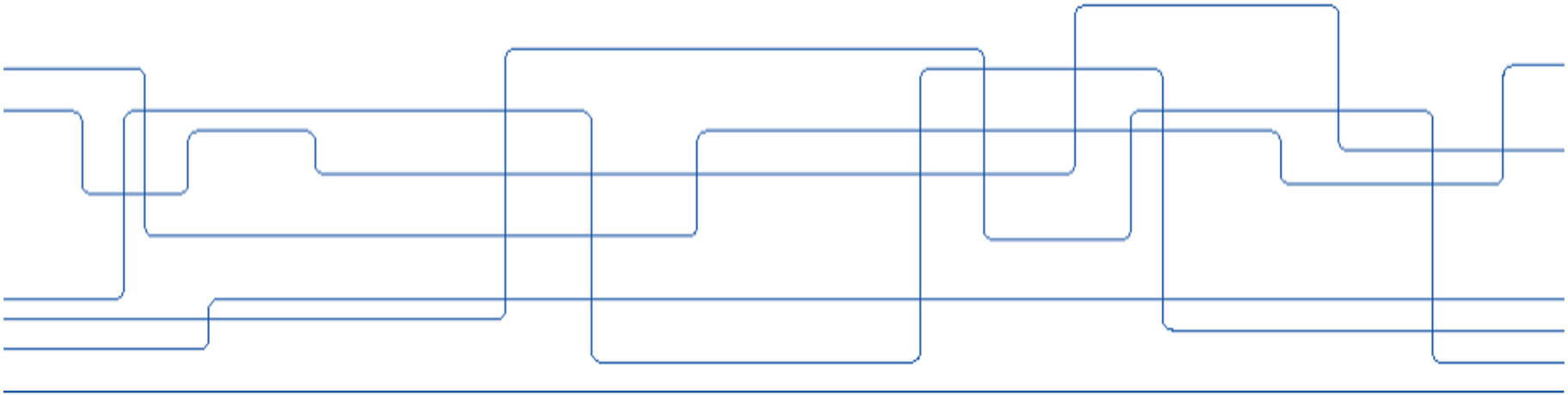
KTH ROYAL INSTITUTE OF TECHNOLOGY

Full-Glow: Fully conditional Glow for more realistic image generation

Moein Sorkhei, Gustav Eje Henter, Hedvig Kjellström



Motivation

Synthesizing images of street scenes e.g. for training autonomous cars:

- Approach 1: Using graphic engines
  - Requires a lot of hand-engineering of features
  - Sometimes the synthesized images are not natural
- Approach 2: Using generative models
  - Ideally they could synthesize diverse images that resemble natural images
  - Image-to-image translation is commonly used to synthesize better target images

Popular generative models:

- GANs are extensively used thanks to their ability in synthesizing sharp images
- Flow-based models
  - Flexible probabilistic models with exact likelihood evaluation thanks to invertibility
  - After the publication of Glow, they have gained more attraction

Kingma, Diederik P., and Prafulla Dhariwal. "Glow: Generative flow with invertible 1x1 convolutions." arXiv preprint arXiv:1807.03039 (2018).

Contributions

Prior work using Glow for image-to-image translation:

- C-Glow: Makes the Glow layers conditional for image generation
- DUAL-Glow: Uses two Glows for medical image modality transfer, where the learned latent distribution of the target Glow is conditional
- C-Flow: Uses two Glows and makes all the coupling layers conditional

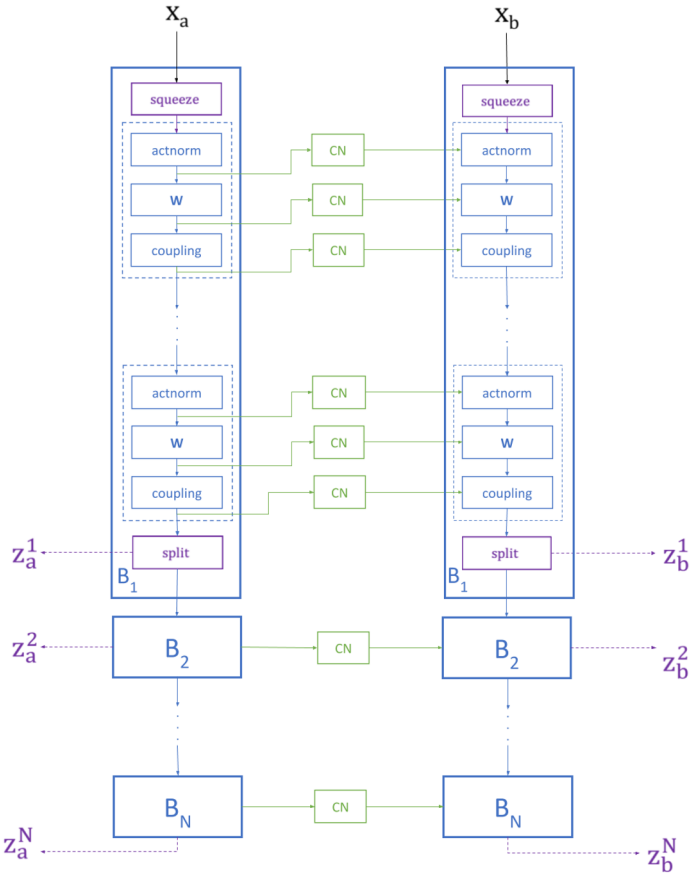
In our work, we build on the previous works by using two Glows and applying conditioning networks to all layers of the network, resulting in a fully conditional model.

