2021 ICCV OCTOBER 11-17 VIRTUAL

PixelPyramids: Exact Inference Models from Lossless Image Pyramids

Shweta Mahajan[1] Stefan Roth[1,2]

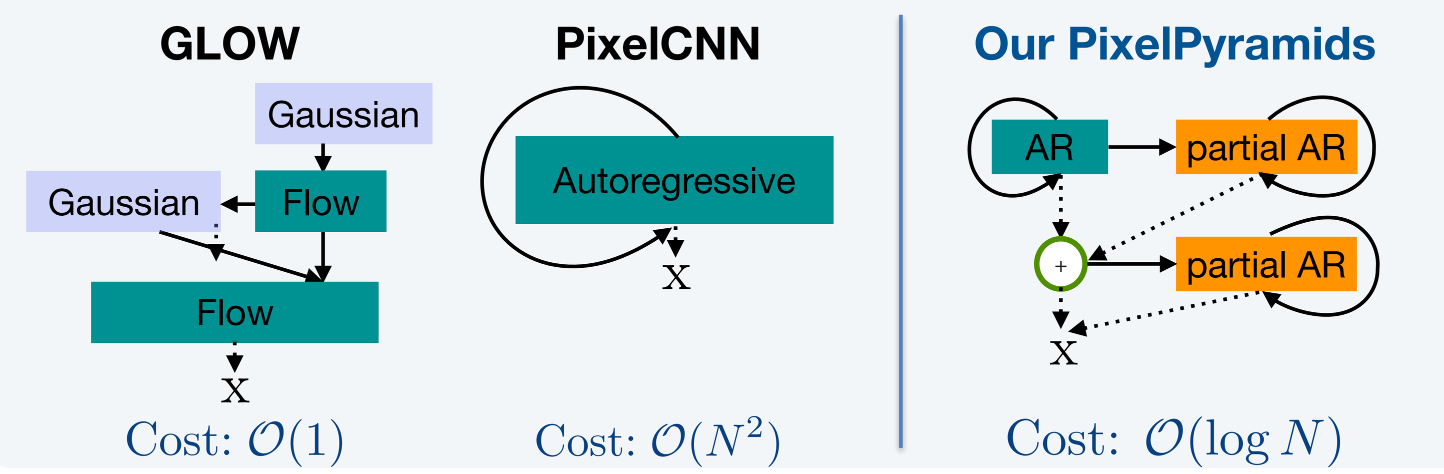
[1]TU Darmstadt [2] hessian.AI

TECHNISCHE UNIVERSITÄT DARMSTADT

hessian.AI

Introduction

- **Goal:** Improve the representational power of exact inference models for images
- Previous work:
  - *Split coupling normalizing flows* — easy to parallelize, but limited modeling capacity
  - *Autoregressive models* — capture complex dependencies, but inefficient sequential sampling



Computationally efficient partial autoregressive structure

Multi-scale representation with Paired Pyramids

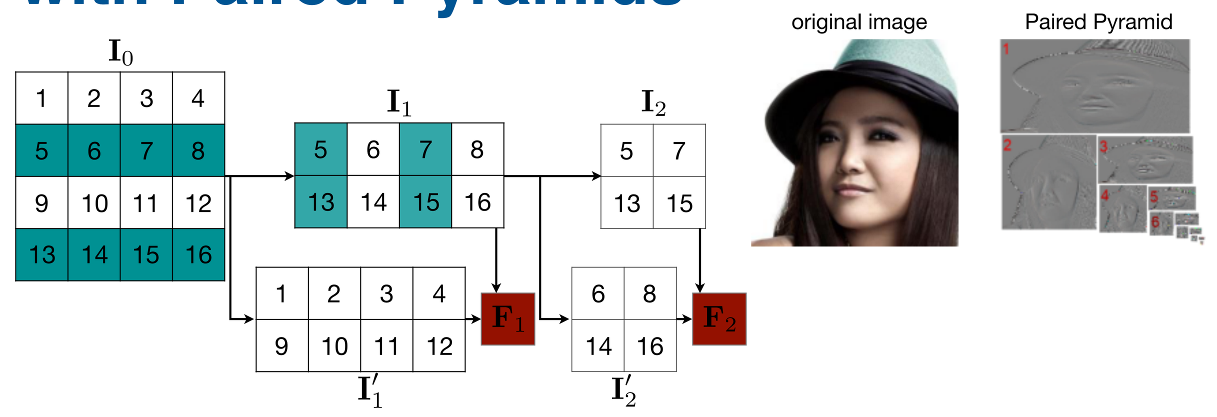
- Fine component at level i

$$\mathbf{F}_i = (\mathbf{I}'_i - \mathbf{I}_i) \bmod 2^b$$

$$(x-y) \bmod K = \begin{cases} (x-y) & \text{if } x \geq y \\ K-(y-x) & \text{otherwise} \end{cases}$$

