Beyond Detection: Visual Realism Assessment of Deepfakes

Luka Dragar[1], **Peter Peer**[1], **Vitomir Štruc**[2], **Borut Batagelj**[1]

[1]*Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia*
[2]*Faculty of Electrical Engineering, University of Ljubljana, Tržaška cesta 25, 1000 Ljubljana, Slovenia*
E-mail: luka.dragar3@gmail.com

Abstract

In the era of rapid digitalization and artificial intelligence advancements, the development of DeepFake technology has posed significant security and privacy concerns. While a considerable amount of work has been done on DeepFake detection and generation techniques, assessing the quality and visual realism of DeepFakes is still underexplored, despite the fact that this is key for the impact a forged video can have in practice. In this paper, we address this gap and present an effective approach for quantifying the visual realism of DeepFake videos. We utilize an ensemble of two Convolutional Neural Network (CNN) models, i.e., Eva and ConvNext, trained on the DeepFake Game Competition (DFGC) 2022 dataset to regress to Mean Opinion Scores (MOS) from DeepFake videos based on features extracted from a sequence of frames. Our method secured the third place in the recent DFGC on Visual Realism Assessment held in conjunction with the 2023 International Joint Conference on Biometrics (IJCB 2023). We provide an overview of the models, data preprocessing, and training procedures. We also report the performance of our models against the competition's baseline model and discuss the main findings.

1 Introduction

DeepFakes have recently emerged as a significant threat to the integrity of digital media. These AI-generated manipulated imagery, commonly containing human faces, represents highly realistic counterfeit content that is increasingly challenging to distinguish from authentic media. This development raises major security and privacy concerns, necessitating effective measures for automatic detection of DeepFakes, but also techniques for assessing the visual realism of the manipulated data. The latter is especially important, since the quality and visual realism of the generated DeepFakes is strongly correlated with the impact a given falsified video can have. However, despite this importance, effective techniques for assessing the visual realism and quality of the generated DeepFakes are still largely missing from the literature.

In this paper, we therefore present a novel approach for predicting the realism of manipulated videos that utilizes the complementary information extracted from the input data by two deep learning models. We train the two models in a regression framework on data from the recent DeepFake Game Competition (DFGC) on Visual Realism Assessment with encouraging results, as also illustrated in Figure 1, and show that our ensemble approach yields highly competitive overall performance.

Figure 1: Face-swap videos with different degrees of realism, annotated with the ground truth Mean Opinion Score MOS (gt) vs. predicted MOS (pred) by our ensemble model.

2 Related work

Deepfake detection, face morphing attack detection, and realism assessment have been a critical focus of research in recent years due to the rapid development of deepfake technologies [2, 3]. Initial studies primarily targeted deepfake detection, aiming to differentiate authentic videos from AI-manipulated ones. With the advent of deep learning, detection methods have seen significant improvements. However, disparities between human and machine perception of deepfakes, as demonstrated by Korshunov and Marcel [4], suggest further research is necessary in this field.

Another key study [5] highlighted the effectiveness of combining human judgement and machine learning models for deepfake detection. The study found that this combination yielded superior performance compared to either one alone and also observed that the ability to process facial visuals, a specialized cognitive capacity, significantly influenced human deepfake detection performance.

Visual Realism Assessment (VRA) is an extension of the detection problem, focusing not only on the authenticity of videos but also on the quality of the generated DeepFakes. Recent studies, such as the DeepFake Game Competition on Visual Realism Assessment (DFGC-VRA)[1], have aimed to develop models that predict the Mean Opinion Score (MOS) for deepfake videos, a measure of subjective quality and realism. Sun *et al.* [6],

in their paper "Visual Realism Assessment for Face-swap Videos," proposed a benchmark for evaluating the effectiveness of various automatic VRA models. They used the DFGC 2022 dataset for their evaluations and demonstrated the feasibility of creating effective VRA models for assessing face-swap videos and emphasized the usefulness of existing deepfake detection features for VRA.

In DFGC 2022, the winning team used an ensemble of models including ConvNext for their deepfake detection solution [7]. This success suggests that models like ConvNext and Vision Transformers can be effectively used not only for deepfake detection but potentially also for Visual Realism Assessment. This underscores the importance of considering visual realism and human cognitive abilities in the development of deepfake detection models and further motivates this work.

3 Methods

Overview. Our approach leverages an ensemble of two distinct Convolutional Neural Network (CNN) models: Eva [10] and ConvNext [9]. Both models are equipped with dedicated regression heads, specifically designed to predict Mean Opinion Scores (MOS) from deepfake videos based on features gathered from sequences of five frames. The models have been trained on the DFGC 2022 dataset and represent our submissions to the recent *DeepFake Game Competition on Visual Realism Assessment*. The competition was held in conjunction with the 2023 IEEE International Joint Conference on Biometrics (IJCB 2023). Our method managed to secure the third place in this competition. The code for the implementation of our models, alongside a detailed technical report, is publicly available from our GitHub repository[1].

