Designing Distillation Losses for Effective Knowledge Transfer from SupCon Representations

Yujin Lee
Department of Artificial Intelligence
Hanbat National University
Daejeon, Republic of Korea
yujin@edu.hanbat.ac.kr

Janghun Hyeon
Department of Artificial Intelligence
Hanbat National University
Daejeon, Republic of Korea
jhhyeon@hanbat.ac.kr

Yunho Jeon†
Department of Artificial Intelligence
Hanbat National University
Daejeon, Republic of Korea
yhjeon@hanbat.ac.kr

Abstract—**This study proposes a knowledge distillation method(KD method) to effectively transfer the expressive structure of a teacher model, pre-trained via Supervised Contrastive Learning (SupCon), to a student model. Existing logit-based KD methods suffer from the limitation of failing to sufficiently convey the rich expressive structure learned by the teacher model. To address this, this study analyzed four KD approaches: *(i)* normalized embedding-based distillation, *(ii)* cosine similarity-based distillation, *(iii)* structural distillation utilizing the Gram matrix, and *(iv)* contrastive-aware distillation directly mimicking the SupCon architecture. Experiments conducted on CIFAR-10 and ImageNet-1k using ResNet-based Teacher-Student settings revealed that the SupCon-based teacher consistently outperformed the conventional Cross-Entropy-based teacher. Furthermore, the proposed structure-preserving KD methods achieved higher accuracy than existing KD approaches. Notably, SupCon-based distillation demonstrated the greatest performance improvement among all methods while reliably transferring complex representation structures. Furthermore, expression analysis utilizing Centered Kernel Alignment (CKA) quantitatively evaluated how each distillation method alters expression similarity between teacher and student models. This research experimentally demonstrates that structure-preserving distillation is essential for the student model to effectively leverage the structural expression advantages possessed by the SupCon-based teacher.**

Index Terms—**Knowledge Distillation, Supervised Contrastive Learning, Contrastive Learning, Representation Alignment, Structure-Preserving Distillation**

I. Introduction

As the scale of deep learning models has expanded rapidly, large-scale models with high expressive power and generalization capabilities are demonstrating outstanding performance across various visual recognition tasks. Alongside this model advancement, learning paradigms for more efficiently utilizing the rich expressive structures learned by large models are also evolving rapidly. Among these, contrastive learning has established itself as a key technique for maximizing expression learning in large models, demonstrating performance surpassing traditional cross-entropy-based learning in both self-supervised and supervised settings.

Supervised Contrastive Learning (SupCon), in particular, learns structural representations by strongly clustering samples of the same class and maximizing the distance between different classes. This characteristic enables SupCon-based teacher models to form high-dimensional embedding spaces that reflect inter-class relationships and directionality, going beyond simple soft labels or logit distributions. Recently, large-scale models have actively adopted such contrastive objectives, leading to a trend where contrastive learning-based teacher models are increasingly utilized [1], [2].

However, the proliferation of SupCon-based teachers introduces new challenges. Existing knowledge distillation methods, designed based on soft label imitation, have limitations in effectively transferring the structural representations learned through contrastive learning to the student model [3]. In particular, while SupCon-based teachers richly contain structural information such as intra-class cohesion, inter-class separation, and directionality, existing KD methods lack mechanisms to utilize this information. Consequently, applying the contrastive teacher directly to distillation yields limited performance gains.

A structure-preserving knowledge distillation method suited to the era of large-scale models trained via contrastive learning is required. The expressive structure of a contrastive learning-based teacher is fundamentally different from that of conventional CE-based teachers, and distillation methods failing to reflect these structural characteristics cannot operate more effectively. Thus, contrastive-aware KD, designed to mimic and preserve contrastive representations, is essential.

We have preliminarily investigated contrastive-aware knowledge distillation methods under SupCon-based teachers, with a focus on the CIFAR-10 and CIFAR-100 benchmarks [4], [5]. In that study, we demonstrated that directly mimicking contrastive structures via SupCon-based KD leads to consistent performance gains over conventional logit-based distillation. However, the analysis was limited to relatively small-scale datasets and primarily centered on contrastive objective alignment.

