AM-RADIO: Agglomerative Vision Foundation Model Reduce All Domains Into One

Abstract

A handful of visual foundation models (VFMs) have recently emerged as the backbones for numerous downstream tasks. VFMs like are trained with distinct objectives, exhibiting unique characteristics for various downstream tasks. We find that despite their conceptual differences, these models can be effectively merged into a unified model through multi-teacher distillation. We name this approach AM-RADIO (Agglomerative Model – Reduce All Domains Into One). This integrative approach not only surpasses the performance of individual teacher models but also amalgamates their distinctive features, such as zero-shot vision-language comprehension, detailed pixel- level understanding, and open vocabulary segmentation capabilities. Additionally, in pursuit of the most hardware-efficient backbone, we evaluated numerous architectures in our multi-teacher distillation pipeline using the same training recipe. This led to the development of a novel architecture (E-RADIO) that exceeds the performance of its predecessors and is at least 6x faster than the teacher models at matched resolution. Our comprehensive benchmarking process covers downstream tasks including ImageNet classification, semantic segmentation linear probing, COCO object detection and integration into LLaVa-1.5.

1 Introduction

Knowledge Distillation has been a very successful and popular technique for transferring the knowledge of a "teacher" model (or ensemble of models) into a typically smaller "student" model. In the original formulation, both the student and the teacher operate on the same in-domain dataset, and the student simultaneously matches the logits of the teacher, and the ground truth labels. Instead of using labeled images, an alternative approach is to train the student model to match the features of the teacher model.

Instead of using a smaller student model, employ an iterative learning procedure with a high-capacity model where a student of equal or greater capacity than the teacher is trained with heavy augmentation applied to the student. Once trained, they expand the dataset by pseudo-labeling new data using the trained student. They then make the student become the teacher, and repeat the process. An important finding in this work is that the student is capable of surpassing the performance of the teacher.

The authors of explore the concept of ensemble distillation, where there are multiple teachers, each of which having restricted domain knowledge. provides an overview of multi-teacher distillation, and proposes that instead of matching the summary of an ensemble of teachers, the student can match the features of each individual teacher via some learned non-shared mapping from the representation space of the student to each teacher. Of interest in their approach is that the student and teacher don't need to share the same architecture, and also that treating teachers individually yields improved performance.

Recently, the concept of Foundation Models (FMs) has emerged, with the general understanding that these models are large, general, and expensive to train. Through training on very large datasets they are broadly applicable to numerous downstream tasks. A seminal example of such models is

.

, which trains on web-scale weakly supervised (image, caption) pairs, and results in exceptional zero-shot performances on a wide array of computer vision benchmarks. While is firmly a FM, another model, has emerged with broad capabilities, often surpassing on dense tasks that require strong spatial features, such as ADE20k and Pascal VOC. Separately, is gaining popularity for its excellent open-vocabulary instance segmentation abilities, whose vision encoder we hypothesize has strong dense feature representations.

We introduce AM-RADIO with the goal of learning from multiple foundational models simultaneously. We observe that, when given a student model of sufficient capacity, it is often able to exceed any of its teachers on important axes. In addition to performing well on representative foundational benchmarks, by virtue of the training framework, our student models are able to mimic their teacher models, and thus are able to perform downstream tasks that are otherwise performed by the teachers. Examples of this include CLIP-ZeroShot applications, since the language model trained by is compatible with our student, and also Segment-Anything tasks, as the student is able to replace the vision encoder and interface with the already-trained mask decoders.

We also study the effect of using a more hardware-efficient model architecture. Most works on efficiency are not directly comparable as they use different training recipes, even when evaluated on the same dataset such as ImageNet-1k, and may be over-tuned. To this end, we evaluate more than 10 promising architectures under the same training recipe for a direct comparison. We reveal that CNN-like architectures are faster but struggle to distill ViT VFMs. This led us to the development of a novel hybrid architecture, E-RADIO, that exceeds the performance of its predecessors and is at least 6x faster than teacher models at matched resolution.

