Confidence Matters: Enhancing Medical Image Classification Through Uncertainty-Driven Contrastive Self-Distillation

Saurabh Sharma[0000−0001−9709−5912], Atul Kumar[0009−0007−9339−9872], and Joydeep Chandra[0000−0001−5994−9024]

Indian Institute of Technology Patna, Patna, Bihar 801103, India
{saurabh_2021cs30,atul_2101ai08,joydeep}@iitp.ac.in

Abstract. The scarcity of data in medical image classification using deep learning often leads to overfitting the training data. Research indicates that self-distillation techniques, particularly those employing mean teacher ensembling, can alleviate this issue. However, directly transferring knowledge distillation (KD) from computer vision to medical image classification yields subpar results due to higher intra-class variance and class imbalance in medical images. This can cause supervised and contrastive learning-based solutions to become biased towards the majority class, resulting in misclassification. To address this, we propose UDCD, an uncertainty-driven contrastive learning-based self-distillation framework that regulates the transfer of contrastive and supervised knowledge, ensuring only relevant knowledge is transferred from the teacher to the student for fine-grained knowledge transfer. By controlling the outcome of the transferable contrastive and teacher's supervised knowledge based on confidence levels, our framework better classifies images under higher intra- and inter-relation constraints with class imbalance raised due to data scarcity, distilling only useful knowledge to the student. Extensive experiments conducted on benchmark datasets such as HAM10000 and APTOS validate the superiority of our proposed method. The code is available at https://github.com/philsaurabh/UDCD_MICCAI.

Keywords: Medical Image Classification · Knowledge Distillation · Uncertainty · Contrastive Learning · Relational Knowledge.

1 Introduction and Background

Recent advances in deep learning have significantly enhanced medical image analysis for computer-aided diagnosis (CAD). Convolutional Neural Networks (CNNs) have become pivotal tools due to their robust feature extraction and classification capabilities, particularly in CAD [21]. However, challenges like data privacy and limited availability often lead CNN to overfit. Knowledge Distillation (KD) [3], involving the regularization of a shallow model with knowledge from a more complex teacher model [11, 23, 15, 17], has shown promise in enhancing diagnostic accuracy in medical image analysis [21, 10, 19] among other prevalent

methods. Among KD techniques, the self-ensembling mean-teacher paradigm has emerged as a leading methodology, offering heightened generalizability even in data-scarce settings [8]. Nonetheless, challenges persist due to high inter-class resemblance, intra-class variance, and class imbalances in medical image datasets, which can confound differentiation and predispose models to overfit.

Recent research efforts aimed at addressing classification challenges have deployed various strategies, including distillation of inter-batch relationships [10], contrastive knowledge preservation [21], and relational knowledge augmentation [19]. These advancements have leveraged contrastive knowledge [21] and relational self-supervision [19]. However, challenges persist for majority classes, where significant class imbalances in positive samples can lead to divergent behavior, particularly evident in [21] where learning may slow down due to the high fraction of positive samples. Additionally, self-supervised learning methods [19, 22] face difficulties in scenarios characterized by high inter-class similarity and the absence of explicit class labels during training, making them susceptible to misclassification. This vulnerability stems from the inherent nature of self-supervised learning, which does not utilize explicit class labels during training [18, 7].

To address the outlined challenges, we introduce UDCD, an Uncertainty-Driven Contrastive Self-Distillation framework. UDCD leverages the mean-teacher method for self-distillation, controlling knowledge transfer based on the confidence levels of the teacher model's predictions. It utilizes Supervised Contrastive Relation Matrices to extract supervised contrastive learning from both models, transferring this knowledge with a weighting contingent upon the teacher's confidence in each prediction. Integrating uncertainty as a measure of confidence enhances label exploitation and facilitates pertinent knowledge transfer, thereby improving model performance in both inter and intra-class relationships. Additionally, UDCD incorporates relational knowledge among contrastive relations, emphasizing relative distances among learned representations of each class rather than distribution gaps. This approach reduces the model's reliance on dataset characteristics during the knowledge transfer phase, bolstering its resilience and generalizability. The proposed UDCD framework contributes significantly by (1) Introducing the extraction of Contrastive Relation Matrices, which contain supervised contrastive learning-based discriminative features crucial and more effective for distinguishing examples across different classes. (2) Proposing a novel Uncertainty-Driven Contrastive Distillation mechanism between the contrastive predictions of the student and a mean teacher model, facilitating the transfer of relevant information while mitigating biased learning. (3) Replacing distribution gap-based knowledge distillation with relative knowledge transfer between the teacher's and student's supervised contrastive knowledge, thereby reducing dependency on dataset characteristics and addressing class imbalance more efficiently. Our methodology undergoes thorough evaluation on two prominent datasets, APTOS [6], and HAM10000 [16, 1]. Through comprehensive analysis, our proposed approach demonstrates a performance superiority ranging from 5% to 11% compared to state-of-the-art techniques across various performance

metrics, particularly excelling in scenarios marked by class imbalance, high inter-class similarity, and intra-class variance. Also, our method maintains robust performance even with increasing class numbers and remains effective in scenarios with limited data availability.

