Densely connected normalizing flows

Matej Grcić, Ivan Grubišić, Siniša Šegvić

University of Zagreb, Faculty of Electrical Engineering and Computing

NeurIPS 2021

Generative modeling

- Assume available dataset D, obtained by sampling an unknown data. distribution p_D
- Our goal is to approximate the unknown p_D using a model p_θ
- Minimize divergence between p_D and p_θ:

$$\min \ \mathrm{KL}(p_D || p_\theta) = \min \ \mathbb{E}_{\mathbf{x} \in D}[-\ln p_\theta(\mathbf{x})]$$

- Various designs of p_θ: Autoregressive factorization Van Oord et al. (2016), Lower bound using variational distribution Kingma and Welling (2014), Unnormalized distribution Salakhutdinov and Hinton (2009), etc.
- We focus on a bijective formulation of p_θ due to exact likelihood and efficient sampling Rezende and Mohamed (2015)

Generative modeling

- Assume available dataset D, obtained by sampling an unknown data. distribution p_D
- Our goal is to approximate the unknown p_D using a model p_θ
- Minimize divergence between p_D and p_θ:

$$\min \ \mathrm{KL}(p_D || p_\theta) = \min \ \mathbb{E}_{\mathbf{x} \in D}[-\ln p_\theta(\mathbf{x})]$$

- Various designs of p_θ: Autoregressive factorization Van Oord et al. (2016), Lower bound using variational distribution Kingma and Welling (2014), Unnormalized distribution Salakhutdinov and Hinton (2009), etc.
- We focus on a bijective formulation of p_θ due to exact likelihood and efficient sampling Rezende and Mohamed (2015)

Normalizing flows

- Given the differentiable bijection $\boldsymbol{f}_\theta$, the change of variable formula is:

$$p_\theta(\boldsymbol{x}) = p(\boldsymbol{z}) \left| \det \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} \right|, \quad \boldsymbol{z} = \boldsymbol{f}_\theta(\boldsymbol{x})$$

- By defining $\boldsymbol{f}_\theta$ as composition $\boldsymbol{f}_\theta = \boldsymbol{f}_{\theta_K} \circ \boldsymbol{f}_{\theta_{K-1}} \circ \dots \circ \boldsymbol{f}_{\theta_1}$, we obtain log-likelihood Dinh et al. (2015) and Rezende and Mohamed (2015):

$$\ln p_\theta(\boldsymbol{x}) = \ln p(\boldsymbol{z}_K) + \sum_{i=1}^{K} \ln |\det \mathbf{J}_{f_i}|.$$

$$\boldsymbol{x} \overset{f_1}{\longleftrightarrow} \boldsymbol{z}_1 \overset{f_2}{\longleftrightarrow} \boldsymbol{z}_2 \overset{f_3}{\longleftrightarrow} \cdots \overset{f_{i-1}}{\longleftrightarrow} \boldsymbol{z}_i \overset{f_i}{\longleftrightarrow} \cdots \overset{f_K}{\longleftrightarrow} \boldsymbol{z}_K, \quad \boldsymbol{z}_K \sim \mathcal{N}(0, \mathrm{I})$$

- Due to the bijective constraint, every $\boldsymbol{z}_i$ has the same dimensionality
- Model expressiveness **is limited** by the input dimensionality

Normalizing flows

- Given the differentiable bijection $\boldsymbol{f}_\theta$, the change of variable formula is:

$$p_\theta(\boldsymbol{x}) = p(\boldsymbol{z}) \left| \det \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} \right|, \quad \boldsymbol{z} = \boldsymbol{f}_\theta(\boldsymbol{x})$$

