Variants of outpainting

2022 하계 세미나

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented By
Ji-Hyun Kim

목차

- Outpainting
- Contextual Outapinting
  - Background: contextual outpainting, contrastive learning
  - “Contextual Outpainting With Object-Level Contrastive Learning." (CVPR 2022)
- Wide-range Blending
  - "Bridging the visual gap: Wide-range image blending. (CVPR 2021)

Outpainting

- Image completion task
  - Outpainting

  - Inpainting

Contextual outpainting

- A variant of outpainting
- Conventional outpainting vs Contextual outpainting
  - Conventional: completing object shapes, extending existing scenery textures
  - Contextual: hallucinating missing background contents based on foreground contents

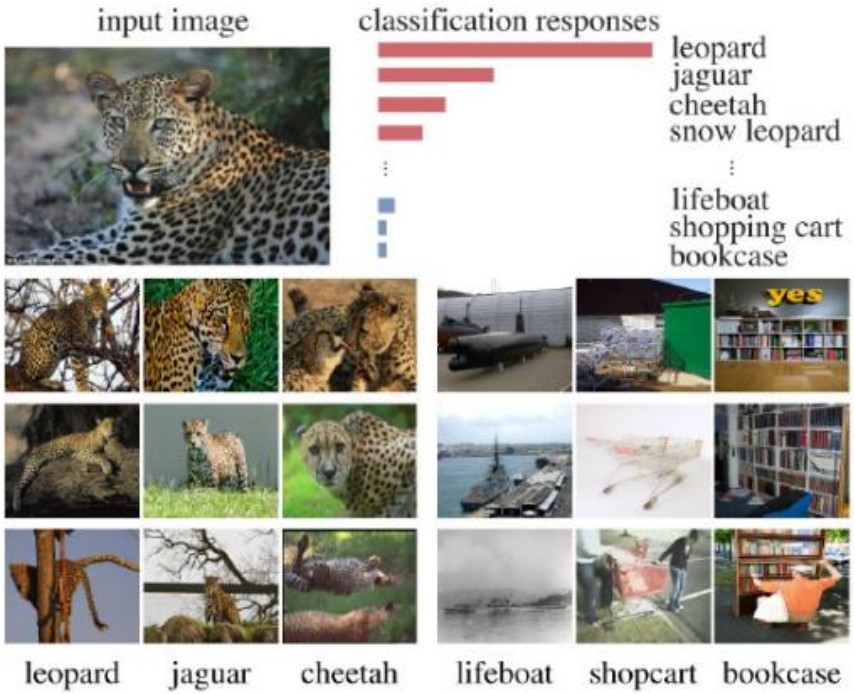
1) Li, Jiacheng, Chang Chen, and Zhiwei Xiong. "Contextual Outpainting With Object-Level Contrastive Learning." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Contextual outpainting

- A more challenging task compared to image completion
  - Lack of information: foreground and background contents share almost nothing