Our main contributions are as follows:

- We describe a general methodology for distilling multiple distinct foundation models into one, including models with incompatible input resolutions.
- We show that these student models are able to outperform their teachers on representative benchmarks.
- We demonstrate that these student models can either drop-in replace their teachers, or their features can be used directly in downstream applications such as providing visual encoding for LLaVA.
- We benchmark a number of efficient architectures and propose a new architecture (E-RADIO) that allows for similar model quality at significant speedups.

2 Related Work

Knowledge Distillation The underpinning of our work is based on the method of Knowledge Distillation which aims to train a "student" model using soft targets produced by an already-trained "teacher" model, using the the teacher's output logits as "soft" labels. Alternatively, distillation can be performed using intermediate network activations. In general, due to the heterogeneous nature of the different teacher foundation models that we employ, we ignore any potential labels coming from the data, and we ignore the logits of teachers, and simply opt to match the feature representations of the teachers before any task-specific processing stages.

Multi-Teacher Distillation There is also a body of work that studies distilling a student model jointly from multiple teacher models simultaneously. Because of the heterogeneous domains that our teacher models cover, we don't apply approaches that marginalize teachers into a unified label, and instead map students to each teacher independently using teacher-specific projection heads from the unified student representation. Although the reason behind this method in is different, we find the same overall strategy to be effective. While doesn't study matching the features of multiple teachers simultaneously, we are able to extend their paradigm via the different projection heads. To preserve drop-in compatibility with teacher frameworks, we eliminate the feature normalization in the loss function.

Distilling Foundation Models Foundation Models are meant to be generalist models that are trained on massive amounts of data, and are typically resource intensive to train from scratch. In the vein of single-teacher distillation, employ self-distillation to train their smaller variants from the larger teacher. distills their model from a teacher. Instead of focusing our energy on one teacher in particular,

we instead grab high-quality versions of (using OpenCLIP), , and . Concurrently with our work, describe a methodology for merging a model into a pretrained model via distillation, which is, in spirit, quite similar to our approach. In contrast to theirs, we include and also simplify the objective to straightforward feature matching. Since we don't rely on the student model to be pre-trained, it also gives us the flexibility to have the student be an architecture distinct from any teacher.

3 Knowledge Agglomeration

We propose a framework to train a vision foundation model from scratch via multi-teacher distillation. We demonstrate that each teacher brings unique properties to the foundational vision model, and the resulting trained model will agglomerate these attributes.

3.1 Overview

As an initial assumption, we expect that the teacher models are capable of representing a broad swath of images found on the internet, coming from datasets such as ImageNet (1k or 21k), LAION-400M or DataComp-1B. With this in mind, we choose to study 3 seminal teacher model families: , , and as they have demonstrated outstanding performance over a broad range of tasks (as in), or specifically strong performance on downstream dense tasks, such as semantic segmentation under linear probe (as in), or open-vocabulary segmentation (as in). Because these teacher models come from such diverse domains, we omit any form of supplemental ground truth guidance and treat the aforementioned datasets simply as sources of images. To assess the quality of our models, we adopt a set of representative metrics across a few broad domains.

- Image level reasoning: (i) k-NN Top-1 accuracy on ImageNet-1K, and (ii) Zero-Shot accuracy using the teacher's language model. k-NN embeds the model's summary feature vector for every image in the training set, and then for each validation image, it uses a weighted sum of the k nearest training vectors to elect a label.
- Pixel-level visual tasks: segmentation mIOU on (i) ADE20K and (ii) Pascal VOC - under the linear probe setting, details in Section 5.3.
- Large Vision-Language Models: we plug our frozen vision encoder model into LLaVA-1.5 and evaluate it on a wide set of tasks including GQA, TextVQA, ScienceQA and VQAv2. Details in Section 5.4.
- SAM-COCO instance segmentation: From , we adopt their COCO instance segmentation methodology to evaluate our ability to replicate SAM visual features.

Results on these tasks, both for teacher models and our AM-RADIO variants, are summarized in Table 1.

