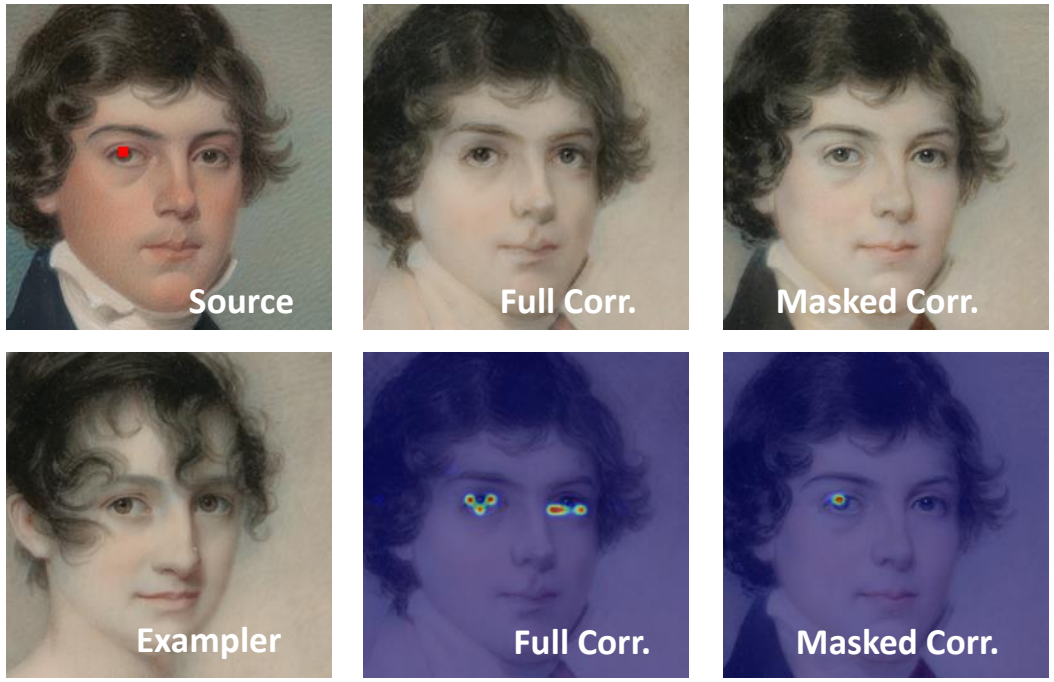
Masked and Adaptive Transformer for Exemplar Based Image Translation

Chang Jiang[1], Fei Gao[1,2]*, Biao Ma[1], Yuhao Lin[1], Nannan Wang[2], Gang Xu[1]

