Using Attention Transfer and Synthetic Data For Unsupervised Few-Sample Feature-Based Knowledge Distillation

Colin Pannikkat
Oregon State University
pannikkc@oregonstate.edu

Ajinkya Gokule
Oregon State University
gokulea@oregonstate.edu

Abstract

Knowledge Distillation (KD) typically relies on large labeled datasets and direct access to a powerful (often "white-box") teacher model. However, in many practical scenarios we are given only a small set of unlabeled samples and a teacher model which is over-parameterized. In this paper, we investigate an unsupervised few-sample feature-based *knowledge distillation framework that addresses both challenges simultaneously. Drawing on the idea of* attention transfer *from intermediate feature maps, as well as* synthetic data generation*, our method trains a small student network to mimic a large teacher using only a few unlabeled samples. We first extract attention maps from teacher and student intermediate layers, then align them using either Euclidean or distribution-based metrics. In parallel, we employ synthetic data expansion—combining MixUp data augmentation and a conditional variational autoencoder (CVAE) trained on few unlabeled images—to further enhance the student's training signal. Experiments on CIFAR-100 and Tiny-ImageNet show that our approach can achieve significant improvements over baseline few-sample methods, approaching the teacher's performance while requiring only black-box access and a handful of unlabeled images. Full training code, scripts, and results are available.*[1]

1. Introduction

Modern deep neural networks achieve remarkable performance across a wide range of applications. However, they are often large and highly resource-demanding, making them impractical for deployment on devices with limited storage or compute. To address this, *knowledge distillation* is frequently used to compress a high-capacity teacher network into a smaller, more efficient student network. Most KD methods assume both (1) ample labeled training data, and (2) full access to the teacher's internal representations (a "white-box" teacher). While these assumptions are not always satisfied in practice, here we specifically focus on the scenario where the teacher's intermediate activations are indeed accessible, but only *few unlabeled samples* are available for training the student.

Under these constraints, a student trained directly on a scarce unlabeled dataset tends to underfit or converge to trivial solutions. In this paper, we propose an approach that leverages the teacher's intermediate feature maps via *attention transfer* [6], coupled with *synthetic data augmentation*, to overcome the lack of labeled data. Our contributions can be summarized as follows:

1. **feature-based distillation:** We align *attention maps* between teacher and student at certain intermediate layers, exploiting our white-box access to the teacher's activations.
2. **Unsupervised few-sample setting:** We operate on only a small set of unlabeled images and use teacher-provided predictions to create student-training targets.
3. **Synthetic data generation:** We integrate MixUp and a conditional variational autoencoder (CVAE) to generate additional unlabeled training images, further boosting the student's performance.

Experiments on CIFAR-100 and Tiny-ImageNet demonstrate that our *attention-transfer + synthetic data* framework significantly outperforms naive few-sample distillation baselines, approaching teacher-level performance despite limited unlabeled data.

2. Related Work

Knowledge Distillation. Knowledge distillation techniques typically follow the paradigm of attempting to train a smaller student model to mimic the outputs, or intermediate representation, of a large teacher to achieve similar accuracy. Most techniques assume abundant labeled data, a white-box teacher, and that the student's training data and teacher's training data are identical. Literature has found

[1] https://github.com/colinpannikkat/ATSynKD.

that the best accuracy is achieved when the student has access to the teacher's training data [3]. This however, may not always be feasible, especially if the teacher model is trained on a vast amount of training data, or if that training data are not publically available, such as recent pretrained LLM's.

Few-Sample and Black-Box KD. Recently, [4] explored few-shot KD where only a small amount of labeled data are available. They utilized feature-based alignment via learned 1x1 convolutions at the end of chosen intermediate student outputs. [5] explored few sample KD in an unsupervised black-box setting, employing synthetic images (via MixUp and generative models) to expand the training set, and querying the teacher for pseudo-labels, achieving promising results. We feel that although a blackbox scenario is realistic, it is limiting in terms of possible KD techniques, and most white-box techniques can be achieved in such a scenario via a surrogate model.

Attention Transfer. In [6], the authors introduced an attention transfer loss that encourages a student network to align its intermediate spatial attention maps with those of a teacher. This can be performed by comparing L^2-normalized feature activations. Later works extended these ideas by exploring different distance metrics or distributions, typically with the assumption that we have direct or partial access to the teacher's features. [1] furthered this in NLP-based tasks, aligning the attention distributions outputs of transformer blocks in LLM's during KD, to great success.

3. Proposed Approach

Our overall goal is to train a compact student S to mimic a teacher T, given:

- A *small* unlabeled dataset $\mathcal{X}$ (few-sample),
- Access to T's activations (we can query outputs and layer activations).

We adopt **feature-based KD** by *attention transfer* [6] and combine it with **synthetic data** generation. Figure 1 (schematic) illustrates the idea.

