FROM RANKING SKELETONS TO TEACHER PREDICTIONS: KNOWLEDGE DISTILLATION ALONG CONTINUOUS TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard logit-based knowledge distillation trains the student to match the teacher's predictions, transferring class rankings and response magnitudes through a single supervision target. This formulation offers limited flexibility in how these two forms of knowledge are distilled. To address this limitation, we propose Ranking-guided Trajectory Knowledge Distillation (RTKD), which extends teacher supervision to a continuous family of targets. We first construct a ranking skeleton that retains the teacher's class ordering while discarding its original response gaps, and then connect the normalized skeleton and teacher prediction via a spherical trajectory. Along this path, class ordering is preserved, while relative response magnitudes gradually change from the skeleton to the teacher prediction. Distillation is performed by integrating the student's discrepancies from targets along the trajectory, with a teacher-dependent prior assigning greater weight to targets closer to the teacher endpoint. An adaptive Huber penalty is applied to geodesic discrepancies to limit the influence of large errors. We evaluate RTKD on two image classification benchmarks, CIFAR-100 and ImageNet-1K, and conduct ablation studies to assess the contributions of its components. On ImageNet-1K, RTKD improves Top-1 accuracy from 70.49% with standard KD to 73.15% when distilling ResNet50 into MobileNet-V1. The code will be released soon.

1 INTRODUCTION

Knowledge Distillation (KD) provides a practical way to improve compact neural networks by transferring knowledge from a stronger teacher. In standard logit-based KD Hinton et al. (2015), the student learns from the teacher's softened predictions, which convey relationships among classes beyond the ground-truth label. Subsequent studies have improved this supervision by decoupling target-class and non-target-class knowledge Zhao et al. (2022), matching prediction correlations Huang et al. (2022), and relaxing constraints on logit magnitudes Sun et al. (2024). These advances suggest that effective distillation depends on how teacher knowledge is presented to the student, as well as on the knowledge itself.

A teacher prediction contains both class ordering and response magnitudes. Class ordering indicates which classes the teacher prefers, while the gaps between responses describe the strength of these preferences. Standard prediction matching transfers both through a single target distribution. However, the same class ordering can be expressed by many different response patterns, and a student with a different architecture or capacity may benefit from supervision beyond the teacher's particular pattern. Retaining only the ordering would discard useful information about response gaps, whereas matching only the original prediction would leave this flexibility unexplored. This motivates a supervision scheme that connects the two forms of knowledge and allows both to contribute to distillation.

To this end, we propose Ranking-guided Trajectory Knowledge Distillation (RTKD), as illustrated in Fig. 1. RTKD constructs a continuous family of targets between a ranking skeleton and the teacher prediction, and trains the student by integrating supervision along this trajectory. The ranking skeleton assigns class scores according to their positions in the teacher's ranking. It therefore preserves the full class ordering without retaining the original response gaps. This provides an ordinal ref-

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.

erence that is more informative than a top-class label, while remaining distinct from the teacher's detailed prediction.

To connect these two representations, we center and normalize the skeleton and teacher logits, and construct the shortest spherical path between them. Centering removes the common logit offset, while normalization places both endpoints on the same scale. Previous work has established the usefulness of normalization for distillation Sun et al. (2024). Here, it also provides a common geometric space for constructing intermediate targets. The spherical path keeps every target normalized and preserves the teacher's class ordering throughout. Relative response magnitudes vary continuously along the path, connecting the rank-based scores to the teacher's response pattern without introducing conflicting class rankings.

The remaining question is how to use this family of targets. Selecting a single point would require choosing one particular response pattern, while matching only the endpoints would omit the intermediate targets. RTKD instead integrates the student's discrepancies from targets along the trajectory. A teacher-dependent prior assigns greater weight to targets near the teacher endpoint, retaining the teacher's detailed responses as the main reference while allowing the ranking skeleton to influence supervision. We measure discrepancies using geodesic distance and apply an adaptive Huber penalty. Its quadratic region supports precise alignment for small errors, while its linear region limits the influence of large errors. The transition threshold follows the trajectory's angular span, with a lower bound for short trajectories. The resulting distillation objective is combined with cross-entropy supervision, and only the student is updated.

We evaluate our RTKD on CIFAR-100 and ImageNet-1K using different teacher-student architectures. On ImageNet-1K, RTKD achieves 73.15% Top-1 accuracy when distilling ResNet50 into MobileNet-V1, improving over standard KD by 2.66 percentage points. Ablation studies further examine the roles of the ranking skeleton, trajectory integration, teacher-dependent prior, and adaptive penalty. Our main contributions are summarized as follows:

- We introduce a ranking-guided construction of continuous distillation targets, connecting ordinal class structure to teacher response magnitudes through an ordering-preserving spherical trajectory.
- We formulate distillation as an integral over this trajectory, combining a teacher-dependent prior with an adaptive Huber geodesic objective.
- We evaluate the framework on CIFAR-100 and ImageNet-1K and analyze its principal components through controlled ablation studies.

