Prior-Guided Deep Inversion for Few-Shot Knowledge Distillation

Jiahe Wang, Yongxian Wei, Tangyu Jiang, and Chun Yuan*
Tsinghua Shenzhen International Graduate School
Tsinghua University
Shenzhen, China
wang-jh24@mails.tsinghua.edu.cn, weiyx23@mails.tsinghua.edu.cn
jiangtangyu@sz.tsinghua.edu.cn, yuanc@sz.tsinghua.edu.cn

***Abstract*—Data-free knowledge distillation (DFKD) enables model compression without access to original training data, but it fundamentally suffers from being an ill-posed inverse problem, as matching Batch Normalization (BN) statistics alone leaves higher-order moments unconstrained. While few-shot scenarios offer valuable priors, effectively leveraging them to regularize this inversion remains an open challenge. To address this, we propose Prior-Guided Deep Inversion (PGDI), a framework that formulates synthesis as a constrained distribution matching problem. We introduce a dual-space guidance mechanism to rigorously shrink the solution space: (i) in the feature space, we minimize the Maximum Mean Discrepancy (MMD) to enforce high-order moment matching in a Reproducing Kernel Hilbert Space (RKHS), ensuring topological consistency; (ii) in the decision space, we employ a hybrid calibration objective that functions as a geometric bias-variance trade-off, balancing class-center estimation (bias reduction) with instance-level diversity (variance preservation). Furthermore, to address the non-stationary nature of the generative process, we design a confidence-aware curriculum that dynamically modulates the distillation trust region, transitioning from high-fidelity priors to the augmented manifold. Extensive experiments on CIFAR-10, CIFAR-100, and Tiny-ImageNet demonstrate that PGDI consistently outperforms state-of-the-art baselines by effectively recovering the target manifold from sparse priors. The code is available at: https://github.com/wang-jh24/PGDI.**

***Index Terms*—Knowledge Distillation, Few-shot Learning, Model Inversion**

I. Introduction

Model compression is critical for deploying deep neural networks on mobile and edge devices constrained by memory and computational resources. Knowledge Distillation (KD) [1] has emerged as a standard paradigm to transfer the generalization capability of a cumbersome teacher to a compact student. Conventional KD assumes access to the original training data. However, in privacy-sensitive domains, original data is often inaccessible due to strict regulations or transmission bottlenecks [2]–[4].

To bypass data dependency, Data-free Knowledge Distillation (DFKD) [5], [6] synthesizes surrogate data by inverting the pre-trained teacher. The prevailing approach, DeepInversion [6], optimizes a generator to recover images that match the Batch Normalization (BN) statistics stored in the teacher.

* Corresponding author.

However, we identify a fundamental theoretical limitation in existing DFKD approaches: Model inversion via BN matching is an ill-posed inverse problem. Mathematically, BN statistics only constrain the first and second-order moments (mean and variance) of the synthesized distribution. This leaves the higher-order moments unconstrained, resulting in an infinitely large solution space that contains many semantically meaningless samples (e.g., "dreamed" images with correct statistics but distorted semantics). While recent works attempt to introduce auxiliary regularizers [7], [8], they struggle to recover the true data topology without external guidance, often leading to *semantic drift* where the student learns from off-distribution artifacts.

In practical scenarios, strictly data-free constraints are rare. A small "support set" (few-shot samples) is often available or can be retained under compliance allowances [9], [10]. Yet, existing methods fail to utilize this valuable data effectively—either overfitting to the few samples or naively mixing them with synthetic noise.

In this work, we propose Prior-Guided Deep Inversion (PGDI), a framework that treats few-shot samples not merely as training data, but as topological priors to regularize the ill-posed inversion problem. We formulate the synthesis process as a *constrained distribution matching* task. Specifically, to rigorously shrink the solution space of the generator, we introduce a dual-guidance mechanism:

- **Feature Space (Topological Consistency):** We propose to match *high-order moments* in a Reproducing Kernel Hilbert Space (RKHS). By minimizing the Maximum Mean Discrepancy (MMD) between synthesized and real features, we ensure the generated manifold is topologically consistent with the real data distribution.
- **Decision Space (Geometric Calibration):** We introduce a geometric *bias-variance trade-off* mechanism. We align synthesized logits with class prototypes to reduce estimation bias while anchoring them to nearest real instances to preserve intra-class variance, thereby preventing mode collapse.

Furthermore, acknowledging that the generator is a non-stationary process with time-varying Signal-to-Noise Ratio (SNR), we design a Confidence-Aware Curriculum. This

strategy dynamically modulates the distillation "trust region," transitioning from high-fidelity real priors (anchoring phase) to the augmented synthetic manifold (exploration phase).

