Complementary Relation Contrastive Distillation

Jinguo Zhu[1†] Shixiang Tang[2] Dapeng Chen[3‡] Shijie Yu[4]
Yakun Liu[3] Mingzhe Rong[1] Aijun Yang[1] Xiaohua Wang[1‡]
[1]Xi'an Jiaotong University [2]The University of Sydney [3]Sensetime Group Limited
[4]Shenzhen Institutes of Advanced Technology, CAS
lechatelia@stu.xjtu.edu.cn tangshixiang@sensetime.com 384822707@qq.com
sj.Yu@siat.ac.cn liuyakun1@sensetime.com {mzrong, yangaijun, xhw}@mail.xjtu.edu.cn

Abstract

Knowledge distillation aims to transfer representation ability from a teacher model to a student model. Previous approaches focus on either individual representation distillation or inter-sample similarity preservation. While we argue that the inter-sample relation conveys abundant information and needs to be distilled in a more effective way. In this paper, we propose a novel knowledge distillation method, namely Complementary Relation Contrastive Distillation (CRCD), to transfer the structural knowledge from the teacher to the student. Specifically, we estimate the mutual relation in an anchor-based way and distill the anchor-student relation under the supervision of its corresponding anchor-teacher relation. To make it more robust, mutual relations are modeled by two complementary elements: the feature and its gradient. Furthermore, the low bound of mutual information between the anchor-teacher relation distribution and the anchor-student relation distribution is maximized via relation contrastive loss, which can distill both the sample representation and the inter-sample relations. Experiments on different benchmarks demonstrate the effectiveness of our proposed CRCD.

1. Introduction

Knowledge distillation aims to transfer the knowledge from one deep learning model (the teacher) to another (the student), such as distilling a large network into a smaller one [19, 49, 2, 48, 12] or ensembling a collection of models into a single model [29, 37, 27, 45]. It has a wide range of applications in the industry especially when a neural network needs to be efficiently deployed on devices with limited computational resources [9, 54, 38]. Although great progress has been achieved in the knowledge distillation regime, there is still no consensus on what kind of knowledge really needs to be preserved in the distillation [14].

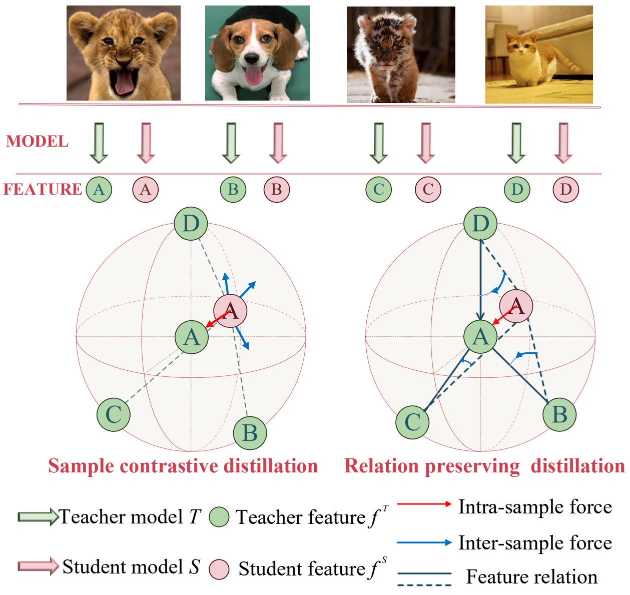


Figure 1: Sample contrastive distillation *vs.* Relation preserving distillation. Four neighboring samples and their corresponding features are displayed, and capital letters are used to identify them. While pulling f_A^S closer to f_A^T, sample contrastive distillation will simultaneously push f_A^S away from f_B^T, f_C^T and f_D^T without distinction, whereas relation preserving distillation preserves the feature relations across the feature space, thus f_A^S can be optimized along the optimal direction.

As one of the most effective distillation methods, CRD [41] holds the view that the representational knowledge is structured. So It tries to capture the correlations and higher-order output dependencies for each sample, which is different from the original KD objective introduced in [19] that treats all dimensions as independent information. CRD

[†]This work was done when Jinguo Zhu was an intern at SenseTime.
[‡]Corresponding authors.

leverages the family of contrastive objectives [20, 40, 46, 4] to maximize a lower-bound of the mutual information between the teacher and student representations. It essentially performs knowledge distillation based on the individual samples, enforcing the representation consistency between the teacher model and the student model.

However, neither CRD nor other sample-based distillation methods can effectively preserve inter-sample relations, which are more valuable than the sample representations themselves in many practical tasks, *e.g.,* retrieval and classification. As shown in Fig. 1, when using sample contrastive distillation methods, *e.g.,* CRD, the optimized forces from other neighbors just push the student representation of sample A away when contrasted negatively, which may not be optimal and can break the latent structural geometry of neighboring samples. Some recent works have shown that transferring the mutual similarity instead of actual representation is beneficial to student representation learning [43, 32, 34, 33]. These methods directly estimate the relations in teacher space by computing the inter-sample similarities, then mimic these similarities in the student space via L_2 loss or KL divergence, ignoring the high-order dependency within the representation in both relation estimation and knowledge distillation.

