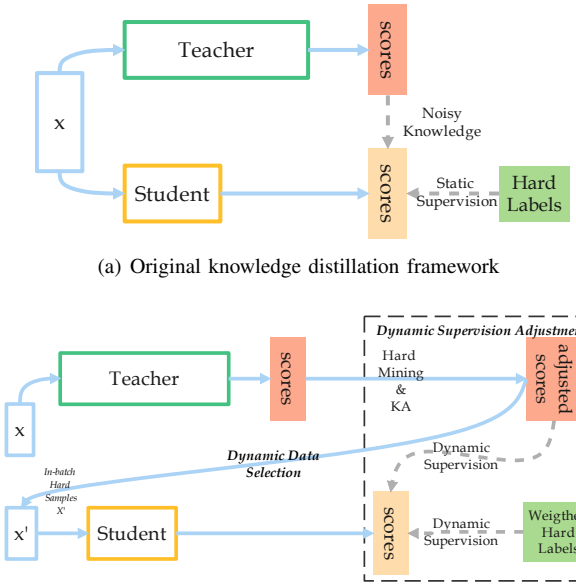
Dynamic Contrastive Distillation for Image-Text Retrieval

Jun Rao*, Liang Ding*, Shuhan Qi†, Meng Fang, Yang Liu, Li Shen, and Dacheng Tao, *Fellow, IEEE*

***Abstract*—The recent advancement in vision-and-language pre-training (VLP) has significantly improved the performance of cross-modal image-text retrieval (ITR) systems. However, the increasing size of VLP models presents a challenge for real-world deployment due to their high latency, making them unsuitable for practical search scenarios. To alleviate this problem, we present a novel plug-in dynamic contrastive distillation (DCD) framework to compress the large VLP models for the ITR task. Technically, we face the following two challenges: 1) the typical uni-modal metric learning approach is difficult to directly apply to cross-modal tasks due to the limited GPU memory to optimize too many negative samples during handling cross-modal fusion features. 2) it is inefficient to static optimize the student network from different hard samples, which affects distillation learning and student network optimization. We propose a method for multi-modal contrastive learning that balances training costs and effects. Our approach involves using a teacher network to identify hard samples for student networks to learn from, allowing the students to leverage the knowledge from pre-trained teachers and effectively learn from hard samples. To learn from hard sample pairs, we propose dynamic distillation to dynamically learn samples of different difficulties to balance better the difficulty of knowledge and students' self-learning ability. We successfully apply our proposed DCD strategy on two state-of-the-art vision-language pretrained models, i.e., ViLT and METER. Extensive experiments on MS-COCO and Flickr30K benchmarks show the effectiveness and efficiency of our DCD framework. We further provide in-depth analyses and discussions that explain how the performance improves.**

***Index Terms*—cross-modal retrieval, neural networks, contrastive learning**



(a) Original knowledge distillation framework

(b) Our dynamic contrastive distillation framework

Fig. 1. Comparison of an existing distillation method and proposed framework. (a) The original distillation method supervises students through hard labels and uncorrected soft labels "Noisy Knowledge". (b) In our DCD framework, we obtain the dynamically adjusted scores by "Hard mining&KA" (Section IV-A) to improve the soft labels provided by the teachers. At the same time we input the new filtered set of hard samples into the student network and learn these more informative samples learning dynamically through the weighted hard labels. Combining these two parts, our framework is greatly improved compared to the original KD (Section V).

I. Introduction

MULTIMODAL learning becomes a surging topic due to the increasing accessibility of multimodal data, such as image, text, video and audio [1], [2]. Also, with the advances of hardware, neural network models are able to scale up their capacity and expressive power to better leverage information from multiple modalities [3], [4], [5], [6].

Multimodal learning (e.g. cross-modal retrieval) becomes a popular research topic [7], [8]. Image-text retrieval focuses on obtaining a set of sentences given a query image, namely measuring cross-modal similarity of image and text. How to accurately measure the similarity of images and texts is at the core of this task. Most image-text retrieval methods adopt the idea of contrastive learning [9], [10], [11]. In this paradigm, different modalities are encoded into a semantic space to obtain modal-agnostic semantic representation, such that the representations of a query and its corresponding matching key are clustered while unmatched key-query pairs are separate.

Constructing a dynamic dictionary might be considered a typical contrastive learning strategy [12], [13], [14]. In the dictionary, each word is a sample embedding encoded by the network. Essentially, such a dynamic dictionary can be regarded as a global negative sample pool to explore informative sample pairs. When the dictionary is large enough and contains enough negative samples, the encoder can extract more discriminative features. In this way, large and consistent dictionaries can be constructed for unsupervised learning with a contrastive loss [15].

Knowledge distillation (KD) [16] is to obtain a much smaller model with comparable performance, while greatly reducing the memory usage and accelerating the model infer-

J. Rao and Y. Liu are with the Harbin Institute of Technology, Shenzhen, China; S. Qi is with Harbin Institute of Technology Shenzhen, and Peng Cheng Laboratory, and Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies, China; L. Ding, L. Shen and D. Tao are with the JD Explore Academy at JD.com, Beijing, China; M. Fang is with the University of Liverpool, Liverpool, the United Kingdom.

∗ Equal contribution. Work was done when Jun was interning at JD Explore Academy.

† Corresponding Author. shuhanqi@cs.hitsz.edu.cn

ence. It has been widely used in recent years in natural language processing (NLP) and computer vision (CV) tasks [17], [18], [19], [20], [21], [22], [23]. However, when applying KD for image-text retrieval, there are incompatibilities between the contrastive learning paradigm and image-text retrieval [24], [25], [26].

On one hand, the dictionary [12] is not suitable for image-text retrieval learning. To acquire multimodal content embedding, most multimodal models must interact with distinct changing modalities of embedding. As a result, the slowly updated dictionary cannot maintain a large and diversified sample pool. On the other hand, the existing image-text retrieval systems are too inefficient to learn the separability of sample pairs [27], [25], [26]. Though state-of-the-art (SOTA) methods, e.g. UNITER [25] and ViLBERT [26] using self-distillation [28], utilize an intermediate model to take a large number of random samples and select the top K sample pairs that are most similar to fine-tune the image-text retrieval model, it is inefficient due to the lack of a stable selection of hard negative samples and appropriate information guidance, as well as the fact that it is computationally intensive for contrastive learning. Because of the limited information and low gradient values of these randomly picked samples, their contributions to the training may be less informative since many of them already satisfy the loss's requirements. Besides, a longer training time is necessary to make the network converge. These problems lead to our first research question (RQ):
RQ1: How to facilitate the contrastive learning paradigm in the distillation of image-text retrieval with informative samples?

Different from those recent SOTA methods [25], [26], we directly introduce a teacher network to train a smaller student using knowledge distillation (KD) [16] to more efficiently learn the differences and similarities of samples within constrained resources, as shown in Figure 1(a). This teacher network may convey knowledge to a student network and select informative samples for training the multimodal interaction layers. A well-learned teacher network is flexible in selecting hard samples, and stabilizes the training process. To make an analogy with the real world, we equate teacher networks with professors and student networks with graduate students. Professors typically have a certain level of knowledge and are aware of which topics are currently challenging, and it is advantageous for students to follow these topics.

Another issue with visual language model compression is how to make greater use of the available information of teacher. However, the vanilla KD approaches as shown in Figure 1(a) are static and unable to actively learn from different samples, consequently failing to learn the teacher's separability of sample pairs well. Another research question arises as a result of this:
RQ2: For the purpose of improving teacher knowledge transfer, is it possible to use the weighting method to dynamically learn diversified content, according to the information of the limited sample pairs?

We design a basic weighting strategy based on the teacher's uncertainty of samples and achieve dynamic distillation by adjusting the sample's contribution to the training. The core of our solution is that, for the distillation loss item, we pay particular attention to the samples that the teacher believes to be mastered. Concerning the task loss item, we place a higher weight on samples that the teacher believes to be confusing so that students can focus on learning these samples through the hard label.

As shown in Figure 1(b), our framework addresses the two research questions described above. Specifically, we first filter out the more informative and difficult samples for students to learn from the samples selected by the teacher. Second, we obtain adjusted scores by knowledge adjustment (KA) and weigh these samples by teacher uncertainty as soft labels for supervising student networks. Finally, we use teacher uncertainty to weigh the hard labels to obtain dynamic supervision signals to enhance students' self-learning ability. In this paper, we validate our approach in different training settings and benchmark datasets upon a single cross-modal fusion layer based on ViLT [29] model. The experimental results indicate that dynamic data selection and supervision weighting improve image-text retrieval performance. At the same time, we use METER [30] with a different architecture using co-attention [26] to achieve the same promising effects. Our contributions are:

- We propose to leverage the teachers' adjusted knowledge to mine hard samples and supervise the students (§IV-A);
- We explore a variety of sample level weighted settings to achieve better teacher knowledge transformation (§IV-B);
- Considering above aspects, we design a plug-in DCD framework to compress VLP models and guarantee competitive results compared to SOTA distillation approaches (§V);
- To the best of our knowledge, we are the first to dynamically distill a pre-trained model based on Transformer architecture with modalities' interaction for image-text retrieval.

