Supplementary Material

Anonymous CVPR submission

Paper ID 2439

1. Codes

Code of our ITrans networks are zipped in codes.zip. This code is not the final version, and will be publicly released after the paper being published.

2. Network Architectures

Our whole ITrans inpainting network contains two parts: edge generation and image inpainting. We describe our inpainting generators in the following tables. Note that the tables only contain generator settings. The settings of edge generation network and discriminators are the same as [3]. Skip connections are widely used in our models. We consider that it is important to combine low-level encoder features. Skip layers are perfectly suited for this task.

3. Results

3.1. Qualitative Comparison

Figure 5 shows our inpainting results by our proposed models. Figure 6 compares images generated by our models to other inpainting approaches. With self-attention applied, our model generates more detailed images than simply using edges and corrupted images.

3.2. Quantitative Comparison

We use four numerical metrics to assess the quality of our models: 1) relative L1 (MAE); 2) structural similarity index (SSIM) [4]; 3) peak signal-to-noise ratio (PSNR); 4) Frechet Inception Distance (FID) [2]. Table 5. shows our testing results on Places test dataset.

References

[1] Zongyu Guo, Zhibo Chen, Tao Yu, Jiale Chen, and Sen Liu. Progressive image inpainting with full-resolution residual network. In *Proceedings of the 27th acm international conference on multimedia*, pages 2496–2504, 2019. 6, 7

[2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1

[3] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1, 6, 7

[4] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

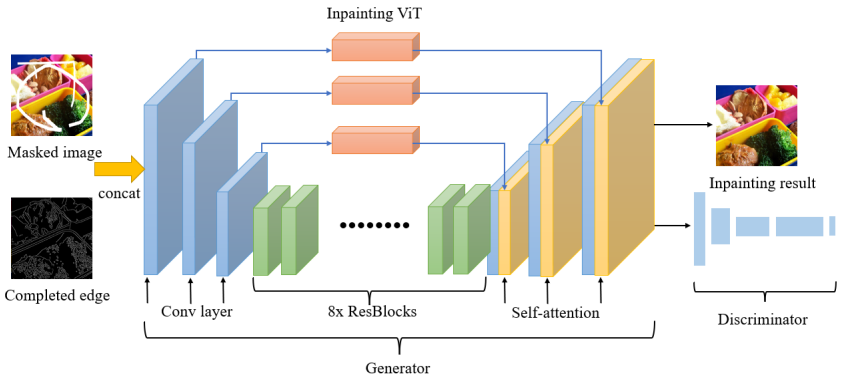


Figure 1. Inpainting ViT U-Net network.

Channels	Kernel size	Stride	Dilation	Padding	Activation
64	7	1	1	0	ReLU
128	4	2	1	1	ReLU
256	4	2	1	1	ReLU
8x ResBlocks					
Skip layers inpainting ViT					
128	4	2	1	1	ReLU
Skip layers inpainting ViT					
64	4	2	1	1	ReLU
Skip layers inpainting ViT					
3	7	1	1	0	Tanh

Table 1. Details of ViT U-Net in Figure 1.

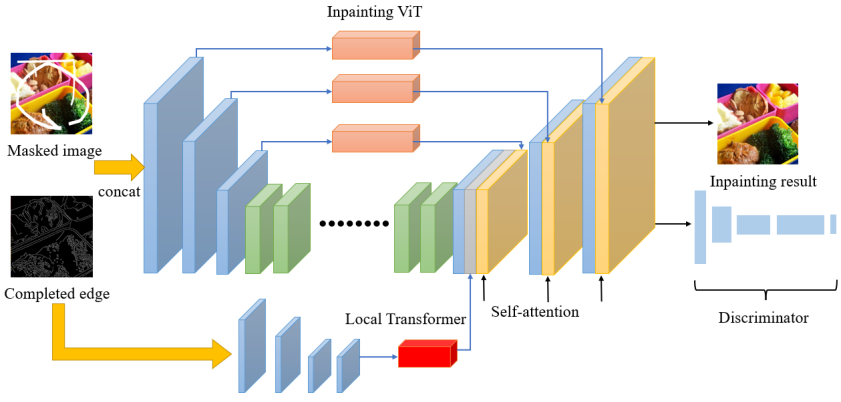
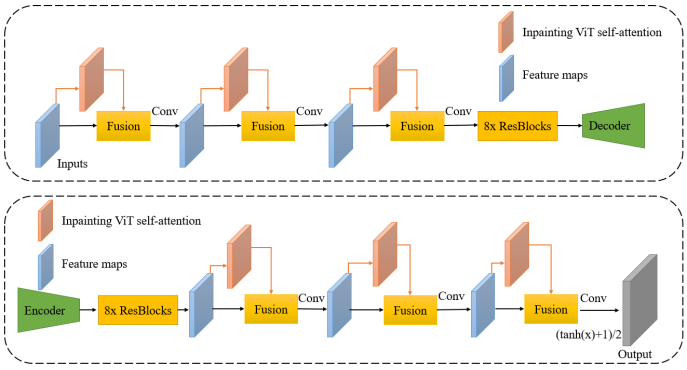


Figure 2. Local Transformer ViT U-Net.