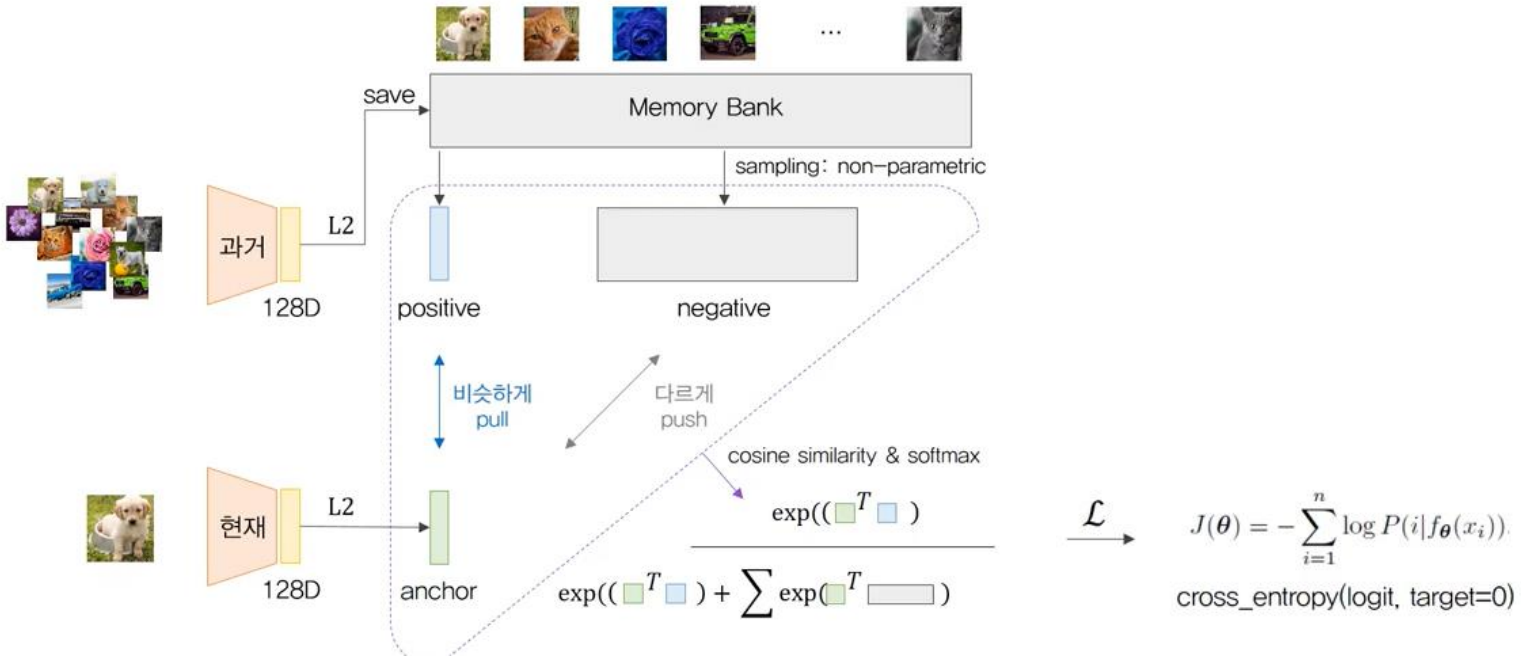
Contrastive learning

- Motivation
  - 지도학습 기반의 이미지 분류 모델 결과, 비슷한 이미지일 때 확률값이 높은 것을 보고 동기를 얻음
  - 모델이 따로 instances 간의 구분(유사성)을 학습하지 않아도, 잘 추출된 feature값은 instances 간의 유사도 정보를 가지고 있을 것



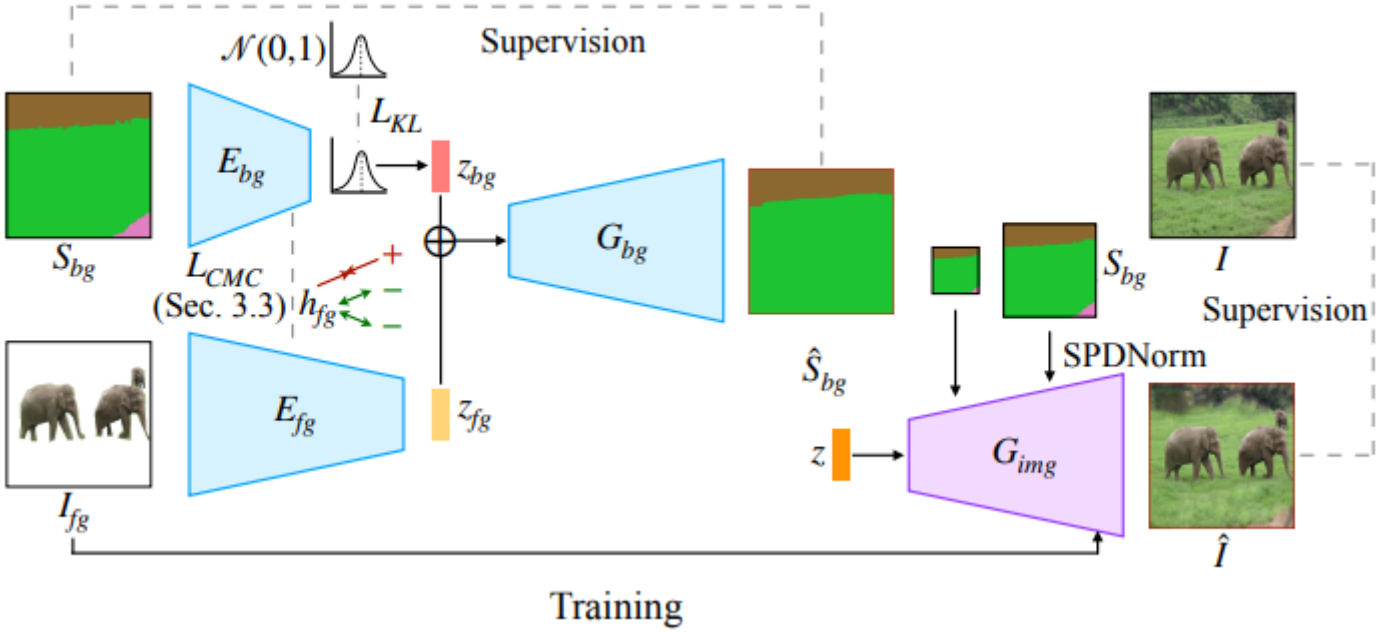
Contrastive learning

- Self-supervised learning
- Learn positive & negative relationships
- Input image to 128 d feature vector