This work extends the previous study in both scope and depth. We scale the evaluation to ImageNet-1k, a large-scale and more challenging benchmark, to examine whether contrastive-aware distillation remains effective under realistic

*This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00240379).

†Corresponding author

high-capacity settings. And we provide a comprehensive analysis of representation alignment using performance metrics and CKA-based similarity analysis, offering deeper insights into the relationship between structural preservation and downstream accuracy.

This research addresses these issues by proposing diverse KD methods designed to enable the student model to effectively reproduce the structural representations of the SupCon-based teacher model. Their effectiveness is empirically analyzed on CIFAR-10 and ImageNet-1k.

II. Related Work

Knowledge Distillation In situations where memory and computational resources are insufficient to train or perform inference with a large model, knowledge distillation (KD) offers a way to reduce the number of parameters and computational cost of a smaller student model while preserving as much performance as possible from a larger teacher model. Hinton et al. [6] proposed training a high-performing teacher model first, and then guiding a lightweight student model to mimic the teacher's soft labels. Through this process, the student model learns not only the hard ground-truth labels but also the teacher's decision boundaries, inter-class relationships, and the probability structure of non-target classes. As a result, KD improves inference speed and reduces memory consumption by shrinking the model size, while still maintaining performance close to that of the teacher model. However, existing KD studies have paid relatively little attention to teacher models trained with contrastive learning, including those trained via Supervised Contrastive Learning (SupCon)

Supervised Contrastive learning Self-supervised learning (SSL) methods such as SimCLR [7] can effectively learn feature representations even in the absence of labeled data. SimCLR constructs a positive pair by applying two different augmentations to the same input sample, while treating all remaining samples within the batch as negative pairs. However, this instance-discrimination objective has a critical limitation in that it cannot exploit other samples belonging to the same semantic class as additional positive pairs. Supervised Contrastive Learning (SupCon) [8] addresses this issue by explicitly leveraging class-label information available in supervised datasets. In SupCon, not only the augmented views of an anchor sample but also all other samples that share its class label are considered positive pairs, whereas samples from different classes are treated as negatives. This formulation enables each anchor to form a richer and more diverse set of positive pairs, facilitating the learning of more robust and discriminative representations. As a result, SupCon encourages tighter intra-class clustering and greater inter-class separation in the embedding space, and has been shown to outperform conventional supervised learning methods that rely solely on cross-entropy loss [8]. SupCon thus forms a structural embedding space that strongly attracts instances of the same class while pushing apart instances of different classes, resulting in a highly organized representation geometry. Such clear class boundary and structural separation may provide a more explicit and well-defined representation target for the student. Motivated by this observation, we investigate suitable knowledge distillation methodologies for SupCon-based teacher.

III. Method

In this section, we describe the distillation methods evaluated in our experiments. In conventional cross entropy-based training, knowledge distillation typically encourages the student model to directly mimic the teacher's logits. However, such an approach is insufficient for transferring the rich representational structure learned by the teacher. To address this limitation, we design a new contrastive-aware distillation loss that more effectively transfers the representational strengths of a teacher model pretrained with supervised contrastive learning. Fig. 1 illustrates the overall distillation architecture and training pipeline.

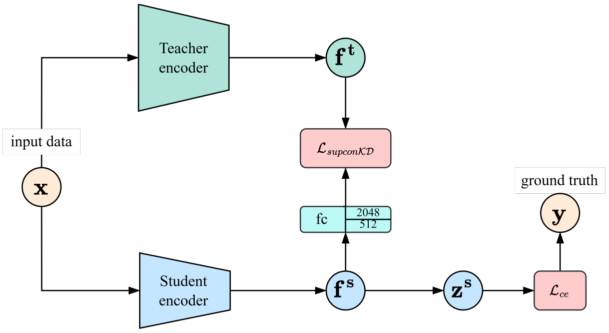

Fig. 1. Network flow illustrating the student embeddings used in the SupCon distillation objective

A. Logit-Based Mean Squared Error (MSE) Distillation

Logit-based distillation using the mean squared error(MSE) between the teacher and student logits is a simple and intuitive baseline for knowledge distillation. This approach encourages the student model to approximate the teacher's output responses by minimizing the squared difference between their logits, thereby transferring class-level prediction information.

