Structural inpaiting

Huy V. Vo
Ngoc Q. K. Duong
Patrick Pérez

Visual inpainting at large

- The task of filling in a plausible way a region in an image
- Variety of forms and names: completion, reconstruction, disocclusion, hallucination, recovery,...
- Numerous applications: restoration and editing of visual content

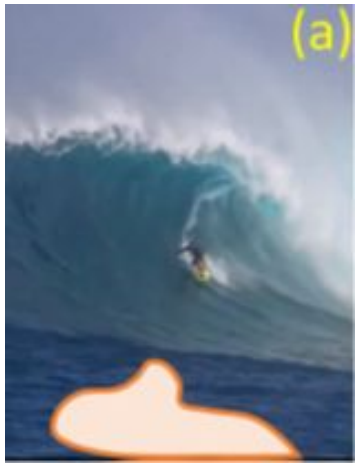
damaged image restored image

Image w. missing region Our inpainted image

technicolor

Visual inpainting at large

(a) Single texture: many satisfactory fillings (with generic tools) exist [1]



(b) Multiple textures, the interface between the textured regions restricts reconstruction freedom

Patch-based inpainting: greedy approaches [2] or iterative optimization-based approaches [3]

[1] Efros and Leung, "Texture synthesis by non-parametric sampling," In Proc. Int. Conf. Computer Vision, 1999
[2] Criminisi et al., "Region filling and object removal by exemplar-based image inpainting," IEEE Trans. Image Processing, 2004
[3] Arias et al., "A variational framework for exemplar-based image inpainting," Int. J. Computer Vision, 2011

technicolor

Visual inpainting at large

(c) Single or multiple structures:
filling-in is very contrived

(d) Content with strong semantics:
the most challenging case

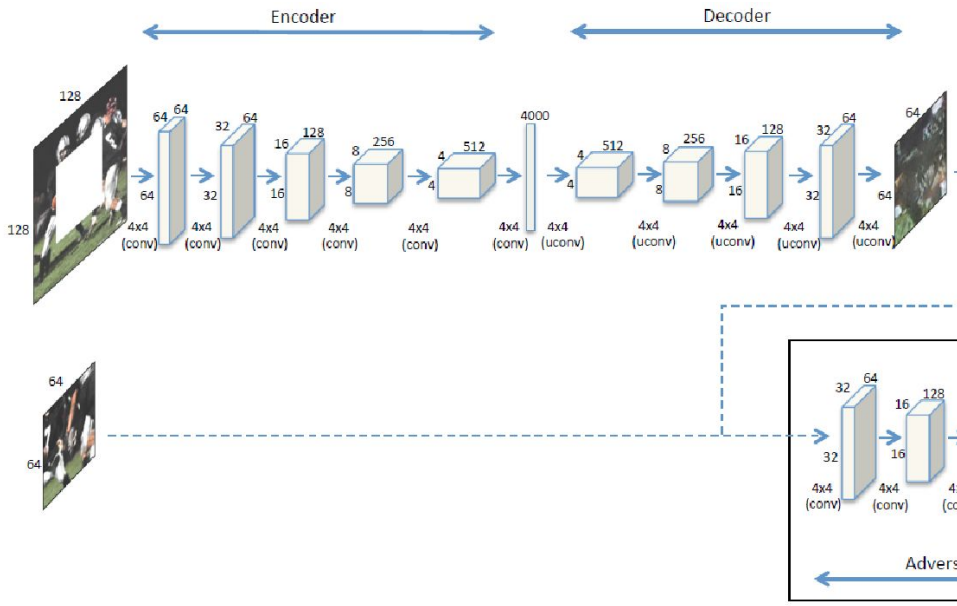
- Patch-based approach [3] or DNN-based approach [4]
- **Class-specific inpainting** [5]: requires the training of a class-specific appearance model

[3] Arias et al., "A variational framework for exemplar-based image inpainting," Int. J. Computer Vision, 2011
[4] Pathak et al., "Context encoders: Feature learning by inpainting," In Proc. CVPR, 2016
[5] Raymond et al., "Semantic Image Inpainting with Deep Generative Models," In Proc. CVPR 2017

technicolor

Context encoder (CE)

- A deep encoder-decoder architecture trained to reconstruct images with missing parts [4]
- Ability to recover complex, semantic structures is impressive in some cases where patch-based approaches are useless!