To robustly distill the structural knowledge of the teacher space, we define a new cross-space relation between two samples and supervise this new relation by its corresponding relation in the teacher representation space. More specifically, given the teacher and student representation of one sample, we select a neighboring sample's representation from the teacher representation space as an anchor. The anchor-student relation is encouraged to be consistent with the anchor-teacher relation. Our method brings at least three merits for distillation. (1) It simultaneously optimizes the representation and relation. When the anchor-student relation is pushed to be consistent with the anchor-teacher relation, the student representation is actually optimized along the optimal direction of representation learning. (2) The anchor-student relation is more effective for distillation compared with the student-student relation (where two representations are both from the student space) in the conventional KD family [43, 32, 34]. The student-student relation is unstable because the two representations in the student space are not well optimized and they will drift significantly during distillation, while the anchor representation within the anchor-student relation is fixed, which can effectively optimize the representation in the student space. (3) As the anchor can be randomly selected from the neighborhood of the considered sample, the student representation of one sample is supervised by multiple relations from different anchors, which guarantees the robustness of the distillation.

The representation relation is modeled by two complementary elements: the feature and its gradient. The feature relation reflects the structural information in representational space, and the gradient relation is computed by the feature gradients after backward propagation. As gradients measure the fastest rate and direction for loss minimization, gradient relation can explore the structural information of optimization kinetics in representational space [18, 39]. During the distillation, we maximize the mutual information between the anchor-teacher relation and the anchor-student relation for both two elements. The maximization problem can further surrogate to maximize the lower bound of mutual information which has been well solved by contrastive learning [46]. Our method is therefore denoted by Complementary Relation Contrastive Distillation(CRCD).

In summary, the main contributions of CRCD are three-fold. First, we define a new anchor-based cross space relation and adopt it to effectively and robustly distill both sample representations and inter-sample relations. Second, the new relation is modeled by two complementary elements, *i.e.*, the feature and its gradients, which capture the structure information of the feature and the optimization kinetics, respectively. Last, we maximize the low bound of mutual information between the anchor-teacher relation and the anchor-student relation and derive an efficient solution in the form of contrastive learning. Extensive experiments empirically validate the effectiveness of CRCD and further improve the current state-of-the-art in various benchmarks.

2. Related Work

Knowledge Distillation. There has been a rising interest in distilling knowledge from one model to another, in which the core issue is that what is the knowledge learned by a teacher and how to best distill the knowledge into a student. In [19], the soft probability distribution is transferred by using a higher temperature value. Compared to the one-hot label, soft targets can contain much more valuable information that defines a rich similarity structure over the data. Furthermore, not only the soft labels but also the hints from intermediate layers are used to train student networks in [35]. Moreover, the attention map [51] and the flow of solution procedure (FSP) [50] are used to transfer knowledge between networks. These works focus on distilling the knowledge modeled by learned presentations of samples themselves, however, ignore the mutual relations between samples, which contain rich structural information learned by the teacher.

There are a few recent works analyzing and exploiting the mutual relation between data samples [28, 34, 32, 33, 6, 5, 7, 25]. In particular, similarity-preserving knowledge [43] proposes to transfer the knowledge presented as similar activation between input pairs. In [34] and [32], the sample relations are modeled explicitly to transfer knowledge. However, these methods all use low-dimensional relation methods, such as cosine similarity [43] or gaus-

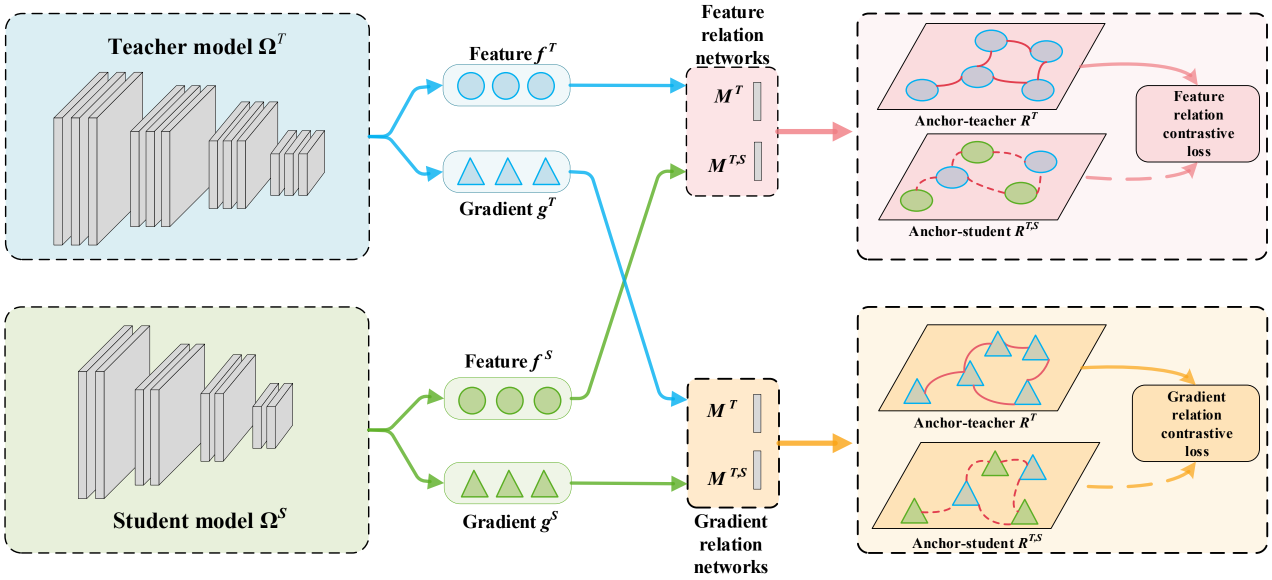


Figure 2: **The Flowchart of CRCD.** To distill the structural knowledge from the teacher model Ω^T to the student model Ω^S, two complementary elements, the feature f and its gradient g, are utilized to model the representation relations. For each element, auxiliary subnetworks M^T and $M^{T,S}$ are used to estimate the anchor-teacher relation R^T in the teacher space and anchor-student relation $R^{T,S}$ across space respectively. Meanwhile, the cross-space $R^{T,S}$ is supervised by its corresponding R^T. By this way, not only the relation estimation but also the representation learning can be achieved.

sian RBF [34] between features, to model the mutual relation, which may be suboptimal for modeling complex inter-sample interdependencies. Instead, in our paper, we design sub-networks to learn the high-dimensional across-space relations which can capture the complex mutual dependencies of deep representations from any two feature spaces.