To achieve this, the student's logits z_s are encouraged to closely match the teacher's logits z_t by minimizing the MSE between the two, as expressed in Eq. (1):

$$\mathcal{L}_{KD}^{base} = \|z_t - z_s\|_2^2 \tag{1}$$

This formulation is simple yet effective, as it enables the student model to learn the structural similarity of the teacher's representations at the logit level rather than relying solely on the output probability distribution.

B. Normalized Embedding-Based Distillation

In supervised Contrastive learning, embedding vectors are typically L2-normalized prior to computing the contrastive loss, projecting all representations onto the unit hypershpere. Motivated by this formulation, one can define a knowledge distillation loss using the MSE between normalized embedding vectors. Let $\bar{z}_t$ denote the normalized teacher embedding and

$\bar{z_s}$ denote the normalized student embedding, Distillation is then performed by minimizing the MSE between these two normalized vectors, as shown in Eq. (2):

$$\mathcal{L}_{KD}^{norm} = \|\bar{z_t} - \bar{z_s}\|_2^2 \tag{2}$$

This metric promotes symmetric and stable alignment between the two representation spaces and serves as an effective distillation strategy, particularly in contrastive learning-based representation learning frameworks.

C. Cosine Similarity-Based Distillation

In supervised contrastive learning, embedding vectors are first L2-normalized, and their inner products are interpreted as cosine similarities in order to maximize the similarity between positive samples. Inspired by this principle, cosine similarity can be employed as a loss function for KD. Cosine similarity measures the directional alignment between two vectors, where values closer to 1 indicate that the vectors point in nearly the same direction. The distillation objective is defined as the negative cosine similarity between these two vectors, as shown in Eq. (3):

$$\mathcal{L}_{KD}^{cosinesimilarity} = \frac{z_t \cdot z_s}{\|z_t\| \, \|z_s\|} \tag{3}$$

This formulation encourages the student representation to align its direction with that of the teacher by maximizing the cosine similarity (equivalently, minimizing its negative). In order words, the goal is to match the orientation of the student embedding to that of the teacher embedding.

Cosine-based distillation is particularly effective in representation learning scenarios where directional information is more relevant than vector magnitude, making it a strong strategy for direction-focused embedding alignment.

D. Gram matrix-based Distillation

Yim et al. [9] demonstrated that using the Gram matrix of features obtained between two layers allows distilling inter-channel correlations—that is, structural relationships between layers—rather than directly matching feature values themselves. This approach was designed to work well even between Teacher-Student models with capacity gaps, using less rigid constraints than directly matching feature maps. The Gram-based structural knowledge distillation loss is defined as follows:

$$\mathcal{L}_{KD}^{gram} = \left\| z_t \cdot {z_t}^{\top} - z_s \cdot {z_s}^{\top} \right\|_2^2 \tag{4}$$

The gram matrix is relatively invariant to changes in channel order or feature scale, so the student model does not need to reproduce the exact same feature shape as the teacher model. This increases the student's expressive freedom and prevents excessive constraints. Furthermore, by summarizing the geometric structure of representations through second-order statistics between features, it can more reliably convey the flow of the teacher's problem-solving process. By learning these structural relationships, the student does not merely mimic the teacher's features but acquires the teacher's own method of processing input. Considering these characteristics, we include Gram-based distillation as one of the structural KD methods in a simple comparative experiment to evaluate its performance.

