Knowledge Transfer from Vision Foundation Models for Efficient Training of Small Task-specific Models

Raviteja Vemulapalli [1] **Hadi Pouransari** [1] **Fartash Faghri** [1] **Sachin Mehta** [1] **Mehrdad Farajtabar** [1]
Mohammad Rastegari [1] **Oncel Tuzel** [1]

Abstract

Vision Foundation Models (VFMs) pretrained on massive datasets exhibit impressive performance on various downstream tasks, especially with limited labeled target data. However, due to their high inference compute cost, these models cannot be deployed for many real-world applications. Motivated by this, we ask the following important question, *"How can we leverage the knowledge from a large VFM to train a small task-specific model for a new target task with limited labeled training data?"*, and propose a simple task-oriented knowledge transfer approach as a highly effective solution to this problem. Our experimental results on five target tasks show that the proposed approach outperforms task-agnostic VFM distillation, web-scale CLIP pretraining, supervised ImageNet pretraining, and self-supervised DINO pretraining by up to 11.6%, 22.1%, 13.7%, and 29.8%, respectively. Furthermore, the proposed approach also demonstrates up to $9\times$, $4\times$ and $15\times$ reduction in pretraining compute cost when compared to task-agnostic VFM distillation, ImageNet pretraining and DINO pretraining, respectively, while outperforming them. We also show that the dataset used for transferring knowledge has a significant effect on the final target task performance, and introduce a retrieval-augmented knowledge transfer strategy that uses web-scale image retrieval to curate effective transfer sets.

1. Introduction

Currently, the computer vision community is witnessing the emergence of various vision and multi-modal foundation models pretrained on massive datasets (Oquab et al., 2023; Radford et al., 2021; Yuan et al., 2021; Li et al., 2023b; Wang et al., 2023b). These models have been shown to work well for many downstream tasks, especially, when task-specific labeled data is limited (Radford et al., 2021). However, they cannot be used for many resource-constrained applications due to their high inference computate cost. Also, many applications such as autonomous driving, medical image diagnostics, and industrial automation, focus on specific tasks or domains and need small task-specific models rather than a large Vision Foundation Model (VFM). This raises an important, timely and yet underexplored question: *How can we leverage the knowledge from a large VFM to effectively train a small task-specific model for a new target task with limited labeled training data?*

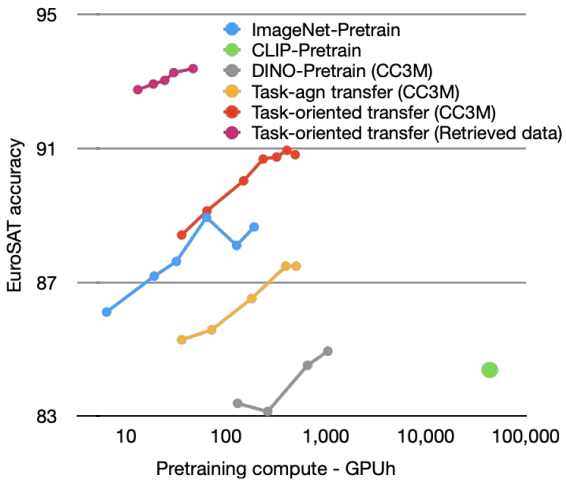


Figure 1. Downstream task performance (EuroSAT classification) of FastViT target model with different pretraining approaches. Here, we finetune the pretrained models using 10 labeled training images per class. Task-oriented knowledge transfer from DINOv2 VFM with generic web data (CC3M) outperforms the popular ImageNet, CLIP and DINO pretraining approaches. Knowledge transfer with a transfer set curated using image retrieval performs significantly better than knowledge transfer with generic web data.

Answering this question requires transferring knowledge from a VFM across both task and model architecture bound-

[1]Apple, USA. Correspondence to: Raviteja Vemulapalli <r_vemulapalli@apple.com>, Hadi Pouransari <mpouransari@apple.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

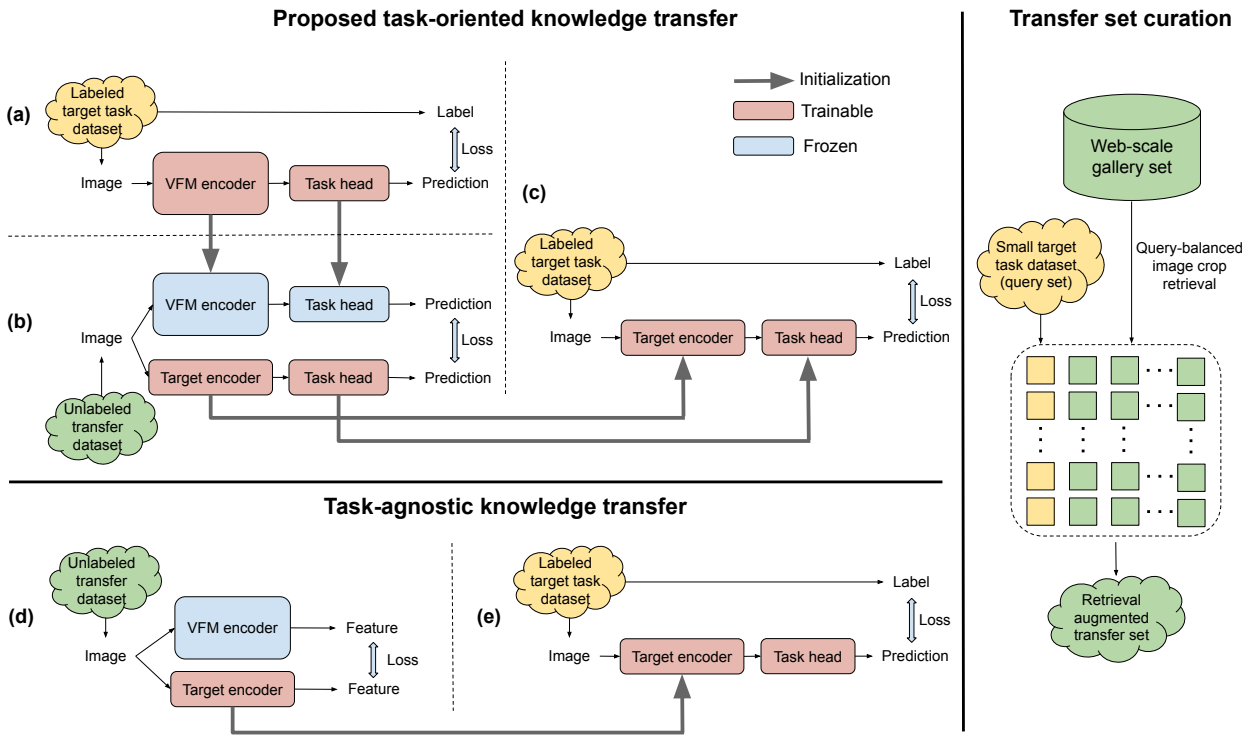


Figure 2. **Top-left:** Proposed task-oriented knowledge transfer approach that (a) first teaches the target task to the VFM using labeled target task data, (b) then uses this VFM to pretrain the target model by matching their target task predictions on an unlabeled transfer dataset, and (c) finally finetunes the target model using labeled target task data. **Bottom-left:** Alternative task-agnostic knowledge transfer approach that (d) first pretrains the target model by matching its features to the features extracted by the VFM on an unlabeled transfer dataset, and (e) then finetunes it using labeled target task data. **Right:** Transfer set curation using query-balanced image crop retrieval with a small target task dataset as the query set and a web-scale gallery set. By retrieving equal number of samples for each query, this approach increases the diversity of the retrieved samples. We perform crop-level retrieval to increase the chances of finding good matches.

aries. This is different from the knowledge distillation setting that only focuses on knowledge transfer between model architectures (Hinton et al., 2015; Tian et al., 2020) and the transfer learning setting that only focuses on knowledge transfer between tasks (Lu et al., 2021).

2. Approach and Contributions

In this work, we propose a simple and highly effective approach for transferring knowledge from a large pretrained VFM to a small task-specific model. This approach, referred to as ***task-oriented knowledge transfer***, first teaches the target task to the VFM using an appropriate task-specific head and limited labeled target task data, and then transfers task-oriented knowledge from the adapted VFM to the target model using the knowledge distillation framework of (Hinton et al., 2015) with a large unlabeled dataset, referred to as the *transfer set*. Finally, the target model is finetuned with limited labeled target task data (see Figure 2 Top-left).

An alternative approach to train a small task-specific model by leveraging a VFM is to first distill the VFM image encoder to the target model image encoder and then finetune the target model using limited labeled target task data (see Figure 2 Bottom-left). We refer to this approach as ***task-agnostic knowledge transfer***. Both task-oriented and task-agnostic knowledge transfer approaches leverage VFMs that have been trained on web-scale datasets. Instead, one could pretrain the small target model directly on a web-scale dataset. However, such pretraining can be extremely expensive. For example, training a MobileViT-V2 model (Mehta & Rastegari, 2023) using 0.7B image-text pairs following the CLIP approach of (Radford et al., 2021) takes around 1.2K A100 GPU hours just for one epoch.

We compare the proposed task-oriented knowledge transfer approach with task-agnostic knowledge transfer from VFMs, web-scale CLIP pretraining, supervised ImageNet pretraining, and self-supervised DINO (Caron et al., 2021) pretraining on five target tasks under limited labeled data settings [1] using two VFMs, namely DINOv2-ViT-L/14 (Oquab et al.,

[1]The target task data splits can be found at https://github.com/apple/ml-vfm-kt/tree/main.

2023) and OpenCLIP-ViT-L/14 (Ilharco et al., 2021), and two target mobile architectures, namely FastViT-S12 (Vasu et al., 2023) and MobileViT-V2-1.0. We experiment with two transfer sets, a generic CC3M (Sharma et al., 2018) transfer set and a target task-related transfer set, and present several insightful findings to the community. To the best of our knowledge, there is no existing work that concretely establishes the below findings about leveraging VFMs for training small task-specific models with limited labeled data.

- Task-oriented knowledge transfer outperforms task-agnostic transfer by a significant margin both in terms of performance (see Table 1) and training cost (see Figure 4). While VFMs can store vast knowledge due to their large capacity, small models may not be able to inherit this vast knowledge. Hence, transferring only task-oriented knowledge works better.
- Task-oriented knowledge transfer from VFMs outperforms the compute-intensive CLIP, widely-used supervised ImageNet and self-supervised DINO pretraining approaches by large margins (see Table 1), and is also computationally efficient (see Figures 1 and 4).
- While most of the existing distillation works use target task datasets as transfer sets, we systematically study the effect of transfer set. We show that task-oriented knowledge transfer outperforms ImageNet and CLIP pretraining without using task-related transfer sets (e.g., by using a large-scale generic web dataset such as CC3M, see Table 1). The performance further improves when large task-related unlabeled datasets are used as transfer sets (see Figure 3).
- When a large task-related transfer set is not readily available, we propose to curate such a dataset using web-scale image retrieval with the limited target task dataset as the query set (see Figure 2-Right). Retrieval-augmented transfer sets outperform the generic CC3M transfer set by a significant margin (see Table 4).

3. Experimental Analysis

3.1. Experimental Setup

We perform the target model training in two stages: *pretraining* followed by *finetuning*. During pretraining, we utilize a VFM by following the task-oriented and task-agnostic knowledge transfer approaches presented in Sec. 2 using a large unlabeled dataset as the transfer set. During finetuning, we train the model on a small labeled target task dataset.

Alternative approaches: *IN-Pretrain:* Supervised pretraining on 1.28M labeled images from the ImageNet-1K dataset (Deng et al., 2009). *CLIP-Pretrain:* Contrastive language image pretraining on 0.7B image-text pairs from (Cao et al., 2023) following the affinity mimicking CLIP distillation approach of (Wu et al., 2023). [2] *DINO-Pretrain:* Self-supervised pretraining on the unlabeled transfer set following the DINO approach of (Caron et al., 2021). We finetune the target model on labeled target task data after IN/CLIP/DINO pretraining.

Target tasks: We present results on five diverse target tasks, namely HAM10K skin disease classification (Tschandl, 2018), EuroSAT land cover classification (Helber et al., 2019), Places365 scene classification (Zhou et al., 2014), ImageNet object classification (Deng et al., 2009), and ADE20K semantic segmentation (Zhou et al., 2017). See Appendix A.1 for details of the corresponding datasets.