2 Methodology

The UDCD framework, depicted in Fig. 1, introduces a student model and a mean-teacher model for optimization. Stochastic gradient descent optimizes the student model, while the teacher weights ω' are updated using exponential moving average (EMA) based on the student weights ω. Image augmentation yields two distinct images, x_s and x_t, each undergoing different perturbations. The student and teacher models extract feature representations $\boldsymbol{z_s}$ and $\boldsymbol{z_t}$ and predict output probabilities p_s and p_t, respectively. Supervised by weighted cross-entropy loss L_{WCE} and KL divergence [3] L_{KL} with respect to p_t, the student's prediction p_s ensures consistency with the teacher. Structural information alignment is enforced using the L_{SCL} loss, extracting supervised contrastive discriminative features. Additionally, a Contrastive Relation Matrix (CRM) $C(x_s)$ $(C(x_t))$ is formed for each model, and the L_{CRA} loss aligns these CRMs between the teacher and student models. The supplementary material provides further details on the algorithm.

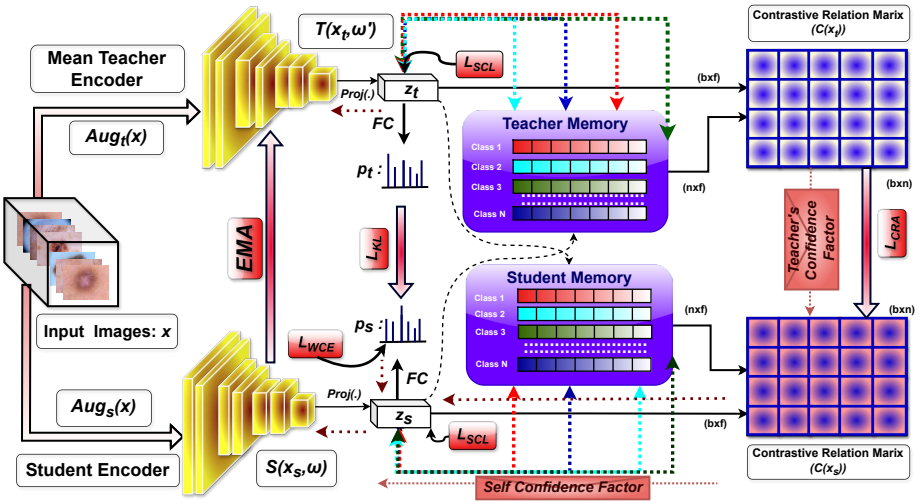


Fig. 1: Proposed UDCD Framework, where directions of the arrows show the flow of the framework from input to output(Dashed Maroon lines denote the back-propagation).

2.1 Contrastive Knowledge Extraction

Our study diverges from recent methods such as CRCKD [21] and SSDKD [19] due to observed limitations regarding erratic behavior stemming from the scarcity or abundance of positive samples. Instead, it focuses on traditional supervised contrastive learning (SCL) [7] integrated with knowledge distillation. SCL enhances contrastive prediction accuracy by incorporating label information alongside samples. The methodology involves applying diverse augmentations to an image, processing it through both teacher and student encoders, and passing it through projection layers for linear transformation. This yields contrastive projection embeddings $\boldsymbol{z_t}$ and $\boldsymbol{z_s}$, normalized to the unit hypersphere via L2 normalization. These embeddings serve two purposes: extracting predictions and distilling logits via KL-Divergence loss [3] and deriving supervised contrastive discriminative features for contrastive knowledge transfer to regularize the consistency of structural knowledge between the teacher and student by enhancing intra-class similarity and inter-class divergence. Inspired by previous work, SCL loss for the teacher and student models, denoted as $L_{SCL}^{(t)}$ and $L_{SCL}^{(s)}$, is defined as follows:

$$L_{SCL}^{(t)} = \sum_{i \in I} L_{SCL,i}^{(t)} = \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{e^{\boldsymbol{z_t(i)} \cdot \boldsymbol{z_p}/\tau}}{\sum_{k \in K_s(i)} e^{\boldsymbol{z_t(i)} \cdot \boldsymbol{z_k}/\tau}}, \tag{1}$$

$$L_{SCL}^{(s)} = \sum_{i \in I} L_{SCL,i}^{(s)} = \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{e^{\boldsymbol{z_s(i)} \cdot \boldsymbol{z_p}/\tau}}{\sum_{k \in K_t(i)} e^{\boldsymbol{z_s(i)} \cdot \boldsymbol{z_k}/\tau}}. \tag{2}$$

In this formulation, $i \in I = \{1, 2, \ldots, 2N\}$ denotes the index of an augmented sample, where $\boldsymbol{z(i)}$ represents its embedding, and $P(i)$ denotes the set of indices of all positive samples in the multi-viewed batch concerning the i^{th} sample, with $|P(i)|$ indicating its cardinality. To access a large sample of negative examples for improved contrastive learning, as suggested in prior works, we adopt the approach of [20] to construct a memory bank $\mathcal{M} \in \mathbb{R}^{b \times d}$ that stores the d-dimensional embeddings of all b training images in a batch ($\mathcal{M}_s$ for the student model, $\mathcal{M}_t$ for the teacher model). Additionally, $K_t(i)$ and $K_s(i)$ represent the set of all negative samples relative to the i^{th} sample in the teacher's and student's memory, respectively. Here, τ denotes the temperature constant.

2.2 Categorical Relation Alignment

To address potential class bias resulting from high class imbalance in supervised contrastive learning, we introduce relational alignment between contrastive features. Unlike CRCKD [21], which focuses on distribution learning, our approach aims to align categorical relations by optimizing the relative distance between category representations of teacher and student models using Huber loss [5]. Each class is represented by a single anchor, irrespective of the number of images in that class, mitigating bias induced by class imbalance, similar to the

approach in [21]. To extract the Contrastive Relation Matrix (CRM), we utilize supervised contrastive embeddings and class embeddings from the memory for each batch. The memory matrix has dimensions $n \times f$ (where n is the number of classes and f is the size of the projection head embedding), and for each batch of size b, we obtain a matrix of dimensions $b \times f$. Anchors are extracted following the approach proposed in [21], aggregating the i^{th} class sample to obtain the i^{th} class anchors. Next, we calculate the contrastive relation to create a Categorical Relation Matrix (CRM) using the similarity of each image projection after applying softmax in a batch, as follows:

$$A_s(i) = \frac{1}{|A(i)|} \sum_{m_s \in A(i)} m_s, \qquad R(x_s, A_s(i)) = \frac{e^{\boldsymbol{z_s(i)} \cdot A_s(i)}}{\sum_{i=1}^{n} e^{\boldsymbol{z_s(i)} \cdot A_s(i)}} \tag{3}$$

Here, $|A(i)|$ denotes the number of samples of the i^{th} class. We do the same with the teacher to get $A(i)$ and $R(x_t, A_s(i))$.

Next, for each triplet $(p, q, r) \in x_s$ for the student and $\in x_t$ for the teacher, we calculate the relational unit vector as follows.

$$R_{pq}(i) = \frac{R(p, A(i)) - R(q, A(i))}{\|R(p, A(i)) - R(q, A(i))\|_2}, \text{ and } R_{pr}(i) = \frac{R(p, A(i)) - R(r, A(i))}{\|R(p, A(i)) - R(r, A(i))\|_2}. \tag{4}$$

Finally, we calculate the Categorical Relational Alignment (CRA) loss L_{CRA} as follows:

$$L_{CRA} = \sum_{(p,q.r) \in x} \delta(\phi(R^t_{pq}(i), R^t_{pr}(i)), \phi(R^s_{pq}(i), R^s_{pr}(i))). \tag{5}$$

Here, R^s and R^t represent the relation units for the teacher and student, respectively, while ϕ and δ denote the distance functions. It is advisable to employ Huber loss for these functions.