E. SupCon-based Distillation

Simply distilling the teacher model's feature in a direct manner is often insufficient for transferring the class-wise structural relationships that the teacher learns through SupCon. To address this limitation, we apply a SupCon-based distillation objective in which the logits of the teacher and student to more faithfully replicate both the inter-class and intra-class representational structures captured by the teacher. The SupCon-based distillation loss is defined as follows:

$$\mathcal{L}_{KD}^{SupCon} = \sum_{t \in T} - \frac{1}{|P(t)|} \sum_{p \in P(t)} \log \frac{\exp\left(z_t \cdot z_p / \tau\right)}{\sum_{s \in S} \exp\left(z_t \cdot z_s / \tau\right)} \tag{5}$$

In Eq. (4):

- T denotes the set of teacher input samples,
- S denotes the set of student input samples,
- $P(t) = s \in S | y_t = y_s$ represents the positive set where teacher and student samples share the same label,
- z_t, z_s and z_p correspond to the teacher logits, student logits, and positive student logits, respectively.

This loss function maximizes the similarity between the teacher embedding z_t (serving as the anchor) and the student embeddings z_p that belong to the same class, while simultaneously increasing their separation form student model is guided to replicate the teacher's contrastive structure-both the positive-negative relationships and the geometric organization of the embedding space-naturally and effectively.

F. Final Loss Function

To enable the student model to effectively imitate the representational structure learned by the teacher through supervised contrastive learning, we propose several forms of knowledge distillation losses. The final objective for the student model is defined as follows:

$$\mathcal{L} = \mathcal{L}_{ce} + \alpha \mathcal{L}_{KD} \tag{6}$$

Here $\mathcal{L}_{ce}$ denotes the cross-entropy loss used to train the classifier module, and $\mathcal{L}_{KD}$ refers to one of the knowledge distillation losses described above. The hyperparameter α controls the relative contribution of the two terms, allowing the student model to balance classification performance and representation learning. Note that unlike simple logit-based KD, Gram/SupCon-based KD loss involves calculating correlations or similarities between representations (e.g., Gram matrices, batch similarity matrices), which may require additional computation and memory usage during training. However, since the structure of the student model remains unchanged, the computational complexity during inference remains identical to the baseline.

IV. Experiments

A. Dataset

The experiments were conducted using two benchmark image datasets: CIFAR-10, which consists of 10 classes, and ImageNet-1k, which contains 1,000 classes.

B. Experimental Model Setup

For the CIFAR-10 experiments, we used ResNet-50 as the teacher model and ResNet-18 as the student model. For the ImageNet-1k experiments, ResNet-50 and ResNet-34 were used as the teacher and the student models, respectively.

The hyperparameter α, which controls the weighting of the knowledge distillation loss, was adjusted separately for each loss type to obtain optimal performance. For CIFAR-10, all experiments were conducted five times with different random seeds, and the reported results correspond to the mean performance, with the values in parentheses indicating the standard deviation. For ImageNet-1k, each experiment was repeated three times, and results are reported as the mean with the standard deviation shown in parentheses.

C. Quantitative Results

TABLE I
Experimental Results on CIFAR-10

Model	**Cross Entropy**	**SupCon**
Teacher (ResNet50)	93.51 (± 0.14)	94.02 (± 0.09)
Student (ResNet18)	92.84 (± 0.44)	93.29 (± 0.21)

TABLE II
Results of KD Methods

Method	**KD from CE teacher**	**KD from SupCon teacher**
Baseline	93.79 (± 0.14)	94.04 (± 0.11)
Norm KD	-	94.10 (± 0.16)
Cosine KD	-	94.32 (± 0.11)
Gram KD	-	94.37 (± 0.13)
SupConKD	-	94.45 (± 0.09)

Table I shows the accuracy of the teacher and student models trained with CE loss and SupCon loss. The results show that the model trained using SupCon loss achieved better performance than CE loss, suggesting its suitability as a teacher model for subsequent knowledge distillation experiments.

