Semantic Image Inpainting with Deep Generative Models

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Raymond A. Yeh* Chen Chen* Teck Yian Lim Alexander G. Schwing Mark Hasegawa-Johnson Minh N. Do

Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign

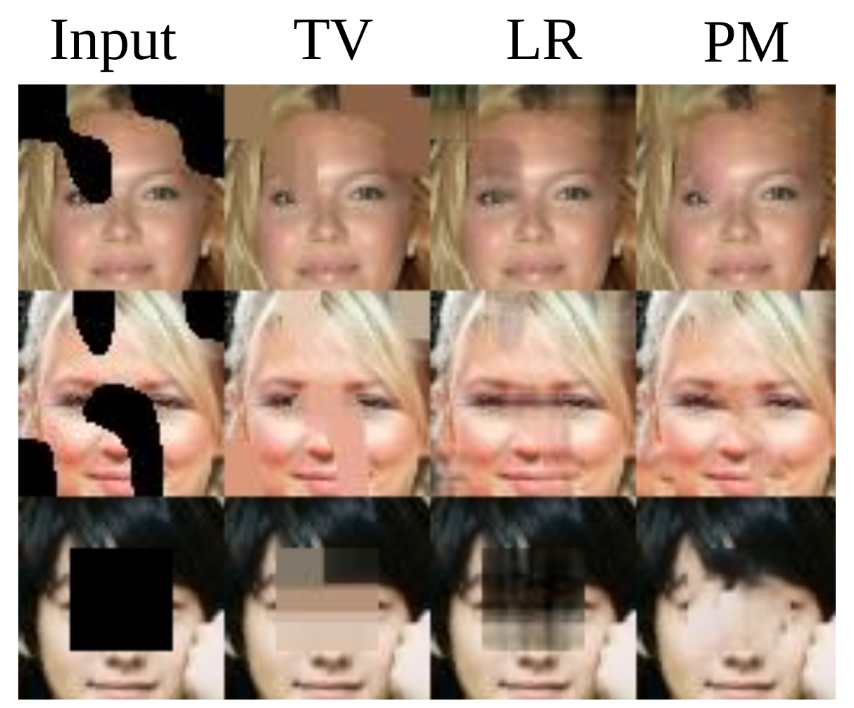
* indicating equal contribution.

CVPR July 21-26 HONOLULU 2017

Motivation

Task: Semantic image inpainting (filling large missing regions)

- ill-posed task
- requires strong prior knowledge on the data
- extracting information from only a single image produces unsatisfactory results



Contributions:

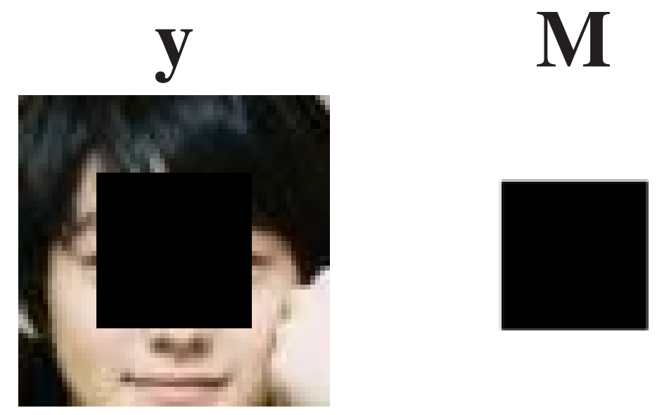
- deep generative models produce missing content by conditioning on available data
- inpainting as constrained optimization problem using **context and prior loss**

Introduction

Problem Formulation:

- Corrupted image: $\mathbf{y}$
- Binary mask: $\mathbf{M}$
- Task: predict uncorrupted version $\hat{\mathbf{x}}$

Given:



Predict:

$\hat{\mathbf{x}} =$

Baselines:

- Total Variation and Low Rank assume smoothness in the pixel space
- Context Encoder is a deep model which treats inpainting as a regression problem

Instead of explicitly defining the prior, we utilize deep generative models to capture prior information.

Generative Adversarial Networks:

- Generator G: deep net mapping perturbation $\mathbf{z}$ to artificial sample
- Discriminator D: deep net discriminating between artificial and real sample, $\mathbf{x}$
- Program:

$$\min_G \max_D V(G,D) = \mathbb{E}_{\mathbf{x}\sim p_{data}}[\log(D(\mathbf{x}))] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{Z}}(\mathbf{z})}[\log(1-D(G(\mathbf{z}))]$$

Our Approach

Intuition of our approach:

- Hypothesis: image that is not from p_{data} (*e.g.*, corrupted data) should not lie on the learned encoding manifold; use manifold can be used as a prior
- Instead of working in the pixel space, we recover the encoding $\hat{\mathbf{z}}$ "closest" to the corrupted image while constrained to the manifold

Solving for the "closest" encoding $\hat{\mathbf{z}}$:

$$\hat{\mathbf{z}} = \arg\min_{\mathbf{z}} \mathcal{L}_c(\mathbf{z}|\mathbf{y},\mathbf{M}) + \mathcal{L}_p(\mathbf{z})$$