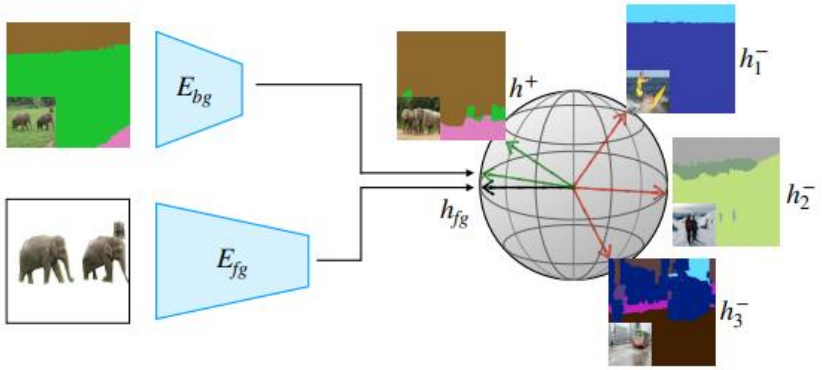
ConTextual Outpainting GAN (CTO-GAN)

- Semantic layout: bridge to synthesize coherent and diverse background contents
- Two stage learning
  - Semantic reasoning: infers semantic layout from the foreground contents
  - Content generation: synthesizes corresponding background contents



Cross-modal contrastive (CMC) loss

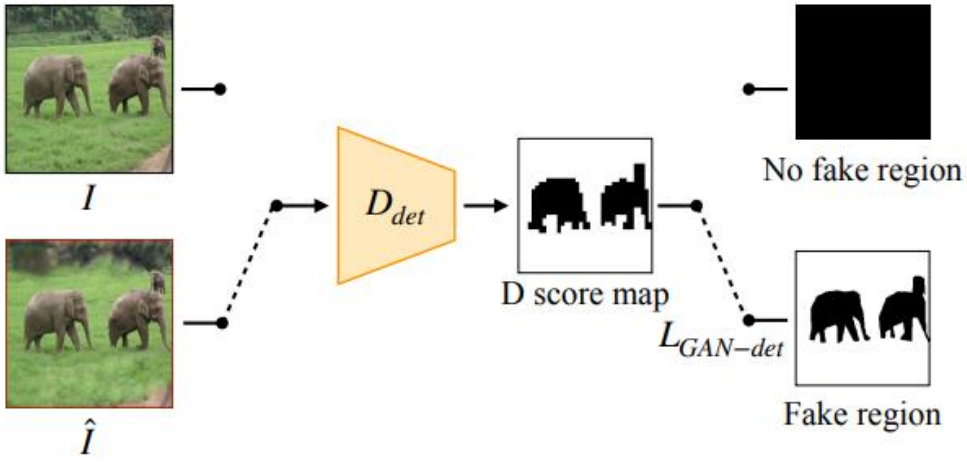
- Foreground & background encoder: encode the features of foreground pixels and background semantic layouts into the same cross-modal embedding space
- Enables foreground encoder to encode foreground images according to their most related background semantics



$$L_{CMC}(h_{fg}, h^+, h^-) = -\log\left[\frac{\exp\left(h_{fg}\cdot h^+/\tau\right)}{\exp\left(h_{fg}\cdot h^+/\tau\right)+\sum_{n=1}^{N}\exp\left(h_{fg}\cdot h_n^-/\tau\right)}\right]$$

Context-aware Discriminator

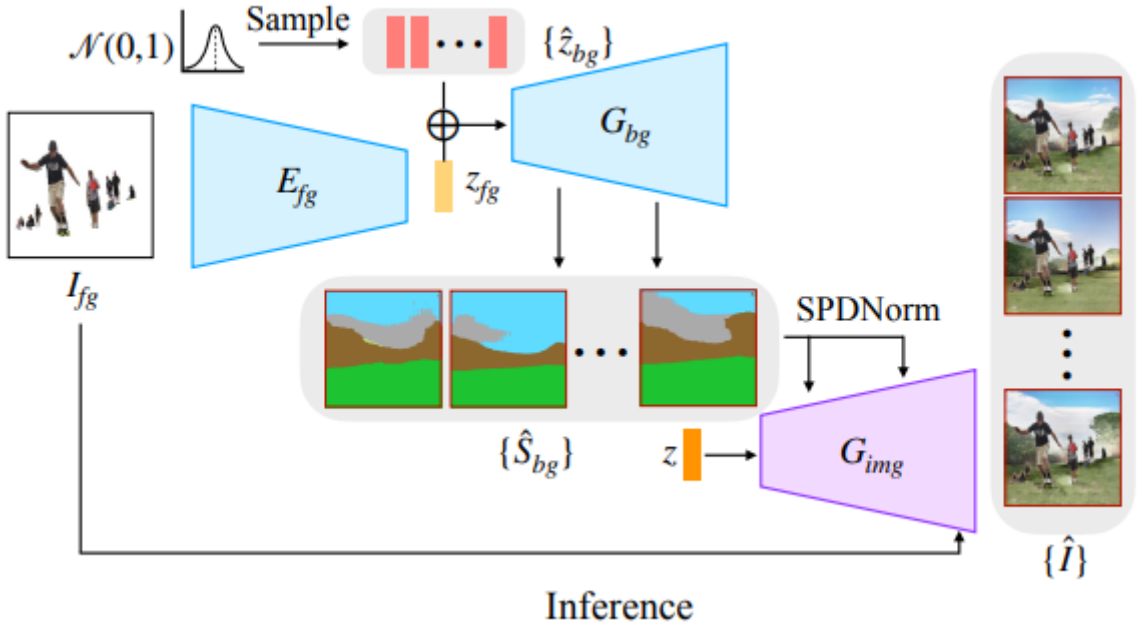
- Detects synthesized region of the generated image
- Applied to content generation stage for context-aware adversarial training
- Predicts a score map, indicating the probability to be real or fake for every spatial location



