Supplementary Material
Adaptive Knowledge Distillation for Classification of Hand Images using Explainable Vision Transformers

Thanh Thi Nguyen[1] (✉), Campbell Wilson[1], and Janis Dalins[2]

[1] AiLECS Lab, Monash University, Melbourne VIC 3800, Australia
{thanh.nguyen9,campbell.wilson}@monash.edu
[2] AiLECS Lab, Australian Federal Police, Melbourne VIC 3800, Australia
janis.dalins@afp.gov.au

1 Additional details of ViT models

Details of the 6 ViT models investigated in this study are presented in Table 1 and 2. The first ViT model was initiated by Google Research (i.e., Google ViT), while DeiT, DINOv2 and ViT-MAE were proposed by Facebook/Meta AI Research. On the other hand, BEiT and Swin Transformer V2 were introduced by Microsoft Research. Repositories for pre-trained/fine-tuned weights of the models are available on the Hugging Face website at https://huggingface.co/.

Table 1: Details of 6 ViT models examined in this research. Each model has approximately from 85M to 87M parameters.

Methods	**Pre-trained/fine-tuned weights**	**Parameters**	**Resolution**
Google ViT	Google/vit-base-patch16-224	85,937,076	224 × 224
DeiT	Facebook/deit-base-patch16-224	85,937,076	224 × 224
BEiT	Microsoft/beit-base-patch16-224	85,900,404	224 × 224
DINOv2	Facebook/dinov2-base	86,857,140	256 × 256
Swin V2	Microsoft/swinv2-base-patch4-window8-256	87,078,316	256 × 256
ViT-MAE	Facebook/vit-mae-base	85,937,076	224 × 224

The BEiT model has the smallest number of parameters, with 85,900,404, while Swin Transformer V2 has the largest, with 87,078,316. The Google ViT, DeiT and ViT-MAE have exactly the same number of parameters, which is 85,937,076. The DINOv2 model has 86,857,140 parameters, placing it in the middle among the investigated models. All the parameters of each of these models are trainable, meaning the *trainable percentage* over the total number of parameters is 100%.

The DINOv2 and Swin Transformer V2 variants used in this research implement a resolution of 256×256 for input images, while the remaining models

Table 2: Resampling methods of the 6 models investigated in this research.

Methods	**Resampling**	**Mean (RGB)**	**SD (RGB)**
Google ViT	BILINEAR	[0.5, 0.5, 0.5]	[0.5, 0.5, 0.5]
DeiT	BILINEAR	[0.5, 0.5, 0.5]	[0.5, 0.5, 0.5]
BEiT	BILINEAR	[0.5, 0.5, 0.5]	[0.5, 0.5, 0.5]
DINOv2	BICUBIC	[0.485, 0.456, 0.406]	[0.229, 0.224, 0.225]
Swin V2	BICUBIC	[0.485, 0.456, 0.406]	[0.229, 0.224, 0.225]
ViT-MAE	BILINEAR	[0.485, 0.456, 0.406]	[0.229, 0.224, 0.225]

employ a resolution of 224×224. The resampling filter used when resizing the input images for DINOv2 and Swin Transformer V2 is the *bicubic resampling* method while the rest employ the *bilinear* method. For the Google ViT, DeiT and BEiT models, the input images are normalized across the RGB channels using a mean of [0.5, 0.5, 0.5] and a standard deviation of [0.5, 0.5, 0.5]. Other models, i.e., DINOv2, Swin Transformer V2 and ViT-MAE, use a mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225], correspondingly to the red, green and blue channels.

2 ViTs versus traditional methods

2.1 Graphical results on the IIT Delhi dataset

A box plot showing comparisons between ViT models and traditional methods in the 100-subject experiment (the column labeled “S=100” in Table 1 of the main paper) on the IIT Delhi dataset is presented in Fig. 1 below.

All ViT models are superior to the traditional methods, which are in the gray color. The results for the traditional methods are extracted from [1], which were obtained from single runs in previous studies, resulting in their boxes being represented as single lines. The performance of Google ViT, DeiT, BEiT and Swin Transformer V2 are equivalent in this experiment. DINOv2 and ViT-MAE are outperformed by other ViT models but still better than traditional methods based on the median accuracy.

In the 137-subject experiment (with results shown in the column labeled “S=137” in Table 1 of the main paper and the box plot in Fig. 2), the best existing result was obtained in [2] using a finger contour profile approach with an accuracy of 0.978. This result is better than two ViT models, which are DINOv2 at 0.976 and ViT-MAE at 0.948. Google ViT and DeiT obtain an excellent accuracy of 0.998. DINOv2 and ViT-MAE are the worst methods in this experiment in terms of both accuracy and stability. The standard deviation results of these two methods are respectively 0.018 and 0.079, the largest values compared with other ViT models.

The BEiT and Swin Transformer V2 methods obtain a nearly perfect accuracy of 0.999 in the 230-subject experiment (with results presented in the column

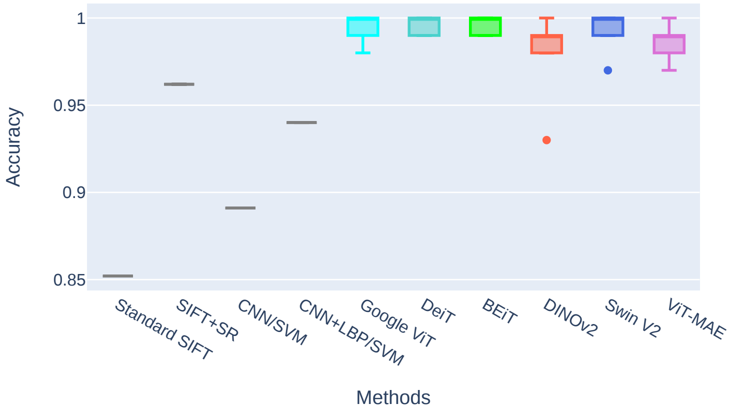


Fig. 1: ViTs and traditional methods using the IIT Delhi dataset in the **100-subject** experiment, where SR stands for the *sparse representation* approach that is combined with SIFT descriptors. Each box demonstrates a distribution of 10 accuracy results obtained from 10 repeated runs.

