[Re] Explaining in Style: Training a GAN to explain a classifier in StyleSpace

Chase van de Geijn, Victor Kyriacou
Irene Papadopoulou, Vasiliki Vasileiou

Graduate School of Informatics
Universiteit van Amsterdam

Goal

Reproduce

- "Explaining in style: Training a gan to explain a classifier in stylespace."
  by Lang O. et al.
  In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 693-702. 2021.

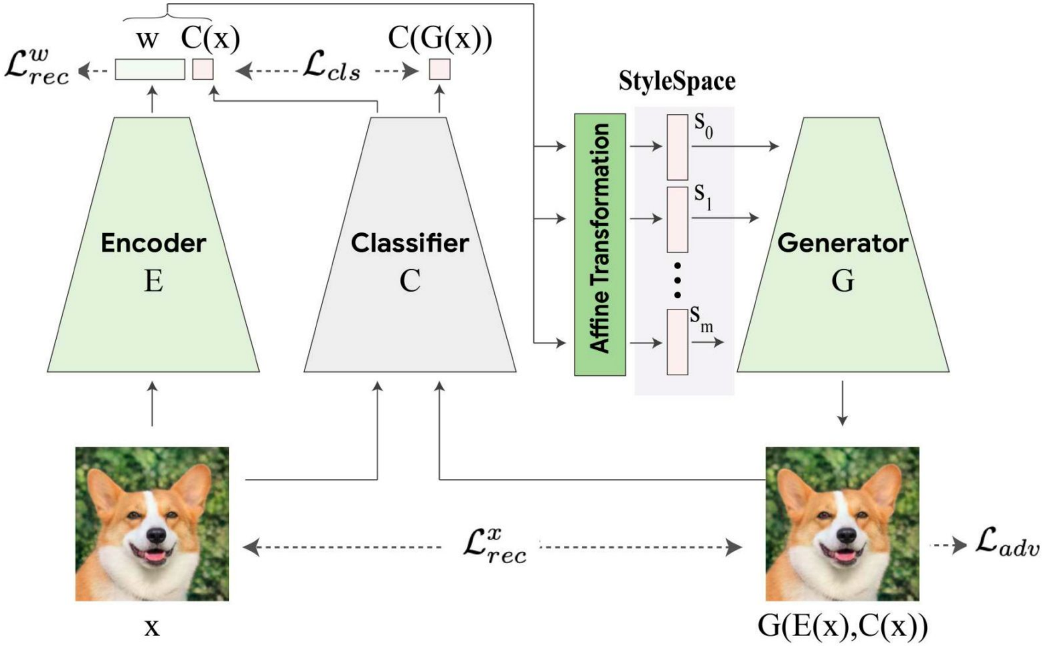
Reproducibility

- ***Reproducibility:*** The ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results.

Reproducibility

- ***Reproducibility:*** The ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results.

StylEx



The above StylEx (Lang et al. 2021) consists of:

- A classifier conditioned generative model that maps an embedding *w* into an output image.
- An encoder that maps any input image into an embedding w, so that the generator can modify attributes in real images.

Scope of reproducibility

Lang et al. assert that by integrating a classifier into the training of StylEx they can obtain principal attributes that are specific for the classification task. Theirs main claims are:

- ***Claim 1:*** They propose the StylEx model for **classifier-based training** of a StyleGAN2, thus driving its StyleSpace **to capture classifier-specific attributes**.
- ***Claim 2:*** A method, *AttFind*, to discover the **principal** classifier-related **attributes** in StyleSpace coordinates, and use these for **counterfactual explanations**.
- ***Claim 3****:* StylEx is applicable for explaining a **large variety of** classifiers and **real-world complex domains**. We show it provides **explanations understood by human users**.

