Detecting Photoshopped Faces by Scripting

Sheng-Yu Wang Oliver Wang Andrew Owens Richard Zhang Alexei A. Efros
UC Berkeley Adobe Research

Kousshik Raj(17CS30022), Suman Pal(19BM6JP22)

IIT Kharagpur

Introduction

- The Problem
- Its Significance
- Proposal
- Dataset

The Problem

- Detecting warping of human faces
- Quite subtle, very hard to detect
- Human accuracy is only 53.5%

Significance of the Problem

- Numerous edited visual content
- When done without user consent, might lead to negative implications
- Body image issues
- Augments fake news

Solution Proposed

- A CNN trained entirely using fake images
- Detects warped faces
- Finding the location of the edits
- Offering possible undo

(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

- A large dataset of face images were obtained from Open Images dataset and Flickr which were then scripted by Face-Aware Liquify(FAL) tool in Adobe Photoshop to generate a variety of face manipulations.

- The FAL tool has 16 parameters which was randomly sampled to create manipulations and it was argued that randomly sampling the space will cover the the space of “realistic” operations.

	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

- Also to test the generalising ability of the model a real artist was contracted to manipulate 50 real photos.

Methodology

- Real or Fake?
- Local Warp Predictor
- Losses

Real or Fake Image?

A global classification model (Dilated Residual Network Variant (DRN-C-26)) have been used as a binary classifier for detecting real or fake images.

The base network of the DRN-C-26 for global classification is initialised with the weights pretrained on local detection task and fine tuned it for global classification task.

Effect of resolution has been studied by training the model with both high and low resolution images.

To increase the robustness of the model, various data augmentation methods including resizing methods, JPEG compression etc has been used.

Local Warp Predictor

Given a manipulated image, the local warp predictor is used to predict where the manipulation has occur and try to get back the original image by reversing it.

To achieve this, first a flow predictor is modelled to predict the “Ground truth” optical flow and thereby reverse it to get the unwarped image.

A DRN-C-26 network has been used here as a 121 per-pixel classifier to predict the optical flow of a manipulated image.

Given a original and manipulated image set, flow values are calculated and rounded off to nearest integer with a cut off value at 5. Hence the flow value u, v $\in$ [-5,5] i.e 121 possible values in total.

Losses

Flow Error (or End point Error): It gives the measure of how accurately the optical flow is predicted with respect “Ground truth” optical flow.

$$\mathcal{L}_{epe}(\mathcal{F}) = ||M \odot \big(\mathcal{F}(X) - U\big)||_2,$$

Here, U is the “ground truth” where as F(X) is the predicted optical flow given a X manipulated image. M is a binary mask applied to those pixels which fails forward-backward consistency test to avoid erroneous flow values.Loss for each pixel is computed and the mean is taken as the final flow error.

Multiscale Loss: It is encouraged to smoothen the flow by considering a multiscale loss on the flow gradient which is given by

$$\mathcal{L}_{ms}(\mathcal{F}) = \sum_{s \in S} \sum_{t \in \{x,y\}} ||M \odot \big(\nabla_t^s(\mathcal{F}(X)) - \nabla_t^s(U)\big)||_2$$

Reconstruction Loss: With the correct flow field being predicted, one can retrieve the original image by inverse warping which naturally gives rise to reconstruction loss given by

$$\mathcal{L}_{rec}(\mathcal{F}) = ||\mathcal{T}\big(X; \mathcal{F}(X)\big) - X_{orig}||_1,$$

where $\mathcal{T}(X;U)$ warps X by resampling with flow U.

The model has been jointly trained by three losses to get a more accurate result.

$$\mathcal{L}_{total} = \lambda_e \mathcal{L}_{epe} + \lambda_m \mathcal{L}_{ms} + \lambda_r \mathcal{L}_{rec}$$

where each losses are weighted by some coefficients. By grid search, appropriate coefficients of the weighted loss has been chosen.

