TRACO-KD: COMPILING TEACHER TRAINING TRAJECTORIES INTO LOGIT KNOWLEDGE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Logit-based knowledge distillation (KD) has emerged as a lightweight alternative to feature-based distillation. Recent logit-based methods mainly improve the alignment of teacher and student outputs, with supervision drawn from a teacher at convergence. However, teacher classification accuracy peaks at the end of training, while relations among non-target classes continue to evolve, making it difficult for a teacher at any single training stage to provide both. Motivated by this observation, we propose Trajectory Compilation Knowledge Distillation (TraCo-KD), which compiles the teacher's full training trajectory offline into a frozen operator acting on converged logits. Specifically, TraCo-KD decomposes logits into a target margin, a non-target scale, and non-target relations, aggregates per-sample trajectory targets for each component, and fits a low-rank operator mapping converged logits to these targets. During distillation, the corrected teacher outputs are combined with the original outputs through per-sample mixing and used directly in the existing loss, requiring only a single teacher forward pass and no additional inference cost. Extensive experiments show that TraCo-KD achieves the best performance among logit-based methods and improves feature-based distillation as a plug-in.

1 INTRODUCTION

The performance of deep neural networks continues to improve with model capacity, but their computational and storage costs limit deployment on mobile devices and embedded systems. Techniques such as pruning, quantization, and coding compress models at the parameter-representation level (Han et al., 2016). Knowledge distillation (KD) instead trains a compact student network using the outputs of a high-capacity teacher network, thereby transferring the teacher's knowledge to the student (Hinton et al., 2014; Gou et al., 2021; Beyer et al., 2022).

Distillation supervision can come from the teacher's intermediate features or output logits. Feature distillation aligns intermediate representations and their relations (Romero et al., 2015; Zagoruyko & Komodakis, 2017; Ahn et al., 2019; Park et al., 2019; Heo et al., 2019; Tian et al., 2020; Chen et al., 2021b;a; 2022). It provides rich information, but requires access to the teacher's intermediate activations and additional adaptation when the teacher and student have heterogeneous architectures. Logit distillation depends only on the teacher's outputs (Hinton et al., 2014), offering a simple interface without requiring structural correspondence between teacher and student. Recent work has improved logit distillation by decoupling target and non-target classes (Zhao et al., 2022), replacing pointwise matching with correlation or rank consistency (Huang et al., 2022; Guan et al., 2025), and normalizing outputs before alignment (Sun et al., 2024; Yang et al., 2023). These methods bring logit distillation close to or even beyond feature distillation. However, they all improve how outputs are aligned, while the object being aligned remains the output of the fully converged teacher.

The converged teacher's output is not necessarily the most effective supervision for distillation. During teacher training, classification accuracy gradually improves, while the relative relations among non-target classes continue to rearrange. Figure 1 shows that the teacher's relation structure is closest to the structure aggregated across training stages after the first learning-rate decay, and then gradually departs from it as training converges. Thus, accuracy is highest at the end of training, whereas the relation structure is closest to the aggregated structure midway through training, and the two evolve asynchronously. Increasing the temperature is the standard way to soften converged

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

outputs (Hinton et al., 2014), but temperature scaling applies a monotonic transformation to each sample's logits and does not change the relative ordering among non-target classes. It therefore cannot recover the relation structure that has drifted during convergence.

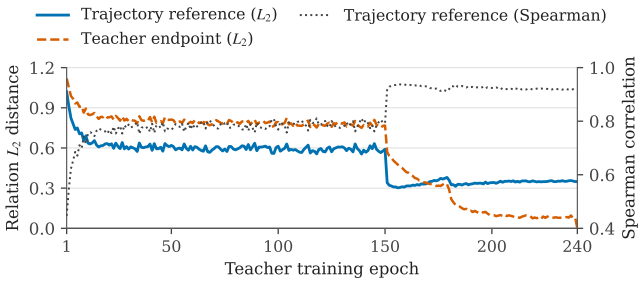

Figure 1: **Evolution of the teacher's non-target relation structure over its training trajectory.** The horizontal axis is the teacher training epoch. The left axis shows the relation L_2 distance to the trajectory target and the converged endpoint, and the right axis shows the Spearman rank correlation with the trajectory target. After epoch 150, the distance to the trajectory target reaches its minimum and slowly increases, while the distance to the converged endpoint continues to decrease.