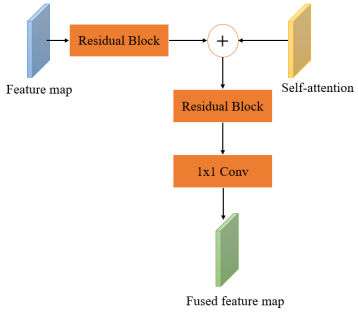
Branch1						Branch 2					
C	K	S	D	P	Ac.	C	K	S	D	P	Ac.
64	7	1	1	0	ReLU	32	4	2	1	1	ReLU
128	4	2	1	1	ReLU	64	4	2	1	1	ReLU
256	4	2	1	1	ReLU	128	3	1	1	1	ReLU
8x ResBlocks						256	3	1	1	1	ReLU
Skip layers Inpainting ViT						Local Transformer layer					
Concatenate											

Channels	Kernel size	Stride	Dilation	Padding	Activation
128	4	2	1	1	ReLU
Skip layers inpainting ViT					
64	4	2	1	1	ReLU
Skip layers inpainting ViT					
3	7	1	1	0	Tanh

Table 2. Details of Local Transfoemer ViT U-Net in Figure 2.



(a) Two applications of fusion block networks.



(b) Feature and self-attention fusion block.

Figure 3. In the fusion network, we demonstrate the fusion block in two ways: in the encoder (upper one) and in the decoder (lower one). In our fusion block, we add two residual blocks to combine feature maps and self-attention together for our inpainting model.

Channels	Kernel size	Stride	Dilation	Padding	Activation
64	7	1	1	0	ReLU
Inpainting ViT, Channels=64					
Fusion Network					
128	4	2	1	1	ReLU
Inpainting ViT, Channels=128					
Fusion Network					
256	4	2	1	1	ReLU
Inpainting ViT, Channels=256					
Fusion Network					
8x ResBlocks					
128	4	2	1	1	ReLU
64	4	2	1	1	ReLU
3	7	1	1	0	Tanh

Table 3. Details of Local Transfoemer ViT U-Net in Figure 3 encoder version.

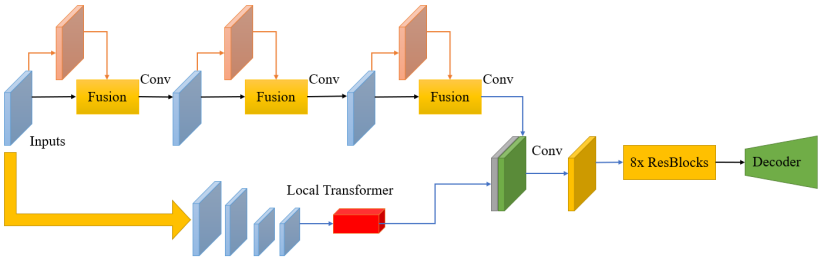
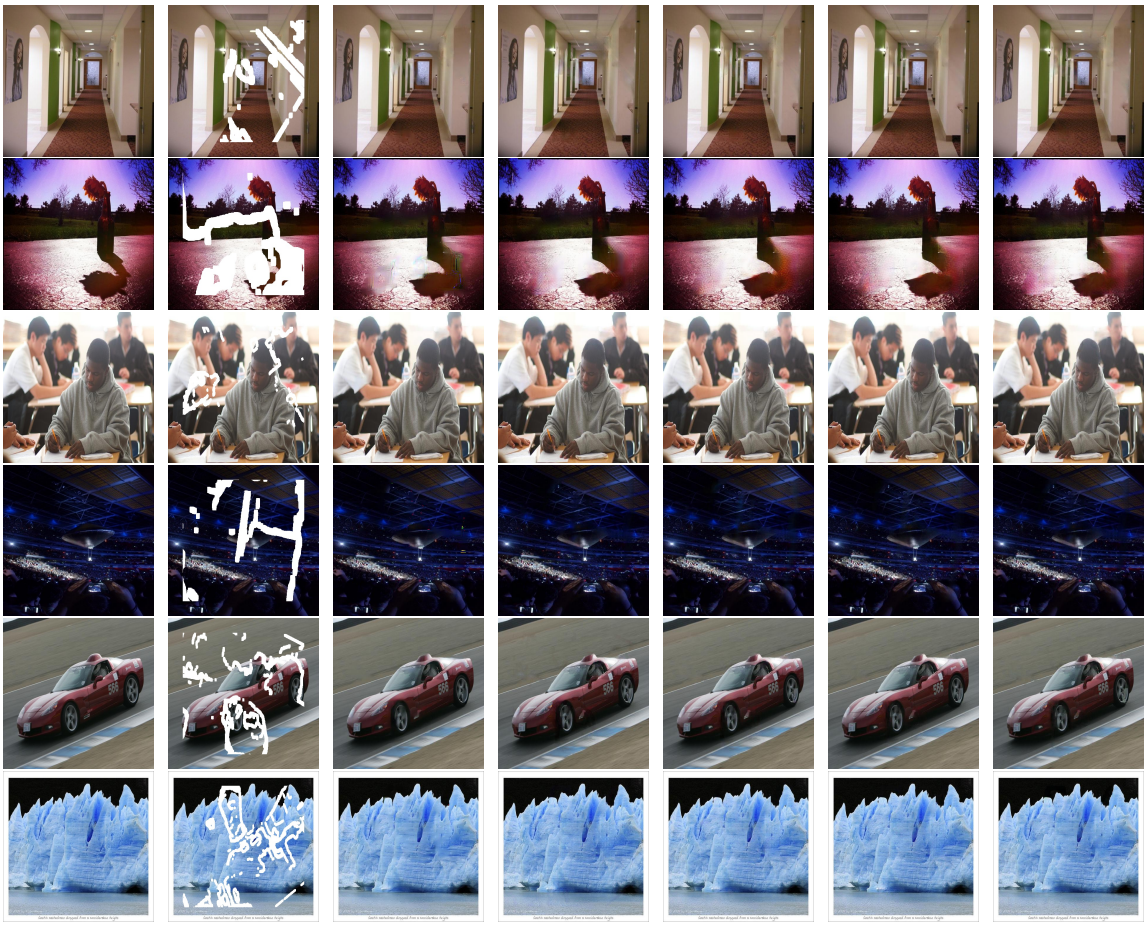