$$\mathcal{L}_{rec}(x) = \|\hat{M} \odot (x - F((1 - \hat{M}) \odot x))\|_2^2$$

$$\mathcal{L} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv}$$

[4] Pathak et al., "Context encoders: Feature learning by inpainting," In Proc. CVPR, 2016

technicolor

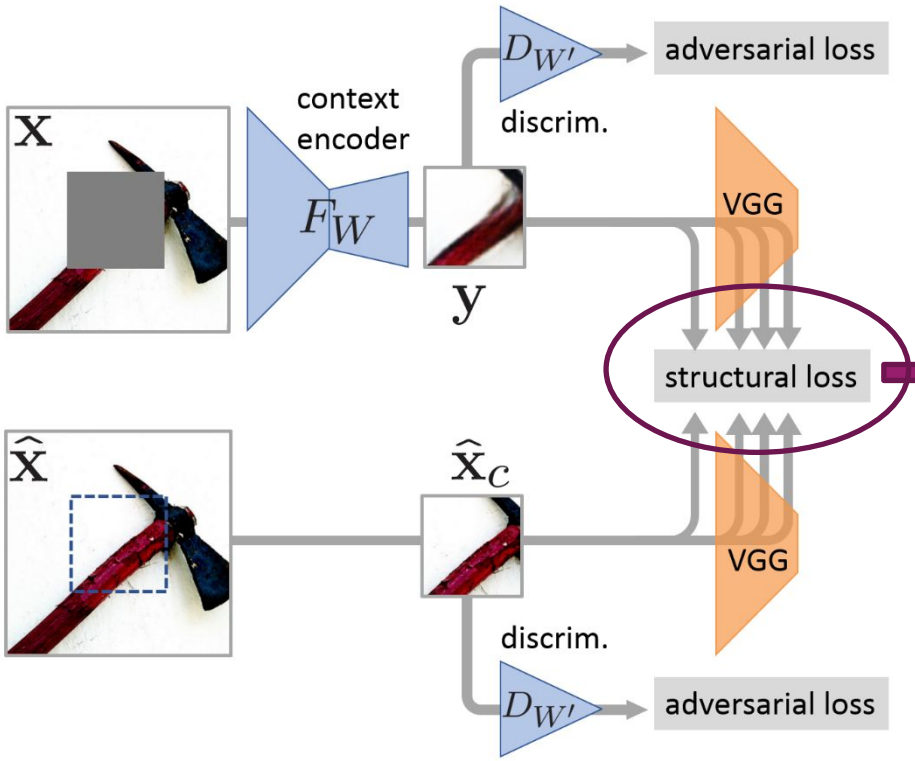
Limitations of the CE

- Surrounding context that CEs actually exploit is mostly local, sometimes only a few pixel wide with no access to visual semantics
- Poor in handling structure, possibly because the adversarial loss contributes way more to the texture than to the structure of the completed scene

[4] Pathak et al., "Context encoders: Feature learning by inpainting," In Proc. CVPR, 2016

technicolor

Proposed structural CE



$$\mathcal{L}_{\text{struct}} = \lambda_0 \mathcal{L}_{\text{pix}} + \sum_\ell \lambda_\ell \mathcal{L}_{\text{feat},\ell},$$

$$\mathcal{L}_{\text{pix}}(\mathbf{y}, \hat{\mathbf{x}}_c) = \|\mathbf{y} - \hat{\mathbf{x}}_c\|_F^2.$$

$$\mathcal{L}_{\text{feat},\ell}(\mathbf{y}, \hat{\mathbf{x}}_c) = \|\phi_\ell(\mathbf{y}) - \phi_\ell(\hat{\mathbf{x}}_c)\|_F^2$$

Training:

$$\min_W \max_{W'} \frac{1}{N} \sum_{n=1}^{N} \Big[\mathcal{L}_{\text{struct}}\big(F_W(\mathbf{x}^{(n)}), \hat{\mathbf{x}}_c^{(n)}\big) + \gamma \mathcal{L}_{\text{adv}}\big(F_W(\mathbf{x}^{(n)}), \hat{\mathbf{x}}_c^{(n)}; W'\big) \Big]$$

[6] Johnson et al., "Perceptual losses for real-time style transfer and super-resolution," In Proc. ECCV, 2016

technicolor

Post-processing

- Optimization-based refinement [7]: built on variational patch-based approach, this refinement seek a reconstruction whose patches have as good matches as possible outside the hole.