- Reconstruction of $\mathbf{I}'_i$

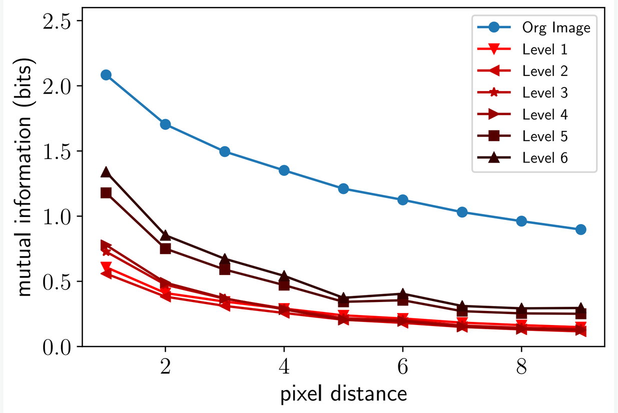
$$\mathbf{I}'_i = (\mathbf{F}_i + \mathbf{I}_i) \bmod 2^b$$

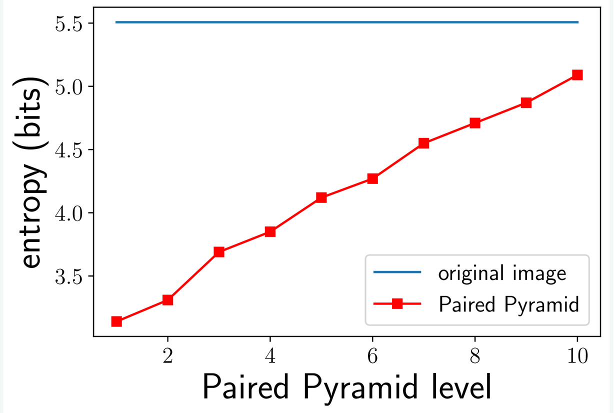


Representational properties:
- Same quantization level as the original image
- Same number of pixels as the original image

- Original image: $\mathbf{I}_0$
- Downsampled image at layer i: $\mathbf{I}_i$
- Residual image at layer i: $\mathbf{I}'_i$
- Number of bits encoding original image: b

Statistical properties:
- Less pixel-to-pixel correlation
- Low entropy





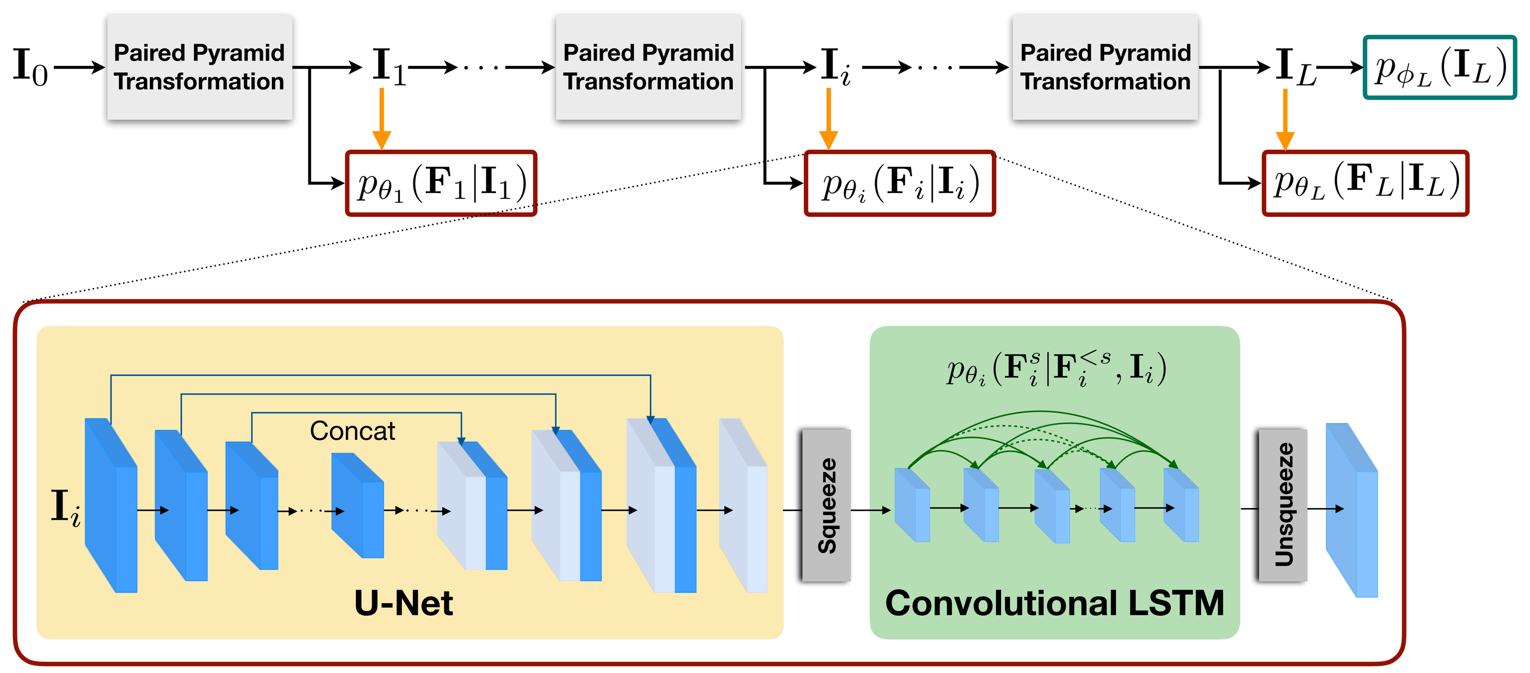
Paired Pyramids: invertible, lossless image coding scheme with localized dependencies

