Attending Generalizability in Course of Deep Fake Detection by Exploring Multi-task Learning

Pranav Balaji
BITS Pilani
Hyderabad, India
f20190040@hyderabad.bits-pilani.ac.in

Abhijit Das
BITS Pilani
Hyderabad, India
abhijitdas2048@gmail.com

Srijan Das
University of North Carolina at Charlotte
North Carolina, United States
sdas24@uncc.edu

Antitza Dantcheva
INRIA
Biot, France
antitza.dantcheva@inria.fr

Abstract

This work explores various ways of exploring multi-task learning (MTL) techniques aimed at classifying videos as original or manipulated in cross-manipulation scenario to attend generalizability in deep fake scenario. The dataset used in our evaluation is FaceForensics++, which features 1000 original videos manipulated by four different techniques, with a total of 5000 videos. We conduct extensive experiments on multi-task learning and contrastive techniques, which are well studied in literature for their generalization benefits. It can be concluded that the proposed detection model is quite generalized, i.e., accurately detects manipulation methods not encountered during training as compared to the state-of-the-art.

1. Introduction

Deepfakes are computer-generated videos, images, or audio recordings that have been manipulated using artificial intelligence and machine learning algorithms to create realistic content of a person doing or saying something they did not actually do [14]. Deepfakes use deep learning techniques, such as neural networks, to manipulate existing content and create something new. The term ”deepfake” is a combination of ”deep learning” and ”fake”. Deepfakes are threats to society because they can be used to spread misinformation, manipulate public opinion, harass individuals, and even blackmail people. They can be particularly dangerous when used in political contexts, where they can be used to damage reputations or influence election outcomes [1]. Deepfakes can be generated in several ways, manipulation methods involve *generative adversarial networks* (GANs) to create realistic videos [33]. Another method involves using facial detection technology to map a person's face onto a face or to *superimpose* their face onto an existing video. One common approach is to train an autoencoder to reconstruct a specific person's face on any given image of a body.

Due to the recent fuel in media tampering techniques such as deep fake, corresponding *manipulation-detection approaches* have been developed. Manipulation detection techniques included image-based [33], video-based [14, 34], or jointly audio and video-based [12] approaches. In the context of image-based deep fake detection to explore the angle of generalizability, second-order local anomaly detection has been used [19] and self-consistency is explored in the work of [38]. Singular frame based detection techniques ensemble predictions across frames of a video [45, 10]. While computationally efficient, they do not exploit the presence of temporal inconsistencies [28]. Hence recently, video-based generalized deep fake detection has been gaining significance. This for the detection of temporal inconsistencies with respect to the lip movement [2], jitters between frames [21] and optical flow [4]. Identity consistency has also been explored, with [18] using a transformer to identify inconsistency between the inner and outer face region based on a database of known identities. Authors of [3] modelled behaviors of world leaders from recorded stock footage and identify behavioral inconsistencies in deepfakes. These techniques usually require prior identity or behavior information about the victim, so they are suited for celebrities but do not scale to civilian victims.

Tolosana et al. [39] reviewed manipulation technique such as DeepFake in 2020 w.r.t. facial regions, and fake

[1] https://tinyurl.com/mrxtxk5z

detection performance and concluded that the generalization of such detection methods is challenging. In other words, when detection methods, such as the presented ones are confronted with adversarial attacks, outside of the training set, such networks have a dramatic drop in performance. Challenge of generalization of deep fake is studied in [14, 34]. In [7] self-supervised learning (SSL) has been explored by learning adversarial examples for generalized deep fake detection. Further, in the same direction of research, a multimodal approach has been adopted in [12, 46] using audio-video analysis. The angle of generalizability for deep fake detection is not much explored and yet remain to be a challenge.

Hence, in order to address the unsolved challenge of generalizability in deep fake detection, we explore the concept of MTL encompassing both supervised and self-supervised learning (SSL) approaches [26]. By employing MTL, we aim to not only detect deep fakes but also identify the specific type of deep fake jointly. MTL involves jointly learning multiple tasks, typically with a shared early layer or common connections at the beginning of the network and individual task-specific layers at the end. This shared layers in early processing allows to improved learning by sharing parameters across tasks. Convolutional Neural Networks (CNNs), based on deep neural networks (DNNs), have demonstrated exceptional performance in simultaneously solving diverse tasks within the MTL framework [29].