[1]Hangzhou Dianzi University [2]Xidian University

https://github.com/AiArt-HDU/MATEBIT

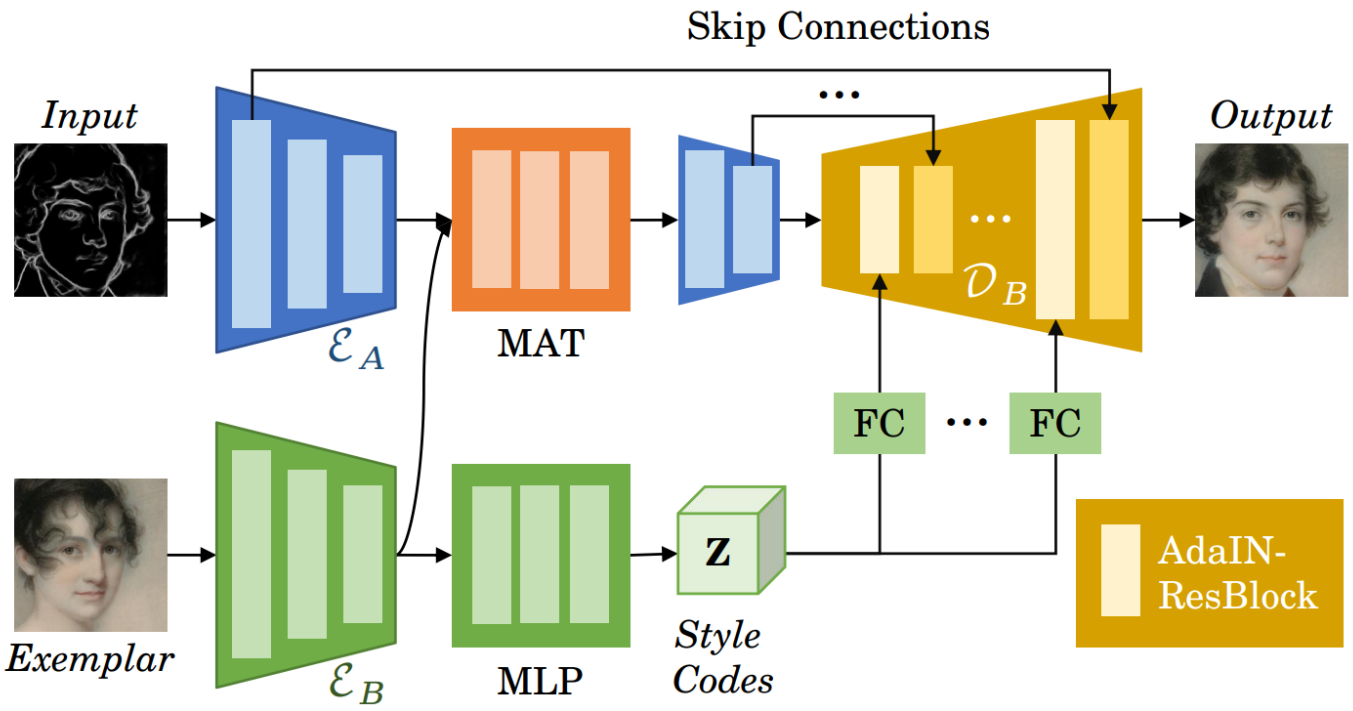
Motivation



- How to build accurate semantic correspondence between conditional inputs and exemplar ?
- How to modulate the style information for global injection?

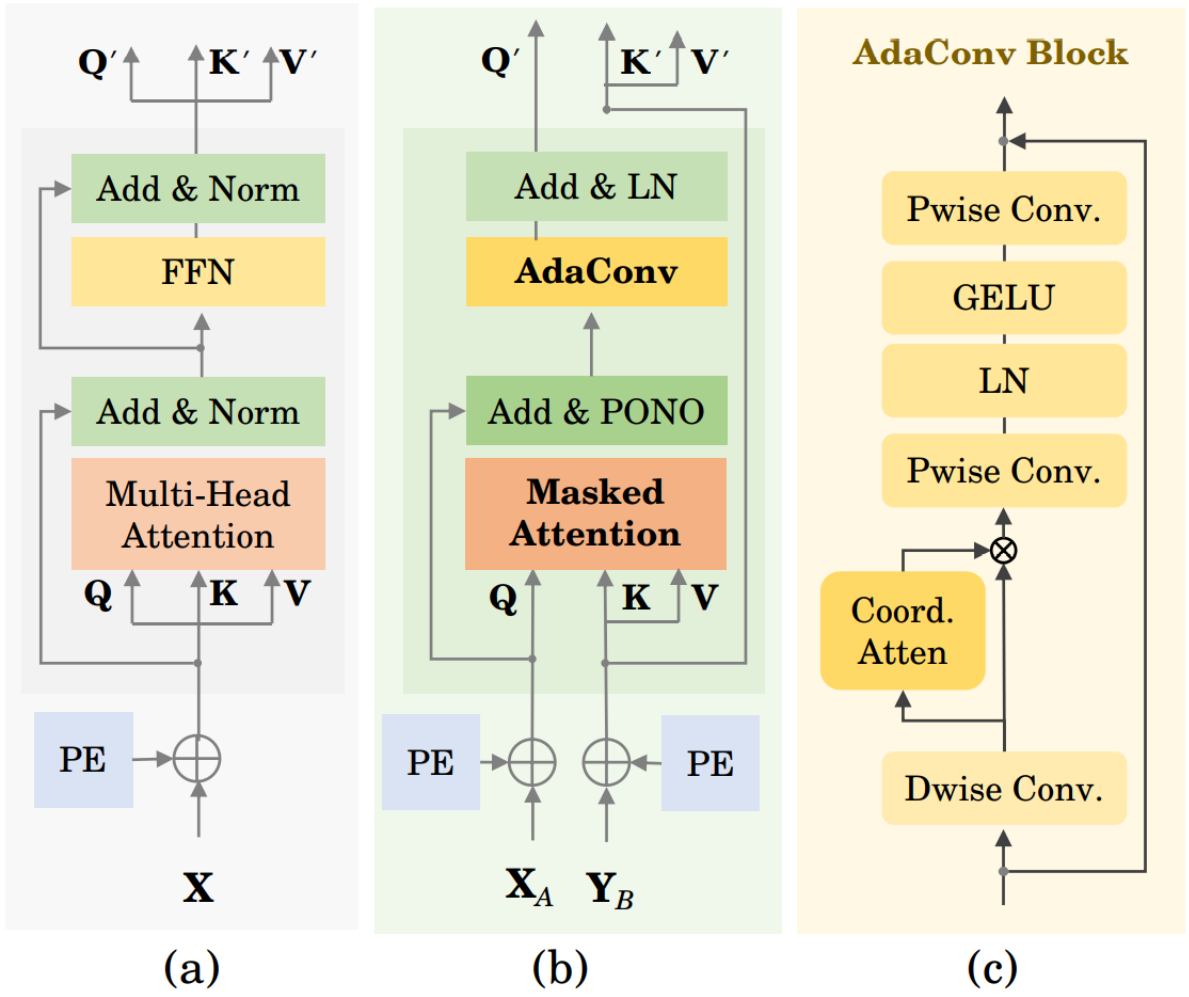
Improve the accuracy of matching on the one hand, and diminish the role of matching in image generation on the other hand.