II. Related work

A. Image-Text Retrieval

It is difficult to represent and match the semantic information of many modalities in image-text retrieval. Recently, numerous existing methods [27], [31], [32], [33], [34], [35] for image-text retrieval encode the features into a semantic space using modality-independent encoders and then perform modal fusion to obtain the corresponding fusion features for cross-modal matching.

Some methods [32], [27] investigate self-attention to improve the feature embedding of intra-modality and then measure distance in a common metric space to use contrastive learning. In practical application scenarios, this type of approach usually allows the encoding of each modality to be calculated in advance, and only involves the calculation of the dot product of each modal vector at the time of retrieval. Thus such an approach is usually more flexible for large-scale retrieval, while these pre-encoded feature vectors of individual modes can be used for other downstream tasks. However, this type of method usually brings worse reproducibility stability and poorer performance [36].

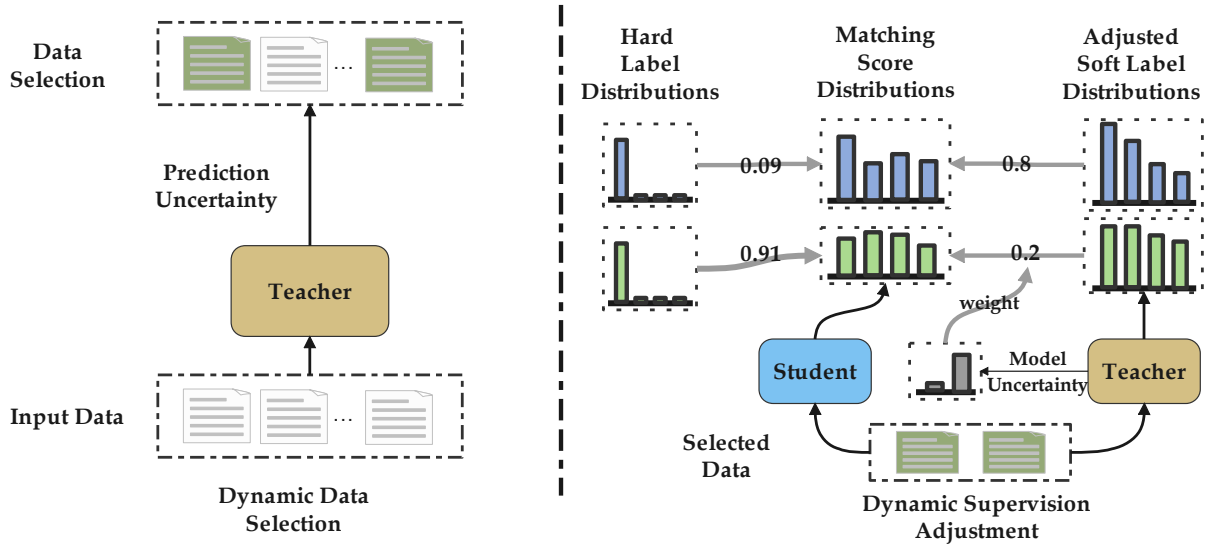


Fig. 2. Illustration of our dynamic contrastive distillation framework (DCD) in two aspects: data selection(*left*) and supervision adjustment(*right*).We obtain more informative data samples through the uncertainty of the teacher network (§IV-A). And we use the dynamic supervision adjustment (§IV-B) to use the selected hard samples and the weighted labels to supervise students dynamically with adjusted scores and weighted hard labels.

Others [24], [37], [29], [26], [25] adopt inter-modality interaction to obtain a robust multi-model representation. With a sophisticated cross-modal attention mechanism or a graph neural network, such methods are able to achieve state of the art in cross-modal tasks. These work demonstrate the importance of multi-modal interaction layers. Existing SOTA methods [25] [30] for ITR have used the transformer architecture for modal interaction between texts and images to obtain multi-modal fusion features. However, this interaction increase the complexity for retrieval (the fusion of the embedding vectors of each modality and too large amount of parameters) and not suitable in practical applications (long inference time). Specifically, if the two modalities have m and n samples, respectively, the complexity of such fusion often has a greater complexity ($O(mn)$), compared to the modality-independent approach [27], [32] ($O(m+n)$).

Therefore we focus on the latter type of approach. If such a heavy modal interaction layer can be compressed and obtained close to the original model, our training time, inference time can be greatly reduced and applied to similarly structured ITR models, which are more practical in real-world deployment.

B. Knowledge Distillation

KD [16] is presently one of the most appealing methods for compression in BERT-like models [38] and can be applied in ITR for compressing the multi-modal interaction layers.

The theory behind KD is that a large teacher model can teach a small student model to imitate the teacher's behavior. In this way, the knowledge contained in the teacher model can be effectively transferred to the student model. A set of methods [39], [19], [18], [40] use intermediate state matching and logit matching to distill a pre-trained language model for downstream tasks, achieving a strong compression effect while performing almost the same as the original model. Nevertheless, these methods still need to use a large-scale unlabeled corpus for distillation, which requires a lot of computational resources [41], [42], [43], [44]. In the multimodal field, there are also several works [45], [46], [47] that introduce KD to compress visual-and-language (VLP) models. Since the baseline models used in these methods are similar to BERT, most of them migrated to the BERT compression method of NLP and obtained good distillation results. After a simple application of KD, only a few works [48], [49], [50], [51], [52] explore how to further transfer the teacher's knowledge to the student. They follow the idea of active learning to choose the data and the degree of learning for each sample.

Inspired by above works, we use ViLT [29] and METER [30], the powerful pre-trained models of the BERT family, and investigate how to perform dynamic and adaptive distillation to achieve better distillation results in multimodal retrieval tasks. Meanwhile, different from the recent weighting methods [53], [50], we introduce the teacher-student framework to obtain the uncertainty score by a well-known teacher network rather than the student model itself.

C. Hard Sample Mining

Although our distillation framework is consistent with the goal in [54] to obtain more informative negative samples to optimize learning, our implementation is quite different from [54] and is more applicable to multi-modal settings.

The key to contrastive learning is using a small number of samples to approximate the distribution of the data. Therefore, hard sample mining to better utilize the informative negative samples has been extensively studied [53], [55], [25], [56], [57]. One of the mainstream methods for hard sample mining is online mining, which uses the loss [56] or different gradients [57] of samples in a batch to decide whether samples are hard. For most image-text retrieval methods [27], [24], [58], the negative samples with the highest similarity in a batch are selected online as the informative hard negative

samples so that there is no need to calculate other negative samples. The SOTA models, ViLBERT [26] and UNITER [25] adopt checkpoints at different stages of training to select hard samples, which help it build diverse sample pools. However, such methods scarifice time and resources to achieve better performance.

In contrast, we achieve the same goal (mining informative hard negatives) in a different way through data selection and knowledge adjustment while striking a balance between time and performance (§IV-A).

III. Preliminary

A. Contrastive Learning

In contrastive learning [15], [59], a representation space is commonly obtained by mixing training with positive and negative samples. Though training, the semantically similar samples are closer together in this space, whereas the semantically dissimilar samples are separated from each other. Unlike typical self-supervised contrastive learning [12], [60], image-text retrieval uses a supervised learning paradigm. Both visual and text are available as queries. In general, considering the image-to-text, for a query image v_i, the matching sentence t_i can be obtained directly according to the data annotation, and the unmatched sentence t_j is obtained by randomly sampling non-labeled related sentences. And the text-to-image scenario is similar. For the convenience of description, we use q to represent a modal embedding query (can be an image v_i or a sentence t_i) and k to represent another modal embedding key.

Given a set of image-text pairs $\{(q_i, k_i)\}_{i=1}^{N}$, our goal here is to use a contrastive learning approach to learn an optimal scoring function such that the scores of the matched image-text pairs (q_i, k_i^+) are higher than the scores of the rest of the unmatched samples $(q_i, k_j^-), j \neq i$.