2 Related Work

Feature-based knowledge distillation. Feature-based methods transfer internal teacher representations through intermediate hints Romero et al. (2014), relational or contrastive objectives Park et al. (2019); Tian et al. (2019), cross-level connections Chen et al. (2021), attention transfer Zagoruyko & Komodakis (2016); Guo et al. (2023), and functional consistency Liu et al. (2023a). FOFA Lin et al. (2025) further addresses feature transfer between heterogeneous architectures. These methods provide useful reference points for our evaluation, but aligning internal representations often requires architecture-dependent components. Output predictions offer a common class space for transferring knowledge across different model families.

Logit-based knowledge distillation. Standard KD Hinton et al. (2015) aligns temperature-softened teacher and student predictions. DKD Zhao et al. (2022) separates target-class and non-target-class supervision, while LSKD Sun et al. (2024) relaxes magnitude constraints through logit standardization. Recent methods further enrich output supervision: EKD Xiang et al. (2025) transfers evidential knowledge through Dirichlet distributions, LDRLD Xu et al. (2025) models local logit relations, and RCKD Wang et al. (2026) aligns relative confidence matrices using cosine similarity. For heterogeneous architectures, OFA Hao et al. (2023) projects intermediate student features into the logit space and enhances target-class supervision. OFA and FOFA are representative methods specifically designed for cross-architecture distillation.

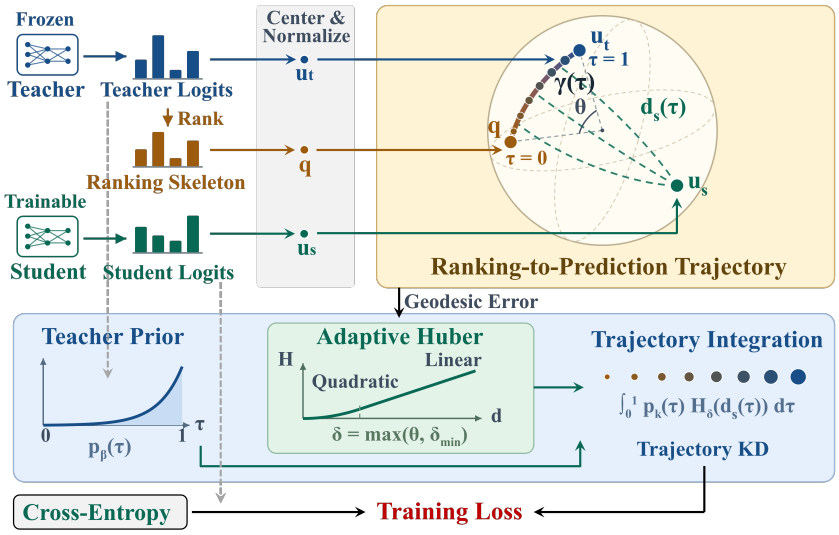


Figure 1: **Overview of our RTKD.** A continuous spherical trajectory connects the ranking skeleton to the teacher prediction. A teacher-dependent prior and an adaptive Huber geodesic loss guide distillation along this trajectory, jointly with cross-entropy supervision.

RTKD constructs a continuous family of targets by connecting a ranking skeleton to the normalized teacher prediction. Class ordering is preserved along the path, while relative response gaps vary. Integrating supervision over this trajectory distinguishes RTKD from directly aligning a fixed prediction or relational representation.

3 Method

We propose Ranking-guided Trajectory Knowledge Distillation (RTKD), which constructs a continuous family of supervision targets from the teacher's class ranking and output prediction. Given a training sample (x, y), let $\mathbf{z}_t, \mathbf{z}_s \in \mathbb{R}^C$ denote the teacher and student logits, where C is the number of classes. RTKD first extracts a ranking skeleton, connects it to the teacher prediction through a spherical trajectory, and integrates the student's discrepancies from targets along this trajectory. The teacher remains fixed throughout training. We omit the dependence on x where the context is clear.

3.1 Ranking Skeleton

Teacher logits describe both the ordering of classes and the gaps between their responses. To obtain a reference that isolates the ordering, we construct a ranking skeleton using rank-based scores. Let r_c denote the descending rank of class c in the teacher logits, with rank 1 assigned to the largest response. The skeleton scores are defined as

$$s_c = C + 1 - r_c, \qquad c = 1, \ldots, C. \tag{1}$$

Classes with tied teacher logits receive the average score of their occupied rank positions. This construction retains the teacher's full class ordering while replacing its original response gaps with a common rank-based spacing. Unlike a top-class label, the skeleton also preserves the ordering among non-top classes. It serves as an ordinal reference for constructing supervision, without assuming that the teacher's response magnitudes should be discarded altogether.

The skeleton scores and teacher logits generally have different offsets and scales. Directly connecting them would make the resulting targets depend on these arbitrary differences. We therefore use the same centering and normalization operation for the skeleton, teacher, and student:

$$\mathcal{N}(\mathbf{v}) = \frac{\mathbf{v} - \bar{v}\mathbf{1}}{\|\mathbf{v} - \bar{v}\mathbf{1}\|_2}, \qquad \bar{v} = \frac{1}{C}\sum_{c=1}^{C} v_c, \tag{2}$$

and write $\mathbf{q} = \mathcal{N}(\mathbf{s})$, $\mathbf{u}_t = \mathcal{N}(\mathbf{z}_t)$, and $\mathbf{u}_s = \mathcal{N}(\mathbf{z}_s)$. Centering removes a common offset, which does not affect class probabilities under softmax, while normalization removes the overall scale. These representations lie on the unit sphere in the centered logit subspace. The distinction between the two teacher-derived representations is thus their pattern of relative responses: $\mathbf{q}$ encodes rank-based spacing, whereas $\mathbf{u}_t$ retains the teacher's normalized response gaps.

3.2 Ranking-to-Prediction Trajectory

The ranking skeleton and teacher prediction provide complementary endpoints, but using either endpoint alone omits the response patterns between them. We connect the two by the shortest spherical path. Let $\theta = \arccos(\mathbf{q}^\top \mathbf{u}_t)$ denote their angular separation. The trajectory is parameterized by

$$\boldsymbol{\gamma}(\tau) = \frac{\sin((1-\tau)\theta)}{\sin\theta}\mathbf{q} + \frac{\sin(\tau\theta)}{\sin\theta}\mathbf{u}_t, \qquad \tau \in [0,1]. \tag{3}$$

Its endpoints satisfy $\boldsymbol{\gamma}(0) = \mathbf{q}$ and $\boldsymbol{\gamma}(1) = \mathbf{u}_t$. When $\theta = 0$, the trajectory reduces to the common endpoint.

The spherical construction keeps all targets centered and normalized, so movement along the path changes relative class responses without introducing an overall scale change. It also provides a direct geometric interpretation of the path parameter: the angular distance from the skeleton is $\tau\theta$. More importantly, the trajectory preserves the teacher's class ordering. The skeleton and teacher prediction have the same ordering, and the two coefficients in Eq. 3 are nonnegative. Consequently, if the teacher ranks class i above class j, then $\gamma_i(\tau) > \gamma_j(\tau)$ throughout the path. The intermediate targets therefore vary the response gaps while maintaining a consistent ordering.

This construction defines a family of targets for each input. Targets near the skeleton emphasize rank-based structure, while those near the teacher retain more of its original response pattern. The parameter τ indexes these targets, and supervision from the entire path is used within each training step.

3.3 Integrated Trajectory Distillation

Having constructed the trajectory, we aggregate its supervision through a weighted integral. Uniform weighting would treat all positions equally, although the teacher endpoint contains response information removed by the skeleton construction. We therefore introduce a teacher-dependent prior that favors targets closer to the teacher as

$$p_\kappa(\tau \mid x) = \frac{\exp(-\sqrt{\kappa}\, c_t(\tau))}{\int_0^1 \exp(-\sqrt{\kappa}\, c_t(v))\, \mathrm{d}v}, \qquad c_t(\tau) = \|\boldsymbol{\gamma}(\tau) - \mathbf{u}_t\|_2 = 2\sin\left(\frac{(1-\tau)\theta}{2}\right). \tag{4}$$

Here, $\kappa \geq 0$ controls the preference for the teacher endpoint, with $\kappa = 0$ yielding a uniform prior. Increasing κ concentrates supervision near the teacher, while decreasing it gives the remaining trajectory a larger relative contribution. The prior depends on the teacher-derived geometry and assigns weights to supervision targets. It does not estimate whether the teacher prediction is correct. Chord distance $c_t(\tau)$ provides a bounded measure of target proximity for this weighting.

For student alignment, we use the geodesic discrepancy

$$d_s(\tau) = \arccos\big(\mathbf{u}_s^\top \boldsymbol{\gamma}(\tau)\big). \tag{5}$$

This measures the angular difference between normalized responses, in the same units as the trajectory span θ. A purely quadratic penalty would allow its slope to grow continuously with this discrepancy. To retain precise alignment for small errors while limiting the influence of large errors, we use

$$H_\delta(d) = \begin{cases} d^2, & 0 \leq d \leq \delta, \\ 2\delta d - \delta^2, & d > \delta, \end{cases} \qquad \delta(x) = \max(\theta(x), \delta_{\min}). \tag{6}$$

The derivative with respect to d is $2\min(d, \delta)$, so large discrepancies have a bounded slope. We use the trajectory span as the transition scale because it describes the angular variation already present in the target family. Since any two points on the trajectory are separated by at most θ, this choice keeps discrepancies within that scale in the quadratic region. The positive lower bound $\delta_{\min}$ prevents

Table 1: Top-1 accuracy (%) on the CIFAR-100 test set with teacher-student pairs from the same architecture family. The best results among feature/logit-based distillation methods are highlighted in **black**/**red**, while the second-best logit-based results are underlined.

