Relational Representation Distillation

Nikolaos Giakoumoglou Tania Stathaki
Imperial College London
{nikos, t.stathaki}@imperial.ac.uk

Code: https://github.com/giakoumoglou/rrd

Abstract

*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.*

1. Introduction

Knowledge Distillation (KD) transfers knowledge from large, high-capacity *teacher* models to compact *student* models [19]. This approach is increasingly relevant as state-of-the-art vision models for image classification [12, 33], object detection [29, 42], and semantic segmentation [6, 7] continue to grow in size and computational cost [14, 23], motivating efficient model compression techniques [3, 41].

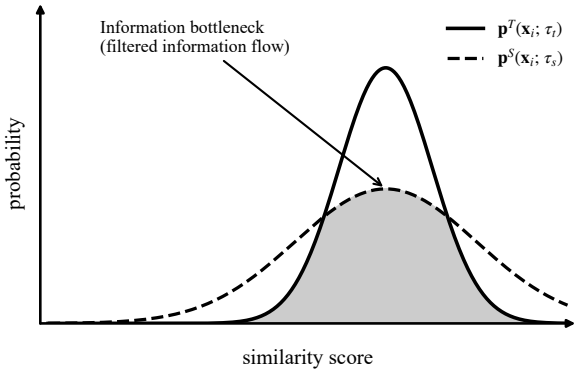


Figure 1. **Visualization of the *information bottleneck* effect.** The teacher produces a sharper similarity distribution $\mathbf{p}^T(\mathbf{x}_i;\tau_t)$ (solid black) highlighting primary relationships, while the student adopts a softer distribution $\mathbf{p}^S(\mathbf{x}_i;\tau_s)$ (dashed black) 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 $I(\mathbf{z}^T;\mathbf{z}^S)$.

The seminal work of Bucilua *et al*. [3] and Hinton *et al*. [19] 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.

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 [40, 43, 54, 56] extend distillation to intermediate representations. However, Tian *et al*. [47] showed that such approaches still neglect the structural knowledge encoded in the teacher's internal representations. To overcome this limitation, Tian *et al*. [47] further adapted the family of contrastive objectives [1, 16, 20, 49] 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"* [1, 26, 53], in which semantically similar samples are undesirably pushed apart due to uniform negative sampling. As shown in our experiments (Tab. 6), 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 ***R**elational **R**epresentation **D**istillation* (RRD), which relaxes rigid contrastive objectives by preserving meaningful *relative relationships* between instances in feature space. 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* (Fig. 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 [47].

Our main **contributions** are as follows:

- 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.
- We establish theoretical connections between our objective, InfoNCE [49], and the KL divergence, showing that InfoNCE arises as a limiting case when $\tau_t \to 0$.
- We empirically demonstrate the advantages of our objective across classification (Tabs. 1 and 2) and detection (Tab. 3) benchmarks, achieving consistent gains over existing methods. RRD achives 75.50% relative improvement [47] over conventional KD and 80.03% when combined with it (Tab. 1).
- We provide quantitative and qualitative analyses of learned representations through correlation alignment (Fig. 2 and Tab. 5), semantic similarity evaluation (Tab. 6), and retrieval-based assessment (Fig. 7 of the *supplementary*), confirming that RRD maintains relational topology between teacher and student embeddings.

2. Related Work

The seminal works of Bucilua *et al*. [3] and Hinton *et al*. [19] established the foundation of knowledge distillation, where compact student networks learn from large teacher models with minimal loss in generalization. Bucilua *et al*. [3] proposed matching model outputs, while Hinton *et al*. [19] 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*. [25] refined this formulation for better transferability. Since then, knowledge distillation has evolved into several branches, most notably *logit-based* [21, 36, 52] and *feature-based* [43, 47, 54, 56] distillation.

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 [52], multi-step training [36], or collaborative learning [57]. Subsequent work refined this process by adjusting how logits are represented or weighted [38, 46, 55, 60]. Further refinements involved dynamic temperature adjustment [27], transformation-based alignment [59], and adaptive teacher calibration [21].

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 [43], attention transfer [56], and feature transformation [22], later extended by functional consistency [30], class-level attention [15], and structural normalization [4, 8, 32]. 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 [54], distance preservation [39], or correlation congruence [40], while contrastive methods [47] maximized mutual information via memory banks. Recent studies introduced redundancy reduction and kernel-based alignment [17, 35]. 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.