- Lu, You, and Bert Huang. "Structured Output Learning with Conditional Generative Flows." *AAAI*. 2020.
- Sun, Haoliang, et al. "Dual-glow: Conditional flow-based generative model for modality transfer." *Proceedings of the IEEE International Conference on Computer Vision*. 2019.
- Pumarola, Albert, et al. "C-flow: Conditional generative flow models for images and 3d point clouds." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

Method

Full-Glow is a fully conditional architecture

$$\frac{1}{N}\left[-\sum_{n=1}^{N}\lambda \log p_{\boldsymbol{\theta}}\left(\mathbf{x}_a^{(n)}\right)-\sum_{n=1}^{N}\log p_{\boldsymbol{\phi}}\left(\mathbf{x}_b^{(n)}\,\middle|\,\mathbf{x}_a^{(n)}\right)\right]$$



Setup of experiments

- Baseline models:
  - C-Glow: We consider two versions/configurations
    - Version 1: with deeper conditioning networks
    - Version 2: with deeper Glow (more flow blocks)
  - DUAL-Glow
  - Pix2pix: Conditional GAN for image-to-image translation
- We used the Cityscapes dataset
  - Natural street-scene images (RGB) with segmentation masks

- Lu, You, and Bert Huang. "Structured Output Learning with Conditional Generative Flows." *AAAI*. 2020.
- Sun, Haoliang, et al. "Dual-glow: Conditional flow-based generative model for modality transfer." *Proceedings of the IEEE International Conference on Computer Vision*. 2019.
- Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
- Cordts, Marius, et al. "The cityscapes dataset for semantic urban scene understanding." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

Quantitative comparison

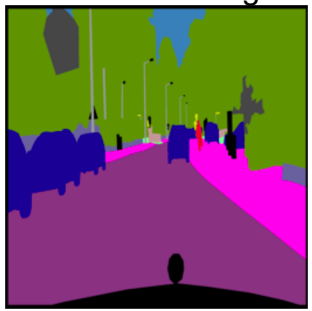
- Bits per dimension (BPD) for likelihood-based models
- For each model, we also resynthesized the validation set three times, conditioned on the segmentation maks
  - PSPNet segmentations of the generated images were then evaluated against the ground-truth segmentations in three ways

Model	Cond. BPD	Mean pixel acc.	Mean class acc.	Mean class IoU
C-Glow v.1 [23]	2.568	35.02 ± 0.56	12.15 ± 0.05	7.33 ± 0.09
C-Glow v.2 [23]	2.363	52.33 ± 0.46	17.37 ± 0.21	12.31 ± 0.24
Dual-Glow [36]	2.585	71.44 ± 0.03	23.91 ± 0.19	18.96 ± 0.17
pix2pix [15]	—	60.56 ± 0.11	22.64 ± 0.21	16.42 ± 0.06
Our model	**2.345**	**73.50 ± 0.13**	**29.13 ± 0.39**	**23.86 ± 0.30**
Ground-truth	—	*95.97*	*84.31*	*77.30*

Zhao, Hengshuang, et al. "Pyramid scene parsing network." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

Visual comparison of model outputs

Conditioning



Ground truth

C-Glow v.1

C-Glow v.2

DUAL-Glow

pix2pix

Full-Glow sample 1

Full-Glow sample 2

Visual comparison of model outputs

Conditioning

Ground truth

C-Glow v.1

C-Glow v.2

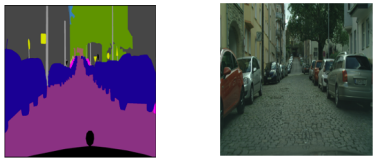
DUAL-Glow

pix2pix

Full-Glow sample 1

Full-Glow sample 2

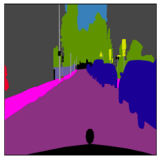
Using the model for content transfer



Desired content along with its segmentation

Invertible model

Extracted z



New structure

Invertible model

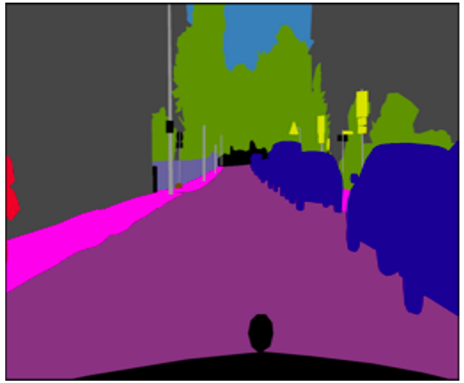
New image

Extracted z

Content transfer example

Desired content

Desired structure



Content applied to structure

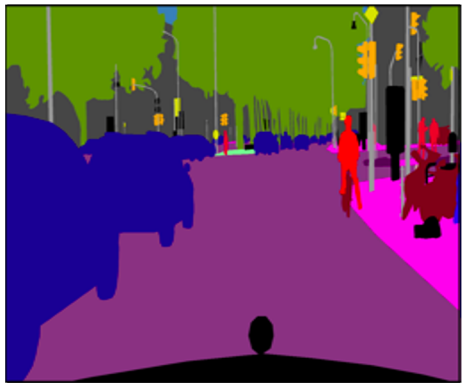
Notice the difference in walls, asphalt, and the cars compared to ground-truth image

Ground-truth image for the given structure

Content transfer example

Desired content

Desired structure



Content applied to structure

Notice the difference in walls, asphalt, and the cars compared to ground-truth image

Ground-truth image for the given structure

Summary

- We proposed a fully conditional Glow-based architecture for more realistic conditional street-scene image generation
- Our improved conditioning allows for generating images that are more interpretable by a semantic classifier
- We synthesized higher-resolution images than previous works (see paper)
- We demonstrated promising results in content transfer
- Flow-based models can provide an alternative to GANs

Thank you for listening!