COMPATLIFT: ANCHORED FEATURE LIFTING FOR REMOTE SENSING OPEN-VOCABULARY SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Remote-Sensing Open-Vocabulary Semantic Segmentation (RS-OVSS) requires fine spatial detail to delineate small objects and narrow structures, while extracting high-resolution features with the backbone incurs substantial computation. Feature upsampling provides a more efficient alternative, but existing learned feature upsamplers are mainly developed for pretrained visual encoders and may produce representations that are incompatible with a frozen encoder–decoder segmentation pipeline. We study this problem in a SAM 3-based RS-OVSS pipeline and observe that directly replacing native features with learned upsamplers substantially degrades segmentation performance, whereas bicubic interpolation largely preserves the behavior of the frozen decoder. Motivated by this, we propose CompatLift, a lightweight feature upsampler framework that learns image-guided residual updates anchored by the bicubic reference. The upsampler module is trained *without* segmentation annotations using high-resolution features from the same frozen SAM 3 encoder, while the backbone and prediction heads remain frozen. We further introduce a simple score calibration step to compensate for the confidence shift caused by feature lifting. Extensive experiments on representative eight remote-sensing benchmarks show that CompatLift improves SAM 3 segmentation performance by lifted higher-resolution, decoder-compatible feature representations.

1 INTRODUCTION

Open-vocabulary semantic segmentation allows the target categories in remote-sensing imagery to be specified by text at inference time (Li et al., 2025a). Remote-sensing scenes often contain small objects and narrow structures, making fine spatial detail important for accurate boundary delineation (Cao et al., 2024; Liang et al., 2024; Tang et al., 2025). However, visual features from pretrained encoders are usually computed on grids much coarser than the input image. Increasing the encoder input resolution can provide finer features, but it also increases the computational cost of the encoder. Feature upsampling provides a more efficient alternative by increasing the spatial resolution of intermediate features without modifying the backbone at inference time.

Recent methods learn image-guided mappings to upsample features from pretrained visual encoders, including FeatUp, AnyUp, and JAFAR (Fu et al., 2024; Wimmer et al., 2026; Couairon et al., 2026). LiFT uses higher-resolution features from the same frozen encoder as training targets (Suri et al., 2024), and ViT-Up studies feature upsampling for vision transformers (Wandel et al., 2026). These methods show that feature upsampling can improve spatial representations. FeatUp and ViT-up studies also report evaluations with frozen probes or prediction heads, but these results do not directly establish whether lifted features remain usable by a pretrained open-vocabulary predictor in a frozen multi-branch pipeline. This question is important for the SAM 3-based SegEarth-OV3 pipeline (Li et al., 2025b), because changing intermediate visual features can shift the representation presented to the prediction components and can increase the cost of downstream processing.

We study this compatibility problem in the frozen SegEarth-OV3 pipeline. Figure 1 illustrates that visual and textual features are processed by a downstream decoder in SAM 3. Separately, Figure 6 shows that reducing the encoder input resolution decreases segmentation accuracy, whereas enlarging the feature grid at a fixed encoder input increases downstream processing time. Under a fixed $2\times$

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.

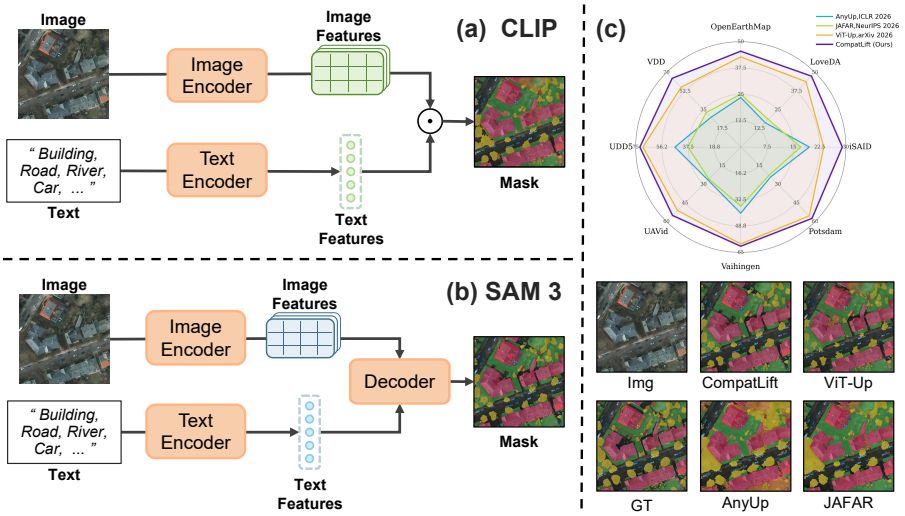


Figure 1: **Open-vocabulary segmentation with CLIP and SAM 3.** (a) CLIP-based methods compare image and text embeddings through dot-product similarity to obtain semantic predictions. (b) SAM 3 combines visual and textual features in a downstream decoder to predict masks, making the frozen decoder sensitive to changes in the image features. (c) Top: per-dataset mIoU (%) of AnyUp, JAFAR, ViT-Up, and CompatLift on eight remote-sensing benchmarks. Bottom: qualitative comparison showing the input image (Img), ground truth (GT), and predictions from CompatLift, ViT-Up, AnyUp, and JAFAR.