$$L_{GAN-det}(D_{det}) = \mathcal{E}(D_{det}(\hat{I}), \mathbf{m}) + \mathcal{E}(D_{det}(I), \mathbf{m}^0),$$
$$L_{GAN-det}(G_{img}) = \mathcal{E}(D_{det}(\hat{I}), \mathbf{m}^0),$$

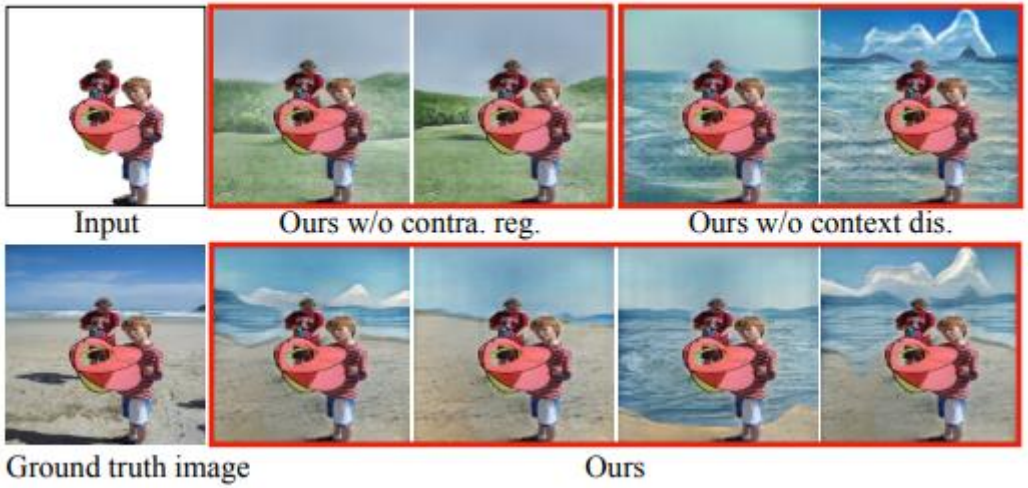
ConTextual Outpainting GAN (CTO-GAN)

- Iinference
  - Multiple background semantic layouts



Ablation studies

- Ablation studies on contrastive regularization and context aware discriminator
- Without contrastive regularization
  - Inaccurate semantic reasoning
  - Foreground images with the same semantic classes are not well grouped
- Without context aware discriminator
  - Mixes different classes (sea and sand) together



