Relational Representation Distillation

Nikolaos Giakoumoglou
Imperial College London
London, UK, SW7 2AZ
n.giakoumoglou23@imperial.ac.uk

Tania Stathaki
Imperial College London
London, UK, SW7 2AZ
t.stathaki@imperial.ac.uk

Abstract

*Knowledge Distillation (KD) is an effective method for transferring knowledge from a large, well-trained teacher model to a smaller, more efficient student model. Despite its success, one of the main challenges in KD is ensuring the efficient transfer of complex knowledge while maintaining the student's computational efficiency. While contrastive learning methods typically push different instances apart and pull similar ones together, applying such constraints to KD can be too restrictive. Contrastive methods focus on instance-level information, but lack attention to relationships between different instances. We propose **R**elational **R**epresentation **D**istillation (RRD), which improves knowledge transfer by maintaining structural relationships between feature representations rather than enforcing strict instance-level matching. Specifically, our method employs sharpened distributions of pairwise similarities among different instances as a relation metric, which is utilized to match the feature embeddings of student and teacher models. Our approach demonstrates superior performance on CIFAR-100 and ImageNet ILSVRC-2012, outperforming traditional KD and sometimes even outperforms the teacher network when combined with KD. It also transfers successfully to other datasets like Tiny ImageNet and STL-10. Code is available at https://github.com/giakoumoglou/distillers.*

1. Introduction

Knowledge Distillation (KD) is a technique that facilitates the transfer of knowledge from a larger, well-trained model (teacher) to a smaller, more efficient model (student). This is achieved by minimizing the Kullback-Leibler (KL) divergence between their outputs, allowing the student model to approximate the performance of the teacher model while maintaining lower computational complexity. This process is particularly beneficial for deployment in resource-constrained environments. A critical aspect of KD is representation learning, which enables the student model to acquire meaningful feature representations that capture the underlying data distribution. Effective representation learning in KD can significantly boost the performance of the student model across various domains, such as natural language processing, computer vision, and speech recognition [20, 28, 42]. Despite these advantages, a major challenge in KD is the efficient transfer of complex knowledge from the teacher to the student model. Ensuring that the student model captures the abstract features and nuanced information present in the teacher model, without the need for similar computational capacity, remains a significant bottleneck.

Recent advancements in KD have increasingly focused on capturing richer, more nuanced knowledge from teacher models to improve the effectiveness and robustness of student models. Relational methods emphasize the importance of maintaining structural relationships between data points by minimizing distance-wise, angle-wise, or pairwise similarity losses [37, 47]. These approaches ensure that the student model learns relational information that mirrors the teacher's representational geometry, thereby enhancing performance on downstream tasks. On the other hand, probability-based methods and methods leveraging attention mechanisms aim to align the internal feature distributions or attention maps of the teacher and student models to ensure a more targeted knowledge transfer [38, 54]. Contrastive methods, particularly Contrastive Representation Distillation (CRD), have introduced the use of memory banks and contrastive learning objectives to better preserve the teacher's representation space [46]. By employing contrastive losses like InfoNCE [48], CRD maximizes the mutual information between teacher and student features while using a memory buffer of negative samples to encourage separation between dissimilar representations. This framework allows the student to achieve high fidelity to the teacher's representation. However, the explicit reliance on negative sampling can introduce constraints that may limit flexibility in certain scenarios. Additionally, some methods focus on directly aligning the prediction relations or outputs of teacher and student models using correlation-based losses or ℓ_2 loss functions [21, 52]. While these methods have achieved notable success, there is a growing interest in distillation techniques that balance the need for capturing relational structures without overly rigid constraints.

Our proposed method, ***R**elational **R**epresentation **D**istillation* (RRD), introduces a novel approach to address these challenges by maintaining relational consistency between the teacher and student models. By leveraging a large memory buffer of teacher samples to align their output distributions, our method ensures consistent relational structures, thereby enhancing the robustness and performance of the student model.

Our **contributions** are threefold:

1. We introduce Relational Representation Distillation (RRD), a KD method that leverages a memory buffer to align the similarity distributions between teacher and student outputs while maintaining structural relationships.
2. We validate the effectiveness of RRD through comprehensive testing on standard benchmarks, showcasing considerable gains in both accuracy and robustness. RRD surpasses other methods with a 23.68% relative improvement[1] over conventional KD. When integrated with KD, it demonstrates a 68.66% relative improvement over standard KD.
3. We provide qualitative evidence through t-SNE visualizations, demonstrating that RRD effectively preserves the spatial relationships in the embedding spaces of both student and teacher models,

The rest of this paper is organized as follows. Section 2 reviews related work in KD and self-supervised learning. Section 3 details our proposed methodology. Section 4 presents our experimental setup and results, and Section 5 concludes the paper.

