STRUCTURAL INPAINTING

Huy V. Vo[1], Ngoc. Q. K. Duong[2], Patrick Pérez[3]

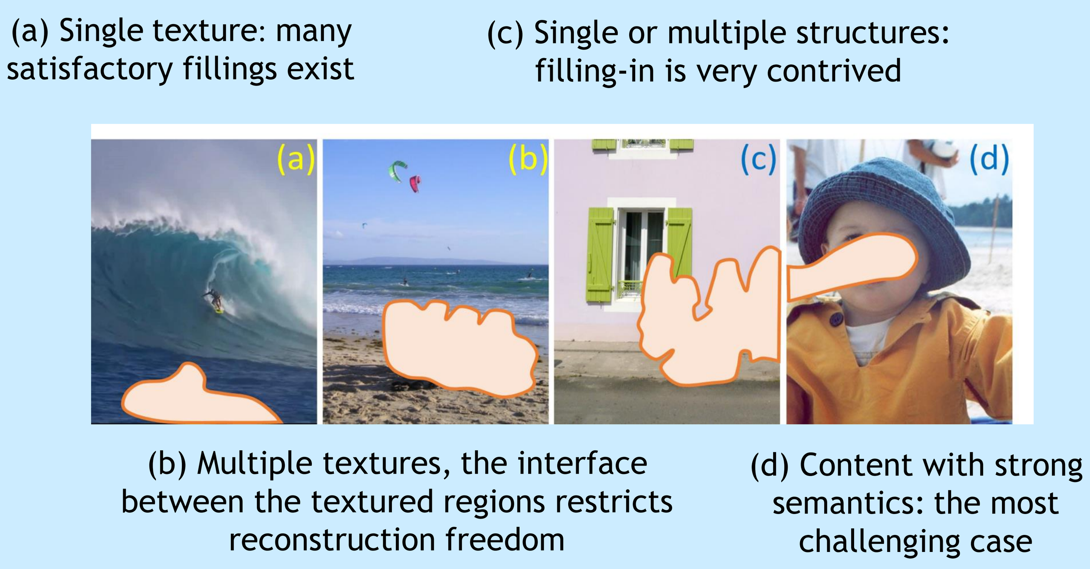
[1] Ecole Polytechnique, France

[2] Technicolor Research & Innovation, France

[3] Valeo.ai, France

Goal: filling in a plausible way a region in an image, better handling structure than the prior art.