Method	Pub.	ResNet56 72.34 ResNet20 69.06	ResNet110 74.31 ResNet32 71.14	ResNet32x4 79.42 ResNet8x4 72.50	WRN-40-2 75.61 WRN-16-2 73.26	WRN-40-2 75.61 WRN-40-1 71.98	VGG13 74.64 VGG8 70.36
FitNet	ICLR'15	69.21	71.06	73.50	73.58	72.24	71.02
RKD	CVPR'19	69.61	71.82	71.90	73.35	72.22	71.48
CRD	ICLR'20	71.16	73.48	75.51	75.48	74.14	73.94
ReviewKD	CVPR'21	**71.89**	73.89	75.63	**76.12**	**75.09**	74.84
SimKD	CVPR'22	71.05	**73.92**	**78.08**	75.53	74.53	**74.89**
CAT-KD	CVPR'23	71.62	73.62	76.91	75.60	74.82	74.65
NORM	ICLR'23	71.35	73.67	76.49	75.65	74.82	73.95
TopKD	ICML'24	71.58	73.77	75.40	75.75	74.43	74.01
KD	ArXiv'15	70.66	73.08	73.33	74.92	73.54	72.98
DKD	CVPR'22	**71.97**	74.11	76.32	76.24	74.81	74.68
CTKD	AAAI'23	71.19	73.52	73.70	75.45	73.93	73.52
NKD	ICCV'23	71.18	73.50	76.35	75.43	74.25	74.86
LSKD	CVPR'24	71.43	74.17	76.62	76.11	74.37	74.36
SDD	CVPR'24	71.52	73.97	75.09	75.86	74.53	73.91
WTTM	ICLR'24	71.92	74.13	76.06	76.37	74.58	74.44
TeKAP	ICLR'25	71.32	73.42	74.79	75.21	73.80	74.00
EKD	ICCV'25	71.48	73.68	77.21	76.15	74.43	74.71
RCKD	AAAI'26	71.83	**74.26**	76.89	**76.60**	74.67	74.86
RTKD	OURS	71.59	73.99	**77.48**	76.12	**74.90**	**74.97**

the penalty scale from vanishing when the skeleton and teacher prediction nearly coincide. The threshold thus adapts to the teacher's target geometry.

Therefore, the final per-sample distillation objective is defined as

$$\mathcal{L}_{\mathrm{RTKD}}(x) = \int_0^1 p_\kappa(\tau \mid x)\, H_{\delta(x)}(d_s(\tau))\, \mathrm{d}\tau. \tag{7}$$

Integrating the discrepancies allows all trajectory positions to contribute according to their prior weights, without selecting a single intermediate target. The prior and penalty play separate roles: the former determines which targets receive greater emphasis, while the latter controls how strongly their discrepancies affect optimization.

For a minibatch $\mathcal{B}$, the final training objective combines trajectory distillation with ground-truth supervision:

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{(x,y)\in\mathcal{B}} \left[\mathcal{L}_{\mathrm{CE}}(\mathbf{z}_s(x), y) + \lambda\, \mathcal{L}_{\mathrm{RTKD}}(x)\right], \tag{8}$$

where λ controls the overall distillation strength and the cross-entropy loss is computed from the original student logits. The ranking skeleton, trajectory, prior, and adaptive threshold are derived from the frozen teacher model, and gradients update only the student. RTKD requires no additional prediction head, and inference uses the student network alone.

4 Experiments

4.1 Main Results

Datasets and experimental setup. We evaluate RTKD on CIFAR-100 Krizhevsky et al. (2009) and ImageNet-1K Russakovsky et al. (2015). CIFAR-100 contains 50,000 training images and 10,000 test images across 100 classes, with an original resolution of 32×32. ImageNet-1K comprises approximately 1.28 million training images and 50,000 validation images from 1,000 classes. For the conventional CNN teacher-student comparisons in Tables 1, 2, and 3, we follow the data preprocessing, augmentation, and optimization settings of DKD Zhao et al. (2022). We report Top-1 accuracy (%) on the CIFAR-100 test set and both Top-1 and Top-5 accuracy (%) on the ImageNet-1K validation set. RTKD is implemented in PyTorch, with pretrained teachers kept frozen during student training. Unless otherwise specified, we use $\delta_{\min} = 0.1$. Our primary comparisons are with

Table 2: Top-1 accuracy (%) on the CIFAR-100 test set with teacher-student pairs from different architecture families. The best results among feature/logit-based distillation methods are highlighted in **black/red**, while the second-best logit-based results are underlined. "–" denotes unavailable results.

Method	Pub.	ResNet32x4 79.42 ShuffleNet-V1 70.50	WRN-40-2 75.61 ShuffleNet-V1 70.50	ResNet50 79.34 MobileNet-V2 64.60	VGG13 74.64 MobileNet-V2 64.60	ResNet32x4 79.42 ShuffleNet-V2 71.82	WRN-40-2 75.61 ResNet8x4 72.50
FitNet	ICLR'15	73.54	73.73	63.16	64.16	73.54	74.61
RKD	CVPR'19	72.28	72.21	64.43	64.52	73.21	75.26
CRD	ICLR'20	75.11	76.05	69.11	69.73	75.65	75.24
ReviewKD	CVPR'21	77.45	77.14	69.89	**70.37**	77.78	74.34
CAT-KD	CVPR'23	**78.26**	**77.35**	**71.36**	69.13	**78.41**	75.39
NORM	ICLR'23	77.42	77.06	70.56	68.94	78.07	75.36
TopKD	ICML'24	–	–	70.11	69.32	76.33	**75.83**
KD	ArXiv'15	74.07	74.83	67.35	67.37	74.45	73.79
DKD	CVPR'22	76.45	76.70	70.35	69.71	77.07	75.56
CTKD	AAAI'23	74.48	75.78	68.47	68.46	75.31	74.61
NKD	ICCV'23	76.31	76.46	70.22	**70.67**	76.92	76.01
LSKD	CVPR'24	75.19	76.03	69.02	68.61	75.56	77.11
SDD	CVPR'24	76.30	76.65	69.55	68.79	76.67	74.83
WTTM	ICLR'24	74.37	75.42	69.59	69.16	76.55	–
TeKAP	ICLR'25	74.92	76.75	69.00	67.39	75.43	74.99
EKD	ICCV'25	–	–	70.38	69.94	77.65	76.75
RCKD	AAAI'26	76.51	76.97	69.67	69.77	77.13	**77.23**
RTKD	OURS	**76.97**	**77.16**	**71.26**	70.24	**77.86**	76.62