Our main contributions are summarized as follows:

1) We identify the ill-posed nature of BN-based inversion as the root cause of semantic drift and propose PGDI to regularize this problem using few-shot priors.
2) We develop a dual-space guidance mechanism that enforces high-order moment matching in RKHS and geometric calibration in the decision space, ensuring both semantic fidelity and diversity.
3) We introduce a trust-region curriculum strategy that handles the non-stationary dynamics of the generator, effectively balancing stability and plasticity.
4) Extensive experiments on CIFAR-10, CIFAR-100, and Tiny-ImageNet demonstrate that PGDI consistently outperforms state-of-the-art DFKD and few-shot KD methods.

II. Related Work

A. Few-shot Knowledge Distillation

Few-shot knowledge distillation (FSKD) transfers teacher knowledge to a compact student from a small target set. Prior works typically focus on architectural modifications (e.g., cross-layer constraints or module grafting) [9]–[11], data augmentation [12], or self-supervised and meta-learning objectives [13], [14] to stabilize transfer. However, most FSKD methods treat the few-shot set as ordinary training data [15]–[19]; principled mechanisms for leveraging these scarce samples as distributional priors or explicitly balancing real and synthesized signals remain under-explored.

B. Data-free Knowledge Distillation

Data-free knowledge distillation (DFKD) synthesizes surrogate data without accessing the original dataset. Foundational approaches rely on model meta-data [5], maximizing teacher responses [20], modeling the output space [21], or matching Batch Normalization (BN) statistics [6]. Subsequent works introduce contrastive or meta-learning signals to improve generative diversity and robustness [7], [8], [22]–[30]. While DFKD provides strong surrogate generation, most approaches overlook the availability of small real dataset. Naive incorporation of such data frequently leads to overfitting, leaving the principled joint use of real and synthesized data under few-shot constraints an open problem.

III. Methodology

A. Problem Formulation and Theoretical Framework

Let $\mathcal{T} : \mathcal{X} \to \mathcal{Y}$ be a pre-trained teacher network with parameters $\theta_{\mathcal{T}}$, and $\mathcal{S}$ be a student network. In the stringent setting of few-shot data-free distillation, the original training distribution $\mathbb{P}_{\mathcal{X}}$ is inaccessible. Instead, we are restricted to a sparse support set $\mathcal{D}_{\text{few}} = \{(x_i, y_i)\}_{i=1}^{N}$, containing only K samples per class ($K \ll |\mathcal{D}_{\text{train}}|$).

The central challenge is that reconstructing $\mathbb{P}_{\mathcal{X}}$ solely from the teacher's internal statistics is an ill-posed inverse problem. Standard data-free methods [6] optimize a generator $\mathcal{G}$ to match the Batch Normalization (BN) statistics of $\mathcal{T}$. While necessary, BN matching only constrains the first and second-order moments of the synthesized distribution $\mathbb{P}_{\mathcal{G}}$. This leaves the higher-order statistics unconstrained, resulting in an infinitely large solution space $\Omega_{\mathcal{G}}$ that includes many semantically meaningless samples (e.g., noisy artifacts with correct means and variances).

To regularize this ill-posed problem, we propose Prior-Guided Deep Inversion (PGDI). As illustrated in Fig. 1, we leverage the few-shot samples $\mathcal{D}_{\text{few}}$ not merely as data, but as *structural priors* to define the topology of the target manifold. We formulate the synthesis process as a *constrained distribution matching* problem:

$$\min_{\mathcal{G}} \mathcal{D}_{\text{RKHS}} \left(\mathbb{P}_{\mathcal{G}} \| \hat{\mathbb{P}}_{\text{real}} \right) + \lambda \mathcal{R}_{\text{geo}}(\mathbb{P}_{\mathcal{G}}), \tag{1}$$

where the first term enforces topological consistency in the feature space, and the second term enforces geometric calibration in the decision space.

B. Stage I: Generative Modeling via Dual-Space Matching

As depicted in Fig. 1(a), the first stage of PGDI involves freezing the teacher $\mathcal{T}$ and optimizing a generator $\mathcal{G}$ to synthesize a proxy dataset $\mathcal{D}_{\text{syn}}$. The fundamental theoretical challenge here is the ill-posedness of data-free inversion: simply matching the teacher's internal statistics (e.g., BN mean and variance) yields an infinitely large solution space $\Omega_{\mathcal{G}}$, containing many samples that satisfy statistical constraints but lack semantic fidelity.