Application: restoration and editing of visual content

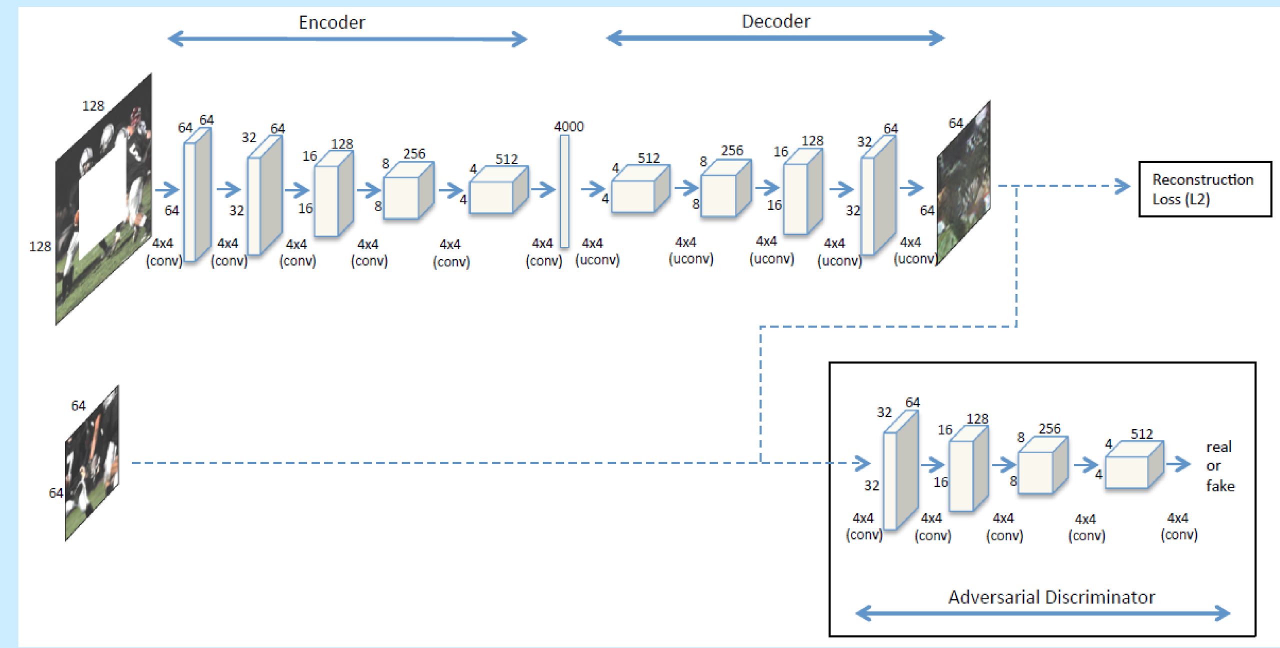


Context Encoder (CE) [1]:

- A deep encoder-decoder architecture trained to reconstruct images with missing parts
- Ability to recover complex scene in some cases where patch-based approaches are useless
- Limitation: (1) poor handling of structures
  (2) little access to visual semantics

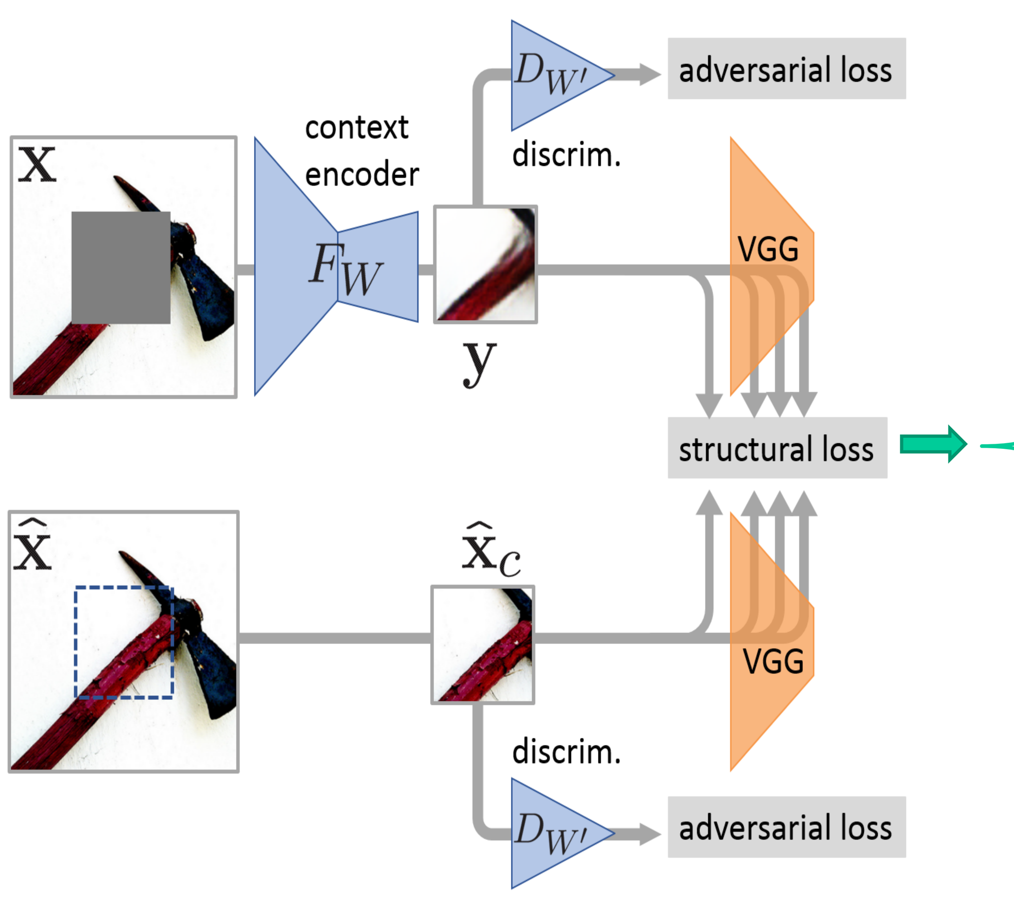
$$\mathcal{L}_{rec}(x) = \|\hat{M} \odot (x - F((1-\hat{M}) \odot x))\|_2^2$$

