Numerical Imaging Project: Review of the article *Globally and Locally Consistent Image Completion* by S. Iizuka

Victoria BRAMI - Clarine VONGPASEUT

Master Mathématiques Vision Apprentissage

victoria.brami@eleves.enpc.fr - clarine.vongpaseut@eleves.enpc.fr

February 2022

Introduction: Inpainting Principles

- **Goal** Complete missing zones of a given image
- **Application:** Painting restoration, Special effects on Images/Videos, Photomontage etc.

Figure: Inpainting used to restore damaged artwork/pictures

Introduction: Different Approaches for Inpainting

- Historically Handmade Techniques
- Computationally based approaches:
  - **Since 2000s:** Patch Propagation Based Models.
  - **Since 2014:** Generative models, like Auto-Encoders and GANs to predict missing parts of the image.

$\Longrightarrow$ We study a generative deep learning based model in our project

Outline

1. Inpainting process using Neural Networks
2. Experiences on this approach
   - Discriminant ablation study
   - Channel ablation study
3. Comparison with Inpainting using a Patch based method
4. Conclusion

Table of Contents

Inpainting process using Neural Networks: Iizuka et al. model

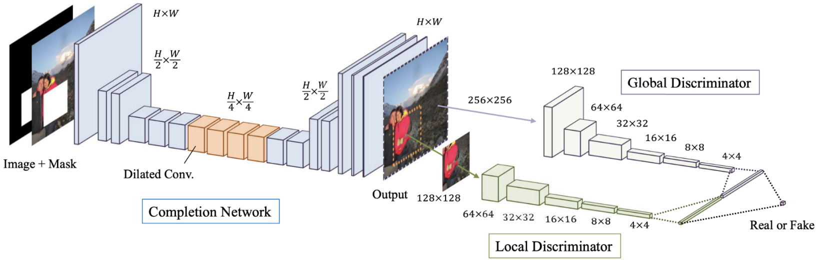

Figure: Architecture of Iizuka et al. model [1]

Inpainting process using Neural Networks: the training process

1st phase

- Completion Network only
- Apply one random mask of dimensions in $[48, 96]^2$ to each 160×160- pixel images
- Back-propagation L2 loss on the area to complete

2nd phase

- Discriminators only
- For the each image genrate two random masks
- BCE loss with images inpainted by the completion Network as fake and the original images as real

Inpainting process using Neural Networks: the training process

3rd phase

- Both networks are trained jointly
- Combining the two loss functions
- Back-propagation for each network using the gradient of the loss function w.r.t. to each network's parameter

Table of Contents

Experiments with Iizuka model: Dataset and metrics

Dataset used for experimentations: CelebA dataset.
Quantitative metrics:

- **Mean Squared Error** (MSE):

$$MSE(I_{GT}, I_{Gen}) = \frac{1}{H}\frac{1}{W}\sum_i \sum_j (I_{GT}(i,j), I_{Gen}(i,j))^2$$

- **Peak-Signal to Noise Ratio** (PSNR):

$$PSNR(I_{GT}, I_{Gen}) = 10\log_{10}(\frac{255^2}{MSE(I_{GT}, I_{Gen})})$$

- **Similarity Index Measure** (SSIM):
  Quantifies image quality degradation.
- **Fréchet Distance** (FID).

Experiments with Iizuka model: FID Score

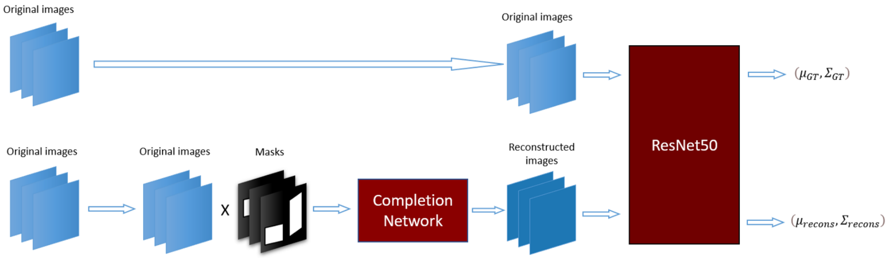

Figure: Computation of Fréchet Distance (FID) Score

Experiments: Discriminator Ablation Study

Tested the model:

1. Without Local discriminator
2. Without Global discriminator.

Training:

Retrained **Phase 2** and **Phase 3**.

