Image Inpainting via Conditional Texture and Structure Dual Generation

(ICCV 2021)

Xiefan Guo, Hongyu Yang, Di Huang (2021)

Image Inpainting

Image Inpainting

Image inpainting is the process of genarating or reconstructing distorted regions of an Image.

Challenges : (Previous methods)

- The cases with large corruptions, generally suffer from distorted results.
- Previous methods have a common drawback in recovering the global structure of the image.
- Previous methods involving structures are sensitive to the accuracy of structures (e.g. edges and contours) which is not easy to guarantee.

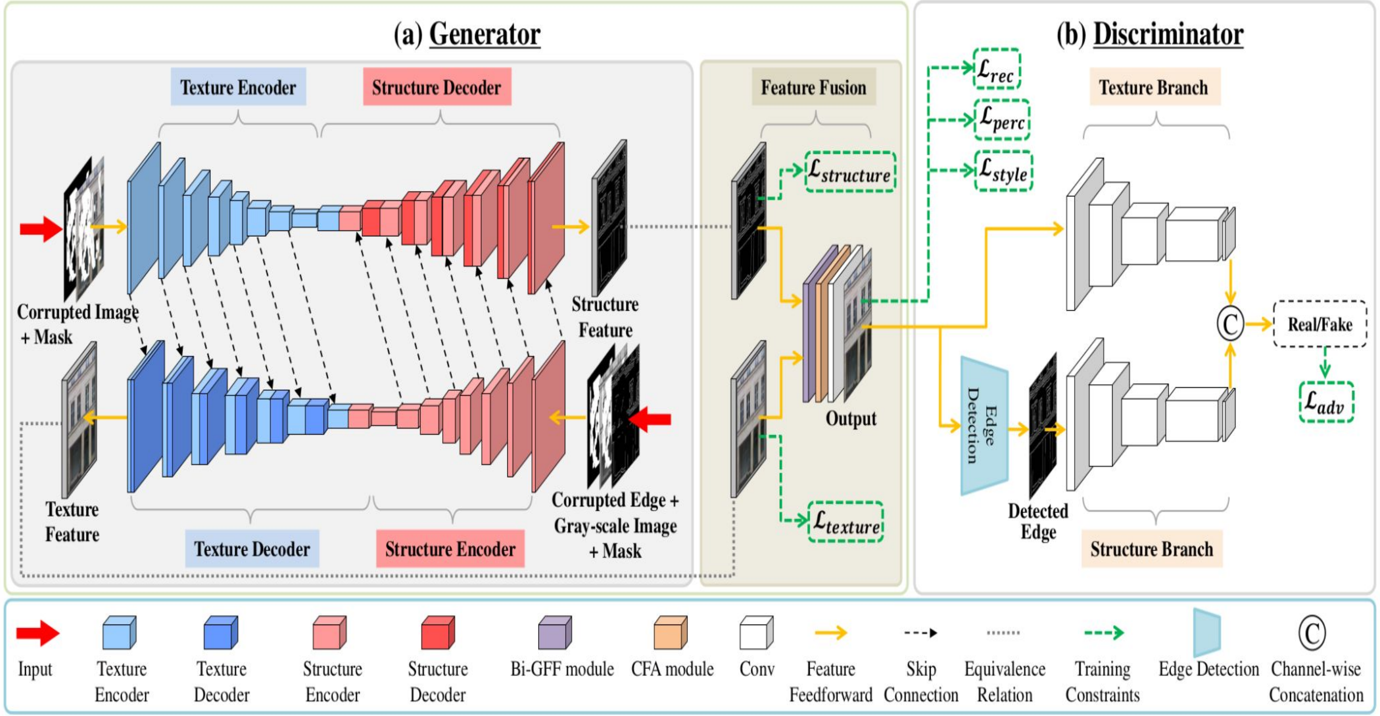
Architecture

Structure-constrained texture synthesis

