Detecting Photoshopped Faces by Scripting Photoshop

ICCV 2019
Sheng-Yu Wang, Oliver Wang, Andrew Owens, Richard Zhang, Alexei A. Efros

Presenter: Seunghwan Choi

DAVIAN Vision Paper Study

2020/03/09

Index

Introduction

- The popular press has mostly focused on “DeepFakes” and other GAN-based methods
  - Such methods may one day be able to convincingly simulate a real person
  - For now, they are prone to degeneracies and exhibit visible artifacts

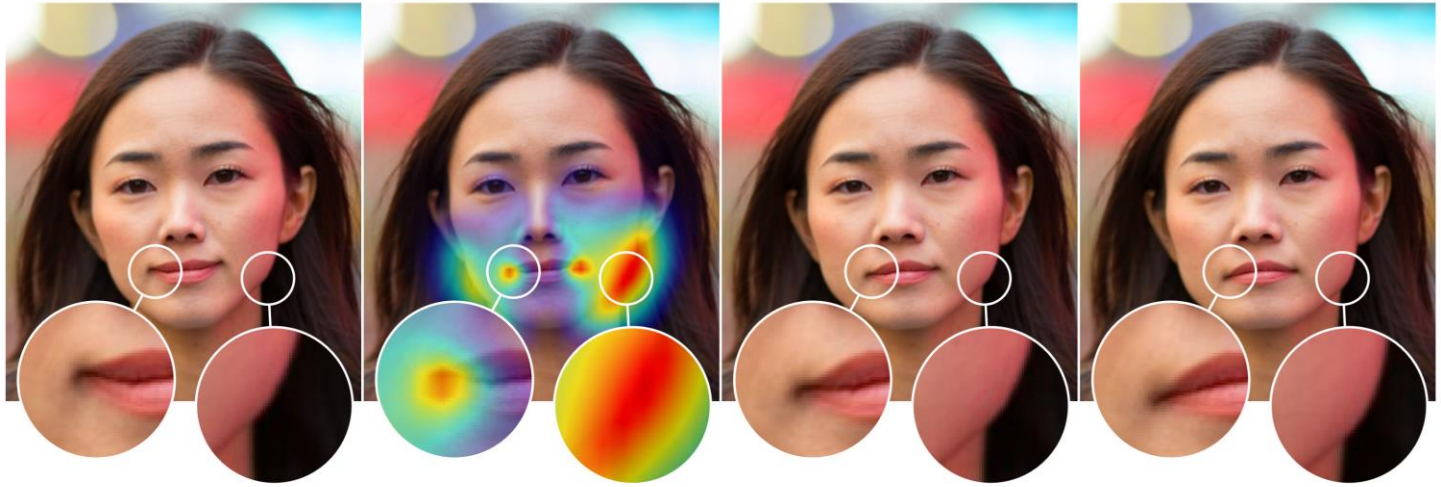
- Rather, the more subtle image manipulations have been the largest contributors to the proliferation of manipulated visual content
  - Most malicious photo manipulations are created using standard image editing tools, such as Adobe Photoshop

- The authors focus on image warping applied to faces

Introduction Contributions

- The authors present a method for detecting Photoshop manipulation

- The authors show that the model outperforms humans at the task of recognizing manipulated images
  - The model can predict the specific location of edits
  - The model can be used to "undo" a manipulation to reconstruct the unedited image



(a) Manipulated photo (b) Detected manipulations (c) Suggested "undo" (d) Original photo

Figure 1: Given an input face (a), our tool can detect that the face has been warped with the Face-Aware Liquify tool from Photoshop, predict where the face has been warped (b), and attempt to "undo" the warp (c) and recover the original image (d).

Dataset

- The authors script the Face-Aware Liquify(FAL) tool in Adobe Photoshop
  - FAL represents manipulation using 16 parameters
  - Each parameter corresponds to higher-level semantics
  - Ex) adjusting the width of the nose, eye distance, chin height, etc.