2. Related Work

Self-supervised learning. Self-supervised learning has significantly impacted representation learning by leveraging unlabeled data. Various pretext tasks have been proposed to learn these representations, such as colorizing an image [56], predicting a missing patch [39], estimating the rotation angle [14], and solving jigsaw puzzles [35, 36]. Contrastive methods further extend this concept by pulling together an anchor and a "positive" sample in the embedding space while pushing apart the anchor from many "negative" samples [24], represented by frameworks such as [7–9, 17] and other variants [23, 51]. Differing from these, [58] introduces relational self-supervised learning, which explores the relationships between data points. These approaches have inspired various KD methods, including our proposed method, which adapts relational consistency from self-supervised learning to the KD framework.

Knowledge distillation. The work by [20] laid the foundation for knowledge distillation (KD) by introducing a method for transferring knowledge from larger teacher models to smaller student models, ensuring that the student retains strong generalization capabilities. This process uses temperature scaling in the teacher's softmax output to produce "soft targets" that are easier for the student to learn from. Many methods extend this approach by adding loss functions that complement standard cross-entropy and KL-divergence losses. Examples include utilizing intermediate representations or "hints" to guide student learning [42], aligning attention maps between teacher and student [54], and preserving relational information between samples [47].

Feature-based knowledge distillation. Feature-based KD methods focus on transferring knowledge through intermediate representations or feature alignments. These approaches often involve aligning the spatial and channel-wise attention maps of teacher and student models, enabling the student to better mimic the teacher's feature extraction [54]. Other techniques directly align the teacher and student's correlation structures [40], or employ variational inference to optimize knowledge transfer [1]. Additionally, relational approaches maintain structural relationships in the feature space, ensuring the student model captures essential relational knowledge [37]. Methods preserving internal dynamics across layers have also shown to be effective in sustaining feature-based information during KD [19, 38].

Contrastive knowledge distillation. Contrastive KD methods combine self-supervised contrastive learning with KD principles to improve knowledge transfer. A notable approach is CRD [46], which applies contrastive loss to maximize mutual information between teacher and student representations. CRD uses a memory buffer [50] to maintain a large set of negative samples, which enables effective contrastive learning and enhances the quality of knowledge transfer. These contrastive approaches aim to preserve the teacher's feature space, resulting in better student model performance in KD tasks.

Relational and structural knowledge distillation. Several KD techniques focus on preserving structural and relational information in the student model. Methods like relational KD capture relationships between data points, enabling the student model to learn the structural organization of the teacher's feature space [37]. Techniques like using orthogonal matrices for intra-batch feature similarity [32] and aligning pairwise feature kernels [15] further

[1] Average relative improvement is calculated as: $\frac{1}{N}\sum_{i=1}^{N}\frac{\text{Acc}^i_{\text{RRD}}-\text{Acc}^i_{\text{KD}}}{\text{Acc}^i_{\text{KD}}-\text{Acc}^i_{\text{van}}}$, where $\text{Acc}^i_{\text{RRD}}$, Acc^i_{KD}, and $\text{Acc}^i_{\text{van}}$ represent the accuracies of RRD, KD, and vanilla training of the i-th student model, respectively [46].

enhance relational knowledge transfer. This category also includes approaches that optimize relational information among outputs, such as preserving relations between predictions [21], aligning teacher and student features using ℓ_2 loss [52], and maintaining alignment in the student's prediction space. Cross-stage connections [6] and direct reuse of the teacher's classifier [3] contribute to better alignment in learned representations, while differentiable meta-learning optimizes layer-wise transfer [13].

Our method, Relational Representation Distillation (RRD), differentiates itself from state-of-the-art methods by focusing on maintaining relational consistency and alignment between the teacher and student models. Like many knowledge distillation techniques, RRD introduces an additional loss function to the training objective [2, 5, 21, 31, 32, 46, 52]. Similar to CRD [46], RRD employs a large memory buffer of teacher samples. However, unlike CRD, which uses instance discrimination, RRD relaxes this constraint to focus on aligning the relational structures of the output distributions. This approach diverges from traditional KD methods that often rely on direct alignment of logits or intermediate features. By leveraging pairwise similarities instead of explicit negative instances, RRD not only enhances the robustness and performance of the student model but also provides a more flexible and scalable solution for KD.

3. Methodology

This section presents our methodology to improve the efficiency and accuracy of KD. Our method, ***R**elational* ***R**epresentation* ***D**istillation* (RRD), improves KD efficiency by maintaining structural relationships between feature representations using a memory buffer of teacher samples to align their output distributions. Figure 1 shows an overview of the proposed RRD method.

3.1. Preliminaries

KD involves transferring knowledge from a high-capacity teacher neural network, denoted as f^T, to a more compact student neural network, f^S. Consider x_i as the input to these networks, typically an image. We represent the outputs at the penultimate layer (just before the final classification layer, or logits) as $\mathbf{z_i^T} = f^T(x_i)$ and $\mathbf{z_i^S} = f^S(x_i)$ for the teacher and student models, respectively. The primary objective of KD 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 mathematically expressed as:

$$\mathcal{L} = \mathcal{L}_{\text{sup}}(y_i, \mathbf{z_i^S}) + \lambda \cdot \mathcal{L}_{\text{distill}}(\mathbf{z_i^T}, \mathbf{z_i^S}) \tag{1}$$

where y_i represents the true label for the input x_i and λ is a hyperparameter that balances the supervised loss and the distillation loss. The loss $\mathcal{L}_{\text{sup}}$ is the alignment error between the network prediction and the annotation. For example, in the image classification task [10, 33, 41, 44], it is normally a cross-entropy loss. For object detection [4, 27], it includes bounding box regression as well. The loss $\mathcal{L}_{\text{distill}}$ is the mimic error of the student network towards a pre-trained teacher network, typically implemented as KL divergence between student and teacher outputs [20].