3.1. Attention Transfer (AT)

We closely follow the attention transfer from [6]. Let A_j^T be the teacher's intermediate activation tensor at layer j, and A_j^S be the student's corresponding layer output. We define an *attention mapping* function

$$F : \mathbb{R}^{C\times H\times W} \longrightarrow \mathbb{R}^{H\times W}, \tag{1}$$

which pools the channels (e.g. summation [6]) to produce a 2D map. We then normalize each map in an ℓ_p sense:

$$Q_j^S = \frac{F\big(A_j^S\big)}{\big\|F\big(A_j^S\big)\big\|_p}, \quad Q_j^T = \frac{F\big(A_j^T\big)}{\big\|F\big(A_j^T\big)\big\|_p}. \tag{2}$$

A general *attention loss* is then defined by choosing a distance function $\mathcal{D}$ between these normalized maps:

$$\mathcal{L}_{\text{AT}} = \sum_{j\in\mathcal{I}} \mathcal{D}\big(Q_j^S,\, Q_j^T\big), \tag{3}$$

where $\mathcal{I}$ indexes the chosen layers.

KL-based variant. We can convert each Q_j^S, Q_j^T into a probability map via a spatial softmax:

$$P_j^S = \text{softmax}\big(Q_j^S\big), \quad P_j^T = \text{softmax}\big(Q_j^T\big).$$

We then measure their discrepancy via Kullback–Leibler divergence:

$$\mathcal{L}_{\text{KLAT}} = \sum_{j\in\mathcal{I}} \text{KL}\big(P_j^T \parallel P_j^S\big). \tag{4}$$

SSIM-based variant. We also propose a version that compares Q_j^S, Q_j^T using the *normalized MSE* (NMSE), derived from the Structural Similarity Index Measure (SSIM) [2]:

$$\text{NMSE}(\mathbf{s},\mathbf{t}) = \frac{\|\mathbf{s}-\mathbf{t}\|_2^2}{\|\mathbf{s}\|_2^2 + \|\mathbf{t}\|_2^2 + c}, \tag{5}$$

where c is a small constant. We then define the SSIM-based attention transfer loss as

$$\mathcal{L}_{\text{SSIMAT}} = \sum_{j\in\mathcal{I}} \text{NMSE}\big(Q_j^T,\, Q_j^S\big). \tag{6}$$

3.2. Synthetic Data Generation

To address the limited size of $\mathcal{X}$, we follow the FS-BBT procedure from [5] to create a larger synthetic training set $\mathcal{X}'$, which combines both *MixUp* and *CVAE* samples. The overall process is:

1. **Generate M MixUp samples.** Randomly pick pairs (x_i, x_j) from $\mathcal{X}$ and sample $\lambda \sim \text{Beta}$ with $\lambda \in [0,1]$. Form a synthetic sample
$$x_{\text{mu}} = \lambda\, x_i + (1-\lambda)\, x_j.$$
2. **Filter out extreme mixes.** Disqualify any x_{mu} for which λ is too close to 0 or 1, e.g. $\lambda < 0.05$ or $\lambda > 0.95$. Let M_1 be the number of qualified MixUp images retained.
3. **Train a CVAE on $\mathcal{X}$.** Using teacher-provided hard labels for the few real samples, we fit a conditional variational autoencoder (CVAE). This model learns to generate new images conditioned on the teacher-assigned class.

4. **Generate M_2 CVAE samples.** Since we keep M_1 MixUp images, let $M_2 = M - M_1$. We sample M_2 latent vectors (some from a normal distribution and some from a uniform distribution) and decode them with the CVAE to obtain synthetic images x_{cvae}. This step produces both in-distribution and out-of-distribution variants.

5. **Union of real and synthetic.** We combine the real images and synthetic samples into

$$\mathcal{X}' = \mathcal{X} \cup \mathcal{X}_{\text{mu}} \cup \mathcal{X}_{\text{cvae}},$$

where $\mathcal{X}_{\text{mu}}$ are the retained MixUp samples and $\mathcal{X}_{\text{cvae}}$ are those from the CVAE.

6. **Label each synthetic sample via the teacher.** For each $x \in (\mathcal{X}_{\text{mu}} \cup \mathcal{X}_{\text{cvae}})$, we query the teacher T to obtain a soft label y_T and hard label $\hat{y}_T$, which is then used in the KD objective.

In this manner, we expand a tiny unlabeled set $\mathcal{X}$ into a sufficiently large pseudo-labeled set $\mathcal{X}'$, ensuring better coverage of the input space and thereby improving the student's training signal.

