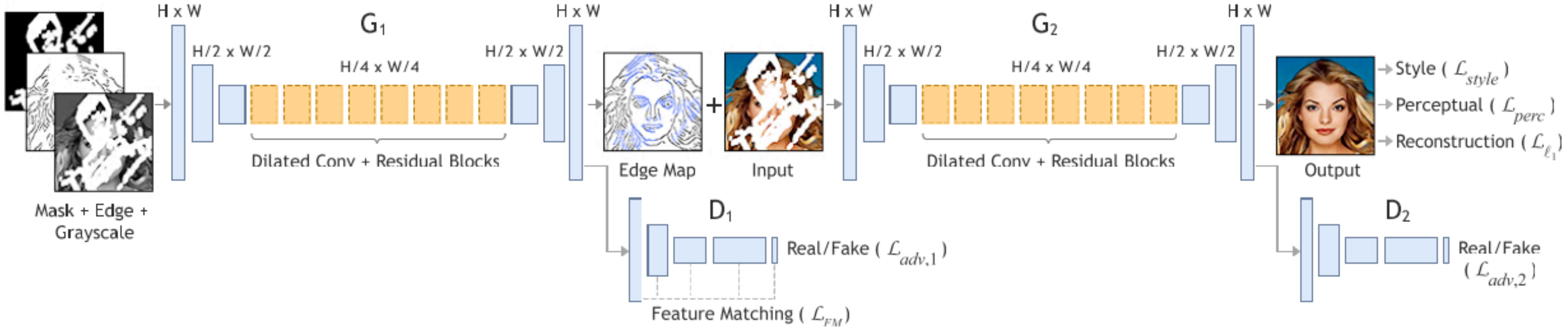
Image Inpainting and Editing with Structural Prior Guidance

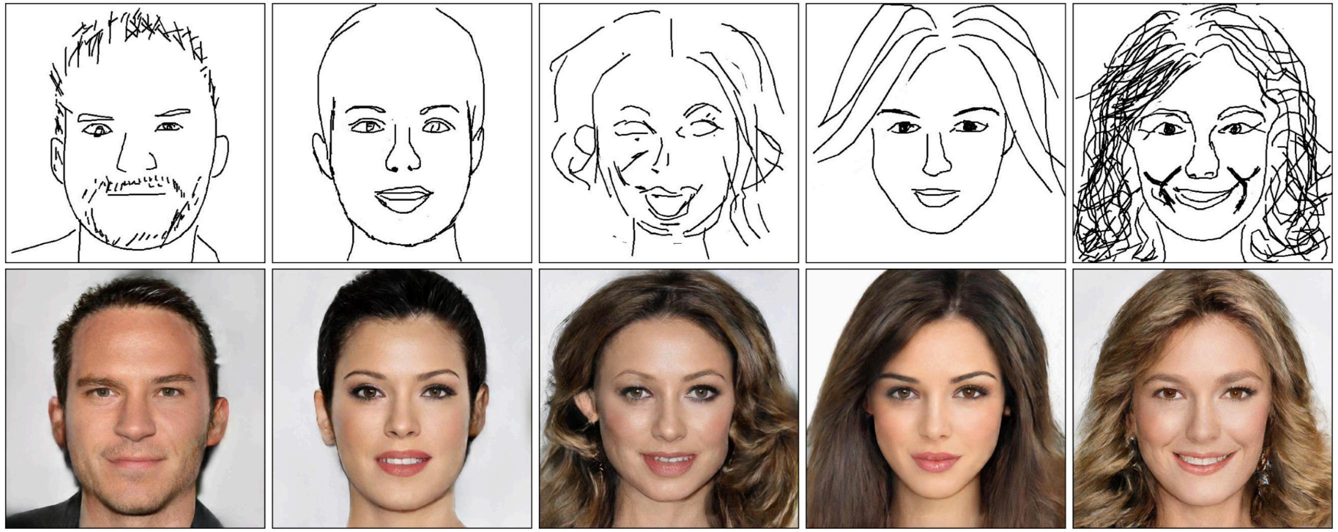
Chenjie Cao,
School of Data Science, Fudan University
ccjdurandal422@163.com

1 Image Inpainting

Task: Sketch/Edge based Image Inpainting/Editing

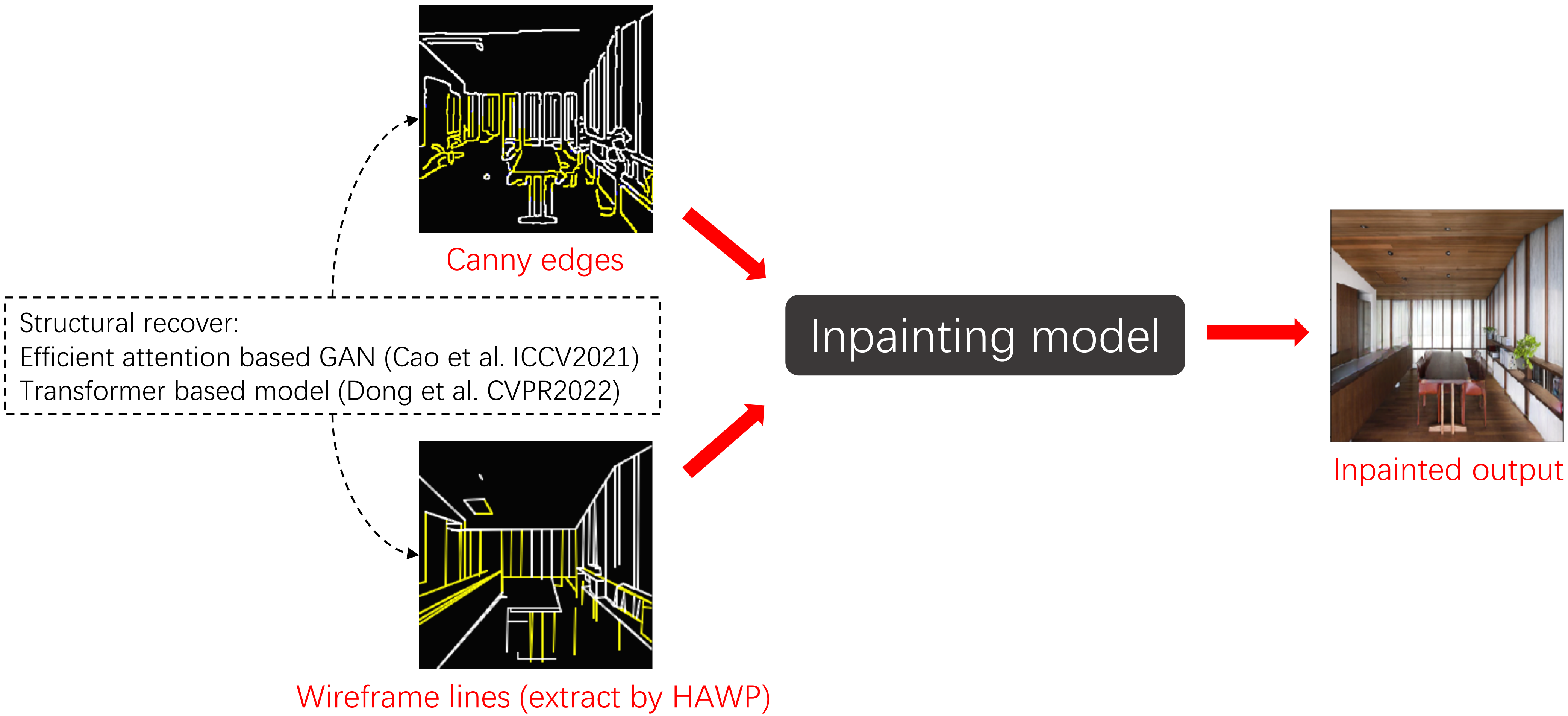


Edgeconnect, Nazeri, Kamyar, et al. ICCV workshop (2019)



DeepFaceDrawing, Chen et al. SIGGRAPH (2020)

Line/Edge priors → inpainting/synthesis



Cao et al, Learning a Sketch Tensor Space for Image Inpainting of Man-made Scenes. ICCV2021
Dong et al, Incremental Transformer Structure Enhanced Image Inpainting with Masking Positional Encoding. CVPR2022
Xue N et al. [HAWP] Holistically-attracted wireframe parsing CVPR2020

Learning a Sketch Tensor Space for Image Inpainting of Man-made Scenes

Chenjie Cao, Yanwei Fu
School of Data Science, Fudan University
{20110980001,yanweifu}@fudan.edu.cn

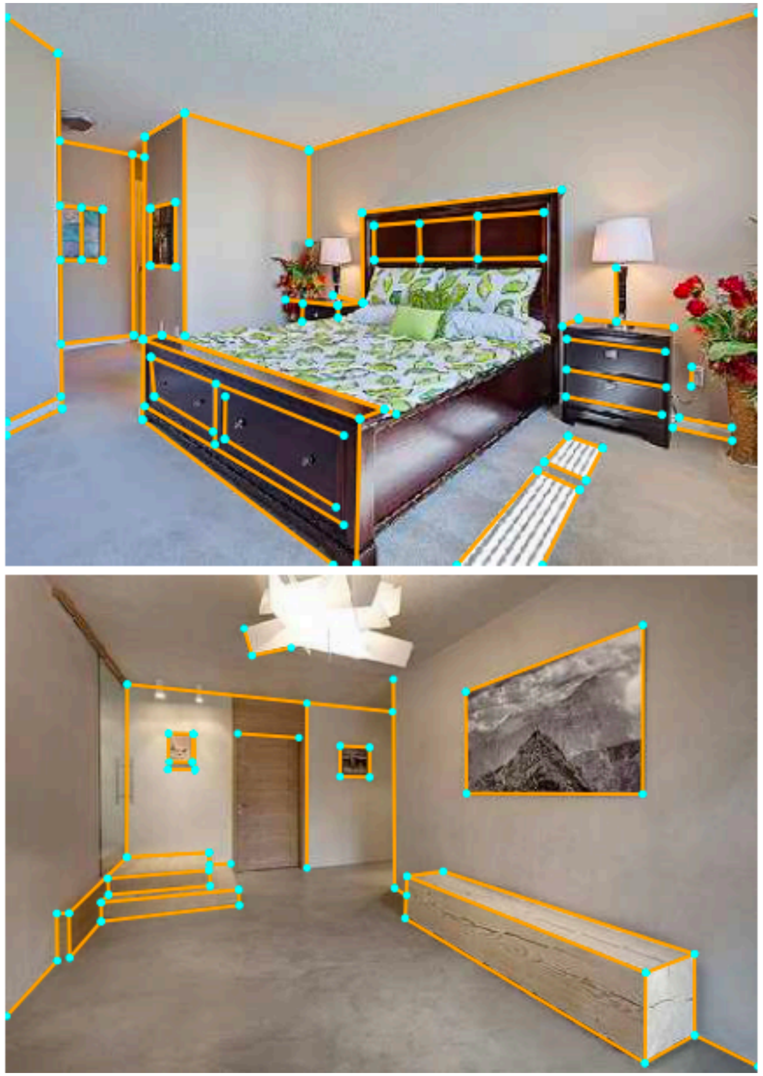
ICCV 2021

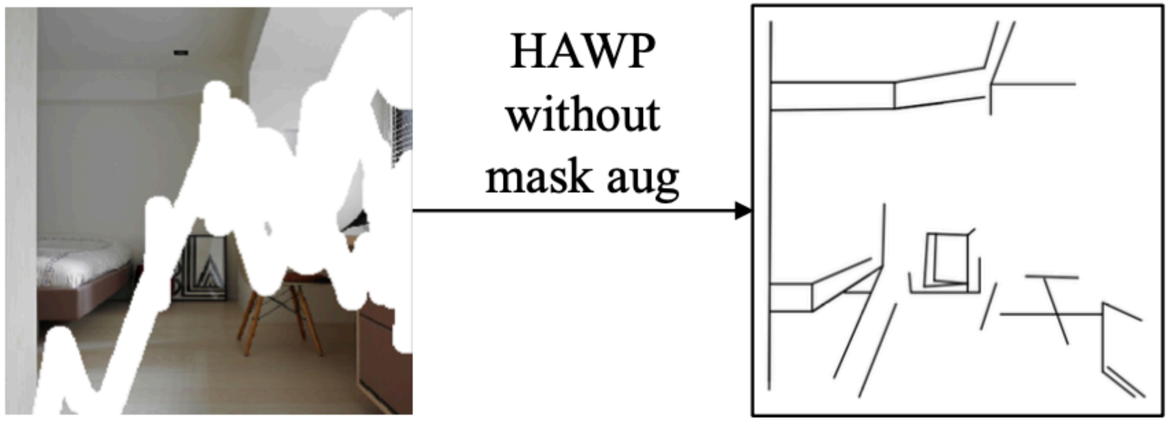
Codes and models are released in **https://ewrfcas.github.io/MST_inpainting**

(a) Inputs (b) Exemplar-based (c) GateConv (d) EdgeConnect (e) Ours

Filling in the Missing critical structures for man-made scenes

Motivation





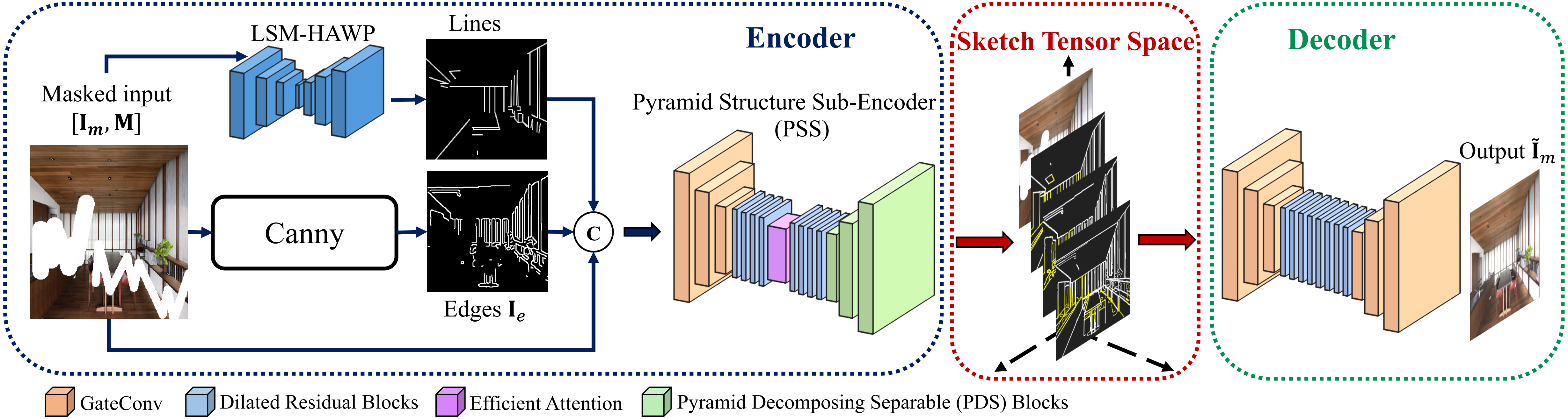
Unreliable pattern transfer for corrupted priors

Motivation:

- Introduce discretely represented wireframes to the image inpainting.
- Learning a more robust prior detector for masked images.
- Improve inpainting performance efficiently.

Xue N, Wu T, Bai S, et al. Holistically-attracted wireframe parsing CVPR2020.