3.2 Adaptor Heads

We opt for simplicity in design of the adaptor heads, and leave alternative architectures as future work. To this end, we employ a simple 2-layer MLP, with a LayerNorm and GELU in between. The input dimension is the student embedding dimension, the intermediate dimension is the maximum embedding dimension of all teachers, and the output dimension matches the specific teacher. For each teacher, we employ two heads, one for the summary vector, and one for the spatial features.

3.3 Distillation Dataset Choice

In table 2 we study the effect of different datasets on downstream metrics. While the highest image classification metrics are achieved using ImageNet-1K as the training dataset, we argue that it doesn't fairly measure "zero shot" performance as the student directly learns the teacher features in the evaluation domain. For this reason, we opt for the DataComp-1B dataset.

3.4 Loss Formulation

Because we don't have ground truth data for each teacher for each image, we instead opt to match the features coming from each teacher's vision encoder. In particular, we distinguish between the

Table 1: Comparison of vision foundation and RADIO models. "Zero-Shot" and k-NN are computed on ImageNet-1K. ADE20K and VOC (PascalVOC2012) refer to linear probe semantic segmentation mIOU. GQA, POPE (popular), TextVQA, and VQAv2 are obtained via LLaVa 1.5 by replacing the vision encoder. COCO is the instance segmentation metric introduced by to evaluate distillation. RADIO attains the best metrics on most benchmarks, and is competitive with the rest, while E-RADIO enables high quality results in resource constrained settings. Note that Zero-Shot and COCO use teacher's decoder head that is not finetuned. Throughput computed using NVIDIA A100 GPU, stated resolution, and TensorRT v8601. *Denotes teachers used to train our final RADIO. :We failed to export DINOv2-g-reg to TensorRT, so we report DINOv2-g here, which should be fairly close. ::We were unable to get zero shot working using their model code.

Model TextVQA	Params (M) VQAv2	Resolution SAM COCO	Throughput	Zero-shot	k-NN	ADE20k	VOC	GQA	
OpenCLIP-H/14	632	224	503	77.19	81.10	40.04	68.03	57.94	
50.48	72.24	-							
MetaCLIP-H/14	632	224	486	80.51	82.12	35.39	62.62	60.57	
53.65	75.71	-							
SigLIP-L/14	428	384	241	82.61	85.16	40.53	70.31	57.70	
56.65	71.94	-							
Intern-ViT-6B	5,902	224	63	83.20	78.43	47.20	76.85	60.18	
52.45	76.75	-							
	5,537	448	14	-	68.64	42.78	74.43	61.19	
60.36	78.83	-							
DFN CLIP-H/14	633	378	170	83.90	85.27	39.00	70.29	61.73	
56.78	78.78	-							
OpenAI CLIP-L/14	305	336	414	75.54	79.80	36.51	67.04	62.20	
57.92	78.49	-							
DINOv2-g/14-reg	1,137	224	294	-	83.41	48.68	82.78	61.88	
47.18	76.23	-							
SAM-H/16	637	1024	12	-	22.12	28.08	34.34	49.92	
43.91	57.65	77.18							
E-RADIO-L (Ours)	391	512	468	80.73	83.89	48.22	81.64	61.70	
51.47	76.73	76.31							
RADIO-ViT-H/16 (Ours)	653	432	158	82.93	86.06	51.34	84.71	63.01	
56.32	79.28	76.23							

Table 2: Ablation study on the choice of training dataset. We use MetaCLIP ViT-H/14 and DINOv2 ViT-g/14 teachers, and a ViT-L/14 student model with CPE. Both "k-NN" and "Zero Shot" are for ImageNet-1k. ADE20k refers to mIOU linear probe on ADE20k.

Dataset	k-NN	Zero Shot	ADE20K
ImageNet 1K	84.79	80.44	48.11
ImageNet 21K	84.61	80.10	48.65
LAION-400M	83.77	77.46	48.6
DataComp-1B	83.91	78.51	49.01

summary feature vector and the spatial feature vectors for each teacher. The summary feature is computed differently based on the model. For and , we use the "class token" as the summary feature vector, and we don't match a summary for .