3.2. Relational Representation Distillation

Traditional contrastive learning uses instance discrimination and relies on $(K+1)$-softmax classification, where different instances are pushed apart, and matching instances are expected to have identical features. This can lead to the class collision problem where semantically similar instances are forced apart even when they should maintain some similarity. Applying such properties to KD imposes overly strict constraints, where a contrastive loss encourages the representations from the teacher and student models for the same input data to be similar, while simultaneously pushing apart representations from different data inputs:

$$\mathcal{L}_{\text{contrastive}}(\mathbf{z_i^T}, \mathbf{z_i^S}) = -\log \frac{\exp(\phi(\mathbf{z_i^T}, \mathbf{z_i^S})/\tau)}{\sum_{k=1}^{M} \exp(\phi(\mathbf{z_i^S}, \mathbf{z_k})/\tau)} \tag{2}$$

where ϕ is a similarity function, typically the dot product between ℓ_2 normalized embeddings $\phi(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T \cdot \mathbf{v}/||\mathbf{u}|| \cdot ||\mathbf{v}||$, τ is a temperature parameter, and M is the number of negative samples typically sampled from a memory buffer $Q = \{\mathbf{z_k}\}_{k=1}^{M}$.

Our method relaxes this contrastive objective by not enforcing strict instance-level matching. Instead of explicitly pushing different instances apart, we focus on preserving pairwise similarity relationships between instances. This relaxation allows the student more flexibility in learning from the teacher where different instances can maintain varying degrees of similarity rather than being strictly separated, and teacher-student pairs do not need to have identical features.

Specifically, to maintain structural relationships between instances, given an input image x_i, and the outputs $\mathbf{z_i^T} = f^T(x_i)$ and $\mathbf{z_i^S} = f^S(x_i)$ for the teacher and student models, respectively, we calculate the similarities between $\mathbf{z_i^T}$ and the j-th instance of the memory bank as $\phi(\mathbf{z_i^T}, \mathbf{z_j})$. A softmax layer can be adopted to process the calculated similarities, which then produces a similarity distribution:

$$\mathbf{p_i^T}(j) = \frac{\exp(\phi(\mathbf{z_i^T}, \mathbf{z_j})/\tau_t)}{\sum_{k=1}^{M} \exp(\phi(\mathbf{z_i^T}, \mathbf{z_k})/\tau_t)} \tag{3}$$

where τ_t is the temperature parameter for the teacher network. At the same time, we can calculate the relationship between $\mathbf{z_i^S}$ and the j-th instance of the memory bank as $\phi(\mathbf{z_i^S}, \mathbf{z_j})$. The resulting similarity distribution is:

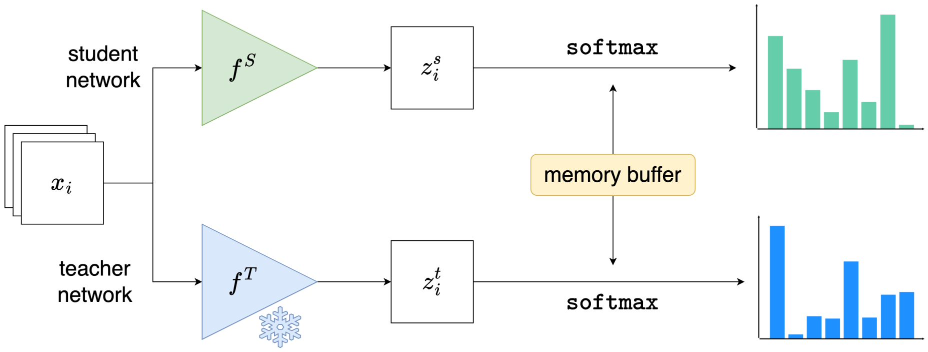


Figure 1. Overview of the RRD method. The student network processes input x_i to produce embeddings $\mathbf{z_i^S}$, while the teacher network, represented as frozen by a snowflake, processes the same input to generate embeddings $\mathbf{z_i^T}$. The embeddings are stored in a memory buffer to align their output distributions via softmax layers. The relationship between the teacher and student models is reinforced by leveraging the stored embeddings to guide the training of the student network.

$$\mathbf{p_i^S}(j) = \frac{\exp(\phi(\mathbf{z_i^S}, \mathbf{z_j})/\tau_s)}{\sum_{k=1}^{M} \exp(\phi(\mathbf{z_i^S}, \mathbf{z_k})/\tau_s)} \tag{4}$$

where τ_s is a different temperature parameter for the student network. We set $\tau_s < \tau_t$ to create sharpened distributions for the student model. That way the student's similarity distribution $\mathbf{p_i^S}$ becomes more peaked around the highest similarity values compared to the teacher's distribution $\mathbf{p_i^T}$. This sharpening helps the student focus on learning the most significant relationships while maintaining flexibility in capturing secondary similarities, thus avoiding the class collision problem that occurs with strict instance discrimination.