Overview



Model Pipeline:

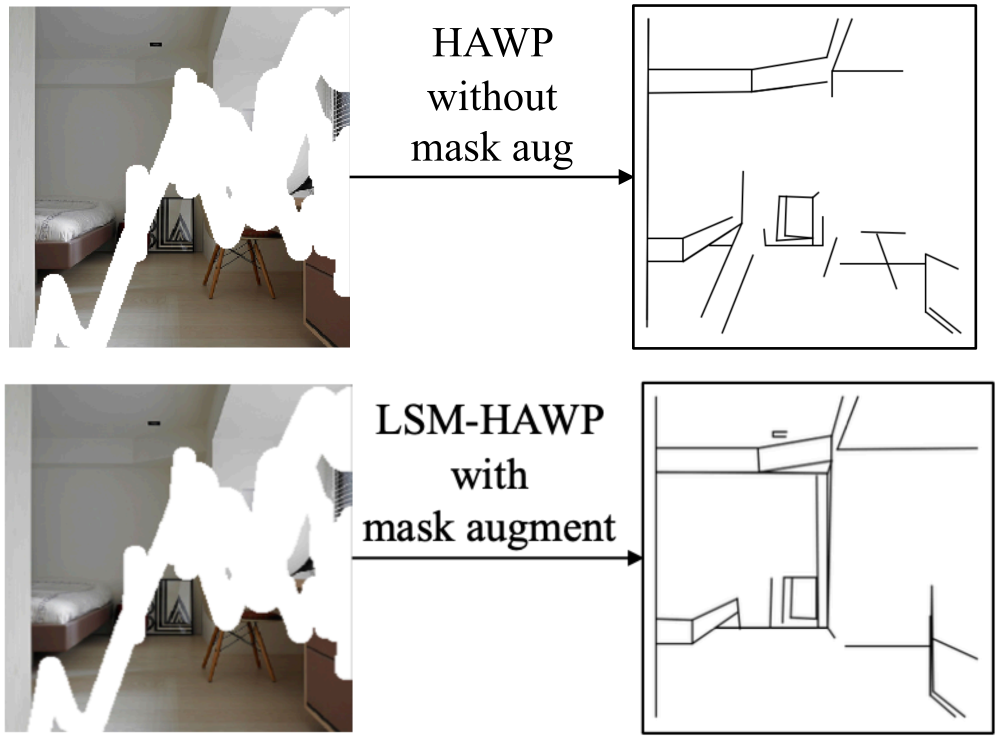
- Use LSM-HAWP and canny detector to extract line and edge maps.
- Refine structures by Pyramid Structure Sub-Encoder (PSS) to sketch tensor space.
- Decoder predicts the final inpainted image.

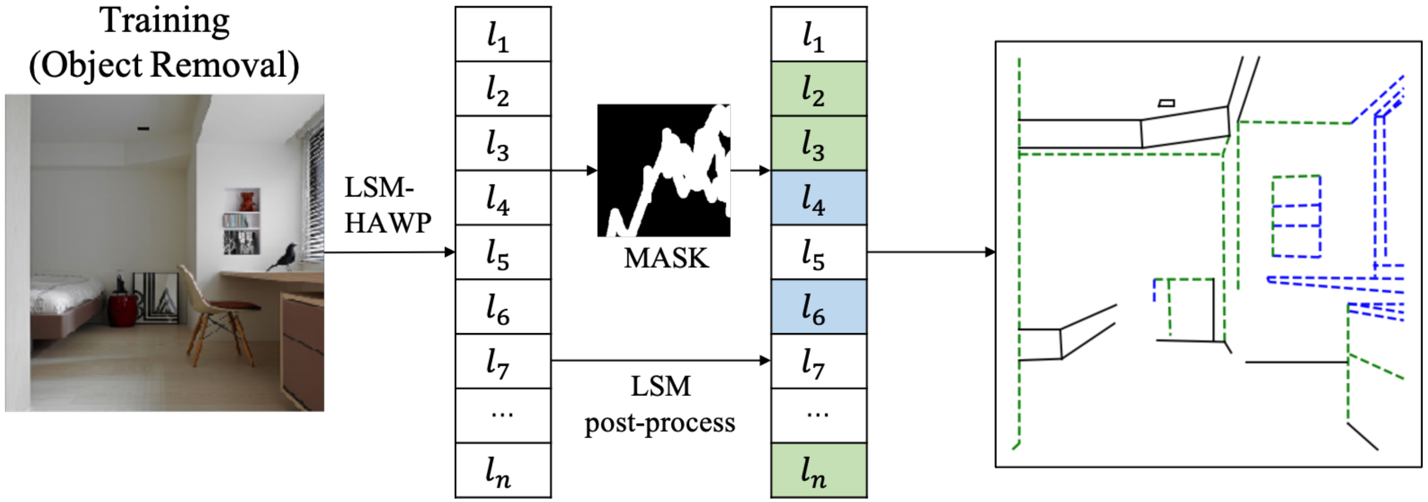
Line Segment Masking (LSM)

Line Segment Masking (LSM):

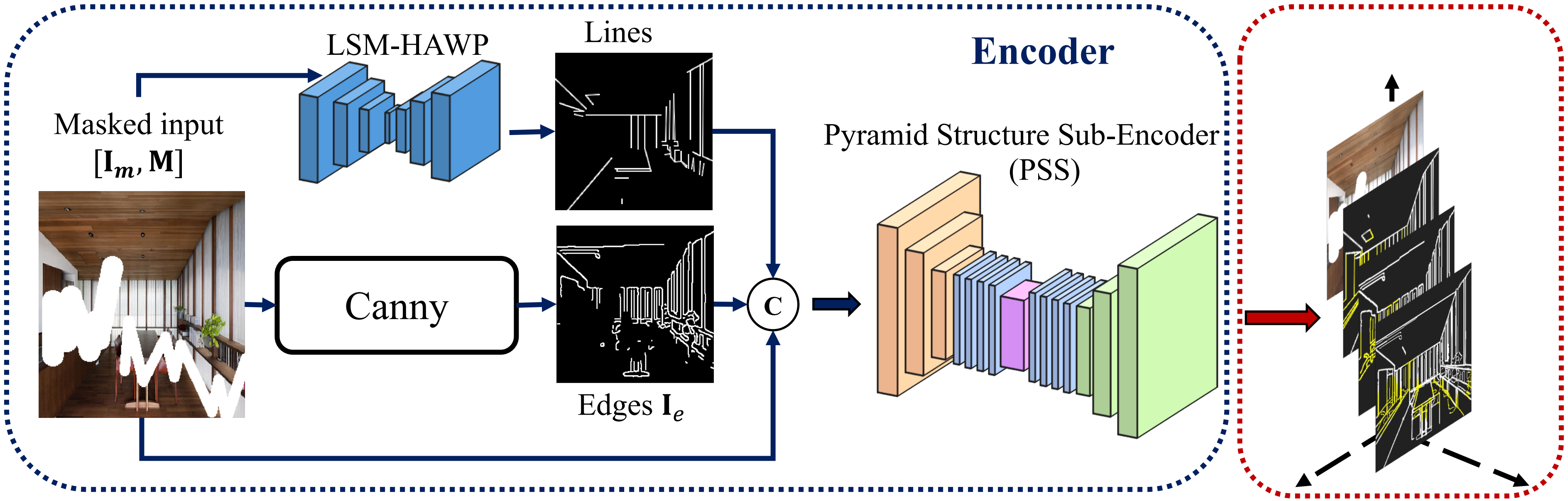
- HAWP failed to directly achieve good results for masked images.
- We use LSM as a data augmentation to improve HAWP as LSM-HAWP.

	unmasked testset			masked testset		
Threshold	5	10	15	5	10	15
HAWP	62.16	65.94	67.64	35.39	38.47	40.15
LSM-HAWP	**63.20**	**67.06**	**68.70**	**48.93**	**53.30**	**55.39**





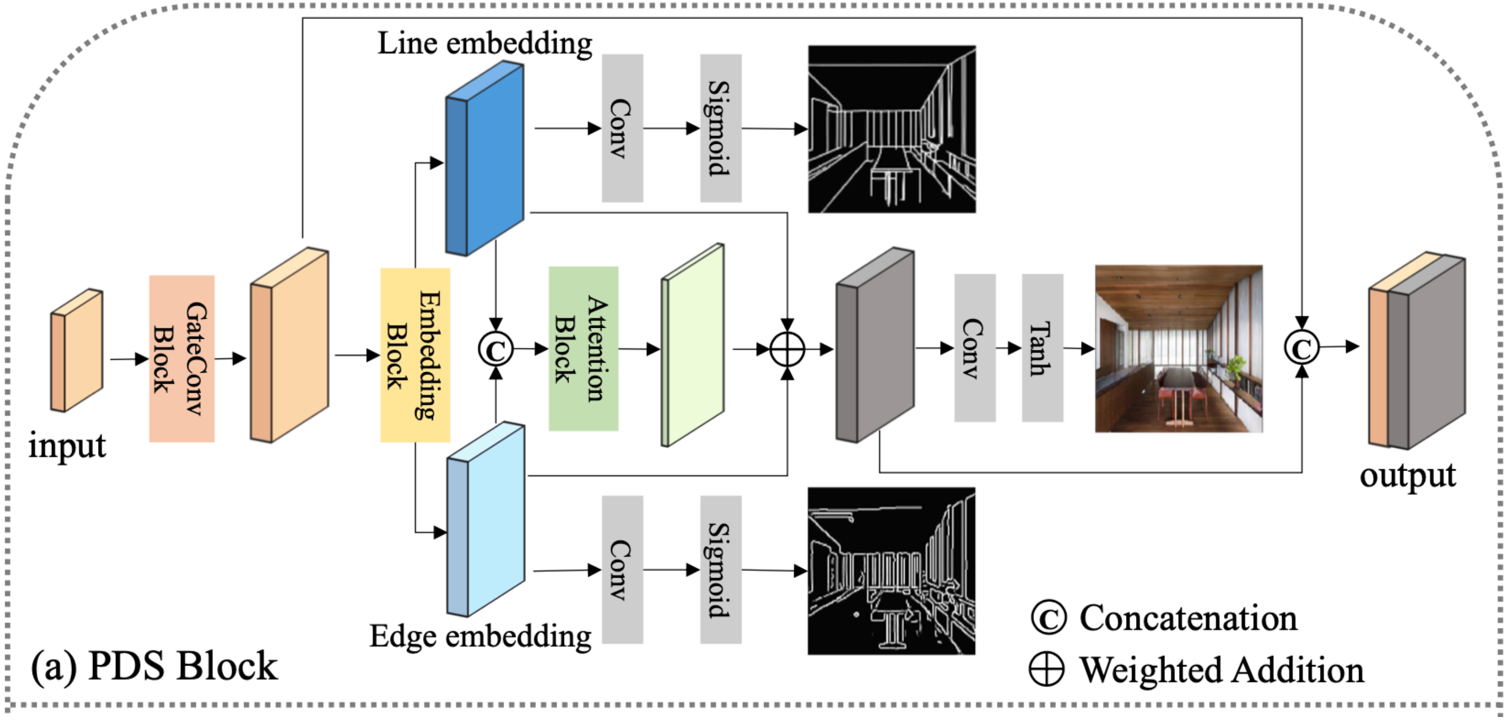
Pyramid Structure Sub-Encoder (PSS)



Pyramid Structure Sub-Encoder:

- Partially Gated Convolutions
- Efficient Attention Block
- Pyramid Decomposing Separable (PDS) Block

Pyramid Decomposing Separable (PDS)



- Learning line and edge embeddings respectively
- Embeddings are combined with a trade-off attention block to predict coarse inpainted results.
- Optimizing multi-scale structures with two discriminators for better decoupling of lines and edges.

Experiments: dataset

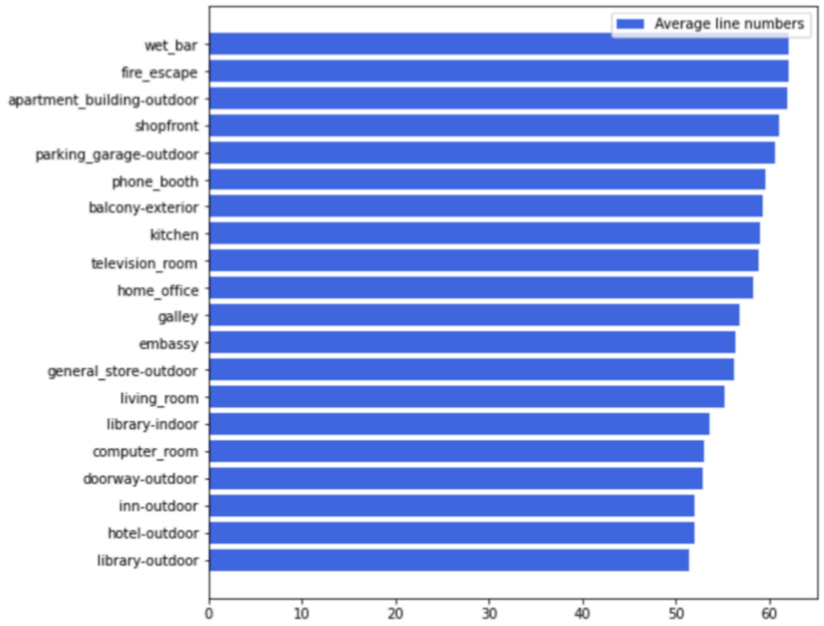


Figure 1. The bar chart of the scenes with top20 average line segment (confidence≥ 0.925) numbers of Places2.

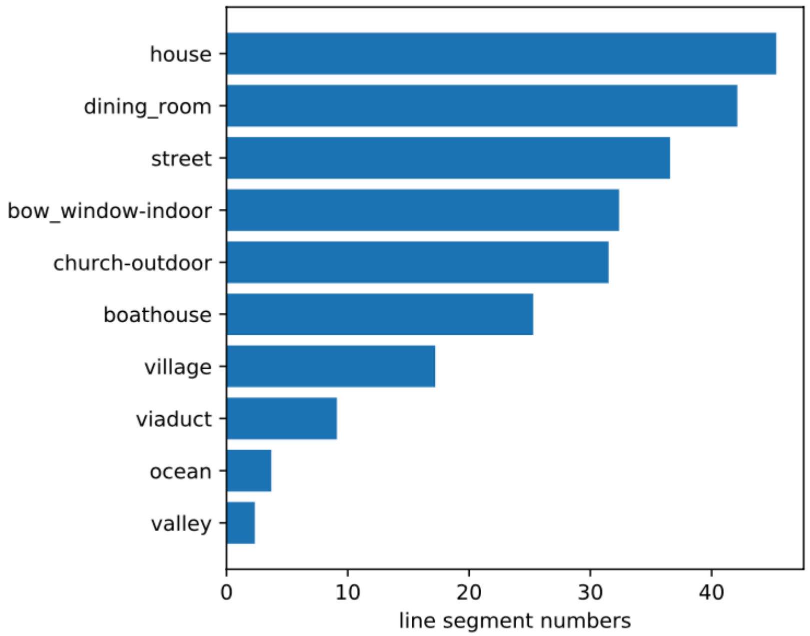


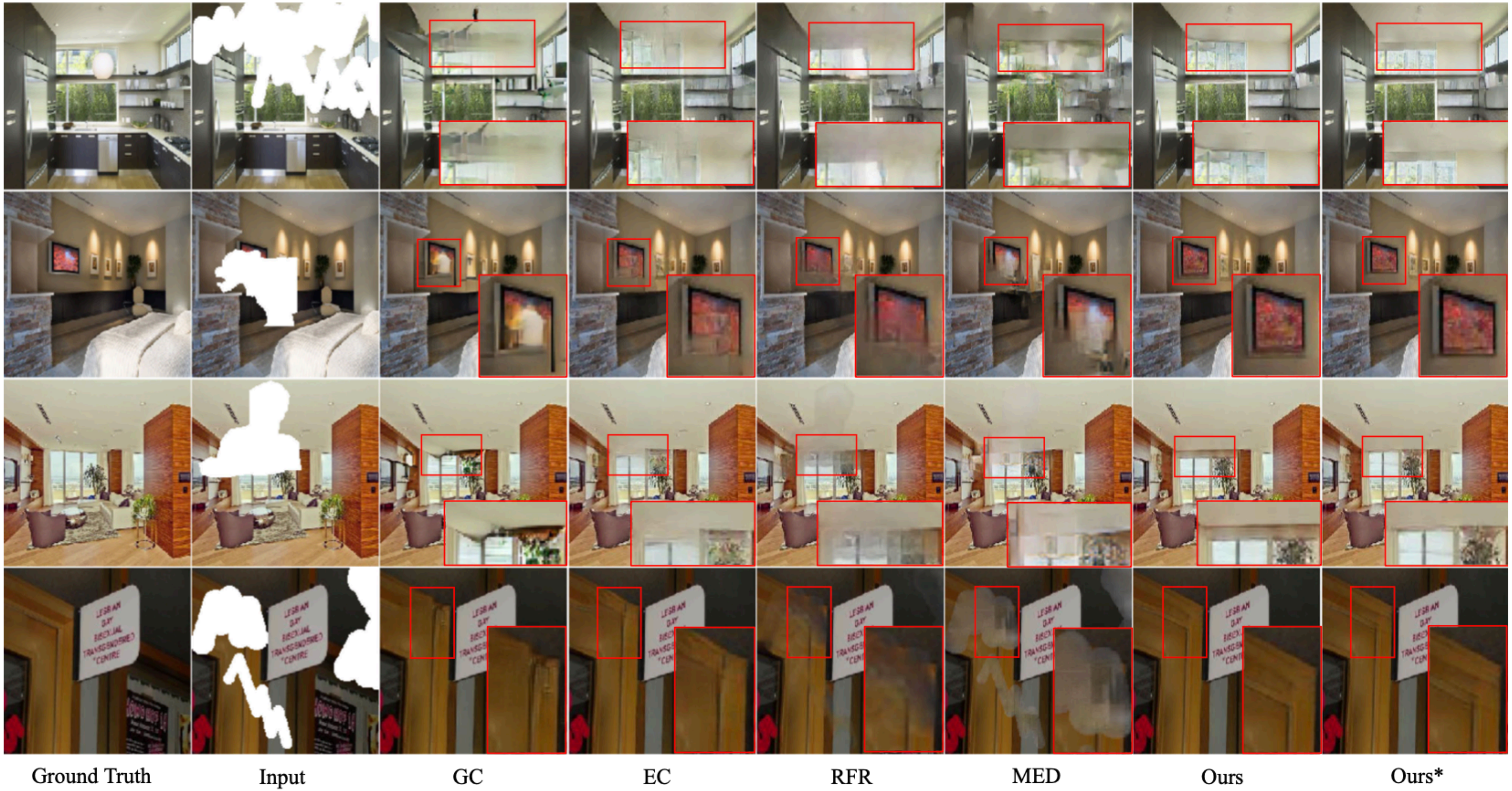
Figure 2. The bar chart of the line segment (confidence≥ 0.925) numbers of the comprehensive Places2 (P2C).

