# FOVEAL RE-ENCODING FOR FROZEN BACKBONES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Human vision is foveated, resolving detail only around the point of gaze and moving it wherever the task needs acuity, whereas a vision foundation model encodes one global view into a low-resolution grid and can look closer only by pushing a crop through the backbone again. Feature upsamplers do not change this, since they are trained to reproduce the backbone's encoding of the same field of view sampled more finely, which leaves the scale of analysis where the single global pass put it. We give a frozen DINOv3 a fovea. Conditioned on a point query, foveal re-encoding (FoRE) runs a lightweight modulator on the coarse map and the queried box of that same global view, predicting the features the backbone would produce by re-encoding the box at native resolution, in 2.8 milliseconds per query and almost six times more cheaply. It reads no pixel outside the view an image-guided upsampler receives, and what it learns is a remapping of that view's encoding rather than a reconstruction from pixels, so it differs from an upsampler in what it predicts rather than in what it reads. Distillation against the backbone's own re-encodings needs no labels or backbone updates. Across 25 datasets, more than fifteen readouts, and three domains, the predicted features are the strongest source apart from the teacher almost everywhere and lead every upsampling baseline by wide margins under in-context matching, with the largest gains where one wide-field encoding collapses most structure. On per-patch readouts they surpass even the re-encoding they were distilled from, an effect we trace to variance reduction and summarize in an averaging law for when it appears and reverses.

## 1 INTRODUCTION

The eye does not analyze a scene at a single resolution. Acuity falls off sharply with distance from the point of gaze, and a saccade brings the fovea to whatever the task needs to see clearly, so a fixed budget of photoreceptors buys detail where it matters and coarse context everywhere else. Vision foundation models such as DINOv2 (Oquab et al., 2023) and DINOv3 (Siméoni et al., 2025) offer no counterpart, since they resize an image to a fixed global view and encode it once into a feature grid far below the pixel grid, committing every location to one wide-field scale of analysis. The loss is most conspicuous in remote sensing, where buildings, roads, and vehicles remain visible in the view yet collapse into a handful of patch tokens. The constraint tightens under a fixed compute budget, as on a satellite or a mobile robot, where the scene dwarfs the encoder, so re-encoding all of it at the resolution a few regions demand is out of reach, though those regions are usually all the task needs.

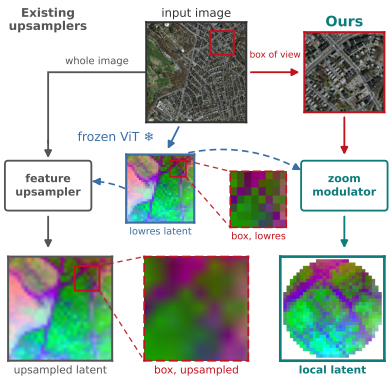


Figure 1: **Motivation.** Both branches see one image and its coarse latent, which an upsampler resamples while our modulator re-encodes the marked box.

No remedy today supplies a fovea, as Figure 1 illustrates. Feature upsamplers (Kopf et al., 2007; Fu et al., 2024) refine the coarse map under image guidance, and the learned ones (Wimmer et al., 2026; Couairon et al., 2026) train on the backbone's own encoding of the same field of view sampled more finely. They therefore add detail at the scale of analysis the wide-field pass has fixed, not at that of



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the region itself. What they forgo is not pixels. Re-encoding a crop of that same global view, which holds nothing an upsampler is not handed, attains a patch cosine of 0.986 with the native-resolution encoding of the region, while interpolating the coarse map reaches 0.635. Re-encoding, however, spends a full forward pass on every glance, and retraining the encoder for flexible resolutions (Beyer et al., 2023; Dehghani et al., 2023) presupposes control of pretraining and scales with area. The eye provides what is missing, cheap high-acuity features at a chosen location on an intact global percept.

This paper builds that mechanism for a frozen backbone, as sketched in Figure 2. A point query, consisting of a normalized center and a normalized radius, plays the role of gaze. Active and glimpse-based vision have long placed resolution where a task needs it (Bajcsy, 1988; Mnih et al., 2014), but by training an encoder end to end rather than retrofitting a fovea onto a frozen checkpoint. A modulator reads the global feature map together with the queried box of that view and predicts the features the backbone would produce by re-encoding the box at native resolution. It reads no pixel an image-guided upsampler does not, yet contributes a learned re-encoding rather than a filter, and distillation against this zoom teacher supervises it without labels or backbone updates. Feature change is structurally confined to a disc around the query, keeping the untouched global map valid as periphery, and once the view is encoded each foveation costs 2.8 milliseconds, almost six times cheaper than re-encoding. A fixed budget therefore buys many glances rather than one. Figure 7 shows this on one aerial tile, where the places worth a closer look cover a small part of the scene, so one wide pass settles where they are and a few queries resolve them without re-encoding the rest.

We evaluate it as a drop-in feature source under the readouts dense tasks already run on frozen features (Amir et al., 2021; Balazevic et al., 2023), across 25 datasets, more than fifteen readouts, and three domains. It is the best source apart from the teacher on ten of eleven segmentation readouts, and its margin over upsampling baselines is widest where the wide-field encoding destroys the most structure, correlating at 0.82 with the headroom a dataset leaves. The recipe reproduces the ordering on three further frozen encoders. Readouts that decide patch by patch even prefer the prediction to the re-encoding it imitates, raising five-shot in-context segmentation from 0.432 to 0.491 mean intersection over union (mIoU) and depth accuracy on 18 of 21 datasets. Distilled students sometimes overtake their teachers (Hinton et al., 2015; Ranzinger et al., 2024), and here the cause is variance reduction, since training over tens of thousands of images shrinks what a re-encoding does not share with nearby inputs. It is not the per-image artifact that registers and test-time decomposition remove inside one encoding (Darcet et al., 2024; Yang et al., 2024a), as Appendix A details. We organize this into an averaging law that predicts where the reversal appears and reverses.

- We introduce foveal re-encoding (FoRE), an interface that addresses a frozen vision foundation model by a location and a scale rather than by a resolution, together with the distillation target that gives the interface its content, the encoding the backbone itself would produce once the queried region fills the frame, learned from unlabeled images alone.
- We separate what re-encoding provides from what resampling reaches, and give an averaging law for when a distilled re-encoding beats its teacher and when it does not.
- We show that the interface is practical rather than diagnostic, since one encoding of the view serves a stream of glances at a small cost each, which is cheaper than re-encoding from the second glance onward and than upsampling the whole frame over a fovea's range.

## 2 Foveal Re-Encoding

### 2.1 Problem setting and geometry

Let $f$ denote a frozen vision transformer that maps a square image to a grid of patch features. Every image is center-cropped to a square canonical frame to which all normalized coordinates refer. The global view $I^g$ is that frame resized to a fixed view resolution, which the backbone encodes in a single pass, and its encoding $F = f(I^g)$ is the coarse map, a patch grid whose resolution lies far below the pixel grid. Throughout the paper superscripts name image kinds and subscripts index patches, with $i$ running over the coarse grid and $j$ over the teacher grid, and the backbone and the view are instantiated in Section 3. A point query

$$q = (\mathbf{c}, r), \qquad \mathbf{c} \in [0,1]^2, \quad r \in [0.10,\, 0.25], \tag{1}$$

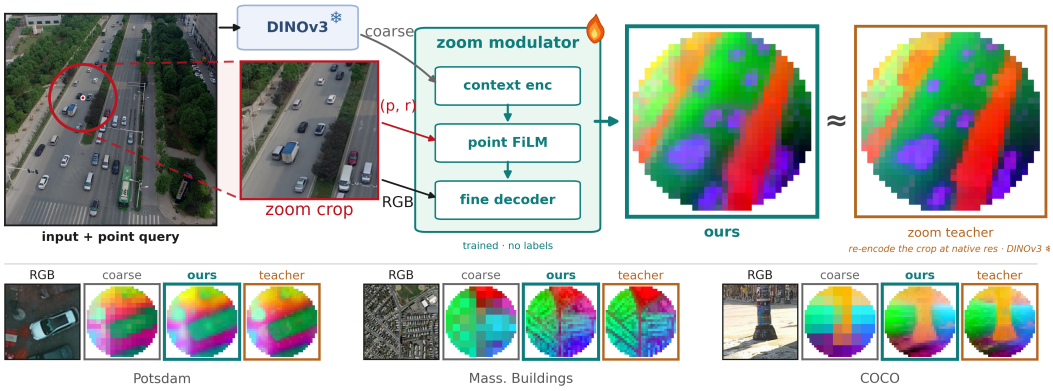


Figure 2: **Method overview.** The frozen backbone encodes the global view once, and a lightweight modulator, conditioned on the point query and the queried box of that view, turns the coarse map into a prediction of the zoom teacher, the features the backbone would produce by re-encoding the crop at native resolution. The strip shows joint PCA of the coarse map, our prediction, and the teacher for three queries. Our outputs are discs because feature change is confined to the query support.

pairs a normalized center with a normalized radius, sampled log-uniformly over that range during training and drawn from the same range at evaluation. The query selects the square box

$$B(q) = \{\mathbf{x} : \|\mathbf{x} - \mathbf{c}\|_\infty \le (1+\beta)\, r\}, \tag{2}$$

in which $\beta = 0.15$ is a context margin. The box is square because the encoder consumes square inputs, and which subset of it a query may change is a separate question, settled in Section 2.2.

Two crops of that box play distinct roles,

$$T = f(I^z), \qquad Y = \Phi_\theta(F, q, I^v). \tag{3}$$

The zoom teacher $T$ re-encodes the box at native resolution, which means that $I^z$ is cut from the canonical image before any view resizing and only afterwards resized to the view resolution. The student modulator $\Phi_\theta$ instead reads $I^v$, the same box cut from $I^g$ and resized in the same way, so that it never receives a pixel outside the view the backbone has already encoded, and its prediction $Y$ lives on the teacher patch grid. Wherever the canonical frame is wider than the view the teacher does resolve detail the view discarded, bounded by the query magnification because the teacher crop is itself resized to the view resolution, and elsewhere the two crops carry the same pixels. The teacher remains equally far from the interpolated coarse map in both regimes, as Appendix B verifies.

Since $B(q)$ spans a fraction $2(1+\beta)\, r$ of the frame and is re-encoded at the same resolution as the view, the query range of Equation (1) corresponds to magnifications

$$\frac{1}{2(1+\beta)\, r} \in [1.7,\ 4.3]. \tag{4}$$

How much of the teacher is reachable from $I^v$ at all? A control teacher $T^v = f(I^v)$ answers this by applying the identical crop-and-encode procedure to the student's own pixel input, measuring what a backbone-equivalent computation recovers from those pixels. In our main instantiation, measured on held-out images whose canonical frames are twice as wide as the view so that the two crops genuinely differ, the cosine between $T^v$ and $T$ is 0.986, a level specific to the zoom range. Almost the entire benefit of looking closer therefore comes from re-encoding the region at the scale of the query rather than from resolving new pixels, which Appendix M confirms on the readouts. It is that scale-conditioned re-encoding, which a transformer performs differently when a structure occupies a few patches and when it fills the frame, that the student learns. In the same measurement our deployed model attains a patch cosine of 0.847 against 0.635 for bilinear sampling of $F$, closing three fifths of the distance to the control teacher. Which of its two inputs carries that content is measurable in the same way, and withholding the coarse map leaves the modulator below plain interpolation while withholding the crop pixels costs almost nothing, so a glance is a query-local remapping of the wide-field encoding rather than a regression of features from pixels, which is what lets it cost a fraction of a forward pass, as Appendix O reports.

### 2.2 ZOOM MODULATOR

**Support.** Which region may a query change? We take the support to be a sublevel set of a distance $\rho$ on the canonical frame,

$$S(q) = \{\mathbf{x} : \rho(\mathbf{x}, \mathbf{c}) \le \kappa r\}, \tag{5}$$

where $\kappa = 1.10$ is a support multiplier, and we ask which $\rho$ the query itself justifies. The support is a function of the query alone, so rotating the image about $\mathbf{c}$ rotates the support with it, while $q$, carrying nothing but a center and a scale, is unchanged by that rotation, so $S(q)$ must be invariant under every rotation about $\mathbf{c}$. A norm whose balls are invariant under all rotations is a multiple of the Euclidean norm, so isotropy alone fixes $\rho(\mathbf{x}, \mathbf{c}) = \|\mathbf{x} - \mathbf{c}\|_2$ and makes $S(q)$ a disc of radius $\kappa r$, whereas the $\ell_1$ diamond and the $\ell_\infty$ box of Equation (2) are invariant only under the rotations that permute the axes. An anisotropic support would resolve detail farther along some image directions than others for a reason present in neither the query nor the image, and it would blunt the scalar $r$, since a box of half-side $\kappa r$ reaches $\sqrt{2}\,\kappa r$ diagonally and so supervises an extent that differs by 41 percent with direction. The same reasoning makes a biological fovea radial.

**Modulated global map.** Inside $\Phi_\theta$, a context encoder of stacked depthwise convolutional blocks processes $F$ strictly locally into a context map $H$. A point-query encoder embeds $q$ through a rasterized support mask and Fourier features of the center and radius, and a spatial feature-wise linear modulation layer (Perez et al., 2018), zero-initialized so training starts from the identity, injects that embedding into $H$ to yield a modulated context map $\tilde{H}$. The frozen map is updated only inside the support,

$$M_i = F_i + s_i\, \xi(\tilde{H}_i), \qquad s_i = \mathbf{1}\big[\mathbf{x}_i \in S(q)\big], \tag{6}$$

with $i$ indexing patches of the coarse grid, $\mathbf{x}_i$ the canonical coordinate of patch $i$, and $\xi$ a learned projection, so that the indicator $\mathbf{1}[\cdot]$ makes the hard gate $s_i$ zero all change outside the disc.

**Fine decoder.** A coordinate-conditioned fine decoder renders $Y$ on the teacher patch grid. With $\mathbf{x}_j$ the canonical coordinate of target patch $j$,

$$Y_j = M(\mathbf{x}_j) + s_j\, \psi\big(\tilde{H}_{\mathcal{N}(j)}, \mathbf{x}_j, q, I^v\big), \tag{7}$$

where $M(\mathbf{x}_j)$ bilinearly samples the modulated map, $\mathcal{N}(j)$ its $3\times3$ neighborhood, $I^v$ enters through a convolutional encoder, $s_j = \mathbf{1}[\mathbf{x}_j \in S(q)]$ applies the support test on the fine grid, and $\psi$ is a per-patch multilayer perceptron. This decoder is the dominant lever, as Appendix O shows.

**Structural locality.** Locality follows from Equations (6) and (7) rather than training. The gates zero all change outside $S(q)$, the context encoder has no global pathway, and the decoder reads only local neighborhoods, so distant regions cannot be altered, and everything the model can modulate lies inside the teacher crop because $\kappa \le 1 + \beta$. The gates are binary rather than graded so Equation (6) leaves the map exactly frozen outside the disc, which keeps the global percept valid with no term in the objective to enforce it. On the teacher grid the residual vanishes outside the disc as well, while the bilinear sample $M(\mathbf{x}_j)$ still reaches a modulated token within one coarse patch of the boundary, so exactness on that grid holds outside a band of that width, which no reported number touches because every readout scores supported positions only. The disc leaves the corners of $B(q)$ unsupervised, which matters only where a support must tile an image, and the box variant covering them costs per-patch matching accuracy, so the disc remains the default, as Appendix T reports.

### 2.3 SELF-SUPERVISED DISTILLATION OBJECTIVE

The objective is a single distillation term, the mean cosine distance to the teacher over supported patches,

$$\mathcal{L} = \frac{\sum_j s_j\big(1 - \cos(Y_j, T_j)\big)}{\sum_j s_j}, \tag{8}$$

with the teacher under stop-gradient. Training reads only unlabeled images and the backbone receives no gradient, so at test time the view is encoded once and each further query costs only the modulator and the decoder. Corpus and optimization details appear in Section 3 and Appendix B.

# 3 EXPERIMENTAL SETUP

**Instantiation.** Our main experiments instantiate the method on DINOv3 ViT-L/16 (Siméoni et al., 2025) with a $512^2$ global view, whose coarse map is a $32 \times 32$ grid of 1024-dimensional patch features, and teacher crops are re-encoded at that same resolution. The modulator carries 9.2M trainable parameters and trains for 24k steps with two queries per image, while the backbone-generality study of Section 4.2 retrains it unchanged on DINOv2 ViT-L/14 (Oquab et al., 2023) and Franca ViT-L/14 (Venkataramanan et al., 2026) with $448^2$ views and on RADIO ViT-L/16 (Ranzinger et al., 2024) with $512^2$ views. Architecture and optimization details are given in Appendix B.

**Training data.** The training corpus holds 15,000 unlabeled images, 6,000 COCO (Lin et al., 2014) photographs and 9,000 aerial and satellite images from six public sources, namely OpenEarthMap (Xia et al., 2023), xView2 (Gupta et al., 2019), LoveDA (Wang et al., 2021), DeepGlobe (Demir et al., 2018), SpaceNet1 (Van Etten et al., 2018), and SpaceNet7 (Van Etten et al., 2021). No annotation reaches it. The corpus shares no image with any evaluation set, few-shot supports included. Every evaluation set is split by image into a fitting portion supplying few-shot supports and a held-out portion carrying every reported number, as Appendix B specifies.

**Feature sources under comparison.** The sources compared throughout are the coarse map itself, denoted lowres, six upsamplers of it, namely bilinear interpolation, joint bilateral upsampling (JBU) (Kopf et al., 2007), and the learned FeatUp (Fu et al., 2024), NAF (Chambon et al., 2026), AnyUp (Wimmer et al., 2026), and JAFAR (Couairon et al., 2026), each matched to our backbone, our student, and the zoom teacher. Every source but the teacher starts from that coarse map and sees no pixel outside the view, whereas the teacher spends a backbone pass per query and cuts its crop at native resolution, which makes it a cost reference, not a competing method.

**Evaluation suite.** Throughout the paper a readout is a fixed, lightweight protocol mapping frozen patch features to task predictions, so varying only the feature source isolates feature quality. The remote sensing suite covers 21 datasets spanning land cover, buildings, roads, and small objects, the natural-image suite covers PASCAL VOC (Everingham et al., 2010), ADE20K (Zhou et al., 2017), and COCO-Stuff (Caesar et al., 2018), and the indoor suite uses NYUv2 (Silberman et al., 2012) with real metric depth, surface normals, and forty-class segmentation. Table 5 names each with its reference. Query crops have radii in $[0.10, 0.25]$ and come from the held-out portion of each dataset. The readouts comprise five few-shot segmentation classifiers at one and five shots, namely prototype matching (Snell et al., 2017), self-support prototypes (Fan et al., 2022), FSSDINO (Zakir & Ho, 2026), INSID3 (Cuttano et al., 2026), and FROST (Park, 2026). They are joined by linear-probe and trained-decoder segmentation, linear and decoder-based depth and surface-normal probing, patch retrieval, unsupervised clustering and segmentation, faithfulness of the class-activation map (CAM), crop classification, temporal label propagation, and semantic correspondence, each named with its reference in Appendix B alongside the seeds and split protocol. Three metrics need a definition. The depth threshold accuracy $\delta_1$ is the fraction of pixels whose predicted depth lies within a factor of 1.25 of the target, and J&F is the mean of region similarity and contour accuracy. The CAM average gain is the rise in classifier confidence for the predicted class when crop features are pooled with the map as weights rather than uniformly, as a percentage of the headroom to full confidence. Significance uses two-sided Wilcoxon signed-rank tests over datasets.

**Protocols.** All main-text results follow a crop-grid protocol that extracts features for identical query crops from every source and runs every readout on the cached features with the same code and seeds. Our ports retain every published hyperparameter and differ from the original repositories only as Appendix B lists, and since every source runs under the identical port those differences cannot bias the comparison between sources. They do put absolute values on a different footing from the numbers published for these classifiers, which predict on a finer full-image grid and threshold at pixel resolution, so every ordering we report is between feature sources rather than between methods. Every dataset contributes two query crops on each of 120 images, half of them fitting whatever a readout needs and half carrying the reported numbers. Scores are computed only on the patches inside the query support, the region our method may modify, so all sources are scored on identical patches, and every few-shot number averages three episode seeds shared across sources.

Table 1: **Few-shot and probe segmentation.** Mean mIoU over the 21 remote sensing datasets, with every readout defined in Section 3. Bold marks the best source excluding the teacher, and the two right columns count the datasets where our student beats the strongest upsampler and the teacher.

| Readout | Shots | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher | vs. upsampler | vs. teacher |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Prototype | 1 | 0.392 | 0.393 | 0.372 | 0.391 | 0.402 | 0.381 | **0.404** | **0.404** | 0.438 | 10/21 | 4/21 |
| | 5 | 0.472 | 0.475 | 0.448 | 0.471 | 0.486 | 0.465 | 0.490 | **0.502** | 0.539 | 15/21 | 4/21 |
| Self-support | 1 | 0.382 | 0.375 | 0.356 | 0.376 | 0.383 | 0.358 | 0.389 | **0.408** | 0.441 | 12/21 | 7/21 |
| | 5 | 0.453 | 0.451 | 0.431 | 0.445 | 0.460 | 0.435 | 0.467 | **0.485** | 0.523 | 15/21 | 5/21 |
| FSSDINO | 1 | 0.388 | 0.396 | 0.371 | 0.391 | **0.403** | 0.382 | 0.402 | 0.393 | 0.426 | 9/21 | 5/21 |
| | 5 | 0.462 | 0.471 | 0.440 | 0.467 | 0.480 | 0.460 | 0.483 | **0.493** | 0.523 | 12/21 | 6/21 |
| INSID3 | 1 | 0.340 | 0.322 | 0.313 | 0.316 | 0.324 | 0.278 | 0.324 | **0.421** | 0.378 | 18/21 | 14/21 |
| | 5 | 0.368 | 0.345 | 0.326 | 0.340 | 0.347 | 0.295 | 0.345 | **0.491** | 0.432 | 21/21 | 13/21 |
| FROST | 1 | 0.187 | 0.174 | 0.131 | 0.155 | 0.166 | 0.168 | 0.149 | **0.399** | 0.379 | 21/21 | 13/21 |
| | 5 | 0.207 | 0.216 | 0.130 | 0.183 | 0.194 | 0.182 | 0.164 | **0.512** | 0.488 | 21/21 | 13/21 |
| Linear probe | – | 0.540 | 0.560 | 0.526 | 0.552 | 0.574 | 0.565 | 0.577 | **0.583** | 0.605 | 13/21 | 7/21 |

Table 2: **Regression, retrieval, and unsupervised readouts.** Means over the 21 remote sensing datasets, except CAM and classification, which need a class label per crop and average 19, and temporal propagation, a single time-series benchmark on SpaceNet7. Arrows give the direction of improvement, bold marks the best source excluding the teacher, and Section 3 defines the metrics.

| Evaluation | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| Depth $\delta_1$ (linear head) ↑ | 0.292 | 0.297 | 0.286 | 0.297 | 0.297 | 0.290 | 0.292 | **0.323** | 0.311 |
| Depth RMSE (linear head) ↓ | 2.212 | 2.207 | 2.275 | 2.209 | 2.223 | 2.246 | 2.229 | **2.037** | 2.137 |
| CAM average gain ↑ | 32.7 | 29.5 | 29.1 | 30.0 | 29.5 | 25.7 | 29.8 | **48.2** | 32.1 |
| Crop classification accuracy ↑ | 0.826 | 0.828 | 0.825 | 0.828 | 0.828 | 0.826 | 0.829 | **0.841** | 0.824 |
| Trained decoder mIoU ↑ | 0.567 | 0.573 | 0.552 | 0.561 | 0.583 | 0.584 | 0.594 | **0.601** | 0.613 |
| Trained decoder depth $\delta_1$ ↑ | 0.359 | 0.361 | 0.350 | 0.357 | 0.358 | 0.362 | 0.360 | **0.390** | 0.413 |
| Retrieval precision at 1 ↑ | 0.616 | 0.644 | 0.594 | 0.643 | 0.652 | 0.623 | 0.631 | **0.659** | 0.696 |
| Clustering normalized mutual information ↑ | 0.244 | 0.252 | 0.234 | 0.252 | 0.268 | 0.255 | 0.269 | **0.279** | 0.299 |
| STEGO unsupervised mIoU ↑ | 0.281 | 0.263 | 0.264 | 0.257 | 0.257 | 0.268 | 0.258 | **0.286** | 0.275 |
| Deep-spectral mIoU ↑ | 0.402 | 0.397 | 0.397 | 0.399 | 0.408 | 0.399 | **0.413** | **0.413** | 0.430 |
| Temporal propagation J&F ↑ | 0.239 | 0.293 | 0.322 | 0.286 | 0.302 | 0.308 | 0.330 | **0.352** | 0.398 |

# 4 RESULTS

## 4.1 MAIN RESULTS UNDER FROZEN READOUTS

Two regimes separate cleanly in Table 1, and they organize the rest of the paper. Under in-context matching readouts, which compare each query patch directly against support patches, the student is not merely the best non-teacher source on every readout but surpasses the teacher on all four of them, at 0.491 against 0.432 for five-shot INSID3 and 0.399 against 0.379 for one-shot FROST. Its lead over the strongest upsampler on FROST holds at a Wilcoxon $p$ below $10^{-6}$ for both shot counts, and its five-shot INSID3 advantage of 0.059 is significant at $p = 0.011$, while the remaining three matching rows are ahead without reaching significance. Under prototype-style and probe readouts, which average support patches into class statistics before classifying, the teacher recovers the lead, although the student is still the best non-teacher source on ten of the eleven rows. Both regimes reappear in Table 2. A shared linear head puts the student ahead of the teacher on depth threshold accuracy on 18 of 21 datasets at $p = 1 \times 10^{-4}$, on the root mean squared error (RMSE) on all 21 at $p < 10^{-6}$, and on CAM faithfulness on all 19, while a trained decoder on the same depth task flips the ordering back. Per-dataset tables and the significance analysis are in Appendices H and K.

## 4.2 GENERALIZATION ACROSS DOMAINS AND BACKBONES

These effects could be artifacts of pseudo-ground-truth targets or of one domain, and three tests rule that out. First, repeating the study on the four benchmarks outside remote sensing returns the pattern of Tables 1 and 2. Our model leads every source under both matching readouts, at 0.402 against the teacher's 0.353 for one-shot FROST and 0.470 against 0.408 for INSID3, and ahead of it on depth and CAM faithfulness, as Table 37 reports. The linear probe is the one row where the teacher recovers, as fitted readouts do in remote sensing. The aerial-only variant, for which photographs and indoor scenes are equally out of domain, tracks it closely and leads on depth and CAM.