Connection to information-theoretic objectives. Our objective is also related to *InfoNCE* [49] and *Noise-Contrastive Estimation (NCE)* [16], which maximize a lower bound on mutual information between representations [20]. While adversarial frameworks [13, 13] 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 (Fig. 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 [3, 19]. 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 [9, 37, 41, 44], it is normally a cross entropy loss. For object detection [5, 31], 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 [19] or ℓ_2 distance between feature maps [43]. 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 (see Sec. 9.1 of the *supplementary* for ablations on memory structure). 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 (see Sec. 8 of the *supplementary* for theoretical analysis).

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:

$$\mathbf{p}^T(\mathbf{x}_i; \theta_T, \mathcal{M}^+) = \left[p_1^T, \ldots, p_{K+1}^T\right] \tag{2}$$

where

$$p_i^T = \frac{\exp(\mathbf{z}_i^T \cdot \mathbf{m}_j / \tau_t)}{\sum_{m \sim \mathcal{M}^+} \exp(\mathbf{z}_i^T \cdot \mathbf{m} / \tau_t)}, \tag{3}$$

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:

$$\mathbf{p}^S(\mathbf{x}_i; \theta_S, \mathcal{M}^+) = \left[p_1^S, \ldots, p_{K+1}^S\right] \tag{4}$$

where

$$p_i^S = \frac{\exp(\mathbf{z}_i^S \cdot \mathbf{m}_j / \tau_s)}{\sum_{m \sim \mathcal{M}^+} \exp(\mathbf{z}_i^S \cdot \mathbf{m} / \tau_s)} \tag{5}$$

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 $\mathbf{x}_i$:

$$\hat{\theta}_S = \arg\min_{\theta_S} \sum_i^N D_{\mathrm{KL}}(\mathbf{p}^T(\mathbf{x}_i;\theta_T,\mathcal{M}^+) \parallel \mathbf{p}^S(\mathbf{x}_i;\theta_S,\mathcal{M}^+))$$
$$= \arg\min_{\theta_S} \sum_i^N H(\mathbf{p}^T(\mathbf{x}_i;\theta_T,\mathcal{M}^+), \mathbf{p}^S(\mathbf{x}_i;\theta_S,\mathcal{M}^+))$$
$$+ H(\mathbf{p}^T(\mathbf{x}_i;\theta_T,\mathcal{M}^+)), \quad \text{constant} \tag{6}$$

where D_{KL} denotes the KL divergence between $\mathbf{p}^T(\mathbf{x}_i;\theta_T,\mathcal{M}^+)$ and $\mathbf{p}^S(\mathbf{x}_i;\theta_S,\mathcal{M}^+)$. Here, $H(\mathbf{p}^T(\mathbf{x}_i;\theta_T,\mathcal{M}^+), \mathbf{p}^S(\mathbf{x}_i;\theta_S,\mathcal{M}^+))$ 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 $H(\mathbf{p}^T(\mathbf{x}_i;\theta_T,\mathcal{M}^+), \mathbf{p}^S(\mathbf{x}_i;\theta_S,\mathcal{M}^+))$:

$$\hat{\theta}_S = \arg\min_{\theta_S} \sum_{i=1}^N \mathcal{L}_{RRD}(\mathbf{x}_i,\theta_S,\theta_T,\mathcal{M}) \tag{7}$$

where

$$\mathcal{L}_{RRD} = -\mathbf{p}^T(\mathbf{x}_i;\theta_T,\mathcal{M}^+) \cdot \log \mathbf{p}^S(\mathbf{x}_i;\theta_S,\mathcal{M}^+)$$
$$= -\sum_{j=1}^{K+1} \frac{\exp(\mathbf{z}_i^T \cdot \mathbf{m}_j/\tau_t)}{\sum_{k=1}^{K+1}\exp(\mathbf{z}_i^T \cdot \mathbf{m}_k/\tau_t)} \log \frac{\exp(\mathbf{z}_i^S \cdot \mathbf{m}_j/\tau_s)}{\sum_{k=1}^{K+1}\exp(\mathbf{z}_i^S \cdot \mathbf{m}_k/\tau_s)}. \tag{8}$$

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 (Sec. 4.4 and Fig. 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.

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:

$$\mathcal{L}_{NCE} = \sum_i^N -\log \frac{\exp(\mathbf{z}_i^T \cdot \mathbf{z}_i^S/\tau_s)}{\sum_{\mathbf{m}\sim\mathcal{M}^+}\exp(\mathbf{z}_i^S \cdot \mathbf{m}/\tau_s)}, \tag{9}$$

which matches the *InfoNCE* loss [49]. 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 Sec. 8 of the *supplementary*.