Experiments

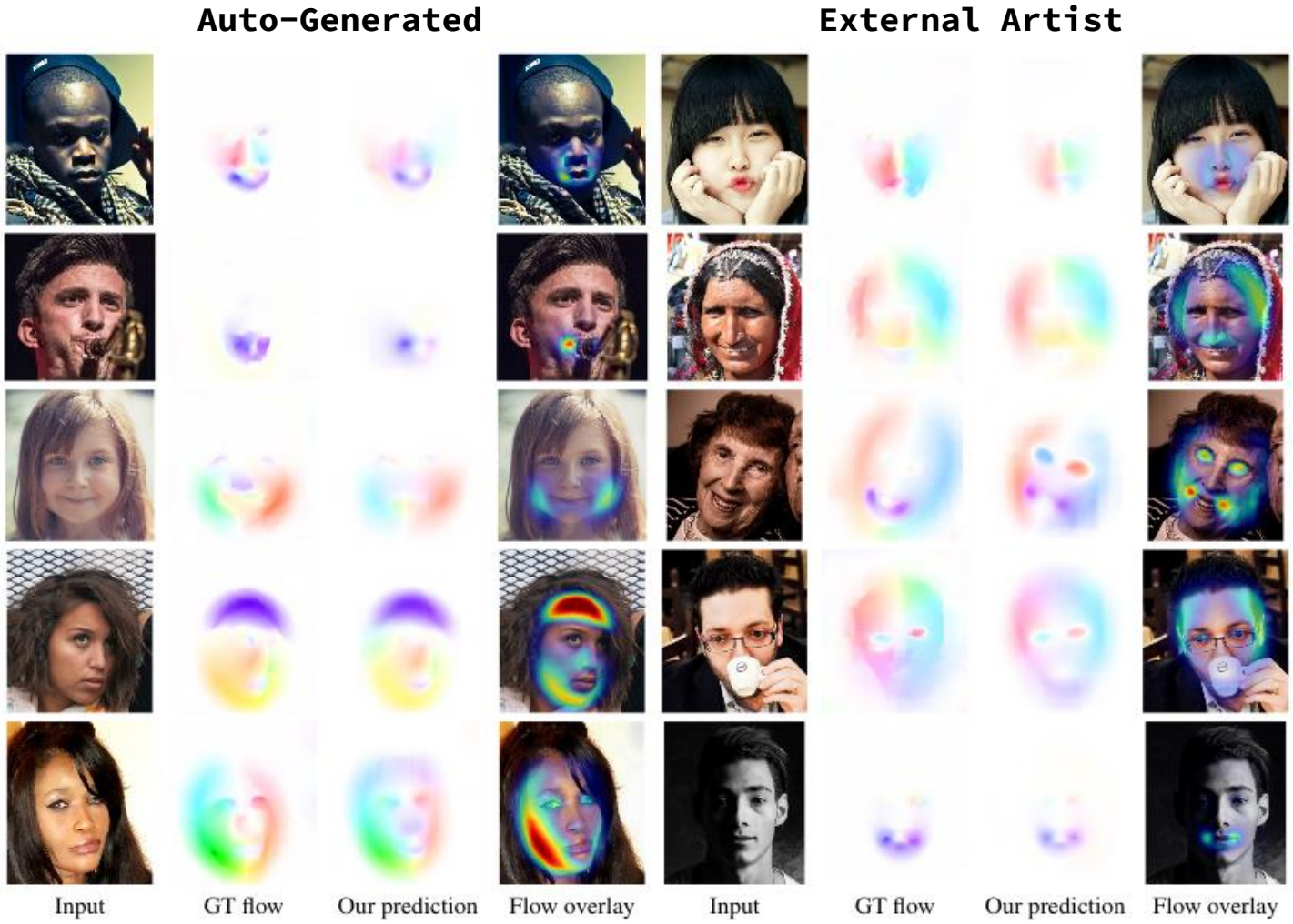
- Real or Fake
- Localising and Undoing Warp
- Other Manipulations

Real or Fake Classification

Algorithm			Validation (Random FAL)					Test (Professional Artist)				
Method	Resol-ution	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++	–	–	51.3	86.3	16.2	52.7	–	50.0	85.7	14.3	55.3	61.9
Self-consistency*	–	–	–	–	–	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**

AP = Area under the Precision vs Recall graph

2AFC = Fraction of time, it assigns a higher manipulation probability to fake, when both real and fake are given


Auto-Generated
External Artist
Input
GT flow
Our prediction
Flow overlay
Input
GT flow
Our prediction
Flow overlay