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



CE for image inpaiting (image is from [1]).

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

Optimization-based refinement [3]

- Built on variational patch-based approach, this refinement seek a reconstruction whose patches have as good matches as possible outside the hole.
- Objective function to be minimized:

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

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

Experimental architecture

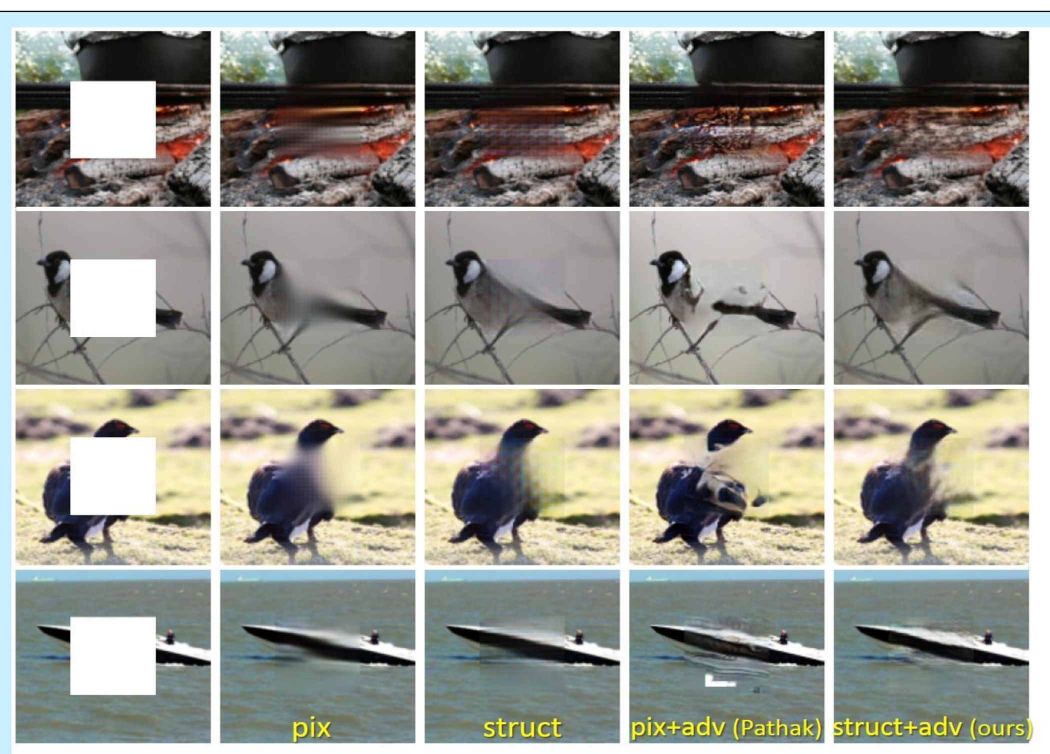
Encoder-decoder network (input is color image of size 128 × 128 × 3, output is color image of size 64 × 64 × 3)

✓ **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)