MTL has been extensively studied in various domains of machine learning and deep learning applications. For instance, it has found application in natural language processing tasks, such as unified representations [11] and representation learning [27]. Additionally, MTL has been employed in speech recognition [16], drug discovery [31], and computer vision-related tasks including face analysis [13, 23], pedestrian detection [6, 43, 37], face alignment [44], attribute prediction [1], among others. Furthermore, MTL has gained significance in recent times for face attribute learning, also known as semantic features, as they provide a more natural description of objects [30] and activities [24]. This approach enables a comprehensive understanding of the visual world by jointly modeling multiple attributes in face-related tasks. In their recent work [20, 13], the authors propose a biasless approach for face attribute analysis. These couple of work on MTL inspires and gives the importance of achieving generalizability in deep fake detection, it becomes crucial to explore MTL. While MTL is commonly studied using Supervised Learning (SL), the exploration of MTL in the context of SSL remains relatively unexplored, making it highly relevant to the problem at hand. Additionally, it is worth noting that deep fake detection has not been extensively investigated within the MTL framework. Therefore, in this paper, we aim to address the identified research gap by exploring deep fake detection in the MTL scenario.

SSL has demonstrated successful applications in both image-based methods such as SimCLR [8], MoCo [26], MAE [25], DINO [5], and video-based approaches like CoCLR [22], VideoMAE [40], SVT [32]. SSL focuses on learning representations by leveraging inherent data structure, eliminating the reliance on labeled data. This paradigm has shown significant improvements in generalizability, with recent studies surpassing the performance of supervised models on zero-shot testing [25]. Building on these advancements, we anticipate that SSL can enable the model to learn better representations for enhancing generalizability in the context of deep fake detection. Specifically, the SSL-based approach can empower the model to effectively discriminate between fake and real instances, irrespective of the manipulation technique employed. This advancement raises several open questions regarding the application of MTL and SSL compared to SL in deepfake detection. Thus, in this paper, we aim to address the following questions:

- Can MTL be effectively employed for deepfake detection?
- How does the choice of loss affect MTL?
- How can a combination of Contrastive and Cross entropy-based learning be utilized to enhance the generalizability of deepfake detection?
- What are the appropriate sub-tasks and their relationships with the primary task within the MTL framework for deepfake detection?

To address the aforementioned inquiries, we conducted a thorough examination of the existing literature on MTL, SL, and SSL. In our experimental investigation, we explored both contrastive learning and SL techniques within the MTL framework. The outcomes of our study indicate that MTL holds significant value for the task of deepfake detection. However, the subsequent question arises: which type of learning, SL or SSL, is most relevant to this particular problem? Our experimental analysis provides insights into this query, revealing that incorporating a combination of SL or/and SSL as a sub-task of manipulation classification yields superior results compared to employing SSL alone for the detection task within the MTL paradigm. Finally, we also introspect the relation of the sub-task (see Figure 1) and the binary classification problem (fake vs real).

Therefore, in this study, we comprehensively explore the optimal strategies for employing MTL in the detection of deepfakes. Through extensive experimental analysis, we demonstrate various configurations of MTL networks that yield exceptional performance in deepfake detection. Furthermore, we offer end users a range of options for selecting the backbones of MTL for traditional deepfake detection networks in both Contrastive and Cross entropy approaches (see Figure 2), tailored to their specific constraints. We firmly believe that this empirical investigation not only contributes to the advancement of deepfake detection and its generalizability but also presents promising avenues for future research in this domain.

2. Proposed methodology

As mentioned previously that Deepfakes can be generated in several ways and several new techniques can evolve. Main manipulation methods involve using generative and superimpose technique. Detecting deepfakes is a challenging problem because they are designed to be convincing and difficult to distinguish from real content. Moreover, generalizability of deep fake detection i.e while they are tested on cross-manipulation technique is an added challenge. To underpin this problem we hypothesized to use explore MTL in both supervised learning (SL) and self-supervised learning (SSL) as a base while learning detection technique and type of deep fake jointly. In lieu of these we assume that in the course of MTL-based training, the model will learn a better representation for generalizability. In other words, the MTL-base will help the model learn how to discriminate between fake and real irrespective of the manipulation used.

2.1. MTL for deep fake detection

The aim of the experiments is to design a training paradigm that enables the core encoder to learn features that are generalizable across various manipulation techniques. To this end, we employ 2 techniques: Multi-task learning and SSL with Momentum contrast (MoCo) with label-informed positive and negative pool construction. Due to popularity and the relevance in the learning aspect MoCo was selected among the SSL techniques available. The following encoder and stream format is employed throughout the paper.

Encoder: In all our experiments, the S3D encoder, pre-trained with CoCLR [22] is used.

Classifier: A 3 layer MLP was used to classify the embedding vector produced by the S3D into the respective classes. This was trained using Cross entropy loss.