From the probabilistic perspective, aligning k_i to q_i is equivalent to maximizing the conditional probability $p(k_i|q_i)$ while minimizing the probability for all negative pairs $p(k_j^-|q_i), j \neq i$. According to [61], $p(k_j|q_i)$ can be approximated as:

$$p\left(k_j \mid q_i\right) \sim \frac{\exp^{s(q_i,k_j)}}{\sum_{m=1}^{N} \exp^{s(q_i,k_m)}} \tag{1}$$

where $s(q_i, k_j)$ is the matching score between q_i and k_j; the denominator is a sum over all possible sentences, which is a partition function for normalization. Therefore, NCE loss [61] can be measured in a softmax fashion:

$$\begin{aligned} \mathcal{L}_{NCE} &= \sum_{i=1}^{N} -\log p\left(k_i \mid q_i\right) \\ &\sim \sum_{i=1}^{N} -\log\left(\frac{\exp^{s(q_i,k_i)}}{\exp^{s(q_i,k_i)} + \sum_{m\neq i} \exp^{s(q_i,k_m)}}\right) \end{aligned} \tag{2}$$

The denominator in Equation 2 requires a sum over all sentences in a dataset, which is intractable in practice. Therefore, we usually compute the NCE loss on a mini-batch of $K(K \ll N)$ image-text pairs sampled from the whole dataset.

B. Knowledge Distillation

For vanilla KD [16], we need a teacher network to guide the student network. We consider a single mini-batch and let $\boldsymbol{z_i^k}$ be the k-th value of the logit vector $\boldsymbol{z_i}$. The initial teacher and student model can be defined as: teacher $\boldsymbol{p}(\theta^t)$ and student $\boldsymbol{p}(\theta^s)$, respectively, where θ is the net parameters and $\boldsymbol{p^k}(\cdot) = \frac{\exp\left(\boldsymbol{z}^k(\theta)/\tau\right)}{\sum_{j=1}^{K}\exp(\boldsymbol{z}^j(\theta)/\tau)}$ is the probability predict of the matching label and K is the number of classes. So the KL divergence distillation loss can be defined as:

$$\mathcal{L}_{KL}\left(\boldsymbol{p}(\tau|\theta^s), \boldsymbol{p}(\tau|\theta^t)\right) = \tau^2 \sum_j \boldsymbol{p}^j(\tau|\theta^t) \cdot \log \frac{\boldsymbol{p}^j(\tau|\theta^t)}{\boldsymbol{p}_j(\tau|\theta^s)} \tag{3}$$

where τ is the temperature factor used in KD, which controls how much to rely on the teacher's soft predictions. For simplicity of notation, we use $\mathcal{L}_{KL}$ to represent $\mathcal{L}_{KL}\left(\boldsymbol{p}(\tau|\theta^s), \boldsymbol{p}(\tau|\theta^t)\right)$.

For a better distillation effect, we follow [62] and use Mean Squared Error (MSE) loss. The MSE Loss can be defined as follows:

$$\mathcal{L}_{MSE} = ||\boldsymbol{z}(\theta^s) - \boldsymbol{z}(\theta^t)||_2^2 \tag{4}$$

We can therefore get the final loss $\mathcal{L}$ of the student network:

$$\mathcal{L} = \alpha\mathcal{L}_{MSE} + (1-\alpha)\mathcal{L}_{task} \tag{5}$$

where α is the hyper-parameter that balances the importance of the task loss $\mathcal{L}_{task}$ and the distillation loss $\mathcal{L}_{MSE}$.

IV. Method

Figure 2 illustrates our DYNAMIC CONTRASTIVE DISTILLATION framework in two aspects. From the data side ("Dynamic Data Selection"), we select the informative samples for students according to the teacher's uncertainty estimation. From the supervision side ("Dynamic Supervision Adjustment"), we use teacher uncertainty to select the level of importance of supervision.

A. Dynamic Data Selection and Knowledge Adjustment

1) Dynamic Data Selection: Contrastive learning [12] benefits from a large batch size [13] and extensive data augmentation [63]. However, the high computational cost of multimodal fusion layers hinders its wide usage.

As defined in Section III-A, $\mathcal{L}_{NCE}$ requires the network to pass all $N \times N$ pairs into the multimodal layers. Assuming that the number of tokens in the image is m, the number of tokens in the text is n, and the dimension of each token is d, then the complexity of the self-attention mechanism is $O\left(d(m+n)^2\right)$. Due to such high computational complexity, most image-text retrieval methods adopt a smaller value of $K(K \ll N)$. NCE loss directly samples positive samples in a mini-batch of K pairs and get the remaining $K \times (K-1)$ mismatched pairs. In typical methods [25], [26], [24], they take only one negative sample. Thus Equation 2 becomes the following:

$$\mathcal{L}_{ITM} = \sum_{i=1}^{K} -\log\left(\frac{\exp^{s\left(q_i,k_i^+\right)}}{\exp^{s\left(q_i,k_i^+\right)} + \exp^{s\left(q_i,k_i^-\right)}}\right) \tag{6}$$

where k^+ is the paired key of the corresponding query, and k^- is the unmatched negative key. In particular, most methods usually use the way of VSE++ [27] to calculate more negative samples, and only update the gradient of the hardest negative sample during back propagation. VSE++ [27] will be constrained by the efficacy of the existing network's learning, and if the present network is not well-optimized, it may result in worse hard negative selection. In contrast, we provide a superior approach by employing a network that has previously been optimized as a selection for the hard negative samples. Motivated by [64] and [55], we select hard examples from the teacher to input into the student for supervised learning.

We get fewer but more useful negative samples by taking larger random negative samples and propagating them through the teacher network. As shown in Figure 1(b), we begin by increasing the number of negative samples obtained by the teacher network to M. Then we calculate M negative sample pairs and a positive sample pair in the teacher network to obtain the logits over the binary class of $M+1$ matching scores. Then we sort the scores, take the $M'+1$ largest scores, and find the corresponding samples and their matching scores. We input these informative samples into the student network, and get the logits of the $M'+1$ matching scores of the student network. The new image-text matching loss function can be defined as follows:

$$\mathcal{L}_{ITM'} = \sum_{i=1}^{K} -\log\left(\frac{\exp^{s\left(q_i,k_i^+\right)}}{\exp^{s\left(q_i,k_i^+\right)}+\sum_{j=1}^{M'}\exp^{s\left(q_i,k_{ij}^-\right)}}\right) \tag{7}$$

Then we use the teacher's logits score to constrain the student network. In this way, we not only reduce the gradient calculation of the student network but also improve the ability of the student network. Naturally, if a larger negative sample value (M) is adopted, the performance of the network will be improved to a certain extent, but at the same time, it will increase the network learning time. We need to adjust this value (M) and the number of negative samples (M') that students need to learn in practical applications.

2) Knowledge Adjustment: In KD and KD-based approaches, the student network is trained under the supervision of teacher predictions, regardless of whether this supervision signal is right or wrong. We select the indistinguishable and hard negative samples according to the teacher network. Therefore, the scores of these negative samples may be higher than those of positive samples. We call this situation a "genetic error". At the same time, if a student continues to learn this erroneous knowledge under the supervision of the teacher, it will further lead to errors in the student network. Therefore, we try to fix the samples of this mini-batch where the teacher's prediction does not match the true label, which we call knowledge adjustment.

For simplicity, we consider a mini-batch with only one matching pair. As described in the aforementioned section, we obtained the matching scores of $M^{'}+1$ samples with the teacher network. The matching scores are sorted in descending order, but the positive samples are not necessarily among the $M^{'}+1$ samples. Based on this consideration, we move the positive sample to the top of the original matching score list to generate a new mapping of samples and matching scores. In this way, we ensure that the sample with the highest confidence score must be a positive sample.

When computing the loss, for implementation convenience, for a batch of positive and negative samples (e.g., one positive sample pair and 15 negative sample pairs), we set the first position in the sample list to be the matching positive sample pair and the remaining positions to be the negative sample pairs. In general, we expect the first score to be the highest (because it is a positive sample pair), so we guarantee the nature of the highest matching score for the positive sample pair by performing such an insertion operation on the list of scores inferred by the teacher model, which can correct some erroneous outputs. This method also keeps the numerical distribution of soft targets, which is helpful in stabilizing the training process.

With such a simple implementation of dynamic data selection and knowledge adjustment, we reduce the computational cost of training as well as bring a more performance and inference time balanced student model, and also bring a simple and effective alternative for practical lightweight ITR model deployment.

B. Dynamic Supervision Adjustment

In the actual data annotation for image-text retrieval, we can know that image and text matching (1) or not matching (0), so we call this label information as "hard labels". We can also get the output of the teacher, which represents the degree of matching (normalized to a value between 0 and 1). This time the more matching samples are closer to 1 for the output of matching scores, and the more not matching samples scores are closer to 0. The output is a score to indicate how an image and a text match. We call this score as "soft labels". In Hindon [16], it shows the information of soft labels is easier to learn compared to hard labels because of the inclusion of inter-class differences. Similarly, in image-text retrieval, soft labels are more representative of how well different sample pairs match, and bring more information worth learning.

We select more valuable learning samples through the teacher network and ensure the relative correctness of the predictions of the teacher, but there are still uncertain samples. Intuitively, for samples with high degree of certainty considered by the teacher, the soft labels provided by the teachers bring more significant learning information than the hard labels. Although these samples can provide a certain amount of information, if the student network completely relies on the guidance of the teacher's judgement at this time, this output may mislead feature learning in the fine-tuning stage and hurt adaptation performance.