correspondence field that maps each pixel in the hole to one outside

Objective function to be minimized:

$$E(\mathbf{x},\psi) = \alpha \sum_{p\in \text{hole}} \sum_{\ell\in L} \left\|\phi_\ell(\mathbf{x},p) - \phi_\ell(\mathbf{x},\psi(p))\right\|_F^2$$
$$+\alpha' \sum_{\ell\in L} \left\|\phi_\ell(\mathbf{x}_c) - \phi_\ell(\mathbf{y})\right\|_F^2 + \beta \text{TV}(\mathbf{x}),$$

[7] Yang et al., "High-Resolution Image Inpainting using Multi-Scale Neural Patch Synthesis," Proc. CVPR, 2016

Experimental architecture

Encoder-decoder network:

- **Input**: Color image of size 128 × 128 × 3
- **Encoder**: Five convolutional layers (4 × 4 filters with stride 2 and ReLU) with 64, 64, 128, 256 and 512 channels, respectively
- **Bottleneck**: A fully connected layer of size 2000 (half size of Pathak's)
- **Decoder**: Four convolutional layers mirroring the last four of the encoder. In order to avoid the checker-board effect that showed up in our first experiments, we replaced the original "deconvolutional" design by the upsampling+convolution alternative proposed in [8]
- **Output**: Color image of size 64 × 64 × 3.

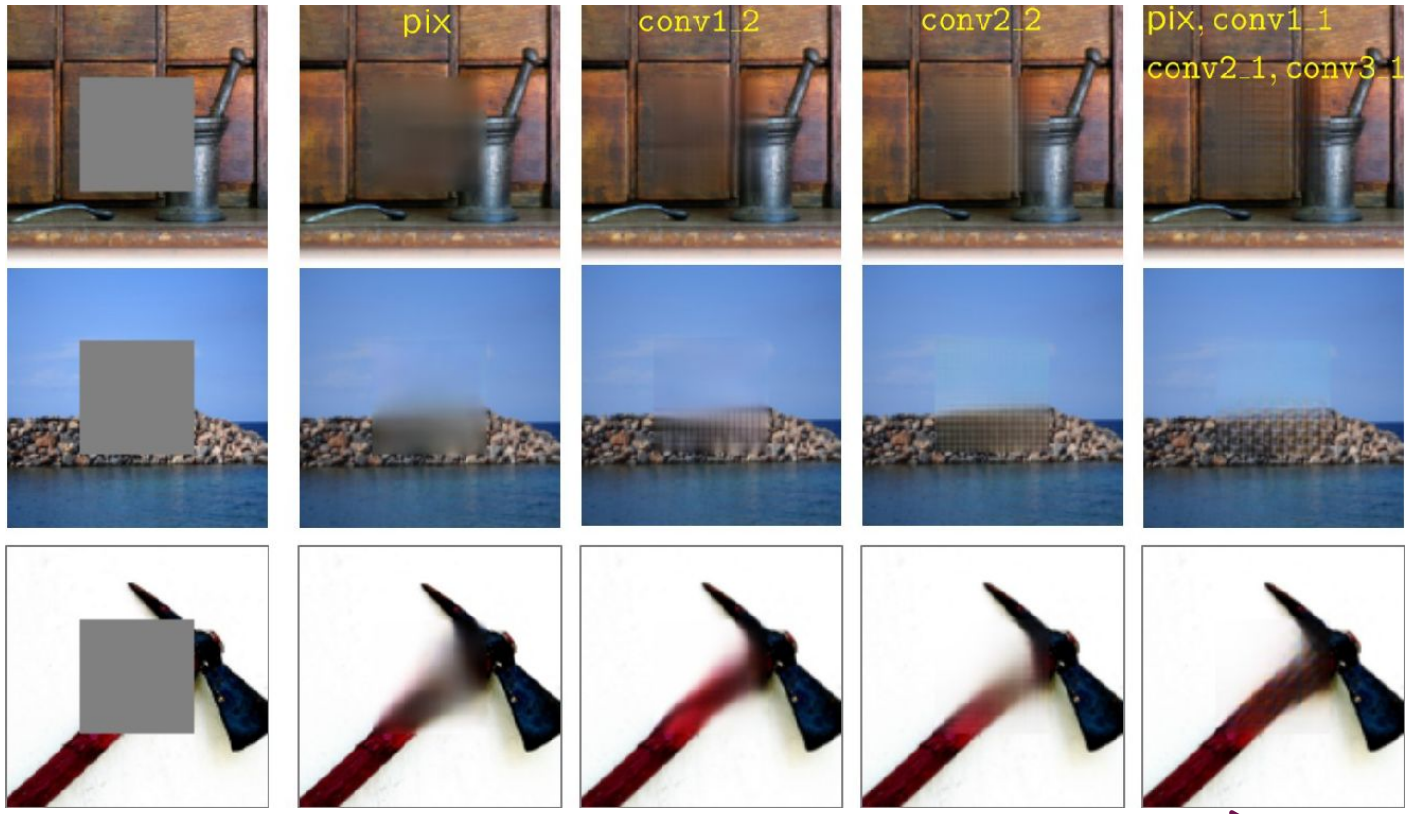
Adversarial network takes 64×64×3 inputs and is composed of four convolutional layers (4 × 4 filters and ReLU). It is lighter than the one in Pathak et al., with four times fewer parameters.