Transfer sets: For each target task, we experiment with two transfer sets. The first one is a generic transfer set consisting of 2.87M unlabeled images from the training split of the CC3M dataset (Sharma et al., 2018), and the second one is a task-related transfer set consisting of unlabeled images from the target task domain. For each task, we use the entire training split of the corresponding dataset as the task-related transfer set. See Appendix A.2 for further details. For ADE20K segmentation and EuroSAT classification tasks, we also experiment with transfer sets curated using image retrieval. See Appendix B for details of the retrieval process.

Vision foundation models: We use DINOv2-ViT-L/14 model (Oquab et al., 2023) and the OpenCLIP-ViT-L/14 model (Ilharco et al., 2021) trained on the DataComp-1B dataset (Gadre et al., 2023) as VFMs. For brevity, we refer to them as DINOv2 and OpenCLIP, respectively.

Target models: We use two recent efficient architectures, namely FastViT-S12 (Vasu et al., 2023) and MobileViT-V2-1.0 (Mehta & Rastegari, 2023), as image encoders for the target models. For brevity, we refer to them as FastViT and MViT-V2, respectively. We present the results for FastViT in this section and the results for MViT-V2 in Appendix C.2.

Task-specific heads: For classification tasks, the task-specific head is a linear classifier. The input to the classification layer is the final CLS token features for DINOv2 and OpenCLIP models, and the output of the final average pooling layer for MViT-V2 and FastViT models. For segmentation tasks, the task-specific head consists of a DeepLabV3 segmentation head (Chen et al., 2017) on top of the final spatial feature map and a spatial upsampling layer. As segmentation tasks requires high resolution features, we modify the image encoders such that their output spatial resolution is $1/8$ of the input image resolution. For DINOv2 and OpenCLIP models, we use a convolution stride of 8 instead of 14 for the patch embedding layer and resize the position

[2] (Wu et al., 2023) has shown that affinity mimicking distillation works better than the CLIP loss of (Radford et al., 2021). We use the best ViT-B/16 model from (Mehta et al., 2022) as teacher.

Improvement over	When using generic CC3M transfer set					When using target task-related transfer set				
	HAM10K	EuroSAT	Places365	ImageNet	ADE20K	HAM10K	EuroSAT	Places365	ImageNet	ADE20K
Task-agnostic transfer	0.38 - 2.91	1.25 - 5.27	1.88 - 4.89	1.90 - **11.6**	3.41 - 10.5	1.01 - 3.53	0.71 - 2.35	1.38 - 3.47	3.40 - **6.30**	1.11 - 6.25
CLIP-Pretrain	2.89 - 6.79	1.81 - 9.55	2.20 - 5.50	2.30 - 11.2	6.52 - **20.9**	5.93 - 11.4	4.24 - 15.5	3.38 - 7.72	5.30 - **22.1**	5.45 - 19.2
IN-Pretrain	0.93 - 4.27	0.21 - 3.73	2.19 - 8.17	-	5.12 - **13.7**	3.97 - 8.93	2.64 - 9.64	3.37 - 9.89	-	3.53 - **12.0**
DINO-Pretrain	0.20 - 2.89	2.59 - 6.94	2.88 - 8.46	2.43 - **29.8**	7.86 - 15.4	4.70 - 8.76	5.57 - 12.6	2.72 - 4.50	4.00 - **21.7**	8.84 - 18.1

Table 1. Performance gain of task-oriented knowledge transfer from VFMs when compared to other training strategies. These values are the minimum and maximum gains observed across all VFM and target model combinations we have experimented with for each dataset. Highest performance gain over each competing approach is highlighted in bold.

embeddings accordingly. For MViT-V2 and FastViT models, we change the last two stride-2 convolutions to stride-1 convolutions.

Adapting VFM to target task: When teaching a target task to a VFM, we consider two options: 1) training only the task-specific head with frozen VFM and 2) finetuning the entire VFM. We compare both models in terms of their target task performance and use the best one for task-oriented knowledge transfer. In our experiments with ViT-L/14 based VFMs, finetuning performed best in all cases (except when only 1% labeled data is available for ImageNet dataset).

Loss functions: For finetuning, we use cross entropy loss, and for matching task predictions, we use KL divergence. For segmentation tasks, these losses are applied at each pixel. The loss function used for matching features depends on the VFM. In the case of OpenCLIP model, we use contrastive loss (Tian et al., 2020). Since DINOv2 is trained to produce good patch features along with global image features, we evaluate both these features. We use contrastive loss with global image features and cosine similarity loss with patch features. See Appendix A.3 for additional training details.

3.2. Effectiveness of Task-oriented Knowledge Transfer

3.2.1. Performance Improvement

Figure 3(a) compares the performance of various approaches for different VFM and transfer set combinations on five downstream tasks with FastViT as the target image encoder. Please see Appendix C.1 for additional results corresponding to FastViT model, and Appendix C.2 for results corresponding to MViT-V2 model. For task-agnostic knowledge transfer from DINOv2, we experiment with both global image features and patch features and report the best results.

The proposed task-oriented knowledge transfer from VFMs outperforms task-agnostic transfer from VFMs, ImageNet, CLIP and DINO pretraining approaches for all target tasks irrespective of whether the generic CC3M transfer set or a target task-related transfer set is used. Table 1 shows the minimum and maximum performance gains of task-oriented knowledge transfer when compared to other pretraining strategies. Specifically, when using the generic CC3M transfer set, the performance gains are up to 11.6%, 20.9%, 13.7% and 29.8% when compared to task-agnostic transfer, CLIP, ImageNet, and DINO pretraining approaches, respectively. When using target task-related transfer sets, the performance gains are up to 6.3%, 22.1%, 12% and 21.7% when compared to task-agnostic transfer, CLIP, ImageNet and DINO pretraining approaches, respectively.

When task-related transfer set is used, task-oriented transfer even outperforms finetuned VFM for some of the tasks when the amount of labeled data is small (see red and black curves in Figure 3)). This could be because the knowledge transfer process uses large-scale unlabeled target domain data that is not used while finetuning VFMs.

3.2.2. Training Efficiency

The main focus of this work is on improving the performance of the target model under limited labeled data settings. Hence, for each competing approach, we train the target model long enough to achieve the best possible performance with the given limited labeled data. Figure 4 compares various approaches in terms of their pretraining compute cost measured using A100 GPU hours (GPUh) with FastViT as the target model. Task-oriented knowledge transfer from VFMs achieves better performance with significantly less training compute when compared to other pretraining strategies. Specifically, it demonstrates up to $9\times$, $4\times$ and $15\times$ reduction in training compute when compared to task-agnostic knowledge transfer, ImageNet pretraining and DINO pretraining, respectively, while outperforming them. Web-scale CLIP pretraining of the target model performs significantly worse than task-oriented knowledge transfer while using significantly more compute (43K GPUh).

Note that the training compute for task-oriented knowledge transfer includes the compute used for adapting the VFM to the target task. Since VFMs are trained only on small labeled datasets, the compute cost of adapting them is significantly lower than the compute cost of distilling them on large unlabeled datasets. For example, while finetuning DINOv2 VFM on Places365 dataset using 250 images per class for 100 epochs takes 32 GPUh, distilling the finetuned VFM on Places365 transfer set for 100 epochs takes 180

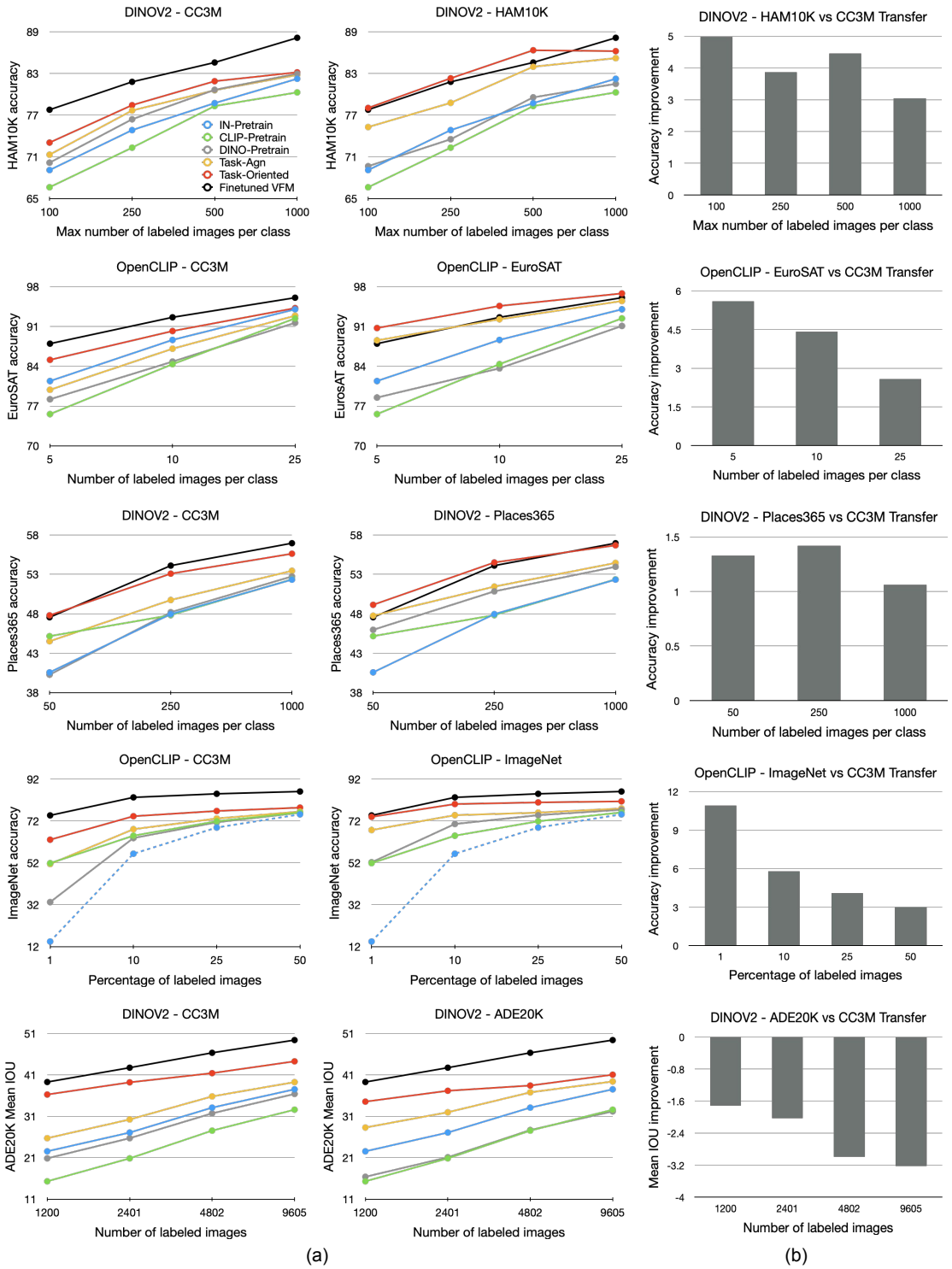


Figure 3. (a) Comparison of various approaches for different (VFM, transfer set) combinations with FastViT as the target image encoder. (b) Performance improvement when unlabelled target task data is used instead of generic CC3M dataset for task-oriented knowledge transfer. The target tasks are HAM10K classification, EuroSAT classification, Places365 classification, ImageNet classification and ADE20K segmentation from top to bottom. Task-oriented knowledge transfer from VFMs (red curves) clearly outperforms alternative training strategies. The performance of finetuned VFMs used for knowledge transfer is also shown here for reference (black curves). When the target task is ImageNet classification, the blue curve corresponds to training from scratch instead of ImageNet pretraining.