We align these distributions by minimizing their KL divergence, a relation metric to quantify the similarity between the teacher and student distributions:

$$\mathcal{L}_{\text{relational}}(\mathbf{z_i^T}, \mathbf{z_i^S}) = D_{\text{KL}}(\mathbf{p_i^T} \parallel \mathbf{p_i^S}) = H(\mathbf{p_i^T}, \mathbf{p_i^S}) - H(\mathbf{p_i^T}) \tag{5}$$

where D_{KL} denotes the KL divergence between $\mathbf{p_i^T}$ and $\mathbf{p_i^S}$. Since $\mathbf{p_i^T}$ will be used as a target, the gradient will be clipped here to avoid model collapse, thus we only minimize the cross-entropy term $H(\mathbf{p_i^T}, \mathbf{p_i^S})$ in our implementation.

The quality of the target similarity distribution $\mathbf{p_i^T}$ is crucial for reliable and stable training, which we achieve by maintaining a large memory buffer to store feature embeddings from teacher batches. The structural relationships between the teacher and student models are preserved by aligning the similarity distributions of their outputs using the KL divergence. To ensure the representations lie on a unit hypersphere, we normalize the outputs $\mathbf{z_i^T}$ and $\mathbf{z_i^S}$ before computing the loss. Furthermore, $\mathcal{L}_{\text{relational}}$ is computed by encoding $\mathbf{z_i^T}$ and $\mathbf{z_i^S}$ through a projection head that matches their dimensions, ensuring compatibility for comparison and alignment. This projection head also facilitates knowledge transfer by implicitly encoding relational information from previous samples [30].

The final objective function, which includes the supervised loss and standard KL divergence, is given by:

$$\mathcal{L} = \mathcal{L}_{\text{sup}}(y_i, \mathbf{z_i^S}) + \lambda \cdot \mathcal{L}_{\text{distill}}(\mathbf{z_i^T}, \mathbf{z_i^S}) + \beta \cdot \mathcal{L}_{\text{relational}}(\mathbf{z_i^T}, \mathbf{z_i^S}) \tag{6}$$

where β is a hyperparameter that balances the proposed loss $\mathcal{L}_{\text{relational}}$.

4. Experiments

We evaluate our RRD framework in the KD task of model compression of a large network to a smaller one [46]. This approach is consistent with the standard practice in the field, allowing for fair comparison across methods.

4.1. Experimental Setup

We implement DCD in PyTorch following the implementation of CRD [46].

Datasets. (1) CIFAR-100 [26] contains 50,000 training images with 500 images per class and 10,000 test images. (2) ImageNet ILSVRC-2012 [12] includes 1.2 million images from 1,000 classes for training and 50,000 for validation. (3) STL-10 [11] consists of a training set of 5,000 labeled images from 10 classes and 100,000 unlabeled images, and a test set of 8,000 images. (4) Tiny ImageNet [12] has 200 classes, each with 500 training images and 50 validation images.

Setup. We experiment on CIFAR-100 and ImageNet with student-teacher combinations of various capacity, such as ResNet [16] or Wide ResNet (WRN) [55], VGG [45], MobileNet [43], and ShuffleNet [29, 57] (more details about the network architectures are described in the supplementary material). We use $M = 16384$ samples (we ablate M in Section 4.7) and set the temperature parameter of the student to $\tau_s = 0.04$ and of the teacher to $\tau_t = 0.07$ (we also ablate both temperatures in Section 4.7). We set $\lambda = 1$ and $\beta = 1$ (we study the impact of β in Section 4.7). Both the student and teacher outputs are projected to a 128-dimensional space. We use a projection head of a single linear layer, followed by ℓ_2 normalization. We train for 240 epochs for CIFAR-100 and 120 for ImageNet. More details on the training details are described in the supplementary material.

Comparison. We compare our approach to the following state-of-the-art methods: (1) KD [20]; (2) FitNets [42]; (3) AT [54]; (4) SP [47]; (5) CC [40]; (6) VID [1]; (7) RKD [37]; (8) PKT [38]; (9) AB [19]; (10) FT [25]; (11) FSP [53]; (12) NST [22]; (13) CRD [46]; (14) OFD [18]; (15) WSLD [59]; (16) IPWD [34]. In the supplementary material, we include additional methods.

Role of the projection layer. Following [30], we use a projection head consisting of a single linear layer, followed by ℓ_2 normalization. The projection layer is not merely a dimension-matching tool but plays a crucial role in KD. It effectively transfers knowledge by implicitly encoding relational information from previous samples [30].