We aim to reduce the negative influence of noisy soft labels by evaluating the credibility of these soft labels for each sample and reweighting the contributions of samples with error-prone predictions in the NCE loss and KD loss. In order to improve learning on such samples, we divide them into the following three parts: uncertainty estimation, weighted hard labels and weighted soft labels.

1) Uncertainty Estimation: A crucial step in achieving dynamic supervision is estimating the prediction uncertainty. Entropy [65] is an information-theory-based method for calculating uncertainty and is commonly employed for this purpose. It is also common to utilize approximate Bayesian inference methods, such as MC-Dropout [66] and Ensemble [67], to model uncertainty. Using the difference in the prediction results of numerous inferences on the same input, these methods evaluate the uncertainty of models. This mode can significantly increase the training time (more model inferences) and is unsuitable for networks with a large number of parameters (e.g., teacher networks used in distillation). The objective of our work is to obtain information about the uncertainty of prediction distribution, which can be obtained by averaging the entropy of multiple matched or unmatched samples' scores and feeding the data directly into the network with negligible computation time. So we use the average predictive entropy to estimate uncertainty.

2) Weighted Hard Labels: We use naive entropy to weight the hard labels in the task loss term, similar to previous works [53], [48] that assign sample-wise weights. In contrast to [49], we focus on self-exploration of task loss and reduce attention to the most difficult samples. Intuitively, the greater the teacher's uncertainty about the output of a sample, the lower the reliability of the output. Therefore, we improve the student model of self-exploration by increasing its attention to the sample of task loss in which the teacher is highly uncertain.

Given N instances in one batch, the corresponding output matching scores probability distribution of the student model over the positive-and-negative pair index y (position 0 is the positive pair, and the others are negative pairs) is $p(y \mid x_i)$ like Equation 1, denoting the model confidence towards the positive pair. The uncertainty score u of teacher about the output of x_i can be defined as Equation 8 with negligible computational overhead:

$$u_{x_i} = -\sum_{y} p(y \mid x_i) \log p(y \mid x_i) \tag{8}$$

Then we normalized the uncertain results to get the weight of the corresponding sample.

$$w_i = -\frac{u_{x_i}}{\sum_{i=1}^{K} u_{x_i}} \tag{9}$$

Finally, we combine the previous ITM loss to get the final weighted task loss as shown in Equation 10.

$$\mathcal{L}_{WITM} = -\sum_{i=1}^{K} w_i \log \left(\frac{\exp^{s\left(q_i, k_i^{+}\right)}}{\exp^{s\left(q_i, k_i^{+}\right)} + \sum_{j=1}^{M'} \exp^{s\left(q_i, k_{ij}^{-}\right)}} \right) \tag{10}$$

3) Weighted Soft Labels: Similar to [50], we found that reducing the weight of the hardest samples in the distillation term is beneficial to students' learning. It is possible that the teacher didn't generate the correct matching scores since these examples weren't well learned. As a result, if we focus on these harder samples, the supervisory information provided by teachers will be limited and may not be correct, and such errors may propagate to the student.

Therefore, we reduce the soft label loss term to pay attention to such samples in order to re-learn the samples that the teacher failed to master in a relatively correct training direction for the student network. However, even if these more hard samples need to be weighted down, they can still provide some useful information for the student. Therefore, even if we reduce the weight of such samples, we should not use the focal loss style weights [53] to make the weight difference between these samples too large. As a result, we defined the following reversed weights c based on the forward weights calculated by teacher entropy:

$$c_i = \frac{\exp^{(1-w_i)^2}}{\sum_{i=1}^{K} \exp^{(1-w_i)^2}} \tag{11}$$

Therefore, we combine the previous MSE loss to get the final weighted distillation loss as shown in Equation 12:

$$\mathcal{L}_{WDS} = \sum_{i=1}^{K} c_i || \boldsymbol{z_i}(\theta^s) - \boldsymbol{z_i}(\theta^t) ||_2^2 \tag{12}$$

C. Overall Learning Objective

The training objective in our method is finding the optimal θ^s by minimizing the combination of the above two weighted losses:

$$\mathcal{L}' = \alpha \mathcal{L}_{WDS} + (1-\alpha)\mathcal{L}_{WITM} \tag{13}$$

V. EXPERIMENTS

A. Datasets

We conducted experiments on two widely-used benchmarks: MS-COCO [68] and Flickr30k [69], which consist of 123,287 and 31,783 images, respectively, and each image has five corresponding sentence descriptions. We closely followed [70] to split the datasets. Concretely, the processed Flickr30k dataset contains 1,000, 1,000, and 29,783 images for testing, validation, and training, respectively. As for MS-COCO, 5,000 images for testing and 5,000 for validation, the rest 113,287 images are left for training.

During inference, the performance for image-text retrieval is reported by Recall at K (R@K), which represents the ranking proportion of queries with ground-truth within the top K. R@1, R@5 and R@10 are our evaluation metrics.

B. Implementation Details

We validated our proposed dynamic contrastive distillation on two state-of-the-art vision-language pretrained models, *ViLT* [29] and *METER* [30], where ViLT is used for the main experiments (§V-C1-§V-C4) to demonstrate the effectiveness of our method, and METER is used in §V-C5 to show the universality. Specifically, in the main experiments, we use ViLT with 12 Transformer layers as a teacher model for all scenarios that require distillation. While we use a 6 layers of Transformer as the student network for both Flickr30k and MS-COCO, we leave their best settings in the original paper [29] as the default. We compress the original model for 40 epochs and for 20 epochs on Flickr30k and MS-COCO

TABLE I
COMPARISONS OF EXISTING METHODS EXPERIMENTAL RESULTS ON FLICKR30K AND MS-COCO TEST SETS. "*" REPRESENTS THE RESULTS OBTAINED FROM [71], WHICH REMOVED THE EXCESSIVELY TIME-CONSUMING ONLINE HARD SAMPLE MINING PROCESS.

Visual Embed	Model	Param	Time (ms)	Flickr30K						MS-COCO					
				Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval		
				R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Region	SCAN [24]	∼73M	∼900	67.4	90.3	95.8	48.6	77.7	85.2	72.7	94.8	98.4	58.8	88.4	94.8
	CAMP [37]	∼94M	∼1000	68.1	89.7	95.2	51.5	77.1	85.3	72.3	94.8	98.3	58.5	87.9	95.0
	VSRN [31]	∼204M	-	71.3	90.6	96.0	54.7	81.8	88.2	76.2	94.8	98.2	62.8	89.7	95.1
	SAEM [32]	∼178M	≫900	69.1	91.0	95.1	52.4	81.1	88.1	71.2	94.1	97.7	57.8	88.6	94.9
	ViLBERT-Base* [26]	∼285M	∼920	76.8	93.7	97.6	59.1	85.7	92.0	77.0	94.1	97.2	62.3	89.5	95.0
	UNITER-Base* [25]	∼174M	∼900	78.3	93.3	96.5	62.9	87.2	92.7	74.4	93.9	97.1	60.7	88.0	93.8
Linear	DCD (Ours)	∼66M	∼7	75.6	91.0	94.6	53.7	81.4	88.2	76.5	94.1	98.0	59.7	89.7	95.7

TABLE II
PERFORMANCE OF THE TEACHER AND STUDENTS WITH DIFFERENT LOSS RE-WEIGHTING METHODS. "*" INDICATES THAT DYNAMIC SAMPLE SELECTION AND KNOWLEDGE ADJUSTMENT ARE USED. "N/A" MEANS THE TRAINING PROCESS DOES NOT CONVERGE, AND IT IS ALMOST IMPOSSIBLE TO RETRIEVE CORRECT RESULTS.

Method	Flickr30K		MS-COCO		Avg.
	TR@1	IR@1	TR@1	IR@1	
ViLT [29] (*12L Teacher*)	83.7	62.2	83.7	68.4	74.5
6L Student					
Directly fine-tuning	63.3	42.0	68.6	47.5	55.4
Vanilla KD [16]	71.8	50.6	72.8	54.7	62.5
WSL KD* [72]	74.4	51.7	74.4	57.2	64.4
Student-Uncertainy*	N/A	N/A	N/A	N/A	N/A
DCD*	75.6	53.7	76.5	59.7	66.4

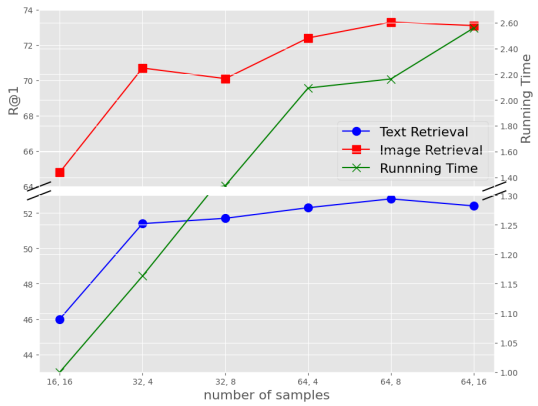


Fig. 3. The effects on retrieval ("R@1") and training time ("Running Time") when inputting the different numbers of samples for teacher and student. "(X, Y)" in the horizontal coordinate indicates the numbers of samples randomly inputting to the teacher and student, respectively. The green line represents the corresponding running time of each setting, while the red and blue lines show R@1 for text and image retrieval, respectively.