Prior work has improved teacher supervision through early stopping or the selection of intermediate checkpoints. An early-stopped teacher can yield better distillation in some settings (Cho & Hariharan, 2019), intermediate checkpoints can outperform the converged model in several settings (Wang et al., 2022), and checkpoints ordered along the optimization path can form an easy-to-hard curriculum (Jin et al., 2019; Rezagholizadeh et al., 2022). Most of these methods obtain additional supervision by selecting or scheduling particular moments along the training trajectory. Our experiments, however, show that no single point on the trajectory simultaneously provides both accuracy and relation structure. An intermediate checkpoint preserves the relation structure of its stage, but also retains the underfitting of that stage. We select five intermediate checkpoints from the same training trajectory as teachers, and none outperforms the converged teacher in distillation (Section 4.5). This indicates that the relation structure needed for distillation is distributed across the entire training trajectory and is difficult to obtain in full from any single point.

Motivated by this observation, we no longer select a single training stage. Instead, we decouple accuracy from relation structure and combine the predictive ability of the converged teacher with relation structure aggregated across training stages. This design rests on the following consideration. The trajectory-aggregated structure and the converged logits are responses of the same network to the same input. They share most of their information, with their differences concentrated in a few degrees of freedom, so a compact mapping on the converged logits can approximate the trajectory target. The training trajectory therefore need not be stored or replayed during distillation, but can be compiled offline into a frozen operator.

We propose TraCo-KD (Figure 2). In the offline stage, TraCo-KD aggregates the teacher's full training trajectory into per-sample targets and fits an operator that takes the converged logits and training labels as input to reconstruct those targets. During distillation, the operator-corrected logits are interpolated with the converged logits for each sample and then used in the original logit distillation loss. The entire distillation process requires only a single forward pass through the converged teacher and introduces neither additional student parameters nor inference cost.

Our main contributions are as follows.

- We show that the teacher's classification accuracy and non-target class relation structure evolve asynchronously during training. No single point on the training trajectory provides both, which imposes an inherent limitation on checkpoint-selection methods.
- We propose TraCo-KD, which compiles the teacher's training trajectory offline into a frozen operator on converged logits. The operator can be placed before any logit distillation loss and can also be combined with feature distillation methods.

- We systematically evaluate TraCo-KD on CIFAR-100 and ImageNet. It achieves the best performance among logit-based methods and consistently improves feature distillation methods.

2 Related Work

Feature distillation. Feature distillation extracts supervision from the teacher's intermediate representations. Its focus has expanded from layerwise feature matching (Romero et al., 2015; Chen et al., 2022) to attention and statistical dependency modeling (Zagoruyko & Komodakis, 2017; Ahn et al., 2019), relational learning across samples (Park et al., 2019; Tian et al., 2020; Xu et al., 2020), and improved feature selection and cross-stage alignment (Heo et al., 2019; Chen et al., 2021b;a). These methods exploit the teacher's rich internal representations, but depend on access to intermediate activations and require additional projection modules when the teacher and student have heterogeneous architectures.

Logit distillation. Logit distillation uses the teacher's softened predictive distribution as supervision (Hinton et al., 2014; Furlanello et al., 2018; Zhang et al., 2018; Lan et al., 2018; Guo et al., 2020). Other work introduces an intermediate-capacity teacher assistant to reduce the teacher–student capacity gap (Mirzadeh et al., 2020). Subsequent loss designs mainly decompose supervision, align relations, or transform outputs. They independently weight target-class and non-target-class knowledge (Zhao et al., 2022), relax absolute-value matching through correlation or rank consistency (Huang et al., 2022; Guan et al., 2025), and normalize or standardize outputs before alignment (Yang et al., 2023; Sun et al., 2024). These alignment methods all modify how teacher and student outputs are matched. Their transformations of teacher outputs preserve the relative ordering among non-target classes, so the inter-class relation structure still comes from the end of training.

Distillation using the teacher training process. Prior work has found that early-stopped teachers and intermediate checkpoints can outperform converged teachers in some settings (Cho & Hariharan, 2019; Wang et al., 2022). Building on this observation, later methods schedule multiple checkpoints in training order (Jin et al., 2019; Rezagholizadeh et al., 2022; Panigrahi et al., 2025; Yang et al., 2019), or ensemble multiple models from the same trajectory (Huang et al., 2017; Wang et al., 2022). These methods treat the training trajectory as a set of candidate teachers, and the transferred supervision depends on models or model combinations at particular moments. In contrast, we aggregate non-target class relations across training stages and compile them into a frozen operator on converged logits, combining the converged teacher's predictive ability with relational information from the training trajectory.

3 Method

TraCo-KD consists of an offline compilation stage and a distillation stage. The offline stage takes the teacher's per-epoch outputs and the training labels as input. It first decomposes each logit vector into three components (Section 3.2), aggregates the full trajectory into a per-sample target on each component (Section 3.3), fits an operator $g(\mathbf{z}, y)$ that maps converged logits to that target (Section 3.4), and fits a per-sample mixing predictor that sets the correction strength (Section 3.5). During distillation both modules are frozen, and only the corrected teacher logits replace the teacher logits in the original loss.