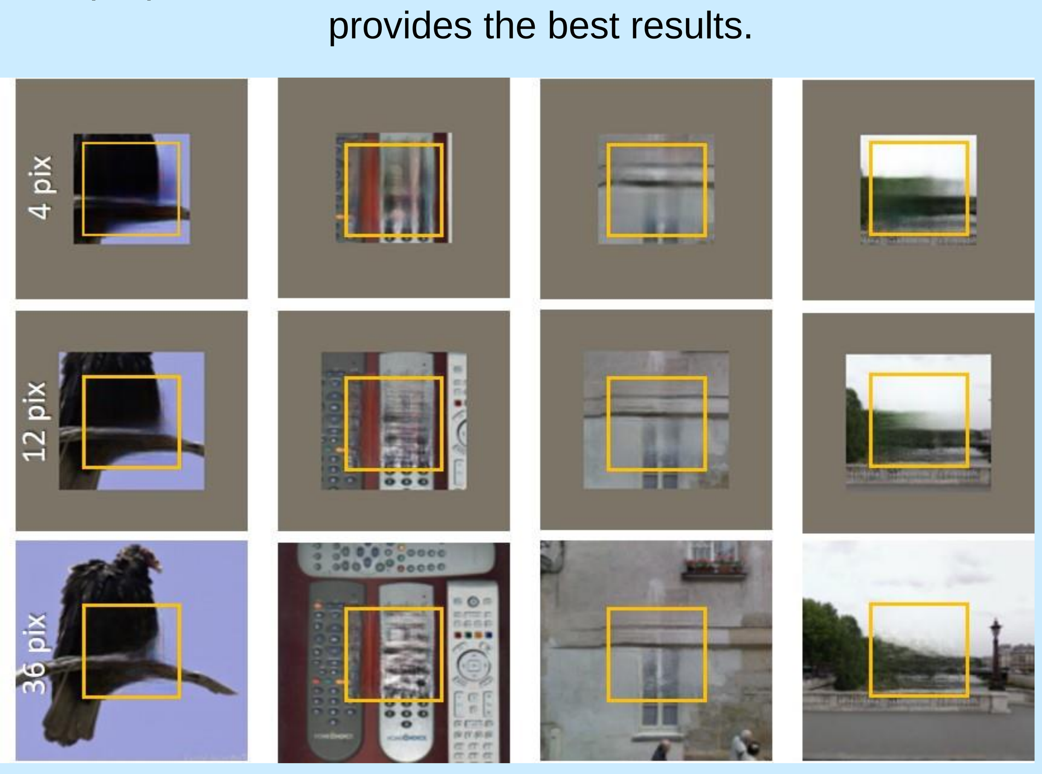
✓ **Decoder:** Four convolutional layers mirroring the last four of the encoder. In order to avoid the checker-board effect that showed up in our first experiments, we replaced the original "deconvolutional" design by the upsampling+convolution alternative proposed in [4]

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.

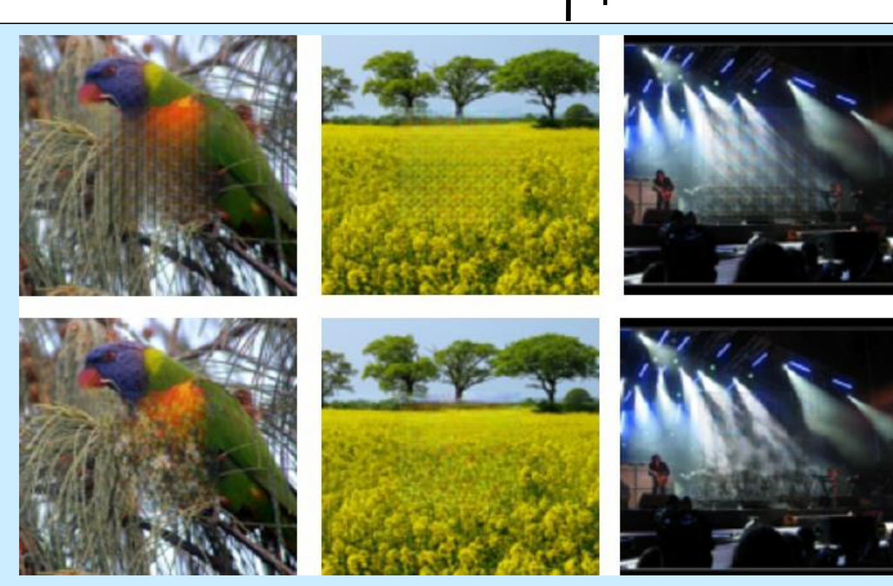


	av. ℓ_1 error	av. ℓ_2 error	PSNR
Pathak (Paris)	8.37%	1.63%	19.57dB
ours (ImageNet)	8.07%	1.49%	19.89dB
ours (Paris)	7.53%	1.35%	20.59dB