4.2. Results on CIFAR-100

Table 1 and Table 2 present the top-1 accuracies of student networks trained using different distillation techniques across various teacher-student architectural pairings. Table 1 examines pairings where both student and teacher models share similar architectural styles, while Table 2 focuses on cross-architecture distillations. Our proposed loss consistently outperforms the conventional KD technique. While the standalone performance of our method is comparable to CRD, its integration with KD not only achieves higher accuracies but in some cases, surpasses the performance of the teacher networks, such as in the distillation of WRN-40-2 to ShuffleNet-v1. The enhanced performance of our distillation method can be credited to multiple factors that collectively improve the transfer of knowledge from the teacher to the student model. Our approach uses a unique loss function that complements KD's primary focus on matching the softened output logits of the teacher and student. We introduce an additional layer of representational alignment that ensures not only the final outputs but also the intermediate feature representations of the student closely match those of the teacher. This dual focus allows the student model to mimic the teacher's outputs and develop more robust and generalizable internal representations.

4.3. Results on ImageNet

Table 3 presents the top-1 accuracies of student networks trained with various distillation techniques across different teacher-student architectural pairings. These findings affirm the scalability of our RRD method on large datasets like ImageNet, highlighting its ability to effectively distill complex models. Our approach achieves competitive results, surpassing KD across all tested architectures. Furthermore, RRD shows improvement across different architectures, demonstrating its effectiveness in various distillation scenarios. The combination of RRD with KD further improves results among the compared techniques in most cases.

4.4. Transferability of Representations

Our study investigates knowledge transfer from a larger teacher network (WRN-40-2) to a smaller student network (WRN-16-2), aiming to develop versatile representations for various tasks and datasets. We apply this technique by having the student network learn directly from CIFAR-100 or via distillation. The student network is employed as a static feature extractor for STL-10 and Tiny ImageNet images, adjusted to 32×32 resolution. We evaluate feature adaptability by training a linear classifier on the final feature layer for classifications with 10 categories (STL-10) and 200 categories (Tiny ImageNet). We document the impact of distillation approaches on feature transferability in Table 4. Our findings reveal that, except for FitNet, all distillation techniques significantly improve feature transferability on both datasets. Notably, while the teacher network achieves highest performance on CIFAR-100, its features show least effective transfer, likely due to training data specificity. Conversely, the student network using RRD and KD distillation not only equates to the teacher's CIFAR-100 performance but also surpasses it in transferability, showing improvements of 3.1% in STL-10 and Tiny ImageNet.

4.5. Capturing Inter-class Correlations

Cross-entropy loss overlooks the relationships among class logits in a teacher network, often resulting in less effective knowledge transfer. Distillation techniques that use ”soft targets”, such as those described by [20], have successfully captured these relationships, improving student model performance. Figure 2 assesses the effectiveness of different distillation methods on the CIFAR-100 KD task using WRN-40-2 as the teacher and WRN-40-1 as the student. We compare students trained without distillation, with attention transfer [54], with KL divergence [20], and with our proposed RRD method. Our findings show that RRD achieves close alignment between teacher and student logits, as evidenced by reduced differences in their correlation matrices. While RRD does not match CRD [46] in terms of exact correlation alignment, it significantly enhances learning efficiency and reduces error rates. The smaller discrepancies

Table 1. Test top-1 accuracy (%) of student networks on CIFAR-100, comparing students and teachers of the same architecture using various distillation methods. The values in bold indicate the maximum of each column and underlined values mark the second best.

Teacher Student	WRN-40-2 WRN-16-2	WRN-40-2 WRN-40-1	resnet-56 resnet-20	resnet-110 resnet-20	resnet-110 resnet-32	resnet-32x4 resnet-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 [20]	74.92	73.54	70.66	70.67	73.08	73.33	72.98
FitNet [42]	73.58	72.24	69.21	68.99	71.06	73.50	71.02
AT [54]	74.08	72.77	70.55	70.22	72.31	73.44	71.43
SP [47]	73.83	72.43	69.67	70.04	72.69	72.94	72.68
CC [40]	73.56	72.21	69.63	69.48	71.48	72.97	70.81
VID [1]	74.11	73.30	70.38	70.16	72.61	73.09	71.23
RKD [37]	73.35	72.22	69.61	69.25	71.82	71.90	71.48
PKT [38]	74.54	73.45	70.34	70.25	72.61	73.64	72.88
AB [19]	72.50	72.38	69.47	69.53	70.98	73.17	70.94
FT [25]	73.25	71.59	69.84	70.22	72.37	72.86	70.58
FSP [53]	72.91	n/a	69.95	70.11	71.89	72.62	70.33
NST [22]	73.68	72.24	69.60	69.53	71.96	73.30	71.53
CRD [46]	75.48	74.14	71.16	71.46	73.48	75.51	73.94
CRD+KD [46]	75.64	74.38	71.63	71.56	73.75	75.46	74.29
OFD [18]	75.24	74.33	70.38	n/a	73.23	74.95	73.95
WSLD [59]	n/a	73.74	71.53	n/a	73.36	74.79	n/a
IPWD [34]	n/a	74.64	71.32	n/a	**73.91**	**76.03**	n/a
RRD (ours)	75.33	73.55	70.71	70.72	73.10	74.48	73.99
RRD+KD (ours)	**75.66**	**74.67**	**72.19**	**71.74**	73.54	75.08	**74.32**

Table 2. Test top-1 accuracy (%) of student networks on CIFAR-100 involving students and teachers from different architectures, using various distillation methods. The values in bold indicate the maximum of each column and underlined values mark the second best.