3.1 Preliminaries

Let $\mathbf{z} \in \mathbb{R}^K$ be the logit vector output by a K-class classifier, z_j its j-th entry, and y the ground-truth label. For the same input the teacher and the student produce logit vectors $\mathbf{z}^{\mathcal{T}}$ and $\mathbf{z}^{\mathcal{S}}$, where $\mathbf{z}^{\mathcal{T}}$ comes from the converged teacher. Write $\mathcal{L}_{\text{logit}}$ for any distillation objective that takes these two logit vectors as input. A generic logit distillation loss is

$$\mathcal{L} \;=\; \lambda_{\text{ce}}\, \mathcal{L}_{\text{ce}}(\mathbf{z}^{\mathcal{S}}, y) \;+\; \lambda_{\text{kd}}\, \mathcal{L}_{\text{logit}}\big(\mathbf{z}^{\mathcal{S}},\, \mathbf{z}^{\mathcal{T}}\big)\,, \tag{1}$$

where $\mathcal{L}_{\text{ce}}$ is the cross-entropy between the student output and the label, and λ_{ce} and λ_{kd} weight the two terms. TraCo-KD replaces only the teacher logits $\mathbf{z}^{\mathcal{T}}$ in equation 1 by a corrected $\mathbf{z}^{\mathcal{T}\prime}$ and

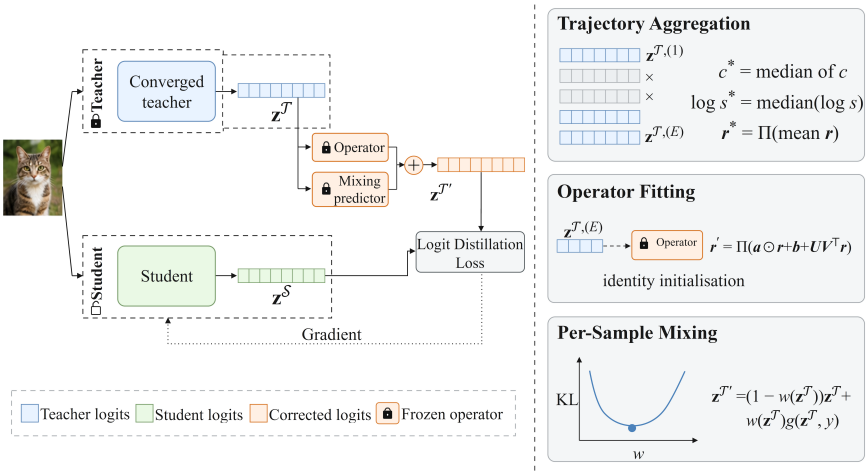

Figure 2: **Overview of TraCo-KD.** The offline stage aggregates the teacher's full trajectory into a per-sample target and fits a frozen operator to reconstruct it. The distillation stage corrects the converged teacher's output with that operator and passes the result to the existing logit distillation loss.

leaves $\mathcal{L}_{\text{logit}}$ untouched, so it composes with any logit distillation objective. In our experiments, $\mathcal{L}_{\text{logit}}$ comprises three terms for predictive distribution alignment, inter-sample relation alignment, and inter-class relation alignment, each evaluated at one base temperature and four auxiliary temperatures and summed. Section 4.6 further applies the same operator to two feature distillation methods.

3.2 Logit Decomposition

Trajectory supervision targets the relative relations among non-target classes. In the teacher logits, these relations are mixed with the target margin and output scale. Because softmax is invariant to a constant shift of its logits, a logit vector can be decomposed without loss into three components, namely the margin of the target class over the rest, the dispersion of the non-target classes, and their relative shape. To explicitly characterize the relative structure among non-target classes, we separate these components and construct trajectory targets and fit the operator in component space, providing the relation component with its own supervision.

Let $\mathcal{N}(y) = \{j : j \neq y\}$ be the non-target classes. We decompose $\mathbf{z}$ into an overall offset μ, a target margin c, a non-target scale s and a non-target relation $\boldsymbol{r}$ as follows,

$$\begin{aligned} \mu &= \frac{1}{K-1} \sum_{j \in \mathcal{N}(y)} z_j, \qquad s = \operatorname{std}_{j \in \mathcal{N}(y)} z_j, \\ c &= z_y - \mu, \qquad r_j = \frac{z_j - \mu}{s} \quad (j \in \mathcal{N}(y)), \qquad r_y = 0. \end{aligned} \tag{2}$$

The map is invertible up to the offset, with $z_j = \mu + s\, r_j$ for $j \in \mathcal{N}(y)$ and $z_y = \mu + c$, and μ does not affect the softmax output. All subsequent analysis is carried out on the relation component $\boldsymbol{r}$, which carries the inter-class structure.