- By defining $\boldsymbol{f}_\theta$ as composition $\boldsymbol{f}_\theta = \boldsymbol{f}_{\theta_K} \circ \boldsymbol{f}_{\theta_{K-1}} \circ \cdots \circ \boldsymbol{f}_{\theta_1}$, we obtain log-likelihood Dinh et al. (2015) and Rezende and Mohamed (2015):

$$\ln p_\theta(\boldsymbol{x}) = \ln p(\boldsymbol{z}_K) + \sum_{i=1}^{K} \ln |\det \mathbf{J}_{f_i}|.$$

$$\boldsymbol{x} \overset{f_1}{\longleftrightarrow} \boldsymbol{z}_1 \overset{f_2}{\longleftrightarrow} \boldsymbol{z}_2 \overset{f_3}{\longleftrightarrow} \cdots \overset{f_{i-1}}{\longleftrightarrow} \boldsymbol{z}_i \overset{f_i}{\longleftrightarrow} \cdots \overset{f_K}{\longleftrightarrow} \boldsymbol{z}_K, \quad \boldsymbol{z}_K \sim \mathcal{N}(0, \mathrm{I})$$

- Due to the bijective constraint, every $\mathbf{z}_i$ has the same dimensionality
- Model expressiveness **is limited** by the input dimensionality

Normalizing flows

- Given the differentiable bijection $\boldsymbol{f}_\theta$, the change of variable formula is:

$$p_\theta(\boldsymbol{x}) = p(\boldsymbol{z}) \left| \det \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} \right|, \quad \boldsymbol{z} = \boldsymbol{f}_\theta(\boldsymbol{x})$$

- By defining $\boldsymbol{f}_\theta$ as composition $\boldsymbol{f}_\theta = \boldsymbol{f}_{\theta_K} \circ \boldsymbol{f}_{\theta_{K-1}} \circ \cdots \circ \boldsymbol{f}_{\theta_1}$, we obtain log-likelihood Dinh et al. (2015) and Rezende and Mohamed (2015):

$$\ln p_\theta(\boldsymbol{x}) = \ln p(\boldsymbol{z}_K) + \sum_{i=1}^{K} \ln |\det \mathbf{J}_{f_i}|.$$

$$\boldsymbol{x} \xleftrightarrow{f_1} \boldsymbol{z}_1 \xleftrightarrow{f_2} \boldsymbol{z}_2 \xleftrightarrow{f_3} \cdots \xleftrightarrow{f_{i-1}} \boldsymbol{z}_i \xleftrightarrow{f_i} \cdots \xleftrightarrow{f_K} \boldsymbol{z}_K, \quad \boldsymbol{z}_K \sim \mathcal{N}(0, \mathrm{I})$$

- Due to the bijective constraint, every $\mathbf{z}_i$ has the same dimensionality
- Model expressiveness **is limited** by the input dimensionality

Intermediate variable augmentation

- At arbitrary step i:

$$\cdots \xleftrightarrow{f_{i-1}} \boldsymbol{z}_i \xrightarrow{\text{aug}} [\boldsymbol{z}_i, \boldsymbol{e}_i] \xrightarrow{h_i} \boldsymbol{z}_i^{(\text{aug})} \xleftrightarrow{f_{i+1}} \boldsymbol{z}_{i+1} \xleftrightarrow{f_{i+2}} \cdots$$

- $aug(\cdot)$ concatenates noise to latent representation $\boldsymbol{z}_i$:

$$aug(\boldsymbol{z}_i) = [\boldsymbol{z}_i, \boldsymbol{e}_i], \quad \boldsymbol{e}_i \sim \mathcal{N}(0, I)$$

- $h_i(\cdot,\cdot)$ transforms the noise based on previous latent variables $\boldsymbol{z}_{<i}$:

$$\boldsymbol{z}_i^{(\text{aug})} = h_i([\boldsymbol{z}_i, \boldsymbol{e}_i], \boldsymbol{z}_{<i}) = [\boldsymbol{z}_i, \boldsymbol{\sigma} \odot \boldsymbol{e}_i + \boldsymbol{\mu}], \quad (\boldsymbol{\mu}, \boldsymbol{\sigma}) = g_i(\boldsymbol{z}_{<i})$$

$$\frac{\partial \boldsymbol{z}_i^{(\text{aug})}}{\partial [\boldsymbol{z}_i, \boldsymbol{e}_i]} = \begin{bmatrix} \mathrm{I} & 0 \\ 0 & \mathrm{diag}(\boldsymbol{\sigma}) \end{bmatrix}$$