Qualitative results

- Semantically coherent contents, vivid textures

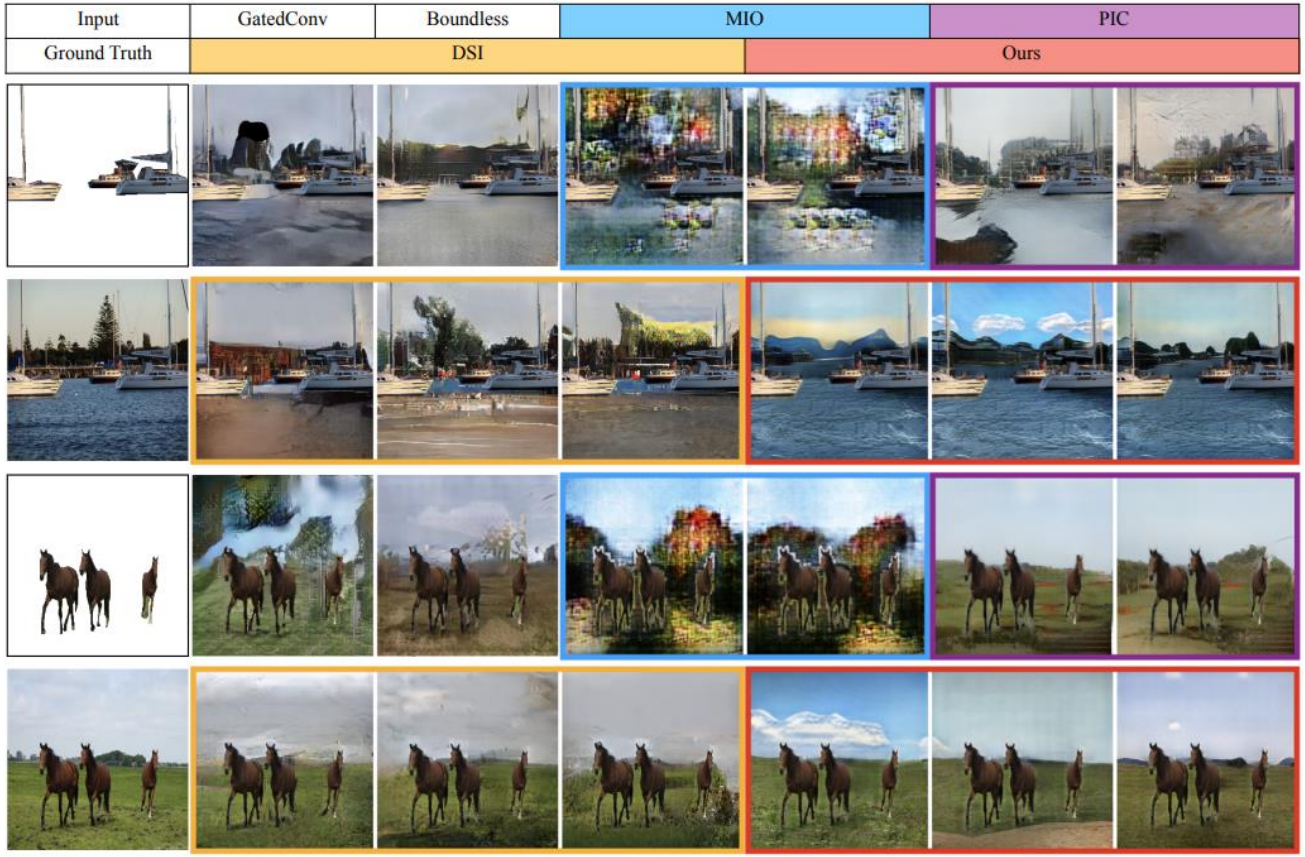


Figure 7. Qualitative comparison with existing methods. For each example, from top to bottom, from left to right, the pictures are: the input foreground image, results of GatedConv [63], Boundless [21], results of MIO [66] (in blue box), results of PIC [71] (in purple box), the ground truth image, results of DSI [39] (in yellow box) and results of our method (in red box).

Quantitative results

- Accomplishment in perceptual quality and semantic coherence
- COCO-Stuff dataset
  - 53,865 training images, 2,252 test images

Metric	Perceptual		Semantic		Subjective	Distortion	
	FID ↓	LPIPS ↓	mIoU ↑	Accu ↑	Avg. Rank. ↓	PSNR ↑	SSIM ↑
GatedConv	40.10	0.436	26.6	38.2	4.25	14.29	0.436
Boundless	31.11	0.411	26.8	38.8	3.40	**15.54**	0.514
MIO	60.33	0.487(0.455)	26.6	31.6	5.39	11.36(12.86)	0.433(0.462)
PIC	33.14	0.417(0.378)	25.4	39.0	3.92	14.37(15.88)	0.467(0.510)
DSI	30.74	0.395(0.351)	26.6	39.1	2.42	14.94(16.22)	0.494(0.542)
Ours	**27.34**	**0.371(0.341)**	**31.5**	**47.0**	**1.61**	14.79(16.01)	**0.529(0.560)**

Bridging the Visual Gap: Wide-Range Blending

- Task which generates smooth transition in the intermediate region between two input images
  - Seamlessly blend them into a novel panoramic image
- Spatial and semantic consistency & visual quality

Bridging the Visual Gap: Wide-Range Blending

- Outpa wards the interm
- Inpair egion is what t
- Exper do not guarai
  - Mi
  - Co

