

000
001
002
003
004
005
006
007
008
009
010
011

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.

012
013

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.

027
028
029
030

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.

036
037

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.

043
044
045

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

Supplementary Material

Anonymous CVPR submission

Paper ID 2439

- [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

054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107

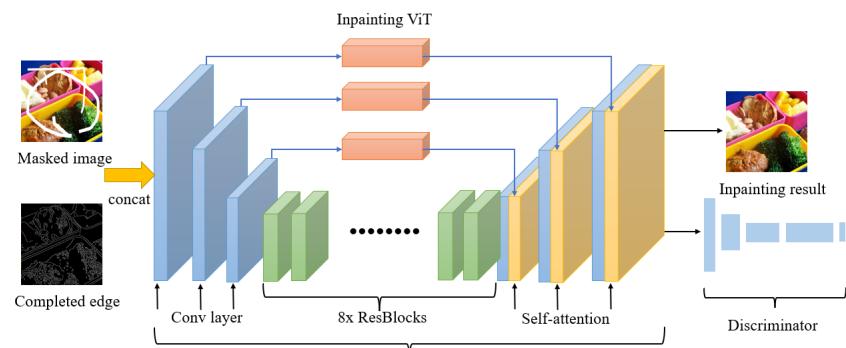


Figure 1. Inpainting ViT U-Net network.

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
<hr/>					
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.

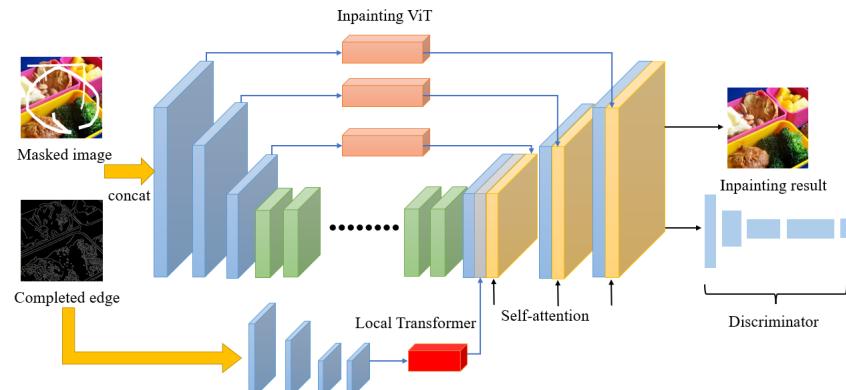


Figure 2. Local Transformer ViT U-Net.

216	Branch1						Branch 2						270
217	C	K	S	D	P	Ac.	C	K	S	D	P	Ac.	271
218	64	7	1	1	0	ReLU	32	4	2	1	1	ReLU	272
219	128	4	2	1	1	ReLU	64	4	2	1	1	ReLU	273
220	256	4	2	1	1	ReLU	128	3	1	1	1	ReLU	274
221	8x ResBlocks						256	3	1	1	1	ReLU	275
222	Skip layers Inpainting ViT						Local Transformer layer						276
223	Concatenate												277
224	Channels	Kernel size	Stride	Dilation	Padding	Activation	278	279	280	281	282	283	284
225	128	4	2	1	1	ReLU	285	286	287	288	289	290	291
226	Skip layers inpainting ViT						292	293	294	295	296	297	298
227	64	4	2	1	1	ReLU	299	300	301	302	303	304	305
228	Skip layers inpainting ViT						306	307	308	309	310	311	312
229	3	7	1	1	0	Tanh	313	314	315	316	317	318	319
230							320	321	322	323	324	325	326
231							327	328	329	330	331	332	333

