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

Mike Ranzinger*, Greg Heinrich*, Jan Kautz, Pavlo Molchanov
NVIDIA
{mranzinger,gheinrich,jkautz,pmolchanov}@nvidia.com

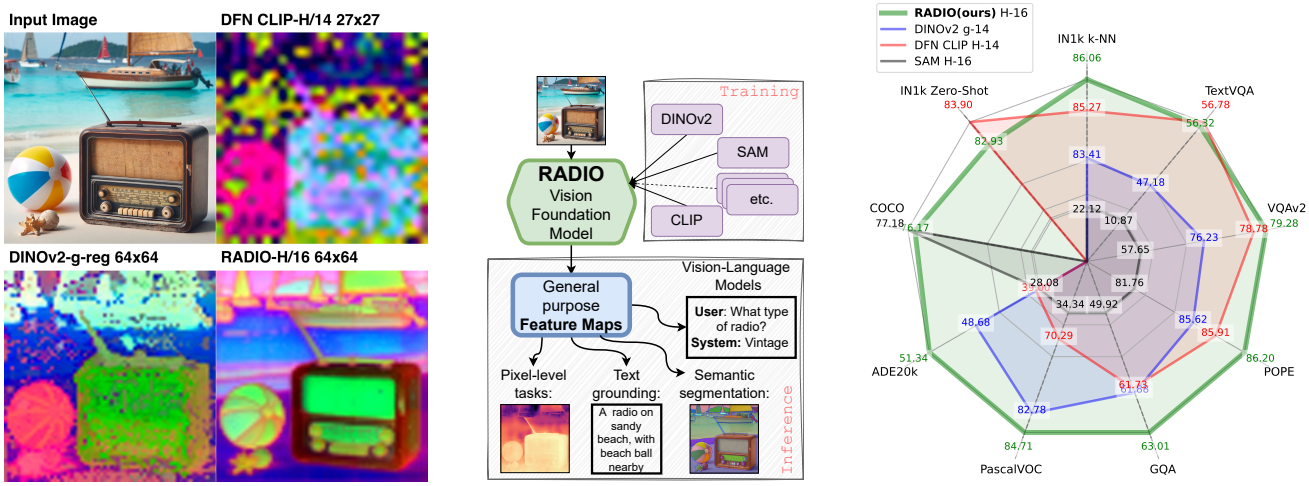


Figure 1. AM-RADIO is a framework to distill multiple pretrained vision foundation models, such as CLIP [51], DINOv2[48], SAM [35], into a single model that we call RADIO. As a result, a single vision foundation model agglomerates unique properties of the original models. This unifying approach obtains state-of-the-art feature representations in a single forward pass while also enabling unique properties such as zero-shot classification (CLIP) or open set instance segmentation (SAM) at negligible additional cost.
Image description: (left) PCA feature visualization of different models. Our proposed RADIO model can process any resolution and aspect ratio, and produces semantically rich dense encodings; (middle) the overview of the AM-RADIO framework; (right) benchmarks on classification, segmentation, and vision-language modeling tasks, see section 5.

Abstract

A handful of visual foundation models (VFMs) have recently emerged as the backbones for numerous downstream tasks. VFMs like CLIP, DINOv2, SAM 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.

Code: https://github.com/NVlabs/RADIO.

1. Introduction

Knowledge Distillation [26] 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

*Equal contribution

Model	Params (M)	Resolution	Throughput	ImageNet1K		Segmentation (linear)		Vision-Language (LLaVa-1.5 [40])				SAM [35]
				Zero-shot	k-NN	ADE20k	VOC	GQA	POPE	TextVQA	VQAv2	COCO
OpenCLIP-H/14 [11]	632	224	503	77.19	81.10	40.04	68.03	57.94	83.61	50.48	72.24	-
MetaCLIP-H/14 [64]	632	224	486	80.51	82.12	35.39	62.62	60.57	84.76	53.65	75.71	-
SigLIP-L/14 [74]	428	384	241	82.61	85.16	40.53	70.31	57.70	84.85	56.65	71.94	-
Intern-ViT-6B [10]	5,902	224	63	83.20††	78.43	47.20	76.85	60.18	84.02	52.45	76.75	-
	5,537	448	14	††	68.64	42.78	74.43	61.19	**87.23**	**60.36**	78.83	-
*DFN CLIP-H/14 [19]	633	378	170	**83.90**	85.27	39.00	70.29	61.73	85.91	56.78	78.78	-
*OpenAI CLIP-L/14 [51]	305	336	414	75.54	79.80	36.51	67.04	62.20	86.09	57.92	78.49	-
*DINOv2-g/14-reg [14]	1,137	224	294†	-	83.41	48.68	82.78	61.88	85.62	47.18	76.23	-
*SAM-H/16 [35]	637	1024	12	-	22.12	28.08	34.34	49.92	81.76	43.91	57.65	**77.18**
E-RADIO-L *(Ours)*	391	512	468	80.73	83.89	48.22	81.64	61.70	85.07	51.47	76.73	76.31
RADIO-ViT-H/16 *(Ours)*	653	432	158	82.93	**86.06**	**51.34**	**84.71**	**63.01**	86.20	56.32	**79.28**	76.23

Table 1. Comparison of vision foundation and RADIO models. "Zero-Shot" and k-NN are computed on ImageNet-1K. ADE20K [77] and VOC (PascalVOC2012) refer to linear probe semantic segmentation mIOU. GQA, POPE (popular), TextVQA, and VQAv2 are obtained via LLaVa 1.5 [40] by replacing the vision encoder. COCO is the instance segmentation metric introduced by [8] to evaluate SAM [35] 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.*

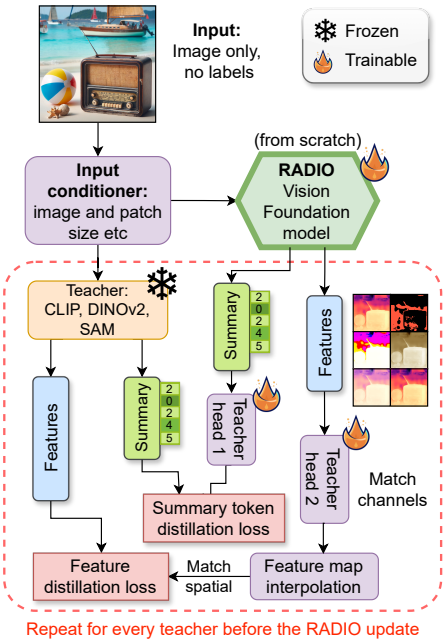


Figure 2. AM-RADIO - is a multi-teacher distillation framework that efficiently trains new vision foundation models of arbitrary architecture. It unifies unique attributes (like zero-shot text grounding, dense correspondence) of each teacher into a single model that even outperforms them on a majority of the tasks.

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 [1, 25, 28, 53, 56, 61, 72].

Instead of using a smaller student model, [63] 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 [26] explore the concept of ensemble distillation, where there are multiple teachers, each of which having restricted domain knowledge. [78] 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) [3] 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 CLIP [51], 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 CLIP is firmly a FM, another model, DINOv2 [48] has emerged with broad capabilities, often surpassing CLIP on dense tasks that require strong spatial features, such as ADE20k [77] and Pascal VOC [18]. Separately, SAM (Segment Anything) [35] 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 CLIP 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 [40, 41].
- 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 [4, 5, 26, 34, 47] 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 [1, 25, 28, 53, 56, 61, 72]. 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 [2, 20, 26, 36, 42, 50, 68, 69, 71, 75, 78]. 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 [78] is different, we find the same overall strategy to be effective. While [61] 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 [3] 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, [48] employ self-distillation to train their smaller variants from the larger teacher. [61] distills their model from a CLIP [51] teacher. Instead of focusing our energy on one teacher *in particular*, we instead grab high-quality versions of CLIP [51] (using OpenCLIP [30]), DINOv2 [48], and SAM [35]. Concurrently with our work, [60] describe a methodology for merging a CLIP model into a pretrained SAM model via distillation, which is, in spirit, quite similar to our approach. In contrast to theirs, we include DINOv2 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 as shown in Figure 2. 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) [15], LAION-400M [54] or DataComp-1B [21]. With this in mind, we choose to study 3 seminal teacher model families: CLIP [51], DINOv2 [48], and SAM [35] as they have demonstrated outstanding performance over a broad range of tasks (as in CLIP), or specifically strong performance on downstream dense tasks, such as semantic segmentation under linear probe (as in DINOv2), or open-vocabulary segmentation (as in SAM). 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 CLIP teacher's language model [51]. k-NN [9, 48, 62] 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 [40] and evaluate it on a wide set of tasks including GQA [29], TextVQA [55], ScienceQA [46] and VQAv2 [23]. Details in Section 5.4.
- **SAM-COCO instance segmentation:** From [8], 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 summary feature vector and the spatial feature vectors for each teacher. The summary feature is computed differently based on the model. For CLIP and DINOv2, we use the "class token" as the summary feature vector, and we don't match a summary for SAM.

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**

Table 2. Ablation study on the choice of training dataset. We use MetaCLIP ViT-H/14 [16] and DINOv2 ViT-g/14 teachers, and a ViT-L/14 student model with CPE [33]. Both "k-NN" and "Zero Shot" are for ImageNet-1k. ADE20k refers to mIOU linear probe on ADE20k.

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 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.