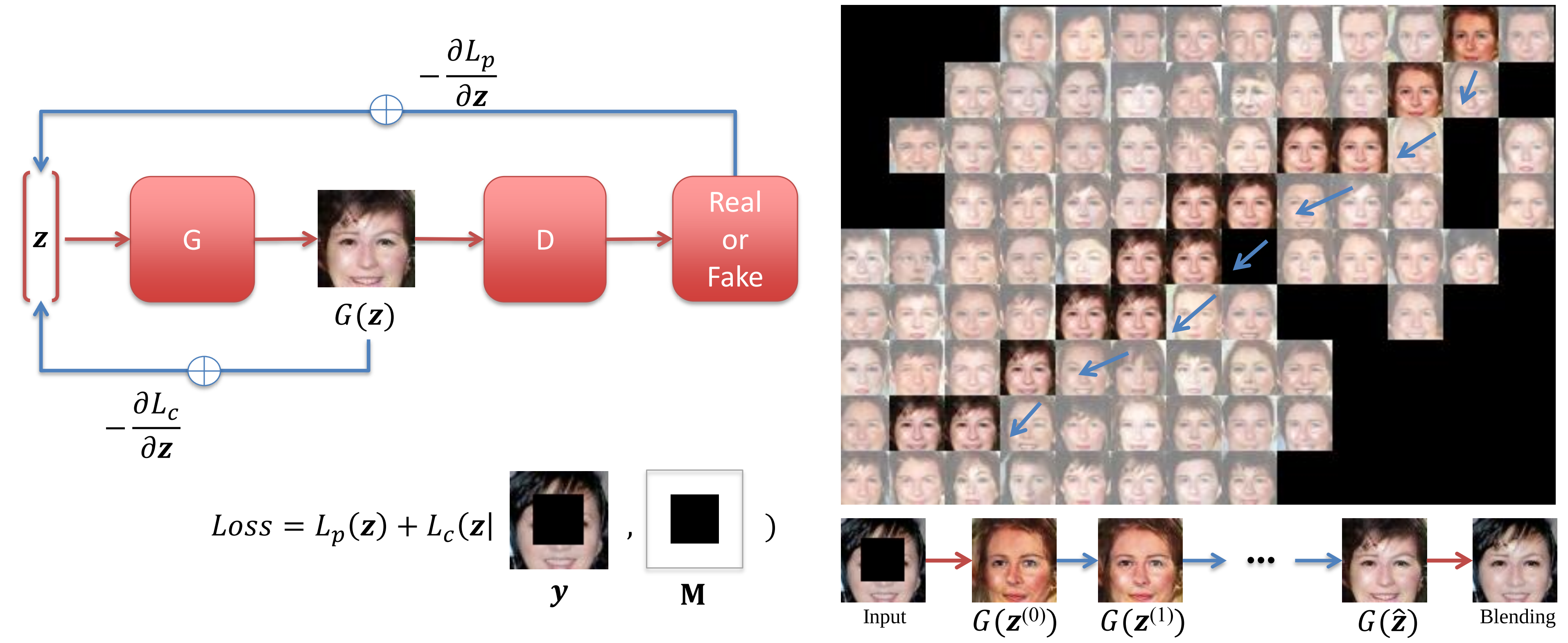
Context Loss: importance weighted metric $\mathbf{W}$ to enforce similarity to the uncorrupted regions:

$$\mathcal{L}_c(\mathbf{z}|\mathbf{y},\mathbf{M}) = \|\mathbf{W}\odot(G(\mathbf{z})-\mathbf{y})\|_1$$

Prior Loss: prior penalizing unrealistic images based on the discriminator:

$$\mathcal{L}_p(\mathbf{z}) = \lambda\log(1-D(G(\mathbf{z})))$$

Illustration of the approach:


$-\frac{\partial L_p}{\partial \mathbf{z}}$
z
G
G(z)
D
Real or Fake
$-\frac{\partial L_c}{\partial \mathbf{z}}$
$Loss = L_p(\mathbf{z}) + L_c(\mathbf{z}|$,)
y
M
Input
$G(\mathbf{z}^{(0)})$
$G(\mathbf{z}^{(1)})$
...
$G(\hat{\mathbf{z}})$
Blending

$$\text{Importance weight metric: } \mathbf{W}_i = \begin{cases} \sum\limits_{j\in N(i)} \frac{(1-\mathbf{M}_j)}{|N(i)|} & \text{if} \quad \mathbf{M}_i \neq 0 \\ 0 & \text{if} \quad \mathbf{M}_i = 0 \end{cases}$$

