Relational Representation Distillation

Nikolaos Giakoumoglou[1], **Tania Stathaki**[1]

[1] Imperial College London

Knowledge distillation transfers knowledge from large, high-capacity teacher models to more compact student networks. The standard approach minimizes the Kullback–Leibler (KL) divergence between the probabilistic outputs of the teacher and student, effectively aligning predictions but neglecting the structural relationships encoded within the teacher's internal representations. Recent advances have adopted contrastive learning objectives to address this limitation; however, such instance-discrimination–based methods inevitably induce a *"class collision problem"*, in which semantically related samples are inappropriately pushed apart despite belonging to similar classes. To overcome this, we propose **R**elational **R**epresentation **D**istillation (RRD) that preserves the *relative relationships* among instances rather than enforcing absolute separation. Our method introduces separate temperature parameters for teacher and student distributions, with a *sharper teacher* (low τ_t) emphasizing primary relationships and a *softer student* (high τ_s) maintaining secondary similarities. This dual-temperature formulation creates an implicit information bottleneck that preserves fine-grained relational structure while avoiding the over-separation characteristic of contrastive losses. We establish theoretical connections showing that InfoNCE emerges as a limiting case of our objective when $\tau_t \to 0$, and empirically demonstrate that this relaxed formulation yields superior relational alignment and generalization across classification and detection tasks.

Date: July 16, 2024
Correspondence: Nikolaos Giakoumoglou <nikos@imperial.ac.uk>
Code: https://github.com/giakoumoglou/rrd

1 Introduction

Knowledge Distillation (KD) transfers knowledge from large, high-capacity *teacher* models to compact *student* models (Hinton et al., 2015). This approach is increasingly relevant as vision models for image classification (Liu et al., 2021; Dosovitskiy et al., 2021), object detection (Ren et al., 2016; Lin et al., 2017), and semantic segmentation (Chen et al., 2017b,c) continue to grow in size and computational cost (Goyal et al., 2019; Kornblith et al., 2019), motivating efficient model compression techniques (Buciluǎ et al., 2006; Polino et al., 2018).

The seminal work of Buciluǎ et al. (2006) and Hinton et al. (2015) introduced the idea of transferring knowledge by minimizing the *Kullback–Leibler* (KL) divergence between teacher and student output distributions. This formulation makes intuitive sense when the output is a categorical probability mass function over classes. However, in many cases, we wish to transfer richer internal knowledge — not just about class probabilities but about the underlying *representations* that encode visual semantics and inter-class relations.

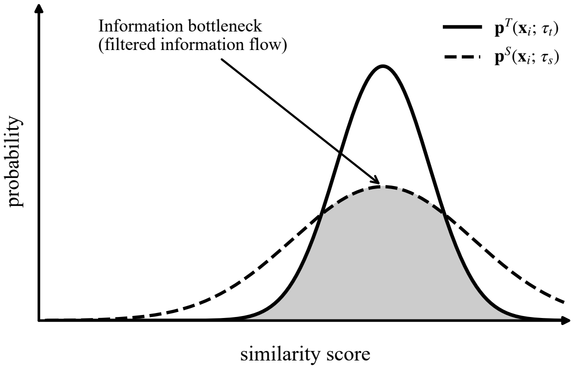


Figure 1 Visualization of the information bottleneck effect. The teacher produces a sharper similarity distribution $\mathbf{p}^T(\mathbf{x}_i;\tau_t)$ (solid black) highlighting primary relationships, while the student adopts a softer distribution $\mathbf{p}^S(\mathbf{x}_i;\tau_s)$ (dashed black) that retains secondary similarities. The gray-shaded overlap region illustrates the *filtered information flow*, where only essential relational cues are transferred from teacher to student, effectively bounding $I(\mathbf{z}^T;\mathbf{z}^S)$.

Representational knowledge is inherently *structured*: feature dimensions exhibit non-trivial correlations and higher-order dependencies. Logit matching alone cannot capture this relational structure. To address this, feature-based methods (Romero et al., 2014; Zagoruyko and Komodakis, 2016; Yim et al., 2017; Peng et al., 2019) extend distillation to intermediate representations. However, Tian et al. (2022) showed that such approaches still neglect the structural knowledge encoded in the teacher's internal representations. To overcome this limitation, Tian et al. (2022) further adapted the family of contrastive objectives (Gutmann and Hyvärinen, 2010; van den Oord et al., 2019; Arora et al., 2019; Hjelm et al., 2019) for distilling structured knowledge between teacher and student networks. These objectives have been highly successful in density estimation and self-supervised representation learning, as they implicitly maximize a lower bound on the mutual information between paired embeddings.

However, such instance-discrimination–based approaches inevitably introduce a *"class-collision problem"* (Arora et al., 2019; Li et al., 2021; Yeh et al., 2022), in which semantically similar samples are undesirably pushed apart due to uniform negative sampling. As shown in our experiments (Table 7), this repulsion degrades semantic coherence, leading to fragmented feature spaces where related instances lose proximity. This occurs because contrastive objectives enforce strict pairwise separation across instances—even when preserving relative similarity would be more semantically meaningful. In the context of distillation, such behavior weakens the alignment between teacher and student relational structures, preventing the student from learning coherent semantic geometry.

To overcome these limitations, we propose a method that relaxes rigid contrastive objectives by preserving meaningful *relative relationships* between instances in feature space, which we term **R**elational **R**epresentation **D**istillation (RRD). For example, given images of a *"cat"*, *"dog"*, and *"plane"*, what matters is not absolute similarity scores but their relative ordering: the cat should be closest to another cat, followed by the dog (another animal), and farthest from the plane. RRD achieves this by aligning the teacher and student similarity distributions through a KL-based loss, rather than enforcing one-hot positive matches. We introduce distinct temperature parameters for the teacher and student distributions: a **sharper teacher** (low τ_t) emphasizes primary relationships, while a **softer student** (high τ_s) retains secondary similarities. This dual-temperature mechanism naturally forms an *information bottleneck* (Figure 1), bounding the information transferred between teacher and student to only the most salient relational cues. As shown empirically, this formulation mitigates class collisions and yields superior structural alignment compared to prior methods such as CRD (Tian et al., 2022).

Our main **contributions** are as follows: **(i)** We propose an objective that preserves structural relationships between feature representations using distinct temperature parameters for the teacher and student, forming an implicit information bottleneck that balances sharp primary alignment with smooth secondary similarities **(ii)** We establish theoretical connections between our objective, InfoNCE (van den Oord et al., 2019), and the KL divergence, showing that InfoNCE arises as a limiting case when $\tau_t \to 0$. **(iii)** We empirically demonstrate the advantages of our objective across classification (Tables 1 to 3) and detection (Table 4) benchmarks, achieving consistent gains over existing methods. RRD achives 75.50% relative improvement (Tian et al., 2022) over conventional KD and 80.03% when combined with it (Tables 1 and 2). **(iv)** We provide quantitative and qualitative analyses of learned representations through correlation alignment (Figure 2 and table 6), semantic similarity evaluation (Table 7), and retrieval-based assessment (Figure 7, confirming that RRD maintains relational topology between teacher and student embeddings.

2 Related Work

The seminal works of Buciluă et al. (2006) and Hinton et al. (2015) established the foundation of knowledge distillation, where compact student networks learn from large teacher models with minimal loss in generalization. Buciluă et al. (2006) proposed matching model outputs, while Hinton et al. (2015) introduced temperature scaling in the softmax to soften class probabilities, revealing richer inter-class relations—often called "*dark knowledge*". Later extensions such as Li et al. (2014) refined this formulation for better transferability. Since then, knowledge distillation has evolved into several branches, most notably *logit-based* (Mirzadeh et al., 2019; Yang et al., 2019; Huang et al., 2022) and *feature-based* (Romero et al., 2014; Zagoruyko and Komodakis, 2016; Yim et al., 2017; Tian et al., 2022) distillation.

2.1 Logit-based Distillation

Logit-based approaches transfer knowledge by matching the *output logits* (pre-softmax scores) of the teacher and student, encouraging the student to mimic the teacher's predictive distribution and class-level semantics. Early methods improved stability and transfer via hierarchical supervision (Yang et al., 2019), multi-step training (Mirzadeh et al., 2019), or collaborative learning (Zhang et al., 2017). Subsequent work refined this process by adjusting how logits are represented or weighted (Zhou et al., 2021; Yuan et al., 2021; Niu et al., 2022; Sun et al., 2024). Further refinements involved dynamic temperature adjustment (Li et al., 2022), transformation-based alignment (Zheng and Yang, 2024), and adaptive teacher calibration (Huang et al., 2022).

2.2 Feature-based Distillation

Feature-based methods transfer richer structural knowledge by aligning intermediate teacher and student representations to capture spatial or semantic relationships. Foundational work explored hint-based supervision (Romero et al., 2014), attention transfer (Zagoruyko and Komodakis, 2016), and feature transformation (Kim et al., 2018), later extended by functional consistency (Liu et al., 2023a), class-level attention (Guo et al., 2023), and structural normalization (Chen et al., 2021c, 2022; Liu et al., 2023b). A major branch focuses on preserving *relational structures* among embeddings, ensuring pairwise or higher-order relationships remain consistent between teacher and student. Early works modeled such relations through inner products (Yim et al., 2017), distance preservation (Park et al., 2019), or correlation congruence (Peng et al., 2019), while contrastive methods (Tian et al., 2022) maximized mutual information via memory banks. Recent studies introduced redundancy reduction and kernel-based alignment (Miles et al., 2024; He and Ozay, 2022). Our objective fits within this broader class of *feature-relational distillation* methods but relaxes hard contrastive constraints, focusing instead on smooth relational distributions that preserve semantic coherence.

2.3 Connection to Information-theoretic Objectives

Our objective is also related to *InfoNCE* (van den Oord et al., 2019) and *Noise-Contrastive Estimation (NCE)* (Gutmann and Hyvärinen, 2010), which maximize a lower bound on mutual information between representations (Hjelm et al., 2019). While adversarial frameworks (Goodfellow et al., 2014) similarly involve divergence minimization, they differ in intent—NCE optimizes a likelihood-based estimation objective rather than adversarial discrimination. In this context, our dual-temperature formulation can be viewed as an information bottleneck that bounds the mutual information $I(\mathbf{z}^T; \mathbf{z}^S)$, ensuring that only essential relational cues are transferred during distillation (Figure 1).

3 Methodology

Here, we introduce our objective which transfers knowledge from a pre-trained teacher network to a student network by leveraging relational cues embedded in their feature representations. Section 3.1 outlines the fundamental principles of knowledge distillation, Section 3.2 details the formulation of our relational objective, and Section 3.3 provides an analytical interpretation of the *information bottleneck* that regulates the flow of relational knowledge.

3.1 Preliminaries on Knowledge Distillation

Knowledge distillation transfers knowledge from a high-capacity teacher network f_θ^T to a compact student network f_θ^S (Hinton et al., 2015; Buciluă et al., 2006). Its primary objective is to enable the student model to approximate the performance of the teacher model while leveraging the student's computational efficiency. The overall distillation process can be formulated as:

$$\hat{\theta}_S = \arg\min_{\theta_S} \sum_{i}^{N} \left(\mathcal{L}_{\text{sup}}(\mathbf{x}_i, \theta_S, y_i) + \mathcal{L}_{\text{distill}}(\mathbf{x}_i, \theta_S, \theta_T)\right), \tag{1}$$

where $\mathbf{x}_i$ is an image, y_i is the corresponding label, θ_S is the parameter set for the student network, and θ_T is the set for the teacher network. The loss $\mathcal{L}_{\text{sup}}$ is the alignment error between the network prediction and the annotation. For example in image classification task (Mishra and Marr, 2017; Shen and Savvides, 2020; Polino et al., 2018; Cho and Hariharan, 2019), it is normally a cross entropy loss. For object detection (Liu et al., 2018; Chen et al., 2017a), it includes bounding box regression as well. The distillation loss $\mathcal{L}_{\text{distill}}$ quantifies how well the student mimics the pre-trained teacher, commonly implemented using KL divergence between softmax outputs (Hinton et al., 2015) or ℓ_2 distance between feature maps (Romero et al., 2014). While this approach demonstrates effectiveness with labeled data, its performance in unsupervised settings remains an open research question.

3.2 Relational Representation Distillation

Given an input image $\mathbf{x}_i$, it is first mapped into features $\mathbf{z}_i^T = f_\theta^T(\mathbf{x}_i)$ and $\mathbf{z}_i^S = f_\theta^S(\mathbf{x}_i)$, where $\mathbf{z}_i^T, \mathbf{z}_i^S \in \mathbb{R}^d$ and f_θ^T, f_θ^S denote the teacher and student networks, respectively. All features are ℓ_2-normalized, *i.e.*, $\mathbf{z}_i^T \leftarrow \frac{\mathbf{z}_i^T}{||\mathbf{z}_i^T||}$ and $\mathbf{z}_i^S \leftarrow \frac{\mathbf{z}_i^S}{||\mathbf{z}_i^S||}$, ensuring they lie on a unit hypersphere.

Let $\mathcal{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_K]$ denote a memory bank where K is the memory length and $\mathbf{m}_k \in \mathbb{R}^d$ is a feature vector. The memory $\mathcal{M}$ stores previous teacher features and is updated following a *first-in-first-out* (FIFO) strategy: we add the teacher's features from the current batch while removing the oldest stored features per iteration (*cf.* Section D.1). This buffer is critical for computing stable similarity distributions that capture relational structures between feature representations. Without stored references, similarity estimation would be limited to the current batch, restricting relational learning. While minimizing cross-entropy between student and teacher similarity distributions using $\mathcal{M}$ allows soft contrasting against random samples, direct teacher alignment is not enforced. To address this, we extend the memory bank to $\mathcal{M}^+ = [\mathbf{m}_1, \ldots, \mathbf{m}_K, \mathbf{m}_{K+1}]$ by appending the teacher embedding $\mathbf{z}_i^T$ as $\mathbf{m}_{K+1}$. This ensures that the teacher's most recent representation is explicitly considered when computing similarity scores (*cf.* Section C).

We define $\mathbf{p}^T(\mathbf{x}_i; \theta_T; \mathcal{M}^+)$ as the teacher similarity scores between the extracted teacher feature $\mathbf{z}_i^T$ and existing memory features $\mathbf{m}_j$ (for $j = 1$ to $K+1$), represented as:

$$\mathbf{p}^T(\mathbf{x}_i; \theta_T, \mathcal{M}^+) = \left[p_1^T, \ldots, p_{K+1}^T\right] \quad \text{where} \quad p_i^T = \frac{\exp(\mathbf{z}_i^T \cdot \mathbf{m}_j/\tau_t)}{\sum\limits_{m \sim \mathcal{M}^+} \exp(\mathbf{z}_i^T \cdot \mathbf{m}/\tau_t)}, \tag{2}$$

and $(\cdot)$ denotes the *inner product*, and τ_t is a temperature parameter for the teacher. Similarly, we define $\mathbf{p}^S(\mathbf{x}_i; \theta_S, \mathcal{M}^+)$ as the student similarity scores between the extracted student feature $\mathbf{z}_i^S$ and existing memory features $\mathbf{m}_j$, represented as:

$$\mathbf{p}^S(\mathbf{x}_i; \theta_S, \mathcal{M}^+) = \left[p_1^S, \ldots, p_{K+1}^S\right] \quad \text{where} \quad p_i^S = \frac{\exp(\mathbf{z}_i^S \cdot \mathbf{m}_j/\tau_s)}{\sum\limits_{m \sim \mathcal{M}^+} \exp(\mathbf{z}_i^S \cdot \mathbf{m}/\tau_s)} \tag{3}$$

and τ_s is a temperature parameter for the student.