2.3 Uncertainty Driven Learning Task

To estimate confidence, we utilize uncertainty as a metric [9]. Two types of confidence measures are considered: 1) Teacher's confidence, computed during knowledge transfer from the Contrastive Relation Matrix (CRM) using L_{CRA}, and 2) Self-confidence, which regulates the student's own contrastive learning to minimize bias towards irrelevant majority class learning. The aim is to encourage the student to rely more on the traditional learning approach when the teacher exhibits less confidence in some examples and also when the student's contrastive learning adversely affects its performance. For each data sample x, the probability distribution of the output class with respect to the class label y for the student version is computed as $P(y|x) = \text{softmax}(S(x))$. The prediction uncertainty, for instance, x, is determined by:

$$u(x) = Entropy(softmax\ (S(x))) = -\sum_y P(y|x) \log P(y|x). \tag{6}$$

We introduce two learnable confidence parameters, ψ_1 and ψ_2, as follows (for teacher and student ψ^t and ψ^s are used respectively):

$$\psi_1 = \frac{u(x)}{U}, \psi_2 = 1 - \frac{u(x)}{U}, \tag{7}$$

where U is the factor utilized to normalize the weight to the range $[0, 1]$. Considering it, for the overall learning process, our method operates as follows:

$$L = (1 + \psi_2^s + \psi_2^t)L_{WCE} + \lambda_1.L_{KL} + \psi_1^t.\lambda_2.L_{SCL}^{(S)} + \psi_1^t.\lambda_3.L_{CRA} \tag{8}$$

In this context, L_{WCE} represents weighted cross-entropy, where each class's weight is inversely proportional to its cardinality. L_{KL} denotes the traditional KL-divergence loss [3], and λ_1 to λ_3 are the corresponding hyperparameters. During testing, both the mean teacher and the projection heads are omitted, ensuring that the inference time matches that of the vanilla student model.

3 Experiments and Results

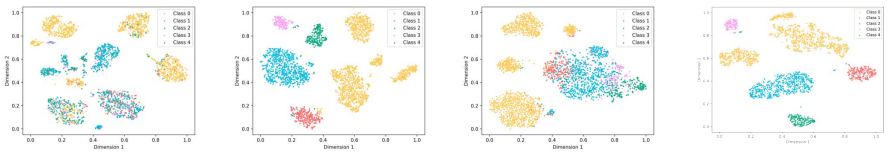
3.1 Experimental Settings

We leverage two widely recognized datasets, APTOS [6] and HAM10000 [16], drawing insights from previous studies [21, 19], to conduct a comparative analysis. Our evaluation encompasses accuracy (ACC), F1-score, balanced mean accuracy (BMA), recall (REC), and mean average precision (AP) metrics against existing baselines. We explore four distinct scenarios to assess the efficacy of UDSD: (i) its impact on inter and intra-class relations, (ii) its performance in addressing class imbalance, (iii) its effectiveness compared to state-of-the-art KD methods, and (iv) its performance under conditions of data scarcity. Using the Densenet architecture for both teacher and student models, we also investigate UDCD's effectiveness with various backbone networks, maintaining methodological consistency with established settings in the literature [21, 19].

3.2 Analysis on Inter and Intra Class Relations with class imbalance

To evaluate the effectiveness of UDCD across different challenges, we utilize two datasets: the APTOS dataset for examining inter-class similarity and intra-class variance and the HAM10000 dataset for assessing performance under high class imbalance. For inter-class similarity and intra-class variance analysis on the APTOS dataset, t-SNE plots are employed, comparing UDCD with medical image-specific baselines CRCKD and SSD-KD [21, 19]. Notably, UDCD demonstrates significant separation in class-wise clusters, particularly excelling in capturing distinct patterns and discriminating minority class outliers as shown in Fig. 2. This proficiency is attributed to confidence-based knowledge transfer, enhancing precise class discrimination. Regarding high class imbalance evaluation on the HAM10000 dataset, various samples with different distributions are generated,

and four variations are implemented to comprehensively examine the impact on model performance using different learning mechanisms. The examination of results in Table 1 indicates a significant improvement in UDCD's methodology over the nearest-performing baseline concerning class imbalance (2-5% in accuracy and 1-11% in AP for different cases), possibly due to the extraction and transfer of dark knowledge [22] facilitated by contrastive learning in UDCD.



(a) Original Data (b) CRCKD (c) SSD-KD (d) UDCD

Fig. 2: t-SNE plots for the APTOS dataset using different training techniques, where the data points of each class are shown in different colors.

Table 1: Results on HAM10000 dataset with different data distributions. **Bold** and Underline represents the best and second-best results respectively.