Dual stream: To encourage the encoder to learn not just to differentiate between original videos and manipulated videos but also to identify the type of manipulation, we consider training it using a stream for Multi-class classification. We proceed to explain MTL in details.

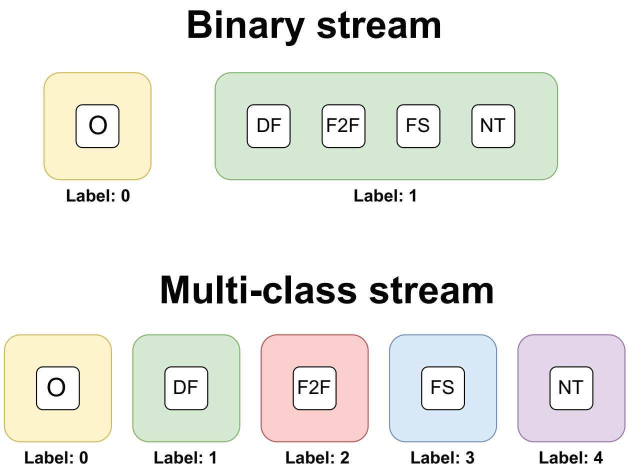


Figure 1. Difference between Binary stream and multi-class stream

2.1.1 Multi-task learning

This involves using two different streams - binary and multi-class, where labels are defined differently.
Binary stream: The task performed is binary classification where 0 represents ”original” media while 1 represents ”manipulated” media. All manipulation techniques - Deepfakes, Face2Face, FaceSwap and NeuralTextures fall under this category. This stream is of ultimate interest.
Multi-class stream: The task performed is multi-class classification where 0 represents ”originals” and each of classes 1-n are assigned to different manipulation techniques. For example, in the case that a model is being trained on Originals, FaceSwap, Face2Face and NeuralTextures, there would be a total of 4 classes. This stream gives the model richer information in the form of *which* manipulation technique is being encountered, which could help in generalizability. Both types of the models are illustrated in Figure 1. This stream is only employed during training and is discarded during testing.

2.1.2 Label-Informed MoCo

Following the success in applying MoCo [26] to labelled data [22], we design the other technique. In essence, it is MoCo for videos applied while the positive and negative pool is constructed not by augmentations, but by indexing the corresponding labels. The positive pool for a given sample, in this case, may consist of more than one instance as opposed to the original MoCo paradigm. Hence, we used the Multi-Instance Info-NCE loss as per the following equa-

