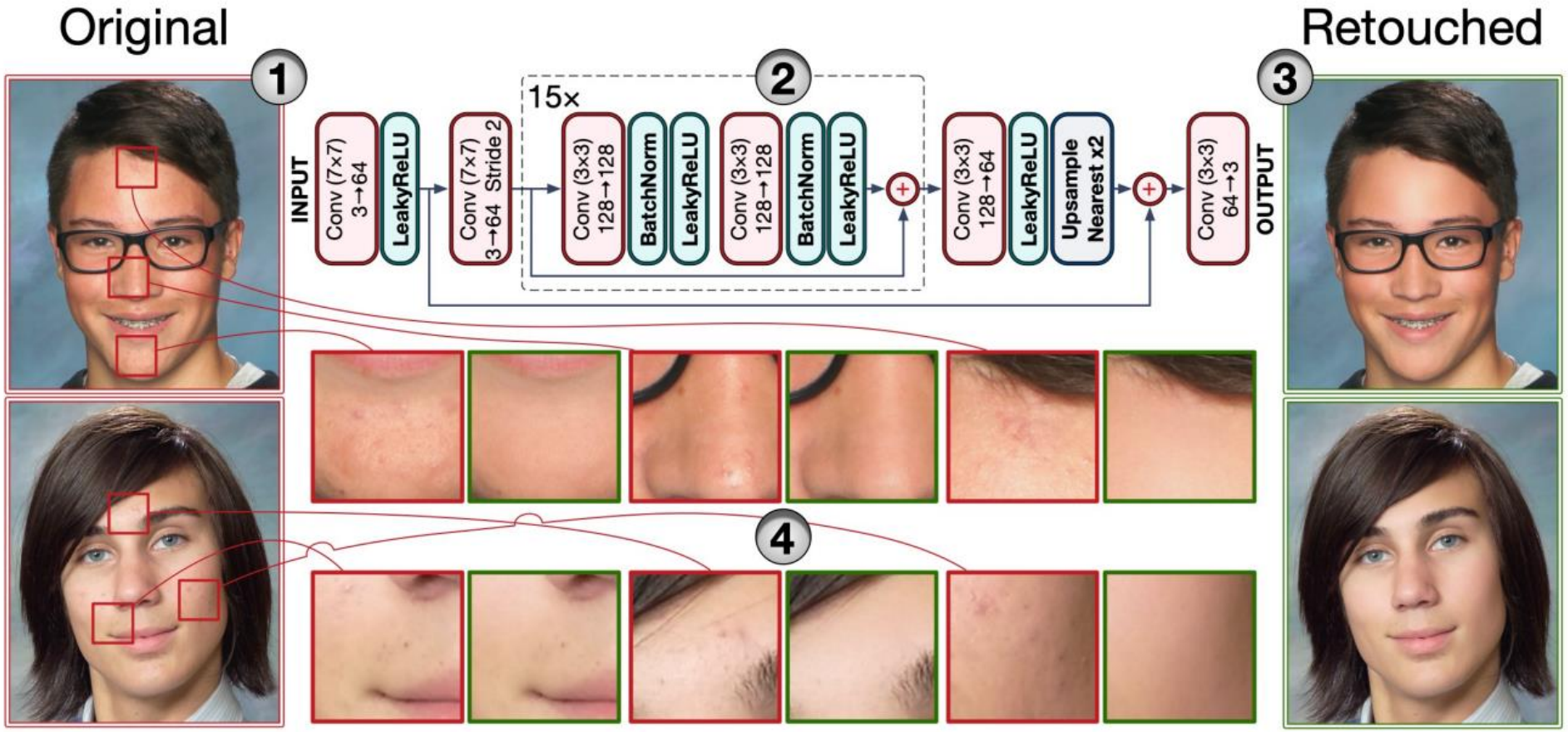
AutoRetouch: Automatic Professional Face Retouching

