CRIPAC), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China.
[2]Beijing University of Civil Engineering and Architecture (BUCEA), China.
[3]Interactive Entertainment Group of Netease Inc, Guangzhou, China.
[4]Huazhong University of Science and Technology (HUST), China.
[5]Faculty of Computer and Information Science, University of Ljubljana (UNILJ), Slovenia.
[6]Faculty of Electrical Engineering, University of Ljubljana (UNILJ), Slovenia.
[7]University of Science and Technology of China (USTC), China.
[8]Institute of Network Technology (INT), Yantai, China.
[9]Nanjing University of Science and Technology (NUST), China.

Abstract

This paper presents the summary report on the DeepFake Game Competition on Visual Realism Assessment (DFGC-VRA). Deep-learning based face-swap videos, also known as deepfakes, are becoming more and more realistic and deceiving. The malicious usage of these face-swap videos has caused wide concerns. There is a ongoing deepfake game between its creators and detectors, with the human in the loop. The research community has been focusing on the automatic detection of these fake videos, but the assessment of their visual realism, as perceived by human eyes, is still an unexplored dimension. Visual realism assessment, or VRA, is essential for assessing the potential impact that may be brought by a specific face-swap video, and it is also useful as a quality metric to compare different face-swap methods. This is the third edition of DFGC competitions, which focuses on the new visual realism assessment topic, different from previous ones that compete creators versus detectors. With this competition, we conduct a comprehensive study of the SOTA performance on the new task. We also release our MindSpore codes to further facilitate research in this field (`https://github.com/bomb2peng/DFGC-VRA-benckmark`).

*Jing Dong (jdong@nlpr.ia.ac.cn) is the corresponding author.

1. Introduction

The word "deepfake" arises around the end of 2017 from a group of Reddit users named "deepfakes" who share deepfakes they created involving celebrities' faces swapped onto the bodies of actresses in pornographic videos. The fake facial images and videos created by deep-learning models are called deepfakes for short, which can also include face reenactment, facial attribute editing, and completely generated faces. Recently, with the new wave of AI-generated contents, or AIGC, which are created by large-scale pre-trained text-to-image generation models, deepfakes can also be seen as a special case of AIGC. Broad concerns about the negative usages of deepfake and AIGC call for the research on automatic detection of these contents. And there has been a lot of works focusing on the detection topic.

As deepfakes are ultimately received and consumed by human viewers, the perceptual quality or visual realism of deepfakes plays an important role in determining their po-

tential social impacts. We note that VRA is basically the same as the quality assessment of deepfake videos, but focusing on the visual realism aspect instead of image sharpness or lossy compression. Apart from evaluating potential social impacts, the deepfake VRA task can also provide meaningful quality evaluation metrics for generated images and generative models [10, 29]. Compared to the objective side of deepfake detection, the deepfake VRA is the subjective side of the coin, but it is a quite under-explored problem compared with deepfake detection and still demands more research attentions.

With the hope to draw more attention to the deepfake VRA task, we organize this new DFGC-VRA competition after our previous DFGC-2021 [23] and DFGC-2022 [25] competitions. The current competition is related to its precedents in that it uses the deepfake video dataset collected from the last year, but it has different topics from the previous DFGC's. Our previous competitions focus on deepfake detection, deepfake creation, and their adversarial game, while this competition put more stress on the new visual realism assessment problem. Although without the cash prize as provided in our previous competitions, the current event still draws relatively wide attentions from the community. In total, we received 54 registration applications, we approved 28 of them that comply with our policy, and 15 teams submitted to our leaderboard. In this paper, we summarize the DFGC-VRA competition: its related previous work, the competition dataset, evaluation metrics, competition procedure, and overall results. Top solutions are also reported and discussed to show effective methods for this new task.

We summarize the contributions of this work as follows:

- The DFGC-VRA is the first competition that focuses on the visual realism assessment task for face-swap deepfake videos.
- We summarize the top solutions used in this competition and compare their technical traits.
- We open-source our re-implementation of top solutions using the MindSpore framework that can be used as a benchmark for future deepfake VRA work.

2. Related Work

Deepfake Detection

Deepfake detection aims at distinguishing whether a face image/video is deepfake or real. With the releasement of recent benchmarks and large datasets [14, 36, 16], deepfake detection models have obtained better performances, by employing self-supervised data augmentations [27], stronger Transformer models [34], audio-visual multimodalities [33] etc.. However, they still struggle in generalizing to detecting unseen deepfake methods, and the lack of explainability also hinders their real-world usage [24] in law enforcement or court of law.