Teacher Student	VGG-13 MobileNet-v2	ResNet-50 MobileNet-v2	ResNet-50 VGG-8	ResNet-32x4 ShuffleNet-v1	ResNet-32x4 ShuffleNet-v2	WRN-40-2 ShuffleNet-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 [20]	67.37	67.35	73.81	74.07	74.45	74.83
FitNet [42]	64.14	63.16	70.69	73.59	73.54	73.73
AT [54]	59.40	58.58	71.84	71.73	72.73	73.32
SP [47]	66.30	68.08	73.34	73.48	74.56	74.52
CC [40]	64.86	65.43	70.25	71.14	71.29	71.38
VID [1]	65.56	67.57	70.30	73.38	73.40	73.61
RKD [37]	64.52	64.43	71.50	72.28	73.21	72.21
PKT [38]	67.13	66.52	73.01	74.10	74.69	73.89
AB [19]	66.06	67.20	70.65	73.55	74.31	73.34
FT [25]	61.78	60.99	70.29	71.75	72.50	72.03
NST [22]	58.16	64.96	71.28	74.12	74.68	76.09
CRD [46]	69.73	69.11	74.30	75.11	75.65	76.05
CRD+KD [46]	69.94	69.54	**74.97**	75.12	76.05	76.27
OFD [18]	69.48	69.04	n/a	75.98	76.82	75.85
WSLD [59]	n/a	68.79	73.80	75.09	n/a	75.23
IPWD [34]	n/a	**70.25**	74.58	76.03	n/a	**76.44**
RRD (ours)	67.93	68.84	74.01	74.11	74.80	74.98
RRD+KD (ours)	**69.98**	69.13	74.26	**76.83**	**76.83**	**76.44**

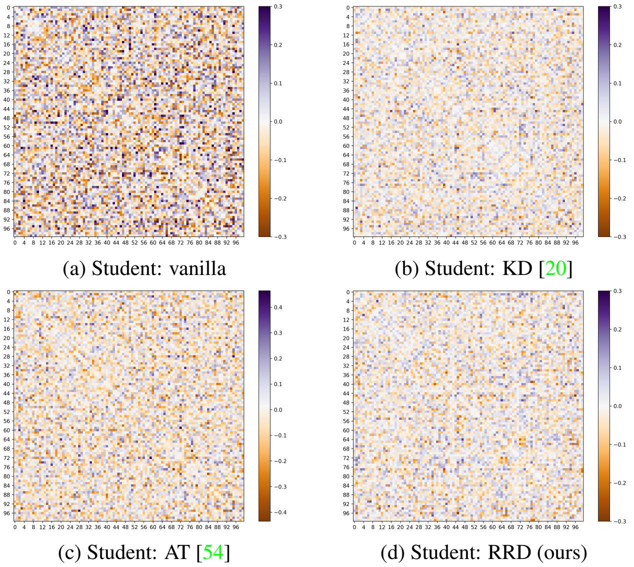
Table 3. Test top-1 (%) on ImageNet validation set using various distillation methods. The table compares students and teachers of the same and different architecture. The values in bold indicate the maximum of each column while underlined values mark the second best.

Teacher Student	ResNet-34 ResNet-18	ResNet-50 ResNet-18	ResNet-50 MobileNet
Teacher	73.31	76.16	76.16
Student	69.75	69.75	69.63
KD [20]	70.67	71.29	70.49
AT [54]	71.03	71.18	70.18
SP [47]	70.62	71.08	n/a
CC [40]	69.96	n/a	n/a
VID [1]	n/a	71.11	n/a
RKD [37]	70.40	n/a	68.50
AB [19]	n/a	n/a	68.89
FT [25]	n/a	n/a	69.88
FSP [53]	70.58	n/a	n/a
NST [22]	70.29	n/a	n/a
CRD [46]	71.17	71.25	69.07
OFD [18]	71.03	n/a	71.33
WSLD [59]	**72.04**	n/a	71.52
IPWD [34]	71.88	n/a	**72.65**
RRD (ours)	71.22	71.33	70.66
RRD+KD (ours)	71.40	**71.51**	71.83

between teacher and student logits indicate that the RRD objective captures a substantial portion of the correlation structure in the logits, resulting in lower error rates, though CRD achieves a slightly closer match.

4.6. Visualization of t-SNE Embeddings

We provide t-SNE [49] visualizations to compare the embeddings generated by various KD methods and the teacher network on the CIFAR-100 dataset. Figure 3 displays the embeddings from the teacher network, a WRN-40-2, and the student network, WRN-40-1, under standard training as well as distillation using AT and RRD where we limit the dataset to the first 10 classes of CIFAR-100 to offer a clearer understanding of the embedding space. We observe improved consistency in the embedding distributions between the teacher and student networks, indicating that RRD effectively transfers the knowledge of the teacher's feature space to the student. Our relational consistency approach ensures that the spatial relationships in the embedding spaces of both the student and teacher models are preserved. This alignment not only enhances the student's performance but also maintains the integrity of the feature representations learned by the teacher.