The University of British Columbia

WACV 2021

https://openaccess.thecvf.com/content/WACV2021/papers/Shafaei_AutoRetouch_Automatic_Professional_Face_Retouching_WACV_2021_paper.pdf

Goal


Original
Retouched
1
2
3
4
15×
INPUT
Conv (7×7) 3→64
LeakyReLU
Conv (7×7) 3→64 Stride 2
Conv (3×3) 128→128
BatchNorm
LeakyReLU
Conv (3×3) 128→128
BatchNorm
LeakyReLU
Conv (3×3) 128→64
LeakyReLU
Upsample Nearest x2
Conv (3×3) 64→3
OUTPUT

Motivations

- Photography studios process **thousands of high-quality portraits a day**. In the USA alone and within the niche market of school photography, studios process over **56 million photos annually**.
- On average, professional skin retouching requires **two minutes** of laborious work per portrait.
- One of the **most time-consuming** steps in a typical studio production pipeline is face retouching.
- Studios that we surveyed **were not happy with the quality of existing automation products**, instead, relying on human labor.

Face Retouching and Beautification tools

- Commercial tools (RELATED WORK):
  - Visage Labs
  - BeautyPlus
  - Meitu
  - Facetune
  - …
- But
  - None of them are **fully automatic**
  - None of them can provide **high-quality** and **high-resolution** face retouching
  - They just apply blind smoothing

Problems to be addressed

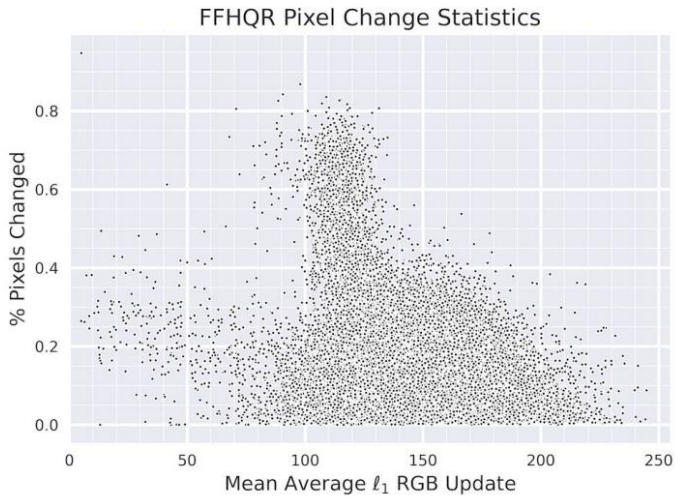
1) The **lack of publicly-available datasets** has stifled the research and development of working solutions

- We release the first large-scale face retouching dataset with our baseline: the *Flickr-Faces-HQ-Retouching* (FFHQR) Dataset

https://github.com/skylab-tech/ffhqr-dataset

FFHQR Dataset

- **70,000** 1 MP face-aligned images that are collected from *Flickr*
- Variety of **ages**, **ethnicity**, **lighting** conditions, and the large number of images that could benefit from face retouching
- We hired a team of professional image editors to retouch the images
- One of the key features of professional **retouching** is that the image updates are **sparse**, most of the pixels do not change
  - In the majority**, less than 40%** of the pixels are edited
  - Most pixel value changes are in the [100, 200] range
- The **studio data** contains **203,725** image pairs of original and retouched images



Problems to be addressed

2) **Input size** (to work with HR images)