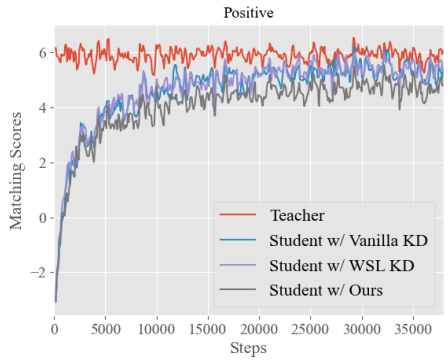
datasets, respectively. We train our models on 8 SuperPoD NVIDIA A100 GPUs. Due to the limitation of computational resources and in order to achieve a better trade-off between training time and retrieval performance (see discussion in Section V-C2), we set the number of negative samples to 63 and 7 for the teacher (i.e. M in §IV-A) and student (i.e. M' in §IV-A), respectively, for dynamic data selection.

C. Results and Analysis

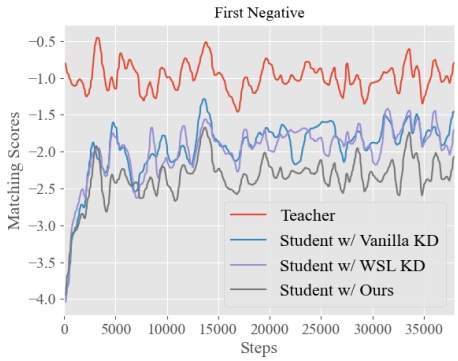
*1) **Image-text Retrieval Results**:* In this section, we compare the two datasets of Flickr30K and MSCOCO to verify the effectiveness of our framework. Table I shows the experimental results of Flickr30K and MSCOCO 1K test sets. For a comprehensive comparison, we list not only the *retrieval results* ("R@K") of the existing image-text retrieval methods, including SCAN [24], CAMP [37], VSRN [31], SAEM [32], ViLBERT [26], and UNITER [25], but also their corresponding *model sizes* ("Param") and *inference latency* ("Time").

a) Time and Param: Our distilled student model is based on patch features [73], which have negligible computational consumption and can be fed directly into the modality interaction transformer to obtain the final multimodal features. The traditional image-text retrieval models [26], [24], [37], [31], [32], [25] involve region supervision (i.e., the pretrained off-the-shelf object detector such as Faster-RCNN [74]) to obtain a better retrieval recall rate but add more time consumption (requiring the time-consuming regional selection). As shown in Table I, the inference speed of our compressed model ("Ours") is significantly faster than that of the existing region-based models, where the speedup is at least **129**× (*7ms* vs. 900*ms*). Also, our approach has the smallest model size among them, achieving up to 4.1× parameter compression (comparing to ViLBERT-Base).

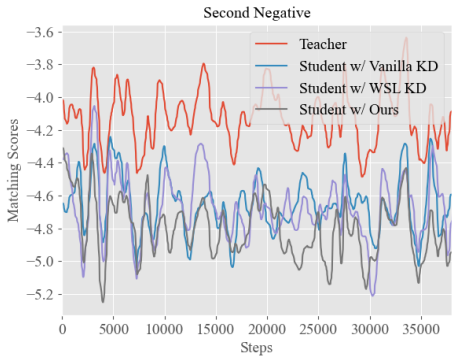
b) Retrieval Results: Regarding retrieval performance, our compressed model achieves competitive, if not better, performance compared to the existing state-of-the-art in image-text retrieval. Specifically, compared with similar-sized method SCAN [24], we achieve 8.2% and 5.1% improvements in R@1 text and image retrieval, respectively, in the Flickr30k dataset. In the MS-COCO 1K dataset, our method also obtains 3.8% and 0.9% improvement in text and image retrieval, respectively. Meanwhile, our compressed model gains consistent improvement compared to those larger non-pretrained



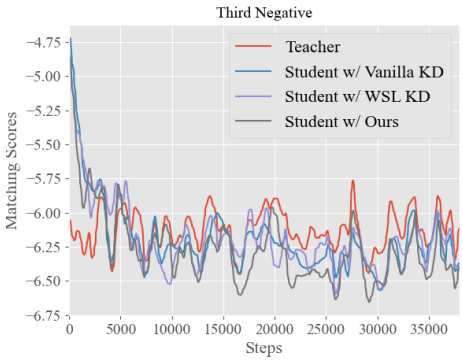
(a) The matching score of the **positive** pair.



(b) The matching score of the **most similar negative** pair.



(c) The matching score of the **second similar negative** pair.



(d) The matching score of the **third similar negative** pair.

Fig. 4. Comparison of different dynamic supervisions, i.e. "Student w/ {Vanilla KD [16]/ WSL KD [72]/ DCD (Ours)}" in terms of the matching score of true positives (a) and true negatives (b, c, d). The outputs of the teacher model are reported as the reference.

modal interaction models, i.e. CAMP [37], VSRN [31], and SAEM [32]. And encouragingly, compared to the state-of-the-art pretrained models, i.e. VILBERT [26] and UNITER [25], our compressed model achieves better performance with significantly fewer parameters, such as TR@1, TR@10, and IR@10 on MS-COCO.

These results demonstrate that DCD can achieve a balance between retrieval performance, the number of parameters (storage consumption), and calculation time (calculation consumption) in comparison to other robust region-based models.

TABLE III
THE IMPACT OF COMPONENTS IN FLICKR30K

Method	Flickr30K					
	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Vanilla KD	71.8	90.3	94.1	50.6	79.8	87.6
+DS&KA	73.7	90.7	94.4	51.5	80.0	87.2
+HW	75.5	90.0	94.2	53.5	81.4	88.4
+SW	74.6	89.7	94.4	51.7	80.1	87.7
+FULL	75.6	91.0	94.6	53.7	81.4	88.2

*2) **Analysis of Dynamic Data Selection**:* Recall that we denote the M and M' by the number of negative samples as input to the teacher and student network, respectively. These two factors may significantly influence both *retrieval performance* (Recall) and *training costs* (Time). To achieve the desired trade-off, we carefully investigate the effects of

TABLE IV
THE IMPACT OF COMPONENTS IN MS-COCO

Method	MS-COCO					
	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Vanilla KD	72.8	92.5	96.6	54.7	86.3	93.8
+DS&KA	73.9	93.2	97.5	57.1	87.8	95.2
+HW	74.5	93.4	97.5	58.1	87.6	94.7
+SW	75.2	93.9	97.6	57.0	88.0	95.0
+FULL	76.5	94.1	98.0	59.7	89.7	95.7

TABLE V
THE GENERALIZABILITY OF DCD UPON THE SOTA VLP MODEL – METER [30]. "*" INDICATES THAT DYNAMIC SAMPLE SELECTION AND KNOWLEDGE ADJUSTMENT ARE USED.

Method	Flickr30K					
	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
METER[30](*6L Teacher*)	94.3	99.6	99.9	82.2	96.3	98.4
3L Student						
Directly fine-tuning	80.3	96.4	98.3	55.2	86.7	93.3
Vanilla KD	92.7	99.4	99.8	79.8	96.4	98.4
WSL KD*	92.9	99.3	99.8	80.9	96.6	98.6
DCD*	93.4	99.3	99.8	81.9	96.6	98.7

them spanning a reasonable range, that is, (M+1, M'+1)[1] $\in$ {(16, 16), (32, 4), (32, 8), (64, 4), (64, 8), (64, 16)} shown in Figure 3. Note that the training costs of adding negative samples to the teacher network is substantially lower than that of the student network. We therefore can reduce the time cost of the student when mining difficult samples by increasing the sample input of the teacher network.

As seen in Figure 3, increasing the number of (negative) samples can basically obtain better text and image retrieval results (see the red and blue lines) but significantly enhance the training costs (see the green line), validating the effectiveness of negative samples in our dynamic contrastive distillation frameworks. We also show several interesting findings: *1)* increasing the number of hard negative samples of students does not improve the retrieval if we set a relatively small number of negative samples for teachers, e.g. (M+1, M'+1) changes from (32, 4) to (32, 8), demonstrating the necessity of setting a relatively large number of negative samples for teachers; *2)* increasing the number of negative samples for students causes the retrieval results to rise first and then decline. It shows that while there are some hard samples in the batch, it also increases the number of simple sample pairs, which hurts the network's final result. This is similar to what VSE++ [27] reported. Based on observations, to achieve the desired trade-off, we set the number of negative samples to 63 and 7 (in total 64 and 8) for the teacher and student, respectively, for hard sample dynamic selection.