Our distillation objective can be formulated as minimizing the KL divergence between the similarity scores of the teacher, $\mathbf{p}^T(\mathbf{x}_i; \theta_T, \mathcal{M}^+)$ and the student, $\mathbf{p}^S(\mathbf{x}_i; \theta_S, \mathcal{M}^+)$, over all the instances $\mathbf{x}_i$:

$$\begin{aligned} \hat{\theta}_S &= \arg\min_{\theta_S} \sum_i^N D_{\text{KL}}(\mathbf{p}^T(\mathbf{x}_i; \theta_T, \mathcal{M}^+) \parallel \mathbf{p}^S(\mathbf{x}_i; \theta_S, \mathcal{M}^+)) \\ &= \arg\min_{\theta_S} \sum_i^N H(\mathbf{p}^T(\mathbf{x}_i; \theta_T, \mathcal{M}^+), \mathbf{p}^S(\mathbf{x}_i; \theta_S, \mathcal{M}^+)) + H(\mathbf{p}^T(\mathbf{x}_i; \theta_T, \mathcal{M}^+)), \end{aligned} \tag{4}$$

where D_{KL} denotes the KL divergence between teacher $\mathbf{p}^T(\mathbf{x}_i; \theta_T, \mathcal{M}^+)$ and student $\mathbf{p}^S(\mathbf{x}_i; \theta_S, \mathcal{M}^+)$ distributions, $H(\mathbf{p}^T(\mathbf{x}_i; \theta_T, \mathcal{M}^+), \mathbf{p}^S(\mathbf{x}_i; \theta_S, \mathcal{M}^+))$ represents the cross-entropy between the teacher's and student's similarity distributions, while $H(\mathbf{p}^T(\mathbf{x}_i;\theta_T,\mathcal{M}^+))$ is the entropy of the teacher's similarity distribution. Since $\mathbf{p}^S(\mathbf{x}_i;\theta_S,\mathcal{M}^+)$ will be used as a target, the gradient is clipped here, thus we only minimize the cross-entropy term $H(\mathbf{p}^T(\mathbf{x}_i;\theta_T,\mathcal{M}^+),\mathbf{p}^S(\mathbf{x}_i;\theta_S,\mathcal{M}^+))$:

$$\hat{\theta}_S = \arg\min_{\theta_S} \sum_{i=1}^{N} \mathcal{L}_{RRD}(\mathbf{x}_i,\theta_S,\theta_T,\mathcal{M}) \tag{5}$$

where

$$\mathcal{L}_{RRD} = -\mathbf{p}^T(\mathbf{x}_i;\theta_T,\mathcal{M}^+)\cdot\log\mathbf{p}^S(\mathbf{x}_i;\theta_S,\mathcal{M}^+) = -\sum_{j=1}^{K+1}\frac{\exp(\mathbf{z}_i^T\cdot\mathbf{m}_j/\tau_t)}{\sum\limits_{k=1}^{K+1}\exp(\mathbf{z}_i^T\cdot\mathbf{m}_k/\tau_t)}\log\frac{\exp(\mathbf{z}_i^S\cdot\mathbf{m}_j/\tau_s)}{\sum\limits_{k=1}^{K+1}\exp(\mathbf{z}_i^S\cdot\mathbf{m}_k/\tau_s)}. \tag{6}$$

Since we keep the teacher network frozen during training, teacher similarity scores p_j^T directly influence corresponding student weights p_j^S. The ℓ_2 normalization ensures similarity between $\mathbf{z}_i^T$ and $\mathbf{m}_{K+1}$ equals $\mathbb{1}$ pre-softmax, making it dominate other p_j^T values. This maximum weight for p_{K+1}^S can be controlled via temperature τ_t. The optimization aligns student features $\mathbf{z}_i^S$ with teacher features while maintaining contrast against memory features.

Note here that appending the current teacher embedding as $\mathbf{m}_{K+1}$ ensures a clear peak in $\mathbf{p}^T$, preventing the KL divergence from degenerating into weak contrastive alignment. As shown in our ablations (Section 4.4, Figure 3b), performance improves with larger K until plateauing around $K = 16384$. Storing 128-d features requires only ∼600MB on ImageNet, allowing the bank to reside on GPU.

3.2.1 Relation to InfoNCE Loss

As $\tau_t \to 0$, the teacher's softmax distribution $\mathbf{p}^T$ becomes a one-hot vector with $p_{K+1}^T = 1$ and zeros elsewhere. This reduces our objective to:

$$\mathcal{L}_{NCE} = \sum_{i}^{N} -\log\frac{\exp(\mathbf{z}_i^T\cdot\mathbf{z}_i^S/\tau_s)}{\sum\limits_{\mathbf{m}\sim\mathcal{M}^+}\exp(\mathbf{z}_i^S\cdot\mathbf{m}/\tau_s)}, \tag{7}$$

which matches the *InfoNCE* loss (van den Oord et al., 2019). This implements instance discrimination through $(K+1)$-way classification, separating different instances while enforcing identical representations for matching pairs. We also provide a theoretical analysis in the Section C.

3.2.2 Relation to Kullback-Leibler Divergence

Hinton et al. (2015) defined the knowledge distillation loss via the *Kullback–Leibler* (KL) divergence between the softened output distributions of the teacher and student networks:

$$\mathcal{L}_{\mathrm{KL}} = \sum_{i=1}^{N}\tau^2 D_{\mathrm{KL}}\Big(\sigma(y_i^T/\tau)\parallel\sigma(y_i^S/\tau)\Big) = \sum_{i=1}^{N}\tau^2\sum_{c=1}^{C}\sigma\left(\frac{y_{i,c}^T}{\tau}\right)\log\frac{\sigma\left(\frac{y_{i,c}^T}{\tau}\right)}{\sigma\left(\frac{y_{i,c}^S}{\tau}\right)} \tag{8}$$

where $\sigma(x)$ denotes the softmax function, and y_i^T, y_i^S represent the logits of the teacher and student networks, respectively, with $y_{i,c}^S$ and $y_{i,c}^T$ referring to their logit values for class c, before applying the softmax function. Both losses use KL divergence to align the teacher and student distributions. However, in Hinton's formulation the softmax is computed over C class logits (representing class predictions), while in our objective it is computed over $(K+1)$ memory bank entries (representing similarity scores between features and memory bank entries).

3.2.3 Full Objective

Consistent with prior work in knowledge distillation, we formulate a full training objective that integrates supervised learning, standard KL divergence–based distillation, and our proposed loss. For clarity, we denote our method as "*RRD*" when using only supervised learning and our relational loss, and "*RRD+KD*" when combining all three loss components (*i.e.*, $\lambda > 0$). The full objective is given by:

$$\hat{\theta}_S = \arg\min_{\theta_S} \sum_{i}^{N} \Big(\mathcal{L}_{\text{sup}}(\mathbf{x}_i, \theta_S, y_i) + \lambda \cdot \mathcal{L}_{\text{KL}}(\mathbf{x}_i, \theta_S, \theta_T) + \beta \cdot \mathcal{L}_{RRD}(\mathbf{x}_i, \theta_S, \theta_T, \mathcal{M}) \Big) \tag{9}$$

where λ and β balance the KL divergence and our proposed loss, respectively. The combination of losses provides complementary supervision: KD's soft targets provide direct class-level supervision through logit-space KL divergence, while our method ensures feature-space consistency. Unlike (Tian et al., 2022) which enforces strict instance-level discrimination that can push semantically similar samples too far apart, our method relaxes this constraint by focusing on preserving relative relationships in the feature space, allowing the student to maintain more nuanced similarities between instances while still learning discriminative representations.

3.3 Information Bottleneck

The dual-temperature formulation in our objective establishes an implicit *information bottleneck* between the teacher and student similarity distributions (*cf.* Figure 1). We set $\tau_s > \tau_t$, where τ_t and τ_s denote the temperature parameters for the teacher and student, respectively. A smaller τ_t produces a *sharper* teacher distribution $\mathbf{p}^T(\mathbf{x}_i; \theta_T, \mathcal{M}^+)$ that emphasizes dominant relational cues, while a larger τ_s yields a *softer* student distribution $\mathbf{p}^S(\mathbf{x}_i; \theta_S, \mathcal{M}^+)$ that retains uncertainty over secondary similarities. This design enforces a selective transfer mechanism—only high-confidence, structurally salient relations from the teacher produce strong gradients in the student, while weaker or noisy relations are attenuated. Consequently, the student acts as a controlled filter, transmitting only the most informative relational cues from the teacher's representation space. This mechanism, which we refer to as the *filtered information flow*, constrains the amount of information shared between teacher and student representations, forming a principled *information bottleneck*. The effect of this temperature asymmetry and its empirical validation are analyzed in our ablation studies (*cf.* Section 4.4, Figure 3a, Section D).

Formally, the mutual information between the teacher and student representations is bounded by the entropy difference between their respective similarity distributions:

$$I(\mathbf{z}^T; \mathbf{z}^S) \leq H(\mathbf{p}^S(\mathbf{x}_i; \theta_S, \mathcal{M}^+)) - H(\mathbf{p}^T(\mathbf{x}_i; \theta_T, \mathcal{M}^+)), \tag{10}$$

where $H(\mathbf{p}^S(\mathbf{x}_i; \theta_S, \mathcal{M}^+)) > H(\mathbf{p}^T(\mathbf{x}_i; \theta_T, \mathcal{M}^+))$ due to the softer student distribution. This entropy gap defines the effective information capacity of the distillation process—the larger the gap, the stronger the bottleneck—ensuring that our objective transmits only essential relational structure.

Intuitively, this bottleneck enforces a *coarse-to-fine* transfer of representational knowledge: the teacher conveys sharp, high-fidelity relational signals, while the student absorbs them through smoother similarity distributions that generalize across related instances. This selective filtering prevents overfitting to instance-level details and encourages the student to preserve the underlying topology of the teacher's relational representation. Empirically, this *filtered information flow* yields improved structural alignment (Table 6) and enhanced semantic organization (Table 7) compared to both conventional and contrastive distillation objectives.

4 Experiments

We conduct extensive experiments across multiple benchmarks to evaluate the effectiveness and generality of our proposed objective. Section 4.1 outlines the experimental setup, Section 4.2 presents quantitative results across benchmarks, and Section 4.3 analyzes the learned representations and structural properties. Further ablations are discussed in Section 4.4 and Section D.

Table 1 Distillation results on CIFAR-100 (same architecture). Test top-1 accuracy (%) for different teacher–student pairs. The best and second-best results are **bolded** and underlined. Results for baselines are from Tian et al. (2022); ours are averaged over five runs. **Architecture abbreviations**: W: WideResNet, rn: resnet.

Teacher Student	WRN-40-2 WRN-16-2	WRN-40-2 WRN-40-1	rn-56 rn-20	rn-110 rn-20	rn-110 rn-32	rn-32x4 rn-8x4	VGG-13 VGG-8
Teacher	75.61	75.61	72.34	74.31	74.31	79.42	74.64
Student	73.26	71.98	69.06	69.06	71.14	72.50	70.36
KD (Hinton et al., 2015)	74.92	73.54	70.66	70.67	73.08	73.33	72.98
FitNet (Romero et al., 2014)	73.58	72.24	69.21	68.99	71.06	73.50	71.02
AT (Zagoruyko and Komodakis, 2016)	74.08	72.77	70.55	70.22	72.31	73.44	71.43
SP (Tung and Mori, 2019)	73.83	72.43	69.67	70.04	72.69	72.94	72.68
CC (Peng et al., 2019)	73.56	72.21	69.63	69.48	71.48	72.97	70.81
RKD (Park et al., 2019)	73.35	72.22	69.61	69.25	71.82	71.90	71.48
FSP (Yim et al., 2017)	72.91	n/a	69.95	70.11	71.89	72.62	70.33
OFD (Heo et al., 2019a)	75.24	74.33	70.38	n/a	73.23	74.95	73.95
CRD (Tian et al., 2022)	75.48	74.14	71.16	71.46	73.48	75.51	73.94
CRD+KD (Tian et al., 2022)	75.64	74.38	71.63	71.56	73.75	75.46	74.29
RRD (ours)	**75.85**	74.61	71.89	**71.92**	73.73	**75.77**	74.01
RRD+KD (ours)	75.67	**74.68**	**72.03**	71.75	**73.96**	75.53	**74.37**

4.1 Experimental Setup

We evaluate the proposed framework on both image classification and object detection tasks using five standard benchmarks: CIFAR-100 (Krizhevsky, 2009), ImageNet ILSVRC-2012 (Deng et al., 2009), STL-10 (Coates and Ng, 2011), Tiny ImageNet (Deng et al., 2009), and MS-COCO (Lin et al., 2015). Following prior work (Tian et al., 2022), we evaluate thirteen teacher–student architecture pairs with varying capacity gaps to assess generalization across different model families. For classification, we adopt the implementation protocol of (Tian et al., 2022), while for detection we follow (Zhao et al., 2022; Chen et al., 2021c). To ensure dimensional consistency and preserve relational information during feature alignment, both teacher and student features are passed through *projection heads*—two-layer MLPs (512 hidden, 128 output)—that nonlinearly project intermediate representations into a shared embedding space for computing relational similarity distributions, improving KL alignment stability and structural coherence as observed in prior work (Tian et al., 2022; Miles and Mikolajczyk, 2024). The memory bank size is fixed at $K = 16{,}384$. The temperature parameters are set to $\tau_t = 0.02$ for the teacher and $\tau_s = 0.1$ for the student, reflecting the asymmetric softening central to our *information bottleneck* formulation (Figure 1 validated in Figure 3a). The distillation weight λ is fixed at 0.9 for CIFAR-100 and 1.0 for ImageNet, while β is set to 1.5 and 1.0 respectively. When combined with standard KD, the temperature for the logit-space KL divergence is $\tau = 4$. Additional implementation details are provided in Section B.

4.2 Main Results

We first evaluate RRD on image classification (CIFAR-100, ImageNet) and object detection (MS-COCO). Across tasks, our objective consistently improves over baseline distillation methods and exhibits strong complementarity when combined with Hinton's KD (Hinton et al., 2015), validating its ability to capture transferable relational structure. Extended comparison is provided in Section E.

4.2.1 Results on CIFAR-100

Tables 1 and 2 report top-1 accuracy for same-architecture and cross-architecture (*i.e.* different) teacher–student pairs, respectively. RRD consistently outperforms classical KD (Hinton et al., 2015) and contrastive-based CRD (Tian et al., 2022), demonstrating that preserving relative similarity distributions offers stronger guidance than strict instance discrimination. When combined with Hinton's KD (Hinton et al., 2015), it achieves further gains—KD provides class-level supervision through logit-space KL divergence, while RRD enforces relational consistency in the feature space.

Table 2 Distillation results on CIFAR-100 (different architecture). Test top-1 accuracy (%) for different teacher–student pairs. The best and second-best results are **bolded** and underlined. Results for baselines are from Tian et al. (2022); ours are averaged over five runs. **Architecture abbreviations**: W: WideResNet, RN: ResNet, MN: MobileNet, SN: ShuffleNet.

Teacher Student	VGG-13 MN-v2	RN-50 MN-v2	RN-50 VGG-8	RN-32x4 SN-v1	RN-32x4 SN-v2	WRN-40-2 SN-v1
Teacher	74.64	79.34	79.34	79.42	79.42	75.61
Student	64.60	64.60	70.36	70.50	71.82	70.50
KD (Hinton et al., 2015)	67.37	67.35	73.81	74.07	74.45	74.83
FitNet (Romero et al., 2014)	64.14	63.16	70.69	73.59	73.54	73.73
AT (Zagoruyko and Komodakis, 2016)	59.40	58.58	71.84	71.73	72.73	73.32
SP (Tung and Mori, 2019)	66.30	68.08	73.34	73.48	74.56	74.52
CC (Peng et al., 2019)	64.86	65.43	70.25	71.14	71.29	71.38
RKD (Park et al., 2019)	64.52	64.43	71.50	72.28	73.21	72.21
FSP (Yim et al., 2017)	58.16	64.96	71.28	74.12	74.68	76.09
OFD (Heo et al., 2019a)	69.48	69.04	n/a	75.98	76.82	75.85
CRD (Tian et al., 2022)	69.73	69.11	74.30	75.11	75.65	76.05
CRD+KD (Tian et al., 2022)	69.94	69.54	**74.97**	75.12	76.05	76.27
RRD (ours)	69.61	**70.11**	74.30	75.60	76.31	75.98
RRD+KD (ours)	**69.99**	69.65	74.53	**76.68**	**76.87**	**76.64**

Table 3 Distillation results on ImageNet. Student top-1 accuracy (%) on the ImageNet validation set under various teacher–student pairs. Results are from Tian et al. (2022); ours are based on a single run.