upsampling factor, directly replacing native features with outputs from previous learned upsamplers can degrade mean intersection over union (mIoU), while bicubic resizing remains close to the native model (see Table 2). These observations indicate that increasing feature-grid resolution alone is insufficient: *the lifted representation must also remain compatible with the frozen prediction pipeline, rather than only upsampling the feature resolution*.

Motivated by this observation, we propose **CompatLift**, a compatibility feature upsampling module for frozen SAM 3-based segmentation. CompatLift predicts an image-guided high-resolution (HR) feature and combines its residual with a bicubic reference using a residual mixing coefficient. The bicubic feature acts as an anchor, preserving the original representation structure, while the learned residual provides additional spatial detail. During training, CompatLift is supervised by HR features extracted from the same frozen encoder. The module is trained on unlabeled Million-AID imagery *without* updating the encoder, text encoder, and prediction components. The HR feature supervision is used only during training and is removed at inference.

Across eight remote-sensing benchmarks, CompatLift achieves an average mIoU of 54.2, compared with 53.4 for native SegEarth-OV3 and 52.8 for bicubic feature resizing. It improves over the native system on seven of the eight datasets, with the largest gains on OpenEarthMap and iSAID. The only decrease is 0.4 mIoU points on Potsdam. Our contributions are as follows:

- We identify feature compatibility as a central challenge when applying learned feature upsampling to open-vocabulary segmentation with frozen prediction components, and show that direct feature replacement can substantially reduce performance.
- We propose CompatLift, which learns image-guided residual updates relative to a bicubic anchor and uses same-encoder HR supervision.
- CompatLift is evaluated on eight representative remote-sensing benchmarks, improving the native SegEarth-OV3 pipeline and producing the best segmentation mIoU performance.

2 Related Work

Open-vocabulary segmentation in remote sensing. Open-vocabulary segmentation methods adapt vision–language representations to dense prediction, allowing categories to be specified through

text (Radford et al., 2021; Li et al., 2022; Ghiasi et al., 2022). For remote-sensing imagery, SegEarth-OV uses SimFeatUp to refine coarse CLIP features and applies the resulting representations across datasets without task-specific training (Li et al., 2025a). SegEarth-OV3 adapts SAM 3 to the same task by combining semantic and instance predictions and using category-presence scores to suppress irrelevant classes (Carion et al., 2026; Li et al., 2025b). Beyond semantic segmentation, foundation models have also been explored for remote-sensing change-detection (Li et al., 2024).

Upsampling pretrained visual features. Joint bilateral upsampling(JBU) uses a high-resolution guidance image to interpolate a coarse signal (Kopf et al., 2007). Learned operators such as CARAFE and SAPA make feature aggregation depend on local content (Wang et al., 2019; Lu et al., 2022). For pretrained visual features, JAFAR uses image-guided cross-attention to interpolate features at arbitrary output resolutions (Couairon et al., 2026). AnyUp supports transfer to unseen feature extractors without retraining the upsampler (Wimmer et al., 2026).

Feature upsamplers also differ in how they obtain training targets. FeatUp learns an upsampler and a downsampler by requiring the upsampled representation to reconstruct encoder features from transformed image views (Fu et al., 2024). LiFT obtains targets from the same frozen encoder applied to higher-resolution versions of the training images (Suri et al., 2024). Our image-guided branch follows FeatUp. We mix its output with a bicubic reference and supervise the mixed representation using scale-consistent feature alignment loss.

Feature upsampling under frozen predictors. Training a prediction head on upsampled features allows it to adapt to the new representation, whereas reusing a head trained on native features tests transfer without that adaptation. FeatUp reports frozen-probe evaluations for segmentation and depth prediction (Fu et al., 2024), and ViT-Up examines feature preservation with prediction heads trained on native low-resolution features (Wandel et al., 2026). We study feature updates within the complete SegEarth-OV3 pipeline, where the modified features undergo further Transformer processing before segmentation. We evaluate the learned update at $2\times$ with all pretrained components frozen, and separately measure how feature-grid expansion affects downstream processing time.

3 METHOD

We study feature upsampling for open-vocabulary segmentation when the downstream prediction components are frozen. The goal is to increase the spatial resolution of the visual features while keeping them compatible with the representation expected by the original prediction pipeline. CompatLift achieves this by combining a bicubic feature reference with a learned image-guided residual update and training the upsampler supervised by HR features extracted from the same frozen encoder.