Let $f\left(x|\Theta_0\right)$ be the student vision encoder with parameters Θ_0, and $y_i^s = h_i^{(s)}(x'|\Theta_i^{(s)})$ be the learned student head matching teacher summary features $z_i^{(s)} = t_i^{(s)}\left(x|\Phi_i\right)$ with student adaptor parameters $\Theta_i^{(s)}$ and teacher parameters Φ_i.

$$x' = f\left(x|\Theta_0\right); \quad y_i^{(s)} = h_i^{(s)}\left(x'|\Theta_i^{(s)}\right); \quad z_i^{(s)} = t_i^{(s)}\left(x|\Phi_i\right), \tag{1}$$

$$L_{\text{summary}}(x) = \sum_i \lambda_i L_{\cos}(y_i^{(s)}, z_i^{(s)}) \tag{2}$$

We found empirically that cosine distance loss produced better models compared to L1, MSE, Smooth-L1 [22]. 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 $h_i^{(v)}(x'|\Theta_i^{(v)})$ be the learned

Method	Zero Shot	k-NN	ADE20K
Naive	70.63	79.50	44.71
Uncertainty [12]	70.92	79.37	44.57
AdaLoss [27]	71.31	79.77	44.36

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.

student head for matching teacher feature vectors, and corresponding $t_i^{(v)}(x|\Phi_i^{(v)})$ be the teacher feature vectors, with $x' = f(x|\Theta_0)$, then the spatial feature loss is:

$$L_{\text{match}}(x,y) = \alpha L_{\cos}(x,y) + \beta L_{\text{smooth-l1}}(x,y)$$
$$L_{\text{features}}(x) = \sum_i \gamma_i L_{\text{match}}\left(h_i^{(v)}(x'|\Theta_i^{(v)}), t_i^{(v)}(x|\Phi_i^{(v)})\right) \quad (3)$$

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 [12], AdaLoss [27] (momentum 0.99) and separately with AMTML-KD [42], 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 CLIP 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.

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 [45] 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 [16] 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 SAM features, we feed the expected resolution of 1024^2. Given that our CLIP and DINOv2 teachers are patch-14 models, we opt to feed the student 432^2 inputs, as that is the same effective resolution as 378^2 for patch-14. We found that interpolating DINOv2 features doesn't degrade results, so the teacher operates at 224px and we upsample the outputs to match the student.

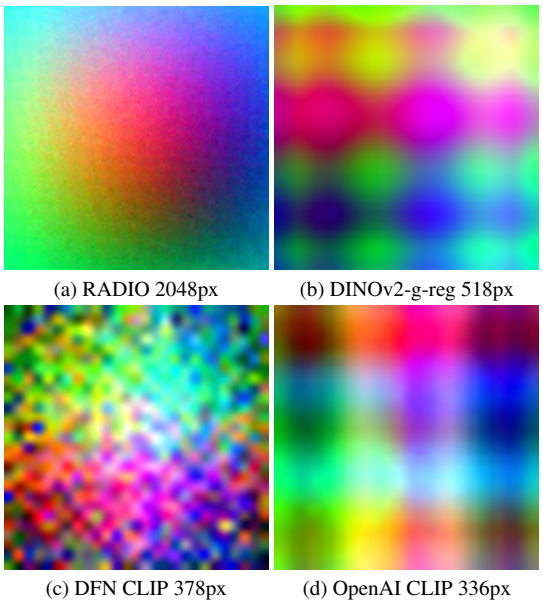
Rank/Teacher Partitioning. We group teacher models by (batch_size, student_resolution), and then distribute the groups to different GPUs, such that each GPU processes a consistent batch size and input resolution. We also sample groups at different rates. For our training setups that include SAM, we train 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 SAM 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 [16] 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) [33] augmentation with the number of positions being equal to 128^2. 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 [6, 66, 67].

High-Resolution ViT Student. In SAM, they employ the ViTDet [37] architecture as a way to reduce the computa-

Method	k-NN	ADE20K
Non-CPE	**82.96**	47.30
CPE	82.84	**48.52**

Table 5. Comparing identical ViT-L/14 student models, with and without CPE [33] formulation. While the student only ever trains at 224^2 resolution, CPE allows us to generalize to 518^2 resolution, not only improving over non-CPE, but even outperforming DINOv2-g itself.



(a) RADIO 2048px (b) DINOv2-g-reg 518px

(c) DFN CLIP 378px (d) OpenAI CLIP 336px

Figure 3. PCA visualization of the position embeddings for various models. The CPE method not only allows RADIO to learn an arbitrarily large absolution position embedding map, but also goes a long way towards regularizing the space and eliminating high frequency artifacts. As seen with the other models, position embeddings normally have regular frequency patterns, leading to undesirable output artifacts from the ViT [6, 66, 67].

tional 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 SAM 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 [73] 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 14^2 feature map, and the teachers a 16^2 feature map. For L_{features} we bilinearly interpolate the outputs to match the larger resolution between the student and teacher features.

	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**

Table 6. Comparing identical ViT models, with CLS token and average pooling summarization.

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.

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 [58], ResNetv2 [57], RegNetY [52], FasterViT [24], EfficientViT [8], ConvNext [44], NFNet [7], SwinV2 [43], MaxViT [59], PoolformerV2 [70] and MViTV2 [38]. 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.

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 7^2 for input resolution of 224^2px, thus hardly

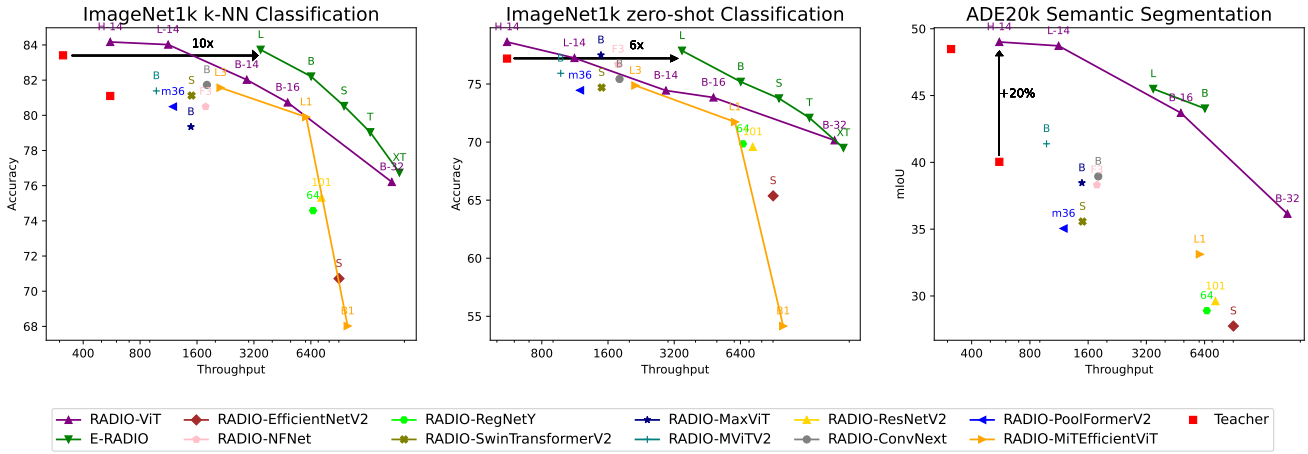


Figure 4. All models followed the same training protocol. The results from three benchmarks show that RADIO and E-RADIO models outperform others in efficiency. This under-performance in other models might be due to overfitting architectures on supervised ImageNet-1K training. E-RADIO notably delivers results 10 times faster and with a 20% improvement over teacher models. We study E-RADIO at 224px resolution, with a window size of 7.

Backbone	Param. Count	Through-put	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
↳ 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

Table 7. Comparison of backbones. Throughput is measured using TensorRT 9.0.1 on A100 in mixed FP16/FP32 precision at batch size 128 on 224^2px 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.

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 [43]), and interleave local windowed attention with "global" windowed attention as done in [24] and ViTDet [37]. 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. Up-/down-sampling is performed by strided convolution with a kernel size 3x3 and a stride of 2. The last idea is borrowed from EdgeViT [49], which uses local-global-local attention. See Appendix for details. Finally, E-RADIO upsamples final feature maps by 2x via a deconvolutional layer and adds them to feature maps from the third stage, resulting in only a 16x spatial resolution reduction. Such upsampling gives an improvement in dense task while being only 10% slower. Results of E-RADIO in Table 7 demonstrate that the proposed architecture significantly outperforms the competition, and can be seen as an efficient replacement for the much slower full ViT.

5.2. Comparison with teachers

A comprehensive set of results is presented in Table 1. We notice that MetaCLIP is better than OpenCLIP, and DFN CLIP better than MetaCLIP. DINOv2 provides important properties for dense tasks: ADE20k and VOC. Our E-RADIO-L model is significantly faster than all ViT models. At the same time, it strongly outperforms MetaCLIP on most metrics at matched throughput, while also enabling Zero-shot capability that is absent in DINOv2 and SAM. Our full model, ViT-H/16, is as fast as the teachers but outperforms them on 6 out of 9 tasks, demonstrating the efficiency of the proposed distillation framework.

Drop-In SAM Replacement. Following [8], we use their evaluation harness to compute the mIOU for instance segmentation using pretrained SAM with vision encoder re-

COCO 2017 drop-in SAM replacement at 1024x1024

Family	Arch	mIOU	Throughput
SAM	Base	75.78	50.94
	Large	77.02	20.62
	Huge	77.18	11.83
E-RADIO (ours)	Large	76.31	121.74
RADIO (ours)	ViTDet-H/16-W8†	76.09	29.09
	ViTDet-H/16-W16†	76.23	27.91

Table 8. We substitute SAM's vision encoder with our RADIO model. RADIO aligns with SAM's features just before the encoder's neck layer. We also examine the impact of varying ViTDet window sizes. Differences in throughput owe to the fact that RADIO doesn't use relative positional embeddings and we reduced shuffling with our patch reordering algorithm (in appendix). Throughput is computed on an NVIDIA A100 GPU using TensorRT and batch size 16. †This is the same model, just with a different window size setting.

placed by our model. Table 8 shows the results of the COCO Instance Segmentation task using the baseline SAM models and RADIO.