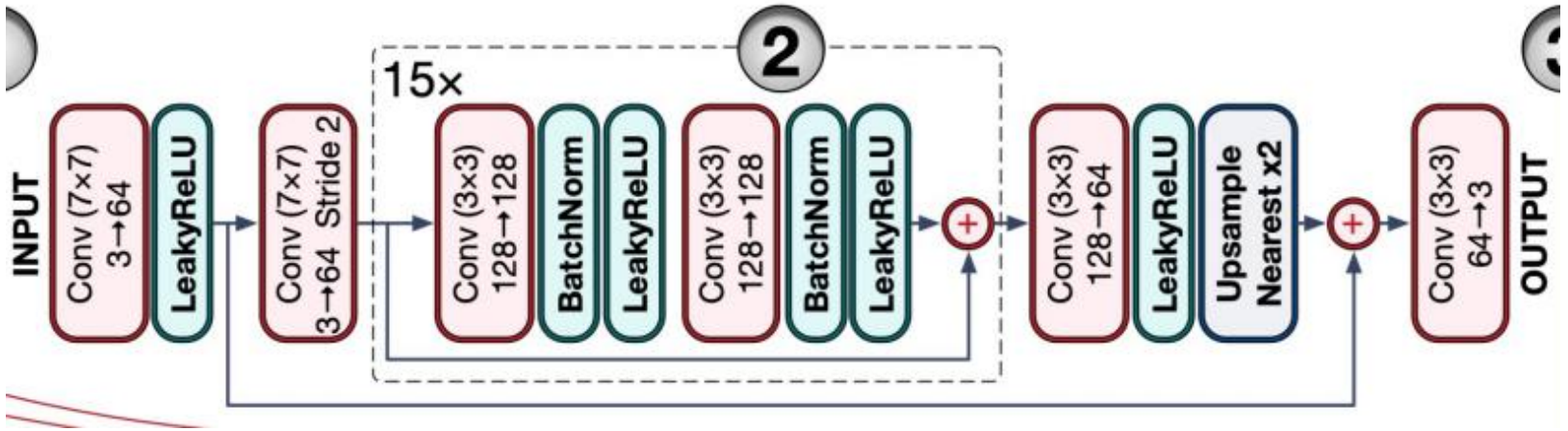
- **Downscaling** the image tensors to a factor smaller than half leads to slight pixel displacement and **loss of the original texture** in the output space.
  - This observation **eliminates most of the typical architectural** choices such as U-Net in Pix2Pix
  - **We limited the downscaling** in our network search to preserve fine details
- We observed that incorporating **additive skip-connection decreases the random jitters in the output**. Since we expect the majority of the pixels to remain unchanged, the skip connections **also allow for more direct transfer of the input to the output**.

Problems to be addressed

3) **GPU memory** (related to the previous one)

- The depth of the architecture and the number of kernels would typically become limited by the available GPU memory
- Face retouching is mostly a local operation – the correction of any pixel does not depend on the far away pixels
- This problem structure **enables a sliding-window strategy** that we exploit with convolutional architectures. Without loss of generality, we utilize more kernels and deeper architectures by limiting the input to a large-enough **w × w sub-window**.

Architecture



Loss function

$$L = \alpha \text{ RAGAN} + \beta \text{ Perceptual} + \gamma \text{ MSE}$$

$\alpha = 10^{-3}$, $\beta = 6 \cdot 10^{-3}$, $\gamma = 0.5$

- **RAGAN** (*Relativistic Average GAN*): running estimate of 300 previous predictions
- **Perceptual**: mean squared-error on the output features of the 14th layer in the 19-layer *VGG* mode
- **MSE**: L1 also works, but MSE is **faster**

RAGAN loss

- **More stable** and **faster training**, higher-quality and higher-resolution images
- The discriminator in the original GAN measures **the probability for a given real sample or a generated sample**.
- The author argues that key relative discriminant information between real data and generated data is missing in original GAN.
- The discriminator in RGAN takes into account that **how a given real sample is more realistic compared to a given random generated sample**

https://arxiv.org/pdf/1906.01529.pdf

Training

- We train the studio data model for approximately **two weeks** on three 2080 Ti GPUs and one 1080 Ti.
- We train a model on the FFHQR dataset in just **five days**.
- We run the training for **500 epochs** and set **w = 150 px**. We set the batch-size to 75, the learning rate of Adam to $2x10^{-4}$ for our model and 10^{-4} for RAGAN discriminator
- **56,000** images for training, **7000** images for validation, and **7000** images for testing.
- **Data augmentation**: mirror augmentation and color perturbation