Methods↓	$\rho = 144$ (Severe Class Imbalance)				$\rho = 58$ (Original Data)			
	ACC(%)	**AP(%)**	**REC(%)**	**F1(%)**	**ACC(%)**	**AP(%)**	**REC(%)**	**F1(%)**
KD [3]	86.56	77.43	71.53	74.36	85.05	74.20	74.62	74.41
SCL-IKD [14]	86.12	78.23	72.64	75.33	85.69	74.91	71.68	75.30
CRCKD [21]	88.11	81.46	76.17	76.29	85.90	76.55	76.63	76.59
SSD-KD [19]	79.23	61.87	74.40	67.56	85.42	73.20	**87.04**	79.52
UDCD	**90.24**	**87.84**	**75.90**	**79.68**	**90.28**	**85.12**	82.91	**83.83**
Methods↓	$\rho = 17$ (Less Class Imbalance)				$\rho = 9$ (Very less Class Imbalance)			
	ACC(%)	**AP(%)**	**REC(%)**	**F1(%)**	**ACC(%)**	**AP(%)**	**REC(%)**	**F1(%)**
KD [3]	76.07	60.95	71.95	65.99	75.21	66.57	77.12	71.54
SCL-IKD [14]	76.69	61.01	70.70	65.50	74.30	66.09	78.82	71.90
CRCKD [21]	78.88	60.03	78.55	66.33	74.43	66.33	**83.37**	72.12
SSD-KD [19]	79.08	60.58	73.95	66.60	81.92	68.71	77.15	**72.96**
UDCD	**81.60**	**72.89**	**80.36**	**70.57**	**84.04**	**69.66**	78.12	72.57

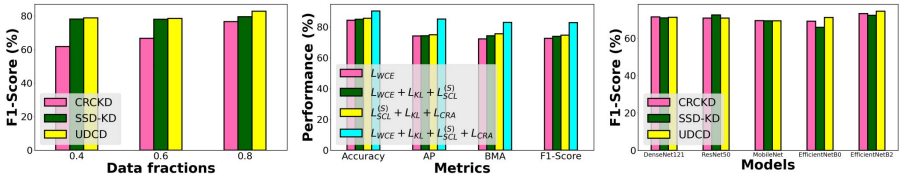
3.3 Other Analyses and studies

We evaluate the effectiveness of UDCD by conducting a comparative analysis with state-of-the-art knowledge distillation methods customized to tackle the aforementioned challenges, employing Densenet121. The outcomes of this evaluation, presented in Table 2, reveal UDCD's significant margin of superiority

(up to 11% gain in precision) over existing baselines owing to its unique learning mechanism. Additionally, we assess the effectiveness of UDCD in addressing data scarcity, particularly in real-world medical datasets like skin lesion classification (Additional visualizations are presented in the supplementary). Figure 3a demonstrates UDCD's superior performance even under significant data scarcity, possibly attributed to relational knowledge transfer with confidence-based knowledge transfer. Furthermore, Figure 3b presents an ablation study highlighting the significance of all components. Additionally, Figure 3c compares the performance of various backbone networks with recent relevant methods CRCKD [21] and SSD-KD [19], which shows significant improvement in F1-score for all different backbones including Resnet50 [2], Efficientnet [12], Mobilenet [13], etc for APTOS dataset.

Table 2: Performance comparison of baselines to the **UDCD** framework. **Bold** and Underline represent the best and second-best results respectively.

Methods↓	APTOS [6]				HAM10000 [16]			
	ACC(%)	**AP(%)**	**REC(%)**	**F1(%)**	**ACC(%)**	**AP(%)**	**REC(%)**	**F1(%)**
Scratch [4]	83.13	70.34	68.16	69.10	84.31	74.18	72.21	72.56
MTG	83.01	70.07	68.59	69.76	85.05	74.20	76.09	74.41
RKD [11]	83.77	71.30	71.26	70.89	88.72	79.60	80.26	79.47
SSKD [22]	83.49	71.15	71.53	71.25	84.42	74.99	**85.95**	79.68
CRCKD [21]	84.07	71.93	71.49	**71.45**	85.90	76.55	78.17	76.59
CRD [15]	84.09	71.75	69.30	70.22	85.35	74.43	76.45	74.81
SCL-IKD [14]	80.45	71.25	68.43	67.74	85.58	74.89	77.68	75.48
SSD-KD [19]	84.53	71.55	**72.90**	70.99	85.42	73.20	85.57	79.52
UDCD	**85.01**	**73.38**	71.42	71.31	**90.28**	**85.12**	82.91	**83.83**



(a) Few Shot Analysis (b) Ablation Study (c) Backbone Analysis

Fig. 3: Figures for other analyses and studies.