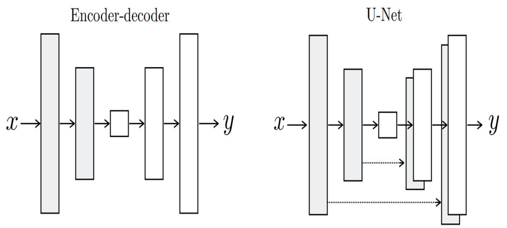
Texture-guided structure reconstruction

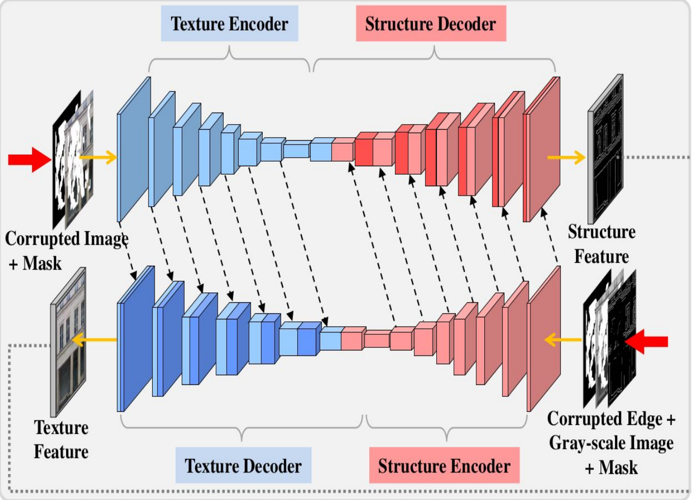
Bi-directional Gated Feature Fusion (Bi-GFF)

Contextual Feature Aggregation (CFA)



Generator :





Generator

Two-stream generator which jointly synthesizes image textures and structures.(Partial Convolutions)

U-Net Variant

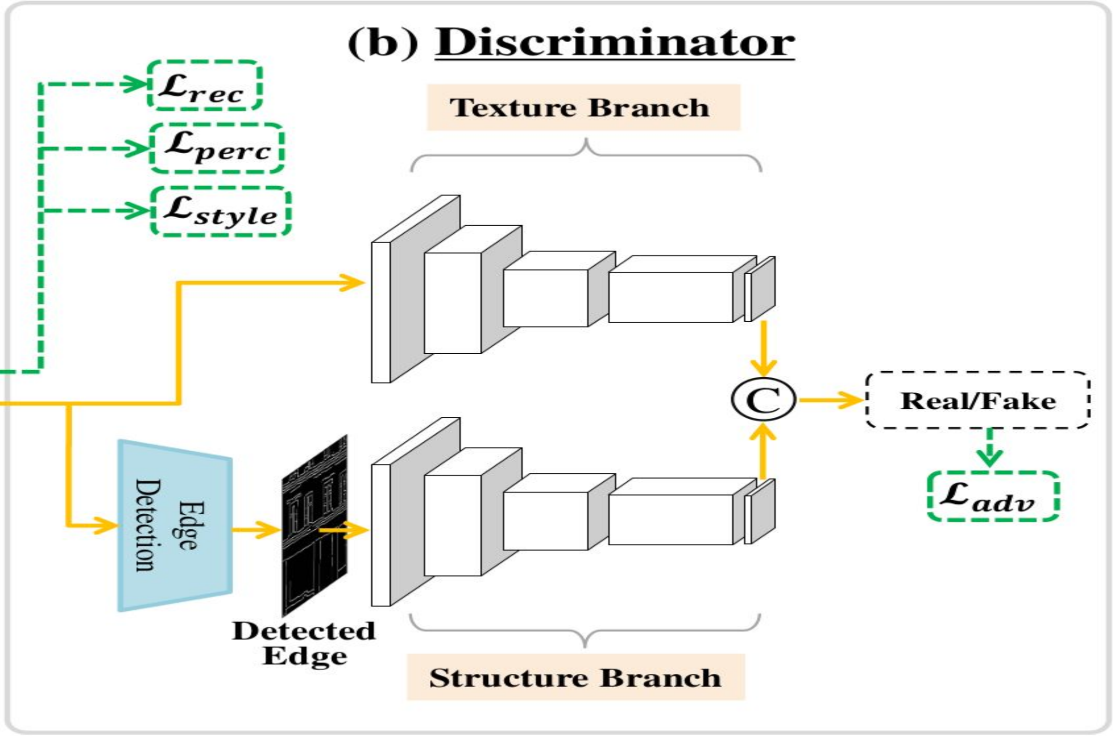
- Encoder-decoder
- Skip Connections

Input : The corrupted image and its corresponding edge map.
Output : Texture and structure Feature map.

Discriminator:

Discriminator

Two-stream discriminator to distinguish images by feature statistics of both texture and structure.

