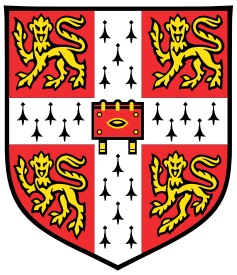
Improving Representation Learning through Generative Modeling

From Diffusion Models to Riemannian Geometry



Georgios Batzolis

Department of Applied Mathematics and Theoretical Physics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Churchill College December 2024

I would like to dedicate this thesis to my loving parents ...

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Georgios Batzolis
December 2024

Acknowledgements

And I would like to acknowledge ...

Abstract

This is where you write your abstract ...

Table of contents

List of figures

List of tables

Chapter 1

Thesis Outline and Contributions

1.1 CAFLOW: Conditional Autoregressive Flows

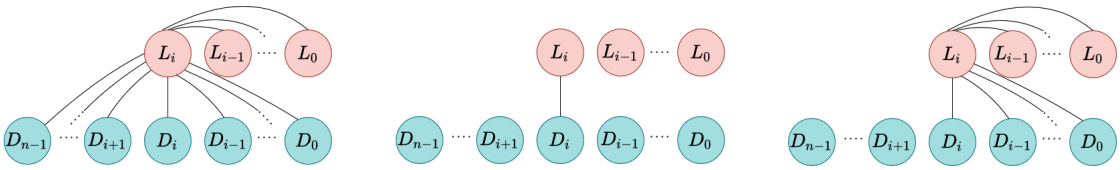


Fig. 1.1 From left to right: ideal dependencies in the i^{th} autoregressive component. Dual-Glow modeling assumption [6]; information is exchanged only between latent spaces having the same dimension. Our modeling assumption; we retain the dependencies between L_i and the latent spaces of lower dimension.

In this chapter, we introduce CAFLOW, a new diverse image-to-image translation model that simultaneously leverages the power of auto-regressive modeling and the modeling efficiency of conditional normalizing flows. We transform the conditioning image into a sequence of latent encodings using a multi-scale normalizing flow and repeat the process for the conditioned image. We model the conditional distribution of the latent encodings by modeling the auto-regressive distributions with an efficient multi-scale normalizing flow, where each conditioning factor affects image synthesis at its respective resolution scale. Our proposed framework performs well on a range of image-to-image translation tasks. It outperforms former designs of conditional flows because of its expressive auto-regressive structure.

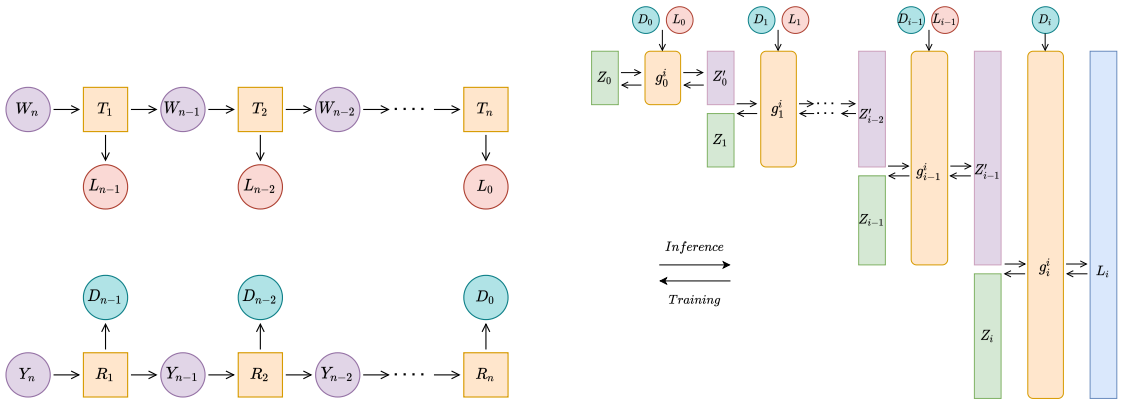


Fig. 1.2 Left: unconditional normalizing flow architecture used to encode conditioning and conditioned images, denoted by $Y_n = Y$ and $W_n = W$ respectively, into a sequence of hierarchical latent variables. Right: design of the conditional transformation G_i^θ that models the i^{th} autoregressive component. The index of the flow i is omitted in both the transformed latent variable Z_j and the intermediate latent variables Z'_j for simplicity.

