

EXTRACTING DEEP LOCAL FEATURES TO DETECT MANIPULATED IMAGES OF HUMAN FACES

Michail Tarasiou, Stefanos Zafeiriou

Imperial College London

ABSTRACT

Recent developments in computer vision and machine learning have made it possible to create realistic manipulated videos of human faces, raising the issue of ensuring adequate protection against the malevolent effects unlocked by such capabilities. In this paper we propose local image features that are shared across manipulated regions are the key element for the automatic detection of manipulated face images. We also design a lightweight architecture with the correct structural biases for extracting such features and derive a multitask training scheme that consistently outperforms image class supervision alone. The trained networks achieve state-of-the-art results in the *FaceForensics++* dataset using significantly reduced number of parameters and are shown to work well in detecting fully generated face images.

1. INTRODUCTION

After decades of exposure to successfully manipulated image content people no longer think of images as self-evidently portraying an accurate representation of reality. Up until recently, video content has been thought of as a reliable source of information, given that realistic tampering required a considerable amount of resources and in most cases was easily identifiable. However, this has changed with the rise of automated techniques enabling successful facial manipulations. While these techniques have the capacity to kickstart a revolution in computer graphics and digital content creation, if misapplied can definitely have applications of malevolent abuse and severe negative impacts on human rights standards (e.g. privacy, stigmatization, discrimination). In addition, previously successful image forensics methods do not generalize to the new techniques or to the artifacts produced from video compression which is commonly applied when a video is uploaded in social media platforms. Thus, it is critical to develop tools that help the automatic authentication of video content. In this paper we make the following contributions towards that goal:

- we propose that manipulation artifacts are local in nature, i.e. there is no high level semantic disparity between real and manipulated images, and are shared

across the manipulated regions of an image. Following these principles we design a Convolutional Neural Network outperforming all networks of similar size or smaller in the *FaceForensics++* (*FF++*) dataset

- we show that adding auxiliary dense classification tasks improves model performance significantly over the base for all examined cases and derive a training scheme with similar gains that does not require having real forgery masks. This leads to state-of-the-art results in the *FF++* dataset while at the same time the number of model parameters is reduced twenty-fold

2. RELATED WORK

Facial image manipulation techniques can broadly be categorized by two factors: 1. the type of manipulation, namely identity or facial expression, 2. whether the system uses computer vision based techniques or machine learning. *FaceSwap* [1] is a computer vision based technique for identity manipulation, while *Deepfakes* [2] use deep autoencoders to achieve the same task. *Face2Face* [3] is a computer vision system for live facial reenactment. Generative Adversarial Networks (GANs) [4] were introduced to model and produce new samples representative of a data distribution. Applied on human faces GANs can create highly realistic results in high resolution [5, 6, 7] and perform image-to-image translations for multiple domains showing impressive results in tasks such as altering the hair color, gender, age, skin tone and emotion of human faces.

Similarly manipulation detection techniques can be split between those employing handcrafted and learnt features. Image forensics and steganalysis focus on detecting a set of predefined manipulation artifacts that result from key image manipulation operations e.g. copy-move, splicing. They generally work great for a single type of forgery but fail in unseen cases and have been shown to deteriorate significantly in performance in the presence of video compression [8]. More recent literature includes many works using deep learning methods. End-to-end trained CNNs have been shown to generally outperform other approaches. Results from training and testing various models provided in literature [9, 10, 11]