logit-based methods, including KD Hinton et al. (2015), DKD Zhao et al. (2022), CTKD Li et al. (2023), NKD Yang et al. (2023), LSKD Sun et al. (2024), SDD Wei et al. (2024), WTTM Zheng & Yang (2024), TeKAP Hossain et al. (2025), EKD Xiang et al. (2025), and RCKD Wang et al. (2026). Feature-based methods, including FitNet Romero et al. (2014), RKD Park et al. (2019), CRD Tian et al. (2019), ReviewKD Chen et al. (2021), SimKD Chen et al. (2022), CAT-KD Guo et al. (2023), NORM Liu et al. (2023b), and TopKD Kim et al. (2024), are included as references. The separate protocol for CNN, Transformer, and MLP combinations is described with those results below.

Results on CIFAR-100. Tables 1 and 2 compare teacher-student pairs within and across CNN architecture families, respectively. Within the same family, RTKD achieves the best accuracy among logit-based methods on three of the six pairs. It reaches 77.48% on ResNet32x4→ResNet8x4, exceeding EKD and RCKD by 0.27 and 0.59 percentage points (pp), respectively, and also leads on WRN-40-2→WRN-40-1 and VGG13→VGG8. Across architecture families, RTKD ranks first among logit-based methods on four of the six pairs and outperforms RCKD on five. In particular, it achieves 71.26% on ResNet50→MobileNet-V2 and 77.86% on ResNet32x4→ShuffleNet-V2, improving over RCKD by 1.59 and 0.73 pp, respectively. NKD and RCKD remain stronger on VGG13→MobileNet-V2 and WRN-40-2→ResNet8x4, respectively. These results support the effectiveness of trajectory supervision across architecture families, while showing that the gains depend on the teacher-student pair.

Results on ImageNet-1K. Table 3 reports Top-1 and Top-5 accuracy on ImageNet-1K. For ResNet34→ResNet18, RTKD obtains 71.76% Top-1 and 90.57% Top-5 accuracy, with the second-best Top-1 result among logit-based methods, although RCKD performs better on both metrics. For ResNet50→MobileNet-V1, RTKD leads the logit-based comparisons with 73.15% Top-1 and 91.35% Top-5 accuracy, exceeding standard KD by 2.66 and 1.43 pp and RCKD by 0.74 and 0.53 pp, respectively. These results extend the effectiveness of trajectory supervision to large-scale classification, with a clearer advantage in the evaluated cross-architecture pair.

Results across CNN, Transformer, and MLP architectures. We separately evaluate ten heterogeneous teacher-student pairs on CIFAR-100 under the experimental protocol of OFA Hao et al. (2023), including upsampling inputs to 224×224. This setting is distinct from the conventional CNN comparisons above. We include OFA and FOFA Lin et al. (2025), two representative methods specifically designed for heterogeneous knowledge distillation, with FOFA serving as a feature-based reference. As shown in Table 4, RTKD achieves the best results among logit-based methods

Table 3: Top-1 and Top-5 accuracy (%) on the ImageNet-1K validation set. The best results among feature/logit-based distillation methods are highlighted in **black**/**red**, while the second-best logit-based results are underlined.

ResNet34 → ResNet18 Teacher: 73.31 / 91.42 Student: 69.75 / 89.07												
	Feature-based					Logit-based						
	AT	CRD	ReviewKD	FCFD	CAT-KD	KD	DKD	LSKD	TeKAP	EKD	RCKD	RTKD
Top-1	70.69	71.17	71.61	**72.24**	71.26	70.66	71.70	71.42	71.35	71.73	**72.02**	71.76
Top-5	90.01	90.13	90.51	**90.74**	90.45	89.88	90.31	90.29	90.54	**90.69**	90.67	90.57
ResNet50 → MobileNet-V1 Teacher: 76.16 / 92.87 Student: 68.87 / 88.76												
	Feature-based					Logit-based						
	AT	CRD	ReviewKD	FCFD	CAT-KD	KD	DKD	LSKD	TeKAP	EKD	RCKD	RTKD
Top-1	70.18	71.32	72.56	**73.37**	72.24	70.49	72.05	72.18	72.87	72.54	72.41	**73.15**
Top-5	89.68	90.41	91.00	**91.35**	91.13	89.92	91.05	90.80	91.05	91.20	90.82	**91.35**

Table 4: Top-1 accuracy (%) on CIFAR-100 using cross-architecture teacher-student pairs. The best results among logit-based methods are shown in **red**.

