

Leveraging Different Learning Styles for Improved Knowledge Distillation in Biomedical Imaging

Usma Niyaz^a, Abhishek Singh Sambyal^a, Deepti R. Bathula^a

^aDepartment of Computer Science and Engineering, Indian Institute of Technology Ropar, Rupnagar, 140001, Punjab, India

Abstract

Learning style refers to a type of training mechanism adopted by an individual to gain new knowledge. As suggested by the VARK model, humans have different learning preferences, like Visual (V), Auditory (A), Read/Write (R), and Kinesthetic (K), for acquiring and effectively processing information. Our work endeavors to leverage this concept of knowledge diversification to improve the performance of model compression techniques like Knowledge Distillation (KD) and Mutual Learning (ML). Consequently, we use a single-teacher and two-student network in a unified framework that not only allows for the transfer of knowledge from teacher to students (KD) but also encourages collaborative learning between students (ML). Unlike the conventional approach, where the teacher shares the same knowledge in the form of predictions or feature representations with the student network, our proposed approach employs a more diversified strategy by training one student with predictions and the other with feature maps from the teacher. We further extend this knowledge diversification by facilitating the exchange of predictions and feature maps between the two student networks, enriching their learning experiences. We have conducted comprehensive experiments with three benchmark datasets for both classification and segmentation tasks using two different network architecture combinations. These experimental results demonstrate that knowledge diversification in a combined KD and ML framework outperforms conventional KD or ML techniques (with similar network configuration) that only use predictions with an average improvement of 2%. Furthermore, consistent improvement in performance across different tasks, with various network architectures, and over state-of-the-art techniques establishes the robustness and generalizability of the proposed model.

Keywords: Feature sharing, Model compression, Learning styles, Knowledge Distillation, Online distillation, Mutual learning, Teacher-student network, Multi-student network

1. Introduction

Undoubtedly, deep learning techniques perform exceptionally well in many domains, including medical diagnosis. However, the quest to achieve state-of-the-art performance has led to the development to highly complex neural networks. This restricts their application in many domains as they are computationally expensive to train and difficult to deploy. Specifically, most medical images are characterized by high resolution. While lightweight networks fail to capture the pixel-level spatial contextual information due to limited capacity, highly parameterized models provide significantly superior performance. However, these giant models are impractical for real-world applications requiring resource-constrained edge or embedded devices deployment. Consequently, model compression is an active area of research that aims to reduce the configuration complexity of state-of-the-art deep networks to enable their deployment in resource-limited domains without significant reduction in performance [12, 13].

It is a misconception that only large and highly complex models achieve the best performance [4, 20]. Many state-of-the-art

approaches have shown that strategically designed lightweight models can provide similar performance [6]. It is now known that a significant percentage of nodes in these models are redundant, and pruning these connections minimally affects the performance [56]. Several model compression techniques have been developed to simplify highly complex networks and substantially reduce the requirement for resources [3, 42]. These include Network Pruning [5], Quantization [13], Low-Rank Matrix Approximation [62], Knowledge Distillation (KD) [17], Deep Mutual Learning (DML) [61], and their variants. Alternatively, scaling methods such as random search optimization [24], and EfficientNet [47] aims to efficiently balance model complexity and performance by systematically adjusting various model dimensions. The ultimate goal of all these techniques is to optimize the network configuration without compromising the model's performance.

Network Pruning involves the elimination of nodes, weights, or layers that do not significantly influence performance. It speeds up the inference process by considering the parameters that matter the most for a specific task. Quantization-based approaches reduce the precision of numbers used to represent the neural network weights, and Low-rank approximation uses tensor decomposition to estimate the informative parameters of a network [8]. Tiny machine learning is another fast-growing

Email addresses: usma.20csz0015@iitrpr.ac.in (Usma Niyaz), abhishek.19csz0001@iitrpr.ac.in (Abhishek Singh Sambyal), bathula@iitrpr.ac.in (Deepti R. Bathula)

field that strives to bring the transformative power of machine learning to resource-constrained devices. It has shown significant potential for low-power applications in the vision and audio space [50].

Knowledge Distillation emerged as a potential and widely used model compression technique. It leverages the knowledge gained by a highly parameterized teacher network by sharing it with a lightweight student network such that the compact model approximates the performance of the teacher network. As an extension of the KD concept, several alternative approaches have been proposed that leverage additional supervision from the pre-trained teacher model, particularly emphasizing the intermediate layers [40, 52, 16]. Some of these approaches involve using spatial attention maps [57], while others explore the use of pairwise similarity patterns [30] or strive to maximize the mutual information between teacher and student features [2, 55]. Other variants of KD, like Evolutionary Knowledge Distillation [59] and Self Distillation [60], were also introduced to enhance the performance of the student networks further. Knowledge distillation is well exploited in many healthcare applications for classification [26, 1, 43, 23, 22, 32], localization [51, 19], and segmentation [39, 58]. It focuses on the problems of multi-modality and small datasets [21, 34] and is often collaboratively used with other approaches to boost the performance of student networks [25, 14, 53].

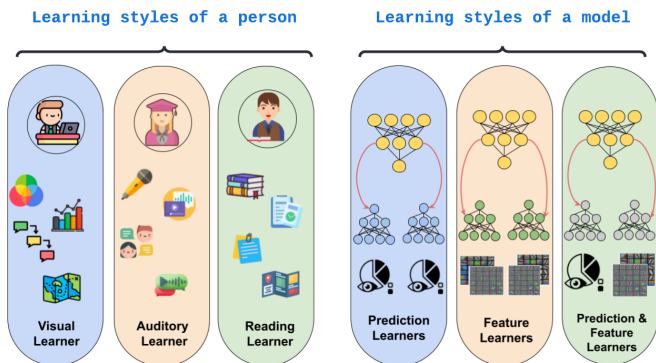


Figure 1: Illustration of a person’s different learning style adopted for training the deep learning model.

In these conventional techniques, knowledge transfer is typically limited to sharing predictions from a larger teacher network to a smaller student network (KD) or between two similar networks (ML). While alignment of predictions helps improve the performance, the logits have limited capacity to encapsulate complex insights. To address this limitation, Chen et al. [7] introduced the use of feature maps in knowledge distillation. Due to their ability to capture high-level semantic information, feature maps were found to be more effective than logits for image classification.

In this work, we extend the knowledge distillation paradigm with the concept of diverse learning styles from classroom dynamics. A learning style refers to a type of training mechanism that an individual prefers to use to gain new knowledge. For example, as depicted in Figure 1, the VARK model assumes four types of learners – Visual, Auditory, Reading & Writing, and

kinesthetic. Taking inspiration from this idea, we propose an enriched knowledge transfer protocol that incorporates the idea of different learning styles in terms of knowledge diversification. Consequently, we combine KD with ML in a single-teacher, multi-student framework to enable collaborative learning where the teacher imparts knowledge to the students, and the students also learn from each other. Unlike conventional KD techniques, where the teacher shares the same knowledge with all the students, we propose to train individual student networks with varying forms of information from the teacher. Similarly, students exchange different types of information in the form of final predictions and intermediate layer features.

The main contributions of this work are summarized as follows:

1. We propose an enhanced knowledge transfer protocol that incorporates the idea of different learning styles in terms of knowledge diversification.
2. In a single-teacher, multi-student network that simulates classroom dynamics, the teacher network imparts knowledge to the student networks in diverse formats, such as predictions to one student and feature maps to another. Further, we enrich each student’s learning process by facilitating the exchange of diversified knowledge among them.
3. Extensive experiments on metastatic tissue classification, brain tumor segmentation, and dermatological segmentation and classification tasks demonstrate the superiority of the proposed method over conventional knowledge distillation approaches.
4. Finally, the improvement afforded by the knowledge diversification strategy is attributed to the increased similarity of learned representations between higher layers of the teacher and student networks as measured by the Centered Kernel Alignment (CKA) metric.

2. Related work

2.1. Knowledge Distillation

Knowledge Distillation [17] is an approach introduced to transfer the knowledge in terms of probability outputs, p_i , from a complex, highly parameterized pre-trained teacher network $f(X, \phi)$ to a simple and compact student network $g(X, \theta)$ to achieve model compression while retaining the high performance of the teacher.

Given a training set with N samples $X = \{\mathbf{x}_i\}_{i=1}^N$ with corresponding labels $Y = \{y_i\}_{i=1}^N$, the teacher network $f(X, \phi)$, is trained on the ground truth labels. The probabilistic output of a teacher network for a sample x_i is defined as p_i given by the extended softmax as:

$$p_i^c = \frac{e^{\mathbf{z}^c / T}}{\sum_{c=1}^C e^{\mathbf{z}^c / T}} \quad \text{for } c = 1, 2, \dots, C \quad (1)$$

where \mathbf{z}^c corresponds to the logits, C is the number of classes, and T is the temperature parameter to get a smoother output

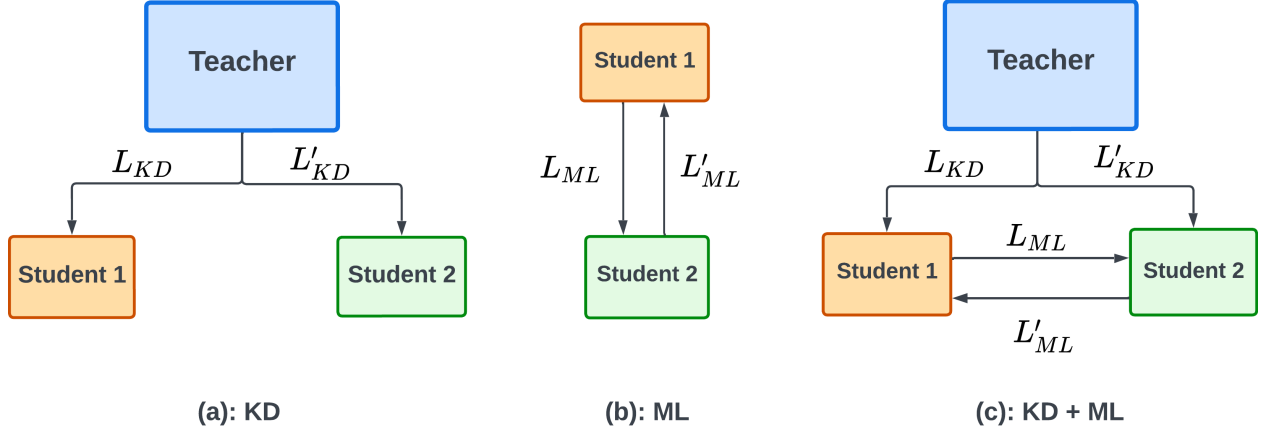


Figure 2: **KD** - Transfers knowledge in terms of soft labels from a large, pre-trained teacher to a compact student network; **ML** - Both networks are considered as students for information exchange; **KD + ML** - Training students with knowledge from teacher network as well as from each other. Conventionally, the same type of knowledge was shared with both student networks with $L_{KD} = L'_{KD}$ and $L_{ML} = L'_{ML}$. (The size of the block represents the complexity of a model, and color represents different architectures.)

Table 1: Combinations of distillation techniques and learning strategies employed: (a) KD – Knowledge Distillation only, (b) ML – Mutual Learning only, (c) combined KD and ML; V1 – sharing of final predictions only, V2 – sharing of features only, and a diverse knowledge paradigm V3 – sharing of predictions and features together.

(a)				(b)			
KD	V1	V2	V3	ML	V1	V2	V3
T → S ₁	Predictions	Features	Predictions	S ₁ → S ₂	Predictions	Features	Predictions
T → S ₂	Predictions	Features	Features	S ₂ → S ₁	Predictions	Features	Feature

(c)			
KD + ML	V1	V2	V3
T → S ₁ , S ₂ → S ₁	Predictions, Predictions	Features, Features	Features, Predictions
T → S ₂ , S ₁ → S ₂	Predictions, Predictions	Features, Features	Predictions, Features

probability distribution of the classes. Generally, the objective function for the teacher network is the standard *Cross-Entropy (CE) error* defined as:

$$L_\phi = L_{CE}(p_i, y_i) = -\sum_{i=1}^N (y_i \log(p_i) + (1 - y_i) \log(1 - p_i)) \quad (2)$$

Now, the student networks are trained on the combined loss of *Cross-Entropy (CE)*, and *Knowledge Distillation (KD)*, where the *CE* helps the student networks to adhere to the ground truth labels and *KD* assists them to align their learning with that of the teacher. Here, *Kullback Leibler (KL) divergence* [28] is used for L_{KD_p} to measure the correspondence between the teacher and student predictions p_i and s_i respectively as:

$$L_{KD_p} = D_{KL}(s_i || p_i) = \sum_{i=1}^N s_i(x_i) \log \frac{s_i(x_i)}{p_i(x_i)} \quad (3)$$

Finally, the loss function for the student network is the weighted (α) summation of the cross entropy (L_{CE}) and knowledge distillation (L_{KD_p}) terms:

$$L_\theta = \alpha L_{CE}(s_i, y_i) + (1 - \alpha) L_{KD_p}(s_i, p_i) \quad (4)$$

where hyperparameter α is used to balance the contributions of the hard target loss (L_{CE}) and soft target loss (L_{KD_p}) during

the distillation process for each student. The knowledge can be transferred in an online or offline manner from the teacher to the student networks. In offline knowledge distillation (KD (off)), training is done in two steps; first, the teacher network is pre-trained on a dataset, and then the knowledge is distilled to train the student network, whereas in online knowledge distillation (KD (on)) [10], both teacher and student networks are trained simultaneously in a single training step.