[8] Odena et al., "Deconvolution and Checkerboard Artifacts," Distill, 2016

technicolor

Results with different choices of structural loss

Trained on 1.2M images from ImageNet.



$$\mathcal{L}_{\text{struct}} = \mathcal{L}_{\text{pix}} + \mathcal{L}_{\text{feat, conv1\_1}} + \mathcal{L}_{\text{feat, conv2\_1}} + \mathcal{L}_{\text{feat, conv3\_1}}$$

technicolor

Benefit of adversarial loss

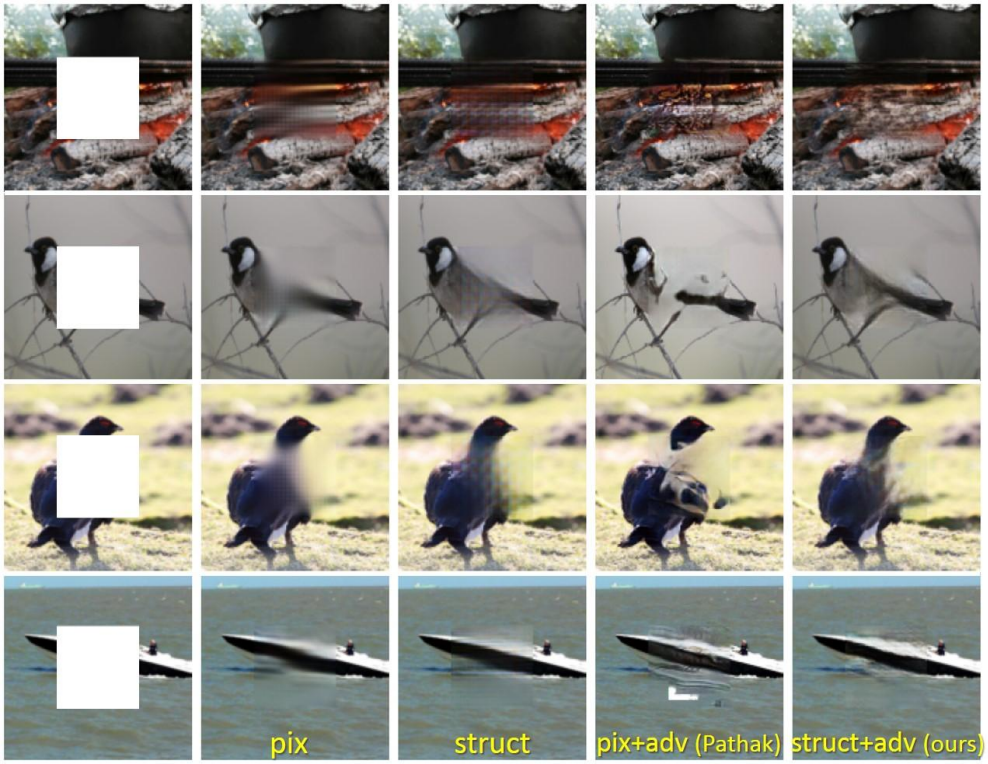
Structural loss alone: grid-like artifacts

Structural loss + adversarial loss

Curriculum learning trick: proceeds with 50 epochs of training with structural loss, followed by 10 epochs of adversarial training.