5.3. Semantic Segmentation Linear Probing

We train a linear head on top of the frozen features of the teachers and students alike and evaluate performance in the MMSeg [13] framework using the mIoU metric on ADE20k and PascalVOC2012 datasets. We use a training and evaluation crop size of 512 for RADIO, 518 for DINOv2, and the native resolution for the others. We use the "slide" evaluation mode with a stride of $\frac{2}{3}$ the crop size. We train the linear head for 160k steps using a total batch size of 16, a base learning rate of 10^{-3} and the AdamW optimizer.

5.4. Visual Question Answering

We replace the vision encoder in a LLaVA 1.5[40] setup with our own encoder. A 2-layer MLP is used to project frozen visual features into the language token space. Under the default LLaVA 1.5 settings, we pretrain a multimodal projection MLP and then run instruction tuning to finetune a Vicuna 7B-1.5 model[76]. We evaluate models using the validation sets of GQA [29], TextVQA [55], POPE [39] (popular), and we score the model on the Test-Dev set of VQAv2 [23] using EvalAI[65]. We use the vision encoder's native input resolution, resizing the long edge and padding the short edge. Experimental results are compiled in Table 1. Owing to the increased input resolution flexibility of RADIO, we resize the long edge of the image to 432px aspect preserving, only padding to the nearest multiple of the patch size. This results in 462 tokens on average, versus the 576 tokens required by the 336px patch-14 encoders, a 20% reduction.

Backbone	Depth	Surface Normals	Multi-view corr.
DFN CLIP-H/14	52.5	23.0	20.3
OpenAI CLIP-L/14	53.7	25.3	20.7
DINOv2-g/14-reg	83.2	59.6	59.9
SAM-H/16	68.2	50.3	45.3
RADIO-ViT-H/16 (ours)	81.0	58.5	62.1

Table 9. Probing 3D Awareness: we use the code from [17] and evaluate our RADIO model and its teachers on monocular depth, surface normals and multi-view correspondance tasks, using the NAVI[31] dataset. For each task we report the accuracy, averaged over all thresholds.

5.5. 3D Awareness Probing

Following the work from [17], we probe our model's ability to extract 3D features such as depth, surface normals and multi-view keypoint correspondance. Our results are summarized in Table 9 and show that our model's performance is on par with the bigger DINOv2-g-14-reg[14] and significantly better than other comparably-sized teachers.

6. Conclusion and Key Insights

Most VFMs have unique properties such as language grounding (CLIP), dense correspondences (DINOv2), and detailed segmentation (SAM), but also large holes in capability. Distillation allows uniting all these properties in a single model that often outperforms any of the teachers. We have also observed that better teachers yield better students, which allows RADIO to absorb and challenge the current SOTA foundation models at a given point in time.

Feature distillation loss. We observe the crucial importance of full feature distillation to boost the performance of the teacher in dense image understanding tasks, such as an 18% relative improvement on ADE20K.

SAM vs DINOv2. We find that, out of the box, SAM is not well-suited for downstream tasks, whereas DINOv2 significantly outperforms in zero- and few-shot tasks. For example, ADE20K segmentation via linear probing is 1.7x better with the latter, and the ImageNet1k k-NN metric is 4x better. SAM excels in detecting edges and segmenting objects but performs poorly in high-level object description and combining the semantics of multiple objects (Figure 4).

Dense features. As seen in figure 1, RADIO is capable of producing high resolution and low-noise features. An issue we identified, however, shown in figure 5 is that RADIO appears to have a latent 'low resolution' and 'high resolution' mode, likely due to the partitioned training between CLIP+DINO and SAM objectives, which we intend to fix in future work.

Efficient backbone. Based on our analysis of distilling efficient backbones, we conclude that most model designs are overly tailored towards supervised training on ImageNet1K,

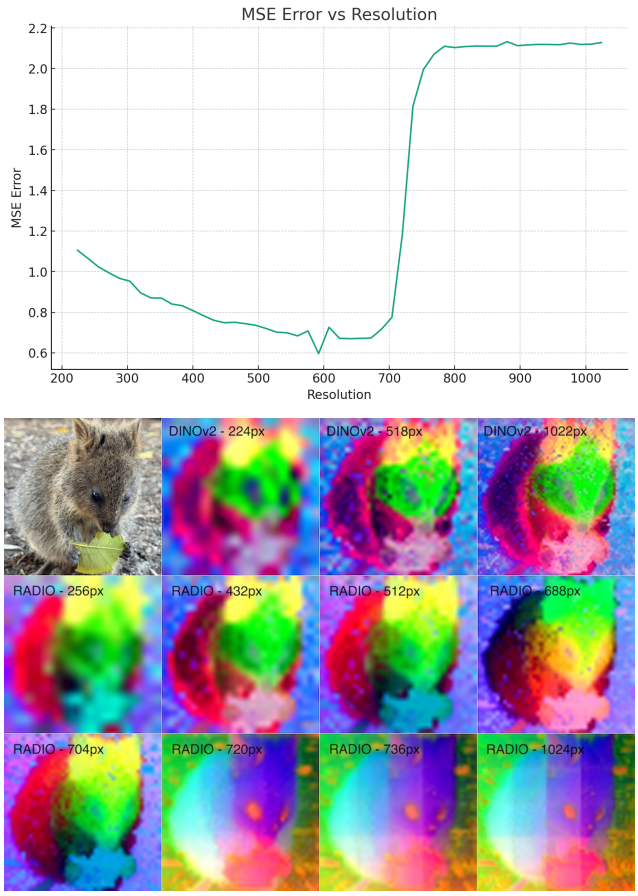


Figure 5. RADIO "mode switches" when resolution is increased. In the plot, we show the MSE error between the RADIO features coming from its DINOv2 head at different resolutions, versus the features actually produced by DINOv2 at 518px. We bilinearly interpolate the RADIO features to match the DINOv2 feature resolution. At 720px, there is a sudden jump in the error, which corresponds with a complete change in color space in the image.

and as a result, do not scale well to VFM settings. We designed a new vision backbone, E-RADIO, with a hybrid CNN-Transformer architecture that improves upon the Pareto frontier.

References

[1] S. Ahn, S. Hu, A. Damianou, N. D. Lawrence, and Z. Dai. Variational information distillation for knowledge transfer. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9155–9163, Los Alamitos, CA, USA, 2019. IEEE Computer Society. 2, 3

[2] Umar Asif, Jianbin Tang, and Stefan Harrer. Ensemble knowledge distillation for learning improved and efficient networks. In *European Conference on Artificial Intelligence*, 2019. 3

[3] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundational models defining a new era in vision: A survey and outlook, 2023. 2, 3

[4] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems*, pages 2654–2662, 2014. 3

[5] L. Beyer, X. Zhai, A. Royer, L. Markeeva, R. Anil, and A. Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10915–10924, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 3

[6] Daniel Bolya, Chaitanya Ryali, Judy Hoffman, and Christoph Feichtenhofer. Window attention is bugged: How not to interpolate position embeddings, 2023. 5, 6

[7] Andrew Brock, Soham De, Samuel L. Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization, 2021. 6

[8] Han Cai, Junyan Li, Muyan Hu, Chuang Gan, and Song Han. Efficientvit: Multi-scale linear attention for high-resolution dense prediction, 2023. 2, 4, 6, 7

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. 4

[10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 2

[11] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning, 2022. 2

[12] R. Cipolla, Y. Gal, and A. Kendall. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7482–7491, Los Alamitos, CA, USA, 2018. IEEE Computer Society. 5, 7

[13] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 8

[14] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2023. 2, 8

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 4, 5

[17] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun,

Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3D Awareness of Visual Foundation Models. In *CVPR*, 2024. 8

[18] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 2

[19] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks, 2023. 2

[20] Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. Efficient knowledge distillation from an ensemble of teachers. In *Interspeech*, 2017. 3

[21] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets, 2023. 3

[22] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 4

[23] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4, 8

[24] Ali Hatamizadeh, Greg Heinrich, Hongxu Yin, Andrew Tao, Jose M. Alvarez, Jan Kautz, and Pavlo Molchanov. Fastervit: Fast vision transformers with hierarchical attention, 2023. 6, 7

[25] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Choi. A comprehensive overhaul of feature distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1921–1930, Los Alamitos, CA, USA, 2019. IEEE Computer Society. 2, 3

[26] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2, 3

[27] Hanzhang Hu, Debadeepta Dey, Martial Hebert, and J. Andrew Bagnell. Learning anytime predictions in neural networks via adaptive loss balancing. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2019. 5, 8

[28] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *CoRR*, abs/1707.01219, 2017. 2, 3

[29] Drew A. Hudson and Christopher D. Manning. GQA: a new dataset for compositional question answering over real-world images. *CoRR*, abs/1902.09506, 2019. 4, 8, 9, 10, 11

[30] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 3

[31] Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, André Araujo, Ricardo Martin-Brualla, Kaushal Patel, Daniel Vlasic, Vittorio Ferrari, Ameesh Makadia, Ce Liu, Yuanzhen Li, and Howard Zhou. Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations, 2023. 8

[32] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. 2

[33] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers, 2023. 4, 5, 6

[34] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 2765–2774, Red Hook, NY, USA, 2018. Curran Associates Inc. 3

[35] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 1, 2, 3

[36] Xu Lan, Xiatian Zhu, and Shaogang Gong. Knowledge distillation by on-the-fly native ensemble, 2018. 3

[37] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection, 2022. 5, 7, 6

[38] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection, 2022. 6

[39] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 8

[40] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2, 3, 4, 8

[41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 3

[42] Yuang Liu, Wei Zhang, and Jun Wang. Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing*, 415:106–113, 2020. 3, 5