Results

- **Neither PSNR nor SSIM are reliable performance indicators** for high-quality face retouching methods
- *Mean-Opinion-Score* (MOS)

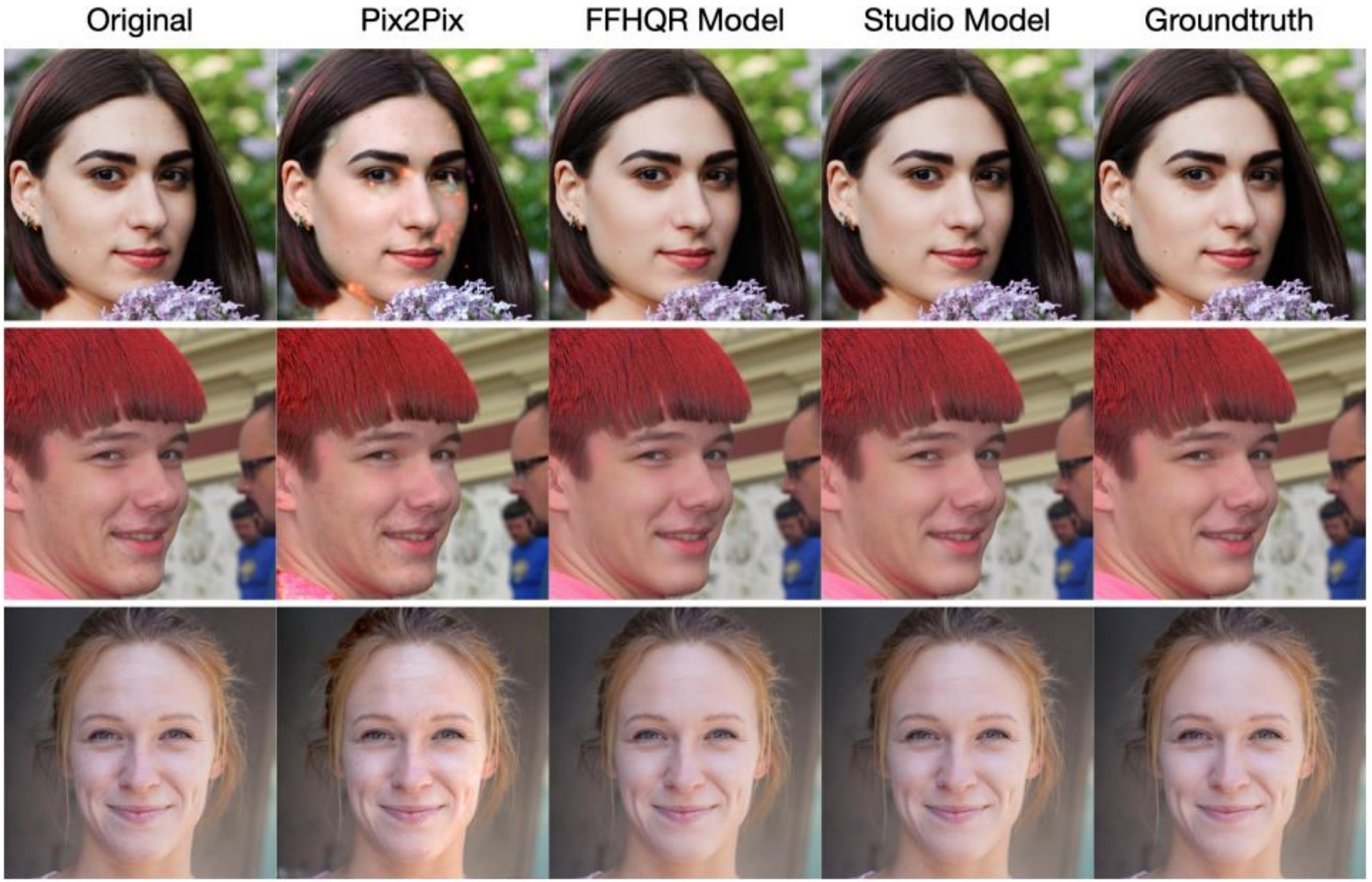
		Studio Data		FFHQR	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
	No edit (input → output)	39.12	0.9807	45.58	0.9938
(1)	MSE	39.58	0.9858	40.53	0.9876
	MSE + Perc.	39.41	**0.9875**	39.04	0.9889
	MSE + Perc. + RAGAN	**41.04**	0.9865	44.82	0.9944
	Pix2Pix [7]	28.12	0.8893	29.58	0.9224
(2)	MSE	40.01	0.9844	**45.88**	0.9949
	MSE + Perc.	39.88	0.9851	45.00	**0.9952**
	MSE + Perc. + RAGAN	39.65	0.9849	44.92	**0.9952**
	Pix2Pix [7]	30.29	0.9152	33.45	0.9585