Contrastive Learning. Contrastive Learning serves as the core idea of several recent works on self-supervised representation learning [8, 16, 30, 20, 15, 44, 42, 13]. Contrastive losses such as NCE [31, 20] measure the similarities of data samples in a deep representation space, which learn representation by contrasting positive and negative representation pairs. For knowledge distillation, CRD [41] is the first study that combines contrastive learning with knowledge distillation, which aims to maximize mutual information [3] between the teacher and student representations. Besides, SSKD [47] proposes to use contrastive tasks as self-supervised pretext tasks, which can facilitate the extraction of richer knowledge from the teacher to the student. From the usage of the contrastive loss, our method is more similar to CRD, but our objective is the mutual relations of deep representations, instead of the representations themselves.

3. Methodology

Fig. 1 presents the overall flowchart of our proposed CRCD. Given a teacher network Ω^T and a student network Ω^S, we denote the representation of an input x produced by the two networks as $\phi^T(x)$ and $\phi^S(x)$, respectively. Let x_i and x_j be two training samples randomly chosen from the sample set X. We denote the relation in the teacher space as $r^T_{i,j}$, where $r^T_{i,j}$ is a vector computed by a sub-network M^T that takes $\phi^T(x_i)$ and $\phi^T(x_j)$ as inputs. We further define a new relation $r^{T,S}_{i,j}$ computed by another sub-network $M^{T,S}$. It is noteworthy that the inputs $\phi^T(x_i)$ and $\phi^S(x_j)$ for $M^{T,S}$ are from different spaces. Regarding $\phi^T(x_i)$ as an anchor representation, the cross-space anchor-student relation $r^{T,S}_{i,j}$ is expected to be consistent with the teacher-space anchor-teacher relation $r^T_{i,j}$, which not only preserves the relation between x_i and x_j, but also drives the $\phi^S(x_j)$ to be consistent with $\phi^T(x_j)$.

In the following sub-sections, we first demonstrate how to use contrastive learning to perform the relation distillation, then two complementary elements are introduced to model the representation relations, and the implementation details and some discussions will be presented at last. The complete mathematical derivation refers to the supplementary materials.

3.1. Relation Contrastive Distillation

Assume that we are given a set of training examples with empirical data distribution $p(X)$, the sampling procedure for the conditional marginal distributions $p(R^T|X)$, $p(R^{T,S}|X)$ are modeled as

$$\begin{aligned} x_i, x_j \sim p(X), &\quad r^T_{i,j} = M^T(\phi^T(x_i), \phi^T(x_j)), \\ x_m, x_n \sim p(X), &\quad r^{T,S}_{m,n} = M^{T,S}(\phi^T(x_m), \phi^S(x_n)) \end{aligned} \tag{1}$$

respectively. While the sampling procedure of the conditional joint distribution $p(R^T, R^{T,S}|X)$ is modeled as:

$$\begin{aligned} x_i, x_j \sim p(X), \quad r^T_{i,j} &= M^T(\phi^T(x_i), \phi^T(x_j)), \\ r^{T,S}_{i,j} &= M^{T,S}(\phi^T(x_i), \phi^S(x_j)). \end{aligned} \tag{2}$$

For ease of notation, we utilize $p(R^T)$, $p(R^{T,S})$ and $p(R^T, R^{T,S})$ to briefly represent $p(R^T|X)$, $p(R^{T,S}|X)$ and $p(R^T, R^{T,S}|X)$. Intuitively, we aim to maximize the mutual information (MI) of the two relation distributions from R^T and $R^{T,S}$, which is

$$I(R^T, R^{T,S}) = \mathbb{E}_{p(R^T, R^{T,S})} \log \frac{p(R^T, R^{T,S})}{p(R^T)p(R^{T,S})}. \tag{3}$$

MI Lower Bound. To derive a solvable loss function, we define a distribution q with latent variable C which indicates whether the relation tuple $(r^T, r^{T,S})$ is drawn from the joint distribution or the product of marginal distributions:

$$\begin{aligned} q(R^T, R^{T,S}|C = 1) = p(R^T, R^{T,S}) \\ q(R^T, R^{T,S}|C = 0) = p(R^T)p(R^{T,S}). \end{aligned} \tag{4}$$

More specifically, $C=1$ means r^T and $r^{T,S}$ are computed based on the same input pair as in Eq. 2, and $C=0$ means r^T and $r^{T,S}$ are independently selected as in Eq. 1. In our data, we provide 1 relevant relation pair ($C = 1$) with N irrelevant relation pair ($C=0$). Then the prior $q(C = 1) = 1/(N + 1)$ and $q(C = 0) = N/(N + 1)$. Combing the priors with the Bayes' rule, the posterior for $C = 1$ is given by:

$$q(C=1|R^T, R^{T,S}) = \frac{p(R^T, R^{T,S})}{p(R^T, R^{T,S}) + Np(R^T)p(R^{T,S})}. \tag{5}$$

By connection to the mutual information, the posterior $\log q(C=1|R^T, R^{T,S}) \leq -\log(N) + \log\left(\frac{p(R^T, R^{T,S})}{p(R^T)p(R^{T,S})}\right)$. Taking the expectation on both sides w.r.t. $p(R^T, R^{T,S})$, which is also equivalent to $q(R^T, R^{T,S}|C = 1)$, we have:

$$\begin{aligned} I(R^T, R^{T,S}) \geq \log(N) + \\ \mathbb{E}_{q(R^T, R^{T,S}|C=1)} \log q(C=1|R^T, R^{T,S}) \end{aligned} \tag{6}$$

where $\log(N)+\mathbb{E}_{q(R^T, R^{T,S}|C=1)} \log q(C=1|R^T, R^{T,S})$ is a lower bound of the mutual information.

