JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Text-guided Unsupervised Latent Transformation for Multi-attribute Image Manipulation

Xiwen Wei[1], Zhen Xu[1], Cheng Liu[2], Si Wu[1,3*] , Zhiwen Yu[1], and Hau San Wong[4]

[1]School of Computer Science and Engineering, South China University of Technology

[2]Department of Computer Science, Shantou University

[3]Peng Cheng Laboratory

[4]Department of Computer Science, City University of Hong Kong

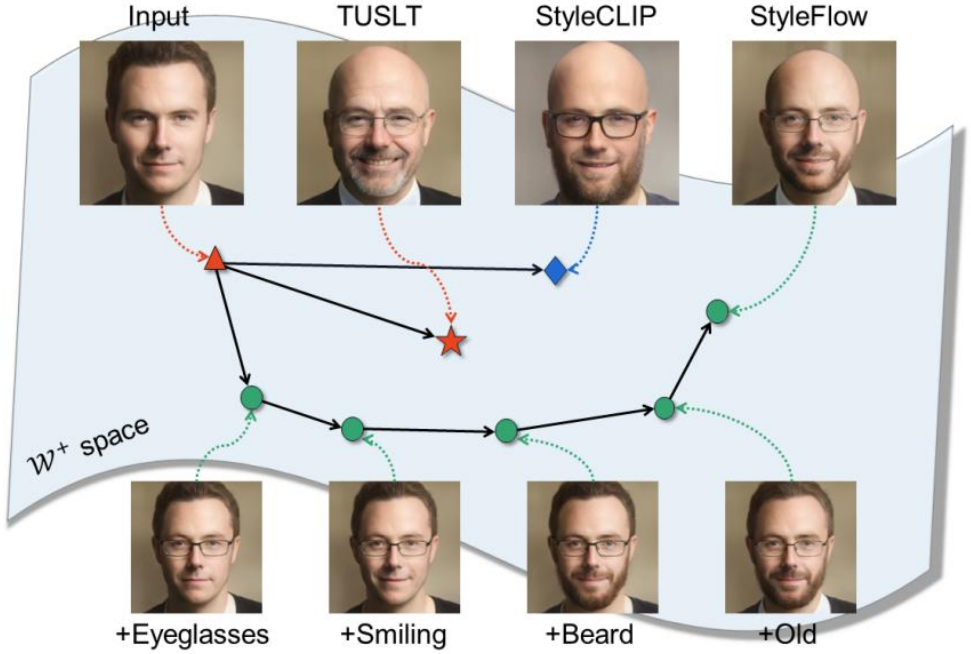
{202021044777, csxuzhen}@mail.scut.edu.cn, cliu@stu.edu.cn, {cswusi, zhwyu}@scut.edu.cn, cshswong@cityu.edu.hk

THU-AM-267

Proposed Approach

The existing image editing methods focus on discovering semantic latent directions associated with individual visual attributes, and a sequential manipulation process is thus needed for multi-attribute manipulation.

- The proposed model infers **a single step** of latent space walk to simultaneously **manipulate multiple attributes**.
- We jointly train a **latent mapping network** with an **auxiliary attribute classifier**.
- Our latent mapping network breaks down the challenging multi-attribute manipulation task into sub-tasks: **inferring diverse semantic directions** and integrating the target-related ones into **a single transformation vector**.