Relation to *Kullback-Leibler* divergence. Hinton *et al.* [19] defined the knowledge distillation loss via the *Kullback–Leibler* (KL) divergence between the softened output distributions of the teacher and student networks:

$$\mathcal{L}_{\mathrm{KL}} = \sum_{i=1}^N \tau^2 D_{\mathrm{KL}}\Big(\sigma(y_i^T/\tau) \parallel \sigma(y_i^S/\tau)\Big)$$
$$= \sum_{i=1}^N \tau^2 \sum_{c=1}^C \sigma\left(\frac{y_{i,c}^T}{\tau}\right) \log \frac{\sigma\left(\frac{y_{i,c}^T}{\tau}\right)}{\sigma\left(\frac{y_{i,c}^S}{\tau}\right)} \tag{10}$$

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).

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:

$$\hat{\theta}_S = \arg\min_{\theta_S} \sum_i^N \Big(\mathcal{L}_{\text{sup}}(\mathbf{x}_i,\theta_S,y_i)$$
$$+ \lambda \cdot \mathcal{L}_{\mathrm{KL}}(\mathbf{x}_i,\theta_S,\theta_T) + \beta \cdot \mathcal{L}_{RRD}(\mathbf{x}_i,\theta_S,\theta_T,\mathcal{M})\Big) \tag{11}$$

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 [47] 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 (see Fig. 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 (Sec. 4.4 and Fig. 3a and Sec. 9 of the *supplementary*).

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

$$I(\mathbf{z}^T; \mathbf{z}^S) \leq H(\mathbf{p}^S(\mathbf{x}_i; \theta_S, \mathcal{M}^+)) - H(\mathbf{p}^T(\mathbf{x}_i; \theta_T, \mathcal{M}^+)), \tag{12}$$

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 (Tab. 5) and enhanced semantic organization (Tab. 6) 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 9 of the *supplementary*.

4.1. Experimental Setup

We evaluate the proposed framework on both image classification and object detection tasks using five standard benchmarks: CIFAR-100 [24], ImageNet ILSVRC-2012 [11], STL-10 [10], Tiny ImageNet [11], and MS-COCO [28]. Following prior work [47], 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 [47], while for detection we follow [8, 58]. 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 [34, 47]. 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 (Fig. 1 validated in Fig. 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 Sec. 7 of the *supplementary*.

4.2. Main Results

We first evaluate RRD on image classification (CIFAR-100, ImageNet) and object detection (MS-COCO). Extended comparison is provided in Sec. 10 of the *supplementary material*.

Results on CIFAR-100. Table 1 reports top-1 accuracy across both same architecture and cross (*i.e.*, different) architecture teacher–student pairs. RRD consistently outperforms classical KD [19] and contrastive-based CRD [47], demonstrating that preserving relative similarity distributions offers stronger guidance than strict instance discrimination. When combined with Hinton's KD [19], it achieves further gains—KD provides class-level supervision through logit-space KL divergence, while RRD enforces relational consistency in the feature space.

Table 1. **Distillation results on CIFAR-100.** Test top-1 accuracy (%) for different teacher–student pairs. The best and second-best results are **bolded** and underlined. Results for baselines are from [47]; ours are averaged over five runs. **Architecture abbreviations**: W: WideResNet, R: ResNet, MN: MobileNet, SN: ShuffleNet.