Datasets: (training/validation)

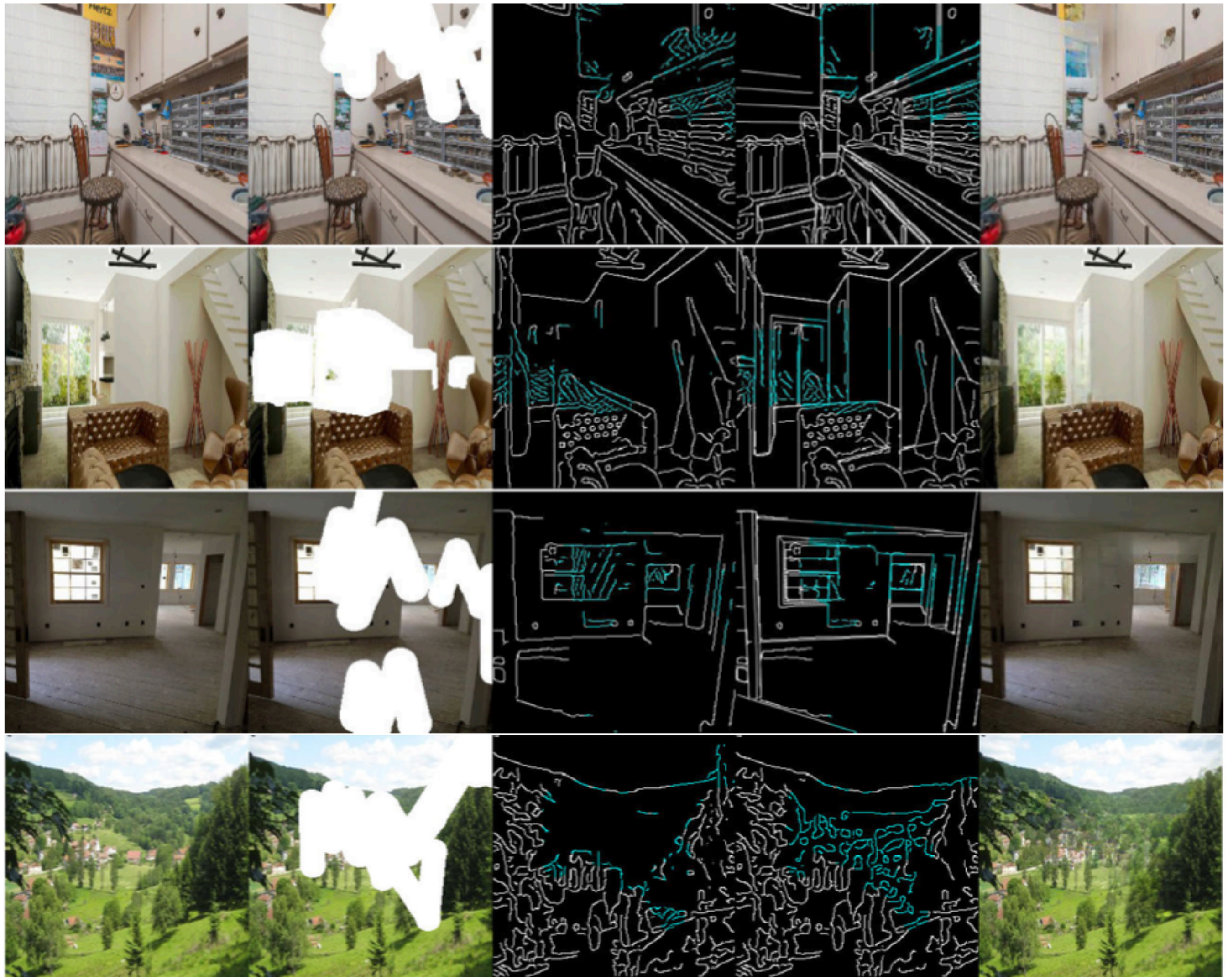
- ShanghaiTech (S.-T.) (5000/462)
- Man-made Places2 (P2M) (50000/1000)
- Comprehensive Places2 (P2C) (50000/1000)
- York Urban (Y.-U.) (-/102)

Experiments: Qualitative Results

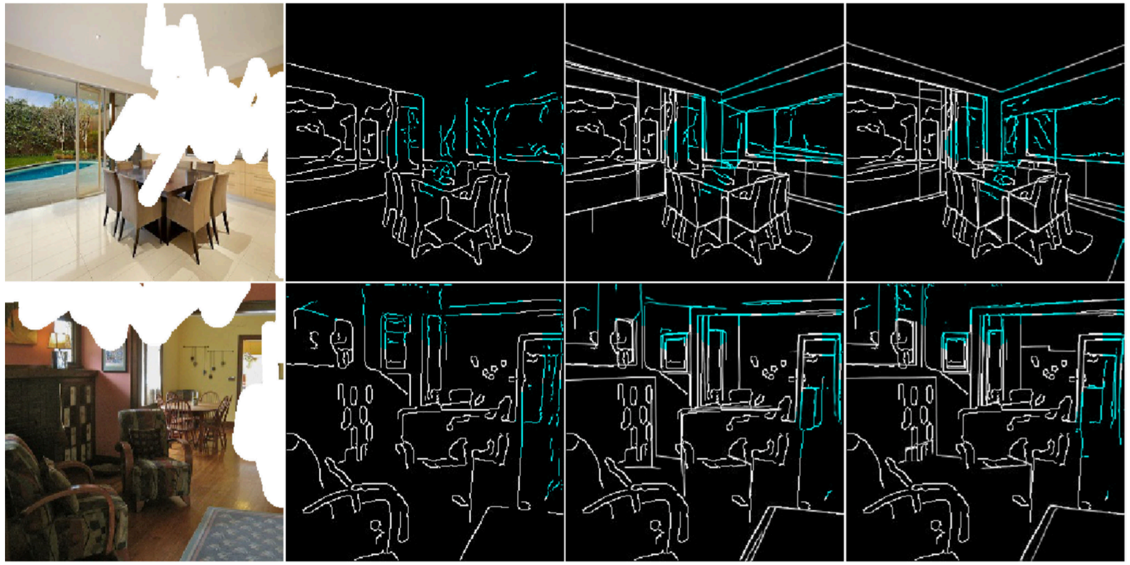
- * means the object removal mode



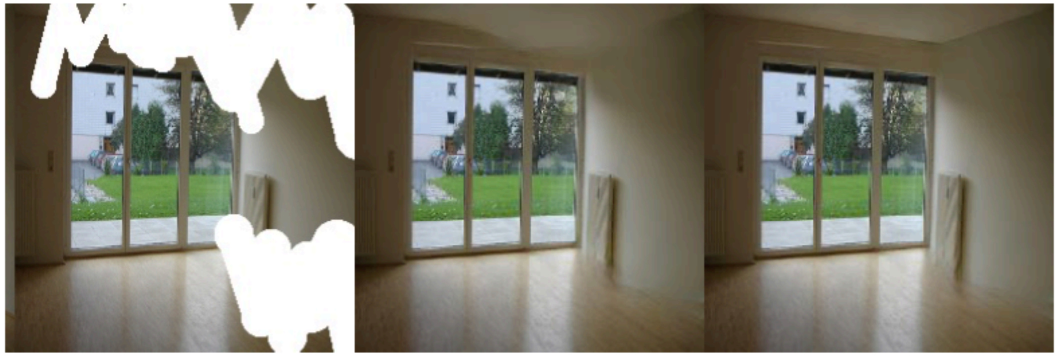
Experiments: Qualitative Results and Ablations



Ground truth | Masked input | EC edge | Ours edge | Ours results



(a) input (b) Ours w/o lines (c) Ours w/o PDS (d) Ours

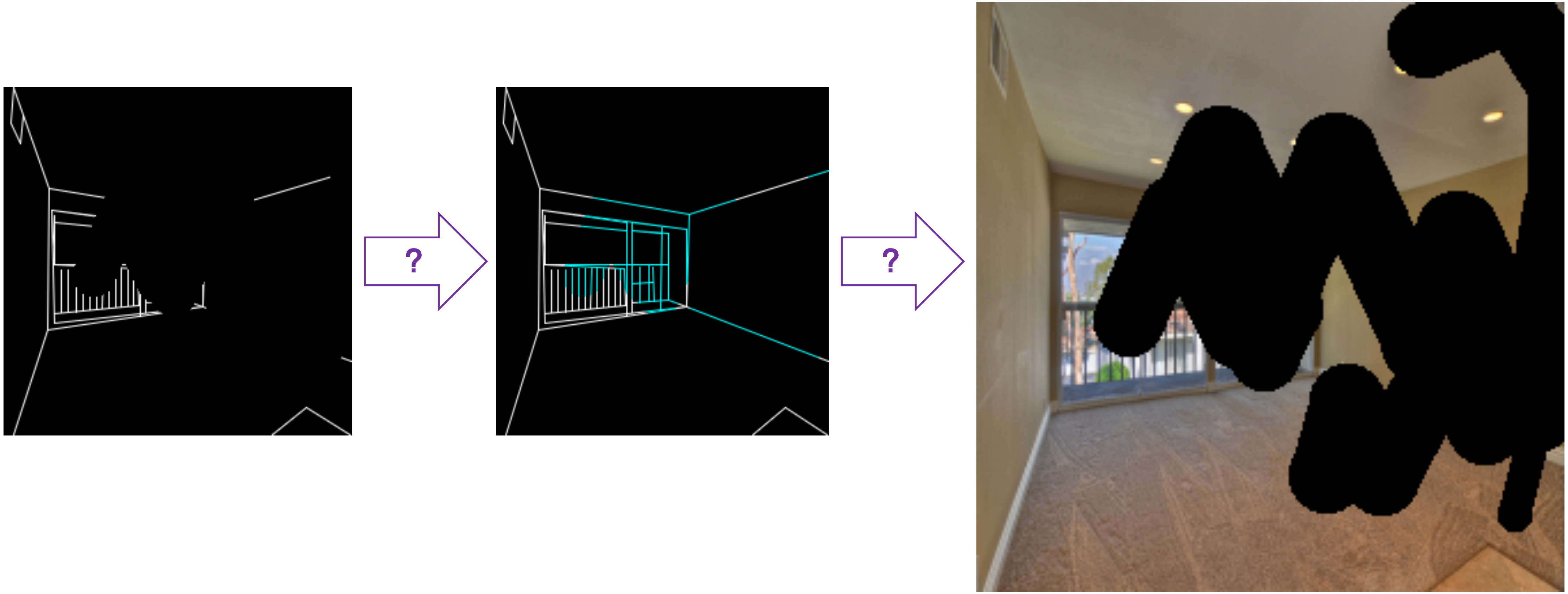


(a) input (b) PSS (edge) (c) PSS (edge+line)

Figure 4. Qualitative results w. and w./o. lines in ShanghaiTech.

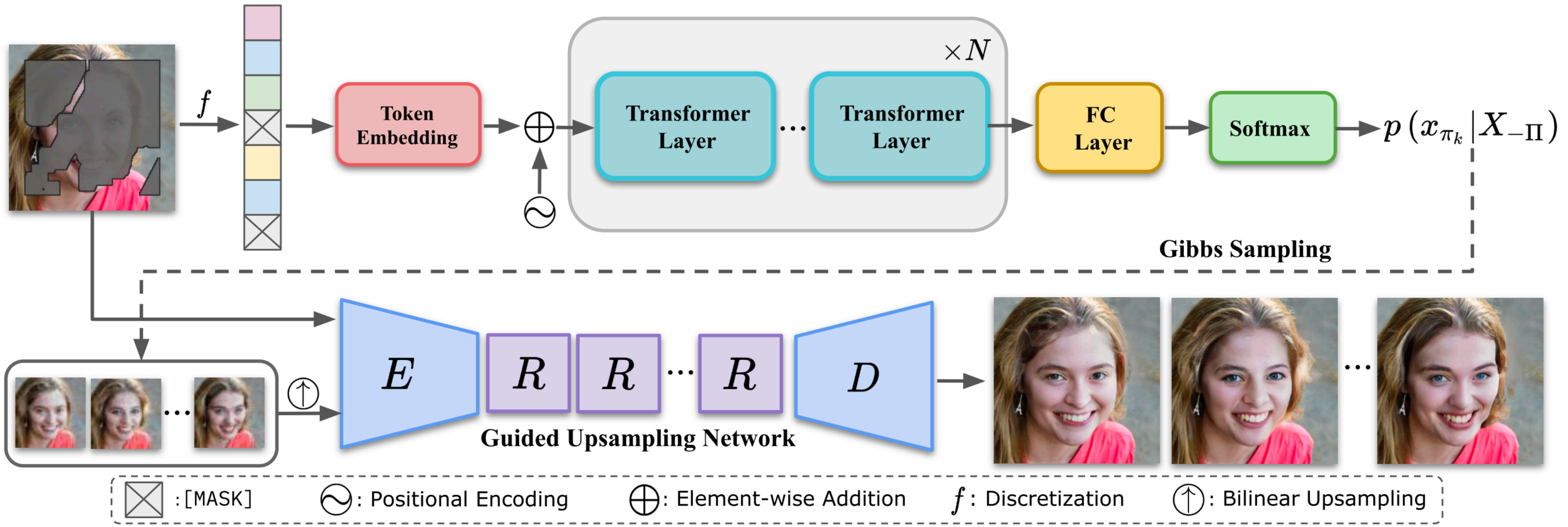
Experiments: Open Problems

- Are CNNs good enough to tackle the structural recovery?
- Can we extend the edge/line to the high-resolution inpainting?



How about modeling the priors with **Transformers**?