Distribution Approximation. As there is no knowledge about the true distribution of $q(C = 1|R^T, R^{T,S})$, we approximate the distribution by fitting a parameterized model h: $\{R^T, R^{T,S}\} \rightarrow [0, 1]$ with the samples from $q(C=1|R^T, R^{T,S})$. The log-likelihood of the sampled data under this model is defined as:

$$\begin{aligned} \mathcal{I}(h) = \mathbb{E}_{q(R^T, R^{T,S}|C=1)}[\log h(R^T, R^{T,S})] \\ + N\mathbb{E}_{q(R^T, R^{T,S}|C=0)}[\log(1 - h(R^T, R^{T,S}))]. \end{aligned} \tag{7}$$

To achieve a good approximation to $q(C = 1|R^T, R^{T,S})$, we need to maximize the log likelihood. Consider the bound in Eq. 6 and the fact that $N\mathbb{E}_{q(R^T, R^{T,S}|C=0)}[\log(1 - h(R^T, R^{T,S}))]$ is non-positive, we have

$$\begin{aligned} I(R^T, R^{T,S}) &\geq \log N + \mathbb{E}_{q(R^T, R^{T,S}|C=1)}[\log h(R^T, R^{T,S})] \\ &\quad + N\mathbb{E}_{q(R^T, R^{T,S}|C=0)}[\log(1 - h(R^T, R^{T,S}))] \\ &\geq \log N + \mathcal{I}(h), \end{aligned} \tag{8}$$

where $\log N + \mathcal{I}(h)$ is the lower bound of the mutual information with the parameterized model h. The maximization of the log-likelihood is also to maximize the lower bound.

Relation Contrastive Loss. In our method, the inputs for the function h are teacher-space relation r^T and cross-space relations $r^{T,S}$, which are the results of the teacher Ω^T, the student Ω^S, and the two sub-networks M^T, $M^{T,S}$. Except the teacher Ω^T, the other three networks Ω^S, M^T and $M^{T,S}$ also need to be optimized during the distillation. We aim to maximize the mutual information, which is equivalent to minimizing the relation contrastive loss $\mathcal{L}_{RC}$:

$$\begin{aligned} \mathcal{L}_{RC}(h, \Omega^S, M^T, M^{T,S}) = &- \sum_{q(C=1)} \log h(r^T, r^{T,S}) \\ &- N \sum_{q(C=0)} \log[1 - h(r^T, r^{T,S})] \end{aligned} \tag{9}$$

where $\{(r^T, r^{T,S})|C = 1\}$ act as positive pairs while $\{(r^T, r^{T,S})|C = 0\}$ act as negative pairs. Due to Eq. 8, the contrastive loss can fit the distribution $q(C|R^T, R^{T,S})$ to increase the lower-bound of mutual information of R^T and $R^{T,S}$, by which not only the parameterized model h, but also the other three networks Ω^S, M^T and $M^{T,S}$ can be jointly optimized.

3.2. Complementary Relation

Modeling relation between sample representations is the prerequisite for distilling the structural information. We therefore propose two learnable sub-networks $M^{T,S}$ and M^T to estimate the relation.

The sub-network $M^{T,S}$ is to compute the anchor-student relation with representation $\phi^T(x_i)$ and $\phi^S(x_j)$:

$$\begin{aligned} r^{T,S}_{i,j} &= M^{T,S}(\phi^T(x_i), \phi^S(x_j)) \\ &= W^A(\sigma(W^A_i \phi^T(x_i) - W^A_j \phi^S(x_j))), \end{aligned} \tag{10}$$

where W^A_i and W^A_j are linear transformations that can solve the dimension mismatch problem. σ is ReLU function and W^A is used for transformation. The anchor-student relation is supervised by the fixed anchor-teacher relation $r^T(x_i, x_j)$, computed by another sub-network M^T:

$$\begin{aligned} r^T_{i,j} &= M^T(\phi^T(x_i), \phi^T(x_j)) \\ &= W^B(\sigma(W^B_i \phi^T(x_i) - W^B_j \phi^T(x_j))). \end{aligned} \tag{11}$$

It is noteworthy that the relations $r^{T,S}$ and r^{T} are not scalar values but high-dimensional vectors. We claim that the high-dimensional relation can more accurately capture the structural information of deep representations than low-dimensional relation *e.g.,* cosine similarity, which will be validated in section 4.2. Furthermore, the small learnable networks also increases relation flexibility.

The relations are modeled by two complementary elements: *feature* f and its *gradient* g. Specifically, the representation $\phi(x)$ in Eq. 10 and Eq. 11 can be either the feature of the teacher/student model or its gradient.