Intermediate variable augmentation

- At arbitrary step i:

$$\cdots \overset{f_{i-1}}{\longleftrightarrow} \boldsymbol{z}_i \xrightarrow{\text{aug}} [\boldsymbol{z}_i, \boldsymbol{e}_i] \xrightarrow{h_i} \boldsymbol{z}_i^{(\text{aug})} \overset{f_{i+1}}{\longleftrightarrow} \boldsymbol{z}_{i+1} \overset{f_{i+2}}{\longleftrightarrow} \cdots$$

- $aug(\cdot)$ concatenates noise to latent representation $\boldsymbol{z}_i$:

$$aug(\boldsymbol{z}_i) = [\boldsymbol{z}_i, \boldsymbol{e}_i], \quad \boldsymbol{e}_i \sim \mathcal{N}(0, I)$$

- $h_i(\cdot, \cdot)$ transforms the noise based on previous latent variables $\boldsymbol{z}_{<i}$:

$$\boldsymbol{z}_i^{(\text{aug})} = h_i([\boldsymbol{z}_i, \boldsymbol{e}_i], \boldsymbol{z}_{<i}) = [\boldsymbol{z}_i, \boldsymbol{\sigma} \odot \boldsymbol{e}_i + \boldsymbol{\mu}], \quad (\boldsymbol{\mu}, \boldsymbol{\sigma}) = g_i(\boldsymbol{z}_{<i})$$

$$\frac{\partial \boldsymbol{z}_i^{(\text{aug})}}{\partial [\boldsymbol{z}_i, \boldsymbol{e}_i]} = \begin{bmatrix} \mathrm{I} & 0 \\ 0 & \mathrm{diag}(\boldsymbol{\sigma}) \end{bmatrix}$$

Cross-unit coupling

- At arbitrary step i:

$$\dots \xleftrightarrow{f_{i-1}} \boldsymbol{z}_i \xrightarrow{\text{(aug)}} [\boldsymbol{z}_i, \boldsymbol{e}_i] \xrightarrow{h_i} \boldsymbol{z}_i^{\text{(aug)}} \xleftrightarrow{f_{i+1}} \boldsymbol{z}_{i+1} \xleftrightarrow{f_{i+2}} \dots$$

- Likelihood lower bound defined as:

$$\ln p(\boldsymbol{z}_i) \geq \mathbb{E}_{\boldsymbol{e}_i \sim p^*(\boldsymbol{e}_i)}[\ln p(\boldsymbol{z}_i^{\text{(aug)}}) - \ln p^*(\boldsymbol{e}_i) + \ln|\det \mathrm{diag}(\boldsymbol{\sigma})|].$$

- Trivial "inverse" - remove noise dimensions:

$$\boldsymbol{z}_i^{\text{(aug)}} = [\boldsymbol{z}_i, \boldsymbol{\sigma} \odot \boldsymbol{e}_i + \boldsymbol{\mu}] \Rightarrow \boldsymbol{z}_i = \boldsymbol{z}_i^{\text{(aug)}}{}_{[:d]}, \quad d = dim(\boldsymbol{z}_i)$$

- Resulting scheme with the increased model width at arbitrary steps:

$$\boldsymbol{x} \xleftrightarrow{f_1} \boldsymbol{z}_1 \xleftrightarrow{f_2, aug, h_2} \boldsymbol{z}_2^{(aug)} \xleftrightarrow{f_3} \dots \xleftrightarrow{f_i, aug, h_i} \boldsymbol{z}_i^{(aug)} \xleftrightarrow{f_{i+1}} \dots \xleftrightarrow{f_K} \boldsymbol{z}_K$$

Cross-unit coupling

- At arbitrary step i:

$$\dots \overset{f_{i-1}}{\longleftrightarrow} \boldsymbol{z}_i \xrightarrow{(\text{aug})} [\boldsymbol{z}_i, \boldsymbol{e}_i] \xrightarrow{h_i} \boldsymbol{z}_i^{(\text{aug})} \overset{f_{i+1}}{\longleftrightarrow} \boldsymbol{z}_{i+1} \overset{f_{i+2}}{\longleftrightarrow} \dots$$