Localizing and Undoing Warp

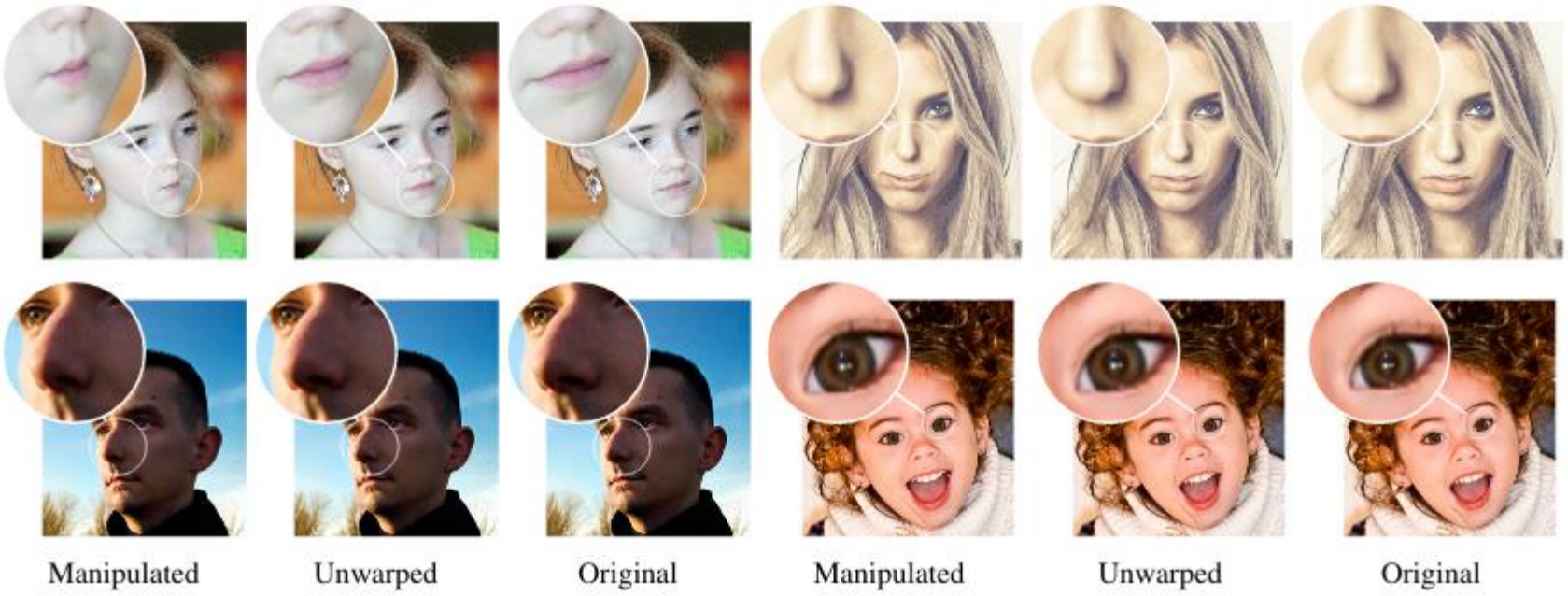
	Losses			Face-Aware Liquify (FAL)						Other Manipulations					
				Val (Rand-FAL)			Artist-FAL			Artist-Liquify			Portrait-to-Life		
	EPE	Multi-scale	Pix ℓ_1	EPE ↓	IOU-3 ↑	ΔPSNR ↑	EPE ↓	IOU-3 ↑	ΔPSNR ↑	EPE ↓	IOU-3 ↑	ΔPSNR ↑	EPE ↓	IOU-3 ↑	ΔPSNR ↑
EPE-only	✓			**0.51**	**0.45**	+2.67	0.74	**0.33**	+2.09	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	–

EPE = End Point Error Loss

IOU-3 = Intersection Over Union after applying a threshold of 3 to the flow magnitudes

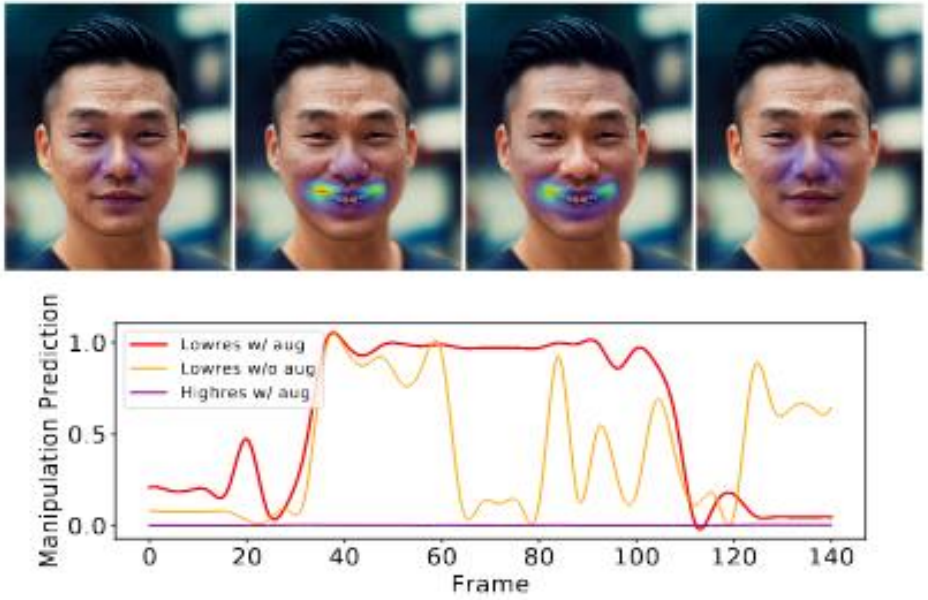
ΔPSNR = Similarity between unwrapped and original image

Artist edited Dataset



Other Manipulations

- **Puppeteering**



Future Works

Future Works

- The model can be generalized for various kinds of image editing by increasing the robustness of the dataset.
- The framework can be extended to body manipulations and photometric edits such as skin smoothening.
- Not suitable for detecting warping in big regions, can improve the multinomial pixel classification.

THANK YOU