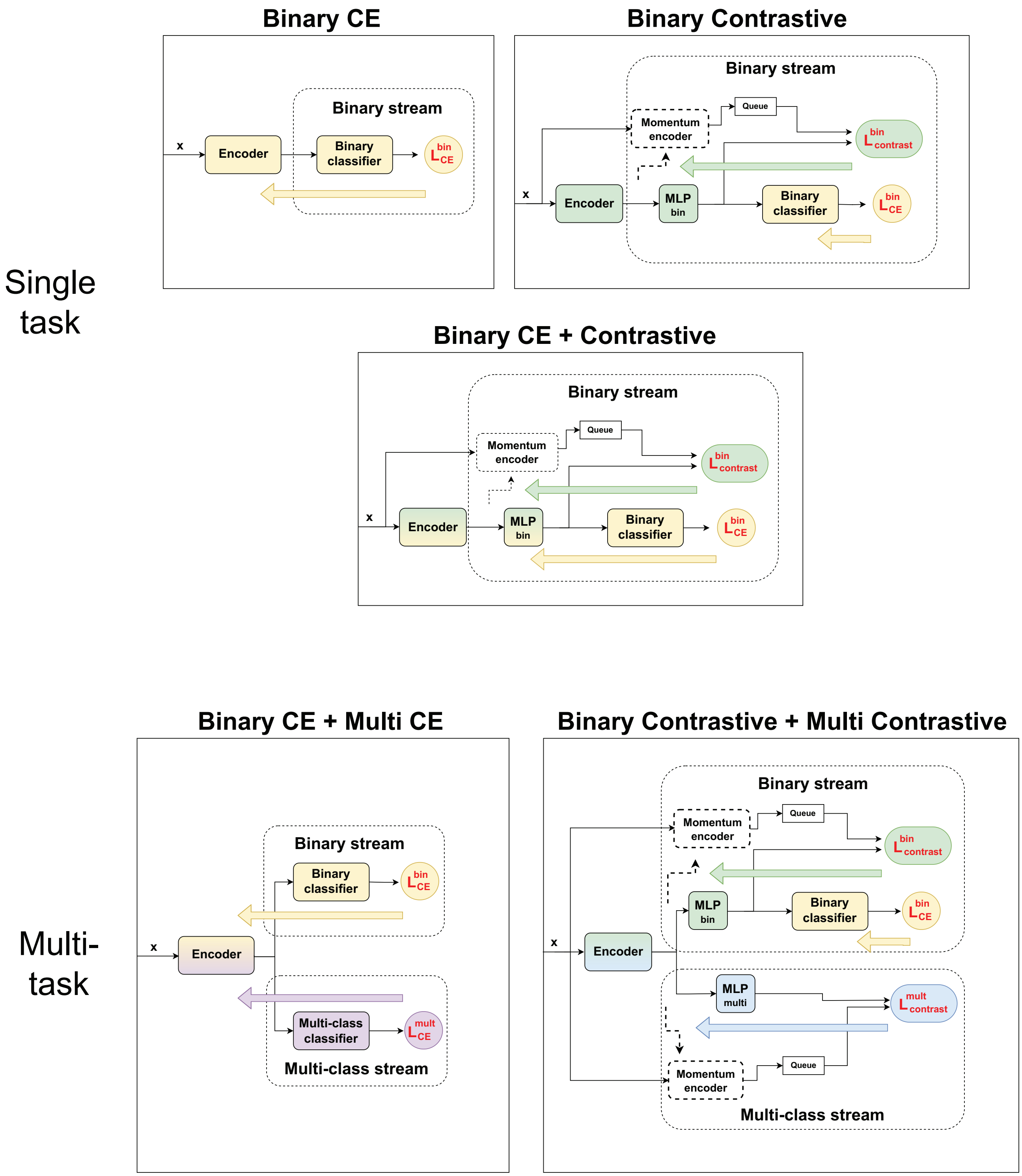


Figure 2. Block representation of different scenarios of MTL proposed for deep fake detection. Colored arrows indicate gradient flow

tion:

$$L_{contrast} = -E[\log \frac{\Sigma_{p \in P_i} \exp(z_i \cdot z_p/\tau)}{\Sigma_{p \in P_i} \exp(z_i \cdot z_p/\tau) + \Sigma_{n \in N_i} \exp(z_i \cdot z_n/\tau)}] \tag{1}$$

Equation 1: Multi-Instance Info-NCE loss. The numerator is a sum of 'similarity' between sample x_i and its positive set P_i. P_i is defined as the set of examples in the queue with the same label as x_i, and N_i is defined to be its complementary set. z_i is the representation of x_i from the main encoder while z_p and z_n are taken from the queue, generated by the EMA encoder in previous iterations. $z_i \cdot z_j$ refers to the similarity between vectors z_i and z_j and is defined to be the dot product. Intuitively, this loss has the effect of pulling the representations of positive pairs together and pushing the negative pairs apart.

The construction of positive and negative pools for MoCo is also based on the definition of labels which is different in the binary and multi-class streams. In either case, positive samples are drawn from the same pool as the anchor while negative samples are drawn from any random pool that is different from the anchor.

Six different cases were considered for training, with respect to gradient propagation to the main encoder:

- **Binary CE:** The encoder is updated using the Binary Cross Entropy (CE) loss L_{CE}^{bin}. This entails standard fine-tuning. The model was pretrained with CoCLR SSL weights.

- **Binary Contrastive:** The encoder is updated only using the binary version of $L_{contrast}$ defined in 1, $L_{contrast}^{bin}$, where positive and negative pairs are defined as in Figure 2. Inspired by MoCo v2 [9], we use an MLP head after the encoder wherever $L_{contrast}$ is involved. The performance is monitored using a standalone Binary Classifier updated using BCE loss. These gradients are cut off at the classifier.

- **Binary CE + Contrastive:** Similar to the previous case, but here the gradients from the Binary CE (BCE) loss are not cut off at the classifier and are allowed to propagate to the main encoder. Hence, the encoder is updated using both L_{CE}^{bin} and $L_{contrast}^{bin}$.

- **Binary CE + Multi CE:** With the addition of multi-tasking, an additional stream performing multi-class classification is introduced. The encoder is updated from the Cross Entropy loss from both streams, L_{CE}^{bin} and L_{CE}^{mult}. The binary stream is of primary focus at the end of training.

- **Binary Contrastive + Multi Contrastive:** The encoder is updated using the two label-informed contrastive losses from the two streams, $L_{contrast}^{bin}$ and $L_{contrast}^{mult}$ where the difference is between positive and negative pair construction as defined in Figure 2. Just like in the case of *Binary Contrastive*, Separate MLP heads are used in each stream and a standalone Binary Classifier is trained with L_{CE}^{bin} to monitor performance.