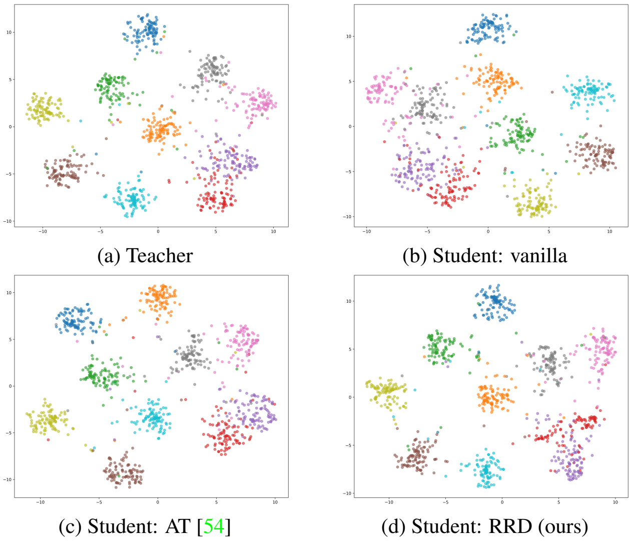
(a) Student: vanilla

(b) Student: KD [20]

(c) Student: AT [54]

(d) Student: RRD (ours)

Figure 2. Comparison of correlation matrix differences between teacher and student logits across various distillation methods on the CIFAR-100 task. Subfigures show results for (a) students trained without distillation, (b) with KL divergence [20], (c) with attention transfer (AT) [54], and (d) with our RRD method, highlighting better matching between student's and teacher's correlations. Results have been re-implemented according to [46].



(a) Teacher

(b) Student: vanilla

(c) Student: AT [54]

(d) Student: RRD (ours)

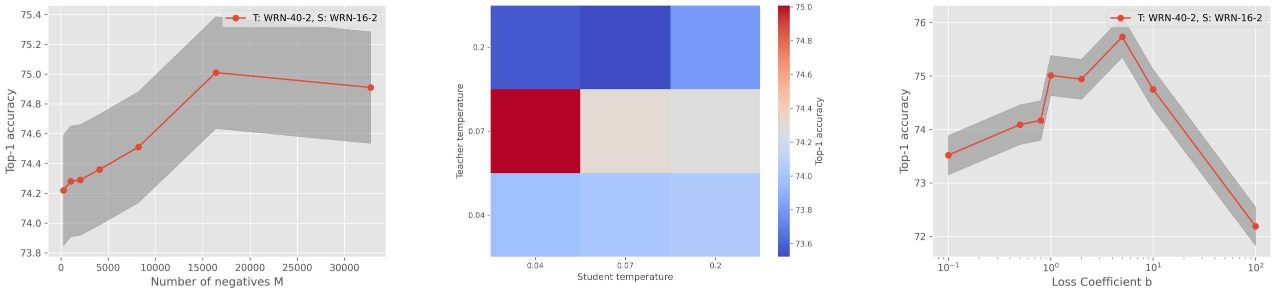
Figure 3. t-SNE visualizations of embeddings from the teacher network and student networks trained using different distillation techniques on the first 10 classes of the CIFAR-100 dataset: (a) teacher, (b) student trained without distillation, (c) student trained with attention transfer (AT) [54], and (d) student trained with RRD.

4.7. Ablation Study

There are two main hyperparameters in our objective: the number of negative samples M in the memory buffer $\mathcal{Q}$ of

Table 4. Test top-1 accuracy (%) of WRN-16-2 (student) distilled from WRN-40-2 (teacher). In this setup, the representations learned from the CIFAR-100 dataset are transferred to the STL-10 and Tiny ImageNet datasets. The network is frozen, and a linear classifier is trained on the last feature layer to perform classification with 10 classes (STL-10) or 200 classes (Tiny ImageNet). Results adapted from [46]. The values in bold indicate the maximum of each row.

	Teacher	*Student*	KD	AT	FitNet	CRD	CRD+KD	RRD	RRD+KD
CIFAR-100→STL-10	68.6	69.7	70.9	70.7	70.3	71.6	**72.2**	71.2	71.4
CIFAR-100→Tiny ImageNet	31.5	33.7	33.9	34.2	33.5	**35.6**	35.5	34.6	34.5



(a) Effect of varying M

(b) Effect of varying τ_s and τ_t

(c) Effect of varying loss coefficient β

Figure 4. Ablation study results on CIFAR-100 using WRN-40-2 as the teacher and WRN-16-2 as the student. (a) Effect of the number of negatives M on performance. Increasing M generally improves performance, but the difference becomes negligible beyond $M = 4096$. (b) Effect of temperature parameters τ_s and τ_t. Lower τ_s than τ_t results in better performance, with extremely high or low temperatures leading to sub-optimal solutions. (c) Effect of loss coefficient β on performance (logarithmic scale). Optimal β values range between 1 and 2.