ConvNext. The ConvNext [9] model we employ is part of a family of convolutional neural networks (ConvNets). It is an evolution of the ResNet architecture, progressively incorporating elements from hierarchical vision Transformers. The ConvNext model leverages the pretrained DFGC-1st-2022-model (the winner of the DFGC2022 competition [7]) as its backbone. The model is used to extract features from the input video frames. The last fully connected layer of the backbone model is replaced with an identity layer and a dropout layer is then introduced to manage model complexity. A defining feature of our ConvNext model is its approach to feature aggregation, specifically designed to handle the mean and standard deviation vectors from the video frames. To accommodate these aggregated features, several fully connected layers are added. The final layer in this structure is specifically designed to output the Mean Opinion Score (MOS).

The forward pass of the ConvNext model begins with an input video sequence. A random starting point is chosen within the video, from which a sequence of 5 consecutive frames is selected. Each frame is processed by the backbone model to extract corresponding features. Following the feature extraction step, the mean and standard deviation of the features are calculated for each frame sequence. The calculation is guided by the principles of average pooling and standard deviation pooling, common techniques in the video quality assessment field. Specifically, the mean f_{mean} and standard deviation f_{std} of the features are computed as follows:

$$f_{\text{mean}} = \frac{1}{n}\sum_{i=1}^{n} f_i, \tag{1}$$

$$f_{\text{std}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(f_i - f_{\text{mean}})^2}. \tag{2}$$

Here, n represents the total number of frames, the feature vector for the i-th frame is denoted as f_i and f_{mean} is the average feature vector. The computed mean and standard-deviation vectors are concatenated to form the video-level features. They are then fed into the fully connected layers of the ConvNext model, culminating in the final output: the Mean Opinion Score (MOS) for the input frame sequence.

It is noteworthy that the weights of the backbone model are not frozen in the ConvNext model during training. This approach allows for the fine-tuning of the entire model, including the backbone and the newly added layers, for the specific task of predicting MOS scores on deepfake videos. The learning objective used is the Root Mean Squared Error (RMSE), which provides a measure of the differences between values predicted by the model and the values actually observed.

The Eva model. In parallel to ConvNext, we also utilize the Eva [10] model, a novel approach to visual representation learning that explores the limits of large-scale masked image modeling (MIM) using publicly accessible data. Eva is a vanilla Vision Transformer (ViT) that is pre-trained to reconstruct masked out image-text aligned vision features conditioned on visible image patches. This unique pretraining task allows Eva to scale efficiently, even up to one billion parameters, setting new records across a wide range of downstream visual tasks, such as image recognition, video action recognition, object detection, instance segmentation, and semantic segmentation, without heavy supervised training [10].

Just as with ConvNext, we use Eva as a feature extractor for predicting Mean Opinion Scores (MOS) on deepfake videos. The architecture is similar to that of ConvNext, with the backbone model replaced by the Eva model. However, unlike ConvNext, the Eva model is not initialized with weights from the winners of the DFGC2022 competition. Instead, we use weights pretrained on ImageNet using the timm library[2].

4 Experiments

Experimental dataset. The DeepFake Game Competition (DFGC) 2022 dataset, which originated from the second DFGC, held in conjunction with IJCB-2022, is composed of 2799 fake and 1595 real face-swap videos,

[1] https://github.com/TheLukaDragar/UNI-LJ-VRA

[2] https://github.com/rwightman/pytorch-image-models/tree/main/results

each approximately 5 seconds in length. More specifically, it contains face-swap videos for 20 pairs of subjects (IDs), and 35 deepfake creation methods. The fake videos are generated using a variety of face-swap methods like DeepFaceLab [8], SimSwap [11], and FaceShifter [12], along with postprocessing operations.

For the competition 1,400 deepfake videos have been annotated by five independent human raters who assessed video realism among other factors, rating it on a scale of 1 (very bad) to 5 (very good) - examples of the average scores are provided and marked as gt in Figure 1. These 1,400 videos are organized into a train set of 700 videos, provided to participants for training, and three test sets, used for the competition itself [1].

Preprocessing. The data preprocessing stage of our work with the DFGC-2022 dataset involved extracting faces from each video frame with the Multi-task Cascaded Convolutional Networks (MTCNN) [13] model and OpenCV [14]. The bounding boxes of the faces were resized and adjusted by a scale factor of 1.3, providing context and improving prediction accuracy [7]. This process generated a new dataset of cropped face images, used as input for subsequent modeling. By preprocessing the data prior to model training, we were able to expedite the training process, save time and computational resources.