To rigorously shrink this solution space and recover the true data manifold, we introduce a dual-guidance mechanism that enforces consistency in two complementary vector spaces: the *Deep Feature Space* (for semantic topology) and the *Decision Logit Space* (for boundary geometry).

1) Manifold Regularization (Base Inversion): We start with the standard inversion objective to ensure the synthesized images $\hat{x}$ fall within the valid input domain recognized by $\mathcal{T}$. This "Base Inversion" (labeled as "Inversion Loss" in Fig. 1(a)) serves as a coarse regularization term. It minimizes the L_2 distance between the running statistics (mean $\mu_l^{\mathcal{T}}$ and variance $\sigma_l^{\mathcal{T}}$) stored in the teacher's Batch Normalization (BN) layers and those of the synthesized batch:

$$\mathcal{L}_{\text{BN}} = \sum_{l} \left(\|\mu_l(\hat{x}) - \mu_l^{\mathcal{T}}\|_2 + \|\sigma_l(\hat{x}) - \sigma_l^{\mathcal{T}}\|_2 \right). \tag{2}$$

Combined with a class-conditional constraint $\mathcal{L}_{\text{cls}}$ and an adversarial entropy term $\mathcal{L}_{\text{adv}}$, this defines a "loose envelope" around the target manifold. However, matching first and second-order moments is insufficient to capture complex non-Gaussian distributions, often leading to "semantic drift" where images possess correct textures but distorted structures.

2) Feature Space: High-Order Moment Matching in RKHS: To enforce strict topological consistency, we introduce Feature-level Alignment (middle block of Fig. 1(a)). We view the feature extraction process as a mapping $\phi : \mathcal{X} \to \mathcal{H}$ into a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$.

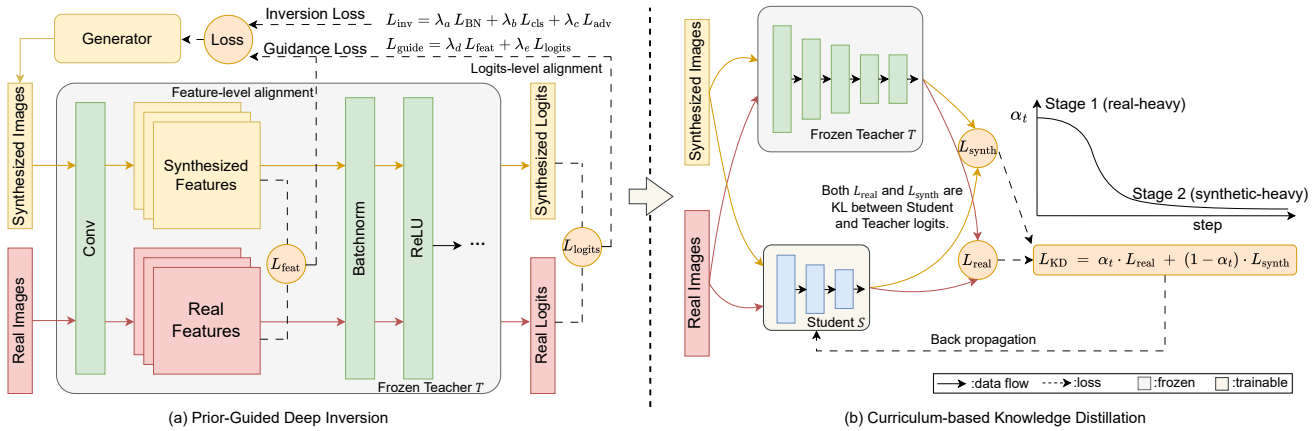

Fig. 1. **The Architectural Framework of PGDI.** (a) **Prior-Constrained Generative Modeling**: To resolve the ill-posed nature of inversion, the generator $\mathcal{G}$ is optimized under dual-space constraints. We introduce *Feature-level Alignment* (yellow block) to minimize distributional discrepancy in RKHS via MMD, and *Logits-level Alignment* (red block) to perform geometric calibration, balancing class-center bias and instance variance. (b) **Confidence-Aware Distillation**: The student $\mathcal{S}$ learns from both Real Images (red flow) and Synthesized Images (yellow flow). Given the non-stationary nature of the generator, a dynamic schedule α_t (curve graph) modulates the distillation loss $\mathcal{L}_{\text{KD}}$, smoothly transitioning the "Trust Region" from high-fidelity priors to the augmented manifold.