To promote the progress of deepfake detection methods for real-world applications, many competitions are proposed. Some well-known ones include the DeepFake Detection Challenge (DFDC) [6] in 2020, the ForgeryNet Challenge [12] in 2021, and the OpenMFC series [11] from 2020 to 2022. Different from these competitions, where datasets are all pre-constructed by only the organizers, our DeepFake Game Competition (DFGC) series [23, 25] draws challenging deepfake data from all the participants that act as counterpart players to the detection side. Hence, its datasets tend to be more diverse and hard to be detected, which better mimic the real-world scenario.

Through the testing of deepfake detection competitions, several effective techniques for improving the detection performance are noted. Some data augmentation skills like regional erasing [1, 35] and blending [2, 27] are helpful for improving generalization ability. Ensemble of multiple advanced high-capacity models improves the stability of detection performance [3]. Although deepfake detection is a different task from deepfake realism assessment, we have found both through this competition and our previous study [28] that they are closely related, in that powerful detection models are also helpful for predicting visual realism.

Image and Video Quality Assessment

Image and video quality assessment, i.e., IQA and VQA, are classical research topics in image processing and multimedia community. They aim at automatically assessing the visual quality of images and videos when they go through some degradation processes, typically lossy compressions and network streaming. We only introduce some no-reference (NR) IQA/VQA methods here as they are the most related to our setting. Many classical IQA methods are based on the Natural Scene Statistics (NSS) model and design hand-crafted features, e.g., BRISQUE [22] and FRIQUEE [9]. Classical VQA also includes statistical features of the motion information, e.g., TL-VQM [15]. Deep-learning based IQA/VQA methods become popular with the end-to-end feature learning ability. RankIQA [19] employs the Siamese network and ranking loss to train on pairs of images and address the problem of limited size of datasets. FastVQA [31] proposes Grid Mini-patch Sampling (GMS) and Fragment Attention Network (FANet) to reduce the computational cost that hinders end-to-end VQA model training.

Traditional IQA/VQA works mainly focus on natural scene images and videos, but there is very little work on the quality assessment of the new form of generated media, e.g., GAN images and deepfake videos. GAN models are commonly evaluated using the Frechet Inception Distance (FID) metric, which measures the distance between real and fake image feature distributions. However, FID cannot indi-

cate the visual quality of each individual image. GIQA [10] addresses this problem by proposing several models for predicting the quality of individual GAN images, with the best model being a Gaussian Mixture Model. The work [29] proposes generalized visual quality assessment for face images generated by various GANs, employing meta-learning and pair-wise ranking on pseudo quality scores to mitigate overfitting. Based on another pseudo score guided learning-based model in GIQA [10], the work [36] uses a Q-Net trained on GAN face images to do quality assessment and quality control in deepfake dataset construction. However, this method is designed and trained on GAN face images, which may not be suitable for assessing the face-swap images and videos.

User studies on deepfake visual realism and human performance on deepfake detection are carried out in recent works [16, 24], across different datasets. These evaluations all use no-reference settings. Notably, our DFGC-2022 [25] conducted user study on the competition dataset with some kind of reference to help making the quality annotations more precise. This is because the dataset has multiple deepfake videos for the same source-target pair, which are created by the participants using different face-swap methods. Our raters evaluate these corresponding deepfake videos in a sequence, and hence comparisons between high and low qualities can be established more informatively. In our primary deepfake VRA work [28], we find that the visual realism assessment task is closely related to the deepfake detection task, in that the features of pretrained deepfake detection models are very useful in predicting visual realism. The method in this previous work also acts as the baseline in this competition.

3. Dataset

This competition uses the dataset from our last year competition DFGC-2022 [25], which was created using various face-swap methods and has diversified degrees of visual realism. More specifically, it contains face-swap videos for 20 pairs of subjects (IDs), and the total number of deepfake creation methods is 35. Each video is about 5 seconds in duration and has 1920x1080 resolution. Based on this dataset, 1,400 deepfake videos have been annotated with Mean Opinion Scores (MOS) on the visual realism, and they are obtained by averaging the ratings of 5 human raters. During the annotation, each human rater views the full-length video and then gives a score in 5 scales: 1 (very bad) to 5 (very good). As detailed in the related work, the annotation process has some kind of reference and hence more accurate. More details on the dataset are in [25, 28].