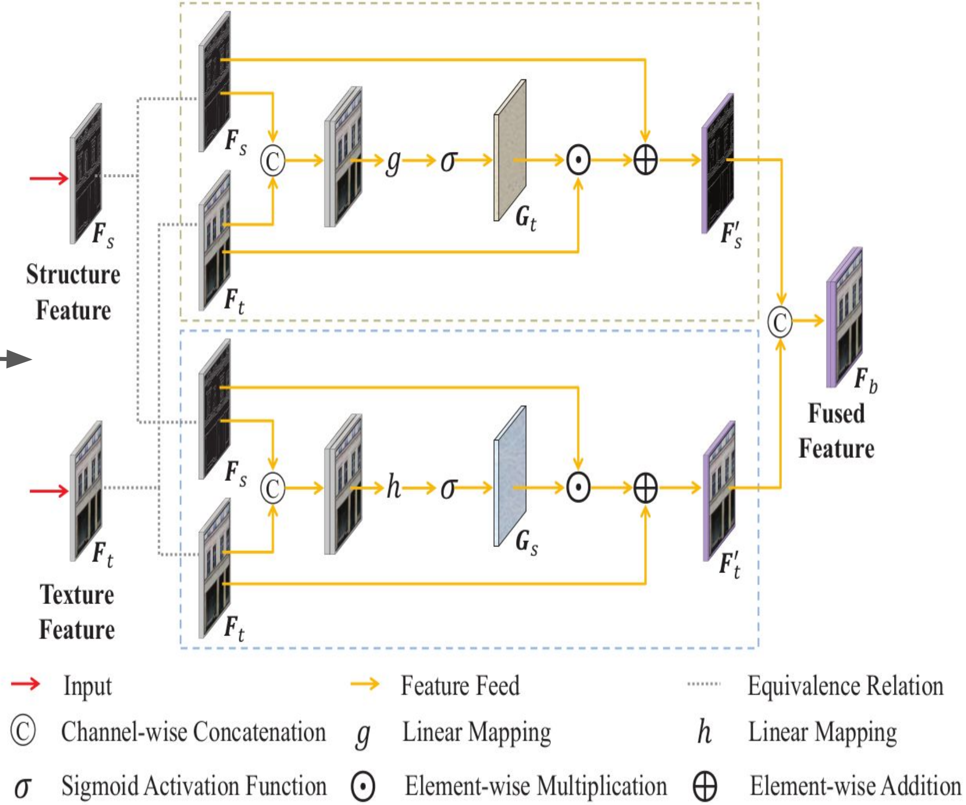


Input : Texture and structure Feature map.

Output : Real/Fake.

Bi-GFF :

Bi-directional Gated Feature Fusion (Bi-GFF)