3.1 FEATURE LIFTING IN FROZEN SEGMENTATION

Let $\mathcal{E}$ denote the frozen SAM 3 image encoder. Given an image I, the encoder produces a feature pyramid $\mathcal{F}(I) = \{F^{(0)}, F^{(1)}, F^{(2)}\}$ and the corresponding positional encodings $\mathcal{P}(I)$. Let $\mathcal{D}$ denote the frozen SegEarth-OV3 prediction pipeline, including the semantic branch, instance decoder, category-presence prediction, and prediction fusion. Given a category prompt q, the native prediction is written as

$$Y_{\text{nat}} = \mathcal{D}\big(\mathcal{F}(I), \mathcal{P}(I), q\big). \tag{1}$$

We apply feature lifting only to the deepest feature level $\ell^\star = 2$, whose channel dimension is $C = 256$. For a low-resolution (LR) input I_L, the corresponding feature is

$$F_L = \text{sg}\left[\Phi_{\ell^\star}(I_L)\right] \in \mathbb{R}^{C\times h\times w}, \tag{2}$$

where $\Phi_{\ell^\star}$ denotes the level-$\ell^\star$ feature produced by the frozen encoder and $\text{sg}(\cdot)$ stops gradients. The encoder, text encoder, prediction heads, and fusion modules remain frozen throughout training and evaluation. Compatibility is therefore evaluated by directly feeding the modified features into the unchanged prediction pipeline.

3.2 BICUBIC-ANCHORED RESIDUAL FEATURE LIFTING

SAM 3 features provide strong semantic representations but remain spatially coarse. Meanwhile, the frozen downstream decoder imposes constraints on feature scale and representation distribution

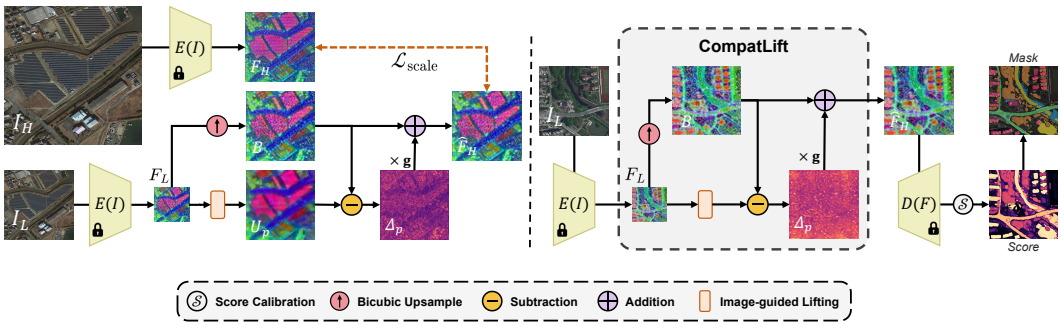


Figure 2: **Training and deployment of CompatLift.** During training (left), paired low- and high-resolution views are encoded by the frozen SAM 3 encoder. CompatLift combines the bicubic anchor B with an image-guided proposal using the residual mixing coefficient g and optimizes the high-resolution feature supervision. During deployment (right), only the low-resolution encoder, image-guided lifting module, frozen prediction components, and score calibration are used.

shifts, motivating us to enhance spatial details through lightweight $2\times$ residual feature lifting while maintaining decoder compatibility. CompatLift uses bicubic interpolation as a stable reference and learns a residual correction for spatial detail enhancement. Given the low-resolution feature F_L, we first obtain the bicubic reference $B = \text{Bic}_{2\times}(F_L)$, and generate a high-resolution proposal through an image-guided lifting branch:

$$F_H^p = U_\theta(F_L, I_L). \tag{3}$$

Instead of directly using F_H^p as the decoder input, we formulate the lifted feature as

$$\widehat{F}_H = B + g\Delta_\theta, \qquad \Delta_\theta = F_H^p - B, \tag{4}$$

where g controls the contribution of the learned correction. This residual formulation allows the lifting branch to introduce additional spatial details while maintaining the representation structure provided by the bicubic reference. Figure 3 provides a schematic view of the feature-space transformation performed by CompatLift.

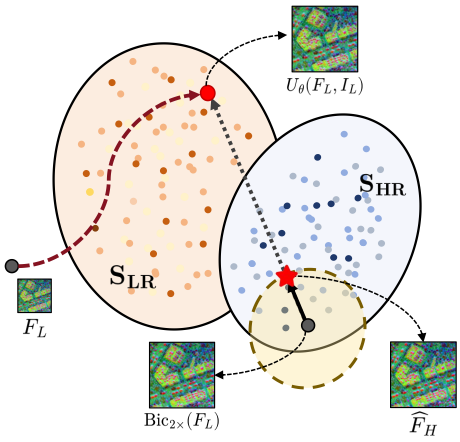


Figure 3: **Illustration of the feature space of CompatLift.** The orange and blue regions represent the low- and high-resolution feature distributions, respectively. Starting from the native feature F_L, the feature upsampler produces high-resolution $U_\theta(F_L, I_L)$. CompatLift combines the bicubic reference $\text{Bic}_{2\times}(F_L)$ and the U_θ through a residual mixing to produce $\widehat{F}_H$.