2.2. Deep Mutual Learning

Unlike knowledge distillation, mutual learning [61] is a two-way sharing of information as both networks are treated as student networks. Both the students can be of the same or different configuration. They teach each other collaboratively throughout the entire training process. The loss functions L_{θ_1} and L_{θ_2} for the two student networks $g_1(X, \theta_1)$ and $g_2(X, \theta_2)$ respectively are defined as

$$L_{\theta_1} = L_{CE}(s_1, \mathbf{y}) + L_{ML_p}(s_1, s_2) \quad (5)$$

$$L_{\theta_2} = L_{CE}(s_2, \mathbf{y}) + L_{ML_p}(s_2, s_1)$$

where s_k , $k \in \{1, 2\}$ is the predictions of the k th student network, and similar to *KD*, L_{ML_p} defined as the *Mutual Learning loss* is the *KL divergence* between the predictions of two students.

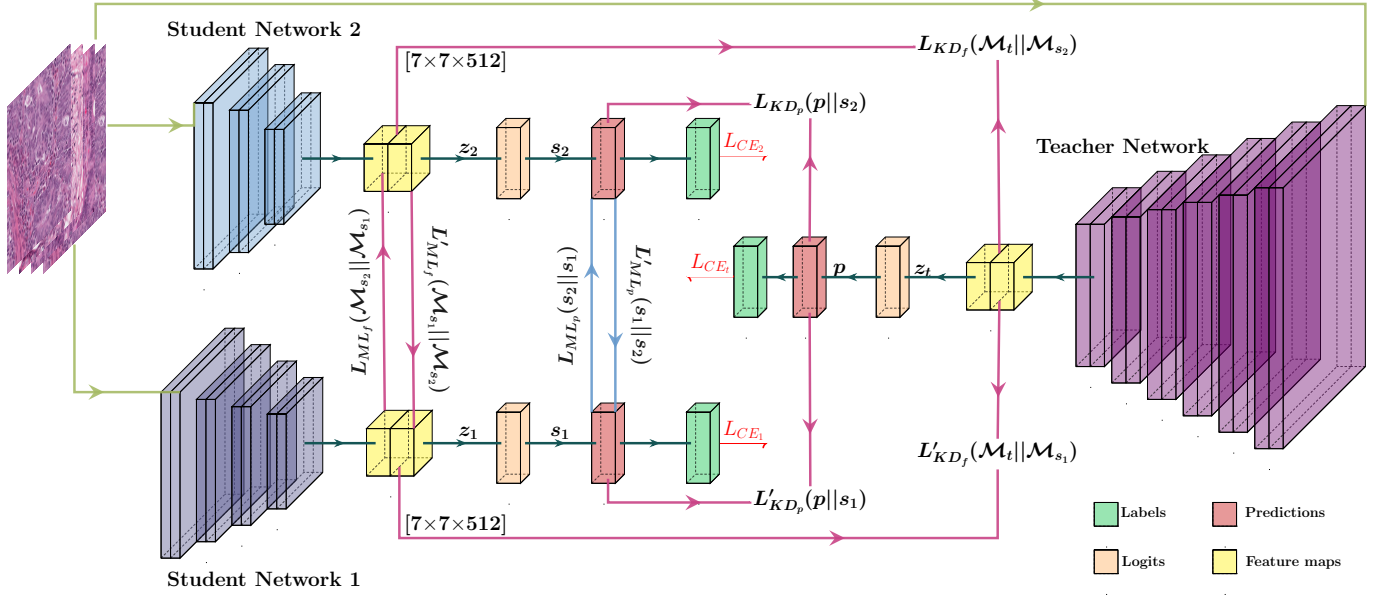


Figure 3: Overview of our proposed model that combines Knowledge Distillation (KD) with Mutual Learning (ML) in a single-teacher, two-student framework, leveraging a diversified knowledge-sharing strategy in the classification task. The loss terms L and L' represent the respective losses for each student, while L_{KD_p} and L_{KD_f} capture the knowledge distillation losses between the teacher and students over predictions and features, respectively. Similarly, L_{ML_p} and L_{ML_f} denote the mutual learning losses between the two students over predictions and features, respectively. Here ResNet50 is used as a teacher network and ResNet18 as a student network

Therefore, $L_{ML_p}(s_{k'}, s_k) = D_{KL}(s_{k'}||s_k)$, where the KL distance from $s_{k'}$ to s_k is computed using equation 3.

3. Method

3.1. Knowledge Distillation and Mutual Learning (KDML) with Knowledge Diversification

Figures 2 (a) and (b) depict the standard distillation techniques used for knowledge distillation and mutual learning. Extension of KD to a single teacher and multiple student networks is also quite standard. Recently, [38] proposed a combination of KD and ML, where in addition to sharing information by the teacher, students also exchange information, as shown in Figure 2 (c). In such configurations, the ensemble of student networks is used for the final prediction. In the current work, we investigate the benefit of knowledge diversification in four different distillation techniques- (i) Offline KD-only framework as shown in Figure 2 (a) (ii) Online KD-only framework (iii) ML-only framework with two students as shown in Figure 2 (b), and (iv) combined KD + ML frameworks with one teacher and two students as shown in Figure 2 (c).

Leveraging the idea that different learning styles can improve the understanding of learners, we propose to train individual student networks with different information from the teacher. The teacher shares final predictions with one student and intermediate layer features with another. Similarly, in the ML framework, students engage in information exchange, sharing predictions and intermediate-layer features with one another. To underscore the significance of knowledge diversification, we train each of the above-mentioned distillation techniques with three different

information-sharing strategies (V1, V2, and V3), as shown in Table 1. Furthermore, we use student networks with identical architectures to emphasize the influence of different learning styles.

We hypothesize that diversifying knowledge has the potential to enhance the efficacy of the three established distillation techniques, namely KD (off), KD (on), and ML. We anticipate that the most optimal performance can be achieved by combining the KD and ML configuration with the ensemble of two student networks trained with distinct information from the teacher and promoting the exchange of diverse knowledge between them.

3.2. Application in Classification Task

In our approach, which uses a combined KD + ML configuration with a knowledge diversification strategy to distill knowledge from the teacher while facilitating mutual learning among students, we define the respective loss functions as follows:

$$L_{\theta_1} = \alpha L_{CE}(s_1, y) + \beta L_{KD_f}(\mathcal{M}_{s_1}, \mathcal{M}_t) + \gamma L_{ML_p}(s_1, s_2) \quad (6)$$

$$L_{\theta_2} = \alpha' L_{CE}(s_2, y) + \beta' L'_{KD_p}(s_2, p) + \gamma' L'_{ML_f}(\mathcal{M}_{s_2}, \mathcal{M}_{s_1}) \quad (7)$$

where L'_{KD_p} and L_{ML_p} represent the same loss terms based on predictions as defined in equations 4 and 5 respectively. To encourage knowledge diversification, we introduce two supplementary loss terms, L_{KD_f} and L'_{ML_f} . These loss terms are constructed based on features shared by the teacher, represented

as \mathcal{M}_t , and the other student, denoted as \mathcal{M}_{s_k} . We define the feature map-based loss function as the *Mean Square Error (MSE)* between the feature maps of the corresponding networks. In general, for n number of feature maps, the *MSE* between two feature maps is defined as $\frac{1}{n} \sum_{i=1}^n (\hat{\mathcal{M}}_i - \mathcal{M}_i)^2$. The detailed depiction of our proposed approach with a knowledge diversification strategy is shown in Figure 3, and the network architecture details are given in Table 6 in the supplementary material. Refer to Table 5 (supplementary material) for a comprehensive view of the different loss terms involved in the distillation methods with various learning styles.

Using Equations 6 and 7, we can derive the loss functions for KD-only and ML-only configurations with knowledge diversification by setting weighing parameters $\gamma = \gamma' = 0$ and $\beta = \beta' = 0$ respectively. As each student is learning from different information, we use separate weighing parameters for individual terms of the loss function and optimize them using grid search. For a test sample x_i , we consider the ensemble classification probability, $\hat{s}(x_i)$, as the highest probability for a particular class across all student network predictions. This is given as $\hat{s}(x_i) = \max\{s_k(x_i)\}_{k=1}^K$.

Although sharing predictions is a common practice, a more detailed explanation is required for the sharing of feature information. With the assumption that the last layers of deep neural networks encode high-level semantic information, we propose to use the output of the teacher network’s last convolution layer as feature information to share with student networks. However, to enable this knowledge transfer, it is necessary to ensure that the student network’s convolutional block has an output feature map with dimensions matching the teacher network’s. This ensures an effective transfer of knowledge between the two networks. In KD configurations, where the teacher and student networks do not have layers with matching dimensions to share or compare the feature map information, an additional convolutional block can be added to the teacher network with an output dimension to match that of the student. This ensures the compactness of the student networks. A similar approach can also be adopted for ML configurations with non-identical student networks.

3.3. Application in Segmentation Task

In general, U-Net [41] is the preferred and most commonly used architecture for image segmentation. To capture essential feature information for sharing with student networks, we employ the output feature map from the initial convolution layer of the teacher’s decoder network, as depicted in Figure 8 in the supplementary material. It was found empirically that this layer, being close to the encoder, contains valuable semantic information. It helps generate more precise predictions when combined with the information from previous layers of the encoder through skip connections. Unlike classification task that uses cross-entropy loss (L_{CE}), we use a combination of *Focal loss (FL)* [31] and *Dice loss (DL)* [35], defined as L_{FD} , to train the networks with ground truth segmentation labels. The loss function, L_{FD} between the predicted $\hat{\mathbf{g}}$ and ground truth mask \mathbf{g} ,

is defined as:

$$L_{FD}(\hat{\mathbf{g}}, \mathbf{g}) = - \underbrace{\sum_{i=1}^N (1 - \hat{g}_i)^\tau \log(\hat{g}_i)}_{\text{Focal Loss}} + 1 - \underbrace{\frac{2 \sum_{i=1}^N \hat{g}_i g_i}{\sum_{i=1}^N \hat{g}_i^2 + \sum_{i=1}^N g_i^2}}_{\text{Dice Loss}} \quad (8)$$

where \hat{g}_i and g_i are the corresponding predicted probability and ground truth, and τ is the focusing parameter that enables the model to prioritize hard examples during training. The L_{FD} is most effective for segmentation as *Dice loss* handles unequal distribution of foreground-background elements, whereas *Focal loss* alleviates the problem of class imbalance by using a down-weighting mechanism to reduce the influence of easy examples that are well-classified and focus on hard examples that require additional attention during training. Consequently, the loss functions for the segmentation task using KD + ML configuration with a knowledge diversification are formulated as follows:

$$L_{\theta_1} = \alpha L_{FD}(s_1, \mathbf{y}) + \beta L_{KD_f}(\mathcal{M}_{s_1}, \mathcal{M}_t) + \gamma L_{ML_p}(s_1, s_2) \quad (9)$$

$$L_{\theta_2} = \alpha' L_{FD}(s_2, \mathbf{y}) + \beta' L'_{KD_p}(s_2, \mathbf{p}) + \gamma' L'_{ML_f}(\mathcal{M}_{s_2}, \mathcal{M}_{s_1}) \quad (10)$$

Similar to the classification task, using Equations 9 and 10, we can derive the segmentation loss functions for knowledge diversified KD-only and ML-only configurations by setting $\gamma = \gamma' = 0$ and $\beta = \beta' = 0$ respectively. Finally, the ensemble prediction mask is calculated as the union of individual student predictions, $\hat{y}(x_i) = \bigcup_{i=1}^n \{s_k(x_i)\}_{k=1}^K$.

4. Experiments

4.1. Dataset details

For the classification task, we used the Histopathologic Cancer Detection Dataset [18] consisting of 220k images for identifying metastatic tissue in histopathologic scans of lymph node sections. Due to limited computational resources, a balanced subset of 20,000 histological images was randomly selected from the repository for our experiment. Each image represents a patch of size 96×96 extracted from histopathologic scans of lymph node sections, annotated with a binary label indicating presence of metastatic tissue. For the segmentation task, we chose the Low-Grade Gliomas (LGG) Segmentation Dataset [29]. This dataset consists of 3929 MR brain images obtained from The Cancer Imaging Archive (TCIA) [48]. Each image has a resolution of 256×256 and a corresponding manual FLAIR abnormality segmentation mask. The datasets for both tasks are split into training, validation, and testing with a 75:10:15 ratio. We have conducted additional experiments using the HAM10000 dataset for both (a) Lesion Classification and (b) Lesion segmentation tasks. A more detailed analysis of these experiments is provided in the supplementary material (Section A.6).

4.2. Implementation details

We used two different network architecture combinations for our single teacher and two student configurations – (a) ResNet50 as a teacher with ResNet18 as students and (b) ResNet50 as a teacher and MobileNet as students. While these networks can be used directly for the classification task, they were used as backbones for the teacher and student networks in the U-Net framework for the segmentation task. ResNet50, ResNet18, and MobileNet are all pre-trained models equipped with approximately 26 million, 11 million, and 4.8 million parameters, respectively. Consequently, the ensemble of two ResNet18 students and the ensemble of two MobileNet students provide model compression of 15% and 63%, respectively, when compared to the ResNet50 teacher network.

The feature-sharing aspect of our proposed classification model requires that the dimensions of the feature maps of teacher and student networks match. The feature map of the last convolutional layer of ResNet50 is $7 \times 7 \times 2048$, ResNet18 and MobileNet have dimensions of $7 \times 7 \times 512$ and $7 \times 7 \times 1024$, respectively. To resolve this discrepancy, an additional convolutional block is introduced in the teacher network containing 1×1 convolutional layer with the number of filters to match the student – 512 for ResNet18 and 1024 for MobileNet. For the segmentation task, as the feature maps extracted from the first convolutional layer in the decoder of both the teacher and student networks have the same dimensions ($16 \times 16 \times 256$), no additional modifications are required.