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

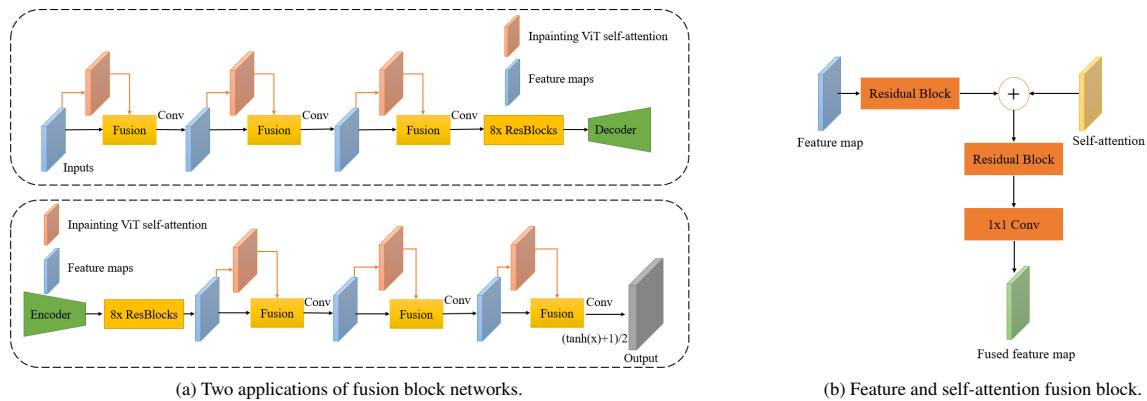


Figure 3. In the fusion network, we demonstrate the fusion block in two ways: in the encoder (upper one) and in the decoder (lower one). In our fusion block, we add two residual blocks to combine feature maps and self-attention together for our inpainting model.

253	Channels	Kernel size	Stride	Dilation	Padding	Activation	307
254	64	7	1	1	0	ReLU	308
Inpainting ViT, Channels=64							309
Fusion Network							310
257	128	4	2	1	1	ReLU	311
Inpainting ViT, Channels=128							312
Fusion Network							313
260	256	4	2	1	1	ReLU	314
Inpainting ViT, Channels=256							315
Fusion Network							316
8x ResBlocks							317
264	128	4	2	1	1	ReLU	318
265	64	4	2	1	1	ReLU	319
266	3	7	1	1	0	Tanh	320
267							321
268							322
269							323

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

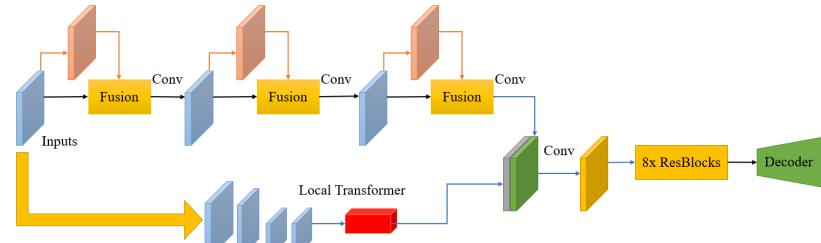


Figure 4. Hybrid network.

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
<u>Inpainting ViT, Channels=64</u>											
<u>Fusion Network</u>											
128	4	2	1	1	ReLU	64	4	2	1	1	ReLU
<u>Inpainting ViT, Channels=128</u>											
<u>Fusion Network</u>											
256	4	2	1	1	ReLU	128	3	1	1	1	ReLU
<u>Inpainting ViT, Channels=256</u>											
<u>Fusion Network</u>											
<u>Concatenate</u>											
Channels	Kernel size		Stride	Dilation	Padding	Activation					
256	3		1	1	1	ReLU					
<u>8x ResBlocks</u>											
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.