Masks:$\approx$ 9% – 36% of the image.

Evaluation:

On CelebA test set.

Masks:$\approx$ 9% – 36% of the image.

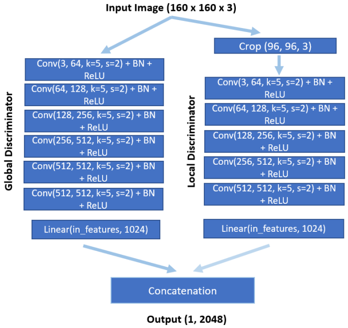

Figure: Discriminators architecture

Experiments: Discriminator Ablation Study

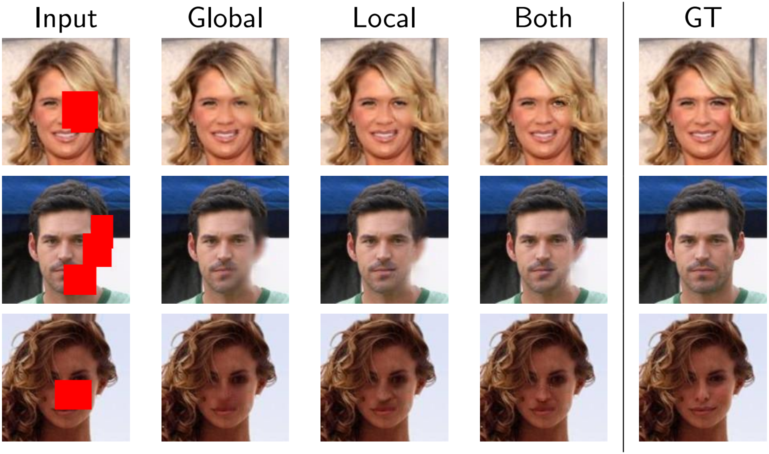

Table: Comparison between the outputs from the 3 models

$\Longrightarrow$ More blurry images when ablating Local discriminator

Discriminant Ablation Study

Quantitative results:

Model	**MSE**	**PSNR**	**SSIM**	**FID Score**
Global only.	0.020	17.221	0.683	$100.75^{\pm 0.2}$
Local only.	0.053	12.942	0.595	$69.57^{\pm 0.08}$
Local and Global	**0.015**	**18.447**	**0.708**	$37.51^{\pm 0.02}$

Table: Evaluation on CelebA test set

$\Longrightarrow$ Combined Context discriminators significantly improves model's performances on all criteriums.

Discriminant Ablation Study: Example

Figure: Ground Truth Only

Discriminator Ablation Study: Example

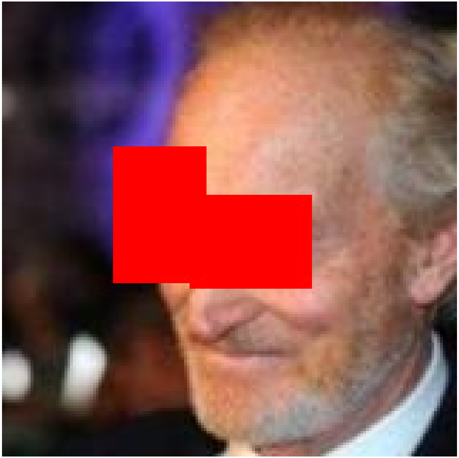

Figure: Input Image

Discriminator Ablation Study: Example

Figure: Global Only

Discriminator Ablation Study: Example

Figure: Local Only

Discriminator Ablation Study: Example

Figure: Local and Global

Experiences on Iizuka Model: Channel Ablation Study

Objective:

Evaluate inner model parameters influence on image completion.

Framework:

Step 1: Remove channels' outputs on the layers of the Completion Network.
Step 2: Evaluate and compare the FID score of the model with the suppressed.