Teacher	Student	From Scratch		Feature-based			Logit-based				
		T. Acc.	S. Acc.	FitNet	CRD	FOFA	KD	DKD	OFA	LDRLD	RTKD
Swin-T	ResNet18	89.26	74.01	78.87	77.63	81.22	78.74	80.26	80.54	82.17	**83.65**
ViT-S	ResNet18	92.04	74.01	77.71	76.60	80.11	77.26	78.10	80.15	80.36	**80.80**
Mixer-B/16	ResNet18	87.29	74.01	77.15	76.42	80.07	77.79	78.67	79.39	**80.69**	80.30
Swin-T	MobileNet-V2	89.26	73.68	74.28	79.80	78.78	74.68	71.07	80.98	81.64	**84.86**
ViT-S	MobileNet-V2	92.04	73.68	73.54	78.14	78.87	72.77	69.80	78.45	79.21	**82.45**
Mixer-B/16	MobileNet-V2	87.29	73.68	73.78	78.15	78.62	73.33	70.20	78.78	80.64	**81.42**
ConvNeXt-T	DeiT-T	88.41	68.00	60.78	65.94	79.59	72.99	74.60	75.76	77.46	**77.87**
Mixer-B/16	DeiT-T	87.29	68.00	71.05	65.35	74.66	71.36	73.44	73.90	75.31	**75.48**
ConvNeXt-T	Swin-P	88.41	72.63	24.06	67.09	80.74	76.44	76.80	78.32	**80.71**	79.49
Mixer-B/16	Swin-P	87.29	72.63	75.20	67.03	78.44	75.93	76.39	78.93	**80.52**	78.31

on seven of the ten pairs, with an average Top-1 accuracy of 80.46%, exceeding OFA and LDRLD by 1.94 and 0.59 pp, respectively. The gains are particularly notable for Transformer-to-CNN distillation: RTKD achieves 84.86% and 82.45% with Swin-T and ViT-S teachers on MobileNet-V2, surpassing OFA by 3.88 and 4.00 pp. With DeiT-T students, it improves over OFA by 2.11 and 1.58 pp using ConvNeXt-T and Mixer-B/16 teachers, respectively. These results show that RTKD transfers knowledge effectively across heterogeneous model families using only output logits.

4.2 Ablation Studies

We evaluate RTKD on CIFAR-100 with ResNet32x4 as the teacher and ResNet8x4 as the student. Table 5 reports Top-1 accuracy for all 16 configurations at a fixed distillation weight of $\lambda = 75$. The full configuration uses $\kappa = 30$ and $\delta_{\min} = 0.1$, while uniform-prior variants remove the preference for the teacher endpoint. Differences are expressed in percentage points (pp). Full RTKD combines the full-ranking trajectory, teacher-dependent prior, and adaptive Huber geodesic penalty. All variants retain the cross-entropy term and modify only the specified components.

Trajectory construction. We compare full RTKD with supervision at the *Teacher endpoint*, *Skeleton endpoint*, or *Prior mean point* $\boldsymbol{\gamma}(\mathbb{E}_{p_\kappa}[\tau])$, as well as a *Two-endpoint mixture* with renormalized prior weights. The point-target variants retain the trajectory-dependent Huber threshold. We also test a *Top-1 skeleton* constructed without the full class ranking. Full RTKD achieves 77.48%, compared with 77.00%, 71.01%, 77.14%, 76.61%, and 75.36% for these five variants, respectively. These results indicate that the full ranking and teacher response gaps both contribute to supervision, while integrating intermediate targets improves upon the tested single-target and endpoint-mixture alternatives.

Teacher prior and robust penalty. *Uniform prior* replaces the teacher-dependent density with $p(\tau) = 1$. *Squared geodesic* replaces the Huber penalty with $d_s(\tau)^2$, and *Uniform prior + squared*

Table 5: Ablation studies on CIFAR-100 with ResNet32x4 → ResNet8x4. Each row modifies the indicated components of full RTKD. Δ denotes the Top-1 accuracy difference from the full method in percentage points (pp). Fixed Huber thresholds δ_0 are in radians.

Component	Configuration	Top-1 (%)	Δ (pp)
Target distribution	Teacher endpoint	77.00	−0.48
	Skeleton endpoint	71.01	−6.47
	Prior mean point	77.14	−0.34
	Two-endpoint mixture	76.61	−0.87
Skeleton and distance	Top-1 skeleton	75.36	−2.12
	Huber chord distance	77.33	−0.15
Prior and penalty	Uniform prior	76.88	−0.60
	Squared geodesic	76.35	−1.13
	Uniform prior + squared geodesic	76.71	−0.77
Huber threshold	Fixed Huber ($\delta_0 = 0.1$)	76.01	−1.47
	Uniform prior + fixed Huber ($\delta_0 = 0.1$)	76.12	−1.36
	Fixed Huber ($\delta_0 = 0.25$)	77.03	−0.45
	Uniform prior + fixed Huber ($\delta_0 = 0.25$)	76.63	−0.85
	Fixed Huber ($\delta_0 = 0.5$)	76.46	−1.02
	Uniform prior + fixed Huber ($\delta_0 = 0.5$)	76.78	−0.70
Full RTKD		**77.48**	–

