Prior Guided GAN Based Semantic Inpainting

CVPR JUNE 14-19 2020

Avisek Lahiri[1*], Arnav Kumar Jain[2*], Sanskar Agrawal[1], Pabitra Mitra[1], Prabir Kumar Biswas[1]

[1]Indian Institute of Technology Kharagpur, [2]Microsoft

Motivation

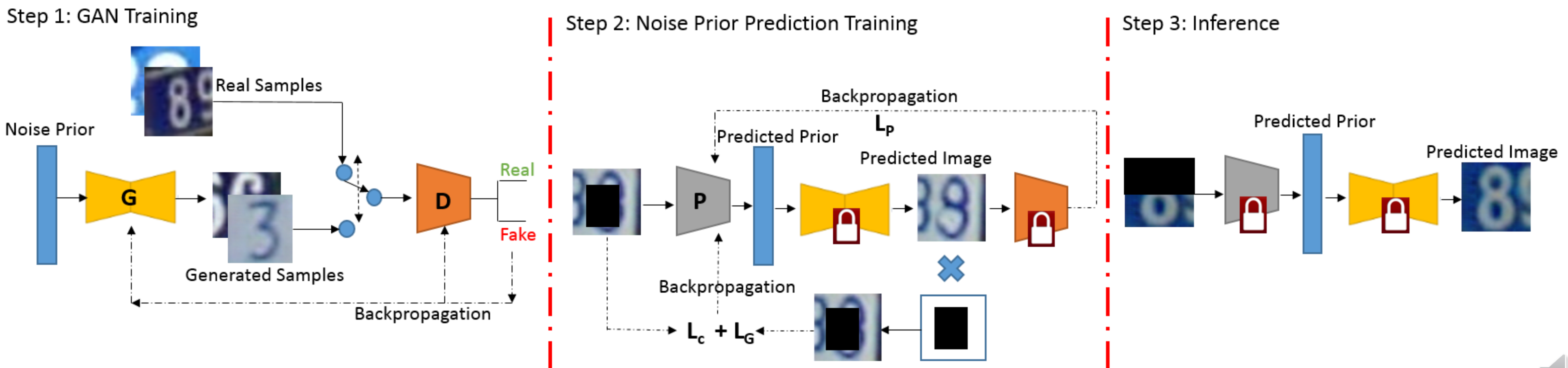
- Recent methods train a single feed-forward network over the masked images
- Another approach is to find the ‘*best-matching*’ latent vector by using a pre-trained generative model*
- High inference time due to iterative optimization and difficulty in scaling to higher resolutions
- Learned a data driven parametric network to directly predict a matching latent prior for a given input
- Regularized the network with structural prior for better preservation of pose and size of the objects
- Leveraged recent high resolution GAN models to scale our inpainting network to 256×256
- Extended our model for sequence reconstruction, using a recurrent net based grouped latent prior learning

** Yeh et al.* "Semantic Image Inpainting with Deep Generative Models*“, CVPR. 2017.*

Noise Prior Prediction Network

<u>Aim: Learn to predict a “good” z vector from just unmasked pixels</u>

- ❑ Step 1: Independent training of GAN (can be any generative model !!!)
- ❑ Step 2: Learn to predict noise prior conditioned on masked image
- ❑ Step 3: Pass the predicted prior through the generator of pre-trained GAN



Structural Prior guided Training

Problem Setup

- ❑ Have structural priors to regularize GAN outputs
- ❑ State-of-the-art landmark detection models fail on masked images