Standard BN matching is limited to low-order statistics. In contrast, we employ the Maximum Mean Discrepancy (MMD) distance. Theoretical results [31] establish that if the kernel $k(\cdot,\cdot)$ is characteristic (e.g., a Gaussian RBF kernel), minimizing MMD is equivalent to matching all higher-order moments of the two distributions. We construct a multi-scale feature representation $\phi(\cdot)$ by concatenating global pooled features from multiple depths of $\mathcal{T}$, ensuring alignment across both low-level texture and high-level semantics.

The objective is to minimize the discrepancy between the synthesized distribution $\mathbb{P}_{\mathcal{G}}$ and the empirical distribution $\hat{\mathbb{P}}_{\text{real}}$ defined by the few-shot priors:

$$\begin{aligned}\mathcal{L}_{\text{MMD}} &= \left\| \mathbb{E}_{\hat{x}\sim\mathbb{P}_{\mathcal{G}}}[\phi(\hat{x})] - \mathbb{E}_{x\sim\hat{\mathbb{P}}_{\text{real}}}[\phi(x)] \right\|_{\mathcal{H}}^2 \\ &= \mathbb{E}[k(\hat{x},\hat{x}')] - 2\mathbb{E}[k(\hat{x},x)] + \mathbb{E}[k(x,x')].\end{aligned} \tag{3}$$

By minimizing Eq. (3), we force the generator to mathematically "anchor" the synthesized manifold to the real few-shot samples in the high-dimensional feature space, effectively eliminating out-of-distribution artifacts.

3) Decision Space: Geometric Calibration via Bias-Variance Trade-off: While feature alignment ensures internal semantic correctness, the decision boundaries in the few-shot regime are highly susceptible to overfitting. A generator trained solely to minimize classification loss tends to suffer from *mode collapse*, producing samples that cluster tightly around the class mean while ignoring the intra-class spread. We address this via Logits-level Alignment (right block of Fig. 1(a)), formulated as a geometric Bias-Variance Trade-off.

Let $\hat{\mathbf{z}} = \mathcal{T}(\hat{x})$ be the synthesized logits and Z_c be the set of real logits for class c. We propose a hybrid loss $\mathcal{L}_{\text{logits}}$ composed of two complementary regularization terms:

$$\mathcal{L}_{\text{logits}} = \beta \cdot \mathcal{L}_{\text{proto}} + (1-\beta) \cdot \mathcal{L}_{\text{anchor}}. \tag{4}$$

(1) Bias Reduction (Global Prototype Matching): The first term, $\mathcal{L}_{\text{proto}}$, enforces Class-level Compactness. We define the prototype $\bar{\mathbf{p}}_c$ as the mean temperature-scaled probability of real samples and minimize the KL divergence:

$$\mathcal{L}_{\text{proto}} = D_{\text{KL}}\left(\sigma(\hat{\mathbf{z}}/\tau) \| \bar{\mathbf{p}}_c\right). \tag{5}$$

From a statistical perspective, this term reduces the *estimation bias* by pulling synthesized samples towards the stable class center. It ensures that the generator captures the dominant semantic direction of the class, preventing the synthesis of ambiguous boundary samples.

(2) Variance Preservation (Local Instance Anchoring): However, solely optimizing compactness leads to a collapsed distribution with zero variance. To counter this, $\mathcal{L}_{\text{anchor}}$ enforces Instance-level Diversity. We force each synthesized logit to anchor to its nearest neighbor in the real support set Z_c:

$$\mathcal{L}_{\text{anchor}} = \min_{\mathbf{z}\in Z_c} \|\hat{\mathbf{z}} - \mathbf{z}\|_2^2. \tag{6}$$

This term prevents mode collapse by ensuring the synthesized distribution "covers" the support of the real data. It explicitly preserves the *intra-class variance* (e.g., variations in pose or background) present in the few-shot priors.

This dual objective can be interpreted as calibrating the *geometry* of the synthesized manifold. $\mathcal{L}_{\text{proto}}$ constrains the "center of mass" (Bias), while $\mathcal{L}_{\text{anchor}}$ maintains the "radius of gyration" (Variance). Together, they produce a high-fidelity proxy dataset that is both accurate and diverse.

C. Stage II: Confidence-Aware Curriculum Distillation

Once the synthetic pool is constructed (or dynamically updated), we proceed to the distillation stage shown in Fig. 1(b). Crucially, the generator $\mathcal{G}$ is a non-stationary process. During the early optimization phases, $\mathcal{G}$ has not yet converged, and the synthesized data $\mathcal{D}_{\text{syn}}$ suffers from a low Signal-to-Noise Ratio (SNR). Conversely, the real data $\mathcal{D}_{\text{few}}$ is high-fidelity but sparse.