Training data. We implemented a method to process each video by selectively extracting sequences of frames. This involved randomly selecting a starting point within each video and subsequently capturing a sequence of five frames from that point. The frames were then transformed in accordance with the requirements of our modeling process. Each sequence was matched with its corresponding Mean Opinion Score (MOS) label.

Next, the dataset was divided into three subsets to facilitate the training, validation, and testing of our models. The distribution of the dataset was as follows: out of the total 700 videos, 70% (490 videos) were allocated for training, 20% (140 videos) for testing, and the remaining 10% (70 videos) were used for validation.

Experimental setup. Our training procedure involved the use of a High-Performance Computing (HPC) infrastructure, facilitated by Pytorch Lightning. The employed loss function was Root Mean Square Error (RMSE), and AdamW was chosen as the optimizer with a learning rate of 2e-5. The learning rate scheduler, ReduceLROnPlateau, was incorporated along with early stopping, both of which were monitored via validation loss. The selected hyperparameters were a batch size of 2, a dropout rate of 0.1, a sequence length of 5, gradient accumulation across 8 batches, and a maximum of 33 epochs. Training was conducted on two Tesla V100S-32GB GPUs using a distributed data parallel (ddp) strategy.

Due to the non-deterministic nature of the training process, multiple models were trained with identical parameters. From these runs, the model with the best performance, as determined by validation loss, was selected for the final submission. Weights & Biases was utilized for logging and real-time tracking of the training process. At the end of the training, the best model checkpoint based on validation loss was selected and further trained with the same hyperparameters to accommodate the remaining data. Once early stopping was triggered, this final model checkpoint was saved for the final predictions.

Predictions and model averaging. The final model checkpoints of both the ConvNext and Eva models were utilized to make predictions on the test sets for our final submission. These test sets, referred to as Test Set 1 (300 videos ID-disjoint with the train set), Test Set 2 (280 videos method-disjoint), and Test Set 3 (120 videos ID&method-disjoint), were extracted from different subsets of the DeepFake Game Competition (DFGC) 2022 dataset.

Our dataloader operates in a stochastic manner, selecting sequences of 5 frames from the videos at random. To mitigate the variance introduced by this randomness, we applied an averaging strategy for each test set: we generated predictions 10 times and then computed the average. This procedure yields more robust predictions that accommodate the inherent randomness of our frame sequence selection process. For the combination of predictions from the two models, we employed a weighted average approach. The final prediction was computed as 0.75 times the ConvNext prediction plus 0.25 times the Eva prediction. This weighting scheme was chosen due to the ConvNext model's superior performance during the training phase.

To assess the consistency of our predictions, we calculated the Root Mean Square Error (RMSE) between pairs of predictions. For the ConvNext model, the average RMSE was found to be 0.16, indicating a reasonable level of consistency in our predictions.

5 Results

Our experimental results are detailed in Tables 1 and 2, and visually depicted in Figures 1 and 2. Table 1 displays the performance of our two distinct models, Eva and ConvNext, and our ensemble model across three test sets. Each model's performance is evaluated using three metrics: Pearson Linear Correlation Coefficient (PLCC), Spearman Rank-Order Correlation Coefficient (SRCC), both being the competition's official metrics[1], and Root Mean Square Error (RMSE).

In Table 2, we compare the final scores of our models against the baseline model established by the competition organizers [6]. These final scores, computed by averaging the PLCC and SRCC, provide a comprehensive measure of each model's effectiveness in visual realism assessment of DeepFake videos. As shown in Table 3, our ensemble model demonstrated competitive performance and secured the third place in the competition[3].

Regarding the data presented in Tables 1 and 2, a noteworthy finding arises in relation to the performance of our Eva model. Despite initially achieving marginally

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

Table 1: Performance Metrics for Each Model on the Test Sets.

Model	Test Set	PLCC↑	SRCC↑	RMSE↓
Eva	1	**0.8305**	**0.7919**	**0.4128**
	2	0.9158	0.9119	0.3622
	3	0.8726	0.8285	**0.4132**
ConvNext	1	0.7899	0.7387	0.4545
	2	0.9279	0.9171	0.3492
	3	0.8647	0.8211	0.4303
Ensemble	1	0.8091	0.7633	0.4352
	2	**0.9287**	**0.9197**	**0.3447**
	3	**0.8746**	**0.8318**	0.4146

Table 2: Final Scores for Each Model.

Model	Final Score
Baseline	0.5470
Eva	**0.8585**
ConvNext	0.8432
Ensemble	0.8545

lower results than ConvNext during the training phase, Eva demonstrated superior performance on test set 1, ultimately attaining a slightly higher overall score. This indicates that Eva exhibits enhanced generalization capabilities compared to ConvNext.