labeled "S=230" in Table 1 of the main paper and the box plot in Fig. 3). Both models are also the most stable techniques when they achieve a standard deviation of 0.002. The best existing result with an accuracy of 0.952 attained in [2] using the finger contour profile approach, which is only better than one ViT model, i.e., the DINOv2 method with an accuracy of 0.912. All other ViT methods outperform traditional methods in this experiment.

2.2 Graphical results on the 11k hands dataset

A comparison between ViT models and CNN-based methods, i.e., CNN/SVM and CNN+LBP/SVM approaches introduced in [1], in the dorsal-80 experiment is presented in Fig. 4. The ViT-MAE model's accuracy is approximate to those of the two CNN-based methods, while other ViT models significantly outperform the CNN-based methods.

Fig. 5 shows a comparison between models in the palm-80 experiment where the performance of DINOv2 is inferior to that of the CNN+LBP/SVM method while all other ViT models perform better than the CNN-based approaches. Fig. 6 presents results in the dorsal-100 experiment, which demonstrate the superior performance of all ViT models against the CNN-based methods. Likewise, the accuracy of the ViT models in the palm-100 experiment is greater than those of the CNN-based methods, as illustrated in Fig. 7.

Figs. 8 and 9 graphically demonstrate the performance of the competing models in the dorsal-120 and palm-120 experiments, respectively. The DINOv2 method is considerably unstable as it has the largest boxes although its median performance is equivalent to that of the traditional CNN+LBP/SVM method.

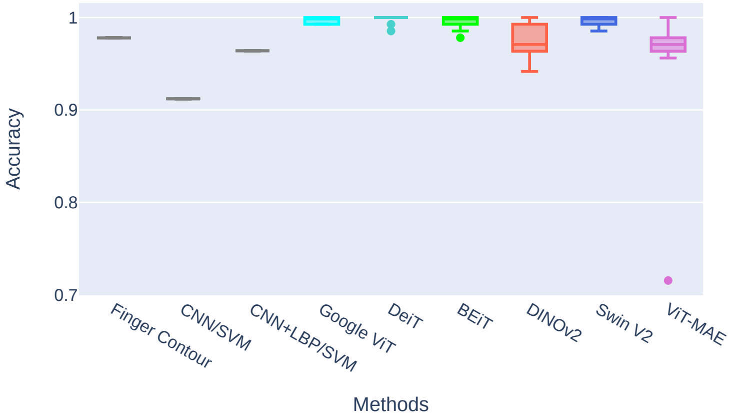


Fig. 2: A comparison between ViTs and traditional methods using the IIT Delhi dataset in the experiment with **137** subjects.

The ViT-MAE model outperforms the two CNN-based methods although it is inferior to Google ViT, DeiT, BEiT and Swin Transformer V2.

Throughout the experiments in this study, it can be seen that most ViT models outperform the CNN-based methods. DINOv2 and ViT-MAE are unstable and demonstrate the worst performance among the 6 investigated ViT models. The Google ViT, DeiT, BEiT and Swin Transformer V2 are comparable with each other and they are the best models for classifying hand images.

3 Additional ViT explainability results with Grad-CAM and DFF

3.1 An IIT Delhi dataset example

An illustration using a Google ViT model to predict subject of a hand image in the IIT Delhi dataset is shown in Fig. 10. The ground truth of the hand image image is subject "026", which is presented on the left side of Fig. 10a. While predicting subject of this image, the subject "019" is shown up as one of the top prediction candidates. We therefore plot a raw image of the subject "019" on the right side of Fig. 10a, aiming to understand the interpretations of the ViT model between the two apparently similar images.

Fig. 10b illustrates Grad-CAM interpretations of the ViT model towards the subject "026" on the left side and subject "019" on the right side of the figure. Basically, the Grad-CAM method helps to answer the question "Where does the ViT model see the subjects "026" and "019" in the input image?" It reveals that the ViT model identifies subject "026" based mainly on the region between the thumb finger and the index figure. On the other hand, the subject "019" is identified mainly based on the region near the hand wrist.

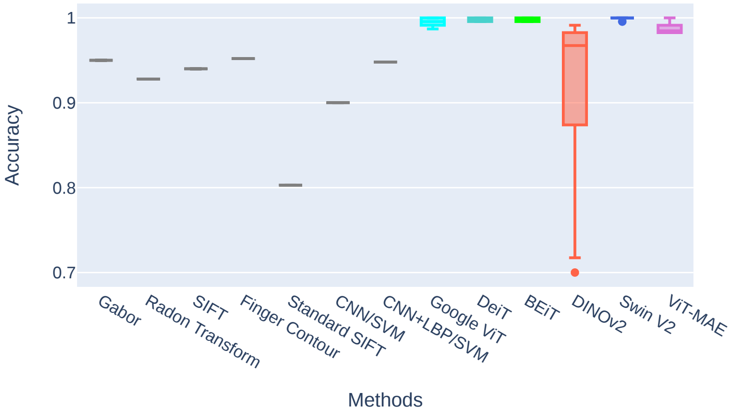


Fig. 3: ViTs and traditional methods using the IIT Delhi dataset in the experiment with **230** subjects. DINOv2 is extremely unstable in this experiment.