*3) **Analysis of Dynamic Supervision Adjustment**:* We first empirically show the superiority of dynamic supervision adjustment, then discuss where the improvement comes from?

*a) **The Empirical Superiority of Our Method**:* In order to investigate the influence of different dynamic supervision adjustment strategies, we carefully compared our approach with existing competitive methods in Table II, including *1)* "**ViLT**" 12 layers ViLT [29] as a teacher to provide soft labels and weights, *2)* "**Directly fine-tuning**" directly finetuning 6 layers ViLT using downstream loss without distillation, *3)* "**Vanilla KD**" 6 layers ViLT distilled by 1), *4)* "**WSL KD**" is an existing strong baseline – weighted soft labels KD [72], which dynamically weights the sample level by combining elements like teacher and student losses, as well as the training step, and *5)* "**Student-Uncertainty**" follows our framework but supervised with the student uncertainty rather than teacher.

Clearly, "vanilla KD" improves the image-text retrieval results by averaged 7.1 points compared to directly fine-tuning the 6 layers ViLT without KD, i.e. "Directly fine-tuning", proving the effectiveness of knowledge distillation. Going a step further, the WSL weighting method "WSL KD" that combines multiple factors can push the effects of distillation to a significantly better level, i.e. averaged 1.9 points of improvements against the static "vanilla KD".

Surprisingly, we discovered that using student uncertainty as a weight for dynamic supervision caused the model to fail in convergence, denoted by "N/A". One possible reason is that students' optimization directions may be incorrect. And such incorrect supervisions are propagated to the students' learning process, exacerbate the errors, and eventually make the networks collapse.

"DCD" that employs the uncertain information from the knowledge-rich teachers to obtain dynamically weighted signals, by contrast, makes the training process stable, leading to further improvements. Compared with the "Vanilla KD" and competitive "WLS KD", DCD brings an average R@1 improvement of 4 and 2 points, respectively, validating the superiority of our approach.

*b) **Where Do the Improvements Come From**:* In order to more intuitively show where the improvements come from, we visualize the learning dynamics of the student network in terms of the matching score of true positives and true negatives on MS-COCO. Figure 4 depicts the matching scores of positive pair and top-3 negative pairs, including the matching score of *a)* the positive pair, *b)* the most similar negative pair, *c)* the second similar negative pair, and *d)* the third similar negative pair. When performing image-text retrieval tasks, we normally expect that matched image-text pairings have higher matching scores while dissimilar pairs have lower matching scores. Namely, a well-trained model is expected to have a high degree of separability between positives and negatives.

Overall, compared to the static "Student w/ vanilla KD" and dynamic supervision method "Student w/ WSL KD", our method (DCD) "Student w/ Ours" obtains **a higher degree of separability between positives and negatives**.

In particular, we show this with the matching score of the POSITIVE pair in Figure 4(a). Although our method is slightly lower than students with vanilla and WSL KD, the difference is not significant, or even comparable.

However, as for NEGATIVE samples, our method significantly reduces the matching score of negative examples, that is, our method could distinguish the true negatives better than other students, as shown in Figure 4(b), 4(c) and 4(d). Matching score observations on true positives and true negatives demonstrate that *our dynamic supervision empowers students with a higher degree of separability between positives and negatives*, thus leading to improvements.

*4) **Ablation Studies**:* To demonstrate the effectiveness of our dynamic contrastive distillation framework, a comprehensive component wise ablation analysis is performed. The results are reported in Table III and IV. Here, we use the vanilla KD described in Section III-B as the baseline. On both datasets, with our data selection and knowledge adjustment ("DS&KA"), we get an average improvement of almost 2 points beyond the baseline. Also, it can be seen that, based on the strategy "DS&KA", our soft-label weighting ("SW") and hard-label weighting ("HW") further obtain consistent improvements on both datasets. In particular, our "HW" gets more improvement on R@1, with averages of 1.9 and 0.8 points on Flickr30K and MS-COCO. Finally, combining all components "FULL", we achieve a further improvement, on average.

*5) **Generalizability of the Dynamic Distillation Framework**:* To verify our framework as a plug-and-play component applicable to other vision-language pretraining (VLP) models that are based on modal interaction fusion, we conduct ex-

[1]"+1" means our settings take one positive sample and the rest are negative samples.

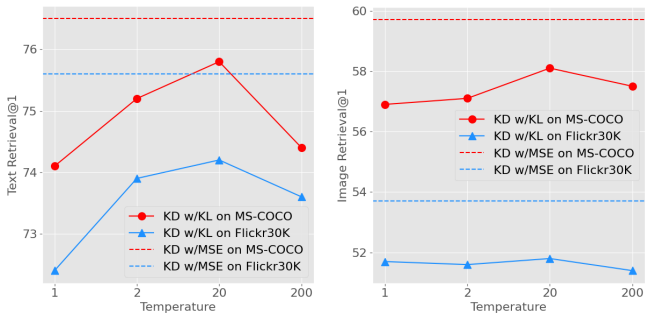
periments upon a current SOTA VLP model METER [30]. METER is a dual-stream architecture that performs training of image-text pair similarity by means of a heavier modal encoder to obtain the encoding of the respective modalities and a heavier modal interaction layer for fusion encoding. Using our framework, its fused modal encoding (co-attention layers [26]) can be compressed, and the results are shown in Table V. We compress the co-attention layer [26] of METER from 6 layers to 3 layers and report the retrieval performance on Flickr30K with fine-tuning only, vanilla KD, WSL [72] weighting, and our distillation framework. The settings are consistent with the original paper [30]. The results show that our distillation framework works well on compression co-attention, demonstrating its universality.

6) ***Hyper-parameter Sensitivity****:* A motivation for DCD is to allow teachers to transmit their knowledge dynamically. In order to have the ability to dynamically adjust during the learning process, it is natural to expect the DCD to be more insensitive and more robust to changes in settings. Here, we evaluate the performance of DCD for different temperatures and different loss weights.

a) ***Temperature****:* In [62], they verified experimentally and theoretically that $\mathcal{L}_{MSE}$ is superior to $\mathcal{L}_{KL}$ in certain circumstances. Using $\mathcal{L}_{MSE}$ instead of $\mathcal{L}_{KL}$ as distillation loss in DCD eliminates the need to set the temperature parameter, thereby reducing the difficulty of hyperparameter adjustment. Based on this motivation, we conduct relevant experiments on distillation for image-text retrieval, try four temperatures with the original $\mathcal{L}_{KL}$ in KD and illustrate the retrieval performance of the student on Flickr30K and MS-COCO, as shown in Figure 5.

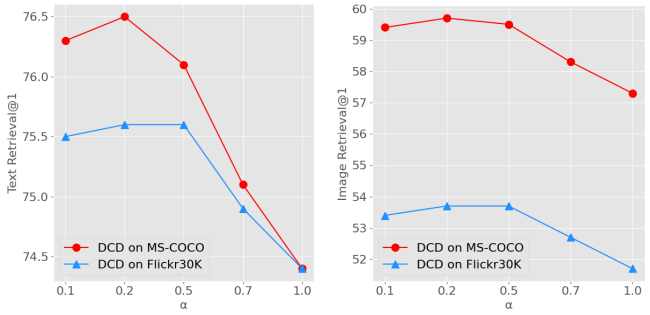
The following findings can be obtained from Figure 5: *1)* As described in Section III-B, $\mathcal{L}_{MSE}$ can produce improved student retrieval results within a specific temperature range, compared to $\mathcal{L}_{KL}$. The best results for Flickr30K text retrieval and image retrieval w/MSE are approximately two points better than w/KL. On MS-COCO, the results of w/MSE were similarly significantly superior to the best results of w/KL, gaining over one point in text retrieval and over two points in image retrieval, respectively. *2)* The results for w/KL are sensitive to temperatures, i.e., the variance of their results is large. On both Flickr30K and MS-COCO, Figure 5 shows that the difference between the best results and the worst R@1 for the students after distillation is close to two points.

b) ***Loss Weight****:* In KD as well as in our DCD, loss weight is an important balancing factor that balances the importance of soft and hard labels. To test the robustness of DCD at different loss weights, we perform experiments using our previous best settings with different α shown in Figure 6 (Equation 13, from 0.1 to 1). It can be found that DCD has good robustness when its weight is small, e.g., 0.1 to 0.3. In the range of 0.1 to 0.3, the difference between the results of the DCD text retrieval are less than 0.5 points (MS-COCO is close to 0.5 points and Flickr30K is only 0.2 points), while the results of the image retrieval are also within one point. This indicates that hard labels and soft labels through our dynamic data selection and dynamic supervision adjustment can have a good balance in this interval to enhance the final



(a) Text retrieval (b) Image retrieval

Fig. 5. Results with different temperature on Flickr30K and MS-COCO. The w/MSE and w/KL mean using $\mathcal{L}_{MSE}$ and $\mathcal{L}_{KL}$ as the distillation loss, respectively.