Originality and Author's Contributions

This chapter is adapted from our published work [1] in the Foundations of Data Science (FoDS) journal. The authors' contributions to this work are as follows:

- **Formulation of Ideas**: The formulation of the CAFLOW framework was entirely my contribution.
- **Experimental Design**: I independently designed all experiments to validate the performance of the proposed framework.
- **Code and Experimental Implementation**: I implemented the entire codebase for the model and conducted all experiments. Christian Etmann provided critical feedback and domain expertise on the training and inference of Normalizing Flows, significantly enhancing the quality of the work.
- **Theory**: I developed the theoretical foundations underpinning the proposed framework, with valuable contributions from Marcelo Carioni, who helped refine and extend the theoretical arguments.
- **Presentation**: Myself and Marcelo Carioni contributed equally to the write-up of this work.

This project was conducted under the supervision of Carola-Bibiane Schönlieb, Christian Etmann, Soroosh Afyouni, and Zoe Kourtzi.

1.2 Non-Uniform Diffusion Models

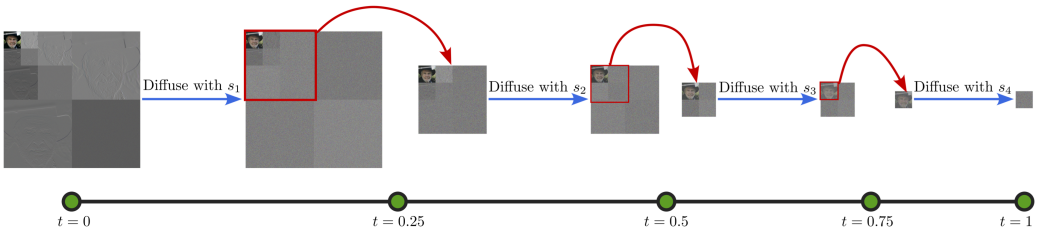


Fig. 1.3 Illustration of a multi-scale diffusion model with three scales. Inspired by multi-scale normalizing flows, this approach diffuses different parts of the image tensor (transformed into multi-level Haar coefficients) at varying speeds. High-frequency detail coefficients diffuse progressively faster, with d_1 diffusing faster than d_2, d_2 faster than d_3, and so on, ensuring that all coefficients reach the same (very low) signal-to-noise ratio (SNR) at their respective terminal diffusion times: $t = 0.25, 0.5, 0.75$, and 1.0, respectively. The low-frequency approximation coefficients a_3 diffuse the slowest, completing their diffusion at $t = 1.0$. The multi-scale structure reduces the dimensionality of the diffusing tensor at each scale, enabling faster computation. Separate neural networks S_1, S_2, S_3, S_4 approximate the score functions at different intervals, leveraging the reduced dimensionality of the intermediate distributions. This hierarchical design mirrors the structure of multi-scale normalizing flows, improving training and sampling efficiency while maintaining high image generation quality.

In this section, we introduce *non-uniform diffusion models*. Unlike standard diffusion approaches that apply the same noise injection schedule to every pixel, non-uniform diffusion models allow different pixels (or groups of pixels) to evolve at varying speeds. This flexibility mirrors the hierarchical approach of multi-scale normalizing flows, where transformations occur at multiple scales, enabling the model to capture image structure more efficiently and produce higher-quality samples in less time. By carefully choosing which parts of the image diffuse faster, non-uniform diffusion opens the door to significantly improved performance, both in terms of image fidelity and computational speed.

We demonstrate that non-uniform diffusion models outperform standard uniform diffusion models by achieving superior FID scores in less training time. Furthermore, these models exhibit remarkable efficiency, sampling up to 4.4 times faster at a resolution of 128×128, with even greater speed-ups anticipated at higher resolutions. Leveraging the adaptability of non-uniform diffusion, we introduce the Conditional Multi-Speed Diffusive Estimator (CMDE), a novel approach derived from a specific choice of non-uniform diffusion. CMDE unifies existing methods for conditional score estimation and delivers performance on par with the widely adopted conditional denoising estimator.