Fig. 10c shows an interpretation based on the Norm component of the ViT (the Norm contributes to the cosine loss in Fig. 1 of the main paper) using the deep feature factorization visualization method. The numbers shown adjacent to the class labels in the figure legend (i.e., “026”) are the softmax values of the concept outputs computed by this method. The top three concepts (candidate classes with highest probabilities) identified by this method are all of the subject “026”, which is the ground truth label of the input hand image. This demonstrates that the information in the Norm component of the Google ViT is useful as it leads to a correct prediction result.

3.2 An 11k hands dataset example

Explainability results on the palm-side left-hand images in the 11k hands dataset are presented in Fig. 11. More specifically, Fig. 11a shows raw images of palm-side left hands of subject “0000041” (left image) and subject “0000010” (right image). The model is asked to predict subject of the left image in Fig. 11a.

Information in the final linear component of the 11th transformer encoder layer of the ViT model is used for the Grad-CAM experiment. The left image of Fig. 11b indicates that the ring on the finger of subject “0000041” is the main feature that the model has learned from hand images of this subject. On the other hand, the right image of Fig. 11b reveals that the model focuses on characteristics of the fingers of the hand images to identify subject “0000010”. The background of the left and right images of Fig. 11b is the same because the model is taking the left image in Fig. 11a as the input. The Grad-CAM method simultaneously shows two interpretations of the ViT model towards the subjects “0000041” (left) and “0000010” (right) because we put these two labels as the two output category targets in the Grad-CAM experiment.

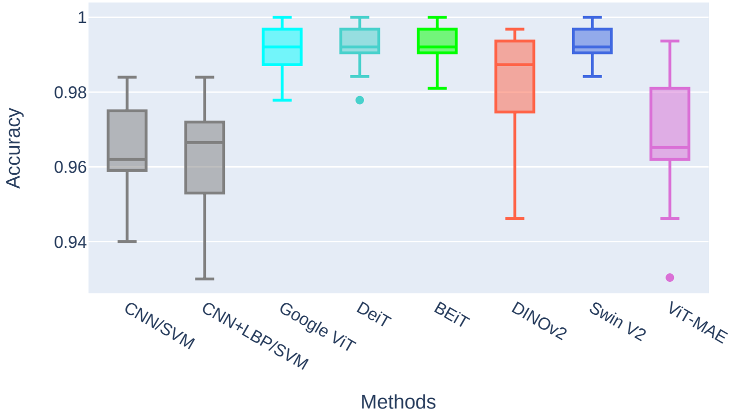


Fig. 4: A box plot illustrating the accuracy of competing methods on the 11k hands dataset in the experiment with **80** subjects using **dorsal** side images.

Fig. 11c displays an explanation for the ViT model based on the deep feature factorization visualization method using information in the Norm component of the ViT. The hand region in the image (in the green color) is clearly identified as of the subject “0000041”, which is commensurate with the ground truth label.

4 Additional results on adaptive distillation methods

Results obtained using the adaptive distillation methods on the IIT Delhi dataset are presented in Table 3. Without distillation, the student model obtains an accuracy of 0.995 on the source domain (i.e., left palm images) and 1.000 on the target domain (i.e., right palm images). This shows that the model does not forget much knowledge learned from the source domain data. Therefore, the three distillation methods (i.e., Hinton et al.’s method, our methods 1 and 2) do not show significant improvement against the no-distillation method (i.e., just simply refine-tuning the student model on the target domain data without consulting to the teacher).

Results for the left to right and right to left domain adaptation experiments on the 11k hands dataset are presented in Tables 4 and 5, respectively. There is also not much knowledge forgetting in these two experiments because the two domains (left and right hands) are quite similar to each other. In Table 4, refine-tuning the student model on the target domain without distillation obtains an accuracy of 0.961 on the source domain and 0.994 on the target domain. These values in Table 5 are 0.949 and 0.990, respectively. The distillation methods show improvement (although not significant) in both experiments, especially on the source domain data (reduce forgetting).

Results from the left-palm to left-dorsal domain adaptation using the 11k hands dataset are presented in Table 6. There is a considerable knowledge for-

Table 3: Results on adapting from *left* palm domain to *right* palm domain in the IIT Delhi dataset.

Accuracy of	Source Domain Left Palm	Target Domain Right Palm
Student before adapting	0.999	0.808
Ensemble teacher	1.000	0.898
Student after adapting - no distil	0.995	1.000
Student after adapting with knowledge distilled from ...		
the ensemble teacher (Hinton et al.)	0.997	0.999
the ensemble teacher (method 1)	0.998	0.999
prior copy as teacher (Hinton et al.)	0.999	1.000
prior copy as teacher (method 1)	0.999	1.000
prior copy as teacher (method 2)	0.999	1.000

Table 4: Results on adapting from *left* palm domain to *right* palm domain in the 11k hands image dataset.

Accuracy of	Source Domain Left Palm	Target Domain Right Palm
Student before adapting	0.985	0.485
Ensemble teacher	0.982	0.403
Student after adapting - no distil	0.961	0.994
Student after adapting with knowledge distilled from ...		
the ensemble teacher (Hinton et al.)	0.978	0.999
the ensemble teacher (method 1)	0.981	0.996
prior copy as teacher (Hinton et al.)	0.986	1.000
prior copy as teacher (method 1)	0.990	1.000
prior copy as teacher (method 2)	0.992	0.999

Table 5: Results on adapting from *right* dorsal domain to *left* dorsal domain in the 11k hands image dataset.