- Likelihood lower bound defined as:

$$\ln p(\boldsymbol{z}_i) \geq \mathbb{E}_{\boldsymbol{e}_i \sim p^*(\boldsymbol{e}_i)}[\ln p(\boldsymbol{z}_i^{(\text{aug})}) - \ln p^*(\boldsymbol{e}_i) + \ln|\det \mathrm{diag}(\boldsymbol{\sigma})|].$$

- Trivial "inverse" - remove noise dimensions:

$$\boldsymbol{z}_i^{(\text{aug})} = [\boldsymbol{z}_i, \boldsymbol{\sigma} \odot \boldsymbol{e}_i + \boldsymbol{\mu}] \Rightarrow \boldsymbol{z}_i = {\boldsymbol{z}_i^{(\text{aug})}}_{[:d]}, \quad d = dim(\boldsymbol{z}_i)$$

- Resulting scheme with the increased model width at arbitrary steps:

$$\boldsymbol{x} \overset{f_1}{\longleftrightarrow} \boldsymbol{z}_1 \overset{f_2,aug,h_2}{\longleftrightarrow} \boldsymbol{z}_2^{(aug)} \overset{f_3}{\longleftrightarrow} \dots \overset{f_i,aug,h_i}{\longleftrightarrow} \boldsymbol{z}_i^{(aug)} \overset{f_{i+1}}{\longleftrightarrow} \dots \overset{f_K}{\longleftrightarrow} \boldsymbol{z}_K$$

Cross-unit coupling

- At arbitrary step i:

$$\cdots \overset{f_{i-1}}{\longleftrightarrow} \boldsymbol{z}_i \overset{\text{(aug)}}{\longrightarrow} [\boldsymbol{z}_i, \boldsymbol{e}_i] \overset{h_i}{\longrightarrow} \boldsymbol{z}_i^{\text{(aug)}} \overset{f_{i+1}}{\longleftrightarrow} \boldsymbol{z}_{i+1} \overset{f_{i+2}}{\longleftrightarrow} \cdots$$

- Likelihood lower bound defined as:

$$\ln p(\boldsymbol{z}_i) \geq \mathbb{E}_{\boldsymbol{e}_i \sim p^*(\boldsymbol{e}_i)}[\ln p(\boldsymbol{z}_i^{\text{(aug)}}) - \ln p^*(\boldsymbol{e}_i) + \ln|\det \mathrm{diag}(\boldsymbol{\sigma})|].$$

- Trivial "inverse" - remove noise dimensions:

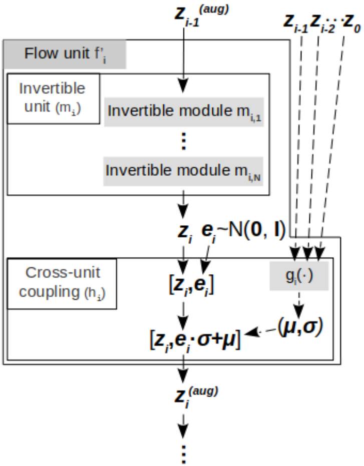
$$\boldsymbol{z}_i^{\text{(aug)}} = [\boldsymbol{z}_i, \boldsymbol{\sigma} \odot \boldsymbol{e}_i + \boldsymbol{\mu}] \Rightarrow \boldsymbol{z}_i = \boldsymbol{z}_i^{\text{(aug)}}{}_{[:d]}, \quad d = dim(\boldsymbol{z}_i)$$

- Resulting scheme with the increased model width at arbitrary steps:

$$\boldsymbol{x} \overset{f_1}{\longleftrightarrow} \boldsymbol{z}_1 \overset{f_2, aug, h_2}{\longleftrightarrow} \boldsymbol{z}_2^{(aug)} \overset{f_3}{\longleftrightarrow} \cdots \overset{f_i, aug, h_i}{\longleftrightarrow} \boldsymbol{z}_i^{(aug)} \overset{f_{i+1}}{\longleftrightarrow} \cdots \overset{f_K}{\longleftrightarrow} \boldsymbol{z}_K$$

