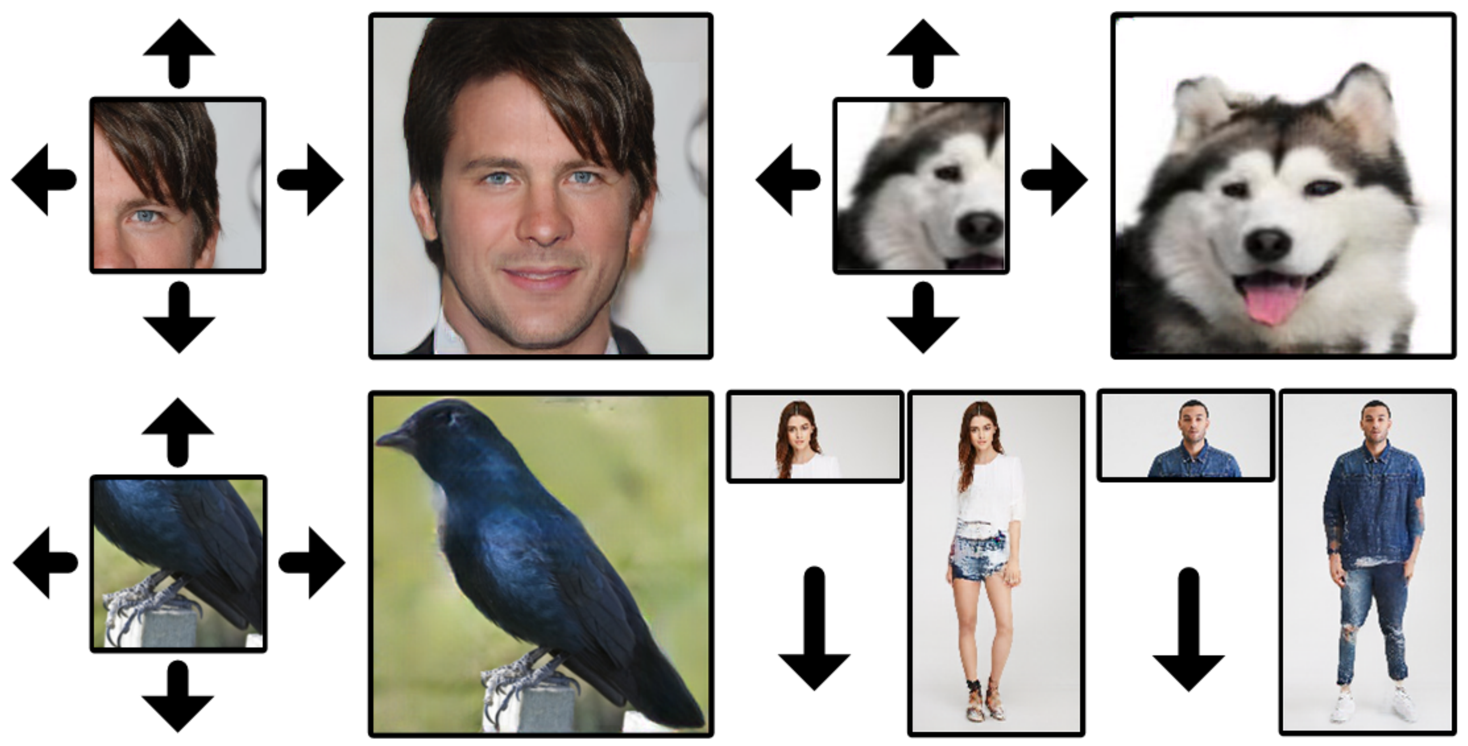
Wide-Context Semantic Image Extrapolation

Yi Wang[1,2], Xin Tao[2], Xiaoyong Shen[2], Jiaya Jia[1,2]
[1] The Chinese University of Hong Kong [2] YouTu Lab, Tencent

Introduction

Target

To infer *unseen* content outside image boundaries, especially *semantically sensitive* and *representative* ones.