Table 3: Top 3 Models, DFGC-VRA 2023 Competition[1].

Model	Test Set	PLCC↑	SRCC↑	Avg↑
OPDAI	1	0.8578	0.8372	0.8851
	2	0.9423	0.9214	
	3	0.8928	0.8592	
HUST	1	0.8117	0.7864	0.8611
	2	0.9281	0.9215	
	3	0.8842	0.8348	
Ours	1	0.8091	0.7633	0.8545
	2	0.9287	0.9197	
	3	0.8746	0.8318	

6 Conclusion

In our study for the DFGC-VRA 2023 challenge[1], we utilized two distinctive CNN models, ConvNext and Eva, to evaluate the visual realism of DeepFake videos. These models, strategically enhanced with pre-processing, feature extraction, and model averaging techniques, were trained to predict Mean Opinion Scores (MOS) on the DFGC 2022 dataset. Notably, our ensemble model earned a third-place ranking in the challenge, underscoring its efficiency in assessing DeepFake video realism.

This research made an important contribution to the evaluation of visual realism in deepfake videos, supporting the pursuit of a safer digital media landscape. As part of our future work, we plan to extend our initial model to further improve performance.

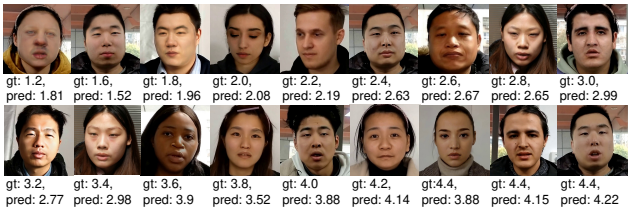

Figure 2: Videos featuring varying levels of realism in face-swapping, annotated according to the Mean Opinion Score (MOS) based on ground truth (gt) and the MOS predicted (pred) by our ensemble model. These videos are arranged in ascending order according to the ground truth MOS.

References

[1] B. Peng, X. Sun, C. Wang, W. Wang, J. Dong, Z. Sun, H. Cong, L. Fu, H. Wang, R. Zhang, Y. Zhang, H. Zhang, X. Zhang, B. Liu, H. Ling, L. Dragar, B. Batagelj, P. Peer, V. Štruc, H. Wang, W. Diao, *DFGC-VRA: DeepFake Game Competition on Visual Realism Assessment*, *In Proceedings of the 2023 IEEE International Joint Conference on Biometrics* (under review).

[2] Masood, Momina and Nawaz, Mariam and Malik, Khalid Mahmood and Javed, Ali and Irtaza, Aun and Malik, Hafiz, *Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward*, *Applied Intelligence*, vol. 53, no. 4, pp. 3974–4026, 2023.

[3] M. Ivanovska, A. Kronovšek, P. Peer, V. Štruc, and B. Batagelj, *Face morphing attack detection using privacy-aware training data*, *In Proceedings of the 31st International Electrotechnical and Computer Science Conference*, 2022.

[4] Pavel Korshunov and Sébastien Marcel, *Deepfake detection: humans vs. machines*, arXiv preprint arXiv:2009.03155, 2020.

[5] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard, *Deepfake detection by human crowds, machines, and machine-informed crowds*, *In Proceedings of the National Academy of Sciences*, Vol. 119, No. 1, 2021.

[6] Xianyun Sun, Beibei Dong, Caiyong Wang, Bo Peng, Jing Dong, *Visual Realism Assessment for Face-swap Videos*, arXiv preprint arXiv:2302.00918, 2023.

[7] Bo Peng, Wei Xiang, Yue Jiang, Wei Wang, Jing Dong, Zhenan Sun, Zhen Lei, Siwei Lyu, *DFGC 2022: The Second DeepFake Game Competition*, arXiv preprint arXiv:2206.15138, 2022.

[8] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr. Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, Pingyu Wu, Bo Zhou, Weiming Zhang, *DeepFaceLab: Integrated, flexible and extensible face-swapping framework*, arXiv preprint arXiv:2005.05535, 2021.

[9] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, Saining Xie, *A ConvNet for the 2020s*, *CoRR*, Vol. abs/2201.03545, 2022.

[10] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, Yue Cao, *EVA: Exploring the Limits of Masked Visual Representation Learning at Scale*, *arXiv preprint arXiv:2211.07636*, 2022.

[11] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge, *Simswap: An efficient framework for high fidelity face swapping*, *In Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[12] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen, *FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping*, arXiv preprint arXiv:1912.13457, 2020.

[13] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, *Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks*, *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[14] G. Bradski, *The OpenCV Library*, *Dr. Dobb's Journal of Software Tools*, 2000.