3.3 Trajectory Aggregation

This decomposition allows outputs from different epochs to be compared at the component level. Teacher training produces one logit vector per training sample per epoch, and standard logit distillation keeps only the final converged state. We aggregate the whole sequence offline into a per-sample target, applying two filters first. The first filter drops the records of epochs in which the teacher misclassifies the sample, so that the relational structure of a wrong prediction is not transferred. The second applies a Tukey fence to the per-sample jump between consecutive epochs, which removes optimization transients such as those at learning rate decays. The margin and the log scale are scalars

and are aggregated by a median, which lowers sensitivity to residual outliers. The relation vector lies on a sphere-like set and is averaged and then re-standardized. If fewer than three valid epochs remain, the sample's trajectory target is set to its converged logits.

For each training sample, let the teacher's per-epoch outputs be $\mathbf{z}^{\mathcal{T}^{(1)}}, \ldots, \mathbf{z}^{\mathcal{T}^{(E)}}$, where E is the total number of teacher training epochs. Applying equation 2 to each epoch's output gives the components $c^{(e)}, s^{(e)}, \boldsymbol{r}^{(e)}$, and $\mathcal{V}$ denotes the set of epochs retained after the two filters. Starred quantities denote the aggregated trajectory target, defined as

$$c^\star = \operatorname{median} c^{(e)}, \qquad \log s^\star = \operatorname{median} \log s^{(e)}, \qquad \boldsymbol{r}^\star = \Pi\left(\frac{1}{|\mathcal{V}|}\sum \boldsymbol{r}^{(e)}\right), \qquad e \in \mathcal{V}. \quad (3)$$

where Π re-standardizes the non-target entries to zero mean and unit variance. Recomposing with the converged teacher's offset $\mu^{(E)}$ gives the target logits $\mathbf{z}^\star$.

3.4 The trajectory operator

The target of equation 3 is indexed by training sample, whereas the teacher outputs during distillation come from augmented views and cannot be retrieved by index. We therefore fit an operator g that takes the converged logits and the training label as input and approximates the trajectory target. The operator adjusts the relation structure through a low-rank cross-class transformation and is initialized at the identity.

The operator acts on each component separately, and primes mark corrected components. With class-indexed parameters $\alpha^c, \beta^c, \alpha^s, \beta^s \in \mathbb{R}^K$, diagonal parameters $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^K$ and a rank-ρ pair $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{K\times\rho}$, where a subscript y indexes the parameter by the sample's ground-truth class and $\odot$ is the elementwise product, it is defined by

$$c' = \alpha^c_y\, c + \beta^c_y, \qquad \log s' = \alpha^s_y\, \log s + \beta^s_y, \quad (4)$$

$$\boldsymbol{r}' = \Pi\big(\boldsymbol{a} \odot \boldsymbol{r} + \boldsymbol{b} + \boldsymbol{U}\boldsymbol{V}^\top \boldsymbol{r}\big)\,, \quad (5)$$

and $g(\mathbf{z}, y)$ is the recomposition of μ, c', s' and $\boldsymbol{r}'$. The operator is initialized at the identity by setting $\alpha^c = \alpha^s = \boldsymbol{a} = \mathbf{1}$, $\beta^c = \beta^s = \boldsymbol{b} = \mathbf{0}$ and $\boldsymbol{U} = \mathbf{0}$, so that initially $g(\mathbf{z}, y) = \mathbf{z}$. The operator holds $6K + 2K\rho$ parameters, which is 3600 for $K = 100$ and $\rho = 15$.

We fit the operator by minimizing the sum of squared errors over the three components,

$$\mathcal{L}_{\text{fit}} = \big\|c' - c^\star\big\|^2 + \big\|\log s' - \log s^\star\big\|^2 + \tfrac{1}{K-1}\big\|\boldsymbol{r}'_{\mathcal{N}(y)} - \boldsymbol{r}^\star_{\mathcal{N}(y)}\big\|^2. \quad (6)$$

where $\|\cdot\|$ denotes the Euclidean norm and the subscript $\mathcal{N}(y)$ selects the non-target entries. One operator is fitted per teacher and is shared by every student distilled from that teacher.