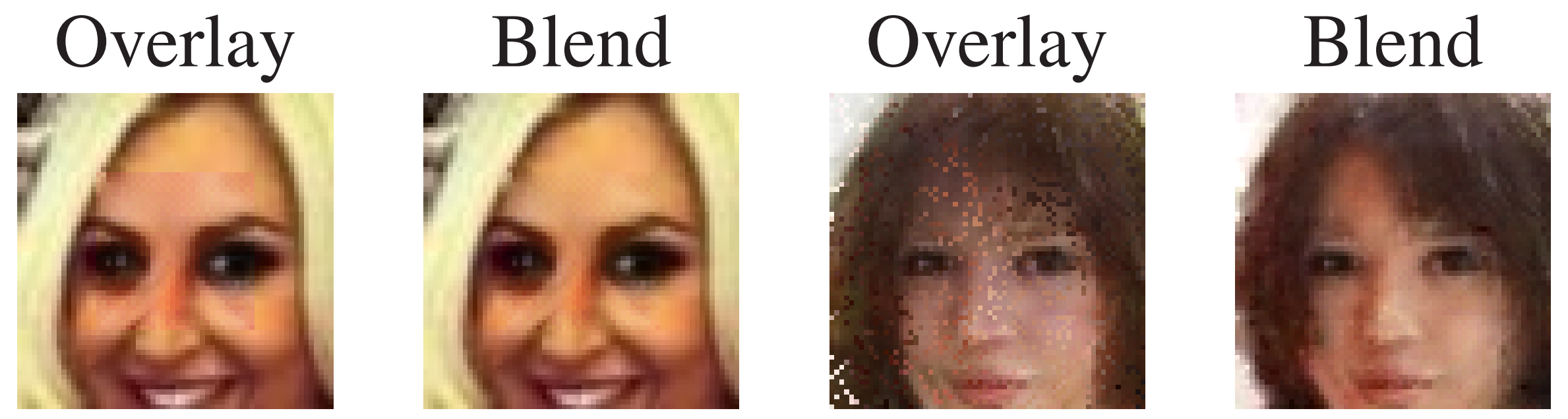
$N(i)$ defines the neighborhood of i

Recovering prediction $\hat{\mathbf{x}}$ via poisson blending rather than simple overlay:

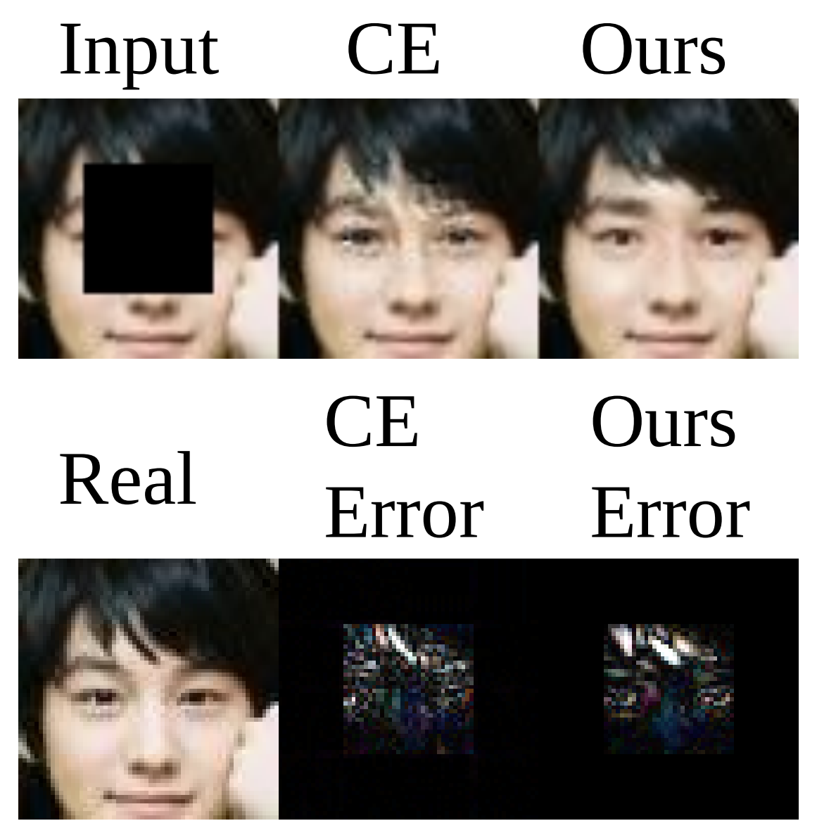
$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\nabla\mathbf{x} - \nabla G(\hat{\mathbf{z}})\|_2^2 \qquad \text{s.t. } \mathbf{x}_i = \mathbf{y}_i \text{ for } \mathbf{M}_i = 1$$

Results

Comparison: Poisson Blending vs. Overlay:



Quantitative Results:



The PSNR values (dB) on the test sets. Left/right results are by Context Encoder (CE)/ours:

Masks/Dataset	CelebA	SVHN	Cars
Center	**21.3**/19.4	**22.3**/19.0	**14.1**/13.5
pattern	**19.2**/17.4	**22.3**/19.8	14.0/**14.1**
random	20.6/**22.8**	24.1/**33.0**	16.1/**18.9**
half	**15.5**/13.7	**19.1**/14.6	**12.6**/11.1

- In the figure above, PSNR for CE is 24.71 dB and ours is 22.98 dB
- Higher PSNR does not mean better visual quality
- The solution is not unique, many hallucinations are reasonable

Qualitative Results: