Representation-Conditioned Diffusion Models for Guided Training Data Generation

Nithesh Chandher Karthikeyan, Jonas Unger, Gabriel Eilertsen
Linköping University
{nithesh.chandher.karthikeyan, jonas.unger, gabriel.eilertsen}@liu.se

Abstract

Data availability remains a critical bottleneck in many deep learning applications. Large-scale datasets are often expensive to collect, curate and annotate, which can limit the scalability and applicability of supervised learning methods. In this work, we evaluate the classification performance of models trained on synthetic image datasets produced by generative deep learning. In particular, we use latent diffusion models conditioned on learned representations from DINOv2, DINOv3, and CLIP. Our results demonstrates that this representation-conditioned formulation significantly outperforms class-conditioned generation by a large margin (+10.76 p.p. top-1 accuracy on ImageNet100), by improving sample quality and mode coverage. Furthermore, by scaling the size of the synthetic dataset, we are able to outperform a classifier trained on the real data (+2.0 p.p. top-1 accuracy).

We also demonstrate how generated images can be used for augmentation purposes, outperforming classical augmentation methods, and how the conditioning space can be used for sample filtering to further improve training value. Collectively, these findings highlight that representation-conditioned diffusion models provide a promising approach for augmenting, complementing, or potentially replacing real-world datasets in large-scale visual learning tasks.

1. Introduction

With continuous advancements in deep learning, there is a constant demand for larger and diverse training datasets. At the same time, generative modeling has experienced significant progress over the last decade, particularly with the emergence of diffusion models [7, 17]. These developments have sparked a growing interest in using these models to expand training datasets via synthetic data generation, e.g., by steering or fine-tuning a general pre-trained text-to-image model [1, 6, 18].

In this paper, we focus on the most challenging scenario for generative dataset synthesis. Specifically, we assume that (1) only the target dataset is available, preventing the use of existing text-to-image models, (2) the downstream model is trained exclusively on the generated dataset (without access to real images), and (3) no additional annotations, such as text prompts, are available to guide generation. Unlike conventional data augmentation, where augmented images only introduce variations to the real samples, training solely on synthetic dataset requires the generative model to capture the full diversity of the real data distribution. Further, this scenario is also relevant in situations where sharing of real data are restricted due to privacy concerns, such as in medical imaging.

Text-to-image diffusion models can produce high quality and diverse samples, with their performance largely benefiting from guidance provided by text prompts, in addition to large models and training dataset. To increase the quality and diversity without text guidance, we consider representation-conditioned diffusion models (RCDMs) [2, 8, 10]. In RCDMs, guidance is provided from extracted representations of each image, which has been demonstrated to significantly increase the quality of generation as compared to unconditioned or class-conditioned generation [10]. In this work, we evaluate how these improvements translate to the case of training data generation.

We provide experiments with a representation-conditioned latent diffusion model [14] trained on ImageNet100, which is a challenging dataset due to its diversity and relatively limited size. We demonstrate that RCDMs show great promise for synthetic dataset generation, surpassing the class-conditioned generation by a large margin. Furthermore, by scaling the size of the synthetic dataset, we are able to outperform the model trained only on the real dataset. We see this as quite remarkable evidence for the promise of RCDMs and using generative deep learning for training data synthesis.

In summary, we present the following contributions:

- We evaluate the classification performance of models trained on synthetic datasets generated by RCDMs, in-