[43] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution, 2022. 6, 7, 2

[44] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022. 6

[45] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 5

[46] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 4

[47] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *AAAI Conference on Artificial Intelligence*, 2019. 3

[48] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 1, 2, 3, 4

[49] Junting Pan, Adrian Bulat, Fuwen Tan, Xiatian Zhu, Lukasz Dudziak, Hongsheng Li, Georgios Tzimiropoulos, and Brais Martinez. Edgevits: Competing light-weight cnns on mobile devices with vision transformers. In *ECCV*, 2022. 7, 2

[50] Seonguk Park and Nojun Kwak. Feature-level ensemble knowledge distillation for aggregating knowledge from multiple networks. In *European Conference on Artificial Intelligence*, 2020. 3

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4

[52] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces, 2020. 6

[53] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *CoRR*, abs/1412.6550, 2014. 2, 3

[54] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021. 3

[55] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4, 8, 12, 13, 14

[56] X. Sun, R. Panda, C. Chen, A. Oliva, R. Feris, and K. Saenko. Dynamic network quantization for efficient video inference. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7355–7365, Los Alamitos, CA, USA, 2021. IEEE Computer Society. 2, 3

[57] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016. 6

[58] Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training. *CoRR*, abs/2104.00298, 2021. 6

[59] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer, 2022. 6

[60] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding, 2023. 3, 7, 8

[61] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation, 2022. 2, 3

[62] Zhirong Wu, Yuanjun Xiong, X Yu Stella, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 4

[63] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2020. 2

[64] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. 2023. 2

[65] Deshraj Yadav, Rishabh Jain, Harsh Agrawal, Prithvijit Chattopadhyay, Taranjeet Singh, Akash Jain, Shiv Baran Singh, Stefan Lee, and Dhruv Batra. Evalai: Towards better evaluation systems for ai agents, 2019. 8

[66] Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, and Yue Wang. Emernerf: Emergent spatial-temporal scene decomposition via self-supervision, 2023. 5, 6

[67] Jiawei Yang, Katie Z Luo, Jiefeng Li, Kilian Q Weinberger, Yonglong Tian, and Yue Wang. Denoising vision transformers, 2024. 5, 6

[68] Ze Yang, Linjun Shou, Ming Gong, Wutao Lin, and Daxin Jiang. Model compression with two-stage multi-teacher knowledge distillation for web question answering system. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, page 690–698, New York, NY, USA, 2020. Association for Computing Machinery. 3

[69] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1285–1294, New York, NY, USA, 2017. Association for Computing Machinery. 3

[70] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision, 2022. 6

[71] Fei Yuan, Linjun Shou, Jian Pei, Wutao Lin, Ming Gong, Yan Fu, and Daxin Jiang. Reinforced multi-teacher selection for knowledge distillation, 2020. 3

[72] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 2, 3

[73] Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, pages 40770–40803. PMLR, 2023. 6

[74] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023. 2

[75] Haoran Zhao, Xin Sun, Junyu Dong, Changrui Chen, and Zihe Dong. Highlight every step: Knowledge distillation via collaborative teaching. *IEEE Transactions on Cybernetics*, 52(4):2070–2081, 2022. 3

[76] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 8

[77] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017. 2

[78] Konrad Zuchniak. Multi-teacher knowledge distillation as an effective method for compressing ensembles of neural networks, 2023. 2, 3

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

Supplementary Material

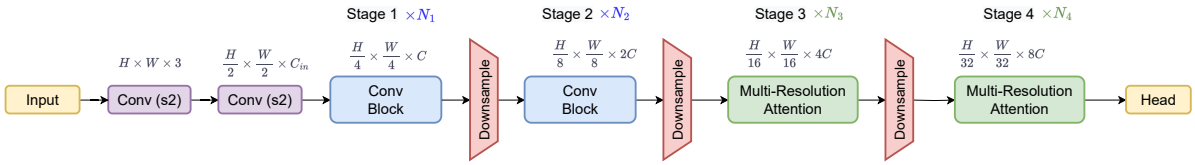


Figure 6. High level architecture of the ERADIO network architecture. Overall architecture is composed of multiple stages: 1) the stem, 2) 2 convolutional blocks from YOLOv8, 3) 2 transformer blocks with multi-resolution windowed self attention.

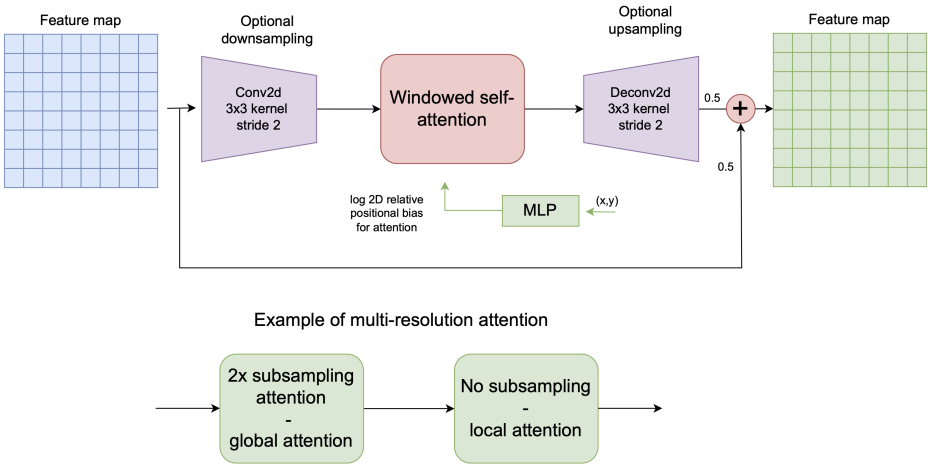


Figure 7. Multi-resolution attention for E-RADIO

A. E-RADIO architecture details

The architecture of E-RADIO is illustrated in Figure 6. It is a hybrid CNN-Transformer architecture. First 2 stages follow convolution paradigm and have the C2f architecture from YOLOv8 model [32]. The last 2 stages have the Transformer architecture with windowed attention and multi-resolution attention (MRA) structure. Every stage, except the last one, are followed by downsample block. We implement it as a strided convolution with 3x3 kernel and stride 2, followed by batch normalization layer.

A.1. Multi-Resolution Attention

Standard transformers struggle to scale with high input image resolution because of quadratic complexity of the attention. SWIN [43] proposed to use windowed attention to reduce the complexity of attention. We reuse windowed attention in the E-RADIO. To address for missing communication between windows, SWIN introduced window shifting, unfortunately, it has non-negligible compute cost. Instead, we propose multi-resolution attention inspired by EdgeViT's Local-Global-Local attention [49]. The idea is illustrated in Figure 7. Every layer in the transformer will have a local windowed attention with optional subsampling via convolutional operator. For example, if susbampling is dissabled, then it is just a standard windowed attention. If the subsampling ratio is 2, then the feature map is downsampled by a factor of 2, windowed attention is performed, and then the feature map is upsampled to the original resolution with deconvolution. For FasterVIT2 models, we interleave subsampled attention with ratio 2 and the normal attention with no subsampling.

A.2. Configurations

All models in the family follow the same configuration except the embedding dimension (hide dimension). We simply scale it up with bigger models. Other parameters:

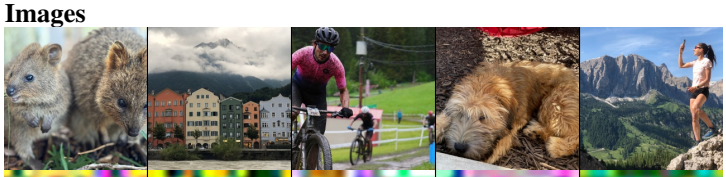
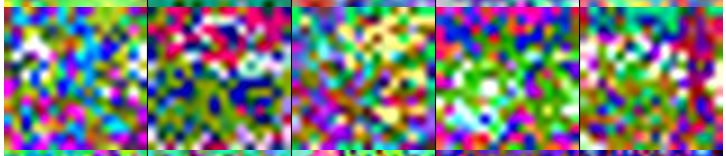
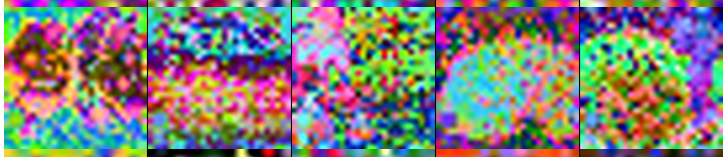
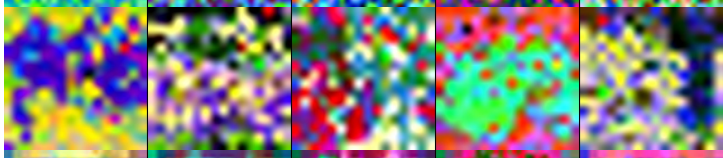
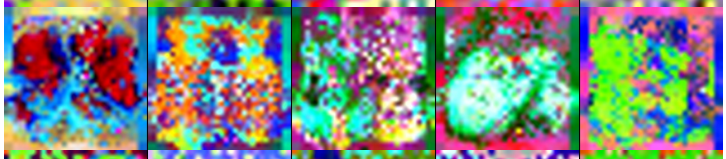
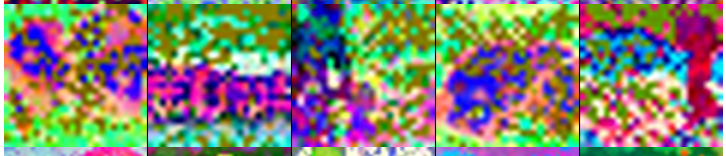
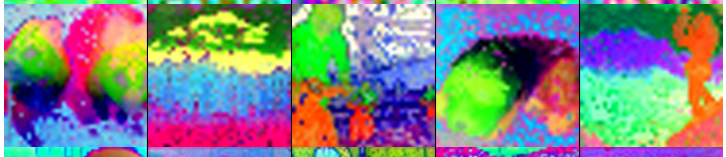
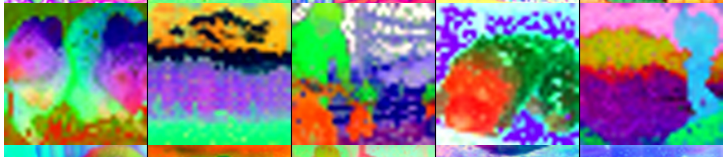
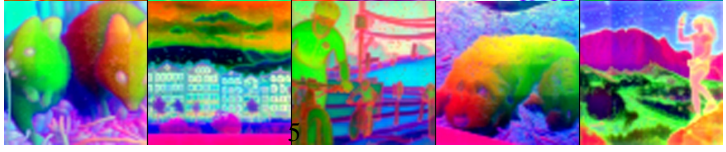
- Input resolution is 224
- In-stem contains 2 3x3 convolutions with stride 2
- Total stages: 2 convolutional and 2 transformer
- First stage takes input feature size of 56x56, has 3 layers with C2f structure from YOLO8 [32].
- Second stage takes input feature size of 28x28, has 3 layers of C2f.