Challenges in this context generation task

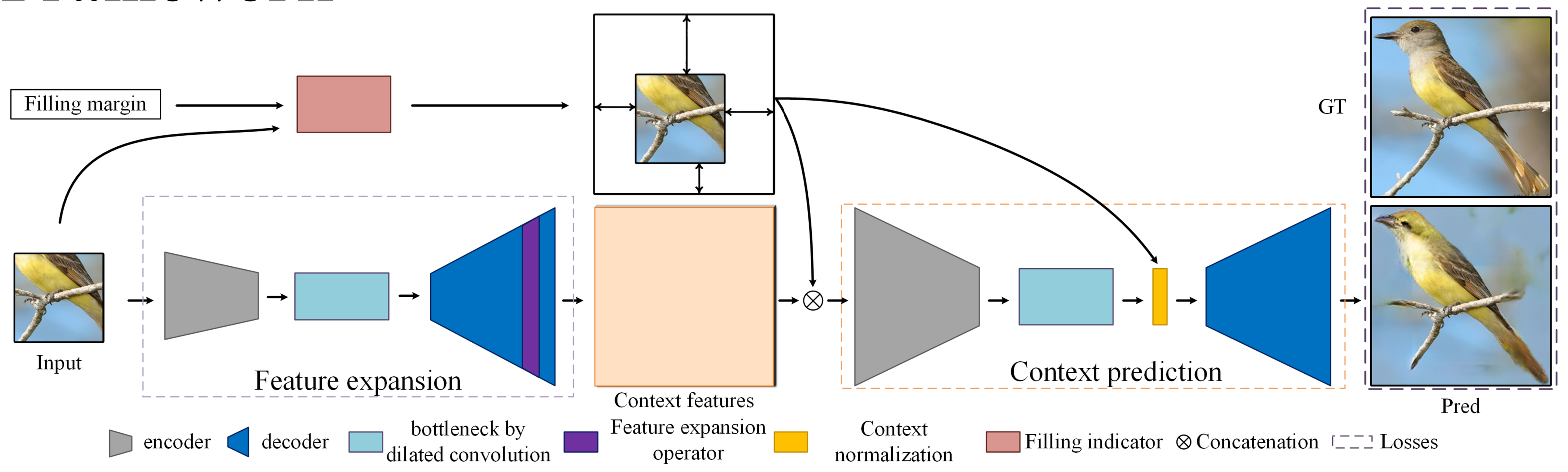
- Image size change: how to *increase image size* beyond boundaries?
  - When / Where to increase image size to target size?
  - Which spatial expansion operator should we use?
- One-sided constraints: the pixels to be predicted away from image border are less constrained than those near border.
  - Potentially accumulating errors / repeated patterns.
  - How to constraint generated contents across *spatial scale*?

Our contribution

- Semantic Regeneration Network (SRN): a deep generative model for image extrapolation.
  - Understanding vastly different context of input incomplete image, and predicting up to *3 times* more unknown pixels than known ones.
  - *Arbitrary-size* semantic generation beyond image boundaries without training multiple models.
  - Various intriguing and important applications.

Our Method

Framework



Our Method

SRN contains two sub-networks: Feature Expansion Network (FEN) and Context Prediction Network (CPN).

- FEN takes small-size images as input and extract features
  - Feature expansion operator: sub-pixel convolution variant

$$s(F)_{i,j,k} = F_{\lfloor i/r_1 \rfloor, \lfloor j/r_2 \rfloor, c' r_2 \cdot mod(i,r_1) + c' \cdot mod(j,r_2) + k}$$

  - It can handle $r_1 \neq r_2$ (body / scene extrapolation)
- CPN decodes these features along with extrapolation indictor into images.
  - Context normalization: maintain style consistency spatially

$$t(f(\mathbf{X}),\rho) = [\rho \cdot n(f(\mathbf{X}_\Omega), f(\mathbf{X}_{\bar{\Omega}})) + (1-\rho) f(\mathbf{X}_\Omega)] \odot \mathbf{M}\downarrow + f(\mathbf{X}_{\bar{\Omega}}) \odot (1 - \mathbf{M}\downarrow)$$

$$n(x_1, x_2) = \frac{x_1 - \mu(x_1)}{\sigma(x_1)} \cdot \sigma(x_2) + \mu(x_2)$$

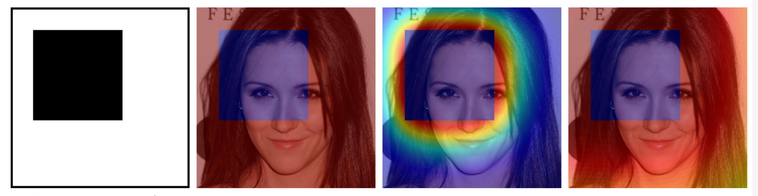
  - Transfer (blend) feature statistics in known areas to unknown ones.

Learning objectives

1. Relative spatial variant loss: incorporate spatial regularization

$$\mathbf{M}_w^i = (g * \overline{\mathbf{M}}^i) \odot \mathbf{M}, \quad \mathbf{M}_w = \mathbf{M}_w^{c-1} / max(\mathbf{M}_w^c, \epsilon)$$

$$\mathcal{L}_s = ||(\mathbf{Y} - G(\mathbf{X}, m; \theta)) \odot \mathbf{M}_w||_1$$



Input mask Use M as Mw Mw in CD [1] Mw in RSV

2. Implicit Diversified MRF loss

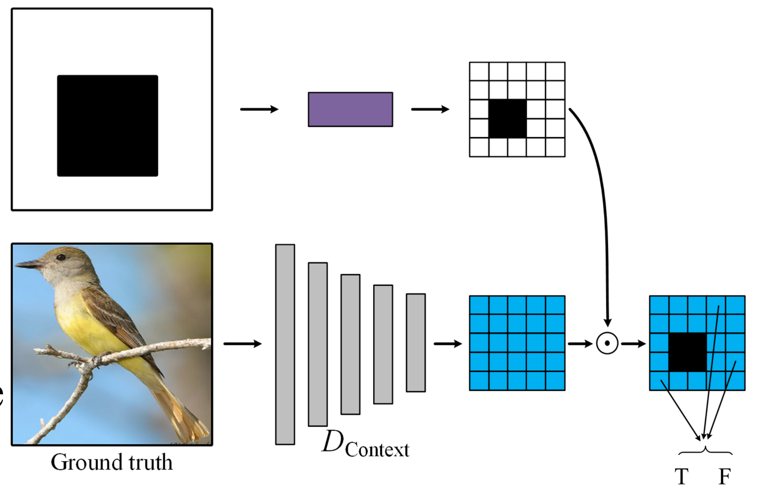
- To create crisp texture by bringing close feature distributions between the generated image and its corresponding ground truth.

3. Context adversarial loss

$$D_{context}(\hat{\mathbf{Y}}) = \frac{\sum_{p \in P(\hat{\mathbf{Y}})} p}{\sum_{q \in \mathbf{M}\downarrow} q},$$

$$w.r.t. \quad P(\hat{\mathbf{Y}}) = d_{context}(\hat{\mathbf{Y}}) \odot \mathbf{M}\downarrow,$$

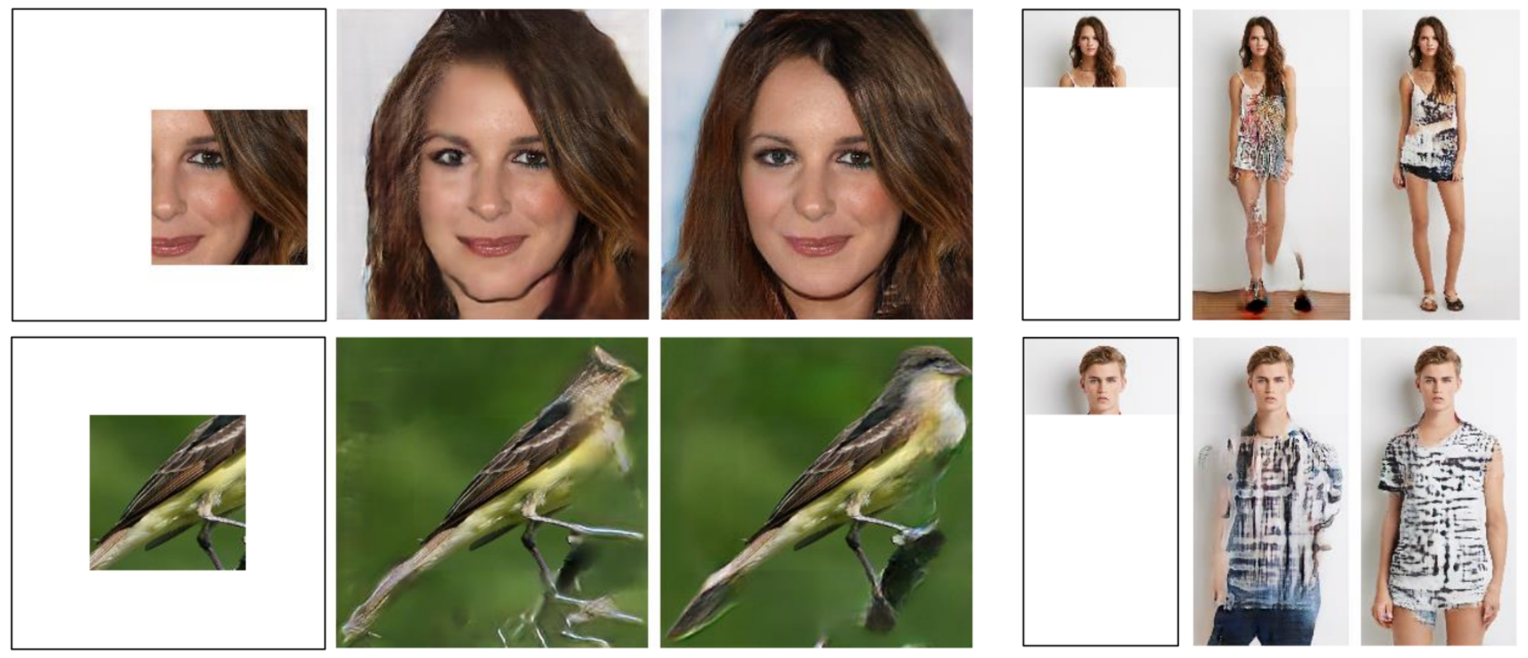
$$\mathcal{L}_{adv}^n = -E_{\mathbf{X} \sim \mathbb{P}_\mathbf{X}}[D_n(G(\mathbf{X};\theta))] + \lambda_{gp} E_{\hat{\mathbf{X}} \sim \mathbb{P}_{\hat{\mathbf{X}}}}[(||\nabla_{\hat{\mathbf{X}}} D_n(\hat{\mathbf{X}}) \odot \mathbf{M}_w||_2 - 1)^2]$$



- Aggregate local regions into a single probability

Experiments

Qualitative Evaluation



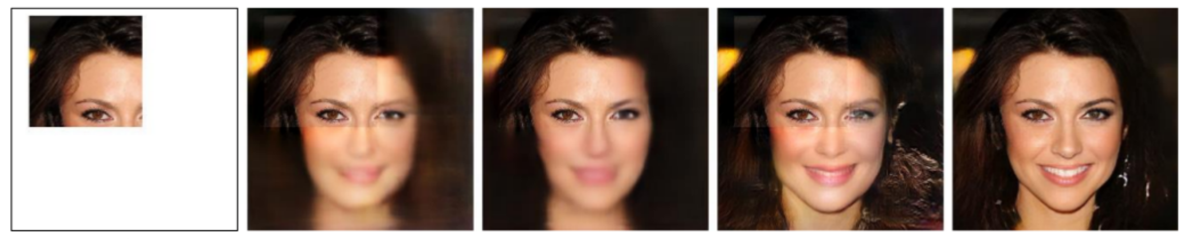
Input images Results of CA[2] Our results Input images Results of CA[2] Our results

[1] Wang, Yi, et al. "Image Inpainting via Generative Multi-column Convolutional Neural Networks." *Advances in Neural Information Processing Systems*. 2018.
[2] Yu, Jiahui, et al. "Generative image inpainting with contextual attention." *IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

Results

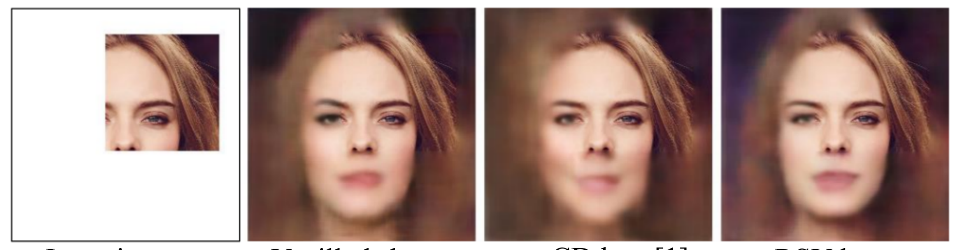
The effectiveness of CN and RSV

- CN



Input images w/o CN in pre-train w/ CN in pre-train w/o CN w/ CN

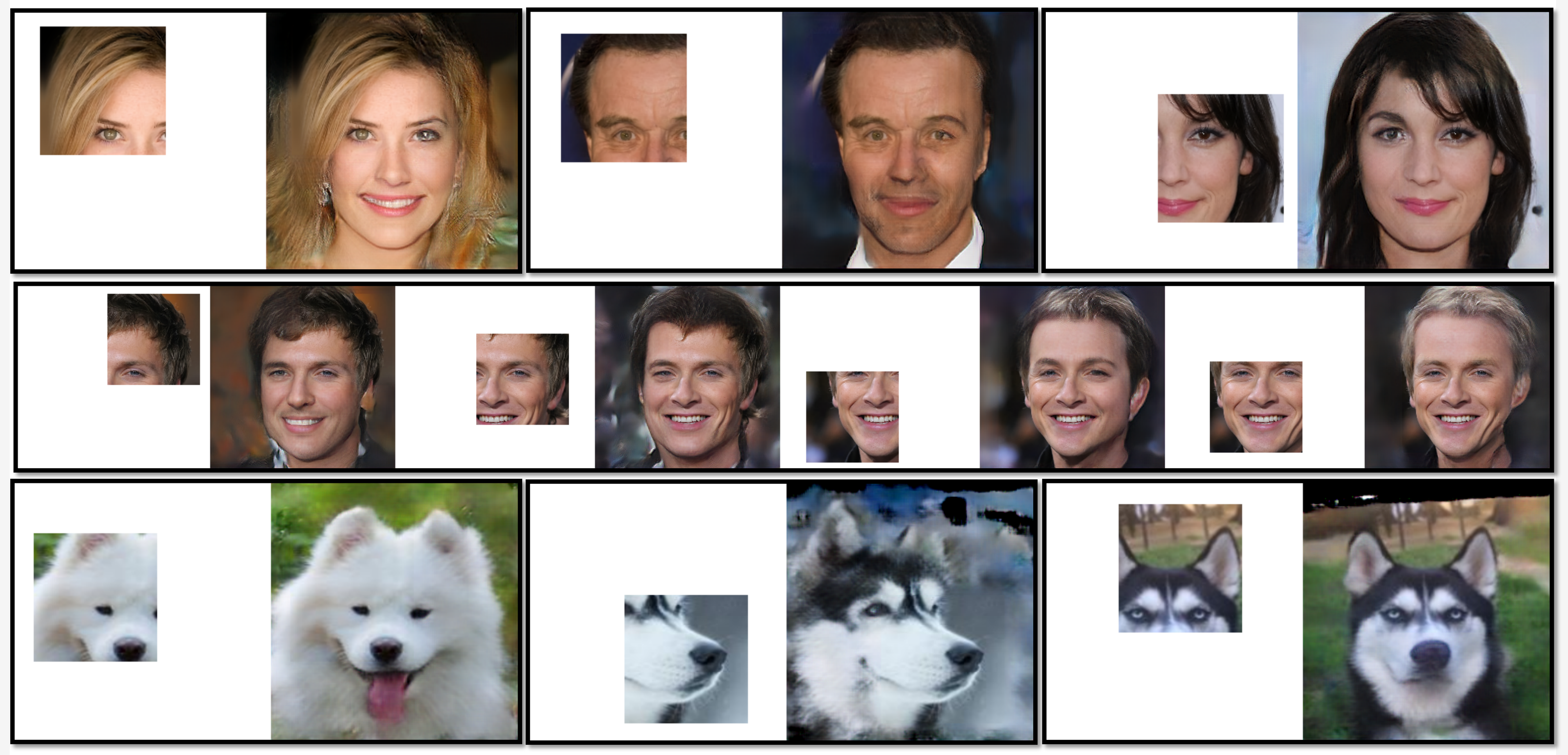
- RSV



Input images Vanilla l_1 loss CD loss [1] RSV loss

More Results

- **Faces**



- **Bodies**

- **Scenes**