Accuracy of	Source Domain Right Dorsal	Target Domain Left Dorsal
Student before adapting	0.992	0.550
Ensemble teacher	0.996	0.668
Student after adapting - no distil	0.949	0.990
Student after adapting with knowledge distilled from ...		
the ensemble teacher (Hinton et al.)	0.969	0.956
the ensemble teacher (method 1)	0.974	0.956
prior copy as teacher (Hinton et al.)	0.983	0.988
prior copy as teacher (method 1)	0.994	0.989
prior copy as teacher (method 2)	0.997	0.986

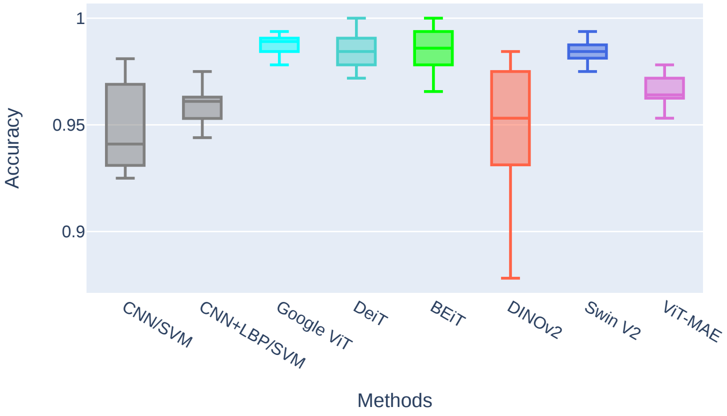


Fig. 5: A box plot illustrating the accuracy of competing methods on the 11k hands dataset in the experiment with **80** subjects using **palm** side images.

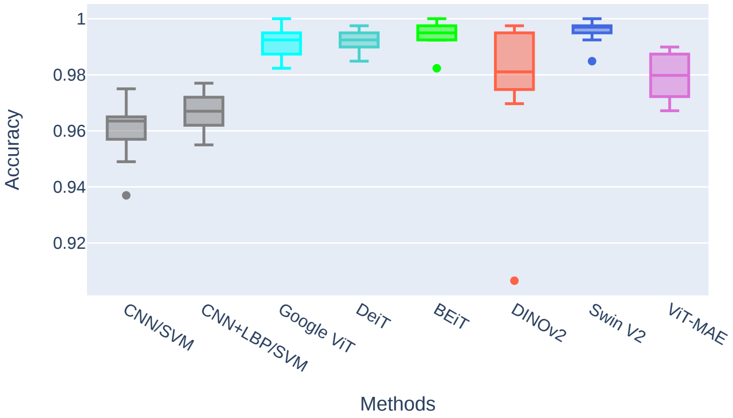


Fig. 6: A box plot illustrating the accuracy of competing methods on the 11k hands dataset in the experiment with **100** subjects using **dorsal** side images.

getting problem in this experiment because the two domains (palm and dorsal) are quite different from each other. After refine-tuning the student model on the target domain data without distillation, the accuracy obtained on the source domain is just 0.496. The distillation methods help reduce the knowledge forgetting problem significantly in this experiment. For example, the Hinton et al.'s method (with the ensemble teacher based on 6 ViT models, i.e., Google ViT, DeiT, BEiT, DINOv2, Swin Transformer V2 and ViT-MAE) improves the accuracy on the source domain to 0.674. Our method 1, with the same ensemble teacher, achieves a higher accuracy on the source domain, reaching 0.738. When using a prior copy of the student model as the teacher, our methods 1 and 2

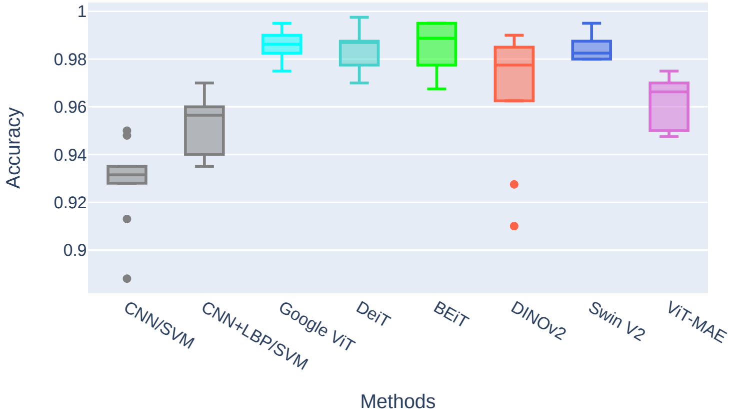


Fig. 7: A box plot illustrating the accuracy of competing methods on the 11k hands dataset in the experiment with **100** subjects using **palm** side images.

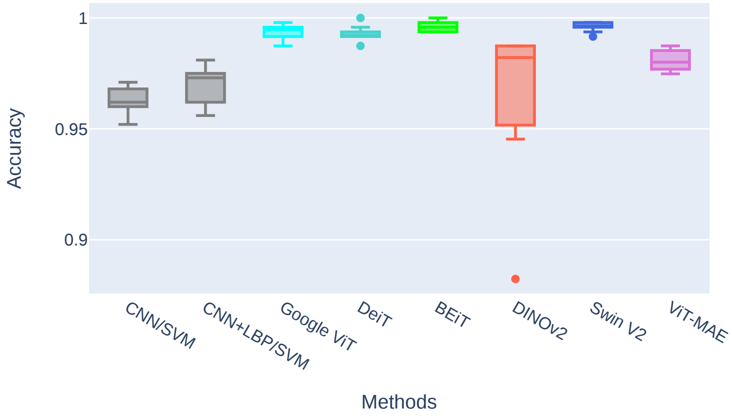


Fig. 8: A box plot illustrating the accuracy of competing methods on the 11k hands dataset in the experiment with **120** subjects using **dorsal** side images.

also outperform the Hinton et al.'s method. This demonstrates the effectiveness of our adaptive distillation methods when applied to solve domain adaptation problems in the hand image classification applications.