(a) Text retrieval (b) Image retrieval

Fig. 6. Results with different loss weight α on Flickr30K and MS-COCO.

distillation results. Also congruent with the findings of [16] is the discovery that soft labels cannot replace hard labels, but do play a role in the optimization of the student network. When employing only soft labels ($\alpha = 1.0$), the distillation results deviated from the best retrieval results on Flickr30K and MS-COCO by roughly 2 and 3 points, respectively.

VI. Conclusion

In this paper, we proposed a play-and-plug dynamic contrastive distillation framework named DCD, which consists of two major aspects, dynamic data selection and dynamic supervision adjustment, for the image-text retrieval task. Extensive experiments upon different VLP models (ViLT and METER) demonstrate that dynamic adjustments in both data and supervision according to teachers' uncertainty estimation can effectively improve student performance and learning efficiency. Further analyses reveal that the improvement comes from 1) fully mining the hard negative samples, and 2) providing a higher degree of separability between positives and negatives. We hope that our method could shed light on more image-text tasks in the future.

Acknowledgment

This research was funded by Science and Technology Innovation 2030 –"Brain Science and Brain-like Research" Major Project (No. 2021ZD0201405), Guang-

dong Provincial Key Laboratory of Novel Security Intelligence Technologies under Grant 2022B1212010005, in part by the Shenzhen Foundational Research Funding under Grant JCYJ20200109113427092, JCYJ20220818102414030, JCYJ20200805173048001, in part by the PINGAN-HITsz Intelligence Finance Research Center, in part by the Ricoh-HITsz Joint Research Center, and in part by the GBase-HITsz Joint Research Center. The computing resources of Pengcheng Cloud Brain are used in this research.

REFERENCES

[1] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, and F. Wei, "Image as a foreign language: Beit pretraining for all vision and vision-language tasks," *CoRR*, vol. abs/2208.10442, 2022.

[2] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and v. . . p. . . y. . . title = OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework, booktitle = ICML.

[3] Z. Kyaw, S. Qi, K. Gao, H. Zhang, L. Zhang, J. Xiao, X. Wang, and T. Chua, "Matryoshka peek: Toward learning fine-grained, robust, discriminative features for product search," *IEEE Trans. Multim.*, vol. 19, no. 6, pp. 1272–1284, 2017.

[4] X. Lu, L. Liu, L. Nie, X. Chang, and H. Zhang, "Semantic-driven interpretable deep multi-modal hashing for large-scale multimedia retrieval," *IEEE Trans. Multim.*, vol. 23, pp. 4541–4554, 2021.

[5] W. Wang, H. Bao, L. Dong, and F. Wei, "Vlmo: Unified vision-language pre-training with mixture-of-modality-experts," in *NeurIPS*, 2022.

[6] Z.-Y. Dou, A. Kamath, Z. Gan, P. Zhang, J. Wang, L. Li, Z. Liu, C. Liu, Y. LeCun, N. Peng, J. Gao, and L. Wang, "Coarse-to-fine vision-language pre-training with fusion in the backbone," in *NeurIPS*, 2022.

[7] C. Li, T. Yan, X. Luo, L. Nie, and X. Xu, "Supervised robust discrete multimodal hashing for cross-media retrieval," *IEEE Trans. Multim.*, vol. 21, no. 11, pp. 2863–2877, 2019.

[8] Y. Zeng, X. Zhang, and H. Li, "Multi-grained vision language pre-training: Aligning texts with visual concepts," in *ICML*, vol. 162, 2022, pp. 25 994–26 009.

[9] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.

[10] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *CVPR*, 2019, pp. 6210–6219.

[11] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *CVPR*, 2021, pp. 9650–9660.

[12] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*. Computer Vision Foundation / IEEE, 2020, pp. 9726–9735.

[13] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020, pp. 1597–1607.

[14] J. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Á. Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent - A new approach to self-supervised learning," in *NeurIPS*, 2020.

[15] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *CVPR*, 2006, pp. 1735–1742.

[16] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NeurIPS*, 2015.

[17] C. Gong, D. Tao, W. Liu, S. J. Maybank, M. Fang, K. Fu, and J. Yang, "Saliency propagation from simple to difficult," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2531–2539.

[18] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling BERT for natural language understanding," in *EMNLP (Findings)*, 2020, pp. 4163–4174.

[19] S. Sun, Y. Cheng, Z. Gan, and J. Liu, "Patient knowledge distillation for BERT model compression," in *EMNLP*, 2019, pp. 4322–4331.

[20] X. Jin, B. Peng, Y. Wu, Y. Liu, J. Liu, D. Liang, J. Yan, and X. Hu, "Knowledge distillation via route constrained optimization," in *ICCV*, 2019, pp. 1345–1354.

[21] P. Chen, S. Liu, H. Zhao, and J. Jia, "Distilling knowledge via knowledge review," in *CVPR*, 2021, pp. 5008–5017.

[22] L. Ding, L. Wang, X. Liu, D. F. Wong, D. Tao, and Z. Tu, "Understanding and improving lexical choice in non-autoregressive translation," in *ICLR*, 2021.

[23] L. Ding, L. Wang, S. Shi, D. Tao, and Z. Tu, "Redistributing low-frequency words: Making the most of monolingual data in non-autoregressive translation," in *ACL*, 2022.

[24] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *ECCV*, 2018, pp. 201–216.

[25] Y. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "UNITER: universal image-text representation learning," in *ECCV*, ser. Lecture Notes in Computer Science, vol. 12375. Springer, 2020, pp. 104–120.

[26] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeurIPS*, 2019.

[27] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: improving visual-semantic embeddings with hard negatives," in *BMVC*, 2018, p. 12.

[28] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *ICCV*. IEEE, 2019, pp. 3712–3721.

[29] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *ICML*, 2021.

[30] Z.-Y. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, C. Zhu, N. V. Peng, Z. Liu, and M. Zeng, "An empirical study of training end-to-end vision-and-language transformers," in *CVPR*, June 2022.

[31] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *ICCV*, 2019, pp. 4653–4661.

[32] Y. Wu, S. Wang, G. Song, and Q. Huang, "Learning fragment self-attention embeddings for image-text matching," in *ACM Multimedia*. ACM, 2019, pp. 2088–2096.

[33] L. Qu, M. Liu, D. Cao, L. Nie, and Q. Tian, "Context-aware multi-view summarization network for image-text matching," in *ACM Multimedia*. ACM, 2020, pp. 1047–1055.

[34] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, "Cross-modal retrieval via deep and bidirectional representation learning," *IEEE Trans. Multim.*, vol. 18, no. 7, pp. 1363–1377, 2016.

[35] Z. Shao, X. Zhang, M. Fang, Z. Lin, J. Wang, and C. Ding, "Learning granularity-unified representations for text-to-image person re-identification," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5566–5574.

[36] J. Rao, F. Wang, L. Ding, S. Qi, Y. Zhan, W. Liu, and D. Tao, "Where does the performance improvement come from - a reproducibility concern about image-text retrieval," in *SIGIR*, 2022.

[37] Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, and J. Shao, "Camp: Cross-modal adaptive message passing for text-image retrieval," in *ICCV*, 2019.

[38] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.

[39] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter," *CoRR*, vol. abs/1910.01108, 2019.

[40] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "Mobilebert: a compact task-agnostic BERT for resource-limited devices," in *ACL*, 2020, pp. 2158–2170.

[41] J. Rao, X. Meng, L. Ding, S. Qi, and D. Tao, "Parameter-efficient and student-friendly knowledge distillation," *CoRR*, vol. abs/2205.15308, 2022.

[42]

[43] L. Ding, D. Wu, and D. Tao, "The usyd-jd speech translation system for iwslt2021," *ArXiv*, vol. abs/2107.11572, 2021.

[44] C. Zan, K. Peng, L. Ding, B. Qiu, B. Liu, S. He, Q. Lu, Z. Zhang, C. Liu, W. Liu, Y. Zhan, and D. Tao, "Vega-mt: The jd explore academy translation system for wmt22," in *WMT@EMNLP*, 2022, pp. 1–12.

[45] Z. Fang, J. Wang, X. Hu, L. Wang, Y. Yang, and Z. Liu, "Compressing visual-linguistic model via knowledge distillation," in *ICCV*, October 2021, pp. 1428–1438.