In the competition protocol, we divide the 1,400 annotated videos into the train set, test-1 set (ID-disjoint with the train set), test-2 set (method-disjoint), and test-3 set (ID&method-disjoint), as illustrated in Fig. 1. The sets

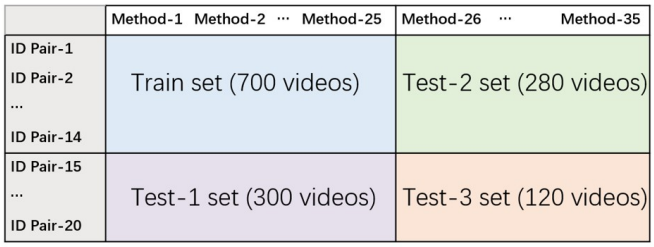

Figure 1: Competition protocol for dataset splits.

contain 700, 300, 280, and 120 video samples, respectively. Participants have access to both videos and groundtruth MOSs of the train set, while the test sets are provided without the groundtruth MOSs. The participants are required to train their VRA models using the provided train set, obtain the predicted MOSs on the test sets, and submit their prediction results to the competition platform [1] for automatic performance evaluation and ranking. We require checking codes after the submission phase to validate the results.

4. Metrics

We use two metrics from the image and video quality assessment literature: Pearson Linear Correlation Coefficient (PLCC) and Spearman's Rank-order Correlation Coefficient (SRCC) to evaluate the prediction linearity and monotonicity with respect to the groundtruth. Since the two metrics are both in the range of [-1, 1], higher the better, we average them to obtain the overall performance metric. The performances on the three test sets are equally weighted and averaged. In summary, the score for a submission is calculated with Equation (1):

$$s = (plcc_1 + srcc_1 + plcc_2 + srcc_2 + plcc_3 + srcc_3)/6 \quad (1)$$

where $plcc_1$ represents the PLCC score on the test-1 set and so on for the other variables.

5. Competition Procedure

The competition procedure can be separated into three phases: registration, submission, and checking. During registration, each team need to first register an account on the Codalab platform and then send an email to the organizers with their institutional email containing information of their members, affiliation, and adviser. Only after we manually checked the registration information, can the team be accepted into this competition. By registering to this competition, the team member are bounded by the terms for using the dataset and also by the competition rules.

During the submission phase, which lasts for one and a half months (Mar. 1-Apr. 15), each team can make up

[1] https://codalab.lisn.upsaclay.fr/competitions/10754#learn_the_details

Table 1: Overview of competition results. LB stands for leaderboard results obtained by re-running the inference code and the teams' model checkpoints, and RP stands for reproduced results by re-running the training/fine-tuning codes by organizers.

Team	LB	RP	Backbone	Pre-train	Fine-tune	Loss	Inference
OPDAI	0.8851	0.8825	Swin-transformer	DFDC Det	DFGC-22	Norm-in-norm, KL divergence	3 frames score fusion
HUST	0.8564	0.8474	ConvNeXt, LSTM	ImageNet	DFGC-22 & extra data	MAE, rank, PLCC	20 frames score fusion, 5 models ensemble
UNILJ	0.8545	0.8501	ConvNeXt, Eva	Deepfake Det, ImageNet	DFGC-22	RMSE	10 clips score fusion, 2 models ensemble
USTC	0.8360	0.8116	ResNet152	self-collected face-swap data	DFGC-22	rank	8 frames score fusion
INT&NUST	0.8257	0.8146	ResNext-Transformer hybrid	ImageNet	DFGC-22	PLCC, MSE	4 frames score fusion, 2 streams fusion
Baseline [28]	0.5470	0.5470	ConvNeXt & Swin-transformer	Deepfake Det	SVR on DFGC-22	MSE	regression on video feature

to one submission to the competition platform and see immediate feed-backs from the leaderboard. The participants run their inference methods to obtain the predictions on the three test sets, which are saved to txt files and then submitted for evaluation. No private sharing between different teams is allowed. One team using multiple submission accounts is not allowed. Participants should not hand-label the released testing data for submission or training. However, they are permitted to use extra datasets for training.

The checking phase begins after the submission closes, during which the participants are required to send their codes and technical reports to the organizers for checking. Only after we can reproduce the leaderboard results within a tolerance level, can their results be officially recognized as valid. We then announce the Top-3 winner teams and their results. We also invite some participants to collaborate on this summary paper if their solutions have high scores or use inspiring methods.

6. Results

Our competition platform received 54 registration requirements, and we approved 28 of them that comply with our registration policy. Over the submission period, 15 teams submitted at least once to the leaderboard, including the organizers' baseline submission. There were 10 teams who achieved higher scores than the baseline [28] score 0.5470. After the submission is closed, there were 7 teams that submitted codes and technical reports for checking. We report the top-5 teams, their results, and methods overview in Table 1. More details of their solutions are described in the next section.