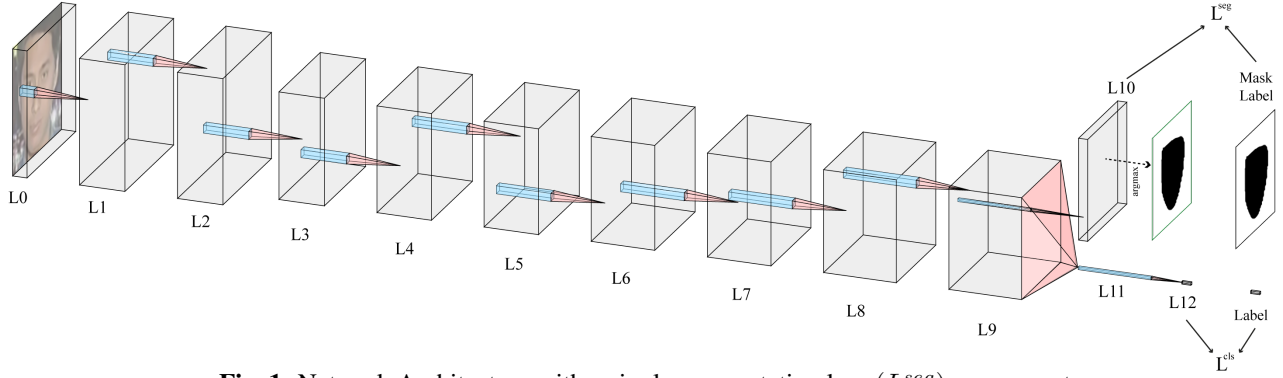


Fig. 1: Network Architecture with a single segmentation loss (L^{seg}) component

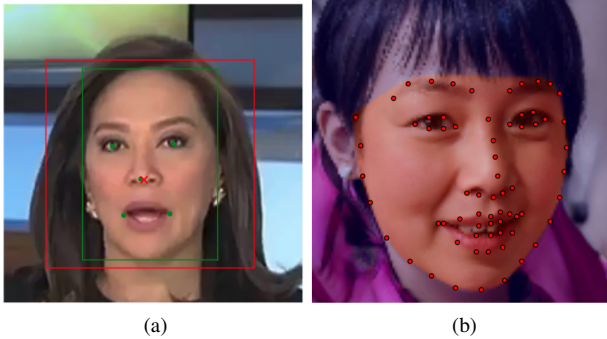


Fig. 2: Data Preprocessing. (a) Face crop. Output of face detector in green, extracted face roi in red, (b) *Convex Hull Mask*. Face landmarks and convex-hull region in red

is presented in [8]. All methods are shown to achieve high accuracy for uncompressed images but performance drops greatly when compression is introduced. In [12] the authors use temporal models on CNN features. For high compression a DenseNet architecture achieves state of the art precision on the binary classification task. The intuition that the final image rescaling and alignment steps common in various *deepfakes* pipelines should introduce face warping artifacts is further explored in [13] who propose a self-supervised technique bypassing the need for expensive positive examples. [14] take advantage of a bias of facial datasets not to include faces with their eyes closed to detect unrealistic blinking patterns in fake videos, while [15] detect head-pose inconsistencies. [16] build person specific models of their speaking mannerisms. Their models work well even in high compression but are person specific and thus can only be applied to specific individuals.

3. PROPOSED APPROACH

3.1. Data Preprocessing

Incorporating prior knowledge into the data preprocessing pipeline has been shown to improve the performance of trained models [8]. For all the experiments that follow data preprocessing involves training on centered face crops rather than full images as shown in Fig.2(a). When forgery masks

are not available, we extract masks artificially as: 1. *Zeros Mask* (ZM) when all image pixels are real, i.e. real images, 2. *Ones Mask* (OM) when all image pixels are generated, i.e. GAN images, 3. *Convex Hull Mask* (CVM) in which case we assign a positive class only to pixels inside the convex hull defined by facial landmarks (Fig.2(b)).

3.2. Model Architecture

In designing the model architecture presented below we follow the intuition that useful features for the classification problem at hand are non-global, non-semantic in nature. Visual examination of manipulated images does not offer sufficient clues. In a human study conducted in [8] humans were shown to be well inferior of most of the models tested, suggesting that successfully manipulated images cannot be identified based on appearance. Instead we assume that useful detection signal has two distinct characteristics: it can be detected in local image patches and it is consistent across the manipulated regions of an image. We design an architecture capable of capturing those features by respectively making the following choices: 1. we restrict the final layer receptive field to 33 pixels or ($\approx 1/16$ of input area), ensuring that no global features are used by the network, 2. we use *Global Average Pooling* (GAP) [17] across spatial locations for the final layer activations, forcing the model to extract features that are shared across image patches rather than patch specific features. The proposed network architecture is illustrated in Fig.1. It consists of a feature extractor with 8 convolutional and one *MaxPool* layers (layers 1-9) and two or more separate classifiers (layers 10, 12) followed by *SoftMax* to derive class probabilities. All convolution steps are followed by *ReLU* activation [18] and Batch Normalization [19]. Convolutional filter strides are set to 1 and *MaxPool* stride is 2 for downsampling. No padding is used throughout the network.