3.5 Per-sample mixing

The operator provides a correction direction, but the required correction strength varies across samples. For samples on which the teacher remains uncertain, the trajectory target carries more information, whereas for samples the teacher classifies confidently from early in training, the trajectory target is close to redundant with the endpoint. A single global strength must trade off between the two regimes, so we use a mixing predictor $w(\cdot)$ to set the interpolation weight for each sample,

$$\mathbf{z}^{\mathcal{T}\prime} \;=\; \big(1 - w(\mathbf{z}^{\mathcal{T}})\big)\,\mathbf{z}^{\mathcal{T}} \;+\; w(\mathbf{z}^{\mathcal{T}})\, g(\mathbf{z}^{\mathcal{T}}, y), \qquad w(\cdot) \in (0,1). \quad (7)$$

Supervision for the mixing coefficient comes from a per-sample oracle. Let σ_τ denote softmax at temperature τ, and let $\boldsymbol{d} = g(\mathbf{z}, y) - \mathbf{z}$ be the correction direction given by the operator. We define

$$w^\star = \arg\min \mathrm{KL}(\sigma_\tau(\mathbf{z}^\star) \,\|\, \sigma_\tau(\mathbf{z} + w\boldsymbol{d}))\,, \qquad w \in (0,1). \quad (8)$$

The terms of this objective that depend on w are a linear term and the log-sum-exp of an affine function. The objective is therefore convex in w, and its minimizer can be found by ternary search. Since $w^\star$ cannot be retrieved by sample index during distillation, we regress it from nine label-free statistics of the teacher logits using a two-layer perceptron with hidden width 32, a sigmoid output, and 353 parameters. The statistics are the top-five softmax probabilities, entropy, the margin between the top two softmax probabilities, the logit standard deviation, and the normalized logit norm. They depend only on the teacher logits and can therefore be applied directly to augmented views.

During distillation, we substitute equation 7 into equation 1 and update only the student parameters.

4 EXPERIMENTS

4.1 SETUP

We evaluate on CIFAR-100 and ImageNet-1K. All distillation runs use the publicly released teacher weights. Because released weights do not come with a training trajectory, we retrain each teacher once under its official recipe, record its per-epoch logits, aggregate the target and fit the operator on that trajectory, and apply the fitted operator directly to the output of the released weights. Student training follows the standard protocol of logit distillation (Tian et al., 2020; Zhao et al., 2022). On CIFAR-100, students are trained with SGD for 240 epochs with batch size 64 and an initial learning rate of 0.025 (0.01 for MobileNet and ShuffleNet students), decayed by 0.1 at epochs 150, 180 and 210. On ImageNet, students are trained for 100 epochs with batch size 512 and an initial learning rate of 0.2, decayed by 0.1 at epochs 30, 60 and 90. Distillation uses the three alignment terms described in Section 3.1, with base temperature 4 on CIFAR-100 and 1 on ImageNet and auxiliary temperatures $\{2, 3, 5, 6\}$. The operator and the mixing predictor are each fitted with Adam for 3000 steps, and all teachers on both datasets share the same fitting configuration.

4.2 CIFAR-100

Tables 1 and 2 report results for homogeneous and heterogeneous teacher–student pairs, respectively. In the homogeneous setting, TraCo-KD surpasses all compared methods on five of the six pairs, improving over the strongest comparison method by 1.48 percentage points on ResNet32×4 to ResNet8×4.

Table 1: **CIFAR-100 top-1 accuracy, homogeneous teacher–student pairs.** Bold marks the best logit-based method in each column.

	Teacher	ResNet32×4	VGG13	WRN-40-2	WRN-40-2	ResNet56	ResNet110
	(acc)	*79.42*	*74.64*	*75.61*	*75.61*	*72.34*	*74.31*
	Student	ResNet8×4	VGG8	WRN-40-1	WRN-16-2	ResNet20	ResNet32
	(acc)	*72.50*	*70.36*	*71.98*	*73.26*	*69.06*	*71.14*
features	FitNet (2015)	73.50	71.02	72.24	73.58	69.21	71.06
features	AT (2017)	73.44	71.43	72.77	74.08	70.55	72.31
features	VID (2019)	73.09	71.23	73.30	74.11	70.38	72.61
features	RKD (2019)	71.90	71.48	72.22	73.35	69.61	71.82
features	OFD (2019)	74.95	73.95	74.33	75.24	70.98	73.23
features	CRD (2020)	75.51	73.94	74.14	75.48	71.16	73.48
features	ReviewKD (2021b)	75.63	74.84	75.09	76.12	71.89	73.89
logits	KD (2014)	73.33	72.98	73.54	74.98	70.66	73.08
logits	TAKD (2020)	73.81	73.23	73.78	75.12	70.83	73.37
logits	DKD (2022)	76.32	74.68	74.81	76.24	**71.97**	74.11
logits	DIST (2022)	76.16	73.74	74.42	75.29	71.55	73.55
logits	NKD (2023)	76.35	74.32	74.43	76.37	71.62	73.79
logits	LSKD (2024)	76.62	74.36	74.37	76.11	71.43	74.17
logits	RKKD (2025)	74.74	74.14	73.93	75.87	71.09	73.09
logits	**Ours**	**78.10**	**75.31**	**75.18**	**76.48**	71.39	**74.48**