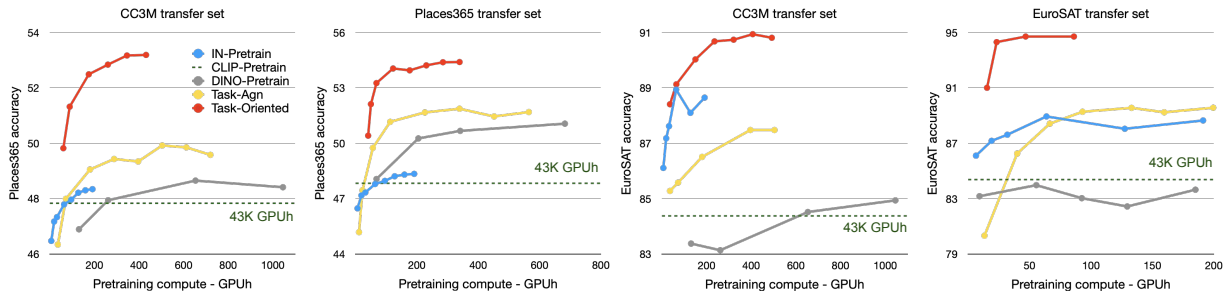


Figure 4. Comparison of various approaches in terms of their pretraining compute. The left two figures correspond to Places365 classification (250 training images per class) and the right two figures correspond to EuroSAT classification (10 training images per class). Here, we use DINOv2 VFM for knowledge transfer and FastViT as the target architecture. Each curve in this figure was obtained by evaluating intermediate checkpoints of one training run. CLIP pretraining is represented using dashed green line.

Dataset	HAM10K	EuroSAT	Places365	ImageNet
Labeled data	100 img/cls	10 img/cls	250 img/cls	50%
VFM	DINOv2	OpenCLIP	OpenCLIP	OpenCLIP
VFM performance				
VFM-LP	71.10	87.33	52.39	84.07
VFM-FT	76.92	92.63	54.48	86.05
Task-oriented knowledge transfer with generic CC3M transfer set				
VFM-LP	69.97	89.59	52.87	78.13
VFM-FT	73.06	90.21	53.33	78.40
Task-oriented knowledge transfer with target task-related transfer set				
VFM-LP	74.60	94.37	54.69	81.01
VFM-FT	78.04	94.63	54.82	81.43

Table 2. Comparison between linear-probed (LP) and finetuned (FT) VFMs in terms of their effectiveness for task-oriented knowledge transfer to FastViT target model.

Pretraining dataset - CC3M transfer set			
ADE20K	Number of labeled training images	1200	2401
	Performance gain	2.19	2.03
ImageNet	Percentage of labeled training images	10	25
	Performance gain	3.8	5.8
HAM10K	Maximum labeled training images per class	250	500
	Performance gain	9.99	7.41
EuroSAT	Number of labeled training images per class	5	10
	Performance gain	3.3	4.9
Places365	Number of labeled training images per class	250	1000
	Performance gain	0.47	0.76
Pretraining dataset - Target task-related transfer set			
ADE20K	Number of labeled training images	1200	2401
	Performance gain	1.23	1.25

Table 3. Performance gain (for FastViT model) due to the final finetuning step of task-oriented knowledge transfer approach.

GPUh. Similarly, while finetuning DINOv2 VFM on EuroSAT dataset using 10 images per class for 200 epochs takes 4 GPUh, distilling the finetuned VFM on EuroSAT transfer set for 5K epochs takes 40 GPUh [3].

3.2.3. VFM Finetuning vs Linear Probing

We consider both linear probing and full finetuning when adapting a VFM to a target task, and use the best-performing VFM for task-oriented knowledge transfer. In almost all of our experiments, finetuning resulted in a better performing VFM when compared to linear probing. In this section, we compare linear-probed and finetuned VFMs in terms of their effectiveness for task-oriented knowledge transfer to FastViT target model (see Table 2). While transferring knowledge from finetuned VFM outperforms transferring from linear-probed VFM by a significant margin in the case of HAM10K dataset ($> 3\%$), the performance gap is small for other datasets ($< 0.5\%$). It is interesting to see that the performance gap between target models distilled from linear-probed and finetuned VFMs is significantly smaller than the performance gap between the linear-probed and finetuned VFMs themselves. This suggests that *full finetuning may not be needed for effective task-oriented knowledge transfer, especially as VFMs become more powerful*.

Note that the small target model exhibits better performance than the teacher VFM on many datasets when a task-related transfer set is used. This is potentially because the small model has been exposed to more task-related samples (additional unlabeled samples) when compared to teacher VFM.

3.2.4. Contributions of Finetuning Step

The last step of the proposed task-oriented knowledge transfer approach is to finetune the target model on limited labeled data (see Figure 2(c)). One may wonder to what extent this finetuning step contributes to the performance, given that the target model learns to follow a VFM finetuned on the target task in the pretraining step. Table 3 shows the

[3] Close to best target task performance is observed after distilling finetuned DINOv2 for about 100 epochs on Places365 transfer set and 5K epochs on EuroSAT transfer set.

performance improvements due to the finetuning step. We see significant gains when generic CC3M transfer set is used during pretraining. When target task-related transfer set is used during pretraining, we see significant gains in the case of ADE20K segmentation task. We did not observe noticeable gains for the classification tasks in this case.

3.3. Importance of Transfer Set

Figure 3(b) shows the improvement in accuracy on five downstream tasks when target task-related transfer sets are used instead of generic CC3M transfer set for task-oriented knowledge transfer from various VFMs to FastViT target model. Please see Appendix C.1 for additional results with FastViT target model and Appendix C.2 for results corresponding to MViT-V2 target model.

Using task-related transfer sets improves the accuracy by up to 5%, 7.3%, 2.3% and 10.9% for HAM10K, EuroSAT, Places365 and ImageNet classification tasks, respectively (including the results presented in Appendices C.1 and C.2). This underscores the importance of the relevance of transfer set to the target task. However, for ADE20K segmentation task, the generic CC3M transfer set performs better. We conjecture that the main reason for this is the size of ADE20K transfer set which has only 19.2K images. We address this issue by curating a larger task-related transfer set using image retrieval as shown in the next section.

3.4. Retrieval Augmented Knowledge Transfer

Our results on HAM10K, EuroSAT, Places365 and ImageNet classification tasks show that using a sufficiently large target task-related transfer set performs better than a generic transfer set such as CC3M. However, such large transfer sets may not be readily available for every task. We propose to address this issue by curating large task-related transfer sets using image retrieval as shown in Figure 2-Right.

Specifically, we use the limited target task labeled dataset as the query set $\mathcal{Q}$ and a large set of images sourced from web as the gallery $\mathcal{G}$. We use an encoder network ϕ to map all the images to a d-dimensional embedding space, and perform retrieval based on Euclidean distances in this space following a query balanced approach. For a query image $x_q \in \mathcal{Q}$, we define k-NN(x_q) to be the set of its k nearest neighbors from $\mathcal{G}$. To retrieve N images in total, we find the smallest k for which $\bigcup_{x_q \in \mathcal{Q}} k\text{-NN}(x_q)$ contains at least N unique images. If $\bigcup_{x_q \in \mathcal{Q}} k\text{-NN}(x_q)$ contains more than N images, we drop the kth neighbor of randomly selected queries until the retrieved set has N images. By giving equal weight to all the queries, this approach increases the diversity of the retrieved samples. We combine the retrieved set with the initial query set and use this retrieval augmented transfer set for task-oriented knowledge transfer.

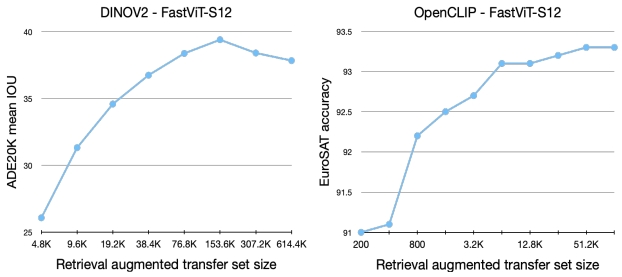


Figure 5. Performance of task-oriented knowledge transfer using retrieval augmented transfer sets of varying sizes. The number of labeled images used for finetuning and also as retrieval queries is 4800 for ADE20K dataset and 100 for EuroSAT dataset.

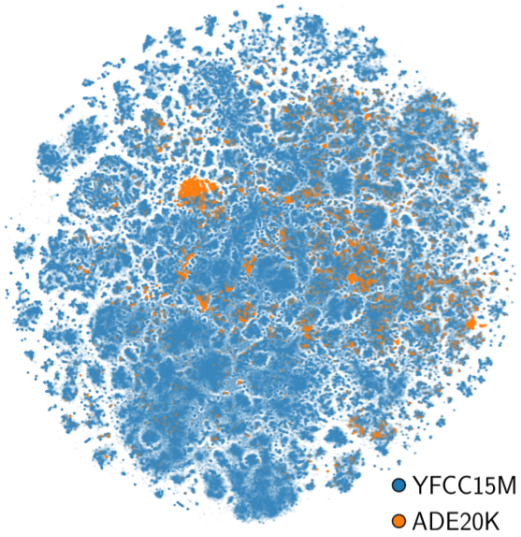


Figure 6. t-SNE (Van der Maaten & Hinton, 2008) visualization of image features of ADE20K dataset and randomly sampled 10% of YFCC15M dataset.

3.4.1. ADE20K Segmentation

We use YFCC15M dataset (Radford et al., 2021) which contains 15M images as the gallery set, OpenCLIP-ViT-L/14 image encoder (Ilharco et al., 2021) trained on the DataComp-1B dataset (Gadre et al., 2023) as the encoder network, and experiment with the combination of DINOv2 VFM and FastViT target model. See Appendix B.1 for details of the retrieval process and Figure 9 in Appendix B.6 for some examples of retrieval results.

Figure 5 (left) shows ADE20K segmentation performance of task-oriented knowledge transfer with retrieval augmented transfer sets of different sizes. Here, we use 4800 labeled images as both finetuning dataset and retrieval query set. The segmentation performance increases with the transfer set size until we reach 154K images and drops after that. Figure 6 displays a t-SNE visualization of image features from the ADE20K dataset and a randomly sampled 10% of

ADE20K mean IOU (transfer from DINOv2 VFM)

Labeled images / Query set size	1200	2401	4802	9605
Full ADE20K transfer (19.2K)	34.57	37.19	38.45	41.07
CC3M transfer (2.87M)	36.28	39.22	41.44	44.29
Retrieval augmented transfer (154K)	37.65	40.40	43.28	44.93

EuroSAT accuracy (transfer from OpenCLIP VFM)

Labeled images / Query set size	50	100	250
Full EuroSAT transfer (2.7K)	90.74	94.63	96.83
CC3M transfer (2.87M)	85.14	90.21	94.25
Retrieval augmented transfer (51K)	89.23	93.25	96.35

Table 4. Comparison between various transfer sets in terms of their effectiveness for task-oriented knowledge transfer.

the YFCC15M dataset. Images from the ADE20K dataset occupy only a small region suggesting that only a small subset of YFCC15M is relevant for ADE20K segmentation. Hence, as we increase the size of retrieved dataset, after certain point, low quality matches will be added to the transfer set resulting in performance drop.

Using 154K as the target transfer set size, we curate different transfer sets by varying the number of query images used for retrieval. Table 4 (top) compares these curated transfer sets with the generic CC3M and the full ADE20K transfer sets. Retrieval augmented transfer sets clearly outperform both ADE20K and generic CC3M transfer sets. [4]

3.4.2. EuroSAT Classification

Since this dataset contains specialized imagery (satellite images), we use a much larger gallery to increase the chances of finding good matches. Specifically, we use the DataComp-1B dataset (Gadre et al., 2023) which contains 1.28B images filtered from Common Crawl (web-crawled data). We take 10 randomly augmented crops from each image and create a gallery of 12.8B images. See Appendix B.2 for additional details about this gallery set and the retrieval process. We show some examples of retrieval results in Figure 7. Our crop-level retrieval strategy enables the extraction of relevant regions from gallery images that are both in-domain and out-of-domain relative to the query. See Figure 10 in Appendix B.6 for additional examples.

Figure 5(right) shows EuroSAT classification performance using task-oriented knowledge transfer with retrieval augmented transfer sets of different sizes. Here, we use a dataset of 100 labeled images (10 images per class) both as fine-tuning dataset and retrieval query set. The performance saturates at about 51K. Using 51K as the target size, we curate different transfer sets by varying the number of query images used for retrieval. Table 4 (bottom) compares these curated transfer sets with the generic CC3M and the full EuroSAT transfer sets. Retrieval augmented transfer sets curated using small query sets outperform the generic CC3M transfer set by a significant margin, and perform competitively when compared to the full EuroSAT transfer set.