Second, the advantage survives a switch to measured ground truth, since on the real metric depth of NYUv2 our model raises $\delta_1$ from the teacher's 0.218 to 0.227 and one-shot FROST on its forty measured classes gives 0.324 against 0.309, as Appendix Q reports.

Third, the pattern survives a change of backbone. Retraining the same modulator and recipe on frozen DINOv2, Franca, and RADIO ViT-L reproduces both regimes on ten datasets and six readouts, as Appendix E reports, so the effect follows the modulator, not the checkpoint or objective.

### 4.3 Queries chosen by a policy

Every query so far has been supplied to the model, so the open question is whether the interface can choose its own from one pass, and under in-context segmentation it can, since the reference images defining a class also rank the coarse map. We score 49 candidate regions by the cosine between their coarse features and a prototype pooled from the reference masks, re-encode the top eight, and leave the map bilinear elsewhere, as Appendix J details. The readout then runs unmodified over whole images from five datasets whose frames are 1.9 to 5.9 times wider than the view. Table 3 averages every scored class on the left and, on the right, the classes whose ground-truth components are small, a criterion fixed by annotation. The policy leads the coarse map in every cell, by 0.026 to 0.038 mIoU over all classes and 0.062 to 0.117 over the small ones, so the gain concentrates where one wide pass loses most. Eight re-encodings cost 11 milliseconds beyond the view and recover 59 to 82 percent of what all 49 buy there, at a sixth of the decoding.

Table 3: **Queries chosen by a policy.** Mean mIoU over five high-resolution datasets, scored on whole images, over every scored class and over the small-object classes alone.

| | All classes | | Small objects | |
|---|---|---|---|---|
| Readout | lowres | Ours | lowres | Ours |
| INSID3, 1 shot | 0.382 | 0.409 | 0.248 | 0.328 |
| INSID3, 5 shots | 0.472 | 0.501 | 0.284 | 0.401 |
| FROST, 1 shot | 0.412 | 0.438 | 0.226 | 0.288 |
| FROST, 5 shots | 0.508 | 0.546 | 0.313 | 0.417 |

### 4.4 Efficiency

Table 4 separates two ways of charging for a query. The amortized cost shares one backbone pass across sixteen queries from an image, the standalone cost charges every query a pass of its own, and the teacher has none to share, so the two coincide. Only the student combines re-encoding-level quality with interpolation-level cost, staying within a factor of two of bilinear sampling, 5.8 times cheaper than the teacher, and 4 to 11 times cheaper than the learned upsamplers, which process the entire map per image. Because one encoding serves any number of glances, the cumulative cost stays below the teacher from the second query onward and below every full-frame upsampler up to 184 queries, as Appendix G measures. The teacher row batches all sixteen crops into one pass, the choice least favorable to us, and Figure 7 shows what 2.8 milliseconds per query buys on one large tile encoded once.

Table 4: **Efficiency.** Milliseconds per point query on an RTX A6000.

| Cost (ms) ↓ | Amortized | Standalone |
|---|---|---|
| bilinear | 1.47 | 23.1 |
| JBU | 19.6 | 314 |
| FeatUp | 16.7 | 267 |
| NAF | 31.9 | 511 |
| AnyUp | 14.7 | 235 |
| JAFAR | 11.3 | 182 |
| Ours | 2.85 | 24.5 |
| Teacher | 16.6 | 16.6 |

### 4.5 Qualitative results

Figure 3 compares eight feature sources on three query crops, with sixteen further datasets in Appendix V. Two observations recur. First, no upsampler recovers structure the coarse map has lost, most strikingly where field boundaries and rows of buildings exist only in the two re-encoding columns. Second, the student reproduces the semantic structure of the teacher while its maps are visibly smoother, although smoothness by itself does not account for the advantage that follows. The three crops are overhead farmland, a dense residential block, and a crowd under umbrellas, so the pattern is not particular to one domain, though it is weakest on the photograph, whose structures the wide-field pass already resolves. What separates the two re-encoding columns from the other six is not the quality of the interpolation but the target it was fitted to, as Section 5.3 argues.

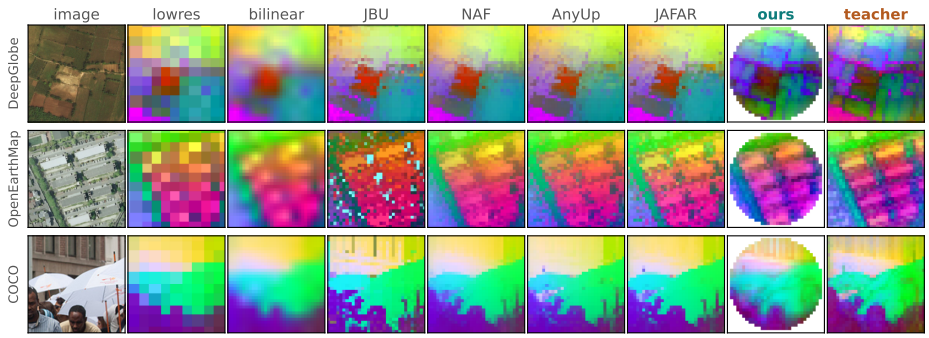


Figure 3: **Qualitative comparison.** Joint PCA of patch features, one query crop per row, computed over the eight sources shown so colors are comparable within a row. The teacher column is the zoom re-encoding, and our output is masked white outside the query support, which it leaves unchanged.

## 5 The Averaging Law

### 5.1 Canonicalization through distillation

Why should a student trained only to imitate its teacher surpass it? A modulator of fixed capacity fits one smooth function across tens of thousands of images, so rather than render each crop on its own terms it regresses the targets of similar inputs together, shrinking whatever varies idiosyncratically between them. The output is a canonicalized re-encoding, a variance-reduced version rather than a degraded copy, and the shrunk part behaves as noise for a readout that compares patches across images. The averaging therefore runs over neighborhoods of the input space rather than over repeated draws of one input, a reading Section 2.1 rules out by measuring how nearly the student's own pixels fix it, and Appendices L and N formalize the account, test it directly, and separate it from the competing ones. Two controls there cut against the obvious reading, since ensembling the teacher over photometric and sub-patch perturbations moves no readout past 0.002 mIoU, against the 0.020 to 0.059 the student gains, and the margin is largest where crops carry identical pixels.

### 5.2 Averaging capacity of the readout

Variance reduction trades a small bias for a large decrease in variance, and whether that pays off is decided downstream, so we group readouts by the internal averaging they perform. Matching readouts such as FROST and INSID3 whiten features and compare individual patches against support patches by winner-take-all voting, so the shrunk part propagates into decisions and the student's variance reduction is worth most, between $+5.0$ and $+13.6$ percent in Figure 4. Per-patch regression with a shared head, as in linear depth probing and CAM, passes that component through in the same way. Fitted readouts average over patches when they estimate class statistics or fit weights, as prototype classifiers, linear probes, retrieval over pooled statistics, and above all a trained decoder do, and that pooling removes the component without the student's help, so the teacher's extra information wins, by 2.0 to 5.3 percent for the probe, decoder, and retrieval and up to 7.7 percent for prototype-style classifiers. The linear probe and linear depth head have identical capacity yet land on opposite sides, which the law explains at the output rather than during fitting, since the shrunk part enters a regression estimate linearly whereas a classification argmax absorbs perturbations below its decision margin. Depth isolates this,

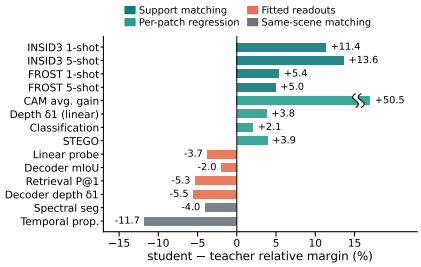


Figure 4: **Averaging law.** Relative margin of the student over the teacher for each readout of Tables 1 and 2, ordered by the internal averaging it performs. Table 31 tests every difference but temporal propagation. The CAM bar is truncated at the glyph, its printed value exact.

since the same features and targets flip from $+3.8$ percent under a linear head to $-5.5$ percent under a trained decoder in Figure 4, so only the readout's capacity changes.

### 5.3 Upsampling versus re-encoding

The best upsampler in Tables 1 and 2 improves on the plain coarse map by at most 0.037 mIoU on any segmentation readout. On three of the four in-context matching readouts it falls below the coarse map by between 0.013 and 0.021, gaining 0.009 only at five-shot FROST, whereas the student adds 0.198 over each dataset's strongest upsampler on FROST at one shot and 0.286 at five. Every source here reads the same coarse map and the same global view, ours included, since it too remaps that map rather than rebuilding features from pixels, so what separates them is the target they are fitted to, and Section 2.1 has already located almost the whole distance between the coarse map and a re-encoding in the narrowing of the field of view rather than in resolution. Consistent with this, our gain over the interpolation baselines tracks task headroom at Pearson correlation 0.82, measured between the teacher-minus-bilinear gap of a dataset and the student-minus-bilinear gain on it over the ten few-shot rows of Table 1 and the 21 datasets, with per-readout correlations from 0.67 to 0.90, so coarse land-cover tasks saturate while buildings, roads, vehicles, and small objects benefit most. The same relation holds on photographs at Pearson correlation 0.78 over 33 dataset-by-readout pairs, where a teacher that fails to beat interpolation leaves no gap to recover while our distance to the teacher barely moves between domains, as Appendix R details. That several upsamplers fall below the coarse map is not an artifact of our ports, and the ordering persists where every learned upsampler runs its authors' released checkpoint, as Appendices F and P show.

### 5.4 The same-scene exception

If the student wins by shrinking what a re-encoding does not share with nearby inputs, it must lose wherever that component carries the signal, which is the sharpest bound on where these features belong. Canonicalization moves every patch toward what its neighbors in the input space have in common, whereas a match confined to one scene turns on the opposite quantity, on what makes these patches unlike one another rather than unlike the patches of other scenes. Temporal label propagation matches patches across frames of one scene, where the teacher leads by a wide margin, 0.398 against 0.352 J&F on SpaceNet7 and 0.688 against 0.568 on DAVIS. On DAVIS it also falls furthest behind the strongest upsamplers, NAF at 0.603 and JAFAR at 0.594, though it stays above bilinear at 0.558. Deep-spectral segmentation, which partitions one crop by its internal affinity structure, shows the same reversed sign, and single-vector correspondence on SPair-71k, which averages on neither side of the match, leaves plain bilinear features the strongest source apart from the teacher, as Appendix S details. These reversals are predictions of the law rather than exceptions, and which regime a task falls in is fixed by whether the match crosses scenes, so a system can route it in advance.

## 6 Conclusion

We have given a frozen vision foundation model a fovea. FoRE keeps the coarse map as periphery and, for any queried location, predicts the features of a native-resolution re-encoding from that map and the queried box of the same view, self-supervised and structurally local. Across 25 datasets and more than fifteen readouts its features lead every upsampling alternative on matched inputs under in-context matching, and on per-patch readouts surpass even the re-encoding they imitate, which our averaging law explains. One global encoding becomes an interactive stream of high-acuity glances at 2.8 milliseconds each, a step toward active perception on frozen foundation models.

**Limitations and outlook.** Four limits bound the claims. The teacher's pixel advantage measured in Section 2.1 persists at the largest magnifications wherever fine detail cannot be inferred from context, as on iSAID small objects. The trained range covers magnifications of 1.7 to 4.3, so a task demanding more must train wider or reapply the model to its own output, which Appendix U shows stays self-consistent for a second step, though only qualitatively. The evaluation relies on crop-level protocols for breadth, so end-to-end validation inside published pipelines remains future work. On real NYUv2 surface normals every source lies within two degrees of mean angular error of every other, so we report them for completeness and base no claim on them, as Appendix Q records. Richer policies than the one Section 4.3 deploys are the natural continuation.

### AI USE STATEMENT

Generative AI tools were used to assist with drafting and revising the manuscript. Specifically, they were used to improve the clarity, structure, and presentation of the text during the writing and editing process. The authors reviewed and verified all AI-assisted content and take full responsibility for the final manuscript.

### REPRODUCIBILITY STATEMENT

Everything needed to reproduce the results is specified in the paper. Section 2 defines the geometry, the architecture, and the training objective, Section 3 names the backbone, the view resolution, the training corpus, and the feature sources under comparison, and Appendix B lists the optimizer settings, the layer widths, the query-sampling ranges, the caching and split protocol, and the exact respects in which our ports of the published readouts differ from the original repositories. The per-dataset numbers behind every averaged row of the main text are tabulated in Appendices H and R. Training reads unlabeled images only and never updates the backbone, so a single frozen public checkpoint together with the listed hyperparameters determines the model. We will release code, trained modulator checkpoints, and the full evaluation suite.

## REFERENCES

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021.

Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988.

Ivana Balazevic, David Steiner, Nikhil Parthasarathy, Relja Arandjelović, and Olivier Henaff. Towards in-context scene understanding. *Advances in Neural Information Processing Systems*, 36: 63758–63778, 2023.

Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14496–14506. IEEE, 2023.

Adrian Boguszewski, Dominik Batorski, Natalia Ziemba-Jankowska, Tomasz Dziedzic, and Anna Zambrzycka. Landcover. ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1102–1110. IEEE, 2021.

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1209–1218. IEEE, 2018.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Loïck Chambon, Paul Couairon, Eloi Zablocki, Alexandre Boulch, Nicolas Thome, and Matthieu Cord. Naf: Zero-shot feature upsampling via neighborhood attention filtering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26604–26613, 2026.

Yu Chen, Yao Wang, Peng Lu, Yisong Chen, and Guoping Wang. Large-scale structure from motion with semantic constraints of aerial images. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 347–359. Springer, 2018.

Paul Couairon, Loick Chambon, Louis Serrano, Jean-Emmanuel Haugeard, Matthieu Cord, and Nicolas Thome. Jafar: Jack up any feature at any resolution. *Advances in Neural Information Processing Systems*, 38:99265–99292, 2026.

Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers, Carlo Masone, and Stefan Roth. INSID3: Training-free in-context segmentation with DINOv3. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, volume 2024, pp. 2632–2652, 2024.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36:2252–2274, 2023.

Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 172–17209. IEEE, 2018.

Patrick Ebel, Mounia El Baz, Junjue Wang, Weihao Xuan, Heli Qi, Zhuo Zheng, Naoto Yokoya, Junghwan Park, Jaewan Park, Arthur Elskens, et al. Artificial intelligence for earthquake response: Outcomes and insights from a global spaceborne rapid mapping challenge. *IEEE Geoscience and Remote Sensing Magazine*, 2026.

Francisco Escobar. Satellite images of water bodies. Kaggle dataset, Sentinel-2 imagery with water masks, 2021. URL `https://www.kaggle.com/datasets/franciscoescobar/satellite-images-of-water-bodies`.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Qi Fan, Wenjie Pei, Yu-Wing Tai, and Chi-Keung Tang. Self-support few-shot semantic segmentation. In *European conference on computer vision*, pp. 701–719. Springer, 2022.

Stephanie Fu, Mark Hamilton, Laura E Brandt, Axel Feldmann, Zhoutong Zhang, and William Freeman. Featup: A model-agnostic framework for features at any resolution. In *International Conference on Learning Representations*, volume 2024, pp. 45324–45350, 2024.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International conference on machine learning*, pp. 1607–1616. PMLR, 2018.

Ritwik Gupta, Bryce Goodman, Nirav Patel, Richard Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howard Choset, and Matthew Gaston. Creating xbd: A dataset for assessing building damage from satellite imagery. 2019.

Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*, 2022.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *European conference on computer vision*, pp. 372–391. Springer, 2022.

Haiwen Huang, Anpei Chen, Volodymyr Havrylov, Andreas Geiger, and Dan Zhang. Loftup: Learning a coordinate-based feature upsampler for vision foundation models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9913–9923. IEEE, 2025.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on geoscience and remote sensing*, 57(1):574–586, 2018.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Johannes Kopf, Michael F Cohen, Dani Lischinski, and Matt Uyttendaele. Joint bilateral upsampling. *ACM Transactions on Graphics (ToG)*, 26(3):96–es, 2007.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. *arXiv preprint arXiv:2305.13310*, 2023.

Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS journal of photogrammetry and remote sensing*, 165:108–119, 2020.

Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8354–8365. IEEE, 2022.

Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv preprint arXiv:1908.10543*, 2019.

Volodymyr Mnih. *Machine learning for aerial image labeling*. University of Toronto (Canada), 2013.

Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. *Advances in neural information processing systems*, 27, 2014.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Junghwan Park. Frost: Training-free few-shot segmentation with frozen features and nonparametric statistics. *arXiv preprint arXiv:2606.31136*, 2026.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Roberson Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021.

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12490–12500. IEEE, 2024.

Caleb Robinson, Kolya Malkin, Nebojsa Jojic, Huijun Chen, Rongjun Qin, Changlin Xiao, Michael Schmitt, Pedram Ghamisi, Ronny Hänsch, and Naoto Yokoya. Global land-cover mapping with weak supervision: Outcome of the 2020 ieee grss data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:3185–3199, 2021.

Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez, and Uwe Breitkopf. The isprs benchmark on urban object classification and 3d building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1:293–298, 2012.

Zhenfeng Shao, Ke Yang, and Weixun Zhou. Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset. *Remote Sensing*, 10(6):964, 2018.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. Does knowledge distillation really work? *Advances in neural information processing systems*, 34: 6906–6919, 2021.

Xin-Yi Tong, Gui-Song Xia, and Xiao Xiang Zhu. Enabling country-scale land cover mapping with meter-resolution satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196: 178–196, 2023.

Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018.

Adam Van Etten, Daniel Hogan, Jesus Martinez Manso, Jacob Shermeyer, Nicholas Weir, and Ryan Lewis. The multi-temporal urban development spacenet dataset. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6394–6403. IEEE, 2021.

Shashanka Venkataramanan, Valentinos Pariza, Mohammadreza Salehi, Lukas Knobel, Elias Ramzi, Spyros Gidaris, Andrei Bursuc, and Yuki M Asano. Franca: Nested matryoshka clustering for scalable visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10533–10544, 2026.

Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021.

Thomas Wimmer, Prune Truong, Marie-Julie Rakotosaona, Michael Oechsle, Federico Tombari, Bernt Schiele, and Jan Eric Lenssen. Anyup: Universal feature upsampling. In *International Conference on Learning Representations*, volume 2026, pp. 140700–140720, 2026.

Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6243–6253. IEEE, 2023.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European conference on computer vision*, pp. 432–448. Springer, 2018.

Jiawei Yang, Katie Z Luo, Jiefeng Li, Congyue Deng, Leonidas Guibas, Dilip Krishnan, Kilian Q Weinberger, Yonglong Tian, and Yue Wang. Denoising vision transformers. In *European Conference on Computer Vision*, pp. 453–469. Springer, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in neural information processing systems*, 37:21875–21911, 2024b.

Hussni Mohd Zakir and Eric Tatt Wei Ho. Revealing the semantic selection gap in dinov3 through training-free few-shot segmentation. *arXiv preprint arXiv:2602.07550*, 2026.

Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. *arXiv preprint arXiv:1905.12886*, 2019.

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)*, pp. 5122–5130. IEEE, 2017.

# A Related Work

**Frozen foundation features for dense prediction.** Self-supervised transformers such as DINO (Caron et al., 2021), DINOv2 (Oquab et al., 2023), and DINOv3 (Siméoni et al., 2025) support segmentation, depth, and correspondence without fine-tuning (Amir et al., 2021; Balazevic et al., 2023), read out by lightweight protocols such as linear probes, prototype classifiers (Snell et al., 2017), self-support matching (Fan et al., 2022), in-context nearest neighbors (Liu et al., 2023), and unsupervised correspondence distillation (Hamilton et al., 2022; Melas-Kyriazi et al., 2022). Such features carry per-image artifacts, which registers (Darcet et al., 2024) and test-time artifact decomposition (Yang et al., 2024a) remove within a single encoding, whereas the variance reduction we analyze in Section 5 arises across an entire corpus and is therefore of a different kind.

**Feature upsampling.** Joint bilateral upsampling (Kopf et al., 2007) sharpens a coarse feature map using the image as guidance, and FeatUp (Fu et al., 2024) popularized learned, backbone-aware upsampling, followed by AnyUp (Wimmer et al., 2026), JAFAR (Couairon et al., 2026), neighborhood attention filtering (Chambon et al., 2026), and LoftUp (Huang et al., 2025), which cross-attends from Fourier features of the guidance image to the coarse tokens at any output resolution.

**What supervises an upsampler.** These methods differ in how they refine the coarse map and, more consequentially here, in what supervises them. JAFAR, AnyUp, and neighborhood attention filtering regress the features the backbone produces when the same image is encoded at a higher input resolution, so their target samples a fixed field of view more finely. FeatUp instead constrains its prediction so that downsampling it reproduces the low-resolution features of several jittered views, some of which are zoomed, so a narrowed field of view does reach its objective, although only through a downsampling operator and never on the grid the prediction is made on. LoftUp adds to that reconstruction term an agreement between a region of its prediction and the backbone's encoding of the corresponding crop of a higher-resolution copy of the image, although that crop constrains one map serving the whole view at a resolution fixed in advance rather than the encoding of a queried region on the grid of that region alone. Since Section 2.1 measures that almost the entire distance between the coarse map and a re-encoding comes from narrowing the field of view rather than from added pixels, the difference between these targets is where the gap lives, and our modulator consumes the identical coarse map and view while being fitted to the target that carries it. A more invasive remedy retrains the encoder itself for flexible patch sizes or native resolutions (Beyer et al., 2023; Dehghani et al., 2023), which presupposes pretraining rather than a frozen checkpoint and still scales with image area.

**Foveated and active vision.** Concentrating resolution where it is needed has a long history in active perception (Bajcsy, 1988), and glimpse-based networks (Mnih et al., 2014; Jaderberg et al., 2015) learn where to crop or warp the input of a small task network trained end to end. Fusing a high-resolution crop with a low-resolution context view is established practice in task-trained dense prediction (Hoyer et al., 2022), and query-conditioned computation is central to promptable segmentation (Kirillov et al., 2023). Our setting differs in that the high-acuity pathway is retrofitted onto a frozen encoder in feature space, with the coarse map serving as peripheral percept and the foveal features supervised by the backbone's own re-encoding instead of by a task loss.

**Knowledge distillation and students that surpass teachers.** Distilled students occasionally outperform their teachers in classification, an effect commonly attributed to soft targets (Hinton et al., 2015; Furlanello et al., 2018), although a controlled study of it finds that a student rarely reproduces the predictive distribution of its teacher even where it scores above it (Stanton et al., 2021), and a student distilled from several heterogeneous foundation teachers can exceed all of them on dense benchmarks (Ranzinger et al., 2024). In our case the single teacher is the student's own frozen backbone applied to a different crop, the target is a feature map rather than a label distribution, and the gain materializes only under specific readouts, a dependence that places the cause in the features rather than in the optimization and connects it to variance reduction under averaging (Breiman, 1996).

## B Implementation Details

**Architecture.** The context encoder consists of three depthwise-separable convolutional blocks of width 256 operating on the $32 \times 32$ frozen feature grid. The point-query encoder concatenates a rasterized support mask, Fourier features of the center and radius with eight frequencies, and the log radius into a 64-dimensional embedding. Spatial feature-wise linear modulation is zero-initialized so that the network is exactly the identity at initialization, which makes the cosine to the teacher start exactly at the bilinear baseline. The coordinate-conditioned fine decoder gathers a $3 \times 3$ token neighborhood around each target coordinate, concatenates a convolutional encoding of the cropped global view at $256^2$ with 96 channels, and applies a four-layer multilayer perceptron of width 1024 per rendered patch. The full modulator has 9,207,393 trainable parameters.

**Training.** Training uses AdamW with learning rate $10^{-4}$, weight decay 0.05, a linear warmup of 1,000 steps, 24,000 steps in total, batch size 8 per GPU on two GPUs, and two queries per image drawn with a minimum separation margin. Radii are sampled log-uniformly from $[0.10, 0.25]$, the crop context margin is $\beta = 0.15$, and the hard-support multiplier is $\kappa = 1.10$. Both views receive photometric augmentation. The backbone is asserted frozen after the first backward pass, and startup validation enforces that view sizes are exact multiples of the patch size and that $\kappa \leq 1 + \beta$, so that every location the model can modulate is under teacher supervision.

**Evaluation crops.** Evaluation crops use radii in $[0.10, 0.25]$ with fixed seeds shared across all feature sources, and every dataset contributes two crops per image. Every few-shot result averages three episode seeds, each of which redraws the support and query assignment, and the seeds are shared across feature sources so that all sources see identical episodes. Cached features are stored in half precision. For every readout and every feature source, scores are computed only on patches inside the query support, which is the region our method is allowed to modify, so that all sources are scored on identical patches.

**Separation of training and evaluation data.** Crops are split by image parity into a fitting half and a held-out evaluation half, so that the two crops of one image always fall on the same side and no image is shared between the halves. This split sits inside a separation that already holds at the level of images, since the training corpus of Section 3 shares no image with any of the 25 evaluation sets. None of the 15,000 images the modulator is distilled on therefore appears in either half of any evaluation set, and no support crop of any few-shot episode comes from an image the modulator was trained on.

**Fitting the readouts.** Every fitted readout, meaning the linear probe, the trained decoder, the depth heads, the normal head, and the CAM classifier, is trained on the fitting half and evaluated on the held-out half. Few-shot episodes draw their support crops from the fitting half and score on query crops from the held-out half, use at most forty query crops per class, and are repeated over the three seeds described above. Depth probing follows the JAFAR protocol with a linear 256-bin head whose prediction is the expectation of the bin centers under its normalized weights, so the readout is a regression, trained for 20 epochs under a scale-invariant logarithmic loss plus a gradient term, using Depth-Anything-V2-Small (Yang et al., 2024b) predictions as targets on datasets without depth annotation and real metric depth on NYUv2. The trained segmentation decoder is a UPerNet-style pyramid decoder (Xiao et al., 2018) trained identically for every source. The remaining readouts of Section 3 are STEGO-style (Hamilton et al., 2022) and deep-spectral (Melas-Kyriazi et al., 2022) unsupervised segmentation, temporal label propagation on SpaceNet7 time series and on DAVIS (Pont-Tuset et al., 2017), and semantic correspondence on SPair-71k (Min et al., 2019).

**Crop-grid ports.** The crop-grid readouts follow the reference implementations at their published hyperparameters, and one difference is common to all of them. The five few-shot classifiers, FROST and INSID3 included, run on the $32 \times 32$ crop grid and are scored by patch-level intersection over union, whereas their references predict on a finer full-image grid, $64 \times 64$ for FROST, and threshold at pixel resolution, so absolute crop-grid numbers are not comparable to published full-image numbers. Reference defaults are otherwise kept throughout, including the whitened features that FROST compares without renormalizing them and the single nearest neighbor per view that INSID3 votes with, and the one further deviation is that our FROST port omits the flip-augmented support

anchors. CAM and crop classification need a class label per crop and therefore drop SpaceNet7 and Massachusetts Roads, whose thin roads and buildings never form the majority label of a crop.

**Canonical resolution versus the view.** Fourteen of the 25 evaluation sets have canonical images that are typically wider than the $512^2$ view, by median factors between 1.3 and 13, so on them the zoom teacher resolves pixel detail the view has discarded. On the remaining eleven, which include every natural and indoor benchmark, the canonical image is at or below the view resolution and the two crops of Section 2.1 carry the same pixels. The teacher's distance to the interpolated coarse map is nonetheless the same in both regimes, at patch cosine 0.63 on LandCover.ai and 0.67 on WHU, whose canonical images are 512 pixels wide, against 0.67 on iSAID, whose canonical images are 3.4 times wider than the view. This is the evidence behind the statement in Section 2.1 that what separates the teacher from the coarse map is not primarily pixels.

## C Evaluation Datasets

Table 5 names the 25 datasets of the evaluation suite in the order the per-dataset tables of Appendix H use, together with what each of them labels. Two further benchmarks lie outside that suite because they are scored per sequence or per image pair rather than per query crop, DAVIS (Pont-Tuset et al., 2017) for temporal label propagation on video, whose numbers are quoted in Section 5.4, and SPair-71k (Min et al., 2019) for semantic correspondence, reported in Table 42. Where the per-dataset tables shorten a name, this table gives it in full, so WBS is the Water Bodies dataset, DLRSD is the Dense Labeling Remote Sensing Dataset, and Mass. is Massachusetts.

Table 5: **The evaluation suite.** The 21 remote sensing datasets, the three photographic ones, and the indoor one, in the order used throughout the appendix. Query crops are drawn from the held-out portion of each of them, as Appendix B describes.

| Dataset | Labels |
|---|---|
| LoveDA (Wang et al., 2021) | urban and rural land cover |
| ISPRS Potsdam (Rottensteiner et al., 2012) | very high resolution urban classes |
| ISPRS Vaihingen (Rottensteiner et al., 2012) | the same classes over a smaller town |
| iSAID (Zamir et al., 2019) | small object categories in satellite images |
| DeepGlobe (Demir et al., 2018) | rural land cover |
| LandCover.ai (Boguszewski et al., 2021) | buildings, woodland, water, and roads |
| FloodNet (Rahnemoonfar et al., 2021) | flooded and intact structures after a hurricane |
| UAVid (Lyu et al., 2020) | oblique urban drone video |
| UDD5 (Chen et al., 2018) | the five-class split of the Urban Drone Dataset |
| DLRSD (Shao et al., 2018) | dense labels over the UC Merced scenes |
| WBS (Escobar, 2021) | water bodies in Sentinel-2 scenes |
| OpenEarthMap (Xia et al., 2023) | global land cover at sub-meter resolution |
| DFC2020 (Robinson et al., 2021) | coarse global land cover from Sentinel imagery |
| Five-Billion-Pixels (Tong et al., 2023) | Gaofen-2 land cover in twenty-four categories |
| SpaceNet1 (Van Etten et al., 2018) | building footprints over Rio de Janeiro |
| SpaceNet2 (Van Etten et al., 2018) | building footprints over four further cities |
| SpaceNet7 (Van Etten et al., 2021) | monthly building footprints, also our time series |
| WHU Building (Ji et al., 2018) | building footprints in aerial imagery |
| Mass. Buildings (Mnih, 2013) | building footprints over Massachusetts |
| Mass. Roads (Mnih, 2013) | road centerlines over Massachusetts |
| xView2 (Gupta et al., 2019) | building footprints and damage grades |
| PASCAL VOC (Everingham et al., 2010) | objects in photographs |
| ADE20K (Zhou et al., 2017) | scene parsing in photographs |
| COCO-Stuff (Caesar et al., 2018) | things and stuff in photographs |
| NYUv2 (Silberman et al., 2012) | indoor depth, normals, and forty classes |

## D Assets and Licenses

All datasets are public research benchmarks used under their respective licenses, and the backbone checkpoints are the official public releases. Learned-upsampler baselines use the authors' public implementations.

## E BACKBONE GENERALITY

The method is defined for any frozen vision transformer, so we repeat the study on three further backbones, DINOv2 ViT-L/14, Franca ViT-L/14, and RADIO ViT-L/16. Each is frozen and each carries a modulator of its own, retrained from scratch because a modulator is tied to the feature space it was distilled in, and every other element of the recipe is the one Section 3 describes. The view is $448^2$ for the two backbones with a patch size of 14 and $512^2$ for the one with a patch size of 16, so that all three decode the same $32 \times 32$ grid and carry the same 9.2M modulator parameters. Every upsampler is likewise matched to the backbone it is evaluated on. The sweep covers ten datasets, eight overhead benchmarks, namely LoveDA, Potsdam, SpaceNet2, iSAID, UAVid, WHU Building, Vaihingen, and DFC2020, together with the two photographic benchmarks PASCAL VOC and ADE20K. Vaihingen and DFC2020 were added after the sweep had already run on the other eight, and including them costs us a row rather than buying one, since it turns a five-shot FROST advantage on RADIO into the tie reported below.

Table 6 reports every readout, and the two regimes of Section 5.2 separate on each backbone as they do on DINOv3. Under in-context matching our model is the best source apart from the teacher on eleven of the twelve rows and ahead of its own teacher on eight of them, one-shot INSID3 on DINOv2 being the row an upsampler wins. One-shot FROST is the readout on which the teacher keeps a lead on all three backbones, and five-shot FROST on RADIO ends level with it at 0.495 against 0.495. Per-patch depth puts it first among the deployable sources and above its own teacher on all three. The linear probe fits weights of its own, so the teacher recovers there exactly as it does throughout the study, and since a probe is also the setting the learned upsamplers are ordinarily tuned for, NAF leads that row on every backbone while our model stays close to its own teacher.

Table 6: **Backbone generality.** Every readout of the study that is defined per dataset, averaged over the ten datasets of this sweep, for three frozen backbones that each carry a modulator of their own trained under the identical recipe. Every upsampler is matched to the backbone it is evaluated on, higher is better throughout, and bold marks the best source excluding the teacher.

| Evaluation | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|
| DINOv2 ViT-L/14 | | | | | | | | |
| INSID3, 1 shot mIoU | 0.265 | 0.340 | 0.329 | 0.343 | **0.375** | 0.307 | 0.360 | 0.258 |
| INSID3, 5 shots mIoU | 0.268 | 0.367 | 0.360 | 0.364 | 0.405 | 0.337 | **0.426** | 0.278 |
| FROST, 1 shot mIoU | 0.256 | 0.269 | 0.221 | 0.261 | 0.261 | 0.208 | **0.384** | 0.388 |
| FROST, 5 shots mIoU | 0.255 | 0.323 | 0.209 | 0.303 | 0.309 | 0.230 | **0.497** | 0.494 |
| Linear probe mIoU | 0.443 | 0.478 | 0.456 | 0.473 | **0.514** | 0.504 | 0.486 | 0.487 |
| Depth $\delta_1$ (linear) | 0.262 | 0.269 | 0.260 | 0.269 | 0.272 | 0.268 | **0.287** | 0.271 |
| Franca ViT-L/14 | | | | | | | | |
| INSID3, 1 shot mIoU | 0.327 | 0.308 | 0.294 | 0.302 | 0.310 | 0.282 | **0.393** | 0.362 |
| INSID3, 5 shots mIoU | 0.349 | 0.336 | 0.311 | 0.329 | 0.329 | 0.309 | **0.435** | 0.420 |
| FROST, 1 shot mIoU | 0.227 | 0.220 | 0.188 | 0.206 | 0.225 | 0.198 | **0.355** | 0.378 |
| FROST, 5 shots mIoU | 0.256 | 0.286 | 0.191 | 0.274 | 0.270 | 0.186 | **0.486** | 0.471 |
| Linear probe mIoU | 0.450 | 0.481 | 0.458 | 0.482 | **0.510** | 0.508 | 0.490 | 0.492 |
| Depth $\delta_1$ (linear) | 0.259 | 0.268 | 0.260 | 0.267 | 0.271 | 0.266 | **0.286** | 0.270 |
| RADIO ViT-L/16 | | | | | | | | |
| INSID3, 1 shot mIoU | 0.329 | 0.354 | 0.337 | 0.360 | 0.363 | 0.309 | **0.432** | 0.375 |
| INSID3, 5 shots mIoU | 0.356 | 0.389 | 0.361 | 0.394 | 0.381 | 0.320 | **0.504** | 0.408 |
| FROST, 1 shot mIoU | 0.233 | 0.235 | 0.190 | 0.219 | 0.219 | 0.188 | **0.375** | 0.391 |
| FROST, 5 shots mIoU | 0.257 | 0.285 | 0.170 | 0.281 | 0.246 | 0.197 | **0.495** | 0.495 |
| Linear probe mIoU | 0.479 | 0.508 | 0.475 | 0.499 | **0.530** | 0.515 | 0.511 | 0.516 |
| Depth $\delta_1$ (linear) | 0.265 | 0.271 | 0.261 | 0.270 | 0.272 | 0.268 | **0.285** | 0.275 |

## F COMPARISON ON DINOV2 VIT-B/14

Table 7 repeats the readouts of Table 6 on a fourth frozen backbone, DINOv2 ViT-B/14, on which every learned upsampler runs a checkpoint released by its own authors. NAF and AnyUp are backbone agnostic and run the single checkpoint each of them publishes, while JAFAR and LoftUp (Huang et al., 2025) each publish one for this backbone. The learned upsamplers of this table are exactly the

four whose released weights cover this backbone. The modulator follows the recipe of Appendix E, the query crops are the ones that sweep uses, and the readouts are unchanged, so the table is read exactly as Table 6 is.

The ordering matches the rest of the study. LoftUp is the strongest of the upsamplers under in-context matching, and our model is ahead of every upsampler on five of the six rows, by 0.156 mIoU at one-shot FROST and 0.219 at five shots, and the exception is the linear probe, where NAF leads at 0.512 against our 0.492, which is the fitted readout the learned upsamplers are ordinarily tuned for. Our model is above its own teacher on all six rows, including per-patch depth at 0.290 against 0.271 and the linear probe, which is the one place this backbone departs from the pattern, since a fitted readout returns the lead to the teacher on every other backbone we ran. The teacher meanwhile falls below interpolation under INSID3, at 0.285 and 0.326 against 0.326 and 0.362, which is the inversion the DINOv2 rows of Table 6 also show.

Table 7: **DINOv2 ViT-B/14.** A fourth frozen backbone, on which every learned upsampler runs a checkpoint released by its own authors, since NAF and AnyUp are backbone agnostic and JAFAR and LoftUp each publish one for this backbone. The modulator follows the recipe of Appendix E and every entry is a mean over the same ten datasets. Higher is better throughout and bold marks the best source excluding the teacher.

| Evaluation | lowres | bilinear | JBU | NAF | AnyUp | JAFAR | LoftUp | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| INSID3, 1 shot mIoU | 0.296 | 0.326 | 0.317 | 0.347 | 0.293 | 0.308 | 0.372 | **0.379** | 0.285 |
| INSID3, 5 shots mIoU | 0.303 | 0.362 | 0.355 | 0.379 | 0.322 | 0.328 | 0.418 | **0.441** | 0.326 |
| FROST, 1 shot mIoU | 0.247 | 0.232 | 0.205 | 0.224 | 0.206 | 0.179 | 0.249 | **0.405** | 0.388 |
| FROST, 5 shots mIoU | 0.243 | 0.286 | 0.197 | 0.279 | 0.231 | 0.184 | 0.271 | **0.505** | 0.485 |
| Linear probe mIoU | 0.444 | 0.478 | 0.467 | **0.512** | 0.511 | 0.490 | 0.445 | 0.492 | 0.486 |
| Depth $\delta_1$ (linear) | 0.261 | 0.267 | 0.259 | 0.269 | 0.265 | 0.262 | 0.264 | **0.290** | 0.271 |

## G Latency Against the Number of Queries

Figure 5 measures the cost of one image as a function of how many of its regions are queried, and three regimes appear. The teacher spends a backbone pass per query, 16.6 milliseconds once its batches are full, and so has the steepest curve. Our model spends 23.0 milliseconds encoding the view and 1.35 per query after it. The learned upsamplers spend that same pass plus one over the whole frame and then serve every query for the price of a gather, which places them on a nearly flat line at a high offset. Our curve is below the teacher from the second query onward and below every full-frame upsampler up to 184 queries, so in the regime the interface is built for, a bounded number of glances at a wide field, foveation is the cheapest route to re-encoded features. At a single query the teacher is the cheaper of the two, at 23.0 milliseconds against 26.4, because one re-encoding costs about what encoding the view costs and only our model also returns the periphery. The standalone column of Table 4 charges a query the view pass plus the per-query cost measured inside a full batch of sixteen, which is why it reads 24.5 milliseconds where a query timed alone here costs 26.4. Past 184 queries the whole frame is being asked for and one dense pass is the better arrangement.

## H Full Per-Dataset Tables

This section expands every row of Tables 1 and 2 to its individual datasets, with one table per readout and the dataset mean repeated at the end of each so that every table can be checked against the corresponding row of the main paper. Appendix H.1 covers the few-shot classifiers and the linear probe, and Appendix H.2 covers the regression, retrieval, and unsupervised readouts. The three natural-image benchmarks are expanded separately in Appendix R, and the backbone comparison of Table 6 is reported as dataset means only. Temporal propagation is omitted because it is a single benchmark rather than a suite, and its per-source numbers already appear in Table 2.

### H.1 Few-shot and probe segmentation

The eleven tables of this section expand every row of Table 1 to its individual datasets. Sources appear in the order used in the main paper, and the final row of each table repeats the dataset mean,

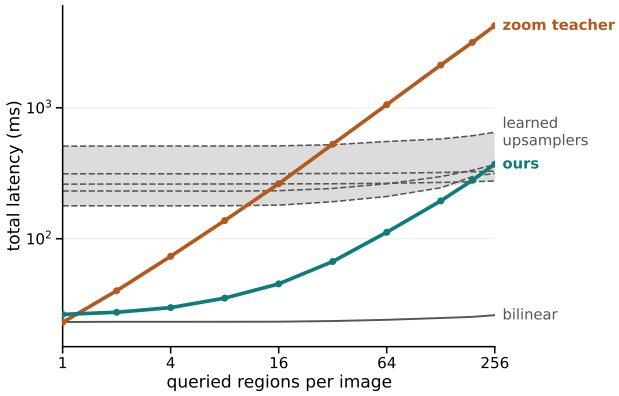


Figure 5: **Latency against the number of queried regions.** Measured time to serve a growing number of point queries on one image on an RTX A6000. The teacher re-encodes once per query in batches of at most sixteen crops, our model encodes the view once and decodes each query from it, and each learned upsampler encodes the view once, upsamples the whole frame once, and then only gathers from the result, which is what puts the five of them in the band. Every source is timed in a process of its own so that no other model is resident on the device and each point is the mean of two repetitions that agree to within 3 percent.

which matches the corresponding entry of Table 1 exactly. Throughout this appendix bold marks the best source excluding the teacher, which re-encodes the crop at native resolution and is reported as a cost reference rather than as a competing method, and Table 5 names every dataset in full, so that Mass. abbreviates Massachusetts, WBS the Water Bodies dataset, and DLRSD the Dense Labeling Remote Sensing Dataset.

Table 8: **Per-dataset mIoU for prototype matching at one shot.**

| Dataset | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| LoveDA | 0.393 | 0.394 | 0.383 | 0.389 | 0.397 | 0.377 | 0.394 | **0.401** | 0.405 |
| Potsdam | 0.505 | 0.509 | 0.498 | **0.511** | 0.510 | 0.501 | 0.510 | 0.443 | 0.440 |
| Vaihingen | 0.435 | **0.438** | 0.375 | 0.425 | 0.433 | 0.409 | 0.428 | 0.395 | 0.524 |
| iSAID | 0.320 | 0.318 | 0.272 | 0.328 | 0.323 | 0.312 | 0.321 | **0.354** | 0.395 |
| DeepGlobe | 0.457 | 0.458 | 0.416 | **0.465** | 0.457 | 0.422 | 0.443 | 0.417 | 0.428 |
| LandCover.ai | 0.566 | 0.567 | 0.571 | 0.588 | **0.589** | 0.559 | 0.578 | 0.564 | 0.584 |
| FloodNet | 0.554 | 0.561 | 0.530 | 0.553 | 0.594 | 0.582 | **0.600** | 0.571 | 0.623 |
| UAVid | 0.297 | 0.294 | 0.272 | 0.294 | 0.296 | 0.276 | 0.294 | **0.323** | 0.362 |
| UDD5 | 0.497 | **0.498** | 0.480 | 0.494 | 0.497 | 0.473 | 0.488 | 0.495 | 0.514 |
| DLRSD | 0.492 | 0.498 | 0.517 | 0.493 | 0.517 | 0.497 | **0.521** | 0.472 | 0.474 |
| WBS | 0.543 | 0.547 | 0.560 | 0.546 | 0.555 | 0.553 | **0.562** | 0.491 | 0.489 |
| OpenEarthMap | 0.408 | 0.407 | 0.396 | 0.398 | 0.415 | 0.397 | 0.420 | **0.438** | 0.450 |
| DFC2020 | 0.291 | 0.287 | **0.315** | 0.293 | 0.311 | 0.311 | 0.306 | 0.276 | 0.265 |
| Five-Billion-Pixels | 0.140 | 0.135 | 0.145 | 0.137 | 0.143 | 0.136 | **0.158** | 0.142 | 0.204 |
| SpaceNet1 | 0.436 | 0.425 | 0.376 | 0.431 | 0.430 | 0.399 | 0.426 | **0.484** | 0.469 |
| SpaceNet2 | 0.525 | **0.542** | 0.471 | 0.526 | 0.525 | 0.438 | 0.492 | 0.457 | 0.542 |
| SpaceNet7 | 0.144 | 0.141 | 0.142 | 0.138 | 0.147 | 0.145 | 0.160 | **0.166** | 0.178 |
| WHU Building | 0.595 | 0.613 | 0.516 | 0.599 | 0.660 | 0.582 | 0.656 | **0.688** | 0.690 |
| Mass. Buildings | 0.221 | 0.218 | 0.198 | 0.218 | 0.223 | 0.218 | 0.239 | **0.288** | 0.427 |
| Mass. Roads | 0.104 | 0.100 | 0.119 | 0.097 | 0.112 | 0.111 | 0.139 | **0.167** | 0.231 |
| xView2 | 0.308 | 0.300 | 0.254 | 0.290 | 0.317 | 0.292 | 0.337 | **0.459** | 0.500 |
| Mean | 0.392 | 0.393 | 0.372 | 0.391 | 0.402 | 0.381 | **0.404** | **0.404** | 0.438 |

Table 9: **Per-dataset mIoU for prototype matching at five shots.**

| Dataset | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| LoveDA | 0.526 | 0.527 | 0.502 | 0.527 | 0.530 | 0.492 | 0.525 | **0.537** | 0.541 |
| Potsdam | 0.578 | 0.579 | 0.563 | **0.580** | **0.580** | 0.564 | 0.578 | 0.517 | 0.509 |
| Vaihingen | 0.553 | 0.564 | 0.488 | 0.547 | **0.565** | 0.535 | **0.565** | 0.556 | 0.650 |
| iSAID | 0.406 | 0.415 | 0.366 | 0.408 | 0.432 | 0.416 | 0.429 | **0.437** | 0.512 |
| DeepGlobe | 0.530 | 0.538 | 0.522 | 0.545 | 0.539 | 0.511 | 0.523 | **0.549** | 0.555 |
| LandCover.ai | 0.704 | 0.711 | 0.703 | 0.701 | **0.729** | 0.728 | 0.725 | 0.708 | 0.715 |
| FloodNet | 0.644 | 0.652 | 0.616 | 0.631 | 0.666 | 0.667 | 0.687 | **0.692** | 0.734 |
| UAVid | 0.434 | 0.430 | 0.386 | 0.422 | 0.444 | 0.407 | 0.445 | **0.501** | 0.553 |
| UDD5 | 0.605 | 0.609 | 0.582 | 0.606 | **0.611** | 0.582 | 0.605 | 0.591 | 0.602 |
| DLRSD | 0.541 | 0.544 | 0.574 | 0.544 | 0.577 | 0.555 | 0.580 | **0.588** | 0.586 |
| WBS | 0.767 | 0.771 | 0.801 | 0.761 | 0.797 | 0.787 | **0.810** | 0.696 | 0.663 |
| OpenEarthMap | 0.451 | 0.444 | 0.415 | 0.441 | 0.452 | 0.420 | 0.455 | **0.478** | 0.499 |
| DFC2020 | 0.394 | 0.396 | 0.402 | 0.404 | 0.402 | 0.403 | 0.408 | **0.423** | 0.414 |
| Five-Billion-Pixels | 0.238 | 0.233 | 0.231 | 0.229 | 0.246 | 0.232 | 0.255 | **0.269** | 0.342 |
| SpaceNet1 | 0.491 | 0.490 | 0.412 | 0.506 | 0.493 | 0.441 | 0.484 | **0.551** | 0.563 |
| SpaceNet2 | 0.543 | 0.562 | 0.511 | 0.538 | 0.562 | 0.513 | **0.572** | 0.557 | 0.627 |
| SpaceNet7 | 0.187 | 0.190 | 0.165 | 0.184 | 0.195 | 0.182 | 0.197 | **0.210** | 0.242 |
| WHU Building | 0.646 | 0.664 | 0.572 | 0.659 | 0.710 | 0.677 | 0.715 | **0.771** | 0.828 |
| Mass. Buildings | 0.220 | 0.220 | 0.194 | 0.220 | 0.223 | 0.220 | 0.242 | **0.282** | 0.406 |
| Mass. Roads | 0.109 | 0.105 | 0.122 | 0.103 | 0.115 | 0.117 | 0.144 | **0.179** | 0.244 |
| xView2 | 0.345 | 0.335 | 0.281 | 0.329 | 0.341 | 0.309 | 0.354 | **0.457** | 0.529 |
| Mean | 0.472 | 0.475 | 0.448 | 0.471 | 0.486 | 0.465 | 0.490 | **0.502** | 0.539 |

Table 10: **Per-dataset mIoU for self-support prototypes at one shot.**

| Dataset | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| LoveDA | 0.371 | 0.361 | 0.352 | 0.365 | 0.358 | 0.331 | 0.353 | **0.397** | 0.393 |
| Potsdam | 0.462 | 0.465 | 0.443 | **0.468** | 0.460 | 0.449 | 0.460 | 0.419 | 0.391 |
| Vaihingen | **0.429** | 0.427 | 0.365 | 0.427 | 0.422 | 0.394 | 0.424 | 0.401 | 0.555 |
| iSAID | 0.337 | 0.334 | 0.264 | 0.327 | 0.324 | 0.305 | 0.330 | **0.375** | 0.445 |
| DeepGlobe | 0.409 | 0.399 | 0.356 | 0.412 | 0.409 | 0.348 | 0.393 | **0.416** | 0.429 |
| LandCover.ai | 0.489 | 0.472 | **0.537** | 0.499 | 0.493 | 0.473 | 0.501 | 0.508 | 0.508 |
| FloodNet | 0.604 | 0.611 | 0.550 | 0.595 | 0.636 | 0.622 | **0.654** | 0.635 | 0.689 |
| UAVid | 0.313 | 0.298 | 0.278 | 0.304 | 0.304 | 0.268 | 0.304 | **0.361** | 0.402 |
| UDD5 | **0.494** | 0.474 | 0.454 | 0.465 | 0.463 | 0.422 | 0.441 | 0.484 | 0.491 |
| DLRSD | 0.455 | 0.451 | **0.493** | 0.452 | 0.466 | 0.442 | 0.486 | 0.476 | 0.473 |
| WBS | 0.498 | 0.489 | 0.508 | 0.497 | 0.502 | 0.508 | **0.524** | 0.418 | 0.389 |
| OpenEarthMap | 0.384 | 0.375 | 0.354 | 0.368 | 0.378 | 0.330 | 0.381 | **0.416** | 0.413 |
| DFC2020 | 0.227 | 0.215 | **0.255** | 0.227 | 0.228 | 0.205 | 0.238 | 0.242 | 0.216 |
| Five-Billion-Pixels | 0.129 | 0.116 | 0.129 | 0.112 | 0.124 | 0.119 | **0.133** | **0.133** | 0.196 |
| SpaceNet1 | 0.496 | 0.502 | 0.424 | 0.498 | 0.501 | 0.478 | 0.493 | **0.539** | 0.533 |
| SpaceNet2 | **0.548** | 0.537 | 0.469 | 0.538 | 0.515 | 0.442 | 0.499 | 0.469 | 0.561 |
| SpaceNet7 | 0.161 | 0.162 | 0.160 | 0.158 | 0.169 | 0.175 | 0.181 | **0.184** | 0.201 |
| WHU Building | 0.640 | 0.647 | 0.555 | 0.641 | 0.721 | 0.649 | 0.727 | **0.769** | 0.805 |
| Mass. Buildings | 0.210 | 0.203 | 0.191 | 0.207 | 0.209 | 0.201 | 0.231 | **0.279** | 0.433 |
| Mass. Roads | 0.097 | 0.093 | 0.114 | 0.089 | 0.103 | 0.106 | 0.123 | **0.170** | 0.225 |
| xView2 | 0.261 | 0.235 | 0.222 | 0.238 | 0.257 | 0.243 | 0.284 | **0.467** | 0.522 |
| Mean | 0.382 | 0.375 | 0.356 | 0.376 | 0.383 | 0.358 | 0.389 | **0.408** | 0.441 |