Preliminaries: Image Completion with Transformers (ICT)

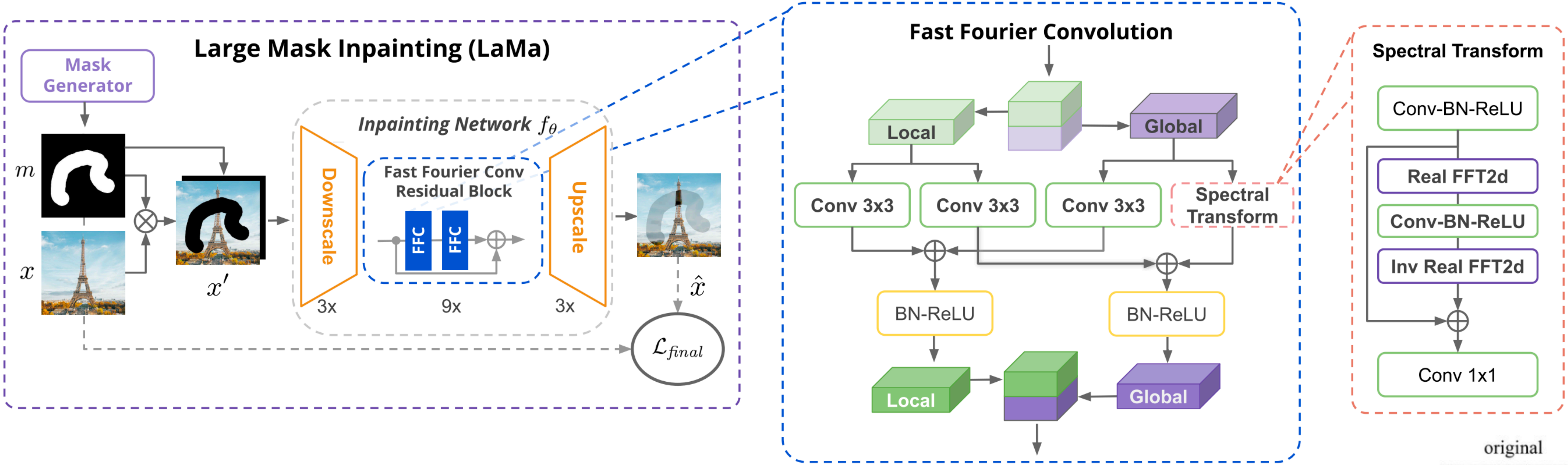


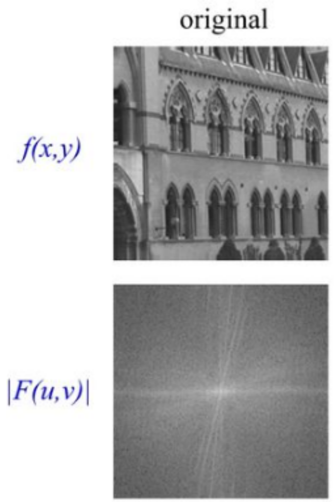
Recovering low-resolution images (priors) with bi-directional transformer; then using the guided Upsampling network (CNN) to recover high-resolution results

Attention is good at recovering structures

Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. ICCV 2021.

Preliminaries: Resolution-robust Inpainting with Fourier Conv (LaMa)





Fourier convolutions are used to for the high-resolution image inpainting

256x256 trained model can be generalized to high-resolution images

Roman Suvorov, Elizaveta Logacheva, et al. Resolution-robust large mask inpainting with fourier convolutions. WACV 2022.

Incremental Transformer Structure Enhanced Image Inpainting with Masking Positional Encoding (ZITS)

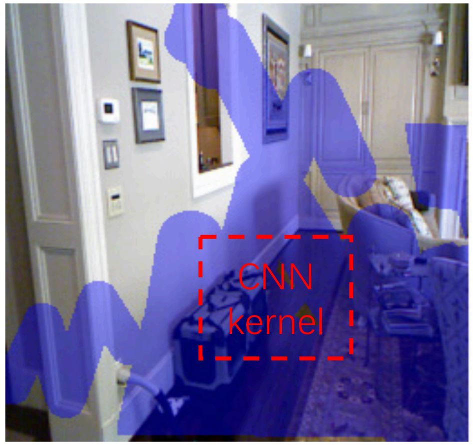
Qiaole Dong,* Chenjie Cao,* Yanwei Fu†
School of Data Science, Fudan University
{18307130096,20110980001,yanweifu}@fudan.edu.cn

CVPR 2022

Codes&Models: https://github.com/DQiaole/ZITS_inpainting

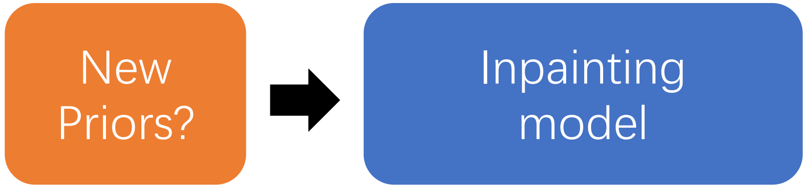
(a) Masked Image (b) LaMa (c) Ours

Challenges

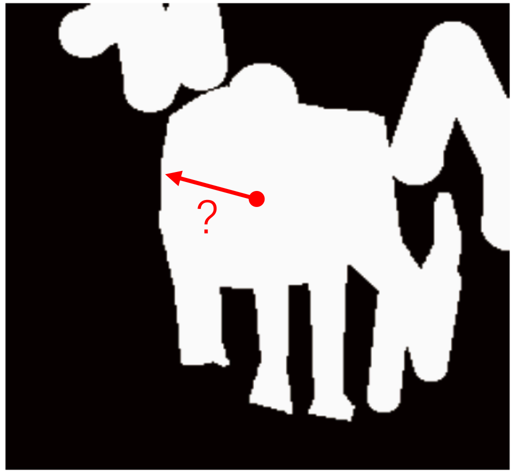


Limited receptive fields

Missing holistic structures

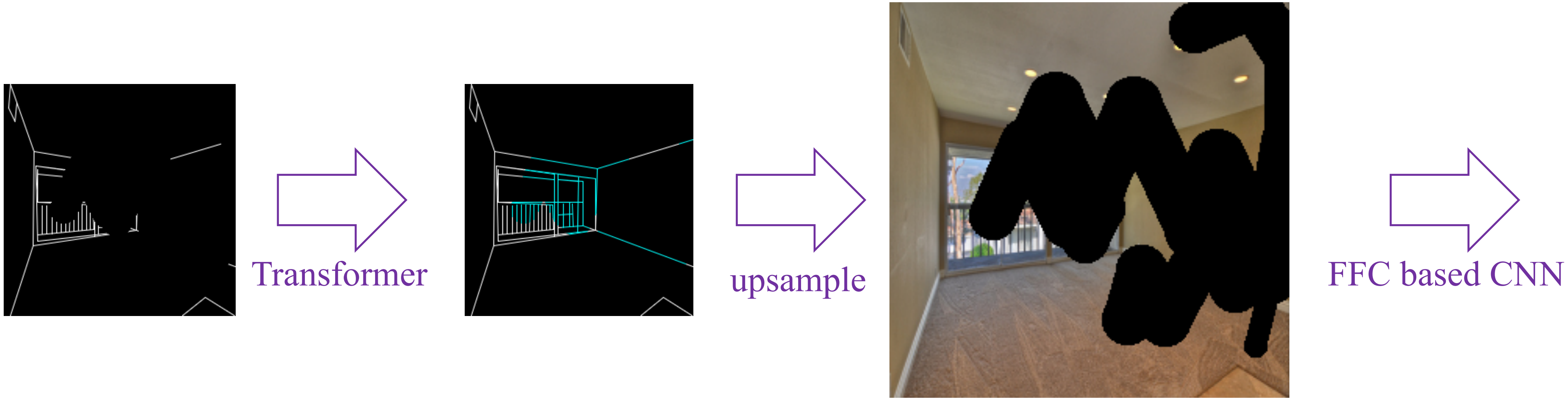


Heavy computations



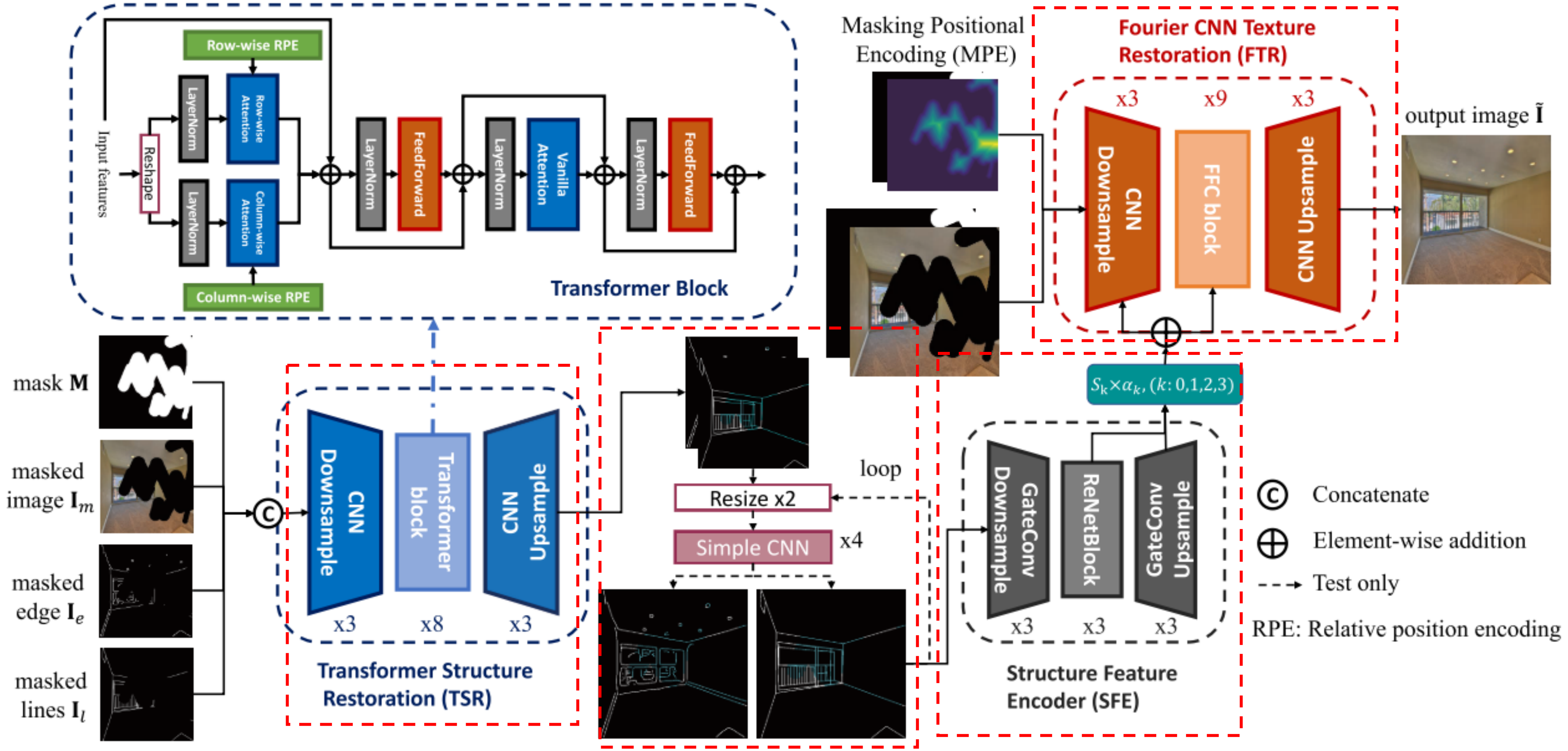
No positional information in masked regions

Motivation



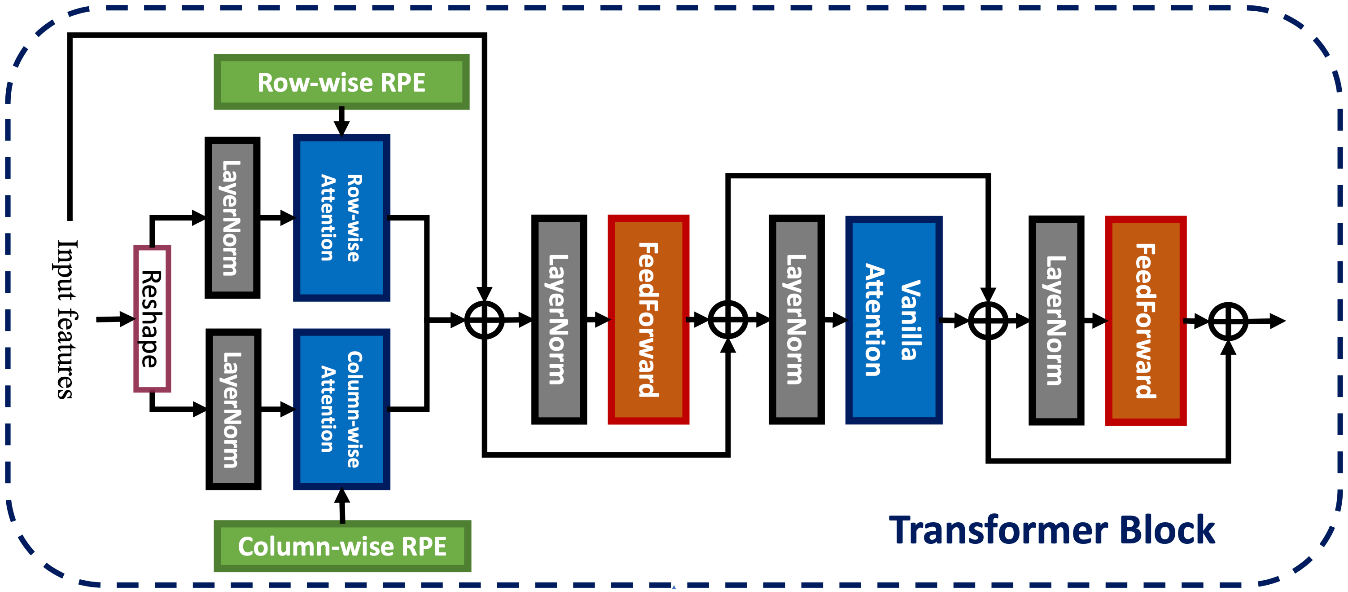
- Using Transformer to recover structures in sketch tensor space.
- Incrementally finetuning pre-trained inpainting models for additional structural priors.
- Introduce the positional encoding for masked regions.

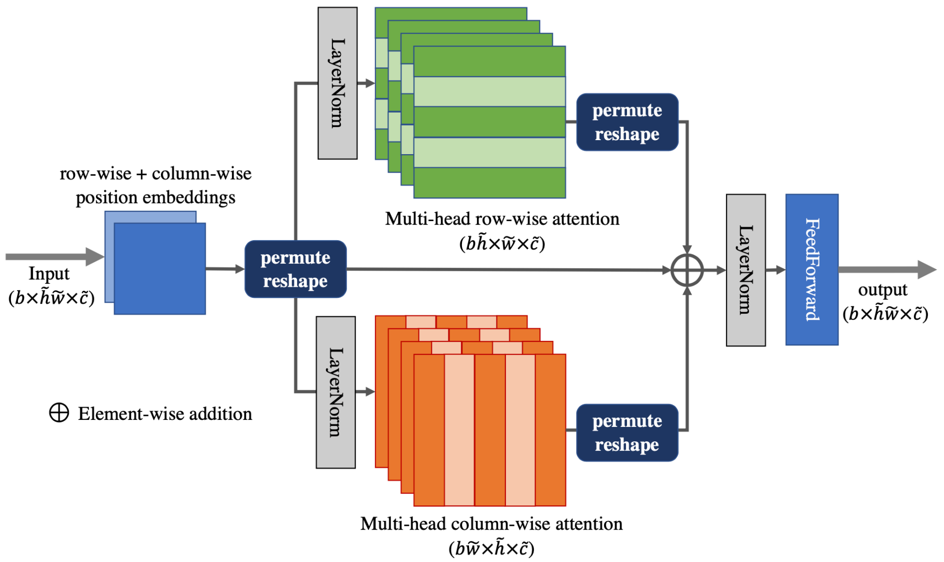
Overview



Input features
Reshape
LayerNorm
Row-wise Attention
Column-wise Attention
Row-wise RPE
Column-wise RPE
FeedForward
Vanilla Attention
Transformer Block
Masking Positional Encoding (MPE)
Fourier CNN Texture Restoration (FTR)
x3
x9
CNN Downsample
FFC block
CNN Upsample
output image $\tilde{\mathbf{I}}$
mask $\mathbf{M}$
masked image $\mathbf{I}_m$
masked edge $\mathbf{I}_e$
masked lines $\mathbf{I}_l$
Transformer block
x8
Transformer Structure Restoration (TSR)
loop
Resize x2
Simple CNN
x4
$S_k \times \alpha_k, (k: 0,1,2,3)$
GateConv Downsample
ReNetBlock
GateConv Upsample
Structure Feature Encoder (SFE)
Concatenate
Element-wise addition
Test only
RPE: Relative position encoding

Transformer Structure Restoration (TSR)