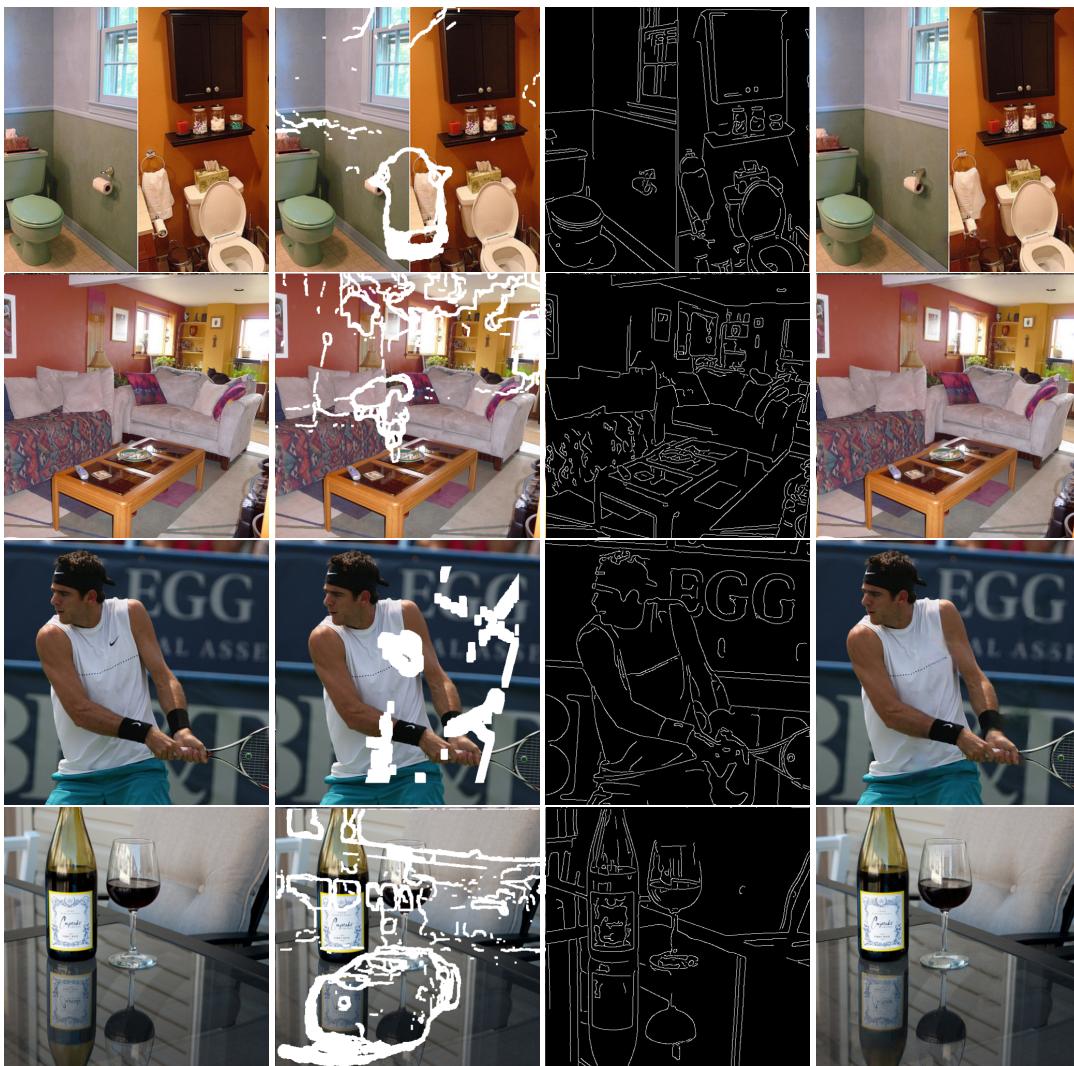


Figure 5. Inpainting Examples. From left to right: Original images, input images, complete edges, inpainting results.

432
433
434
435
436
437
438
439
440441
442
443
444
445
446
447
448
449450
451
452
453
454
455
456
457458
459
460
461
462
463
464
465466
467
468
469
470
471
472
473474
475
476
477
478
479
480
481
482
483
484
485486
487
488
489
490
491
492
493
494495
496
497
498
499
500
501
502
503504
505
506
507
508
509
510
511512
513
514
515
516
517
518
519520
521
522
523
524
525
526
527528
529
530
531
532
533
534
535
536
537
538
539