Table 11: **Per-dataset mIoU for self-support prototypes at five shots.**

| Dataset | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| LoveDA | **0.523** | 0.520 | 0.486 | 0.514 | 0.520 | 0.474 | 0.514 | **0.523** | 0.540 |
| Potsdam | 0.567 | 0.566 | 0.551 | **0.569** | 0.568 | 0.553 | **0.569** | 0.479 | 0.469 |
| Vaihingen | 0.509 | 0.502 | 0.454 | 0.490 | 0.504 | 0.482 | 0.515 | **0.535** | 0.635 |
| iSAID | 0.397 | 0.389 | 0.313 | 0.367 | 0.385 | 0.351 | 0.383 | **0.428** | 0.511 |
| DeepGlobe | 0.511 | 0.520 | 0.499 | **0.529** | 0.525 | 0.478 | 0.506 | 0.520 | 0.538 |
| LandCover.ai | 0.579 | 0.573 | **0.630** | 0.568 | 0.590 | 0.577 | 0.604 | 0.600 | 0.609 |
| FloodNet | 0.646 | 0.655 | 0.611 | 0.642 | 0.671 | 0.651 | 0.683 | **0.692** | 0.743 |
| UAVid | 0.436 | 0.423 | 0.380 | 0.414 | 0.432 | 0.397 | 0.425 | **0.505** | 0.574 |
| UDD5 | 0.566 | 0.555 | 0.547 | 0.551 | 0.559 | 0.542 | 0.561 | **0.577** | 0.570 |
| DLRSD | 0.532 | 0.515 | 0.575 | 0.517 | 0.551 | 0.505 | 0.573 | **0.576** | 0.570 |
| WBS | 0.717 | 0.703 | 0.738 | 0.700 | 0.729 | 0.728 | **0.743** | 0.686 | 0.686 |
| OpenEarthMap | 0.414 | 0.398 | 0.374 | 0.390 | 0.407 | 0.369 | 0.412 | **0.462** | 0.478 |
| DFC2020 | 0.302 | 0.300 | **0.333** | 0.299 | 0.306 | 0.297 | 0.315 | 0.319 | 0.314 |
| Five-Billion-Pixels | 0.222 | 0.227 | 0.232 | 0.226 | **0.246** | 0.211 | 0.242 | 0.221 | 0.276 |
| SpaceNet1 | 0.519 | 0.537 | 0.445 | 0.523 | 0.533 | 0.508 | 0.527 | **0.568** | 0.566 |
| SpaceNet2 | 0.568 | 0.583 | 0.526 | 0.566 | 0.567 | 0.509 | 0.568 | **0.591** | 0.634 |
| SpaceNet7 | 0.182 | 0.186 | 0.162 | 0.181 | 0.178 | 0.170 | 0.188 | **0.210** | 0.242 |
| WHU Building | 0.651 | 0.671 | 0.583 | 0.658 | 0.727 | 0.705 | 0.741 | **0.783** | 0.836 |
| Mass. Buildings | 0.216 | 0.213 | 0.200 | 0.215 | 0.215 | 0.208 | 0.237 | **0.275** | 0.406 |
| Mass. Roads | 0.102 | 0.099 | 0.111 | 0.094 | 0.106 | 0.108 | 0.124 | **0.168** | 0.233 |
| xView2 | 0.362 | 0.337 | 0.290 | 0.327 | 0.349 | 0.318 | 0.368 | **0.473** | 0.553 |
| Mean | 0.453 | 0.451 | 0.431 | 0.445 | 0.460 | 0.435 | 0.467 | **0.485** | 0.523 |

Table 12: **Per-dataset mIoU for FSSDINO at one shot.**

| Dataset | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| LoveDA | 0.399 | 0.402 | 0.397 | 0.397 | **0.409** | 0.396 | 0.401 | 0.393 | 0.395 |
| Potsdam | 0.488 | 0.508 | 0.490 | **0.509** | 0.508 | 0.494 | 0.507 | 0.430 | 0.422 |
| Vaihingen | 0.444 | **0.454** | 0.391 | 0.443 | 0.452 | 0.425 | 0.448 | 0.396 | 0.527 |
| iSAID | 0.323 | 0.313 | 0.278 | 0.312 | 0.309 | 0.301 | 0.312 | **0.349** | 0.401 |
| DeepGlobe | 0.431 | 0.438 | 0.410 | **0.453** | 0.436 | 0.417 | 0.433 | 0.390 | 0.391 |
| LandCover.ai | 0.532 | 0.554 | 0.559 | **0.573** | 0.564 | 0.541 | 0.560 | 0.521 | 0.497 |
| FloodNet | 0.556 | 0.555 | 0.528 | 0.534 | **0.575** | 0.555 | 0.570 | 0.537 | 0.577 |
| UAVid | 0.303 | 0.309 | 0.297 | 0.310 | 0.313 | 0.304 | 0.310 | **0.331** | 0.381 |
| UDD5 | **0.501** | 0.490 | 0.469 | 0.488 | 0.474 | 0.457 | 0.466 | 0.472 | 0.484 |
| DLRSD | 0.496 | 0.513 | **0.536** | 0.497 | 0.533 | 0.524 | **0.536** | 0.473 | 0.480 |
| WBS | 0.559 | 0.594 | 0.613 | 0.575 | 0.613 | **0.629** | 0.627 | 0.451 | 0.428 |
| OpenEarthMap | 0.409 | 0.420 | 0.401 | 0.409 | 0.422 | 0.404 | 0.424 | **0.430** | 0.454 |
| DFC2020 | 0.293 | 0.296 | **0.315** | 0.302 | 0.298 | 0.297 | 0.304 | 0.271 | 0.253 |
| Five-Billion-Pixels | 0.130 | 0.128 | **0.151** | 0.135 | 0.132 | 0.129 | 0.149 | 0.135 | 0.194 |
| SpaceNet1 | 0.441 | 0.431 | 0.377 | 0.429 | 0.432 | 0.388 | 0.415 | **0.470** | 0.466 |
| SpaceNet2 | 0.489 | **0.494** | 0.409 | 0.467 | 0.477 | 0.398 | 0.450 | 0.413 | 0.494 |
| SpaceNet7 | 0.130 | 0.141 | 0.124 | 0.134 | 0.146 | 0.130 | 0.146 | **0.156** | 0.175 |
| WHU Building | 0.590 | 0.624 | 0.518 | 0.614 | 0.689 | 0.573 | 0.664 | **0.693** | 0.724 |
| Mass. Buildings | 0.223 | 0.225 | 0.131 | 0.222 | 0.225 | 0.218 | 0.236 | **0.300** | 0.446 |
| Mass. Roads | 0.101 | 0.098 | 0.117 | 0.095 | 0.110 | 0.107 | 0.132 | **0.161** | 0.215 |
| xView2 | 0.316 | 0.321 | 0.288 | 0.312 | 0.341 | 0.333 | 0.356 | **0.486** | 0.549 |
| Mean | 0.388 | 0.396 | 0.371 | 0.391 | **0.403** | 0.382 | 0.402 | 0.393 | 0.426 |

Table 13: **Per-dataset mIoU for FSSDINO at five shots.**

| Dataset | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| LoveDA | 0.510 | 0.519 | 0.491 | 0.514 | 0.518 | 0.484 | 0.513 | **0.532** | 0.525 |
| Potsdam | 0.579 | 0.596 | 0.584 | 0.595 | **0.598** | 0.587 | **0.598** | 0.500 | 0.484 |
| Vaihingen | 0.552 | **0.565** | 0.491 | 0.544 | 0.558 | 0.527 | 0.538 | 0.547 | 0.656 |
| iSAID | 0.390 | 0.405 | 0.349 | 0.398 | 0.420 | 0.390 | 0.411 | **0.449** | 0.525 |
| DeepGlobe | 0.512 | 0.528 | 0.514 | **0.544** | 0.542 | 0.526 | 0.524 | 0.523 | 0.537 |
| LandCover.ai | 0.648 | 0.657 | 0.624 | 0.669 | 0.680 | **0.688** | 0.678 | 0.684 | 0.678 |
| FloodNet | 0.630 | 0.644 | 0.605 | 0.624 | 0.654 | 0.631 | 0.670 | **0.679** | 0.707 |
| UAVid | 0.445 | 0.440 | 0.404 | 0.434 | 0.448 | 0.420 | 0.445 | **0.497** | 0.552 |
| UDD5 | 0.605 | 0.621 | 0.593 | 0.613 | 0.620 | 0.598 | **0.626** | 0.578 | 0.579 |
| DLRSD | 0.542 | 0.556 | 0.571 | 0.561 | 0.582 | 0.566 | 0.584 | **0.585** | 0.579 |
| WBS | 0.745 | 0.755 | 0.774 | 0.754 | 0.769 | **0.778** | **0.778** | 0.670 | 0.642 |
| OpenEarthMap | 0.439 | 0.448 | 0.420 | 0.444 | 0.453 | 0.428 | **0.460** | 0.456 | 0.474 |
| DFC2020 | 0.350 | 0.353 | 0.358 | 0.369 | 0.368 | 0.368 | 0.381 | **0.385** | 0.358 |
| Five-Billion-Pixels | 0.225 | 0.229 | 0.214 | 0.221 | 0.247 | 0.227 | **0.259** | 0.256 | 0.312 |
| SpaceNet1 | 0.494 | 0.507 | 0.439 | 0.508 | 0.523 | 0.476 | 0.503 | **0.549** | 0.553 |
| SpaceNet2 | 0.506 | **0.510** | 0.452 | 0.476 | 0.497 | 0.453 | 0.508 | 0.499 | 0.564 |
| SpaceNet7 | 0.184 | 0.182 | 0.180 | 0.185 | 0.180 | 0.183 | 0.198 | **0.223** | 0.260 |
| WHU Building | 0.641 | 0.676 | 0.579 | 0.660 | 0.723 | 0.686 | 0.715 | **0.765** | 0.825 |
| Mass. Buildings | 0.219 | 0.220 | 0.177 | 0.217 | 0.223 | 0.222 | 0.247 | **0.306** | 0.368 |
| Mass. Roads | 0.106 | 0.101 | 0.120 | 0.099 | 0.110 | 0.110 | 0.128 | **0.161** | 0.227 |
| xView2 | 0.370 | 0.379 | 0.303 | 0.368 | 0.370 | 0.323 | 0.382 | **0.508** | 0.568 |
| Mean | 0.462 | 0.471 | 0.440 | 0.467 | 0.480 | 0.460 | 0.483 | **0.493** | 0.523 |

Table 14: **Per-dataset mIoU for INSID3 at one shot.**

| Dataset | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| LoveDA | 0.369 | 0.346 | 0.343 | 0.343 | 0.359 | 0.293 | 0.341 | **0.428** | 0.341 |
| Potsdam | 0.477 | 0.490 | 0.475 | **0.516** | 0.505 | 0.473 | 0.494 | 0.501 | 0.309 |
| Vaihingen | 0.408 | 0.387 | 0.333 | 0.395 | 0.391 | 0.321 | 0.403 | **0.469** | 0.511 |
| iSAID | 0.264 | 0.228 | 0.213 | 0.218 | 0.242 | 0.197 | 0.248 | **0.434** | 0.512 |
| DeepGlobe | 0.455 | 0.419 | 0.390 | 0.422 | 0.422 | 0.382 | 0.404 | **0.474** | 0.435 |
| LandCover.ai | 0.536 | 0.530 | 0.507 | 0.503 | 0.502 | 0.469 | 0.511 | **0.676** | 0.523 |
| FloodNet | 0.528 | 0.534 | 0.497 | 0.496 | 0.543 | 0.426 | 0.538 | **0.720** | 0.664 |
| UAVid | 0.258 | 0.268 | 0.247 | 0.266 | 0.279 | 0.238 | 0.261 | **0.383** | 0.378 |
| UDD5 | 0.395 | 0.423 | 0.415 | 0.455 | 0.402 | 0.412 | 0.417 | **0.538** | 0.367 |
| DLRSD | 0.469 | 0.473 | 0.503 | 0.462 | 0.484 | 0.414 | 0.471 | **0.520** | 0.434 |
| WBS | 0.449 | 0.366 | **0.470** | 0.335 | 0.432 | 0.376 | 0.427 | 0.446 | 0.282 |
| OpenEarthMap | 0.312 | 0.373 | 0.342 | 0.363 | 0.367 | 0.337 | 0.359 | **0.425** | 0.289 |
| DFC2020 | 0.245 | **0.262** | 0.229 | 0.255 | 0.220 | 0.209 | 0.250 | 0.261 | 0.185 |
| Five-Billion-Pixels | 0.124 | 0.119 | 0.119 | 0.117 | 0.120 | 0.119 | 0.127 | **0.136** | 0.203 |
| SpaceNet1 | 0.340 | 0.292 | 0.257 | 0.285 | 0.253 | 0.208 | 0.309 | **0.463** | 0.450 |
| SpaceNet2 | 0.358 | 0.323 | 0.294 | 0.289 | 0.298 | 0.212 | 0.302 | **0.399** | 0.358 |
| SpaceNet7 | 0.098 | 0.089 | 0.090 | 0.079 | 0.099 | 0.075 | 0.085 | **0.140** | 0.164 |
| WHU Building | 0.573 | 0.369 | 0.405 | 0.374 | 0.415 | 0.244 | 0.409 | **0.679** | 0.637 |
| Mass. Buildings | 0.197 | 0.197 | 0.167 | 0.194 | 0.194 | 0.182 | 0.183 | **0.215** | 0.246 |
| Mass. Roads | 0.077 | 0.069 | 0.079 | 0.070 | 0.075 | 0.072 | 0.078 | **0.121** | 0.182 |
| xView2 | 0.220 | 0.205 | 0.189 | 0.207 | 0.195 | 0.184 | 0.190 | **0.412** | 0.471 |
| Mean | 0.340 | 0.322 | 0.313 | 0.316 | 0.324 | 0.278 | 0.324 | **0.421** | 0.378 |

Table 15: **Per-dataset mIoU for INSID3 at five shots.**

| Dataset | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| LoveDA | 0.385 | 0.385 | 0.366 | 0.405 | 0.374 | 0.320 | 0.397 | **0.517** | 0.394 |
| Potsdam | 0.507 | 0.515 | 0.490 | 0.531 | 0.517 | 0.461 | 0.514 | **0.576** | 0.338 |
| Vaihingen | 0.429 | 0.411 | 0.381 | 0.414 | 0.430 | 0.354 | 0.442 | **0.578** | 0.632 |
| iSAID | 0.297 | 0.290 | 0.323 | 0.271 | 0.294 | 0.265 | 0.325 | **0.509** | 0.530 |
| DeepGlobe | 0.478 | 0.448 | 0.446 | 0.461 | 0.441 | 0.405 | 0.442 | **0.545** | 0.447 |
| LandCover.ai | 0.560 | 0.576 | 0.557 | 0.531 | 0.562 | 0.530 | 0.541 | **0.748** | 0.601 |
| FloodNet | 0.587 | 0.573 | 0.548 | 0.522 | 0.564 | 0.475 | 0.563 | **0.781** | 0.702 |
| UAVid | 0.301 | 0.317 | 0.270 | 0.321 | 0.323 | 0.280 | 0.312 | **0.495** | 0.459 |
| UDD5 | 0.449 | 0.455 | 0.474 | 0.477 | 0.455 | 0.453 | 0.476 | **0.616** | 0.452 |
| DLRSD | 0.512 | 0.477 | 0.475 | 0.479 | 0.483 | 0.414 | 0.460 | **0.615** | 0.531 |
| WBS | 0.546 | 0.464 | 0.505 | 0.372 | 0.517 | 0.419 | 0.401 | **0.651** | 0.401 |
| OpenEarthMap | 0.349 | 0.329 | 0.288 | 0.323 | 0.315 | 0.254 | 0.293 | **0.433** | 0.325 |
| DFC2020 | 0.236 | 0.249 | 0.219 | 0.242 | 0.197 | 0.190 | 0.224 | **0.292** | 0.162 |
| Five-Billion-Pixels | 0.147 | 0.151 | 0.117 | 0.163 | 0.184 | 0.159 | 0.163 | **0.211** | 0.260 |
| SpaceNet1 | 0.365 | 0.334 | 0.319 | 0.323 | 0.314 | 0.232 | 0.370 | **0.550** | 0.487 |
| SpaceNet2 | 0.402 | 0.322 | 0.290 | 0.317 | 0.313 | 0.223 | 0.317 | **0.495** | 0.390 |
| SpaceNet7 | 0.102 | 0.103 | 0.061 | 0.109 | 0.103 | 0.083 | 0.106 | **0.152** | 0.197 |
| WHU Building | 0.595 | 0.375 | 0.387 | 0.373 | 0.429 | 0.242 | 0.430 | **0.746** | 0.776 |
| Mass. Buildings | 0.177 | 0.199 | 0.101 | 0.197 | 0.192 | 0.166 | 0.183 | **0.206** | 0.262 |
| Mass. Roads | 0.063 | 0.069 | 0.034 | 0.078 | 0.081 | 0.082 | 0.081 | **0.125** | 0.204 |
| xView2 | 0.239 | 0.202 | 0.199 | 0.224 | 0.203 | 0.192 | 0.197 | **0.462** | 0.518 |
| Mean | 0.368 | 0.345 | 0.326 | 0.340 | 0.347 | 0.295 | 0.345 | **0.491** | 0.432 |

Table 16: **Per-dataset mIoU for FROST at one shot.**

| Dataset | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| LoveDA | 0.230 | 0.229 | 0.134 | 0.183 | 0.220 | 0.143 | 0.163 | **0.408** | 0.356 |
| Potsdam | 0.305 | 0.288 | 0.272 | 0.252 | 0.308 | 0.315 | 0.299 | **0.503** | 0.410 |
| Vaihingen | 0.279 | 0.305 | 0.200 | 0.267 | 0.278 | 0.206 | 0.233 | **0.475** | 0.536 |
| iSAID | 0.222 | 0.217 | 0.185 | 0.170 | 0.266 | 0.238 | 0.178 | **0.361** | 0.454 |
| DeepGlobe | 0.238 | 0.203 | 0.155 | 0.222 | 0.179 | 0.184 | 0.238 | **0.423** | 0.437 |
| LandCover.ai | 0.254 | 0.287 | 0.230 | 0.286 | 0.245 | 0.276 | 0.256 | **0.650** | 0.489 |
| FloodNet | 0.329 | 0.330 | 0.221 | 0.319 | 0.277 | 0.287 | 0.232 | **0.612** | 0.602 |
| UAVid | 0.177 | 0.217 | 0.119 | 0.195 | 0.204 | 0.143 | 0.143 | **0.362** | 0.439 |
| UDD5 | 0.269 | 0.265 | 0.186 | 0.246 | 0.247 | 0.242 | 0.270 | **0.487** | 0.427 |
| DLRSD | 0.241 | 0.245 | 0.193 | 0.229 | 0.239 | 0.258 | 0.268 | **0.468** | 0.435 |
| WBS | 0.307 | 0.238 | 0.131 | 0.170 | 0.209 | 0.235 | 0.116 | **0.484** | 0.311 |
| OpenEarthMap | 0.161 | 0.167 | 0.113 | 0.148 | 0.174 | 0.205 | 0.173 | **0.423** | 0.329 |
| DFC2020 | 0.089 | 0.069 | 0.069 | 0.069 | 0.035 | 0.088 | 0.071 | **0.210** | 0.180 |
| Five-Billion-Pixels | 0.075 | 0.057 | 0.041 | 0.060 | 0.045 | 0.055 | 0.072 | **0.126** | 0.198 |
| SpaceNet1 | 0.181 | 0.088 | 0.115 | 0.025 | 0.064 | 0.089 | 0.091 | **0.486** | 0.415 |
| SpaceNet2 | 0.162 | 0.049 | 0.045 | 0.017 | 0.109 | 0.046 | 0.048 | **0.407** | 0.318 |
| SpaceNet7 | 0.006 | 0.000 | 0.005 | 0.000 | 0.001 | 0.000 | 0.002 | **0.022** | 0.038 |
| WHU Building | 0.186 | 0.254 | 0.223 | 0.326 | 0.141 | 0.197 | 0.087 | **0.627** | 0.599 |
| Mass. Buildings | 0.075 | 0.038 | 0.009 | 0.003 | 0.022 | 0.064 | 0.075 | **0.209** | 0.307 |
| Mass. Roads | 0.024 | 0.016 | 0.000 | 0.015 | 0.005 | 0.007 | 0.003 | **0.133** | 0.172 |
| xView2 | 0.114 | 0.083 | 0.096 | 0.059 | 0.206 | 0.252 | 0.123 | **0.506** | 0.497 |
| Mean | 0.187 | 0.174 | 0.131 | 0.155 | 0.166 | 0.168 | 0.149 | **0.399** | 0.379 |

Table 17: **Per-dataset mIoU for FROST at five shots.**

| Dataset | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| LoveDA | 0.182 | 0.210 | 0.123 | 0.194 | 0.180 | 0.181 | 0.193 | **0.514** | 0.431 |
| Potsdam | 0.314 | 0.289 | 0.237 | 0.250 | 0.260 | 0.269 | 0.266 | **0.571** | 0.477 |
| Vaihingen | 0.317 | 0.387 | 0.167 | 0.324 | 0.316 | 0.237 | 0.260 | **0.621** | 0.715 |
| iSAID | 0.226 | 0.237 | 0.213 | 0.199 | 0.325 | 0.255 | 0.249 | **0.505** | 0.572 |
| DeepGlobe | 0.296 | 0.316 | 0.221 | 0.308 | 0.264 | 0.233 | 0.232 | **0.540** | 0.521 |
| LandCover.ai | 0.287 | 0.311 | 0.191 | 0.300 | 0.290 | 0.285 | 0.266 | **0.779** | 0.689 |
| FloodNet | 0.438 | 0.461 | 0.285 | 0.437 | 0.393 | 0.384 | 0.424 | **0.838** | 0.786 |
| UAVid | 0.208 | 0.265 | 0.139 | 0.232 | 0.240 | 0.193 | 0.202 | **0.556** | 0.576 |
| UDD5 | 0.454 | 0.455 | 0.332 | 0.430 | 0.390 | 0.399 | 0.369 | **0.641** | 0.584 |
| DLRSD | 0.335 | 0.342 | 0.295 | 0.310 | 0.335 | 0.343 | 0.326 | **0.657** | 0.622 |
| WBS | 0.221 | 0.181 | 0.086 | 0.150 | 0.142 | 0.136 | 0.067 | **0.760** | 0.490 |
| OpenEarthMap | 0.149 | 0.180 | 0.081 | 0.140 | 0.143 | 0.176 | 0.146 | **0.446** | 0.404 |
| DFC2020 | 0.036 | 0.031 | 0.021 | 0.031 | 0.022 | 0.016 | 0.028 | **0.267** | 0.179 |
| Five-Billion-Pixels | 0.053 | 0.045 | 0.022 | 0.042 | 0.034 | 0.041 | 0.025 | **0.162** | 0.243 |
| SpaceNet1 | 0.325 | 0.241 | 0.150 | 0.092 | 0.289 | 0.304 | 0.127 | **0.573** | 0.560 |
| SpaceNet2 | 0.144 | 0.116 | 0.044 | 0.081 | 0.113 | 0.078 | 0.043 | **0.595** | 0.476 |
| SpaceNet7 | 0.000 | 0.004 | 0.004 | 0.001 | 0.001 | 0.007 | 0.000 | **0.071** | 0.077 |
| WHU Building | 0.285 | 0.376 | 0.069 | 0.292 | 0.188 | 0.135 | 0.132 | **0.787** | 0.820 |
| Mass. Buildings | 0.012 | 0.006 | 0.002 | 0.006 | 0.012 | 0.001 | 0.010 | **0.216** | 0.348 |
| Mass. Roads | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.004 | 0.000 | **0.126** | 0.185 |
| xView2 | 0.062 | 0.083 | 0.047 | 0.028 | 0.127 | 0.141 | 0.074 | **0.540** | 0.495 |
| Mean | 0.207 | 0.216 | 0.130 | 0.183 | 0.194 | 0.182 | 0.164 | **0.512** | 0.488 |

Table 18: **Per-dataset mIoU for linear-probe segmentation.**

| Dataset | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| LoveDA | 0.510 | 0.523 | 0.499 | 0.510 | **0.538** | 0.519 | 0.534 | 0.531 | 0.534 |
| Potsdam | 0.597 | 0.613 | 0.593 | 0.610 | **0.618** | 0.615 | 0.615 | 0.598 | 0.543 |
| Vaihingen | 0.518 | 0.557 | 0.470 | 0.545 | 0.570 | 0.546 | 0.567 | **0.575** | 0.623 |
| iSAID | 0.273 | 0.307 | 0.287 | 0.303 | 0.343 | 0.334 | 0.326 | **0.360** | 0.453 |
| DeepGlobe | 0.447 | 0.456 | 0.446 | 0.454 | 0.460 | **0.463** | 0.456 | 0.446 | 0.464 |
| LandCover.ai | 0.701 | 0.733 | 0.691 | 0.728 | 0.739 | 0.736 | 0.738 | **0.740** | 0.687 |
| FloodNet | 0.559 | 0.591 | 0.536 | 0.577 | 0.605 | 0.611 | **0.621** | 0.595 | 0.670 |
| UAVid | 0.512 | 0.537 | 0.460 | 0.526 | 0.564 | 0.543 | 0.556 | **0.596** | 0.661 |
| UDD5 | 0.609 | 0.626 | 0.590 | 0.633 | 0.631 | 0.614 | 0.622 | **0.643** | 0.638 |
| DLRSD | 0.606 | 0.623 | 0.626 | 0.604 | 0.647 | 0.639 | **0.651** | 0.595 | 0.532 |
| WBS | 0.793 | 0.807 | 0.813 | 0.802 | **0.818** | 0.812 | **0.818** | 0.799 | 0.750 |
| OpenEarthMap | 0.520 | 0.544 | 0.517 | 0.540 | 0.566 | 0.548 | **0.569** | 0.544 | 0.575 |
| DFC2020 | 0.443 | 0.449 | 0.453 | 0.449 | 0.458 | **0.462** | **0.462** | 0.445 | 0.437 |
| Five-Billion-Pixels | 0.118 | 0.120 | 0.121 | 0.116 | 0.135 | 0.133 | 0.143 | **0.147** | 0.195 |
| SpaceNet1 | 0.661 | 0.689 | 0.610 | 0.676 | 0.685 | 0.664 | 0.667 | **0.700** | 0.700 |
| SpaceNet2 | 0.735 | 0.765 | 0.707 | 0.752 | 0.760 | 0.735 | 0.750 | **0.772** | 0.790 |
| SpaceNet7 | 0.497 | 0.506 | 0.504 | 0.502 | 0.520 | 0.522 | 0.530 | **0.539** | 0.567 |
| WHU Building | 0.776 | 0.822 | 0.721 | 0.803 | 0.853 | 0.824 | 0.837 | **0.874** | 0.916 |
| Mass. Buildings | 0.420 | 0.430 | 0.418 | 0.425 | 0.460 | 0.474 | 0.506 | **0.551** | 0.672 |
| Mass. Roads | 0.396 | 0.390 | 0.381 | 0.381 | 0.408 | 0.414 | 0.456 | **0.466** | 0.527 |
| xView2 | 0.648 | 0.672 | 0.608 | 0.659 | 0.682 | 0.659 | 0.684 | **0.725** | 0.778 |
| Mean | 0.540 | 0.560 | 0.526 | 0.552 | 0.574 | 0.565 | 0.577 | **0.583** | 0.605 |

### H.2 Regression, retrieval, and unsupervised readouts

The ten tables of this section expand the rows of Table 2 that are defined per dataset. Temporal propagation has no such expansion because it is a single benchmark on SpaceNet7 time series. Higher is better in every table except the one reporting root mean squared error.