We employed standard data augmentation techniques [37] widely recognized in the field of machine learning. These include horizontal flips, vertical flips, rotation, transpose, shift, and scale, along with normalization to introduce variations in orientation, position, scale, and intensity. By increasing the diversity, these techniques help counteract over-fitting and increase the robustness of the models. The optimal values of all the hyper-parameters for different distillation techniques and learning strategies were identified using a grid search. All the models were trained with Adam optimizer with a learning rate of 0.0001, batch size as 8, and the temperature parameter T as 2. These parameters are selected empirically. We report our models’ average and standard deviation of 3 runs for a more robust evaluation.

4.3. Experimental setup

We conducted a comprehensive set of experiments to evaluate the robustness and generalizability of our model. Firstly, we compared four different distillation techniques – (i) ML, (ii) KD (on), (iii) KD (off), and (iv) Combined KD and ML (KD + ML). Secondly, we explored the performance of these four techniques using three different learning strategies – V1 (predictions only), V2 (features only), and a diverse knowledge paradigm, V3 (both predictions and features). Different combinations of distillation and learning strategies lead to a total of 12 different variants. The performance of these 12 variants was evaluated for classification and segmentation tasks on three standard datasets. Finally, these experiments were repeated using two different teacher-student network architecture combinations, ResNet50-ResNet18 and ResNet50-MobileNet.

4.4. Evaluation metrics

To assess the effectiveness of the proposed knowledge distillation approach, we used accuracy for the classification models and Intersection-over-Union (IoU) and F1-score for the segmentation task as defined below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \quad (12)$$

$$IoU = \frac{TP}{TP + FP + FN} \quad (13)$$

Where TP, TN, FP, and FN represent the number of True Positive, True Negative, False Positive, and False Negative predictions, respectively. Accuracy measures the proportion of correct predictions to the total number of predictions, and IoU assesses the overlap between a model’s predicted and the ground truth segmentations.

Table 2: Performance comparison of baseline models on metastatic tissue classification and LGG Segmentation.

Model	Classification	Segmentation	
	Accuracy	IOU	F-score
ResNet50	94.35 ± 0.763	76.86 ± 0.791	86.93 ± 0.537
ResNet18	94.07 ± 0.436	75.06 ± 0.792	85.13 ± 0.561
MobileNet	91.38 ± 0.491	68.15 ± 0.287	80.24 ± 0.225

5. Results and Discussion

For a baseline comparison, we first evaluated the performance of standalone pre-trained ResNet50, ResNet18, and MobileNet models for both classification and segmentation tasks, as shown in Table 2. Additional baseline performance metrics on the HAM10000 dataset are presented in Table 13 in the supplementary material.

5.1. Classification Task

The results of metastatic tissue classification using ResNet50-ResNet18 combination for teacher and student networks are shown in Table 3. We observe that for each of the individual distillation techniques, the knowledge diversification paradigm (V3), where both predictions and features are shared, provides the best performance in terms of classification accuracy. Similarly, a comparison across different distillation techniques for a specific learning strategy shows that the combined KD + ML approach is superior, corroborating the findings of [38]. Expectedly, the combined KD + ML model trained with the knowledge diversification paradigm (V3) outperforms all other models. When compared with the conventional KD-only or ML-only models that share only predictions (V1), the proposed model provides an average improvement of 2% in classification accuracy. Lastly, it can also be observed that in addition to the

Table 3: Performance comparison of ResNet50-ResNet18 for classification accuracy for Histopathologic Cancer Detection dataset using four different distillation techniques: ML - Mutual Learning, KD (on) - online Knowledge Distillation, KD (off) - offline Knowledge Distillation, and KD + ML - combined KD & ML; and three different learning strategies - V1 (predictions only), V2 (features only) and a diverse knowledge paradigm, V3 (both predictions and features).

	V1 ($\alpha = 0.2, \alpha' = 0.2$) ($\beta = 0, \beta' = 0$) ($\gamma = 0.8, \gamma' = 0.8$)	V2 ($\alpha = 0.2, \alpha' = 0.2$) ($\beta = 0, \beta' = 0$) ($\gamma = 0.8, \gamma' = 0.8$)	V3 ($\alpha = 0.1, \alpha' = 0.2$) ($\beta = 0, \beta' = 0$) ($\gamma = 0.9, \gamma' = 0.8$)
ML			
S1	94.25 \pm 0.042	92.52 \pm 0.487	95.04 \pm 0.095
S2	94.15 \pm 0.442	93.52 \pm 0.219	94.89 \pm 0.245
Ensemble	94.22 \pm 0.155	94.19 \pm 0.106	95.36 \pm 0.239
	V1 ($\alpha = 0.2, \alpha' = 0.2$) ($\beta = 0.8, \beta' = 0.8$) ($\gamma = 0, \gamma' = 0$)	V2 ($\alpha = 0.2, \alpha' = 0.2$) ($\beta = 0.8, \beta' = 0.8$) ($\gamma = 0, \gamma' = 0$)	V3 ($\alpha = 0.2, \alpha' = 0.2$) ($\beta = 0.8, \beta' = 0.8$) ($\gamma = 0, \gamma' = 0$)
KD (off)			
T	94.35 \pm 0.763	94.35 \pm 0.763	94.35 \pm 0.763
S1	94.02 \pm 0.576	94.32 \pm 0.176	94.59 \pm 0.127
S2	94.31 \pm 0.134	94.25 \pm 0.245	94.87 \pm 0.530
Ensemble	94.75 \pm 0.954	94.68 \pm 0.176	95.45 \pm 0.490
	V1 ($\alpha = 0.2, \alpha' = 0.2$) ($\beta = 0.8, \beta' = 0.8$) ($\gamma = 0, \gamma' = 0$)	V2 ($\alpha = 0.2, \alpha' = 0.2$) ($\beta = 0.8, \beta' = 0.8$) ($\gamma = 0, \gamma' = 0$)	V3 ($\alpha = 0.1, \alpha' = 0.2$) ($\beta = 0.9, \beta' = 0.8$) ($\gamma = 0, \gamma' = 0$)
KD (on)			
T	94.43 \pm 0.353	94.15 \pm 1.312	95.43 \pm 0.707
S1	94.74 \pm 0.191	93.23 \pm 0.869	95.75 \pm 1.079
S2	94.55 \pm 0.281	94.38 \pm 0.912	95.23 \pm 0.438
Ensemble	95.18 \pm 0.162	95.06 \pm 0.776	96.15 \pm 0.707
	V1 ($\alpha = 0.1, \alpha' = 0.2$) ($\beta = 0.45, \beta' = 0.4$) ($\gamma = 0.45, \gamma' = 0.4$)	V2 ($\alpha = 0.2, \alpha' = 0.2$) ($\beta = 0.4, \beta' = 0.4$) ($\gamma = 0.4, \gamma' = 0.4$)	V3 ($\alpha = 0.2, \alpha' = 0.4$) ($\beta = 0.4, \beta' = 0.3$) ($\gamma = 0.4, \gamma' = 0.3$)
KD + ML			
T	93.21 \pm 2.341	94.46 \pm 0.309	95.75 \pm 0.756
S1	94.37 \pm 0.883	95.46 \pm 0.487	95.85 \pm 0.968
S2	94.71 \pm 0.518	95.06 \pm 0.353	95.37 \pm 0.353
Ensemble	95.25 \pm 0.360	96.15 \pm 0.219	96.68 \pm 0.584

ensemble accuracy, the V3 learning strategy also improves the performance of individual student networks.

To demonstrate the generalizability of our proposed method across different datasets, we repeated the above experiments using the HAM10000 dataset for the lesion classification task. Results of this experiment, presented in Table 14 of the supplementary material, depict similar trends in the performance of various models. Furthermore, to evaluate the robustness of our proposed approach across different architectures, the metastatic tissue classification experiments were repeated using the ResNet50-MobileNet combination for teacher and student networks. Results of these experiments are shown in Table 8 of the supplementary material. Notably, our proposed model demonstrated consistent improvement in performance across different network architectures and datasets, showcasing similar trends in the context of various distillation techniques, learning strategies, and their combinations.

5.2. Segmentation Task

The corresponding results of U-Net with ResNet50-ResNet18 network architectures for the segmentation task are reported in

Table 4. These results depict similar trends observed for the classification task, further emphasizing the importance of combining KD with ML and the influence of knowledge diversification. Moreover, it establishes the generalizability of the proposed approach to more than one type of task.

To better appreciate the significance of the proposed approach, we provide a qualitative comparison of different distillation techniques and information-sharing strategies using individual student predictions for some hard-to-segment test samples. Figure 4 extensively compares all the models for three test samples. In general, it can be observed that KD-only and ML-only models struggle with the segmentation of small regions of interest. It can also be noticed that for conventional models, only one of the two students manages to predict these small regions. In our proposed approach, an ensemble of student predictions is used for the final predictions, facilitating the student networks to consistently predict the region of interest and yield optimal performance. Moreover, student networks trained using a diverse knowledge paradigm demonstrate a superior ability to discern finer structures compared to other models.

We demonstrate the importance of combining KD with ML by comparing all models trained with a knowledge diversification paradigm in Figure 5. It can be noticed that the combined KD + ML model successfully detects these small regions of interest from these hard sample images where KD-only and ML-only models fail. Similarly, to underscore the significance of knowledge diversification over other learning strategies, we show sample predictions from the combined KD + ML model trained with V1 (predictions only), V2 (features only), and a diverse knowledge paradigm, V3 (both predictions and features) (Figure 9 in the supplementary material) where we observe that the V3 strategy helps detect small and fine regions of interest better than V1 and V2.

We repeated the above experiments using the HAM10000 dataset for the lesion segmentation task, and the corresponding results are presented in Table 15 of the supplementary material. Moreover, the LGG segmentation task was repeated using the ResNet50-MobileNet combination for the teacher-student networks in the U-Net. Consistent improvement in performance observed in these additional experiments further emphasizes the reliability of the KD + ML model with a knowledge diversification strategy.

5.3. Comparison with existing KD techniques

Table 10 in the supplementary material compares our proposed approach with existing knowledge distillation methods. For the classification task, we conducted a comprehensive comparison against a range of knowledge distillation approaches, including Fitnet [40], AT [57], SP [30], SRRL [16], VID [2], SimKD [55], and SemCKD [52]. A consistent setup was maintained to ensure fairness, employing a ResNet50 model as the teacher and ResNet18 as the student model for all existing methodologies.

In addition, for the segmentation task, we compared our approach with well-known segmentation methods like SKD [33], IFVD [54], CWD [44], and DSD [9], originally designed for computer vision semantic segmentation tasks. To provide a

Table 4: Performance comparison of U-Net with ResNet50-ResNet18 encoder for segmentation task using LGG dataset with IoU and F-score metrics using four different distillation techniques: ML - Mutual Learning, KD (on) - online Knowledge Distillation, KD (off) - offline Knowledge Distillation, and KD + ML – combined KD & ML; and three different learning strategies - V1 (predictions only), V2 (features only) and a diverse knowledge paradigm, V3 (both predictions and features).

ML	V1 $\alpha = 0.1, \alpha' = 0.1$ $\beta = 0, \beta' = 0$ $\gamma = 0.9, \gamma' = 0.9$		V2 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0, \beta' = 0$ $\gamma = 0.8, \gamma' = 0.8$		V3 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0, \beta' = 0$ $\gamma = 0.8, \gamma' = 0.8$	
	IoU	F-score	IoU	F-score	IoU	F-score
S1	76.69 ± 0.989	86.44 ± 0.622	74.08 ± 1.812	85.51 ± 0.121	77.93 ± 0.537	87.33 ± 0.509
S2	75.28 ± 1.661	85.18 ± 1.060	75.18 ± 1.541	85.90 ± 0.782	77.26 ± 0.756	87.13 ± 0.381
Ensemble	77.26 ± 0.438	87.49 ± 0.593	75.70 ± 1.503	86.02 ± 0.974	78.24 ± 0.494	87.63 ± 0.614
KD (off)	V1 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.8, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$		V2 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.8, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$		V3 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.8, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$	
	IoU	F-score	IoU	F-score	IoU	F-score
T	76.86 ± 0.791	86.93 ± 0.537	76.86 ± 0.791	86.93 ± 0.537	76.86 ± 0.791	86.93 ± 0.537
S1	75.81 ± 0.643	86.22 ± 0.381	75.45 ± 0.410	86.01 ± 0.261	76.81 ± 1.432	86.65 ± 1.381
S2	75.77 ± 0.410	86.11 ± 0.956	76.29 ± 1.576	86.50 ± 0.989	77.59 ± 0.931	86.99 ± 0.728
Ensemble	77.28 ± 0.356	87.56 ± 0.274	77.79 ± 1.283	87.29 ± 1.576	78.87 ± 0.452	87.93 ± 0.390
KD (on)	V1 $\alpha = 0.1, \alpha' = 0.1$ $\beta = 0.9, \beta' = 0.9$ $\gamma = 0, \gamma' = 0$		V2 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.8, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$		V3 $\alpha = 0.1, \alpha' = 0.2$ $\beta = 0.9, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$	
	IoU	F-score	IoU	F-score	IoU	F-score
T	75.09 ± 0.848	85.11 ± 0.558	76.64 ± 0.593	86.77 ± 0.381	76.27 ± 0.516	86.71 ± 0.190
S1	76.85 ± 0.945	86.91 ± 0.601	77.74 ± 0.254	87.47 ± 0.162	78.51 ± 0.473	88.09 ± 0.452
S2	77.52 ± 0.466	86.65 ± 0.360	77.51 ± 0.367	87.33 ± 0.212	78.04 ± 0.226	87.01 ± 0.967
Ensemble	78.76 ± 0.565	88.23 ± 0.763	78.41 ± 0.141	87.90 ± 0.130	79.23 ± 0.494	88.37 ± 0.322
KD + ML	V1 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.4, \beta' = 0.4$ $\gamma = 0.4, \gamma' = 0.4$		V2 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.4, \beta' = 0.4$ $\gamma = 0.4, \gamma' = 0.4$		V3 $\alpha = 0.1, \alpha' = 0.1$ $\beta = 0.45, \beta' = 0.45$ $\gamma = 0.45, \gamma' = 0.45$	
	IoU	F-score	IoU	F-score	IoU	F-score
T	76.20 ± 0.247	86.08 ± 0.749	78.03 ± 0.982	87.70 ± 0.685	78.06 ± 1.130	87.61 ± 1.330
S1	76.31 ± 0.883	86.56 ± 0.565	78.20 ± 0.792	87.76 ± 0.459	78.66 ± 0.542	88.46 ± 0.506
S2	77.00 ± 0.707	87.01 ± 0.473	78.94 ± 1.440	88.23 ± 0.905	79.39 ± 0.491	88.84 ± 0.367
Ensemble	78.83 ± 0.275	88.09 ± 0.070	79.63 ± 0.700	88.31 ± 0.638	80.12 ± 0.597	89.52 ± 0.556

more comprehensive analysis, we also implemented AT [57] and Fitnet [40], even though they were not initially designed for segmentation tasks. Additionally, we included a domain-specific segmentation approach, EMKD [39]. Throughout this evaluation, we maintained consistency by using a ResNet50 as the backbone of U-Net for the teacher model and ResNet18 for the student model.