All block diagrams for all the cases are illustrated in Figure 2.

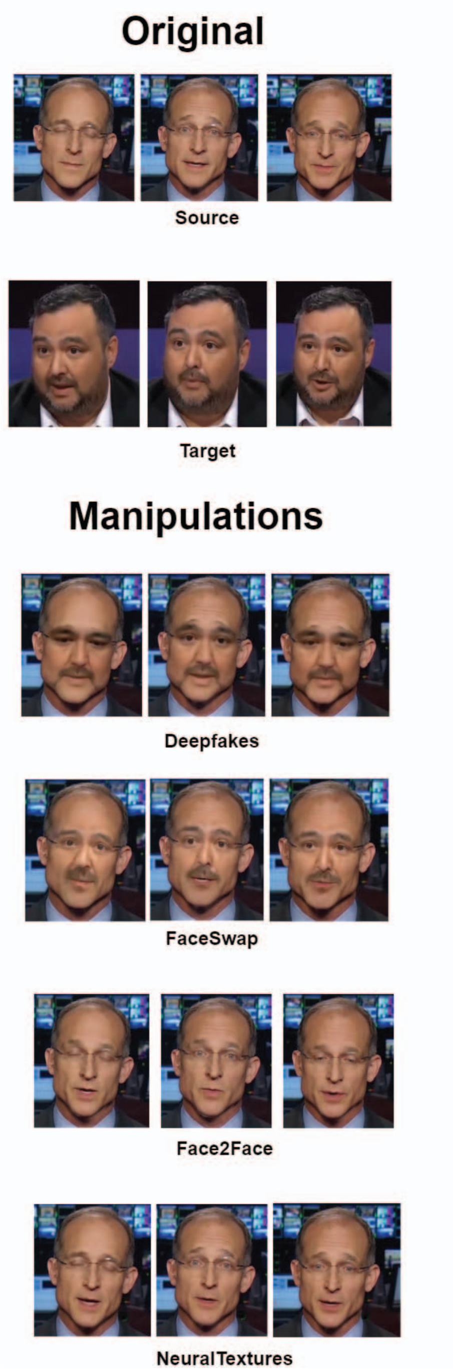


Figure 3. Examples of Manipulation techniques in FaceForensics++

3. Experiment results and discussion

In this section we proceed to explain the employed database and the experimental results achieved to validate our proposed methodology.

Table 1. Statistics of different manipulating techniques available in Faceforensics++

Split	DeepFake	Face2Face	FaceSwap	NeuralTextures	Original	Total
Train	720	720	720	720	720	3600
Val	140	140	140	140	140	700
Test	140	140	140	140	140	700
Total	1000	1000	1000	1000	1000	5000

3.1. Dataset

We employed a publicly available and most popular dataset on Deep fake.

FaceForensics++: The FaceForensics++(FF++) dataset [33] is a large-scale benchmark dataset for face manipulation detection, which was created to help develop automated tools that can detect deepfakes and other forms of facial manipulation. The dataset consists of more than 1,000 high-quality videos with a total of over 500,000 frames, which were generated using various manipulation techniques such as facial reenactment, face swapping, and deepfake generation.

The videos in the dataset are divided into four categories, each corresponding to a different manipulation technique: Deepfakes(DF), Face2Face(F2F), FaceSwap(FS), and NeuralTextures(NT). Deepfakes use machine learning algorithms to generate realistic-looking fake videos, while Face2Face and FaceSwap involve manipulating the facial expressions and identity of a person in a video. NeuralTextures uses a different approach by altering the texture of a face to make it appear different. The dataset includes both real and manipulated videos, with each manipulation technique applied to multiple individuals. Examples are shown in Figure 3.

3.2. Discussion

Evidently, MTL scenarios seem to be outperforming the base line i.e. S3D initialized with Supervised weights and the state-of-the-art for deepfake detection, the winner [35] of the DFDC Challenge [17]. [35] differs from our models in 3 key ways: Architecture (Image based EfficientNet B7 [36] vs our video based S3D [42]), Pre-training (Noisy student [41] vs our CoCLR [22]) and Image resolution (380 vs our 128). Our models outperform [35] in the majority of the scenarios. Only for the *Deepfakes* cross-manipulation scenario,[35] narrowly outperformed our best model by 0.5%.