	Train	Val	Test
Source	OpenImage & Flickr		Flickr
Total Images	1.1M	10k	100
Unmanipulated images	157k	5k	50
Manipulated images	942k	5k	50
Manipulations	Random FAL		Pro Artist

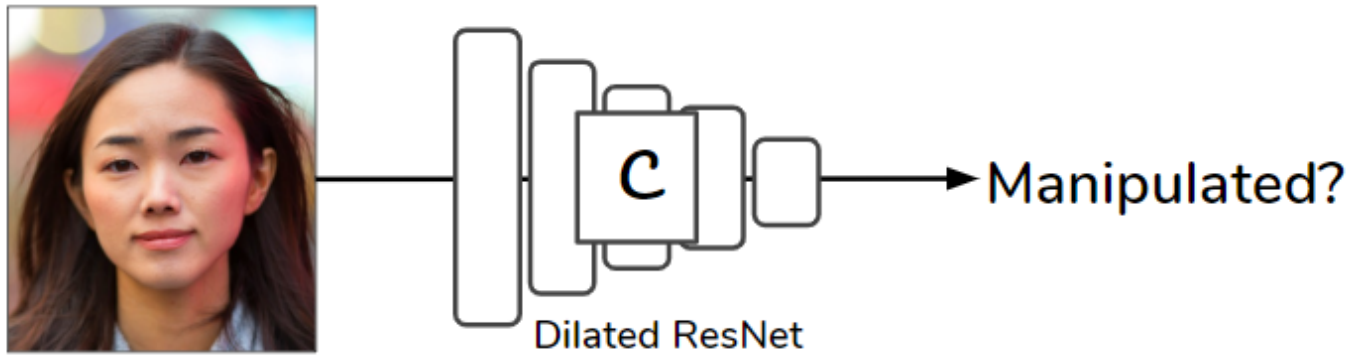
Table 1: **Dataset statistics.** This includes our own automatically created data as well as a smaller test set of manipulations created by a professional artist.

(a) Real images

(b) Manipulated images

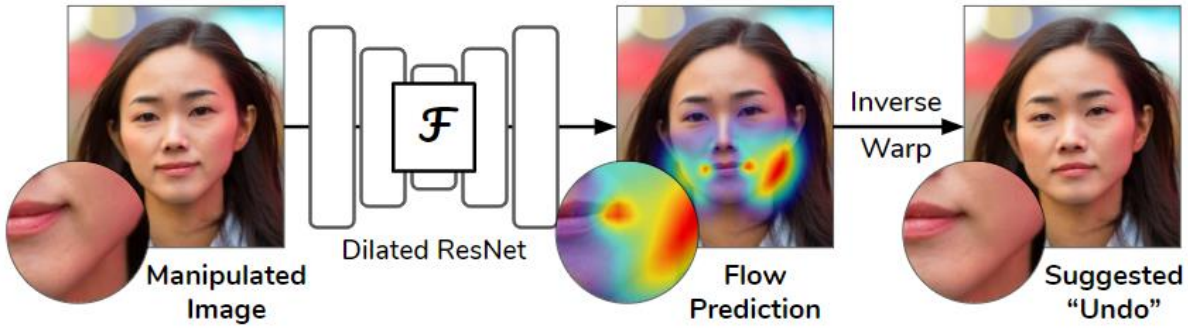
Methods Binary Classification

- Two models are presented to detect facial manipulations
  - A global classification model
  - A local warp predictor
- A binary classifier is trained to address the question "has this image been manipulated?"
  - The model adopts a Dilated Residual Network variant (DRN-C-26)
  - To increase robustness, more aggressive data augmentation is considered



Methods Warping Field Prediction

- Notations
  - F: a flow prediction model
  - X_{orig}: an original image
  - X: a warped image
  - U: an approximate ground truth flow field



- Training loss
  - $\mathcal{L}_{epe}(\mathcal{F}) = ||M \odot \big(\mathcal{F}(X) - U\big)||_2$
  - $\mathcal{L}_{ms}(\mathcal{F}) = \sum_{s \in S} \sum_{t \in \{x,y\}} ||M \odot \big(\nabla_t^s(\mathcal{F}(X)) - \nabla_t^s(U)\big)||_2$
  - $\mathcal{L}_{rec}(\mathcal{F}) = ||\mathcal{T}\big(X; \mathcal{F}(X)\big) - X_{orig}||_1$
  - $\mathcal{L}_{total} = \lambda_e \mathcal{L}_{epe} + \lambda_m \mathcal{L}_{ms} + \lambda_r \mathcal{L}_{rec}$