Table 19: **Per-dataset linear-probe depth threshold accuracy $\delta_1$.**

| Dataset | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| LoveDA | 0.293 | 0.296 | 0.290 | 0.296 | 0.296 | 0.291 | 0.293 | **0.319** | 0.310 |
| Potsdam | 0.240 | 0.245 | 0.239 | 0.244 | 0.244 | 0.240 | 0.243 | **0.261** | 0.263 |
| Vaihingen | 0.224 | 0.231 | 0.208 | 0.227 | 0.227 | 0.218 | 0.224 | **0.253** | 0.234 |
| iSAID | 0.254 | 0.260 | 0.249 | 0.259 | 0.260 | 0.250 | 0.254 | **0.270** | 0.267 |
| DeepGlobe | 0.324 | 0.336 | 0.329 | 0.338 | 0.336 | 0.327 | 0.325 | **0.344** | 0.348 |
| LandCover.ai | 0.327 | 0.331 | 0.327 | 0.329 | 0.330 | 0.327 | 0.326 | **0.362** | 0.349 |
| FloodNet | 0.296 | 0.303 | 0.280 | 0.302 | 0.303 | 0.289 | 0.297 | **0.326** | 0.315 |
| UAVid | 0.318 | 0.319 | 0.309 | 0.320 | 0.318 | 0.315 | 0.316 | **0.392** | 0.359 |
| UDD5 | 0.274 | 0.279 | 0.262 | 0.280 | 0.277 | 0.271 | 0.274 | **0.287** | 0.271 |
| DLRSD | 0.256 | 0.263 | 0.255 | 0.262 | 0.264 | 0.257 | 0.259 | **0.288** | 0.269 |
| WBS | 0.270 | 0.280 | 0.263 | 0.278 | 0.277 | 0.270 | 0.274 | **0.295** | 0.284 |
| OpenEarthMap | 0.309 | 0.314 | 0.300 | 0.312 | 0.313 | 0.302 | 0.307 | **0.337** | 0.326 |
| DFC2020 | 0.333 | 0.337 | 0.331 | 0.337 | 0.338 | 0.336 | 0.335 | **0.379** | 0.367 |
| Five-Billion-Pixels | 0.270 | 0.275 | **0.276** | 0.274 | 0.273 | 0.269 | 0.269 | 0.271 | 0.274 |
| SpaceNet1 | 0.341 | 0.347 | 0.337 | 0.346 | 0.347 | 0.341 | 0.341 | **0.424** | 0.386 |
| SpaceNet2 | 0.324 | 0.327 | 0.315 | 0.327 | 0.329 | 0.324 | 0.326 | **0.369** | 0.356 |
| SpaceNet7 | 0.297 | 0.302 | 0.301 | 0.300 | 0.303 | 0.294 | 0.296 | **0.306** | 0.305 |
| WHU Building | 0.285 | 0.296 | 0.261 | 0.293 | 0.293 | 0.278 | 0.282 | **0.336** | 0.310 |
| Mass. Buildings | 0.277 | 0.277 | 0.278 | 0.276 | 0.278 | 0.277 | 0.279 | **0.296** | 0.294 |
| Mass. Roads | 0.282 | 0.287 | 0.274 | 0.286 | 0.283 | 0.277 | 0.277 | **0.303** | 0.292 |
| xView2 | 0.329 | 0.343 | 0.326 | 0.342 | 0.341 | 0.332 | 0.329 | **0.355** | 0.344 |
| Mean | 0.292 | 0.297 | 0.286 | 0.297 | 0.297 | 0.290 | 0.292 | **0.323** | 0.311 |

Table 20: **Per-dataset linear-probe depth root mean squared error, where lower is better.**

| Dataset | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| LoveDA | 2.264 | 2.260 | 2.300 | 2.260 | 2.275 | 2.293 | 2.283 | **2.114** | 2.155 |
| Potsdam | 2.564 | 2.558 | 2.572 | 2.559 | 2.561 | 2.578 | 2.563 | **2.420** | 2.445 |
| Vaihingen | 2.319 | 2.304 | 2.424 | 2.315 | 2.360 | 2.434 | 2.369 | **2.112** | 2.293 |
| iSAID | 2.244 | 2.235 | 2.298 | 2.234 | 2.240 | 2.286 | 2.257 | **2.117** | 2.214 |
| DeepGlobe | 1.994 | 1.971 | 2.004 | 1.972 | 1.976 | 2.006 | 2.004 | **1.883** | 1.910 |
| LandCover.ai | 2.128 | 2.117 | 2.155 | 2.110 | 2.124 | 2.134 | 2.127 | **1.937** | 2.010 |
| FloodNet | 2.222 | 2.218 | 2.276 | 2.216 | 2.246 | 2.272 | 2.251 | **2.015** | 2.147 |
| UAVid | 2.240 | 2.212 | 2.277 | 2.211 | 2.221 | 2.272 | 2.243 | **1.875** | 2.015 |
| UDD5 | 2.322 | 2.318 | 2.374 | 2.314 | 2.330 | 2.359 | 2.333 | **2.184** | 2.302 |
| DLRSD | 2.412 | 2.388 | 2.441 | 2.388 | 2.395 | 2.425 | 2.410 | **2.251** | 2.305 |
| WBS | 2.389 | 2.403 | 2.404 | 2.394 | 2.430 | 2.424 | 2.415 | **2.134** | 2.302 |
| OpenEarthMap | 2.045 | 2.057 | 2.131 | 2.067 | 2.059 | 2.095 | 2.056 | **1.877** | 1.978 |
| DFC2020 | 2.097 | 2.093 | 2.125 | 2.093 | 2.105 | 2.115 | 2.115 | **1.931** | 1.980 |
| Five-Billion-Pixels | 2.133 | 2.170 | 2.282 | 2.177 | 2.180 | 2.170 | 2.164 | **2.113** | 2.168 |
| SpaceNet1 | 2.075 | 2.054 | 2.127 | 2.059 | 2.064 | 2.092 | 2.079 | **1.681** | 1.886 |
| SpaceNet2 | 2.191 | 2.171 | 2.187 | 2.174 | 2.198 | 2.203 | 2.202 | **1.909** | 2.067 |
| SpaceNet7 | 2.190 | 2.179 | 2.218 | 2.181 | 2.190 | 2.204 | 2.203 | **2.135** | 2.157 |
| WHU Building | 2.139 | 2.123 | 2.404 | 2.137 | 2.158 | 2.215 | 2.183 | **1.902** | 2.056 |
| Mass. Buildings | 2.117 | 2.138 | 2.279 | 2.128 | 2.153 | 2.129 | 2.134 | **2.043** | 2.145 |
| Mass. Roads | 2.246 | 2.271 | 2.336 | 2.271 | 2.292 | 2.307 | 2.291 | **2.145** | 2.272 |
| xView2 | 2.119 | 2.117 | 2.169 | 2.121 | 2.133 | 2.145 | 2.137 | **2.002** | 2.080 |
| Mean | 2.212 | 2.207 | 2.275 | 2.209 | 2.223 | 2.246 | 2.229 | **2.037** | 2.137 |

Table 21: **Per-dataset class-activation-map average gain.** SpaceNet7 and Massachusetts Roads are excluded because their buildings and roads never form the majority label of a crop.

| Dataset | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| LoveDA | 27.5 | 24.6 | 20.5 | 25.1 | 23.6 | 19.2 | 23.0 | **44.7** | 26.4 |
| Potsdam | 33.5 | 31.5 | 29.6 | 32.5 | 31.3 | 28.7 | 31.7 | **47.9** | 20.3 |
| Vaihingen | 67.7 | 63.7 | 59.7 | 63.9 | 64.4 | 57.5 | 65.1 | **73.6** | 66.3 |
| iSAID | 20.4 | 17.5 | 17.8 | 17.5 | 17.3 | 14.5 | 17.4 | **35.2** | 18.9 |
| DeepGlobe | 39.0 | 34.0 | 31.5 | 34.2 | 32.4 | 25.0 | 30.4 | **55.0** | 41.8 |
| LandCover.ai | 27.5 | 23.8 | 22.3 | 24.7 | 23.7 | 21.3 | 24.2 | **42.2** | 28.6 |
| FloodNet | 37.3 | 34.7 | 31.4 | 35.3 | 35.9 | 31.0 | 36.0 | **55.5** | 38.6 |
| UAVid | 46.6 | 43.5 | 43.7 | 45.2 | 44.9 | 39.9 | 44.7 | **64.3** | 50.9 |
| UDD5 | 37.6 | 35.1 | 35.5 | 36.0 | 37.0 | 34.4 | 37.3 | **49.4** | 37.3 |
| DLRSD | 26.0 | 23.4 | 22.9 | 25.6 | 24.3 | 22.0 | 25.2 | **45.0** | 23.3 |
| WBS | 43.7 | 41.0 | 42.1 | 41.6 | 41.6 | 39.1 | 42.2 | **54.9** | 41.0 |
| OpenEarthMap | 28.4 | 23.9 | 24.4 | 24.9 | 22.9 | 18.3 | 24.5 | **51.7** | 30.3 |
| DFC2020 | 24.0 | 21.7 | 22.9 | 21.6 | 21.5 | 18.4 | 21.9 | **37.2** | 20.4 |
| Five-Billion-Pixels | 10.6 | 9.5 | 9.4 | 9.8 | 9.4 | 8.5 | 9.3 | **21.0** | 17.7 |
| SpaceNet1 | 37.1 | 33.4 | 30.1 | 34.2 | 32.8 | 27.7 | 33.6 | **61.1** | 36.8 |
| SpaceNet2 | 37.0 | 33.9 | 29.5 | 34.5 | 32.2 | 24.6 | 31.4 | **40.4** | 31.8 |
| WHU Building | 34.8 | 29.3 | 35.8 | 29.8 | 30.7 | 26.9 | 32.1 | **56.1** | 36.4 |
| Mass. Buildings | 18.7 | 16.0 | 28.2 | 14.9 | 15.2 | 13.3 | 15.1 | **39.7** | 20.2 |
| xView2 | 23.3 | 20.7 | 15.6 | 18.7 | 20.4 | 17.4 | 20.9 | **41.9** | 22.2 |
| Mean | 32.7 | 29.5 | 29.1 | 30.0 | 29.5 | 25.7 | 29.8 | **48.2** | 32.1 |

Table 22: **Per-dataset crop classification accuracy.** SpaceNet7 and Massachusetts Roads are excluded because their buildings and roads never form the majority label of a crop.

| Dataset | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| LoveDA | 0.675 | 0.675 | 0.658 | **0.683** | 0.667 | 0.658 | 0.658 | 0.675 | 0.700 |
| Potsdam | 0.783 | 0.775 | **0.792** | 0.775 | 0.767 | 0.783 | 0.767 | 0.767 | 0.650 |
| Vaihingen | 0.742 | 0.733 | 0.758 | 0.750 | 0.767 | 0.733 | 0.758 | **0.783** | 0.792 |
| iSAID | **1.000** | **1.000** | **1.000** | 0.992 | **1.000** | **1.000** | **1.000** | 0.983 | 1.000 |
| DeepGlobe | 0.867 | 0.867 | 0.858 | 0.867 | 0.867 | 0.867 | 0.867 | **0.883** | 0.825 |
| LandCover.ai | 0.950 | 0.950 | 0.950 | 0.942 | **0.958** | **0.958** | 0.950 | 0.950 | 0.908 |
| FloodNet | **0.858** | 0.850 | 0.842 | **0.858** | **0.858** | 0.842 | 0.842 | **0.858** | 0.842 |
| UAVid | 0.767 | 0.767 | 0.767 | 0.783 | 0.775 | 0.775 | 0.792 | **0.825** | 0.800 |
| UDD5 | 0.842 | **0.850** | 0.817 | 0.825 | 0.842 | 0.825 | **0.850** | 0.800 | 0.825 |
| DLRSD | 0.725 | 0.742 | 0.733 | **0.758** | 0.733 | 0.733 | 0.733 | 0.733 | 0.700 |
| WBS | 0.900 | 0.900 | 0.900 | 0.883 | 0.892 | 0.900 | 0.900 | **0.908** | 0.867 |
| OpenEarthMap | 0.550 | 0.567 | 0.550 | 0.567 | 0.567 | 0.567 | 0.550 | **0.625** | 0.583 |
| DFC2020 | 0.700 | 0.708 | 0.717 | 0.700 | 0.708 | 0.725 | 0.717 | **0.767** | 0.717 |
| Five-Billion-Pixels | 0.450 | 0.450 | 0.450 | 0.475 | 0.450 | 0.450 | 0.475 | **0.525** | 0.575 |
| SpaceNet1 | 0.967 | 0.967 | 0.967 | 0.967 | 0.967 | 0.967 | 0.967 | **0.975** | 0.967 |
| SpaceNet2 | 0.958 | 0.958 | 0.958 | 0.950 | 0.958 | 0.958 | **0.967** | 0.942 | 0.950 |
| WHU Building | 0.983 | **0.992** | 0.975 | 0.983 | 0.983 | 0.975 | 0.983 | **0.992** | 0.983 |
| Mass. Buildings | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | 1.000 |
| xView2 | 0.975 | 0.975 | 0.975 | 0.967 | 0.975 | 0.975 | 0.975 | **0.992** | 0.975 |
| Mean | 0.826 | 0.828 | 0.825 | 0.828 | 0.828 | 0.826 | 0.829 | **0.841** | 0.824 |

Table 23: **Per-dataset trained-decoder segmentation mIoU.**

| Dataset | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| LoveDA | 0.537 | 0.542 | 0.527 | 0.525 | 0.541 | 0.534 | 0.546 | **0.574** | 0.558 |
| Potsdam | 0.624 | 0.619 | 0.614 | 0.619 | **0.630** | 0.628 | 0.629 | 0.605 | 0.542 |
| Vaihingen | 0.548 | 0.556 | 0.514 | 0.533 | 0.569 | 0.556 | 0.573 | **0.589** | 0.611 |
| iSAID | 0.333 | 0.362 | 0.269 | 0.329 | 0.357 | 0.342 | 0.344 | **0.381** | 0.466 |
| DeepGlobe | 0.427 | 0.429 | 0.397 | 0.433 | 0.425 | **0.470** | 0.450 | 0.462 | 0.459 |
| LandCover.ai | 0.792 | **0.811** | 0.756 | 0.743 | 0.788 | 0.771 | 0.780 | 0.794 | 0.728 |
| FloodNet | 0.638 | 0.647 | 0.612 | 0.638 | 0.657 | 0.674 | 0.687 | **0.688** | 0.675 |
| UAVid | 0.551 | 0.542 | 0.519 | 0.528 | 0.601 | 0.559 | 0.569 | **0.602** | 0.670 |
| UDD5 | 0.640 | 0.640 | 0.614 | **0.644** | 0.640 | 0.616 | 0.622 | 0.638 | 0.659 |
| DLRSD | 0.649 | 0.648 | 0.641 | 0.649 | 0.664 | 0.667 | **0.677** | 0.662 | 0.532 |
| WBS | 0.800 | 0.810 | 0.810 | 0.791 | 0.811 | **0.819** | 0.815 | **0.819** | 0.731 |
| OpenEarthMap | 0.514 | 0.495 | 0.512 | 0.494 | **0.570** | 0.542 | 0.562 | 0.539 | 0.577 |
| DFC2020 | 0.449 | 0.433 | 0.448 | 0.431 | 0.461 | 0.455 | 0.458 | **0.470** | 0.426 |
| Five-Billion-Pixels | 0.111 | 0.110 | 0.122 | 0.115 | 0.113 | 0.110 | 0.118 | **0.134** | 0.176 |
| SpaceNet1 | 0.687 | 0.684 | 0.679 | 0.683 | **0.701** | 0.685 | **0.701** | **0.701** | 0.703 |
| SpaceNet2 | 0.783 | **0.788** | 0.752 | 0.767 | 0.786 | 0.774 | 0.773 | 0.775 | 0.783 |
| SpaceNet7 | 0.478 | 0.507 | 0.504 | 0.499 | 0.499 | **0.535** | 0.526 | 0.521 | 0.605 |
| WHU Building | 0.844 | 0.842 | 0.790 | 0.826 | **0.878** | 0.846 | 0.865 | 0.870 | 0.926 |
| Mass. Buildings | 0.399 | 0.453 | 0.427 | 0.448 | 0.393 | 0.528 | 0.558 | **0.564** | 0.690 |
| Mass. Roads | 0.416 | 0.428 | 0.431 | 0.402 | 0.443 | 0.445 | 0.486 | **0.493** | 0.542 |
| xView2 | 0.691 | 0.684 | 0.657 | 0.679 | 0.711 | 0.702 | 0.727 | **0.733** | 0.811 |
| Mean | 0.567 | 0.573 | 0.552 | 0.561 | 0.583 | 0.584 | 0.594 | **0.601** | 0.613 |

Table 24: **Per-dataset trained-decoder depth threshold accuracy $\delta_1$.**

| Dataset | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| LoveDA | 0.335 | 0.331 | **0.336** | 0.334 | 0.320 | 0.312 | 0.327 | 0.332 | 0.369 |
| Potsdam | 0.294 | 0.297 | 0.293 | 0.296 | 0.286 | 0.309 | 0.279 | **0.325** | 0.357 |
| Vaihingen | 0.362 | 0.374 | 0.320 | 0.341 | 0.363 | 0.352 | 0.359 | **0.402** | 0.420 |
| iSAID | 0.285 | 0.311 | 0.303 | 0.297 | 0.290 | 0.311 | 0.311 | **0.328** | 0.330 |
| DeepGlobe | 0.334 | 0.317 | 0.320 | 0.328 | 0.347 | 0.356 | 0.337 | **0.386** | 0.373 |
| LandCover.ai | 0.366 | 0.356 | 0.388 | 0.379 | 0.355 | **0.391** | 0.374 | 0.368 | 0.448 |
| FloodNet | 0.422 | 0.441 | 0.417 | 0.427 | 0.445 | 0.427 | 0.429 | **0.463** | 0.480 |
| UAVid | 0.606 | 0.636 | 0.559 | 0.611 | 0.602 | 0.628 | 0.614 | **0.686** | 0.687 |
| UDD5 | 0.382 | 0.386 | 0.376 | 0.370 | 0.374 | **0.388** | 0.376 | 0.384 | 0.403 |
| DLRSD | 0.323 | 0.315 | 0.311 | 0.319 | 0.305 | 0.312 | 0.307 | **0.346** | 0.367 |
| WBS | 0.334 | 0.320 | 0.320 | 0.330 | 0.308 | 0.305 | 0.330 | **0.350** | 0.327 |
| OpenEarthMap | 0.367 | 0.356 | 0.327 | 0.340 | **0.372** | 0.340 | 0.357 | 0.353 | 0.405 |
| DFC2020 | 0.345 | 0.368 | 0.322 | **0.369** | 0.352 | 0.347 | 0.357 | 0.367 | 0.428 |
| Five-Billion-Pixels | 0.268 | 0.273 | **0.292** | 0.258 | 0.279 | 0.287 | 0.284 | 0.272 | 0.289 |
| SpaceNet1 | 0.519 | 0.503 | 0.508 | 0.487 | 0.514 | 0.517 | 0.523 | **0.560** | 0.609 |
| SpaceNet2 | 0.393 | 0.390 | 0.397 | 0.408 | 0.403 | 0.391 | 0.392 | **0.437** | 0.466 |
| SpaceNet7 | 0.275 | 0.287 | **0.309** | 0.298 | 0.283 | 0.298 | 0.288 | 0.292 | 0.307 |
| WHU Building | 0.413 | 0.401 | 0.364 | 0.416 | 0.409 | 0.413 | 0.423 | **0.455** | 0.474 |
| Mass. Buildings | 0.308 | 0.294 | 0.283 | 0.302 | 0.307 | 0.322 | 0.299 | **0.351** | 0.359 |
| Mass. Roads | 0.255 | 0.265 | 0.264 | 0.269 | 0.266 | 0.263 | 0.263 | **0.343** | 0.352 |
| xView2 | 0.345 | 0.352 | 0.338 | 0.322 | 0.334 | 0.328 | 0.329 | **0.384** | 0.415 |
| Mean | 0.359 | 0.361 | 0.350 | 0.357 | 0.358 | 0.362 | 0.360 | **0.390** | 0.413 |

Table 25: **Per-dataset patch retrieval precision at 1.**

| Dataset | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| LoveDA | 0.571 | **0.602** | 0.539 | 0.586 | 0.594 | 0.543 | 0.549 | 0.566 | 0.590 |
| Potsdam | 0.661 | 0.676 | 0.649 | **0.683** | 0.673 | 0.670 | 0.653 | 0.628 | 0.571 |
| Vaihingen | 0.652 | 0.732 | 0.617 | 0.713 | **0.735** | 0.667 | 0.693 | 0.733 | 0.839 |
| iSAID | 0.455 | 0.431 | 0.408 | 0.468 | **0.550** | 0.442 | 0.508 | 0.468 | 0.652 |
| DeepGlobe | 0.476 | 0.516 | 0.485 | 0.528 | 0.523 | **0.539** | 0.503 | 0.485 | 0.519 |
| LandCover.ai | 0.795 | 0.838 | 0.763 | **0.853** | 0.842 | 0.762 | 0.765 | 0.838 | 0.840 |
| FloodNet | 0.584 | 0.661 | 0.535 | **0.673** | 0.662 | 0.600 | 0.625 | 0.670 | 0.696 |
| UAVid | 0.549 | 0.604 | 0.495 | 0.606 | 0.624 | 0.580 | 0.609 | **0.662** | 0.745 |
| UDD5 | 0.699 | **0.751** | 0.662 | 0.741 | 0.733 | 0.700 | 0.715 | 0.733 | 0.731 |
| DLRSD | 0.652 | 0.698 | 0.670 | 0.699 | **0.707** | 0.670 | 0.669 | 0.680 | 0.685 |
| WBS | 0.842 | 0.857 | 0.870 | 0.845 | 0.873 | 0.865 | **0.877** | 0.825 | 0.812 |
| OpenEarthMap | 0.507 | 0.530 | 0.478 | 0.561 | 0.558 | 0.518 | 0.540 | **0.587** | 0.624 |
| DFC2020 | 0.451 | 0.461 | 0.457 | **0.486** | 0.475 | 0.472 | 0.466 | 0.476 | 0.445 |
| Five-Billion-Pixels | 0.103 | 0.121 | 0.118 | 0.124 | 0.132 | 0.124 | 0.131 | **0.138** | 0.234 |
| SpaceNet1 | 0.787 | 0.780 | 0.722 | 0.730 | 0.762 | 0.730 | 0.700 | **0.800** | 0.822 |
| SpaceNet2 | 0.802 | 0.835 | 0.767 | **0.840** | 0.820 | 0.767 | 0.762 | 0.837 | 0.855 |
| SpaceNet7 | 0.600 | 0.612 | 0.620 | 0.597 | 0.627 | **0.642** | 0.625 | 0.640 | 0.675 |
| WHU Building | 0.847 | 0.895 | 0.762 | 0.880 | 0.910 | 0.867 | 0.860 | **0.942** | 0.945 |
| Mass. Buildings | 0.612 | 0.575 | 0.565 | 0.593 | 0.575 | 0.635 | 0.637 | **0.672** | 0.787 |
| Mass. Roads | 0.560 | 0.567 | 0.562 | 0.555 | 0.540 | 0.562 | 0.590 | **0.642** | 0.712 |
| xView2 | 0.732 | 0.777 | 0.725 | 0.740 | 0.778 | 0.735 | 0.772 | **0.820** | 0.840 |
| Mean | 0.616 | 0.644 | 0.594 | 0.643 | 0.652 | 0.623 | 0.631 | **0.659** | 0.696 |

Table 26: **Per-dataset clustering normalized mutual information.**

| Dataset | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| LoveDA | 0.346 | 0.358 | 0.314 | 0.355 | 0.362 | 0.342 | 0.344 | **0.368** | 0.366 |
| Potsdam | 0.405 | 0.419 | 0.412 | 0.409 | **0.445** | 0.438 | 0.439 | 0.387 | 0.410 |
| Vaihingen | 0.413 | 0.440 | 0.349 | 0.425 | 0.454 | 0.446 | **0.460** | 0.438 | 0.521 |
| iSAID | 0.127 | 0.132 | 0.120 | 0.124 | 0.146 | 0.138 | 0.144 | **0.163** | 0.178 |
| DeepGlobe | 0.254 | 0.263 | 0.214 | 0.269 | 0.259 | 0.218 | 0.239 | **0.280** | 0.279 |
| LandCover.ai | 0.299 | 0.297 | 0.257 | **0.308** | 0.286 | 0.273 | 0.275 | 0.285 | 0.297 |
| FloodNet | 0.428 | 0.450 | 0.397 | 0.456 | 0.486 | 0.454 | 0.486 | **0.504** | 0.500 |
| UAVid | 0.376 | 0.401 | 0.342 | 0.404 | 0.432 | 0.407 | 0.422 | **0.450** | 0.478 |
| UDD5 | 0.390 | 0.427 | 0.395 | 0.424 | **0.445** | 0.440 | 0.441 | 0.429 | 0.418 |
| DLRSD | 0.373 | 0.383 | 0.420 | 0.376 | 0.426 | 0.414 | **0.444** | 0.419 | 0.442 |
| WBS | 0.297 | 0.311 | **0.386** | 0.305 | 0.364 | 0.383 | 0.373 | 0.352 | 0.362 |
| OpenEarthMap | 0.269 | 0.271 | 0.247 | 0.267 | 0.293 | 0.270 | 0.306 | **0.333** | 0.386 |
| DFC2020 | 0.177 | 0.180 | 0.216 | 0.183 | 0.205 | 0.213 | **0.223** | 0.206 | 0.205 |
| Five-Billion-Pixels | 0.166 | 0.170 | 0.177 | 0.171 | 0.195 | 0.195 | **0.212** | 0.190 | 0.251 |
| SpaceNet1 | 0.160 | 0.170 | 0.112 | 0.174 | 0.163 | 0.136 | 0.158 | **0.193** | 0.193 |
| SpaceNet2 | 0.205 | 0.183 | 0.177 | 0.199 | 0.185 | 0.148 | 0.178 | **0.207** | 0.252 |
| SpaceNet7 | 0.044 | 0.043 | 0.045 | 0.042 | 0.047 | 0.044 | 0.050 | **0.061** | 0.070 |
| WHU Building | 0.272 | 0.262 | 0.218 | 0.263 | 0.311 | 0.267 | 0.317 | **0.392** | 0.425 |
| Mass. Buildings | 0.043 | 0.044 | 0.030 | 0.045 | 0.042 | 0.041 | 0.048 | **0.049** | 0.072 |
| Mass. Roads | 0.027 | 0.025 | 0.028 | 0.023 | 0.028 | 0.024 | 0.033 | **0.036** | 0.046 |
| xView2 | 0.062 | 0.063 | 0.053 | 0.067 | 0.062 | 0.058 | 0.067 | **0.109** | 0.119 |
| Mean | 0.244 | 0.252 | 0.234 | 0.252 | 0.268 | 0.255 | 0.269 | **0.279** | 0.299 |