[Paired Pyramids] H. H. Torbey and H. E. Meadows. System for lossless digital image coding/decoding. In Visual Communications and Image Processing IV, volume 1199, pages 989–1002. International Society for Optics and Photonics, SPIE, 1989.
[PixelCNN++] Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In ICML, 2016.
[Glow] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In NeurIPS 2018.
[SPN] Jacob Menick and Nal Kalchbrenner. Generating high fidelity images with subscale pixel networks and multidimensional upscaling. In ICLR, 2019.
[MaCoW] Xuezhe Ma, Xiang Kong, Shanghang Zhang, and Eduard H.Hovy. MaCow: Masked convolutional generative flow. In NeurIPS, 2019.
[WaveletFlow] Jason J. Yu, Konstantinos G. Derpanis, and Marcus A. Brubaker. Wavelet Flow: Fast training of high resolution normalizing flows. In NeurIPS, 2020.

This work has been supported by the German Research Foundation as part of the Research Training Group *Adaptive Preparation of Information from Heterogeneous Sources (AIPHES)* under grant No. GRK 1994/1.
Further, this project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 866008).

erc AIPHES

PixelPyramids architecture



- The joint distribution over the pixels factorizes as

$$p_{\theta,\phi_L}(\mathbf{I}_0) = p_{\phi_L}(\mathbf{I}_L) \prod_{i=1}^{L} p_{\theta_i}(\mathbf{F}_i \mid \mathbf{I}_i)$$

- The conditional distribution over the fine components decomposes as

$$p_{\theta_i}(\mathbf{F}_i \mid \mathbf{I}_i) = p_{\theta_i}(\mathbf{F}_i^1 \mid \mathbf{I}_i) \prod_{s=2}^{S_i} p_{\theta_i}(\mathbf{F}_i^s \mid \mathbf{F}_i^{<s}, \mathbf{I}_i)$$

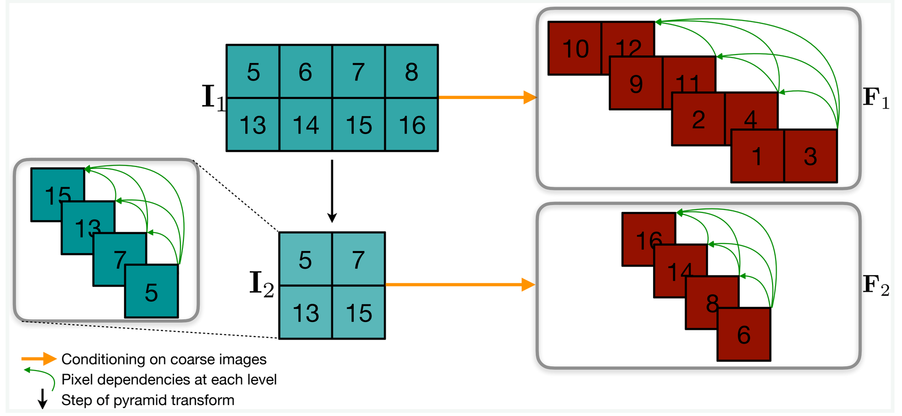
- Number of levels: L
- Number of subsampled images at layer i: S_i