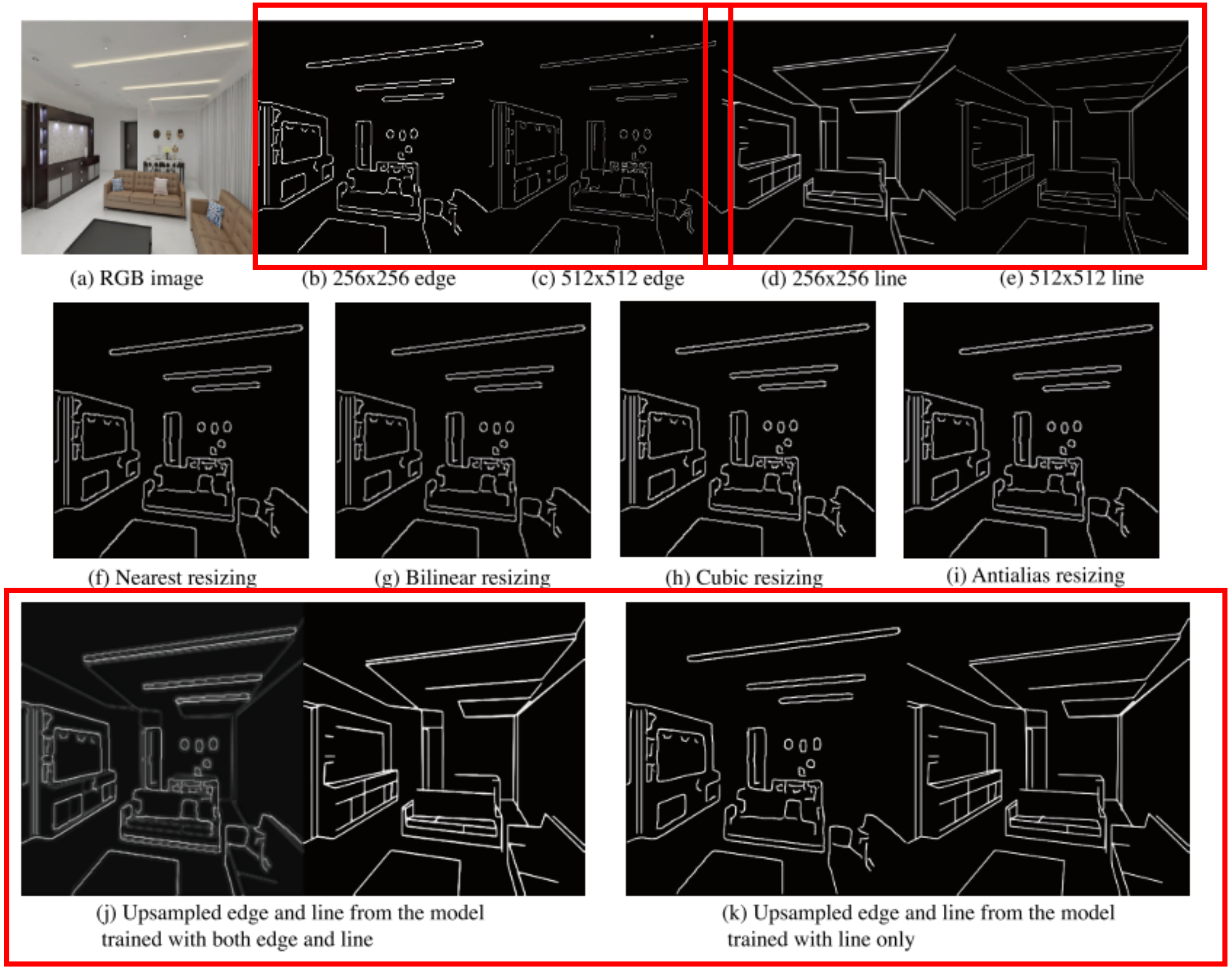


Using interleaved axial-transformer and vanilla-transformer to save computation and improve the performance.

	FPS	GPU Memory (MB)
w./o. Axial	6.41	14845
with Axial	**7.89**	**10547**

Axial	RPE	Edge P.	Edge R.	Edge F1	Line P.	Line R.	Line F1	Avg F1
		38.27	33.12	34.78	52.93	65.79	57.73	46.26
✔		**38.30**	32.90	34.64	52.74	**66.48**	57.87	46.26
✔	✔	37.34	**34.25**	**35.10**	**53.60**	66.23	**58.35**	**46.72**

Simple Structure Upsampler (SSU)

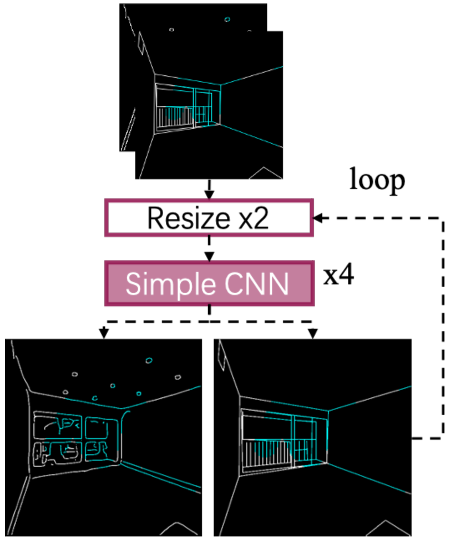


(a) RGB image (b) 256x256 edge (c) 512x512 edge (d) 256x256 line (e) 512x512 line

(f) Nearest resizing (g) Bilinear resizing (h) Cubic resizing (i) Antialias resizing

(j) Upsampled edge and line from the model trained with both edge and line

(k) Upsampled edge and line from the model trained with line only



Ambiguities between 256 canny edges (b) and 512 canny edges (c).

Discrete lines are consistent in both 256x256 and 512x512.

Optimized by **discrete lines** (k) works better than **lines and edges** (j).

Masked Positional Encoding (MPE)

Table 3. Ablation studies of MPE on 512×512 Places2 finetuned with dynamic resolutions from 256 to 512.

	PSNR↑	SSIM↑	FID↓	LPIPS↓
with MPE	**24.23**	**0.881**	**26.08**	**0.133**
w./o. MPE	24.20	0.880	26.29	0.135

Figure 9. Ablations of 512×512 Places2 with and without MPE.

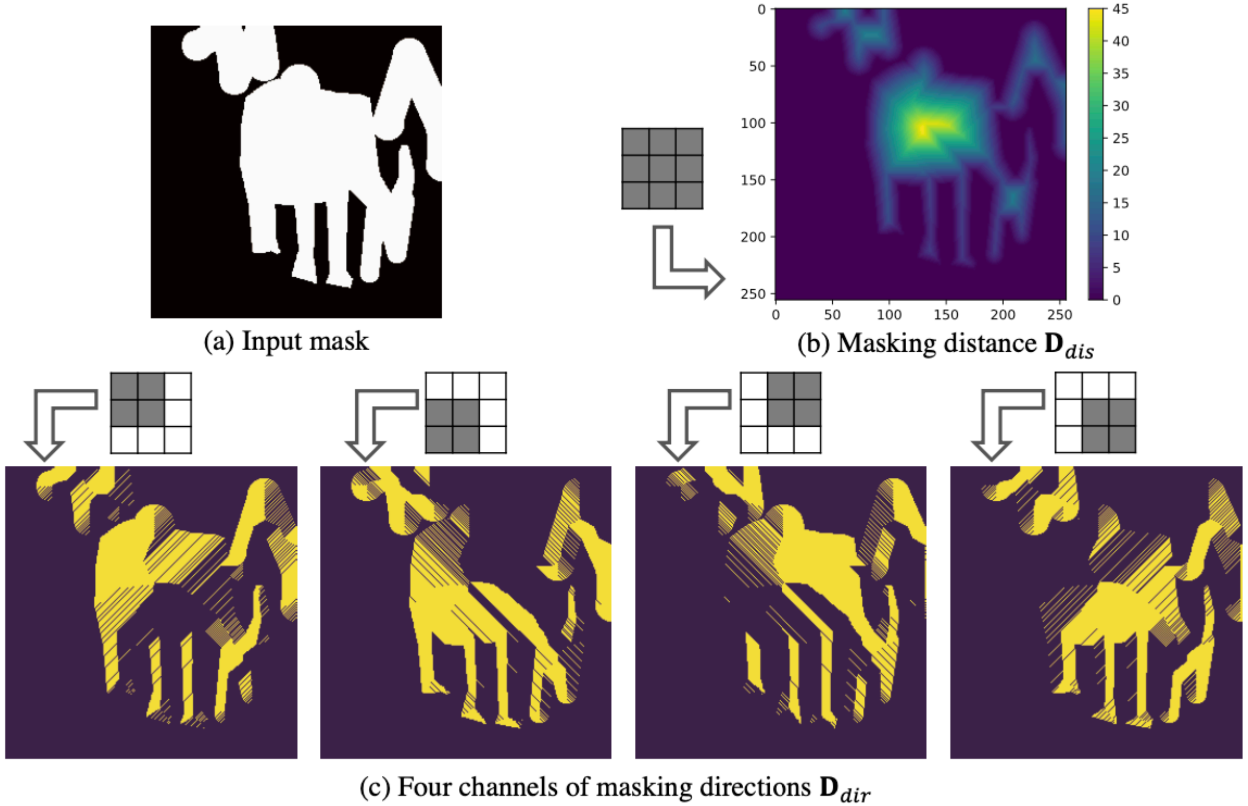
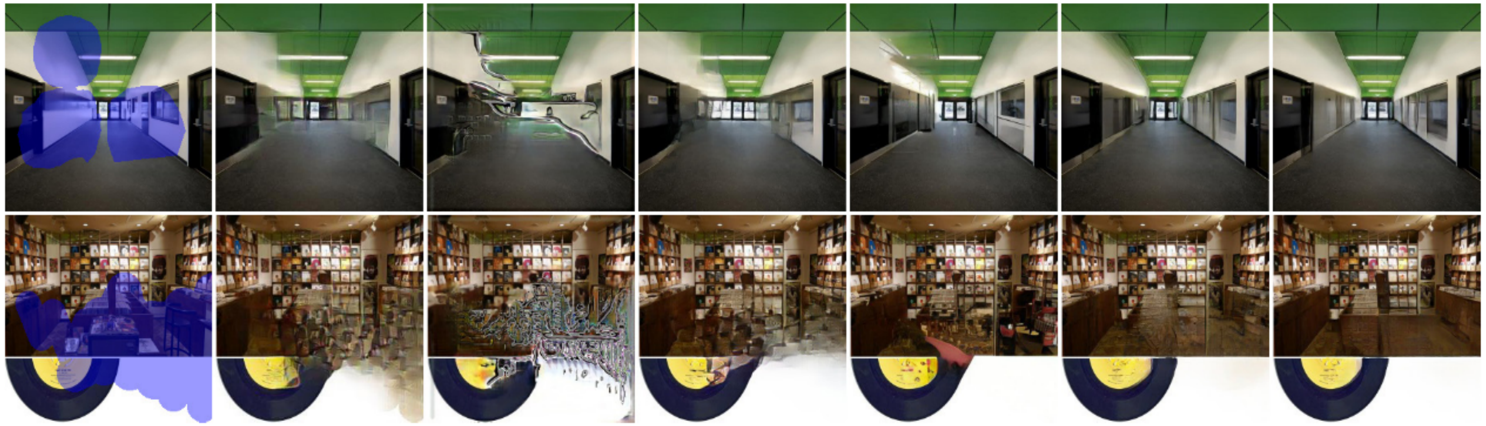


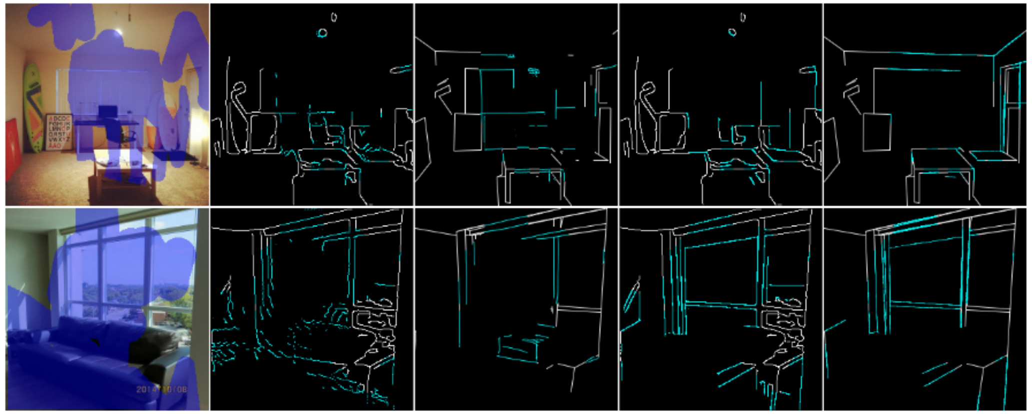
Figure 4. The illustration of our masking relative position encoding. (a) Input mask, (b) masking distance $\mathbf{D}_{dis}$ and the all-one 3×3 kernel, (c) masking directions $\mathbf{D}_{dir}$ and their kernels.

Qualitative results

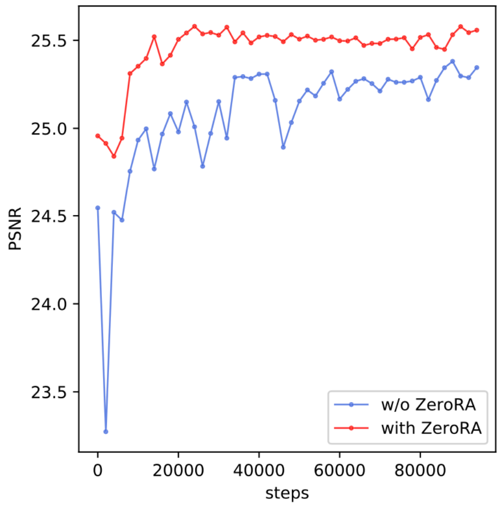
(a) Masked Input (b) EC (c) MST (d) LaMa (e) Ours

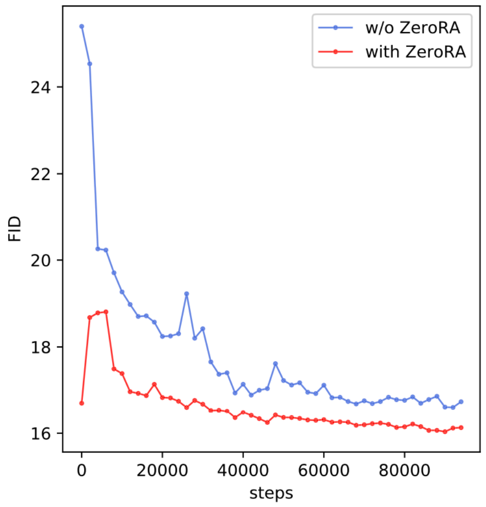


(a) Masked Input (b) EC (c) HiFill (d) MST (e) Co-Mod (f) LaMa (g) Ours

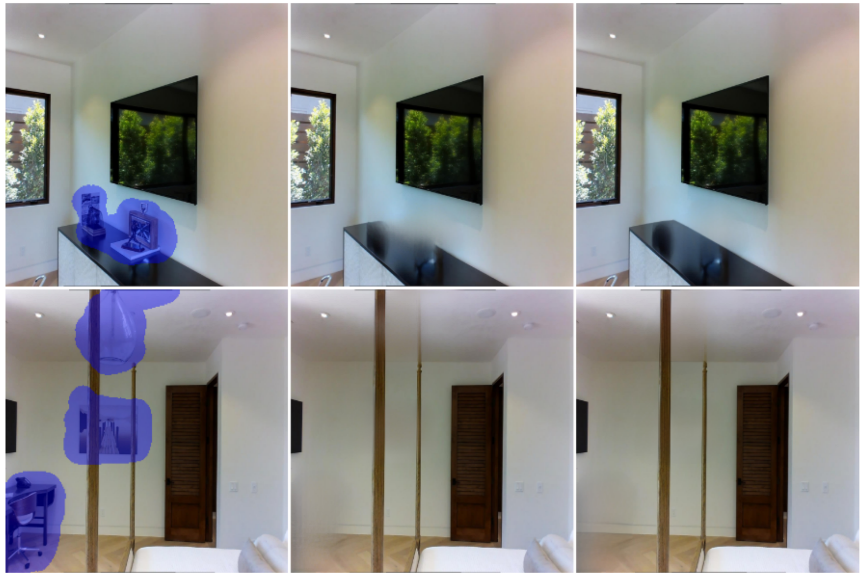


(a) Masked input (b) Edges from MST (c) Lines from MST (d) Edges from ours (e) Lines from ours