Table 27: **Per-dataset STEGO-style unsupervised segmentation mIoU.**

| Dataset | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| LoveDA | **0.217** | 0.198 | 0.179 | 0.172 | 0.174 | 0.177 | 0.177 | 0.194 | 0.192 |
| Potsdam | **0.332** | 0.295 | 0.279 | 0.285 | 0.286 | 0.249 | 0.273 | 0.285 | 0.215 |
| Vaihingen | 0.118 | 0.097 | 0.112 | 0.119 | 0.140 | 0.133 | 0.135 | **0.236** | 0.321 |
| iSAID | 0.053 | 0.053 | 0.055 | 0.038 | **0.056** | 0.055 | 0.050 | 0.054 | 0.062 |
| DeepGlobe | 0.144 | 0.194 | 0.191 | 0.211 | 0.199 | **0.238** | 0.220 | 0.190 | 0.214 |
| LandCover.ai | **0.453** | 0.339 | 0.314 | 0.345 | 0.321 | 0.365 | 0.318 | 0.328 | 0.285 |
| FloodNet | 0.239 | 0.228 | 0.179 | 0.155 | 0.197 | 0.249 | 0.188 | **0.283** | 0.236 |
| UAVid | 0.197 | 0.190 | 0.173 | 0.179 | 0.176 | 0.154 | 0.179 | **0.212** | 0.234 |
| UDD5 | 0.287 | **0.325** | 0.323 | 0.280 | 0.318 | 0.267 | 0.256 | 0.288 | 0.346 |
| DLRSD | 0.288 | **0.291** | 0.276 | 0.273 | 0.251 | 0.236 | 0.281 | 0.228 | 0.189 |
| WBS | **0.538** | 0.367 | 0.367 | 0.372 | 0.367 | 0.369 | 0.367 | 0.484 | 0.469 |
| OpenEarthMap | 0.201 | **0.217** | 0.208 | 0.160 | 0.204 | 0.183 | 0.215 | 0.193 | 0.190 |
| DFC2020 | 0.208 | 0.173 | 0.164 | 0.225 | 0.162 | 0.154 | 0.178 | **0.236** | 0.279 |
| Five-Billion-Pixels | 0.056 | 0.056 | 0.063 | 0.056 | 0.062 | **0.107** | 0.070 | 0.078 | 0.105 |
| SpaceNet1 | 0.441 | 0.374 | 0.377 | 0.373 | 0.374 | 0.375 | 0.374 | **0.471** | 0.410 |
| SpaceNet2 | 0.334 | 0.329 | 0.332 | 0.334 | 0.333 | 0.333 | 0.317 | **0.338** | 0.295 |
| SpaceNet7 | 0.384 | 0.377 | **0.480** | 0.372 | 0.375 | 0.384 | 0.377 | 0.374 | 0.332 |
| WHU Building | 0.256 | 0.269 | 0.278 | 0.288 | 0.271 | 0.307 | 0.277 | **0.325** | 0.321 |
| Mass. Buildings | 0.370 | 0.371 | 0.382 | 0.374 | 0.375 | **0.435** | 0.375 | 0.371 | 0.374 |
| Mass. Roads | 0.419 | 0.420 | **0.426** | 0.424 | 0.424 | 0.425 | **0.426** | **0.426** | 0.326 |
| xView2 | 0.369 | 0.369 | 0.383 | 0.365 | 0.325 | **0.444** | 0.356 | 0.413 | 0.384 |
| Mean | 0.281 | 0.263 | 0.264 | 0.257 | 0.257 | 0.268 | 0.258 | **0.286** | 0.275 |

Table 28: **Per-dataset deep-spectral unsupervised segmentation mIoU.**

| Dataset | lowres | bilinear | JBU | FeatUp | NAF | AnyUp | JAFAR | Ours | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| LoveDA | 0.445 | 0.442 | 0.430 | 0.442 | 0.447 | 0.432 | 0.444 | **0.455** | 0.460 |
| Potsdam | 0.527 | 0.525 | 0.529 | 0.522 | 0.538 | 0.535 | **0.546** | 0.500 | 0.525 |
| Vaihingen | 0.412 | 0.414 | 0.368 | 0.412 | 0.421 | 0.407 | **0.423** | 0.398 | 0.467 |
| iSAID | 0.290 | 0.281 | 0.285 | 0.279 | 0.290 | 0.284 | 0.294 | **0.299** | 0.311 |
| DeepGlobe | 0.420 | 0.411 | 0.397 | 0.414 | 0.417 | 0.380 | 0.407 | **0.422** | 0.433 |
| LandCover.ai | 0.523 | 0.521 | 0.500 | **0.525** | 0.521 | 0.516 | 0.517 | 0.520 | 0.526 |
| FloodNet | 0.529 | 0.515 | 0.504 | 0.530 | 0.549 | 0.523 | 0.554 | **0.556** | 0.577 |
| UAVid | 0.336 | 0.334 | 0.316 | 0.333 | 0.355 | 0.340 | 0.353 | **0.360** | 0.386 |
| UDD5 | 0.475 | 0.476 | 0.481 | 0.472 | 0.496 | 0.486 | **0.498** | 0.474 | 0.474 |
| DLRSD | 0.503 | 0.498 | 0.532 | 0.497 | 0.521 | 0.520 | **0.535** | 0.516 | 0.534 |
| WBS | 0.561 | 0.556 | **0.596** | 0.559 | 0.579 | 0.594 | 0.595 | 0.577 | 0.591 |
| OpenEarthMap | 0.302 | 0.288 | 0.287 | 0.283 | 0.301 | 0.293 | 0.314 | **0.330** | 0.357 |
| DFC2020 | 0.183 | 0.184 | **0.203** | 0.185 | 0.195 | 0.197 | 0.202 | 0.190 | 0.189 |
| Five-Billion-Pixels | 0.116 | 0.113 | 0.118 | 0.115 | 0.123 | 0.127 | **0.133** | 0.123 | 0.152 |
| SpaceNet1 | 0.442 | 0.445 | 0.423 | 0.450 | 0.446 | 0.442 | 0.444 | **0.464** | 0.472 |
| SpaceNet2 | 0.471 | 0.457 | 0.462 | 0.472 | 0.458 | 0.434 | 0.459 | **0.478** | 0.499 |
| SpaceNet7 | 0.329 | 0.325 | **0.338** | 0.326 | 0.328 | 0.328 | 0.335 | 0.335 | 0.347 |
| WHU Building | 0.543 | 0.527 | 0.501 | 0.526 | 0.552 | 0.519 | 0.562 | **0.597** | 0.628 |
| Mass. Buildings | 0.344 | 0.342 | 0.346 | 0.341 | 0.343 | 0.339 | **0.348** | 0.340 | 0.351 |
| Mass. Roads | 0.326 | 0.316 | **0.333** | 0.319 | 0.323 | 0.320 | 0.329 | **0.333** | 0.341 |
| xView2 | 0.371 | 0.365 | 0.377 | 0.367 | 0.365 | 0.366 | 0.379 | **0.402** | 0.416 |
| Mean | 0.402 | 0.397 | 0.397 | 0.399 | 0.408 | 0.399 | **0.413** | **0.413** | 0.430 |

# I Queries Supplied by the Task

Every other protocol in this paper draws query centers uniformly, which measures how well an arbitrary region is served. This section repeats the study with queries of the kind an application issues (Ebel et al., 2026), so that the region is one a task has already selected. The arm differs from the main protocol in the query alone, since the images, the readouts, the episode seeds, and the feature sources are those of Section 3, and the uniform arm reported beside it runs the main protocol unchanged and reproduces its numbers for every source.

Queries are placed from the annotation without any of it reaching the model, which still receives only the global view, the point query, and the queried box of that view. Within each image we take the connected components of every foreground class, center a query on the bounding box of a component, and give it the smallest radius whose support disc contains that box, clipped into the trained range of Section 2.1. A component counts as resolvable when at least eight patches of its class fall inside the support on the prediction grid, which excludes objects finer than that grid can carry, and we keep up to two resolvable components per image under the separation rule the sampler already enforces. Object scale asks for the top of the trained range, since 93 percent of these queries sit at the smallest radius and their mean magnification is 4.3 against 3.0 for uniform queries. Five datasets qualify because their annotation delineates individual objects, whereas a connected component of a land-cover or road map is a region of arbitrary extent whose bounding box carries no scale an application would request.

Two consequences of the construction bound what the comparison can say. The arms score different crop populations, since 57 of the 600 images hold no resolvable component and contribute nothing to the object arm, so the two halves of Table 29 are not two measurements of one quantity and the claim we draw is the ordering of sources inside each arm. The teacher also gains from the change, as the limitations of Section 6 anticipate, because these queries sit at the magnifications where fine detail cannot be inferred from context, so this protocol speaks about the sources that can be deployed at interactive cost rather than about the teacher.

Table 29: **Queries placed by the task.** Every readout on each of the five remote sensing datasets whose annotation delineates individual objects, with the mean over them closing each block. Queries are drawn uniformly in the left half, as everywhere else in the paper, and centered on an annotated object in the right half, with the radius that fits the object into the support. Bold marks the best source excluding the teacher.

| | Queries drawn uniformly | | | | Queries placed on objects | | | |
|---|---|---|---|---|---|---|---|---|
| Dataset | lowres | bilinear | Ours | Teacher | lowres | bilinear | Ours | Teacher |
| INSID3, 1 shot | | | | | | | | |
| iSAID | 0.264 | 0.228 | **0.434** | 0.512 | 0.193 | 0.187 | **0.307** | 0.426 |
| WHU Building | 0.573 | 0.369 | **0.679** | 0.637 | 0.447 | 0.339 | **0.762** | 0.765 |
| SpaceNet2 | 0.358 | 0.323 | **0.399** | 0.358 | 0.370 | 0.305 | **0.443** | 0.441 |
| Mass. Buildings | 0.197 | 0.197 | **0.215** | 0.246 | 0.195 | 0.191 | **0.217** | 0.293 |
| xView2 | 0.220 | 0.205 | **0.412** | 0.471 | 0.211 | 0.193 | **0.397** | 0.390 |
| Mean | 0.322 | 0.264 | **0.428** | 0.445 | 0.283 | 0.243 | **0.425** | 0.463 |
| INSID3, 5 shots | | | | | | | | |
| iSAID | 0.297 | 0.290 | **0.509** | 0.530 | 0.232 | 0.243 | **0.408** | 0.481 |
| WHU Building | 0.595 | 0.375 | **0.746** | 0.776 | 0.502 | 0.336 | **0.762** | 0.808 |
| SpaceNet2 | 0.402 | 0.322 | **0.495** | 0.390 | 0.347 | 0.315 | **0.477** | 0.515 |
| Mass. Buildings | 0.177 | 0.199 | **0.206** | 0.262 | 0.076 | 0.187 | **0.205** | 0.298 |
| xView2 | 0.239 | 0.202 | **0.462** | 0.518 | 0.252 | 0.197 | **0.453** | 0.550 |
| Mean | 0.342 | 0.278 | **0.484** | 0.495 | 0.282 | 0.256 | **0.461** | 0.530 |
| FROST, 1 shot | | | | | | | | |
| iSAID | 0.222 | 0.217 | **0.361** | 0.454 | 0.171 | 0.122 | **0.300** | 0.414 |
| WHU Building | 0.186 | 0.254 | **0.627** | 0.599 | 0.295 | 0.248 | **0.750** | 0.722 |
| SpaceNet2 | 0.162 | 0.049 | **0.407** | 0.318 | 0.262 | 0.225 | **0.462** | 0.474 |
| Mass. Buildings | 0.075 | 0.038 | **0.209** | 0.307 | 0.052 | 0.074 | **0.315** | 0.427 |
| xView2 | 0.114 | 0.083 | **0.506** | 0.497 | 0.110 | 0.194 | **0.401** | 0.511 |
| Mean | 0.152 | 0.128 | **0.422** | 0.435 | 0.178 | 0.172 | **0.446** | 0.510 |
| FROST, 5 shots | | | | | | | | |
| iSAID | 0.226 | 0.237 | **0.505** | 0.572 | 0.179 | 0.168 | **0.415** | 0.557 |
| WHU Building | 0.285 | 0.376 | **0.787** | 0.820 | 0.257 | 0.220 | **0.779** | 0.836 |
| SpaceNet2 | 0.144 | 0.116 | **0.595** | 0.476 | 0.240 | 0.200 | **0.535** | 0.577 |
| Mass. Buildings | 0.012 | 0.006 | **0.216** | 0.348 | 0.002 | 0.002 | **0.307** | 0.471 |
| xView2 | 0.062 | 0.083 | **0.540** | 0.495 | 0.085 | 0.088 | **0.543** | 0.615 |
| Mean | 0.146 | 0.163 | **0.529** | 0.542 | 0.153 | 0.136 | **0.516** | 0.611 |
| Linear probe | | | | | | | | |
| iSAID | 0.273 | 0.307 | **0.360** | 0.453 | 0.247 | 0.258 | **0.291** | 0.396 |
| WHU Building | 0.776 | 0.822 | **0.874** | 0.916 | 0.727 | 0.784 | **0.859** | 0.905 |
| SpaceNet2 | 0.735 | 0.765 | **0.772** | 0.790 | 0.683 | 0.718 | **0.742** | 0.756 |
| Mass. Buildings | 0.420 | 0.430 | **0.551** | 0.672 | 0.394 | 0.401 | **0.582** | 0.707 |
| xView2 | 0.648 | 0.672 | **0.725** | 0.778 | 0.622 | 0.654 | **0.732** | 0.802 |
| Mean | 0.570 | 0.599 | **0.656** | 0.722 | 0.535 | 0.563 | **0.641** | 0.713 |

# J CHOOSING WHERE TO LOOK

Section 4.3 closes the loop the rest of the paper leaves open, so the system picks its own queries instead of receiving them. This section gives the construction and the controls.

**Candidates and assembly.** A fixed grid of $7 \times 7$ point queries covers the frame, each at radius $0.125/(1+\beta)$, so its box spans a quarter of the frame and its magnification is 4.0, near the top of the trained range. Each candidate is decoded on a $16^2$ grid and one candidate owns every cell of a $64^2$ canvas, an assignment that leaves the canvas coordinates at the uniform grid of cell centers whichever candidate owns them, so all sources land on identical coordinates and the ground truth is a nearest-neighbor resize of the canonical mask. Selecting a subset of candidates is then exact masking over one full assembly, which is why a single extraction serves every budget and every policy.

**The policy.** A prototype is the mean of the $\ell_2$-normalized coarse features inside the reference masks, pooled over the reference images, and a candidate scores as the largest cosine to that prototype among the coarse cells it owns. Both quantities come from the pass that produced the coarse map, so ranking the candidates costs no further backbone work. The reference images are re-encoded at every candidate, since a reference and its mask are given in advance and can be prepared offline.

**Evaluation.** Five datasets whose canonical frames exceed the view carry the study, iSAID, UAVid, Vaihingen, FloodNet, and UDD5, at 120 images each, split by image parity into a reference half and a scored half. Every readout runs unmodified at the canvas resolution and is scored over the whole image, with three episode seeds and 25 query images, 15 under INSID3. A class counts as small-object when its area-weighted mean ground-truth component, $\sum_k a_k^2 / \sum_k a_k$ over components of area $a_k$, is at most 64 canvas cells. A median cannot serve, since roads and trees also break into components of a few cells, whereas the area-weighted mean separates the vehicle and small-structure classes of each suite from everything else by more than an order of magnitude.

**What the budget buys.** Table 30 sweeps the budget. Eight regions cost 11 milliseconds of decoding beyond the 23 the view costs and recover 59 to 82 percent of what re-encoding all forty-nine buys on the small-object classes, at a sixth of that decoding; over all classes, where the labels cover the frame, they recover 31 to 55 percent. Placing the same eight from the annotation rather than from the prototype gains 0.009 to 0.055 mIoU, and the eight the policy picks hold 51 percent of the ground-truth foreground against 73 percent for that oracle, so the policy concedes coverage but little accuracy, and sixteen regions match the oracle under FROST. The teacher row is a ceiling rather than an alternative, since it spends a backbone pass per region, and its lead here is the one Appendix N predicts, every dataset of this study having a canonical frame wider than the view.

**Both sides of the match must be re-encoded.** The gain is not a property of the reference side. Re-encoding the references while leaving the query on the coarse map moves the small-object mean by $-0.035$ to $+0.004$, whereas re-encoding the query as well moves it by $+0.027$ to $+0.221$. Re-encoding the query against coarse references recovers much of that, and adding the reference side on top of it gains a further 0.001 to 0.069, most of all on the vehicle classes. The two sides therefore interact, and a readout that compares individual patches wants both of its arguments at the scale it compares them at.

**Scope.** These numbers compare feature maps under one port and are not a ranking of methods, as every crop-grid result in this paper is. The teacher is read at the canvas grid rather than at its own finer one, which is the comparison the canvas defines but understates what a backbone pass per region resolves.

# K SIGNIFICANCE TESTS

Table 31 reports the full test results underlying the significance marks in the main paper. The pattern separates cleanly by readout family, with matching and per-patch regression readouts favoring

Table 30: **Choosing where to look, every budget.** Mean mIoU over the five datasets, scored on whole images, with each readout at one and five shots. The left half averages every scored class and the right half the small-object classes, with I for INSID3, F for FROST, and the digit giving the shot count. The four- and sixteen-region rows were run under FROST alone. Bold marks the best row that neither consults the annotation nor spends a backbone pass per region.

| | All classes | | | | Small objects | | | |
|---|---|---|---|---|---|---|---|---|
| Feature map | I-1 | I-5 | F-1 | F-5 | I-1 | I-5 | F-1 | F-5 |
| lowres | 0.382 | 0.472 | 0.412 | 0.508 | 0.248 | 0.284 | 0.226 | 0.313 |
| bilinear | 0.349 | 0.424 | 0.425 | 0.523 | 0.163 | 0.180 | 0.261 | 0.340 |
| Ours, four regions | – | – | 0.428 | 0.531 | – | – | 0.269 | 0.389 |
| Ours, eight regions | 0.409 | 0.501 | 0.438 | 0.546 | 0.328 | 0.401 | 0.288 | 0.417 |
| Ours, sixteen regions | – | – | 0.451 | 0.563 | – | – | 0.304 | 0.434 |
| Eight regions from the annotation | 0.445 | 0.529 | 0.448 | 0.555 | 0.383 | 0.428 | 0.304 | 0.430 |
| Ours, all forty-nine | **0.470** | **0.564** | **0.465** | **0.577** | **0.368** | **0.426** | **0.331** | **0.442** |
| Teacher, all forty-nine | 0.471 | 0.566 | 0.480 | 0.588 | 0.430 | 0.500 | 0.401 | 0.500 |

Table 31: **Significance tests.** Two-sided Wilcoxon signed-rank tests over the remote sensing datasets. The middle block compares our student against the teacher and the bottom block against whichever upsampler, bilinear interpolation included, is strongest on the dataset at hand, so its mean difference need not equal the difference between the means of Table 1. Positive mean differences favor the student, and for depth RMSE the sign is flipped so that positive also favors the student.

| | | vs. teacher | | | vs. best upsampler | | |
|---|---|---|---|---|---|---|---|
| Metric | Shots | wins | difference | $p$ | wins | difference | $p$ |
| INSID3 | 1 | 14/21 | $+0.043$ | 0.065 | 18/21 | $+0.082$ | $4\times10^{-5}$ |
| | 5 | 13/21 | $+0.059$ | 0.011 | 21/21 | $+0.128$ | $10^{-6}$ |
| FROST | 1 | 13/21 | $+0.020$ | 0.37 | 21/21 | $+0.198$ | $10^{-6}$ |
| | 5 | 13/21 | $+0.024$ | 0.29 | 21/21 | $+0.286$ | $10^{-6}$ |
| Linear probe | – | 7/21 | $-0.022$ | 0.089 | 13/21 | $+0.001$ | 0.84 |
| Trained decoder | – | 8/21 | $-0.012$ | 0.36 | 12/21 | $-0.001$ | 0.81 |
| Depth $\delta_1$ | – | 18/21 | $+0.012$ | $10^{-4}$ | 20/21 | $+0.025$ | $3\times10^{-6}$ |
| Depth RMSE | – | 21/21 | $+0.100$ | $10^{-6}$ | 21/21 | $+0.169$ | $10^{-6}$ |
| CAM average gain | – | 19/19 | $+16.2$ | $4\times10^{-6}$ | 19/19 | $+16.8$ | $4\times10^{-6}$ |
| Classification | – | 12/19 | $+0.017$ | 0.067 | 9/19 | $+0.005$ | 0.50 |

the student and fitted readouts favoring the teacher, in agreement with the averaging law. The comparison against the strongest upsampler is significant for every matching and per-patch regression metric and indistinguishable from zero for the fitted readouts, where the upsamplers already sit close to the teacher. Read row by row against the teacher the evidence is uneven, and it is worth saying where it is strong. The reversal is individually significant under five-shot INSID3 and under all three per-patch regression metrics, holding on 18 of 21 datasets for depth threshold accuracy, on 21 of 21 for depth error and on 19 of 19 for CAM faithfulness, whereas one-shot INSID3 and both FROST rows lead in the mean and on 13 or 14 of the 21 datasets without reaching significance on 21 paired samples. What the averaging law predicts is which family a readout belongs to rather than the size of any single margin, so we read the table by family and claim no significance for the matching rows one at a time.

## L A Sufficient Condition for the Reversal

Section 5.1 attributes the reversal to variance reduction across a corpus, and that account can be written as a condition on the readout, which is what lets it predict an ordering rather than summarize one. Write the teacher as the student plus a residual, $\eta = T - Y$, which is an identity and not an assumption, so that every comparison between the two sources is a question about how a readout treats $\eta$. What makes the identity informative is where the residual comes from. Let $\Pi$ denote the projection, in mean square under the distribution of views and queries the modulator is trained on, onto the class of maps a modulator of its capacity can realize, a projection the objective of Equation (8) fits because a cosine distance between two features is an affine function of the squared distance between their normalized versions. A converged student approximates $\Pi T$, and $\eta$ is then the part of the teacher that no member of that class reaches. Two properties follow and neither is stochastic. The residual is a deterministic function of the crop, so it is not a sampling error that repetition or ensembling could remove, and it is orthogonal to the class, so it differs between inputs that the class maps to nearly the same output, which is to say between crops that look alike in the view. The proposition below is stated for the residual itself, so it holds for whatever function the modulator converges to and does not rest on how closely that function approaches the projection, and the quantities it turns on are measurable, as Table 32 reports.

**Proposition 1.** *Let a readout decide from patch features in the metric it whitens them into, and let the teacher be the student plus $\eta$ as above. Then the following hold. (a) If the readout labels a query patch by its nearest support patch and the support patches come from other images, the teacher scores below the student whenever $\eta$ raises the squared distance between patches of one class by more than it raises the squared distance between patches of different classes. (b) Let the readout instead average $n$ patches per class, or fit a linear map by least squares over patches from many images, and let the values of $\eta$ on those patches be approximately uncorrelated across images. Then the part of $\eta$ uncorrelated with the label enters the decision in proportion to $n^{-1/2}$ while the part correlated with it does not diminish, so the teacher gains on the student as $n$ grows and overtakes it once that correlated part is nonzero. (c) If support and query patches come from the same scene, $\eta$ is common to both and therefore raises between-class distance without raising within-class distance, which reverses the inequality of (a).*

The three parts follow from one observation each. A nearest-neighbor decision errs when a within-class distance exceeds a between-class distance, so an increment that enlarges the former by more than the latter can only raise the error rate, which gives (a). An increment that is reproducible within one scene cancels in the difference between two patches of the same object there and does not cancel between patches of different objects, which gives (c). For (b), an average of $n$ terms whose fluctuations are uncorrelated shrinks those fluctuations by $n^{-1/2}$ while leaving any component they share intact, and a least-squares fit is such an average in the patches it is fitted on. What the proposition supplies is the mechanism and the direction, not the assignment of a readout to a case, which we make from the protocol of each readout by how broadly it pools before deciding.

Each condition is met by a different part of the evidence. Condition (a) holds in the matching family of Table 1, and Appendix N rules out the readings of $\eta$ under which it would fail, since neither ensembling nor smoothing reproduces the effect and the margin is largest where the teacher sees exactly the pixels the student sees, so $\eta$ is a limit of the class rather than information the input withholds. Condition (b) is the recovery of the fitted and pooled readouts in Section 5.2, where the teacher leads by 2.0 to 5.3 percent and the prototype classifiers, which average their support before

matching, give it the largest margins. Condition (c) is the same-scene exception of Section 5.4, where propagation within one time series or one video is scored against supports drawn from the same scene.

**Measuring the residual.** Those are arguments from outcomes, and the conditions are also statements about $\eta$ itself, which is observable. We take the teacher and the student on the same query crops of every evaluation set, apply the positional debiasing the matching readouts apply, and whiten by the pooled within-class covariance, which is the metric those readouts decide in, and in that metric we give each source a class separation, the mean squared distance between patches of different classes minus the mean squared distance between patches of one class. Computing it over patch pairs drawn from different images turns condition (a) into the statement that the student separates classes better than the teacher, since the teacher is the student plus $\eta$, and computing it over pairs drawn from one image turns condition (c) into the statement that this comparison reverses. Table 32 reports both over the 21 remote sensing datasets.

Across images the student separates classes better on 17 of the 21 datasets, by 0.027 on a teacher separation of 0.141, at Wilcoxon $p = 0.0009$. Within one image the difference turns negative, at $-0.011$ on 14 of the 21, and the swing between the two regimes runs the same way on every dataset without exception, by 0.039 at $p = 10^{-6}$, which is the sharper statement of the two because it is measured on the same patches in the same metric with the same pair of sources and so isolates the pairing rather than the data. The cross-image statistic also does more than carry the right sign, since it orders the datasets, correlating with the margin the student actually holds over the teacher across the four matching readouts at Pearson 0.47 with $p = 0.03$, and at 0.54 with $p = 0.01$ under one-shot INSID3. Where the residual costs a nearest-neighbor decision most is therefore where the reversal is largest. The four benchmarks outside remote sensing point the same way, with the cross-image difference positive on all four and the swing between the two regimes positive on all four, although four paired samples settle nothing on their own.