Results

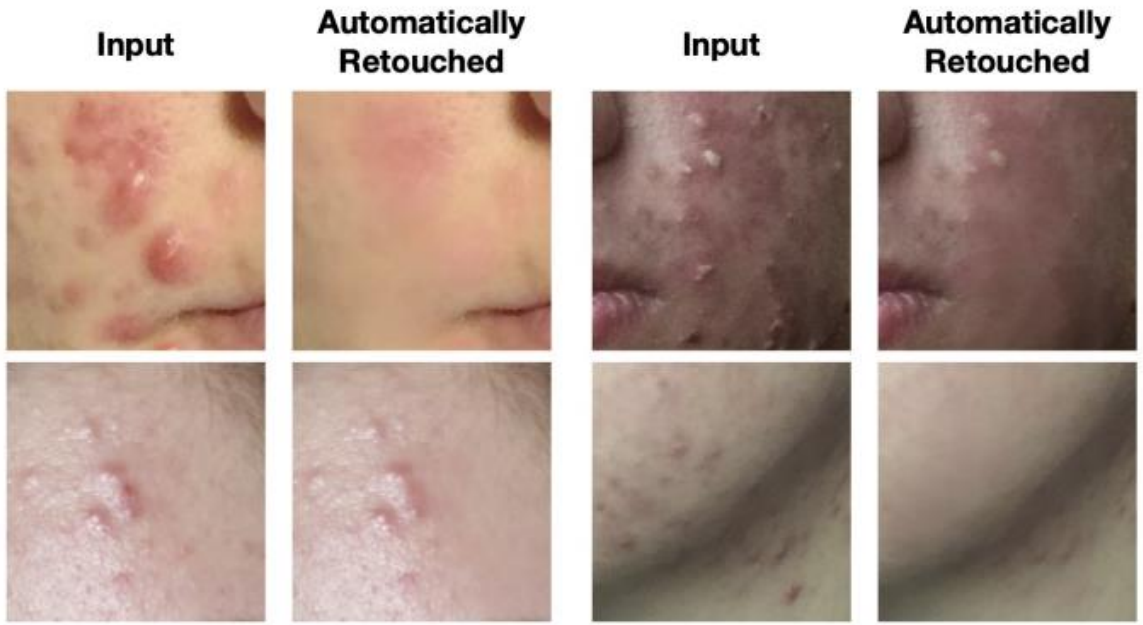
#	Method 1	Method 2	Voted 1 ↑		Voted 2 ↑		Undecided		n
			%	Median Time (s)	%	Median Time (s)	%	Median Time (s)	
(1)	MSE+Perc.+RAGAN	MSE	**50.1%**	28.0	17.9%	24.0	32.0%	26.3	730
	MSE+Perc.+RAGAN	MSE+Perc.	**28.3%**	30.9	22.9%	30.3	48.8%	26.1	791
	MSE+Perc.+RAGAN	Pix2Pix [7]	**100.0%**	1.9	0%	-	0%	-	508
(2)	MSE+Perc.+RAGAN	MSE	**94.6%**	12.2	4.4%	19.5	1.0%	19.6	802
	MSE+Perc.+RAGAN	MSE+Perc.	**95.7%**	9.8	3.4%	17.5	0.9%	15.7	898
	MSE+Perc.+RAGAN	Pix2Pix [7]	**99.9%**	1.3	0.1%	0.9	0%	-	1008
	MSE+Perc.+RAGAN	Groundtruth	**76.3%**	8.4	18.7%	12.0	5.0%	19.0	672