Figure 4. Hybrid network.

Fusion Network Branch						Local transformer branch					
C	K	S	D	P	Ac.	C	K	S	D	P	Ac.
64	7	1	1	0	ReLU	32	4	2	1	1	ReLU
Inpainting ViT, Channels=64						64	4	2	1	1	ReLU
Fusion Network						128	3	1	1	1	ReLU
128	4	2	1	1	ReLU	256	3	1	1	1	ReLU
Inpainting ViT, Channels=128						Local Transformer					
Fusion Network											
256	4	2	1	1	ReLU						
Inpainting ViT, Channels=256											
Fusion Network											
Concatenate											
Channels		Kernel size			Stride	Dilation		Padding			Activation
256		3			1	1		1			ReLU
8x ResBlocks											
128		4			2	1		1			ReLU
64		4			2	1		1			ReLU
3		7			1	1		0			Tanh

Table 4. Details of Hybrid in Figure 4.

Figure 5. Inpainting Examples. From left to right: Original images, input images, complete edges, inpainting results.



(a) GT (b) Input (c) FRRN (d) EC (e) ViT (f) LT ViT (g) Hybrid

Figure 6. Qualitative comparison with current models. (a) Ground truth. (b) Ground truth with mask. (c) FRRN [1]. (d) EdgeConnect [3]. (e) Ours ViT U-Net. (f) Ours local transformer ViT U-Net. (g) Ours hybrid network.

Method	Mask	FRRN	EC	ViT	LT ViT	Hybrid
PSNR	0-10%	31.33	31.92	31.99	**32.15**	31.96
	10-20%	27.73	27.60	27.66	**27.88**	27.68
	20-30%	24.53	24.70	24.73	**24.98**	24.79
	30-40%	22.16	22.50	22.53	**22.79**	22.61
	40-50%	20.06	20.69	20.74	**21.05**	20.83
	50-60%	15.58	17.87	18.03	**18.27**	17.95
	all	23.83	24.36	24.43	**24.68**	24.45
SSIM	0-10%	0.9677	0.9712	0.9718	**0.9726**	0.9714
	10-20%	0.9358	0.9342	0.9353	**0.9375**	0.9347
	20-30%	0.8795	0.8808	0.8824	**0.8868**	0.8819
	30-40%	0.8077	0.8132	0.8150	**0.8222**	0.8154
	40-50%	0.7199	0.7324	0.7345	**0.7456**	0.7364
	50-60%	0.5445	0.5789	0.5809	**0.5963**	0.5859
	all	0.8113	0.8203	0.8216	**0.8288**	0.8226
MAE(%)	0-10%	1.78	1.56	1.55	**1.54**	1.56
	10-20%	2.22	2.28	2.26	**2.22**	2.26
	20-30%	3.23	3.26	3.23	**3.14**	3.22
	30-40%	4.51	4.46	4.40	**4.25**	4.37
	40-50%	6.19	5.91	5.83	**5.61**	5.77
	50-60%	11.76	9.10	8.97	**8.63**	9.05
	all	4.86	4.36	4.32	**4.17**	4.33
FID	0-10%	4.49	4.08	**4.05**	4.06	4.15
	10-20%	11.52	10.69	10.90	**10.31**	11.21
	20-30%	22.94	20.43	20.72	**19.14**	21.49
	30-40%	38.04	34.77	34.80	**31.84**	35.94
	40-50%	58.73	53.94	53.00	**49.02**	54.65
	50-60%	99.18	90.01	91.56	**89.06**	87.17
	all	17.10	16.24	16.49	**15.10**	16.42

Table 5. Quantitative results on Places2 dataset. We compare our networks (ViT, LT VIT, and Hybrid) with FRRN [1] and EdgeConnect [3]. For evaluating metrics, PSNR and SSIM are higher the better, while MAE and FID are lower the better. The best metric is boldfaced.