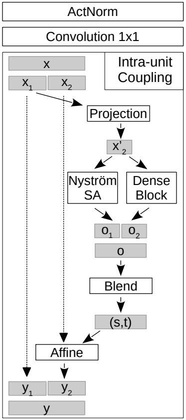
Cross-unit coupling - scheme

- **Invertible unit:** arbitrary composition of differentiable bijections
- **Cross-unit coupling:** modular coupling layer over latent representations in multiple stages

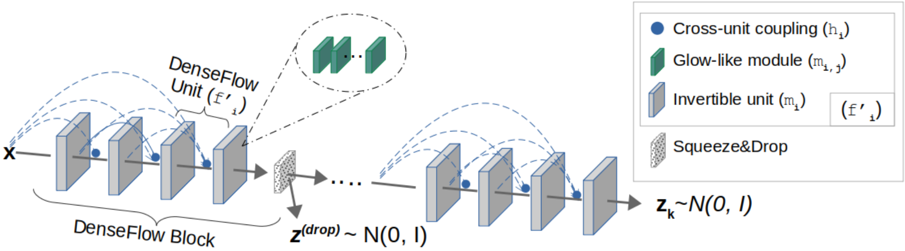


Intra-unit coupling

- Based on Glow modules Kingma and Dhariwal (2018)
- Coupling network fuses:
  - Local correlations produced by Dense block Huang et al. (2017)
  - Global context captured by Nyström Self-attention Xiong et al. (2021)
- More efficient than Flow++ coupling Ho et al. (2019)
- **Intra-unit coupling:** second level of skip connections



- Image-oriented multi-scale architecture
- Dense skip connections provided by cross-unit and intra-unit couplings



Density estimation

	Method	CIFAR-10 32x32	ImageNet 32x32	CelebA 64x64	ImageNet 64x64
Variational Autoencoders	Conv Draw Gregor et al. (2016)	3.58	4.40	-	4.10
	DVAE++ Vahdat et al. (2018)	3.38	-	-	-
	IAF-VAE Kingma et al. (2016)	3.11	-	-	-
	BIVA Maaløe et al. (2019)	3.08	3.96	2.48	-
	Imp. DDPM Nichol and Dhariwal (2021)	2.94	-	-	3.53
Autoregressive Models	Gated PixelCNNOord et al. (2016)	3.03	3.83	-	3.57
	PixelRNN Van Oord et al. (2016)	3.00	3.86	-	3.63
	PixelCNN++ Salimans et al. (2017)	2.92	-	-	-
	Image Transformer Parmar et al. (2018)	2.90	3.77	2.61	-
	PixelSNAIL Chen et al. (2018)	2.85	3.80	-	-
	SPN Menick and Kalchbrenner (2019)	-	3.85	-	3.53
	Routing transformer Roy et al. (2021)	2.95	-	-	3.43
Normalizing Flows	Real NVP Dinh et al. (2017)	3.49	4.28	3.02	3.98
	GLOW Kingma and Dhariwal (2018)	3.35	4.09	-	3.81
	Residual Flow Chen et al. (2019)	3.28	4.01	-	3.78
	i-DenseNet Perugachi-Diaz et al. (2021)	3.25	3.98	-	-
	Flow++ Ho et al. (2019)	3.08	3.86	-	3.69
	ANF Huang et al. (2020)	3.05	3.92	-	3.66
	VFlow Chen et al. (2020)	2.98	3.83	-	3.66
Hybrid Architectures	MaCow Ma et al. (2019)	3.16	-	-	3.69
	SurVAE Flow Nielsen et al. (2020)	3.08	4.00	-	3.70
	NVAE Vahdat and Kautz (2020)	2.91	3.92	2.03	-
	PixelVAE++ Sadeghi et al. (2019)	2.90	-	-	-
	δ-VAE Razavi et al. (2019)	**2.83**	3.77	-	-
	DenseFlow-74-10 (ours)	2.98	**3.63**	**1.99**	**3.35**

Computational complexity

- Our DenseFlow uses **only one** GPU for training!
  - Without gradient checkpointing
  - Without mixed precision