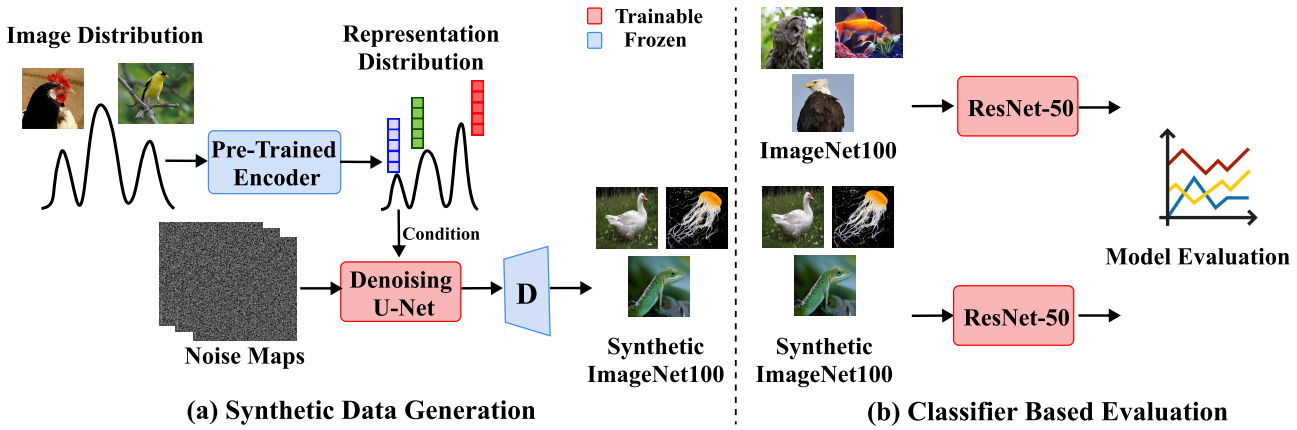

Figure 1. Overall pipeline of the methodology, which consist of two parts: (a) Synthetic Data Generation using Representation Conditioned Diffusion Model (RCDM) and (b) Classifier Based Evaluation for Comparative analysis.

vestigating the impact of dataset size and representation used for conditioning (DINOv2 [11], DINOv3 [16], and CLIP [12]).

- We show that, by scaling the synthetic dataset, RCDMs can surpass the performance of models trained solely on real data, highlighting their potential for large-scale dataset generation.
- We further demonstrate that RCDM generated images are effective for data augmentation, outperforming classical techniques such as RandAugment [4] and MixUp [19], and how the representation space used for conditioning allows simple sample filtering.

While this work demonstrates a first case study of using RCDMs for dataset generation, the results open up for many possibilities in future research, for improving or replacing datasets by means of generative synthesis.

2. Method

In this section, we describe the pipeline used for generation and validation of synthetic training data. The pipeline consists of two main stages: synthetic data generation using Representation Conditioned Diffusion Models (RCDMs) and evaluation using a classifier. Figure 1 illustrates the overall workflow, which is described in detail in the following subsections.

2.1. Synthetic Data Generation using RCDM

Training the RCDM involves two stages. First, input images are projected into a lower-dimensional representation space using a pre-trained encoder. Second, a latent diffusion model (LDM) [14] is trained using these representations as conditioning signals [9].

For the experiments, we train three RCDM models conditioned on representations extracted from the image encoder of CLIP [12] (also called unCLIP [13]), DINOv2 [11], and DINOv3 [16]. Specifically, we use the CLS token embeddings from the DINOv2 ViT-B/14 and DINOv3 ViT-B/16 architectures, each producing 768-dimensional representations, and embeddings from CLIP ViT-B/32 with a dimensionality of 512. Training is performed on the ImageNet-100 dataset, a subset of ImageNet containing 100 classes [15]. The dataset contains 130k images, all resized to a resolution of 256×256. Each model was trained for 125 epochs with a batch size of 16, resulting in approximately 1.16 million training steps.

For sampling, we use the Denoising Diffusion Probabilistic Model (DDPM) [7] with 100 inference steps. To increase dataset diversity beyond the original 130k images, images are generated using different random seeds, each corresponding to a distinct latent noise initialization. This produces diverse images while preserving semantic content through representation conditioning [9].

2.2. Classifier Based Evaluation

To evaluate the quality of the synthetic ImageNet-100 data in relation to the real dataset, we train separate classifiers on the real and synthetic datasets using the ResNet-50 [5] architecture. This evaluation measures how effectively models trained on synthetic data can learn discriminative features comparable to those learned from real data.