3.3. Loss Function

To further enforce the contribution of local features we use a joint classification-segmentation objective function. For the segmentation tasks we use real segmentation labels where available or construct approximations as discussed in section

3.1. To extract segmentation labels for training we choose the indices corresponding to pixels at the center of each final layer activation’s receptive field. For a mask label y_{ij} and predicted class probabilities p_{ij} at location (i, j) the segmentation loss is defined as the average *Cross-Entropy Loss* $H(y_{ij}, p_{ij})$ over all locations $(i, j) \in \{1, \dots, N\}$:

$$L^{seg} = \frac{1}{N^2} \sum_i \sum_j H(y_{ij}, p_{ij})$$

At the same time we apply GAP and a linear layer followed by *SoftMax* to derive class probabilities for the image p . Similarly, these are used to define a *Cross-Entropy Loss* for labels y and predicted probabilities p :

$$L^{cls} = H(y, p)$$

The joint loss function for the model is defined for k segmentation objectives:

$$L = \lambda^{cls} L^{cls} + \sum_k \lambda_k^{seg} L_k^{seg}$$

using hyperparameters:

$$\lambda^{cls}, \lambda_k^{seg} : 0 \leq \lambda^{cls}, \lambda_k^{seg} \leq 1.0, \lambda^{cls} + \sum_k \lambda_k^{seg} = 1.0$$

4. EXPERIMENTS

4.1. Datasets

In the ablation study presented in section 4.2 we use the *FF++* dataset [8] which extends *FaceForensics* [20] to include samples generated from *Deepfakes* [2] and *FaceSwap* [1] containing 1000 videos per manipulation method. Additionally, in section 4.3 we use the *DFDD* [21], *DFDC* [22] and *Celeb-DF* [23] large scale deepfakes datasets. For the GAN images detection experiments, pretrained networks based on [5] and [6] were used to generate images of faces. From each technique 70k faces were selected such that a face could be found using a reasonably high detection threshold in a pretrained face detector [24] and these were split into 56k, 7k, 7k train, validation and test sets respectively. For real images the *Flickr-Faces-HQ* (FFHQ) dataset [6] of human faces was used. In order to examine the effect of compression on the detection performance all images were compressed using H.264 coding at levels similar to the ones used in the *FF++* dataset.

4.2. Effect of Forgery Localization Loss

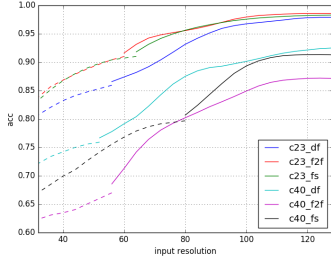
To assess the contribution of forgery localization on classification performance we formulate our loss function using a single segmentation objective: to predict manipulated pixels. We perform an ablation study varying λ_{seg} while using

Table 1: Binary Classification Accuracy - (TOP) Manipulation Method-specific Training, (BOTTOM) Training with All Manipulation Methods

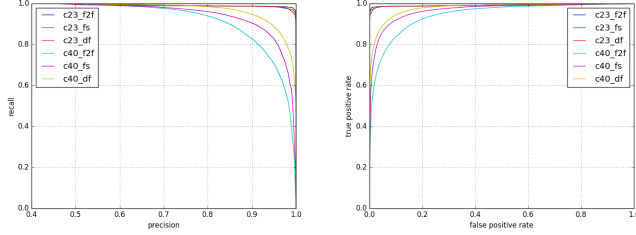
λ^{seg}	C23 Compression			C40 Compression		
	DF	F2F	FS	DF	F2F	FS
0.0	96.80	97.72	97.57	91.60	84.47	89.72
0.2	97.90	98.37	97.90	91.68	84.88	89.69
0.3	97.86	98.58	98.32	92.40	86.32	89.77
0.4	97.86	98.44	98.10	91.95	86.20	90.15
0.5	97.75	98.43	98.24	91.80	86.40	90.23
0.6	97.55	98.45	98.09	91.83	87.11	90.56
0.7	97.51	98.38	98.17	91.71	86.81	91.26
[10]	90.18	94.93	93.14	80.95	77.30	76.83
[11]	95.26	95.84	93.43	89.52	84.44	83.56
[8]	<u>98.85</u>	98.36	98.23	94.28	91.56	93.70
[12]	-	-	-	<u>96.70</u>	<u>93.21</u>	<u>95.80</u>
0.0	94.54	94.73	94.16	85.92	82.89	83.49
0.2	95.64	96.28	95.41	86.44	83.05	84.27
0.3	95.77	96.23	95.97	85.96	83.96	84.82
0.4	95.81	96.79	95.51	87.06	83.52	84.55
0.5	96.14	96.54	95.96	86.23	83.74	84.19
0.6	96.78	96.91	95.97	85.88	84.70	85.17
0.7	96.65	97.13	95.92	86.83	84.98	84.77
[10]	90.25	93.96	87.74	86.93	83.66	74.28
[11]	89.55	88.60	81.24	80.43	69.06	59.16
[8]	<u>97.49</u>	<u>97.69</u>	<u>96.79</u>	<u>93.36</u>	<u>88.09</u>	<u>87.42</u>