3.3. Overall Training Objective

We define a combined objective that penalizes both mismatches in the teacher-student outputs and mismatches in their attention maps. Specifically, for an input x, we first form a KD loss term:

$$\mathcal{L}_{\text{KD}}(x) = \alpha \text{CE}(y_S, y_T) + (1-\alpha)\text{CE}(\hat{y}_S, \hat{y}_T) \quad (7)$$

where the first cross-entropy uses the teacher's softmaxed output y_T as a *soft* target, and the second cross-entropy uses $\hat{y} = \arg\max(y_T)$ as a *hard* target. The parameter $\alpha \in [0, 1]$ controls the balance between soft and hard targets.

For attention transfer, we compute an attention loss $\mathcal{L}_{\text{AT}}$ across teacher-student pairs on the same (possibly synthetic) inputs x, based on a chosen distance metric (e.g., KL or NMSE) over normalized attention maps.

Putting these together, each input x from the union of original and synthetic sets $\left(\mathcal{X} \cup \mathcal{X}_{\text{mu}} \cup \mathcal{X}_{\text{cvae}}\right)$ incurs:

$$\mathcal{L}_{\text{ours}}(x) = \lambda \mathcal{L}_{\text{KD}}(x) + \left(1-\lambda\right) \mathcal{L}_{\text{AT}}(x) \quad (8)$$

where $\lambda \in [0, 1]$ is a hyperparameter controlling how much we weight the KD term versus the attention-transfer term. We then average $\mathcal{L}_{\text{ours}}(x)$ using it as an estimator for generalization, training loss and validation loss. Note that only the student network is updated via backpropagation; the teacher is kept fixed.

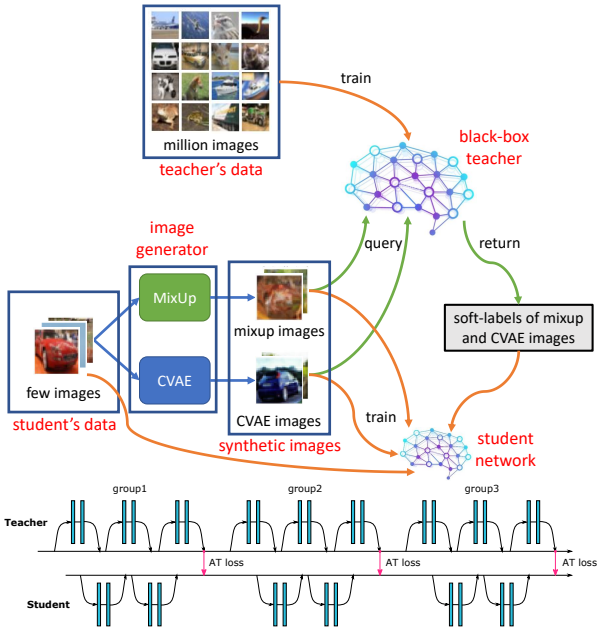

Figure 1. **Schematic of our approach.** Adapted from [5] and [6]. We generate MixUp and CVAE images from a small unlabeled set $\mathcal{X}$. The teacher T is queried for each synthetic sample, yielding $T(x)$ as a soft label. The student S is then trained via cross-entropy on $T(x)$ plus an attention-transfer loss that forces S's feature attentions to match T's.

4. Experiments and Results

We evaluate our approach on **CIFAR-100** and **Tiny-ImageNet**, following the baseline setup from [5, 6]. For each dataset, we create a *few-sample* subset by taking 50 unlabeled images from each class. We adopt **ResNet-32** as the teacher and **ResNet-20** as the student. We do AutoAugment, horizontal flips, and standard normalization. We train the CVAE using the teacher's hard labels on these few images. We generate $\sim 40K$ synthetic images for CIFAR100 and $\sim 50K$ for Tiny-Imagenet via MixUp + CVAE, query the teacher for each, and train the student with the final loss in Section 3.3.

4.1. Quantitative Results

Table 1 shows results on CIFAR-100. Using only 5K unlabeled images, the *Baseline Student* alone obtains 32.85% test accuracy, while *Baseline KD* (student trained on few unlabeled images that were labeled with teacher's soft labels) improves to 47.54%. Methods that leverage *attention transfer* from limited data, e.g. *EuclidAT*, reach 36.70%. Our approach that uses **KL-based attention transfer** (KLAT KD) or **SSIM-based attention transfer** (SSIMAT KD) can further improve to the mid-40% range. Finally, adding **synthetic data** (the MixUp + CVAE expansions) yields $\approx 56\%$ KD accuracy and $\approx 51\%$ test accuracy, 3 percentage point

CIFAR-100 ($N = 5000$, $M = 40000$)		
Method	**KD Accuracy**	**Test Accuracy**
Baseline Teacher[(1)]	-	66.99%
Baseline Student[(1)]	-	64.17%
Student Alone[(2)]	-	32.85%
Baseline KD[(2)]	51.34%	47.54%
EuclidAT KD[(2,3)]	40.27%	36.70%
KLAT KD[(2,4)]	49.17%	45.75%
SSIMAT KD[(2,5)]	50.14%	45.33%
KLAT KD + Synth[(2,4)]	**56.70**%	**51.27**%
SSIMAT KD + Synth[(2,5)]	**56.80**%	**50.18**%

Table 1. KD (w.r.t. teacher predicted labels) and test (w.r.t ground-truth labels) validation accuracy on baslines, previous methods, and our methods on CIFAR-100.