The training procedure follows the standard configuration used for training ResNet-50 on ImageNet in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Specifically, the classifier is trained using the same architectural setup, data preprocessing pipeline, and optimization strategy adopted in ILSVRC. By maintaining an identical training configuration across both datasets, the comparison provides a consistent and fair assessment of the representational quality and class-discriminative structure present in the generated synthetic images.

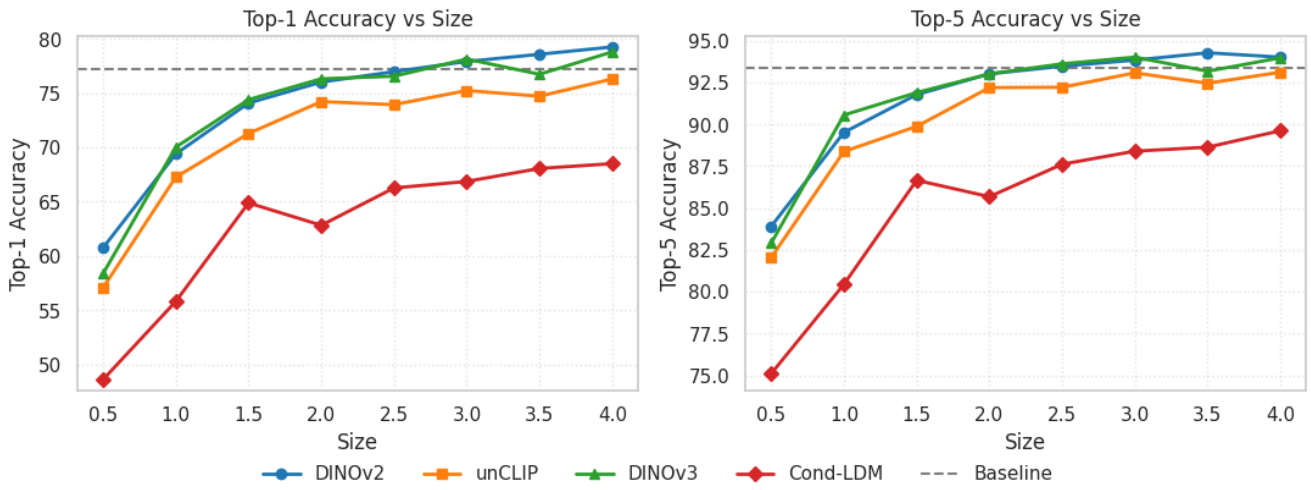

Figure 2. Scaling synthetic datasets for classifier training. Top-1 (left) and Top-5 (right) accuracy of a ResNet-50 trained on synthetic ImageNet-100 datasets generated using representation-conditioned diffusion models with DINOv2, DINOv3, and CLIP representations, compared to a class-conditioned LDM baseline and ResNet-50 trained on real ImageNet100 baseline.

3. Results

This section presents the experimental evaluation of the proposed framework. We assess the quality and usefulness of the generated synthetic images using quantitative downstream classification experiments on the ImageNet-100 dataset.

3.1. Synthetic Dataset Scaling

In this experiment, we investigate whether the performance gap between classifiers trained on real and synthetic data can be reduced using synthetic datasets generated by RCDMs, and further analyze this by scaling the synthetic dataset beyond the size of the real dataset to evaluate the impact of increased scale and diversity on downstream classification performance.

For this evaluation, we train classifiers based on the ResNet50 architecture under multiple training datasets. The baseline model is trained on the real ImageNet100 dataset, which achieves a Top-1 accuracy of 77.26% and a Top-5 accuracy of 93.38%. In addition, three classifiers are trained on synthetic datasets generated using RCDMs conditioned on representations extracted from DINOv2, DINOv3, and CLIP. As an additional baseline for synthetic data generation, we include a class-conditioned latent diffusion model (Cond-LDM [14]), which generates images conditioned solely on class labels.