This extensive comparative analysis offers valuable insights into the effectiveness and performance of our proposed method in relation to a diverse set of state-of-the-art knowledge distillation and segmentation techniques

5.4. Explainability

To understand the effect of using different learning styles for the classification task, we used the popular Centered Kernel Alignment (CKA) metric [27] to measure and compare the similarity of learned representations of different layers of the teacher

and student networks. As highlighted (blue box) in Figure 6, the CKA plots show increased similarity between the higher layers of the teacher and final layers of the student networks for the knowledge diversification paradigm (V3). As the deeper layers of a network are considered task-specific, this increased similarity could potentially explain the improved performance of KD + ML with V3 compared to all other knowledge distillation and sharing strategies (Figure 10 and 11 in the supplementary material). Finally, Figure 7 shows the ensemble segmentation masks of different knowledge-sharing models trained with the knowledge diversification paradigm (V3).

In addition, fidelity is another important criterion for evaluating knowledge distillation models, where we expect the student predictions to match the teacher rather than achieve higher accuracy. We conducted some preliminary experiments to see how fidelity between students and teachers is affected by KD + ML as well as different learning styles. We observed that fidelity is not

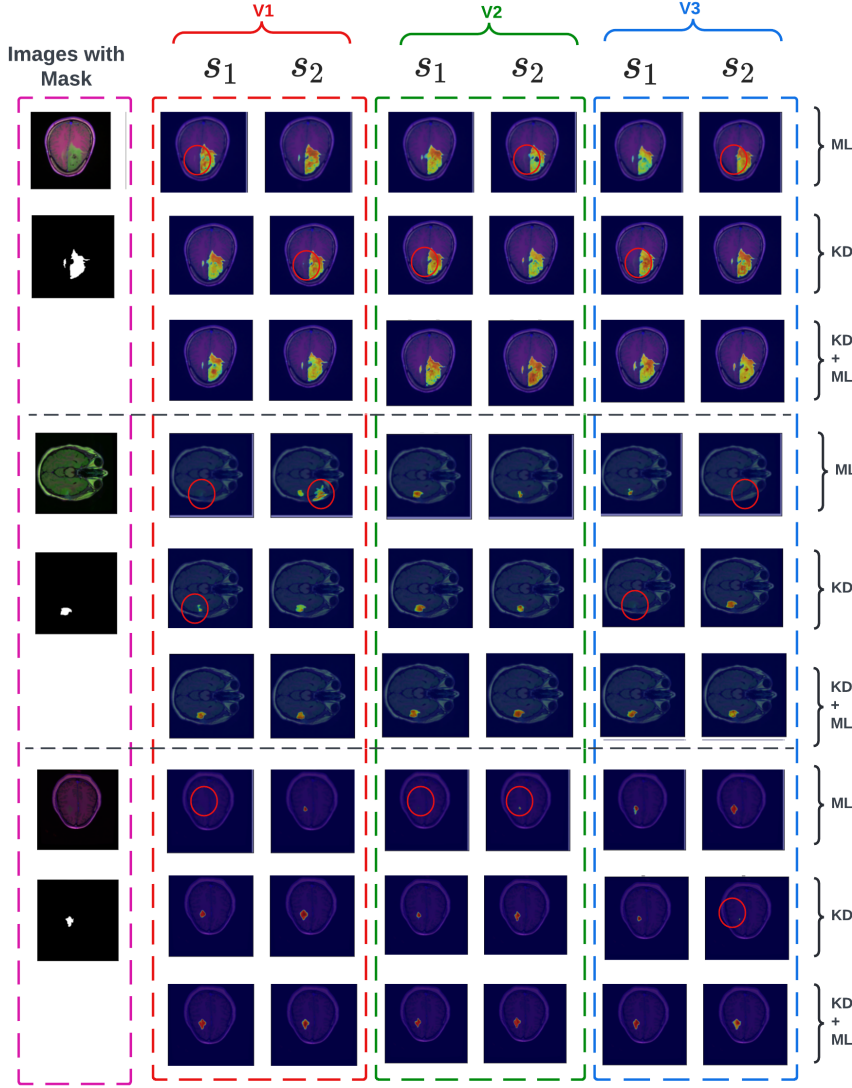


Figure 4: Heatmap visualization of individual student predictions (s_1 and s_2) for three hard-to-segment examples. Predictions are shown for KD only, ML only, and KD + ML models trained with only V1 (predictions only), V2 (features only), and a diverse knowledge paradigm, V3 (both predictions and features), along with the original input image and the corresponding ground-truth mask. Red circles highlight regions where models failed to detect tumor segments.

consistently improved for all combinations of experiments (Table 11 and 12 in the supplementary material). This is expected as the students are forced to not just match with the teacher but also with other students. Future efforts can be directed towards a combined increase in accuracy as well as fidelity of the student networks.

6. Conclusions

In this work, we propose an enhanced knowledge transfer protocol that embraces the concept of different learning styles through knowledge diversification. Within the framework of a single-teacher, multi-student network that emulates classroom dynamics, the teacher network imparts knowledge to the student networks in various forms, such as predictions to one student and feature maps to another. Furthermore, we enhance individual

student learning by fostering the exchange of diversified knowledge among students. Extensive experiments were conducted on metastatic tissue classification, brain tumor segmentation, and dermatological classification and segmentation tasks involving four different distillation techniques and three distinct learning strategies. These experiments showcase the superiority of our proposed approach over the existing knowledge distillation methods. In particular, our combined KD and ML model, trained with knowledge diversification, yields an average improvement of 2% in classification accuracy compared to conventional KD or ML techniques with similar network configurations, which rely solely on predictions. The improvement in performance provided by the proposed approach is significant in the context of model compression. In practice, our proposed model can be adapted to a range of medical imaging tasks that necessitate lightweight networks for deployment in real-world scenarios without compromising performance.

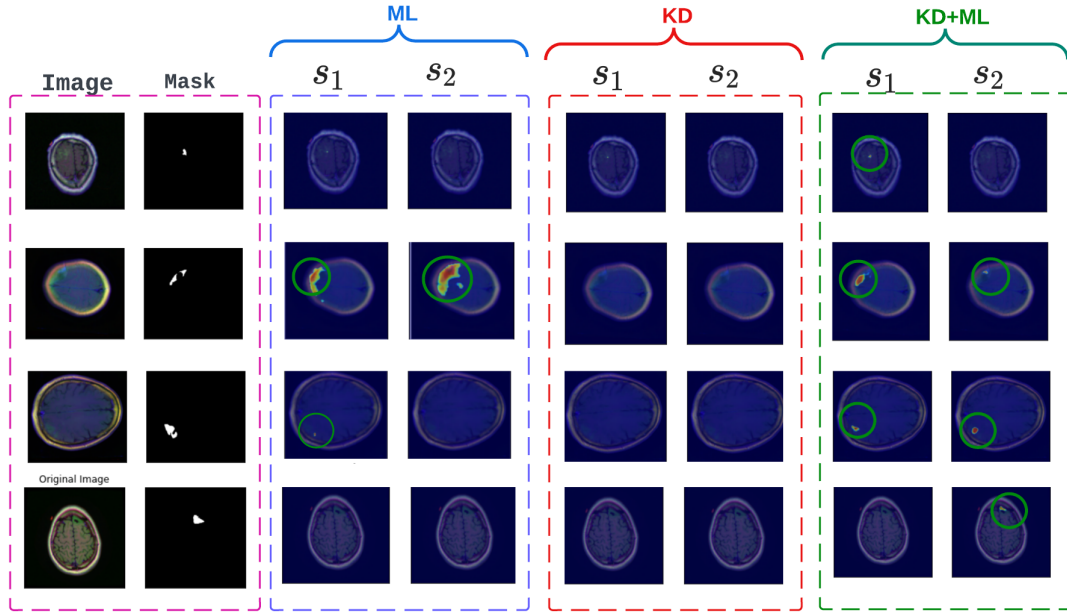


Figure 5: Heatmap visualization of individual student predictions (s_1 and s_2) for four hard-to-segment examples. Predictions are shown for KD only, ML only, and KD + ML models trained using diverse knowledge paradigm V3 (both predictions and features) along with the original input image and corresponding ground-truth mask. Green circles highlight the tumor segments detected.

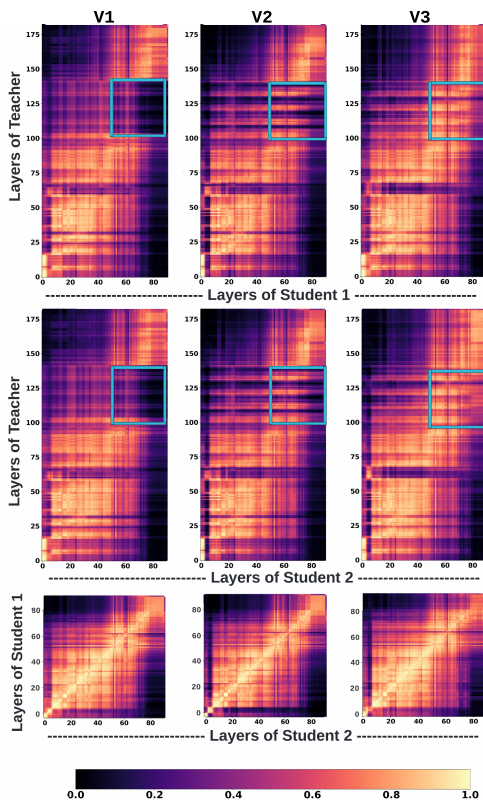


Figure 6: CKA plots visualizing the similarity of learned representation between different layers of teacher and student networks using KD + ML (V3) model.

References

- [1] Anil Adepu, Subin Sahayam, Umarani Jayaraman, and Rashmika Arram-raj. Melanoma classification from dermatoscopy images using knowledge distillation for highly imbalanced data. *Computers in biology and medicine*, 154, 01 2023. doi: 10.1016/j.combiomed.2023.106571. 2
- [2] Sungsoo Ahn and Taesup Lee. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 7, 17
- [3] Ali Alqahtani, Xianghua Xie, and Mark Jones. Literature review of deep network compression. *Informatics*, 8:77, 11 2021. doi: 10.3390/informatics8040077. 1
- [4] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *NIPS*, 2014. 1
- [5] Davis W. Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John V. Guttag. What is the state of neural network pruning? *ArXiv*, abs/2003.03033, 2020. 1
- [6] Cristian Bucilunundefined, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 535–541, 2006. ISBN 1595933395. doi: 10.1145/1150402.1150464. URL <https://doi.org/10.1145/1150402.1150464>. 1
- [7] Wei-Chun Chen, Chia-Che Chang, and Che-Rung Lee. Knowledge distillation with feature maps for image classification. In C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision – ACCV 2018*, pages 200–215, Cham, 2019. Springer International Publishing. ISBN 978-3-030-20893-6. 2
- [8] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. Model compression and acceleration for deep neural networks: The principles, progress, and challenges. *IEEE Signal Processing Magazine*, 35(1):126–136, 2018. doi: 10.1109/MSP.2017.2765695. 1
- [9] Yingchao Feng, Xian Sun, Wenhui Diao, Jihao Li, and Xin Gao. Double similarity distillation for semantic image segmentation. *IEEE Transactions on Image Processing*, 30:5363–5376, 2021. 7, 17
- [10] Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. Online knowledge distillation via collaborative learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11017–11026, 2020. 3
- [11] HAM10000 Lesion Segmentations. <https://www.kaggle.com/datasets/tschandl/ham10000-lesion-segmentations/>. 22
- [12] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights

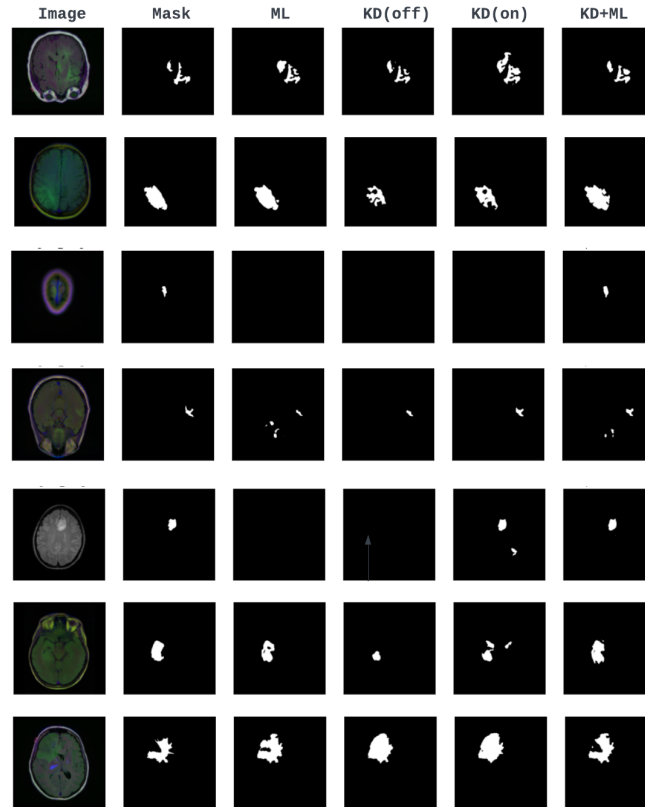


Figure 7: Comparison of the quality of segmentations obtained using different knowledge-sharing models trained with a diverse knowledge paradigm, V3 (both predictions and features). The proposed approach generates segmentation masks that are much finer than other models and closer to the ground truth.