[46] J. Rao, T. Qian, S. Qi, Y. Wu, Q. Liao, and X. Wang, "Student can also be a good teacher: Extracting knowledge from vision-and-language model for cross-modal retrieval," in *CIKM*, 2021.

[47] X. Gu, T. Lin, W. Kuo, and Y. Cui, "Zero-shot detection via vision and language knowledge distillation," *CoRR*, vol. abs/2104.13921, 2021.

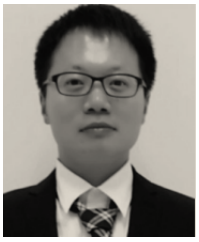
[48] L. Li, Y. Lin, S. Ren, P. Li, J. Zhou, and X. Sun, "Dynamic knowledge distillation for pre-trained language models," in *EMNLP*, 2021, pp. 379–389.

[49] S. Tang, L. Feng, W. Shao, Z. Kuang, W. Zhang, and Z. Lu, "Learning efficient detector with semi-supervised adaptive distillation," in *BMVC*, 2019, p. 215.
[50] Y. Zhang, Z. Lan, Y. Dai, F. Zeng, Y. Bai, J. Chang, and Y. Wei, "Prime-aware adaptive distillation," in *ECCV*, ser. Lecture Notes in Computer Science, vol. 12364. Springer, 2020, pp. 658–674.
[51] M. Fang, Y. Li, and T. Cohn, "Learning how to active learn: A deep reinforcement learning approach," in *EMNLP*, 2017, pp. 595–605.
[52] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao, "Panda: Prompt transfer meets knowledge distillation for efficient model adaptation," *ArXiv*, vol. abs/2208.10160, 2022.
[53] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *ICCV*, 2017, pp. 2999–3007.
[54] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *NIPS*, 2016, pp. 1849–1857.
[55] W. Li, C. Gao, G. Niu, X. Xiao, H. Liu, J. Liu, H. Wu, and H. Wang, "UNIMO: towards unified-modal understanding and generation via cross-modal contrastive learning," in *ACL*, 2021, pp. 2592–2607.
[56] A. Shrivastava, A. Gupta, and R. B. Girshick, "Training region-based object detectors with online hard example mining," in *CVPR*, 2016, pp. 761–769.
[57] B. Li, Y. Liu, and X. Wang, "Gradient harmonized single-stage detector," in *AAAI*, 2019, pp. 8577–8584.
[58] Q. Zhang, Z. Lei, Z. Zhang, and S. Z. Li, "Context-aware attention network for image-text retrieval," in *CVPR*, 2020, pp. 3533–3542.
[59] B. Wang, L. Ding, Q. Zhong, X. Li, and D. Tao, "A contrastive cross-channel data augmentation framework for aspect-based sentiment analysis," *ArXiv*, 2022.
[60] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao, "E2s2: Encoding-enhanced sequence-to-sequence pretraining for language understanding and generation," *ArXiv*, vol. abs/2205.14912, 2022.
[61] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *AISTATS*, 2010, pp. 297–304.
[62] T. Kim, J. Oh, N. Kim, S. Cho, and S. Yun, "Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation," in *IJCAI*, 2021, pp. 2628–2635.
[63]
[64] J. D. Robinson, C. Chuang, S. Sra, and S. Jegelka, "Contrastive learning with hard negative samples," in *ICLR*. OpenReview.net, 2021.
[65] B. Settles, "Active learning literature survey," 2009.
[66] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *ICML*, 2016, pp. 1050–1059.
[67] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *NIPS*, vol. 30, 2017.
[68] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *ECCV*, vol. 8693. Springer, 2014, pp. 740–755.
[69] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, 2014.
[70] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 664–676, 2017.
[71] B. Zhang, H. Hu, V. Jain, E. Ie, and F. Sha, "Learning to represent image and text with denotation graph," in *EMNLP*, 2020, pp. 823–839.
[72] H. Zhou, L. Song, J. Chen, Y. Zhou, G. Wang, J. Yuan, and Q. Zhang, "Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective," in *ICLR*. OpenReview.net, 2021.
[73] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*. OpenReview.net, 2020.
[74] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

Jun Rao is currently pursuing the M.S. degree in Harbin Institute of Technology, shenzhen. His long-term research goal is to build socially intelligent embodied agents with the ability to perceive and engage in **multimodal** human communication. As steps towards this goal, his research focuses on 1) the fundamentals of **multimodal learning**, specifically the representation, translation, fusion, and alignment of heterogeneous data sources, 2) human-centered **language, vision**, and their applications, 3) the real-world deployment of **efficiency** involves both the amount of **computation** and **data** required for (pre-)training and using NLP models.

Liang Ding received Ph.D. from the University of Sydney. He is currently an algorithm scientist with JD.com and leading the NLP research group at JD Explore Academy. He works on deep learning for NLP, including language model pretraining, language understanding, generation, and translation. He published over 30 research papers in NLP/AI, including ACL, EMNLP, NAACL, COLING, ICLR, AAAI, SIGIR, and CVPR, and importantly, some of his works were successfully applied to the industry. He served as Area Chair and Session Chair for ACL 2022 and SDM 2021. He won many AI challenges, including SuperGLUE/ GLUE, WMT2022, IWSLT 2021, WMT 2020, and WMT 2019. Liang led the team to be the first to outperform human performance (in Dec. 2021) on two challenging tasks and then got first place (in Jan. 2022) with an average score of 91.3 on the general language understanding evaluation (GLUE) benchmark. Afterward, Liang's team developed the Vega-v2 model, which sat atop the SuperGLUE leaderboard (in Oct. 2022).

Shuhan Qi Shuhan Qi received his M.S. and Ph.D. from Harbin Institute of Technology, and was a visiting scholar at the National University of Sinjapore. Now he is an associate researcher in the School of Computer Science and Technology, Harbin Institute of Technology. His main part-time positions include double-appointed assistant researcher in Network Intelligence Department of Pengcheng Laboratory and deputy director of Internet Application Technology Engineering Laboratory of Shenzhen Development and Reform Commission. He works at the Computer Application Research Center of Harbin Institute of Technology. His main research area is multimedia information retrieval and machine learning, and he has been engaged in the research of multimedia brand information analysis for a long time. He has published more than 30 papers in IEEE Transaction on Multimedia, SIGIR, ACM MM and other important international journals and conferences. He is also a member and reviewer of many famous international conferences and journals, including IEEE TMM, IEEE TNNLS, IEEE TKDE, IJCAI and other international top journals and conferences.



Meng Fang received the Ph.D. degree from the University of Technology, Sydney, Australia. He was a research fellow with the School of Computing and Information Systems, University of Melbourne. He is currently an assistant professor with the Department of Computer Science, University of Liverpool, the United Kingdom. His research focuses on natural language processing and reinforcement learning.

Yang Liu received the B.Eng. degree in computer science from Ocean University of China, Qingdao, China, the M.Sc. degree in software engineering from Peking University, Beijing, China, and the Ph.D. degree in computer science from University of Oxford, Oxford, U.K., in July 2018, under the guidance of Prof. A. Simpson. He is currently an Assistant Professor with the Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. He is interested in security and privacy problems and, in particular, the privacy issues related to mobile and IoT devices.

Li Shen received his Ph.D. in school of mathematics, South China University of Technology in 2017. He is currently a research scientist at JD Explore Academy, China. Previously, he was a research scientist at Tencent AI Lab, China. His research interests include theory and algorithms for large scale convex/nonconvex/minimax optimization problems, and their applications in statistical machine learning, deep learning, reinforcement learning, and game theory.

Dacheng Tao (F'15) is currently the president of JD Explore Academy and a Senior Vice President of JD.com. He is also an advisor and chief scientist of the digital science institute at the University of Sydney. He mainly applies statistics and mathematics to artificial intelligence and data science. His research interests range across computer vision, data science, image processing, machine learning, and video surveillance. His research results have been presented in one monograph and 500+ publications in prestigious journals and at prominent conferences, such as IEEE T-PAMI, T-NNLS, T-IP, JMLR, IJCV, NIPS, ICML, CVPR, ICCV, ECCV, ICDM, and ACM SIGKDD, with several Best Paper awards, such as the Best Theory/Algorithm Paper Runner Up Award at IEEE ICDM07, the Best Student Paper Award at IEEE ICDM13, the Distinguished Student Paper Award at the 2017 IJCAI, the 2014 ICDM 10-Year Highest-Impact Paper Award, and the 2017 IEEE Signal Processing Society Best Paper Award. He was a recipient of the 2015 Australian Scopus-Eureka Prize, the 2015 ACS Gold Disruptor Award, and the 2015 UTS Vice-Chancellor's Medal for Exceptional Research. He is a Fellow of AAAS, OSA, IAPR, and SPIE.