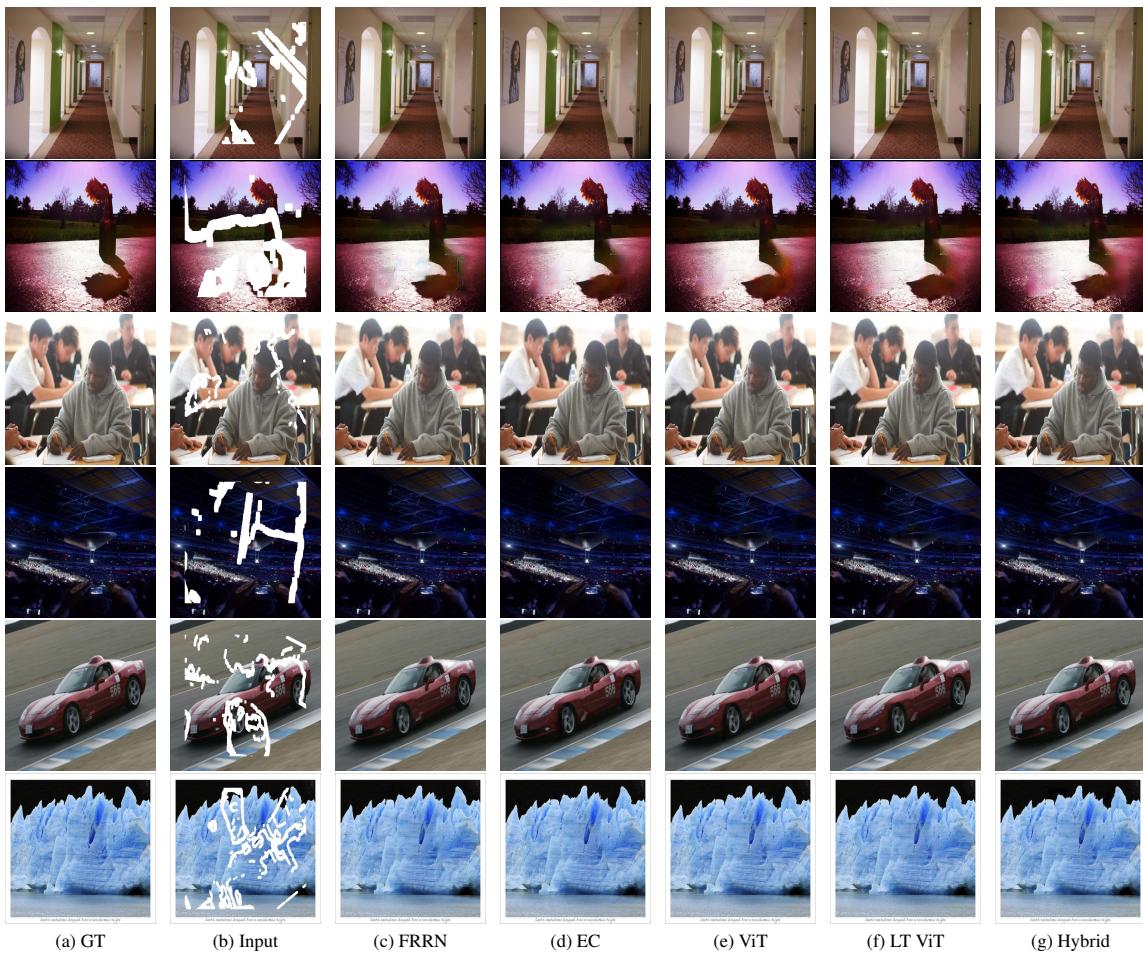


Figure 6. Qualitative comparison with current models. (a) Ground truth. (b) Ground truth with mask. (c) FRRN [1]. (d) EdgeConnect [3]. (e) Ours ViT U-Net. (f) Ours local transformer ViT U-Net. (g) Ours hybrid network.

583
584
585
586
587
588
589
590
591
592
593

635
636
637
638
639
640
641
642
643
644
645
646
647

	Method	Mask	FRRN	EC	ViT	LT ViT	Hybrid	
659	PSNR	0-10%	31.33	31.92	31.99	32.15	31.96	702
		10-20%	27.73	27.60	27.66	27.88	27.68	703
		20-30%	24.53	24.70	24.73	24.98	24.79	704
		30-40%	22.16	22.50	22.53	22.79	22.61	705
		40-50%	20.06	20.69	20.74	21.05	20.83	706
		50-60%	15.58	17.87	18.03	18.27	17.95	707
		all	23.83	24.36	24.43	24.68	24.45	708
660	SSIM	0-10%	0.9677	0.9712	0.9718	0.9726	0.9714	709
		10-20%	0.9358	0.9342	0.9353	0.9375	0.9347	710
		20-30%	0.8795	0.8808	0.8824	0.8868	0.8819	711
		30-40%	0.8077	0.8132	0.8150	0.8222	0.8154	712
		40-50%	0.7199	0.7324	0.7345	0.7456	0.7364	713
		50-60%	0.5445	0.5789	0.5809	0.5963	0.5859	714
		all	0.8113	0.8203	0.8216	0.8288	0.8226	715
661	MAE(%)	0-10%	1.78	1.56	1.55	1.54	1.56	716
		10-20%	2.22	2.28	2.26	2.22	2.26	717
		20-30%	3.23	3.26	3.23	3.14	3.22	718
		30-40%	4.51	4.46	4.40	4.25	4.37	719
		40-50%	6.19	5.91	5.83	5.61	5.77	720
		50-60%	11.76	9.10	8.97	8.63	9.05	721
		all	4.86	4.36	4.32	4.17	4.33	722
662	FID	0-10%	4.49	4.08	4.05	4.06	4.15	723
		10-20%	11.52	10.69	10.90	10.31	11.21	724
		20-30%	22.94	20.43	20.72	19.14	21.49	725
		30-40%	38.04	34.77	34.80	31.84	35.94	726
		40-50%	58.73	53.94	53.00	49.02	54.65	727
		50-60%	99.18	90.01	91.56	89.06	87.17	728
		all	17.10	16.24	16.49	15.10	16.42	729

689 Table 5. Quantitative results on Places2 dataset. We compare our networks (ViT, LT VIT, and Hybrid) with FRRN [1] and EdgeConnect [3].
690 For evaluating metrics, PSNR and SSIM are higher the better, while MAE and FID are lower the better. The best metric is boldfaced.
691

692

693

694

695

696

697

698

699

700

701

743

744

745

746

747

748

749

750

751

752

753

754

755