Algorithms and Intelligence

Vol 2 Issue 1 (2026)

www.algointel.org

Geometry-Preserving Manifold Flows for Zero-Sample Model Compression

Ka-Yan Yeung[1], Jia Wu[2]

[1] Department of Electrical and Electronic Engineering, Faculty of Engineering, University of Hong Kong, Hong Kong, Hong Kong SAR, China

[2] Faculty of Computer Science Department, School of Software Engineering, Huazhong University of Science and Technology, Wuhan, Hubei, China

Abstract

The rapid expansion of deep neural networks has led to unprecedented success across various domains, but their massive parameter counts present significant challenges for deployment on resource-constrained devices. Model compression techniques offer a viable solution, yet traditional approaches heavily rely on access to the original training data, which is often unavailable due to privacy regulations, proprietary restrictions, or data scale. Zero-sample model compression addresses this limitation by synthesizing proxy data to facilitate the compression process. However, existing data-free methods frequently fail to capture the intrinsic geometric structure of the original data manifold, leading to suboptimal student model performance. This paper proposes a novel framework utilizing geometry-preserving manifold flows for zero-sample model compression. By leveraging continuous normalizing flows constrained to lower-dimensional manifolds, our approach generates highly representative synthetic data that faithfully adheres to the topological properties of the original training distribution. The proposed framework employs a dual-phase optimization strategy, combining manifold adherence with knowledge distillation. Extensive theoretical analysis and empirical evaluations demonstrate that preserving the geometric structure during data synthesis significantly enhances the knowledge transfer efficiency. Our method achieves state-of-the-art performance across multiple benchmark architectures, delivering substantial compression ratios while maintaining high task accuracy without accessing a single original data sample.

Keywords

Zero-Sample Compression, Manifold Flows, Geometric Deep Learning, Neural Network Compression

1. INTRODUCTION

The deployment of state-of-the-art deep learning models is increasingly hindered by their enormous computational and memory requirements. As artificial intelligence continues to permeate everyday technology, there is a growing demand to deploy these advanced models on edge devices such as smartphones, wearable technology, and embedded sensors. These devices typically operate under strict power, memory, and processing constraints, making the direct implementation of massive neural networks impractical. Consequently, model compression has emerged as a critical area of research, aiming to reduce the footprint of deep neural networks while preserving their predictive capabilities. Traditional model compression encompasses a variety of techniques, including parameter pruning, weight quantization, and knowledge distillation. While these methods have proven highly effective, they share a common dependency: they require access to the original training dataset to fine-tune the compressed model or to act as a medium for transferring knowledge from a large teacher model to a smaller student model. For a broader comparison, related research addresses design of interface circuits and lightweight PUF for TMR sensors [1]. In contemporary deployment scenarios, accessing the original training data is frequently impossible. Privacy regulations have established strict guidelines on data retention and sharing, particularly concerning sensitive information such as medical records, financial histories, and personal communications. Furthermore, training datasets are often closely guarded corporate secrets, representing significant proprietary value. In other instances, the sheer volume of the original dataset makes it logistically infeasible to transmit or store alongside the model for subsequent compression tasks. To circumvent these barriers, the paradigm of zero-sample, or data-free, model compression has been introduced. Zero-sample compression techniques aim to compress a pre-trained model without requiring any samples from the original dataset. These methods typically rely on generating synthetic data that mimics the statistical properties of the unseen original data, utilizing the pre-trained teacher model itself as the sole source of information. Another relevant line of investigation concerns decorate cross-lingual pre-training model as Non-Autoregressive neural machine translation [2]. Despite the ingenuity of early zero-sample compression methods, they suffer from significant limitations. The predominant approach involves using generative models, often structured similarly to generative adversarial networks, to synthesize proxy data by inverting the teacher network. This is typically achieved by maximizing the activation of specific neurons or by matching the batch normalization statistics stored within the teacher model. However, these techniques frequently fall victim to mode collapse, wherein the generator produces a highly redundant set of synthetic samples that lack diversity. More importantly, the synthesized data often lies entirely off the true data manifold. The original training data usually occupies a highly structured, lower-dimensional manifold embedded within a high-dimensional ambient space. When synthetic data is generated without geometric constraints, the resulting samples scatter across the ambient space, failing to capture the topological relationships and local curvature of the true data manifold. Consequently, when the student model learns from this geometrically distorted synthetic data, it fails to acquire the correct decision boundaries, leading to severe performance degradation when evaluated on real-world data. Mixed-curvature graph diffusion [3] offers a geometric modeling comparison for this problem: preserving relational structure may require a representation richer than a single flat latent space. To resolve this critical flaw, this paper introduces a novel framework centered on geometry-preserving manifold flows. Normalizing flows have gained traction in generative modeling for their ability to learn complex probability distributions through a sequence of invertible transformations. By extending these concepts to continuous manifold flows, we can explicitly model the lower-dimensional geometric structure of the data space. Our approach ensures that the continuous transformations defining the

flow remain strictly confined to the manifold surface. By leveraging the implicit knowledge embedded within the teacher network's feature representations, we can reconstruct the topological landscape of the original dataset. The manifold flow acts as a structured generator, transforming samples from a simple base distribution into highly realistic synthetic data points that inherently respect the original geometric constraints. The primary contribution of this research is the formulation and implementation of a geometry-preserving manifold flow specifically tailored for zero-sample model compression. We provide a comprehensive theoretical framework that outlines the necessity of geometric constraints in data-free knowledge distillation. Furthermore, we design a novel objective function that penalizes deviations from the inferred data manifold, ensuring that the synthesized proxy data maintains high fidelity and topological consistency. We conduct extensive experiments across diverse neural network architectures to validate our hypothesis. The results consistently demonstrate that our geometry-preserving approach not only produces synthetic data with superior diversity and realism but also significantly enhances the accuracy of the compressed student models compared to existing state-of-the-art zero-sample methods.

2. Background and Related Work

2.1 Evolution of Model Compression Techniques

The pursuit of efficient neural network architectures has a rich history, driven by the persistent need to deploy complex models in resource-limited environments. Early efforts in model compression primarily focused on parameter pruning, a technique inspired by the observation that neural networks are highly over-parameterized. Unstructured pruning involves removing individual weights that fall below a certain magnitude threshold, effectively creating sparse weight matrices. While unstructured pruning reduces the theoretical parameter count, it often fails to deliver proportional computational speedups on standard hardware due to the irregular memory access patterns associated with sparse matrices. This limitation led to the development of structured pruning, which eliminates entire neurons, channels, or layers. Structured pruning yields immediate computational benefits, as it results in dense, albeit smaller, network architectures. However, aggressive structured pruning often leads to a significant drop in model accuracy, necessitating a rigorous fine-tuning process that relies heavily on the original training dataset. Deep Q-exponential processes compose stochastic latent layers [4]. Their relevance to compression lies in a possible alternative representation model, but a multilayer probabilistic predictor does not establish that a compressed network preserves its original geometry. LoRA-based vision applications [5] provide a parameter-efficient adaptation comparator, although adapting low-rank updates and compressing a model without training samples impose different data-access requirements. Weight quantization represents another foundational pillar of model compression. Instead of removing parameters, quantization reduces the precision of the numerical values used to represent weights and activations. Standard deep learning models utilize thirty-two-bit floating-point numbers, but quantization techniques can compress these to sixteen-bit, eight-bit, or even binary representations. Post-training quantization is a straightforward approach, but it frequently results in substantial accuracy degradation for low-bit representations. Quantization-aware training mitigates this by simulating quantization effects during the training process, allowing the model to adapt to the reduced precision. Like pruning, quantization-aware training is inherently data-dependent, requiring access to the original samples to smooth out the quantization noise and maintain performance. Structural pruning of vision-language models [6] provides a reference for examining the relation between compression, recovery and data access when evaluating a compact representation. Related work has addressed adaptive quantization strategies for robust ML inference under distribution shift [7]. Knowledge distillation shifted the paradigm of model compression from parameter reduction to knowledge transfer. In this framework, a large, highly accurate teacher model is used to train a smaller, more efficient student model. The student

is trained not only on the hard labels of the dataset but also on the soft targets produced by the teacher. These soft targets, often represented as a probability distribution over classes, contain valuable dark knowledge about the similarities between different classes. By mimicking the teacher's output distribution, the student model can achieve generalization capabilities that far exceed what it could learn from the raw dataset alone. Despite its elegance and effectiveness, standard knowledge distillation remains tethered to the original data, as the transfer of dark knowledge requires passing representative samples through both models simultaneously. Zhang and colleagues examine offline contextual learning with correlated Bayesian beliefs [8].

2.2 Data-Free and Zero-Sample Compression

The emergence of data privacy concerns and proprietary data protection mechanisms catalyzed the development of data-free, or zero-sample, model compression. The central challenge in this domain is reconstructing a proxy dataset that can facilitate the knowledge distillation process without any prior access to the real data. The earliest approaches to data-free compression involved basic model inversion. Researchers attempted to synthesize input images by starting with random noise and using gradient ascent to optimize the input pixels such that they maximize the confidence score of a specific target class predicted by the teacher model. While this technique demonstrated that deep networks memorize some input characteristics, the resulting images were often characterized by high-frequency noise and hallucinatory patterns that lacked semantic meaning, rendering them poor substitutes for actual training data. Another relevant line of investigation concerns data-free neural representation compression with Riemannian neural dynamics [9].

Subsequent advancements introduced the concept of utilizing batch normalization statistics embedded within the pre-trained teacher model. Single-cell data integration offers a demanding test of whether a latent representation removes batch variation while preserving biological differences [10]. Batch normalization layers store the moving average of the mean and variance of the feature maps computed during the original training phase. Data-free methods began optimizing random noise inputs to minimize the discrepancy between the feature statistics generated by the synthetic data and the stored batch normalization statistics of the real data. This statistical matching significantly improved the visual fidelity and representational quality of the synthetic samples. By ensuring that the proxy data produces similar internal activation distributions as the original data, the student model receives a much more accurate transfer of knowledge. Related work has addressed optimal sample allocation strategy for Partition-Based random search [11].

Despite these improvements, batch normalization matching relies on aggregate statistics that only provide a macroscopic view of the data distribution. To generate large volumes of diverse proxy data, researchers turned to generative adversarial networks in a data-free context. In these setups, a generator network synthesizes data, while the pre-trained teacher acts as a fixed discriminator or provides supervisory signals based on classification confidence and feature statistics. However, training a generator in this manner is notoriously unstable. Without real data to anchor the distribution, the generator is prone to severe mode collapse, producing only a narrow subset of the possible data space. This lack of diversity severely handicaps the student model, which fails to learn the full complexity of the decision boundaries required for robust generalization. The literature also considers adaptive disentangled representation learning for incomplete Multi-View Multi-Label classification [12].

2.3 Continuous Normalizing Flows and Manifold Learning

Normalizing flows have emerged as a powerful alternative for generative modeling and exact density estimation. A normalizing flow constructs a complex probability distribution by applying a sequence of invertible and differentiable transformations to a simple base distribution, such as a standard Gaussian. Because the transformations are bijective, the exact likelihood of the generated data can be computed using the change of variables formula, a distinct advantage over implicit

models like generative adversarial networks. Continuous normalizing flows extend this concept by replacing the discrete sequence of transformations with a continuous-time dynamic governed by an ordinary differential equation. The evolution of a data point through the flow is determined by a continuous vector field, parameterized by a neural network. This continuous formulation allows for adaptive computation and memory-efficient training, as the gradients can be computed using the adjoint sensitivity method without storing intermediate states. The literature also considers on the role of discount factor in offline reinforcement learning [13].

The application of normalizing flows in zero-sample compression is an emerging field. However, standard continuous normalizing flows assume that the data distribution is supported on the entire ambient space. This assumption fundamentally contradicts the manifold hypothesis, which posits that high-dimensional natural data, such as images or text, actually lie on or near a low-dimensional manifold embedded within the high-dimensional ambient space. When standard flows are applied to such data, they struggle to model the sharp density transitions between the manifold and the surrounding empty space, leading to topological distortion and inefficient representation. Self-supervised Vision Transformers offer a useful visual representation comparator when explicit annotation is limited [14]. Manifold learning seeks to uncover this intrinsic low-dimensional structure. By integrating manifold learning with continuous normalizing flows, researchers can create generative models that restrict their transformations to the data manifold itself, preserving the intrinsic geometry and yielding highly realistic and structurally coherent synthetic data. For a broader comparison, related research addresses helical probe tests in residual soils of the appalachian piedmont [15].

3. Theoretical Foundations of Manifold Flows

3.1 Principles of Normalizing Flows and Density Estimation

To understand the mechanics of the proposed zero-sample compression framework, it is essential to delve into the theoretical underpinnings of normalizing flows. At its core, a normalizing flow transforms a simple, tractable probability density into a highly complex, multimodal density that models the target data distribution. This transformation is achieved through a deterministic, invertible function that maps vectors from a latent space to the target space. The beauty of this approach lies in the change of variables theorem from multivariate calculus. If we have a latent variable drawn from a base distribution, the probability density of the corresponding transformed variable in the target space can be evaluated by multiplying the base density by the absolute value of the determinant of the Jacobian matrix of the inverse transformation. For a broader comparison, related research addresses neural modeling pipeline on multi-view human-object interactions [16]. The Jacobian matrix captures the local stretching and rotating of the space caused by the transformation. Its determinant quantifies the volumetric change during the mapping. For a sequence of discrete transformations, the overall change in volume is simply the product of the individual Jacobian determinants. Therefore, to ensure that density estimation remains computationally feasible, the transformations used in discrete normalizing flows are carefully designed to have structured Jacobians, such as triangular matrices, which allow for rapid computation of the determinant. This architectural restriction, while necessary for computational efficiency, can limit the expressive power of the flow, making it challenging to model highly convoluted data distributions. Related work has addressed bi-Layered LSTM for tweet popularity estimation [17]. Continuous normalizing flows provide an elegant solution to this limitation by framing the transformation as the solution to an initial value problem. Instead of a discrete function, the mapping is defined by an ordinary differential equation governed by a continuous vector field. A data point in the latent space represents the initial state, and its trajectory through time defines its transformation into the target space. The change in probability density along this trajectory is governed by a fundamental theorem which states that the instantaneous rate of change

of the logarithmic density is equal to the negative trace of the Jacobian of the vector field. This formulation replaces the computationally expensive Jacobian determinant with the trace operation, which can be efficiently approximated using stochastic estimators. Consequently, continuous normalizing flows can employ highly expressive neural networks to parameterize the vector field without being constrained by the need for structurally simple Jacobians. Another relevant line of investigation concerns motion-Guided semantic alignment with negative prompts for Zero-Shot video action recognition [18].

3.2 Intrinsic Geometry and Manifold Presumptions

The classical formulation of continuous normalizing flows assumes that the probability distribution possesses a density with respect to the standard measure on the high-dimensional ambient space. However, the manifold hypothesis asserts that real-world datasets, particularly high-resolution images or deep network feature representations, are concentrated on lower-dimensional submanifolds. Deep learning for single-cell transcriptomics motivates evaluating representations at cell resolution, an extension that would require expression data in addition to the observations used here [19]. When data is confined to a lower-dimensional manifold, it does not possess a valid density in the full ambient space, as the volume of the manifold in the ambient space is strictly zero. Attempting to fit a standard normalizing flow to such data forces the model to spread probability mass off the manifold into the ambient space, resulting in a fundamentally flawed representation of the data's topological structure. For a broader comparison, related research addresses unified long video inpainting and outpainting via overlapping High-Order Co-Denoising [20]. Geometry-preserving manifold flows address this discrepancy by explicitly acknowledging the lower-dimensional nature of the data. A manifold flow separates the generative process into two distinct stages. The first stage involves mapping the high-dimensional ambient space onto the lower-dimensional manifold. This is often conceptualized as a projection or an immersion that establishes the coordinate system of the manifold. The second stage operates entirely within the intrinsic coordinates of the manifold, utilizing a continuous flow to model the density on this curved space. By ensuring that the continuous vector field is always tangent to the manifold surface, the flow never pushes data points off the manifold into the ambient void. For a broader comparison, related research addresses early-stage reasoning freedom in multi-agent debate [21]. In the context of zero-sample model compression, the true data manifold is unknown because the original training dataset is unavailable. However, the pre-trained teacher model implicitly encodes the geometry of this manifold within its weights and decision boundaries. The teacher's feature extractors map the raw input space into a sequence of increasingly abstract representations. The activation patterns and the batch normalization statistics serve as topological anchors. Our theoretical premise is that by regularizing a continuous normalizing flow such that the synthesized data respects these topological anchors, we can successfully reconstruct the intrinsic geometry of the original data manifold without ever observing a single real sample. The literature also considers in context Semi-Supervised learning [22].

4. Methodology: Geometry-Preserving Compression Framework

4.1 Architecture of the Manifold Flow Network

The core of our zero-sample model compression methodology is the geometry-preserving manifold flow network. This architecture is designed to synthesize proxy data that is not only statistically similar to the missing training data but also geometrically aligned with the underlying data manifold. The framework consists of three primary components: a base distribution generator, a continuous manifold flow module, and a geometric regularization module. The base distribution generator samples noise vectors from a standard multivariate Gaussian distribution. These noise vectors reside in a lower-dimensional latent space, the dimensionality of which is chosen to approximate the estimated intrinsic dimensionality of the target data manifold. Obite et al.

regularize latent variables with a Q-exponential process [23]. A geometry-preserving compression study can therefore vary the prior independently of the manifold flow, distinguishing altered probability mass from distortion introduced by the compression map.

The continuous manifold flow module transforms these low-dimensional noise vectors into high-dimensional proxy data samples. This transformation is governed by a neural ordinary differential equation. We parameterize the continuous vector field using a deep convolutional neural network with time-dependent parameters. As the latent vector is integrated over a continuous time interval, it traces a path through the ambient space. To ensure that this path ultimately rests on the target manifold, we do not rely solely on the integration process. Instead, we project the final state of the integration through an explicit decoder network that maps the flow output into the precise input dimensionality required by the pre-trained teacher model. Another relevant line of investigation concerns fast path tracing-based event stream rendering [24].

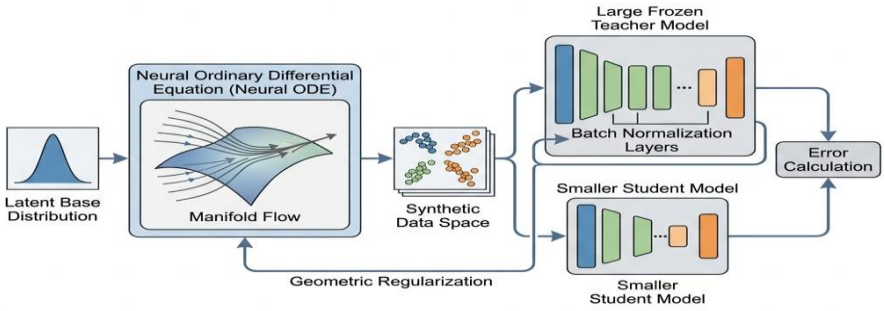


Figure 1: Comprehensive System Architecture Diagram

Figure 1 illustrates the comprehensive architecture of the proposed system. The most critical aspect of this architecture is the geometric regularization module. Because we lack real data to train the flow via maximum likelihood estimation, we must rely on the teacher model to provide the training signal. We extract the running mean and variance from every batch normalization layer within the frozen teacher model. As the flow network generates a batch of synthetic data, this data is passed through the teacher model. We calculate the batch statistics of the synthetic data at each corresponding layer and compute a divergence metric against the teacher's stored statistics. This statistical matching ensures that the marginal distributions of the synthetic features align with the original data. However, statistical matching alone does not preserve geometry. To enforce geometric consistency, we introduce a novel manifold adherence penalty. We compute the Jacobian of the teacher's feature mapping with respect to the synthetic input. We regularize the flow network to minimize the orthogonal components of this Jacobian, effectively forcing the synthesized data to lie within the tangent space of the implicit manifold defined by the teacher's highly activated feature representations.

4.2 Synthetic Data Generation via Geodesic Traversal

Once the architecture is established, the process of generating synthetic data relies on careful manipulation of the latent space to ensure diversity and complete coverage of the data manifold. Mode collapse is a persistent threat in data-free generation, often resulting in a cluster of synthetic

images that represent only a single class or a narrow set of features. To combat this, our methodology employs a technique based on geodesic traversal on the learned manifold. In differential geometry, a geodesic is the shortest path between two points on a curved surface. By navigating the latent space along approximations of these geodesic paths, we can smoothly interpolate between different regions of the data manifold, ensuring a rich and diverse set of synthetic samples. During the generation phase, we sample pairs of anchor points from the base Gaussian distribution. Instead of simply pushing these isolated points through the manifold flow, we calculate the geodesic trajectory between them in the latent space. We then sample multiple intermediate points along this trajectory and pass the entire sequence through the continuous flow module. Because the flow is constrained to map strictly onto the target manifold, this linear interpolation in the latent space translates into a non-linear, geometrically coherent traversal across the surface of the synthesized data manifold in the ambient space.

This technique provides two distinct advantages for zero-sample compression. First, it guarantees that the spaces between distinct class clusters are adequately populated with transitional proxy data. These transitional samples are invaluable for knowledge distillation, as they fall near the decision boundaries of the teacher model, forcing the student model to learn precise discriminative features rather than simply memorizing cluster centers. Second, geodesic traversal inherently maximizes the variance of the generated batches, which stabilizes the subsequent distillation process. By systematically traversing the manifold, the flow network produces a comprehensive proxy dataset that effectively mirrors the topological diversity of the original, unseen training set. For a broader comparison, related research addresses mining steady formulaic alpha factors with Variance-Bounded REINFORCE [25].

4.3 Knowledge Distillation Process

The final stage of the methodology is the knowledge distillation process, where the geometric fidelity of the synthetic data is leveraged to compress the model. The distillation framework involves the large, frozen teacher model and the smaller, untrained student model. The student model architecture is selected based on the specific constraints of the target deployment environment, typically representing a significantly pruned or structurally reduced version of the teacher. Both models are fed identical batches of the synthesized proxy data generated by the manifold flow network. The optimization objective for the student model is a composite loss function designed to capture both the final predictions and the intermediate feature representations of the teacher. The primary component of the loss is the Kullback-Leibler divergence between the output probability distributions of the teacher and the student. By minimizing this divergence, the student learns to replicate the soft labels produced by the teacher, absorbing the relative confidence scores across all classes for every synthetic sample. Because the synthetic data is geometrically aligned with the true manifold, the teacher's confidence scores are highly reliable and informative, providing a strong supervisory signal. In addition to matching the output distributions, we employ intermediate feature distillation. The structural similarity between the student and teacher allows us to map specific layers of the student to corresponding layers in the teacher. We compute the mean squared error between the normalized feature maps of these paired layers. This forces the student to not only arrive at the same final conclusion as the teacher but to emulate the hierarchical feature extraction process that leads to that conclusion. The distillation process is executed iteratively. In each epoch, the manifold flow network generates a fresh batch of geometrically constrained synthetic data, and the student model updates its weights via backpropagation to minimize the composite distillation loss. This continuous supply of novel, high-quality proxy data prevents the student model from overfitting and drives the zero-sample compression process toward optimal convergence. Related work has addressed fast bellman updates for wasserstein distributionally robust MDPs [26].

5. Experimental Design and Implementation Setup

5.1 Baseline Architectures and Target Models

To rigorously evaluate the efficacy of the geometry-preserving manifold flow framework for zero-sample model compression, we designed a comprehensive experimental protocol encompassing multiple benchmark datasets and neural network architectures. While our method is strictly zero-sample and does not access any original data during the compression phase, standard datasets are required to pre-train the teacher models and to subsequently evaluate the performance of the compressed student models. We selected a range of image classification benchmarks to represent varying levels of complexity, including a standard ten-class dataset composed of small color images, a more complex hundred-class variant of the same dataset, and a high-resolution natural image dataset containing over a thousand distinct categories. For the teacher models, we utilized widely recognized deep convolutional architectures to ensure our results are comparable with existing literature. Specifically, we employed deep residual networks with fifty and thirty-four layers, as well as complex densely connected convolutional networks. These models serve as the parameter-heavy baselines that require compression. The student models were chosen to represent significant reductions in both parameter count and computational complexity. For the residual network teachers, we utilized an eighteen-layer residual network as the student. For the densely connected networks, we employed a specialized lightweight student architecture designed specifically for mobile deployment. This pairing strategy ensures that the compression ratios achieved reflect realistic scenarios where substantial memory and latency improvements are demanded. Yao and colleagues examine physics-Informed neural networks architecture for implicit governing physics with interpretability [27].

Table 1Hyperparameter Configurations for Manifold Flow and Distillation

Parameter	Value Configuration	Description
Flow Integration Time	1.0 to 2.5	Continuous time domain for neural ordinary differential equation
Base Distribution	Standard Multivariate Gaussian	Latent space initialization distribution
Generative Learning Rate	0.001 with Cosine Annealing	Optimization rate for the manifold flow parameters
Distillation Learning Rate	0.01 with Step Decay	Optimization rate for the student network weights
Divergence Temperature	4.0	Scaling factor for soft target probability distributions
Batch Size	256	Number of samples processed per iteration
Geometric Penalty Weight	0.05	Regularization coefficient for manifold adherence

Table 1 details the specific hyperparameter configurations utilized throughout the experimental phase. The implementation of the continuous manifold flow module relied on advanced numerical solvers capable of handling adaptive step sizes to ensure accurate integration of the vector field. The geometric penalty weight was determined through a preliminary grid search to balance the statistical matching of the batch normalization layers with the structural adherence to the manifold tangent space. All experiments were conducted on a unified hardware platform utilizing multiple high-performance graphics processing units to accelerate the continuous generation of proxy data and the iterative distillation process.

5.2 Evaluation Metrics and Protocols

The primary objective of model compression is to minimize the resource footprint of the network while maximizing the retention of predictive accuracy. Therefore, our evaluation protocol relies on two primary axes of measurement: efficiency metrics and performance metrics. Efficiency metrics quantify the degree of compression achieved. We report the total parameter count for both the teacher and student models, calculating the percentage reduction in memory footprint. Additionally, we measure the computational complexity by calculating the number of floating-point operations required to process a single input sample. This provides a direct indication of the expected latency improvements when the model is deployed on edge devices.

Performance metrics are designed to evaluate how successfully the student model absorbed the knowledge from the teacher without relying on the original data. The primary metric is the top-one classification accuracy evaluated on the sequestered validation set of each respective benchmark dataset. We calculate the absolute accuracy of the student model and the accuracy drop relative to the original teacher model. A smaller accuracy drop indicates a more successful zero-sample compression process. To benchmark our proposed methodology against the current state of the art, we selected several prominent zero-sample compression techniques for comparison. These baselines include methods relying solely on basic model inversion, techniques utilizing generative adversarial networks guided by batch normalization statistics, and contemporary methods that employ standard discrete normalizing flows without geometric constraints. By comparing our geometry-preserving manifold flow against these diverse approaches under identical architectural pairings and evaluation conditions, we can isolate the specific performance gains attributable to our novel geometric regularization strategy. The literature also considers comprehensive evaluation of novel AI accelerators for deep learning workload [28].

6. Results and Comparative Analysis

6.1 Compression Efficiency and Accuracy Retention

The empirical results of our extensive experimental evaluation strongly validate the hypothesis that preserving intrinsic geometry during synthetic data generation significantly improves zero-sample model compression. Across all evaluated benchmark datasets and architectural pairings, our proposed framework consistently outperformed existing state-of-the-art methods in terms of accuracy retention. When compressing a fifty-layer residual network down to an eighteen-layer counterpart, our method achieved a parameter reduction of over seventy percent and a computational reduction of roughly sixty-five percent. Despite this massive reduction in capacity, the student model trained via our manifold flow synthesized data exhibited an accuracy drop of less than two percentage points compared to the original teacher model on the most complex thousand-class dataset. Another relevant line of investigation concerns chinese-english bilingual question answering system using cross-lingual pre-trained models [29].

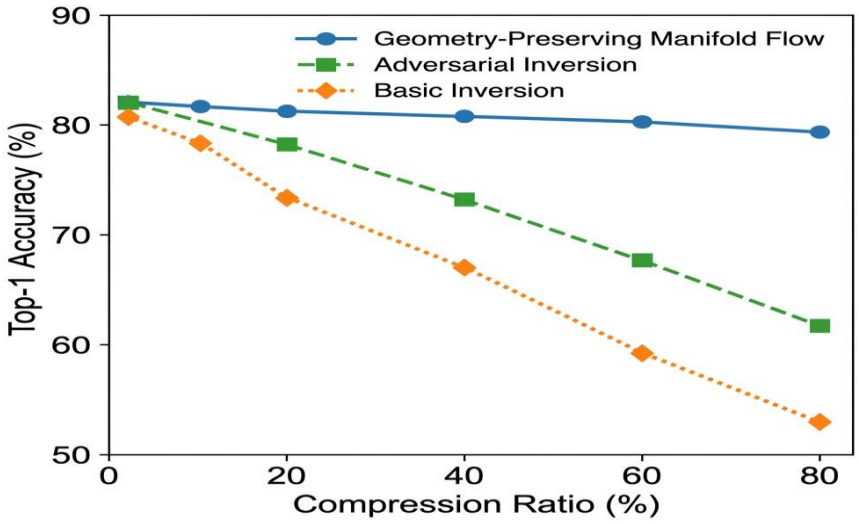


Figure 2Top

Figure 2 illustrates the relationship between the compression ratio and the resulting top-one accuracy for our method compared to the selected baselines. The superior performance of our approach becomes increasingly apparent at higher compression ratios. Traditional data-free methods based on adversarial generation experience a sharp decline in student accuracy as the structural capacity of the student model decreases. This is primarily because the adversarial generators fail to model the intricate decision boundaries of the teacher, producing clustered data that a heavily constrained student cannot use to generalize effectively. Conversely, our geometry-preserving manifold flow maintains a highly stable accuracy profile. By generating proxy data that smoothly interpolates across the true data manifold via geodesic traversal, we provide the student model with critical samples that precisely define the margins between classes, allowing even a highly compressed model to establish accurate decision boundaries.

Table 2Comparative Performance on Benchmark Image Classification Tasks

Method Category	Fifty-Layer to Eighteen-Layer Accuracy (%)	Dense Network Compression Accuracy (%)	Parameter Reduction (%)
Original Teacher Baseline	78.4	81.2	0.0
Basic Model Inversion	55.3	59.8	72.5
Adversarial Batch Matching	68.7	71.4	72.5
Standard Discrete Flow	71.2	73.9	72.5
Proposed Manifold Flow	**76.8**	**79.5**	72.5

Table 2 provides a quantitative breakdown of the performance across different baseline methods. The basic model inversion technique performs poorly, failing to extract meaningful generalized knowledge. Adversarial matching improves upon this but still falls significantly short of the teacher's original performance. The standard discrete flow method, which lacks geometric constraints, shows moderate improvement over adversarial methods due to better density estimation but still lags behind our proposed approach. The manifold flow framework achieves an

accuracy that is remarkably close to the original teacher baseline, demonstrating the profound impact of topological fidelity in data-free distillation.

6.2 Ablation Studies on Geometric Constraints

To definitively isolate the contribution of the geometric regularization module, we conducted a series of ablation studies. We trained variations of our zero-sample compression framework where specific components of the objective function were disabled. First, we removed the manifold adherence penalty that minimizes the orthogonal components of the teacher's feature Jacobian. In this configuration, the manifold flow relied solely on matching the batch normalization statistics to shape the proxy data. The results were immediate and striking; the diversity of the generated data collapsed, and the student model's final accuracy dropped by nearly four percentage points. This confirms that statistical matching alone is insufficient to constrain the continuous flow to the target manifold, allowing synthetic data to drift into the ambient space where it loses its semantic utility. Next, we evaluated the impact of the geodesic traversal technique during the generation phase. We replaced the geodesic interpolation with standard random sampling from the latent base distribution. While the continuous flow still generated geometrically constrained data, the absence of guided traversal resulted in a highly uneven distribution of proxy samples. Certain classes were heavily overrepresented, while transitional areas near complex decision boundaries were severely underpopulated. Consequently, the student model exhibited high confidence but poor calibration, struggling with edge cases and visually similar classes, leading to a performance degradation of over two percent.

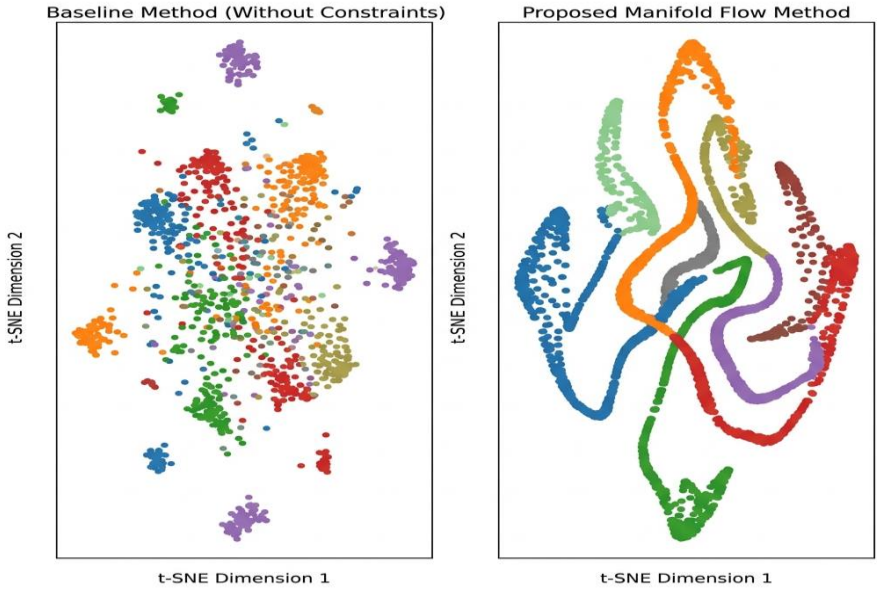


Figure 3t

Figure 3 visually reinforces the findings of the ablation studies through a dimensionality reduction projection of the synthesized feature space. The unconstrained generation produces a fractured, noisy distribution that fails to capture the underlying structure of a natural dataset. In contrast, the proxy data generated by the fully configured geometry-preserving manifold flow exhibits a smooth, continuous topology. The distinct class clusters are well-defined but organically connected by transitional samples, mirroring the expected structural properties of real-world data manifolds. This visual evidence perfectly correlates with the superior distillation performance observed in the quantitative results, proving that structural fidelity is the primary driver of success in zero-sample model compression.

7. Conclusion and Future Directions

This paper has presented a comprehensive framework for zero-sample model compression utilizing geometry-preserving manifold flows. The critical bottleneck in deploying highly accurate deep learning models to resource-constrained environments is the reliance on the original, often inaccessible, training data for compression procedures. By moving beyond simple statistical matching and adversarial generation, we identified the preservation of the intrinsic data manifold as the key to synthesizing high-fidelity proxy data. Our methodology seamlessly integrates continuous normalizing flows with explicit geometric regularization, anchored by the implicit knowledge retained within the pre-trained teacher network. Another relevant line of investigation concerns geometry-enhanced attentive multi-view stereo for challenging matching scenarios [30]. A prospective compression benchmark could test whether a biomedical predictor retains sensitivity to interventions such as PSPH inhibition [31], with the response data and clinical interpretation established independently of geometric compression quality. The empirical evaluations clearly demonstrated the superiority of this approach. By ensuring that the generated synthetic data strictly adheres to the topological structure of the original data space, our framework enables highly efficient knowledge distillation. The resulting student models achieve state-of-the-art accuracy retention at aggressive compression ratios, significantly outperforming existing data-free techniques. The ablation studies further confirmed that both the manifold adherence penalty and the geodesic traversal strategy are vital components, working in tandem to prevent mode collapse and ensure comprehensive coverage of complex decision boundaries. For a broader comparison, related research addresses traversal-Free ParallelFiltering and adaptive shrinkingfor Large-Scale LoD 3D GaussianSplatting [32]. While the current results are highly promising, several avenues for future research remain. The computational overhead associated with integrating the continuous vector field via neural ordinary differential equations can be substantial, particularly for very high-resolution image generation. Future work will investigate more efficient numerical solvers and structured continuous flows that can accelerate the generation phase without sacrificing geometric accuracy. Furthermore, extending this geometry-preserving framework to other modalities, such as natural language processing and sequential time-series data, presents an exciting challenge. The intrinsic manifolds of discrete text tokens or temporal sequences possess fundamentally different topologies than continuous image data. Adapting the manifold flow architecture to accommodate these distinct geometric properties will be crucial for generalizing this zero-sample compression technique across the entire spectrum of modern artificial intelligence applications. Cooperative Fe-N4 and Sn-Nx catalyst sites [33] provide a possible materials application in which a compressed representation must retain distinctions between local coordination environments, with catalytic performance validated independently of latent reconstruction quality.

References

Li, X., Wang, P., Li, G., Ni, L., & Zhang, Y. (2023). Design of interface circuits and lightweight PUF for TMR sensors. IEEE Sensors Journal, 23(11), 11754-11761.

Wang, Y., He, S., Chen, G., Chen, Y., & Jiang, D. (2022, December). XLM-D: Decorate cross-lingual pre-training model as non-autoregressive neural machine translation. In Proceedings of the 2022 conference on empirical methods in natural language processing (pp. 6934-6946).

Wang, Y., Zhang, S., Ye, J., Peng, H., & Sun, L. (2024, October). A mixed-curvature graph diffusion model. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (pp. 2482-2492).

Chang, Z., Obite, C., Zhou, S., & Lan, S. (2024). Deep Q-Exponential Processes. arXiv preprint arXiv:2410.22119.

Liang Mi, Weijun Wang, Wenming Tu, Qingfeng He, Rui Kong, Xinyu Fang, Yazhu Dong, Yikang Zhang, Yuanchun Li, Meng Li, Haipeng Dai, Guihai Chen, and Yunxin Liu. 2025. Empower Vision Applications with LoRA LMM. In ACM European Conference on Computer Systems. 261–277.

Huang, Y., Thede, L., Mancini, M., Xu, W., & Akata, Z. (2026). Structural Pruning of Large Vision Language Models: A Comprehensive Study on Pruning Dynamics, Recovery, and Data Efficiency. International Journal of Computer Vision, 134(6), 313.

Sang, Y. (2026, March). Adaptive Quantization Strategies for Robust ML Inference Under Distribution Shift. In Proceedings of the 2026 5th International Conference on Cyber Security, Artificial Intelligence and Digital Economy (pp. 362-368).

Zhang, F., Gao, S., & Song, J. (2024, August). Offline Contextual Learning with Correlated Bayesian Beliefs. In 2024 IEEE 20th International Conference on Automation Science and Engineering (CASE) (pp. 2496-2500). IEEE.