Overview of translation network

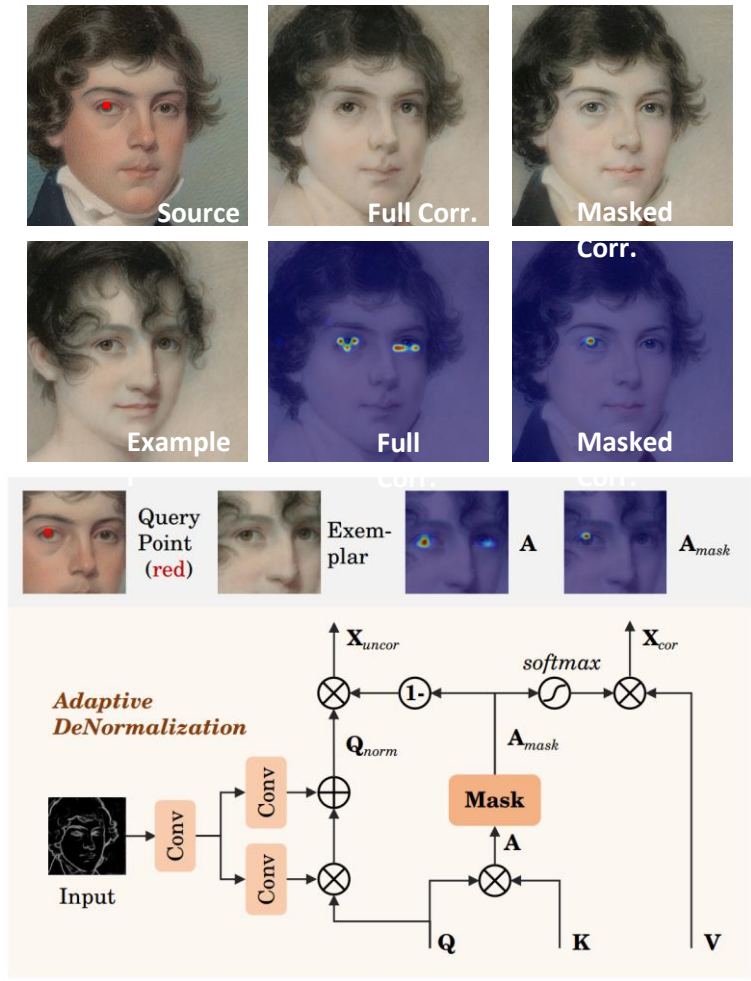


- Accurate semantic matching of correspondences
- Global style information modulation
- Significantly improve the image quality
- General translation solution

MAT Block



(a) Vanilla Transformer block, (b) MAT block,
(c) AdaConv block



Masked Attention

MAT Block

Constructing conditional input and example image starting correspondence using the cosine attention mechanism.