$$F'_t = \beta(G_s \odot F_s) \oplus F_t,$$

$$F'_s = \alpha(G_t \odot F_t) \oplus F_s,$$

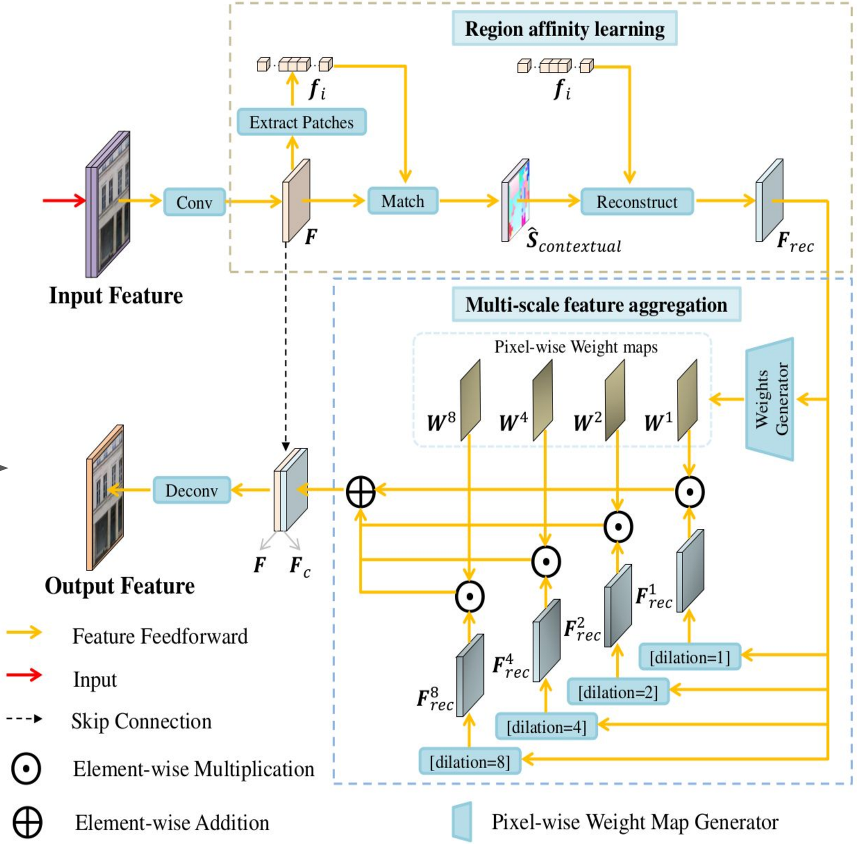
$$F_b = \text{Concat}(F'_s, F'_t).$$

CFA :

$$S_{contextual}^{i,j} = \left\langle \frac{\boldsymbol{f}_i}{\|\boldsymbol{f}_i\|_2}, \frac{\boldsymbol{f}_j}{\|\boldsymbol{f}_j\|_2} \right\rangle, \quad \hat{S}_{contextual}^{i,j} = \frac{\exp\left(S_{contextual}^{i,j}\right)}{\sum_{j=1}^{N} \exp\left(S_{contextual}^{i,j}\right)}.$$

$$\tilde{\boldsymbol{f}}_i = \sum_{j=1}^{N} \boldsymbol{f}_j \cdot \hat{S}_{contextual}^{i,j}, \qquad \boldsymbol{F}_{rec}^{k} = \mathrm{Conv}_k\left(\boldsymbol{F}_{rec}\right),$$

Contextual Feature Aggregation (CFA)



$$\boldsymbol{W} = \mathrm{Softmax}\left(G_w\left(\boldsymbol{F}_{rec}\right)\right),$$
$$\boldsymbol{W}^1, \boldsymbol{W}^2, \boldsymbol{W}^4, \boldsymbol{W}^8 = \mathrm{Slice}(\boldsymbol{W}),$$

$$\boldsymbol{F}_c = \left(\boldsymbol{F}_{rec}^1 \odot \boldsymbol{W}^1\right) \oplus \left(\boldsymbol{F}_{rec}^2 \odot \boldsymbol{W}^2\right) \oplus \left(\boldsymbol{F}_{rec}^4 \odot \boldsymbol{W}^4\right) \oplus \left(\boldsymbol{F}_{rec}^8 \odot \boldsymbol{W}^8\right).$$

Loss Functions

- **Reconstruction Loss :** $\mathcal{L}_{rec} = \mathbb{E}\left[\|\boldsymbol{I}_{out} - \boldsymbol{I}_{gt}\|_1\right].$

- **Perceptual Loss :** $\mathcal{L}_{perc} = \mathbb{E}\left[\sum_i \|\phi_i(\boldsymbol{I}_{out}) - \phi_i(\boldsymbol{I}_{gt})\|_1\right],$

- **Style Loss :** $\mathcal{L}_{style} = \mathbb{E}\left[\sum_i \|(\psi_i(\boldsymbol{I}_{out}) - \psi_i(\boldsymbol{I}_{gt}))\|_1\right],$

- **Adversarial Loss :** $$\mathcal{L}_{adv} = \min_G \max_D \mathbb{E}_{\boldsymbol{I}_{gt}, \boldsymbol{E}_{gt}}\left[\log D(\boldsymbol{I}_{gt}, \boldsymbol{E}_{gt})\right] + \mathbb{E}_{\boldsymbol{I}_{out}, \boldsymbol{E}_{out}} \log\left[1 - D(\boldsymbol{I}_{out}, \boldsymbol{E}_{out})\right].$$

- **Intermediate Loss :** $$\mathcal{L}_{inter} = \mathcal{L}_{structure} + \mathcal{L}_{texture} = \mathrm{BCE}(\boldsymbol{E}_{gt}, \mathcal{P}_s(\boldsymbol{F}_s)) + \ell_1(\boldsymbol{I}_{gt}, \mathcal{P}_t(\boldsymbol{F}_t)),$$

- **Joint Loss :** $$\mathcal{L}_{joint} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{style}\mathcal{L}_{style} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{inter}\mathcal{L}_{inter},$$