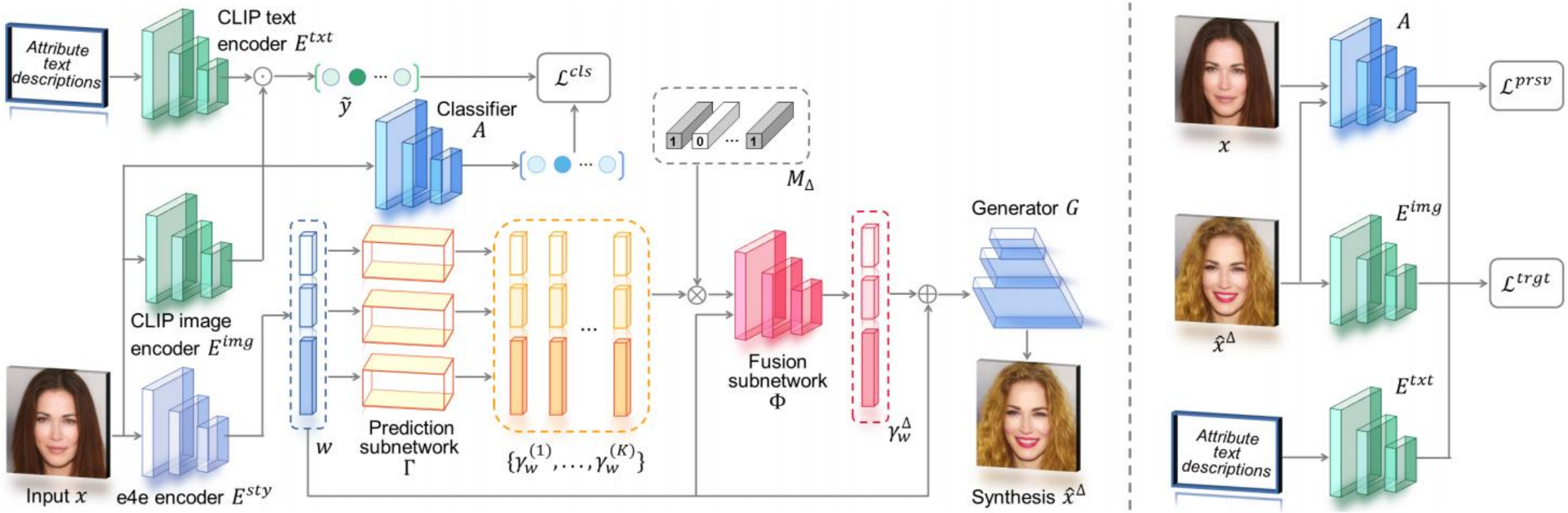


Structure of TUSLT

- The key of TUSLT is to jointly learn a mapping network to infer the latent transformation and an auxiliary attribute classifier to assess manipulation quality.
- The proposed model consists of two learnable components: an auxiliary attribute classifier A trained on the CLIP-based labeled data, and a latent mapping network $\{\Gamma, \Phi\}$ and three pretrianed components:a generator G, an e4e encoder E^{sty} and CLIP encoder $\{E^{txt}, E^{img}\}$.