	Same architecture							Different architecture					
Teacher	W-40-2	W-40-2	R-56	R-110	R-110	R-32x4	VGG-13	VGG-13	R-50	R-50	R-32x4	R-32x4	W-40-2
Student	W-16-2	W-40-1	R-20	R-20	R-32	R-8x4	VGG-8	MN-v2	MN-v2	VGG-8	SN-v1	SN-v2	SN-v1
Teacher	75.61	75.61	72.34	74.31	74.31	79.42	74.64	74.64	79.34	79.34	79.42	79.42	75.61
Student	73.26	71.98	69.06	69.06	71.14	72.50	70.36	64.60	64.60	70.36	70.50	71.82	70.50
KD [19]	74.92	73.54	70.66	70.67	73.08	73.33	72.98	67.37	67.35	73.81	74.07	74.45	74.83
FitNet [43]	73.58	72.24	69.21	68.99	71.06	73.50	71.02	64.14	63.16	70.69	73.59	73.54	73.73
AT [56]	74.08	72.77	70.55	70.22	72.31	73.44	71.43	59.40	58.58	71.84	71.73	72.73	73.32
SP [48]	73.83	72.43	69.67	70.04	72.69	72.94	72.68	66.30	68.08	73.34	73.48	74.56	74.52
CC [40]	73.56	72.21	69.63	69.48	71.48	72.97	70.81	64.86	65.43	70.25	71.14	71.29	71.38
RKD [39]	73.35	72.22	69.61	69.25	71.82	71.90	71.48	64.52	64.43	71.50	72.28	73.21	72.21
FSP [54]	72.91	n/a	69.95	70.11	71.89	72.62	70.33	58.16	64.96	71.28	74.12	74.68	76.09
OFD [18]	75.24	74.33	70.38	n/a	73.23	74.95	73.95	69.48	69.04	n/a	75.98	76.82	75.85
CRD [47]	75.48	74.14	71.16	71.46	73.48	75.51	73.94	69.73	69.11	74.30	75.11	75.65	76.05
CRD+KD [47]	75.64	74.38	71.63	71.56	73.75	75.46	74.29	69.94	69.54	**74.97**	75.12	76.05	76.27
RRD (ours)	**75.85**	74.61	71.89	**71.92**	73.73	**75.77**	74.01	69.61	**70.11**	74.30	75.60	76.31	75.98
RRD+KD (ours)	75.67	**74.68**	**72.03**	71.75	**73.96**	75.53	**74.37**	**69.99**	69.65	74.53	**76.68**	**76.87**	**76.64**

Table 2. **Distillation results on ImageNet.** Student top-1 accuracy (%) on the ImageNet validation set under various teacher–student pairs. Baseline results are from [47]; ours are single-run evaluations.

	Teacher	*Student*	KD [19]	AT [56]	SP [48]	CC [40]	RKD [39]	CRD [47]	RRD	RRD+KD
ResNet-34→ResNet-18	73.31	69.75	70.67	71.03	70.62	69.96	70.40	71.17	**72.03**	71.99
ResNet-50→ResNet-18	76.16	69.75	71.29	71.18	71.08	n/a	n/a	71.25	**71.97**	71.88
ResNet-50→MobileNet-v2	76.16	69.63	70.49	70.18	n/a	n/a	68.50	69.07	71.54	**71.56**

Results on ImageNet. Table 2 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 [19], CRD [47], and their combination across all evaluated settings, even when applied independently, underscoring its robustness and broad applicability to real-world distillation scenarios.

Results on COCO. Table 3 extends our analysis to object detection, evaluating Faster R-CNN models trained with different distillation methods on MS-COCO [28]. 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 [8] and DKD [58], underscoring that relational structure consistently benefits spatially localized prediction tasks.

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 Sec. 10 of the *supplementary*. Additional visualizations (t-SNE, nearest-neighbor retrieval) are provided in Sec. 11 of the *supplementary*.

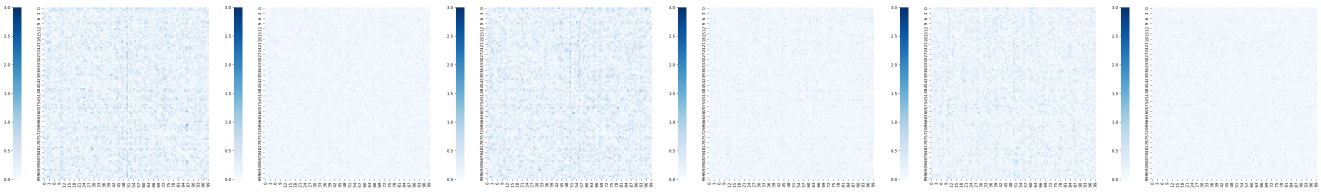
Visualization of inter-class correlations. Figure 2 compares the correlation matrix differences between teacher and student logits. Our objective achieves better alignment of correlation structures compared to models trained without distillation or with alternative methods [19, 47]. Standalone, it outperforms CRD [47], demonstrating stronger structural preservation. When combined with Hinton's KD [19], it further improves alignment. This structural coherence indicates that RRD helps students internalize the teacher's feature geometry rather than memorizing isolated logits.

Table 3. **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`. Baseline results follow [58]; ours are from a single run.