- and connections for efficient neural networks. In *Advances in Neural Information Processing Systems*, pages 1135–1143, 2015. 1
- [13] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *International Conference on Learning Representations (ICLR)*, 2016. 1
- [14] Taimur Hassan, Muhammad Shafay, Bilal Hassan, Muhammad Usman Akram, Ayman ElBaz, and Naoufel Werghi. Knowledge distillation driven instance segmentation for grading prostate cancer. *Computers in Biology and Medicine*, 150:106124, 2022. ISSN 0010-4825. doi: <https://doi.org/10.1016/j.combiomed.2022.106124>. 2
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 13
- [16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Knowledge distillation via softmax regression representation learning. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015. 2, 7, 17
- [17] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *NIPS*, abs/1503.02531, 2015. 1, 2
- [18] Histopathologic Cancer Detection: Modified version of the Patch-Camelyon (PCam) Benchmark Dataset. <https://www.kaggle.com/competitions/histopathologic-cancer-detection/>. 5
- [19] Thi Kieu Khanh Ho and Jeonghwan Gwak. Utilizing knowledge distillation in deep learning for classification of chest x-ray abnormalities. *IEEE Access*, 8:160749–160761, 2020. doi: 10.1109/ACCESS.2020.3020802. 2
- [20] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1382–1391, 2017. 1, 14
- [21] Minhao Hu, Matthias Maillard, Ya Zhang, Tommaso Ciceri, Giammarco Barbera, Isabelle Bloch, and Pietro Gori. Knowledge distillation from multi-modal to mono-modal segmentation networks. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, 06 2020. 2
- [22] Zina M. Ibrahim, Daniel Bean, Thomas Searle, Linglong Qian, Honghan Wu, Anthony Shek, Zeljko Kraljevic, James Galloway, Sam Norton, James T. H. Teo, and Richard JB Dobson. A knowledge distillation ensemble framework for predicting short- and long-term hospitalization outcomes from electronic health records data. *IEEE Journal of Biomedical and Health Informatics*, 26(1):423–435, 2022. doi: 10.1109/JBHI.2021.3089287. 2
- [23] Sajid Javed, Arif Mahmood, Talha Qaiser, and Naoufel Werghi. Knowledge distillation in histology landscape by multi-layer features supervision. *IEEE Journal of Biomedical and Health Informatics*, pages 1–11, 2023. doi: 10.1109/JBHI.2023.3237749. 2
- [24] Irena Jekova and Vessela Krasteva. Optimization of end-to-end convolutional neural networks for analysis of out-of-hospital cardiac arrest rhythms during cardiopulmonary resuscitation. *Sensors*, 21(12), 2021. ISSN 1424-8220. doi: 10.3390/s21124105. URL <https://www.mdpi.com/1424-8220/21/12/4105>. 1
- [25] Lie Ju, Xin Wang, Xin Zhao, Huimin Lu, Dwarikanath Mahapatra, Paul Bonnington, and Zongyuan Ge. Synergic adversarial label learning for grading retinal diseases via knowledge distillation and multi-task learning. *IEEE journal of biomedical and health informatics*, PP, 01 2021. doi: 10.1109/JBHI.2021.3052916. 2
- [26] Md Shakib Khan, Kazi Nabiul Alam, Abdur Dhruva, Hasib Zunair, and Nabeel Mohammed. Knowledge distillation approach towards melanoma detection. *Computers in Biology and Medicine*, 146:105581, 05 2022. doi: 10.1016/j.combiomed.2022.105581. 2
- [27] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019. 8
- [28] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. 3
- [29] LGG MRI Segmentation. <https://www.kaggle.com/datasets/mateuszbeda/lgg-mri-segmentation/>. 5
- [30] Yangming Li, Naiyan Wang, and Jianping Liu. Similarity-preserving

- knowledge distillation. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2017. 2, 7, 17
- [31] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5
- [32] Mucan Liu, Chonghui Guo, and Sijia Guo. An explainable knowledge distillation method with xgboost for icu mortality prediction. *Computers in Biology and Medicine*, 152:106466, 2023. ISSN 0010-4825. doi: <https://doi.org/10.1016/j.compbiomed.2022.106466>. 2
- [33] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2599–2608, 2019. doi: 10.1109/CVPR.2019.00271. 7, 17
- [34] Jia Mi, LiFang Wang, Yang Liu, and Jiong Zhang. Kde-gan: A multimodal medical image-fusion model based on knowledge distillation and explainable ai modules. *Computers in Biology and Medicine*, 151:106273, 2022. ISSN 0010-4825. doi: <https://doi.org/10.1016/j.compbiomed.2022.106273>. 2
- [35] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 565–571. IEEE, 2016. 5
- [36] Classification models. https://github.com/qubvel/classification_models. 14
- [37] Jakub Nalepa, Michal Marcinkiewicz, and Michal Kawulok. Data augmentation for brain-tumor segmentation: A review. *Frontiers in Computational Neuroscience*, 13, 2019. ISSN 1662-5188. doi: 10.3389/fncom.2019.00083. URL <https://www.frontiersin.org/articles/10.3389/fncom.2019.00083>. 6
- [38] Usma Niyaz and Deepti R. Bathula. Augmenting knowledge distillation with peer-to-peer mutual learning for model compression. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4, 2022. doi: 10.1109/ISBI52829.2022.9761511. 4, 6
- [39] Dian Qin, Jiajun Bu, Zhe Liu, Xin Shen, Sheng Zhou, Jing-Jun Gu, Zhihong Wang, Lei Wu, and Hui-Fen Dai. Efficient medical image segmentation based on knowledge distillation. *IEEE Transactions on Medical Imaging*, 40:3820–3831, 2021. 2, 8, 17
- [40] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chas-sang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *CoRR*, abs/1412.6550, 2015. 2, 7, 8, 17
- [41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5, 14
- [42] Fahad Sarfraz, Elahe Arani, and Bahram Zonooz. Knowledge distillation beyond model compression. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6136–6143, 2021. doi: 10.1109/ICPR48806.2021.9413016. 1
- [43] Majid Sepahvand and Fardin Abdali-Mohammadi. Joint learning method with teacher–student knowledge distillation for on-device breast cancer image classification. *Computers in Biology and Medicine*, 12 2022. doi: 10.1016/j.compbiomed.2022.106476. 2
- [44] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5291–5300, 2021. doi: 10.1109/ICCV48922.2021.00526. 7, 17
- [45] Skin Cancer MNIST: HAM10000. <https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000/>, . 22
- [46] Skin Cancer MNIST: HAM10000. <https://challenge.isic-archive.com/landing/2018/>, . 22
- [47] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 1
- [48] TCIA. <https://www.cancerimagingarchive.net/>. 5
- [49] Philipp Tschandl, Noel Codella, Allan Halpern, Susana Puig, Zoi Apalla, Christoph Rinner, Peter Soyer, Cliff Rosendahl, Josep Malvehy, Iris Zalaudek, Giuseppe Argenziano, Caterina Longo, and Harald Kittler. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26, 08 2020. doi: 10.1038/s41591-020-0942-0. 22
- [50] Tiny ML. <https://www.tinyml.org/about/>. 2
- [51] Tom van Sonsbeek, Xiantong Zhen, Dwarikanath Mahapatra, and Marcel Worring. Probabilistic integration of object level annotations in chest x-ray classification. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3619–3629, 2023. doi: 10.1109/WACV56688.2023.00362. 2
- [52] Chen-Hao Wang, Yu Liu, Wei Wu, and Ming-Hsuan Yang. Cross-layer distillation with semantic calibration. In *Proceedings of the European Conference on Computer Vision*, 2020. 2, 7, 17
- [53] Yongwei Wang, Yuheng Wang, Jiayue Cai, Tim Lee, Chunyan Miao, and Z. Wang. Ssd-kd: A self-supervised diverse knowledge distillation method for lightweight skin lesion classification using dermoscopic images. *Medical Image Analysis*, 84:102693, 11 2022. doi: 10.1016/j.media.2022.102693. 2
- [54] Yukang Wang, Wei Zhou, Tao Jiang, Xiang Bai, and Yongchao Xu. Intra-class feature variation distillation for semantic segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 346–362, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58571-6. 7, 17
- [55] Kaichao You, Yangming Li, Jiawei Xu, Yunhong Wang, and Yimin Luo. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2, 7, 17
- [56] Ruichi Yu, Ang Li, Chun-Fu Chen, Jui-Hsin Lai, Vlad I. Morariu, Xintong Han, Mingfei Gao, Ching-Yung Lin, and Larry S. Davis. Nisp: Pruning networks using neuron importance score propagation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9194–9203, 2017. 1
- [57] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *Proceedings of the 5th International Conference on Learning Representations*, 2017. 2, 7, 8, 17
- [58] Shuwei Zhai, Guotai Wang, Xiangde Luo, Qian Yue, Kang Li, and Shaoting Zhang. Pa-seg: Learning from point annotations for 3d medical image segmentation using contextual regularization and cross knowledge distillation. *ArXiv*, abs/2208.05669, 2022. 2
- [59] Kangkai Zhang, Chunhui Zhang, Shikun Li, Dan Zeng, and Shiming Ge. Student network learning via evolutionary knowledge distillation. *IEEE Transactions on Circuits and Systems for Video Technology*, abs/2103.13811, 2021. 2
- [60] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3712–3721, 2019. 2
- [61] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018. 1, 3
- [62] Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Stephen Boyd, and Peter W Glynn. Stochastic mirror descent in variationally coherent optimization problems. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/e6ba70fc093b4ce912d769ede1ceeba8-Paper.pdf. 1

A. Supplementary Material

A.1. Teacher-Student information sharing

Table 5: A systematic representation of the different loss terms combinations that generate various combinations of distillation techniques and learning strategies.

Method	Version	L_{KD}		L'_{KD}		L_{ML}		L'_{ML}	
		L_{KD_p}	L_{KD_f}	L'_{KD_p}	L'_{KD_f}	L_{ML_p}	L_{ML_f}	L'_{ML_p}	L'_{ML_f}
KD	V1	✓	×	✓	×	×	×	×	×
	V2	×	✓	×	✓	×	×	×	×
	V3	✓	×	×	✓	×	×	×	×
ML	V1	×	×	×	×	✓	×	✓	×
	V2	×	×	×	×	×	✓	×	✓
	V3	×	×	×	×	✓	×	×	✓
KD + ML	V1	✓	×	✓	×	✓	×	✓	×
	V2	×	✓	×	✓	×	✓	×	✓
	V3	×	✓	✓	×	✓	×	×	✓

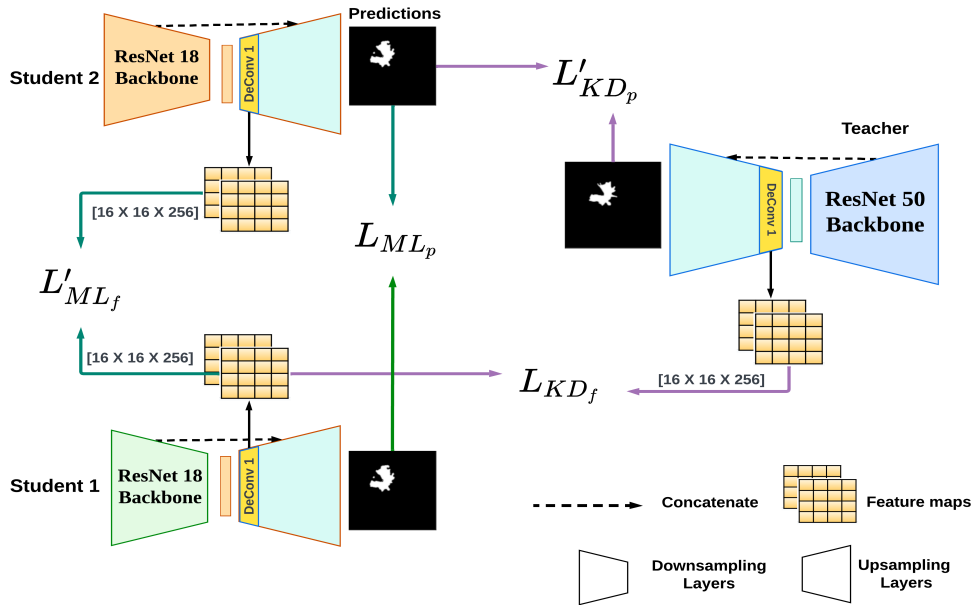


Figure 8: Overview of our proposed model that combines Knowledge Distillation (KD) with Mutual Learning (ML) in a single-teacher, two-student framework, leveraging a diversified knowledge-sharing strategy in the segmentation task. The loss terms L and L' represent the respective losses for each student, while L'_{KD_p} and L_{KD_f} capture the knowledge distillation losses between the teacher and concerned student over predictions and features, respectively. Similarly, L_{ML_p} and L'_{ML_f} denote the mutual learning losses between the two students over predictions and features, respectively. Here ResNet50 is used as a teacher network and ResNet18 as a student network

A.2. Teacher-Student network details

ResNet 50 as a teacher model: ResNet-50, a deep convolutional neural network (CNN) architecture, was introduced in 2015 [15] and has been highly influential in the field of image classification since then. It incorporates a unique concept called residual learning, which employs skip connections to enable learning residual functions instead of directly fitting the desired mapping. The model is pretrained on the ImageNet dataset, benefiting from pre-initialized weights, rendering it a powerful architecture for image analysis tasks. The decision to use ResNet-50 as the teacher model is driven by its exceptional ability to capture complex features.