- Third stage takes features of size 14x14, has 5x multi-resolution attention, window size 7.
- Forth stage takes features of size 7x7, has 5x windowed attention of window size 7.
- Embedding dimension for different model variants: XT - 64, T - 80, S - 96, B - 128, L - 192. The smallest XT and T models have [1, 3, 4, 5] layers for each of 4 stages.
- Output features have resolution of 14x14 and are obtained by upsampling the features of stage 4 by 2x with deconvolution and adding to stage 3 features of size 14x14.

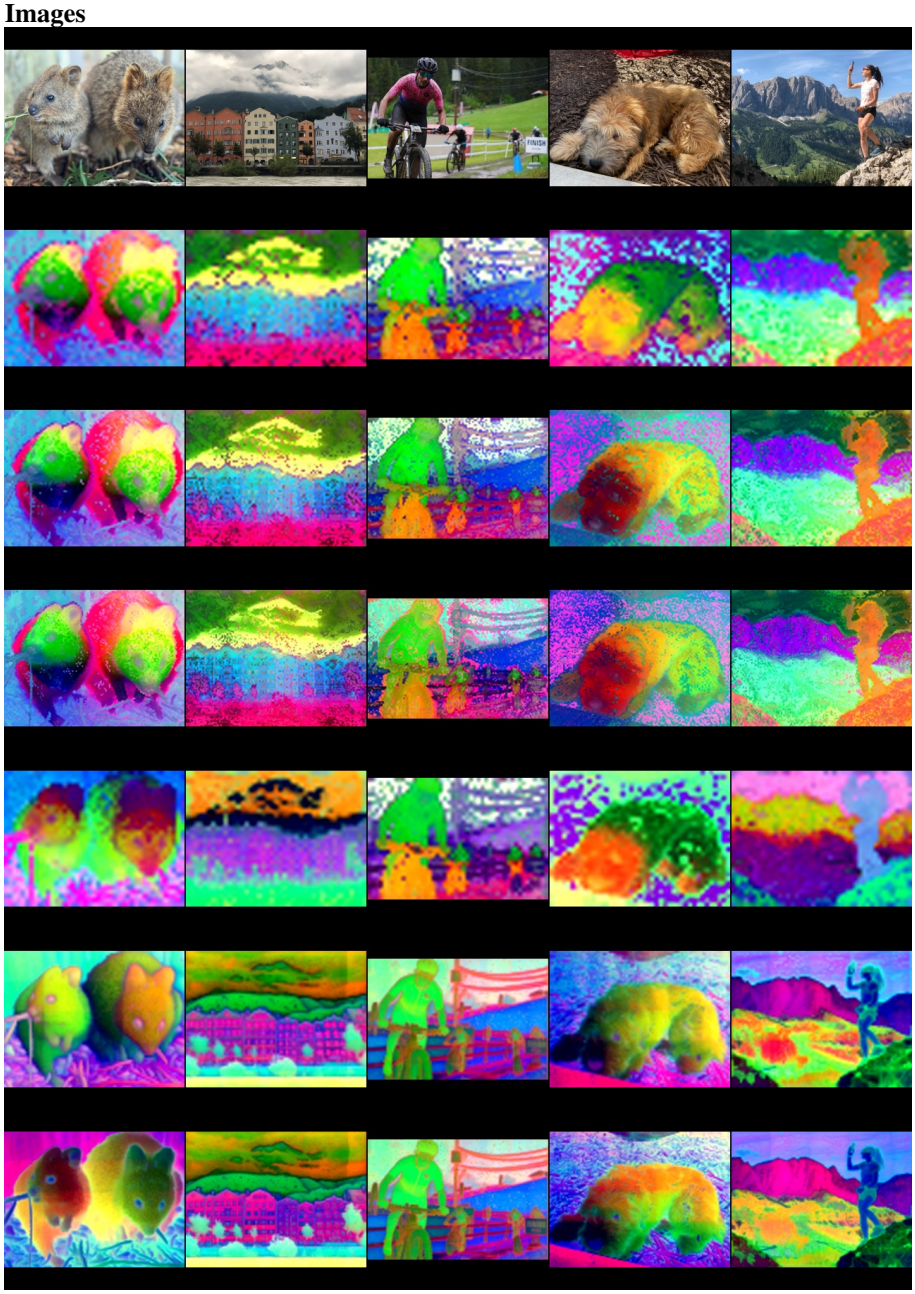
B. PCA Visualizations

We visualize various models using PCA to reduce the model's spatial feature dimensionality down to 3 dimensions, and directly map those to RGB. Most models are only able to handle square inputs at fixed resolutions, however DINOv2 and RADIO can handle arbitrary resolutions and aspect ratios, so we visualize them in both settings.

B.1. Square Models

Model	Resolution	Images
		
OpenCLIP-H/14	224	
MetaCLIP-H/14	224	
SigLIP-M/14	384	
InternViT-6B	224	
	448	
DFN CLIP	378	
OpenAI CLIP	336	
DINOv2-g	518	
SAM-H	1024	
RADIO	512	
	1024	

B.2. Flexible Models

Model	Resolution	Images
DINOv2-g	518	
	1022	
	2044	
RADIO	512	
	1024	
	2048	

C. ViTDet Augmentation

The following python code shows how the alternating window/global architecture of ViTDet [37] can be applied to a transformer. We take advantage of the fact that transformers are permutation invariant *after position encodings have been applied*, and thus it's easy to organize the patch order such that contiguous chunks of patches belong to the same window. Once reordered in this way, alternating between windowed and global attention is achieved simply by absorbing the windows into the batch dimension or returning to the original shape respectively. We also enforce that the final transformer layer always applies global attention.

```
from einops import rearrange
def reorder_patches(patches: torch.Tensor,
```

```
                    patched_size: Tuple[int, int],
                    window_size: int):
    p_idxs = torch.arange(patches.shape[1])
    p_idxs = rearrange(p_idxs, '(wy y wx x) -> (wy wx y x)',
                       wy=patched_size[0] // window_size, y=window_size,
                       wx=patched_size[1] // window_size, x=window_size)
    p_idxs = p_idxs.reshape(1, -1, 1).expand_as(patches)

    return torch.gather(patches, p_idxs), p_idxs

def vitdet_aug(blocks: nn.Sequential,
               patches: torch.Tensor,
               patched_size: Tuple[int, int],
               window_sizes: List[int],
               num_windowed: int):
    B, T, C = patches.shape
    window_size = sample(window_sizes)
    sq_window_size = window_size ** 2
    patches, p_idxs = reorder_patches(patches, patched_size, window_size)
    period = num_windowed + 1
    for i, block in enumerate(blocks[:-1]):
        if i % period == 0:
            patches = patches.reshape(B * sq_window_size, -1, C)
        elif i % period == num_windowed:
            patches = patches.reshape(B, T, C)
        patches = block(patches)

    # Always use global attention with the last block
    patches = patches.reshape(B, T, C)
    patches = blocks[-1](patches)

    # Finally, put the patches back in input order
    ret = torch.empty_like(patches)
    ret = ret.scatter(dim=1, index=p_idxs, src=patches)
    return ret
```

D. Comparison with SAM-CLIP [60]

Concurrently with our work, SAM-CLIP was introduced as a method of fusing SAM and CLIP into a single model. Due to the concurrency of effort, we don't compare our model with the full suite of metrics demonstrated in their method, however, we do have some overlap in key metrics such as Zero-Shot ImageNet-1k, and ADE20k semantic segmentation via linear probing. We present the comparison in table 10, however we note that there are enough differences between these two models that we can't conclude one way or another what is the superior approach. Instead we'll argue that DINOv2 does a better job of ADE20k linear probing than SAM, and thus our significantly higher quality on this metric is likely due to the inclusion of DINOv2, which is a key introduction with our approach.