The results shown in Figure 2 demonstrate that classifiers trained on synthetic datasets generated using representations from DINOv2 and DINOv3 outperform the baseline model trained on real data when the dataset size is increased to three times the size of the real dataset. The best result is obtained using the dataset generated with DINOv2 representations, achieving 79.26% Top-1 accuracy, when the dataset is scaled to four times the size of the real dataset. The dataset generated with DINOv3 representations performs comparably, reaching 78.8% Top-1 accuracy at the same scale. Although the synthetic dataset generated using CLIP representations does not exceed the baseline accuracy, it achieves a competitive Top-1 accuracy of 76.3% at four times the size of the real dataset and has the potential to surpass the baseline score based on the observed trend.

Another notable observation is the superior performance of datasets generated using RCDMs compared to the class-conditioned diffusion baseline. The classifier trained on data generated by Cond-LDM achieves a highest Top-1 accuracy of 68.5%, leaving a significant gap relative to the baseline model trained on real data. This highlights the importance of rich semantic conditioning via pre-trained vision encoders, which enables the diffusion model to better capture the underlying structure of the data distribution.

3.2. Filtering in Representation Space

To study the effect of filtering in representation space, we first generate synthetic datasets of twice the size of the real dataset using RCDMs conditioned on DINOv2, DINOv3, and CLIP. The generated images are then mapped back to the representation space, where class centroids are computed. For each class, the 25% nearest and 25% farthest representations are removed, retaining only the middle samples. The corresponding images form a filtered dataset with the same size as the real dataset. As shown in Table 1, this filtering achieves slight improvements over using unfiltered synthetic data, although a noticeable gap to the real-data baseline remains. Nevertheless, this result suggests

Table 1. Classification performance comparison on ImageNet100 with different synthetic datasets and sample filtering strategies.

Experiment	Size	Top-1 (%)	Top-5 (%)
Baseline	1.0	77.26	93.38
DINOv2-ResNet50	1.0	69.40	89.52
DINOv3-ResNet50	1.0	70.02	90.56
CLIP-ResNet50	1.0	67.26	88.38
Sample Filtering (DINOv2)	1.0	69.98	90.14
Sample Filtering (DINOv3)	**1.0**	**71.10**	**90.10**
Sample Filtering (CLIP)	1.0	66.22	88.16

that representation-space filtering can be a promising sampling strategy, which is difficult to attain in other generative frameworks such as text-to-image models.

3.3. Data Augmentation vs Synthetic Data

To further assess the effectiveness of the generated synthetic datasets for data augmentation, we compare their performance with traditional data augmentation techniques. Along with the baseline, we evaluate classifiers trained using common augmentation strategies including AutoAug [3], RandAug [4], Mixup [19], and ME-ADA [20]. These methods are known to improve generalization by increasing the diversity of the training data through transformations applied directly to real images.

As shown in Table 2, traditional augmentation techniques significantly improve the performance of the baseline model. In particular, Mixup achieves the highest Top-1 accuracy of 81.58%, followed closely by AutoAug with 81.28%. RandAug and ME-ADA also provide improvements over the baseline, reaching 79.88% and 80.14% Top-1 accuracy, respectively.

Classifiers trained on real data augmented with synthetic images generated by RCDMs achieve comparable or superior performance to standard augmentation methods. In particular, adding synthetic data generated with DINOv2 representations (twice the size of real dataset, marked as x2 in Table 2) achieves the best result with 82.22% Top-1 and 95.86% Top-5 accuracy, outperforming all evaluated augmentation strategies. Synthetic data generated using DINOv3 and CLIP representations also achieve competitive results, highlighting the effectiveness of representation-conditioned diffusion for providing an alternative to conventional data augmentation methods.