Dataset	Model	GPU type	GPUs	Duration (h)	Likelihood (bpd)
CIFAR-10	VFlow Chen et al. (2020)	RTX 2080Ti	16	∼500	2.98
	NVAE Vahdat and Kautz (2020)	Tesla V100	8	55	2.91
	DenseFlow-74-10 (ours)	RTX 3090	1	250	2.98
ImageNet32	VFlow Chen et al. (2020)	Tesla V100	16	∼1440	3.83
	NVAE Vahdat and Kautz (2020)	Tesla V100	24	70	3.92
	DenseFlow-74-10 (ours)	Tesla V100	1	310	3.63
CelebA	VFlow Chen et al. (2020)	n/a	n/a	n/a	-
	NVAE Vahdat and Kautz (2020)	Tesla V100	8	92	2.03
	DenseFlow-74-10 (ours)	Tesla V100	1	224	1.99

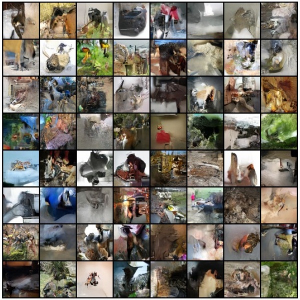
Visual quality - FID score

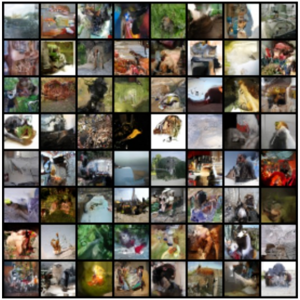
- Competitive visual quality on CIFAR10

	Model	FID ↓
Autoregressive Models	PixelCNN Ostrovski et al. (2018) and Van Oord et al. (2016)	65.93
	PixelIQN Ostrovski et al. (2018)	49.46
Normalizing Flows	i-ResNet Behrmann et al. (2019)	65.01
	Glow Kingma and Dhariwal (2018)	46.90
	Residual flow Chen et al. (2019)	46.37
GANs	DCGAN Ostrovski et al. (2018) and Radford et al. (2016)	37.11
	WGAN-GP Gulrajani et al. (2017) and Ostrovski et al. (2018)	36.40
	DA-StyleGAN V2 Zhao et al. (2020)	5.79
Hybrid Architectures	SurVAE-flow Nielsen et al. (2020)	49.03
	VAEBM Xiao et al. (2020)	12.19
	DenseFlow-74-10 (ours)	34.90

Visual samples

- Samples generation:
  - Sample the latent distribution to obtain $\mathbf{z}$: $\mathbf{z} \sim \mathcal{N}(0, \mathrm{I})$
  - Apply the inverse transformation $\mathbf{x} = \mathbf{f}_\theta^{-1}(\mathbf{z})$





Takeaways

- Expressiveness of a NF does not only depend on latent dimensionality but also on its distribution across the model depth
- Expressiveness of a NF can also be improved by conditioning the introduced noise with the proposed densely connected cross-unit coupling
- Combining these insights with Nystrom self attention and the proposed intra-unit coupling increases the NF performance while reducing computational requirements

- GitHub: matejgrcic/DenseFlow
- ArXiv: abs/2106.04627
- Contact: matej.grcic@fer.hr
- Questions: email or new issue

References I

Behrmann, Jens et al. (2019). "Invertible residual networks". In: *International Conference on Machine Learning*. PMLR, pp. 573–582.

Chen, Jianfei et al. (2020). "Vflow: More expressive generative flows with variational data augmentation". In: *International Conference on Machine Learning*. PMLR, pp. 1660–1669.

Chen, Tian Qi et al. (2019). "Residual Flows for Invertible Generative Modeling". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 9913–9923.

Chen, Xi et al. (2018). "PixelSNAIL: An Improved Autoregressive Generative Model". In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 863–871.

Dinh, Laurent, David Krueger, and Yoshua Bengio (2015). "NICE: Non-linear Independent Components Estimation". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.

Dinh, Laurent, Jascha Sohl-Dickstein, and Samy Bengio (2017). "Density estimation using Real NVP". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

References II

Gregor, Karol et al. (2016). "Towards Conceptual Compression". In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3549–3557.

Gulrajani, Ishaan et al. (2017). "Improved Training of Wasserstein GANs". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5767–5777.