Table II summarizes the results of training student models using various knowledge distillation loss functions applied in this study, including the existing KD baseline, with the teacher model trained via Supervised Contrastive Learning as the baseline. The KD baseline consistently outperforms the baseline student model, confirming that knowledge distillation is effective in improving the accuracy of the student model. Furthermore, when knowledge distillation was performed from a teacher model trained via supervised contrastive learning, the student model's accuracy improved by an additional 0.25% compared to using a teacher model trained via cross-entropy. All distillation methods showed performance improvements over the baseline student model. Notably, applying the supervised contrastive knowledge distillation loss function achieved a 1.16% accuracy improvement over the baseline student model. These results suggest that supervised contrastive learning enables the teacher model to learn a richer representational structure, which can be effectively transferred to the student model through knowledge distillation.

TABLE III
Experimental Results on ImageNet-1k

Model	**CE**	**SupCon**
Teacher (ResNet50)	75.82 (± 0.01)	76.34 (± 0.19)
Student (ResNet34)	73.44 (± 0.13)	73.95 (± 0.04)

TABLE IV
Results of KD Methods

Method	**KD from CE teacher**	**KD from SupCon teacher**
Baseline	73.93 (± 0.03)	73.7 (± 0.11)
Norm KD	-	73.84 (± 0.11)
Cosine KD	-	74.05 (± 0.09)
Gram KD	-	74.07 (± 0.00)
SupConKD	-	74.08 (± 0.00)

Table III shows the results comparing the performance of the teacher model (ResNet50) and the student model (ResNet34) on ImageNet-1k using CE-based learning versus SupCon-based learning. The teacher trained with SupCon achieved +0.52% higher accuracy than the CE-based teacher, confirming that the contrastive objective learns richer representational structures. The student model also showed a +0.51% performance improvement when trained with SupCon compared to CE. This suggests that the representation structure learned by SupCon provides a direct benefit to the student model even on complex datasets like ImageNet.

Table IV summarizes the experimental results for various knowledge distillation (KD) loss functions. For SupCon-based knowledge distillation, normalization was experimentally applied to the student model's feature vectors to reflect the characteristics of contrastive learning. And the effect of the normalization is analyzed in Section IV-D.

Applying the KD base resulted in a +0.26% performance improvement over the baseline student when using the CE-based teacher, but when using the SupCon-based teacher, performance actually decreased by 0.25% compared to the baseline student (73.95% → 73.70%). This occurred because the SupCon-based teacher learned a more complex and high-dimensional representation space than the CE teacher, and the existing soft-label-based KD failed to effectively utilize this information. In other words, a significant capacity gap emerged between the teacher and student on ImageNet, revealing the limitations of existing KD methods that directly transfer SupCon-based representations. ImageNet features a large number of classes, extremely limited positive pairs, and an explosive increase in negative pairs. This structure forms a high-dimensional space that maximizes inter-class separation and strongly demands intra-class compactness in the teacher embedding space. Conversely, CIFAR-10 has fewer classes and abundant positive pairs, resulting in a relatively simple teacher structure that the student could readily accommodate. Due to these structural differences, KD bases were effective on CIFAR-10, but on ImageNet, the student struggled to handle the teacher's contrastive representations, leading to decreased performance.

Norm-based KD achieved performance similar to the baseline model (73.84%). In contrast, structure-based KD methods generally showed improved performance compared to the baseline student model. Cosine similarity-based KD achieved an accuracy of 74.05% by reliably aligning the directional information of the student representation. Gram matrix-based KD induced stable knowledge transfer through structural constraints based on second-order statistics between features, achieving an accuracy of 74.07%. The highest accuracy was observed in SupCon-based KD (74.08%), which can be interpreted as a result of more directly transferring the structural representation of the teacher model, learned through contrastive learning, to the student model.

Overall, the ImageNet experimental results demonstrate that the structural advantages possessed by the SupCon-based teacher are difficult to fully leverage with existing KD bases, suggesting that structure-preserving KD is a more suitable strategy for utilizing contrastive teachers.