The lifting branch U_θ is implemented with a single-stage image-guided JBU operator inspired by FeatUp (Fu et al., 2024). Different from previous feature upsampling settings that often require multiple lifting stages for large scale factors, CompatLift only performs a $2\times$ expansion of the deepest SAM 3 feature. Therefore, a lightweight single-stage aggregation is sufficient to inject spatial details while avoiding unnecessary computation.

To better adapt the lifting operation to remote-sensing images, we extend the RGB guidance used in JBU with additional spatial cues:

$$G(I_L) = [I_L, \text{HP}(I_L), \nabla_x I_L, \nabla_y I_L], \tag{5}$$

where $\text{HP}(\cdot)$ denotes high-frequency components obtained by subtracting a mean-filtered image with reflection padding from I_L and ∇_x, ∇_y represent image gradients computed as one-pixel forward differences along the horizontal and vertical directions, respectively. The guidance representation $G(I_L)$ is encoded to predict spatially adaptive kernels $K_\theta(x, y; I_L)$, which determine how neighboring low-resolution features are aggregated to produce the high-resolution proposal

$$F_H^p(x) = \sum_{y \in \mathcal{N}(x)} K_\theta(x, y; G(I_L)) F_L(y), \tag{6}$$

where K_θ determines the contribution of neighboring features during local aggregation. The additional spatial cues enable the predicted kernels to better adapt to object boundaries and high-frequency regions in remote-sensing images. The resulting proposal is finally integrated with the bicubic reference through the residual formulation in Eq. 4. The exact guidance construction and boundary handling are provided in the supplementary material.

3.3 Training with frozen encoder supervision

The lifted feature is trained without segmentation annotations by using the frozen SAM 3 encoder as a scale-consistent feature teacher. Given the same scene at two resolutions, we obtain the high-resolution target feature as

$$F_H^T = \mathrm{sg}\Big(\Phi^{(\ell^*)}(I_H)\Big), \quad (7)$$

where I_H is the high-resolution view and $\mathrm{sg}(\cdot)$ denotes gradient stopping. Since both F_H^T and the lifted feature $\widehat{F}_H$ are produced by the same frozen encoder space, the supervision encourages feature enhancement while preserving the representation space expected by the downstream decoder.

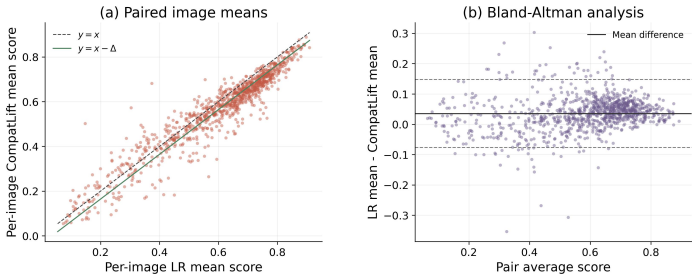


Figure 4: **Confidence shift after feature lifting.** Statistics are computed from paired native LR and CompatLift predictions on 1,000 unlabeled Million-AID images. (a) Per-image mean scores show a near-unit-slope relation with a negative offset. (b) The Bland–Altman plot shows the difference between native and CompatLift scores across their pairwise mean, motivating a constant correction δ before background rejection.

To align the lifted feature with the teacher, we combine three complementary objectives. First, cosine similarity preserves the spatial feature direction,

$$\mathcal{L}_{\cos} = 1 - \frac{1}{|\Omega|}\sum_{p\in\Omega}\Big\langle n_\epsilon(\widehat{F}_{H,p}), n_\epsilon(F_{H,p}^T)\Big\rangle, \quad (8)$$

where $n_\epsilon(x) = x/\max(\|x\|_2, \epsilon)$. Second, we match channel-wise statistics to maintain the feature distribution,

$$\mathcal{L}_{\mathrm{dist}} = \frac{1}{C}\sum_{c=1}^{C}\Big[(\mu_c(\widehat{F}_H) - \mu_c(F_H^T))^2 + (\sigma_c(\widehat{F}_H) - \sigma_c(F_H^T))^2\Big]. \quad (9)$$

Finally, point-wise reconstruction is measured by the mean squared error (MSE) $\mathcal{L}_{\mathrm{MSE}} = \mathrm{MSE}(\widehat{F}_H, F_H^T)$. The overall scale-consistent feature alignment loss is

$$\mathcal{L}_{\mathrm{scale}} = 0.6\mathcal{L}_{\cos} + 0.3\mathcal{L}_{\mathrm{dist}} + 0.1\mathcal{L}_{\mathrm{MSE}}. \quad (10)$$

At inference stage, the high-resolution teacher feature are discarded. We set $\widehat{f}_2 = \widehat{F}_H$ and bicubically double the spatial size of the remaining feature pyramid network (FPN) features and all positional encodings. The shared visual feature is updated to the lifted deepest level before applying the original prediction heads and fusion. Consequently, the learned intervention is confined to the deepest FPN feature, while geometric resizing also affects the other levels.