$$A(u,v) = \frac{\tilde{Q}(u)\tilde{K}(v)^T}{\|\tilde{Q}(u)\|\|\tilde{K}(v)\|},$$

$$\tilde{Q}(u) = Q(u) - \bar{Q}(u),\ \tilde{K}(u) = K(u) - \bar{K}(u)$$

The mask interval is divided by the ReLU mechanism, A_{mask} is the post-mask probability matrix, and the post-mask confidence matrix is obtained by summing up the j dimensions.

$$X_{cor} = \tilde{A}_{mask} V, \text{with} \tilde{A}_{mask} = \text{softmax}(\alpha \cdot A_{mask})$$

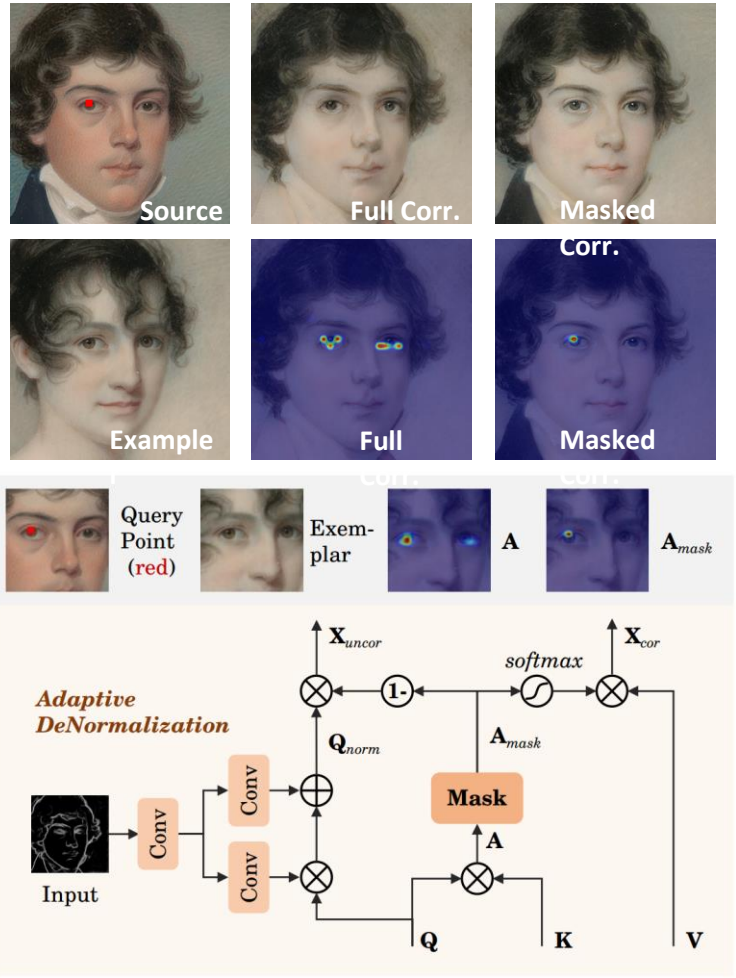
$$X_{uncor} = \left(1 - \sum_j A_{mask}\right) \odot Q_{norm}$$