Feature Element. The feature element is the ℓ_2 normalized output of teacher/student's backbone. With the feature element f, the representations $\phi^T(x)$ and $\phi^S(x)$ reflect the direct activation relative to the input x:

$$\phi^T(x) = f^T(x); \quad \phi^S(x) = f^S(x) \tag{12}$$

Gradient Element. The gradient element is the gradient with respect to the feature. It reflects the optimization kinetics in the feature space, encoding important structural information. Given an input sample x into a teacher/student network Ω, the gradient of task loss $\mathcal{L}_{cls}$ relative to the feature f is computed as:

$$g(x) = \frac{\partial}{\partial f}\mathcal{L}_{cls}(\Omega, x). \tag{13}$$

With gradient elements, the representation $\phi^T(x)$ and $\phi^S(x)$ can reflect the optimization kinetics:

$$\phi^T(x) = g^T(x); \quad \phi^S(x) = g^S(x) \tag{14}$$

Element Combination. *Complementary relation* is modeled to leverage feature and gradient elements simultaneously. Specifically, after the one-sided relations: *feature relation* r^f and *gradient relation* r^g , are computed with feature and gradient elements respectively, their corresponding relation contrastive losses can also be calculated by Eq. 9. By optimizing these two losses simultaneously, these two elements can both be utilized.

3.3. Implementation

Critic Function. We specify the parameterized critic function h in Eq. 7 to distinguish whether the relation pair $(r^T, r^{T,S})$ is sampled from the joint distribution $p(R^T, R^{T,S})$ or the product of marginal distribution $p(R^T)p(R^{T,S})$. The formulation is similar to NCE [46]:

$$h(r^T, r^{T,S}) = \frac{e^{h_1(r^T)h_2(r^{T,S})/\tau}}{e^{1/\tau}} \tag{15}$$

where τ is a temperature hyperparameter, and h_1 and h_2 first perform the linear transformation on relations, then normalize the transformed relations with ℓ_2 norm.

Sampling Policy. We adopt the following sampling policy: in each forward-propagation, the anchor relation r^T_{ij} and positive relation $r^{T,S}_{ij}$ are calculated using representations from any two samples x_i and x_j in the current mini-batch, while the negative relations $r^{T,S}_{ik}$ are calculated using the anchor representation from x_i and the representations (indexed with k) sampled from the buffer where features and gradients are stored. Considering a B-size min-batch, we construct the anchor/positive relation for each sample pair thus the number of these two relations can be B^2. For each anchor relation, we sample N feature/gradient from the buffer to construct N negatives for contrastive learning.

To make the feature/gradient buffer reflect the current network state better, we propose a queuing sampling method instead of a randomly sampling strategy. The queue records the N sample indices from the immediate preceding mini-batches and is updated after each forward-propagation by replacing the oldest indices with the current mini-batch. According to these recorded indices, the representations of these samples are used to calculated relation contrastive loss, whose effectiveness will be studied in Sec. 4.2.

Loss Function. To achieve the superior performance and conduct a fair comparison with other methods, we also incorporate the naive knowledge distillation loss $\mathcal{L}_{kd}$ [19] along with our relation contrastive loss. Given the pre-softmax logits z^T and z^S for teacher and student, the naive KD loss can be expressed as

$$\mathcal{L}_{kd} = \rho^2\mathcal{H}(\sigma(z^T/\rho), \sigma(z^S/\rho)) \tag{16}$$

where ρ is the temperature, $\mathcal{H}$ refers to the cross-entropy and σ is softmax function. The complete objective is:

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha\mathcal{L}_{KD} + \beta_1\mathcal{L}^f_{RC} + \beta_2\mathcal{L}^g_{RC} \tag{17}$$

where $\mathcal{L}^f_{RC}$ and $\mathcal{L}^g_{RC}$ are the relation contrastive loss computed with the feature (f) and gradient (g), respectively. $\mathcal{L}_{cls}$ is the cross entropy loss for classification. We set hyper-parameters to $\alpha = 1$ and $\beta_1 = \beta_2 = 0.5$ empirically.

Discussion. CRD [41] aims to maximize the mutual information between the representations of the sample themselves from teacher/student models. Meanwhile, the proposed CRCD seeks the consistency between the teacher-space relation and cross-space relation. Indeed, if $i = j$ in Eq. 9, the loss of CRCD essentially optimizes the cross-space relation of one sample, which degrades to the loss of CRD. Moreover, the number of pair-wise relations is at quadratic level relative to the number of samples, which also increases the optimized stability of contrastive loss.

4. Experiments

4.1. Datasets and Experimental Setup

Datasets. Our experiments are conducted on two widely used classification datasets, *i.e.*, CIFAR100 [24] and ImageNet [10]. CIFAR100 contains 60000 images for 100 classes, and there are 500 and 100 images per class for

Table 1: Testing accuracy (%) on CIFAR100 with different relation modeling methods. $\mathcal{L}_2$ loss and relation contrastive loss $\mathcal{L}_{RC}$ are used to distill the feature relation r^f.

teacher student	resnet56 resnet20	resnet110 resnet20	ResNet50 vgg8
RKD [32]	70.54	70.98	73.65
CC [34]	71.42	70.96	73.76
SP [43]	71.59	71.15	73.95
PKT [33]	71.68	71.08	74.01
$r^f + \mathcal{L}_2$	71.93	71.54	74.15
$r^f + \mathcal{L}_{RC}$	72.70	72.02	74.69

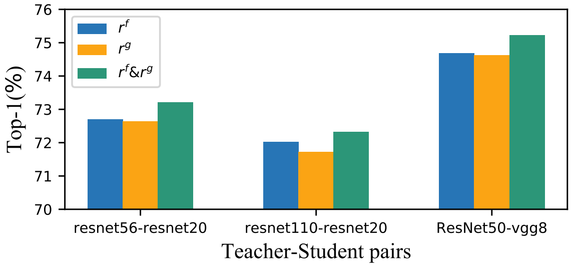


Figure 3: Accuracy of different relation elements. The feature relation r^f, gradident relation r^g, and complementary relation $r^f \& r^g$ are distilled on three teacher-student pairs.

training and testing respectively. ImageNet is a well-known large-scale image classification benchmark with 1000 classes, consisting of 1281167 images for training and 50000 images for testing.