What was provided

The authors provided a notebook that showcased their method:

- ***A trained model:*** A pre-trained StylEx that was trained on the FFHQ dataset

- ***Attribute Find**:* Their proposed *AttFind* method for finding the dominant features of the classifiers decision

- ***Examples**:* In their paper, they provide examples of generated images on a variety of datasets

However

The authors did **not** provide:

- ***Training Code:*** is not provided because of internal dependencies.
- ***Architecture choice***: is not detailed in the paper.
- ***Experimental setup***: such as learning rate, latent space size, importance weighting of losses, hardware requirements, and training time
- ***Dataset Labels***: The pretrained model was trained on an internal annotated version of FFHQ
- ***Full Training Procedure***: The true training procedure used a mix of sampling a random latent variable in addition to the encoded ones

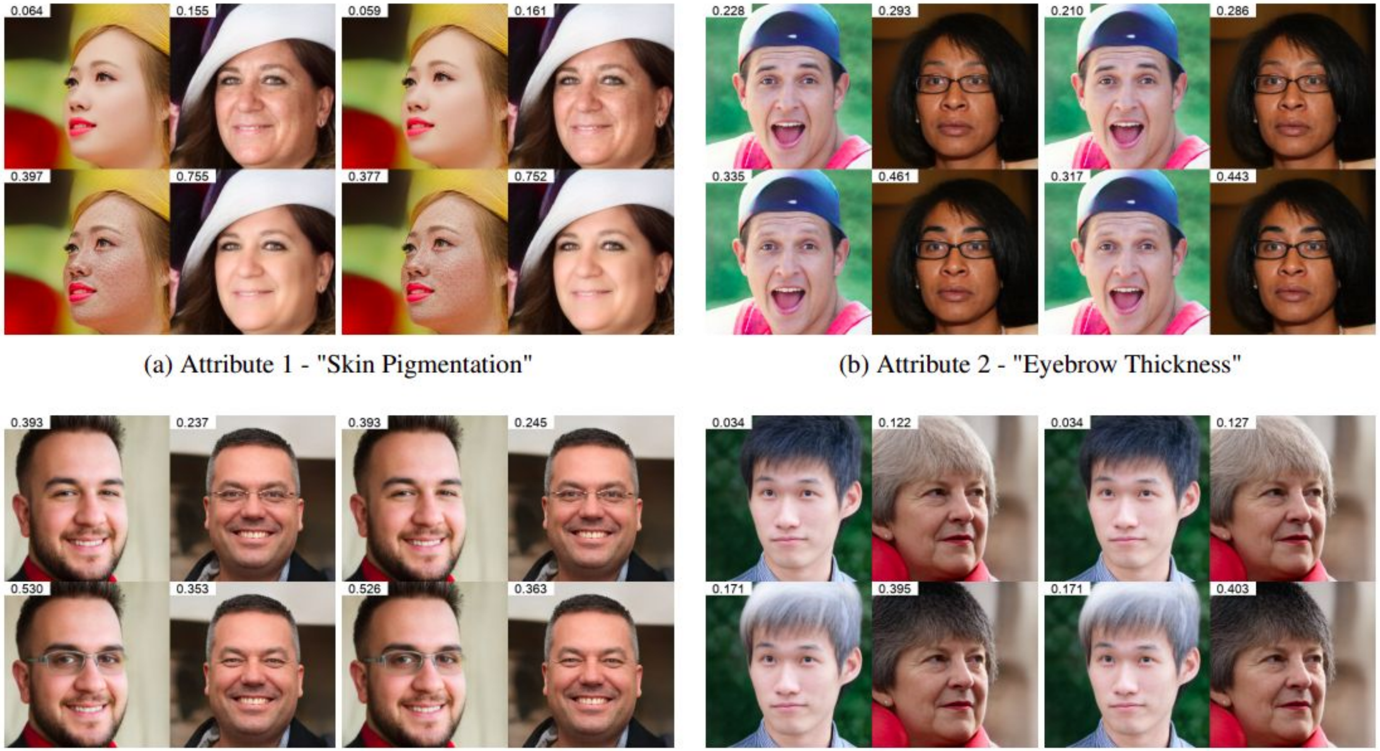
What we did

Because **full reproduction is impossible given** what the authors' provide, to verify the previous claims our goals are to:

- **Evaluate whether the principal attributes** we obtain **match their results** using their pre-trained weights.
- Retrain on datasets of smaller images and **analyze the scalability** of their method using fewer training steps and **smaller architectures**.
- **Conduct two user studies** on visual coherence and distinctness **to assess** whether attributes extracted are **interpretable by humans**

Pre-trained model

Paper vs Pre-trained AttFind results:



(a) Attribute 1 - "Skin Pigmentation"

(b) Attribute 2 - "Eyebrow Thickness"

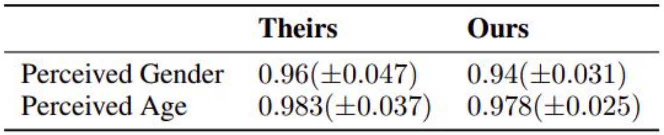
(c) Attribute 3 - "Add/Remove glasses"

(d) Attribute 4 - "Dark/White hair"

Figure 2: **Comparison of top 4 detected attributes for the perceived age classifier. Theirs (left images) vs ours (right images).** These images show how the probability of classifying a person as young or old changes based on the each attribute. On the first column of each image we display the probability of the person being perceived as old and on the second column the probability of them being perceived as young

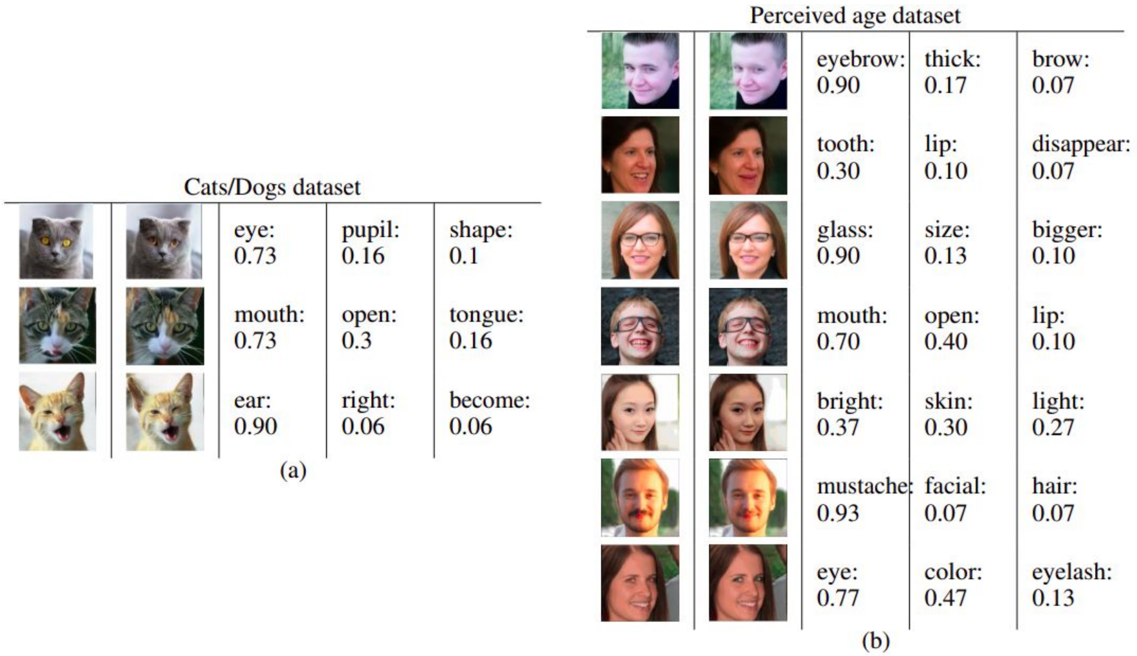
User Study

Quantitative evaluation results:



	Theirs	Ours
Perceived Gender	0.96(±0.047)	0.94(±0.031)
Perceived Age	0.983(±0.037)	0.978(±0.025)

Table 1: **Comparison of user study results.** Correct identification of the top-6 attributes.



Cats/Dogs dataset

eye: 0.73	pupil: 0.16	shape: 0.1
mouth: 0.73	open: 0.3	tongue: 0.16
ear: 0.90	right: 0.06	become: 0.06

(a)