As can be seen from Table 1, there are discrepancies between the reproduced results (RP) and their leaderboard results (LB) for all teams. Most discrepancies are quite small and may be due to the randomness and software/hardware differences between the organizers' and participants' machines. The leaderboard results are all better than the reproduced results, since the submission procedure tends to select out the most over-fitted randomness. The relatively large discrepancy in our reproduced result of the USTC solution is due to the Apex library incompatibility problem on our machine.

All methods in Table 1 use advanced large models that have high capacities and are pretrained on ImageNet or deepfake detection (Det) datasets to mitigate over-fitting. The models (apart from the baseline) are then fine-tuned on the DFGC-22 VRA dataset, using either cropped frames or clips (a group of frames) as inputs. The training losses range from traditional root of mean squared error (RMSE) loss to ranking losses proposed in the image quality assessment literature. For inference, most methods fuse the scores of multiple frames/clips from a testing video, and some methods also ensemble multiple models' results. More details of these solutions are described in the next section. We note the large improvements of top solutions over the baseline in Table 1 and conjecture that fine-tuning the model end-to-end may play a key role.

7. Top Solutions

As a baseline method, in [28] we propose to employ the DFGC-2022 first-place deepfake detection model [3] as a fixed feature extractor and train a Support Vector Regression model (SVR) to predict the realism score. The method first extracts per-frame features and calculates their mean and standard deviation as the video-level feature. The SVR takes video features as input and predict the MOS scores. More details can be found in [28]. In the following, we mainly describe the participants' solutions.

7.1. OPDAI

The team OPDAI is from the Interactive Entertainment Group of Netease Inc.

The model used is the Swin-transformer v2 (`swinv2_large_window12to16_192to256_22kft1k`) [20]. It is first pretrained on the DFDC dataset [6] for deepfake detection, and then finetuned on the competition training data (DFGC-22). For data pre-processing, each video is extracted with 10 frames, and face detection and cropping is conducted. Some data augmentation operations are used, including random erasing, random horizontal flip, random color adjustment, random contrast adjustment.

The pretraining on the DFDC deepfake detection task uses the MSE loss. For the finetuning on the competition VRA task, two losses are used, i.e., the Norm-in-norm loss [17] for image quality assessment and the KL-divergence loss. The Norm-in-norm loss uses normalization to speed-up convergence and to encourage linear predictions with respect to groundtruth scores. Given label Q and prediction $\hat{Q}$, the Norm-in-norm loss is defined as:

$$L_{NIN}(Q,\hat{Q}) = \sum_{i=1}^{N} \left| \hat{S}_i - S_i \right| \tag{2}$$

$$S_i = \frac{Q_i - \frac{1}{N}\sum_{i=1}^{N} Q_i}{(\sum_{i=1}^{N} \left| Q_i - \frac{1}{N}\sum_{i=1}^{N} Q_i \right|^q)^{\frac{1}{q}}} \tag{3}$$

where $\hat{S}_i$ can be calculated similarly with S_i in Equation (3) that are normalized versions of original scores, and the parameter q is set to 2. The KL-divergence loss is defined as:

$$L_{KLD}(Q,\hat{Q}) = \sum_{i=1}^{N} \hat{W}_i \times \log \frac{\hat{W}_i}{W_i} \tag{4}$$

$$W_i = \frac{\exp(Q_i)}{\sum_{i=1}^{N} \exp(Q_i)} \tag{5}$$

where $\hat{W}_i$ can be calculated similarly with W_i in Equation (5) that are Softmax-normalized versions of original scores. Finally, the total loss is the sum of the two losses:

$$L(Q,\hat{Q}) = L_{NIN}(Q,\hat{Q}) + L_{KLD}(Q,\hat{Q}) \tag{6}$$

The released training dataset is randomly divided into 600 videos for training and 100 videos for validation. In each training epoch, a random frame from each video is selected for gradient updates. Drop path and data augmentations are used to alleviate overfitting. Learning rate warm-up is also used for automatically selecting appropriate learning rates. The best model on the validation set is used for submission. For inference, three frames at the 0.25, 0.5, and 0.75 positional points of a video are used for prediction and then averaged. Test time augmentation based on left-right flipping is also adopted.