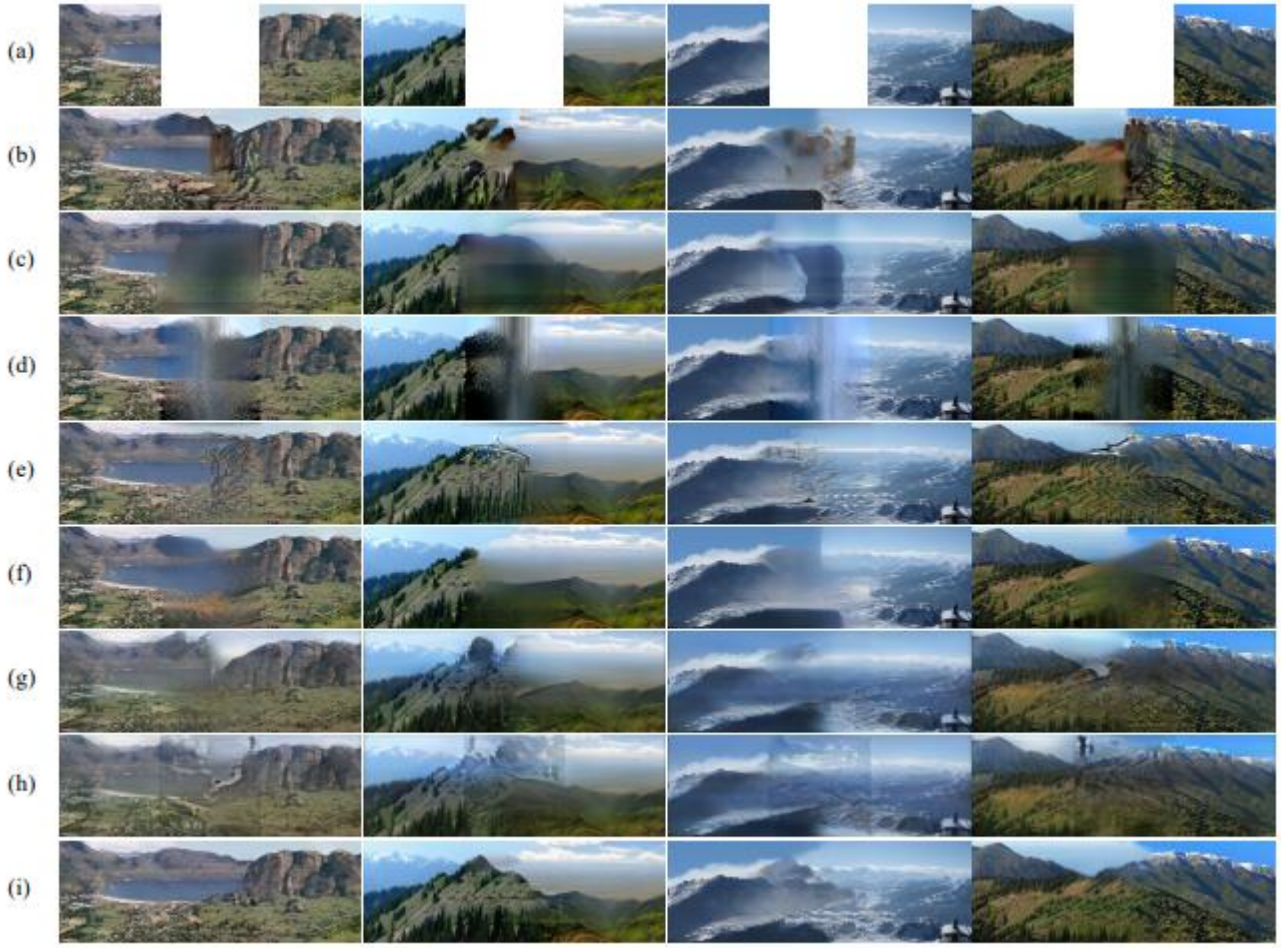
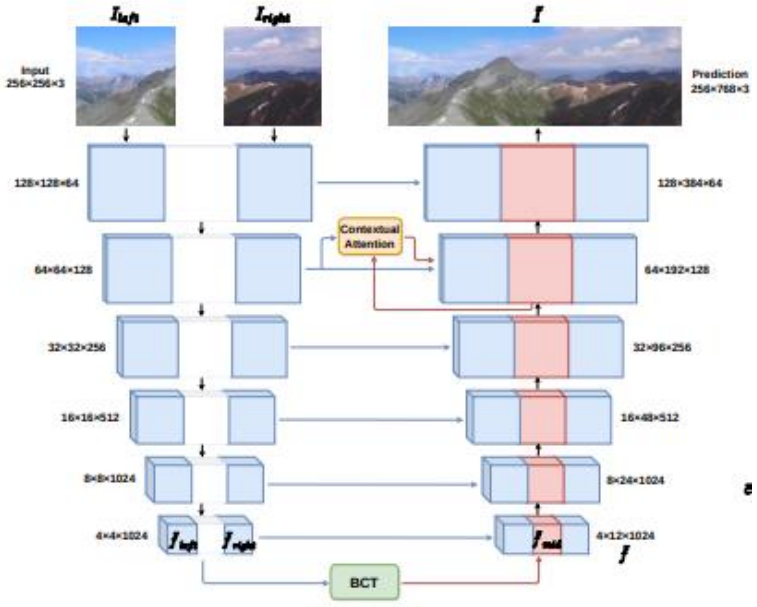


Figure 5: Qualitative comparison with baselines of image inpainting and image outpainting: (a) input images, (b) CA [21], (c) PEN-Net [22], (d) StructureFlow [12], (e) HiFill [20], (f) ProFill [23], (g) SRN [17], (h) Yang *et al.* [19], and (i) Ours.