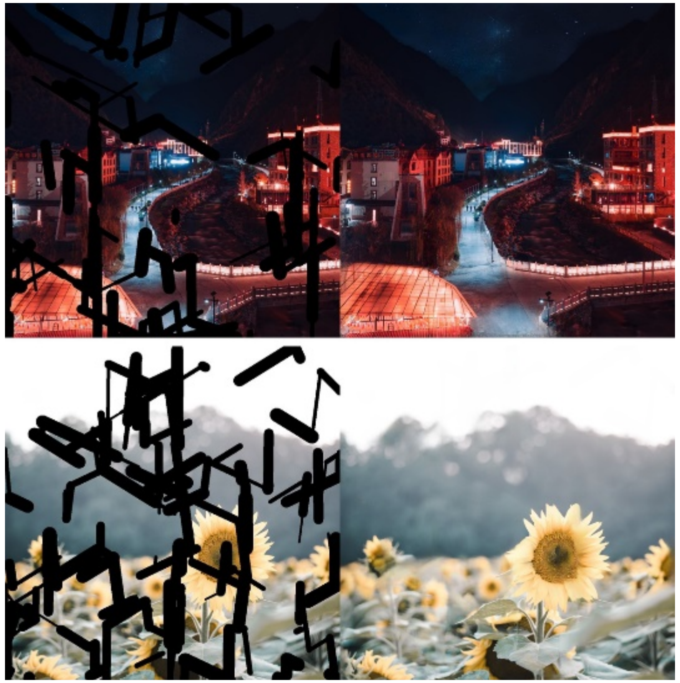




1024x1024 Inpainting Results



(a) Masked Input (b) LaMa (c) Ours

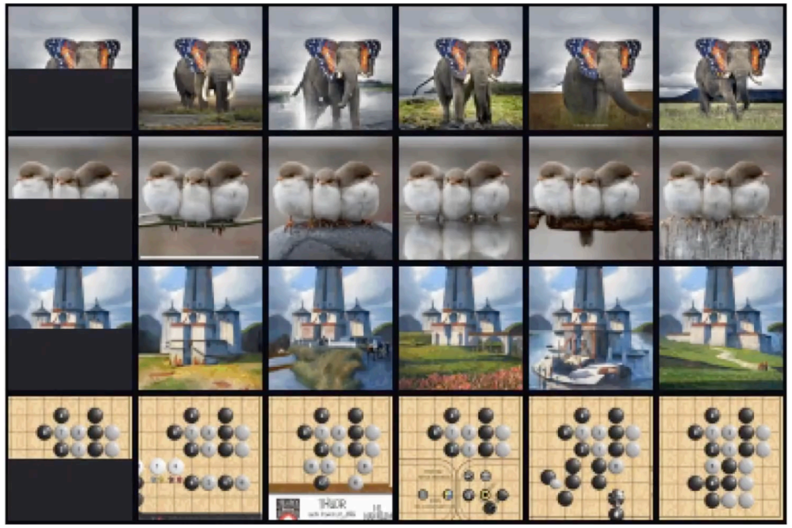


(a) Masked Image (b) LaMa (c) Ours

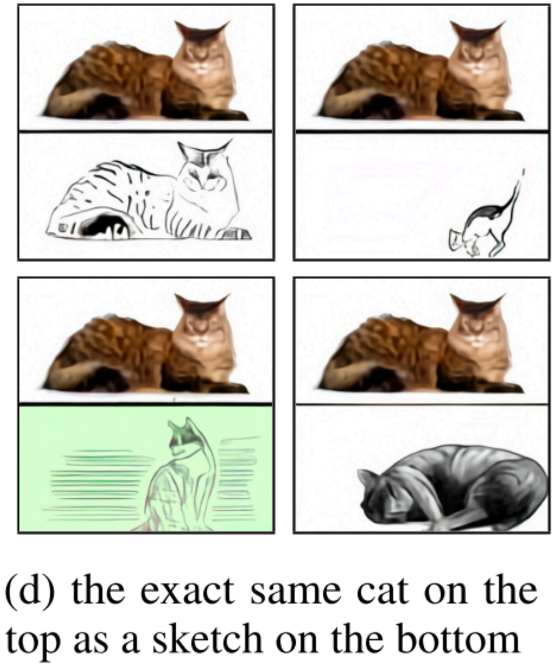
Image Editing

2

Transformer-based image generation



(a) iGPT[1]



(b) DALLE[2]

(c) Taming[3]

[1] Chen M, Radford A, Child R, et al. Generative pretraining from pixels[C] PMLR, 2020.

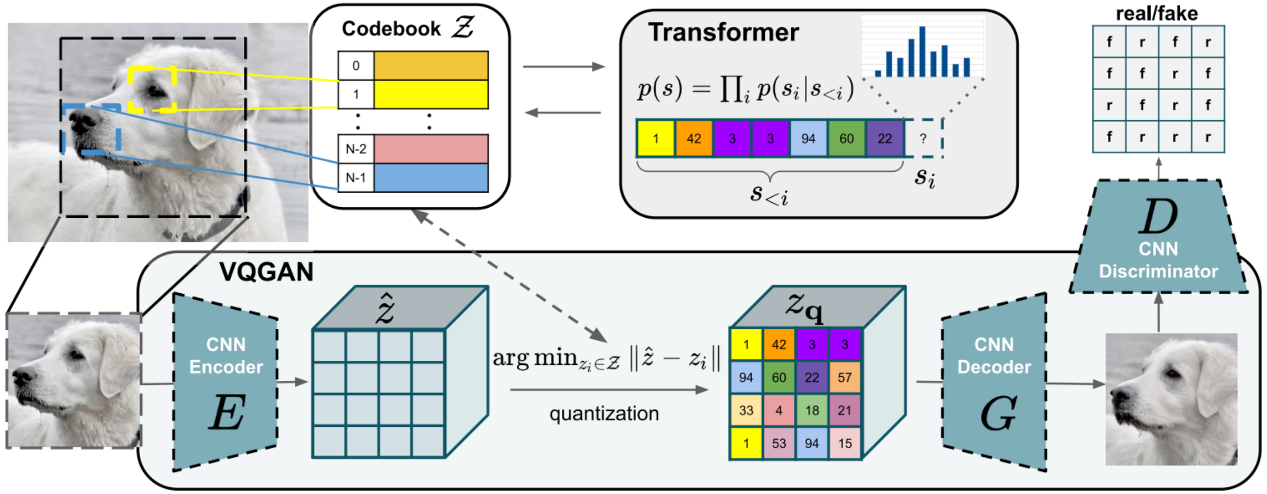
[2] Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation[J]. arXiv preprint arXiv:2102.12092, 2021.

[3] Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis[C] CVPR, 2021.

Two important mechanisms in Transformer generation

- 1. Patch-wise Autoregressive Generation

- 2. Discreate Learning (DALLE, Taming, NvWA)



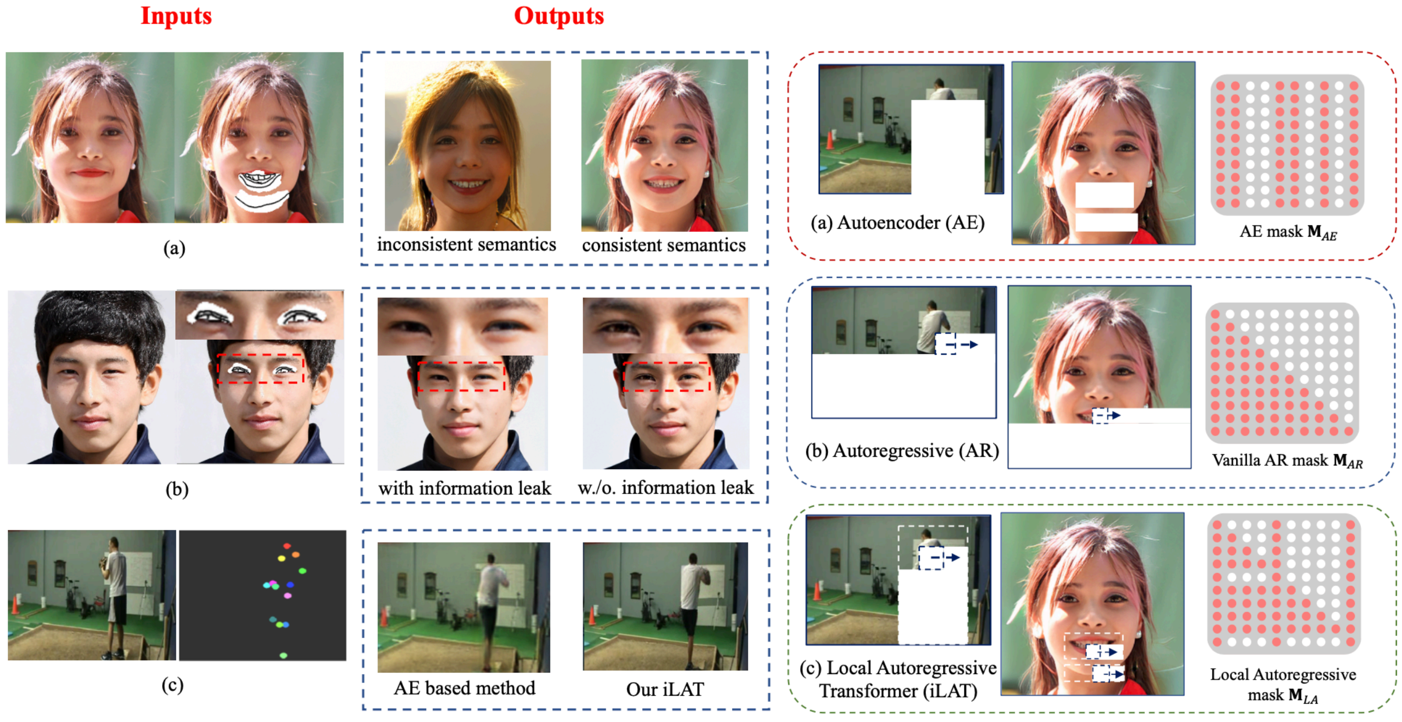
Esser, Patrick, Robin Rombach, and Bjorn Ommer. "Taming transformers for high-resolution image synthesis." CVPR 2021.

The Image Local Autoregressive Transformer

Chenjie Cao, Yuxin Hong, Xiang Li, Chengrong Wang, Chengming Xu, Yanwei Fu,* Xiangyang Xue
School of Data Science
Fudan University
{20110980001,yanweifu}@fudan.edu.cn

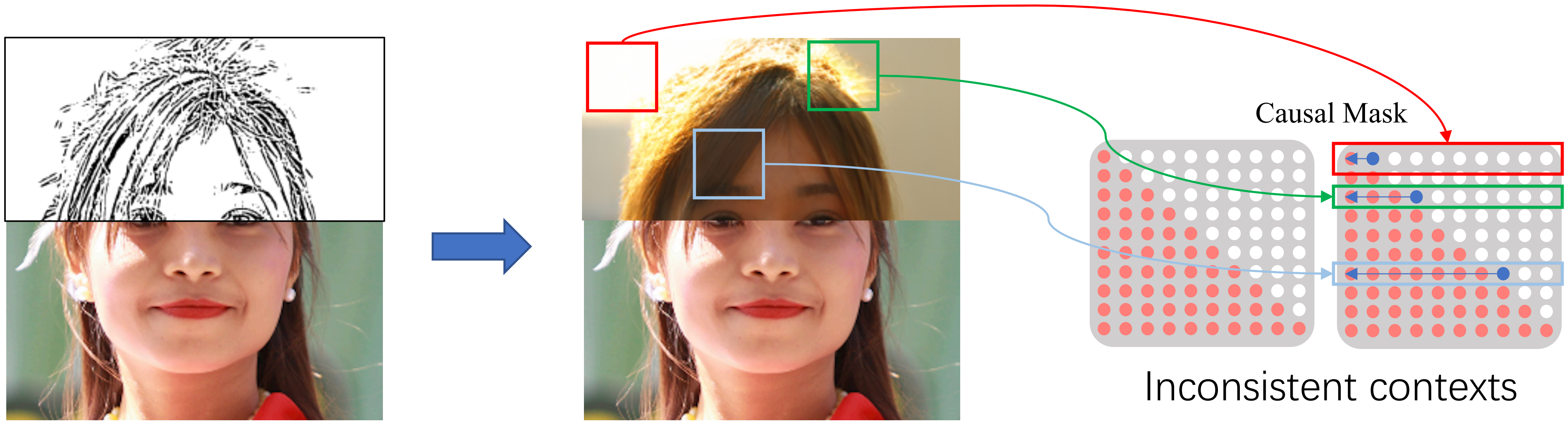
NeurIPS2021

Codes&Models: **https://github.com/ewrfcas/iLAT**

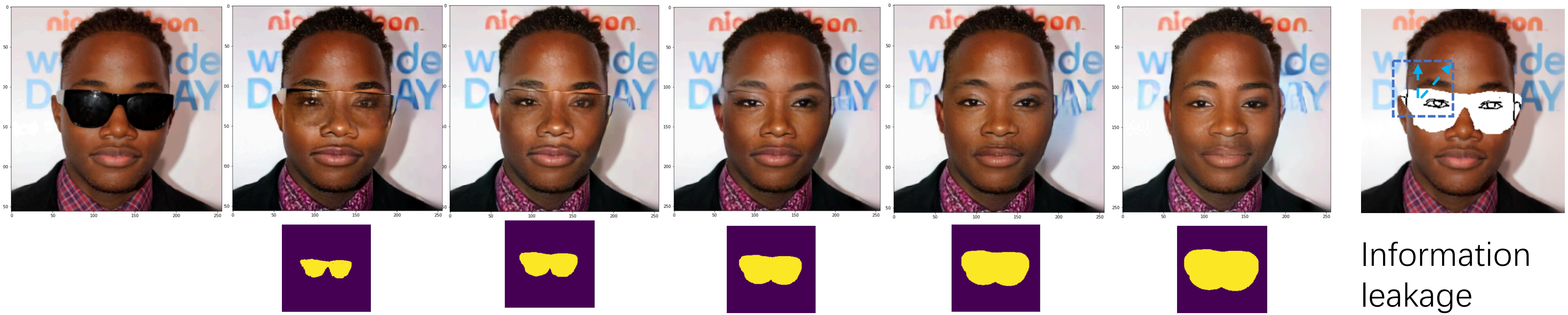


(A) Inputs and outputs of local generation compared with previous works (B) Comparison of different generative modes