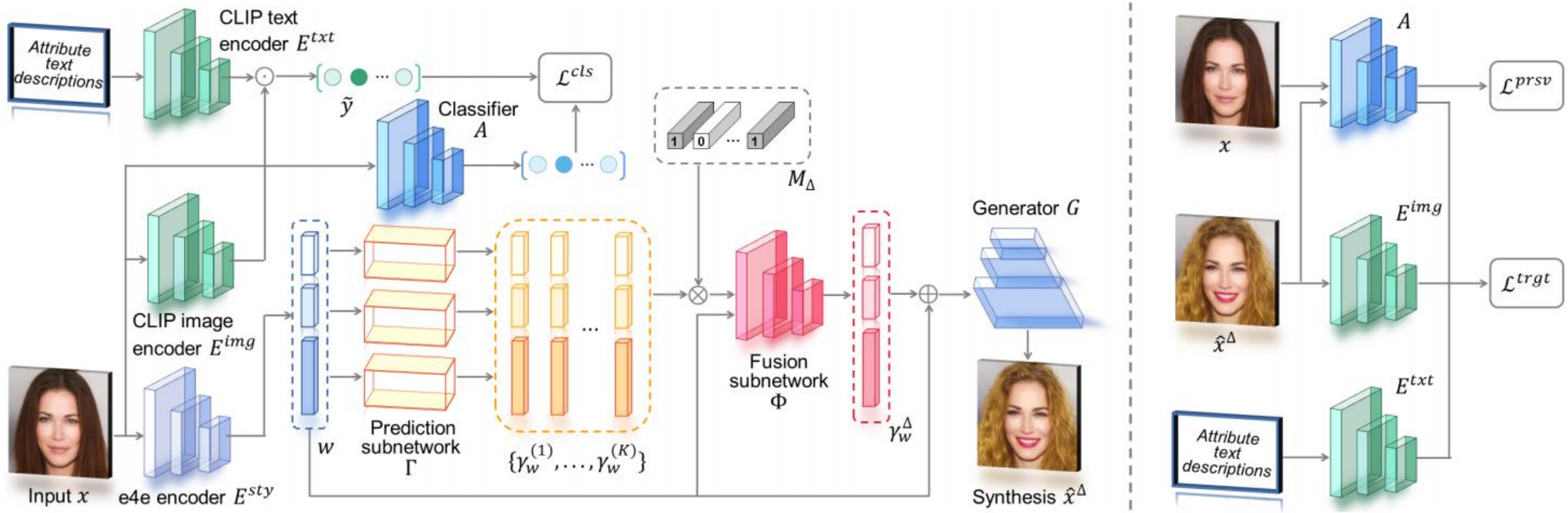
Structure of TUSLT

- Firstly, We employ the Contrastive Language-Image Pre-training(CLIP) model to generate pseudo-labeled data by measuring the semantic similarities between attribute text descriptions and training images.



Structure of TUSLT

- Further, we adopt a two-stage architecture for the mapping network: the earlier stage employs a prediction subnetwork to infer a set of semantic directions, and the latter stage operates on the resulting directions and nonlinearly fuses the target-related ones. The intermediate semantic directions are associated with preset attributes and tailored for the input image.



Auxiliary Attribute Classifier

- Let $T = \{t^{(1)}, \ldots, t^{(K)}\}$ denote a set of text prompts, and $t^{(i)}$ describes the i-th preset attribute. To identify the attributes reflected in images, we embed training images and T in the shared embedding space, and measure the semantic similarity as:

$$S^{(i)}(x) = cos(E^{txt}(t^{(i)}), E^{img}(x))$$

- where $cos(\cdot, \cdot)$ denotes the cosine distance between input vectors. $S^{(i)}(x)$ should be larger when $t^{(i)}$ and x represent the same attribute. At this point, we pseudo-annotate training images, and the corresponding label $\tilde{y}$ is defined as:

$$\tilde{y}^{(i)} = \begin{cases} 1, if\ S^{(i)}(x) > \tau \\ 0, \quad otherwise \end{cases}$$

Latent Mapping Network

- It is challenging to directly infer the latent transformations for a variety of attribute combinations. To address this problem, we design a two-stage architecture for our latent mapping network.
- The first stage is based on a direction prediction subnetwork Γ that produces latent directions denoted by $\Gamma(\mathrm{w}) = \{\gamma_w^{(1)}, \ldots, \gamma_w^{(K)}\}$, where $\gamma_w^{(i)}$ associates with the preset attribute i, conditioned on w.
- At the second stage, a fusion subnetwork Φ operates on the produced directions. We define a binary vector $\Delta \in \{0, 1\}^{\mathrm{K}}$ to indicate target attributes. Φ learns to integrate the directions as indicated by Δ, and infers a residual vector γ_w^{Δ} defined as:

$$\gamma_w^{\Delta} = \Phi(w, M_{\Delta} \otimes \Gamma(w))$$

Latent Mapping Network

- As a result, multi-attribute manipulation can be carried out by simply adding the initial latent code to the residual vector. The generator G of StyleGAN is employed to decode the resultinglatent vector as:

$$\hat{x}^{\Delta} = G(w + \alpha\gamma_{w}^{\Delta})$$

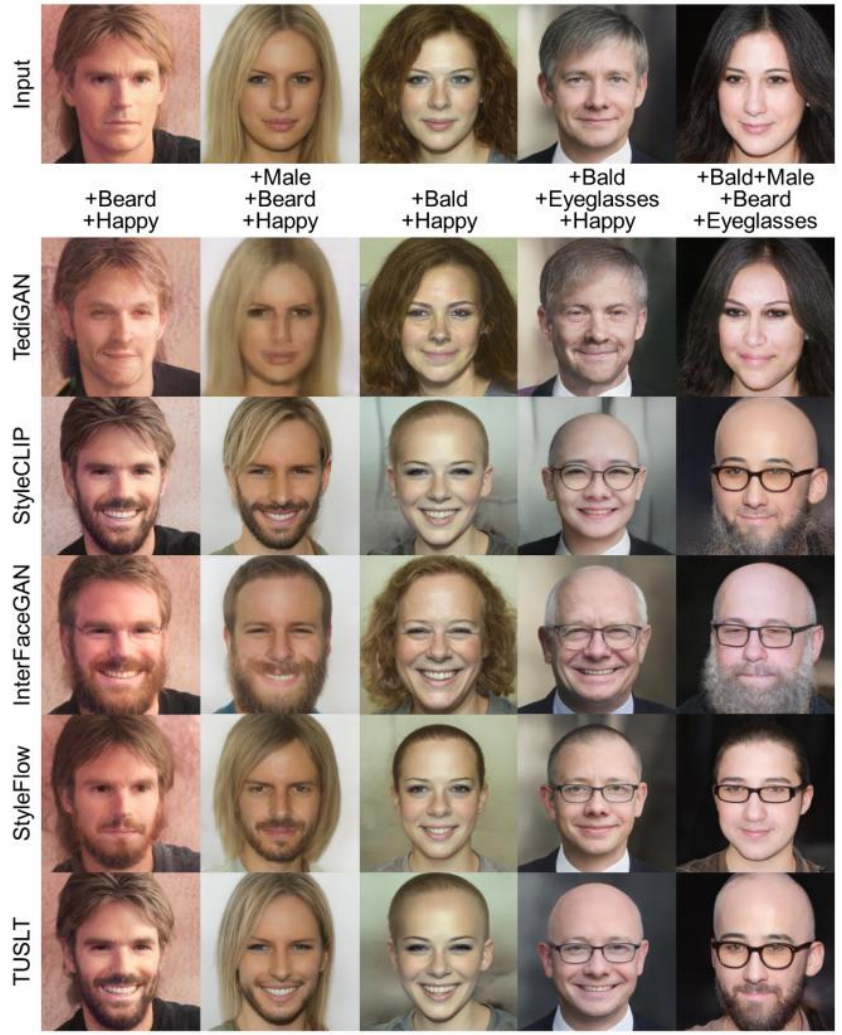
- where α controls the manipulation strength. Although our mapping network stacks two stages, each stage has access to the latent code of the input image.

Semantically meaningful directions

- Single-attribute transformation results of StyleCLIP and TUSLT.

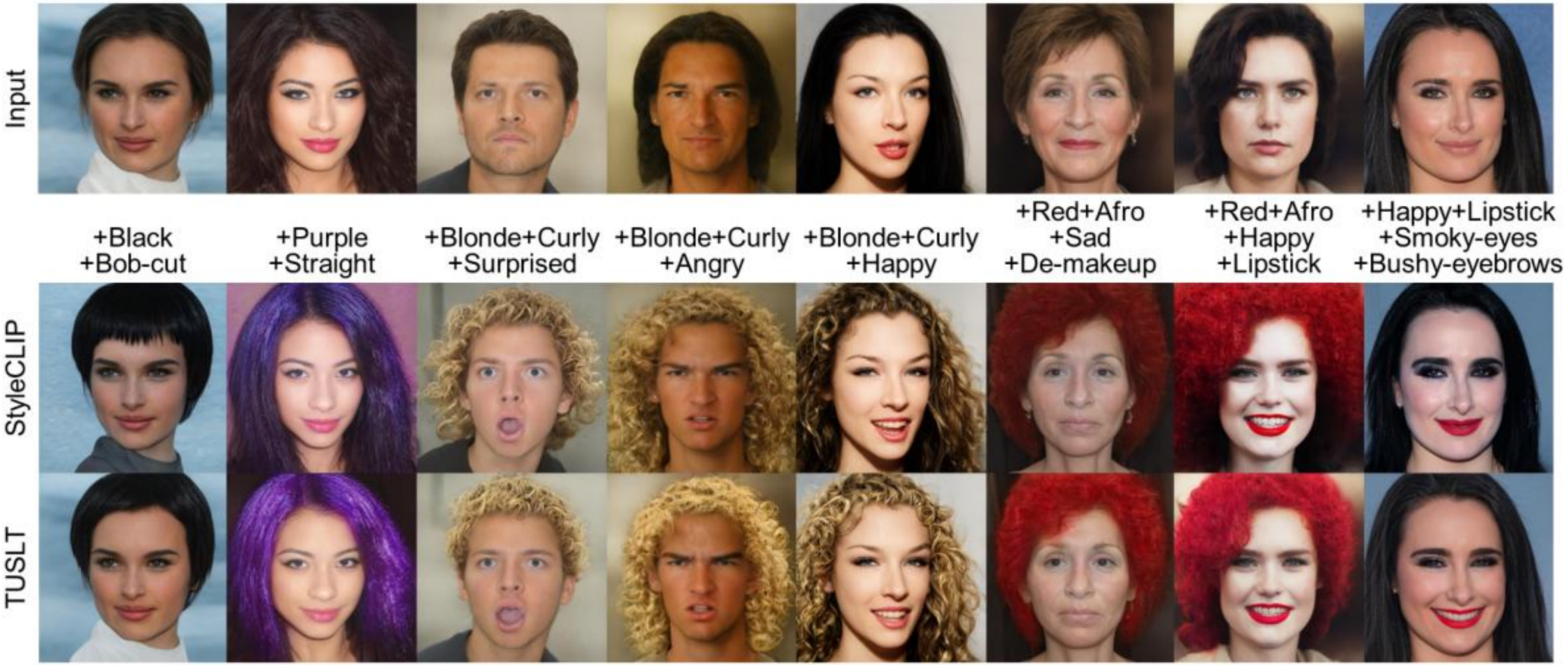
Precise manipulation on multiple attributes