Ho, Jonathan et al. (2019). "Flow++: Improving flow-based generative models with variational dequantization and architecture design". In: *International Conference on Machine Learning*. PMLR, pp. 2722–2730.

Huang, Chin-Wei, Laurent Dinh, and Aaron Courville (2020). "Augmented normalizing flows: Bridging the gap between generative flows and latent variable models". In: *arXiv preprint arXiv:2002.07101*.

Huang, Gao et al. (2017). "Densely Connected Convolutional Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, pp. 2261–2269.

Kingma, Diederik P. and Prafulla Dhariwal (2018). "Glow: Generative Flow with Invertible 1x1 Convolutions". In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 10236–10245.

References III

Kingma, Diederik P. and Max Welling (2014). "Auto-Encoding Variational Bayes". In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Kingma, Diederik P et al. (2016). "Improving variational inference with inverse autoregressive flow". In: *arXiv preprint arXiv:1606.04934*.

Ma, Xuezhe et al. (2019). "MaCow: Masked Convolutional Generative Flow". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 5891–5900.

Maaløe, Lars et al. (2019). "BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 6548–6558.

Menick, Jacob and Nal Kalchbrenner (2019). "Generating High fidelity Images with subscale pixel Networks and Multidimensional Upscaling". In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Nichol, Alex and Prafulla Dhariwal (2021). "Improved Denoising Diffusion Probabilistic Models". In: *CoRR* abs/2102.09672. arXiv: 2102.09672. URL: https://arxiv.org/abs/2102.09672.

References IV

Nielsen, Didrik et al. (2020). "SurVAE Flows: Surjections to Bridge the Gap between VAEs and Flows". In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

Oord, Aäron van den et al. (2016). "Conditional Image Generation with PixelCNN Decoders". In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain,* pp. 4790–4798.

Ostrovski, Georg, Will Dabney, and Rémi Munos (2018). "Autoregressive Quantile Networks for Generative Modeling". In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 3933–3942.

Parmar, Niki et al. (2018). "Image Transformer". In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 4052–4061.

Perugachi-Diaz, Yura, Jakub M Tomczak, and Sandjai Bhulai (2021). "Invertible DenseNets with Concatenated LipSwish". In: *arXiv preprint arXiv:2102.02694.*

Radford, Alec, Luke Metz, and Soumith Chintala (2016). “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks”. In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Razavi, Ali et al. (2019). “Preventing Posterior Collapse with delta-VAEs”. In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Rezende, Danilo Jimenez and Shakir Mohamed (2015). “Variational Inference with Normalizing Flows”. In: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. Vol. 37. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 1530–1538.

Roy, Aurko et al. (2021). “Efficient Content-Based Sparse Attention with Routing Transformers”. In: *Trans. Assoc. Comput. Linguistics* 9, pp. 53–68.

Sadeghi, Hossein et al. (2019). “PixelVAE++: Improved PixelVAE with Discrete Prior”. In: *CoRR* abs/1908.09948.

Salakhutdinov, Ruslan and Geoffrey Hinton (2009). “Deep Boltzmann Machines”. In: *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*. Vol. 5. Proceedings of Machine Learning Research. Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA: PMLR, pp. 448–455.

References VI

Salimans, Tim et al. (2017). "PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Vahdat, Arash and Jan Kautz (2020). "NVAE: A Deep Hierarchical Variational Autoencoder". In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Vahdat, Arash et al. (2018). "DVAE++: Discrete Variational Autoencoders with Overlapping Transformations". In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 5042–5051.

Van Oord, Aaron, Nal Kalchbrenner, and Koray Kavukcuoglu (2016). "Pixel recurrent neural networks". In: *International Conference on Machine Learning*. PMLR, pp. 1747–1756.

Xiao, Zhisheng et al. (2020). "VAEBM: A Symbiosis between Variational Autoencoders and Energy-based Models". In: *CoRR* abs/2010.00654.

Xiong, Yunyang et al. (2021). "Nyströmformer: A Nyström-Based Algorithm for Approximating Self-Attention". In: *CoRR* abs/2102.03902.

References VII

Zhao, Shengyu et al. (2020). "Differentiable Augmentation for Data-Efficient GAN Training". In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.