7.2. HUST

Apart from the 700 deepfake videos in the competition training set, this team also used 2 extra datasets with their own labeled MOS for training. The extra-1 dataset contains 119 deepfake videos, including 85 videos from the FaceForensics++ dataset [26] and 34 videos from the Celeb-DF v2 dataset [18]. The extra-2 dataset contains 90 deepfake videos, including 61 from the FaceForensics++ and 29 from the Celeb-DF v2. Each video is extracted with 20 frames and cropped around faces. Data augmentation includes Resize, HorizontalFlip, ShiftScaleRotate and RandomRotate90 from the Albumentation library.

Five base models are trained for ensemble, which are all based on the ConvNeXt model [21] pretrained on the ImageNet dataset (`convnext_tiny_384_in22ft1k`). Detailed training strategy of these models are shown in Table 2. Apart from the Model-4, all models use ImageNet-pretrained ConvNeXt and fine-tune on the train set and extra sets, and the video score is obtained by averaging 20 frame scores. For the Model-4, it uses the fixed Model-2 to extract frame features and trains a two layer LSTM model on top of the frame features to predict video scores. For result submission, the team also used a trick that model ensemble weights are separately tuned for each of the three test sets.

Table 2: Training strategies of the 5 base models in the HUST solution. "Train" means the competition train set.

ID	Backbone	Training Data	BatchSize	Iters	Pre-trained
1	ConvNeXt	80% Train + Extra-1	32	3000	ImageNet
2	ConvNeXt	80% Train + Extra-1	32	5500	ImageNet
3	ConvNeXt	100% Train + Extra-1 + Extra-2	32	9000	ImageNet
4	LSTM + ConvNeXt	100% Train + Extra-1 + Extra-2	32	540	Model-2
5	ConvNeXt	5-fold Train	48	best of each fold	ImageNet

The training loss is a combination of three terms:

$$L = L_{MAE} + \alpha \cdot L_{PLCC} + \beta \cdot L_{rank} \tag{7}$$

where $\alpha = 0.5$ and $\beta = 1$. The first part is the Mean Absolute Error (MAE) loss, also known as the L1 loss, which is not so sensitive to outliers compared to the L2 loss. The second part is the PLCC loss, since PLCC is one of the competition metric and is also a differentiable function. It is defined as:

$$L_{PLCC} = 1 - abs(PLCC(Q,\hat{Q})) \tag{8}$$

The third part is the pair-wise ranking loss [30, 19], which pulls the estimated quality difference of two images closer

to the margin. It is defined as:

$$L_{rank}^{ij} = \max(0, margin - e(Q_i, Q_j) \cdot (\hat{Q}_i - \hat{Q}_j)) \quad (9)$$

$$margin = |Q_i - Q_j| \quad (10)$$

$$e(Q_i, Q_j) = \begin{cases} 1, Q_i \geq Q_j \\ -1, Q_i < Q_j \end{cases} \quad (11)$$

7.3. UNILJ

For complete details about this method, we refer to this team's new work [7], and we only describe the overview of the solution here. Two models are trained and ensembled in this method, which are the ConvNeXt model [21] (`convnext_xlarge_384_in22ft1k`) and a scaled-up Vision Transformer, i.e., the Eva model [8] `eva_large_patch14_336.in22k_ft_in22k_in1k`. The ConvNeXt model is initialized with the last year's DFGC-2022 competition first-place deepfake detection model weights [4], which was trained on a collection of 9 deepfake datasets. The Eva model is initialized with its pretrained weights in the timm library.

In data pre-processing, faces are detected and cropped for every frame in the video. Considering the temporal nature of videos, 5 consecutive frames from a randomly selected starting point is selected as a clip and input to the model. Each frame separately goes through the model to obtain 5 feature vectors. The mean and standard deviation of these extracted features are then concatenated and fed to several fully connected layers to output the predicted MOS. Here, the usage of mean and standard deviation of frame features to represent a clip is similar to the baseline method [28]. The training loss for both models is the Root of Mean Squared Error (RMSE) loss. The AdamW optimizer is used, the initial learning rate is 2e-5 and reduced on plateau. The provided training set is randomly divided to train (70%), validation (10%) and test (20%) sets for parameter tuning, and the final model is then trained again on the whole training set with early stopping.

For inference, the ensemble weights for the ConvNeXt and the Eva model are 0.75 and 0.25, respectively. The video prediction of each model is the average of predictions for 10 clips randomly selected from the testing video.