In the heterogeneous setting, TraCo-KD achieves the best logit-based result on all six pairs, improving over the strongest logit-based method by 1.73 percentage points on ResNet32×4 to SHN-V2. The operator is fitted on the teacher alone, without involving the student architecture. Its consistent lead on all six heterogeneous pairs shows that it also works when teacher and student architectures differ. ReviewKD is slightly ahead in the three italicized columns of Table 2. Section 4.6 further examines combining the trajectory operator with feature distillation methods.

4.3 IMAGENET

We evaluate on ImageNet with the homogeneous pair ResNet34 to ResNet18 and the heterogeneous pair ResNet50 to MobileNet-V1, with results in Table 3. TraCo-KD surpasses all compared methods in top-1 accuracy on both pairs, improving over the strongest comparison method on each pair by

Table 2: **CIFAR-100 top-1 accuracy, heterogeneous teacher–student pairs.** Bold marks the best logit-based method in each column, and italics mark columns where a feature method leads all logit methods.

	Teacher	ResNet32×4	ResNet32×4	WRN-40-2	WRN-40-2	VGG13	ResNet50
	(acc)	*79.42*	*79.42*	*75.61*	*75.61*	*74.64*	*79.34*
	Student	SHN-V2	SHN-V1	ResNet8×4	SHN-V1	MN-V2	MN-V2
	(acc)	*71.82*	*70.50*	*72.50*	*70.50*	*64.60*	*64.60*
features	FitNet (2015)	73.54	73.59	74.61	73.73	64.16	63.16
features	AT (2017)	72.73	71.73	74.11	73.32	59.40	58.58
features	VID (2019)	73.57	73.38	74.65	73.61	65.56	65.79
features	RKD (2019)	73.21	72.28	75.26	72.21	64.52	64.43
features	OFD (2019)	76.82	75.98	74.36	75.85	69.48	69.04
features	CRD (2020)	75.65	75.11	75.24	76.05	69.73	69.11
features	ReviewKD (2021b)	77.78	*77.45*	74.34	*77.14*	*70.37*	69.89
logits	KD (2014)	74.45	74.07	73.79	74.83	67.37	67.35
logits	TAKD (2020)	72.12	74.53	74.03	75.34	67.91	68.02
logits	DKD (2022)	77.07	76.45	75.56	76.70	69.71	70.35
logits	DIST (2022)	76.35	75.23	75.67	75.73	68.48	69.14
logits	NKD (2023)	76.26	75.31	76.01	75.96	68.72	69.39
logits	LSKD (2024)	75.56	75.19	77.11	76.03	68.61	69.02
logits	RKKD (2025)	75.35	75.98	76.25	76.13	68.23	69.45
logits	**Ours**	**78.80**	**77.37**	**77.16**	**77.09**	**70.30**	**70.69**

0.36 and 0.27 percentage points, respectively. ImageNet has ten times as many classes as CIFAR-100, and the operator's parameter count grows linearly with the number of classes. Both datasets use the same fitting configuration for the operator and mixing predictor, showing that trajectory compilation extends directly to a larger label space. On the heterogeneous pair, TraCo-KD surpasses ReviewKD, which requires intermediate-layer projection, so trajectory supervision that relies on the teacher output alone remains competitive with feature distillation at scale.

Table 3: **ImageNet top-1 and top-5 accuracy.** Bold marks the best.

		ResNet34 → ResNet18		ResNet50 → MN-V1	
		top-1	top-5	top-1	top-5
	Teacher	73.31	91.42	76.16	92.86
	Student	69.75	89.07	68.87	88.76
feat.	AT (2017)	70.69	90.01	69.56	89.33
feat.	OFD (2019)	70.81	89.98	71.25	90.34
feat.	CRD (2020)	71.17	90.13	71.37	90.41
feat.	ReviewKD (2021b)	71.61	90.51	72.56	91.00
logits	KD (2014)	71.03	90.05	70.50	89.80
logits	DKD (2022)	71.70	90.41	72.05	91.05
logits	LSKD (2024)	71.42	90.29	72.18	90.80
logits	RKKD (2025)	71.27	90.16	71.54	90.84
logits	**Ours**	**72.06**	**90.64**	**72.83**	**91.06**

4.4 Ablation

We add the operator and the per-sample mixing coefficient in turn on ResNet32×4 to ResNet8×4, with results in Table 4. The baseline uses the same distillation configuration as TraCo-KD without the operator. The operator alone contributes most of the gain, which shows that the target obtained by trajectory compilation is itself the main source of improvement. Adding the per-sample mixing coefficient on top brings a further gain, which shows that adapting the correction strength to the teacher's uncertainty is better than applying a single strength to every sample, consistent with the analysis of Section 3.5 that a single global strength must compromise between confident and uncertain samples.