Resnet18 as a student model: ResNet-18 serves as a suitable student model in our approach. It is a lighter variant within the ResNet architecture, with fewer layers. Despite its reduced complexity, ResNet-18 demonstrates commendable performance in classification tasks. Similar to ResNet-50, we initialize the model weights using pre-trained weights from the ImageNet dataset [36], enabling effective knowledge transfer. The architecture details for both ResNet50 and ResNet18 are defined in Table 6.

MobileNet as a student model: MobileNet [20] is a family of lightweight deep neural network architectures tailored for mobile and embedded vision applications. It utilizes depthwise separable convolutions, employing separate convolutional filters for each input channel followed by a 1x1 pointwise convolution to reduce parameters and computation significantly. This approach achieves efficient inference on resource-constrained devices while maintaining reasonable accuracy.

U-Net with ResNet-50/ResNet-18 as Backbone: By leveraging pre-trained ResNet-50/ResNet-18 as the foundational architecture for U-Net[41], we harness their robust feature representation capabilities to enhance the accuracy of segmentation tasks significantly. The strategic use of skip connections within the U-Net framework plays a crucial role in preserving intricate details and spatial information during upsampling, leading to superior segmentation performance. U-Net augmented with ResNet-50/ResNet-18 backbones prove to be a versatile and effective choice for diverse image segmentation applications, including medical, semantic, and instance segmentation. Our methodology adopts U-Net with ResNet-50 as the teacher model, encompassing around 33 million parameters, and U-Net with ResNet-18 as the student network, comprising around 14 million parameters. The architecture details for U-Net are defined in Table 7.

U-Net with MobileNet as backbone: Integrating pre-trained MobileNet as the backbone of U-Net, the architecture leverages MobileNet’s computational performance while utilizing U-Net’s superior segmentation capabilities. This combination aims to achieve efficient image segmentation, making it well-suited for applications in resource-constrained environments or real-time processing scenarios. In this study, U-Net with MobileNet as the student network, comprising around 8 million parameters.

Table 6: Architecture of ResNet-50 and Resnet-18 network

layer name	Output Shape	18-Layer	50-Layer
Conv1	112×112	$7 \times 7, 64, \text{stride } 2$ $3 \times 3, \text{maxpool, stride } 2$	
Conv2.x	56×56	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
Conv3.x	28×28	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 3$
Conv4.x	14×14	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$
Conv5.x	7×7	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
Conv6	7×7	-	$\begin{bmatrix} 1 \times 1, 512 \end{bmatrix} \times 1$

Table 7: Architecture of U-Net with ResNet-50 and Resnet-18 as a backbone

Downsampling Layers				Upsampling Layers		
layer name	Output Shape	18-Layer	50-Layer	Layer name	Output Shape	Upsample Block
Conv1	128×128	$3 \times 3, 64, \text{stride } 2$		DeConv 1	16×16	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$
Conv2.x	64×64	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	DeConv 2	32×32	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$
Conv3.x	32×32	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 3$	DeConv 3	64×64	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$
Conv4.x	16×16	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	DeConv 4	128×128	$\begin{bmatrix} 3 \times 3, 32 \\ 3 \times 3, 32 \end{bmatrix} \times 2$
Conv5.x	8×8	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	DeConv 5	256×256	$\begin{bmatrix} 3 \times 3, 16 \\ 3 \times 3, 16 \end{bmatrix} \times 2$

A.3. Results:

Table 8: Performance comparison of classification accuracy for Histopathologic Cancer Detection dataset using four different distillation techniques: ML - Mutual Learning, KD (on) - online Knowledge Distillation, KD (off) - offline Knowledge Distillation, and KD + ML – combined KD & ML; and three different learning strategies - V1 (predictions only), V2 (features only) and a diverse knowledge paradigm, V3 (both predictions and features). Here, ResNet50 is used as the teacher network and MobileNet as the student network

	V1	V2	V3
ML	$(\alpha = 0.2, \alpha' = 0.2)$ $(\beta = 0, \beta' = 0)$ $(\gamma = 0.8, \gamma' = 0.8)$	$(\alpha = 0.1, \alpha' = 0.2)$ $(\beta = 0, \beta' = 0)$ $(\gamma = 0.9, \gamma' = 0.8)$	$(\alpha = 0.1, \alpha' = 0.2)$ $(\beta = 0, \beta' = 0)$ $(\gamma = 0.9, \gamma' = 0.8)$
S1	92.67 ± 0.268	91.95 ± 0.523	92.21 ± 0.410
S2	92.83 ± 0.142	91.69 ± 0.654	92.99 ± 0.325
Ensemble	93.43 ± 0.170	92.18 ± 0.675	93.53 ± 0.127
	V1	V2	V3
KD(off)	$(\alpha = 0.2, \alpha' = 0.2)$ $(\beta = 0.8, \beta' = 0.8)$ $(\gamma = 0, \gamma' = 0)$	$(\alpha = 0.2, \alpha' = 0.2)$ $(\beta = 0.8, \beta' = 0.8)$ $(\gamma = 0, \gamma' = 0)$	$(\alpha = 0.2, \alpha' = 0.2)$ $(\beta = 0.8, \beta' = 0.8)$ $(\gamma = 0, \gamma' = 0)$
T	94.35 ± 0.763	94.35 ± 0.763	94.35 ± 0.763
S1	93.39 ± 0.070	92.73 ± 0.664	93.61 ± 0.353
S2	93.73 ± 0.219	93.72 ± 0.339	93.33 ± 0.155
Ensemble	94.02 ± 0.091	93.97 ± 0.120	94.23 ± 0.155
	V1	V2	V3
KD(on)	$(\alpha = 0.1, \alpha' = 0.2)$ $(\beta = 0.9, \beta' = 0.8)$ $(\gamma = 0, \gamma' = 0)$	$(\alpha = 0.2, \alpha' = 0.2)$ $(\beta = 0.8, \beta' = 0.8)$ $(\gamma = 0, \gamma' = 0)$	$(\alpha = 0.2, \alpha' = 0.2)$ $(\beta = 0.8, \beta' = 0.8)$ $(\gamma = 0, \gamma' = 0)$
T	94.52 ± 0.128	94.45 ± 0.268	94.43 ± 0.212
S1	94.07 ± 0.091	94.02 ± 0.070	93.89 ± 0.466
S2	94.1 ± 0.226	93.35 ± 0.170	94.15 ± 0.079
Ensemble	94.12 ± 0.028	94.04 ± 0.012	94.47 ± 0.593
	V1	V2	V3
KD + ML	$(\alpha = 0.1, \alpha' = 0.2)$ $(\beta = 0.45, \beta' = 0.4)$ $(\gamma = 0.45, \gamma' = 0.4)$	$(\alpha = 0.2, \alpha' = 0.2)$ $(\beta = 0.4, \beta' = 0.4)$ $(\beta = 0.4, \beta' = 0.4)$	$(\alpha = 0.2, \alpha' = 0.4)$ $(\beta = 0.4, \beta' = 0.3)$ $(\gamma = 0.4, \gamma' = 0.3)$
T	94.68 ± 0.311	94.72 ± 0.213	95.05 ± 0.005
S1	94.33 ± 0.127	93.81 ± 0.270	95.03 ± 0.070
S2	94.16 ± 0.184	94.04 ± 0.512	94.29 ± 0.342
Ensemble	94.95 ± 0.077	94.74 ± 0.121	95.61 ± 0.174

Table 9: Performance comparison for segmentation task using LGG dataset with IoU and F-score metrics using four different distillation techniques: ML - Mutual Learning, KD (on) - online Knowledge Distillation, KD (off) - offline Knowledge Distillation, and KD + ML – combined KD & ML; and three different learning strategies - (V1 (predictions only), V2 (features only) and a diverse knowledge paradigm, V3 (both predictions and features). Here, ResNet50 is used as the teacher network and MobileNet as the student network

ML	V1 $\alpha = 0.1, \alpha' = 0.1$ $\beta = 0, \beta' = 0$ $\gamma = 0.9, \gamma' = 0.9$		V2 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0, \beta' = 0$ $\gamma = 0.8, \gamma' = 0.8$		V3 $\alpha = 0.1, \alpha' = 0.2$ $\beta = 0, \beta' = 0$ $\gamma = 0.9, \gamma' = 0.8$	
	IoU	F-score	IoU	F-score	IoU	F-score
S1	70.35 ± 0.655	81.90 ± 0.601	69.65 ± 0.315	81.90 ± 0.671	70.76 ± 0.445	82.87 ± 0.304
S2	69.65 ± 0.902	82.10 ± 0.322	68.55 ± 1.180	80.11 ± 0.234	69.81 ± 1.302	82.97 ± 1.005
Ensemble	71.48 ± 0.484	83.11 ± 1.048	70.14 ± 1.712	82.15 ± 0.987	73.05 ± 0.894	83.95 ± 0.903
KD (off)	V1 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.8, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$		V2 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.8, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$		V3 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.8, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$	
	IoU	F-score	IoU	F-score	IoU	F-score
T	76.86 ± 0.691	86.93 ± 0.537	76.86 ± 0.691	86.93 ± 0.537	76.86 ± 0.791	86.93 ± 0.537
S1	70.53 ± 1.025	81.83 ± 0.622	70.09 ± 0.247	81.43 ± 0.169	71.51 ± 0.647	82.74 ± 0.169
S2	70.29 ± 0.994	82.47 ± 1.036	70.18 ± 0.247	82.06 ± 0.882	72.09 ± 0.329	81.66 ± 0.438
Ensemble	71.51 ± 0.477	83.38 ± 0.311	71.09 ± 0.869	82.49 ± 1.586	73.19 ± 0.566	84.57 ± 0.718
KD (on)	V1 $\alpha = 0.1, \alpha' = 0.1$ $\beta = 0.9, \beta' = 0.9$ $\gamma = 0, \gamma' = 0$		V2 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.8, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$		V3 $\alpha = 0.1, \alpha' = 0.2$ $\beta = 0.9, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$	
	IoU	F-score	IoU	F-score	IoU	F-score
T	76.43 ± 0.742	86.64 ± 0.473	76.11 ± 0.325	86.07 ± 0.205	76.57 ± 0.606	86.07 ± 0.887
S1	70.41 ± 0.753	81.93 ± 0.223	70.06 ± 0.304	80.96 ± 0.856	71.49 ± 0.767	83.27 ± 0.163
S2	70.66 ± 0.435	82.25 ± 0.260	70.67 ± 0.441	81.09 ± 0.570	71.77 ± 0.876	83.56 ± 0.537
Ensemble	72.76 ± 0.491	84.23 ± 0.782	71.41 ± 0.366	81.31 ± 0.528	73.92 ± 0.622	85.03 ± 0.410
KD + ML	V1 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.4, \beta' = 0.4$ $\gamma = 0.4, \gamma' = 0.4$		V2 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.4, \beta' = 0.4$ $\gamma = 0.4, \gamma' = 0.4$		V3 $\alpha = 0.1, \alpha' = 0.1$ $\beta = 0.45, \beta' = 0.45$ $\gamma = 0.45, \gamma' = 0.45$	
	IoU	F-score	IoU	F-score	IoU	F-score
T	76.92 ± 0.223	86.71 ± 0.735	76.25 ± 0.911	87.89 ± 0.786	76.40 ± 0.374	87.81 ± 0.240
S1	71.77 ± 0.732	82.64 ± 1.106	70.81 ± 0.417	82.31 ± 0.642	72.13 ± 0.657	83.76 ± 0.526
S2	71.53 ± 1.601	81.61 ± 0.756	71.38 ± 0.791	81.92 ± 0.551	71.59 ± 1.031	82.94 ± 0.767
Ensemble	73.14 ± 0.047	84.48 ± 0.070	72.12 ± 0.110	83.53 ± 0.247	74.31 ± 0.507	85.56 ± 0.332

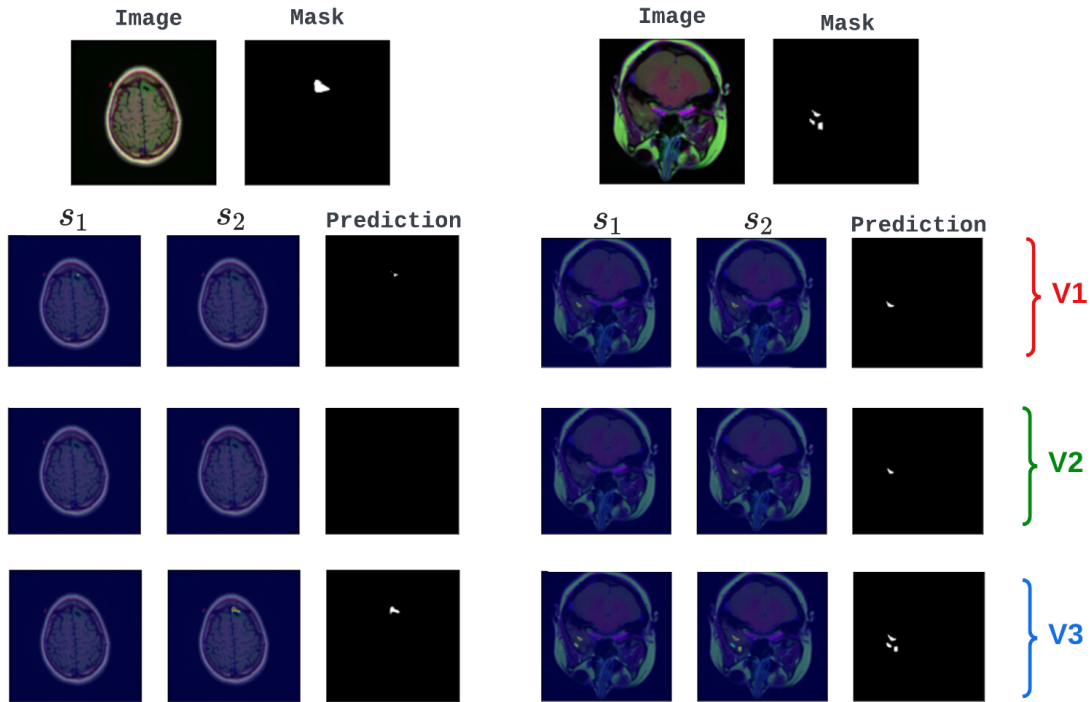


Figure 9: Heatmaps of individual student predictions for some hard-to-segment examples with KD + ML model trained using V1 (predictions only), V2 (features only) and a diverse knowledge paradigm, V3 (both predictions and features). The heatmaps highlight the significance of integrating both predictions and features in the combined model (V3) over predictions only (V1) and feature only (V2)