7.4. USTC

To mitigate over-fitting on the provided training set, the ResNet152 backbone model is firstly pretrained on a self-collected face-swap dataset for the quality ranking task. The self-collected dataset is augmented from the CelebA-HQ dataset using face swap methods such as SimSwap, FaceShifter, InfoSwap, and MegaFS. The pseudo-groundtruth of quality labels is obtained based on the similarity of face-swap images to the target images in terms of pose, expression, etc., where higher quality face-swap images are deemed to have higher similarity in these aspects. Pairs of images are then fed into the Siamese network, and the same ranking loss as in Equation (9) is employed.

The team tried to use as much data as possible for training and incorporated the released test videos into the training procedure. Also the test set labels are hidden, this may be thought of as a leak in our competition design. The last-round trained model is used to predict the realism scores on the test videos, and they are then used as the pseudo-groundtruth labels for the next round training. The training loss for fine-tuning on the competition dataset is similar to that used in pretraining, except that an extra inaccuracy term is added to the margin in consideration of the inaccurate test set pseudo-label. More formally, Equation (10) is adapted to the following:

$$margin = |Q_i - Q_j| + k \cdot M_{inacc} \quad (12)$$

where $k = 0, 1, 2$ is the number of images in the pair that come from the test set. The inaccuracy term M_{inacc} is estimated as the maximum prediction error on the train set. For final inference of video scores, this method uses the average image scores of 8 frames extracted from the video. For the

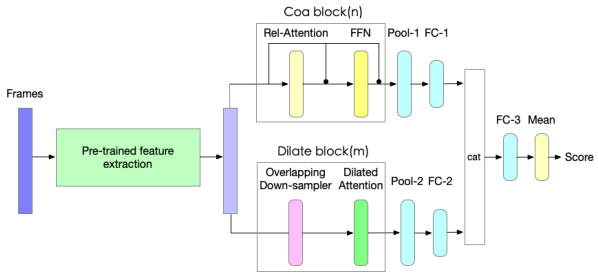

Figure 2: Model structure of the INT&NUST team.

training of the model on the competition dataset, PLCC loss and MSE loss are used.

$$L = w_1 \times L_{PLCC} + w_2 \times L_{MSE} \quad (13)$$

where the weights w_1 and w_2 are set to 1 and 0.001, respectively. The LION optimizer is used for optimization, and learning rate warm up is also used.

7.5. INT&NUST

The model structure of this method is shown in Fig. 2. The input data is a clip that includes cropped faces from 4 random frames of a video. The feature extraction module, which is the ResNext-101 model [32] pretrained on ImageNet, extracts feature maps for each frame. The frame feature map is then input to a dual-stream both designed for score prediction (FC-1 and FC-2), the frame score is then obtained by regression on the two stream scores (FC-3), and the final clip score is obtained by taking the mean over frames (Mean). The first Coa-block stream is adapted from

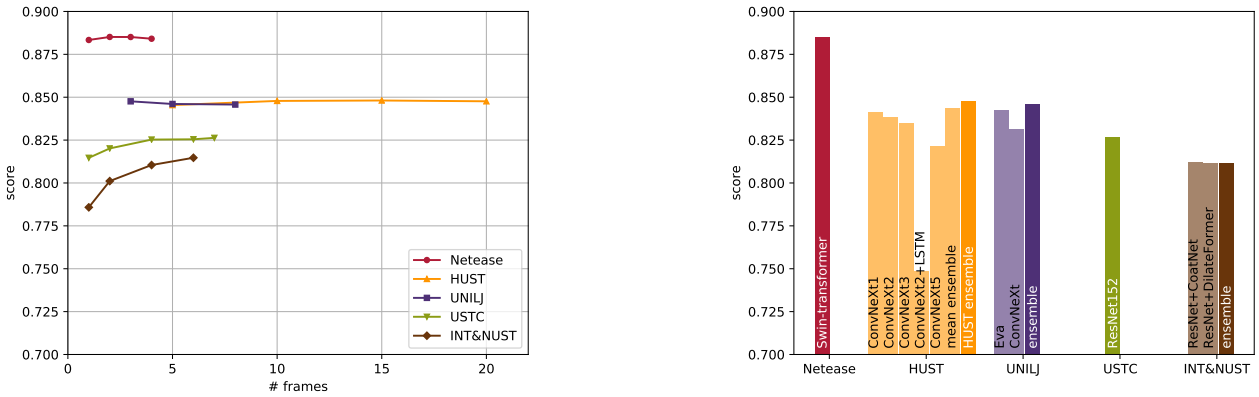

Figure 3: Analysis of ensemble strategies of different number of frames/clips (left) and different models (right).

the self-attention module of the CoAtNet [5], which combines CNNs and Transformers. The second Dilate-block stream is from the Dilateformer [13], where the Sliding Window Dilated Attention (SWDA) is used to achieve locality and sparsity, improving the Vision Transformers. The number of blocks in the two streams is $n = 2$ and $m = [4, 2]$ which means 4 dilate blocks + 2 normal transformer blocks.