[4]The best performance in Figure 5(left) is lower than the performance for curated transfer set corresponding to 4800 query images in Table 4(top). This is because, we used shorter training runs (60K steps) to get the results in Figure 5(left), and once we identified the best transfer set size, we used longer training runs (200K steps) to get the results in Table 4(top).

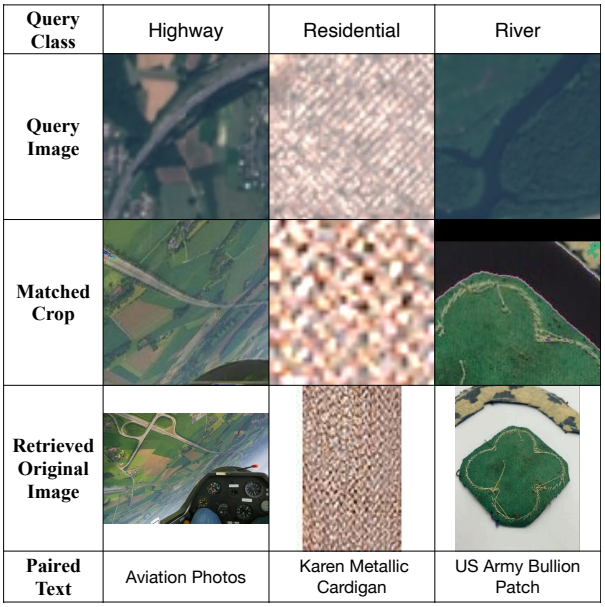


Figure 7. Images retrieved from DataComp-1B dataset for queries from the EuroSAT dataset. Proposed cropping and augmentation-based retrieval strategy selects only the relevant regions from gallery images (see the Highway example). It also enables the identification of regions with similar patterns from out-of-domain images (see the Residential and River examples).

4. Related Works

Knowledge distillation is a widely-used approach for transferring knowledge between model architectures by training one model to mimic the outputs of another model. Numerous distillation approaches have been proposed over the past decade based on various knowledge representations such as task logits (Hinton et al., 2015), intermediate features or embeddings (Heo et al., 2019; Tian et al., 2020), relations between samples (Park et al., 2019), attention maps (Zagoruyko & Komodakis, 2017), etc. Please refer to Wang & Yoon (2022); Hu et al. (2023) for an overview of existing distillation approaches. Some recent works have specifically focused on multi-modal distillation of image-language models (Fang et al., 2021; Li et al., 2023c; Wang et al., 2023a; Sun et al., 2023; Yang et al., 2023; Vasu et al., 2024). In addition to transferring knowledge between model architectures, this work also focuses on transferring knowledge between tasks.

Transfer learning, where a model is first pretrained on a data-rich task before being partially or fully finetuned on a downstream task, has been well studied over the past decade (Niu et al., 2020; Lu et al., 2021). Recently, (Entezari et al., 2023) compared various pretraining approaches and showed that supervised ImageNet training and large-scale CLIP training are effective pretraining strategies for several downstream vision tasks. While the standard transfer learning setting focuses on transferring knowledge only between tasks, this work focuses on transferring knowledge between both tasks and model architectures.

Image retrieval has been used by various recent works to curate training datasets (Udandarao et al., 2022; Li et al., 2023a; Xu et al., 2023; Wallingford et al., 2023; Liu et al., 2023). While (Li et al., 2023a) focuses on self-supervised learning, the remaining works focus on training or adapting vision-language models. Different from these works, we use retrieval to curate task-related datasets for transferring knowledge from VFMs to small task-specific models.

Self-supervised learning, which uses unlabeled data to obtain a good initial feature representation, has received significant attention in the recent past, and several approaches have been proposed based on contrastive learning (Chen et al., 2020; He et al., 2020), distillation (Grill et al., 2020; Chen & He, 2021; Caron et al., 2021), redundancy reduction (Zbontar et al., 2021), clustering (Caron et al., 2018; 2020) and image inpainting (He et al., 2022; Bao et al., 2022). Please refer to Ozbulak et al. (2023) for a detailed review of existing self-supervised learning approaches.

Semi-supervised learning approaches leverage both labeled and unlabeled data to improve the final task performance. They focus on effectively propagating label information from a labeled dataset to an unlabeled dataset (Lee, 2013; Xie et al., 2020b), and training the network using consistency constraints on the unlabeled samples (Tarvainen & Valpola, 2017; Berthelot et al., 2019; Xie et al., 2020a; Sohn et al., 2020; Verma et al., 2022). Please refer to (Chen et al., 2022) for a recent survey of semi-supervised approaches.

Task-agnostic knowledge transfer from VFMs has been explored in recent works such as DINOv2 (Oquab et al., 2023) and GSD (Huang et al., 2023). However, they did not evaluate the effectiveness of task-agnostic transfer in limited labeled data settings. Also, they use distillation approaches specifically designed for transformer architectures.

Task-oriented knowledge transfer has been recently explored in the context of large langugae models (LLMs) by (Hsieh et al., 2023; Fu et al., 2023). These approaches use the rationales extracted from LLMs by chain-of-thought prompting to train small task-specific models. In this work, we focus on vision foundation models.

Recently, (Borup et al., 2023) has also focused on training small task-specific models with limited labeled data. They use a pool of small CNN models trained on small-scale classification datasets as source models and focus on finding a set of suitable source models to distill for a given target task. Different from this work, we focus on knowledge transfer from recent VFMs that have been trained on massive datasets, and show that transferring knowledge from a single VFM outperforms various popular pretraining strategies. While (Borup et al., 2023) assumes the existence of large-scale unlabeled target domain data, we show that knowledge transfer from VFMs is effective even when using generic web data such as CC3M. We also propose a retrieval-based approach for curating effective task-related transfer sets.

5. Conclusions

In this work, we proposed a simple yet highly effective task-oriented knowledge transfer approach for training small task-specific models by leveraging pretrained VFMs. Our experimental results on five target tasks show that the proposed approach outperforms task-agnostic VFM distillation, web-scale CLIP pretraining, supervised ImageNet pretraining, and self-supervised DINO pretraining approaches by a significant margin both in terms of target task performance and training efficiency. We showed that while transferring knowledge from VFMs using task-related transfer sets works best, using general web data such as CC3M is also highly effective when compared to popular ImageNet, CLIP and DINO pretraining approaches. We also proposed a crop-based retrieval strategy to curate large task-related transfer sets, and experimentally demonstrated its effectiveness. In summary, this work advocates for *pretraining small task-specific models by transferring task-oriented knowledge from VFMs using large task-related transfer sets when they are available and retrieval-augmented transfer sets in the absence of large task-related transfer sets.*

Future work: In this work, we only used labeled target task data to finetune VFMs. A potential future work is to leverage additional large-scale unlabeled data to better adapt VFMs to the target task/domain, thereby eventually improving the small task-specific model trained with knowledge transfer from VFMs. However, finetuning VFMs on large scale datasets could be computationally expensive. Another potential future work is to explore more sophisticated transfer set curation strategies based on active learning and adversarial filtering that may result in better transfer sets that lead to better target model performance.

Limitations: Though our crop-based retrieval strategy enables fine-grained retrieval from a diverse set of images, curating large transfer sets could still be difficult for some specialized domains such as industrial automation that are not well covered by web data.

Acknowledgements

We thank Ting-Yao Hu, Karren Yang, Mohammad Samragh Razlighi, Pavan Kumar Anasosalu Vasu, and Barry Theobald from Apple for their valuable suggestions.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Bao, H., Dong, L., Piao, S., and Wei, F. Beit: BERT pre-training of image transformers. In *ICLR*, 2022.

Berthelot, D., Carlini, N., Goodfellow, I. J., Papernot, N., Oliver, A., and Raffel, C. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019.

Borup, K., Phoo, C. P., and Hariharan, B. Distilling from similar tasks for transfer learning on a budget. *arXiv:2304.12314*, 2023.

Cao, L., Zhang, B., Chen, C., Yang, Y., Du, X., Zhang, W., Lu, Z., and Zheng, Y. Less is more: Removing text-regions improves CLIP training efficiency and robustness. *arXiv:2305.05095*, 2023.

Caron, M., Bojanowski, P., Joulin, A., and Douze, M. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Chen, L., Papandreou, G., Schroff, F., and Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv:1706.05587*, 2017.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Chen, X. and He, K. Exploring simple siamese representation learning. In *CVPR*, 2021.

Chen, Y., Mancini, M., Zhu, X., and Akata, Z. Semi-supervised and unsupervised deep visual learning: A survey. *arXiv:2208.11296*, 2022.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR Workshops*, 2020.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Entezari, R., Wortsman, M., Saukh, O., Shariatnia, M., Sedghi, H., and Schmidt, L. The role of pre-training data in transfer learning. *arXiv:2302.13602*, 2023.

Fang, Z., Wang, J., Hu, X., Wang, L., Yang, Y., and Liu, Z. Compressing visual-linguistic model via knowledge distillation. In *ICCV*, 2021.

Fu, Y., Peng, H., Ou, L., Sabharwal, A., and Khot, T. Specializing smaller language models towards multi-step reasoning. In *ICML*, 2023.

Gadre, S. Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., Orgad, E., Entezari, R., Daras, G., Pratt, S., Ramanujan, V., Bitton, Y., Marathe, K., Mussmann, S., Vencu, R., Cherti, M., Krishna, R., Koh, P. W., and' Alexander Ratner, O. S., Song, S., Hajishirzi, H., Farhadi, A., Beaumont, R., Oh, S., Dimakis, A., Jitsev, J., Carmon, Y., Shankar, V., and Schmidt, L. Datacomp: In search of the next generation of multimodal datasets. *arXiv:2304.14108*, 2023.

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

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. B. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. B. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

Helber, P., Bischke, B., Dengel, A., and Borth, D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 12(7):2217–2226, 2019.

Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., and Choi, J. Y. A comprehensive overhaul of feature distillation. In *ICCV*, 2019.

Hinton, G. E., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015.

Hsieh, C., Li, C., Yeh, C., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C., and Pfister, T. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *ACL*, 2023.

Hu, C., Li, X., Liu, D., Wu, H., Chen, X., Wang, J., and Liu, X. Teacher-student architecture for knowledge distillation: A survey. *arXiv:2308.04268*, 2023.

Huang, W., Peng, Z., Dong, L., Wei, F., Jiao, J., and Ye, Q. Generic-to-specific distillation of masked autoencoders. In *CVPR*, 2023.

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., and Schmidt, L. Openclip. 2021. URL https://doi.org/10.5281/zenodo.5143773.

Lee, D.-H. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshops*, 2013.

Li, A. C., Brown, E. L., Efros, A. A., and Pathak, D. Internet explorer: Targeted representation learning on the open web. In *ICML*, 2023a.

Li, J., Li, D., Savarese, S., and Hoi, S. C. H. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023b.

Li, X., Fang, Y., Liu, M., Ling, Z., Tu, Z., and Su, H. Distilling large vision-language model with out-of-distribution generalizability. *arXiv:2307.03135*, 2023c.

Liu, H., Son, K., Yang, J., Liu, C., Gao, J., Lee, Y. J., and Li, C. Learning customized visual models with retrieval-augmented knowledge. In *CVPR*, 2023.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *ICLR*, 2019.

Lu, Y., Luo, L., Huang, D., Wang, Y., and Chen, L. Knowledge transfer in vision recognition: A survey. *ACM Comput. Surv.*, 53(2):37:1–37:35, 2021.

Mehta, S. and Rastegari, M. Separable self-attention for mobile vision transformers. *Trans. Mach. Learn. Res.*, 2023.

Mehta, S., Naderiparizi, S., Faghri, F., Horton, M., Chen, L., Farhadi, A., Tuzel, O., and Rastegari, M. Rangeaugment: Efficient online augmentation with range learning. *arXiv:2212.10553*, 2022.

Niu, S., Liu, Y., Wang, J., and Song, H. A decade survey of transfer learning (2010-2020). *IEEE Trans. Artif. Intell.*, 1(2):151–166, 2020.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2023.

Ozbulak, U., Lee, H. J., Boga, B., Anzaku, E. T., Park, H., Messem, A. V., Neve, W. D., and Vankerschaver, J. Know your self-supervised learning: A survey on image-based generative and discriminative training. *arXiv:2305.13689*, 2023.