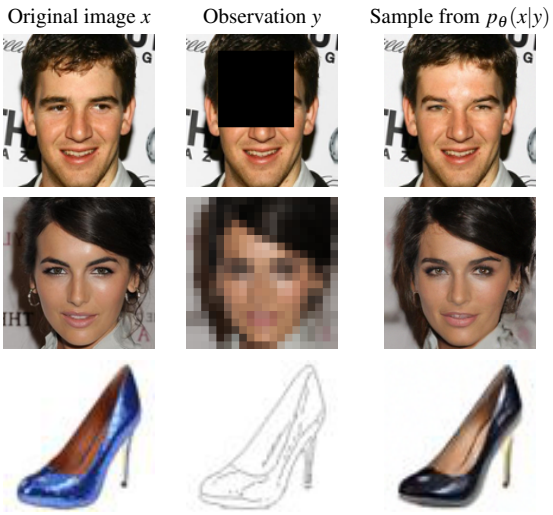


Fig. 1.4 Results from our conditional multi-speed diffusive estimator.

On the theoretical side, we introduce a principled objective for training non-uniform diffusion models and provide a proof of consistency for the conditional denoising estimator, thereby establishing the reliability of the most widely adopted approach to training conditional diffusion models.

Beyond these theoretical and methodological advances, we conduct a comprehensive empirical evaluation of different approaches to training conditional diffusion methods, comparing their effectiveness across tasks such as super-resolution, inpainting, and edge-to-image translation. Finally, to encourage future research and practical adoption, we release MSDiff, an open-source library dedicated to non-uniform and conditional diffusion models, enabling other researchers and practitioners to easily experiment with and build upon our work.

Originality and Author's Contributions

This chapter is adapted from [3]. The author's contributions to this work are as follows:

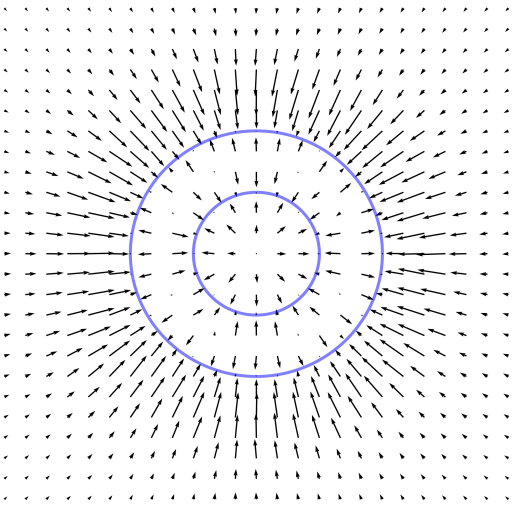
- **Formulation of Ideas**: I came up with the idea of non-uniform diffusion and how to train non-uniform diffusion models. This idea led to multi-scale diffusion and the CMDE estimator which can be used for training conditional diffusion models. Jan Stanczuk noticed that CMDE is an interpolation between CDE and CdiffE.

- **Experimental Design**: Myself and Jan Stanczuk had equal contribution in the experimental design.

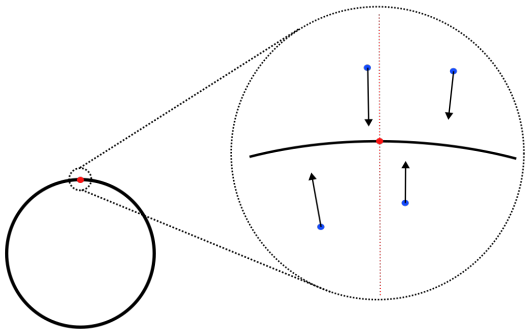
- **Code and Experimental Implementation**: I implemented the experiments presented in the paper. Myself and Jan Stanczuk had equal contribution in the delevopment of the codebase.
- **Theory**: The ideas for the proofs of Theorem 1 and 2 were found together. Final technical ideas were completed by Jan Stanczuk. Jan Stanczuk also compiled the proof of Theorem 3.
- **Presentation**: The design of the paper was discussed together. Jan Stanczuk led the write up of the paper. I contributed to experimental sections of the write up.

This project was conducted under the supervision of Carola-Bibiane Schönlieb and Christian Etmann.

1.3 Diffusion Models Encode the Intrinsic Dimension of Data Manifolds



(a) The data manifold (in blue) and the neural approximation of the score field $\nabla_{\mathbf{x}} \ln p_{t_0}(\mathbf{x})$ obtained from a diffusion model. Near the manifold, the score field is perpendicular to the manifold surface.



(b) The red dot shows a point $\mathbf{x}_0$ on the data manifold where we wish to estimate the dimension. We sample K blue points $\mathbf{x}_t^{(i)}$ in a close neighborhood of the red point and evaluate the score field. The resulting vectors $s_\theta(\mathbf{x}_\varepsilon^{(i)}, \varepsilon)$ point in the normal direction. We put the vectors into a matrix and perform SVD to detect the dimension of the normal space. The dimension of the manifold equals the number of (almost) vanishing singular values.