3.4 Score calibration for decoder compatibility

Feature lifting can change the confidence scores produced by the prediction pipeline and consequently affect background rejection. We define the rejection confidence as the pixel-wise maximum of the final fused category scores. Figure 4 compares the per-image mean rejection confidences of the native and CompatLift pipelines and shows an approximately constant downward shift for CompatLift. This shift can cause the original rejection threshold to suppress valid foreground predictions.

Table 1: **Open-vocabulary semantic segmentation on eight remote-sensing benchmarks.** Values are mIoU (%), and Avg. is the unweighted mean over the eight datasets. The first block contains compared methods trained with remote-sensing segmentation data, and the second block contains target-dataset training-free methods. Published rows provide context. SegEarth-OV3 and CompatLift in the final rows are evaluated with the frozen SAM 3 pipeline. CompatLift is trained only on unlabeled imagery. "–" denotes an unavailable result. The best results are highlighted in bold and the second best results are underlined.

Method	OpenEarthMap	LoveDA	iSAID	Potsdam	Vaihingen	UAVid	UDD5	VDD	Avg.
Training on remote-sensing segmentation data									
SCAN$_{\text{CVPR 2024}}$	–	23.2	44.3	27.5	15.2	20.3	34.1	29.2	–
SAN$_{\text{CVPR 2023}}$	–	25.3	49.6	37.3	39.2	23.5	37.2	35.8	–
SED$_{\text{CVPR 2024}}$	–	24.6	51.2	29.4	39.0	21.3	35.7	32.5	–
Cat-Seg$_{\text{CVPR 2024}}$	–	28.6	53.3	35.8	42.3	25.7	40.2	39.1	–
OVRS$_{\text{TGRS 2025}}$	–	31.5	52.7	36.4	43.5	24.1	40.8	37.2	–
GSNet$_{\text{AAAI 2025}}$	–	32.5	<u>53.7</u>	37.9	44.1	24.2	40.9	37.3	–
RSKT-Seg$_{\text{AAAI 2026}}$	–	33.2	**54.3**	38.4	42.7	25.7	42.1	39.7	–
SkySense-O$_{\text{CVPR 2025}}$	40.8	38.3	43.9	54.1	51.6	–	–	–	–
Target-dataset training-free									
CLIP$_{\text{ICML 2021}}$	12.0	12.4	7.5	15.6	10.8	10.9	9.5	14.2	11.6
MaskCLIP$_{\text{ECCV 2022}}$	25.1	27.8	14.5	33.9	29.9	28.6	32.4	32.9	28.1
SCLIP$_{\text{ECCV 2024}}$	29.3	30.4	16.1	39.6	35.9	31.4	38.7	37.9	32.4
GEM$_{\text{CVPR 2024}}$	33.9	31.6	17.7	39.1	36.4	33.4	41.2	39.5	34.1
ClearCLIP$_{\text{ECCV 2024}}$	31.0	32.4	18.2	42.0	36.2	36.2	41.8	39.3	34.6
SegEarth-OV$_{\text{CVPR 2025}}$	40.3	36.9	21.7	48.5	40.0	42.5	50.6	45.3	40.7
ProxyCLIP$_{\text{ECCV 2024}}$	38.9	34.3	21.8	49.0	47.5	35.8	40.8	47.8	39.5
CorrCLIP$_{\text{ICCV 2025}}$	32.9	36.9	25.5	51.9	47.0	38.3	46.1	47.3	40.7
SegEarth-OV3$_{\text{arXiv 2025}}$	<u>42.9</u>	<u>47.4</u>	27.6	**57.8**	<u>60.8</u>	<u>54.7</u>	<u>71.7</u>	<u>64.5</u>	<u>53.4</u>
CompatLift (ours)	**45.3**	**47.9**	29.6	<u>57.4</u>	**61.1**	**55.3**	**72.0**	**65.0**	**54.2**

We therefore apply a simple additive correction,

$$\widetilde{m}_{\mathrm{CL}}(p) = m_{\mathrm{CL}}(p) + \delta = \max_{c \in \mathcal{C}} s_c^{\mathrm{CL}}(p) + \delta, \tag{11}$$

where $s_c^{\mathrm{CL}}(p)$ is the final fused score for category c at pixel p, and $\mathcal{C}$ is the set of candidate foreground categories. To estimate δ, we process 1,000 unlabeled Million-AID training images through both pipelines. For each image, we compute the spatial mean of the rejection confidence and subtract the CompatLift mean from the native mean. Averaging these differences yields a single offset.