Method	ResNet-101 → ResNet-18			ResNet-101 → ResNet-50			ResNet-50 → MobileNet-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 [19]	33.97	54.66	36.62	38.35	59.41	41.71	30.13	50.28	31.35
FitNet [43]	34.13	54.16	36.71	38.76	59.62	41.80	30.20	49.80	31.69
ReviewKD [8]	36.75	56.72	34.00	**40.36**	60.97	**44.08**	33.71	53.15	36.13
DKD [58]	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**

Table 4. **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 [47]; ours are averaged over five runs.

	Teacher	*Student*	KD [19]	AT [56]	FitNet [43]	CRD [47]	CRD+KD [47]	RRD	RRD+KD
CIFAR-100→STL-10	68.6	69.7	70.9	70.7	70.3	71.6	**72.2**	72.0	72.0
CIFAR-100→Tiny ImageNet	31.5	33.7	33.9	34.2	33.5	**35.6**	35.5	35.5	35.2



(a) Vanilla Mean: 0.24, Max: 1.66 | (b) KD [19] Mean: 0.09, Max: 0.49 | (c) CRD [47] Mean: 0.23, Max: 1.56 | (d) CRD+KD [47] Mean: 0.10, Max: 0.57 | (e) RRD (ours) Mean: 0.18, Max: 0.99 | (f) RRD+KD (ours) 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.

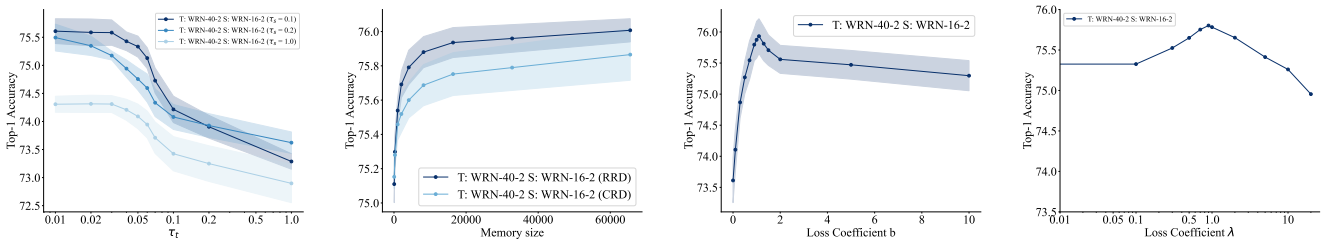
Transferability of representations. Table 4 evaluates a WRN-16-2 student distilled from a WRN-40-2 teacher as a frozen feature extractor. RRD achieves strong and consistent transfer performance on unseen datasets, indicating that its representations effectively encode relational semantics rather than overfitting to task-specific decision boundaries or local class structures.

Correlation matrix alignment. Table 5 quantifies how well student models preserve the relational structure of their teachers on CIFAR-100, using Frobenius distance [45], Pearson correlation [2], and structural similarity index (SSIM) [51]. RRD maintains strong structural alignment, and when combined with KD [19], 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.

Semantic structure preservation. Table 6 evaluates the semantic organization of learned features on CIFAR-100 through intra-class compactness, inter-class separation, normalized mutual information (NMI) [50], and retrieval precision measured by mean average precision at 5 (mAP@5) [58]. RRD improves semantic clustering compared to CRD [47], 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 further analyze the influence of individual design choices on CIFAR-100 using WRN-40-2 (teacher) and WRN-16-2 (student). Each experiment is repeated three times to ensure consistency and reliability; full ablations and additional analyses are provided in Sec. 9 of the *supplementary*.



(a) Temperatures τ_s and τ_t (b) Memory size K (c) Loss coefficient β (d) Loss coefficient λ

Figure 3. **Ablation study results on CIFAR-100 using WRN-40-2 as the teacher and WRN-16-2 as the student.** 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. Each configuration is averaged over three runs, and all curves are smoothed using Savitzky–Golay filtering for visual clarity.

Table 5. **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 [19]	<u>2.868</u>	<u>0.994</u>	<u>0.992</u>
CRD [47]	8.862	0.946	0.938
CRD+KD [47]	3.265	0.992	0.989
RRD (ours)	6.969	0.970	0.964
RRD+KD (ours)	**2.720**	**0.995**	**0.993**

Table 6. **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 [19]	<u>0.914</u>	<u>1.012</u>	<u>0.546</u>	<u>85.4</u>
CRD [47]	0.910	0.915	0.509	84.6
CRD+KD [47]	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**

Temperature parameter. 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, with a sharp drop when $\tau_t > \tau_s$, indicating that softer teacher distributions than student harm distillation and aligning with our information bottleneck interpretation. Excessively soft student distributions ($\tau_s = 1.0$) consistently underperform. Note that $\tau_t \to 0$ corresponds to the *argmax* operation, producing a one-hot target, while higher values of τ_t produce softer distributions that weaken teacher-student alignment.