Let $f(x|\Theta_0)$ be the student vision encoder with parameters Θ_0, and $y_i = h_i(x_1|\Theta_i)$ be the learned student head matching teacher summary features $z_i = t_i(x|\Phi_i)$ with student adaptor parameters Θ_i and teacher parameters Φ_i.

$$x_1 = f(x|\Theta_0); z_i = t_i(x|\Phi_i), y_i = h_i(x_1|\Theta_i); L_{summary}(x) = \sum_i \lambda_i L_{cos}(y_i, z_i) \tag{1}$$

We found empirically that cosine distance loss produced better models compared to L1, MSE, Smooth-L1. Additionally, supervising the spatial features of the model by matching the teacher was not only important for downstream dense tasks, but also improved the holistic quality of our model.

For matching the spatial features, we employ a combination of cosine similarity and smooth L1. Similar to equation (2) where we found that cosine similarity produced the best results, we found the same to be true for the spatial features. However, we want to allow our student model to be a drop-in replacement in the teacher frameworks, thus it's important that we match the magnitude of the teacher vectors, and so we include smooth L1. In (3) we show the formulation of this loss. Let $h_i(x_1|\Theta_i)$ be the learned student head for matching teacher feature vectors, and corresponding $t_i(x|\Phi_i)$ be the teacher feature vectors, with $x_1 = f(x|\Theta_0)$, then the spatial feature loss is:

$$L_{match}(x, y) = \alpha L_{cos}(x, y) + \beta L_{smooth-l1}(x, y) \tag{2}$$

$$L_{features}(x) = \sum_i \gamma_i L_{match}(h_i(x_1|\Theta_i), t_i(x|\Phi_i)) \tag{3}$$

We choose $\alpha = 0.9$ and $\beta = 0.1$ to mostly rely on the empirically better cosine distance, but to also match vector magnitudes.

3.4.1 Loss Balancing

Due to the number of possible combinations of loss weights between the different teachers, and even which teachers, and possible formulations of loss functions, we mostly opted toward naive loss balancing with all teachers equally weighted for spatial features ($\gamma_i = 1$). For summary features, we have $\lambda_{CLIP} = \lambda_{DINO} = 1$ and $\lambda_{SAM} = 0$.

We did experiment with automatic loss balancing using predicted uncertainty, AdaLoss (momentum 0.99) and separately with AMTML-KD, as ways to learn the balance of λ_i and γ_i. In the case of AMTML-KD, the model would always collapse its entire weight around the teacher and would yield worse results than naive manual balancing. Based on the results in table 4, there is very little advantage to the more exotic balancing schemes, so we opt for the "Naive" method throughout the rest of the paper.

Table 3: Ablation over which teachers we supervise the spatial features. We use a ViT-L/14 student model and train on the LAION-400M dataset. Adding this loss term is always beneficial. DINOv2 appears to provide better spatial features than CLIP, but training the student to match both teachers produces the best results. We don't ablate SAM as we solely want it for its spatial features.

Teachers	Zero Shot	k-NN	ADE20K
None	75.77	82.59	41.18
CLIP	75.64	82.60	44.42
DINOv2	74.68	83.02	47.05
Both	74.85	82.96	48.13

Table 4: Loss term balancing methods comparison. We use a ViT-B/14 student, and CLIP+DINOv2 teachers. We found that AdaLoss produces the best results on the ImageNet tasks, but the worst on ADE20K.

Method	Zero Shot	k-NN	ADE20K
Naive	70.63	79.50	44.71
Uncertainty	70.92	79.37	44.57
AdaLoss	71.31	79.77	44.36

4 Implementation Details

Performing heterogeneous multi-teacher distillation is not trivial due to a mismatch in feature dimensions, input resolutions, concepts for loss computation, and downsampling ratios, as well as challenges in fitting multiple teachers into a single GPU.