4 Experiments

4.1 Experimental Setup

Dataset. Following SegEarth-OV (Li et al., 2025a) and SegEarth-OV3 (Li et al., 2025b), we separate the unlabeled training imagery from the segmentation benchmarks used for evaluation. CompatLift is trained only with Million-AID imagery (Long et al., 2021). No segmentation labels, category prompts, or decoder outputs are used for training. We evaluate on eight remote-sensing open-vocabulary segmentation benchmarks: OpenEarthMap (OEM) (Xia et al., 2023), LoveDA (Wang et al., 2021), iSAID (Zamir et al., 2019), Potsdam, Vaihingen[1], UAVid (Lyu et al., 2020), UDD5 (Chen et al., 2018), and VDD (Cai et al., 2025). These datasets cover the multi-class remote-sensing segmentation setting used by SegEarth-OV3 and provide the test scenes for all reported comparisons.

Implementation Details. We use the frozen SAM 3 image encoder, text encoder, and SegEarth-OV3 prediction pipeline. CompatLift learns one $2\times$ lifting stage on the deepest FPN feature level, increasing the feature resolution from 36×36 to 72×72 with 256 channels. The remaining FPN features and positional encodings are resized using bicubic interpolation. Training uses paired 504×504 and 1008×1008 views, a batch size of 20, NAdam with a learning rate of 10^{-4}, and 2,000 optimization steps. All experiments are conducted on a single NVIDIA A100 GPU. The residual mixing coefficient is set to 0.2. The score calibration δ is set to 0.04. We report mIoU and the

[1] `https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab`.

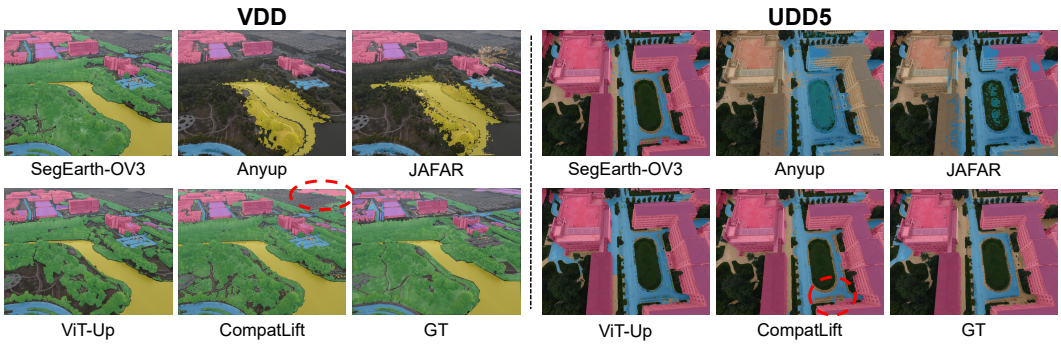


Figure 5: **Qualitative comparison on VDD and UDD5.** Each group shows the same scene for SegEarth-OV3, AnyUp, JAFAR, ViT-Up, CompatLift, and the ground truth. Red dashed circles mark a distant small building in VDD and parked bicycles in UDD5, illustrating cases where CompatLift better separates fine structures from surrounding classes.

Table 2: **Compatibility of feature lifting under the frozen SegEarth-OV3 pipeline.** Values are mIoU (%). The native and bicubic rows are references. Each "+ Res mixing" row applies the residual mixing coefficient g to the preceding baseline without updating its learned weights. Arrows show the signed mIoU change from that baseline. AnyUp uses released weights. JAFAR, ViT-Up, and CompatLift are adapted to SAM 3. Residual mixing is inserted at the deepest FPN feature for AnyUp and JAFAR and before the FPN for ViT-Up. All rows use the frozen SegEarth-OV3 prediction pipeline.