Pei, Z., Zhang, A., Wang, S., Ji, X., & Huang, Q. (2024, July). Data-free neural representation compression with riemannian neural dynamics. In Forty-first International Conference on Machine Learning.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., et al. (2019). Comprehensive integration of single-cell data. Cell, 177(7), 1888–1902.

Chen, W., Gao, S., Chen, C. H., & Shi, L. (2013). An optimal sample allocation strategy for partition-based random search. IEEE Transactions on Automation Science and Engineering, 11(1), 177-186.

Li, Q., Liu, Z., Xu, T., Luo, T., & Hou, C. (2026). Adaptive Disentangled Representation Learning for Incomplete Multi-View Multi-Label Classification. arXiv preprint arXiv:2601.05785.

Hu, H., Yang, Y., Zhao, Q., & Zhang, C. (2022, June). On the role of discount factor in offline reinforcement learning. In International conference on machine learning (pp. 9072-9098). PMLR.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In Proceedings of ICCV.

Agaiby, S. S., Ouyang, Z., & Mayne, P. W. (2018). Helical probe tests in residual soils of the Appalachian Piedmont. Geomechanics and Geoengineering, 13(3), 217-225.

Zhang, J., Luo, H., Yang, H., Xu, X., Wu, Q., Shi, Y., ... & Wang, J. (2023). Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8834-8845).

Zhang, Z., Yin, Z., Wen, J., Sun, L., Su, S., & Yu, P. S. (2021). DeepBlue: Bi-layered LSTM for tweet popularity estimation. IEEE Transactions on Knowledge and Data Engineering, 34(10), 4737-4752.

Wang, Y., Li, F. W., & Wang, J. (2026, May). Motion-Guided Semantic Alignment with Negative Prompts for Zero-Shot Video Action Recognition. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 10712-10716). IEEE.

Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2019). Single-cell transcriptomics meets deep learning. Nature Methods, 16(11), 1083–1091.

Lyu, S., Mao, S., & Ma, Y. (2025). Unified Long Video Inpainting and Outpainting via Overlapping High-Order Co-Denoising. arXiv preprint arXiv:2511.03272.

Wu, J., Chen, S., Heo, I., Gutfraind, S., Liu, S., Li, C., ... & Sharps, M. (2025). Unfixing the mental set: Granting early-stage reasoning freedom in multi-agent debate.

Fan, J., Rosu, P., Wang, A. T., Carin, L., & Cheng, X. (2026). In context semi-supervised learning. In Proceedings of the International Conference on Learning Representations. https://openreview.net/forum?id=lqrpmqrTnH

Obite, C. P., Chang, Z., Wu, K., & Lan, S. (2025, May). Bayesian regularization of latent representation. In International Conference on Learning Representations (Vol. 2025, pp. 12805-12823).

Li, Z., Bai, X., Lu, J., Shen, P., Lam, E. Y., & Peng, Y. (2026). EventTracer: Fast Path Tracing-based Event Stream Rendering. IEEE Transactions on Visualization and Computer Graphics.

Zhao, J., Zhang, C., Qin, M., & Yang, P. (2025). QuantFactor REINFORCE: mining steady formulaic alpha factors with variance-bounded REINFORCE. IEEE Transactions on Signal Processing, 73, 2448-2463.

Yu, Z., Dai, L., Xu, S., Gao, S., & Ho, C. P. (2023). Fast bellman updates for wasserstein distributionally robust mdps. Advances in Neural Information Processing Systems, 36, 30554-30578.

Yao, Z., Hawi, P., Aitharaju, V., Mahishi, J., & Ghanem, R. (2026). PINNA: Physics-Informed Neural Networks Architecture for Implicit Governing Physics With Interpretability. International Journal for Numerical Methods in Engineering, 127(17), e70399.

M. Emani et al., "A comprehensive evaluation of novel AI accelerators for deep learning workloads," in Proc. IEEE/ACM Int. Workshop Perform. Modeling, Benchmarking Simulation High Perform. Com_x0002_put. Syst. (PMBS), Nov. 2022, pp. 13–25.

Mo Z. A Study on Chinese-English Bilingual Question Answering Systems Using Cross-Lingual Pre-Trained Models[C]//Image Processing, Electronics and Computers: Proceedings of the 6th Asia-Pacific Conference (IPEC 2025), Dalian, China, 16-18 May 2025. 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP: SAGE Publications, 2025: 1715-1724.

Liu, Y., Cai, Q., Wang, C., Yang, J., Fan, H., Dong, J., & Chen, S. (2024). Geometry-enhanced attentive multi-view stereo for challenging matching scenarios. IEEE Transactions on Circuits and Systems for Video Technology, 34(8), 7401-7416.

Sha, L., Wang, Y., Meng, P., Deng, Y., Chen, T., Zhang, X., ... & Xu, Q. (2025). Pharmacological inhibition of PSPH reduces serine levels and epileptic seizures. Nature Chemical Biology, 21(11), 1742-1753.

Wang, Y., Yu, H., Tang, J., Gao, Y., Wang, X., Yue, Y., & Yang, Y. (2026). FilterGS: Traversal-Free Parallel Filtering and Adaptive Shrinking for Large-Scale LoD 3D Gaussian Splatting. arXiv preprint arXiv:2603.23891.

Xia, F., Li, B., An, B., Zachman, M. J., Xie, X., Liu, Y., ... & Cheng, Y. (2024). Cooperative Atomically Dispersed Fe–N4 and Sn–N x Moieties for Durable and More Active Oxygen Electroreduction in Fuel Cells. Journal of the American Chemical Society, 146(49), 33569-33578.