$\lambda_{seg} = 0.0$ as the base case. Test set classification accuracies on *FF++* can be seen in Table 1. Overall there is a noticeable performance boost with up to +2.6% improvement over the base case which is always the worst performer. We note that all models work well for *c23* compression and achieve state of the art results (underlined numbers) using a significantly smaller network ($\sim 1M$ compared to $\sim 22M$ trainable parameters). For higher compression we note a significant performance deterioration for all models. We also note a performance drop in every case when the training set includes samples from all manipulation methods as opposed to method-specific training. Because our model does not employ fully connected layers it can be trained at a specified resolution and use variable size input images during inference. In practice this is performed by cropping the 128×128 original input. Reducing image size leads to a graceful performance degradation which introduces a trade-off between inference speed and performance (Fig.3(a)). Dotted and solid lines correspond to using the mean probability of pixel forgery from the segmentation prediction - which works better in the small resolution regime - and the output of the image classifier. Additional performance metrics are presented in Fig.3(b-c).

Per class classification accuracies for detecting generated images can be seen in Table 2. We notice that all networks perform better in this task and that they perform worse for the more realistic *StyleGan* images. Here, including segmenta-



(a) Inference Classification accuracy with variable input size



(b) Precision-Recall

(c) ROC

Fig. 3: Manipulation detection - Test Metrics

tion loss using OMs was not found to improve results, however, using CVMs we see a noticeable jump in performance in every case.

4.3. Effect of Segmentation Objective

In section 4.2 we showed that networks trained to detect fully generated images benefit from learning to separate foreground face pixels from the background even though all pixels in these images are generated. To assess whether these benefits can be attributed to learning to detect manipulated pixels or to separate face pixels from background and how well CVMs can replace real forgery masks in the proposed training scheme we compare results from the following experiments using the *FF++*, *DFDD*, *DFDC* and *Celeb-DF* datasets: 1. base case from section 4.2, 2. best model from section 4.2, 3. use CVMs instead of real forgery masks to predict manipulated pixels, 4. use CVMs to predict face region instead of manipulated pixels, 5. use CVMs to predict face region and real masks to predict manipulated pixels ($k = 2$). 6. use CVMs to predict face region and manipulated pixels ($k = 2$), 7. use real masks to predict face region and manipulated pixels ($k = 2$). Results are presented in Table 3. We note that learning to detect manipulated pixels appears to be the more important task, however, including both segmentation objectives is shown to work better overall. In every case including a segmentation component in the loss function boosts classification performance, except for case 4 where for the more challenging *DFDC* and *Celeb-DF* datasets it underperforms the base case. Surprisingly, using CVMs for segmenting both face and manipulated pixels appears to work

Table 2: Binary Classification Accuracy - (TOP) Manipulation Method-specific Training, (BOTTOM) Training with All Manipulation Methods. (LEFT) OM, (RIGHT) CVM

λ^{seg}	C23 Compression				C40 Compression			
	PGAN	STYLEGAN	PGAN	STYLEGAN	PGAN	STYLEGAN	PGAN	STYLEGAN
0.0	99.96	99.96	99.55	99.55	99.76	99.76	97.66	97.66
0.2	99.97	99.97	99.55	99.77	99.77	99.87	97.59	98.31
0.3	99.98	99.99	99.59	99.78	99.81	99.93	97.60	98.68
0.4	99.96	99.99	99.55	99.87	99.73	99.93	97.32	98.28
0.5	99.94	99.99	99.52	99.87	99.78	99.91	97.28	98.51
0.6	99.97	99.99	99.56	99.90	99.74	99.95	97.43	98.56
0.7	99.94	99.98	99.50	99.85	99.74	99.95	97.27	98.52
0.0	99.66	99.66	99.23	99.23	97.31	97.31	95.04	95.04
0.2	99.62	99.78	99.32	99.54	97.21	97.79	94.78	96.42
0.3	99.63	99.88	99.31	99.60	97.33	98.59	94.89	97.46
0.4	99.65	99.86	99.32	99.68	97.02	98.21	94.94	96.98
0.5	99.53	99.87	99.33	99.70	97.17	98.27	94.75	96.93
0.6	99.61	99.86	99.20	99.80	96.99	98.41	94.84	97.29
0.7	99.65	99.89	99.24	99.79	97.03	98.53	94.89	97.28

Table 3: Binary Classification Accuracy - Different Segmentation Objectives

Case #	Mask		Dataset			
	Face	Fake	FF-DF	DFDD	DFDC ¹	Celeb-DF
1	None	None	96.80	94.41	85.76	91.05
2	None	Real	97.90	97.14	-	92.57
3	None	CVM	97.67	96.82	88.20	92.29
4	CVM	None	97.20	95.30	83.26	90.79
5	CVM	Real	97.74	96.92	-	91.42
6	CVM	CVM	98.03	97.03	88.76	92.62
7	Real	Real	97.97	96.16	-	91.50

better than using real segmentation masks for both tasks. This could be explained by the fact that CVMs cover a larger area than real masks and thus cover the information rich boundary between manipulated and non-manipulated pixels. If a single objective were to be selected that would be case 2 which was the training scheme used in experiments presented in section 4.2.