Directly training on a fixed mixture of these sources leads to suboptimal convergence. To handle this dynamic SNR, we propose a Dynamic Trust-Region Strategy (Curriculum

TABLE I
CIFAR-10/100 AND TINY-IMAGENET RESULTS (TOP-1%). **T.** DENOTES THE TEACHER ACCURACY, AND THE **BEST** RESULTS FOR EACH ROW ARE HIGHLIGHTED IN BOLD.

Dataset	Teacher	Student	Accuracy (%)									
			T.	DeepInv	DAFL	ZSKD	FAST	CMI	DFQ	SSD-KD	PGDI_{5s}	PGDI_{10s}
CIFAR-10	ResNet-34	ResNet-18	95.70	93.26	92.22	93.32	94.05	94.84	94.61	94.26	94.74	**95.10**
	VGG-11	ResNet-18	92.25	90.36	81.10	89.46	90.53	91.13	90.84	90.67	91.22	**91.32**
	WRN-40-2	WRN-16-1	94.87	83.04	65.71	83.74	89.29	**90.01**	86.14	89.96	88.45	89.15
	WRN-40-2	WRN-40-1	94.87	86.85	81.33	86.07	92.51	92.78	91.69	93.23	92.89	**93.23**
	WRN-40-2	WRN-16-2	94.87	89.72	81.55	89.66	92.45	92.52	92.01	**93.11**	92.73	92.83
CIFAR-100	ResNet-34	ResNet-18	78.05	61.32	74.47	67.74	74.34	77.04	77.01	75.16	76.96	**77.13**
	VGG-11	ResNet-18	71.32	54.13	54.16	54.31	67.44	70.56	66.21	68.77	70.15	**70.97**
	WRN-40-2	WRN-16-1	75.83	53.77	20.88	36.66	54.02	**57.91**	51.27	55.61	53.65	53.97
	WRN-40-2	WRN-40-1	75.83	61.33	42.83	53.60	63.91	68.88	54.43	64.57	69.46	**69.55**
	WRN-40-2	WRN-16-2	75.83	61.34	43.70	54.59	65.12	68.75	64.79	65.28	69.24	**69.40**
Tiny-ImageNet	ResNet-34	ResNet-18	-	-	-	-	-	64.01	63.73	-	64.13	**64.17**

Learning). We view distillation as information fusion with a time-varying confidence coefficient α_t. As plotted in the "step vs. α_t" graph in Fig. 1(b), we define the distillation loss $\mathcal{L}_{\text{KD}}$ as:

$$\mathcal{L}_{\text{KD}} = \alpha_t(t) \cdot \mathcal{L}_{\text{real}} + (1 - \alpha_t(t)) \cdot \mathcal{L}_{\text{synth}}, \tag{7}$$

where $\mathcal{L}_{\text{real}}$ and $\mathcal{L}_{\text{synth}}$ are KL divergences on logits. We employ a cosine annealing schedule for α_t.

This schedule creates a smooth curriculum that aligns with the student's learning dynamics:

- **High Trust-Region Phase (Real-Heavy):** Initially $\alpha_t \approx 1$. The student prioritizes the "Real Images" path (red arrows in Fig. 1(b)). This establishes reliable decision boundaries based on high-confidence priors, preventing the student from overfitting to early-stage generative artifacts.
- **Exploration Phase (Synthetic-Heavy):** As training progresses, $\alpha_t \to 0$. The generator converges, and the SNR of "Synthesized Images" improves. The student shifts focus to the synthetic path (yellow arrows) to explore the augmented manifold. This phase fills the gaps between sparse real samples, improving generalization and robustness.

This mechanism ensures the student effectively exploits the "infinite" sampling capability of the generator without being misled by optimization noise, achieving a balance between stability and plasticity.

IV. EXPERIMENTS

A. Experimental Setup

Datasets and Models. We conduct comprehensive experiments on three standard benchmarks: CIFAR-10, CIFAR-100 [32], and Tiny-ImageNet [33]. We investigate five representative pairs covering homogeneous distillation (e.g., ResNet-34 $\to$ ResNet-18), heterogeneous distillation (e.g., VGG-11 $\to$ ResNet-18), and distillation across varying depth and width configurations (e.g., WRN-40-2 $\to$ WRN-16-1). All teacher models are pre-trained on the original datasets and kept frozen during the distillation process.