Parameter Setting. For CIFAR, mini-batch size is set to 64 in 1 GPU. SGD optimizer is used with weight decay and momentum of 0.0001 and 0.9 respectively. And the learning rate and schedule strategy follow [41], which is included in supplementary materials. For ImageNet, batchsize is set to 256 in 8 GPUs, and the standard training settings for ImageNet is adopted. For other competing methods, we use the implementation settings in papers or official shared codes. The relation dimension computed by sub-networks $M^{T,S}$ and M^T is set to 256-d since the representation dimension in most of our experimental networks is 256-d.

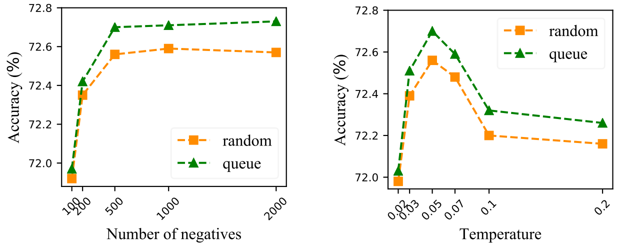
4.2. Ablation Study

Three teacher-student pairs are selected for ablation study. Their model names and top-1 accuracy (%) when trained individually on CIFAR100 are shown below:

teacher	resnet56 73.25	resnet110 73.89	ResNet50 79.04
student	resnet20 69.06	resnet20 69.06	vgg8 70.71

Table 2: Testing accuracy (%) on CIFAR100 with different transformations for critic function h. IM: identity mapping; LP: linear projection; NP: nonlinear projection. The transformation dimensions are appended as subscripts.

teacher student	resnet56 resnet20	resnet110 resnet20	ResNet50 vgg8
IM	72.35	71.84	74.25
NP_{256}	72.52	71.98	74.49
LP_{64}	72.45	71.92	74.34
LP_{128}	**72.70**	72.02	**74.69**
LP_{256}	72.65	**72.12**	74.57



(a) Effects of varying N

(b) Effects of varying τ

Figure 4: Accuracy of varying negative number N and temperature τ with different sample policies.

The first two are with similar architectures, while the last one is with a very different architecture. These experiments are conducted on CIFAR100, and results are averaged over 3 runs.

Effectiveness of relation modeling method. We first demonstrate the effectiveness of anchor-based relation. In contrast to conventional modeling methods, our relation is cross-space and high-dimensional. To verify its superiority, we compare it with four methods using low-dimensional relations: 1) RKD [32]; 2) CC [34]; 3) SP [43]; and 4) PKT [33]. For a fair comparison, we also use $\mathcal{L}_2$ loss to preserve representation relations and only feature relation is involved. The results are shown in Tab. 1. Over all three teacher-student pairs, our proposed relation boosts the test accuracy by a large margin even with $\mathcal{L}_2$ loss, which means that our relation modelling method is superior.

Effectiveness of complementary relation elements. We propose two elements: feature and its gradient, to model representation relation. To verify their complementarity, we test the distilling accuracy of these two elements when used alone and when used simultaneously. As Fig. 3 shows, their combination can get the best result, which indicates that the feature and the gradient are complementary to each other and can more comprehensively present the representation interdependences.

Table 3: Contrastive loss functions. To simplify, the anchor relation r_{ij}^{T}, positive relation $r_{ij}^{T,S}$, and negative relation $r_{ik_{j\neq k}}^{T,S}$ after critic transformation are denoted as u, v^{+} and v^{-} respectively. All relations are ℓ_2 normalized before inner product. τ is the temperature weight, and m is the margin parameter. Additionally, σ is *sigmoid* function.

Name	Loss function
$\mathcal{L}_{MT}$[36]	$max(u^{'}v^{-} - u^{'}v^{+} + m, 0)$
$\mathcal{L}_{CL}$ [26]	$-\log\sigma(u^{'}v^{+}/\tau) - \log(1 - \sigma(u^{'}v^{-}/\tau))$
$\mathcal{L}_{NCE}$ [31]	$-u^{'}v^{+}/\tau + \log\sum e^{u^{'}v^{-}/\tau}$
$\mathcal{L}_{RC}$	$-\log\frac{e^{u^{'}v^{+}/\tau}}{e^{1/\tau}} - N\sum\log(1 - \frac{e^{u^{'}v^{-}/\tau}}{e^{1/\tau}})$

Table 4: Testing accuracy (%) on CIFAR100 with different contrastive loss functions.

teacher		resnet56	resnet110	ResNet50
student		resnet20	resnet20	vgg8
$\mathcal{L}_2$		71.93	71.54	73.89
$\mathcal{L}_{MT}$	$m = 0.4$	72.21	71.83	74.25
$\mathcal{L}_{CL}$	$\tau = 0.05$	72.15	71.72	74.07
$\mathcal{L}_{NCE}$	$\tau = 0.05$	72.53	**72.09**	74.44
$\mathcal{L}_{RC}$	$\tau = 0.05$	**72.70**	72.02	**74.69**

Effectiveness of critic function h. We propose the critic function h in Eq. 15 to estimate the distribution $q(C = 1|R^{T}, R^{T,S})$. To investigate the effectiveness of h_1 and h_2 selection, we conduct three experiments, including specifying the h_1 and h_2 functions with identity mapping, nonlinear projection and linear transformation(default). In particular, h is degraded to cosine similarity estimation when identity mapping is adopted. For nonlinear projection, we use a MLP with one hidden layer $h(r) = W^{(2)}\sigma(W^{(1)}r)$ where r is input relation and σ is a ReLU nonlinearity.