E. Automatic Loss Balancing

E.1. Uncertainty

Following [12], we have:

$$L(x) = \sum_k \frac{1}{2\sigma_k^2} L_k(x) + \log \sigma_k \tag{4}$$

Family	Model	Zero-Shot	ADE20k
SAM	ViTDet-H/16		28.2
DFN CLIP	ViT-H/14	**83.9**	31.7
SAM-CLIP	ViTDet-B/16	71.7	38.4
RADIO	ViT-H/14	82.7	**51.3**

Table 10. We compare our common key metrics with those demonstrated in SAM-CLIP [60]. We note that there are numerous differences between the two approaches, including model capacity and architecture. SAM-CLIP uses the ViT-B variant of SAM as a starting point, which implies it's a ViTDet-B/16 architecture. As a result of this choice, their metrics are computed at a resolution of 1024. RADIO trains a vanilla ViT-H/14 from scratch, and as a result of the flexibility gained via the CPE method, we evaluate Zero-Shot ImageNet1k at a resolution of 432, and we run ADE20k linear probing at a resolution of 512 using the exact same weights. We note that Zero-Shot quality is largely determined by the quality of the CLIP teacher and the capacity of the student. We attribute our superior quality on ADE20k semantic segmentation largely to our inclusion of DINOv2 as a teacher.

where the σ_k values are predicted by the student. In practice, the student predicts $b := \log \sigma_k^2$ for numerical stability, to avoid division by zero, and to regress unconstrained scalar values.

We make some minor modifications to (4) to make training a bit more stable in our setting. We replace the manual λ scalars with the learned uncertainty weights, and add the loss term for large uncertainties. Altogether, this yields:

$$\begin{aligned} \lambda_k &= \frac{e^{-b_k}}{2} \\ L(x) &= \sum_k \lambda_k L_k(x) + \frac{b_k}{2} \end{aligned} \tag{5}$$

Let $b_i^{(s|v)}(x'|\Theta_i^{(s)})$ be a learned function predicting balance parameters for teacher i and summary weight (s) or feature vector weight (v), we transform equation (5) slighty to:

$$\begin{aligned} \psi(x) &= \log(1 + e^x) \\ \lambda_i^{(m)} &= e^{-b_i^{(m)}(x')} \\ L(x) &= \sum_i \sum_{m \in \{s,v\}} \lambda_i^{(m)} L_i^{(m)}(x) + \psi\left(b_i^{(m)}(x')\right) \end{aligned} \tag{6}$$

The function $\psi(x)$ is the familiar "softplus" nonlinear activation function. We drop the division by 2 on the left because, assuming outputs are initially $b \sim \mathcal{N}(0, \sigma^2)$, then the loss weights will initially have an expected value of 1, matching the naive weighting. On the right, we replace $\frac{b_k}{2}$ with $\psi(x)$ for a few reasons:

- When $x \gtrapprox 4$, then $\psi(x) \approx x$, yielding the same expression as before.
- When $x \approx 0$, then $\psi'(x) \approx \frac{1}{2}$, yielding the same expression as before.
- When $x < 0$, which translates to a loss weight > 1, $\psi'(x) \to 0$, improving stability as the weight gets larger.
- It has range $(0, \infty)$ which aesthetically enforces the loss to be greater than zero.

E.2. AdaLoss

In addition to uncertainty auto-balancing, we also explored AdaLoss [27]. In this formulation, we have:

$$\begin{aligned} \lambda_i^{(m)} &= \frac{1}{\mathbb{E}(L_i^{(m)})} \\ L(x) &= \sum_i \sum_{m \in \{s,v\}} \lambda_i^{(m)} L_i^{(m)}(x) \end{aligned} \tag{7}$$

F. Visual Question Answering Samples

Figures 9 to 13 show sample questions from our Visual Question Answering datasets, together with sample answers when using our vision encoders in a LLaVA setup.

Figure 8. Visualization of the LLaVA attention maps over the visual features produced by a RADIO encoder. We use one sample image from the GQA[29] validation set and one associated question: "What color is the helmet in the middle of the image?". For each layer in the language model, we retrieve attention scores for all positions of the visual tokens, average them over all attention heads, and overlay corresponding heat maps with the input image. We can see that as we progress through the layers, the model's attention focuses on the relevant part of the image. The model's answer is "Blue".

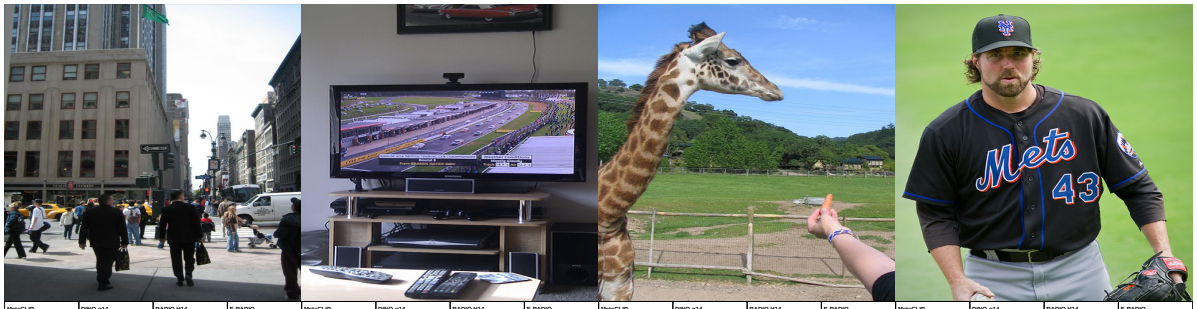
MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Does the boat to the left of the flag look small or large? A: small			
Small	Small	Small	Small
Q:Is the boat to the left of the other boat small and white? A: yes			
No	Yes	Yes	Yes

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Which kind of furniture is right of the drawers? A: shelves			
Shelves	Shelf	Table	Shelf
Q:Is the lamp to the right or to the left of the pot? A: right			
Right	Right	Right	Right
Q:Is the pan to the right of a bowl? A: no			
No	No	No	No
Q:How are the items of furniture to the right of the vegetables that are not rotten called? A: shelves			
Shelves	Shelves	Shelves	Shelves
Q:Which kind of appliance is to the right of the shelf? A: stove			
Oven	Blender	Oven	Oven
Q:Are there stoves to the left of the shelf? A: no			
No	No	No	No
Q:Which kind of furniture is to the left of the pot? A: shelf			
Shelf	Shelf	Shelf	Shelf
Q:Which kind of material makes up the stove? A: wood			
Wood	Wood	Wood	Brick
Q:Is the floor gray or red? A: gray			
Gray	Gray	Gray	Gray
Q:Is the wood stove to the right or to the left of the shelf? A: right			
Right	Right	Right	Right
Q:Which kind of furniture is to the left of the shelves? A: drawers			
Shelf	Shelf	Shelf	Shelf
Q:What is the color of the container that is filled with water? A: brown			
Brown	Brown	Black	Blue
Q:Are the shelves and the drawers made of the same material? A: yes			
No	Yes	Yes	No
Q:Is the wood table to the left of the appliance the pot is on? A: no			
No	No	No	No
Q:Is there any kettle on the appliance that the vegetable is to the left of? A: no			
Yes	No	Yes	No
Q:Is the container that looks brown filled with water? A: yes			
Yes	Yes	Yes	No
Q:What are the shelves made of? A: wood			
Wood	Wood	Wood	Wood
Q:On which side of the image is the pot? A: right			
Left	Right	Right	Right
Q:In which part of the picture is the container, the top or the bottom? A: bottom			
Top	Top	Top	Top
Q:Are there any lamps or beds in the photograph? A: yes			
No	No	No	No
Q:What is on the stove? A: pot			
Pot	Pot	Pot	Pot
Q:Are the drawers in the bottom part or in the top of the picture? A: top			
Bottom	Top	Top	Bottom
Q:What is the floor made of? A: brick			
Brick	Stone	Stone	Stone
Q:Is the floor made of hardwood? A: no			
No	No	No	No
Q:Is the wood stove to the right of the wood table? A: no			
Yes	No	No	Yes
Q:What is the item of furniture to the left of the shelves made of wood? A: shelf			
Chair	Shelf	Shelf	Shelf
Q:What is the stove made of? A: wood			
Wood	Wood	Wood	Wood

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Who is in front of the gray building? A: passengers			
People	People	People	People
Q:What are the passengers in front of? A: building			
Building	Building	Building	Airport
Q:Is there a window that is not large? A: yes			
No	No	No	No
Q:Are there any doors or cars in the picture? A: no			
No	Yes	No	Yes
Q:Who is before the building? A: passengers			
People	People	People	People
Q:Who is boarding the airplane? A: passengers			
People	People	People	People
Q:Who is boarding the plane? A: passengers			
People	People	People	People
Q:Are there any trucks? A: no			
No	No	No	No
Q:Is the gray building behind the passengers that are boarding the airplane? A: yes			
Yes	Yes	Yes	Yes
Q:Does the logo look white and large? A: no			
Yes	Yes	Yes	Yes

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What is the color of the fence made of metal? A: green			
Brown	Green	Green	Green
Q:Which kind of animal is to the right of the sheep? A: cow			
Goat	Goat	Goat	Goat
Q:What is the small animal? A: cow			
Goat	Sheep	Sheep	Goat
Q:Is the sheep white or black? A: black			
Black	Black	White	White
Q:Are there both fences and goats in this photograph? A: yes			
No	Yes	Yes	Yes
Q:Are there either fences or toilets? A: yes			
No	Yes	Yes	Yes
Q:What type of animal is to the right of the black animal? A: cow			
Sheep	Sheep	Sheep	Sheep
Q:Do the animals in this photo have different species? A: yes			
Yes	Yes	No	No
Q:What is the animal to the right of the goat on the left? A: cow			
Sheep	Sheep	Sheep	Sheep
Q:Are there either horses or sheep that are not black? A: no			
No	No	No	No
Q:Are there both a fence and a goat in this photo? A: yes			
No	No	Yes	Yes
Q:Does the goat's tail look small and gray? A: yes			
Yes	Yes	Yes	Yes
Q:Is the sheep to the right or to the left of the animal that is not big? A: left			
Right	Right	Right	Right
Q:What animal is to the left of the small animal? A: goat			
Goat	Sheep	Sheep	Goat
Q:Are there any horses or cows that are not small? A: no			
No	No	No	No

Figure 9. Sample questions from the GQA[29] and their answers from our LLaVA models, using various image encoders. Answers are painted green when they match the ground truth, pink otherwise.