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.

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 [47] 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 representations through a controlled *information bottleneck*. Extensive experiments across image classification, transfer learning, and object detection benchmarks demonstrate that RRD consistently outperforms both conventional and contrastive distillation approaches across diverse datasets and architectures. By emphasizing relative relationships rather than absolute feature matching, RRD preserves the structural and semantic integrity of learned representations, resulting in more robust, generalizable, and transferable models that faithfully capture the teacher's relational geometry while delivering superior downstream performance, enhanced stability, and improved representational consistency across tasks and scales.

Acknowledgments

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

[1] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning, 2019. 1, 2

[2] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009. 7

[3] 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. 1, 2, 3

[4] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier, 2022. 2

[5] 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, 2017. 3

[6] 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, 2017. 1

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017. 1

[8] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review, 2021. 2, 5, 6, 7

[9] 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. 3

[10] 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*, page 921–928, Madison, WI, USA, 2011. Omnipress. 5

[11] 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. 5

[12] 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. 1

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 3

[14] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning, 2019. 1

[15] Ziyao Guo, Haonan Yan, Hui Li, and Xiaodong Lin. Class attention transfer based knowledge distillation, 2023. 2

[16] 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. 1, 3

[17] Bobby He and Mete Ozay. Feature kernel distillation. In *International Conference on Learning Representations*, 2022. 2

[18] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation, 2019. 6

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 1, 2, 3, 4, 5, 6, 7, 8

[20] 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. 1, 3

[21] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher, 2022. 2

[22] 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. 2

[23] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better?, 2019. 1

[24] Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009. 5

[25] Jinyu Li, Rui Zhao, Jui-Ting Huang, and Yifan Gong. Learning small-size dnn with output-distribution-based criteria. In *Interspeech*, 2014. 2

[26] Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations, 2021. 2

[27] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation, 2022. 2

[28] 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. 5, 6

[29] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017. 1

[30] Dongyang Liu, Meina Kan, Shiguang Shan, and Xilin Chen. Function-consistent feature distillation, 2023. 2

[31] 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. 3

[32] Xiaolong Liu, Lujun Li, Chao Li, and Anbang Yao. Norm: Knowledge distillation via n-to-one representation matching, 2023. 2

[33] 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. 1

[34] Roy Miles and Krystian Mikolajczyk. Understanding the role of the projector in knowledge distillation, 2024. 5

[35] Roy Miles, Ismail Elezi, and Jiankang Deng. v_kd : improving knowledge distillation using orthogonal projections, 2024. 2

[36] Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant, 2019. 2

[37] Arun Mishra and Debbie Marr. Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. In *International Conference on Learning Representations*, 2017. 3

[38] Yulei Niu, Long Chen, Chang Zhou, and Hanwang Zhang. Respecting transfer gap in knowledge distillation, 2022. 2

[39] 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. 2, 6

[40] 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. 1, 2, 6

[41] Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. In *International Conference on Learning Representations*, 2018. 1, 3

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. 1

[43] 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. 1, 2, 3, 6, 7

[44] 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. 3

[45] Amit Singhal. Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin*, 24(4):35–43, 2001. 7

[46] Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation, 2024. 2

[47] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation, 2022. 1, 2, 5, 6, 7, 8

[48] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1365–1374, 2019. 6

[49] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 1, 2, 3, 4

[50] 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. 7

[51] 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. 7

[52] Lei Yang, Xiaohang Zhan, Dapeng Chen, Junjie Yan, Chen Change Loy, and Dahua Lin. Learning to cluster faces on an affinity graph, 2019. 2

[53] Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning, 2022. 2

[54] 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. 1, 2, 6

[55] Li Yuan, Francis E. H. Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization, 2021. 2

[56] 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. 1, 2, 6, 7

[57] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning, 2017. 2

[58] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation, 2022. 5, 6, 7

[59] Kaixiang Zheng and En-Hui Yang. Knowledge distillation based on transformed teacher matching, 2024. 2

[60] 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. 2