Model design

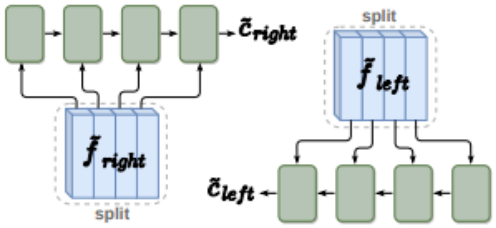
- Output wide-range image is obtained by horizontally concatenating three portions($\tilde{I}_{left}, \tilde{I}_{mid}, \tilde{I}_{right}$)
  - $\tilde{I}_{left}$ and $\tilde{I}_{right}$ should be identical to corresponding I_{left} and I_{right}
  - $\tilde{I}_{mid}$ should provide smooth transition between $\tilde{I}_{left}$ and $\tilde{I}_{right}$
- Bidirectional Content Transfer module in the bottleneck
- Contextual attention on skip connection



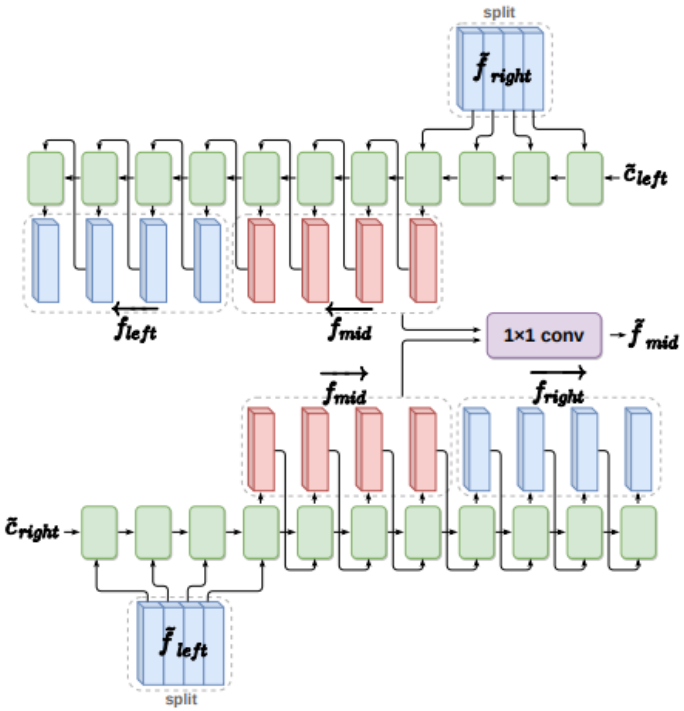
(a) Full model.

Bidirectional Content Transfer (BCT)

- For smooth transition
- Long Short-Term Memory (LSTM) encoder and a conditional LSTM decoder
- Performs sequential prediction for generating $\tilde{f}_{mid}$, conditioned on $\tilde{f}_{left}$ or $\tilde{f}_{right}$



(b) LSTM encoder.



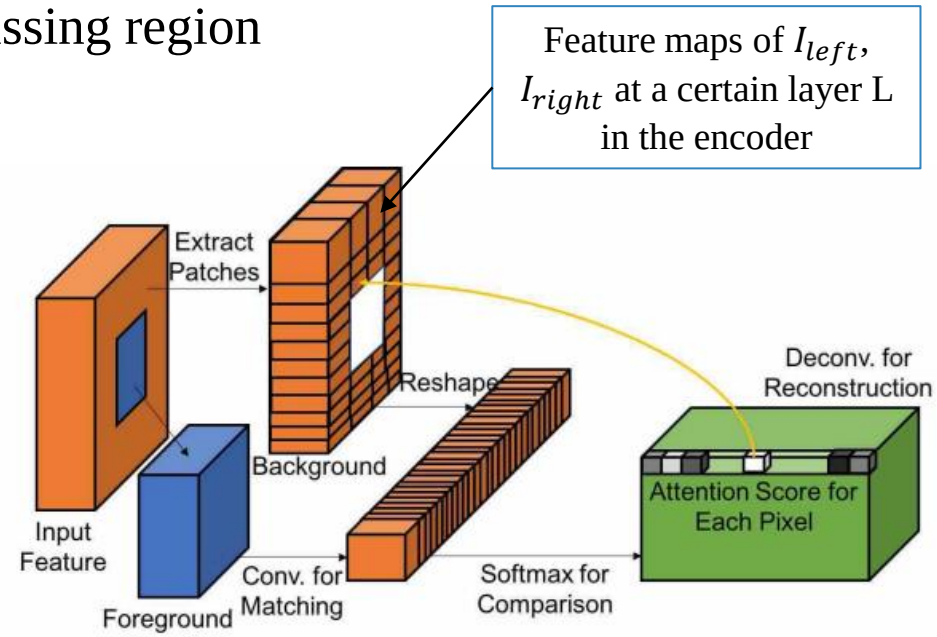
(c) LSTM decoder.