Park, W., Kim, D., Lu, Y., and Cho, M. Relational knowledge distillation. In *CVPR*, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C., Cubuk, E. D., Kurakin, A., and Li, C. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020.

Sun, X., Zhang, P., Zhang, P., Shah, H., Saenko, K., and Xia, X. DIME-FM: distilling multimodal and efficient foundation models. *arXiv:2303.18232*, 2023.

Tarvainen, A. and Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017.

Tian, Y., Krishnan, D., and Isola, P. Contrastive representation distillation. In *ICLR*, 2020.

Tschandl, P. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. 2018. doi: 10.7910/DVN/DBW86T. URL https://doi.org/10.7910/DVN/DBW86T.

Udandarao, V., Gupta, A., and Albanie, S. Sus-x: Training-free name-only transfer of vision-language models. *arXiv:2211.16198*, 2022.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Vasu, P. K. A., Gabriel, J., Zhu, J., Tuzel, O., and Ranjan, A. Fastvit: A fast hybrid vision transformer using structural reparameterization. *arXiv:2303.14189*, 2023.

Vasu, P. K. A., Pouransari, H., Faghri, F., Vemulapalli, R., and Tuzel, O. Mobileclip: Fast image-text models through multi-modal reinforced training. *CVPR*, 2024.

Verma, V., Kawaguchi, K., Lamb, A., Kannala, J., Solin, A., Bengio, Y., and Lopez-Paz, D. Interpolation consistency training for semi-supervised learning. *Neural Networks*, 145:90–106, 2022.

Wallingford, M., Ramanujan, V., Fang, A., Kusupati, A., Mottaghi, R., Kembhavi, A., Schmidt, L., and Farhadi, A. Neural priming for sample-efficient adaptation. *arXiv:2306.10191*, 2023.

Wang, L. and Yoon, K. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(6): 3048–3068, 2022.

Wang, T., Zhou, W., Zeng, Y., and Zhang, X. Efficientvlm: Fast and accurate vision-language models via knowledge distillation and modal-adaptive pruning. In *ACL*, 2023a.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., and Wei, F. Image as a foreign language: BEIT pretraining for vision and vision-language tasks. In *CVPR*, 2023b.

Wu, K., Peng, H., Zhou, Z., Xiao, B., Liu, M., Yuan, L., Xuan, H., Valenzuela, M., Chen, X., Wang, X., Chao, H., and Hu, H. Tinyclip: CLIP distillation via affinity mimicking and weight inheritance. In *ICCV*, 2023.

Xie, Q., Dai, Z., Hovy, E. H., Luong, T., and Le, Q. Unsupervised data augmentation for consistency training. In *NeurIPS*, 2020a.

Xie, Q., Luong, M., Hovy, E. H., and Le, Q. V. Self-training with noisy student improves imagenet classification. In *CVPR*, 2020b.

Xu, H., Xie, S., Huang, P., Yu, L., Howes, R., Ghosh, G., Zettlemoyer, L., and Feichtenhofer, C. Cit: Curation in training for effective vision-language data. *arXiv:2301.02241*, 2023.

Yang, C., An, Z., Huang, L., Bi, J., Yu, X., Yang, H., and Xu, Y. CLIP-KD: an empirical study of distilling CLIP models. *arXiv:2307.12732*, 2023.

Yuan, L., Chen, D., Chen, Y., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., Liu, C., Liu, M., Liu, Z., Lu, Y., Shi, Y., Wang, L., Wang, J., Xiao, B., Xiao, Z., Yang, J., Zeng, M., Zhou, L., and Zhang, P. Florence: A new foundation model for computer vision. *arXiv:2111.11432*, 2021.

Zagoruyko, S. and Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021.

Zhou, B., Lapedriza, À., Xiao, J., Torralba, A., and Oliva, A. Learning deep features for scene recognition using places database. In *NeurIPS*, 2014.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. Scene parsing through ADE20K dataset. In *CVPR*, 2017.

A. Experimental Setup

A.1. Target Task Datasets

Places365 scene classification (Zhou et al., 2014): This dataset has 1.8M training and 36.5K validation images. We split the original validation set into two subsets consisting of 3.65K and 32.85K images, and use them for validation and testing, respectively. This is a 365-class classification task.

HAM10K skin lesion disease classification (Tschandl, 2018): This dataset consists of 10K training, 193 validation and 1.5K test images. HAM10K dataset is highly imbalanced with just 115 training images in the smallest class and 6705 training images in the largest class. When experimenting with N training images per class, if a class does not have N images, we just use all the images from that class. This is a 7-class classification task.

EuroSAT land cover classification (Helber et al., 2019): This dataset consists of 27K images. We follow the most difficult setting in (Helber et al., 2019) and use 10% of the dataset (2.7K images) as training split. We split the remaining 90% dataset into two splits consisting of 5.4K and 18.9K images, and use them for validation and testing, respectively. This is a 10-class classification task.

ADE20K semantic segmentation (Zhou et al., 2017): This dataset consists of 20.2K training and 2K validation images. We split the original training set into two subsets with 19.2K and 1K images, and use them for training and validation, respectively. We use the original 2K validation set as the test set. This is a semantic segmentation task with 150 semantic classes.

ImageNet (Deng et al., 2009): This dataset consists of 1.28M training images and 50K validation images. We report final accuracy results on the validation split. This is a 1000-class classification task.

The training, validation and test splits used in this work can be found at `https://github.com/apple/ml-vfm-kt/tree/main`.

A.2. Transfer Sets

The generic CC3M transfer set consists of 2.87M images. For each target task, we use the entire training split of the corresponding dataset as the task-related transfer set. So, the Places365, ImageNet, AD20K, HAM10K and EuroSAT transfer sets consist of 1.8M, 1.28M, 19.2K, 10K and 2.7K images, respectively.

A.3. Training Details

We use AdamW optimizer (Loshchilov & Hutter, 2019) with cosine learning rate decay in all our experiments. We use input resolutions of 256×256 and 512×512 for classification and segmentation tasks, respectively.

Loss functions: The loss function used for matching features depends on the VFM. In the case of OpenCLIP model, we use contrastive loss (Tian et al., 2020) with a linear projection layer on top of the target model output to match its dimensionality with the CLIP embedding dimensionality ($d = 768$). In the case of DINOv2 global image features, we use contrastive loss with a linear projection layer ($d = 768$) on outputs of both DINOv2 and target models. In the case of DINOv2 patch features, we use cosine similarity loss with a linear projection layer on top of the target model features to match their dimensionality with DINOv2 feature dimensionality ($d = 1024$). We also resize DINOv2 patch features to match the spatial resolution of the target model features.

Augmentations and resolution: We adopt the advanced image augmentations used for ImageNet-1k in (Mehta & Rastegari, 2023) and an input resolution of 256×256 for ImageNet pretraining, task-agnostic VFM distillation, VFM classification finetuning, classification task logits distillation, and target classification model finetuning. We adopt the segmentation task-related augmentations used in (Mehta & Rastegari, 2023) and an input resolution of 512×512 for VFM segmentation finetuning, segmentation task logits distillation and target segmentation model finetuning. We use the augmentations from (Caron et al., 2021) for DINO pretraining and the augmentations from (Radford et al., 2021) for CLIP pretraining with an input resolution of 256×256.

VFM finetuning: When finetuning on Places365 dataset, we use a batch size of 512. We train for 200 epochs when 50 labeled images per class are used and 100 epochs when 250/1000 labeled images per class are used. When finetuning on HAM10K dataset, we use a batch size of 128 and train for 200 epochs. When finetuning on EuroSAT dataset, we train for

500 epochs with a batch size of 50 when using 5 labeled images per class, 400 epochs with a batch size of 100 when using 10 labeled images per class, and 300 epochs with a batch size of 128 when using 25 labeled images per class. When finetuning on ImageNet dataset, we use a batch size of 512 and train for 25 epochs when using 10%, 25% and 50% labeled data. When using 1% labeled data, we only train the linear classifier head for 100 epochs with a batch size of 1024. When finetuning on ADE20K dataset, we use a batch size of 32 and train for 300/250/200/100 epochs when using 1200/2401/4802/9605 labeled images. We run each finetuning experiment with several learning rates from $7e^{-6}$ to $3e^{-3}$ and report results corresponding to best ones.

ImageNet pretraining: We use a learning rate of $2e^{-3}$ and train with a batch size of 1024 for 300 epochs.

CLIP pretraining: We train for 200K iterations on 0.7B image-text pairs from (Cao et al., 2023) using a learning rate of $5e^{-4}$ and a batch size of 65.5K.

DINO pretraining: We use a learning rate of $3e^{-3}$ and train for 100, 200, 300, 10K, 15K and 20K epochs on CC3M, Places, ImageNet, EuroSAT, ADE20K and HAM10K datasets, respectively, with a batch size of 1024.

Task-agnostic VFM feature distillation: We train for 100 epochs with a batch size of 2048 when using CC3M transfer set. We train for 200 epochs with a batch size of 2048 when using Places365 and ImageNet transfer sets. We train for 10K and 20K epochs with a batch size of 1024 when using ADE20K and HAM10K transfer sets, respectively. When using EuroSAT transfer set, we train for 10K epochs using a batch size of 2048. We use a learning rate $1e^{-3}$ for these distillation experiments.

Distillation of finetuned VFM: When using CC3M transfer set, we train for 100 epochs with a learning rate of $7e^{-4}$ and batch size of 512 when distilling VFMs finetuned for Places365 and HAM10K classification tasks, 100 epochs with a learning rate of $7e^{-4}$ and a batch size of 1024 when distilling VFMs finetuned for EuroSAT classification task, 150 epochs with a learning rate of $3e^{-3}$ and a batch size of 1024 when distilling VFMs finetuned for ImageNet classification task, 15 epochs with a learning rate of $7e^{-4}$ and a batch size of 128 when distilling VFMs finetuned for ADE20K segmentation task. When using task-related transfer sets, we train for 200 epochs with a learning rate of $7e^{-4}$ and batch size of 512 when distilling VFMs finetuned for Places365 classification task, 7K epochs with a learning rate of $7e^{-4}$ and a batch size of 128 when distilling VFMs finetuned for HAM10K classification task, 10K epochs with a learning rate of $7e^{-4}$ and a batch size of 1024 when distilling VFMs finetuned for EuroSAT classification task, 300 epochs with a learning rate of $3e^{-3}$ and a batch size of 1024 when distilling VFMs finetuned for ImageNet classification task, 2K epochs with a learning rate of $7e^{-4}$ and a batch size of 128 when distilling VFMs finetuned for ADE20K segmentation task. When using curated transfer sets, we train for 550K steps with a learning rate of $7e^{-4}$ and a batch size of 128 when distilling VFMs finetuned for HAM10K classification task, 200K steps with a learning rate of $7e^{-4}$ and a batch size of 128 when distilling VFMs finetuned for ADE20K segmentation task, and 30K steps with a learning rate of $7e^{-4}$ and a batch size of 1024 when distilling VFMs finetuned for EuroSAT classification task.

Target model finetuning: When finetuning on Places365 dataset, we use a batch size of 512. We train for 200 epochs when 50 labeled images per class are used and 100 epochs when 250/1000 labeled images per class are used. When finetuning on HAM10K dataset, we use a batch size of 128 and train for 200 epochs. When finetuning on EuroSAT dataset, we train for 800 epochs with a batch size of 50 when using 5 labeled images per class, 600 epochs with a batch size of 100 when using 10 labeled images per class, and 400 epochs with a batch size of 128 when using 25 labeled images per class. When finetuning on ImageNet dataset, we use a batch size of 512 and train for 50 epochs. When finetuning on ADE20K dataset, we use a batch size of 32 and train for 500/400/300/200 epochs when using 1200/2401/4802/9605 labeled images. We run each finetuning experiment with several learning rates from $7e^{-6}$ to $3e^{-3}$ and report results corresponding to best ones.

B. Image Retrieval Details

In this section, we describe the retrieval process used for curating task-related transfer sets. Given a query set $\mathcal{Q}$ and a gallery set $\mathcal{G}$, we find N images from $\mathcal{G}$ that best match the images in $\mathcal{Q}$ in a query-balanced fashion, as explained in Section 3.4. In our experiments, $\mathcal{Q}$ consists of a few images per class from the domain of interest. For example, in the case of EuroSAT with 5 images per class, $\mathcal{Q}$ consists of a total of $5 \times 10 = 50$ images. We performed both image-level and crop-level retrievals, as explained below.