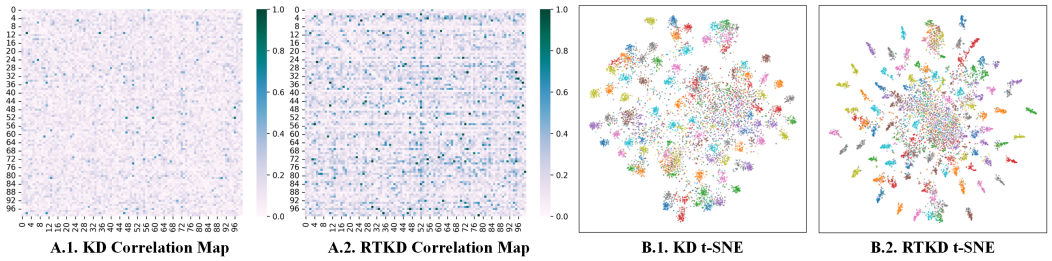


Figure 2: Class-wise teacher-student logit discrepancy maps (left) and t-SNE visualizations of student outputs (right), comparing KD and RTKD in each pair using ResNet8x4 (distilled by ResNet32x4).

geodesic applies both changes. These variants obtain 76.88%, 76.35%, and 76.71%, respectively. With adaptive Huber, teacher-dependent weighting improves over the uniform prior by 0.60 pp. But with squared geodesic energy, it instead lowers accuracy by 0.36 pp. Conversely, the gain from Huber is 1.13 pp under the teacher prior but only 0.17 pp under the uniform prior. This dependence supports pairing teacher-proximal weighting with a penalty that limits the influence of large angular discrepancies, rather than attributing the improvement to weighting alone.

Adaptive threshold and distance geometry. *Fixed Huber* replaces $\delta(x) = \max\{\theta(x), \delta_{\min}\}$ with a sample-independent threshold $\delta_0 \in \{0.1, 0.25, 0.5\}$; its *Uniform prior* counterparts additionally remove teacher-dependent weighting. *Huber chord distance* replaces the geodesic discrepancy with $\|\mathbf{u}_s - \boldsymbol{\gamma}(\tau)\|_2$ while retaining the Huber form and mapping its threshold to chord distance. Under the teacher prior, the best tested fixed threshold is 0.25, yielding 77.03%, and under the uniform prior, it is 0.5, yielding 76.78%. Adaptive Huber improves on these results by 0.45 and 0.10 pp, respectively. At the same distillation weight, these comparisons favor tying the penalty's transition scale to the trajectory span. The chord distance variant reaches 77.33%, close to the full method's 77.48%, indicating that the framework's accuracy is largely retained under this alternative spherical distance.

4.3 Additional Analyses

Visualization analysis. Fig. 2 compares class-wise teacher-student logit discrepancies and t-SNE projections of student outputs. RTKD exhibits larger discrepancies in class-averaged non-target logits than KD, yet achieves higher classification accuracy, reaching 77.48%. This indicates that closer matching of raw teacher logits does not necessarily yield better student performance. Meanwhile,

RTKD displays several more compact and distinct peripheral clusters in the t-SNE projection, although central overlap remains. These observations are consistent with the motivation of RTKD that transferring class ordering and relative response structure through trajectory supervision allows the student to learn useful prediction patterns without exactly reproducing the teacher's response magnitudes.

Sensitivity to κ. We examine the effect of κ, which controls the concentration of the trajectory prior toward the teacher prediction. As shown in Fig. 3, both teacher-student pairs achieve their highest accuracy at $\kappa = 30$, reaching 77.48% for ResNet32x4→ResNet8x4 and 77.86% for ResNet32x4→ShuffleNet-V2. Accuracy varies non-monotonically with κ, and larger values do not provide further gains. These results favor a moderate preference for the teacher endpoint while retaining supervision from the rest of the trajectory. We therefore use $\kappa = 30$ unless otherwise specified.

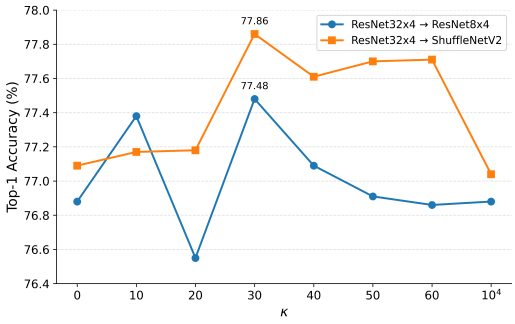


Figure 3: Sensitivity to discrete values of κ.

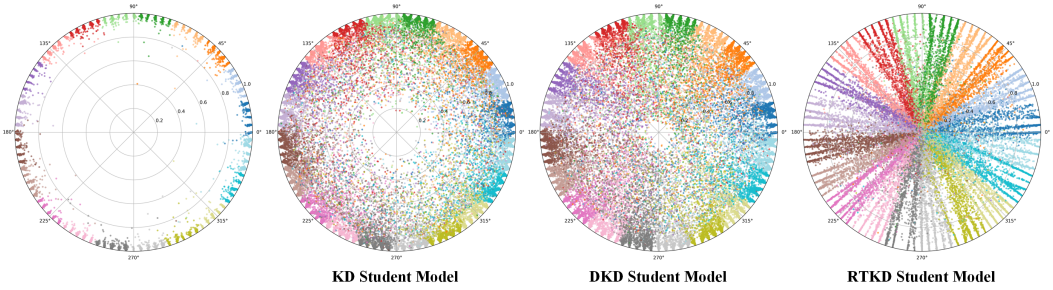


Figure 4: Polar visualizations of teacher and student predictions on the CIFAR-100 training set, with radius indicating one minus normalized prediction entropy and angle representing the probability-weighted circular mean of class directions.