Teacher Student	RN-34 RN-18	RN-50 RN-18	RN-50 MN-v2
Teacher	73.31	76.16	76.16
Student	69.75	69.75	69.63
KD (Hinton et al., 2015)	70.67	71.29	70.49
AT (Zagoruyko and Komodakis, 2016)	71.03	71.18	70.18
SP (Tung and Mori, 2019)	70.62	71.08	n/a
CC (Peng et al., 2019)	69.96	n/a	n/a
RKD (Park et al., 2019)	70.40	n/a	68.50
CRD (Tian et al., 2022)	71.17	71.25	69.07
RRD (ours)	**72.03**	**71.97**	71.54
RRD+KD (ours)	71.99	71.88	**71.56**

4.2.2 Results on ImageNet

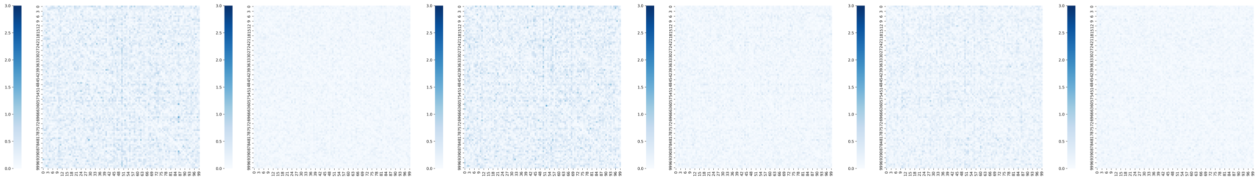
Table 3 shows that RRD scales effectively to large-scale benchmarks such as ImageNet, maintaining strong and consistent performance across diverse architectures. Its advantage remains evident for both same- and cross-architecture teacher–student pairs. RRD surpasses KD (Hinton et al., 2015), CRD (Tian et al., 2022), and their combination across all evaluated settings, even when applied independently, underscoring its robustness and broad applicability to real-world distillation scenarios.

4.2.3 Results on COCO

Table 4 extends our analysis to object detection, evaluating Faster R-CNN models trained with different distillation methods on MS-COCO (Lin et al., 2015). By transferring relational cues from teacher feature maps, RRD improves detection accuracy across teacher–student pairs while maintaining stability during fine-tuning and optimization. Its performance remains competitive with or superior to advanced methods such as ReviewKD (Chen et al., 2021c) and DKD (Zhao et al., 2022), underscoring that relational structure consistently benefits spatially localized prediction tasks.

Table 4 Object detection distillation results on COCO. Detection performance (AP, AP_{50}, AP_{75}) of student detectors trained with different distillation methods using Faster R-CNN on COCO `val2017`. Results are from Zhao et al. (2022); ours are based on a single run. **Architecture abbreviations**: RN: ResNet, MN: MobileNet.

Method	RN-101 → RN-18			RN-101 → RN-50			RN-50 → MN-v2		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Teacher	42.04	62.48	45.88	42.04	62.48	45.88	40.22	61.02	43.81
Student	33.26	53.61	35.26	37.93	58.84	41.05	29.47	48.87	30.90
KD (Hinton et al., 2015)	33.97	54.66	36.62	38.35	59.41	41.71	30.13	50.28	31.35
FitNet (Romero et al., 2014)	34.13	54.16	36.71	38.76	59.62	41.80	30.20	49.80	31.69
ReviewKD (Chen et al., 2021c)	36.75	56.72	34.00	**40.36**	60.97	**44.08**	33.71	53.15	36.13
DKD (Zhao et al., 2022)	35.05	56.60	37.54	39.25	60.90	42.73	32.34	53.77	34.01
RRD (ours)	**36.85**	**57.10**	**39.20**	40.15	**61.00**	43.90	**33.90**	**54.20**	**36.00**



(a) Vanilla
Mean: 0.24
Max: 1.66

(b) KD
Mean: 0.09
Max: 0.49

(c) CRD
Mean: 0.23
Max: 1.56

(d) CRD+KD
Mean: 0.10
Max: 0.57

(e) RRD
Mean: 0.18
Max: 0.99

(f) RRD+KD
Mean: 0.07
Max: 0.55

Figure 2 Correlation alignment on CIFAR-100. Correlation matrix comparison between teacher (WRN-40-2) and student (WRN-40-1) logits. Lower values indicate stronger alignment of inter-class relations.

Table 5 Transfer learning performance results with distilled students. Top-1 accuracy (%) of a WRN-16-2 student distilled from WRN-40-2, evaluated on STL-10 and Tiny ImageNet. Results for baselines are from Tian et al. (2022); ours are averaged over five runs. **Abbreviations**: C: CIFAR, TIN: Tiny ImageNet.

	Teacher	*Student*	KD	AT	FitNet	CRD	CRD+KD	RRD	RRD+KD
C-100→STL-10	68.6	69.7	70.9	70.7	70.3	71.6	**72.2**	72.0	72.0
C-100→TIN	31.5	33.7	33.9	34.2	33.5	**35.6**	35.5	35.5	35.2

4.3 Representation Analysis

We next analyze the learned feature representations to assess structural and semantic fidelity. This includes examining transferability, inter-class correlation alignment, and semantic organization. Extended comparison in provided in Section E. Additional visualizations (t-SNE, nearest-neighbor retrieval) are provided in Section F.

4.3.1 Visualization of Inter-class Correlations

Figure 2 compares the correlation matrix differences between teacher and student logits. Our objective achieves better alignment of correlation structures compared to models trained without distillation or with alternative methods (Hinton et al., 2015; Tian et al., 2022). Standalone, it outperforms CRD (Tian et al., 2022), demonstrating stronger structural preservation. When combined with Hinton's KD (Hinton et al., 2015), it further improves alignment. This structural coherence indicates that RRD helps students internalize the teacher's feature geometry rather than memorizing isolated logits.

4.3.2 Transferability of Representations

Table 5 evaluates a WRN-16-2 student distilled from a WRN-40-2 teacher as a frozen feature extractor. RRD achieves strong and consistent transfer performance on unseen datasets, indicating that its representations

Table 6 Correlation matrix alignment metrics on CIFAR-100. Quantitative comparison of relational alignment between teacher (WRN-40-2) and student (WRN-40-1) embeddings using Frobenius distance, Pearson correlation, and SSIM.

Method	Frobenius ↓	Pearson ↑	SSIM ↑
Vanilla	10.491	0.951	0.935
KD (Hinton et al., 2015)	2.868	0.994	0.992
CRD (Tian et al., 2022)	8.862	0.946	0.938
CRD+KD (Tian et al., 2022)	3.265	0.992	0.989
RRD (ours)	6.969	0.970	0.964
RRD+KD (ours)	**2.720**	**0.995**	**0.993**

Table 7 Semantic similarity preservation on CIFAR-100. Evaluation of intra-/inter-class structure, NMI, and retrieval mAP@5 of student embeddings under differet distillation methods.

Method	Intra ↓	Inter ↑	NMI ↑	mAP@5 ↑
Vanilla	0.945	0.866	0.551	81.0
KD (Hinton et al., 2015)	0.914	1.012	0.546	85.4
CRD (Tian et al., 2022)	0.910	0.915	0.509	84.6
CRD+KD (Tian et al., 2022)	0.973	0.926	0.463	85.2
RRD (ours)	0.928	0.909	0.544	83.9
RRD+KD (ours)	**0.902**	**1.023**	**0.562**	**85.9**

effectively encode relational semantics rather than overfitting to task-specific decision boundaries or local class structures.

4.3.3 Correlation Matrix Alignment

Table 6 quantifies how well student models preserve the relational structure of their teachers on CIFAR-100, using Frobenius distance (Singhal, 2001), Pearson correlation (Benesty et al., 2009), and structural similarity index (SSIM) (Wang et al., 2004). RRD maintains strong structural alignment, and when combined with KD (Hinton et al., 2015), it achieves the best overall correspondence. Although KD slightly outperforms standalone RRD on some metrics due to its direct logit matching, combining with Hinton's KD objective leverages global relational cues from RRD and local soft-target supervision from KD, resulting in the most faithful teacher–student alignment.

4.3.4 Semantic structure preservation.

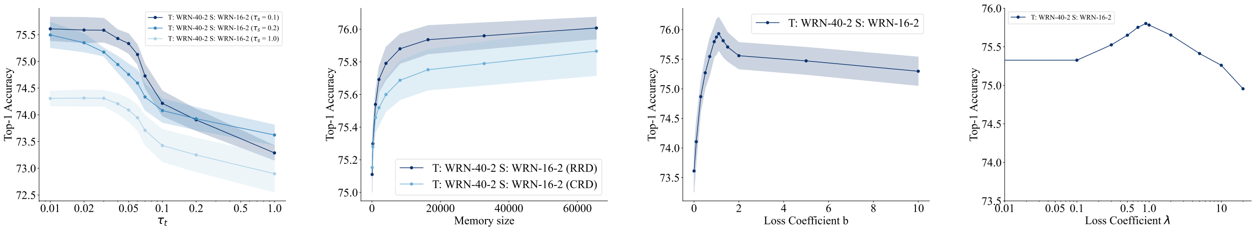
Table 7 evaluates the semantic organization of learned features on CIFAR-100 through intra-class compactness, inter-class separation, normalized mutual information (NMI) (Vinh et al., 2010), and retrieval precision measured by mean average precision at 5 (mAP@5) (Zhao et al., 2022). RRD improves semantic clustering compared to CRD (Tian et al., 2022), confirming that the *information bottleneck*—introduced by asymmetric temperatures—filters noise while retaining essential relational structure. This leads to more coherent embedding spaces where similar classes remain proximally organized.

4.4 Ablation Study

We provide an ablation study on CIFAR-100 using WRN-40-2 (teacher) and WRN-16-2 (student). Each experiment is repeated three times to ensure consistency; full ablations and additional analyses are provided in Section D.

4.4.1 Temperature Parameters

To verify the effectiveness of τ_s and τ_t, we fixed $\tau_s \in \{0.1, 0.2, 1.0\}$ and varied τ_t from 0.01 to 1.0. Figure 3a shows optimal performance at $\tau_s = 0.1$, $\tau_t = 0.02$. Performance degrades as τ_t increases, dropping sharply when



(a) Temperatures τ_s, τ_t

(b) Memory size K

(c) Loss coefficient β

(d) Loss coefficient λ

Figure 3 **Ablation study results.** We analyze the influence of individual design choices of RRD, including temperature parameters (τ_t, τ_s), memory size (K), and loss coefficients (β, λ) to analyze their impact on performance.

$\tau_t > \tau_s$, indicating softer teacher distributions harm distillation, consistent with our information bottleneck interpretation. Excessively soft student distributions $(\tau_s = 1.0)$ consistently underperform. As $\tau_t \to 0$ corresponds to the *argmax* operation producing a one-hot target, while higher τ_t yields softer distributions weakening teacher-student alignment.

4.4.2 Memory Size

We tested memory sizes from $K = 64$ (batch size limitation, *i.e.*, no memory bank) to $65,536$ (Figure 3b) and observed that performance improves for both RRD and CRD (*repr.*) methods, plateauing around $K = 16384$ with minimal gains thereafter. Using separate memory banks for teacher and student would fail because the KL divergence would align distributions over different feature sets without direct feature matching. Storing student features in the memory bank would create unstable training targets as student representations continuously evolve during training, undermining consistent supervision.

4.4.3 Loss Weighting

We investigated the impact of loss coefficient β by varying it from 0 to 10. As shown in Figure 3c, values of β between 0.5 and 1.5 work best, similar to (Tian et al., 2022) findings. Finally, although we typically fix $\lambda = 1$, Figure 3d illustrates the effect of varying it from 0 to 100.

5 Conclusion

We introduced a distillation framework that transfers knowledge by aligning relational similarity structures between teacher and student through a controlled *information bottleneck*. Experiments across image classification, transfer learning, and object detection demonstrate that RRD consistently outperforms conventional and contrastive distillation approaches across diverse datasets and architectures. By emphasizing relative relationships rather than absolute feature matching, RRD preserves the structural integrity of learned representations, yielding robust, generalizable models that capture the teacher's relational geometry while delivering superior downstream performance and representational consistency across tasks and scales.

Acknowledgements

We acknowledge the computational resources and support provided by the Imperial College Research Computing Service (http://doi.org/10.14469/hpc/2232), which enabled our experiments.

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 Conference on Computer Vision and Pattern Recognition*, pages 9163–9171, 2019.

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning, 2019.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006.

Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration, 2021a. https://arxiv.org/abs/2012.03236.

Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier, 2022. https://arxiv.org/abs/2203.14001.

Gongfan Chen, Yuting Wang, Jiajun Xu, Zhe Du, Qionghai Dai, Shiyang Geng, and Tao Mei. Learning efficient object detection models with knowledge distillation. In *Advances in Neural Information Processing Systems*, pages 742–751, 2017a.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, 2017b.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017c.

Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein contrastive representation distillation, 2021b. https://arxiv.org/abs/2012.08674.

Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review, 2021c. https://arxiv.org/abs/2104.09044.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4794–4802, 2019.

Adam Coates and Andrew Y. Ng. The importance of encoding versus training with sparse coding and vector quantization. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 921–928, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. https://api.semanticscholar.org/CorpusID:57246310.

Xueqing Deng, Dawei Sun, Shawn Newsam, and Peng Wang. Distpro: Searching a fast knowledge distillation process via meta optimization, 2022. https://arxiv.org/abs/2204.05547.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. https://arxiv.org/abs/1406.2661.

Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning, 2019.

Ziyao Guo, Haonan Yan, Hui Li, and Xiaodong Lin. Class attention transfer based knowledge distillation, 2023. https://arxiv.org/abs/2304.12777.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

Bobby He and Mete Ozay. Feature kernel distillation. In *International Conference on Learning Representations*, 2022. https://openreview.net/forum?id=tBIQEvApZK5.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation, 2019a. https://arxiv.org/abs/1904.01866.

Byeongho Heo, Minsik Lee, Seong Joon Yun, Jin Young Choi, and In So Kweon. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3779–3787, 2019b.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization, 2019.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher, 2022. https://arxiv.org/abs/2205.10536.

Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. In *Advances in Neural Information Processing Systems*, pages 185–195, 2017.

Jangho Kim, Seongwon Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In *Advances in Neural Information Processing Systems*, pages 2760–2769, 2018.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better?, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Jinyu Li, Rui Zhao, Jui-Ting Huang, and Yifan Gong. Learning small-size dnn with output-distribution-based criteria. In *Interspeech*, September 2014. https://www.microsoft.com/en-us/research/publication/learning-small-size-dnn-with-output-distribution-based-criteria/.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations, 2021.

Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation, 2022. https://arxiv.org/abs/2211.16231.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017.

Dongyang Liu, Meina Kan, Shiguang Shan, and Xilin Chen. Function-consistent feature distillation, 2023a. https://arxiv.org/abs/2304.11832.

Wei Liu, Andrew Rabinovich, and Alexander C Berg. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 618–634, 2018.

Xiaolong Liu, Lujun Li, Chao Li, and Anbang Yao. Norm: Knowledge distillation via n-to-one representation matching, 2023b. https://arxiv.org/abs/2305.13803.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 116–131, 2018.

Roy Miles and Krystian Mikolajczyk. Understanding the role of the projector in knowledge distillation, 2024. https://arxiv.org/abs/2303.11098.

Roy Miles, Adrian Lopez Rodriguez, and Krystian Mikolajczyk. Information theoretic representation distillation, 2022. https://arxiv.org/abs/2112.00459.

Roy Miles, Ismail Elezi, and Jiankang Deng. v_kd : improving knowledge distillation using orthogonal projections, 2024. https://arxiv.org/abs/2403.06213.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant, 2019. https://arxiv.org/abs/1902.03393.

Arun Mishra and Debbie Marr. Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. In *International Conference on Learning Representations*, 2017.

Yulei Niu, Long Chen, Chang Zhou, and Hanwang Zhang. Respecting transfer gap in knowledge distillation, 2022. https://arxiv.org/abs/2210.12787.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019.

Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 268–284, 2018.

Baoyun Peng, Xi Li, Yifan Wu, Yizhou Fan, Bo Wang, Qi Tian, and Jun Liang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5007–5016, 2019.

Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. In *International Conference on Learning Representations*, 2018.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *Proceedings of the 4th International Conference on Learning Representations*, 2014.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

Li Shen and Marios Savvides. Amalgamating knowledge towards comprehensive classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1687–1696, 2020.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

Amit Singhal. Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin*, 24(4):35–43, 2001.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation, 2024. https://arxiv.org/abs/2403.01427.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation, 2022.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1365–1374, 2019.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (86):2579–2605, 2008. http://jmlr.org/papers/v9/vandermaaten08a.html.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. In *ICML*, pages 283–290, 2010.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

Jing Yang, Brais Martinez, Adrian Bulat, and Georgios Tzimiropoulos. Knowledge distillation via softmax regression representation learning. In *International Conference on Learning Representations*, 2021. https://openreview.net/forum?id=ZzwDy_wiWv.

Lei Yang, Xiaohang Zhan, Dapeng Chen, Junjie Yan, Chen Change Loy, and Dahua Lin. Learning to cluster faces on an affinity graph, 2019.

Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning, 2022.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017.

Li Yuan, Francis E. H. Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization, 2021. https://arxiv.org/abs/1909.11723.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *Proceedings of the 5th International Conference on Learning Representations*, 2016.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks, 2017.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.

Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning, 2017. https://arxiv.org/abs/1706.00384.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation, 2022. https://arxiv.org/abs/2203.08679.

Kaixiang Zheng and En-Hui Yang. Knowledge distillation based on transformed teacher matching, 2024. https://arxiv.org/abs/2402.11148.

Helong Zhou, Liangchen Song, Jiajie Chen, Ye Zhou, Guoli Wang, Junsong Yuan, and Qian Zhang. Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective, 2021. https://arxiv.org/abs/2102.00650.

Jinguo Zhu, Shixiang Tang, Dapeng Chen, Shijie Yu, Yakun Liu, Aijun Yang, Mingzhe Rong, and Xiaohua Wang. Complementary relation contrastive distillation, 2021a. https://arxiv.org/abs/2103.16367.

Jinguo Zhu, Shixiang Tang, Dapeng Chen, Shijie Yu, Yakun Liu, Aijun Yang, Mingzhe Rong, and Xiaohua Wang. Complementary relation contrastive distillation, 2021b. https://arxiv.org/abs/2103.16367.

Appendix

A Algorithm

We present the pseudocode of our method in Algorithm 1.

Algorithm 1 Pseudocode in a PyTorch-like style.

```
# f_s, f_t: student and teacher networks
# queue: memory of K features (CxK)
# t_s, t_t: temperature for student and teacher

for x in loader: # load a minibatch x with N samples
    s = f_s.forward(x) # student embeddings: NxC
    s = normalize(s, dim=1) # L2 normalization

    with torch.no_grad(): # no gradients
        t = f_t.forward(x) # teacher embeddings: NxC
        t = normalize(t, dim=1) # L2 normalization

    # enqueue the current minibatch
    enqueue(queue, t)

    # student similarities
    out_s = mm(s.view(N, C), queue.view(C, K))

    # teacher similarities
    out_t = mm(t.view(N, C), queue.view(C, K))

    # relational loss using softmax and log-softmax
    loss = -sum(
        softmax(out_t / t_t, dim=1) *
        log_softmax(out_s / t_s, dim=1), dim=1
    ).mean()

    # SGD update: student network only
    loss.backward()
    update(f_s.params)

    # dequeue the earliest minibatch
    dequeue(queue)
```

mm: matrix multiplication; softmax: exponential normalization across dimensions; log_softmax: logarithm of softmax; normalize: ℓ_2 normalization.

B Implementation Details

We implement RRD in PyTorch following the protocol of Tian et al. (2022)[1]. This protocol has become a common standard and is widely used among many papers to demonstrate their knowledge distillation methods (Hinton et al., 2015; Romero et al., 2014; Zagoruyko and Komodakis, 2016; Tung and Mori, 2019; Peng et al., 2019; Ahn et al., 2019; Park et al., 2019; Passalis and Tefas, 2018; Heo et al., 2019b; Kim et al., 2018; Yim et al., 2017; Huang and Wang, 2017; Tian et al., 2022). This approach is consistent with the standard practice in the field, allowing for fair comparison across methods. For visualization, we follow the implementation of (Sun et al., 2024)[2].

B.1 Baseline Methods

We compare our approach to the following state-of-the-art methods from the literature: (1) Knowledge Distillation (KD) (Hinton et al., 2015); (2) FitNets: Hints for Thin Deep Nets (Romero et al., 2014); (3) Attention Transfer (AT) (Zagoruyko and Komodakis, 2016); (4) Similarity-Preserving Knowledge Distillation (SP) (Tung and Mori, 2019); (5) Correlation Congruence (CC) (Peng et al., 2019); (6) Variational Information Distillation for Knowledge Transfer (VID) (Ahn et al., 2019); (7) Relational Knowledge Distillation (RKD) (Park et al., 2019); (8) Learning Deep Representations with Probabilistic Knowledge Transfer (PKT) (Passalis and Tefas, 2018); (9) Knowledge Transfer via Distillation of Activation Boundaries Formed by Hidden Neurons (AB) (Heo et al., 2019b); (10) Paraphrasing Complex Network: Network Compression via Factor Transfer

[1] Available at: https://github.com/HobbitLong/RepDistiller.

[2] Available at: https://github.com/sunshangquan/logit-standardization-KD.

(FT) (Kim et al., 2018); (11) A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning (FSP) (Yim et al., 2017); (12) Like What You Like: Knowledge Distill via Neuron Selectivity Transfer (NST) (Huang and Wang, 2017); (13) Contrastive Representation Distillation (CRD) (Tian et al., 2022); (14) A Comprehensive Overhaul of Feature Distillation (OFD) (Heo et al., 2019a); (15) Rethinking Soft Labels for Knowledge Distillation: A Bias-Variance Tradeoff Perspective (WSLD) (Zhou et al., 2021); (16) Respecting Transfer Gap in Knowledge Distillation (IPWD) (Niu et al., 2022); (17) Knowledge Distillation via Softmax Regression Representation Learning (SRRL) (Yang et al., 2021); (18) Cross-Layer Distillation with Semantic Calibration (SemCKD) (Chen et al., 2021a); (19) Distilling Knowledge via Knowledge Review (ReviewKD) (Chen et al., 2021c); (20) Knowledge Distillation with the Reused Teacher Classifier (SimKD) (Chen et al., 2022); (21) Searching A Fast Knowledge Distillation Process via Meta Optimization (DistPro) (Deng et al., 2022); (22) Knowledge Distillation via N-to-One Representation Matching (NORM) (Liu et al., 2023b); (23) Wasserstein Contrastive Representation Distillation (WCoRD) (Chen et al., 2021b); (24) Complementary Relation Contrastive Distillation (CRCD) (Zhu et al., 2021b); (25) Feature Kernel Distillation (FKD) (He and Ozay, 2022); (26) Information Theoretic Representation (ITRD) (Miles et al., 2022); (27) Test-Time Memory Distillation (TTM) (Zheng and Yang, 2024); (28) Weighted Test-Time Memory Distillation (WTTM) (Zheng and Yang, 2024); (29) Decoupled Knowledge Distillation (DKD) (Zhao et al., 2022); (30) Feature Correlation-based Feature Distillation (FCFD) (Liu et al., 2023a); (31) Category-aligned Knowledge Distillation (CAT-KD) (Guo et al., 2023); (32) Distillation with Instance-wise Selective Teaching (DIST) (Huang et al., 2022); (33) Cross-layer Transfer Knowledge Distillation (CTKD) (Li et al., 2022); (34) Relational Representation Distillation (RRD) (ours).

B.2 Datasets

We take five widely researched datasets: (1) CIFAR-100 (Krizhevsky, 2009) is a standard benchmark for knowledge distillation and contains 50,000 training images of size 32×32 with 500 images per class and 10,000 test images. (2) ImageNet ILSVRC-2012 (Deng et al., 2009), which is more challenging that CIFAR, and includes 1.2 million images from 1,000 classes for training and 50,000 for validation. (3) STL-10 (Coates and Ng, 2011) consists of a training set of 5,000 labeled images from 10 classes, and a test set of 8,000 images. (4) Tiny ImageNet (Deng et al., 2009) has 200 classes, each with 500 training images and 50 validation images. (5) MS-COCO (Lin et al., 2015) is an 80-category general object detection dataset. The `train2017` split contains 118,000 images, and the `val2017` split contains 5,000 images.

B.3 Network Architectures

We use the following network architectures as described in (Tian et al., 2022): (1) Wide Residual Network (WRN) (Zagoruyko and Komodakis, 2017), where WRN-d-w represents a wide ResNet with depth d and width factor w; (2) ResNet (rn) (He et al., 2015), where rn-d represents a CIFAR-style ResNet with 3 groups of basic blocks having 16, 32, and 64 channels, respectively, and rn-8 $\times 4$ and rn-32 $\times 4$ indicate a 4-times wider network with 64, 128, and 256 channels; (3) ResNet (RN) (He et al., 2015), where RN-d represents an ImageNet-style ResNet with Bottleneck blocks and more channels; (4) MobileNet-v2 (MN-v2) (Sandler et al., 2018), using a width multiplier of 0.5 in our experiments; (5) VGG (Simonyan and Zisserman, 2015), where the VGG network used is adapted from its original ImageNet counterpart; and (6) ShuffleNet-v1 (SN-v1) (Zhang et al., 2018) and ShuffleNet-v2 (SN-v2) (Ma et al., 2018), which are adapted for efficient training with input sizes of 32×32. To ensure dimensional consistency and preserve relational information during feature alignment, both teacher and student features are passed through lightweight *projection heads*—two-layer MLPs (512 hidden, 128 output)—that nonlinearly project intermediate representations into a shared embedding space for computing relational similarity distributions, while being discarded after training to avoid any additional inference cost. This projection head facilitates knowledge transfer by implicitly encoding relational information from previous samples (Miles and Mikolajczyk, 2024).

B.4 Optimization

We closely follow the protocol of Tian et al. (2022). All methods evaluated in our experiments use SGD with 0.9 Nesterov momentum. For CIFAR-100, we initialize the learning rate as 0.05, and decay it by 0.1 every 30 epochs after the first 150 epochs until the last 240 epoch. For MobileNet-v2, ShuffleNet-v1, and

ShuffleNet-v2, we use a learning rate of 0.01 as this learning rate is optimal for these models in a grid search, while 0.05 is optimal for other models. The batch size is set to 64 for CIFAR-100, and the weight decay is set to 5×10^{-4}. For ImageNet[3], the initial learning rate is set to 0.1 and then divided by 10 at the 30th, 60th, and 90th epochs of the total 120 training epochs. The mini-batch size is set to 256, and the weight decay is set to 1×10^{-4}. All results are reported as means over five trials, except for the results on ImageNet, which are reported in a single trial. Our implementation for MS-COCO follows the settings in (Zhao et al., 2022). We use the two-stage method Faster R-CNN (Ren et al., 2016) with Feature Pyramid Network (FPN) (Lin et al., 2017) as the detection framework. We evaluate three teacher–student settings: ResNet-101 to ResNet-18, ResNet-101 to ResNet-50, and ResNet-50 to MobileNet-V2 (Sandler et al., 2018). All students are trained with the 1× scheduler (schedulers and task-specific loss weights follow Detectron2 (Wu et al., 2019)). We take one RTX 6000 GPU to train the model on CIFAR-100 and four L40 GPUs on ImageNet and MS-COCO.

C Theoretical Analysis

We propose a learning objective with dual aims: aligning the student model's encoding $\mathbf{z}_i^S$ with the teacher model's $\mathbf{z}_i^T$, while enabling $\mathbf{z}_i^S$ to softly contrast with memory-stored random samples $\mathcal{M}$. This can be expressed by minimizing both the ℓ_2 distance between $\mathbf{z}_i^T$ and $\mathbf{z}_i^S$, and the cross-entropy using $\mathcal{M}$:

$$\begin{aligned}\mathcal{L} &= \frac{1}{N}\sum_i^N \Big\{\phi \cdot |\mathbf{z}_i^T - \mathbf{z}_i^S|_2 - \xi \cdot \mathbf{p}^T(\mathbf{x}_i;\theta_T,\mathcal{M}) \cdot \log \mathbf{p}^S(\mathbf{x}_i;\theta_S,\mathcal{M})\Big\} \\ &= \sum_i^N \Big\{ - \phi \cdot \mathbf{z}_i^T \cdot \mathbf{z}_i^S - \xi \cdot \sum_j^K \frac{\exp(\mathbf{z}_i^T \cdot \mathbf{m}_j/\tau_t)}{\sum\limits_{\mathbf{m}\sim\mathcal{M}} \exp(\mathbf{z}_i^T \cdot \mathbf{m}/\tau_t)} \cdot \log \frac{\exp(\mathbf{z}_i^S \cdot \mathbf{m}_j/\tau_s)}{\sum\limits_{\mathbf{m}\sim\mathcal{M}} \exp(\mathbf{z}_i^S \cdot \mathbf{m}/\tau_s)}\Big\}.\end{aligned} \tag{11}$$

Direct optimization of Equation (11) presents challenges in finding optimal hyperparameters (ϕ, ξ, τ_t and τ_s). Our objective using $\mathcal{M}^+$ serves as a simplified approximated upper-bound:

$$\begin{aligned}\mathcal{L}_{\text{RRD}} &= \frac{1}{N}\sum_i^N -\mathbf{p}^T(\mathbf{x}_i;\theta_T,\mathcal{M}^+) \cdot \log \mathbf{p}^S(\mathbf{x}_i;\theta_S,\mathcal{M}^+) \\ &= \sum_i^N \sum_j^{K+1} - \underbrace{\frac{\exp(\mathbf{z}_i^T \cdot \mathbf{m}_j/\tau_t)}{\sum\limits_{\mathbf{m}\sim\mathcal{M}^+} \exp(\mathbf{z}_i^T \cdot \mathbf{m}/\tau_t)}}_{\mathbf{a}_j^i} \cdot \log \frac{\exp(\mathbf{z}_i^S \cdot \mathbf{m}_j/\tau_s)}{\sum\limits_{\mathbf{m}\sim\mathcal{M}^+} \exp(\mathbf{z}_i^S \cdot \mathbf{m}/\tau_s)}.\end{aligned} \tag{12}$$

Here, α_j^i represents the temperature-controlled weighting term under τ_t. Since $\mathbf{z}_i^T$ is added as the $(K+1)$-th element in $\mathcal{M}^+$, we can expand the objective as:

$$\begin{aligned}\mathcal{L}_{\text{RRD}} =& \frac{1}{N}\sum_i^N \Big\{\mathbf{a}_{K+1}^i \cdot \big(- \mathbf{z}_i^S \cdot \mathbf{z}_i^T/\tau_s + \log \sum_{\mathbf{m}\sim\mathcal{M}^+} \exp(\mathbf{z}_i^S \cdot \mathbf{m}/\tau_s)\big) \\ &+ \sum_{j=1}^K \mathbf{a}_j^i \cdot \big(- \mathbf{z}_i^S \cdot \mathbf{m}_j/\tau_s + \log \sum_{\mathbf{m}\sim\mathcal{M}^+} \exp(\mathbf{z}_i^S \cdot \mathbf{m}/\tau_s)\big)\Big\}.\end{aligned} \tag{13}$$

The LSE function $\text{LSE}(\mathcal{M}^+, \mathbf{z}_i^S) = \log \sum_{m\sim\mathcal{M}^+} \exp(\mathbf{z}_i^S \cdot \mathbf{m}/\tau_s)$ denotes the Log-Sum-Exp. The LSE term in the first line maintains non-negativity since inner products between $\mathbf{z}^S$ and $\mathcal{M}$ fall within $[-1, +1]$:

$$\text{LSE}(\mathcal{M}^+, \mathbf{z}_i^S) \geq \log \big(M \cdot \exp(-1/\tau_s)\big) = \log \big(M \cdot \exp(-10)\big) > 0, \tag{14}$$

[3] Available at: https://www.image-net.org/.

where M represents the cardinality of the queue $\mathcal{M}^+$ with fixed $\tau_s = 0.1$.

The second line's LSE term follows:

$$\mathrm{LSE}(\mathcal{M}^+, \mathbf{z}_i^S) \geq \mathrm{LSE}(\mathcal{M}, \mathbf{z}_i^S). \tag{15}$$

This demonstrates that Equation (12) minimizes a relaxed upper-bound of Equation (11):

$$\begin{aligned}
\mathcal{L}_{\mathrm{RRD}} &= \frac{1}{N}\sum_i^N -\mathbf{p}^T(\mathbf{x}_i;\theta_T,\mathcal{M}^+)\cdot\log\mathbf{p}^S(\mathbf{x}_i;\theta_S,\mathcal{M}^+) \\
&\geq \frac{1}{N}\sum_i^N \Big\{\mathbf{a}^i_{K+1}\cdot(-\mathbf{z}_i^S\cdot\mathbf{z}_i^T/\tau_s) + \sum_{j=1}^K \mathbf{a}^i_j\cdot\big(-\mathbf{z}_i^S\cdot\mathbf{m}_j/\tau_s + \log\sum_{\mathbf{m}\sim\mathcal{M}}\exp(\mathbf{z}_i^S\cdot\mathbf{m}/\tau_s)\big)\Big\} \\
&= \frac{1}{N}\sum_i^N \Big\{-\frac{\mathbf{a}^i_{K+1}}{\tau_s}\cdot\mathbf{z}_i^S\cdot\mathbf{z}_i^T - \mathbf{p}^T(\mathbf{x}_i;\theta_T,\mathcal{M})\cdot\log\mathbf{p}^S(\mathbf{x}_i;\theta_S,\mathcal{M})\Big\}.
\end{aligned} \tag{16}$$

This establishes that $\mathcal{L}_{\mathrm{RRD}}$ connects to the intuitive distillation form in Equation (11) (ℓ_2 + cross entropy loss), inherently incorporating alignment and contrast objectives. Our approach simplifies the training process, and empirical evidence shows that both loss formulations yield comparable results when properly tuning τ_t.

C.1 Relation to InfoNCE

Our distillation approach extends InfoNCE (van den Oord et al., 2019) with probabilistic negative sample handling. InfoNCE treats negatives equally, while we apply probability-based weights. The standard InfoNCE formulation consists of pulling teacher encodings close while pushing away from random samples:

$$\hat{\theta}_S = \arg\min_{\theta_S}\mathcal{L}_{NCE} = \arg\min_{\theta_S}\sum_i^N -\log\frac{\exp(\mathbf{z}_i^T\cdot\mathbf{z}_i^S/\tau)}{\sum_{m\sim\mathcal{M}}\exp(\mathbf{z}_i^S\cdot\mathbf{m}/\tau)}. \tag{17}$$

This decomposes into *"pull"* and *"push"* forces - pull acts as positive sample alignment by drawing similar encodings together, while push provides contrast by moving away from negative instances:

$$\mathcal{L}_{NCE} = \sum_i^N\Big\{\underbrace{-\mathbf{z}_i^S\cdot\mathbf{z}_i^T/\tau}_{\text{pull}} + \underbrace{\log\sum_{m\sim\mathcal{M}}\exp(\mathbf{z}_i^S\cdot\mathbf{m}/\tau)}_{\text{push}}\Big\}. \tag{18}$$

Similarly, our objective follows this structure with added weighting:

$$\begin{aligned}
\mathcal{L}_{\mathrm{RRD}} &= \frac{1}{N\cdot M}\sum_i^N\sum_j^{K+1} -\frac{\exp(\mathbf{z}_i^T\cdot\mathbf{m}_j/\tau_t)}{\sum_{\mathbf{m}\sim\mathcal{M}^+}\exp(\mathbf{z}_i^T\cdot\mathbf{m}/\tau_t)}\cdot\log\frac{\exp(\mathbf{z}_i^S\cdot\mathbf{m}_j/\tau_s)}{\sum_{\mathbf{m}\sim\mathcal{M}^+}\exp(\mathbf{z}_i^S\cdot\mathbf{m})/\tau_s} \\
&= \frac{1}{N\cdot M}\sum_i^N\sum_j^{K+1}\underbrace{\frac{\exp(\mathbf{z}_i^T\cdot\mathbf{m}_j/\tau_t)}{\sum_{\mathbf{m}\sim\mathcal{M}^+}\exp(\mathbf{z}_i^T\cdot\mathbf{m}/\tau_t)}}_{\mathbf{a}^i_j}\cdot(\underbrace{-\mathbf{z}_i^S\cdot\mathbf{z}_i^T/\tau_s}_{\text{pull}} + \underbrace{\log\sum_{\mathbf{m}\sim\mathcal{M}}\exp(\mathbf{z}_i^S\cdot\mathbf{m}/\tau^S)}_{\text{push}})).
\end{aligned} \tag{19}$$

The soft labels $\mathbf{A}^i = [\mathbf{a}^i_1, ..., \mathbf{a}^i_{K+1}]$ give us:

Table 8 Memory update strategy ablation. Top-1 accuracy (%) on CIFAR-100 with WRN-40-2 → WRN-16-2 comparing FIFO and momentum ($\alpha = 0.999$) queues. ↑/↓ denote improvement/degradation relative to CRD (Tian et al., 2022). Results averaged over five runs.

Method	CRD	RRD	CRD	RRD	CRD+KD	RRD+KD	CRD+KD	RRD+KD
Momentum	✗	✗	✓	✓	✗	✗	✓	✓
Top-1	74.46	75.44 (↑)	73.55	73.56 (↑)	75.40	75.84 (↑)	75.21	75.45 (↑)

$$\mathcal{L}_{\text{RRD}} = \frac{1}{N \cdot M} \sum_{i}^{N} \sum_{j}^{K+1} \mathbf{a}_j^i \cdot \Big\{ - \mathbf{z}_i^S \cdot \mathbf{z}_i^T / \tau_s + \log \sum_{m \sim \mathcal{M}} \exp(\mathbf{z}_i^S \cdot \mathbf{m} / \tau^S)) \Big\}. \tag{20}$$

At low temperature τ_t, $\mathbf{A}^i$ becomes one-hot ($\mathbf{a}_{K+1}^i = 1$), reducing our objective to standard InfoNCE pull-push dynamics (Equation (18)). However, higher τ_t values improve performance - soft learning shows better results on CIFAR and similar datasets (Krizhevsky, 2009).

D Extended Ablation Results

This section provides the full tabular results underlying the ablation figures in Section 4.4. We examine memory bank configurations and update strategies (Section D.1), detailed sweeps over temperatures, memory size, and loss coefficients (Section D.2), a controlled comparison between RRD and CRD across queue types (Section D.3), and complete loss coefficient grids across all teacher–student pairs (Section D.4).

D.1 Memory Structure

The memory bank $\mathcal{M}$ extends the support of the similarity distributions beyond the current batch, simulating large-batch training without the associated cost and enabling the model to capture relationships that span batch boundaries. Two update strategies are considered: (i) a *First-In-First-Out* (FIFO) queue that replaces the oldest features with the current batch's teacher embeddings at each iteration, preserving the instantaneous relational geometry; and (ii) a *momentum* queue that updates stored features via an exponential moving average $\mathbf{m}_k \leftarrow \alpha \cdot \mathbf{m}_k + (1 - \alpha) \cdot \mathbf{z}_i^T$ with $\alpha \in [0, 1)$, producing smoother but potentially blurred relational targets. In both cases, appending the current teacher embedding as $\mathbf{m}_{K+1}$ ensures a clear peak in $\mathbf{p}^T$, preventing degeneration into weak contrastive alignment. We use a shared teacher memory bank rather than separate banks: separate banks would break the one-to-one correspondence required by the KL divergence, while storing evolving student features would produce unstable training targets.

As shown in Table 8, RRD outperforms CRD under both queue types, with the FIFO configuration yielding the strongest results. The FIFO queue preserves fine-grained relational distinctions that the EMA update tends to smooth out, resulting in sharper and more informative teacher distributions. This advantage holds both standalone and when combined with KD.

D.2 Temperature, Memory Size, and Loss Coefficients

Tables 9 to 12 provide the full numerical values behind Figure 3. The temperature sweep (Table 9) confirms that performance peaks when $\tau_t < \tau_s$ and degrades sharply when this asymmetry is reversed, validating the information bottleneck design. Excessively soft student distributions ($\tau_s = 1.0$) consistently underperform. The memory size sweep (Table 10) shows that both RRD and CRD improve with larger K, plateauing around K=16384; notably, RRD maintains stronger performance at smaller memory sizes, whereas CRD requires $K \geq 16384$ to become competitive. For loss coefficients, $\beta \in [0.5, 1.5]$ yields stable performance (Table 11), and combining with KD via $\lambda \approx 1.0$ provides a modest additional gain (Table 12). We default to β=1 when λ=0 and β=1.5 when $\lambda > 0$.

Table 9 Temperature sweep. Top-1 accuracy (%) on CIFAR-100 with WRN-40-2 → WRN-16-2, varying τ_t for fixed $\tau_s \in \{0.1, 0.2, 1.0\}$. Fixed: K=16384, β=1, λ=0. Mean over three runs.

τ_t	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.1	0.2	1.0
$\tau_s = 0.1$										
RRD (ours)	75.37	75.99	75.48	75.47	75.40	75.37	75.10	74.64	73.80	73.32
	(±0.14)	(±0.30)	(±0.17)	(±0.27)	(±0.10)	(±0.06)	(±0.32)	(±0.26)	(±0.18)	(±0.11)
$\tau_s = 0.2$										
RRD (ours)	75.33	75.57	75.24	74.95	74.93	74.64	74.50	74.31	73.89	73.62
	(±0.33)	(±0.06)	(±0.08)	(±0.12)	(±0.08)	(±0.36)	(±0.23)	(±0.31)	(±0.06)	(±0.27)
$\tau_s = 1.0$										
RRD (ours)	73.87	73.95	73.91	73.84	73.76	73.62	73.48	73.21	72.84	72.43
	(±0.25)	(±0.22)	(±0.23)	(±0.25)	(±0.23)	(±0.27)	(±0.29)	(±0.32)	(±0.34)	(±0.38)

Table 10 Memory size sweep. Top-1 accuracy (%) on CIFAR-100 with WRN-40-2 → WRN-16-2, varying K for both RRD and CRD. Fixed: τ_t=0.02, τ_s=0.1, β=1, λ=0. Mean over three runs.

K	64	256	1024	2048	4096	8192	16384	32768	65536
CRD	75.10	75.36	75.41	75.51	75.56	75.61	75.73	75.85	75.82
	(±0.10)	(±0.02)	(±0.16)	(±0.12)	(±0.25)	(±0.22)	(±0.11)	(±0.32)	(±0.32)
RRD (ours)	75.15	75.19	75.59	75.64	75.73	75.80	75.99	76.00	75.96
	(±0.45)	(±0.14)	(±0.09)	(±0.35)	(±0.11)	(±0.14)	(±0.30)	(±0.18)	(±0.06)

Table 11 Loss coefficient β sweep. Top-1 accuracy (%) on CIFAR-100 with WRN-40-2 → WRN-16-2. Fixed: K=16384, τ_t=0.02, τ_s=0.1, λ=0. Mean over three runs.

β	0	0.1	0.3	0.5	0.7	0.9	1.0	1.1	1.3	1.5	2.0	5.0	10
RRD	73.54	74.06	74.89	74.92	75.60	75.67	75.99	75.92	75.91	75.90	75.51	75.50	75.32
	(±0.52)	(±0.21)	(±0.11)	(±0.52)	(±0.18)	(±0.24)	(±0.30)	(±0.30)	(±0.33)	(±0.18)	(±0.28)	(±0.18)	(±0.26)

Table 12 Loss coefficient λ sweep. Top-1 accuracy (%) on CIFAR-100 with WRN-40-2 → WRN-16-2. Fixed: K=16384, τ_t=0.02, τ_s=0.1, β=1. Mean over three runs.

λ	0	0.1	0.3	0.5	0.7	0.9	1.0	2.0	5.0	10
RRD (ours)	75.99	75.45	75.73	75.73	75.75	75.81	75.86	75.59	75.59	75.04
	(±0.30)	(±0.09)	(±0.08)	(±0.19)	(±0.22)	(±0.24)	(±0.25)	(±0.11)	(±0.24)	(±0.11)

D.3 Comparison of CRD and RRD

Tables 13 and 14 provide a controlled comparison between RRD and CRD (Tian et al., 2022) across queue types, with shared settings (K=16384, β=1, λ=0.9 when combined with KD; τ_t=0.02, τ_s=0.1 for RRD; τ=0.07 for CRD; five runs). Several consistent patterns emerge. With a momentum queue, RRD outperforms CRD on all same-architecture pairs by +0.88 to +1.44 pp and on all cross-architecture pairs by +0.17 to +2.55 pp. With a FIFO queue and no KD, RRD leads on most same-architecture pairs but CRD is stronger on some cross-architecture settings, indicating that the two objectives interact differently with the queue mechanism. Adding KD narrows the gap between methods in both queue types, as logit-level supervision partially compensates for weaker feature-level alignment. Comparing best configurations per method (momentum for CRD, FIFO for RRD), standalone RRD achieves an average +2.95% gain; with KD, the margin reduces to +1.05%, confirming complementarity between relational and logit-based objectives.

Table 13 RRD vs CRD across queue types (same architecture). Top-1 accuracy (%) on CIFAR-100 for same-architecture teacher–student pairs under FIFO and momentum queues. Δ: absolute difference in percentage points. ↑/↓ denote improvement/degradation over CRD. Mean over five runs.

Teacher Student	WRN-40-2 WRN-16-2	WRN-40-2 WRN-40-1	rn-56 rn-20	rn-110 rn-20	rn-110 rn-32	rn-32x4 rn-8x4	VGG-13 VGG-8
			Momentum queue				
CRD	74.46	73.17	70.26	70.31	72.36	74.46	72.93
	(±0.21)	(±0.36)	(±0.18)	(±0.19)	(±0.16)	(±0.15)	(±0.24)
RRD (ours)	75.44	74.21	71.70	71.57	73.57	75.54	73.81
	(±0.23)	(±0.22)	(±0.09)	(±0.21)	(±0.26)	(±0.15)	(±0.12)
Δ	+0.98 (↑)	+1.04 (↑)	+1.44 (↑)	+1.26 (↑)	+1.21 (↑)	+1.08 (↑)	+0.88 (↑)
			FIFO queue				
CRD	73.55	71.14	68.10	68.05	70.38	73.47	70.76
	(±0.29)	(±0.35)	(±0.27)	(±0.28)	(±0.34)	(±0.16)	(±0.18)
RRD (ours)	73.56	71.86	69.82	69.68	71.55	72.94	70.75
	(±0.32)	(±0.17)	(±0.26)	(±0.31)	(±0.25)	(±0.34)	(±0.15)
Δ	+0.01 (↑)	+0.72 (↑)	+1.72 (↑)	+1.63 (↑)	+1.17 (↑)	-0.53 (↓)	-0.01 (↓)
			Momentum queue + KD				
CRD+KD	75.40	74.41	71.48	71.23	73.58	74.93	73.90
	(±0.24)	(±0.32)	(±0.18)	(±0.22)	(±0.27)	(±0.25)	(±0.29)
RRD+KD (ours)	75.84	74.72	71.83	71.44	73.59	74.78	74.08
	(±0.34)	(±0.20)	(±0.27)	(±0.45)	(±0.15)	(±0.23)	(±0.13)
Δ	+0.44 (↑)	+0.31 (↑)	+0.35 (↑)	+0.21 (↑)	+0.01 (↑)	-0.15 (↓)	+0.18 (↑)
			FIFO queue + KD				
CRD+KD	75.21	73.23	71.33	70.79	73.22	74.78	72.82
	(±0.33)	(±0.36)	(±0.33)	(±0.18)	(±0.24)	(±0.16)	(±0.24)
RRD+KD (ours)	75.45	73.98	71.39	71.06	73.54	74.39	73.45
	(±0.16)	(±0.20)	(±0.15)	(±0.45)	(±0.25)	(±0.18)	(±0.24)
Δ	+0.24 (↑)	+0.75 (↑)	+0.06 (↑)	+0.27 (↑)	+0.32 (↑)	-0.39 (↓)	+0.63 (↑)

D.4 Full Ablations of RRD on Loss Coefficients

Tables 15 and 16 report results across all 13 teacher–student pairs for five (β, λ) configurations. Standalone RRD with β=1 provides a strong baseline across same-architecture pairs, while β=1.5 improves cross-architecture transfer where larger capacity gaps demand stronger relational supervision. When combining with KD, λ=0.9 and λ=1.0 with β=1 yield comparable results. The configuration β=1.5, λ=0.9 achieves the best overall performance on cross-architecture pairs, confirming that stronger relational weighting is beneficial when the architectural mismatch is large.

E Extended Results

We supplement the main experiments with additional benchmarks and evaluation metrics. Section E.1 reports computational costs, Section E.2 and Section E.3 provide full comparisons on ImageNet and CIFAR-100 against all baselines listed in Section B.1, and Sections E.4 to E.6 extend the representation analyses with broader method coverage.

E.1 Efficiency Analysis

Table 17 reports per-batch training time on CIFAR-100 using ResNet-32x4 → ResNet-8x4, measured on an RTX 6000 GPU with the MDistiller framework[4]. We record both cold-start (first epoch) and warm-up

[4] Available at: https://github.com/megvii-research/mdistiller.

Table 14 RRD vs CRD across queue types (cross architecture). Top-1 accuracy (%) on CIFAR-100 for cross-architecture teacher–student pairs under FIFO and momentum queues. Δ: absolute difference in percentage points. ↑/↓ denote improvement/degradation over CRD. Mean over five runs.

Teacher	VGG-13	RN-50	RN-50	RN-32x4	RN-32x4	WRN-40-2
Student	MN-v2	MN-v2	VGG-8	SN-v1	SN-v2	SN-v1
Momentum queue						
CRD	67.44 (±0.44)	67.11 (±0.54)	72.55 (±0.27)	74.75 (±0.25)	75.91 (±0.31)	74.83 (±0.25)
RRD (ours)	69.34 (±0.38)	69.66 (±0.45)	74.15 (±0.25)	75.17 (±0.52)	76.08 (±0.45)	75.79 (±0.29)
Δ	+1.90 (↑)	+2.55 (↑)	+1.60 (↑)	+0.42 (↑)	+0.17 (↑)	+0.96 (↑)
FIFO queue						
CRD	65.79 (±0.40)	66.04 (±0.26)	70.69 (±0.18)	72.88 (±0.13)	73.68 (±0.18)	73.13 (±0.20)
RRD (ours)	65.57 (±0.33)	65.47 (±0.32)	71.16 (±0.51)	71.95 (±0.16)	73.13 (±0.39)	71.94 (±0.30)
Δ	-0.22 (↓)	-0.57 (↓)	+0.47 (↑)	-0.93 (↓)	-0.55 (↓)	-1.19 (↓)
Momentum queue + KD						
CRD+KD	67.86 (±0.16)	68.26 (±0.41)	73.03 (±0.24)	74.16 (±0.11)	74.92 (±0.23)	75.01 (±0.20)
RRD+KD (ours)	68.16 (±0.25)	68.14 (±0.80)	73.34 (±0.27)	74.29 (±0.10)	75.57 (±0.25)	75.57 (±0.33)
Δ	+0.30 (↑)	-0.12 (↓)	+0.31 (↑)	+0.13 (↑)	+0.65 (↑)	+0.56 (↑)
FIFO queue + KD						
CRD+KD	69.36 (±0.18)	69.38 (±0.80)	74.05 (±0.15)	75.51 (±0.23)	76.56 (±0.13)	76.30 (±0.28)
RRD+KD (ours)	69.75 (±0.34)	69.92 (±0.29)	74.20 (±0.27)	75.37 (±0.29)	76.51 (±0.33)	76.30 (±0.44)
Δ	+0.39 (↑)	+0.54 (↑)	+0.15 (↑)	-0.14 (↓)	-0.05 (↓)	0.00 (↑)

(subsequent epochs) timings. RRD adds 2.16M parameters from the projection heads and incurs moderate overhead relative to parameter-free methods such as KD and DKD, yet remains lighter than CRD (12.87M additional parameters). The cost is justified by the consistent accuracy gains reported across all benchmarks.

E.2 Results on ImageNet

Table 18 extends the ImageNet evaluation to additional teacher–student pairs and baselines beyond those in Table 3. RRD maintains its advantage across architectures, confirming that the relational objective scales to large-scale settings without requiring dataset-specific tuning beyond the adjustments noted in Section B.

E.3 Results on CIFAR-100

Tables 19 and 20 present the complete CIFAR-100 comparison against all 34 baselines for same-architecture and cross-architecture pairs, respectively. Relative improvements over KD are indicated with arrows. These tables serve as the exhaustive reference for the subset reported in Tables 1 and 2.

E.4 Results on Transferability

Table 21 broadens the transfer evaluation from Table 5 by including all baselines. A WRN-16-2 student distilled from WRN-40-2 on CIFAR-100 is used as a frozen feature extractor; only the final linear layer is retrained for 120 epochs on STL-10 and Tiny ImageNet. RRD yields the strongest overall transfer accuracy,

Table 15 Loss coefficient grid (same architecture). Top-1 accuracy (%) on CIFAR-100 for same-architecture pairs under different (β, λ) configurations. KD and CRD baselines from (Tian et al., 2022). Mean, std, and max over five runs.

Teacher Student	WRN-40-2 WRN-16-2	WRN-40-2 WRN-40-1	rn-56 rn-20	rn-110 rn-20	rn-110 rn-32	rn-32x4 rn-8x4	VGG-13 VGG-8
KD	74.92	73.54	70.66	70.67	73.08	73.33	72.98
CRD	75.48	74.14	71.16	71.46	73.48	75.51	73.94
CRD+KD	75.64	74.38	71.63	71.56	73.75	75.46	74.29
$\beta = 1$							
RRD (ours)	75.85	74.61	71.72	71.92	73.73	75.68	73.99
std	(±0.15)	(±0.21)	(±0.34)	(±0.08)	(±0.13)	(±0.18)	(±0.13)
max	76.10	74.96	72.41	72.03	73.89	76.10	74.18
$\beta = 1.5$							
RRD (ours)	75.74	74.47	71.48	71.60	73.56	75.97	73.94
std	(±0.17)	(±0.35)	(±0.35)	(±0.26)	(±0.12)	(±0.15)	(±0.21)
max	75.99	74.95	72.07	71.90	73.66	76.15	74.30
$\beta = 1, \lambda = 0.9$							
RRD+KD (ours)	75.60	74.70	71.73	71.37	73.80	75.25	73.99
std	(±0.19)	(±0.20)	(±0.16)	(±0.24)	(±0.28)	(±0.19)	(±0.28)
max	75.79	74.89	71.97	71.68	74.19	75.46	74.32
$\beta = 1, \lambda = 1$							
RRD+KD (ours)	75.53	74.51	71.57	71.34	73.76	75.43	73.89
std	(±0.13)	(±0.14)	(±0.30)	(±0.13)	(±0.11)	(±0.17)	(±0.15)
max	75.70	74.63	71.91	71.54	73.87	75.66	74.06
$\beta = 1.5, \lambda = 0.9$							
RRD+KD (ours)	75.67	74.68	72.03	71.75	73.96	75.53	74.37
std	(±0.36)	(±0.22)	(±0.31)	(±0.18)	(±0.09)	(±0.29)	(±0.38)
max	76.22	74.83	71.98	71.83	74.09	75.72	74.67

suggesting that the relational objective encourages generalizable feature geometry rather than task-specific shortcuts.

E.5 Correlation Evaluation

Table 22 extends the correlation alignment analysis of Table 6 to all baselines, using WRN-40-2 → WRN-40-1 on CIFAR-100. Frobenius distance, Pearson correlation, and SSIM are computed between teacher and student inter-class correlation matrices. RRD+KD achieves the closest structural match, while standalone RRD already surpasses all contrastive and most logit-based alternatives.

E.6 Semantic Evaluation

Table 23 provides the full-baseline version of the semantic analysis in Table 7, evaluated on CIFAR-100 with WRN-40-2 → WRN-40-1. Penultimate-layer features are ℓ_2-normalized before computing intra-class compactness, inter-class separation, NMI against the CIFAR-100 superclass taxonomy, and mAP@5. RRD+KD achieves the best trade-off between cluster tightness and retrieval precision across all evaluated methods.

F Visualization

We provide three complementary visualizations that illustrate the structural and semantic properties of distilled representations beyond the quantitative metrics reported in Section 4.3: inter-class correlation structures (Section F.1), t-SNE projections of feature embeddings (Section F.2), and nearest-neighbor retrievals (Section F.3).

Table 16 Loss coefficient grid (cross architecture). Top-1 accuracy (%) on CIFAR-100 for cross-architecture pairs under different (β, λ) configurations. KD and CRD baselines from (Tian et al., 2022). Mean, std, and max over five runs.

Teacher	VGG-13	RN-50	RN-50	RN-32x4	RN-32x4	WRN-40-2
Student	MN-v2	MN-v2	VGG-8	SN-v1	SN-v2	SN-v1
KD	67.37	67.35	73.81	74.07	74.45	74.83
CRD	69.73	69.11	74.30	75.11	75.65	76.05
CRD+KD	69.94	69.54	74.97	75.12	76.05	76.27
$\beta = 1$						
RRD (ours)	69.59	70.03	74.29	75.60	76.25	75.98
std	(±0.24)	(±0.22)	(±0.06)	(±0.24)	(±0.13)	(±0.07)
max	69.92	70.41	74.39	75.85	76.43	76.10
$\beta = 1.5$						
RRD (ours)	69.78	70.34	74.61	75.68	76.48	75.95
std	(±0.43)	(±0.41)	(±0.31)	(±0.16)	(±0.22)	(±0.21)
max	70.43	70.90	75.11	75.87	76.70	76.17
$\beta = 1, \lambda = 0.9$						
RRD+KD (ours)	69.80	69.73	74.18	75.49	76.31	76.39
std	(±0.26)	(±0.14)	(±0.10)	(±0.50)	(±0.23)	(±0.36)
max	70.25	69.90	74.30	75.93	76.70	76.65
$\beta = 1, \lambda = 1$						
RRD+KD (ours)	69.80	69.44	74.32	75.57	76.44	76.44
std	(±0.20)	(±0.87)	(±0.22)	(±0.30)	(±0.16)	(±0.10)
max	70.08	70.22	74.55	75.82	76.58	76.56
$\beta = 1.5, \lambda = 0.9$						
RRD+KD (ours)	69.99	69.65	74.53	76.68	76.87	76.64
std	(±0.85)	(±1.08)	(±0.12)	(±0.25)	(±0.08)	(±0.29)
max	70.16	70.62	74.66	76.04	76.95	76.97

Table 17 Efficiency analysis. Training time per batch (ms) on CIFAR-100. We set ResNet-32x4 as the teacher and ResNet-8x4 as the student. Experiments are conducted using the MDistiller framework on an RTX 6000 GPU. Time measurements compare the first epoch (cold start) to subsequent epochs (epoch 2 onwards) after warm-up.

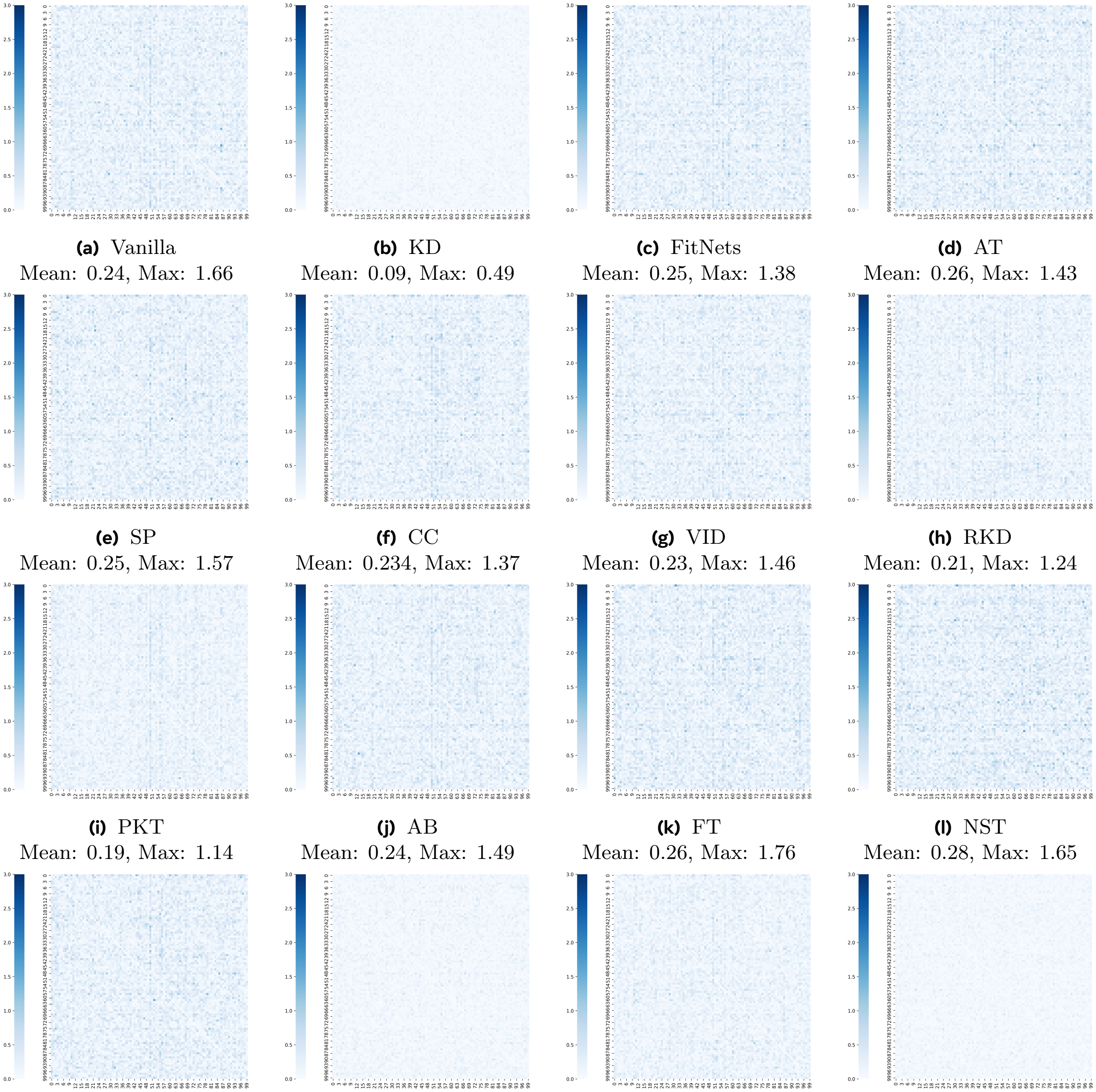
Method	Time (Epoch 1)	Time (Epoch 2)	Params
KD (Hinton et al., 2015)	13	8	0
RKD (Park et al., 2019)	22	17	0
CRD (Tian et al., 2022)	23	19	12,865,797
ReviewKD (Chen et al., 2021c)	19	14	1,808,774
DKD (Zhao et al., 2022)	13	9	0
RRD (ours)	20	15	2,162,944

F.1 Inter-class Correlation Matrices

Figure 4 displays the full set of absolute correlation-matrix differences between teacher (WRN-40-2) and student (WRN-40-1) logits on CIFAR-100, extending the selected subset in Figure 2 to all baselines. Class-wise mean logits are standardized to zero mean and unit variance before computing 100×100 Pearson correlation matrices; off-diagonal entries capture inter-class relational structure. RRD+KD exhibits the smallest deviations from the teacher, with visibly smoother residual maps compared to both contrastive and logit-based alternatives.

F.2 t-SNE Embeddings

Figures 5 and 6 show t-SNE (van der Maaten and Hinton, 2008) projections of ℓ_2-normalized penultimate features from WRN-40-2 (teacher) and WRN-40-1 (student) on the CIFAR-100 validation set (perplexity 40,



(a) Vanilla
Mean: 0.24, Max: 1.66

(b) KD
Mean: 0.09, Max: 0.49

(c) FitNets
Mean: 0.25, Max: 1.38

(d) AT
Mean: 0.26, Max: 1.43

(e) SP
Mean: 0.25, Max: 1.57

(f) CC
Mean: 0.234, Max: 1.37

(g) VID
Mean: 0.23, Max: 1.46

(h) RKD
Mean: 0.21, Max: 1.24

(i) PKT
Mean: 0.19, Max: 1.14

(j) AB
Mean: 0.24, Max: 1.49

(k) FT
Mean: 0.26, Max: 1.76

(l) NST
Mean: 0.28, Max: 1.65

(m) CRD
Mean: 0.23, Max: 1.56

(n) CRD+KD
Mean: 0.10, Max: 0.57

(o) RRD (ours)
Mean: 0.18, Max: 0.99

(p) RRD+KD (ours)
Mean: 0.07, Max: 0.55

Figure 4 Correlation alignment on CIFAR-100. Correlation matrix comparison between teacher (WRN-40-2) and student (WRN-40-1) logits. Lower values indicate stronger alignment of inter-class relations. Methods have been re-implemented accordingly. The abbreviations are listed in Section B.1.

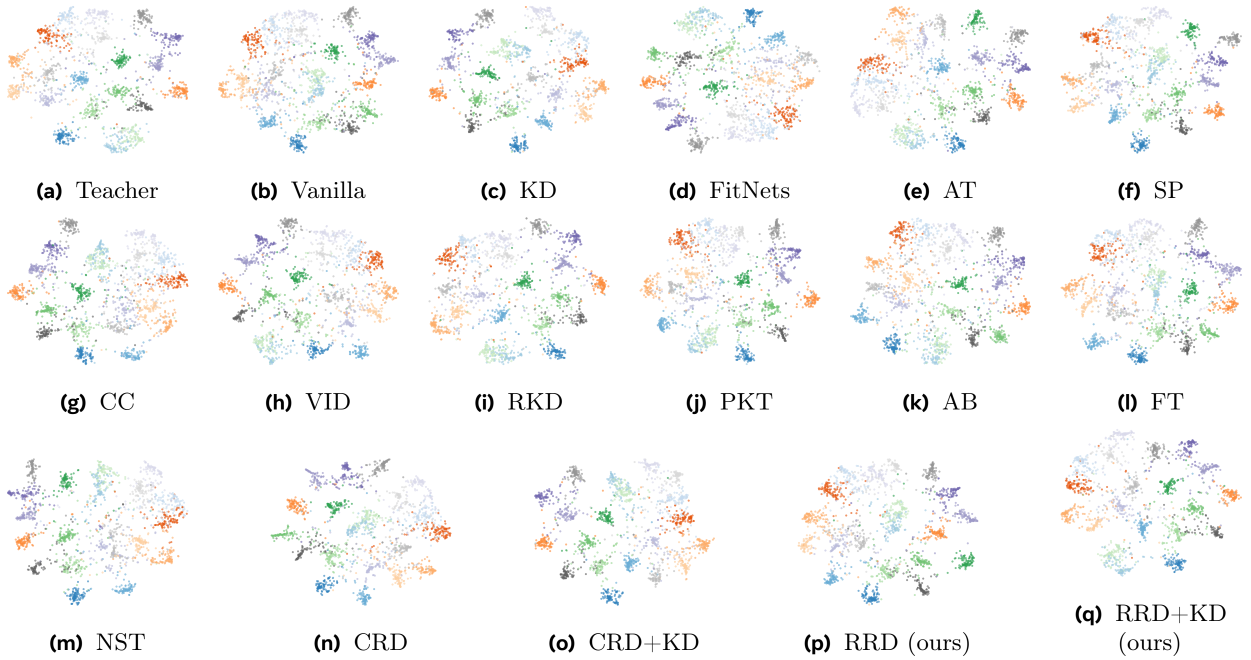

(a) Teacher (b) Vanilla (c) KD (d) FitNets (e) AT (f) SP (g) CC (h) VID (i) RKD (j) PKT (k) AB (l) FT (m) NST (n) CRD (o) CRD+KD (p) RRD (ours) (q) RRD+KD (ours)

Figure 5 t-SNE visualizations of embeddings from teacher and student networks on CIFAR-100 (first 20 classes). We use WRN-40-2 as the teacher and WRN-40-1 as the student. Methods have been re-implemented accordingly. The abbreviations are listed in Section B.1.

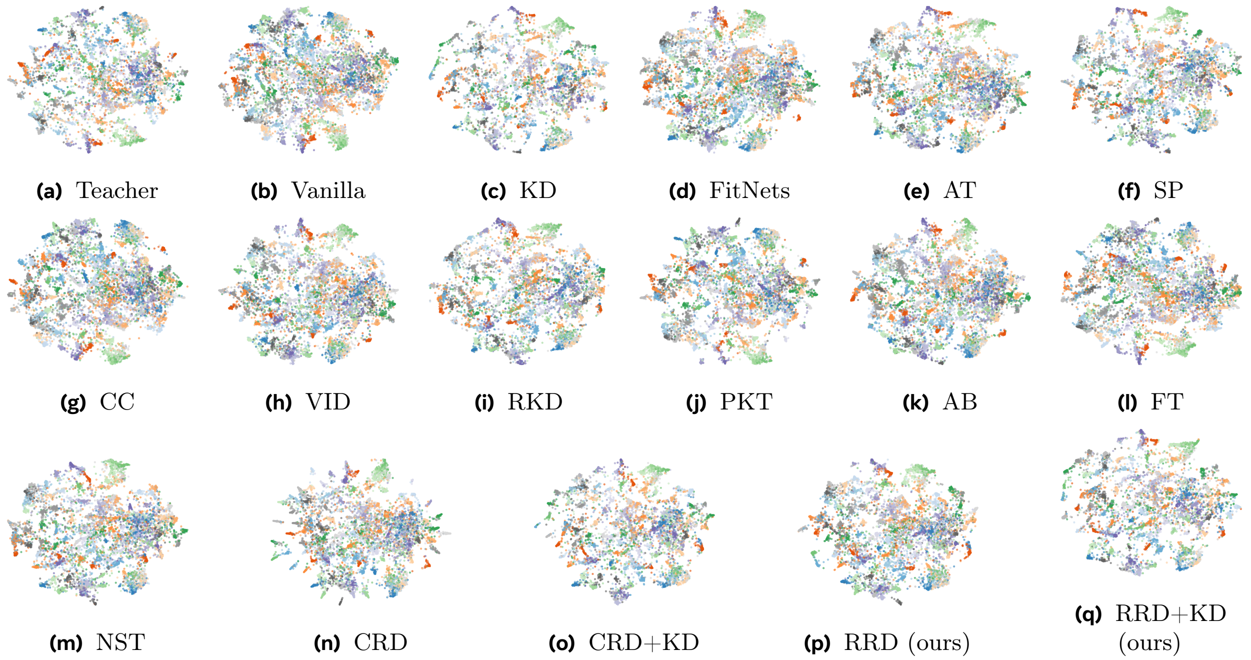

(a) Teacher (b) Vanilla (c) KD (d) FitNets (e) AT (f) SP (g) CC (h) VID (i) RKD (j) PKT (k) AB (l) FT (m) NST (n) CRD (o) CRD+KD (p) RRD (ours) (q) RRD+KD (ours)

Figure 6 t-SNE visualization of embeddings from teacher and student networks on the full CIFAR-100 dataset. WRN-40-2 is used as the teacher and WRN-40-1 as the student. All methods are re-implemented accordingly. The abbreviations are listed in Section B.1.

Table 18 Extended distillation results on ImageNet. Student top-1 accuracy (%) on the ImageNet validation set under various teacher–student pairs. The abbreviations are listed in Section B.1.

Teacher Student	ResNet-34 ResNet-18	ResNet-50 ResNet-18	ResNet-50 MobileNet-v2
Teacher	73.31	76.16	76.16
Student	69.75	69.75	69.63
KD (Hinton et al., 2015)	70.67	71.29	70.49
AT (Zagoruyko and Komodakis, 2016)	71.03	71.18	70.18
SP (Tung and Mori, 2019)	70.62	71.08	n/a
CC (Peng et al., 2019)	69.96	n/a	n/a
RKD (Park et al., 2019)	70.40	n/a	68.50
FT (Kim et al., 2018)	n/a	n/a	69.88
FSP (Yim et al., 2017)	70.58	n/a	n/a
CRD (Tian et al., 2022)	71.17	71.25	69.07
OFD (Heo et al., 2019a)	71.03	n/a	71.33
RRD (ours)	72.03	71.97	71.54
RRD+KD (ours)	71.99	71.88	71.56

300 iterations). The first set restricts to 20 classes for visual clarity; the second covers all 100 classes. RRD produces cluster arrangements that closely mirror the teacher's topology, with tighter intra-class grouping and cleaner inter-class boundaries than CRD or KD alone. The combined RRD+KD variant best reproduces the teacher's global embedding layout.

F.3 Nearest-Neighbor Retrieval

Figure 7 presents qualitative 5-NN retrieval examples on CIFAR-100 using cosine similarity over ℓ_2-normalized penultimate features (WRN-40-2 → WRN-40-1). For each query, the five most similar validation images (excluding the query) are shown. RRD retrieves semantically consistent neighbors that align with the teacher's ranking, while CRD occasionally surfaces off-class samples due to the over-separation effect discussed in Section 1. RRD+KD achieves near-perfect class-consistent retrieval across the displayed examples.

G Discussion

We focus ablation studies on CIFAR-100 for computational tractability, validating scalability on ImageNet. Since RRD can be viewed as a relaxation of CRD's contrastive framework (Section C.1), we adopt a compatible hyperparameter regime and find that performance remains stable across a wide range of settings (Section D). Full comparisons against all 34 baselines are provided in Section E.3; here we highlight key distinctions from related relational approaches. CRCD (Zhu et al., 2021a) learns 256-d relation vectors via MLP sub-networks and optimizes a binary classification objective with a single temperature $\tau = 0.05$; in contrast, RRD computes similarities directly via dot products, aligns distributions through KL divergence with asymmetric temperatures $\tau_t < \tau_s$, and admits InfoNCE as a limiting case when $\tau_t \to 0$ (Section C)—a theoretical connection absent in CRCD. CC (Peng et al., 2019) aligns channel-wise correlation matrices via Frobenius distance at the batch level, whereas RRD operates on instance-to-instance similarity distributions over a 16K memory bank with a dual-temperature information bottleneck—a fundamental difference in granularity and mechanism that accounts for the observed performance gap.

H Broader Impact

The presented research should be categorized as research in the field of knowledge distillation. The primary goal is to reduce computational demands, thereby lowering the energy requirements of AI systems and contributing to more sustainable technology deployment. However, this technique also harbors risks, notably

Table 19 Extended distillation results on CIFAR-100 (same architecture). Test top-1 accuracy (%) for different teacher–student pairs. ↑ denotes outperformance over KD and ↓ denotes underperformance. Results of our method are averaged over five runs. For RRD we set $\beta = 1$ and $\lambda = 0$. For RRD+KD we set $\beta = 1.5$ and $\lambda = 0.9$. The abbreviations are listed in Section B.1.

Teacher	WRN-40-2	WRN-40-2	rn-56	rn-110	rn-110	rn-32x4	VGG-13
Student	WRN-16-2	WRN-40-1	rn-20	rn-20	rn-32	rn-8x4	VGG-8
Teacher	75.61	75.61	72.34	74.31	74.31	79.42	74.64
Student	73.26	71.98	69.06	69.06	71.14	72.50	70.36
KD (Hinton et al., 2015)	74.92	73.54	70.66	70.67	73.08	73.33	72.98
FitNet (Romero et al., 2014)	73.58 (↓)	72.24 (↓)	69.21 (↓)	68.99 (↓)	71.06 (↓)	73.50 (↑)	71.02 (↓)
AT (Zagoruyko and Komodakis, 2016)	74.08 (↓)	72.77 (↓)	70.55 (↓)	70.22 (↓)	72.31 (↓)	73.44 (↑)	71.43 (↓)
SP (Tung and Mori, 2019)	73.83 (↓)	72.43 (↓)	69.67 (↓)	70.04 (↓)	72.69 (↓)	72.94 (↓)	72.68 (↓)
CC (Peng et al., 2019)	73.56 (↓)	72.21 (↓)	69.63 (↓)	69.48 (↓)	71.48 (↓)	72.97 (↓)	70.81 (↓)
VID (Ahn et al., 2019)	74.11 (↓)	73.30 (↓)	70.38 (↓)	70.16 (↓)	72.61 (↓)	73.09 (↓)	71.23 (↓)
RKD (Park et al., 2019)	73.35 (↓)	72.22 (↓)	69.61 (↓)	69.25 (↓)	71.82 (↓)	71.90 (↓)	71.48 (↓)
PKT (Passalis and Tefas, 2018)	74.54 (↓)	73.45 (↓)	70.34 (↓)	70.25 (↓)	72.61 (↓)	73.64 (↑)	72.88 (↓)
AB (Heo et al., 2019b)	72.50 (↓)	72.38 (↓)	69.47 (↓)	69.53 (↓)	70.98 (↓)	73.17 (↓)	70.94 (↓)
FT (Kim et al., 2018)	73.25 (↓)	71.59 (↓)	69.84 (↓)	70.22 (↓)	72.37 (↓)	72.86 (↓)	70.58 (↓)
FSP (Yim et al., 2017)	72.91 (↓)	n/a	69.95 (↓)	70.11 (↓)	71.89 (↓)	72.62 (↓)	70.33 (↓)
NST (Huang and Wang, 2017)	73.68 (↓)	72.24 (↓)	69.60 (↓)	69.53 (↓)	71.96 (↓)	73.30 (↓)	71.53 (↓)
CRD (Tian et al., 2022)	75.48 (↑)	74.14 (↑)	71.16 (↑)	71.46 (↑)	73.48 (↑)	75.51 (↑)	73.94 (↑)
CRD+KD (Tian et al., 2022)	75.64 (↑)	74.38 (↑)	71.63 (↑)	71.56 (↑)	73.75 (↑)	75.46 (↑)	74.29 (↑)
OFD (Heo et al., 2019a)	75.24 (↑)	74.33 (↑)	70.38 (↓)	n/a	73.23 (↑)	74.95 (↑)	73.95 (↑)
WSLD (Zhou et al., 2021)	n/a	73.74 (↑)	71.53 (↑)	n/a	73.36 (↑)	74.79 (↑)	n/a
IPWD (Niu et al., 2022)	n/a	74.64 (↑)	71.32 (↑)	n/a	73.91 (↑)	76.03 (↑)	n/a
SRRL (Yang et al., 2021)	n/a	74.64 (↑)	n/a	n/a	n/a	75.39 (↑)	n/a
SemCKD (Chen et al., 2021a)	n/a	74.41 (↑)	n/a	n/a	n/a	76.23 (↑)	n/a
ReviewKD (Chen et al., 2021c)	76.12 (↑)	75.09 (↑)	71.89 (↑)	n/a	73.89 (↑)	75.63 (↑)	74.84 (↑)
SimKD (Chen et al., 2022)	n/a	75.56 (↑)	n/a	n/a	n/a	78.08 (↑)	n/a
DistPro (Deng et al., 2022)	76.36 (↑)	n/a	72.03 (↑)	n/a	73.74 (↑)	n/a	n/a
NORM (Liu et al., 2023b)	75.65 (↑)	74.82 (↑)	71.35 (↑)	71.55 (↑)	73.67 (↑)	76.49 (↑)	73.95 (↑)
NORM+KD (Liu et al., 2023b)	76.26 (↑)	75.42 (↑)	71.61 (↑)	72.00 (↑)	74.95 (↑)	76.98 (↑)	74.46 (↑)
NORM+CRD (Liu et al., 2023b)	76.02 (↑)	75.37 (↑)	71.51 (↑)	71.90 (↑)	73.81 (↑)	76.49 (↑)	73.58 (↑)
WCoRD (Chen et al., 2021b)	75.88 (↑)	74.73 (↑)	71.56 (↑)	71.57 (↑)	73.81 (↑)	75.95 (↑)	74.55 (↑)
WCoRD+KD (Chen et al., 2021b)	76.11 (↑)	74.72 (↑)	71.92 (↑)	71.88 (↑)	74.20 (↑)	76.15 (↑)	74.72 (↑)
CRCD (Zhu et al., 2021b)	76.67 (↑)	75.95 (↑)	73.21 (↑)	72.33 (↑)	74.98 (↑)	76.42 (↑)	74.97 (↑)
FKD (He and Ozay, 2022)	n/a	n/a	n/a	n/a	n/a	75.57 (↑)	73.78 (↑)
ITRD (corr) (Miles et al., 2022)	75.85 (↑)	74.90 (↑)	71.45 (↑)	71.77 (↑)	74.02 (↑)	75.63 (↑)	74.70 (↑)
ITRD (corr+mi) (Miles et al., 2022)	76.12 (↑)	75.18 (↑)	71.47 (↑)	71.99 (↑)	74.26 (↑)	76.19 (↑)	74.93 (↑)
TTM (Zheng and Yang, 2024)	76.23 (↑)	74.32 (↑)	71.83 (↑)	71.46 (↑)	73.97 (↑)	76.17 (↑)	74.33 (↑)
WTTM (Zheng and Yang, 2024)	76.37 (↑)	74.58 (↑)	71.92 (↑)	71.67 (↑)	74.13 (↑)	76.06 (↑)	74.44 (↑)
WTTM+CRD (Zheng and Yang, 2024)	76.61 (↑)	74.94 (↑)	72.20 (↑)	72.13 (↑)	74.52 (↑)	76.65 (↑)	74.71 (↑)
WTTM+ITRD (Zheng and Yang, 2024)	76.65 (↑)	75.34 (↑)	72.16 (↑)	72.20 (↑)	74.36 (↑)	77.36 (↑)	75.13 (↑)
DKD (Zhao et al., 2022)	76.24 (↑)	74.81 (↑)	71.97 (↑)	n/a	74.11 (↑)	76.32 (↑)	74.68 (↑)
FCFD (Liu et al., 2023a)	76.34 (↑)	75.43 (↑)	71.68 (↑)	n/a	n/a	76.80 (↑)	74.86 (↑)
FCFD+KD (Liu et al., 2023a)	76.43 (↑)	75.46 (↑)	71.96 (↑)	n/a	n/a	76.62 (↑)	75.22 (↑)
CAT-KD (Guo et al., 2023)	75.60 (↑)	74.82 (↑)	71.62 (↑)	n/a	73.62 (↑)	76.91 (↑)	74.65 (↑)
DIST (Huang et al., 2022)	n/a	74.73 (↑)	71.75 (↑)	n/a	n/a	76.31 (↑)	n/a
CTKD (Li et al., 2022)	75.45 (↑)	73.93 (↑)	71.19 (↑)	70.99 (↑)	73.52 (↑)	n/a	73.52 (↑)
RRD (ours)	75.85 (↑)	74.61 (↑)	71.89 (↑)	71.92 (↑)	73.73 (↑)	75.77 (↑)	74.01 (↑)
RRD+KD (ours)	75.67 (↑)	74.68 (↑)	72.03 (↑)	71.75 (↑)	73.96 (↑)	75.53 (↑)	74.37 (↑)

Table 20 Extended distillation results on CIFAR-100 (different architecture). Test top-1 accuracy (%) for different teacher–student pairs. ↑ denotes outperformance over KD and ↓ denotes underperformance. Results of our method are averaged over five runs. For RRD we set $\beta = 1$ and $\lambda = 0$. For RRD+KD we set $\beta = 1.5$ and $\lambda = 0.9$. The abbreviations are listed in Section B.1.