Condition (b) needs $\eta$ to carry a component that follows the label, since a residual independent of it would be pooled away without returning anything to the teacher. It carries one. The fraction of the variance of $\eta$ that its class means explain averages 0.016 over the same datasets against at most 0.0004 when those labels are permuted, and a linear classifier fitted on $\eta$ alone, with access to neither source, reaches 0.512 mean per-class accuracy against a chance level of 0.269 and beats chance on all 21. What the teacher holds and the student does not is therefore structure that a readout pooling many patches can recover and a winner-take-all match against single support patches cannot, which is the asymmetry the law names.

Two things stay assumptions. The identification of the trained modulator with the projection $\Pi$ motivates the two properties of $\eta$ but is not itself measured, which is why the proposition is stated for the residual and nothing above rests on it. And the lack of correlation across images that part (b) assumes of the label-free component of $\eta$ is an idealization, so the $n^{-1/2}$ rate gives the direction of that effect rather than its size.

## M The Control Teacher on Downstream Readouts

Section 2.1 rests the case that looking closer pays through re-encoding rather than through extra pixels on a patch cosine of 0.986 between the control teacher $T^v$, which is the frozen backbone applied to the query box cut out of the global view, and the zoom teacher $T$, which cuts the same box at native resolution. A cosine that high need not imply that the difference is downstream-irrelevant, and this paper supplies its own counterexample, since the gate variants of Appendix T differ by thousandths of a cosine at the center of the support and by 0.130 mIoU under FROST. We therefore put $T^v$ through the readouts themselves.

The six sets are the ones on which the question can be settled, since on the eleven sets whose canonical images are at or below the view resolution the two crops carry the same pixels and the two teachers coincide by construction. Ordered by the width of the canonical frame in units of the view, they are FloodNet at 5.9, DeepGlobe at 4.8, UAVid at 4.2, Massachusetts Buildings and Massachusetts Roads at 2.9, and LoveDA at 2.0, the width at which Section 2.1 measures its cosine. Over the query crops of the evaluation suite that cosine falls with the width, from 0.981 on LoveDA to 0.931 on FloodNet, so these are the conditions under which the argument of Section 2.1 is weakest.

Table 32: **Geometry of the residual.** Per evaluation set, the class separation of each teacher and student, meaning the mean squared distance between patches of different classes minus the mean squared distance between patches of the same class, measured in the metric the matching readouts decide in. The first three columns take both patches of a pair from different images and the fourth takes both from one image. The label energy is the fraction of the variance of the residual explained by its class means, which never exceeds 0.0004 when those labels are permuted. The last column is the mean margin of the student over the teacher across the four in-context matching readouts of Table 1.

| | Across images | | | One image | Label | Matching |
|---|---|---|---|---|---|---|
| Dataset | teacher | student | difference | difference | energy | margin |
| DeepGlobe | 0.163 | 0.171 | +0.008 | -0.012 | 0.0102 | +0.036 |
| DFC2020 | 0.177 | 0.218 | +0.041 | +0.003 | 0.0472 | +0.081 |
| DLRSD | 0.192 | 0.261 | +0.069 | -0.002 | 0.0325 | +0.060 |
| Five-Billion-Pixels | 0.139 | 0.098 | -0.040 | -0.049 | 0.0672 | -0.067 |
| FloodNet | 0.317 | 0.377 | +0.060 | +0.029 | 0.0230 | +0.049 |
| iSAID | 0.154 | 0.197 | +0.043 | +0.000 | 0.0100 | -0.064 |
| LandCover.ai | 0.157 | 0.188 | +0.031 | -0.003 | 0.0113 | +0.138 |
| LoveDA | 0.121 | 0.161 | +0.041 | +0.004 | 0.0138 | +0.086 |
| Mass. Buildings | 0.020 | 0.004 | -0.016 | -0.035 | 0.0037 | -0.079 |
| Mass. Roads | 0.030 | 0.032 | +0.002 | -0.027 | 0.0010 | -0.060 |
| OpenEarthMap | 0.191 | 0.225 | +0.033 | -0.029 | 0.0081 | +0.095 |
| Potsdam | 0.097 | 0.152 | +0.055 | -0.022 | 0.0302 | +0.154 |
| SpaceNet1 | 0.054 | 0.073 | +0.020 | -0.018 | 0.0017 | +0.040 |
| SpaceNet2 | 0.078 | 0.072 | -0.006 | -0.039 | 0.0048 | +0.088 |
| SpaceNet7 | 0.028 | 0.030 | +0.001 | -0.008 | 0.0009 | -0.023 |
| UAVid | 0.256 | 0.309 | +0.053 | +0.011 | 0.0109 | -0.014 |
| UDD5 | 0.183 | 0.248 | +0.064 | +0.022 | 0.0083 | +0.113 |
| Vaihingen | 0.236 | 0.219 | -0.017 | -0.050 | 0.0379 | -0.063 |
| WBS | 0.068 | 0.098 | +0.031 | -0.040 | 0.0041 | +0.214 |
| WHU Building | 0.149 | 0.210 | +0.061 | +0.043 | 0.0071 | +0.002 |
| xView2 | 0.147 | 0.189 | +0.043 | -0.016 | 0.0010 | -0.015 |
| Mean | 0.141 | 0.168 | +0.027 | -0.011 | 0.0160 | +0.037 |

Table 33: **The control teacher downstream.** Means over the six evaluation sets whose canonical images are widest relative to the view, where the two crops differ most and the comparison therefore has the most to reveal. The control teacher re-encodes the query box cut from the global view and so sees only the pixels the student sees, whereas the teacher cuts the same box at native resolution. Few-shot segmentation and the probe report mIoU and depth reports $\delta_1$, all higher-better. No source is marked best, because the control teacher is a diagnostic rather than a deployable one, spending a backbone pass per query as the teacher does.

| Readout | bilinear | Ours (main) | Control teacher | Teacher |
|---|---|---|---|---|
| INSID3, 1 shot | 0.306 | 0.390 | 0.365 | 0.374 |
| INSID3, 5 shots | 0.332 | 0.445 | 0.406 | 0.411 |
| FROST, 1 shot | 0.172 | 0.358 | 0.373 | 0.386 |
| FROST, 5 shots | 0.210 | 0.465 | 0.457 | 0.475 |
| Linear probe | 0.488 | 0.531 | 0.574 | 0.588 |
| Depth $\delta_1$ | 0.303 | 0.330 | 0.320 | 0.320 |

Table 34: **Controls for the reversal.** Every row is a feature source evaluated under the protocol of Table 1, averaged over the 21 remote sensing datasets of Table 1. Few-shot rows report mIoU at one and five shots and the probe column reports linear-probe mIoU, all higher-is-better, and the last column is the mean patch cosine to the zoom teacher. The ensembles average eight encodings of the same crop under photometric jitter, under photometric jitter and sub-patch translation, and under displacement and rescaling of the query box. The matched low-pass row uses the blur width at which the teacher carries the same high-frequency energy as ours, and the row below it filters by a full patch.

| Source | FROST 1 | FROST 5 | INSID3 1 | INSID3 5 | Probe | Cosine |
|---|---|---|---|---|---|---|
| Bilinear | 0.174 | 0.216 | 0.322 | 0.345 | 0.560 | 0.6362 |
| Zoom teacher | 0.379 | 0.488 | 0.378 | 0.432 | 0.605 | 1.0000 |
| Ours | 0.399 | 0.512 | 0.421 | 0.491 | 0.583 | 0.7837 |
| *The zoom teacher, modified* | | | | | | |
| Photometric ensemble | 0.378 | 0.488 | 0.380 | 0.431 | 0.605 | 0.9995 |
| Photometric and shift ensemble | 0.365 | 0.473 | 0.374 | 0.428 | 0.605 | 0.9933 |
| Query-box ensemble | 0.341 | 0.437 | 0.368 | 0.416 | 0.600 | 0.9779 |
| Low-pass matched to ours | 0.365 | 0.471 | 0.375 | 0.425 | 0.604 | 0.9963 |
| Low-pass, one patch | 0.303 | 0.398 | 0.344 | 0.380 | 0.577 | 0.9562 |

Table 33 nonetheless finds the two teachers close. Of the lead the teacher holds over interpolation of the coarse map, the control teacher keeps 94 percent at one-shot FROST and 93 percent at five shots across all six sets, and 92 and 93 percent under INSID3 across the four sets where that lead exceeds 0.02 mIoU and the ratio is meaningful at all. Per-patch depth is closer still, with the two teachers within 0.0003 in $\delta_1$ on every set while interpolation trails by up to 0.040. The extra pixels are not worth nothing, and the linear probe is where they show, with the control teacher keeping 78 percent of the lead and falling 0.055 mIoU behind the teacher on FloodNet, the widest set of the six. The largest single gap on any row is 0.064 mIoU at five-shot FROST on Massachusetts Buildings.

Two cautions apply to reading this table. It selects the sets where the canonical frame is widest, which is also where our own margin over the teacher is smallest, so the ordering between the student and the teacher here is not the one the 21-set means of Table 1 report and nothing about the averaging law should be read off these columns. And the split it measures is between re-encoding at the scale of the query and resolving finer pixels, not between our method and the teacher, since the student is trained against $T$ throughout.

## N Controls for the Reversal

The student is smoother than the teacher, and the readouts that prefer it whiten features and then commit to a single support patch, so several accounts of the reversal compete with the one Section 5.1 gives. Table 34 tests them by constructing further feature sources and pushing each through the protocol of Table 1 without any other change, so that every number in it is comparable to Table 1 row for row.

The first account is that the teacher carries encoder noise which an explicit ensemble would remove as distillation does. It does not survive measurement, because the teacher is a deterministic function of its crop, so its variability exists only relative to nuisance transformations of that crop and those turn out to be small. Averaging eight encodings under photometric jitter leaves the mean at cosine 0.9995 of the single encoding, adding sub-patch translation leaves it at 0.9933, and displacing and rescaling the query box leaves it at 0.9779. The student sits at 0.7837 from the teacher,[1] so the teacher moves under nuisance by about a tenth of the distance the student covers. Under 8 percent of the squared distance between the student and the teacher lies along the directions these ensembles average out, and no ensemble improves any readout by more than 0.002 mIoU, whereas the student improves the four matching readouts by 0.020 to 0.059.

[1]This value averages the 21 remote sensing datasets of Table 1 under the query protocol of that table. The 0.847 quoted in Section 2.1 is a different measurement, taken on held-out images whose canonical frames are twice as wide as the view so that the two crops of Equation (3) genuinely differ.

The second account is smoothness. Measuring roughness as the squared distance from each patch to the mean of its four neighbors divided by the spread of the map about its own mean, the teacher sits at 0.195 and the student at 0.101. Filtering the teacher down to 0.101 costs it accuracy on all four matching readouts instead of gaining any, as does filtering it by a full patch, so the smoothing that distillation produces is not what a matching readout rewards. Whitening likewise absorbs isotropic high-frequency content, so these readouts are not sensitive to roughness as such, and the visible smoothing in Figure 3 is a by-product of the averaging rather than its mechanism.

The third account is a position-indexed artifact of the kind that registers and test-time decomposition remove. Registers are not at issue, since the backbone we distill from already carries four of them and what they remove is therefore absent by construction. The remaining possibility is tested by subtracting the position-dependent part of the teacher's mean over crops, an estimate that uses no labels and nothing from our model. Over the six datasets on which we ran it the correction trades one matching family for the other, gaining 0.045 and 0.068 mIoU under INSID3 at one and five shots while losing 0.071 and 0.118 under FROST. Applying the same correction to every source rather than to the teacher alone shows why it cannot arbitrate. On those same six datasets, whose baselines differ from the 21-dataset means of Table 1 because of the subset, it moves the student's one-shot FROST accuracy from 0.438 to 0.125 and the coarse map's from 0.199 to 0.176, while moving the teacher's only from 0.400 to 0.329. These pipelines are therefore sensitive to whatever patches share at a grid position, and shifting it perturbs sources by amounts unrelated to what they encode. No modification of the teacher that we tried improves both matching families at once, which the student does.

A last control tests our own account rather than a competitor, because the averaging admits two readings and the suite separates them. Were the averaged quantity something the student's input leaves undetermined, the reversal would be largest where the teacher sees most that the student cannot, and Appendix B identifies those datasets as the fourteen whose canonical images are wider than the view. The measurement inverts that prediction. Averaged over the four in-context matching readouts, our margin over the teacher is 0.006 mIoU on those fourteen and 0.095 on the eleven whose canonical images are at or below the view resolution, where the two crops of Equation (3) carry the same pixels and the teacher is a deterministic function of an image the student itself receives. A rank-sum test puts that difference at $p = 0.011$, and the margin falls with the ratio of canonical width to view width at Spearman correlation $-0.59$ over the 25 datasets at $p = 0.002$. The ordering also survives inside remote sensing alone, at $-0.011$ mIoU against 0.084 under one-shot FROST across the fourteen wider datasets and the seven whose resolution matches the view, so it is not a difference between domains. Per-patch depth agrees in direction and more weakly, at 0.009 against 0.016 in $\delta_1$ without reaching significance. What distillation removes is therefore not a component the input fails to determine but one that a smooth function of fixed capacity cannot fit crop by crop, and its removal pays most where no missing pixel competes with it. This also explains why no perturbation of a single crop reproduces the effect, since the average that matters is taken across the corpus rather than within an image.

## O Component Ablations

Table 35 decomposes the prediction rather than ranking variants against one another, since each row removes one thing the decoder reads and the columns record what that removal costs. Reconstruction alone cannot settle such a question, because Appendix T shows that a difference invisible in cosine can still decide a readout, so every variant is carried through one readout of each kind that Section 5.2 distinguishes, in-context matching at both shot counts, a fitted linear probe, and per-patch depth regression under two metrics.

Three conclusions follow, and the first concerns where the content comes from. Withholding the coarse map, so that the crop pixels together with the coordinates and the query are the only inputs while every tensor shape and therefore the parameter count is left unchanged, does not merely degrade the prediction but places it below interpolation of that same map on all five readouts, at 0.358 against 0.613 mIoU under the probe and 2.480 against 2.278 in depth error. A network of our capacity that sees the queried pixels and is told where they lie cannot reach what bilinear sampling reaches with no training at all. The wide-field encoding is therefore what supplies the content. Read together with Appendix M, where the pixels the teacher sees and the student does not turn out to

Table 35: **Component ablations.** Every trained variant runs for a matched 4k steps except in the last row, and the identity row equals bilinear sampling of the coarse map. The first column is the patch cosine to the zoom teacher on held-out crops, and the remaining five carry the same variant through readouts of Section 3, averaged over SpaceNet2, Vaihingen, Potsdam, WHU Building, and iSAID. Bold marks the best entry in each column.

| | | FROST mIoU | | Linear probe | Depth | |
|---|---|---|---|---|---|---|
| Variant | Cosine ↑ | 1 shot ↑ | 5 shots ↑ | mIoU ↑ | $\delta_1$ ↑ | RMSE ↓ |
| Decoding without the coarse map | 0.628 | 0.244 | 0.295 | 0.358 | 0.249 | 2.480 |
| Identity initialization (bilinear) | 0.635 | 0.222 | 0.281 | 0.613 | 0.272 | 2.278 |
| Modulation without coordinate decoder | 0.709 | 0.364 | 0.467 | 0.615 | 0.276 | 2.236 |
| Query-free decoding | 0.812 | 0.422 | 0.572 | 0.629 | 0.290 | 2.144 |
| Full method, 4k steps | 0.813 | 0.424 | 0.574 | 0.630 | 0.292 | 2.141 |
| Decoding without the crop pixels | 0.817 | 0.436 | 0.543 | 0.621 | 0.295 | 2.126 |
| Full method, converged | **0.847** | **0.478** | **0.623** | **0.638** | **0.302** | **2.068** |

be worth little, this fixes what the problem actually is. The encoding of a region at the scale of the query is largely reachable from the encoding of the whole view, so the modulator has to remap that encoding under the query rather than rebuild it from the crop, which is why a glance costs a fraction of a forward pass. It does not make the prediction an upsampled coarse map, since every upsampler in Table 1 reads the same two inputs and is separated from ours by the target it is fitted to rather than by the input it is given, as Section 5.3 argues.

The second concerns what the coordinate-conditioned decoder contributes. Replacing it by feature-wise modulation alone costs 0.104 cosine, and the readouts locate that loss precisely, since the probe falls by 0.015 mIoU while five-shot FROST falls by 0.107. The decoder is thus not a generic improvement in feature quality, which a readout that pools many patches would have registered, but the component that makes individual patches discriminable against one another, exactly the regime in which Section 5.2 predicts such a component to matter.

The third concerns the query and the pixels, and reading it needs the spread between runs. A repetition of the matched 4k recipe, trained alongside the ablations, lands 0.006 from its twin in cosine and 0.021 at one-shot FROST, so differences below that are not read here. Conditioning the decoding on the point query beyond crop selection then moves no column, which confirms rather than weakens the account of Section 2.2. The query is not idle under that removal, since it still fixes which patches may change through the gate of Equation (6), the grid the prediction is rendered on, and the scale at which the target is defined. What the ablation removes is the further embedding the decoder receives, and the interface loses nothing by it, because the query addresses a region instead of shaping the features inside it. Withholding the crop pixels leaves the two depth columns and the cosine unchanged within that spread and costs 0.031 mIoU at five-shot FROST, so that path refines the fine directions that patch matching reads rather than supplying content, and we retain it in the deployed model. Training to convergence improves every column, and the deployed model leads all six.

Two further ablations follow, a support-gate shape ablation that examines the isotropy argument of Section 2.2 in Appendix T, and a recursion test showing that the model stays self-consistent when its own output is fed back for a second zoom step in Appendix U.

## P Upsamplers Below the Coarse Map

Under the matching readouts of Table 1 several upsamplers land below the plain coarse map, which invites the suspicion that our ports disadvantage them, and two observations rule that out. The effect is confined to one readout family, since in Table 1 the same learned upsamplers do improve on the coarse map under prototype classifiers and linear probes, which is the regime they are normally evaluated in. And it has the mechanism the averaging law already supplies. An interpolated vector on a boundary is a blend of the tokens on either side, and whitening rescales exactly the low-variance directions in which such blends differ from real features. A winner-take-all match then has to commit to one support patch, whereas a readout that pools many patches absorbs the blend. Interpolation therefore introduces blended vectors in proportion to how much fine structure an image holds, which

Table 36: **NYUv2 with real ground truth.** An indoor benchmark whose annotations are measured rather than estimated. Depth uses real metric depth in meters, normals are computed from ground-truth depth, and segmentation uses the forty-class labels. Ours (main) is the model used throughout the paper, trained on the aerial pool with COCO photographs mixed in, and Ours (aerial) is the variant trained on the aerial pool alone, which has therefore never seen an indoor or photographic scene. Few-shot segmentation rows report mIoU, where higher is better, and bold marks the best source excluding the teacher.

| Evaluation | Shots | bilinear | FeatUp | NAF | AnyUp | JAFAR | Ours (aerial) | Ours (main) | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| Prototype | 1 | **0.374** | 0.373 | 0.371 | 0.348 | 0.366 | 0.276 | 0.335 | 0.368 |
| | 5 | 0.505 | **0.508** | 0.503 | 0.460 | 0.502 | 0.381 | 0.461 | 0.509 |
| INSID3 | 1 | 0.394 | **0.407** | 0.375 | 0.315 | 0.369 | 0.366 | **0.407** | 0.362 |
| | 5 | 0.420 | 0.407 | 0.398 | 0.311 | 0.412 | 0.447 | **0.488** | 0.404 |
| FROST | 1 | 0.188 | 0.179 | 0.180 | 0.158 | 0.193 | **0.331** | 0.324 | 0.309 |
| | 5 | 0.244 | 0.229 | 0.213 | 0.165 | 0.176 | 0.444 | **0.467** | 0.428 |
| Linear probe mIoU ↑ | – | 0.398 | 0.384 | **0.406** | 0.375 | 0.394 | 0.352 | 0.377 | 0.363 |
| Trained decoder mIoU ↑ | – | 0.411 | **0.416** | 0.402 | 0.402 | 0.398 | 0.372 | 0.385 | 0.361 |
| Real depth $\delta_1$ ↑ | – | 0.199 | 0.199 | 0.198 | 0.198 | 0.199 | **0.259** | 0.227 | 0.218 |
| Real depth RMSE (m) ↓ | – | 1.954 | 1.940 | 1.953 | 1.938 | 1.924 | **1.589** | 1.763 | 1.898 |
| Real normal below $11.25^\circ$ (%) ↑ | – | 7.32 | 7.43 | 7.43 | 7.13 | 7.48 | 7.19 | **7.50** | 7.19 |
| Retrieval precision at 1 ↑ | – | 0.517 | 0.512 | **0.524** | 0.497 | 0.492 | 0.358 | 0.447 | 0.462 |
| CAM average gain ↑ | – | 23.4 | 24.7 | 24.1 | 20.8 | 23.5 | **34.4** | 34.1 | 21.8 |

Table 37: **Transfer beyond remote sensing.** Means over VOC, ADE20K, COCO-Stuff, and NYUv2. Ours (aerial) is the aerial-pool variant and bold marks the best source excluding the teacher.

| Evaluation | bilinear | FeatUp | NAF | AnyUp | JAFAR | Ours (aerial) | Ours (main) | Teacher |
|---|---|---|---|---|---|---|---|---|
| FROST, 1 shot mIoU ↑ | 0.229 | 0.212 | 0.217 | 0.200 | 0.207 | 0.399 | **0.402** | 0.353 |
| FROST, 5 shots mIoU ↑ | 0.278 | 0.242 | 0.250 | 0.197 | 0.203 | 0.572 | **0.575** | 0.463 |
| INSID3, 1 shot mIoU ↑ | 0.440 | 0.438 | 0.443 | 0.418 | 0.428 | 0.426 | **0.470** | 0.408 |
| INSID3, 5 shots mIoU ↑ | 0.492 | 0.479 | 0.455 | 0.417 | 0.465 | 0.508 | **0.541** | 0.404 |
| Depth $\delta_1$ (linear) ↑ | 0.212 | 0.212 | 0.213 | 0.211 | 0.213 | **0.232** | 0.226 | 0.219 |
| CAM average gain ↑ | 14.5 | 15.8 | 14.8 | 12.9 | 14.6 | **25.3** | 24.1 | 14.1 |
| Linear probe mIoU ↑ | 0.316 | 0.316 | **0.320** | 0.309 | 0.315 | 0.246 | 0.274 | 0.281 |

is why the cost is largest in remote sensing and why iSAID, whose objects are the smallest in the suite, is where no upsampler escapes it.

## Q NYUv2 with Real Ground Truth

Table 36 gives the complete indoor results, and three observations matter. First, the per-patch advantage over the teacher survives real ground truth. Both variants improve metric depth over the teacher, from $\delta_1$ of 0.218 and 1.898 meters of error to 0.227 and 1.763 for the main model and to 0.259 and 1.589 for the aerial-only variant, and both lead the teacher under FROST segmentation on the forty real classes. The same holds for the pseudo-ground-truth depth of Table 2, so the depth conclusion does not depend on the target being estimated. Second, fitted and pooled readouts pay a visible cross-domain tax when photographs are absent from training, which mixing COCO into the corpus largely removes, consistent with pooled statistics being the component most sensitive to domain shift. Third, real surface normals do not separate the sources, since every source lies between 38.4 and 40.2 degrees of mean angular error and within four tenths of a point in the fraction of accurate normals, so we report normals for completeness and exclude them from every claim.

## R Natural-Image Suite

Table 38 extends the domain-transfer summary of the main paper. The per-patch identity transfers, since even the aerial-only variant beats the teacher under FROST on photographs it has never seen, while pooled and fitted readouts show the cross-domain tax that including photographs removes. Mixing weak-domain samples into the corpus is therefore a positive training lever that costs nothing measurable in domain, which is why the model used throughout the paper is trained on the mixed corpus.

Table 38: **Natural-image suite.** Results averaged over PASCAL VOC, ADE20K, and COCO-Stuff. Ours (main) is the model used throughout the paper, whose corpus mixes COCO photographs into the aerial pool, and Ours (aerial) is the variant trained on the aerial pool alone and therefore evaluated fully out of domain, and every few-shot classifier is reported at both shot counts. Few-shot segmentation rows report mIoU, where higher is better. Bold marks the best source excluding the teacher.

| Evaluation | Shots | bilinear | FeatUp | NAF | AnyUp | JAFAR | Ours (aerial) | Ours (main) | Teacher |
|---|---|---|---|---|---|---|---|---|---|
| Prototype | 1 | 0.478 | 0.476 | **0.482** | 0.455 | 0.478 | 0.373 | 0.454 | 0.477 |
| | 5 | 0.615 | 0.616 | **0.631** | 0.607 | 0.626 | 0.527 | 0.612 | 0.636 |
| Self-support | 1 | 0.502 | 0.502 | **0.505** | 0.458 | 0.493 | 0.358 | 0.469 | 0.496 |
| | 5 | 0.660 | 0.645 | **0.670** | 0.641 | 0.667 | 0.542 | 0.629 | 0.662 |
| FSSDINO | 1 | 0.487 | 0.479 | **0.489** | 0.471 | 0.484 | 0.358 | 0.435 | 0.458 |
| | 5 | 0.627 | 0.620 | **0.637** | 0.624 | 0.634 | 0.510 | 0.611 | 0.617 |
| INSID3 | 1 | 0.455 | 0.448 | 0.465 | 0.453 | 0.448 | 0.446 | **0.491** | 0.424 |
| | 5 | 0.516 | 0.504 | 0.474 | 0.453 | 0.482 | 0.529 | **0.559** | 0.404 |
| FROST | 1 | 0.243 | 0.223 | 0.229 | 0.213 | 0.212 | 0.422 | **0.428** | 0.368 |
| | 5 | 0.289 | 0.247 | 0.263 | 0.207 | 0.212 | **0.614** | 0.611 | 0.475 |
| Linear probe mIoU ↑ | – | 0.288 | **0.293** | 0.292 | 0.287 | 0.289 | 0.211 | 0.240 | 0.254 |
| Depth $\delta_1$ (linear) ↑ | – | 0.217 | 0.217 | 0.218 | 0.215 | 0.218 | 0.223 | **0.226** | 0.220 |
| CAM average gain ↑ | – | 11.5 | 12.8 | 11.7 | 10.3 | 11.6 | **22.2** | 20.7 | 11.6 |

The residual pattern on photographs is the headroom relation of Section 5.3 rather than a domain failure, and the teacher makes this visible because it is the ceiling of what any re-encoding can deliver. Averaged over the eleven segmentation readouts of each benchmark, the teacher exceeds bilinear interpolation on 8 of 11 readouts on ADE20K but on only 4 of 11 on VOC and 3 of 11 on COCO-Stuff, whose objects commonly occupy a large part of the frame and are therefore already resolved by the wide-field encoding. Our margins follow that ceiling. Across the 33 dataset-by-readout pairs the correlation between the teacher-minus-bilinear headroom and our gain over bilinear is 0.78, matching the 0.82 measured on remote sensing. Splitting the pairs by the sign of the headroom gives $+0.103$ mIoU mean gain on the fifteen pairs where re-encoding helps against $-0.023$ on the eighteen where it does not. The distance to the teacher, in contrast, is domain independent. On the prototype-style and probe readouts, the family the averaging law places on the teacher's side, our model sits 0.022 mIoU below the teacher on the natural suite and 0.032 below it on the same readouts in remote sensing. The headroom of those readouts meanwhile falls from $+0.053$ in remote sensing to $-0.008$ on photographs, and our standing against bilinear falls with it by 0.051. The apparent weakness on photographs is therefore a property of what the benchmarks leave to recover rather than of how well the modulator predicts a re-encoding, and it locates the interface where it belongs, on imagery whose structures are small relative to the field of view.