It is interesting to note that all video-based models significantly outperform [35] on the *cross FaceSwap* and *cross NeuralTextures* cases even though they operate on less than half the frame resolution. This leads to the conclusion that at least for these manipulation techniques, ***temporal information is much more relevant for identifying forgery than frame resolution.***

Moreover, all methods involving MTL or contrastive learning suffer on the *cross Face2Face* and *cross NeuralTextures* scenario compared to regular BCE training. Analyzing the difference between the various manipulation techniques in Figure 3, it is clear that Face2Face and NeuralTextures work by changing the lip movement to match the target, rather than transplant the target's face on the source video. The general difficulty in detecting these techniques in cross training evaluation suggest that these techniques need special consideration and cannot be generalized to, from the other techniques.

As long as MTL is not involved, Contrastive loss leads to better generalization on Deepfakes and FaceSwap, while CE loss leads to better generalization on the other two.

Now we proceed to answer the question raised in the introduction section:

- **Can MTL be effectively employed for deepfake detection?** MTL helps in certain cases as it can be evident from Table 3 that Binary CE + Multi CE produces the best average results in cross-manipulation scenario. Although, just Binary CE training comes very close. Therefore, ***Yes in certain cases, but narrowly.***

- **How does the choice of loss affect MTL?**

  Supervised learning with Cross Entropy loss performs best, as it can be evident from Table 3. In CE training, Binary CE + Multi CE produces the best average results while Binary CE comes close.
  However, in the contrastive loss case, MTL significantly degrades the results as shown by Binary Contrastive + Multi Contrastive being significantly worse than Binary Contrastive training.
  Therefore, ***Cross Entropy loss is suited for Multi-task learning while Contrastive loss is not.***

- **How can a combination of Contrastive and Cross entropy-based learning be utilized to enhance the generalizability of deepfake detection?** It is evident from Table 3 that jointly training on CE and Contrastive loss is worse than training with a single type of loss. This leads us to the conclusion that the two losses produce conflicting gradients that lead the model to a sub optimal parameter space.

[2] The same model used in DFDC's winning solution, but without the DFDC pre-training

Table 2. FF++ results for cross manipulation scenario.

Train on	Test on	Type of training	Accuracy
F2F, FS, NT	DF	Baseline(S3D)	76.08%
F2F, FS, NT	DF	Binary CE	80.08%
F2F, FS, NT	DF	Binary Contrastive	84.77%
F2F, FS, NT	DF	Binary CE + Contrastive	81.64%
F2F, FS, NT	DF	Binary CE + Multi CE	**85.16%**
F2F, FS, NT	DF	Binary Contrastive + Multi Contrastive	83.20%
F2F, FS, NT	DF	DFDC Winner [2] [35]	**85.71%**
DF, FS, NT	F2F	Baseline(S3D)	66.08%
DF, FS, NT	F2F	Binary CE	**73.44%**
DF, FS, NT	F2F	Binary Contrastive	68.36%
DF, FS, NT	F2F	Binary CE + Contrastive	66.02%
DF, FS, NT	F2F	Binary CE + Multi CE	68.75%
DF, FS, NT	F2F	Binary Contrastive + Multi Contrastive	58.98%
DF, FS, NT	F2F	DFDC Winner [35]	72.86%
DF, F2F, NT	FS	Baseline(S3D)	75.12%
DF, F2F, NT	FS	Binary CE	79.30%
DF, F2F, NT	FS	Binary Contrastive	82.42%
DF, F2F, NT	FS	Binary CE + Contrastive	**85.55%**
DF, F2F, NT	FS	Binary CE + Multi CE	83.59%
DF, F2F, NT	FS	Binary Contrastive + Multi Contrastive	80.08%
DF, F2F, NT	FS	DFDC Winner [35]	51.07%
DF, F2F, FS	NT	Baseline(S3D)	69.18%
DF, F2F, FS	NT	Binary CE	**75.78%**
DF, F2F, FS	NT	Binary Contrastive	69.92%
DF, F2F, FS	NT	Binary CE + Contrastive	70.31%
DF, F2F, FS	NT	Binary CE + Multi CE	73.44%
DF, F2F, FS	NT	Binary Contrastive + Multi Contrastive	64.06%
DF, F2F, FS	NT	DFDC Winner [35]	60.00%