Method	OEM	LoveDA	iSAID	Potsdam	Vaihingen	UAVid	UDD5	VDD	Avg.
SegEarth-OV3	42.9	47.4	27.6	**57.8**	60.8	54.7	71.7	64.5	53.4
+ Bicubic 2×	40.9	47.1	27.2	57.3	60.3	54.4	70.9	64.5	52.8
AnyUp	23.5	16.4	19.6	24.1	40.8	26.5	47.0	28.0	28.2
+ Res mixing	40.4↑ 16.9	47.1↑ 30.7	27.9↑ 8.3	57.0↑ 32.9	60.3↑ 19.5	54.4↑ 27.9	71.2↑ 24.2	61.3↑ 33.3	52.5↑ 24.3
JAFAR	25.1	18.5	17.1	22.8	36.5	25.7	37.7	32.5	27.0
+ Res mixing	40.8↑ 15.7	47.0↑ 28.5	28.1↑ 11.0	56.9↑ 34.1	60.2↑ 23.7	54.4↑ 28.7	70.9↑ 33.2	61.8↑ 29.3	52.5↑ 25.5
ViT-Up	42.8	43.9	23.7	55.4	59.8	51.1	69.8	56.2	50.3
+ Res mixing	40.5↓ 2.3	46.4↑ 2.5	27.9↑ 4.2	56.3↑ 0.9	59.5↓ 0.3	53.3↑ 2.2	71.7↑ 1.9	62.7↑ 6.5	52.3↑ 2.0
CompatLift	**45.3**	**47.9**	**29.6**	57.4	**61.1**	**55.3**	**72.0**	**65.0**	**54.2**

unweighted mean over eight benchmarks. All methods use the frozen SegEarth-OV3 decoder and the same evaluation pipeline. For CompatLift, the score offset is estimated from 1,000 unlabeled Million-AID images.

Benchmark. To evaluate the performance of our CompatLift, we compared it with various state-of-the-art methods. The matched group contains native inference, bicubic resizing, AnyUp (Wimmer et al., 2026), JAFAR (Couairon et al., 2026), ViT-Up (Wandel et al., 2026), and CompatLift. AnyUp uses its released weights, whereas JAFAR, and ViT-Up are adapted to SAM 3. For JAFAR, we replace its feature extractor with a frozen SAM 3 backbone and upsample the deepest 256-channel FPN feature using paired 504×504 and 1008×1008 inputs. For ViT-Up, we replace the DINO backbone with a frozen SAM 3 trunk adapter using layers $\{0, 8, 16, 24, 32\}$, and reconstruct the native FPN representation through the frozen SAM 3 neck. Only the corresponding upsampling modules are trained. The SAM 3 decoder and SegEarth-OV3 pipeline remain frozen for all methods. Additional comparison methods and adaptation details are provided in the supplementary material.

4.2 Open-vocabulary segmentation performance

Table 1 reports the quantitative comparison on eight remote-sensing open-vocabulary segmentation benchmarks. Among target-dataset training-free methods, SAM 3-based approaches, SegEarth-OV3 and proposed CompatLift, substantially outperform other CLIP-based methods. They also surpass methods trained on remote-sensing segmentation data on all benchmarks except iSAID. Within the frozen SegEarth-OV3 pipeline, CompatLift improves the native baseline from 53.4 to 54.2,

Table 3: **Component ablation of CompatLift.** Values are mIoU on eight remote-sensing datasets, and Avg. is their unweighted mean. All variants use the frozen SAM 3 encoder and SegEarth-OV3 prediction components.

Method	OEM	LoveDA	iSAID	Potsdam	Vaihingen	UAVid	UDD5	VDD	Avg.
V1	40.9	47.1	27.2	57.3	60.3	54.4	70.9	64.5	52.8
V2	44.2	46.6	28.8	**57.5**	61.1	54.6	72.0	64.5	53.7
V3	32.3	28.6	18.3	34.2	51.3	33.5	49.2	52.8	37.5
V4	43.1	47.7	29.4	56.8	60.0	55.0	71.6	65.0	53.6
V5	**45.3**	**47.9**	**29.6**	57.4	**61.1**	**55.3**	**72.0**	**65.0**	**54.2**

corresponding to a gain of 0.8 mIoU. It also outperforms bicubic feature resizing, which obtains 52.8. CompatLift improves over native SegEarth-OV3 on seven of the eight datasets. The largest gains are observed on OpenEarthMap ($+2.4$) and iSAID ($+2.0$), while performance on Potsdam decreases by 0.4. These results suggest that a decoder-compatible feature update can improve open-vocabulary segmentation while keeping the pretrained SegEarth-OV3 prediction pipeline frozen.

4.3 Compatibility Analysis

Existing feature upsampling methods increase the spatial resolution of visual features, but their compatibility with a fixed prediction pipeline is not guaranteed. We therefore insert the output of each upsampler at its corresponding feature interface and evaluate all methods using the frozen SegEarth-OV3 decoder.

Table 4: Parameter counts of the feature-lifting modules. The encoder, text encoder, and prediction components are frozen.