Problems of the Autoregressive Generation

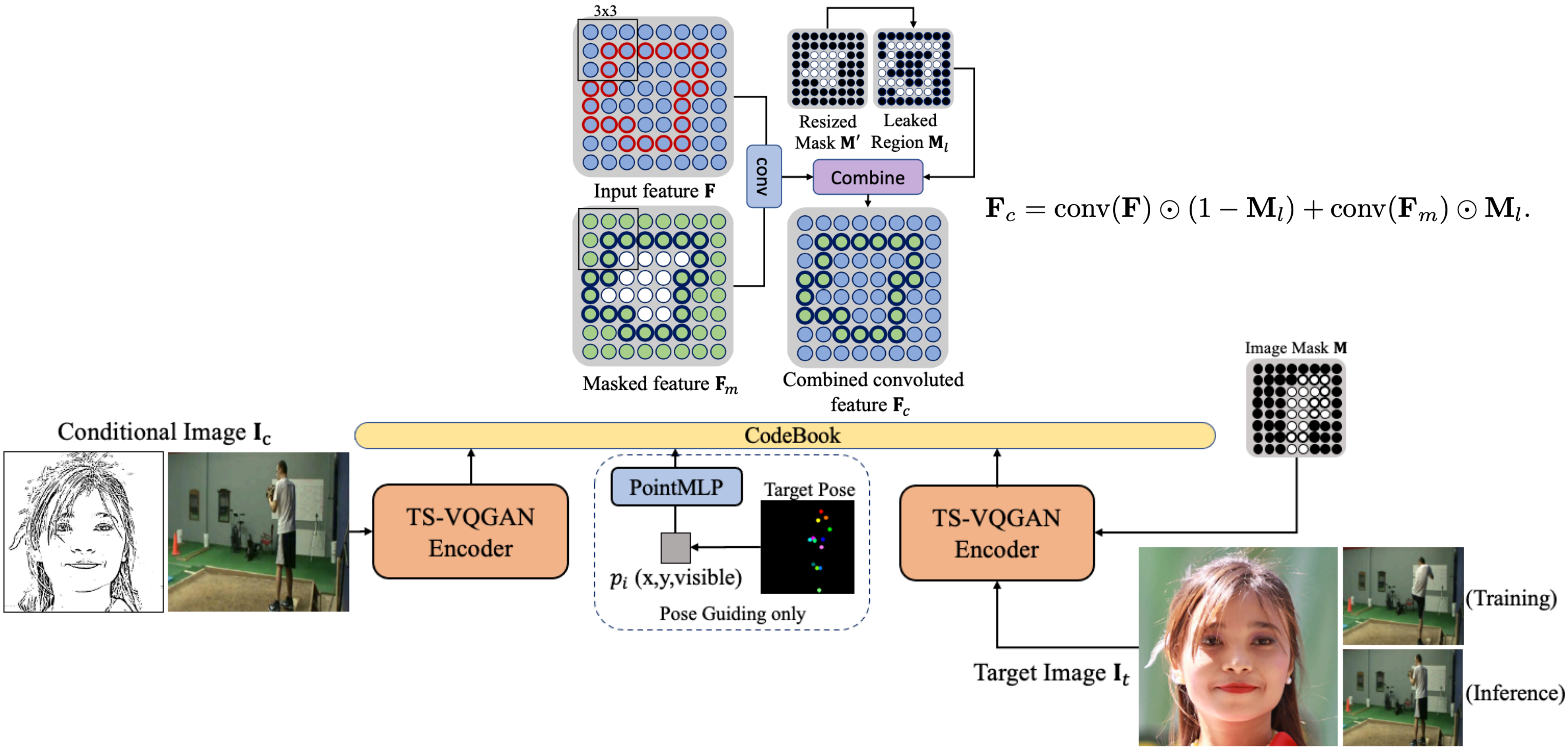


Inconsistent contexts



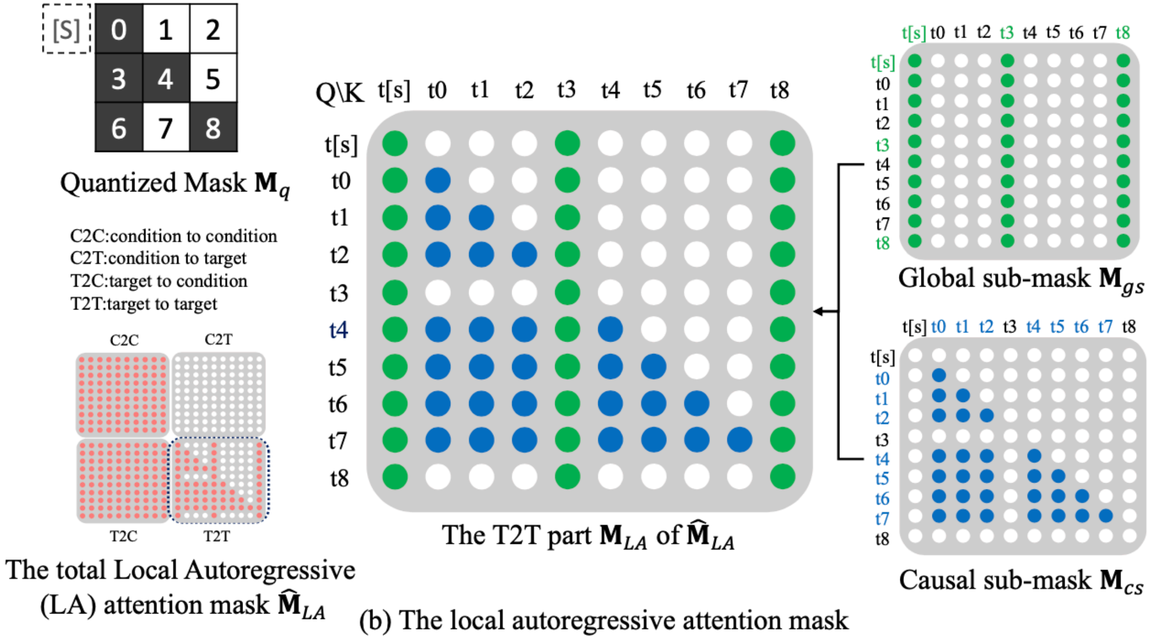
Pipeline (VQGAN->TS-VQGAN)

Two-stream convolution based VQGAN



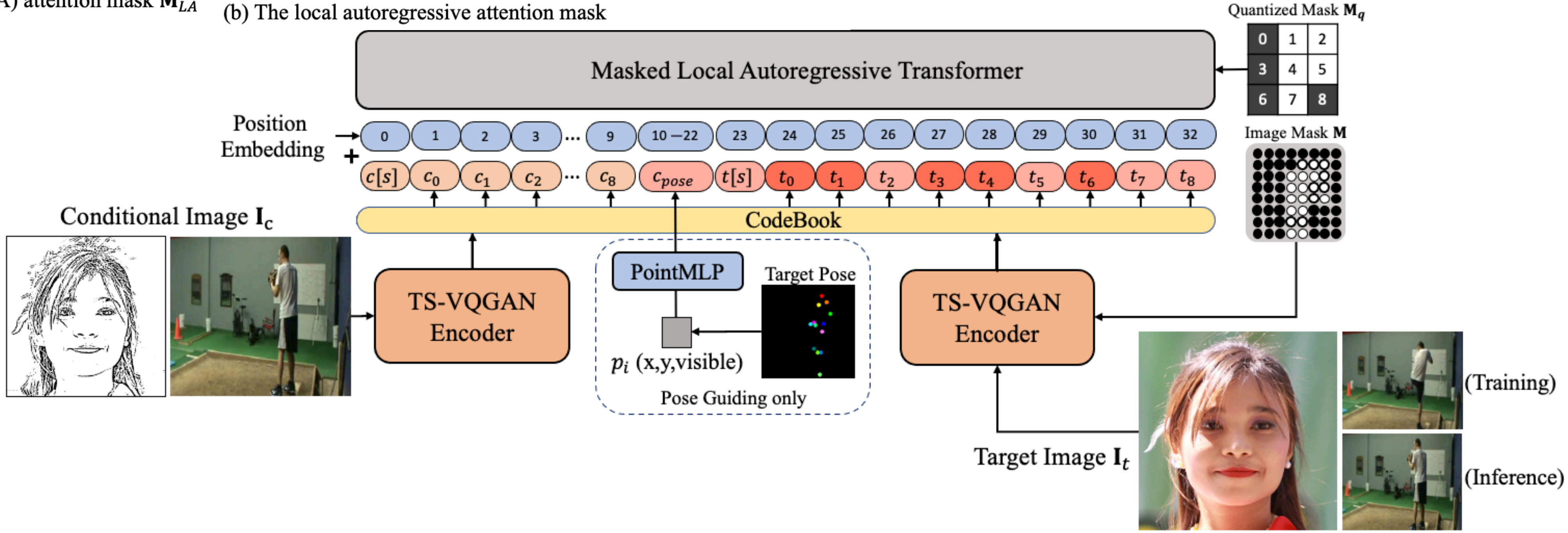
$$\mathbf{F}_c = \mathrm{conv}(\mathbf{F}) \odot (1 - \mathbf{M}_l) + \mathrm{conv}(\mathbf{F}_m) \odot \mathbf{M}_l.$$

Pipeline (Local Autoregressive Transformer)

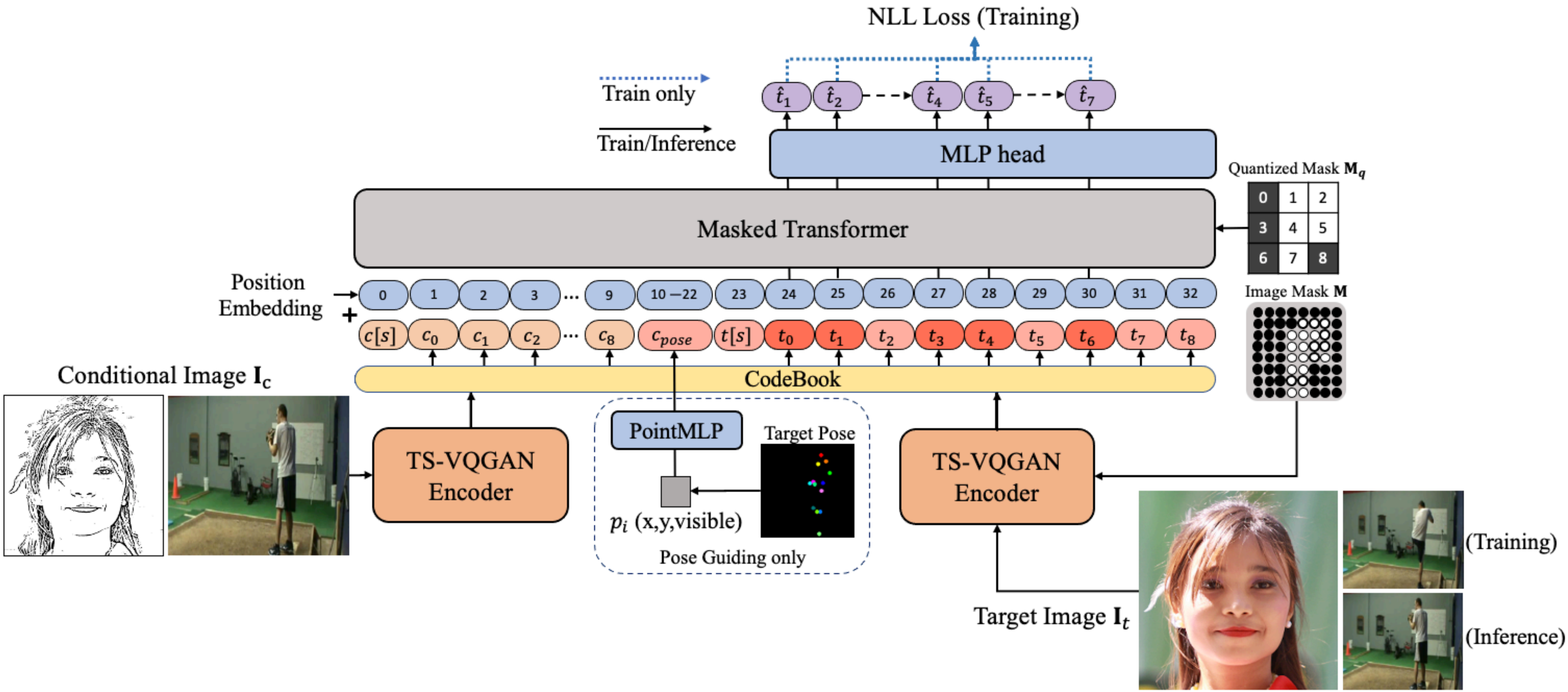


Tokens are splited into global tokens and causal tokens.

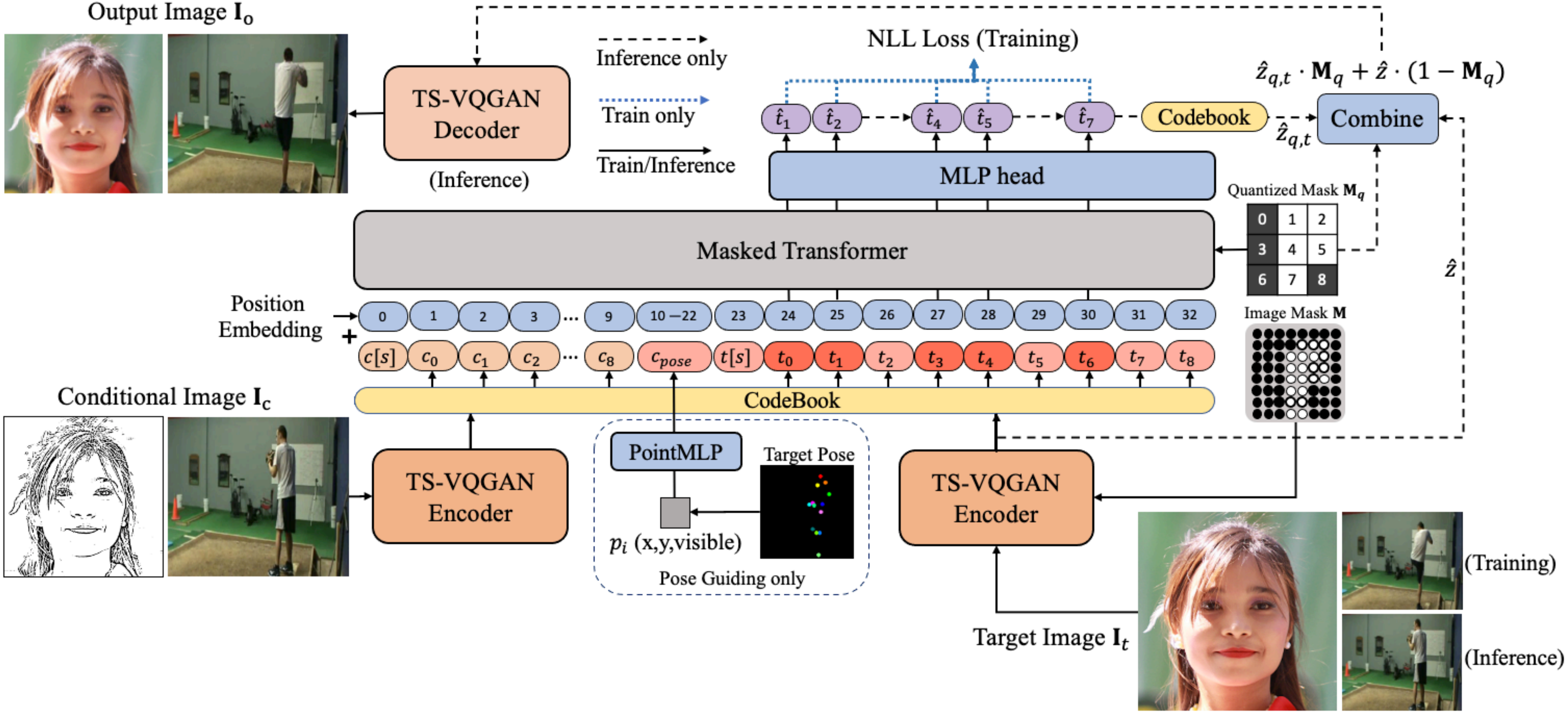
$$p(t_m|c, t_u) = \prod_j p(t_{(m,j)}|c, t_u, t_{(m,<j)}).$$



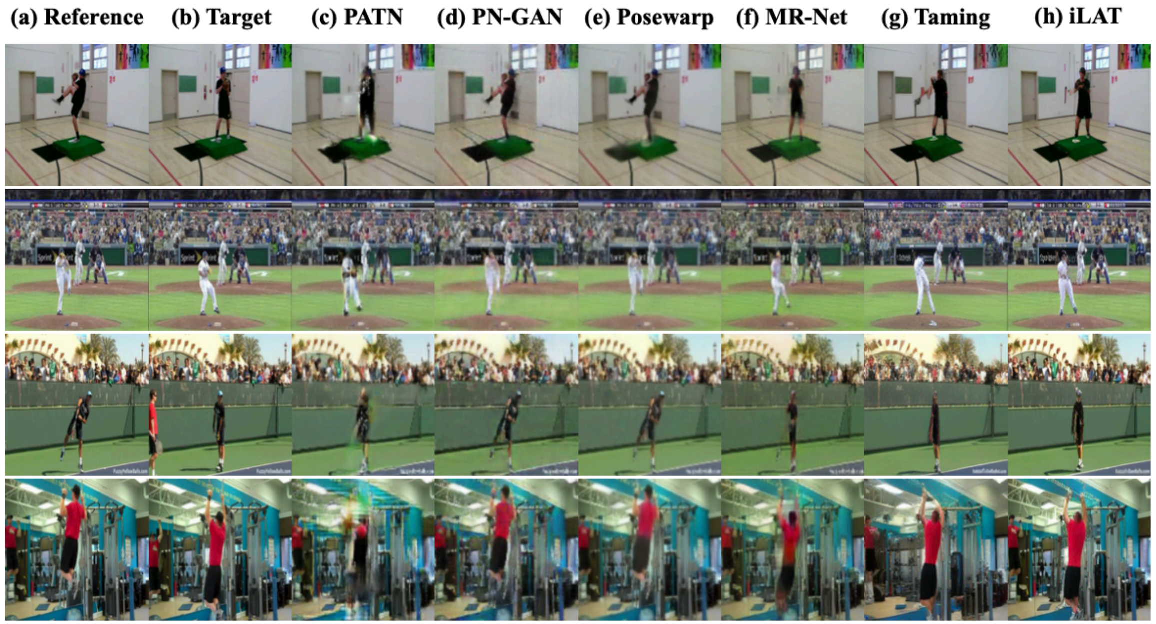
Pipeline (Training Loss)