General. We train all student models using the AdamW optimizer, batch size 1024, cosine annealing learning rate schedule and base learning rate of 0.001. We train for 600k steps, resulting in 614M total examples seen. For our best student model, we train using DFN CLIP ViT-H/14 378px, OpenAI CLIP ViT-L/14 336px, DINOv2 ViT-g/14 224px, and SAM ViTDet-H 1024px. We apply random scale + cropping to both student and teacher inputs. We chose the DataComp-1B dataset due to it having the highest quality results of the web-scale datasets we had access to. We train in two stages, first with CLIP+DINOv2 for 300k steps at 256px, and second with CLIP+DINOv2 at 432px plus SAM at 1024px for 300k steps.

Student architecture. We study two settings for student model architecture:

- Standard ViT architecture to match the architecture of teachers. Our best model is a ViT-H/16.
- Efficient architecture variants prioritizing high throughput on GPUs. See Section 5.1.

Multi-scale Teachers. We choose ViT-H/16 architecture for our student model. To match resolution of features, we feed the expected resolution of 10242. Given that our and teachers are patch-14 models, we opt to feed the student 4322 inputs, as that is the same effective resolution as 3782 for patch-14. We found that interpolating features doesn't degrade results, so the teacher operates at 224px and we upsample the outputs to match the student.

Rank/Teacher Partitioning. We group teacher models by (batch size, student resolution), and then distribute the groups to different GPUs, such that each GPU processes a consistent batch size and input resolution. We also sample groups at different rates. For our training setups that include , we train with 64 GPUs, half of which get the CLIP+DINOv2 group with batch size 32 per GPU and input resolution 432, and the other half get with batch size 2 per GPU and input resolution 1024. This results in an effective batch size of 1,152. For CLIP+DINOv2 training, we use 32 GPUs, resulting in batch size 1024.

Multi-Resolution ViTs. Many of our student models use ViT as the base vision architecture. Traditionally, ViTs use a learned position embedding for each input patch in an image, which in turn enforces that the model always operates at a constant resolution. We employ the Cropped Position Embedding (CPE) augmentation with the number of positions being equal to 1282. The position embeddings are then randomly cropped and interpolated to match the number of input patches for the student model. Even when training with CLIP+DINOv2 at 224 resolution, we found that this technique results in a negligible drop (Table 5) in summary metrics, but improved semantic segmentation linear probing mIOU. For heterogeneous-resolution students, this is a seamless technique that allows ViT to operate at arbitrary resolutions within some envelope. In addition to enabling arbitrary resolutions, as shown in figure 3, CPE reduces the noise artifacts in the position embeddings as compared to other ViT models.

High-Resolution ViT Student. In , they employ the ViTDet architecture as a way to reduce the computational and memory burden of ViT models at high-resolution. We reformulate this arch instead into a training augmentation, where we sample a window size from a set of possible window sizes. This allows us to reduce the computational burden of training the student model with the teacher, and, as we make the window size flexible, it provides an additional throughput scaling mechanism during inference. Table 8 demonstrates our ability to replace SAM's encoder. Separately, we found that high resolution training was unstable, so we apply spectral reparametrization and a weight decay of 0.02 to prevent attention entropy collapse.

Student/Teacher Resolution Mismatch. When the student and teacher downsample images through their processing stack at different rates, it results in the output feature vectors having different resolutions. For example, if the teachers use a ViT-H/14 architecture and student a ViT-H/16, it means that the student outputs a 142 feature map, and the teachers a 162 feature map. For $L_{features}$ we bilinearly interpolate the outputs to match the larger resolution between the student and teacher features.

Feature Summarization. In 3.4 we explained how teacher summary features are extracted using the "class token" of their respective ViT models. We now turn our attention to the summarization of student features. ViTs have 2 options: (i) a separate summarization "CLS" token or (ii) average pooling patch tokens. We evaluate both options in Table 6. We observe that average pooling improves

summary loss, but has a more significant detrimental effect on the feature loss. Given the importance of the latter we choose to use separate CLS tokens.