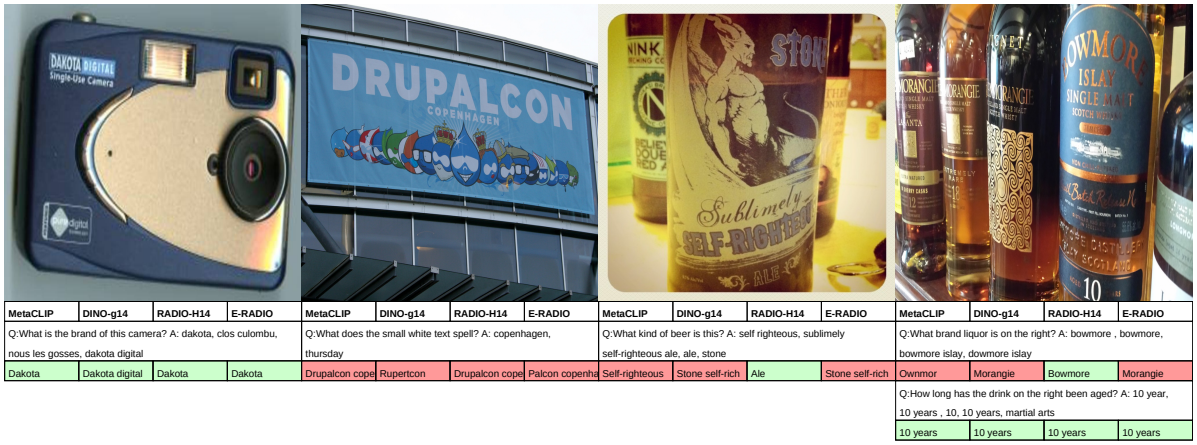
MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Is the woman to the left or to the right of the man that is wearing trousers? A: right			
Left	Left	Right	Right
Q:Who is wearing a shirt? A: man			
Man	Man	Man	Woman
Q:Is the bag in the top part or in the bottom of the photo? A: bottom			
Bottom	Bottom	Bottom	Bottom
Q:Is the bag to the right of the other bag tan or black? A: black			
Black	Tan	Tan	Tan
Q:Who is wearing the glasses? A: man			
Man	Man	Man	Woman
Q:What is the color of the logo? A: black			
White	White	White	Red
Q:Do you see women to the left of the man that is wearing pants? A: no			
No	Yes	No	Yes
Q:What does the man to the left of the traffic light wear? A: suit			
Jacket	Jacket	Jacket	Jacket
Q:What vehicle is the black logo painted on? A: van			
Van	Van	Van	Sign
Q:What's painted on the door? A: logo			
Nothing	Nothing	Nothing	Logo
Q:What is the color of the car? A: yellow			
White	Yellow	Yellow	White
Q:Is the jacket different in color than the logo? A: no			
Yes	Yes	No	No
Q:Which color is the sweater the woman wears? A: white			
Black	Black	Brown	Black
Q:What does the woman wear? A: sweater			
Jacket	Jacket	Skirt	Jacket
Q:Is the blue bag in the bottom or in the top? A: bottom			
Bottom	Bottom	Bottom	Bottom
Q:What is the color of the jacket the man wears? A: black			
Black	Black	Black	Black
Q:What is painted on the van to the right of the man? A: logo			
Logo	Logo	Logo	Logo
Q:Is the car in the top part of the photo? A: no			
No	No	No	No
Q:Is the yellow vehicle to the right or to the left of the man that wears jeans? A: left			
Left	Left	Left	Left
Q:Who is wearing a jacket? A: man			
People	Man	Man	People
Q:Is the woman to the right of the man carrying a bag? A: yes			
Yes	Yes	Yes	No
Q:What is the woman that is to the right of the man carrying? A: bag			
Bag	Bag	Bag	Bag
Q:Does the door look white? A: yes			
Yes	No	No	Yes
Q:Who is wearing jeans? A: man			
People	Man	People	People
Q:Are there either chairs or bags? A: yes			
Yes	Yes	Yes	No
Q:Are there blue bags or cars? A: yes			
Yes	Yes	Yes	Yes
Q:Which side of the picture is the van on? A: right			
Right	Right	Right	Right
Q:On which side of the picture is the woman? A: left			
Left	Left	Right	Left
Q:Is the blue bag to the right or to the left of the woman that wears a sweater? A: left			
Left	Left	Left	Left
Q:Is the man to the left of the bun wearing shorts? A: no			
No	No	No	No
Q:Who is wearing the pants? A: man			
Man	Man	Man	Woman
Q:Is the car different in color than the jacket? A: yes			
Yes	Yes	Yes	Yes
Q:Is the color of the door different than the van? A: no			
No	No	No	No

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Does the remote control to the left of the other remote control look black? A: yes			
Yes	Yes	Yes	Yes
Q:What is the device above the box? A: controller			
Television	Television	Television	Television
Q:Which kind of device is above the box? A: controller			
Television	Television	Television	Television
Q:Is the controller above the container to the left of the speaker? A: yes			
Yes	Yes	Yes	Yes
Q:Are there any tables or couches that are not tan? A: no			
No	No	No	No
Q:What is the color of the table? A: tan			
Brown	Brown	Brown	Brown
Q:What is the device on the carpet? A: speaker			
Remote control	Remote control	Speaker	Television
Q:Is there any speaker on the carpet? A: yes			
Yes	Yes	Yes	No
Q:What device is on the carpet? A: speaker			
Remote control	Remote control	Speaker	Television
Q:Which color is the speaker that is to the right of the remote? A: black			
Black	Black	Black	Black
Q:Is the remote in the top part or in the bottom of the photo? A: bottom			
Bottom	Bottom	Bottom	Bottom
Q:What size is the device that is in the top of the image? A: small			
Large	Large	Large	Large
Q:Is it outdoors? A: no			
No	No	No	No
Q:In which part of the photo is the box, the bottom or the top? A: bottom			
Bottom	Bottom	Bottom	Bottom
Q:In which part is the small device, the bottom or the top? A: top			
Bottom	Bottom	Bottom	Bottom
Q:What is on the tan table? A: remote control			
Remote control	Remote	Remote	Television
Q:Which kind of device is on the table? A: remote control			
Remote control	Remote control	Remote control	Television
Q:What type of device is on the table? A: remote control			
Remote control	Remote control	Remote	Television
Q:What color is the carpet? A: gray			
Gray	Brown	Gray	White

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Is it indoors or outdoors? A: outdoors			
Outdoors	Outdoors	Outdoors	Outdoors
Q:Which side of the photo is the orange vegetable on? A: right			
Right	Right	Right	Right
Q:Does the shirt look green? A: no			
No	No	No	No
Q:On which side of the picture is the person? A: right			
Right	Right	Right	Right
Q:Does the bracelet look blue? A: yes			
Yes	Yes	Yes	Yes
Q:What animal is to the left of the house? A: giraffe			
Giraffe	Giraffe	Giraffe	Giraffe
Q:Is the giraffe to the left or to the right of the person that is on the right? A: left			
Left	Left	Left	Left
Q:Which place is it? A: field			
Field	Zoo	Field	Field
Q:Are there both fences and giraffes in the image? A: yes			
Yes	Yes	Yes	Yes
Q:Which kind of animal is this, a giraffe or a bear? A: giraffe			
Giraffe	Giraffe	Giraffe	Giraffe
Q:Is the person to the right or to the left of the animal near the fence? A: right			
Right	Right	Right	Right
Q:Is the person to the right of the animal near the fence? A: yes			
Yes	Yes	Yes	Yes
Q:Is the house on the right side? A: yes			
Yes	Yes	Yes	Yes
Q:What kind of vegetable is to the right of the giraffe? A: carrot			
Carrot	Carrot	Carrot	Carrot

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Does the belt look lan? A: no			
No	No	No	No
Q:Is the baseball the same color as the field? A: no			
Yes	Yes	Yes	Yes
Q:Which kind of clothing is gray? A: trunks			
Pants	Pants	Pants	Pants
Q:Which place is it? A: field			
Field	Field	Field	Field
Q:Is the shirt the same color as the belt? A: yes			
Yes	Yes	Yes	Yes
Q:Is the cap black? A: yes			
Yes	No	Yes	Yes
Q:What color are the trunks, gray or yellow? A: gray			
Gray	Gray	Gray	Gray
Q:What are the gray clothing items called? A: trunks			
Pants	Pants	Pants	Pants
Q:What color does the shirt have? A: black			
Black	Blue	Black	Black
Q:Do the baseball and the number have a different colors? A: yes			
No	No	Yes	No

Figure 10. Sample questions from the GQA[29] and their answers from our LLaVA models, using various image encoders. Answers are painted green when they match the ground truth, pink otherwise.