Experiments

Algorithm			Validation (Random FAL)					Test (Professional Artist)				
Method	Resolution	with Aug?	Accuracy			AP	2AFC	Accuracy			AP	2AFC
			Total	Orig	Mod			Total	Orig	Mod		
Chance	–	–	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Human	–	–	–	–	–	–	53.5	–	–	–	–	71.1
FaceForensics++ [31]	–	–	51.3	86.3	16.2	52.7	–	50.0	85.7	14.3	55.3	61.9
Self-consistency* [15]	–	–	–	–	–	53.7	–	–	–	–	56.4	72.0
Low-res no aug.	400		97.0	97.2	96.9	99.7	99.5	89.0	86.0	92.0	96.8	**98.0**
Low-res with aug.	400	✓	93.7	91.6	95.7	98.9	98.9	83.0	74.0	92.0	94.4	96.0
High-res with aug.	700	✓	**97.1**	99.8	94.5	**99.8**	**100.0**	**90.0**	96.0	84.0	**97.4**	**98.0**

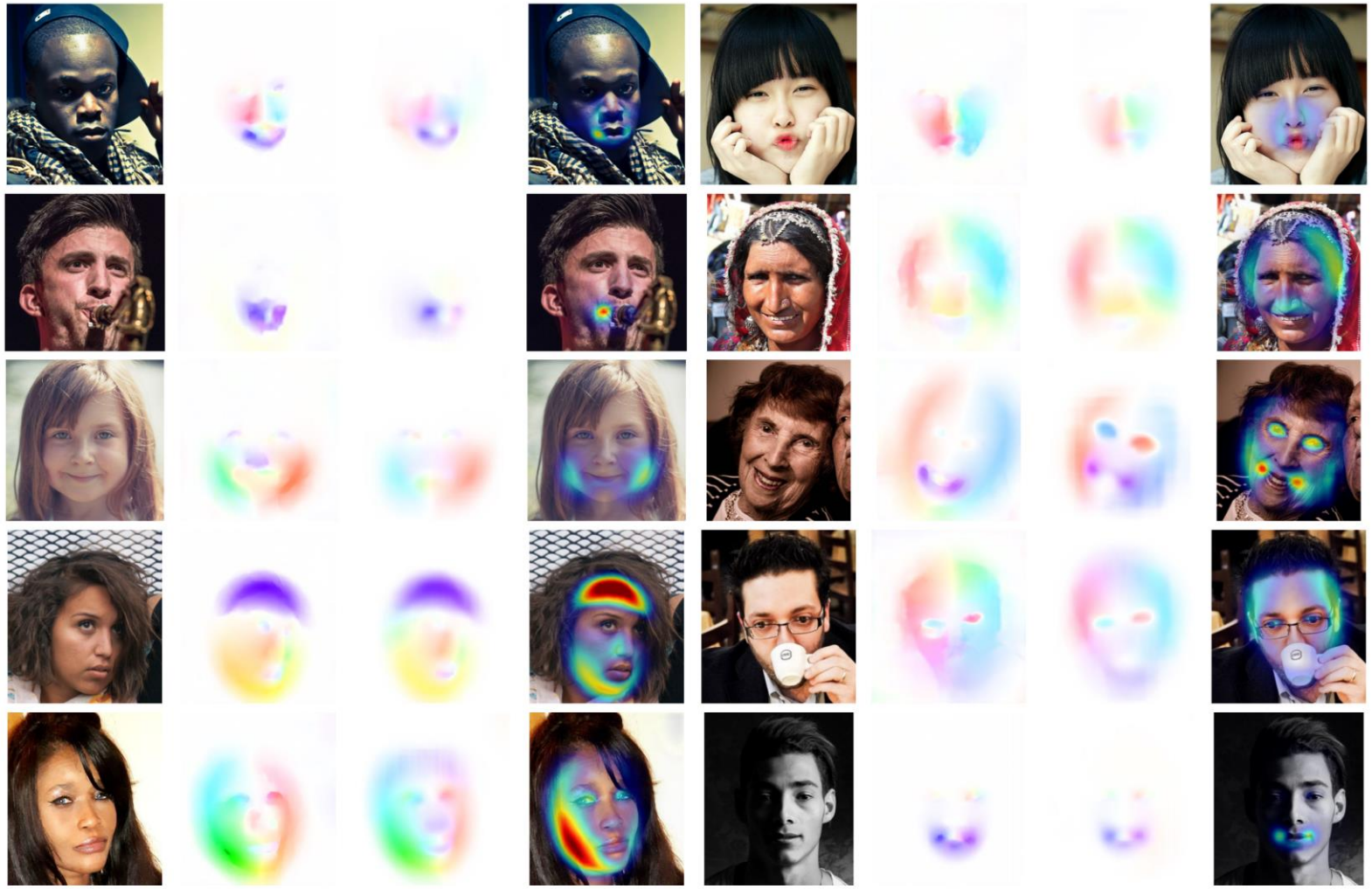
Table 2: **Real-or-fake classifier performance.** We tested models with FAL warping applied both by automated scripting and a professional artist. We observe that training with high-resolution inputs performs the best among the three. In addition, training without augmentation performs better in this domain, but adding augmentation makes the model more robust to corruptions, both within and outside of the augmentation set (see Appendix A3). *Self-consistency was tested on a 2k random subset of the validation set due to running time constraints.