CE inpainting with different losses

The proposed combination of adversarial and structural losses provides the best results



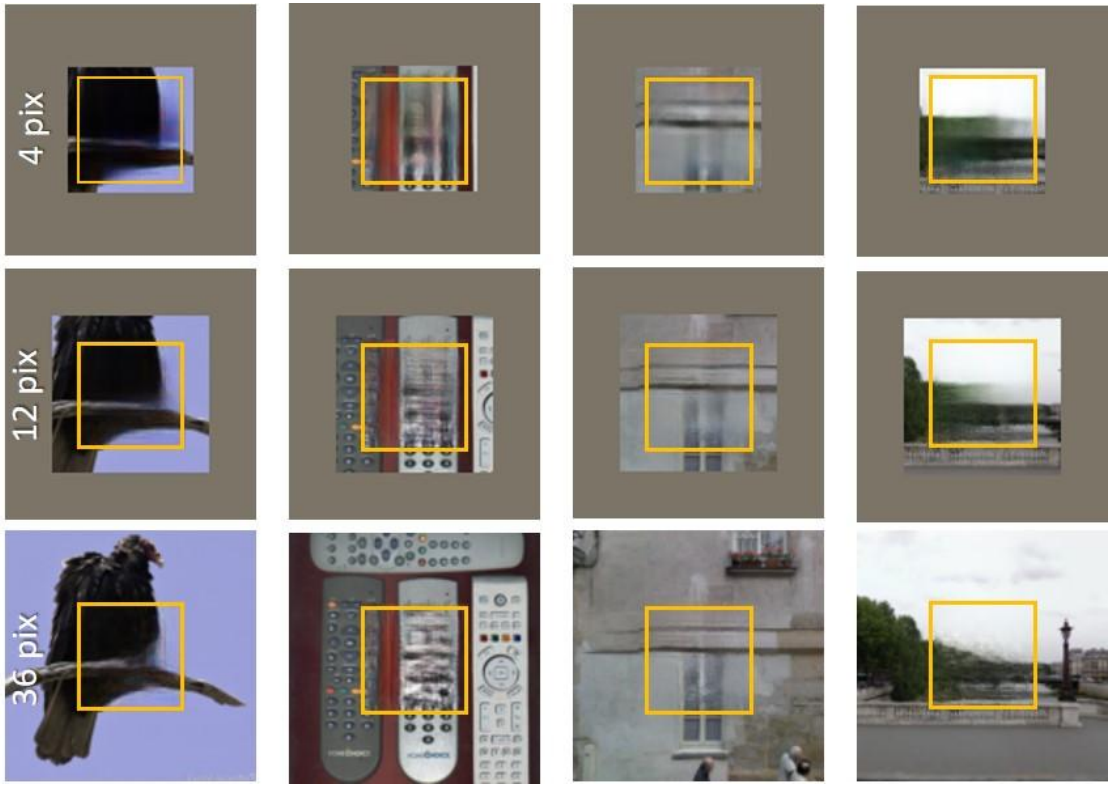
Qualitative results.

	av. ℓ_1 error	av. ℓ_2 error	PSNR
Pathak (Paris)	8.37%	1.63%	19.57dB
ours (ImageNet)	8.07%	1.49%	19.89dB
ours (Paris)	7.53%	1.35%	20.59dB

Quantitative results on 100 ParisStreetView images.

technicolor

Effective context



pix.	4	8	12	16	24	36
ℓ_1	11.31	8.67	8.74	8.08	7.71	7.53
ℓ_2	2.11	1.54	1.54	1.42	1.38	1.35
PSNR	17.38	19.36	19.36	19.98	20.39	20.59