Tiny-ImageNet ($N = 10000$, $M = 50000$)		
Method	**KD Accuracy**	**Test Accuracy**
Baseline Teacher[(1)]	-	53.33%
Baseline Student[(1)]	-	48.86%
Student Alone[(2)]	-	23.29%
Baseline KD[(2)]	37.66%	30.80%
EuclidAT KD[(2,3)]	32.28%	25.92%
KLAT KD[(2,4)]	47.73%	37.63%
SSIMAT KD[(2,5)]	49.20%	37.69%
KLAT KD + Synth[(2,4)]	**??.??**%	**??.??**%
SSIMAT KD + Synth[(2,5)]	**52.92**%	**39.98**%

Table 2. KD (w.r.t. teacher predicted labels) and test (w.r.t ground-truth labels) validation accuracy on baslines, previous methods, and our methods on CIFAR-100.

[(1)]Trained on full dataset.
[(2)]Trained on an unlabeled few-sample dataset.
[(3)]Method from [6].
[(4)]Trained using unsupervised KD loss and KL divergence.
[(5)]Trained using unsupervised KD loss and NMSE.

more than baseline KD. A similar pattern holds for Tiny-ImageNet, as summarized in Table 2, with our final approach achieving $\approx$ 10–15% absolute gain over baseline KD.

4.2. Discussion

A central question is whether mere reliance on teacher outputs, with only a handful of unlabeled samples, provides enough supervision for a student to succeed. The table suggests it does not: student performance remains limited when it matches only the teacher's predictions. Another natural question is whether data augmentation alone—through, for example, MixUp or generative models—can close the gap. Again, the data shows that such augmentation helps but remains insufficient on its own, pointing to the importance of aligning student and teacher features more explicitly.

So does incorporating attention transfer meaningfully enhance few-sample distillation, especially for complex datasets and deeper student/teacher networks? The results indicate that attention-based alignment consistently boosts performance, presumably because it enforces more structured representation matching than predictions alone can provide. Training with synthetic data in addition to attention transfer reinforces this benefit and leads to a stronger overall student, hinting that these two approaches—feature-level alignment and data expansion are complementary. Comparing results across datasets of varying complexity and network capacities underscores that, as both problem difficulty and model expressiveness grow, combining attention constraints with synthetic data could be key to unlocking superior few-sample distillation outcomes.

5. Conclusion

We presented an *unsupervised few-sample knowledge distillation* approach that uses *attention transfer* and *synthetic data augmentation* to train a compact student from a teacher using only a handful of unlabeled samples. Our experiments on CIFAR-100 and Tiny-ImageNet confirm that attention-based feature alignment plus a carefully designed synthetic image pipeline (MixUp + CVAE) can significantly outperform simpler baselines, approaching teacher-level accuracy in highly data-limited settings.

In future work, we aim to scale these techniques to larger datasets (e.g. full ImageNet) and more varied tasks such as object detection or language modeling. We also intend to explore different CVAE reconstruction losses (e.g. SSIM or Frechet) in place of MSE, investigate more flexible choices for the distillation and attention-transfer tradeoffs (such as dynamic λ), and study potential benefits of pretraining student attention layers in isolation. Expanding beyond label conditioning, e.g. conditioning on other model-derived signals, and employing diffusion-based synthetic image generation are further promising directions to increase coverage of data manifolds for few-sample KD.

References

[1] G. Aguilar, Y. Ling, Y. Zhang, B. Yao, X. Fan, and C. Guo. Knowledge distillation from internal representations, 2020.

[2] D. Brunet, E. R. Vrscay, and Z. Wang. On the mathematical properties of the structural similarity index. *IEEE Transactions on Image Processing*, 21(4):1488–1499, 2012.

[3] A. Chawla, H. Yin, P. Molchanov, and J. Alvarez. Data-free knowledge distillation for object detection. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3288–3297, 2021.

[4] T. Li, J. Li, Z. Liu, and C. Zhang. Few sample knowledge distillation for efficient network compression, 2020.

[5] D. Nguyen, S. Gupta, K. Do, and S. Venkatesh. Black-box few-shot knowledge distillation, 2022.

[6] S. Zagoruyko and N. Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer, 2017.