B.1. Image-level Retrieval from YFCC15M

In this setup, we use the YFCC15M dataset (Radford et al., 2021), which contains 15M images, as the gallery set and the OpenCLIP's (Ilharco et al., 2021) ViT-L/14 image encoder, trained on the DataComp-1B dataset (Gadre et al., 2023), as the encoder network. To obtain image-level features for both gallery and query sets, we resize and center-crop them to 224×224 and run the image encoder. We used this retrieval strategy to obtain retrieval-augmented transfer sets for the ADE20K (Zhou et al., 2017) query sets.

B.2. Crop-level Retrieval from DataComp

In this setup, we use the DataComp-1B dataset (Gadre et al., 2023), which contains 1.28B (image, text) pairs (the best pool filtered of 12B images) as the gallery set. For each image in the gallery set, we consider 10 random crops using PyTorch's `RandomResizeCrop` augmentation with scale parameters (0.08, 1.0) and a final crop size of 224×224. We further apply RandAug (Cubuk et al., 2020) to each crop to increase the diversity of gallery crops. Our cropping and augmentation strategy results in a rich and diverse set of 12.8B gallery crops, allowing us to retrieve target task related crops from seemingly out-of-domain images (see Figure 10 in Appendix B.6).

For each crop in the gallery, we store two normalized features obtained by OpenCLIP's (Ilharco et al., 2021) ViT-L/14 image encoders trained with OpenAI and Datacomp-1B pretraining datasets. For efficient retrieval, we use the cached feature store as in the dataset reinforcement strategy of Vasu et al. (2024). For query images, we resize and center-crop to 224×224 and run the same image encoders. For a pair of a gallery crop and a query image, we compute their similarity (for the k-NN search described in Section 3.4) by averaging the cosine similarity of their features computed by the two image encoders. We used this retrieval strategy to obtain retrieval-augmented transfer sets for the EuroSAT (Helber et al., 2019) query sets.

B.3. De-duplication

Web-scale datasets like DataComp, which we use to generate retrieval-augmented transfer sets, may include duplicated images. We found that the original de-duplication performed in DataComp (Gadre et al., 2023) is not sufficient to remove all duplicates. To enhance the quality and diversity of the retrieved sets, we implement a de-duplication process as follows.

For every pair of retrieved crops, we compute the features of their original images using the same two ViT-L/14 image encoders from OpenCLIP (Ilharco et al., 2021), after resizing and center-cropping to 224×224. If the average similarity between the original images of a pair exceeds 0.99, we consider this pair a duplicate. We then cluster all duplicate pairs and ensure that only one instance from each cluster is included in the final retrieved set.

B.4. De-contamination

In our experiments, we use only a subset of a target task dataset as the query set to perform retrieval. It is possible that some other images from the target task dataset might have "leaked" into our gallery set (i.e., DataComp). To ensure we are not retrieving such images, we follow a de-contamination process as described below.

First, we form a set $\mathcal{S}$ of all images present in a given target task dataset (including train, val and test splits). We then compute two image-level features for all images in $\mathcal{S}$ using the same two ViT-L/14 image encoders from OpenCLIP (Ilharco et al., 2021), after resizing and center-cropping to 224×224. Similarly, we compute two image-level features for the original images of the retrieved crops. For every pair of images (x_s, x_r), where $x_s \in \mathcal{S}$ and x_r is the original image of a retrieved crop, we examine the maximum similarity between their two features. If the maximum similarity exceeds 0.95, we mark x_r as a "possible leak". Next, for every x_r identified as a "possible leak", we find its 5 nearest neighbors in $\mathcal{S}$ and perform a human visual inspection to verify whether it is an actual leak. Finally, we remove all retrieved crops corresponding to a leaked image.

B.5. Retrieval Ablations

Let $\mathcal{Q}$ denote the query set, $\mathcal{G}$ denote the gallery set and ϕ denote the encoder network. Initially, we evaluated the transfer sets created with the following four retrieval strategies under task-agnostic knowledge transfer setting.

- **Random:** Randomly select images from the gallery.
- **Best-matches**: For each image $x \in \mathcal{G}$, we use $\min_{x_q \in \mathcal{Q}} \|\phi(x) - \phi(x_q)\|_2$ as its distance to the query set $\mathcal{Q}$. We

retrieve images from $\mathcal{G}$ in the increasing order of their distance to $\mathcal{Q}$.

- **Query-balanced (Image)**: For a query image $x_q \in \mathcal{Q}$, we define k-NN(x_q) to be the set of its k nearest neighbors from the gallery $\mathcal{G}$. To retrieve N images in total, we find the smallest k for which $\bigcup_{x_q \in \mathcal{Q}} k\text{-NN}(x_q)$ contains at least N images. If $\bigcup_{x_q \in \mathcal{Q}} k\text{-NN}(x_q)$ contains more than N images, we drop the kth neighbor of randomly selected queries until the retrieved set contains N images.
- **Query-balanced (Text):** First, we convert the class names in the target task into text descriptions using the templates from (Radford et al., 2021) and encode these text descriptions using the text encoder of the OpenCLIP model used to encode images. Then, we follow the above query balanced retrieval strategy using text queries instead of image queries.

Figure 8 shows ADE20K segmentation performance for task-agnostic knowledge transfer using transfer sets curated by different retrieval strategies. Here, we use 4800 labeled images for finetuning the target model and use the same 4800 images as the query set for retrieval. Query-balanced retrieval based on image queries performs the best. By giving equal weight to all queries, this approach increases diversity in the retrieved samples when compared to the best-matches strategy. Following these initial results, we adopted the query-balanced retrieval strategy for all our experiments.

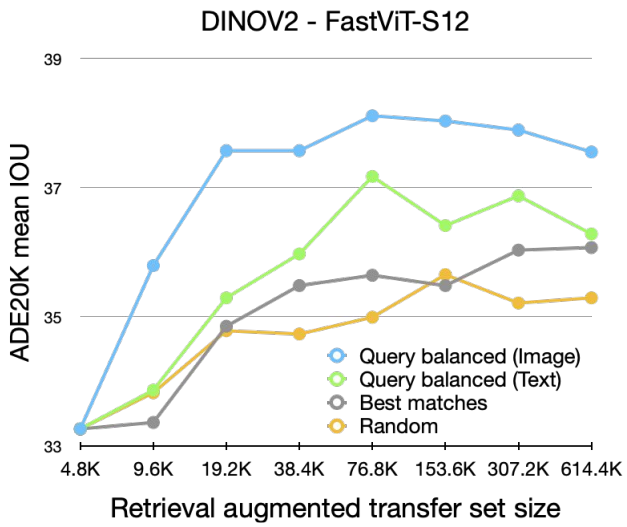


Figure 8. Performance of task-agnostic transfer using transfer sets curated by different retrieval strategies. Here, we use 4800 labeled ADE20K images both as the finetuning dataset and the query set.

B.6. Retrieval Visualizations

In this section, we present some (query, retrieved) image pairs. Figure 9 illustrates image-level retrieval from the YFCC15M dataset for queries from the ADE20K dataset, and Figure 10 shows crop-level retrieval from the DataComp-1B for queries from the EuroSAT dataset. Crop-level retrieval allows for patch extraction from a broader range of images, even those not directly related to the task.

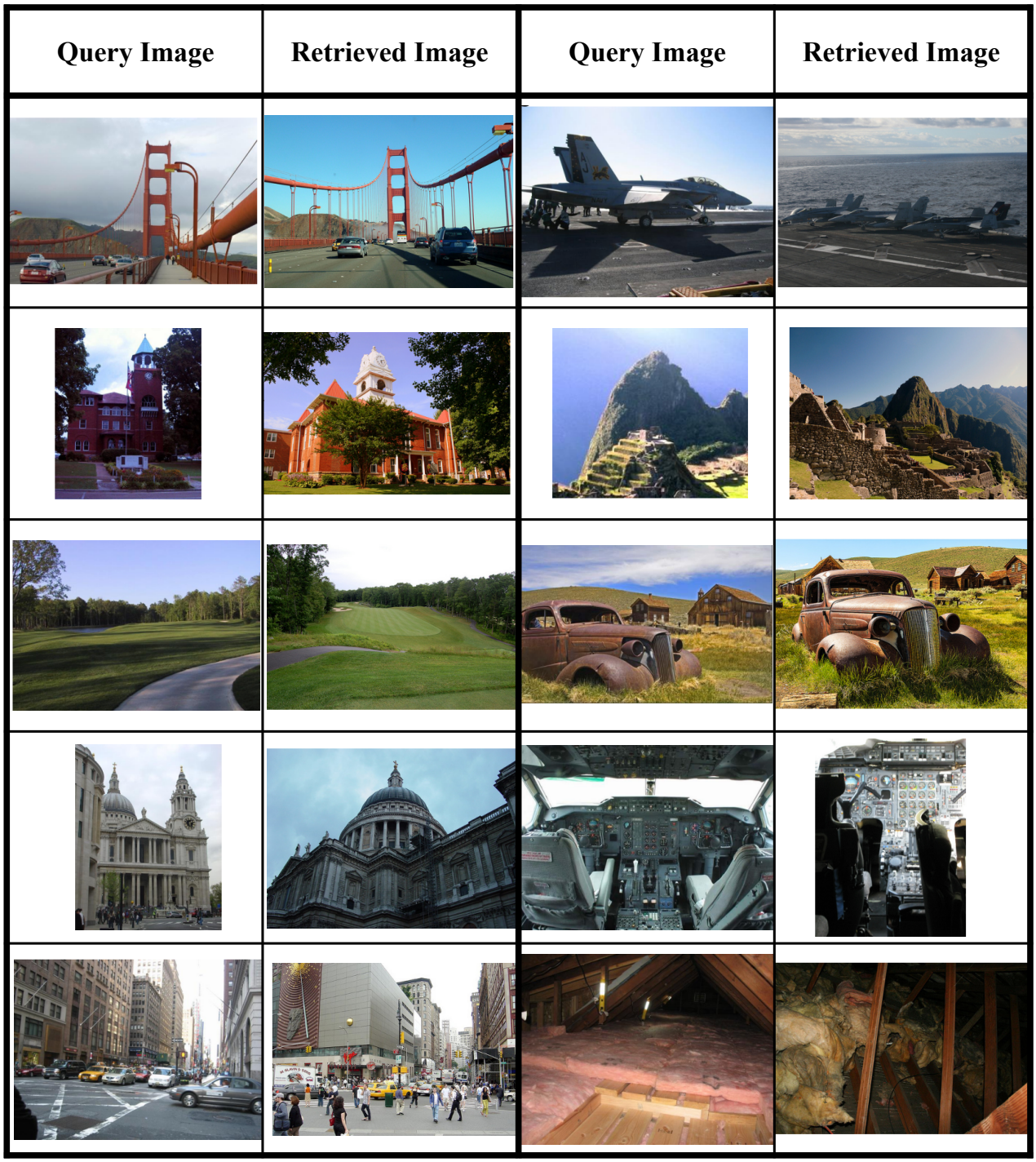


Figure 9. Pairs of ADE20K images (query) and their matched retrieved images from the YFCC15M dataset using the image-level retrieval algorithm presented in Appendix B.1. Since YFCC15M and ADE20K come from similar distributions, image-level retrieval works well.

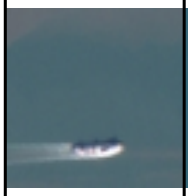
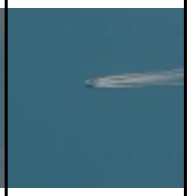
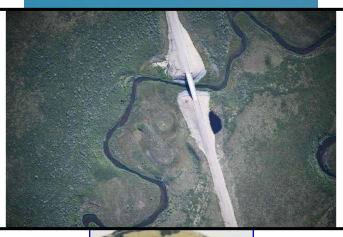
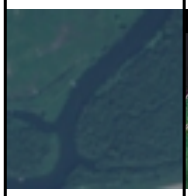
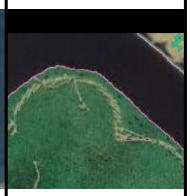
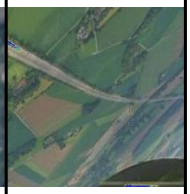
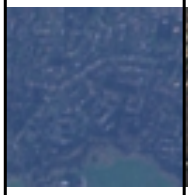
Query Class	Query Image	Matched Crop	Retrieved Original Image	Paired Text
Sea lake				Top View Of Fast Travel Boat Going Straight Forward In The Sea Aerial Shot
River				ACUASI completes remote sensing map of Canada's first highway to the Arctic
River				Original WW2 US Army 88th Infantry Division Bullion Patch With Blue Devil Tab
Highway				Wonderful Aviation Photos (55 pics)
Residential				Window seat view of mid-Channel airliner wing and airspace over west London.
Residential				Karen Metallic Cardigan, view 4, click to view large image.