Experiments

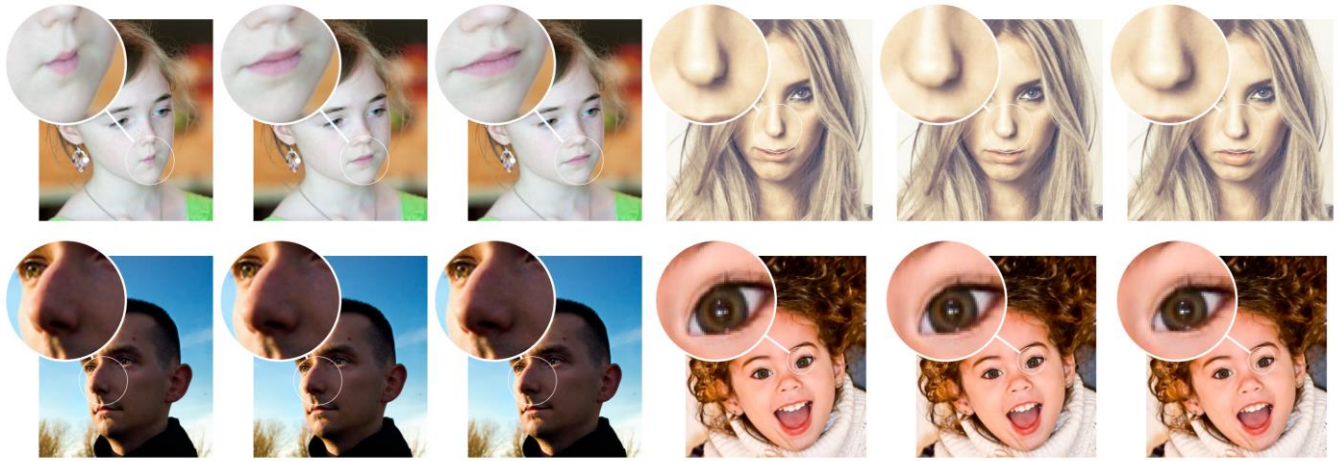
	Losses			Face-Aware Liquify (FAL)						Other Manipulations					
				Val (Rand-FAL)			Artist-FAL			Artist-Liquify			Portrait-to-Life [6]		
	EPE	Multi-scale	Pix ℓ_1	EPE ↓	IOU-3 ↑	ΔPSNR ↑	EPE ↓	IOU-3 ↑	ΔPSNR ↑	EPE ↓	IOU-3 ↑	ΔPSNR ↑	EPE ↓	IOU-3 ↑	ΔPSNR ↑
EPE-only	✓			**0.51**	**0.45**	+2.67	0.74	**0.33**	+2.09	0.63	**0.12**	-1.21	**1.74**	**0.42**	–
MultiG	✓	✓		0.53	0.42	+2.38	0.75	0.30	+2.07	0.59	0.11	-0.84	1.75	0.41	–
Full	✓	✓	✓	0.52	0.43	**+2.69**	**0.73**	0.28	**+2.21**	**0.56**	**0.12**	**-0.72**	**1.74**	0.40	–