In this study, the output dimension of linear or nonlinear transformation are both 256. Table 2 shows testing results using different transformations. We observe that both the linear and nonlinear projection achieve better results than identity mapping under the same projection dimension, which means that critic function with learnable parameters can better fit the distribution $q(C = 1|R^{T}, R^{T,S})$.

Effectiveness of relation contrastive loss. We compare our relation contrastive loss $\mathcal{L}_{RC}$ with other commonly used contrastive loss, such as triplet loss with margin ($\mathcal{L}_{MT}$) [36] and contrastive logistic loss ($\mathcal{L}_{CL}$) [26, 11]. Tab. 3 shows the formulations of four contrastive loss function.

To better analyze the loss function, we only use the feature element and gradient is not employed. The hyperparameters in these losses, *i.e.,* temperature τ and margin m, are tuned to achieve the best results. Results reported in Tab. 4 show that, $\mathcal{L}_{NCE}$ and $\mathcal{L}_{RC}$ can significantly outperform $\mathcal{L}_{MT}$ and $\mathcal{L}_{CL}$, because they can benefit from large number of negative samples. While our objective function $\mathcal{L}_{RC}$ is better than $\mathcal{L}_{NCE}$ in most of teacher-student combinations.

4.3. Hyper-parameter analysis

Several hyper-parameters are worth investigating in our proposed CRCD method. (1) The number of negative samples N; (2) The temperature used to scale the critic scores in Eq. 15; (3) The sampling policy to construct negative relations; (4) The projection dimension of critic function h. We adopt resnet56-resnet20 pair on CIFAR100 for analysis.

Number of negative samples. We validate different N: 100, 200, 500, 1000, 2000. As shown in Tab. 4a, increasing the negative number leads to better performance, and the performance is saturated when $n > 500$. We therefore utilize $N = 500$ in all other experiments to save computational cost. Compared to CRD [41], our CRCD requires fewer negative features to reduce the need of memory. This is because CRCD can utilize few samples to generate a large number of relations, while CRD only depends on the number of samples.

Temperature τ. Fig. 4b reports the results when τ varies from 0.02 to 0.2. We find that both extremely high or low temperature leads to inferior performance. In general, a temperature between 0.03 to 0.07 works well. We set $\tau = 0.05$ for all other experiments.

Sampling policy. To ensure that negative samples are as up-to-date as possible, we store features and gradients in a queue way which will remove the oldest sample when adding the latest sample. We compare the randomly sampling policy and the queuing sampling policy in Fig. 4. The queuing sampling policy (denoted as queue) can consistently outperform the naive randomly sampling policy (denoted as random) when varying negative number N and temperature τ.

Projection dimension. We investigate the influence of output dimension for critic function h by setting output dimension to 64, 128, and 256 (the input relation dimension is 256-d). As shown in Tab. 2, compared to 128-d or 256-d, transformation with lower dimension (64-d) has some accuracy degradation. We utilize the 128-d linear transformation to make a trade-off between effectiveness and computational cost.

4.4. Comparison with State-of-the-arts

CIFAR100. We compare our CRCD with other advanced knowledge distillation methods in Tab. 5. Various modern CNN architectures [17, 21, 53, 52] are selected as teacher networks or student networks. For a fair comparison, we combine all distillation methods with conventional KD [19]. From Tab. 5, we can observe that our distillation

Table 5: **The top-1 accuracies (%) of seven different student-teacher pairs on CIFAR100.** The accuracies of the teachers' and students' performance when they are trained individually are presented in the second partition after the header. FRCD (or GRCD) is the incomplete version of CRCD which means that only feature relation (or gradient relation) is employed in distillation. The best results are **bolded** and the best in competing methods are underlined.

Teacher Student	WRN-40-2 WRN-16-2	WRN-40-2 WRN-40-1	resnet56 resnet20	resnet110 resnet20	resnet110 resnet32	resnet32x4 resnet8x4	vgg13 vgg8
Teacher	76.64	76.64	73.25	73.89	73.89	79.61	75.00
Student	73.53	72.33	69.06	69.06	72.31	72.57	70.71
KD [19]	75.11	73.59	71.08	70.92	73.07	73.19	72.85
FitNet [35]	75.37	73.71	71.65	70.95	73.21	73.42	73.24
AT [51]	75.92	73.92	71.69	71.03	73.29	73.29	73.16
SP [43]	75.84	73.85	71.59	71.15	73.12	73.36	73.29
CC [34]	75.89	73.69	71.42	70.96	73.06	73.52	73.06
VID [1]	75.53	73.95	71.32	70.93	73.19	73.75	73.13
RKD [32]	75.20	73.76	71.54	70.98	73.25	73.51	73.09
PKT [33]	75.67	73.89	71.68	71.08	73.32	73.63	73.28
AB [18]	71.31	73.76	71.29	70.95	73.16	73.43	73.02
FT [23]	75.78	74.02	71.52	71.03	73.21	73.28	73.19
NST [22]	74.51	73.62	71.47	71.14	73.21	73.58	73.14
CRD [41]	75.97	74.47	71.75	71.52	73.81	75.62	74.42
SSKD [47]	75.39	75.30	70.29	71.48	73.64	75.53	74.51
FRCD	76.18	75.26	72.70	72.02	74.65	75.99	74.54
GRCD	76.27	75.24	72.64	71.73	74.48	75.57	74.32
CRCD	**76.67**	**75.95**	**73.21**	**72.33**	**74.98**	**76.42**	**74.97**

Table 6: **Top-1 and Top-5 error rate (%) on ImageNet validation set.** We compare our CRCD with competing methods including AT [51], KD[19], SP [43], CC [34], CRD [41] and SSKD [47], and folow the training settings in [41].