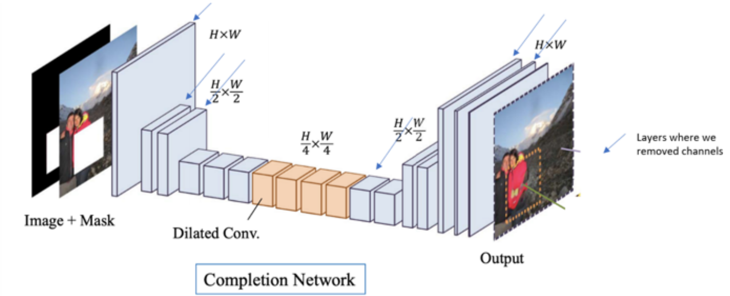

Figure: Layers where channels were been suppressed

Experiences on Iizuka Model: Channel Ablation Study

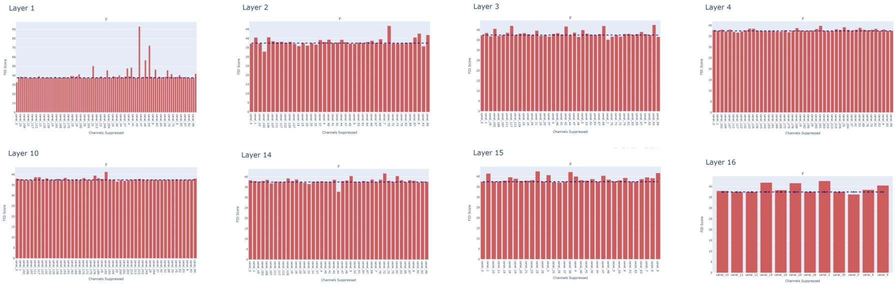

Table: FID scores obtained after removing some channels

- $\Longrightarrow$ Significant increase of FID score on first Conv. layer, on channels 44 and 53 ($FID = 93.0, 72.3$ when normal model is at 37.5).
- $\Longrightarrow$ Decrease of FID score on Conv1 channel 0, conv2 channel 102 and conv14 channel 44 ($FID = 32.5, 32.7$ and 32.8).

Experiences on Iizuka Model: Channel Ablation Study

Visual results

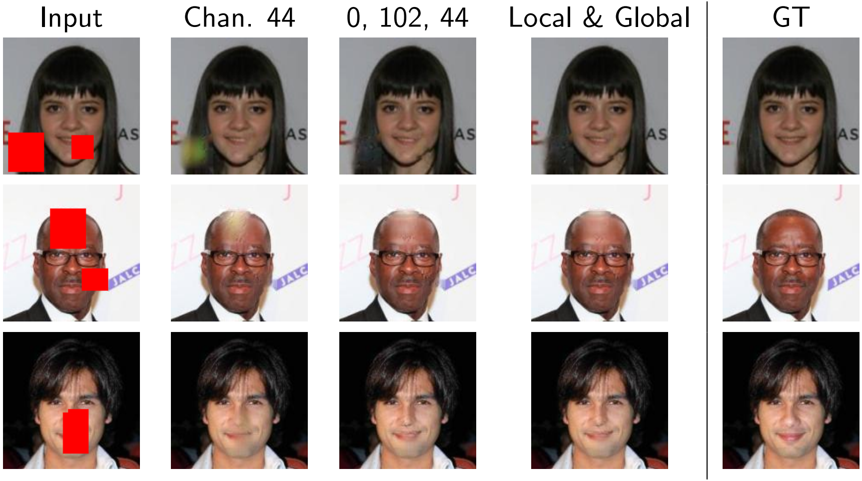

Table: Results on CelebA when removing some channels in conv layers

Experiences on Iizuka Model: Channel Ablation Study

Figure: Ground Truth image

Experiences on Iizuka Model: Channel Ablation Study

Figure: Input image

Experiences on Iizuka Model: Channel Ablation Study

Figure: Conv1 channel 44 removed

Experiences on Iizuka Model: Channel Ablation Study

Figure: Conv1 channel 0, Conv2 channel 102, Conv14 channel 44 removed

Experiences on Iizuka Model: Channel Ablation Study

Figure: Local and Global

Experiments: Channel Ablation Study

Model	mse↓	psnr↑	ssim↑	fid↓
Iizuka (Global)	0.0011	31.938	0.972	8.83
Iizuka (Local)	0.0011	31.907	0.971	8.643
Iizuka (remove Channel 44)	0.0016	29.706	0.964	10.891
Iizuka (remove 3 Channels)	**0.0010**	31.486	0.970	8.061
Iizuka	**0.0010**	**31.983**	**0.973**	**7.743**

Table: Scores of the different models on a batch of 280 images from CelebA test dataset ($2.5 - 25.0\%$ occlusions)

$\Longrightarrow$ Removing some channels in the Completion Net does not implies a huge changes in outputs realism (see FID).