A.4. Comparison with existing KD techniques

Table 10: Comparison with other Knowledge Distillation methods in both Classification and Segmentation tasks. We used the ResNet50 and ResNet18 as the backbone architectures for teacher and student network respectively

Classification		Segmentation		
Method	Accuracy	Method	IoU	F-score
Fitnet[40]	94.78 ± 0.197	Fitnet[40]	75.24 ± 0.261	86.09 ± 0.141
SRRL[16]	91.70 ± 1.074	AT[57]	75.71 ± 0.197	86.18 ± 0.134
AT[57]	94.99 ± 0.014	SKD[33]	77.17 ± 0.254	87.51 ± 0.494
SP[30]	94.56 ± 0.311	IFVD[54]	78.48 ± 0.325	88.58 ± 0.183
VID[2]	95.34 ± 0.374	CWD[44]	75.87 ± 0.657	86.58 ± 0.862
SimKD[55]	84.01 ± 0.212	DSD[9]	77.47 ± 0.084	87.62 ± 0.494
SemCKD[52]	95.41 ± 0.204	EMKD[39]	73.62 ± 0.656	84.29 ± 1.576
Ours	96.68 ± 0.584	Ours	80.12 ± 0.597	89.52 ± 0.556

A.5. Explainability

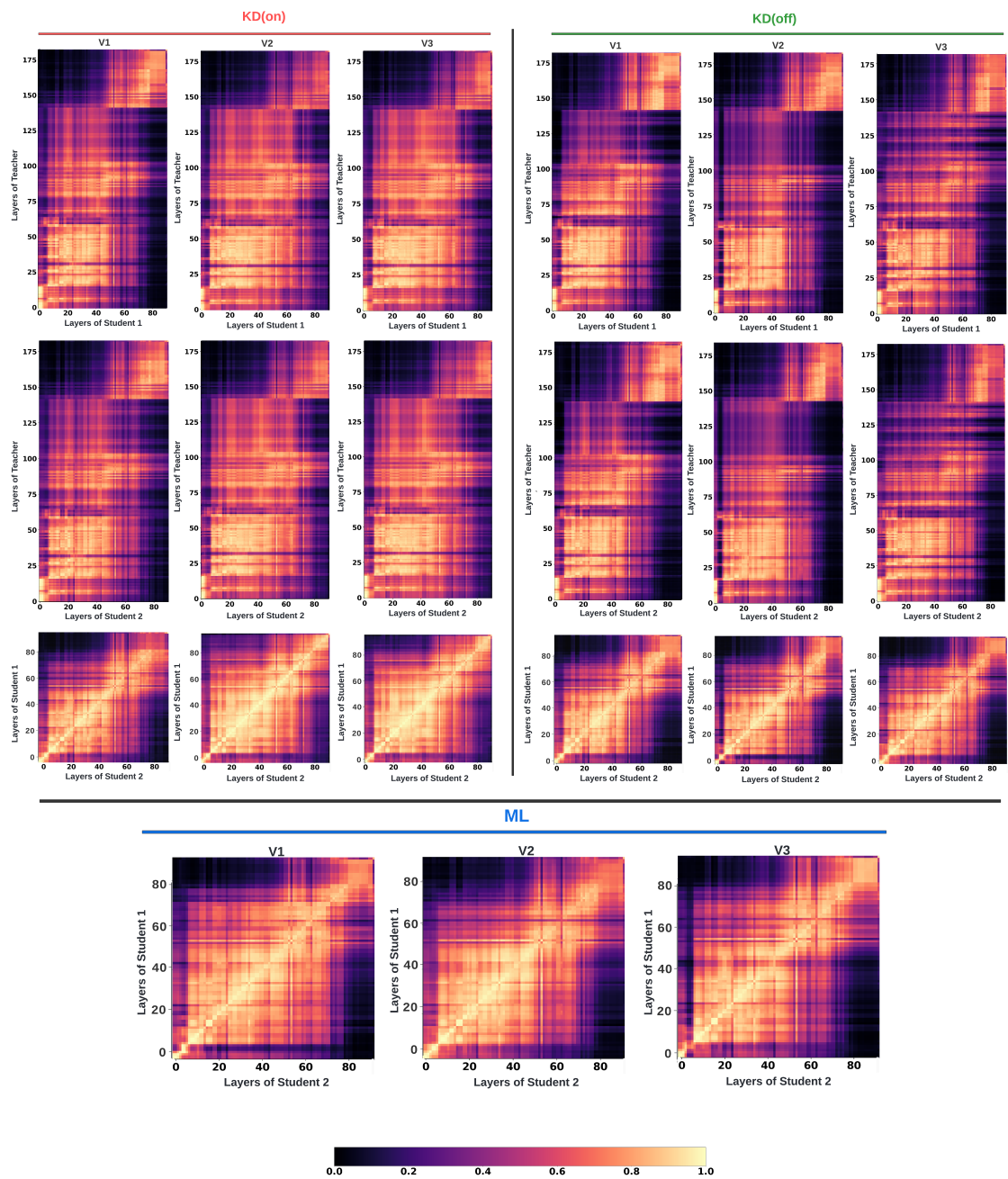


Figure 10: CKA plots visualizing the similarity of learned representation between different layers of teacher and student networks using KD (on), KD (off), and ML models.

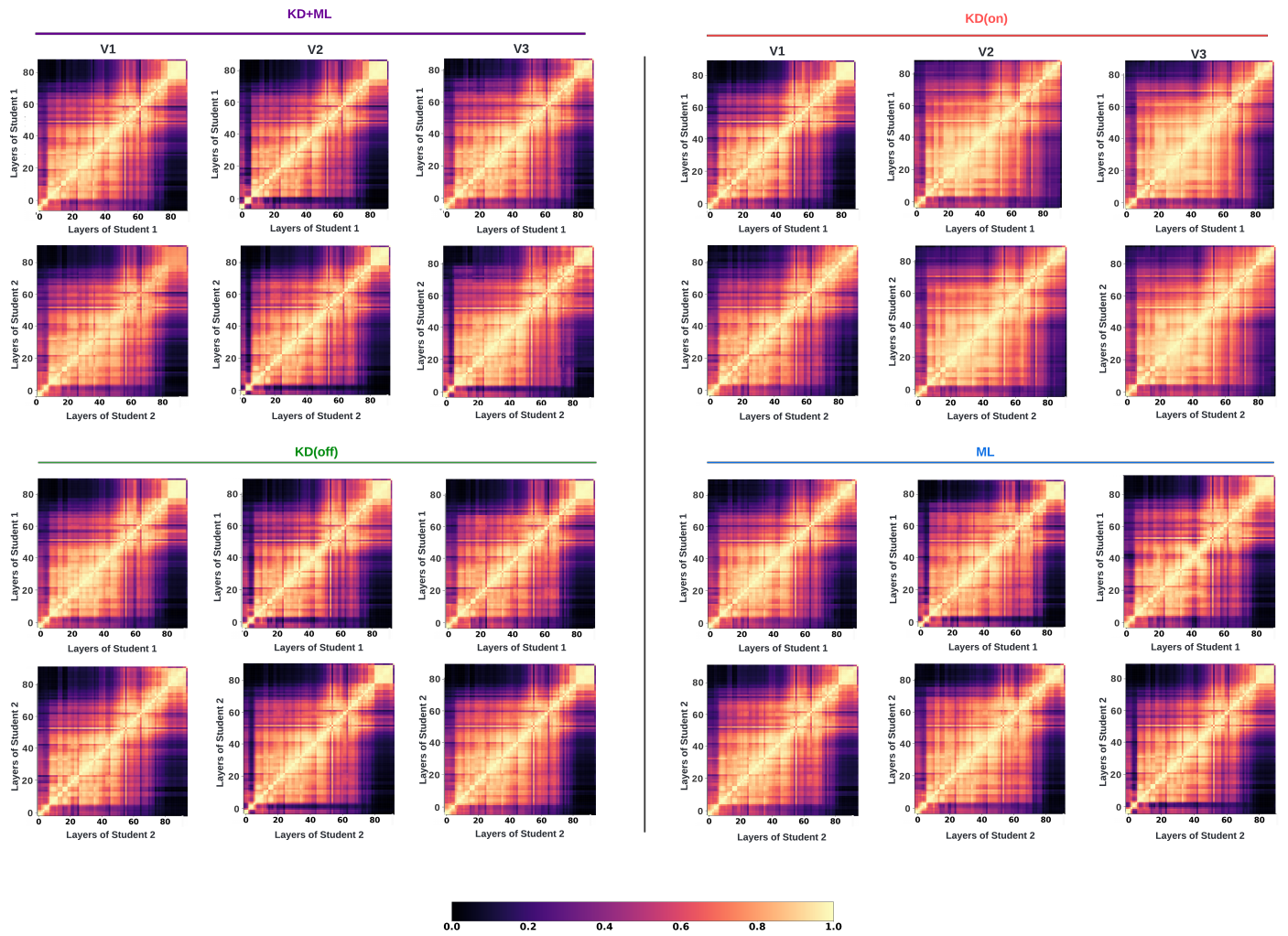


Figure 11: CKA plots visualizing the similarity of learned representation within different layers of student networks using KD+ML, KD (on), KD (off), and ML models.

Table 11: Fidelity comparison of ResNet50-ResNet18 for classification task of Histopathologic Cancer Detection dataset using four different distillation techniques: ML - Mutual Learning, KD (on) - online Knowledge Distillation, KD (off) - offline Knowledge Distillation, and KD + ML – combined KD & ML; and three different learning strategies - V1 (predictions only), V2 (features only) and a diverse knowledge paradigm, V3 (both predictions and features).

	V1 ($\alpha = 0.2, \alpha' = 0.2$) ($\beta = 0.8, \beta' = 0.8$) ($\gamma = 0, \gamma' = 0$)	V2 ($\alpha = 0.2, \alpha' = 0.2$) ($\beta = 0.8, \beta' = 0.8$) ($\gamma = 0, \gamma' = 0$)	V3 ($\alpha = 0.2, \alpha' = 0.2$) ($\beta = 0.8, \beta' = 0.8$) ($\gamma = 0, \gamma' = 0$)
S1	95.82 ± 0.106	95.19 ± 0.183	95.3 ± 0.141
S2	95.57 ± 0.106	95.67 ± 0.176	95.47 ± 0.247
Ensemble	95.82 ± 0.212	95.55 ± 0.637	95.95 ± 0.141
	V1 ($\alpha = 0.2, \alpha' = 0.2$) ($\beta = 0.8, \beta' = 0.8$) ($\gamma = 0, \gamma' = 0$)	V2 ($\alpha = 0.2, \alpha' = 0.2$) ($\beta = 0.8, \beta' = 0.8$) ($\gamma = 0, \gamma' = 0$)	V3 ($\alpha = 0.1, \alpha' = 0.2$) ($\beta = 0.9, \beta' = 0.8$) ($\gamma = 0, \gamma' = 0$)
S1	95.42 ± 0.367	95.44 ± 0.961	96.03 ± 0.466
S2	95.14 ± 0.121	95.89 ± 0.466	95.47 ± 0.339
Ensemble	95.92 ± 0.311	96.01 ± 0.480	96.52 ± 0.207
	V1 ($\alpha = 0.1, \alpha' = 0.2$) ($\beta = 0.45, \beta' = 0.4$) ($\gamma = 0.45, \gamma' = 0.4$)	V2 ($\alpha = 0.2, \alpha' = 0.2$) ($\beta = 0.4, \beta' = 0.4$) ($\beta = 0.4, \beta' = 0.4$)	V3 ($\alpha = 0.2, \alpha' = 0.4$) ($\beta = 0.4, \beta' = 0.3$) ($\gamma = 0.4, \gamma' = 0.3$)
S1	94.05 ± 0.848	95.71 ± 0.466	96.23 ± 0.636
S2	93.92 ± 1.308	95.98 ± 0.028	95.58 ± 0.608
Ensemble	94.60 ± 1.272	96.5 ± 0.231	96.41 ± 0.777

Table 12: Fidelity comparison of U-Net with ResNet50-ResNet18 encoder for segmentation task using LGG dataset with IoU and F-score metrics using four different distillation techniques: ML - Mutual Learning, KD (on) - online Knowledge Distillation, KD (off) - offline Knowledge Distillation, and KD + ML – combined KD & ML; and three different learning strategies - V1 (predictions only), V2 (features only) and a diverse knowledge paradigm, V3 (both predictions and features).