	Teacher	Student	AT	KD	SP	CC	CRD	SSKD	CRCD
Top-1	26.69	30.25	29.30	29.34	29.38	30.04	28.62	28.38	**28.04**
Top-5	8.58	10.93	10.00	10.12	10.20	10.83	9.51	9.33	**9.06**

method CRCD can consistently outperform all other distillation methods with a large margin, including the recent state-of-the-arts, CRD and SSKD. Additionally, even only one element (feature or its gradient) is used in the relation distillation, our method can still achieve the competing accuracy when compared to CRD or SSKD. When the feature and its gradient are employed in the representation relation distillation simultaneously, our CRCD can significantly outperform the other methods. In particular, the accuracy gap between CRCD and the other best performing method is 0.9% (averaged over 7 pairs in Tab. 5).

To evaluate the distillation effectiveness across very different network architectures, we also carry out detailed comparisons in supplementary materials.

ImageNet. Following [41, 47], we adopt the ResNet34-ResNet18 pair to evaluate the effectiveness of CRCD on ImageNet. As shown in Tab. 6, the Top-1 and Top-5 accuracy between the teacher and student without distillation is 3.56% and 2.43%. Our CRCD reaches the best distillation performance by narrowing the performance gap by 2.21% and 1.87% respectively. Results on ImageNet demonstrates the scalability of our CRCD to large-scale benchmarks.

5. Conclusion

In this work, we have proposed a novel knowledge distillation method, CRCD, to distill important structural information from a teacher to a student. To better distill the relation knowledge, two sub-networks are used to estimate the cross-space relation and teacher-space relation, respectively. We maximized the mutual information between the two kinds of relations by a newly proposed relation contrastive distillation loss, and utilized two complementary elements, the feature and its gradient, to enhance the representative ability of the relation. With the design of the loss function, the inter-sample relation and representation learning can be optimized simultaneously. Extensive experiments demonstrate the effectiveness of our approach and suggest that the structural information of deep representation can be better exploited during distillation.

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. 8

[2] Vasileios Belagiannis, Azade Farshad, and Fabio Galasso. Adversarial network compression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 1

[3] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018. 3

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 2

[5] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[6] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep crf for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[7] Dapeng Chen, Zejian Yuan, Badong Chen, and Nanning Zheng. Similarity learning with spatial constraints for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 3

[9] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 1

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[11] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 7

[12] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *International Conference on Learning Representations*, 2020. 1

[13] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *Advances in Neural Information Processing Systems*, 2020. 3

[14] Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. Knowledge distillation: A survey. *arXiv preprint arXiv:2006.05525*, 2020. 1

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 3

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[18] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3779–3787, 2019. 2, 8

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2, 5, 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. *arXiv preprint arXiv:1808.06670*, 2018. 2, 3

[21] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 7

[22] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017. 8

[23] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In *Advances in neural information processing systems*, pages 2760–2769, 2018. 8

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[25] Suichan Li, Dapeng Chen, Bin Liu, Nenghai Yu, and Rui Zhao. Memory-based neighbourhood embedding for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[26] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014. 7

[27] Iou-Jen Liu, Jian Peng, and Alexander G Schwing. Knowledge flow: Improve upon your teachers. *arXiv preprint arXiv:1904.05878*, 2019. 1

[28] Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. Knowledge distillation via instance relationship graph. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7096–7104, 2019. 2

[29] Andrey Malinin, Bruno Mlodozeniec, and Mark Gales. Ensemble distribution distillation. *arXiv preprint arXiv:1905.00076*, 2019. 1

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3, 7

[32] 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, 8

[33] 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. 2, 6, 8

[34] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5007–5016, 2019. 2, 3, 6, 8

[35] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 2, 8

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 7

[37] Zhiqiang Shen, Zhankui He, and Xiangyang Xue. Meal: Multi-model ensemble via adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4886–4893, 2019. 1

[38] Shixiang Tang, Dapeng Chen, Lei Bai, Yixiao Ge, and Wanli Ouyang. Mutual crf-gnn for few shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1

[39] Shixiang Tang, Dapeng Chen, Jinguo Zhu, Shijie Yu, and Wanli Ouyang. Layerwise optimization by gradient decomposition for continual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[40] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 2

[41] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019. 1, 3, 5, 6, 7, 8

[42] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020. 3

[43] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1365–1374, 2019. 2, 6, 8

[44] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *arXiv preprint arXiv:2005.10242*, 2020. 3

[45] Ancong Wu, Wei-Shi Zheng, Xiaowei Guo, and Jian-Huang Lai. Distilled person re-identification: Towards a more scalable system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1187–1196, 2019. 1

[46] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 2, 5

[47] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge distillation meets self-supervision. *arXiv preprint arXiv:2006.07114*, 2020. 3, 8

[48] Zheng Xu, Yen-Chang Hsu, and Jiawei Huang. Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks. *arXiv preprint arXiv:1709.00513*, 2017. 1

[49] 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

[50] 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. 2

[51] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 2, 8

[52] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 7

[53] 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. 7

[54] Jinguo Zhu, Zejian Yuan, Chong Zhang, Wanchao Chi, Yonggen Ling, et al. Crowded human detection via an anchor-pair network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1391–1399, 2020. 1