Figure 10. Pairs of EuroSAT images (query) and their matched retrieved images from the DataComp-1B dataset using a crop-level retrieval algorithm presented in Appendix B.2. For each image, we show the matching crop and the original image, as well as the paired text from the gallery set. With our proposed crop-based retrieval, we can extract only in-domain parts of images (e.g., rows 2, 4, and 5). Furthermore, the combination of cropping and augmentation provides us with in-domain crops from unrelated images (e.g., rows 3 and 6).

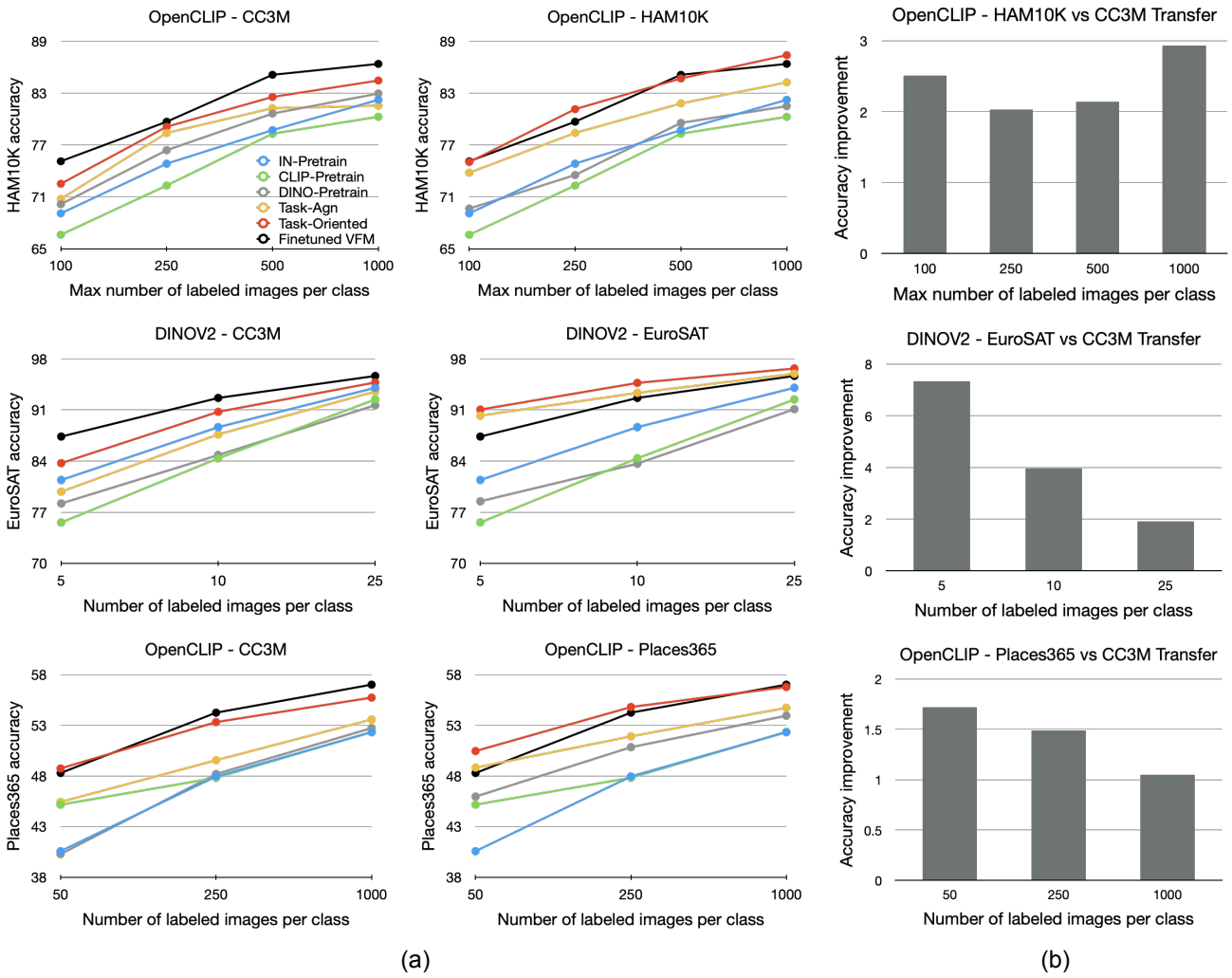


Figure 11. (a) Comparison of various approaches for different VFM - Transfer set combinations with FastViT as the target image encoder. (b) Performance improvement when target task dataset is used instead of generic CC3M transfer set for task-oriented knowledge transfer. The target tasks are HAM10K, EuroSAT and Places365 classification from top to bottom. Task-oriented knowledge transfer from VFMs clearly outperforms various popular alternative training strategies.

C. Additional Results

C.1. FastViT Results

Figure 3 in the main paper compares various approaches for different VFM and transfer set combinations on five downstream tasks with FastViT as the target image encoder. Figure 11 in this Appendix presents additional results corresponding to few more combinations of VFMs and transfer sets. Task-oriented knowledge transfer from VFMs clearly outperforms various popular alternative training strategies.

Tables 5 to 9 present our experimental results on ImageNet, EuroSAT, Places365, ADE20K and HAM10K datasets, respectively, in tabular format. The values in these tables correspond to the plots in Figures 3 and 11.

Percentage of labeled training images		1	10	25	50
Finetuned VFM (OpenCLIP)		74.72*	83.31	84.98	86.05
From scratch		14.46	56.35	68.90	75.17
CLIP-Pretrain		51.94*	65.00*	71.93	76.05
DINO-Pretrain (CC3M)		33.29	63.89	71.36	75.97
DINO-Pretrain (ImageNet)		52.27*	70.64	74.79	77.41
OpenCLIP (CC3M transfer)	Task-Agn	51.47*	68.12	73.24	76.53
	Task-Oriented	63.13	74.29	76.79	78.40
OpenCLIP (ImageNet transfer)	Task-Agn	67.74*	74.81*	76.01	77.97
	Task-Oriented	74.03	80.08	80.91	81.43

Table 5. ImageNet classification accuracy for FastViT target model. The results marked with * are obtained by training only the classification layer instead of the entire model in the finetuning stage. Full finetuning produced inferior results in these cases.

Labeled training images per class		5	10	25
Finetuned VFM (DINOv2)		87.36	92.65	95.66
Finetuned VFM (OpenCLIP)		87.99	92.63	96.08
IN-Pretrain		81.41	88.65	94.04
CLIP-Pretrain		75.59	84.38	92.44
DINO-Pretrain (CC3M)		78.20	84.83	91.66
DINO-Pretrain (EuroSAT)		78.50	83.65	91.11
DINOv2 (CC3M transfer)	Task-Agn (Patch)	76.52	86.11	92.66
	Task-Agn (Image)	79.81	87.66	93.52
	Task-Oriented	83.72	90.75	94.77
DINOv2 (EuroSAT transfer)	Task-Agn (Patch)	90.23	93.34	95.97
	Task-Agn (Image)	86.54	89.56	94.19
	Task-Oriented	91.05	94.71	96.68
OpenCLIP (CC3M transfer)	Task-Agn	79.87	87.10	92.91
	Task-Oriented	85.14	90.21	94.25
OpenCLIP (EuroSAT transfer)	Task-Agn	88.57	92.28	95.5
	Task-Oriented	90.74	94.63	96.83

Table 6. EuroSAT classification accuracy for FastViT target model.

Labeled training images per class		50	250	1000
Finetuned VFM (DINOv2)		47.56 ± 0.02	54.11 ± 0.27	56.95 ± 0.11
Finetuned VFM (OpenCLIP)		48.30 ± 0.59	54.26 ± 0.19	57.03 ± 0.17
IN-Pretrain		40.58 ± 0.14	47.96 ± 0.09	52.33 ± 0.14
CLIP-Pretrain		$45.17 \pm 0.03^*$	47.83 ± 0.30	52.37 ± 0.47
DINO-Pretrain (CC3M)		40.29 ± 0.26	48.20 ± 0.16	52.74 ± 0.12
DINO-Pretrain (Places365)		45.97 ± 0.18	50.85 ± 0.12	53.96 ± 0.06
DINOv2 (CC3M transfer)	Task-Agn (Patch)	$42.46 \pm 0.02^*$	49.60 ± 0.11	53.49 ± 0.06
	Task-Agn (Image)	$44.52 \pm 0.13^*$	49.75 ± 0.07	53.45 ± 0.04
	Task-Oriented	47.81 ± 0.05	53.09 ± 0.05	55.62 ± 0.05
DINOv2 (Places365 transfer)	Task-Agn (Patch)	$46.45 \pm 0.05^*$	51.26 ± 0.20	54.43 ± 0.03
	Task-Agn (Image)	$47.76 \pm 0.04^*$	51.45 ± 0.32	54.45 ± 0.21
	Task-Oriented	49.14 ± 0.02	54.51 ± 0.05	56.68 ± 0.01
OpenCLIP (CC3M transfer)	Task-Agn	$45.44 \pm 0.03^*$	49.56 ± 0.05	53.59 ± 0.09
	Task-Oriented	48.75 ± 0.01	53.33 ± 0.05	55.75 ± 0.06
OpenCLIP (Places365 transfer)	Task-Agn	$48.83 \pm 0.07^*$	51.92 ± 0.26	54.74 ± 0.10
	Task-Oriented	50.47 ± 0.01	54.82 ± 0.04	56.80 ± 0.02

Table 7. Places365 classification accuracy for FastViT target model. The results marked with * are obtained by training only the classification layer instead of the entire model in the finetuning stage. Full finetuning produced inferior results in these cases.

Labeled training images		1200	2401	4802	9605
Finetuned VFM (DINOv2)		39.32 ± 0.11	42.76 ± 0.08	46.35 ± 0.09	49.42 ± 0.14
IN-Pretrain		22.58 ± 0.07	27.11 ± 0.06	33.12 ± 0.25	37.54 ± 0.02
CLIP-Pretrain		15.34 ± 0.20	20.88 ± 0.26	27.57 ± 0.11	32.63 ± 0.08
DINO-Pretrain (CC3M)		20.88 ± 0.27	25.75 ± 0.25	31.74 ± 0.08	36.43 ± 0.02
DINO-Pretrain (ADE20K)		16.43 ± 0.09	21.14 ± 0.20	27.71 ± 0.23	32.23 ± 0.23
DINOv2 (CC3M transfer)	Task-Agn (Patch)	25.75 ± 0.55	30.27 ± 0.14	35.83 ± 0.11	39.28 ± 0.54
	Task-Oriented	36.28 ± 0.01	39.22 ± 0.19	41.44 ± 0.15	44.29 ± 0.03
DINOv2 (ADE20K transfer)	Task-Agn (Patch)	28.32 ± 0.46	32.0 ± 0.19	36.81 ± 0.08	39.44 ± 0.22
	Task-Oriented	34.57 ± 0.02	37.19 ± 0.10	38.45 ± 0.05	41.07 ± 0.21

Table 8. ADE20K mean IOU for FastViT target model.