8. More Analysis on Results

As many approaches use either multiple input frames/clips for score fusion or use multiple models in ensemble,we conduct an analysis of ensemble strategies in inference. It includes the fusion of different number of frames/clips and the ensemble of different models in each solution. The comparison results can be seen in Fig. 3. For the analysis of different frames/clips, under each number of frames, five runs of randomly selected frames are obtained and averaged. The results indicate that relatively lower-performing models benefit more from using larger numbers of frames for ensemble, e.g., the INT&NUST method and the USTC method. Relatively higher-performing models do not have obvious gains from using more frames.

For the analysis of different models, the teams HUST, UNILJ, and INT&NUST employ multiple models, and the former two teams obtain higher scores with the model ensemble, although the improvements are marginal, while the INT&NUST team does not obtain improvements from the ensemble. We test two ensemble strategies for the HUST models, where the "mean" ensemble is the normal way of directly taking the mean of the five models and the "HUST" ensemble is using their original trick of separately tuning weights for each of the three test sets. The "HUST" ensemble achieves a higher score compared to the "mean" ensemble, but it is tuned particularly on these test sets and not a common practice.

We also analyze the results on each separate test sets as they have different disjoints with the train set, as shown in Fig. 1. We first calculate the average score of all teams' leader-board results, and they are $0.5533, 0.7585, 0.6200$ for the Test-1, Test-2, and Test-3 sets, respectively. The raw data for this can be found on the competition results page [2]. We also calculate the average from only the top-5 teams' results, and they are $0.7857, 0.9214, 0.8547$, respectively. The relative orders obtained by these two ways are the same, i.e., the Test-1 set is the most hard to predict, followed by the Test-3 set, and the Test-2 set is the most easy one. This shows that, for our competition data protocol, the VRA of unseen faces is more challenging than that of unseen creation methods. Comparing the results on Test-1 and Test-3, which are both unseen-ID test sets, it shows that the influence of unseen creation methods is not significant.

9. Conclusions

In this paper, we present the summary report of our DFGC-VRA competition, which is the first competition on the visual realism assessment of face-swap deepfake videos. It draws wide attentions from the deepfake detection research community. We received 15 valid submissions from research teams across the world, among which 10 submissions surpassed our provided baseline approach.

The top five solutions are discussed in detail in this paper. They typically use recent large deep-learning models like vision transformers and ConvNeXt pretrained on deepfake detection dataset or ImageNet. The pretrained models are then finetuned on the competition training set using frames or clips as input, employing effective losses borrowed from the image and video quality assessment literature. Their im-

[2] `https://codalab.lisn.upsaclay.fr/competitions/10754#results` and click "Test".

provements over the baseline points to the effectiveness of end-to-end finetuning. The video prediction scores are then obtained by fusing the scores of multiple frames or clips and some also train multiple models in ensemble. Further analyses show that these ensemble strategies have some improvements on some result but are not always effective for all solutions. We are also working on re-implementing these methods using a common framework, and it can be used as a common benchmark and starting point for future research.

There are remaining work to be done in the future, which includes more extensive analysis of the effective component in existing VRA methods, validation of the methods on larger and more diverse deepfake datasets, and exploring the application of VRA on evaluating and improving the quality of generated videos. More research efforts on the deepfake VRA task are needed, and the quality assessment of new AIGC contents are also in the scope of future work.

Acknowledgements

We thank all the participants for their time and efforts in making this competition.

This work is supported by the National Key Research and Development Program of China under Grant No. 2021YFC3320103, the National Natural Science Foundation of China (NSFC) under Grants 62272460, U19B2038, 61972395, 62106015, a grant from Young Elite Scientists Sponsorship Program by CAST (YESS), CAAI-Huawei MindSpore Open Fund, the Pyramid Talent Training Project of BUCEA (No. JDYC20220819), and the BUCEA Post Graduate Innovation Project (PG2023090).

References

[1] DFDC 1st Place Solution. `https://github.com/selimsef/dfdc_deepfake_challenge`.

[2] DFGC-2021 1st Place Solution. `https://github.com/beibuwandeluori/DFGC_Detection`.