Perceived age dataset

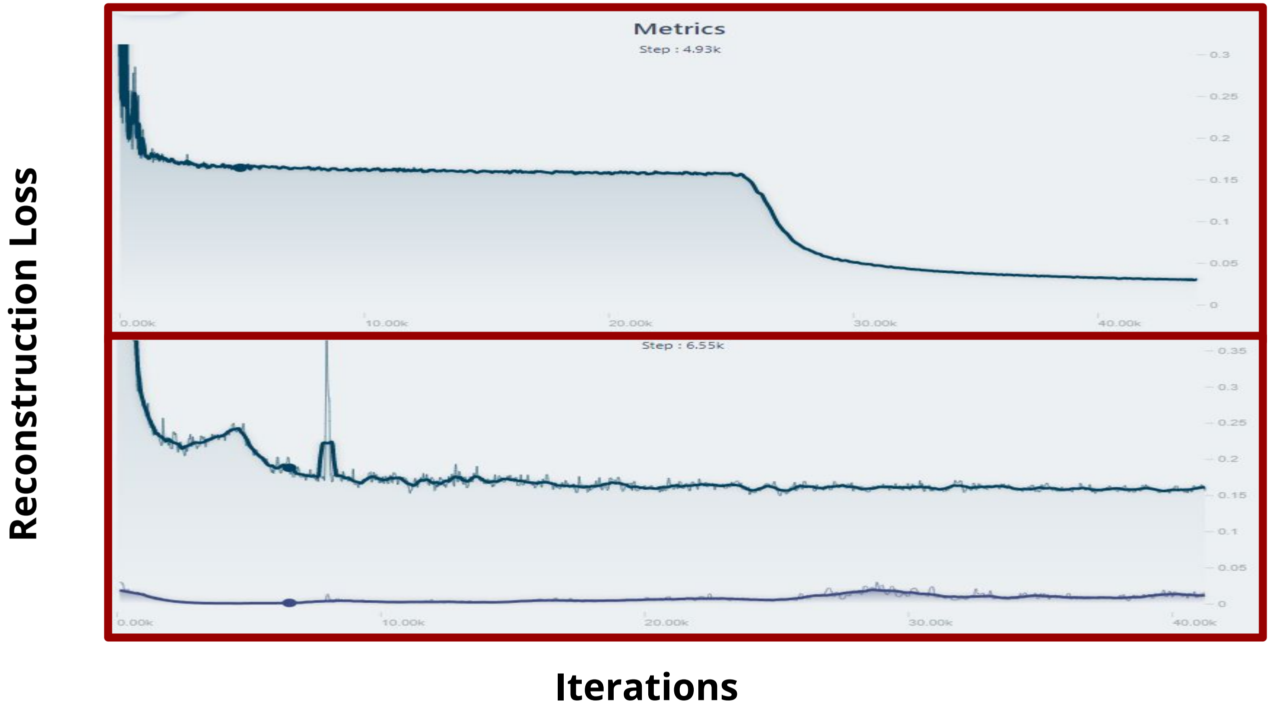
eyebrow: 0.90	thick: 0.17	brow: 0.07
tooth: 0.30	lip: 0.10	disappear: 0.07
glass: 0.90	size: 0.13	bigger: 0.10
mouth: 0.70	open: 0.40	lip: 0.10
bright: 0.37	skin: 0.30	light: 0.27
mustache: 0.93	facial: 0.07	hair: 0.07
eye: 0.77	color: 0.47	eyelash: 0.13

(b)

Table 2: **Verbal description study results.** The 3 most common words in users descriptions for the Cat/Dogs (a) and Perceived age (b) datasets.

Training the model

- **We train** our model on a **low resolution** (MNIST) **dataset**.
- We decreased the latent space dimensionality
- It **still took 9 hours to train for 50,000 iterations**



Conclusion

Reproducibility

- ***Reproducibility:*** The ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results.

Were the claims verifiable?

- ***Claim 1:*** ***We could not verify***
  - Training all three models will take 24 hours for 50,000 iterations on one GPU even for the simple MNIST dataset.
  - The reconstruction error stagnates in a local minimum before suddenly dipping.
  - The model was not always able to escape the local minima within 50,000 iterations.

- ***Claim 2:*** ***We could verify***
  - AttFind can discover:
    - classifier-related attributes in StyleSpace
    - globally and the locally important attributes.
  - Top-4 attributes are the same for the perceived age classifier.

- ***Claim 3:*** ***We could partially verify***
  - We confirm their quantitative results using two domains which shows the counterfactual examples were interpretable by humans
  - Could not train from scratch

Thank you for your attention!