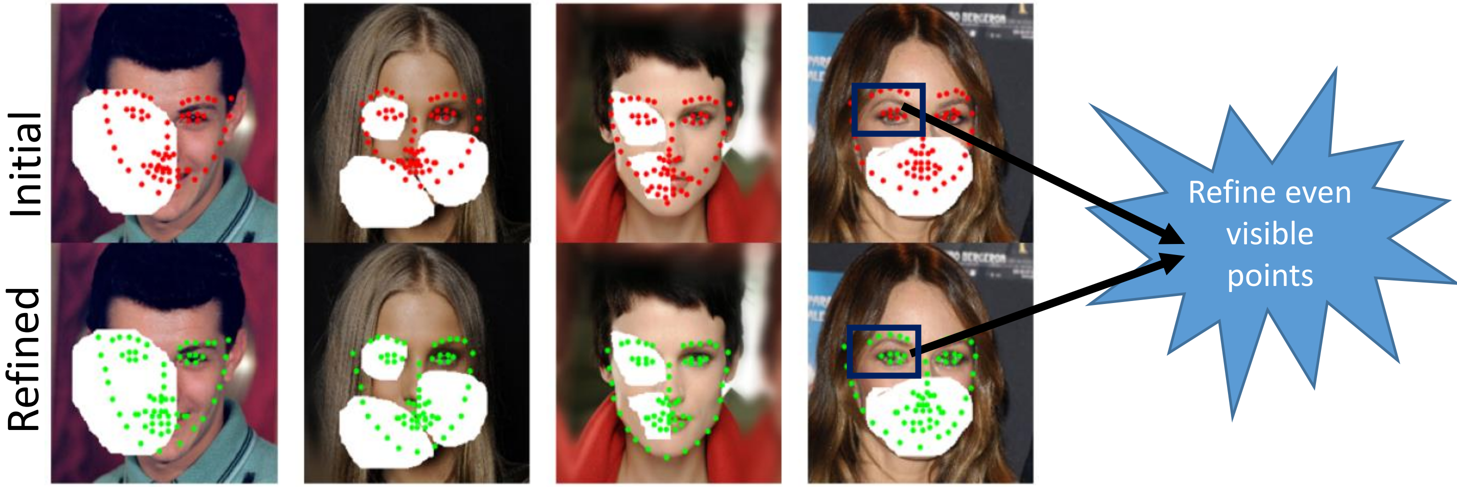
Input:

a) Predicted set $S = \{(x_{pred}, y_{pred})\}^{68\times2}$

b) Target set $\mathrm{T} = \{(x_{actual}, y_{actual})\}^{68\times2}$

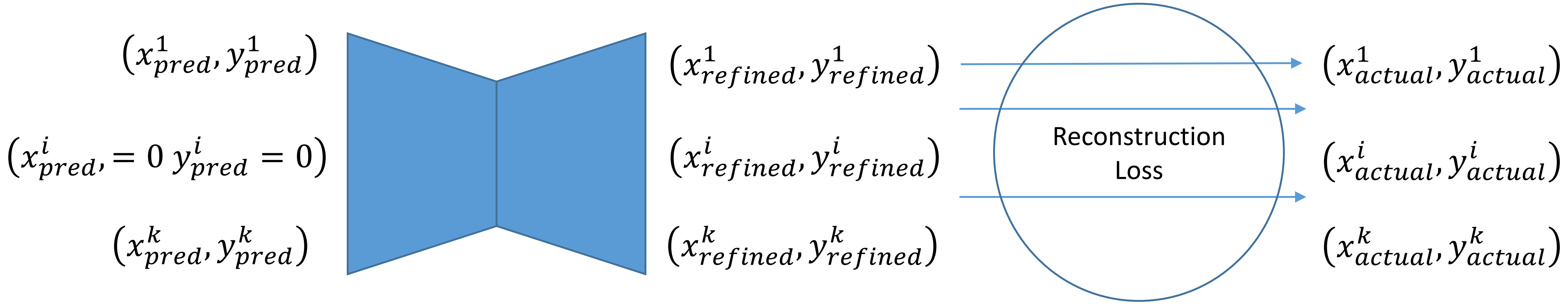
Output:

Refined Set $\mathrm{R} = \{(x_{refined}, y_{refined})\}^{68\times2}$



Refine even visible points

Learning with AutoEncoder Framework



(x^1_{pred}, y^1_{pred})

$(x^i_{pred}, = 0\ y^i_{pred} = 0)$

(x^k_{pred}, y^k_{pred})