U-Net provides global context from the conditioning pixels for encoding scale-specific image details

A conv-LSTM models long-range sequential dependencies between the subsampled images

PixelPyramids afford a sparser dependency structure to encode the joint distribution of image pixels

Sampling and computational cost



Method	Parameters	Training	Sampling
PixelCNN++ [Salimans et. al. 2017]	70 M	0.135	6 x10^6
SPN [Menick et. al. 2019]	**50 M**	—	—
Glow [Kingma et. al. 2018]	171 M	1.79	531
MaCoW [Ma et. al. 2019]	177 M	—	434
WaveletFlow [Yu et. al. 2020]	52 M	**0.0147**	998
PixelPyramids	166 M	0.0198	**70**

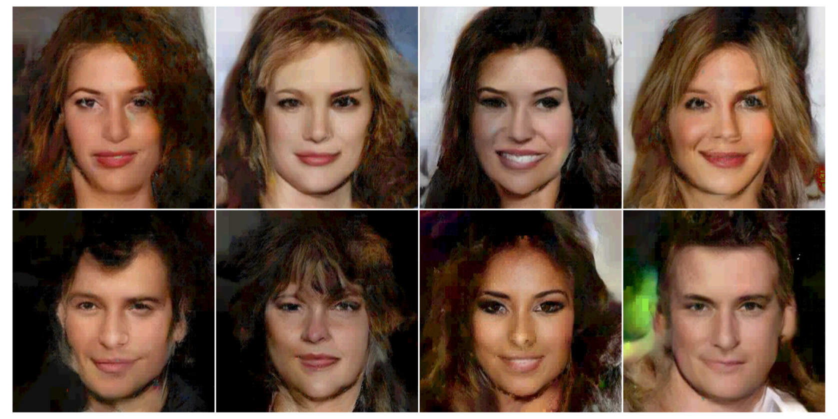
Training and sampling speeds (ms/image) on CelebA-HQ (256x 256)

Let the sampled image $\mathbf{x}$ be of resolution $[C, N, N]$, then the worst-case number of sampling steps required by PixelPyramids is T = $\mathcal{O}(\log N)$

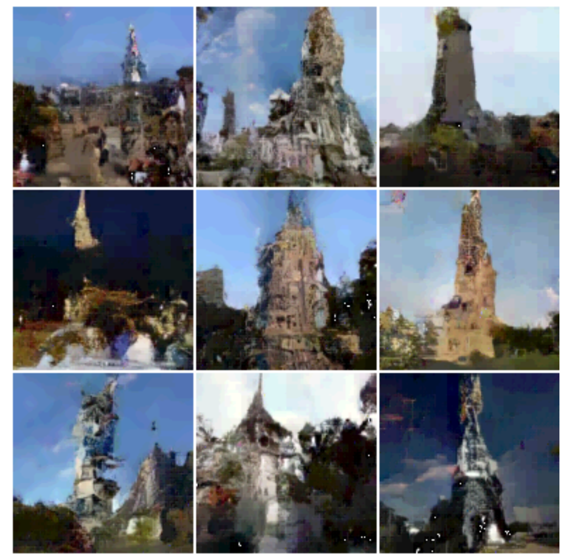
Experiments: Density estimation and image synthesis

Evaluation (in bits/dim) on 5-bit CelebA-HQ (256 x 256) and LSUN (128 x 128)

		LSUN		
Method	CelebA-HQ	Bedroom	Church	Tower
SPN [Menick et. al. 2019]	**0.61**	—	—	—
Glow [Kingma et. al. 2018]	1.03	1.20	—	—
MaCoW [Ma et. al. 2019]	0.67	0.98	1.09	1.02
WaveletFlow [Yu et. al. 2020]	0.94	—	—	—
PixelPyramids	**0.61**	**0.88**	**1.07**	**0.95**


CelebA-HQ (256 x 256)


Bedroom


Tower


Church

State-of-the-art results for density estimation on various datasets

Evaluation on 8-bit ImageNet (128 x 128)

Method	bits/dim
ParallelWavelet [Reed et. al. 2020]	3.55
SPN [Menick et. al. 2019]	**3.08**
PixelPyramids	3.40


ImageNet (128 x 128)

Evaluation on 8-bit CelebA-HQ (1024 x 1024)

Method	bits/dim
WaveletFlow [Yu et. al. 2020]	1.34
PixelPyramids	**0.58**


CelebA-HQ (1024 x 1024)

Code

PixelPyramids generate high-quality and diverse samples of high resolution