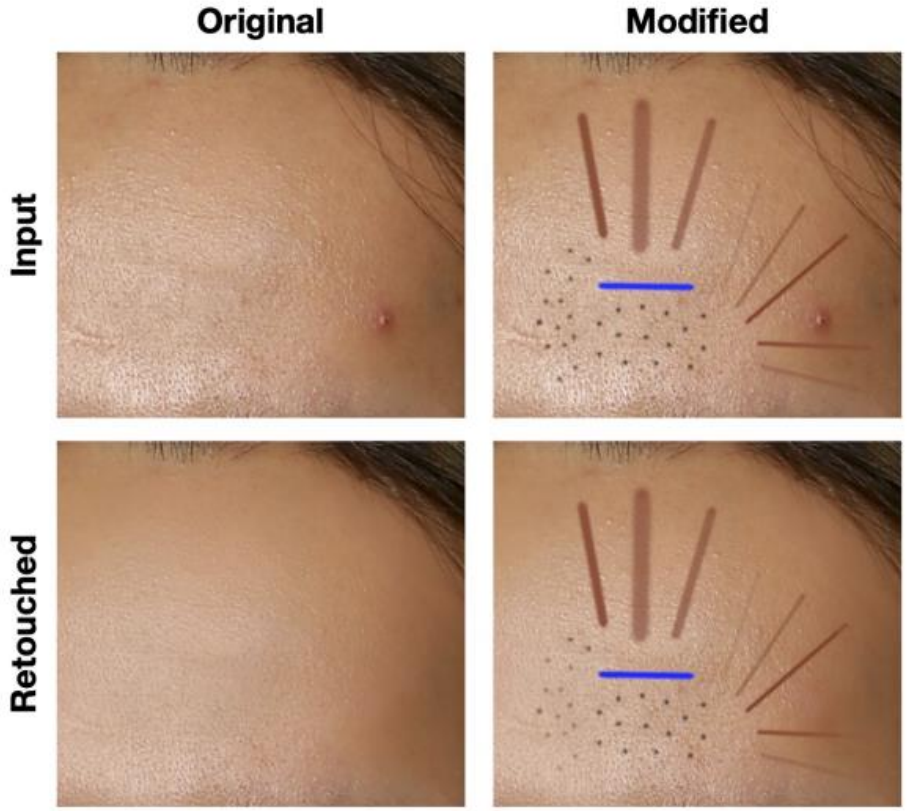
When we add the RAGAN loss, **the network (seemingly) learns to minimize the output space changes**, which results in minimal retouching that preserves the input details as much as possible.

Results



Results


Input
Automatically Retouched
Input
Automatically Retouched


Original
Modified
Input
Retouched

Issues

- **Code has not been yet released**, even though authors have stated it
- It seems the method is based on patches of the face during the training and on the whole face during testing:
  - Is it possible?

- To the best of our knowledge, there is no scalable and objective measure of quality for our application at the time of writing

From authors

- **One model** for all patches.
- We randomly sample window patches from the input image. We randomly sample **around the head**, so sometimes we sample from the background too.
- **The architecture is fully convolutional**. Once the training is over, **we can feed the entire image through the network** and it'll produce the corresponding output.