- Multi-attribute manipulation results of TUSLT and competing methods.



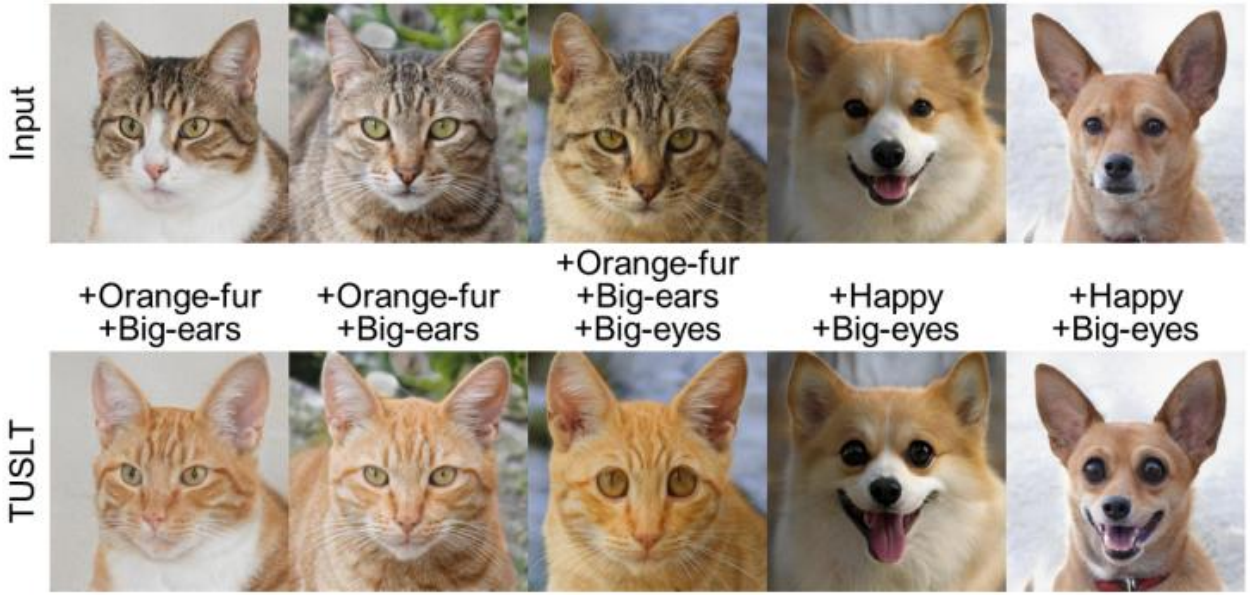
Precise manipulation on multiple attributes

- Diverse image synthesis results of TUSLT and StyleCLIP.



Results on AFHQ and AnimeFace

- We also show the ability of the proposed model to manipulate multiple attributes on AFHQ-cats/dogs and AnimeFace.



Conclusion

- We propose a text-guided unsupervised multi-attribute manipulation model to edit images in **a single latent transformation step.**
- Benefiting from the cross-modal image and text representation of CLIP, we can **jointly train an auxiliary attribute classifier and a latent mapping network for precise attribute manipulation.**
- This work significantly **increases the scalability** of StyleGAN-based image attribute manipulation **without causing any manual annotation cost.**

Thank you