4. Discussion and Conclusion

We have demonstrated how RCDMs can pose as a powerful training data generation engine, which can be used to produce purely synthetic datasets or for augmentation purposes. The representation conditioning provides an effective self-supervised guidance mechanism for increasing quality and diversity, similar to how text prompts guide text-to-image models, and we have demonstrated that this improvement translates to improved training data value. In particular, representations from DINO-based encoders provide strong semantic conditioning signals that help reduce the domain gap between real and synthetic data and significantly improve the usefulness of generated datasets for downstream classification tasks. By scaling the size of a purely synthetic ImageNet100 dataset we are able to outperform training on the real images, which is a promising achievement for generative training data synthesis considering the complexity of the dataset. Despite the recent progress in generative deep learning, until now it has been difficult to reach similar or better performance compared to training on real data. For example, a class-conditioned diffusion model is still far behind, and representation conditioning provides a promising strategy for increasing training value.

Table 2. Classification performance comparison on ImageNet100 using different data augmentation methods and synthetic datasets generated by representation conditioned diffusion models.

Experiment	Size	Top-1 (%)	Top-5 (%)
Baseline	1.0	77.26	93.38
AutoAug	1.0	81.28	95.54
RandAug	1.0	79.88	95.14
Mixup	1.0	81.58	95.54
ME-ADA	1.0	80.14	94.62
Baseline + DINOv2	2.0	81.90	95.80
Baseline + DINOv3	2.0	81.08	95.36
Baseline + CLIP	2.0	81.78	95.30
Baseline + Cond-LDM	2.0	79.62	94.58
Baseline + x2 DINOv2	**3.0**	**82.22**	**95.86**
Baseline + x2 DINOv3	3.0	81.70	95.76
Baseline + x2 CLIP	3.0	80.74	95
Baseline + x2 Cond-LDM	3.0	80.68	94.94

We see many possible directions for future research using RCDMs for training data generation. Most centrally, there are many possibilities in improving the sampling of RCDMs through the representation space used for conditioning. For example, it is possible to transform representations to further improve training value, e.g., by performing interpolation, hard example mining or MixUp in this space. Also, while in general RCDMs provide high sample quality, there are still some low-quality generations such that a mechanism for removing outliers would likely improve the synthetic dataset. Additionally, further experiments are required for verifying the performance of RCDMs for training data generation on other datasets and under different data availability scenarios.

References

[1] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. *Transactions on Machine Learning Research*, 2023. 1

[2] Florian Bordes, Randall Balestriero, and Pascal Vincent. High fidelity visualization of what your self-supervised representation knows about. *arXiv preprint arXiv:2112.09164*, 2021. 1

[3] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. 4

[4] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 2, 4

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[6] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and XIAOJUAN QI. Is synthetic data from generative models ready for image recognition? In *The Eleventh International Conference on Learning Representations*, 2023. 1

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2

[8] Guillermo Jimenez-Perez, Pedro Osório, Josef Cersovsky, Javier Montalt-Tordera, Jens Hooge, Steffen Vogler, and Sadegh Mohammadi. Dino-diffusion: Scaling medical diffusion models via self-supervised pre-training. In *Annual Conference on Medical Image Understanding and Analysis*, pages 257–274. Springer, 2025. 1

[9] Nithesh Chandher Karthikeyan, Jonas Unger, and Gabriel Eilertsen. Evaluating representation conditioned diffusion models: A comparative study of representation encoders. *Available at SSRN 5772685*. 2

[10] Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method. *Advances in Neural Information Processing Systems*, 37:125441–125468, 2024. 1

[11] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[13] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 2

[16] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 2

[17] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015. 1

[18] Brandon Trabucco, Kyle Doherty, Max A Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[19] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2, 4

[20] Long Zhao, Ting Liu, Xi Peng, and Dimitris Metaxas. Maximum-entropy adversarial data augmentation for improved generalization and robustness. *Advances in Neural Information Processing Systems*, 33:14435–14447, 2020. 4