An Analysis of Various Image Inpainting Methods

Yuxing Chen[1], Zhefan Wang[2]

Symbolic Systems Program, Stanford University[1]; Department of Electrical Engineering, Stanford University[2]

ABSTRACT

Inpainting missing regions makes possible applications such as removal of unwanted objects from an image and reconstruction of occluded regions in image-based 3D scenes. However, filling in the holes of an incomplete image with reasonable contents that is consistent both globally and locally is a very challenging task.

DATASET

- **Microsoft COCO 2017 Train/Val/Test**, 256×256 RGB images
- Normalized and resized to 128×128
- Center 64×64 square cropped

Sample Batch of Preprocessed Real Images

Sample Batch of Preprocessed Cropped Images

BACKGROUND

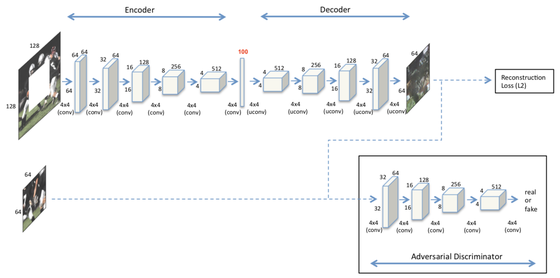
Existing inpainting approaches mainly fall into three categories:

- **Diffusion-based approaches:**
  - can only fill-in tiny holes
  - focus on low-level features $\rightarrow$ less powerful
- **Patch-based approaches:**
  - cannot generate novel objects that do not exist in the source image
  - focus on low-level features $\rightarrow$ less powerful
- **Unsupervised visual feature learning:**
  - can fill-in big holes
  - can generate novel objects
  - can generate more plausible completion

  $\Longrightarrow$ **Most promising category!**

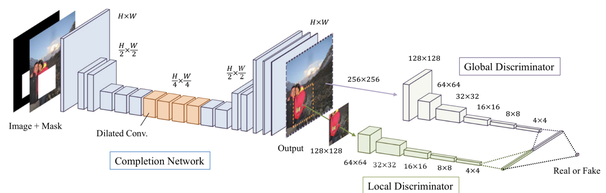
MODELS

Context Encoder



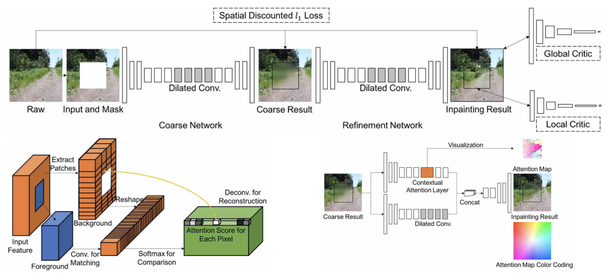
- **Encoder/Decoder** connected via a channel-wise fully-connected layer: each unit in the decoder can reason about the entire image content
- Simple regression towards the ground truth ($\mathcal{L}_{rec}$) is not sufficient: multiple context-consistent ways to fill-in the hole
- Solution: **define joint loss** $\mathcal{L} = \lambda_{\mathbf{rec}}\mathcal{L}_{\mathbf{rec}} + \lambda_{\mathbf{adv}}\mathcal{L}_{\mathbf{adv}}$. λs indicates the "weights" of the losses.

Globally and Locally Consistent Image Completion



- Completion Network: encoder-decoder structure
- **Dual discriminators (global + local):** generate novel objects, and being consistent with the image
- Loss $\mathcal{L} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv}$
- Training split into three phases: train completion network; train the discriminators; train jointly
- **Poisson blending:** remove color inconsistencies

Contextual Attention



- **Contextual attention layer:** explicitly attend on related feature patches at distant spatial locations, ie. matching features of missing pixels (foreground) to surroundings (background)
- Measurement: cosine similarity

  $s_{x,y,x',y'} = \langle \frac{f_{x,y}}{\|f_{x,y}\|}, \frac{b_{x',y'}}{\|b_{x',y'}\|} \rangle$
- Pixel-wise attention score:

  $s^*_{x,y,x',y'} = softmax_{x',y'}(\lambda s_{x,y,x',y'})$
- Further encourage coherency of attention by propagation (fusion)

FUTURE WORK

- Further train current models
- Improve current models by introducing a two-stage course-to-fine network architecture
- Test whether WGAN-GP loss could improve performance
- Add a contextual attention layer (model)
- Do more analysis work

REFERENCES

[1] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. *Proceedings of the IEEE Conference on CVPR*.

[2] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *TOG*, 36(4):107, 2017.

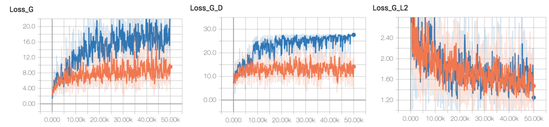
[3] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. *arXiv preprint*, 2018.

EXPERIMENTS & ANALYSIS

- **Input:** center cropped image
- **Output:** input image with all holes completed

Context Encoder

We trained our baseline Context Encoder model (with only L2 loss) for 40 epochs on GPU using 5000 training images. Then we trained another Context Encoder model with joint loss for 40 epochs on the same dataset.



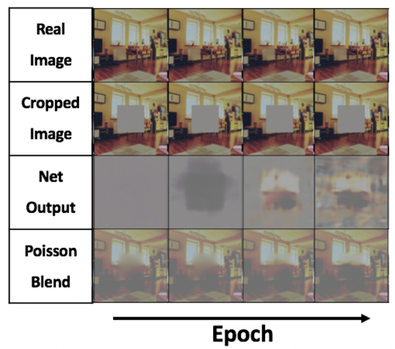
Blue: Only reconstruction loss (L2); Orange: Joint loss (L2+adversarial)

We randomly sample 8 validation images to show the performance progress of the two models over epochs.

Global-Local Net

So far, we trained our Global-Local Net for 6+3+25=34 epochs on GPU using 5000 training images.

Below is a random validation image inferenced at different epochs. Initially, generator highly prefers gray pixels to conservatively lower $\mathcal{L}_{rec}$. Gradually, gradients are learned. Later on, colors are learned.



Global-Local Net is much larger than baseline model and hence need longer training session. But it seems to be promising and we expect its performance to surpass the baseline.