KD (off)	V1 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.8, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$		V2 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.8, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$		V3 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.8, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$	
	IoU	F-score	IoU	F-score	IoU	F-score
S1	81.97 ± 1.681	80.57 ± 1.697	83.94 ± 0.968	91.27 ± 0.565	82.17 ± 0.205	90.21 ± 0.110
S2	82.19 ± 0.779	89.99 ± 1.414	82.85 ± 1.315	90.51 ± 0.933	82.28 ± 0.813	90.28 ± 0.480
Ensemble	83.74 ± 1.800	90.34 ± 1.244	85.25 ± 0.958	92.05 ± 0.551	83.98 ± 0.277	90.67 ± 0.629
KD (on)	V1 $\alpha = 0.1, \alpha' = 0.1$ $\beta = 0.9, \beta' = 0.9$ $\gamma = 0, \gamma' = 0$		V2 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.8, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$		V3 $\alpha = 0.1, \alpha' = 0.2$ $\beta = 0.9, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$	
	IoU	F-score	IoU	F-score	IoU	F-score
S1	80.89 ± 0.431	89.44 ± 0.268	80.89 ± 0.014	89.44 ± 0.021	80.53 ± 1.043	89.20 ± 1.251
S2	79.43 ± 0.070	88.53 ± 0.049	81.62 ± 0.084	89.97 ± 0.049	80.85 ± 0.799	89.43 ± 0.480
Ensemble	79.79 ± 1.732	88.75 ± 1.080	81.04 ± 0.608	89.52 ± 0.367	79.80 ± 0.388	89.26 ± 0.473
KD + ML	V1 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.4, \beta' = 0.4$ $\gamma = 0.4, \gamma' = 0.4$		V2 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.4, \beta' = 0.4$ $\gamma = 0.4, \gamma' = 0.4$		V3 $\alpha = 0.1, \alpha' = 0.1$ $\beta = 0.45, \beta' = 0.45$ $\gamma = 0.45, \gamma' = 0.45$	
	IoU	F-score	IoU	F-score	IoU	F-score
S1	80.22 ± 2.660	89.01 ± 1.640	82.87 ± 1.280	90.62 ± 0.770	82.14 ± 0.134	90.19 ± 0.077
S2	80.60 ± 1.746	89.25 ± 1.074	81.82 ± 0.183	90.06 ± 0.113	82.62 ± 0.905	90.48 ± 0.537
Ensemble	79.97 ± 2.19	88.23 ± 1.994	82.57 ± 1.060	90.45 ± 0.636	81.60 ± 0.077	89.93 ± 0.134

A.6. Classification dataset: HAM10000

To rigorously evaluate the effectiveness of our proposed approach, we conducted a comprehensive series of experiments using the HAM10000 dataset [45]. This dataset comprises 10,015 images distributed across seven distinct classes and was originally sourced from [46]. Each image has a resolution of 450×600 downsampled to 224×224 . To ensure a fair and robust evaluation, we partitioned the dataset into training, validation, and test sets, closely mirroring the original distribution to maintain the dataset’s integrity. We addressed the class imbalance challenge by employing a balanced combination of under-sampling and over-sampling techniques exclusively on the training data.

A.6.1. Evaluation metrics

To ensure a fair evaluation of the multiclass HAM10000 classification, in addition to accuracy, we also used precision and recall, defined below:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (14)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (15)$$

Where TP, TN, FP, and FN represent the number of True Positive, True Negative, False Positive, and False Negative predictions, respectively.

A.7. Segmentation dataset: HAM10000

In the context of the segmentation task, we conducted an evaluation of our approach using the HAM10000 dataset, as referenced in [45]. The segmentation masks for this dataset were sourced from [49] and [11]. It’s worth noting that we maintained consistent preprocessing procedures with those employed in the classification task.

Table 13: Performance comparison of baseline models for HAM10000 classification and segmentation tasks.

Model	Classification			Segmentation	
	Precision	Recall	Accuracy	IOU	F-score
ResNet50	70.64 ± 0.993	72.48 ± 0.404	83.63 ± 0.065	87.77 ± 0.339	93.34 ± 0.261
ResNet18	65.57 ± 0.249	70.95 ± 0.335	82.08 ± 0.291	85.30 ± 0.228	89.88 ± 0.118

Notes

Table 14: Performance comparison for Classification task using HAM10000 with Precision, Recall and Accuracy metrics using four different network configurations: ML - Mutual Learning, KD (on) - online Knowledge Distillation, KD (off) - offline Knowledge Distillation, and KD + ML – combined KD & ML; and three different learning strategies - V1 (predictions only), V2 (features only) and a diverse knowledge paradigm, V3 (both predictions and features).

ML	V1 $\alpha = 0.1, \alpha' = 0.1$ $\beta = 0, \beta' = 0$ $\gamma = 0.9, \gamma' = 0.9$			V2 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0, \beta' = 0$ $\gamma = 0.8, \gamma' = 0.8$			V3 $\alpha = 0.2, \alpha' = 0.4$ $\beta = 0, \beta' = 0$ $\gamma = 0.8, \gamma' = 0.6$		
	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy
S1	66.03 ± 0.636	68.23 ± 0.238	81.78 ± 0.912	65.18 ± 0.937	72.63 ± 0.449	82.19 ± 0.854	67.46 ± 1.356	72.06 ± 1.334	83.21 ± 0.415
S2	65.14 ± 0.238	71.27 ± 1.097	82.78 ± 0.778	66.80 ± 0.945	70.24 ± 0.516	81.87 ± 0.864	65.41 ± 0.811	71.74 ± 0.276	82.30 ± 0.739
Ensemble	67.78 ± 1.119	72.46 ± 1.986	83.15 ± 1.025	67.44 ± 1.098	73.14 ± 0.988	82.33 ± 0.138	69.28 ± 0.190	74.19 ± 0.395	83.57 ± 0.095
KD (off)	V1 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.8, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$			V2 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.8, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$			V3 $\alpha = 0.1, \alpha' = 0.2$ $\beta = 0.9, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$		
	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy
T	70.64 ± 0.993	72.48 ± 0.404	83.63 ± 0.065	70.64 ± 0.993	72.48 ± 0.404	83.63 ± 0.065	70.64 ± 0.993	72.48 ± 0.404	83.63 ± 0.065
S1	67.10 ± 1.513	71.92 ± 0.950	81.81 ± 0.732	67.78 ± 0.844	70.79 ± 0.844	81.99 ± 0.522	69.96 ± 0.735	70.49 ± 0.587	82.88 ± 0.180
S2	66.62 ± 1.419	71.17 ± 0.274	81.30 ± 0.343	68.42 ± 0.744	69.42 ± 0.744	81.64 ± 0.663	67.04 ± 0.637	73.04 ± 0.556	82.62 ± 0.535
Ensemble	68.24 ± 0.960	72.07 ± 0.461	82.62 ± 1.003	69.08 ± 0.499	71.58 ± 0.499	82.25 ± 0.443	71.35 ± 0.289	73.35 ± 0.550	83.33 ± 0.190
KD (on)	V1 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.8, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$			V2 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.8, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$			V3 $\alpha = 0.1, \alpha' = 0.2$ $\beta = 0.9, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$		
	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy
T	69.45 ± 0.390	70.22 ± 0.577	83.24 ± 1.180	71.48 ± 0.997	70.90 ± 0.327	82.06 ± 0.725	70.26 ± 0.729	70.48 ± 0.514	83.42 ± 1.380
S1	69.16 ± 0.513	70.38 ± 0.907	81.72 ± 0.997	68.05 ± 0.793	72.81 ± 0.543	82.09 ± 0.997	69.10 ± 0.848	72.49 ± 0.787	82.24 ± 0.374
S2	65.62 ± 1.734	72.14 ± 0.416	82.35 ± 1.106	67.42 ± 0.424	72.30 ± 0.814	82.87 ± 0.519	69.87 ± 1.195	71.51 ± 0.652	83.51 ± 0.816
Ensemble	68.86 ± 1.600	73.61 ± 0.466	83.78 ± 1.154	69.24 ± 0.512	73.23 ± 0.586	83.16 ± 0.636	72.51 ± 1.166	74.86 ± 0.555	84.42 ± 0.364
KD + ML	V1 $\alpha = 0.1, \alpha' = 0.2$ $\beta = 0.45, \beta' = 0.4$ $\gamma = 0.45, \gamma' = 0.4$			V2 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.4, \beta' = 0.4$ $\gamma = 0.4, \gamma' = 0.4$			V3 $\alpha = 0.2, \alpha' = 0.4$ $\beta = 0.4, \beta' = 0.3$ $\gamma = 0.4, \gamma' = 0.3$		
	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy
T	70.33 ± 1.162	72.96 ± 1.186	83.69 ± 0.274	68.12 ± 0.814	72.16 ± 0.289	82.60 ± 1.154	69.49 ± 0.782	72.71 ± 0.569	84.78 ± 0.903
S1	68.33 ± 1.256	71.12 ± 0.886	83.33 ± 0.261	67.59 ± 0.509	74.57 ± 0.879	83.25 ± 0.476	72.15 ± 0.601	75.11 ± 0.306	83.69 ± 0.113
S2	68.65 ± 0.923	73.74 ± 0.793	83.29 ± 0.590	67.88 ± 0.262	71.87 ± 0.347	82.43 ± 1.085	73.71 ± 0.796	74.88 ± 1.516	84.03 ± 0.226
Ensemble	71.02 ± 0.820	75.75 ± 0.583	84.51 ± 0.344	69.35 ± 0.505	74.91 ± 0.584	83.87 ± 0.481	74.42 ± 0.606	76.97 ± 0.564	85.22 ± 0.742

Table 15: Performance comparison of U-Net with ResNet50-ResNet18 encoder for segmentation task using HAM10000 dataset with IoU and F-score metrics using four different distillation techniques: ML - Mutual Learning, KD (on) - online Knowledge Distillation, KD (off) - offline Knowledge Distillation, and KD + ML - combined KD & ML; and three different learning strategies - V1 (predictions only), V2 (features only) and a diverse knowledge paradigm, V3 (both predictions and features).

ML	V1 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0, \beta' = 0$ $\gamma = 0.8, \gamma' = 0.8$		V2 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0, \beta' = 0$ $\gamma = 0.8, \gamma' = 0.8$		V3 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0, \beta' = 0$ $\gamma = 0.8, \gamma' = 0.8$	
	IoU	F-score	IoU	F-score	IoU	F-score
S1	87.16 ± 0.700	93.10 ± 0.431	86.12 ± 0.322	92.57 ± 0.325	87.05 ± 0.289	93.64 ± 0.162
S2	87.06 ± 0.346	93.36 ± 0.190	86.20 ± 0.583	92.69 ± 0.516	87.12 ± 0.275	93.48 ± 0.155
Ensemble	87.36 ± 0.054	93.51 ± 0.125	86.78 ± 0.516	92.96 ± 0.516	88.05 ± 0.410	93.73 ± 0.233
KD (off)	V1 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.8, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$		V2 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.8, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$		V3 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.8, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$	
	IoU	F-score	IoU	F-score	IoU	F-score
T	87.77 ± 0.339	93.34 ± 0.261	87.77 ± 0.339	93.34 ± 0.261	87.77 ± 0.339	93.34 ± 0.261
S1	87.32 ± 0.254	93.24 ± 0.156	86.83 ± 0.247	93.01 ± 0.565	86.62 ± 1.064	92.82 ± 1.187
S2	87.11 ± 0.480	93.40 ± 0.275	86.73 ± 0.063	93.06 ± 0.035	87.93 ± 0.299	93.28 ± 0.452
Ensemble	87.44 ± 0.226	93.66 ± 0.127	87.06 ± 0.363	93.15 ± 0.253	87.65 ± 0.728	93.58 ± 0.028
KD (on)	V1 $\alpha = 0.1, \alpha' = 0.1$ $\beta = 0.9, \beta' = 0.9$ $\gamma = 0, \gamma' = 0$		V2 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.8, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$		V3 $\alpha = 0.1, \alpha' = 0.2$ $\beta = 0.9, \beta' = 0.8$ $\gamma = 0, \gamma' = 0$	
	IoU	F-score	IoU	F-score	IoU	F-score
T	87.55 ± 0.247	93.36 ± 0.141	86.76 ± 0.636	92.90 ± 0.366	87.32 ± 0.417	93.23 ± 0.240
S1	87.66 ± 0.558	93.42 ± 0.311	87.32 ± 0.494	93.23 ± 0.282	86.86 ± 1.499	92.96 ± 0.855
S2	87.65 ± 0.586	93.42 ± 0.332	87.05 ± 0.551	93.07 ± 0.311	87.65 ± 0.346	93.42 ± 0.197
Ensemble	87.93 ± 0.346	93.58 ± 0.197	87.40 ± 0.071	93.28 ± 0.042	87.73 ± 0.445	93.46 ± 0.254
KD + ML	V1 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.4, \beta' = 0.4$ $\gamma = 0.4, \gamma' = 0.4$		V2 $\alpha = 0.2, \alpha' = 0.2$ $\beta = 0.4, \beta' = 0.4$ $\gamma = 0.4, \gamma' = 0.4$		V3 $\alpha = 0.1, \alpha' = 0.1$ $\beta = 0.45, \beta' = 0.45$ $\gamma = 0.45, \gamma' = 0.45$	
	IoU	F-score	IoU	F-score	IoU	F-score
T	87.39 ± 0.975	93.19 ± 0.452	87.45 ± 0.671	93.30 ± 0.381	88.12 ± 0.979	93.15 ± 0.966
S1	88.08 ± 0.183	93.66 ± 0.098	87.64 ± 1.096	93.41 ± 0.622	88.86 ± 1.004	94.06 ± 0.509
S2	87.80 ± 0.530	93.50 ± 0.304	87.68 ± 0.784	93.43 ± 0.445	88.91 ± 0.954	94.09 ± 0.480
Ensemble	88.37 ± 0.311	93.82 ± 0.176	87.98 ± 0.776	93.60 ± 0.445	89.40 ± 0.799	94.27 ± 0.700