G : Generator
D : Discriminator
I_{gt} : Ground-truth image
E_{gt} : Complete Edge Map
Y_{gt} : Gray-scale image
M_{in} : Initial binary mask
I_{in} : $I_{gt} \odot M_{in}$ (damaged image)
E_{in} : $E_{gt} \odot M_{in}$ (damaged edge map)
Y_{in} : Y_{gt} M_{in} (damaged gray image)
I_{out}, E_{out} = $G(I_{in}, E_{in}, Y_{in}, M_{in})$
$\bigcirc$: element-wise multiplication

Training : Experimental Settings

- CelebA, ParisStreetView and Places2 with their original training, testing and validation split.
- All the images and corresponding masks are resized to 256 X 256 pixels.
- The model is implemented in PyTorch. Training is launched on a single NVIDIA 1080TI GPU (11GB) with batch size of 6.
- we first use a learning rate of 2×10^{-4} initial training, then finetune the model with a learning rate of 5×10^{-5}.
- The discriminator is trained with a learning rate of 1/10 of the generator.
- It takes around 4 days to train the models on CelebA and Paris StreetView and 10 days on Places2.

Results

- Qualitative and quantitative experiments on the CelebA, ParisStreetView and Places2.

Metrics	LPIPS†			PSNR¶			SSIM¶			User Study¶
Mask Ratio	0-20%	20-40%	40-60%	0-20%	20-40%	40-60%	0-20%	20-40%	40-60%	0-60%
PatchMatch [2]	0.074	0.183	0.332	30.02	24.77	20.51	0.864	0.680	0.487	2.7%
PConv [13]	0.065	0.134	0.283	30.19	25.18	21.20	0.885	0.730	0.527	4.0%
DeepFillv2 [36]	0.056	0.123	0.266	30.32	25.34	21.48	0.889	0.735	0.531	14.0%
RFR [11]	0.048	0.101	0.239	30.74	25.80	21.99	0.899	0.750	0.553	23.3%
EdgeConnect [18]	0.061	0.131	0.268	30.28	25.30	21.39	0.886	0.737	0.535	4.7%
PRVS [10]	0.057	0.124	0.257	30.30	25.39	21.50	0.893	0.742	0.541	5.3%
MED [14]	0.053	0.120	0.248	30.41	25.45	21.63	0.895	0.745	0.547	6.0%
Ours	**0.042**	**0.095**	**0.227**	**30.81**	**25.97**	**22.23**	**0.904**	**0.759**	**0.561**	**40.0%**

Table 1: Objective quantitative comparison and user study on Places2 (†Lower is better; ¶Higher is better).

Results

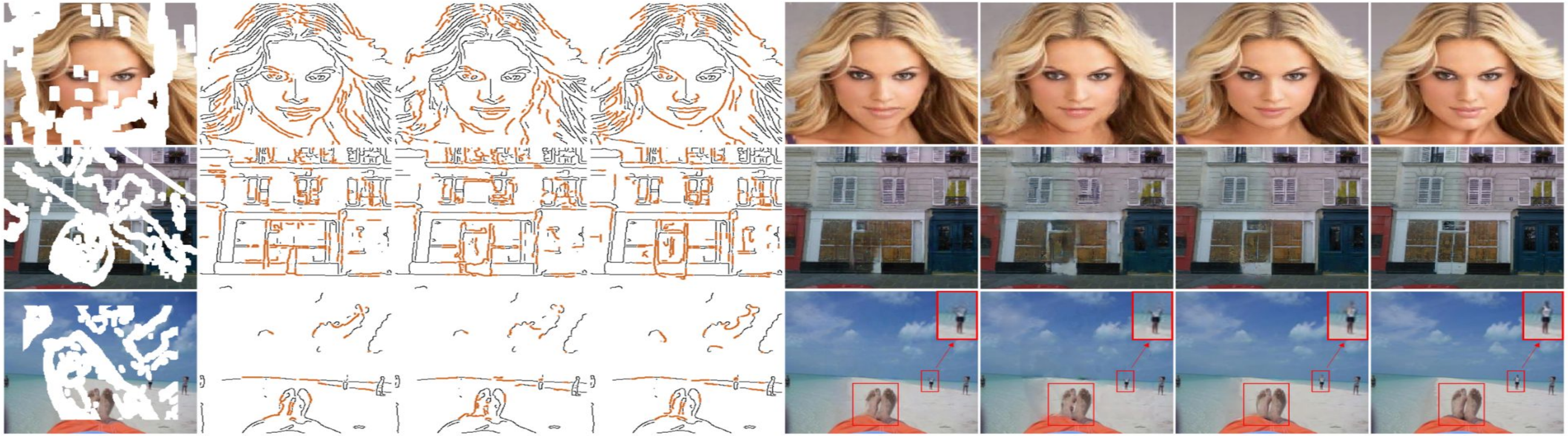
Metrics	LPIPS†			PSNR¶			SSIM¶		
Mask Ratio	0-20%	20-40%	40-60%	0-20%	20-40%	40-60%	0-20%	20-40%	40-60%
PatchMatch [1]	0.059	0.202	0.371	29.81	23.49	18.77	0.878	0.704	0.516
PConv [4]	0.046	0.122	0.221	31.89	26.48	21.32	0.899	0.750	0.558
DeepFillv2 [7]	0.040	0.107	0.214	32.48	26.93	21.70	0.906	0.757	0.569
RFR [3]	0.031	0.090	0.185	33.50	27.63	22.69	0.916	0.780	0.603
EdgeConnect [6]	0.042	0.117	0.215	32.12	26.79	21.66	0.904	0.758	0.566
PRVS [2]	0.039	0.112	0.209	32.34	26.89	21.78	0.908	0.762	0.573
MED [5]	0.037	0.106	0.203	32.68	27.01	21.86	0.907	0.763	0.575
Ours	**0.028**	**0.081**	**0.179**	**33.91**	**27.73**	**22.70**	**0.920**	**0.788**	**0.609**