$$Q_{norm} = \gamma(x_A)\frac{Q - \mu(Q)}{\sigma(Q)} + \beta(x_A)$$

$$X_{agg} = PONO(X_{cor} + X_{uncor} + Q)$$

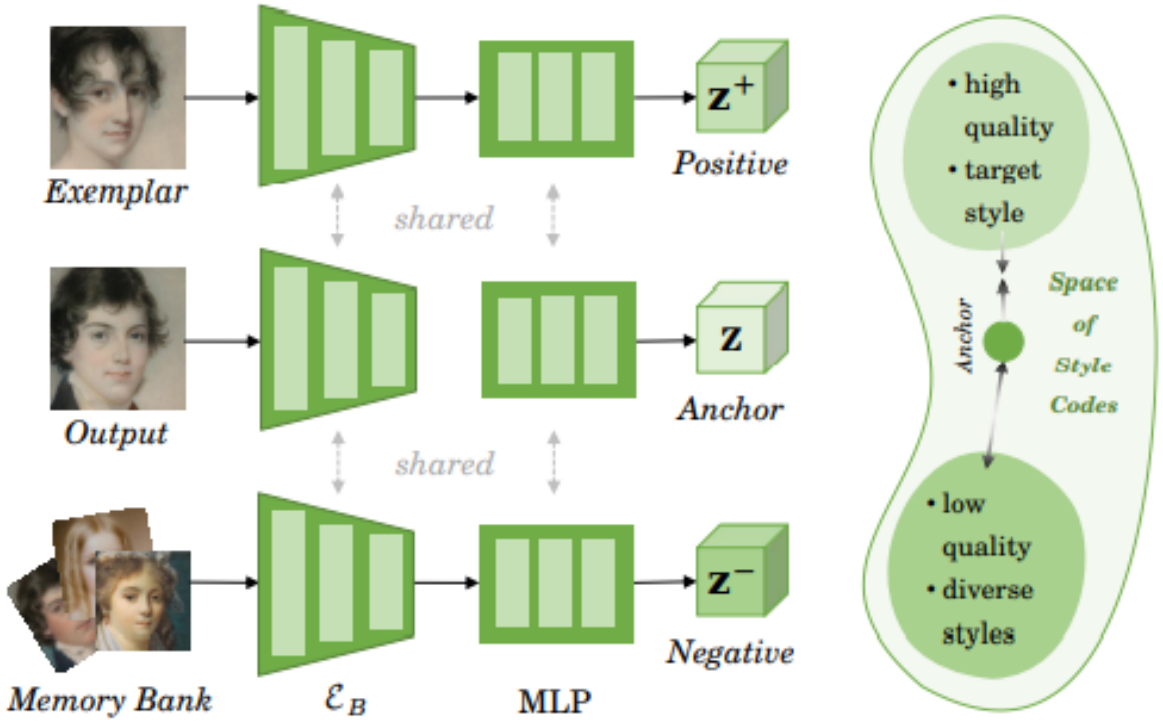
Fusing semantic information by adaptive convolution to generate a more detailed warp image.