D. Ablation Study: Effect of Embedding Normalization in SupCon-Based Distillation

TABLE V
IMPACT OF EMBEDDING NORMALIZATION ON SUPCON-BASED DISTILLATION. NORMALIZATION STABILIZES TRAINING, BUT THE CLASSIFIER HEAD REMAINS UNDERTRAINED DUE TO LIMITED LOGIT SUPERVISION.

Configuration	**Top-1 Accuracy**
Without normalization	NaN
Normalized teacher feature	NaN
Normalized student feature	74.08
Both normalized	73.70

Table V presents the results of an ablation study analyzing performance changes based on the application location of embedding normalization in SupCon-based knowledge distillation. In contrast learning, training relies on cosine similarity, which is sensitive to the scale of feature vectors. Therefore, embedding normalization is commonly used as a preprocessing step to ensure learning stability. Accordingly, this experiment systematically compared whether normalization was applied to the feature vectors of the teacher and student models.

Without normalization, learning failed to converge and NaN values occurred, demonstrating that the SupCon-based loss is highly sensitive to changes in embedding scale. Learning instability was similarly observed when only the teacher features were normalized, suggesting that stable distillation is difficult when the distribution of the student embeddings is uncontrolled. Conversely, applying normalization only to the student feature vectors yielded the highest performance (74.08%), confirming that normalizing the student embeddings plays a crucial role in SupCon-based distillation.

Interestingly, performance actually decreased (73.70%) when both teacher and student features were normalized. This is interpreted as the directionality in the embedding space being stably aligned through normalization, but the classifier head failing to learn sufficiently due to insufficient direct supervision signals for the output logits. In other words, embedding normalization in SupCon-based distillation contributes to learning stability, but it demonstrates that simple output alignment alone has limitations in fully conveying the structural representation information learned by the teacher.

E. Representation Analysis

To analyze how expression alignment between teacher and student changes with knowledge distillation loss, we employed Centered Kernel Alignment (CKA), a quantitative metric for expression similarity. CKA evaluates the similarity of expression structures based on preserving pairwise relationships between two feature spaces. It enables stable comparisons even when architectures or learning objectives differ, making it widely used in neural representation analysis. A higher CKA score indicates that the student model more faithfully reproduces the teacher's representational geometry. In this analysis, we quantitatively evaluated Teacher–Student representation alignment for three knowledge distillation methods that demonstrated improved performance compared to the baseline student model.

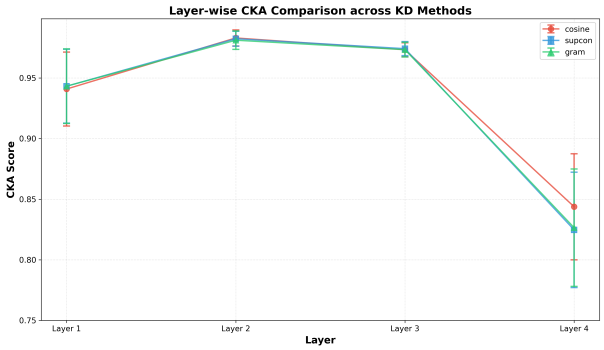

Fig. 2. Layer-wise CKA comparison across KD methods. Error bars indicate variation across blocks.

- **Cosine Similarity-based KD** achieves high similarity between teacher and student representations by inducing direct directional alignment between embeddings. However, this high CKA may arise from over-aligning representations, potentially simplifying the complex representation distributions or relational structures learned by the teacher.

- **Gram matrix-based KD** aligns relationships between representations through the structure of second-order statistics (covariance) across channels. It maintains relatively high CKA while imposing less stringent alignment constraints than cosine similarity-based KD, demonstrating the characteristic of preserving a certain level of structural information while avoiding excessive alignment.
- **SupCon-based KD** exhibited the lowest CKA value. This is because it constrains only the relative relational structure between representations through contrastive learning objectives, rather than directly aligning teacher representations one-to-one. This loose alignment approach can be interpreted as allowing a more flexible alignment centered on relationships within the representation space, rather than directly mimicking the teacher representations.