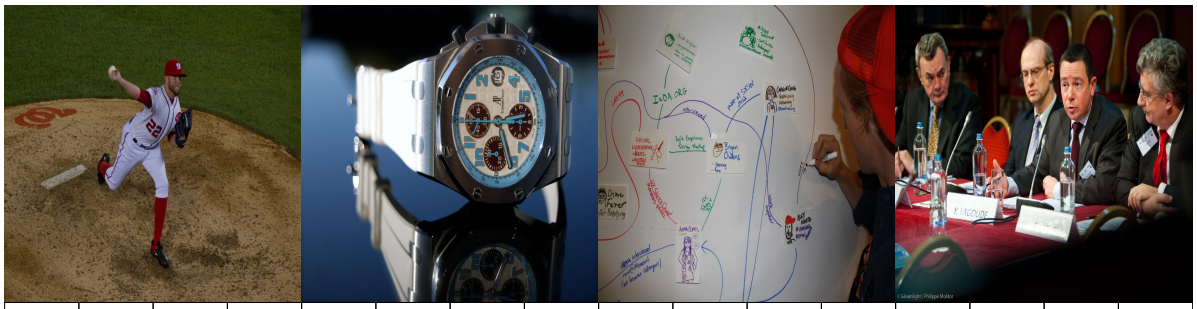


MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What is the brand of this camera? A: dakota, clos culombu, nous les gosses, dakota digital			
Dakota	Dakota digital	Dakota	Dakota

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What does the small white text spell? A: copenhagen, thursday			
Drupalcon cope	Rupertcon	Drupalcon cope	Palcon copenha

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What kind of beer is this? A: self righteous, sublimely self-righteous ale, ale, stone			
Self-righteous	Stone self-rich	Ale	Stone self-rich

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What brand liquor is on the right? A: bowmore , bowmore, bowmore islay, dowmore islay			
Ownmor	Morangie	Bowmore	Morangie
Q:How long has the drink on the right been aged? A: 10 year, 10 years , 10, 10 years, martial arts			
10 years	10 years	10 years	10 years



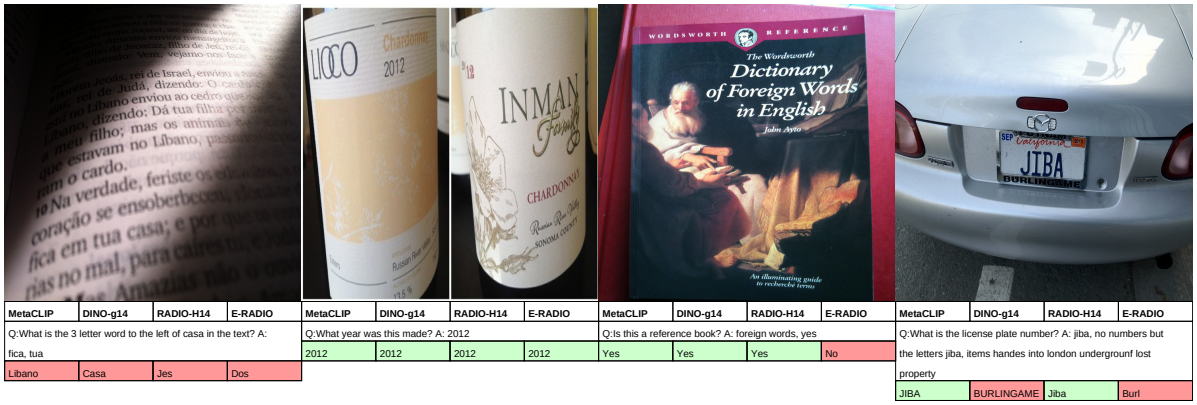
MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What number is on the player's jersey? A: 22			
22	22	22	22

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What is the time? A: 5:42, 5:41, 8:00, 5:40			
11:00	11:00	11:55	11:00
Q:What brand of watch is that? A: unanswerable, audemars, ap, af			
Tissot	Tissot	Tudor	Rolex

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Who is at the center of all of this? A: agile experience design makeup, bryan owens, alexa curtis, mahou			
Aithell	Man	Chris O'Leary	Owens

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Who was the photographer? A: philippe molitor, philippe molitar, no, phillipe meltow, l, clardajne, phillipe molida			
Philippe molitor	Philippe molitor	Philippe molitor	Philippe molitor

Figure 11. Sample questions from the TextVQA [55] dataset and their answers from our LLaVA models, using various image encoders. Answers are painted green when they match the ground truth, pink otherwise.



MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What is the 3 letter word to the left of casa in the text? A:			
fica, tua			
Libano	Casa	Jes	Dos

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What year was this made? A: 2012			
2012	2012	2012	2012

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Is this a reference book? A: foreign words, yes			
Yes	Yes	Yes	No

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What is the license plate number? A: jiba, no numbers but			
the letters jiba, items handes into london undergrounf lost			
property			
JIBA	BURLINGAME	Jiba	Burl

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What is the alcohol content? A: 9%, 2009, 9.0, 9.0% alc/vol.,			
9, smashed pumpkin, 9.0%, lego			
9.0%	9.0%	9.0%	9.0%
Q:What is the name of this ale? A: smashed pumpkin,			
shipyard, shipyard smashed pumpkin			
Smashed pump	Shipyard smash	Shipyard	Shipyard smash

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What is the beer brand front center? A: coors light, coors,			
coors light , secret			
Coors light	Coors light	Coors light	Coors light
Q:What is the company name to the left of the coors logo? A:			
safeway, calculator, safeway			
Coors	Coors	Safeway	Pg&e

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:Who is usa today's bestselling author? A: cathy williams			
Cathy williams	Cathy williams	Cathy williams	Cathy williams
Q:What is the name of this bestselling books? A: secrets of a			
ruthless tycoon, cathy williams, secret of ruthless tycoon			
Cathy williams	Harlequin Prese	Cathy williams	Harlequin prese

MetaCLIP	DINO-g14	RADIO-H14	E-RADIO
Q:What is this food place selling? A: bratwurst, wurst,			
krainerwurst, burenwurst, hotdogs, krainerwurst and			
burenwurst, krainerwurst burenwurst, krainerwurst,			
burenwurst			
Hot dog	Hot dog	Fran Debrezine	Hot dog
Q:What is the top word on the sign on the left? A: krainerwurst			
Krainerwurst	Hot dog	Krainerwurst	Hot dog

Figure 12. Sample questions from the TextVQA [55] dataset and their answers from our LLaVA models, using various image encoders. Answers are painted green when they match the ground truth, pink otherwise.

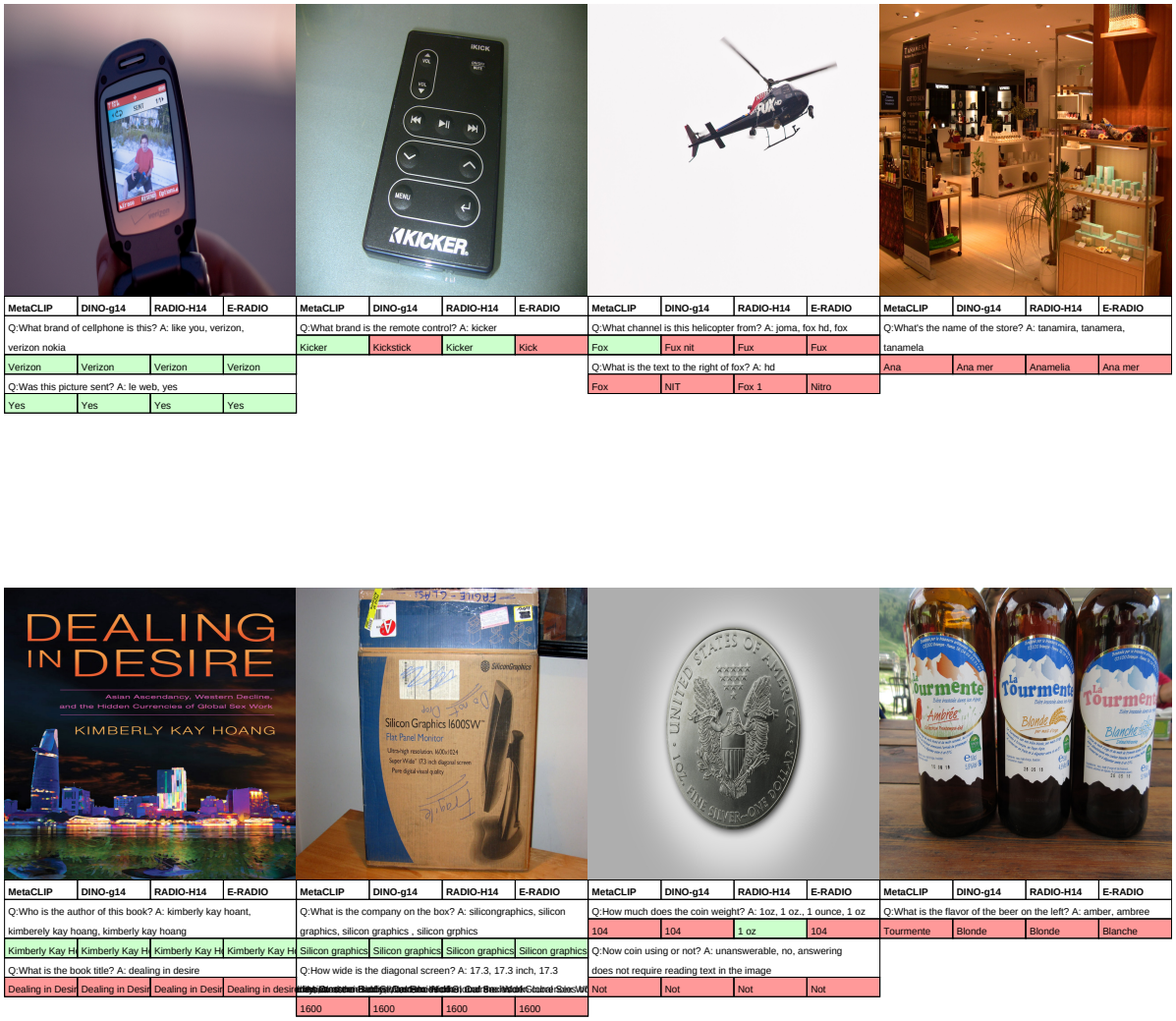


Figure 13. Sample questions from the TextVQA [55] dataset and their answers from our LLaVA models, using various image encoders. Answers are painted green when they match the ground truth, pink otherwise.