Table 3: **Warping localization and undoing performance.** We show performance of our local prediction models across several evaluations: (1) EPE, which measures average flow accuracy, (2) IOU-3, which measures flow magnitude prediction accuracy and (3) ΔPSNR, which measures how closely the predicted unwarping recovers the original image from the manipulated; ↑, ↓ indicate if higher or lower is better. Our full method with all losses (flow prediction, multiscale flow gradient, and pixel-wise reconstruction) performs more strongly than ablations, both across datasets which use Face-Aware Liquify and other manipulations.

Experiments



Experiments

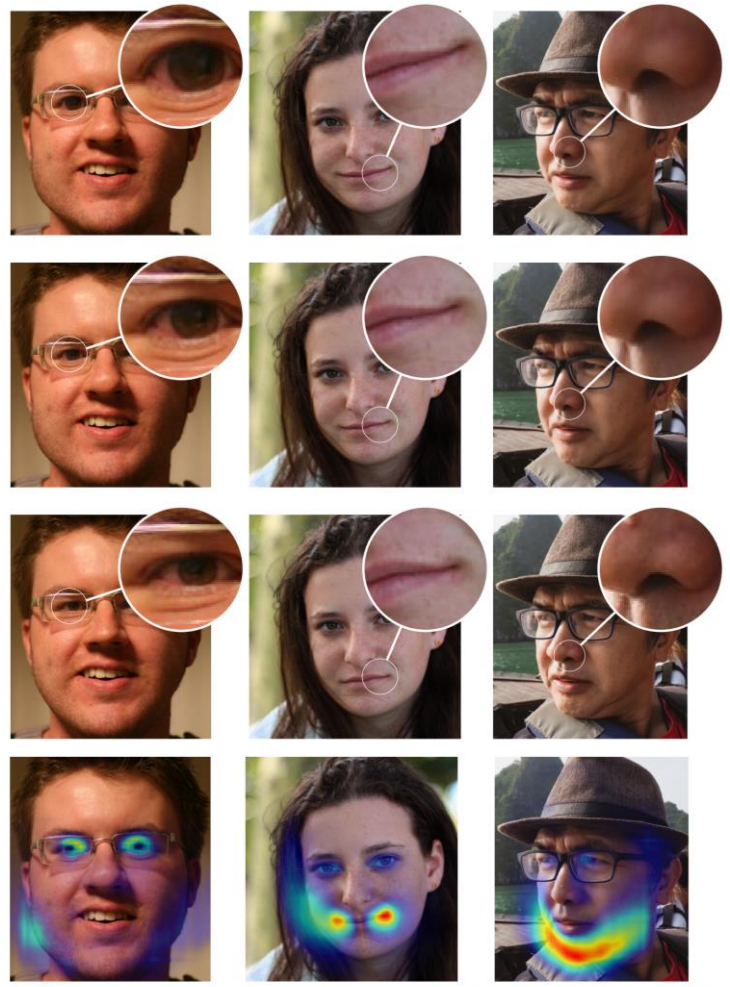


Manipulated Unwarped Original Manipulated Unwarped Original

Figure 4: **Unwarping results.** These images show results from the artist edited test dataset, where the manipulations are reversed by our model. Among other edits, the mouth and nose in the top row were expanded. In the bottom row, the nose shape was made less round and the eye was shrunk.

GT edits Predicted edits Unwarped diff

Figure 7: **Limitations**. When manipulations are too far outside the training distribution, as with the general Liquify tool experiment. Our local prediction model fails to correctly identify warped regions. This is visible in the overlay as well as in the unwarped image (difference to ground truth after unwarping is shown on the right, darker is worse).



Lens Studio Facetune Facetune (airbrushed)

Thank You

Presenter: Seunghwan Choi

DAVIAN Vision Paper Study

2020/03/09