The CKA comparison across layers shows that while all methods maintain high alignment in the initial layers, notable differences emerge in deeper layers. Specifically, the decrease in CKA observed in the upper layers for SupCon-based KD can be interpreted as an intentional relaxation of alignment constraints, granting the student greater expressive freedom at more abstract levels. In contrast, Gram matrix-based KD explicitly aligns structural correlations across layers, resulting in relatively high CKA values even in deeper representations. This difference highlights a key distinction between the two approaches: Gram KD enforces explicit structural matching via secondary statistics, while SupCon-based KD achieves implicit structural alignment through contrasting embedding interactions. Therefore, the lower CKA values in upper layers of SupCon-based KD do not indicate inferior alignment; rather, they reflect a flexible knowledge transfer mechanism accommodating the student's limited capacity.

V. Conclusion

This study analyzed various knowledge distillation techniques to effectively transfer the structural representation advantages learned by a SupCon-based teacher model to a student model and verified their effectiveness. Experimental results confirmed that existing soft label-based KD may lead to performance degradation by failing to fully utilize the complex representation space learned by the SupCon teacher model. Conversely, knowledge distillation techniques preserving structural information achieved more stable performance improvements by enhancing expression alignment between teacher and student models.

Comparing distillation methods with different structural constraints revealed that Cosine Similarity KD tended to exhibit unstable learning or inconsistent performance gains due to its strong alignment constraint. Gram matrix-based KD enabled stable knowledge transfer through structural constraints based on inter-channel correlations, proving effective for classification performance improvement. Conversely, SupCon-based KD exhibited relatively low CKA, which can be interpreted as a result of relaxed structural constraints. This expressive freedom acted more favorably within the student model's limited capacity, ultimately leading to performance gains. This analysis suggests that in scenarios with a capacity gap between teacher and student models, balancing constraints and expressive freedom is more important than excessive alignment.

In summary, our experiments experimentally demonstrate that effectively utilizing the high-dimensional, structured representations provided by SupCon-based teachers requires knowledge distillation strategies centered on structural alignment, rather than simple logit imitation. This implies that distillation design considering both the nature of the representation structure and the capacity of the student model is crucial in future distillation settings utilizing large-scale models based on contrastive learning as teachers. It also provides meaningful direction for research on contrastive learning, nonlinear representation structures, and knowledge distillation between teachers and students.

References

[1] J. Lee, D. Das, M. Hayat, S. Choi, K. Hwang, and F. Porikli, "Customkd: Customizing large vision foundation for edge model improvement via knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[2] S. Yu, S. Kwak, H. Jang, J. Jeong, J. Huang, J. Shin, and S. Xie, "Representation alignment for generation: Training diffusion transformers is easier than you think," in *Proc. ICLR*, 2025. arXiv:2410.06940.

[3] W. Zhu, X. Zhou, P. Zhu, Y. Wang, and Q. Hu, "Ckd: Contrastive knowledge distillation from a sample-wise perspective," *arXiv preprint arXiv:2404.14109*, 2024.

[4] Y. Lee and Y. Jeon, "A study on knowledge distillation from teachers trained by supervised contrastive learning," in *Proceedings of the Annual Conference of the Institute of Electronics and Information Engineers*, (Jeju, Republic of Korea), June 2024.

[5] Y. Lee and Y. Jeon, "A study on effective knowledge distillation using a teacher trained with supervised contrastive learning," in *Proceedings of the Annual Conference of the Institute of Electronics and Information Engineers*, (Jeju, Republic of Korea), June 2025.

[6] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[8] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *arXiv preprint arXiv:2004.11362*, 2021.

[9] J. Yim, D. Joo, J. Bae, and J. 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 (CVPR)*, pp. 4133–4141, 2017.