Baselines. We compare PGDI against two categories of state-of-the-art approaches: (1) DFKD methods, which synthesize data without accessing real samples. We select representative methods including DeepInv [6], CMI [7], DAFL [20], ZSKD [21], DFQ [24], FAST [8], and SSD-KD [34]. (2) FSKD methods, which utilize limited real data, including FSKD [9], Cross Distillation [11], and GraftNet [10].

Implementation Details. We strictly follow the standard training protocols used in prior works [6], [7]. The student networks are optimized using SGD with a Nesterov momentum of 0.9 and a weight decay of 5×10^{-4}. The initial learning rate is set to 0.1 and decays following a cosine annealing schedule. We set the mini-batch size to 128 for all experiments. In the few-shot settings, we randomly sample K images per class ($K = 5$ or 10) from the original training set to serve as real data priors. We report the Top-1 classification accuracy on the test set as the primary evaluation metric.

B. Comparison with Data-free Knowledge Distillation (DFKD)

We evaluate PGDI against representative DFKD methods, including DeepInv [6], DAFL [20], ZSKD [21], DFQ [24], CMI [7], FAST [8], and SSD-KD [34]. DFKD baselines are evaluated in the *strictly data-free* regime. PGDI follows the realistic scenario where a tiny labeled set is available: we use *5/10 shots per class* primarily as *priors* to steer model inversion and synthesize proxy data. Training protocols are matched as applicable.

As presented in Table I, PGDI demonstrates superior distillation performance, achieving the highest student accuracy in the majority of settings (7/10 on CIFAR-10/100). Specifically, PGDI dominates across all ResNet and VGG series pairs on both datasets. Regarding the WideResNet families, SSD-KD and CMI exhibit strong competitiveness, yielding marginally higher accuracy than PGDI on the WRN-16-1 student or the

TABLE II
COMPARISON WITH STATE-OF-THE-ART FEW-SHOT KNOWLEDGE DISTILLATION (FSKD) METHODS.

Method	Shot	CIFAR-10	CIFAR-100
Teacher	–	92.83	69.82
FSKD	5	18.60	21.60
Cross Distillation	5	17.45	17.83
GraftNet	5	79.31	49.62
PGDI (Ours)	5	**92.55**	**68.24**
FSKD	10	19.20	29.33
Cross Distillation	10	18.42	18.77
GraftNet	10	86.24	56.89
PGDI (Ours)	10	**92.62**	**68.93**

CIFAR-10 WRN-16-2 pair. However, it is worth noting that while these pure DFKD methods perform well on the simpler CIFAR-10 dataset, their performance tends to drop on the more challenging CIFAR-100 benchmark. In contrast, PGDI maintains a robust lead on CIFAR-100 (e.g., surpassing SSD-KD by a significant margin on WRN pairs) and achieves the best results on Tiny-ImageNet. This trend suggests that while purely generative approaches like SSD-KD are effective for low-complexity tasks, the integration of real-data priors in PGDI provides crucial guidance that prevents distribution shift, ensuring better scalability and generalization for complex classification problems.

C. Comparison with Few-shot Knowledge Distillation (FSKD)

This section presents a direct comparison between our PGDI and state-of-the-art few-shot distillation methods. Approaches such as FSKD [9], Cross Distillation [11], and GraftNet [10] address model compression by training directly on a small set of real samples. In our experiments, the teacher is VGG16 and the student is VGG16-half, following the setups used in FSKD and GraftNet. Detailed results are reported in Table II.

As shown in Table II, PGDI consistently outperforms traditional FSKD-style methods across all few-shot settings, with performance approaching that of the teacher. The primary reason is that batch normalization constitutes a key bottleneck in few-shot distillation. Classical FSKD and Cross Distillation pipelines often include structured pruning, which fundamentally alters the inputs to BN layers while those layers still normalize with stale running statistics, producing a severe statistical mismatch. The initial errors are then strongly amplified by nonlinearities such as ReLU, leading to widespread neuron inactivity and irreversible information loss, and ultimately catastrophic degradation. Although GraftNet avoids pruning through block-wise distillation, the statistics induced by few-shot data do not fully match the original training distribution, which likewise degrades performance. In contrast, PGDI synthesizes high-quality proxy data via model inversion and forgoes direct training on scarce real samples, thereby sidestepping BN-statistics mismatch. This indirect transfer stabilizes optimization and enables the student to learn from diverse, high-fidelity synthetic data, culminating in superior generalization.

TABLE III
ABLATION ON RESNET-34→RESNET-18 (TOP-1%).

Method	CIFAR-10	CIFAR-100
Baseline	94.05	74.34
w/o L_{feat}	94.85	74.96
w/o L_{logits}	94.91	75.87
w/o Curriculum	94.86	76.87
Full Method	**95.10**	**77.13**

D. Ablation Studies: Validating Dual-Space Constraints

We validate the theoretical contribution of each component on CIFAR-100 (Table III).

1) RKHS Moment Matching ($\mathcal{L}_{\text{MMD}}$): Removing feature guidance causes a significant drop (77.13% → 74.96%). This performance gap confirms that matching only BN statistics is an ill-posed problem, leading to semantic drift. $\mathcal{L}_{\text{MMD}}$ is essential for constraining high-order moments to recover the true manifold topology.

2) Geometric Calibration ($\mathcal{L}_{\text{logits}}$): Omitting logits alignment degrades accuracy to 75.87%. This verifies the necessity of our Bias-Variance trade-off. Without the counterbalancing forces of prototype matching (bias reduction) and instance anchoring (variance preservation), the synthesized decision boundary becomes either shifted or collapsed.

3) Confidence-Aware Curriculum: Using a static weight (fixing $\alpha = 0.5$) yields suboptimal results (76.87%). This corroborates the non-stationarity of the generative process, proving that a dynamic trust-region strategy is required to balance early-stage stability with late-stage exploration.

E. Sensitivity Analysis: Regularization Dynamics

We analyze the sensitivity of feature guidance weight λ_{mmd}, logits guidance weight λ_{logit}, and initial curriculum ratio α_{initial} on CIFAR-100 (5-shot, ResNet-34→ResNet-18), where λ_{mmd} weights the feature-distribution prior (MMD), and λ_{logit} weights the logits-level calibration loss.

1) Feature Guidance (λ_{mmd}): Accuracy peaks at $\lambda_{\text{mmd}} = 100$ (Table IV). Lower weights ($\lambda_{\text{mmd}} < 10$) fail to sufficiently shrink the solution space (under-regularization), while excessive weights ($\lambda_{\text{mmd}} > 100$) cause overfitting to the sparse priors. This confirms $\mathcal{L}_{\text{MMD}}$ acts as a critical topological regularizer.

2) Logits Guidance (λ_{logit}): Performance favors moderate calibration at $\lambda_{\text{logit}} = 1$ (Table IV). Stronger constraints ($\lambda_{\text{logit}} \geq 10$) degrade performance, suggesting that restricting the generator too tightly to few-shot anchors hinders manifold interpolation (mode collapse), limiting the diversity needed for robust student generalization.

3) Initial Trust Region (α_{initial}): Starting with high real-data reliance ($\alpha_{\text{initial}} = 0.9$) outperforms pure synthetic initialization ($\alpha = 0.0$) (Table V). This validates that anchoring the optimization trajectory in a "basin of attraction" defined by high-fidelity priors is crucial for navigating the non-convex landscape.

TABLE IV
ACCURACY (%) W.R.T. GUIDANCE WEIGHTS λ_{mmd} AND λ_{logit}.

Weight	0.1	1	10	100	200
λ_{mmd} (Fix $\lambda_{\text{logit}} = 1$)	94.19	94.53	94.46	**94.74**	94.66
λ_{logit} (Fix $\lambda_{\text{mmd}} = 100$)	94.12	**94.74**	94.56	94.28	94.22

TABLE V
ACCURACY (%) W.R.T. INITIAL CURRICULUM RATIO α_{INITIAL}.

Ratio (α_{initial})	0.0	0.3	0.5	0.7	0.9	1.0
Accuracy	94.45	94.46	94.61	94.63	**94.74**	94.62

V. CONCLUSION

We presented PGDI, which leverages a tiny real set as priors to steer model inversion, applies dual-level guidance (feature + logits), and adopts a curriculum that progressively balances real and synthetic data. Across CIFAR-10/100 and multiple teacher→student pairs, PGDI achieves state-of-the-art student accuracy, and ablations verify the contribution of each component. PGDI is well-suited to privacy-sensitive or data-scarce model compression scenarios where only a few labeled examples are available.

REFERENCES

[1] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[2] X. He, C. Tan, B. Liu, L. Si, W. Yao, L. Zhao, D. Liu, Q. Zhangli, Q. Chang, K. Li *et al.*, "Dealing with heterogeneous 3d mr knee images: A federated few-shot learning method with dual knowledge distillation," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023, pp. 1–5.

[3] R. Cai, Y. Cui, Z. Li, Z. Yu, H. Li, Y. Hu, and A. Kot, "Rehearsal-free domain continual face anti-spoofing: Generalize more and forget less," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8037–8048.