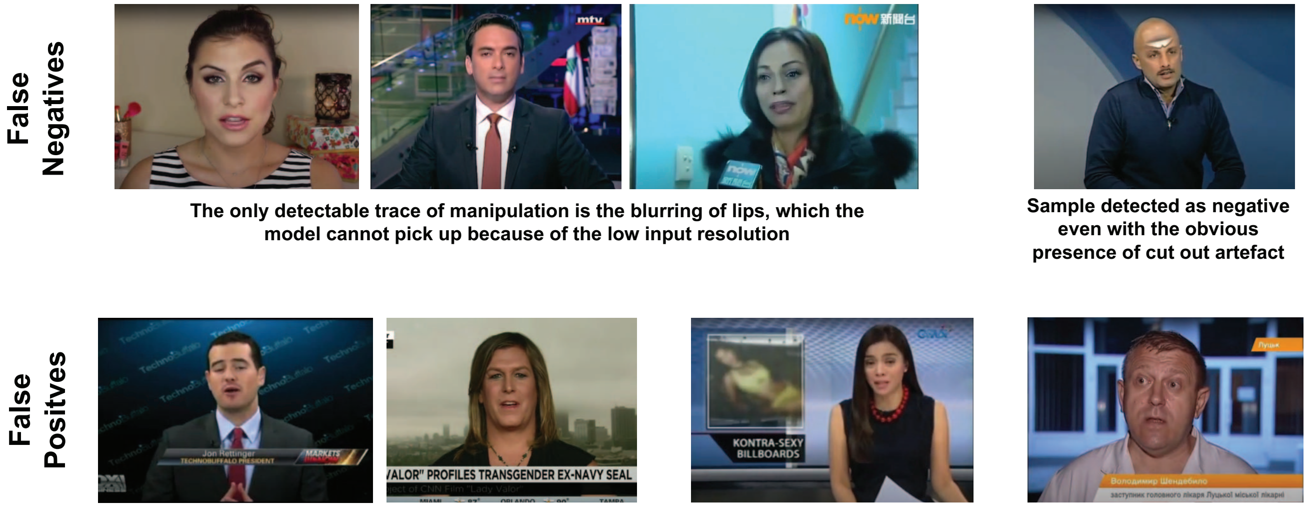
Table 3. Average of all FF++ results for cross manipulation scenario.

Type of training	Avg. Accuracy
Baseline(S3D)	71.62%
Binary CE	77.15%
Binary Contrastive	76.17%
Binary CE + Contrastive	75.98%
Binary CE + Multi CE	**77.74%**
Binary Contrastive + Multi Contrastive	71.58%
DFDC Winner [35]	67.41%

Therefore, ***It is not good to jointly train on Cross Entropy and Contrastive loss.***

- **What are the appropriate sub-tasks and their relationships with the primary task within the MTL framework for deepfake detection?** From Table 3, it is evident that identifying the type of manipulation is a good sub-task when using Cross Entropy loss, but not when using Contrastive loss.
  However, there could be better sub-tasks for generalization such as non-contrastive SSL approaches used as Self-Supervised Auxiliary Training [15], some examples being SVT [32] and VideoMAE [40], which we leave to future work. Therefore for now, ***Finding the type of manipulation is a good sub task for Cross Entropy training.***

A few examples of mis-classifications are shown in Figure 4. Analyzing the false negatives show that sometimes, the only detectable trace of manipulation is the blurring of lips, and the models are unable to pick it up because of the low input resolution. A straightforward fix to this would be to use a higher resolution model, at the cost of higher com-



The only detectable trace of manipulation is the blurring of lips, which the model cannot pick up because of the low input resolution

Sample detected as negative even with the obvious presence of cut out artefact

Samples detected as positive possibly due to the artefacts around the head from the green-screen

Sample detected as positive possibly due to the presence of another blurry face

Sample detected as positive with no apparent signs of forgery

Figure 4. Examples of mis-classifications.

pute.
False positives on the other hand happen due to various reasons. Sometimes the green-screen effect is present around the head region, which the model could mistake as artefacts from forgery techniques. This brings into question if such media should really be treated as *original*. If not, a larger training set containing more of such examples could be used. Sometimes, there are other blurred faces in the video which could confuse the model. This could be resolved by robust face detection and alignment, which we leave to future work.

4. Conclusion

Overall, this study shows that treating deep fake detection naively as a video classification problem rather than image classification greatly helps in generalizability. Manipulation detection and classifying the type of manipulation present are proved to be related tasks by the observation that one serves as a good sub-task for the other. Moreover, we show that while both Cross Entropy training and Contrastive training are good, jointly training on the two losses seems to be counterproductive. When it comes to the contrastive loss framework, we have shown that it plays poorly into multi-task learning, at least for our chosen subtask of identifying the type of manipulation present. While our proposed methods did not surpass the state-of-the-art on FaceForensics++ for all scenarios, we have shown that our models trained on a subset of forgery techniques can generalize to never-before-seen manipulations much better than the state-of-the-art.

5. Acknowledgements

We acknowledge the high performance computing facility, Sharanga, at the Birla Institute of Technology and Science - Pilani, Hyderabad campus primarily used for our experiments. We also acknowledge the support provided by INRIA as part of an internship to Pranav Balaji.