$$\tilde{c}_{right} = \mathcal{E}_{BCT}(\{f^k_{right}\}^K_{k=1})$$

$$\left(\{\overrightarrow{f^k_{mid}}\}^K_{k=1}, \{\overrightarrow{f^k_{right}}\}^K_{k=1}\right) = \mathcal{D}_{BCT}(\{f^k_{left}\}^K_{k=1}, \tilde{c}_{right})$$

Contextual Attention on Skip Connection

- For textures and details
- Matching scores between patches extracted from the surrounding regions and the missing region are computed by cosine similarity
- Softmax is applied on matching scores to get attention score for each patch in the missing region



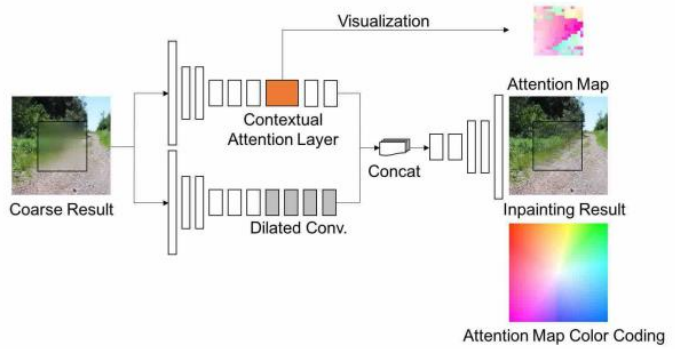


Figure 4: Based on coarse result from the first encoder-decoder network, two parallel encoders are introduced and then merged to single decoder to get inpainting result. For visualization of attention map, color indicates relative location of the most interested background patch for each pixel in foreground. For examples, white (center of color coding map) means the pixel attends on itself, pink on bottom-left, green means on top-right.

Two-stage training

- 1) Self-reconstruction stage
  - Learn to generate high quality images through the fully-guided supervised learning upon self-reconstruction
  - I_{mid} is given (full image provided)

- 2) Fine-Tuning stage
  - Enhance the ability of blending distinct images
  - I_{mid} is not given (two different images provided)

Training objectives

- Pixel reconstruction loss
  - self-reconstruction stage

$$\mathcal{L}_{pixel}^{\text{SR}} = \sum \|\tilde{I}_{left} - I_{left}\|_2 + \|\tilde{I}_{right} - I_{right}\|_2 + \|M \odot (\tilde{I}_{mid} - I_{mid})\|_2,$$

  - Based on finetuning stage

$$\mathcal{L}_{pixel}^{\text{FT}} = \sum \|\tilde{I}_{left} - I_{left}\|_2 + \|\tilde{I}_{right} - I_{right}\|_2$$

Training objectives

- Feature reconstruction loss
  - Self-reconstruction stage
  - GT of the intermediate region is available
  - Extract feature map of GT via image encoder

$$\mathcal{L}^{\mathrm{SR}}_{feat\_rec} = \sum \|\tilde{f}_{mid} - \mathcal{E}(I_{mid})\|_2.$$

- Feature consistency loss
  - Bidirectional Content Transfer module predicts $\{\overrightarrow{f_{mid}}, \overrightarrow{f_{right}}\}$ from $\tilde{f}_{left}$ and vice versa
  - Ideally feature maps $\{\tilde{f}_{left}, \overrightarrow{f_{mid}}, \overrightarrow{f_{right}}\}$ and $\{\overleftarrow{f_{left}}, \overleftarrow{f_{mid}}, \tilde{f}_{right}\}$should be consistent to each other

$$\mathcal{L}_{feat\_con} = \sum \|\tilde{f}_{left} - \overleftarrow{f_{left}}\|_2 + \|\overrightarrow{f_{mid}} - \overleftarrow{f_{mid}}\|_2 + \|\overrightarrow{f_{right}} - \tilde{f}_{right}\|_2.$$

Quantitative results

- Scenery dataset
  - 5040 training images
  - 1000 testing images

Method		FID(↓)	KID(↓) mean	KID(↓) std
Inpainting	CA [21]	91.87	0.0745	0.0022
	PEN-Net [22]	159.70	0.1151	0.0020
	StructureFlow [12]	138.13	0.2168	0.0023
	HiFill [20]	139.39	0.1230	0.0028
	ProFill [23]	46.53	0.0326	0.0011
Outpainting	SRN [17]	70.94	0.0392	0.0012
	Yang *et al.* [19]	82.69	0.0446	0.0012
Ours		**36.13**	**0.0116**	**0.0005**

Table 2: Quantitative comparison with respect to various baselines from image impainting and outpainting.