Table 4: **Ablation** on ResNet32×4 to ResNet8×4, CIFAR-100.

Configuration	Trajectory operator	Per-sample w	Top-1	Δ
Base loss			77.22	–
+ operator	✓		77.72	+0.50
+ operator + w	✓	✓	**78.10**	+0.88

4.5 Comparison with intermediate-checkpoint teachers

We compare trajectory compilation with using intermediate checkpoints directly as teachers. From the same ResNet32×4 trajectory used to fit the operator, we take the weights at epochs 150, 165, 180, 195 and 210 and use each as the teacher, with the student training configuration identical to the baseline and no operator, with results in Table 5. All five intermediate-checkpoint teachers perform worse than the converged teacher. Early checkpoints lack accuracy themselves and student accuracy drops accordingly, while later checkpoints approach the converged teacher but do not surpass it. This result reflects the trade-off between the stage-specific relational structure retained by intermediate checkpoints and their degree of convergence. TraCo-KD builds on converged logits and incorporates relational structure aggregated across training stages. Within the range examined here, it yields higher student accuracy than both the converged-teacher baseline and all intermediate checkpoints.

Table 5: **Intermediate checkpoints as teachers** on ResNet32×4 to ResNet8×4, CIFAR-100. All rows use the base loss in Table 4. The endpoint baseline and TraCo-KD use released teacher weights, and only TraCo-KD applies the trajectory operator. Δ is the student accuracy change from the endpoint baseline.

Distillation setting	Teacher top-1	Student top-1	Δ
Released teacher + base loss	79.42	77.22	–
Checkpoint, epoch 150	61.00	61.92	−15.30
Checkpoint, epoch 165	77.82	76.93	−0.29
Checkpoint, epoch 180	77.11	75.67	−1.55
Checkpoint, epoch 195	79.26	77.06	−0.16
Checkpoint, epoch 210	79.11	76.75	−0.47
TraCo-KD	79.42	**78.10**	+0.88

4.6 Combining the trajectory operator with feature-based methods

Because the operator corrects only the teacher logits and does not touch intermediate features, it can be attached to feature distillation methods as a drop-in component. We add an operator-corrected logit distillation term to the original feature loss of each of FitNet and RKD, retaining their training recipes and hyperparameters and using a shared weight and temperature for the added term across teacher–student pairs. Table 6 reports the results. All twelve combinations of the two methods on six pairs improve. These results show that the information recovered through trajectory compilation complements feature distillation and can further improve its performance.

Table 6: **Combining the trajectory operator with feature-based methods**, CIFAR-100 top-1 accuracy on six homogeneous pairs.

Method	R56→R20	R110→R32	R32×4→R8×4	WRN40-2→16-2	WRN40-2→40-1	VGG13→VGG8	Avg
FitNet	69.21	71.06	73.50	73.58	72.24	71.02	71.77
FitNet + Ours	**69.55**	**72.27**	**75.82**	**74.58**	**73.57**	**73.07**	**73.14**
Δ	+0.34	+1.21	+2.32	+1.00	+1.33	+2.05	+1.38
RKD	69.61	71.82	71.90	73.35	72.22	71.48	71.73
RKD + Ours	**70.77**	**73.28**	**74.59**	**75.56**	**73.79**	**73.64**	**73.61**
Δ	+1.16	+1.46	+2.69	+2.21	+1.57	+2.16	+1.88

4.7 MECHANISM ANALYSIS

We measure the transfer of trajectory structure on the unaugmented CIFAR-100 training set with a single forward pass in fixed order. All measurements use the non-target relation component $\boldsymbol{r}$ of equation 2, and the trajectory target is the aggregate $\mathbf{z}^\star$ of equation 3.

Figure 3 compares students from six methods in relation space, namely TraCo-KD, DKD, LSKD, ReviewKD, FitNet and RKD. The TraCo-KD student is closest to the trajectory target, but not to the converged endpoint, to which the DKD and LSKD students are closer. This shows that TraCo-KD guides the student's non-target relation structure toward the trajectory target. Students trained with feature distillation are farther from both targets, consistent with their supervision not acting on logit relation structure.