Prediction distribution visualization. Fig. 4 visualizes the teacher and student predictions in polar coordinates, where the radius represents one minus the normalized prediction entropy and the angle is the probability-weighted circular mean of fixed class directions. The teacher predictions concentrate near the outer boundary, indicating low predictive entropy. KD and DKD students exhibit broader angular dispersion, whereas RTKD produces more distinct radial bands across a wide range of radii. This pattern suggests that RTKD produces more concentrated class-associated directions in this projection while allowing variation in prediction entropy. The observation is consistent with its training objective, which transfers ranking and relative response structure through normalized logits without requiring the student to reproduce the teacher's overall logit scale.

5 Conclusion

In this paper, we proposed Ranking-guided Trajectory Knowledge Distillation (RTKD), which extends teacher supervision from a single prediction to a continuous family of targets. A ranking skeleton provides an ordinal reference, and a spherical trajectory connects it to the normalized teacher prediction while preserving class ordering. Knowledge is transferred by integrating the student's geodesic discrepancies along this trajectory, with a teacher-dependent prior controlling target contributions and an adaptive Huber penalty limiting the influence of large errors. Experiments on CIFAR-100 and ImageNet-1K demonstrate competitive performance across diverse teacher-student pairs, with particularly consistent gains in the evaluated cross-architecture settings. Ablation studies further support the contributions of the ranking skeleton, trajectory integration, and adaptive penalty. These results show that connecting class ordering with relative response magnitudes provides an effective form of output-level supervision.

AI USE STATEMENT

We used generative AI tools for language polishing, partial code implementation, and suggestions on experimental design. All AI-assisted content was reviewed and verified by the authors. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work complies with the ICLR Code of Ethics. It does not involve human subjects, private or sensitive data, or other ethical concerns requiring special consideration. All datasets and models used in our experiments are publicly available and are used in accordance with their intended research purposes.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, experimental settings, implementation details, and hyperparameters in the main paper. The source code will be made publicly available upon acceptance.

REFERENCES

Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11923–11932. IEEE, 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*, pp. 5008–5017, 2021.

Ziyao Guo, Haonan Yan, Hui Li, and Xiaodong Lin. Class attention transfer based knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11868–11877, 2023.

Zhiwei Hao, Jianyuan Guo, Kai Han, Yehui Tang, Han Hu, Yunhe Wang, and Chang Xu. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. *Advances in Neural Information Processing Systems*, 36:79570–79582, 2023.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Md Imtiaz Hossain, Sharmen Akhter, Choong Seon Hong, and Eui-Nam Huh. Single teacher, multiple perspectives: Teacher knowledge augmentation for enhanced knowledge distillation. In *International Conference on Learning Representations*, volume 2025, pp. 91583–91594, 2025.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *Advances in neural information processing systems*, 35:33716–33727, 2022.

Jungeun Kim, Junwon You, Dongjin Lee, Ha Young Kim, and Jae-Hun Jung. Do topological characteristics help in knowledge distillation? In *Forty-first International Conference on Machine Learning*, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 1504–1512, 2023.

Jhe-Hao Lin, Yi Yao, Chan-Feng Hsu, Hongxia Xie, Hong-Han Shuai, and Wen-Huang Cheng. Feature-based one-for-all: A universal framework for heterogeneous knowledge distillation. *arXiv preprint arXiv:2501.08885*, 2025.

Dongyang Liu, Meina Kan, Shiguang Shan, and Xilin Chen. Function-consistent feature distillation. *arXiv preprint arXiv:2304.11832*, 2023a.

Xiaolong Liu, Lujun Li, Chao Li, and Anbang Yao. Norm: Knowledge distillation via n-to-one representation matching. *arXiv preprint arXiv:2305.13803*, 2023b.

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*, pp. 3967–3976, 2019.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

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

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

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

Yingchao Wang, Wenqi Niu, Xingshan Yao, Li You, and Weilun Fei. Rethinking the dark knowledge and kullback-leibler divergence loss in knowledge distillation under capacity mismatching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26688–26696, 2026.

Shicai Wei, Chunbo Luo, and Yang Luo. Scale decoupled distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15975–15983. IEEE, 2024.

Liangyu Xiang, Junyu Gao, and Changsheng Xu. Evidential knowledge distillation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2814–2824. IEEE, 2025.

Liuchi Xu, Kang Liu, Jinshuai Liu, Lu Wang, Lisheng Xu, and Jun Cheng. Local dense logit relations for enhanced knowledge distillation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4539–4549. IEEE, 2025.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17139–17148. IEEE, 2023.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

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*, pp. 11953–11962, 2022.

Kaixiang Zheng and En-Hui Yang. Knowledge distillation based on transformed teacher matching. In *International Conference on Learning Representations*, volume 2024, pp. 44963–44979, 2024.