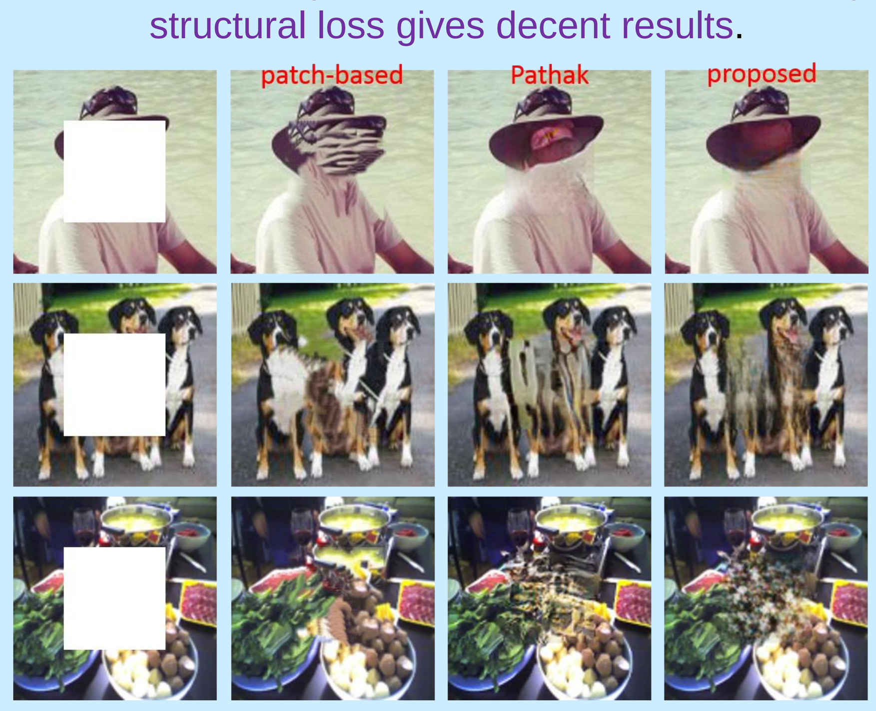
CE inpainting with different losses: qualitative results (above) and quantitative results on 100 ParisStreetView images (below). The proposed combination of adversarial and structural losses provides the best results.



Effective context: inpainting with context of 4, 12, and 36 pixels from the border. Structure completions are possible even with as few as 4 pixels known by the CE → CEs contain only little object or scene-specific knowledge!



Benefit of adversarial loss: Structural loss alone (above) gives grid-like artifacts; Structural loss + adversarial loss(below). Note: adversarial loss is only added after the CE trained only with structural loss gives decent results.