Table 2: Objective quantitative comparison on CelebA (†Lower is better; ¶Higher is better).

Metrics	LPIPS†			PSNR¶			SSIM¶		
Mask Ratio	0-20%	20-40%	40-60%	0-20%	20-40%	40-60%	0-20%	20-40%	40-60%
PatchMatch [1]	0.078	0.195	0.362	30.70	25.31	20.59	0.881	0.689	0.499
PConv [4]	0.058	0.133	0.273	32.05	26.66	22.17	0.898	0.741	0.538
DeepFillv2 [7]	0.050	0.128	0.269	32.31	26.92	22.48	0.905	0.752	0.551
RFR [3]	0.041	0.112	0.234	32.69	27.33	22.76	0.919	0.772	0.568
EdgeConnect [6]	0.053	0.129	0.262	31.98	26.70	22.39	0.903	0.757	0.554
PRVS [2]	0.051	0.125	0.254	32.23	26.89	22.50	0.910	0.762	0.563
MED [5]	0.050	0.122	0.248	32.36	26.97	22.44	0.915	0.760	0.559
Ours	**0.039**	**0.107**	**0.226**	**32.93**	**27.48**	**22.89**	**0.923**	**0.777**	**0.573**

Table 3: Objective quantitative comparison on Paris StreetView (†Lower is better; ¶Higher is better).

Comparisons



(a) Input (b) EdgeConnect (c) PRVS (d) Ours (e) EdgeConnect (f) PRVS (g) Ours (h) Ground-truth

(a) Input (b) Ours (c) G_t (d) G_s (e) attention map (f) W^1 (g) W^2 (h) W^4 (i) W^8

Figure 1: Visualization of the feature maps learned by the network.

Comparisons

(a) Input (b) w/o structure (c) single stream (d) w/o Bi-GFF (e) w/o CFA (f) w/ CA (g) Ours (h) Ground-truth

Figure 7: Visualization of the effects of network architecture and individual modules on Paris StreetView.

Metrics	LPIPS†			PSNR¶			SSIM¶		
Mask Ratio	0-20%	20-40%	40-60%	0-20%	20-40%	40-60%	0-20%	20-40%	40-60%
w/o structure priors	0.054	0.129	0.251	31.72	26.71	22.22	0.909	0.755	0.550
single-stream	0.051	0.122	0.245	32.27	27.03	22.59	0.913	0.764	0.558
w/o Bi-GFF	0.045	0.114	0.236	32.61	27.20	22.75	0.919	0.772	0.567
w/o CFA	0.049	0.119	0.243	32.34	27.09	22.64	0.914	0.766	0.561
w/ CA	0.043	0.115	0.240	32.54	27.15	22.69	0.920	0.769	0.566
Ours	**0.039**	**0.107**	**0.226**	**32.93**	**27.48**	**22.89**	**0.923**	**0.777**	**0.573**

Table 2: Quantitative ablation study on Paris StreetView.

Thank You