[3] DFGC-2022 1st Place Solution. `https://github.com/chenhanch/DFGC-2022-1st-place`.

[4] DFGC-2022 first-place solution of the detection track. https://github.com/chenhanch/DFGC-2022-1st-place.

[5] Z. Dai, H. Liu, Q. V. Le, and M. Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021.

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

[7] L. Dragar, B. Batagelj, P. Peer, and V. Štruc. Beyond detection: Visual realism assessment of deepfakes. In *Proceedings of the 32nd International Electrotechnical and Computer Science Conference (under review)*, 2023.

[8] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022.

[9] D. Ghadiyaram and A. C. Bovik. Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of vision*, 17(1):32–32, 2017.

[10] S. Gu, J. Bao, D. Chen, and F. Wen. Giqa: Generated image quality assessment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 369–385. Springer, 2020.

[11] H. Guan, Y. Lee, L. Diduch, J. Zhang, I. G. Bajgiran, T. Kheyrkhah, P. Fontana, and J. G. Fiscus. Open media forensics challenge (openmfc) 2020-2021: Past, present, and future. 2021.

[12] Y. He, L. Sheng, J. Shao, Z. Liu, Z. Zou, Z. Guo, S. Jiang, C. Sun, G. Zhang, K. Wang, et al. Forgerynet–face forgery analysis challenge 2021: Methods and results. *arXiv preprint arXiv:2112.08325*, 2021.

[13] J. Jiao, Y.-M. Tang, K.-Y. Lin, Y. Gao, J. Ma, Y. Wang, and W.-S. Zheng. Dilateformer: Multi-scale dilated transformer for visual recognition. *IEEE Transactions on Multimedia*, 2023.

[14] H. Khalid, S. Tariq, M. Kim, and S. S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021.

[15] J. Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28(12):5923–5938, 2019.

[16] T.-N. Le, H. H. Nguyen, J. Yamagishi, and I. Echizen. Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10117–10127, 2021.

[17] D. Li, T. Jiang, and M. Jiang. Norm-in-norm loss with faster convergence and better performance for image quality assessment. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 789–797, 2020.

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

[19] X. Liu, J. Van De Weijer, and A. D. Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE international conference on computer vision*, pages 1040–1049, 2017.

[20] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022.

[21] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[22] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

[23] B. Peng, H. Fan, W. Wang, J. Dong, Y. Li, S. Lyu, Q. Li, Z. Sun, H. Chen, B. Chen, et al. Dfgc 2021: A deepfake game competition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021.

[24] B. Peng, S. Lyu, W. Wang, and J. Dong. Counterfactual image enhancement for explanation of face swap deepfakes. In *Pattern Recognition and Computer Vision: 5th Chinese Conference, PRCV 2022, Shenzhen, China, November 4–7, 2022, Proceedings, Part II*, pages 492–508. Springer, 2022.

[25] B. Peng, W. Xiang, Y. Jiang, W. Wang, J. Dong, Z. Sun, Z. Lei, and S. Lyu. Dfgc 2022: The second deepfake game competition. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2022.

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

[27] K. Shiohara and T. Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022.

[28] X. Sun, B. Dong, C. Wang, B. Peng, and J. Dong. Visual realism assessment for face-swap videos. *to appear at International Conference on Image and Graphics (ICIG 2023), arXiv preprint arXiv:2302.00918*, 2023.

[29] Y. Tian, Z. Ni, B. Chen, S. Wang, H. Wang, and S. Kwong. Generalized visual quality assessment of gan-generated face images. *arXiv preprint arXiv:2201.11975*, 2022.

[30] S. Wen and J. Wang. A strong baseline for image and video quality assessment. *arXiv preprint arXiv:2111.07104*, 2021.

[31] H. Wu, C. Chen, J. Hou, L. Liao, A. Wang, W. Sun, Q. Yan, and W. Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 538–554. Springer, 2022.

[32] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[33] W. Yang, X. Zhou, Z. Chen, B. Guo, Z. Ba, Z. Xia, X. Cao, and K. Ren. Avoid-df: Audio-visual joint learning for detecting deepfake. *IEEE Transactions on Information Forensics and Security*, 18:2015–2029, 2023.

[34] C. Zhao, C. Wang, G. Hu, H. Chen, C. Liu, and J. Tang. Istvt: interpretable spatial-temporal video transformer for deepfake detection. *IEEE Transactions on Information Forensics and Security*, 18:1335–1348, 2023.

[35] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021.

[36] T. Zhou, W. Wang, Z. Liang, and J. Shen. Face forensics in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5778–5788, 2021.