Maximum labeled training images per class		100	250	500	1000
Finetuned VFM (DINOv2)		77.78 ± 0.61	81.79 ± 0.97	84.57 ± 1.88	88.12 ± 0.56
Finetuned VFM (OpenCLIP)		75.13 ± 0.82	79.70 ± 1.19	85.12 ± 0.35	86.38 ± 0.74
IN-Pretrain		69.11 ± 1.56	74.85 ± 0.82	78.73 ± 1.88	82.23 ± 1.37
CLIP-Pretrain		66.64 ± 1.75	72.33 ± 1.11	78.31 ± 1.99	80.27 ± 2.08
DINO-Pretrain (CC3M)		70.17 ± 1.60	76.41 ± 1.00	80.64 ± 0.76	82.96 ± 0.31
DINO-Pretrain (HAM10K)		69.66 ± 2.52	73.54 ± 1.86	79.56 ± 0.81	81.50 ± 1.46
DINOv2 (CC3M transfer)	Task-Agn (Patch)	71.76 ± 2.28	75.68 ± 1.83	80.34 ± 0.74	80.95 ± 1.16
	Task-Agn (Image)	71.32 ± 1.94	77.67 ± 1.48	80.58 ± 0.53	82.78 ± 1.01
	Task-Oriented	73.06 ± 0.79	78.44 ± 0.41	81.88 ± 0.64	83.16 ± 0.83
DINOv2 (HAM10K transfer)	Task-Agn (Patch)	75.29 ± 1.22	78.77 ± 1.55	83.96 ± 0.10	85.19 ± 0.40
	Task-Agn (Image)	72.18 ± 2.57	76.85 ± 0.74	81.68 ± 0.85	83.64 ± 0.59
	Task-Oriented	78.04 ± 0.05	82.30 ± 0.14	86.33 ± 0.16	86.20 ± 0.32
OpenCLIP (CC3M transfer)	Task-Agn	70.77 ± 1.00	78.40 ± 0.77	81.26 ± 0.37	81.55 ± 1.31
	Task-Oriented	72.53 ± 0.63	79.12 ± 1.27	82.56 ± 0.4	84.46 ± 0.65
OpenCLIP (HAM10K transfer)	Task-Agn	73.81 ± 0.44	78.40 ± 0.27	81.81 ± 0.82	84.24 ± 1.53
	Task-Oriented	75.04 ± 0.06	81.15 ± 0.01	84.70 ± 0.08	87.39 ± 0.08

Table 9. HAM10K classification accuracy for FastViT target model.

C.2. MViT-V2 Results

The experimental results with FastViT as the target image encoder are presented in the main paper and Appendix C.1. Figure 12 in this Appendix presents results for some VFM and transfer set combinations with MViT-V2 as the target image encoder. Task-oriented knowledge transfer from VFMs clearly outperforms various popular alternative training strategies. For CLIP pretraining of MViT-V2 model, we use the original CLIP loss from (Radford et al., 2021) instead of the affinity mimicking distillation loss of (Wu et al., 2023).

Tables 10 and 11 present our experimental results on Places365 and ADE20K datasets, respectively, in tabular format. The values in these tables correspond to the plots in Figure 12.

Labeled training images		1200	2401	4802	9605
Finetuned VFM (DINOv2)		39.32 ± 0.11	42.76 ± 0.08	46.35 ± 0.09	49.42 ± 0.14
IN-Pretrain		22.36 ± 0.18	26.60 ± 0.26	30.95 ± 0.41	35.02 ± 0.01
CLIP-Pretrain		19.58 ± 0.22	24.66 ± 0.09	29.06 ± 0.24	33.62 ± 0.14
DINOv2 (CC3M transfer)	Task-Agn (Patch)	25.42 ± 0.18	29.52 ± 0.10	32.76 ± 0.28	36.73 ± 0.17
	Task-Oriented	33.35 ± 0.02	35.68 ± 0.08	37.32 ± 0.11	40.14 ± 0.05
DINOv2 (ADE20K transfer)	Task-Agn (Patch)	30.03 ± 0.15	33.12 ± 0.27	35.41 ± 0.15	37.96 ± 0.39
	Task-Oriented	33.19 ± 0.11	35.01 ± 0.09	37.33 ± 0.03	39.07 ± 0.09

Table 10. ADE20K mean IOU for MViT-V2 target model.

Labeled images per class		50	250	1000
Finetuned VFM (DINOv2)		47.56 ± 0.02	54.11 ± 0.27	56.95 ± 0.11
Finetuned VFM (OpenCLIP)		48.30 ± 0.59	54.26 ± 0.19	57.03 ± 0.17
IN-Pretrain		40.46 ± 0.07	47.93 ± 0.15	52.73 ± 0.04
CLIP-Pretrain		$42.42 \pm 0.05^*$	48.06 ± 0.16	52.72 ± 0.08
DINOv2 (CC3M transfer)	Task-Agn (Patch)	41.81 ± 0.16	48.90 ± 0.12	53.15 ± 0.08
	Task-Agn (Image)	42.55 ± 0.19	48.89 ± 0.09	53.05 ± 0.09
	Task-Oriented	47.44 ± 0.02	52.44 ± 0.04	54.93 ± 0.04
DINOv2 (Places365 transfer)	Task-Agn (Patch)	45.14 ± 0.08	50.37 ± 0.06	53.72 ± 0.08
	Task-Agn (Image)	45.73 ± 0.02	50.61 ± 0.05	53.43 ± 0.03
	Task-Oriented	49.20 ± 0.04	54.02 ± 0.05	56.10 ± 0.03
OpenCLIP (CC3M transfer)	Task-Agn	$43.23 \pm 0.09^*$	49.07 ± 0.04	53.04 ± 0.14
	Task-Oriented	47.82 ± 0.02	52.82 ± 0.05	54.92 ± 0.03
OpenCLIP (Places365 transfer)	Task-Agn	$47.61 \pm 0.05^*$	51.39 ± 0.10	54.17 ± 0.11
	Task-Oriented	50.14 ± 0.04	54.30 ± 0.02	56.43 ± 0.04

Table 11. Places365 classification accuracy for MViT-V2 target model. The results marked with * are obtained by training only the classification layer instead of the entire model in the finetuning stage. Full finetuning produced inferior results in these cases.

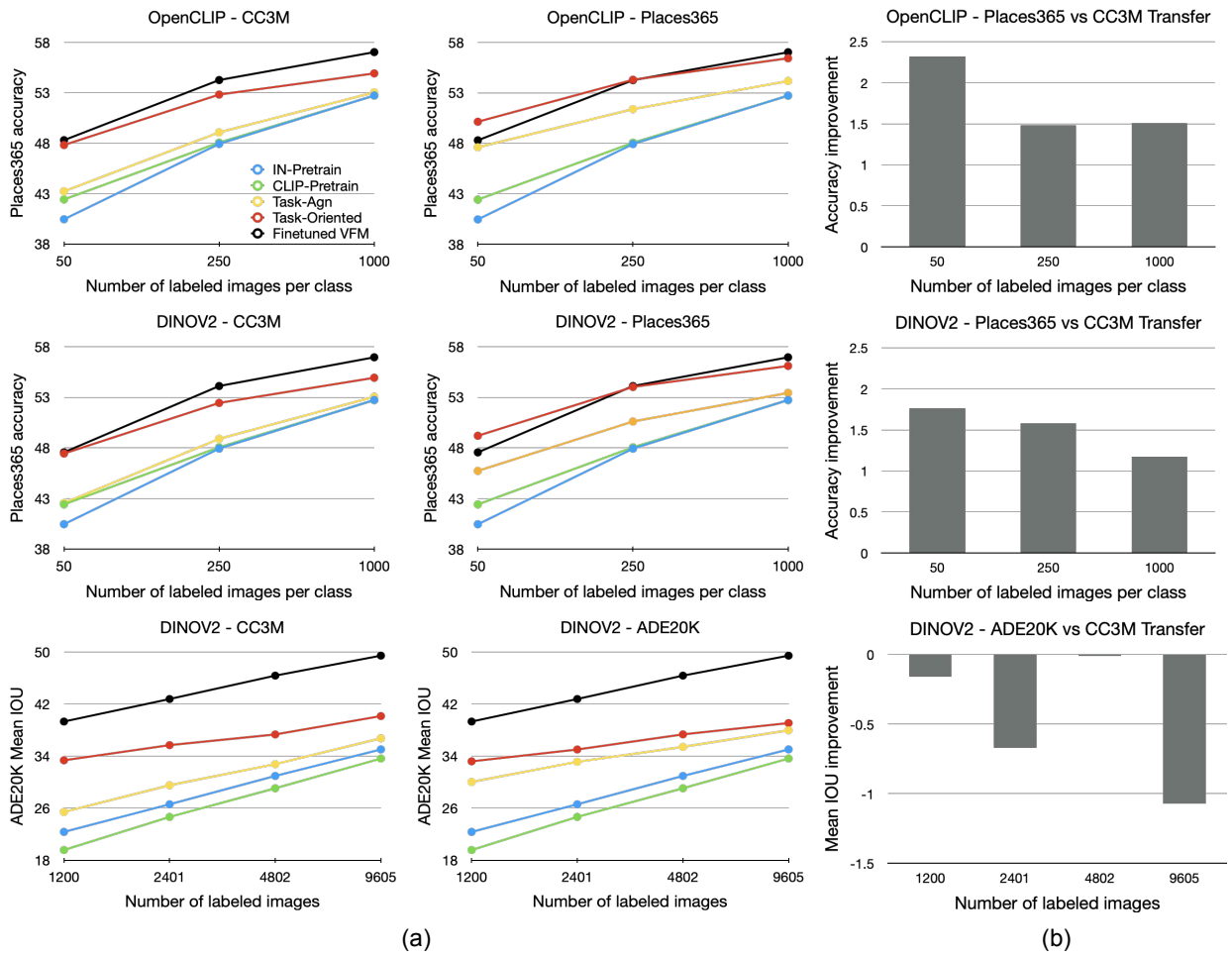


Figure 12. (a) Comparison of various approaches for different VFM - Transfer set combinations with MViT-V2 as the target image encoder. (b) Performance improvement when target task dataset is used instead of generic CC3M transfer set for task-oriented knowledge transfer. The target tasks are Places365 classification and ADE20K segmentation from top to bottom. Task-oriented knowledge transfer from VFMs clearly outperforms various popular alternative training strategies.

Dataset	Labeled images per class	Semi-supervised learning			Proposed
		Round 1	Round 2	Round 3	
EuroSAT	5	81.4	87.2	87.8	91.0
HAM10K	500	81.6	82.4	-	86.3

Table 12. Comparison with pseudo labeling-based semi-supervised learning using FastViT target model.

Dataset	VFM	From scratch	IN-Pretrain	CLIP-Pretrain	DINO-Pretrain	Task-Agn	Task-Oriented
ImageNet	OpenCLIP	79.45	-	79.35	80.04	79.70	81.50
Places365	DINOv2	-	56.74	56.59	57.15	57.51	57.54

Table 13. Performance of FastViT target model trained with various approaches under 100% labeled training data setting.

C.3. Comparison with Pseudo-labeling based Semi-supervised Learning

Table 12 shows the performance of a pseudo labeling-based semi-supervised approach with multiple rounds of training. In the first round, we obtain an initial model by training it with labeled data. In the subsequent rounds, we use the current model to generate pseudo labels for the unlabeled data and train a new model using both labeled and pseudo labeled data. In each round, the model is initialized with ImageNet pretrained checkpoint in the case of EuroSAT dataset and DINO(CC3M)-pretrained checkpoint in the case of HAM10K dataset. We use ImageNet pretraining for EuroSAT and DINO pretraining for HAM10K since these are the best performing baseline pretraining strategies for these datasets. In the case of EuroSAT classification, the performance increases when pseudo labeled data is used for the first time (round 2). However, the improvement is not significant in the next round. The final performance (87.8%) after three rounds of training is still significantly lower than 91% achieved by the proposed approach. In the case of HAM10K dataset, using pseudo labeled data results in negligible performance gain, and the proposed approach performs significantly better.

C.4. Performance with Large Amount of Labeled Training Data

The main focus of this work is on improving the target task performance with limited amount of labeled data. Hence, most of our experiments focus on limited labeled data settings. In this section, we compare the performance of different approaches when large amount of target task labeled data is available.

Table 13 shows the performance achieved by various approaches on ImageNet and Places365 classification tasks when the full target task training dataset is used for both knowledge transfer and supervised finetuning (1.28M images in the case of ImageNet and 1.8M images in the case of Places365). We use FastViT as the target model for these experiments. The proposed task-oriented knowledge transfer approach outperforms all the other pretraining strategies by a significant margin in the case of ImageNet classification, and outperforms ImageNet/CLIP pretraining approaches by a significant margin in the case of Places365 classification.