References

[1] Abrar H Abdulnabi, Gang Wang, Jiwen Lu, and Kui Jia. Multi-task cnn model for attribute prediction. *IEEE Transactions on Multimedia*, 17(11):1949–1959, 2015. 2

[2] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 660–661, 2020. 1

[3] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *CVPR workshops*, volume 1, page 38, 2019. 1

[4] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 1

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[6] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In

European conference on computer vision, pages 109–122. Springer, 2014. 2

[7] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18710–18719, 2022. 2

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 5

[10] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. Combining efficientnet and vision transformers for video deepfake detection. In *International conference on image analysis and processing*, pages 219–229. Springer, 2022. 1

[11] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167, 2008. 2

[12] Emanuele Conti, Davide Salvi, Clara Borrelli, Brian Hosler, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, Matthew C Stamm, and Stefano Tubaro. Deepfake speech detection through emotion recognition: a semantic approach. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8962–8966. IEEE, 2022. 1, 2

[13] Abhijit Das, Antitza Dantcheva, and Francois Bremond. Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 2

[14] Abhijit Das, Srijan Das, and Antitza Dantcheva. Demystifying attention mechanisms for deepfake detection. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–7. IEEE, 2021. 1, 2

[15] Srijan Das, Tanmay Jain, Dominick Reilly, Soumyajit Karmakar, Shyam Marjit, Xiang Li, and Michael Ryoo. From few to more: Enhancing vit performance on limited data. 2023. 7

[16] Li Deng, Geoffrey Hinton, and Brian Kingsbury. New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8599–8603. IEEE, 2013. 2

[17] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020. 6

[18] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. Protecting celebrities from deepfake with identity consistency transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9468–9478, 2022. 1

[19] Jianwei Fei, Yunshu Dai, Peipeng Yu, Tianrun Shen, Zhihua Xia, and Jian Weng. Learning second order local anomaly for general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20270–20280, 2022. 1

[20] Pasquale Foggia, Antonio Greco, Alessia Saggese, and Mario Vento. Multi-task learning on the edge for effective gender, age, ethnicity and emotion recognition. *Engineering Applications of Artificial Intelligence*, 118:105651, 2023. 2

[21] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2018. 1

[22] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33:5679–5690, 2020. 2, 3, 6

[23] SL Happy, A Dantcheva, A Das, F Bremond, R Zeghari, and P Robert. Apathy classification by exploiting task relatedness. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 733–738. 2

[24] SL Happy, Antitza Dantcheva, Abhijit Das, Radia Zeghari, Philippe Robert, and Francois Bremond. Characterizing the state of apathy with facial expression and motion analysis. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019. 2

[25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2, 3

[27] Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-yi Wang. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 912–921, 2015. 2

[28] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1):1–41, 2021. 1

[29] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3994–4003, 2016. 2

[30] Xuesong Niu, Xingyuan Zhao, Hu Han, Abhijit Das, Antitza Dantcheva, Shiguang Shan, and Xilin Chen. Robust remote heart rate estimation from face utilizing spatial-temporal at-

tention. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019. 2

[31] Bharath Ramsundar, Steven M. Kearnes, Patrick Riley, Dale Webster, David E. Konerding, and Vijay S. Pande. Massively multitask networks for drug discovery. *ArXiv*, abs/1502.02072, 2015. 2

[32] Kanchana Ranasinghe, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Michael S Ryoo. Self-supervised video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2874–2884, 2022. 2, 7

[33] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019. 1, 6

[34] Ritaban Roy, Indu Joshi, Abhijit Das, and Antitza Dantcheva. 3d cnn architectures and attention mechanisms for deepfake detection. In *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*, pages 213–234. Springer International Publishing Cham, 2022. 1, 2

[35] Selim Seferbekov. Dfdc 1st place solution, 2020. 6, 7

[36] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 6

[37] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Pedestrian detection aided by deep learning semantic tasks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5079–5087, 2015. 2

[38] Hitika Tiwari, Vinod K Kurmi, KS Venkatesh, and Yong-Sheng Chen. Occlusion resistant network for 3d face reconstruction. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 813–822, 2022. 1

[39] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020. 1

[40] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 2, 7

[41] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 6

[42] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018. 6

[43] Xi Yin and Xiaoming Liu. Multi-task convolutional neural network for pose-invariant face recognition. *IEEE Transactions on Image Processing*, 27(2):964–975, 2017. 2

[44] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 2

[45] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021. 1

[46] Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14800–14809, 2021. 2