4 Conclusion

This paper introduces the Uncertainty-driven Contrastive Self-Distillation (UDCD) framework tailored for medical image classification, addressing challenges such

as high inter-class similarity and class imbalance. Three key innovations are proposed: 1) Contrastive Relation Matrix (CRM) for efficient contrastive feature extraction, 2) Relational Alignment for effective handling of data imbalance, and 3) Uncertainty-driven Supervised Contrastive Knowledge Transfer, mitigating irrelevant knowledge propagation from both student and teacher models. These innovations facilitate the distillation of rich structural knowledge from the mean-teacher model. Experimental evaluations on the HAM10000 [16] and APTOS [6] datasets demonstrate the superior effectiveness of UDCD compared to other knowledge distillation paradigms. While UDCD primarily focuses on image classification, its ability in other paradigms like NLP remains unexplored, and we leave this as a future scope for research.

Acknowledgments. The authors of this paper wish to express their gratitude for the assistance provided by the Prime Minister's Research Fellowship (PMRF) scheme of the Government of India, which facilitated the execution of this research work.

Disclosure of Interests. The authors have no competing interests to declare relevant to this article's content.

References

1. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 168–172. IEEE (2018)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
3. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. CoRR **abs/1503.02531** (2015), `http://arxiv.org/abs/1503.02531`
4. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
5. Huber, P.J.: Robust estimation of a location parameter. Annals of Mathematical Statistics **35**, 492–518 (1964)
6. Karthik, Maggie, S.D.: Aptos 2019 blindness detection (2019), `https://kaggle.com/competitions/aptos2019-blindness-detection`
7. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. p. . (2020)
8. Lee, Y., Willette, J.R., Kim, J., Lee, J., Hwang, S.J.: Exploring the role of mean teachers in self-supervised masked auto-encoders. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net (2023), `https://openreview.net/pdf?id=7sn6Vxp92xV`

9. Li, L., Lin, Y., Ren, S., Li, P., Zhou, J., Sun, X.: Dynamic knowledge distillation for pre-trained language models. In: Moens, M., Huang, X., Specia, L., Yih, S.W. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. pp. 379–389. Association for Computational Linguistics (2021). https://doi.org/10.18653/V1/2021.EMNLP-MAIN.31, https://doi.org/10.18653/v1/2021.emnlp-main.31
10. Liu, Q., Yu, L., Luo, L., Dou, Q., Heng, P.A.: Semi-supervised medical image classification with relation-driven self-ensembling model. IEEE transactions on medical imaging **39**(11), 3429–3440 (2020)
11. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3967–3976 (2019)
12. Qin, D., Bu, J.J., Liu, Z., Shen, X., Zhou, S., Gu, J.J., Wang, Z.H., Wu, L., Dai, H.F.: Efficient medical image segmentation based on knowledge distillation. IEEE Transactions on Medical Imaging **40**(12), 3820–3831 (2021)
13. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
14. Sharma, S., Lodhi, S.S., Chandra, J.: Scl-ikd: intermediate knowledge distillation via supervised contrastive representation learning. Applied Intelligence pp. 1–22 (2023)
15. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. p. . OpenReview.net (2020), https://openreview.net/forum?id=SkgpBJrtvS
16. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)
17. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1365–1374 (2019)
18. Wang, G., Wang, K., Wang, G., Torr, P.H., Lin, L.: Solving inefficiency of self-supervised representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9505–9515 (October 2021)
19. Wang, Y., Wang, Y., Cai, J., Lee, T.K., Miao, C., Wang, Z.J.: SSD-KD: A self-supervised diverse knowledge distillation method for lightweight skin lesion classification using dermoscopic images. Medical Image Anal. **84**, 102693 (2023). https://doi.org/10.1016/j.media.2022.102693, https://doi.org/10.1016/j.media.2022.102693
20. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 3733–3742. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00393
21. Xing, X., Hou, Y., Li, H., Yuan, Y., Li, H., Meng, M.Q.H.: Categorical relation-preserving contrastive knowledge distillation for medical image classification. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. pp. 163–173. Springer (2021)

22. Xu, G., Liu, Z., Li, X., Loy, C.C.: Knowledge distillation meets self-supervision. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX. pp. 588–604. Springer (2020)
23. Zhang, J., Xie, Y., Xia, Y., Shen, C.: Attention residual learning for skin lesion classification. IEEE transactions on medical imaging **38**(9), 2092–2103 (2019)