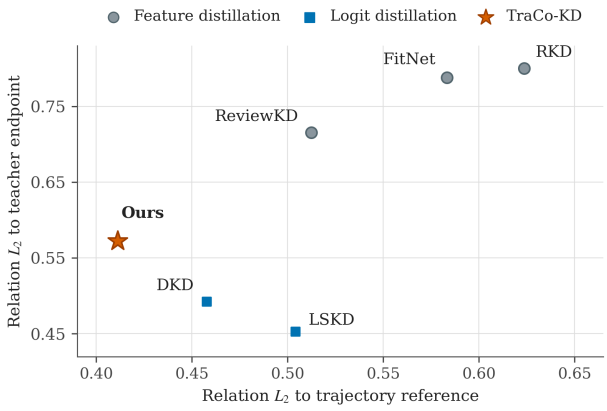

Figure 3: **Students of six methods in non-target relation space** on ResNet32×4 to ResNet8×4, CIFAR-100. The horizontal axis is the relation L_2 distance to the trajectory target and the vertical axis is the relation L_2 distance to the converged endpoint. Gray circles denote feature distillation methods, and blue squares denote logit distillation methods.

5 CONCLUSION

This work investigates the source of supervision in logit distillation. We observe that teacher classification accuracy and non-target relational structure do not evolve in synchrony during training. Accuracy peaks at the training endpoint, whereas the relational structure needed for distillation is distributed across the full trajectory, making it difficult for a teacher at any single training stage to provide both. Based on this observation, we propose TraCo-KD, which decouples accuracy from relational structure by compiling relations aggregated across training stages offline into a frozen operator acting on converged logits. The operator is placed before the existing distillation loss, and distillation requires only one forward pass through the converged teacher. On CIFAR-100 and ImageNet, TraCo-KD achieves the best performance among logit-based methods and consistently improves feature-based distillation. These results show that the teacher's training trajectory can provide supervision beyond the converged teacher and that this supervision can be exploited as a function of the converged output.

AI USE STATEMENT

In this work, we used generative AI tools for translating the manuscript from Chinese into English. We have not used generative AI tools for research ideation, methodology or experiment design, method implementation, or interpretation of experimental results, and synthetic dataset generation and mathematical proof development are not applicable to this work. Additionally, we used generative AI tools for drafting portions of the manuscript, revising and polishing the writing, and searching for and identifying related literature. We have reviewed all AI-assisted work. We take

responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

Code is available at `https://anonymous.4open.science/r/test12345-00BB/`.

References

Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D. Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10925–10934, June 2022.

Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7028–7036, 2021a.

Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11933–11942, June 2022.

Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5008–5017, June 2021b.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1607–1616. PMLR, 2018.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

Yuchen Guan, Runxi Cheng, Kang Liu, and Chun Yuan. Enhancing logits distillation with Plug&Play Kendall's τ ranking loss. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 20537–20552. PMLR, 2025.

Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. Online knowledge distillation via collaborative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. In *International Conference on Learning Representations (ICLR)*, 2016.

Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1921–1930, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2014.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get M for free. In *International Conference on Learning Representations (ICLR)*, 2017.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 33716–33727. Curran Associates, Inc., 2022.

Xiao Jin, Baoyun Peng, Yichao Wu, Yu Liu, Jiaheng Liu, Ding Liang, Junjie Yan, and Xiaolin Hu. Knowledge distillation via route constrained optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Xu Lan, Xiatian Zhu, and Shaogang Gong. Knowledge distillation by on-the-fly native ensemble. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5191–5198, 2020.

Abhishek Panigrahi, Bingbin Liu, Sadhika Malladi, Andrej Risteski, and Surbhi Goel. Progressive distillation induces an implicit curriculum. In *International Conference on Learning Representations (ICLR)*, 2025.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Mehdi Rezagholizadeh, Aref Jafari, Puneeth S. M. Saladi, Pranav Sharma, Ali Saheb Pasand, and Ali Ghodsi. Pro-KD: Progressive distillation by following the footsteps of the teacher. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 4714–4727, 2022.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations (ICLR)*, 2015.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15731–15740, June 2024.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations (ICLR)*, 2020.

Chaofei Wang, Qisen Yang, Rui Huang, Shiji Song, and Gao Huang. Efficient knowledge distillation from model checkpoints. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 607–619. Curran Associates, Inc., 2022.

Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge distillation meets self-supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 588–604. Springer, 2020.

Chenglin Yang, Lingxi Xie, Chi Su, and Alan L. Yuille. Snapshot distillation: Teacher-student optimization in one generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Zhendong Yang, Ailing Zeng, Zhe Li, Tianke Zhang, Chun Yuan, and Yu Li. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17185–17194, October 2023.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations (ICLR)*, 2017.

Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11953–11962, June 2022.