[4] H. Shen and E. Kurshan, "Temporal knowledge distillation for time-sensitive financial services applications," *arXiv preprint arXiv:2312.16799*, 2023.

[5] R. G. Lopes, S. Fenu, and T. Starner, "Data-free knowledge distillation for deep neural networks," *arXiv preprint arXiv:1710.07535*, 2017.

[6] H. Yin, P. Molchanov, J. M. Alvarez, Z. Li, A. Mallya, D. Hoiem, N. K. Jha, and J. Kautz, "Dreaming to distill: Data-free knowledge transfer via deepinversion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8715–8724.

[7] G. Fang, J. Song, X. Wang, C. Shen, X. Wang, and M. Song, "Contrastive model inversion for data-free knowledge distillation," *arXiv preprint arXiv:2105.08584*, 2021.

[8] G. Fang, K. Mo, X. Wang, J. Song, S. Bei, H. Zhang, and M. Song, "Up to 100x faster data-free knowledge distillation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 6, 2022, pp. 6597–6604.

[9] T. Li, J. Li, Z. Liu, and C. Zhang, "Few sample knowledge distillation for efficient network compression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 14 639–14 647.

[10] C. Shen, X. Wang, Y. Yin, J. Song, S. Luo, and M. Song, "Progressive network grafting for few-shot knowledge distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2541–2549.

[11] H. Bai, J. Wu, I. King, and M. Lyu, "Few shot network compression via cross distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 3203–3210.

[12] D. Nguyen, S. Gupta, K. Do, and S. Venkatesh, "Black-box few-shot knowledge distillation," in *European conference on computer vision*. Springer, 2022, pp. 196–211.

[13] J. Rajasegaran, S. Khan, M. Hayat, F. S. Khan, and M. Shah, "Self-supervised knowledge distillation for few-shot learning," *arXiv preprint arXiv:2006.09785*, 2020.

[14] Y. Wu, S. Chanda, M. Hosseinzadeh, Z. Liu, and Y. Wang, "Few-shot learning of compact models via task-specific meta distillation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 6265–6274.

[15] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *International Conference on Learning Representations (ICLR)*, 2015.

[16] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[17] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3967–3976.

[18] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," *arXiv preprint arXiv:1910.10699*, 2019.

[19] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 11 953–11 962.

[20] H. Chen, Y. Wang, C. Xu, Z. Yang, C. Liu, B. Shi, C. Xu, C. Xu, and Q. Tian, "Data-free learning of student networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3514–3522.

[21] G. K. Nayak, K. R. Mopuri, V. Shaj, V. B. Radhakrishnan, and A. Chakraborty, "Zero-shot knowledge distillation in deep networks," in *International conference on machine learning*. PMLR, 2019, pp. 4743–4751.

[22] P. Han, J. Park, S. Wang, and Y. Liu, "Robustness and diversity seeking data-free knowledge distillation," in *ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2740–2744.

[23] X. Qu, J. Wang, and J. Xiao, "Enhancing data-free adversarial distillation with activation regularization and virtual interpolation," in *ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3340–3344.

[24] Y. Choi, J. Choi, M. El-Khamy, and J. Lee, "Data-free network quantization with adversarial knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 710–711.

[25] M. Nagel, M. v. Baalen, T. Blankevoort, and M. Welling, "Data-free quantization through weight equalization and bias correction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1325–1334.

[26] P. Micaelli and A. J. Storkey, "Zero-shot knowledge transfer via adversarial belief matching," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[27] G. Fang, Y. Bao, J. Song, X. Wang, D. Xie, C. Shen, and M. Song, "Mosaicking to distill: Knowledge distillation from out-of-domain data," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 920–11 932, 2021.

[28] K. Do, T. H. Le, D. Nguyen, D. Nguyen, H. Harikumar, T. Tran, S. Rana, and S. Venkatesh, "Momentum adversarial distillation: Handling large distribution shifts in data-free knowledge distillation," *Advances in neural information processing systems*, vol. 35, pp. 10 055–10 067, 2022.

[29] T. Huang, S. You, F. Wang, C. Qian, and C. Xu, "Knowledge distillation from a stronger teacher," *Advances in Neural Information Processing Systems*, vol. 35, pp. 33 716–33 727, 2022.

[30] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International journal of computer vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[31] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The journal of machine learning research*, vol. 13, no. 1, pp. 723–773, 2012.

[32] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 248–255.

[34] H. Liu, Y. Wang, H. Liu, F. Sun, and A. Yao, "Small scale data-free knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6008–6016.