Table of Contents

Patch based method used

- Optimization problem : minimizing distances between patches
- Accounts for texture
- Dependant on patch size, here 7 x 7

Figure: Example where the texture is well reconstructed [2]

Advantages

- Performs well with masks covering the background
- Idem with masks occluding textured regions such as hair

Input	Patch	Iizuka	GT

Table: Comparison of the two methods on images with the background and/or hair occluded

Comparison with a patch based method

Figure: Input

Frame Title

Figure: Patch based method

Figure: Iizuka

Comparison with a patch based method

Figure: Ground truth

Comparison with a patch based method

Disadvantages

- Can't construct structural parts of the face if it's missing
- Long computation time

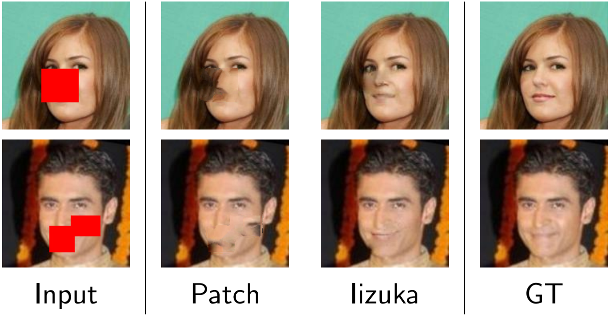

Table: Comparison of the two methods on images with the nose or mouth occluded

Comparison with a Patch Based method

Model	MSE	PSNR	SSIM	FID
Patch-based method [2]	0.0036	26.373	0.943	33.296
Iizuka	**0.0010**	**31.983**	**0.973**	**7.743**

Table: Different metrics evaluated on 280 images of Celeb A test set

Table of Contents

Conclusion

- Importance of both discriminators
- Importance of the first convolution layer
- Removing specific channels seems to improve the results in some cases
- Better performances with Iizuka et al. model than with the patch-based method used for comparison

Perspectives: Towards a More Consistent Model?

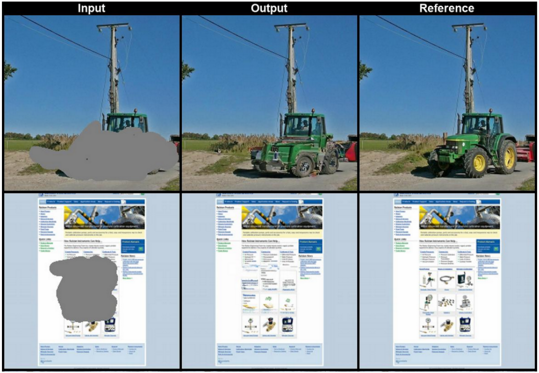

Figure: Palette diffusion model [Saharia et al. 2021] [3]

- **Palette**: U-Net with self attention layers + noised masks in input

Perspectives: Towards a More Consistent Model?

Palette Outputs examples

Figure: Palette samples diversity. (Inputs of the mode on the left)

Figure: GT

References

S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and Locally Consistent Image Completion," *ACM Transactions on Graphics (Proc. of SIGGRAPH 2017)*, vol. 36, no. 4, p. 107, 2017.

A. Newson, A. Almansa, Y. Gousseau, and P. Pérez, "Non-Local Patch-Based Image Inpainting," *Image Processing On Line*, vol. 7, pp. 373–385, 2017.
https://doi.org/10.5201/ipol.2017.189.

C. Saharia, W. Chan, H. Chang, C. A. Lee, J. Ho, T. Salimans, D. J. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," *arXiv preprint arXiv:2111.05826*, 2021.