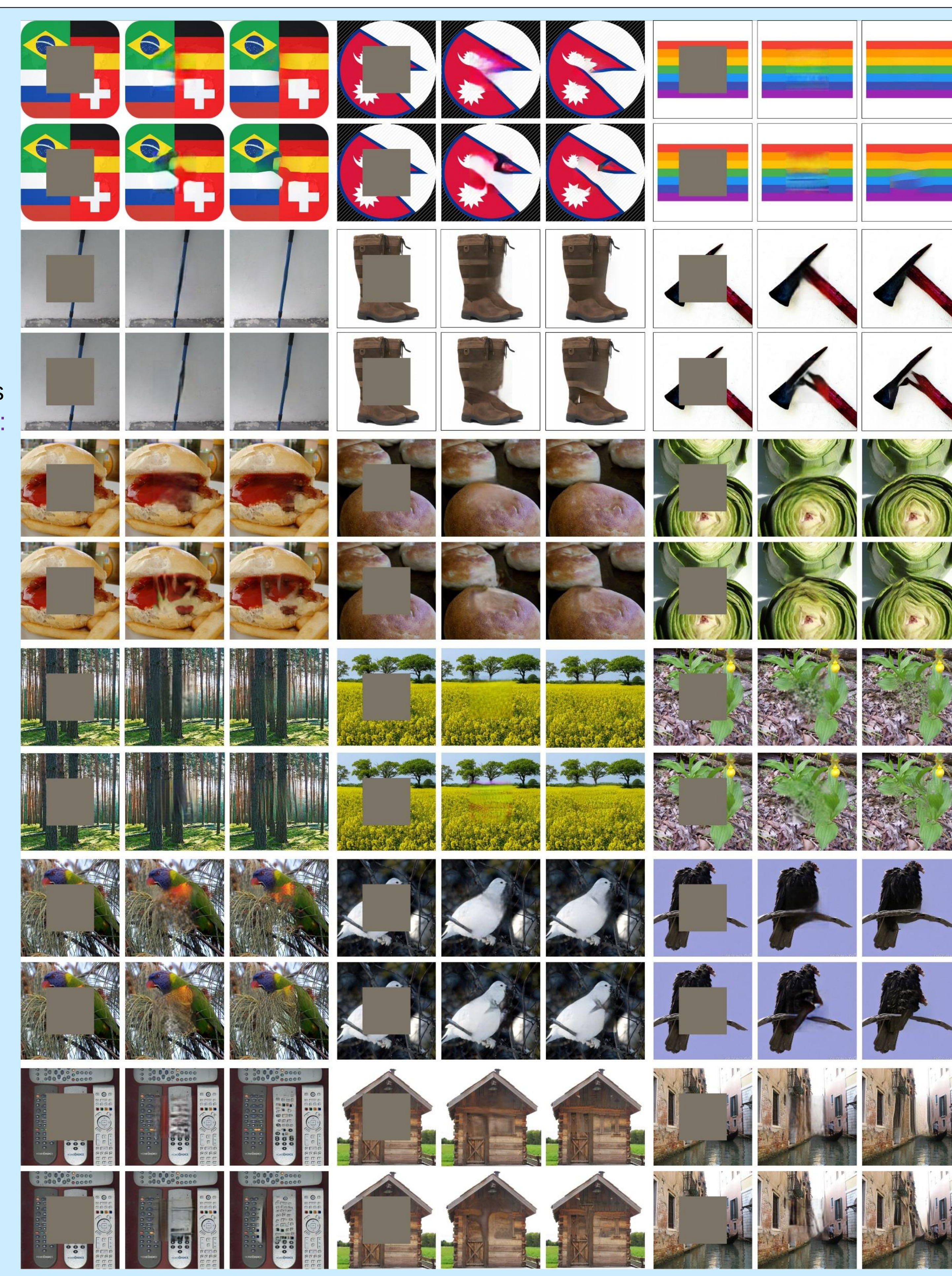


Failure examples: Visual/semantic complexity of the scene defeats both CEs, and patch-based methods.

User study: with 35 participants of various ages and occupations, for more than 83% of the tested ImageNet images, our reconstruction was more often preferred than Pathak's CE. Other user studies about the quality of inpainted images can be found in the paper.

Conclusion

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



CE inpainting followed by 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). Each row contains scenes that are related in a way: Flag graphics; Simple rigid structures; Natural non-rigid objects; Multi-texture scenes; Birds on branches; More complex rigid structures.

References

[1] Pathak et al., "Context encoders: Feature learning by inpainting," In Proc. CVPR, 2016.
[2] Johnson et al., "Perceptual losses for real-time style transfer and super-resolution," In Proc. ECCV, 2016
[3] Yang et al., "High-Resolution Image Inpainting using Multi-Scale Neural Patch Synthesis," Proc. CVPR, 2016
[4] Odena et al., "Deconvolution and Checkerboard Artifacts," Distill, 2016.