Fig. 1.5 (Left) Visualization of the score field near the data manifold. (Right) Visualisation of the estimation of the manifold dimension using the trained diffusion model.

In this chapter, we provide a mathematical proof that diffusion models encode data manifolds by approximating their normal bundles. Based on this observation we propose a novel method for extracting the intrinsic dimension of the data manifold from a trained diffusion model. Our insights are based on the fact that a diffusion model approximates the score function i.e. the gradient of the log density of a noise-corrupted version of the target distribution for varying levels of corruption. We prove that as the level of corruption decreases, the score function points towards the manifold, as this direction becomes the direction of maximal likelihood increase. Therefore, at low noise levels, the diffusion model provides us with an approximation of the manifold's normal bundle, allowing for an estimation of the manifold's intrinsic dimension. To the best of our knowledge our method is the first estimator of intrinsic dimension based on diffusion models and it outperforms well established estimators in controlled experiments on both Euclidean and image data.

Originality and Author's Contributions

This chapter is adapted from our published work presented at ICML 2024 on intrinsic dimension estimation with diffusion models [5]. The authors' contributions to this work are as follows:

- **Formulation of Ideas**: Myself and Jan Stanczuk equally contributed to the formulation of ideas.
- **Experimental Design**: Myself and Jan Stanczuk shared equal contributions to the experimental design.
- **Code and Experimental Implementation**: Myself and Jan Stanczuk had equal contribution in the developement of the code framework used to run the experiments. I implemented the following experiments: synthetic image manifolds and MNIST dimensionality estimation. Jan Stanczuk implemented the following experiments: k-spheres, line manifold, spaghetti line, comparison with auto-encoder on MNIST, robustness analysis, and all benchmark methods.
- **Theory**: Jan Stanczuk derived the theoretical results with help from Teo Deveney.
- **Presentation**: Jan Stanczuk and I equally contributed to the write-up, with Teo Deveney contributing to the theoretical sections.

This project was conducted under the supervision of Carola-Bibiane Schönlieb.

1.4 Variational Diffusion Auto-encoder: Latent Space Extraction from Pre-Trained Diffusion Models

Fig. 1.6 Comparison of original and reconstructed images on the FFHQ dataset using our ScoreVAE framework. The left panel presents the original images from the FFHQ dataset, while the right panel displays the corresponding reconstructions generated by ScoreVAE. The results highlight the effectiveness of ScoreVAE in capturing intricate details and preserving high fidelity, overcoming the limitations of traditional VAE models.

In this paper, we introduce **ScoreVAE**, a novel approach that advances the Variational Autoencoder (VAE) framework by addressing fundamental limitations of conventional VAEs. Traditional VAEs model the reconstruction distribution $p(\mathbf{x}|\mathbf{z})$ as a Gaussian, which often leads to overly smoothed and blurry reconstructions. This limitation arises because the Gaussian assumption fails to capture the complexity and multimodality of real-world data distributions, making it difficult for the model to accurately represent intricate details and sharp features. ScoreVAE addresses this issue by combining a diffusion-time-dependent encoder and an unconditional diffusion model. By employing Bayes' rule for score functions, we analytically derive a robust and flexible model for reconstruction distribution $p(\mathbf{x}|\mathbf{z})$. Our approach bypasses the unrealistic Gaussian assumption, resulting in significantly improved image reconstruction quality.

The ScoreVAE framework also simplifies the training dynamics by decoupling the training of the diffusion model and the encoder. This decoupling enables the use of powerful pre-trained diffusion models that can be readily updated or swapped without retraining the entire system. By separating the prior (diffusion model) from the encoder, ScoreVAE achieves higher fidelity reconstructions compared to traditional VAEs and diffusion decoders.

Our experiments on the CIFAR10 and CelebA datasets demonstrate ScoreVAE's superiority in producing sharper images and reducing reconstruction error. These results underscore the practical advantages of ScoreVAE in handling complex, high-dimensional data, and highlight its potential for improved representation learning and controllable generative modeling.