5 Loss functions during training

This section presents values of the loss and accuracy metrics during the refine-tuning process to adapt a student model to a new domain using the three competing methods: the baseline Hinton et al.'s method [5], our methods 1 and 2. The Hinton et al.'s method and our method 1 use two loss functions: hard loss

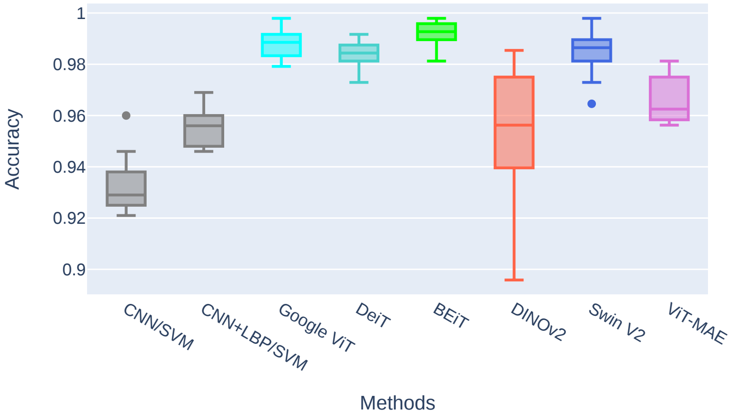


Fig. 9: A box plot illustrating the accuracy of competing methods on the 11k hands dataset in the experiment with **120** subjects using **palm** side images.

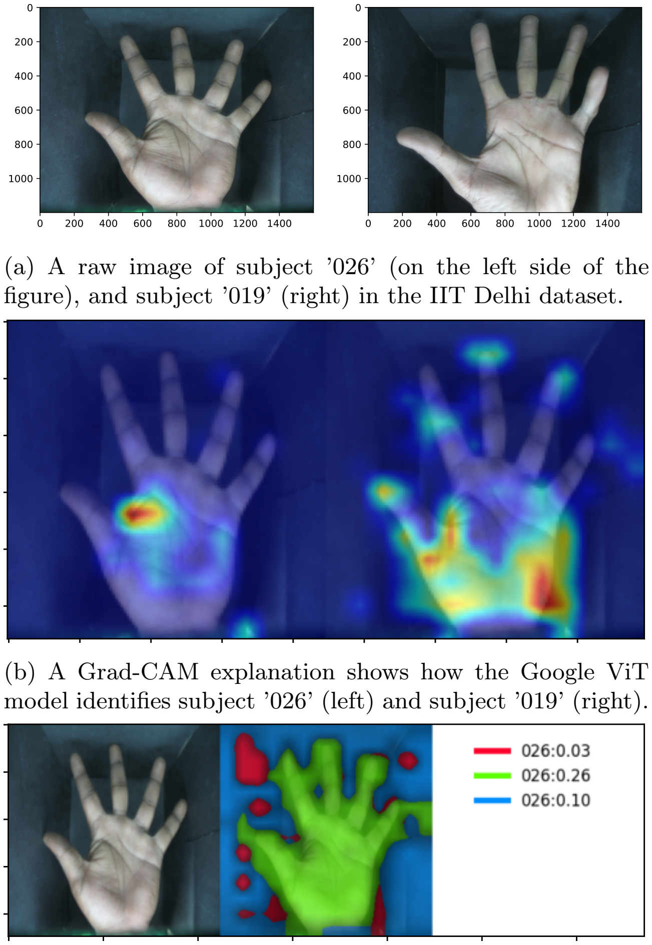
Table 6: Results on adapting from left *palm* domain to left *dorsal* domain in the 11k hands image dataset.

Accuracy of	Source Domain Left Palm	Target Domain Left Dorsal
Student before adapting	0.985	0.044
Ensemble teacher	0.982	0.083
Student after adapting - no distil	0.496	0.961
Student after adapting with knowledge distilled from ...		
the ensemble teacher (Hinton et al.)	0.674	0.938
the ensemble teacher (method 1)	0.738	0.946
prior copy as teacher (Hinton et al.)	0.769	0.958
prior copy as teacher (method 1)	0.829	0.965
prior copy as teacher (method 2)	0.858	0.965

and soft loss, while our method 2 uses three losses: hard loss, soft loss and cosine loss.

Figs. 12-16 show an illustration of training losses and accuracy of our adaptive distillation methods in comparison to the Hinton et al.'s method when refine-tuning the same Google ViT model to adapt from the left-palm domain to left-dorsal domain of the 11k hands dataset.

The three methods produce equivalent *hard loss* values (Fig. 12) across training epochs with the hard loss of the Hinton et al. method being slightly smaller than that of our methods in the first half the training process, i.e., the first 25 epochs out of the total 50 epochs. On the other hand, the *soft loss* (Fig. 13) is considerably different between the Hinton et al. method and our methods: the former renders much larger soft loss values than those of our methods. This is because, during the early stages of training, our methods predominantly empha-



(a) A raw image of subject ’026’ (on the left side of the figure), and subject ’019’ (right) in the IIT Delhi dataset.

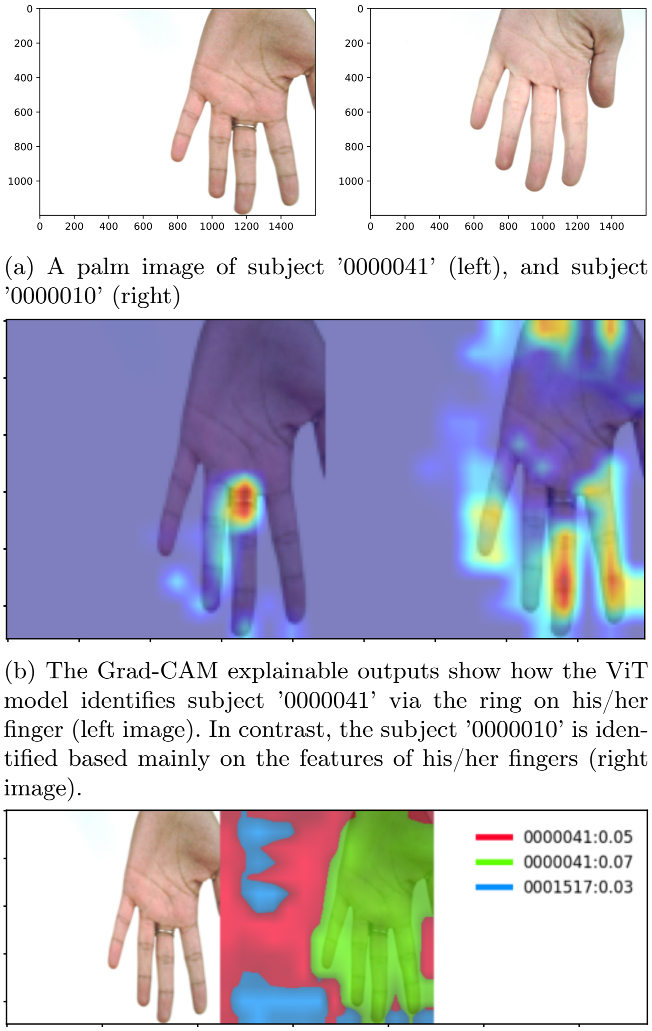
(b) A Grad-CAM explanation shows how the Google ViT model identifies subject ’026’ (left) and subject ’019’ (right).

(c) Explainability with the deep feature factorization visualization method.

Fig. 10: Explainability of a Google ViT model using the PyTorch library for CAM methods.

size the soft loss (i.e., sticking closely with the teacher) over the hard loss (i.e., learning from the ground truth labels of the new domain).

The hard loss and soft loss for method 1 and 2 are approximate, but the cosine loss values between these methods are significantly different (Fig. 14). This is because method 2 also includes the cosine loss during training. That results in the cosine loss of method 2 being the smallest among the three methods. The cosine loss of method 2 increases to a maximum value around the 25th epoch and then starts to decrease until the end of the learning process. As the Hinton et al. method and method 1 do not incorporate the cosine loss, the values of this loss



(a) A palm image of subject ’0000041’ (left), and subject ’0000010’ (right)

(b) The Grad-CAM explainable outputs show how the ViT model identifies subject ’0000041’ via the ring on his/her finger (left image). In contrast, the subject ’0000010’ is identified based mainly on the features of his/her fingers (right image).

(c) Results obtained using the deep feature factorization visualization [3].

Fig. 11: Explainability of a Google ViT model using the PyTorch library for CAM methods [4].

for these two methods increase considerably until the end of the training process. The magnitude of the cosine loss increases more in the Hinton et al. method than in method 1. This is because, even though method 1 does not include the cosine loss, it adheres more closely to the teacher compared to Hinton et al.’s method. This explains why method 1 performs better than the Hinton et al. method on the source domain data.

The overall loss of the methods 1 and 2 is started with a larger value but it decreases quickly compared to that of the Hinton et al. method (Fig. 15). The overall loss of the Hinton et al. method also decreases, but the decreasing magnitude is much less compared with our methods, and it therefore ends up equivalent to the overall loss of method 1 but still larger than that of method 2.

The training accuracy patterns on the target domain of the three methods are shown in Fig. 16. The three methods produce the equivalent training accuracy values. However, with the difference in soft loss and cosine loss values, the methods 1 and 2 produce better accuracy on the source domain data when compared to the Hinton et al.'s method.

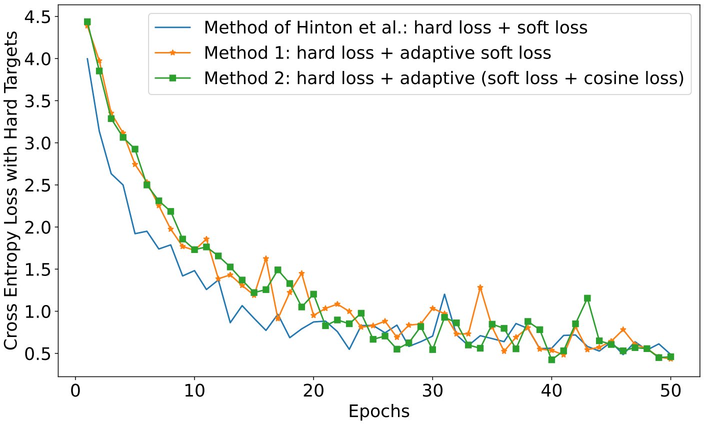


Fig. 12: Hard loss between student against ground truth in the left-palm to left-dorsal experiment of the 11k hands dataset.

Figs. 17-21 present values of loss functions and accuracy during training of the three methods in the adaptation process from right-dorsal to right-palm on the 11k hands dataset. The patterns of these values are similar to those in the left-palm to left-dorsal experiment (shown in Figs. 12-16). More specifically, our methods have smaller soft loss and cosine loss values during the training process compared to those of the Hinton et al.'s method. The overall loss of our methods starts at a larger value compared to that of the Hinton et al.'s method (Eq. 4 in the main paper) due to the inclusion of the adaptive component $\sqrt{\frac{E}{e}}$ as in Eqs. 5 and 7 in the main paper. This contributes to the difference in values of the soft loss and cosine loss between our methods and the Hinton et al.'s method, leading to the better performance of our methods on the source domain data compared to that of the Hinton et al.'s method.