$$X_{MAT} = \text{AdaConv}(X_{agg}) + X_{agg}$$



Masked Attention

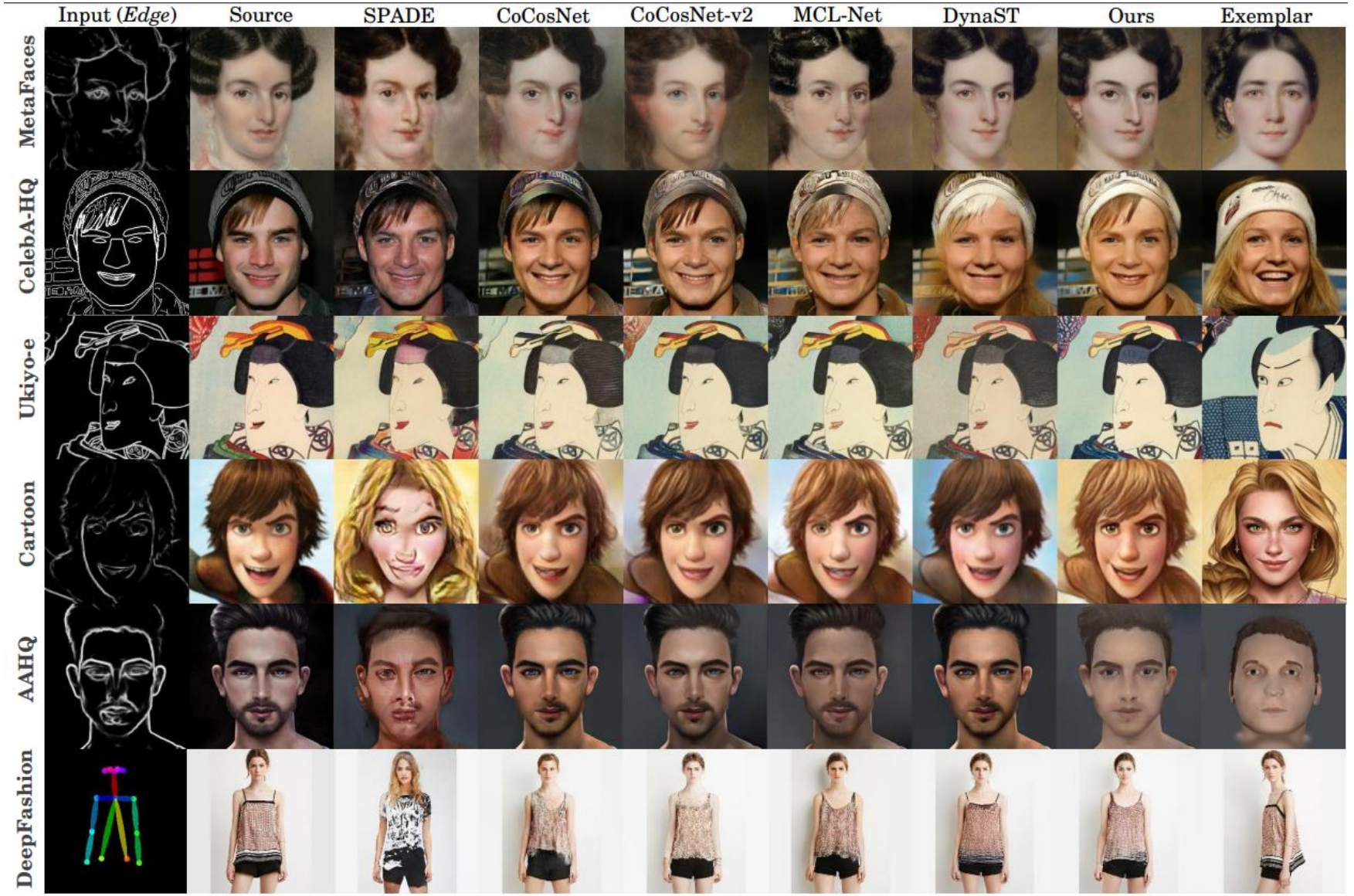
Contrastive style learning



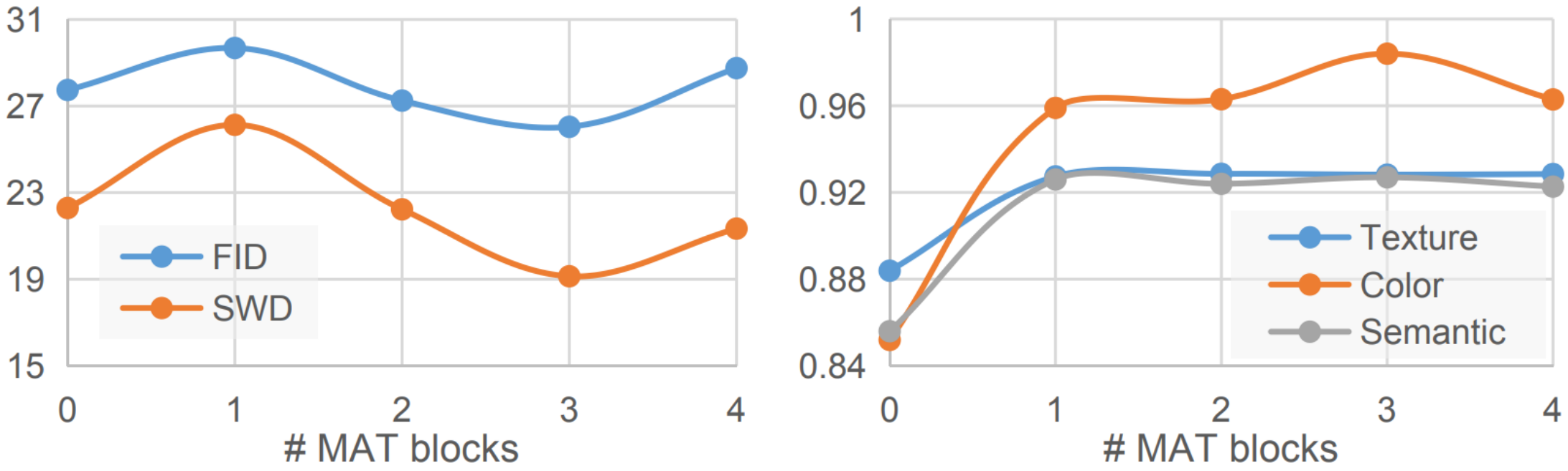
- Push to generate high-quality images
- Global style injection

$$\mathcal{L}_{style} = -\log \frac{\exp(\frac{\mathbf{z}^T \mathbf{z}^+}{\tau})}{\exp(\frac{\mathbf{z}^T \mathbf{z}^+}{\tau}) + \sum_{j=1}^{m} \exp(\frac{\mathbf{z}^T \mathbf{z}_j^-}{\tau})},$$

Comparison with state-of-the-art



Impacts of MAT



Impact of the number of MAT blocks on performance. For example, style, texture, semantic information, etc.