Originality and Author's Contributions

This chapter is adapted from [2]. The authors' contributions to this work are as follows:

- **Formulation of Ideas**: Myself and Jan Stanczuk contributed equally to the formulation of ideas.
- **Experimental Design**: Myself and Jan Stanczuk had equal contributions to the experimental design.
- **Code and Experimental Implementation**: I implemented the code for ScoreVAE model (encoder-only, corrector), adaptation of discrete prior models to continuous time framework and the following experiments: reconstructions with score-VAE and Diffusion Decoder. Jan Stanczuk implemented the beta-VAE and code for semantic manipulation.
- **Theory**: I suggested the sketch of the proof that entailed connecting the ScoreVAE loss to marginal likelihood [4], and Jan Stanczuk derived the rigorous proof.
- **Presentation**: Myself and Jan Stanczuk contributed equally to the final write-up.

This project was supervised by Carola-Bibiane Schönlieb.

1.5 Score-Based Pullback Riemannian Geometry

In this work, we introduce a score-based pullback Riemannian metric that encodes the intrinsic dimensionality and geometry of data under certain distributional assumptions. We show that this data-driven metric can be constructed in practice by modifying normalizing flows with anisotropic base distribution and isometry regularization. This approach yields a scalable framework for computing manifold maps—such as geodesics, exponential maps, distances, and curvature—in closed form. Building on this metric, we additionally construct a Riemannian Auto-encoder (RAE) that recovers the true manifold dimension, offers a global chart of the manifold, and yields an interpretable latent representation thanks to isometry regularization.

(a) Hemisphere (2,3) (b) Sinusoid (1,3) (c) Sinusoid (2,3)

Fig. 1.7 Approximate data manifolds learned by the Riemannian autoencoder generated by score-based pullback Riemannian geometry for three datasets. The orange surfaces represent the manifolds learned by the model, while the blue points correspond to the training data. Each manifold provides a convincing low-dimensional representation of the data, isometric to its respective latent space.

Originality and Author's Contributions

This chapter is adapted from the published work on score-based pullback Riemannian geometry.

- **Formulation of Ideas**: Willem Diepeveen proposed the idea of extracting the data manifold's geometry using a score-based pullback metric. He developed the theoretical motivation of using the score-based metric and showed that the constructed metric allows the computation of manifold maps such as geodesics, exponential map, logarithmic map and distance in closed form. Moreover, he showed that the trained normalizing flow used for the construction of the pullback metric is a Riemannian Auto-encoder.

  I figured out that the proposed metric can be constructed in practice by adapting the framework of Normalizing flows with base distribution anisotropy and l_2 regularisation. Moreover, I discovered the existence of a hessian vector product term that can negatively impact the performance of the Riemannian Auto-encoder if not taken into consideration. Including the hessian vector product term in the regularisation allowed us to use non-affine flows reliably which led to the improvement of the scalability and performance of the method.

- **Experimental Design**: I led the experimental design, incorporating suggestions from Willem Diepeveen and Zakhar Shumaylov.

- **Code and Experimental Implementation**: I implemented the training code and conducted all hyperparameter tuning and experiments. Willem implemented the

functionality for constructing the data-driven manifold maps and the Riemannian autoencoder from a trained model.

- **Theory**: Willem developed the theoretical results. I contributed by explaining the necessity of Hessian-vector product regularization.
- **Presentation**: Zakhar and Willem wrote all sections of the paper except for the experimental section, which I authored.

This project was conducted under the supervision of Carola-Bibiane Schönlieb.

References

[1] Batzolis, G., Carioni, M., Etmann, C., Afyouni, S., Kourtzi, Z., and Schönlieb, C.-B. (2024). Caflow: Conditional autoregressive flows. *Foundations of Data Science*, 6(4):553–583. The first author is supported by GSK. Early access: June 2024.

[2] Batzolis, G., Stanczuk, J., and Schönlieb, C.-B. (2023). Variational diffusion auto-encoder: Latent space extraction from pre-trained diffusion models. *arXiv preprint arXiv:2304.12141*.

[3] Batzolis, G., Stanczuk, J., Schönlieb, C.-B., and Etmann, C. (2022). Non-uniform diffusion models.

[4] Song, Y., Durkan, C., Murray, I., and Ermon, S. (2021). Maximum likelihood training of score-based diffusion models.

[5] Stanczuk, J. P., Batzolis, G., Deveney, T., and Schönlieb, C.-B. (2024). Diffusion models encode the intrinsic dimension of data manifolds. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 46412–46440. PMLR.

[6] Sun, H., Mehta, R., Zhou, H. H., Huang, Z., Johnson, S. C., Prabhakaran, V., and Singh, V. (2019). Dual-glow: Conditional flow-based generative model for modality transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.