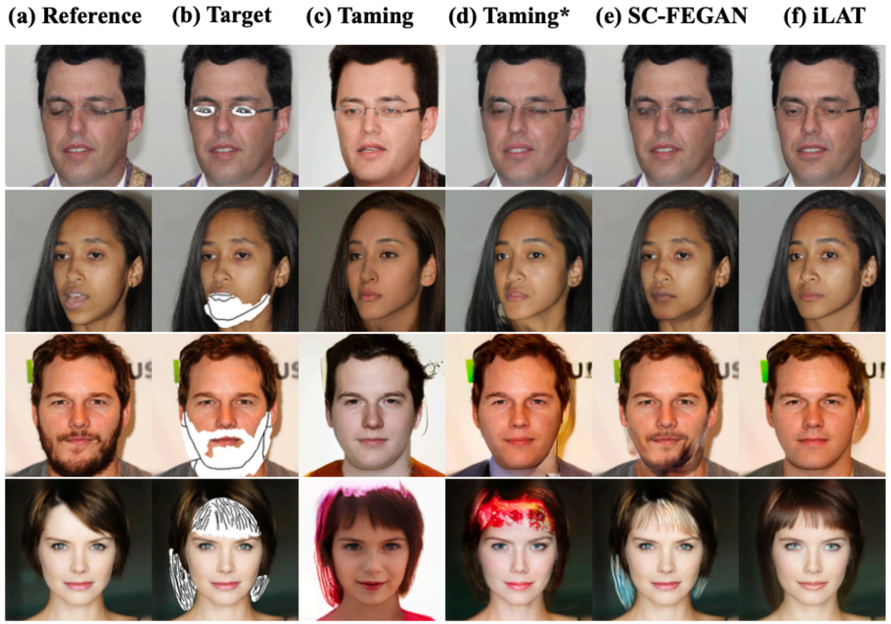
Pipeline (Inference)



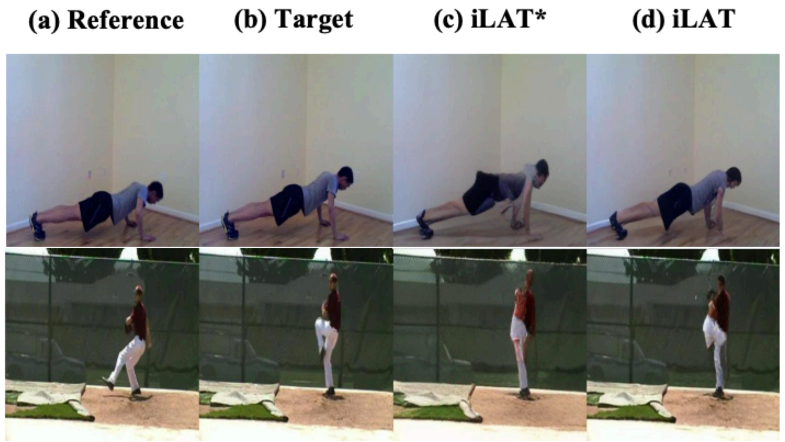
Qualitative Results and Ablations



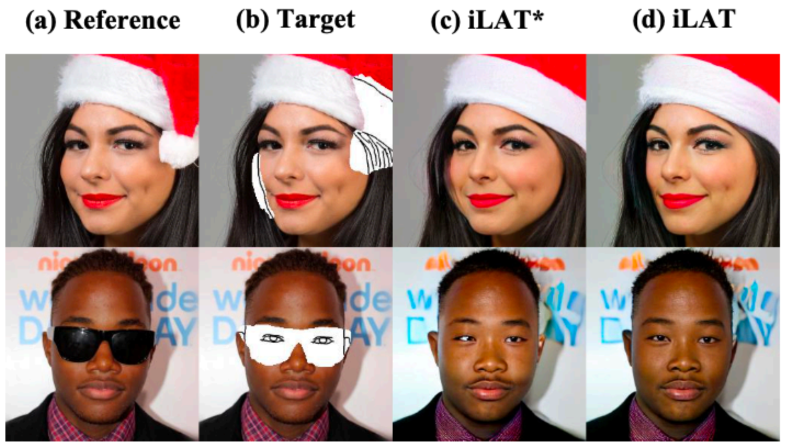
(A) Pose-Guided Generation in PA.



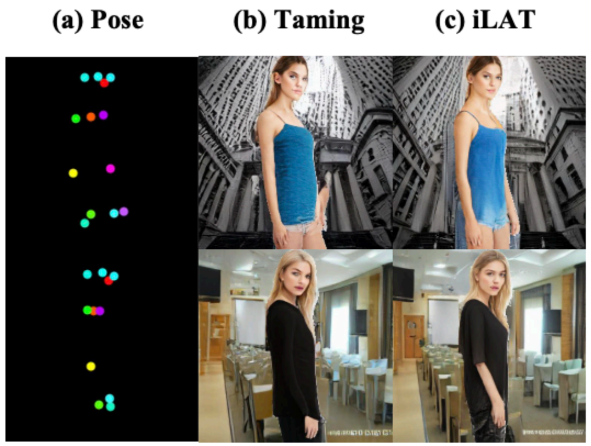
(B) FFHQ (row 1, 2) and CelebA (row 3, 4).



(A) Ablation in pose guiding



(B) Ablation in face editing



(C) Qualitative results in SDF

High-fidelity Portrait Editing via Exploring Differentiable Guided Sketches from the Latent Space

Chengrong Wang[*] *Chenjie Cao*[†] *Yanwei Fu*[†] *Xiangyang Xue*[†*]

[*] School of Computer Science, Fudan University, Shanghai, China
[†] School of Data Science, Fudan University, Shanghai, China

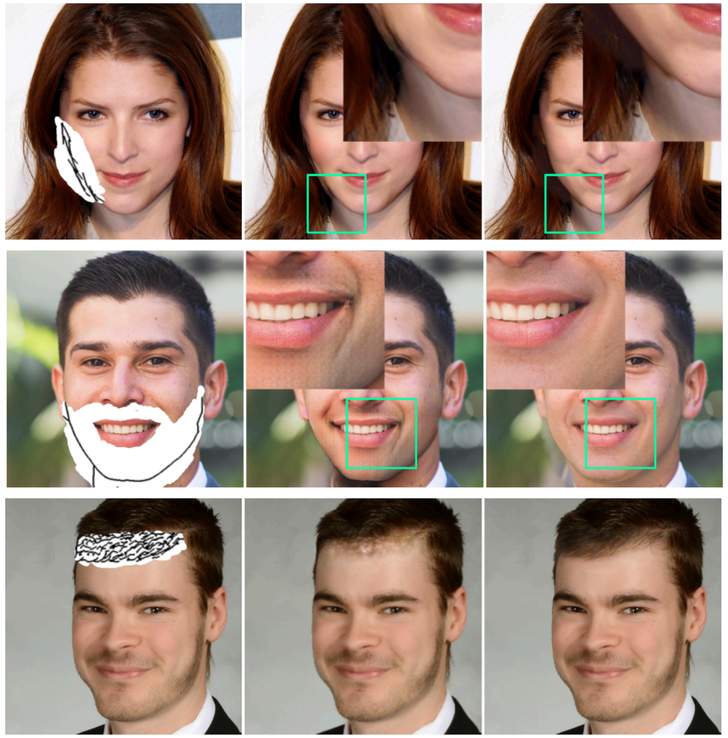
ICASSP2022

(a) Input

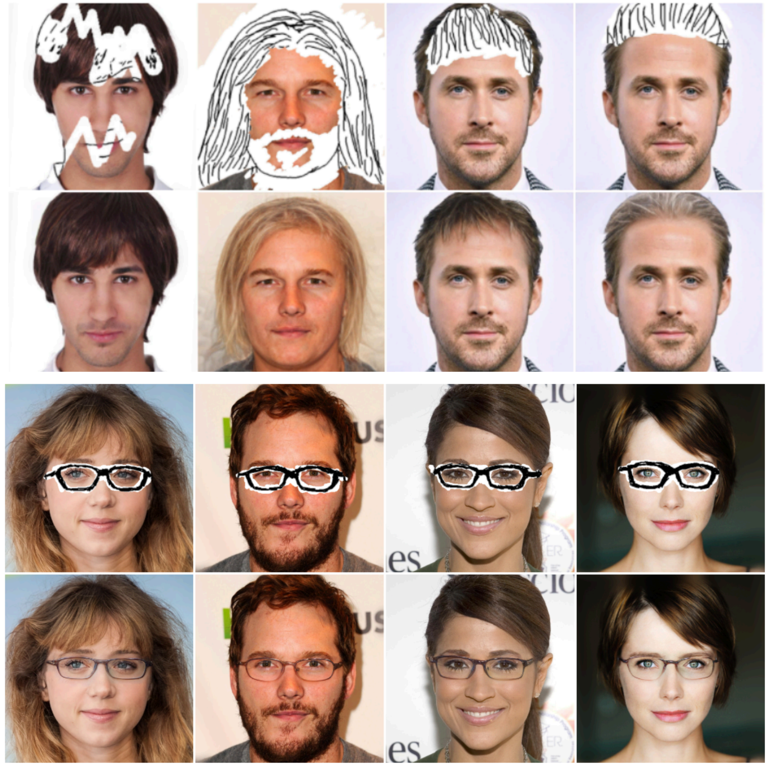
(b) Result by SC-FEGAN

(c) Result by DeepPS

(d) Result by our method



Input DeepPS Ours



Preliminaries: GAN inversion

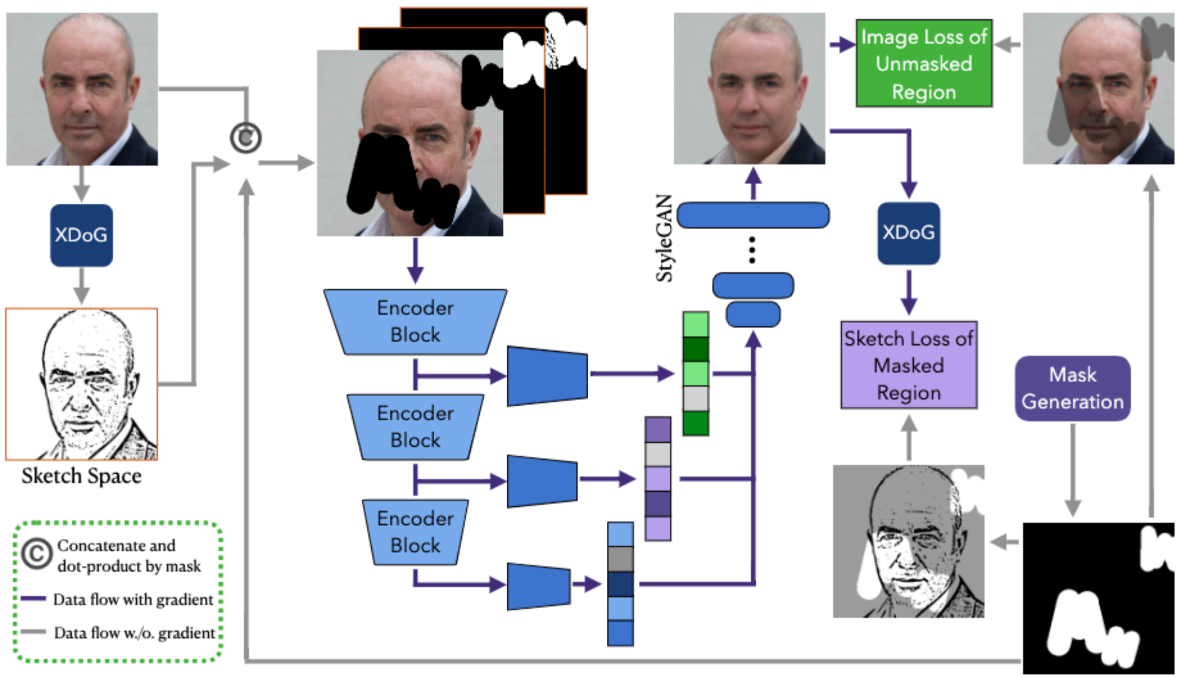
- We could optimize the latent code of a pre-trained GAN (StyleGAN) for a high-quality generation

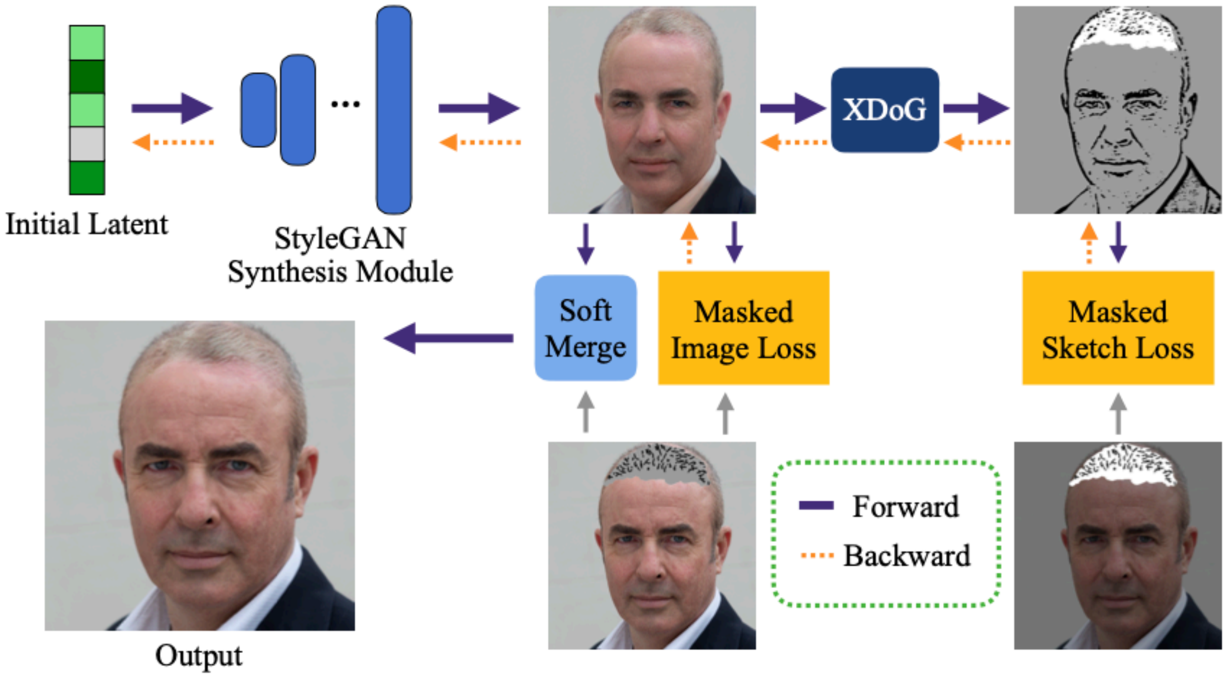
Up: origin image, bottom: generated image from StyleGAN with optimized latent codes

Style fusion with GAN inversion

Pictures are from Image2style (Rameen Abdal, et al. 2019)

Methods





$$\mathcal{L}_{perc}(\mathbf{I_1}, \mathbf{I_2}) = \| \sum_{j=1}^{5} \frac{\lambda_j}{N_j} (\mathbf{F}_j(\mathbf{I_1}) - \mathbf{F}_j(\mathbf{I_2})) \|_2^2$$

Perceptual loss (unmasked regions)

$$\mathcal{D}_{sketch}(\mathbf{S_1}, \mathbf{S_2}) = \sum_{j} \| (\mathbf{P}_j(\mathbf{S_1}) - \mathbf{P}_j(\mathbf{S_2})) \|_1$$

Multi-scale sketch loss (masked regions)

Wang, Chengrong, et al. "High-Fidelity Portrait Editing Via Exploring Differentiable Guided Sketches from the Latent Space." *ICASSP*2022.

Thanks!

Chenjie Cao,
School of Data Science, Fudan University
ccjdurandal163@gmail.com