Teacher Student	VGG-13 MN-v2	RN-50 MN-v2	RN-50 VGG-8	RN-32x4 SN-v1	RN-32x4 SN-v2	WRN-40-2 SN-v1
Teacher	74.64	79.34	79.34	79.42	79.42	75.61
Student	64.60	64.60	70.36	70.5	71.82	70.5
KD (Hinton et al., 2015)	67.37	67.35	73.81	74.07	74.45	74.83
FitNet (Romero et al., 2014)	64.14 (↓)	63.16 (↓)	70.69 (↓)	73.59 (↓)	73.54 (↓)	73.73 (↓)
AT (Zagoruyko and Komodakis, 2016)	59.40 (↓)	58.58 (↓)	71.84 (↓)	71.73 (↓)	72.73 (↓)	73.32 (↓)
SP (Tung and Mori, 2019)	66.30 (↓)	68.08 (↑)	73.34 (↓)	73.48 (↓)	74.56 (↑)	74.52 (↓)
CC (Peng et al., 2019)	64.86 (↓)	65.43 (↓)	70.25 (↓)	71.14 (↓)	71.29 (↓)	71.38 (↓)
VID (Ahn et al., 2019)	65.56 (↓)	67.57 (↑)	70.30 (↓)	73.38 (↓)	73.40 (↓)	73.61 (↓)
RKD (Park et al., 2019)	64.52 (↓)	64.43 (↓)	71.50 (↓)	72.28 (↓)	73.21 (↓)	72.21 (↓)
PKT (Passalis and Tefas, 2018)	67.13 (↓)	66.52 (↓)	73.01 (↓)	74.10 (↑)	74.69 (↑)	73.89 (↓)
AB (Heo et al., 2019b)	66.06 (↓)	67.20 (↓)	70.65 (↓)	73.55 (↓)	74.31 (↓)	73.34 (↓)
FT (Kim et al., 2018)	61.78 (↓)	60.99 (↓)	70.29 (↓)	71.75 (↓)	72.50 (↓)	72.03 (↓)
NST (Huang and Wang, 2017)	58.16 (↓)	64.96 (↓)	71.28 (↓)	74.12 (↑)	74.68 (↑)	76.09 (↑)
CRD (Tian et al., 2022)	69.73 (↑)	69.11 (↑)	74.3 (↑)	75.11 (↑)	75.65 (↑)	76.05 (↑)
CRD+KD (Tian et al., 2022)	69.94 (↑)	69.54 (↑)	74.58 (↑)	75.12 (↑)	76.05 (↑)	76.27 (↑)
OFD (Heo et al., 2019a)	69.48 (↑)	69.04 (↑)	n/a	75.98 (↑)	76.82 (↑)	75.85 (↑)
WSLD (Zhou et al., 2021)	n/a	68.79 (↑)	73.80 (↓)	75.09 (↑)	n/a	75.23 (↑)
IPWD (Niu et al., 2022)	n/a	70.25 (↑)	74.97 (↑)	76.03 (↑)	n/a	76.44 (↑)
SRRL (Yang et al., 2021)	n/a	n/a	n/a	75.18 (↑)	n/a	n/a
SemCKD (Chen et al., 2021a)	n/a	n/a	n/a	n/a	77.62 (↑)	n/a
ReviewKD (Chen et al., 2021c)	70.37 (↑)	69.89 (↑)	n/a	77.45 (↑)	77.78 (↑)	77.14 (↑)
SimKD (Chen et al., 2022)	n/a	n/a	n/a	77.18 (↑)	n/a	n/a
DistPro (Deng et al., 2022)	n/a	n/a	n/a	77.18 (↑)	77.54 (↑)	77.24 (↑)
NORM (Liu et al., 2023b)	68.94 (↑)	70.56 (↑)	75.17 (↑)	77.42 (↑)	78.07 (↑)	77.06 (↑)
NORM+KD (Liu et al., 2023b)	69.38 (↑)	71.17 (↑)	75.67 (↑)	77.79 (↑)	78.32 (↑)	77.63 (↑)
NORM+CRD (Liu et al., 2023b)	69.17 (↑)	71.08 (↑)	75.51 (↑)	77.50 (↑)	77.96 (↑)	77.09 (↑)
WCoRD (Chen et al., 2021b)	69.47 (↑)	70.45 (↑)	74.86 (↑)	75.40 (↑)	75.96 (↑)	76.32 (↑)
WCoRD+KD (Chen et al., 2021b)	70.02 (↑)	70.12 (↑)	74.68 (↑)	75.77 (↑)	76.48 (↑)	76.68 (↑)
CRCD (Zhu et al., 2021b)	n/a	n/a	n/a	n/a	n/a	n/a
FKD (He and Ozay, 2022)	n/a	n/a	74.61 (↑)	75 (↑)	n/a	n/a
ITRD (corr) (Miles et al., 2022)	69.97 (↑)	71.41 (↑)	75.71 (↑)	76.8 (↑)	77.27 (↑)	77.35 (↑)
ITRD (corr+mi) (Miles et al., 2022)	70.39 (↑)	71.34 (↑)	75.49 (↑)	76.91 (↑)	77.40 (↑)	77.09 (↑)
TTM (Zheng and Yang, 2024)	68.98 (↑)	69.24 (↑)	74.87 (↑)	74.18 (↑)	76.57 (↑)	75.39 (↑)
WTTM (Zheng and Yang, 2024)	69.16 (↑)	69.59 (↑)	74.82 (↑)	74.37 (↑)	76.55 (↑)	75.42 (↑)
WTTM+CRD (Zheng and Yang, 2024)	70.30 (↑)	70.84 (↑)	75.30 (↑)	75.82 (↑)	77.04 (↑)	76.86 (↑)
WTTM+ITRD (Zheng and Yang, 2024)	70.70 (↑)	71.56 (↑)	76.00 (↑)	77.03 (↑)	77.68 (↑)	77.44 (↑)
DKD (Zhao et al., 2022)	69.71 (↑)	70.35 (↑)	n/a	76.45 (↑)	77.07 (↑)	76.70 (↑)
FCFD (Liu et al., 2023a)	70.67 (↑)	71.07 (↑)	n/a	78.12 (↑)	78.20 (↑)	77.81 (↑)
FCFD+KD (Liu et al., 2023a)	70.65 (↑)	71.00 (↑)	n/a	78.12 (↑)	78.18 (↑)	77.99 (↑)
CAT-KD (Guo et al., 2023)	69.13 (↑)	71.36 (↑)	n/a	78.26 (↑)	78.41 (↑)	77.35 (↑)
DIST (Huang et al., 2022)	n/a	68.66 (↑)	n/a	76.34 (↑)	77.35 (↑)	n/a
CTKD (Li et al., 2022)	68.46 (↑)	68.47 (↑)	n/a	74.78 (↑)	75.31 (↑)	75.78 (↑)
RRD (ours)	69.61 (↑)	70.11 (↑)	74.30 (↑)	75.60 (↑)	76.31 (↑)	75.98 (↑)
RRD+KD (ours)	69.99 (↑)	69.65 (↑)	74.53 (↑)	76.68 (↑)	76.87 (↑)	76.64 (↑)

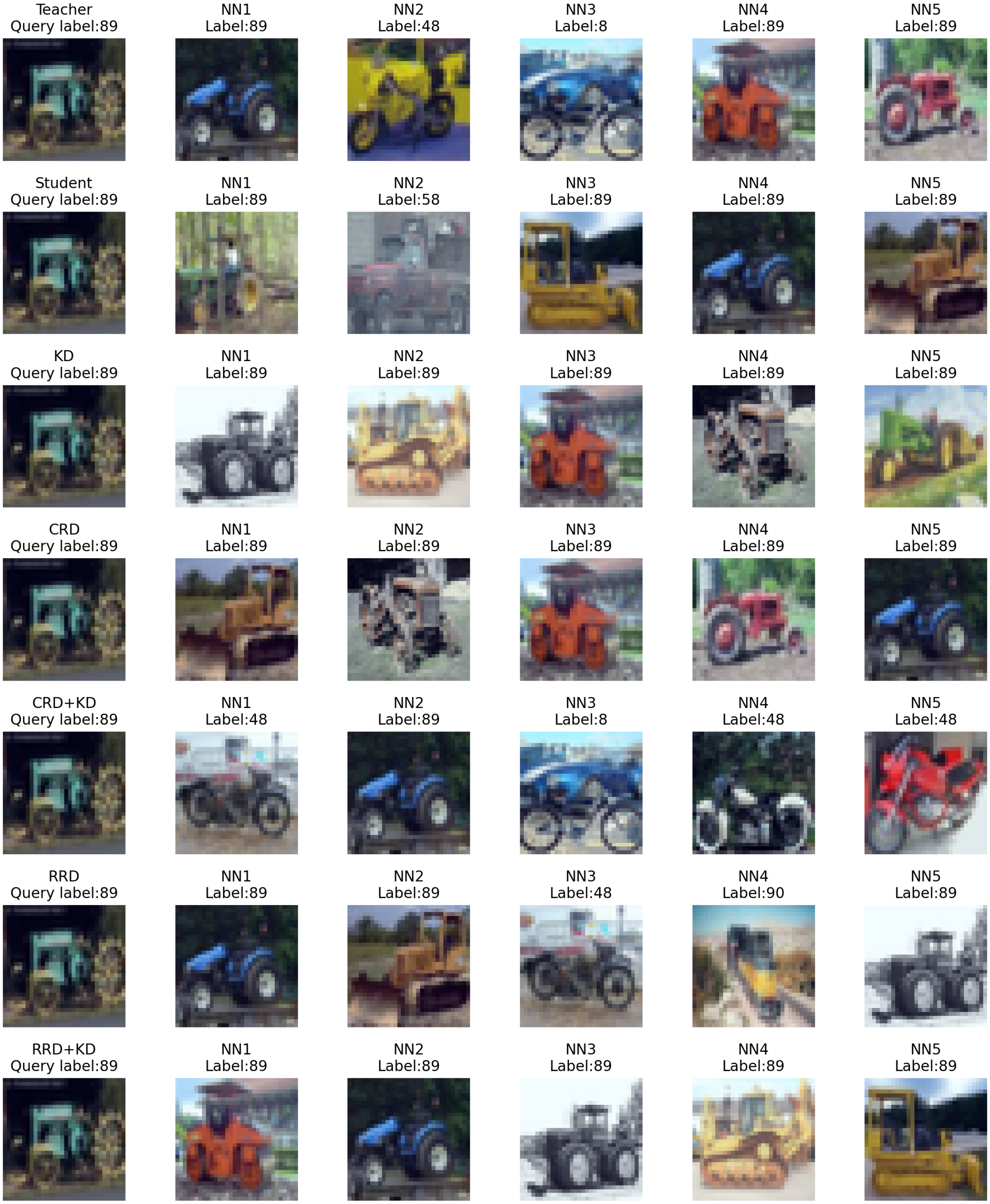


Figure 7 Nearest-neighbor retrieval on CIFAR-100. WRN-40-2 is used as the teacher and WRN-40-1 as the student. Qualitative 5-NN retrieval results showing semantic consistency of student representations distilled with different methods. The abbreviations are listed in Section B.1.

Table 21 Extended transfer learning performance results with distilled students. Top-1 accuracy (%) of a WRN-16-2 student distilled from WRN-40-2, evaluated on STL-10 and Tiny ImageNet. Methods have been re-implemented accordingly. The abbreviations are listed in Section B.1.

	λ	β	CIFAR-100 → STL-10	CIFAR-100 → Tiny ImageNet
Teacher	n/a	n/a	68.6	31.5
Student	n/a	n/a	69.7	33.7
			without KL divergence	
CRD	n/a	n/a	71.6	35.6
RRD (ours)	0	1	72.0 ± 0.65	35.5 ± 0.36
RRD (ours)	0	1.5	72.3 ± 0.31	35.5 ± 0.42
			with KL divergence	
CRD+KD	n/a	n/a	72.2	35.5
RRD+KD (ours)	0.9	1	71.9 ± 0.46	35.0 ± 0.13
RRD+KD (ours)	0.9	1.5	72.0 ± 0.28	35.2 ± 0.38

Table 22 Extended correlation matrix alignment metrics on CIFAR-100. Quantitative comparison of relational alignment between teacher (WRN-40-2) and student (WRN-40-1) embeddings using Frobenius distance, Pearson correlation, and SSIM. Methods have been re-implemented accordingly. The abbreviations are listed in Section B.1.

Method	Frobenius ↓	Pearson ↑	SSIM ↑
Vanilla	10.491	0.951	0.935
KD (Hinton et al., 2015)	2.868	0.994	0.992
FitNet (Romero et al., 2014)	10.491	0.950	0.934
AT (Zagoruyko and Komodakis, 2016)	9.450	0.953	0.941
SP (Tung and Mori, 2019)	8.932	0.955	0.945
CC (Peng et al., 2019)	11.148	0.944	0.928
VID (Ahn et al., 2019)	9.180	0.960	0.950
RKD (Park et al., 2019)	8.994	0.963	0.953
PKT (Passalis and Tefas, 2018)	6.285	0.982	0.973
ABOUND (Heo et al., 2019b)	9.523	0.956	0.943
FT (Kim et al., 2018)	12.584	0.932	0.913
NST (Huang and Wang, 2017)	10.574	0.938	0.923
CRD (Tian et al., 2022)	8.862	0.946	0.938
CRD+KD (Tian et al., 2022)	3.265	0.992	0.989
RRD (ours)	6.969	0.970	0.964
RRD+KD (ours)	2.720	0.995	0.993

the potential to perpetuate existing biases present in teacher models. Such biases could have profound ethical implications, as in sensitive applications. Furthermore, the versatility of the algorithms developed here enables their application across a broad spectrum of vision-related tasks, but this versatility also introduces the dual-use dilemma, where the technology might yield both beneficial and adverse impacts. Thus, careful consideration is needed when deploying these methods to ensure they align with ethical guidelines and promote fairness in AI applications.

I Reproducibility Statement

Our code is publicly available at `https://github.com/giakoumoglou/rrd`. The implementation builds on the RepDistiller (Tian et al., 2022) codebase, a widely adopted standard for knowledge distillation evaluation, and follows its experimental protocol to ensure fair comparison across all baselines. Pseudocode is provided in Algorithm 1, all hyperparameters—temperature settings (τ_t, τ_s), memory size (K), and loss coefficients (β, λ)—are detailed in Section B, and comprehensive ablation studies are reported in Section D.

Table 23 Extended semantic similarity preservation on CIFAR-100. Evaluation of intra-/inter-class structure, NMI, and retrieval mAP@5 of student embeddings under differet distillation methods. The abbreviations are listed in Section B.1.

Method	Intra ↓	Inter ↑	NMI ↑	mAP@5 ↑
Vanilla	0.945	0.866	0.551	81.1
KD (Hinton et al., 2015)	0.914	1.012	0.546	84.2
FitNet (Romero et al., 2014)	0.945	0.873	0.556	82.0
AT (Zagoruyko and Komodakis, 2016)	0.882	0.891	0.565	82.8
SP (Tung and Mori, 2019)	0.768	0.905	0.568	84.2
CC (Peng et al., 2019)	0.948	0.872	0.562	81.2
VID (Ahn et al., 2019)	0.905	0.892	0.562	82.6
RKD (Park et al., 2019)	0.985	0.897	0.549	82.5
PKT (Passalis and Tefas, 2018)	0.882	0.942	0.581	84.9
ABOUND (Heo et al., 2019b)	0.944	0.878	0.557	82.1
FT (Kim et al., 2018)	0.882	0.887	0.550	81.9
NST (Huang and Wang, 2017)	0.831	0.879	0.564	80.8
CRD (Tian et al., 2022)	0.910	0.915	0.509	84.6
CRD+KD (Tian et al., 2022)	0.973	0.926	0.463	85.2
RRD (ours)	0.928	0.909	0.544	83.9
RRD+KD (ours)	0.902	1.023	0.562	85.9