Table 5: Comparing identical ViT models, with CLS token and average pooling summarization.

	Zero Shot	k-NN	ADE20K	VOC	VQAv2
CLS token	78.55	83.91	49.01	83.51	77.66
Avgpool	80.12	83.83	38.36	77.04	78.28

5 Results

In this section, we analyze models obtained with the proposed AM-RADIO framework. First, we touch upon backbone efficiency, then compare with the original teachers (CLIP, DINOv2, SAM), and benchmark models under vision question answering in the LLaVa framework. We will see that the proposed models outperform the original teachers in multiple metrics, including throughput. Results are shown in Figure 1 and Table 1.

5.1 Efficient Students

We aim to find an efficient model architecture to speed up the inference of VFM. There are a number of architectural designs aimed at high throughput on GPU devices. We use our distillation framework to evaluate several backbones with no change in training hyperparameters.

Upon reviewing the literature on efficient vision backbones focused for high GPU throughput, we pick the following list of architectures: EfficientNetV2, ResNetv2, RegNetY, FasterViT, EfficientViT, ConvNext, NFNet, SwinV2, MaxViT, PoolformerV2 and MViTV2. We train all the backbones via distillation on the ImageNet-21k dataset, using OpenCLIP ViT-H/14 (laion2B-s32B-b79K) and DINOv2 g/14 as teachers. Results are compiled in Table 7.

Table 6: Comparison of backbones. Throughput is measured using TensorRT 9.0.1 on A100 in mixed FP16/FP32 precision at batch size 128 on 2242px resolution. Sorted by descending throughput order. FD loss is the Feature Distillation training loss against the DINOv2 teacher, it exhibits high correlation with the ADE20k mIoU. Bolded models form the speed/quality Pareto front.

Backbone	Param. Count	Throughput	Zero Shot	k-NN	ADE20k	FD loss
Teachers						
DINOv2 G/14	1.14B	313	N/A	83.41	47.53	
OpenCLIP H/14	632M	556	77.19	81.10	40.04	
Existing Efficient Models						
EfficientNetV2-S	21M	9017	65.37	70.72	27.75	0.415
ResNetv2-101	44M	7283	69.58	75.32	29.61	0.405
RegNetY-064	30M	6573	69.84	74.59	28.9	0.394
EfficientViT-L1	38M	6048	71.73	79.90	33.12	0.376
ConvNext-B	88M	1805	75.43	81.73	38.95	0.358
NFNet-F3	254M	1777	76.93	80.50	38.31	0.340
SwinV2-S	49M	1497	74.70	81.12	35.57	0.364
MaxViT-B	119M	1486	77.49	79.34	38.46	0.340
PoolformerV2-M36	56M	1194	74.46	80.49	35.05	0.377
MViTV2-B	51M	975	75.92	81.39	41.39	0.345
Proposed architecture						
E-RADIO-B	118M	6422	75.19	82.21	44.03	0.319
E-RADIO-B w/o upsample	113M	7040	75.45	82.05	41.26	0.353
E-RADIO-L	265M	3472	77.87	83.73	45.5	0.265

We observe that many models lag behind teachers. Additionally, CNN-like models are significantly faster than ViTs, while the latter are more accurate. The relatively low performance of existing

efficient backbones on the dense ADE20k segmentation task is not unexpected since all of them apply a spatial dimension reduction factor of 32 for final feature maps of size 72 for input resolution of 2242px, thus hardly capable of capturing fine-grain spatial information.

E-RADIO: To overcome this issue, we propose a novel hybrid architecture, named E-RADIO (Efficient RADIO). This design borrows ideas from existing literature and includes an input stem with strided convolutions to downsample the input image by 4x. It then proceeds with 2 stages of YOLOv8 C2f convolution blocks and 2 stages of transformer. For the transformer variant we pick windowed attention (like in SWIN), and interleave local windowed attention with "global" windowed attention as done in and ViTDet. To perform "global" attention we first downsample the feature map by 2x, apply windowed attention, and then upsample the feature maps back to the original resolution.