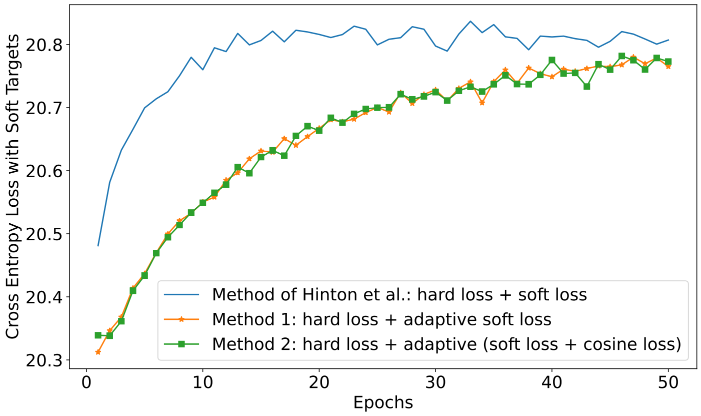


Fig. 13: Soft loss between outputs of student and teacher in the left-palm to left-dorsal experiment of the 11k hands dataset.

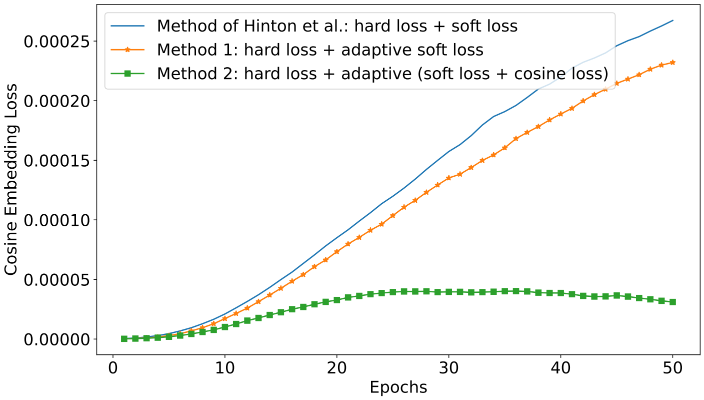


Fig. 14: Cosine loss between internals of student and teacher in the left-palm to left-dorsal experiment of the 11k hands dataset.

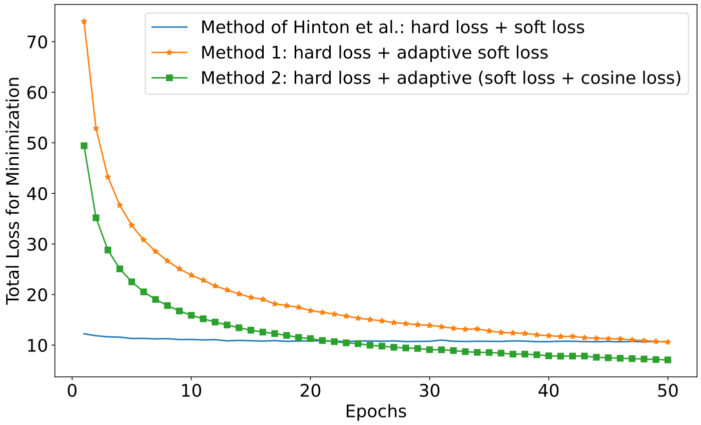


Fig. 15: Overall loss for fine-tuning student parameters in the left-palm to left-dorsal experiment of the 11k hands dataset.

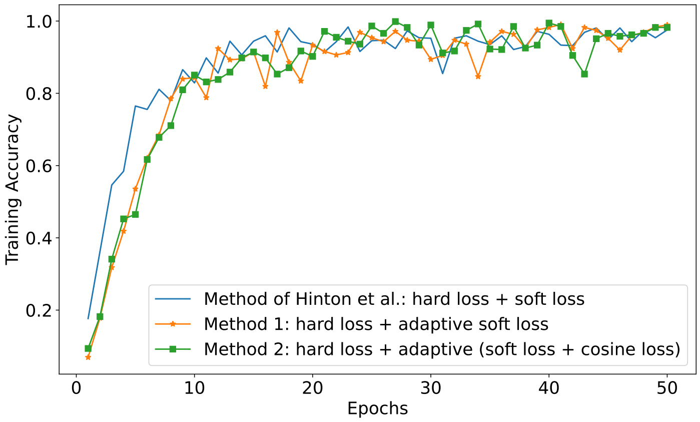


Fig. 16: Accuracy averaged across total number of batches in the left-palm to left-dorsal experiment of the 11k hands dataset.

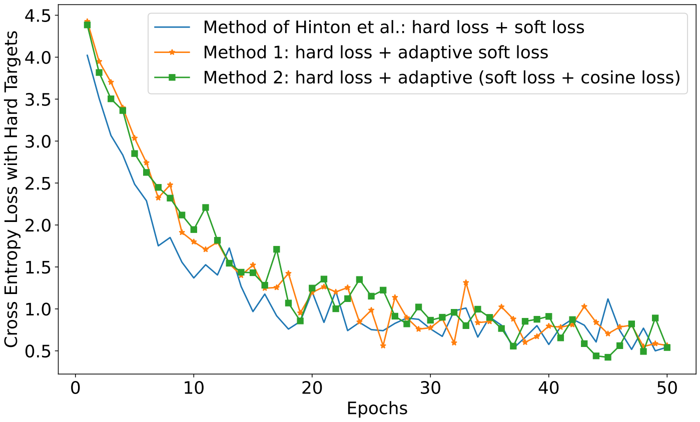


Fig. 17: Hard loss between student against ground truth in the right-dorsal to right-palm experiment of the 11k dataset.