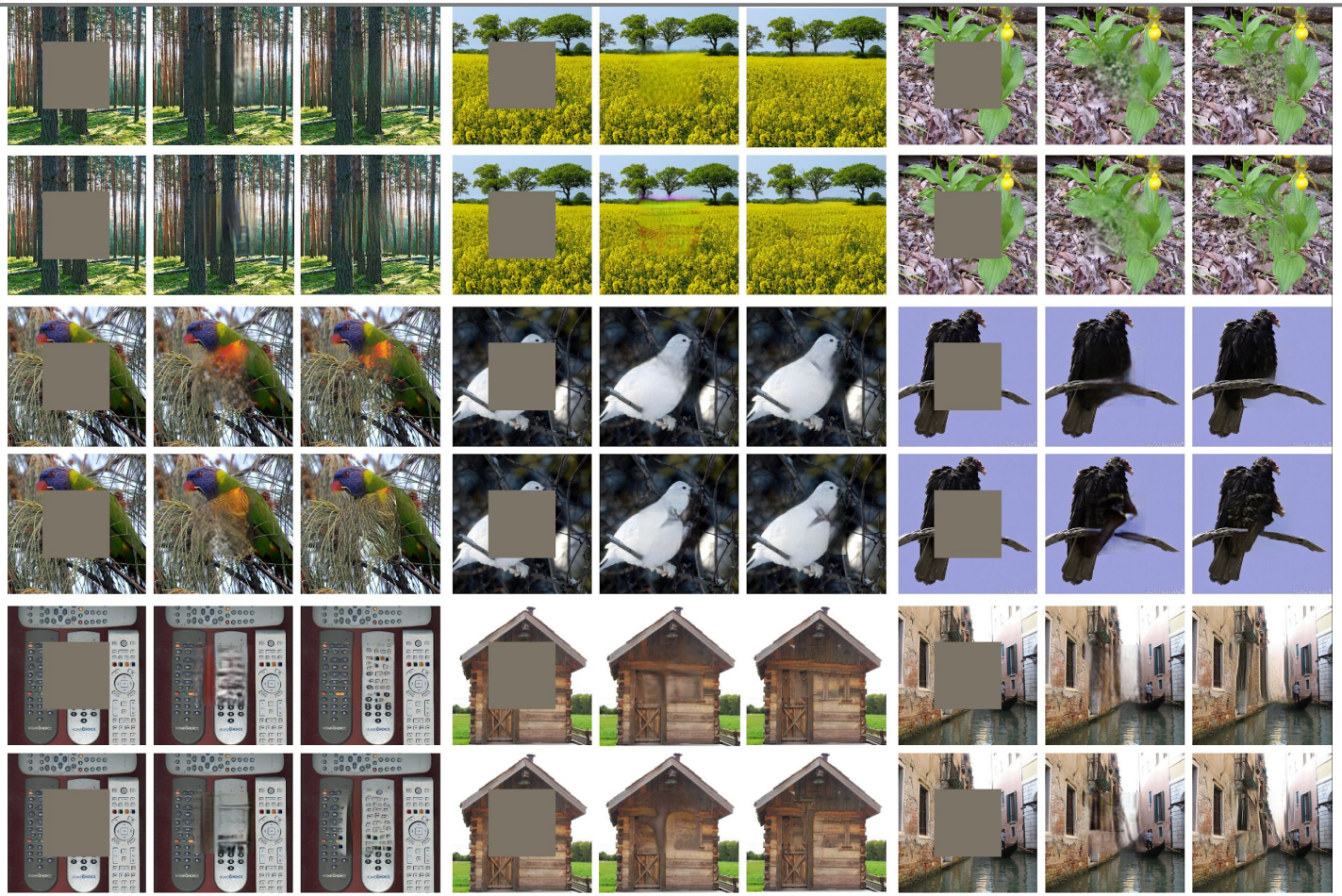
Inpaiting with context of 4, 12, and 36 pixels from the border.

- Robustness: structure completions are possible even with as few as 4 pixels known by the CE
- CEs contain only little object or scene-specific knowledge.

technicolor

Results with optimization-based refinement

For each input image, inpainting by the proposed CE, before and after optimization-based refinement (top) and same for Pathak et al.'s CE (bottom).



- For more than 83% of the images, our reconstruction was more often preferred in the user test.
- 51% (resp. 39%) of the images inpainted by our method (resp. Pathak's method) were considered as natural by at least 50% of participants.

Failure examples

Visual/semantic complexity of the scene defeats both CEs, and patch-based methods.

technicolor

Conclusion

- CE with structural loss is able to complete even complex structures
- Semantics is playing a limited role in the CE
- Inpaiting quality is significantly enhanced by optimization-based refinement

Future work:

- A deeper use of automatic scene understanding
- Relaxing current geometric constraints (inpainting a square domain), incorporating user's input in a seamless fashion.

Original input Inpainted image

Thank you for your attention!

Original image with missing region

Inpainted image

technicolor