5. CONCLUSION

In this paper we examined local features in the task of detecting manipulated images of human faces and proposed a convolutional architecture and training scheme that target finding such features. Adding dense localization components in the objective function was shown to boost classification performance. Using that approach we achieved a strong improvement over our baselines and state-of-the-art results in the *FF++* dataset. Finally, we proposed a methodology for automatically generating all required training labels eliminating the need for real-fake video pairs.

¹No real forgery masks are provided for *DFDC*

6. REFERENCES

- [1] M. Kowalski, *FaceSwap*, <https://github.com/MarekKowalski/FaceSwap>.
- [2] *Deepfakes github*, <https://github.com/deepfakes/faceswap>.
- [3] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Niessner, “Face2face: Real-time face capture and reenactment of rgb videos,” *Commun. ACM*, vol. 62, no. 1, pp. 96–104, Dec. 2018.
- [4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., pp. 2672–2680. Curran Associates, Inc., 2014.
- [5] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, “Progressive growing of gans for improved quality, stability, and variation,” *CoRR*, vol. abs/1710.10196, 2017.
- [6] Tero Karras, Samuli Laine, and Timo Aila, “A style-based generator architecture for generative adversarial networks,” *CoRR*, vol. abs/1812.04948, 2018.
- [7] <https://thispersondoesnotexist.com/>.
- [8] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner, “Faceforensics++: Learning to detect manipulated facial images,” *CoRR*, vol. abs/1901.08971, 2019.
- [9] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, “Distinguishing computer graphics from natural images using convolution neural networks,” in *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, Dec 2017, pp. 1–6.
- [10] Belhassen Bayar and Matthew C. Stamm, “A deep learning approach to universal image manipulation detection using a new convolutional layer,” in *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, New York, NY, USA, 2016, IH&MMSec ’16, pp. 5–10, ACM.
- [11] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen, “Mesonet: a compact facial video forgery detection network,” *CoRR*, vol. abs/1809.00888, 2018.
- [12] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan, “Recurrent convolutional strategies for face manipulation detection in videos,” *CoRR*, vol. abs/1905.00582, 2019.
- [13] Yuezun Li and Siwei Lyu, “Exposing deepfake videos by detecting face warping artifacts,” *CoRR*, vol. abs/1811.00656, 2018.
- [14] Y. Li, M. Chang, and S. Lyu, “In ictu oculi: Exposing ai created fake videos by detecting eye blinking,” in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, Dec 2018, pp. 1–7.
- [15] Xin Yang, Yuezun Li, and Siwei Lyu, “Exposing deep fakes using inconsistent head poses,” *CoRR*, vol. abs/1811.00661, 2018.
- [16] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li, “Protecting World Leaders Against Deep Fakes,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Long Beach, CA, June 2019, p. 8, IEEE.
- [17] Min Lin, Qiang Chen, and Shuicheng Yan, “Network in network,” 2013, cite arxiv:1312.4400Comment: 10 pages, 4 figures, for iclr2014.
- [18] Vinod Nair and Geoffrey E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, USA, 2010, ICML’10, pp. 807–814, Omnipress.
- [19] Sergey Ioffe and Christian Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *CoRR*, vol. abs/1502.03167, 2015.
- [20] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner, “Faceforensics: A large-scale video dataset for forgery detection in human faces,” *CoRR*, vol. abs/1803.09179, 2018.
- [21] Per Karlsson Alexey Vic-tor Vorbyov Thomas Leung Jeremiah Childs Nicholas Dufour, Andrew Gully and ChristophBregler, “Deepfakes detection dataset by google and jigsaw,” 2019.
- [22] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer, “The deepfake detection challenge (dfdc) preview dataset,” 2019.
- [23] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu, “Celeb-df: A new dataset for deepfake forensics,” 2019.

- [24] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou, “Retinaface: Single-stage dense face localisation in the wild,” *CoRR*, vol. abs/1905.00641, 2019.