The three tables of this section report each of them separately, with the same sources as well as a trained decoder and patch retrieval that the averaged table omits, so that the behavior on photographs can be read per dataset, and the two student variants are named as in Table 38.

## S SPair-71k Correspondence

Table 42 shows single-vector nearest-neighbor correspondence, a readout with no averaging on either side of the match. As the averaging law predicts, plain bilinear features are the strongest source at the two looser thresholds, and only the teacher improves on them at the strictest one. Every learned upsampler falls below bilinear as well, so this is a property of the task rather than a weakness specific to our student, which itself lands within 0.010 of bilinear at the strictest threshold once photographs are part of its corpus.

## T Support-Gate Shape Ablation

The disc support of Section 2.2 is the only rotation-invariant choice, but an axis-aligned box has one practical advantage, since it covers the decoded crop exactly and therefore tiles a whole image without leaving gaps. We accordingly train the box variant under the same recipe and measure what

Table 39: **PASCAL VOC.** Per-readout results with the sources of Table 38, where every segmentation row reports mIoU.

| Evaluation | bilinear | FeatUp | NAF | AnyUp | JAFAR | Ours (aerial) | Ours (main) | Teacher |
|---|---|---|---|---|---|---|---|---|
| Prototype, 1 shot | 0.687 | 0.680 | 0.697 | 0.675 | **0.705** | 0.483 | 0.635 | 0.667 |
| Prototype, 5 shots | 0.748 | 0.727 | **0.774** | 0.756 | **0.774** | 0.581 | 0.730 | 0.732 |
| Self-support, 1 shot | 0.668 | 0.658 | **0.681** | 0.638 | 0.675 | 0.416 | 0.627 | 0.635 |
| Self-support, 5 shots | 0.805 | 0.781 | **0.815** | 0.795 | 0.806 | 0.648 | 0.790 | 0.821 |
| FSSDINO, 1 shot | 0.685 | 0.672 | 0.701 | 0.689 | **0.710** | 0.468 | 0.612 | 0.636 |
| FSSDINO, 5 shots | 0.744 | 0.726 | **0.766** | 0.754 | 0.765 | 0.598 | 0.746 | 0.752 |
| INSID3, 1 shot | 0.565 | 0.543 | 0.604 | **0.666** | 0.592 | 0.496 | 0.542 | 0.437 |
| INSID3, 5 shots | **0.601** | 0.545 | 0.430 | 0.553 | 0.542 | 0.586 | 0.558 | 0.416 |
| FROST, 1 shot | 0.231 | 0.202 | 0.221 | 0.213 | 0.203 | 0.550 | **0.560** | 0.406 |
| FROST, 5 shots | 0.238 | 0.146 | 0.230 | 0.118 | 0.144 | **0.701** | 0.653 | 0.393 |
| Linear probe | 0.531 | 0.549 | 0.538 | **0.558** | 0.542 | 0.377 | 0.428 | 0.447 |
| Trained decoder | 0.495 | 0.491 | **0.532** | 0.494 | 0.529 | 0.435 | 0.483 | 0.467 |
| Depth $\delta_1$ (linear) | 0.231 | 0.231 | 0.233 | 0.231 | 0.233 | 0.223 | **0.235** | 0.231 |
| Retrieval precision at 1 | 0.744 | **0.750** | 0.732 | 0.719 | 0.713 | 0.443 | 0.586 | 0.663 |
| CAM gain | 8.9 | 9.8 | 8.7 | 7.6 | 8.4 | **16.7** | 12.6 | 9.6 |

Table 40: **ADE20K.** Per-readout results with the sources of Table 38, where every segmentation row reports mIoU.

| Evaluation | bilinear | FeatUp | NAF | AnyUp | JAFAR | Ours (aerial) | Ours (main) | Teacher |
|---|---|---|---|---|---|---|---|---|
| Prototype, 1 shot | 0.341 | 0.338 | 0.338 | 0.297 | 0.331 | 0.300 | **0.342** | 0.369 |
| Prototype, 5 shots | 0.530 | 0.542 | 0.539 | 0.500 | 0.530 | 0.519 | **0.555** | 0.593 |
| Self-support, 1 shot | **0.406** | 0.400 | 0.398 | 0.319 | 0.380 | 0.323 | 0.385 | 0.427 |
| Self-support, 5 shots | 0.563 | 0.548 | **0.577** | 0.534 | 0.574 | 0.493 | 0.538 | 0.585 |
| FSSDINO, 1 shot | **0.362** | 0.348 | 0.345 | 0.320 | 0.334 | 0.286 | 0.320 | 0.350 |
| FSSDINO, 5 shots | 0.568 | 0.556 | **0.572** | 0.538 | 0.555 | 0.464 | 0.528 | 0.536 |
| INSID3, 1 shot | 0.378 | 0.387 | 0.363 | 0.306 | 0.338 | 0.440 | **0.476** | 0.452 |
| INSID3, 5 shots | 0.435 | 0.458 | 0.457 | 0.383 | 0.461 | 0.519 | **0.564** | 0.400 |
| FROST, 1 shot | 0.283 | 0.249 | 0.249 | 0.232 | 0.232 | 0.352 | **0.368** | 0.395 |
| FROST, 5 shots | 0.350 | 0.345 | 0.337 | 0.308 | 0.312 | 0.587 | **0.600** | 0.570 |
| Linear probe | 0.175 | 0.169 | **0.177** | 0.154 | 0.171 | 0.136 | 0.157 | 0.185 |
| Trained decoder | **0.178** | **0.178** | 0.172 | 0.158 | 0.166 | 0.148 | 0.162 | 0.177 |
| Depth $\delta_1$ (linear) | 0.211 | 0.211 | 0.211 | 0.209 | 0.212 | **0.224** | **0.224** | 0.215 |
| Retrieval precision at 1 | 0.347 | **0.351** | 0.346 | 0.278 | 0.306 | 0.246 | 0.306 | 0.338 |
| CAM gain | 16.8 | 17.9 | 17.1 | 15.5 | 17.4 | 31.1 | **31.6** | 15.5 |

Table 41: **COCO-Stuff.** Per-readout results with the sources of Table 38, where every segmentation row reports mIoU.

| Evaluation | bilinear | FeatUp | NAF | AnyUp | JAFAR | Ours (aerial) | Ours (main) | Teacher |
|---|---|---|---|---|---|---|---|---|
| Prototype, 1 shot | 0.405 | 0.409 | **0.411** | 0.394 | 0.400 | 0.335 | 0.385 | 0.395 |
| Prototype, 5 shots | 0.568 | 0.579 | **0.580** | 0.565 | 0.575 | 0.482 | 0.549 | 0.584 |
| Self-support, 1 shot | 0.430 | **0.450** | 0.435 | 0.418 | 0.425 | 0.335 | 0.394 | 0.425 |
| Self-support, 5 shots | 0.611 | 0.607 | 0.620 | 0.593 | **0.621** | 0.486 | 0.560 | 0.579 |
| FSSDINO, 1 shot | 0.415 | 0.418 | **0.419** | 0.405 | 0.408 | 0.319 | 0.372 | 0.388 |
| FSSDINO, 5 shots | 0.570 | 0.579 | 0.572 | 0.580 | **0.582** | 0.468 | 0.560 | 0.564 |
| INSID3, 1 shot | 0.422 | 0.414 | 0.429 | 0.386 | 0.413 | 0.401 | **0.454** | 0.383 |
| INSID3, 5 shots | 0.512 | 0.508 | 0.537 | 0.422 | 0.444 | 0.481 | **0.556** | 0.397 |
| FROST, 1 shot | 0.216 | 0.218 | 0.217 | 0.195 | 0.201 | **0.364** | 0.357 | 0.304 |
| FROST, 5 shots | 0.280 | 0.249 | 0.221 | 0.196 | 0.179 | 0.553 | **0.582** | 0.462 |
| Linear probe | 0.159 | **0.162** | 0.160 | 0.149 | 0.153 | 0.120 | 0.135 | 0.131 |
| Trained decoder | **0.164** | 0.150 | 0.157 | 0.143 | 0.157 | 0.146 | 0.143 | 0.130 |
| Depth $\delta_1$ (linear) | 0.209 | 0.209 | 0.210 | 0.206 | 0.209 | **0.222** | 0.220 | 0.213 |
| Retrieval precision at 1 | 0.311 | **0.330** | 0.309 | 0.288 | 0.282 | 0.201 | 0.259 | 0.264 |
| CAM gain | 8.9 | 10.6 | 9.4 | 7.8 | 8.9 | **18.9** | 17.8 | 9.7 |

Table 42: **SPair-71k correspondence.** Measured on 360 test pairs as the percentage of correct keypoints (PCK) at thresholds of 0.05, 0.10, and 0.15 of the bounding-box size.

| Source | PCK 0.05 | PCK 0.10 | PCK 0.15 |
|---|---|---|---|
| lowres | 0.402 | 0.582 | 0.676 |
| bilinear | 0.428 | **0.597** | **0.679** |
| JBU | 0.268 | 0.491 | 0.608 |
| FeatUp | 0.399 | 0.576 | 0.670 |
| NAF | 0.386 | 0.558 | 0.653 |
| AnyUp | 0.270 | 0.460 | 0.576 |
| JAFAR | 0.313 | 0.499 | 0.604 |
| Ours (aerial) | 0.306 | 0.448 | 0.540 |
| Ours (main) | 0.418 | 0.569 | 0.656 |
| Teacher | **0.464** | 0.583 | 0.652 |

Table 43: **Support-gate shape ablation.** Disc against box support gates under the same recipe, reported as FROST one-shot mIoU on four datasets. The fine-tuned row initializes the box-gate model from the disc-gate weights before retraining with the box objective, and bold marks the best of the three gates.

| Gate | LoveDA | iSAID | SpaceNet2 | VOC | Mean |
|---|---|---|---|---|---|
| bilinear | 0.229 | 0.217 | 0.049 | 0.231 | 0.181 |
| Disc (deployed) | **0.408** | 0.361 | **0.407** | **0.560** | **0.434** |
| Box | 0.311 | 0.360 | 0.196 | 0.349 | 0.304 |
| Box, fine-tuned | 0.325 | **0.362** | 0.238 | 0.357 | 0.321 |
| Teacher | 0.356 | 0.454 | 0.318 | 0.406 | 0.384 |

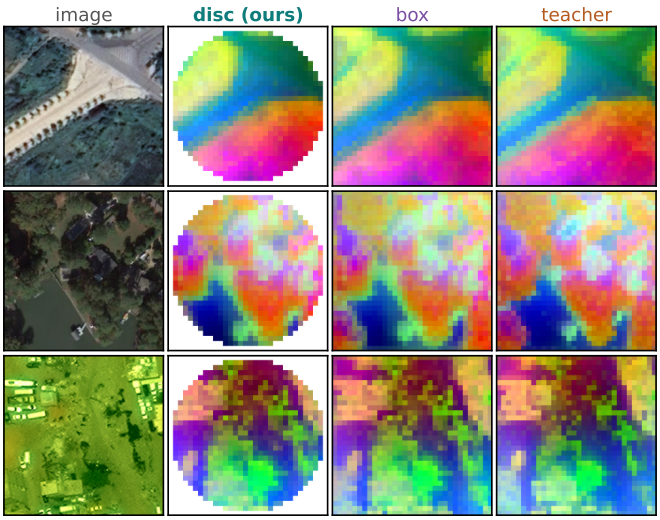


Figure 6: **Support-gate shape ablation.** Feature panels share a joint PCA per row with the teacher, and the disc output is masked white outside its support. The box model is a valid re-encoder that covers the full crop, but its per-patch discriminative sharpness at the center is slightly lower, which matching readouts amplify, as Table 43 quantifies.

isotropy is worth. The box takes the multiplier $\kappa = 1+\beta$ that makes it coincide with $B(q)$, so replacing the disc by it supervises 39 percent more area with the same 9.2M parameters. Table 43 shows the consequence. Under a fitted linear probe the two gates are equivalent, at 0.523 mIoU for the disc against 0.522 for the box averaged over the same four datasets, but under winner-take-all matching readouts the box gate loses 0.130 mIoU on average and drops below the teacher. Diagnostics locate the cause at the center of the support rather than at its rim. Averaged over the four datasets, the box model loses 0.0017 cosine to the teacher at the center while gaining 0.0004 at the rim. After whitening the features with fixed statistics the central loss doubles to 0.0033 while the rim still gains, so the deficit sits in low-variance discriminative directions rather than in the extra area the box has to cover. The whitening-based candidate voting of FROST and INSID3 reads exactly that axis and amplifies it, while prototype-averaging readouts cancel it, which is the averaging law operating inside the few-shot family itself. Fine-tuning the box model from the disc weights recovers only an eighth of the deficit and still trails the teacher, so the loss is intrinsic to the box objective, which dilutes the central gradient share across the harder corners, rather than an optimization artifact. Figure 6 compares the two students visually.

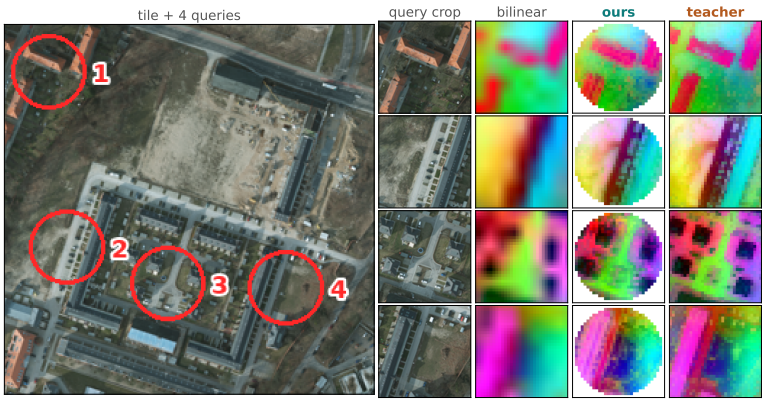


Figure 7: **Interactive demonstration.** Point queries placed on one large aerial tile, each of them served by the single backbone pass that encoded the tile and decoded at the cost reported in Table 4. PCA colors are joint per row, and feature change is confined to the query support.

## U Recursive Zoom

Figure 8 probes self-consistency under recursion. Decoding a second zoom step from the student's own first-step output produces features nearly indistinguishable from the direct one-step zoom to the same final region, and both recover the teacher's structure while removing its speckle, so the student's outputs are valid inputs for further zooming and hierarchical or iterative use does not degrade them.

## V Additional Qualitative Comparisons

Figure 7 shows the interface in the regime it is built for, in which one encoded tile is browsed with successive point queries. Figure 9 then extends the qualitative comparison of the main paper to nine further satellite and aerial datasets, and Figure 10 covers drone-view, natural, and indoor imagery.

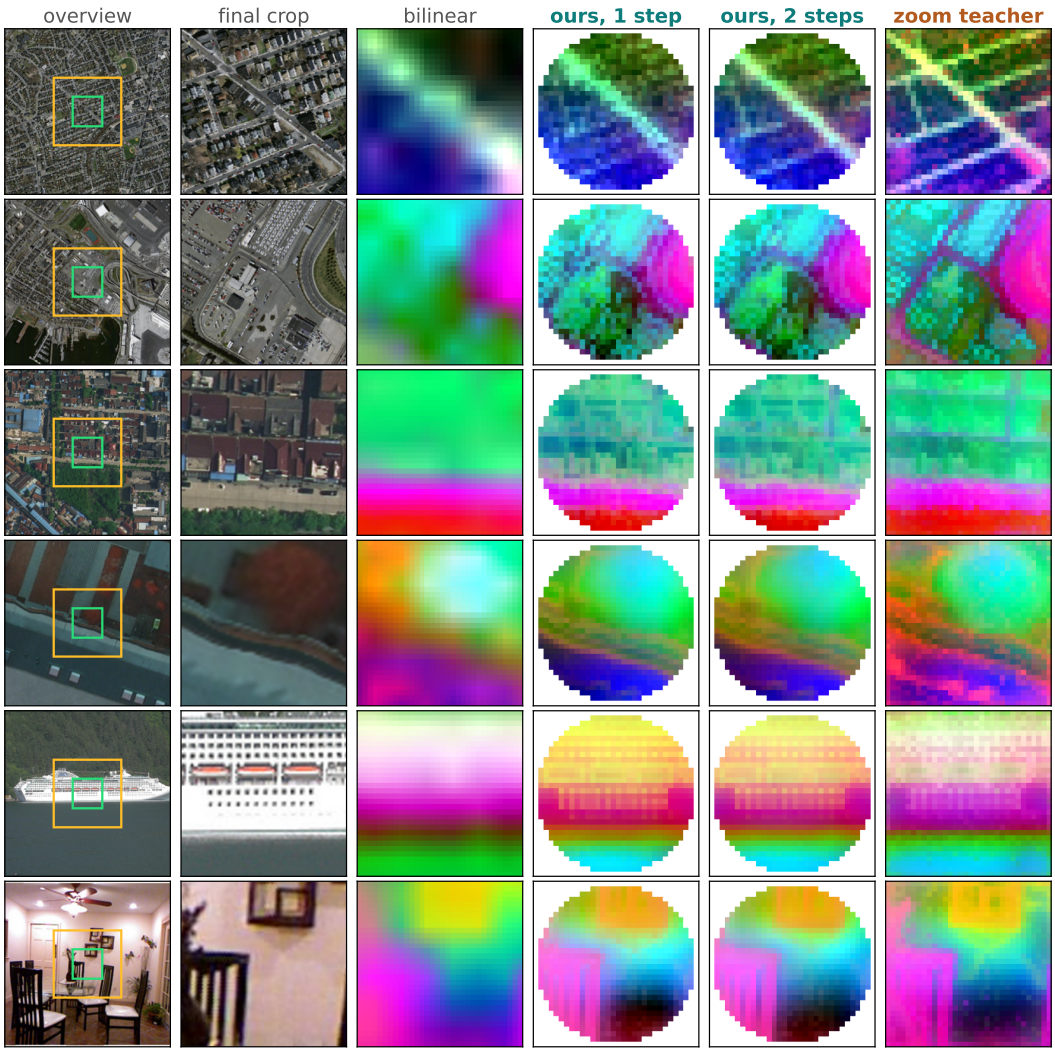


Figure 8: **Recursive zoom.** The student's first zoom output is fed back to the same model as a new global feature map, and a second zoom into a central subregion is decoded from it. Columns show the overview with the first-step region in yellow and the second-step region in green, the crop of the final region, bilinear features, the direct one-step zoom to the final region, the recursive two-step zoom, and the true teacher re-encoding. Joint PCA colors are shared within each row and the region outside the disc of the query support of our outputs is masked white. Rows span aerial and satellite scenes, a natural image, and an indoor scene.

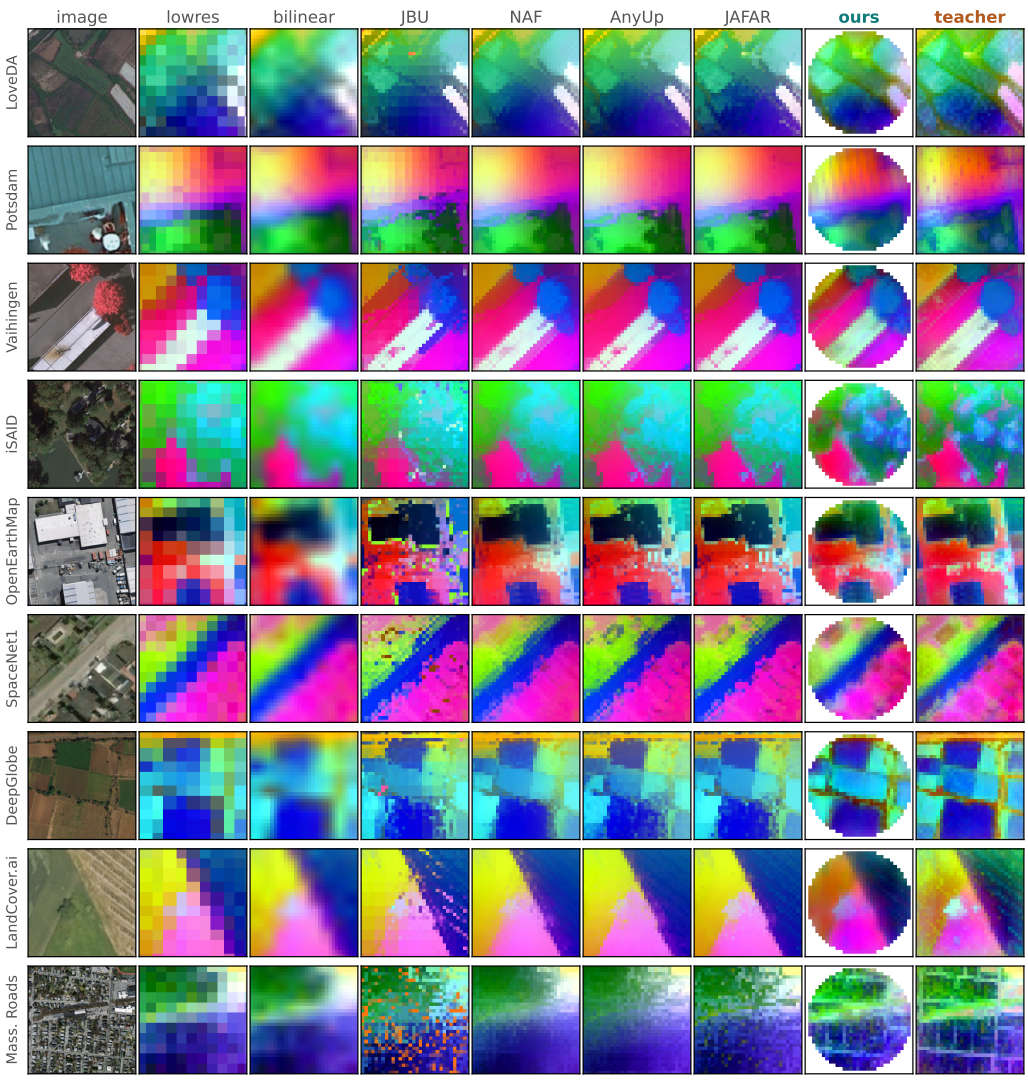


Figure 9: **Qualitative gallery, overhead imagery.** Joint PCA of patch features for one query crop per dataset across nine satellite and aerial benchmarks, computed jointly over the eight sources shown so that colors are comparable within a row. The teacher column is the zoom re-encoding. Only the two re-encoding columns recover structure absent from the coarse map, such as the road crossing on Potsdam, the agricultural field grid on DeepGlobe, and the street grid on Massachusetts Roads, while our prediction additionally removes the teacher's high-frequency speckle. Feature change is confined to the query support, and the region outside that disc, which our method leaves unchanged by construction, is masked white for clarity.

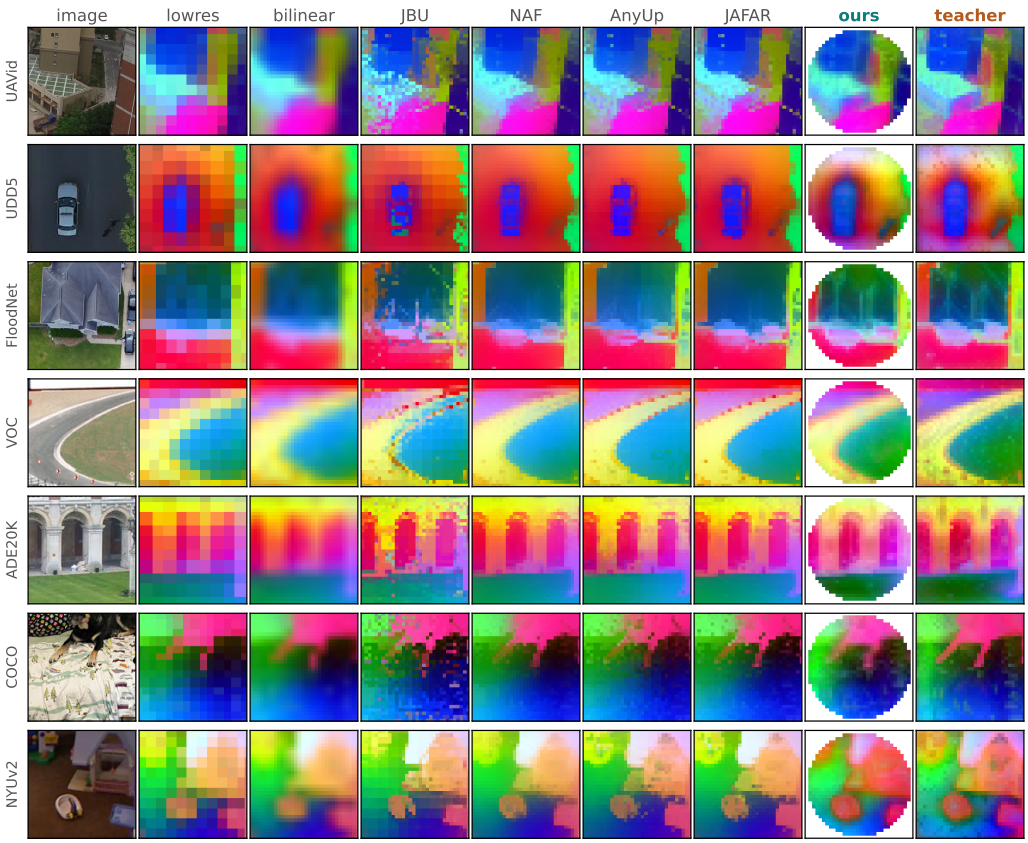


Figure 10: **Qualitative gallery beyond overhead imagery.** Joint PCA of patch features for one query crop per dataset on drone-view, natural, and indoor benchmarks, computed jointly over the eight sources shown so that colors are comparable within a row. The pattern of the satellite gallery repeats, as on the vehicle on UDD5, the arcades on ADE20K, and the toys on NYUv2, so the foveal prediction is not specific to overhead imagery.