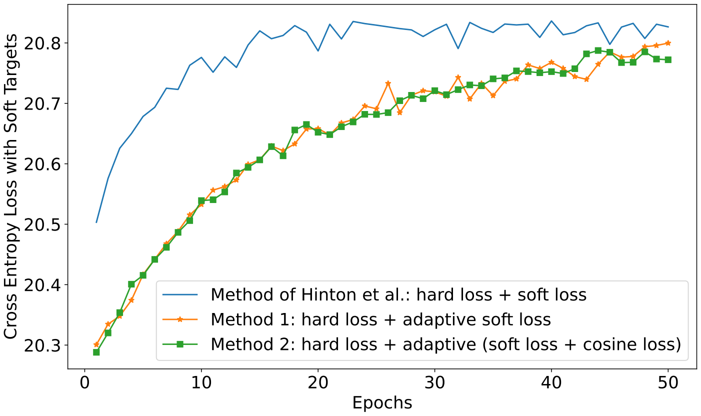


Fig. 18: Soft loss between outputs of student and teacher in the right-dorsal to right-palm experiment of the 11k dataset.

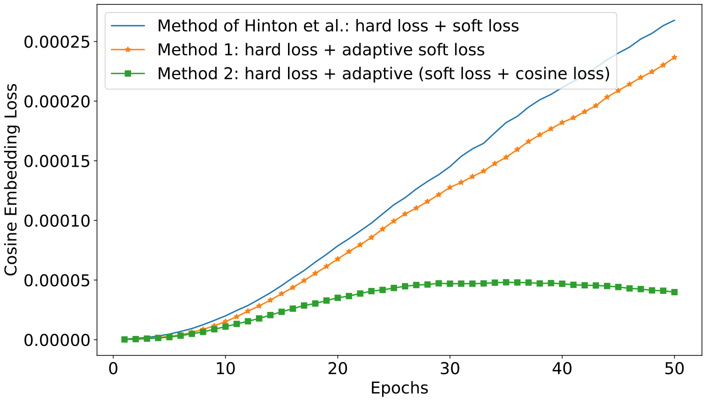


Fig. 19: Cosine loss between internals of student and teacher in the right-dorsal to right-palm experiment of the 11k dataset.

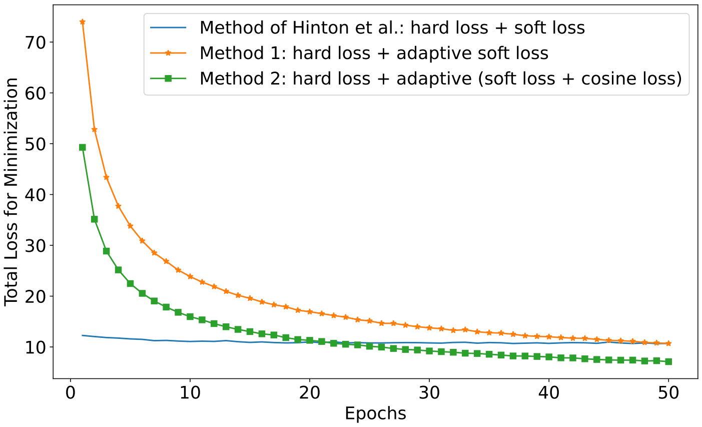


Fig. 20: Overall loss for fine-tuning student parameters in the right-dorsal to right-palm experiment of the 11k dataset.

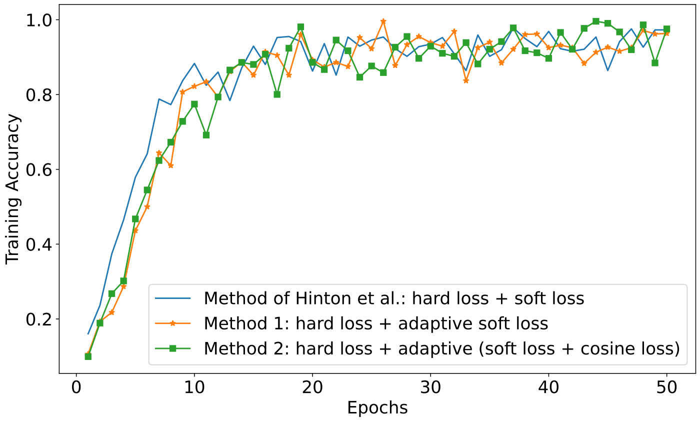


Fig. 21: Accuracy averaged across total number of batches in the right-dorsal to right-palm experiment of the 11k dataset.

References

1. Afifi, M.: 11k hands: Gender recognition and biometric identification using a large dataset of hand images. Multimedia Tools and Applications **78**, 20835–20854 (2019)
2. Bera, A., Bhattacharjee, D., Nasipuri, M.: Finger contour profile based hand biometric recognition. Multimedia Tools and Applications **76**, 21451–21479 (2017)
3. Collins, E., Achanta, R., Susstrunk, S.: Deep feature factorization for concept discovery. In: European Conference on Computer Vision (ECCV). pp. 336–352 (2018)
4. Gildenblat, J., contributors: Pytorch library for CAM methods. https://github.com/jacobgil/pytorch-grad-cam (2021)
5. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: Neural Information Processing System Deep Learning Workshop (2015), arXiv preprint arXiv:1503.02531