Equation (3) and Equation (4); and temperature parameters τ_s and τ_t of Equation (3) and Equation (4) that modulate the softmax probability. We also ablate the hyperparameter β that balances the KD loss. For the ablation study, we adopt WRN-40-2 as the teacher and WRN-16-2 as the student. Experiments are conducted on CIFAR-100, and the results are shown in Figure 4.

Ablation on number of negatives M. We validated different values for M: 256, 1024, 2048, 4096, 8192, 16384, and 32768. As shown in Figure 4a, increasing M leads to improved performance. However, the difference in error rate between $M = 4096$ and $M = 16384$ is less 0.5%. Therefore, we use $M = 16384$ for reporting the accuracy, while in practice lower M should suffice. Going beyond $M = 16384$ proves to harms performance.

Ablation on temperatures τ_s and τ_t. We varied τ_s and τ_t, considering all the permutations of the numbers 0.04, 0.07, and 0.2 taken two at a time, considering that order matters. As Figure 4b illustrates, both extremely high or low temperatures lead to a sub-optimal solution. Also, as expected, a lower τ_s than τ_t leads to better performance. This improvement is attributed to the sharper predictions from the student model with a lower temperature, resulting in more confident and distinct class probabilities. This sharpness helps the student model align better with the teacher's guidance, enhance generalization, provide a more informative learning signal, and reduce overfitting. The sweep spot lies at $\tau_s = 0.04$ and $\tau_t = 0.07$ (red area of Figure 4b).

Ablation on loss coefficient β. We varied β from 0.1 to 100. As Figure 4c illustrates, both extremely high or low β lead to a sub-optimal solution. In general, β between 1 and 2 works well on CIFAR-100, as it balances the KD loss effectively without overwhelming other components of the training objective.

5. Conclusions

Our method advances KD by maintaining relational consistency between teacher and student models. RRD leverages a large memory buffer of teacher samples to align output distributions, ensuring consistent relational structures throughout learning. Unlike traditional approaches, RRD uses pairwise similarities and a relaxed contrastive loss to improve robustness and performance of the student model without explicit negatives. Extensive experiments on CIFAR-100, ImageNet, STL-10, and Tiny ImageNet show that RRD consistently outperforms state-of-the-art KD methods, demonstrating improved accuracy and transferability of representations.

References

[1] 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. 2, 5, 6, 7

[2] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration, 2021. 3

[3] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier, 2022. 3

[4] 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

[5] Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein contrastive representation distillation, 2021. 3

[6] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review, 2021. 3

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. 2

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020.

[9] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021. 2

[10] 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

[11] 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. 4

[12] 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. 4

[13] Xueqing Deng, Dawei Sun, Shawn Newsam, and Peng Wang. Distpro: Searching a fast knowledge distillation process via meta optimization, 2022. 3

[14] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations, 2018. 2

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

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 4

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020. 2

[18] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation, 2019. 5, 6, 7

[19] 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, 2019. 2, 5, 6, 7

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

[21] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *arXiv preprint arXiv:2205.10536*, 2022. 1, 3

[22] 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. 5, 6, 7

[23] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning, 2020. 2

[24] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021. 2

[25] 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. 5, 6, 7

[26] Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009. 4

[27] 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

[28] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2604–2613, 2019. 1

[29] 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. 4

[30] Roy Miles and Krystian Mikolajczyk. Understanding the role of the projector in knowledge distillation, 2024. 4, 5

[31] Roy Miles, Adrian Lopez Rodriguez, and Krystian Mikolajczyk. Information theoretic representation distillation, 2022. 3

[32] Roy Miles, Ismail Elezi, and Jiankang Deng. $v_k d$: improving knowledge distillation using orthogonal projections, 2024. 2, 3

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

[34] Yulei Niu, Long Chen, Chang Zhou, and Hanwang Zhang. Respecting transfer gap in knowledge distillation, 2022. 5, 6, 7

[35] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles, 2016. 2

[36] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. *CoRR*, abs/1805.00385, 2018. 2

[37] 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. 1, 2, 5, 6, 7

[38] 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. 1, 2, 5, 6

[39] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. *CoRR*, abs/1604.07379, 2016. 2

[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. 2, 5, 6, 7

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

[42] 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, 5, 6

[43] 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. 4

[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] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 4

[46] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[47] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1365–1374, 2019. 1, 2, 5, 6, 7

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

[49] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (Nov):2579–2605, 2008. 7

[50] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination, 2018. 2

[51] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised learning with swin transformers, 2021. 2

[52] Jing Yang, Brais Martinez, Adrian Bulat, and Georgios Tzimiropoulos. Knowledge distillation via softmax regression representation learning. In *International Conference on Learning Representations*, 2021. 1, 3

[53] 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. 5, 6, 7

[54] 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, 5, 6, 7

[55] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks, 2017. 4

[56] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization, 2016. 2

[57] 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. 4

[58] Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Ressl: Relational self-supervised learning with weak augmentation, 2021. 2

[59] 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. 5, 6, 7