Conclusion: both the semantic consistency and style realism broadly improve with the number of MAT blocks and peak at three.

More Results

(a) Metfaces

(b) AAHQ

(C) CelebA-HQ

(d) Ukiyo-e-face

More Results

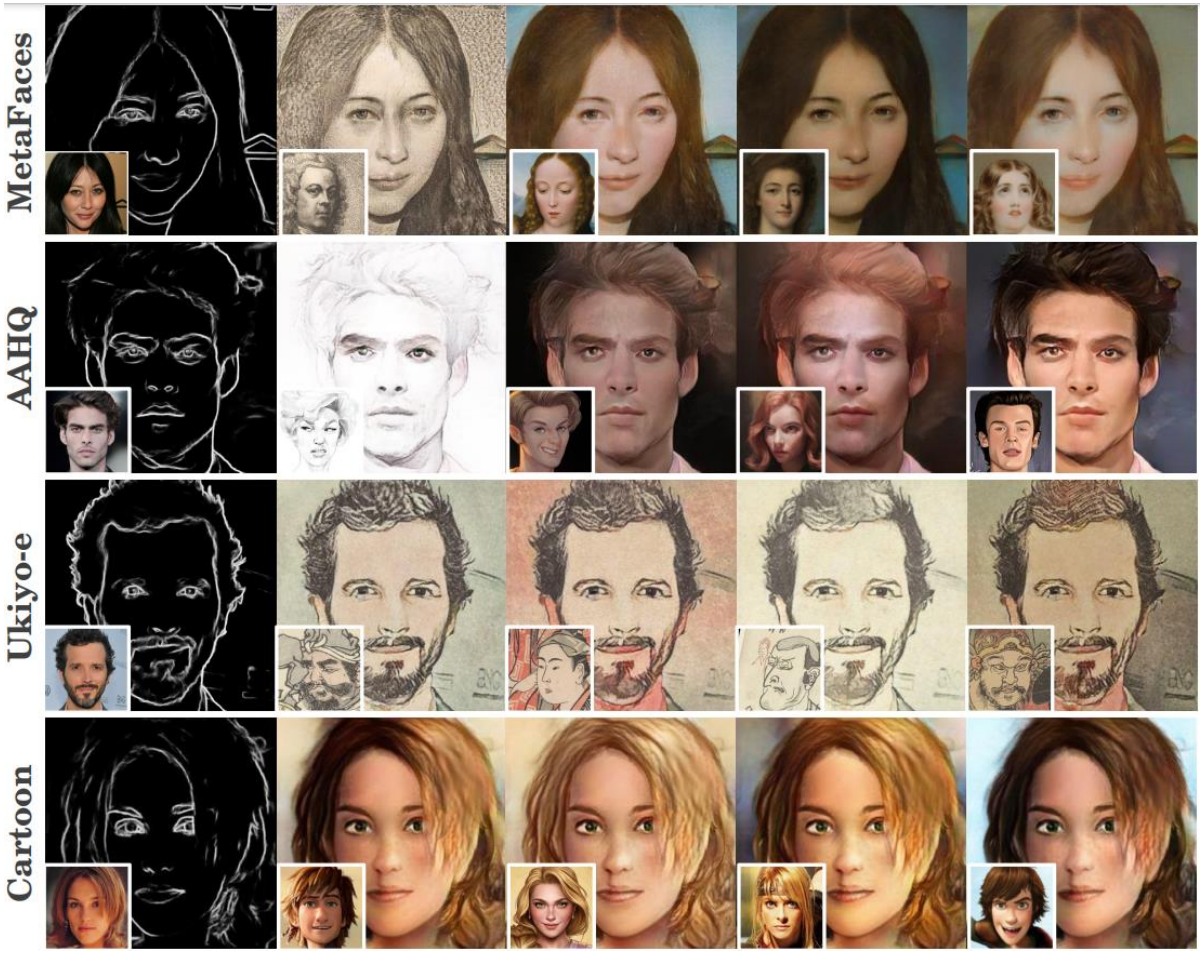
(e) Cartoon

(f) Chinese Ink Painting

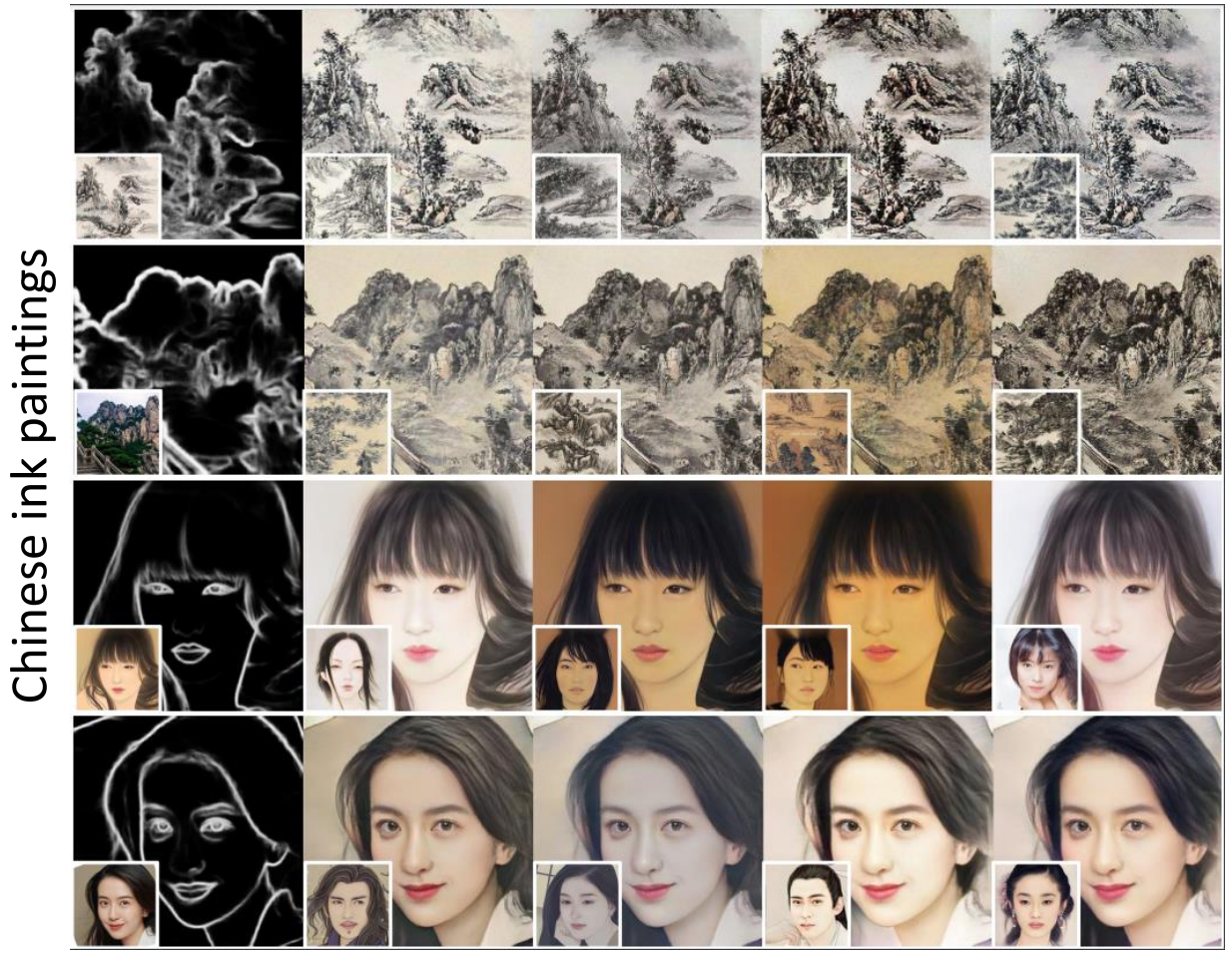
(g) DeepFashion

(h) Landscape

Applications



(a) Artistic Portrait Generation



(b) Chinese Ink Painting Generation