$(x^1_{refined}, y^1_{refined})$

$(x^i_{refined}, y^i_{refined})$

$(x^k_{refined}, y^k_{refined})$

Reconstruction Loss

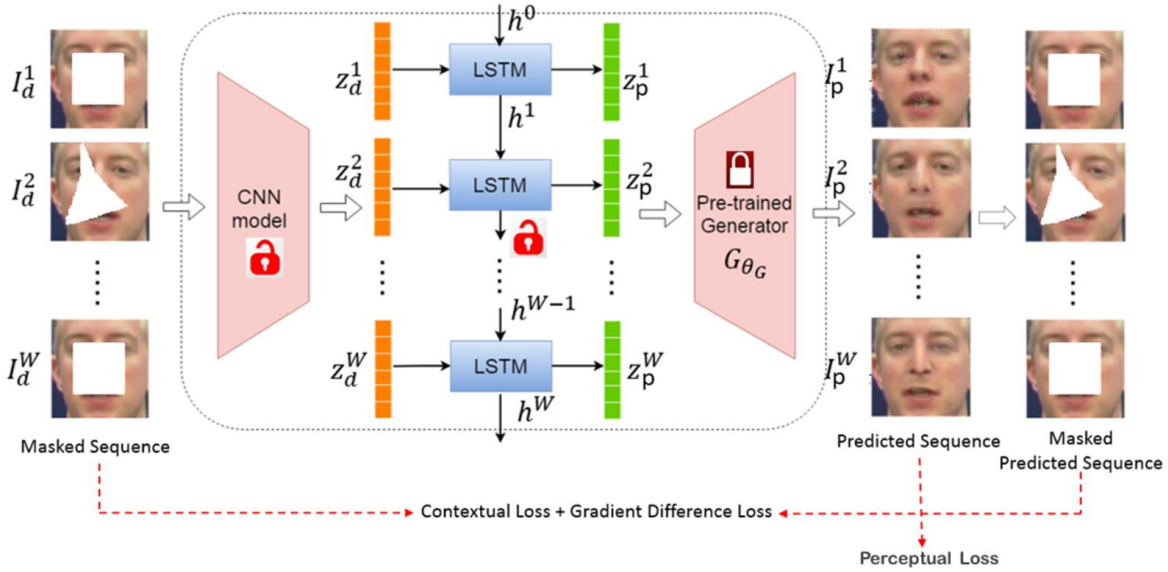
$(x^1_{actual}, y^1_{actual})$

$(x^i_{actual}, y^i_{actual})$

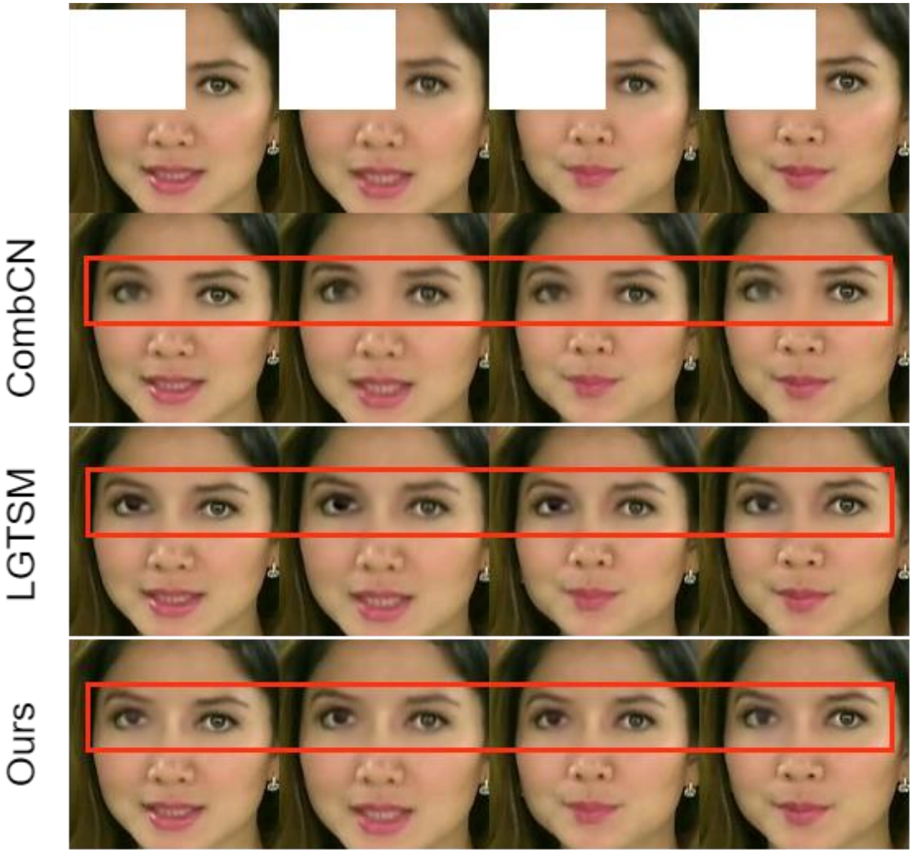
$(x^k_{actual}, y^k_{actual})$

Grouped Prior for Video Inpainting

- ❑ For videos, we need both static picture quality and temporal coherence
- ❑ Independent prediction of z on each frame can leads for temporal jittering
- ❑ Can we learn a group of z vectors together ?



Grouped prior prediction framework for video inpainting



Results: Improvements over iterative Baseline

We convert the iterative framework to a single pass inference model

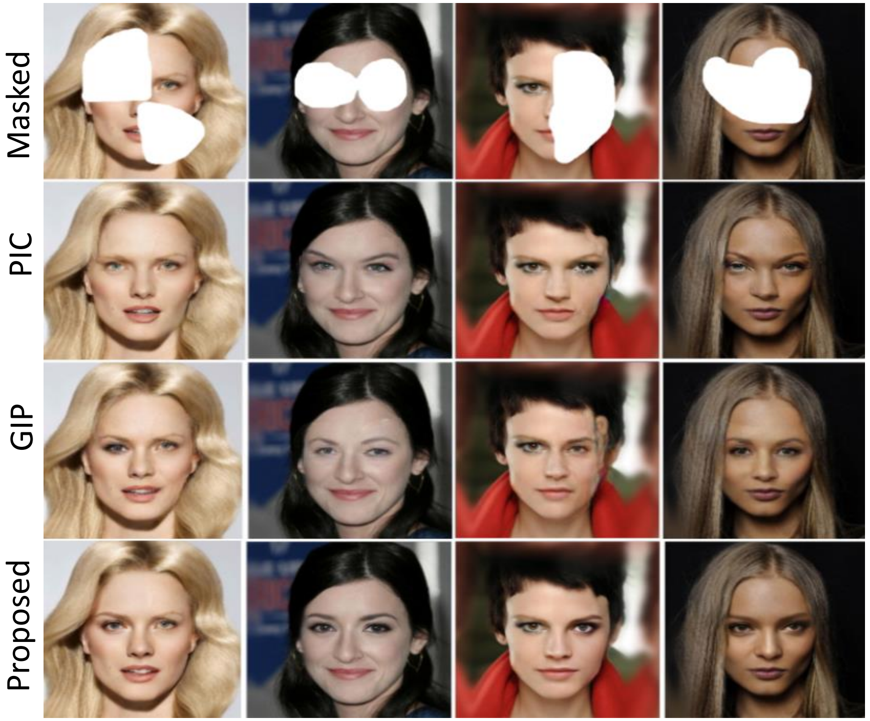
❑ Single pass through our network is the final output

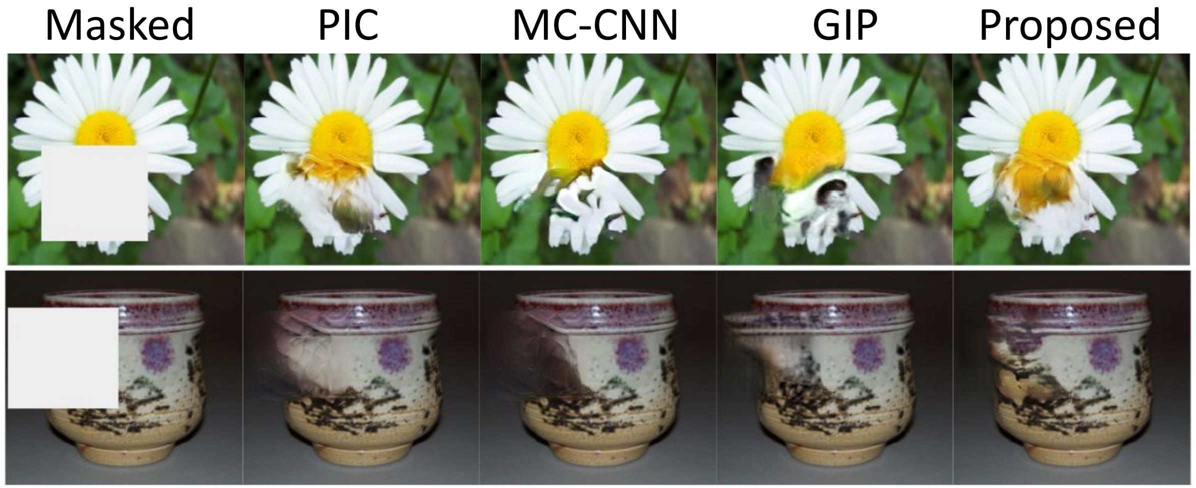
❑ Single pass through Yeh et al.* is far from acceptable quality (requires 1000-1500 iterations)

Resolution	Yeh et al.	Ours : M_z	Ours: M_{z+S}
64X64	2175	2.7	2.8
128X128	10750	11.0	13.2

Inference time (milli-seconds) for inpainting at different resolutions

Approx. 800X Speedup





* *Yeh et al.* "Semantic Image Inpainting with Deep Generative Models", *CVPR. 2017.*