Method	AnyUp	JAFAR	ViT-Up	CompatLift
parameters (M)	0.878	0.596	114.720	0.111

As shown in Table 2, direct feature replacement can substantially degrade segmentation performance. AnyUp, JAFAR, and ViT-Up obtain average mIoU scores of 28.2, 27.0, and 50.3, respectively, compared with 53.4 for native SegEarth-OV3. These results show that increasing feature resolution alone does not guarantee compatibility with the frozen prediction heads. Bicubic resizing obtains 52.8, indicating that it introduces a smaller change to the feature representation. For the external baselines, the "+ Res Mixing" rows apply the same residual coefficient $g = 0.2$ without updating their learned weights. This control improves AnyUp, JAFAR, and ViT-Up to 52.5, 52.5, and 52.3, respectively. Among the residual mixed results, CompatLift achieves the highest average mIoU of 54.2. The remaining performance gap suggests that learning the residual update relative to the bicubic anchor provides an additional benefit beyond applying a mix coefficient alone.

Figure 5 further compares the predictions on VDD and UDD5. In the VDD example, the red dashed circle marks a small, distant building. CompatLift identifies this building as a distinct region and recovers its local extent, whereas SegEarth-OV3, AnyUp, JAFAR, and ViT-Up do not produce a clear building prediction in this area. In the UDD5 example, the highlighted region contains parked bicycles close to road and building regions. CompatLift separates the bicycle pixels from both the road and building predictions. AnyUp and JAFAR separate part of this region, but their boundaries are substantially distorted. ViT-Up preserves the coarse segmentation structure, yet fails to recover the fine bicycle details. These examples provide qualitative evidence that the residual feature update improves local object discrimination and boundary quality. Additional qualitative comparisons are provided in the supplementary material.

Table 4 compares the parameter counts of the feature-lifting modules. CompatLift uses only 0.111M parameters, fewer than AnyUp (0.878M), JAFAR (0.596M), and ViT-Up (114.720M). Despite using the smallest lifting module, CompatLift achieves the best average mIoU in Table 2. These values describe module size and should not be interpreted as end-to-end parameters.

4.4 Ablation Study

To evaluate the contributions of learned feature lifting, residual mixing, and score calibration, we compare five configurations: V1) SegEarth-OV3 with bicubic feature upsampling. V2) V1 with score

calibration. V3) CompatLift without residual mixing, retaining score calibration. V4) CompatLift without score calibration, retaining residual mixing. V5) The full CompatLift.

Table 3 reports the results on eight datasets. Score calibration improves the bicubic baseline from 52.8 (V1) to 53.7 (V2), showing that adjusting background rejection benefits even parameter-free feature upsampling. Without calibration, CompatLift (V4) achieves 53.6, exceeding V1 by 0.8 mIoU. Adding calibration further improves its performance to 54.2 (V5), surpassing the calibrated bicubic baseline by 0.5. Removing residual mixing (V3) reduces the average mIoU to 37.5, highlighting its importance for integrating learned features into the frozen prediction pipeline. Together, these comparisons demonstrate that anchored feature refinement and score calibration each contribute to the final performance, and that their combination is beneficial.

4.5 Discussion

Effect of spatial resolution. Figure 6 examines the effect of resolution and its computational cost. The left panel shows that reducing the input resolution from $1\times$ to $0.5\times$ substantially degrades segmentation performance, whereas increasing it to $4\times$ does not provide further improvement over $2\times$. The right panel shows that enlarging the feature grid increases downstream processing time even when the encoder input is fixed. Together, these results suggest that $2\times$ provides a reasonable accuracy–cost trade-off for feature lifting.

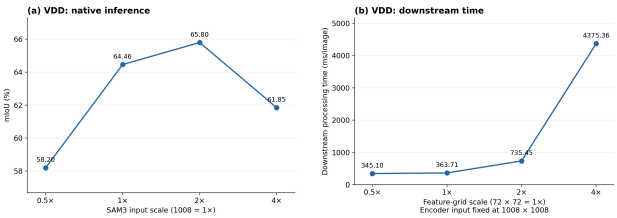


Figure 6: **Resolution–cost trade-off on VDD.** (a) Native SegEarth-OV3 mIoU (%) as the SAM 3 input scale varies relative to 1008×1008. Lower scales are obtained by downsampling the original 4000×3000 image. (b) Downstream processing time (ms/image) as the feature-grid scale varies with the encoder input fixed at 1008×1008.

Decoder confidence shift and calibration. Figure 4 shows that CompatLift largely preserves the relative confidence ordering but introduces a nearly constant downward shift in decoder scores. We therefore add a fixed offset estimated from 1,000 unlabeled Million-AID images before background rejection. This correction preserves category rankings. Its effect is limited to restoring the confidence level used by the rejection threshold. The ablation in Table 3 confirms its contribution, with score calibration improving the average mIoU from 53.6 to 54.2.

4.6 Limitations

Our evaluation focuses on $2\times$ feature lifting within the frozen SAM 3-based SegEarth-OV3 pipeline. Extending CompatLift to larger lifting factors or other encoder–decoder architectures requires further investigation of the feature interfaces and training configuration. In addition, the evaluated benchmarks primarily contain RGB optical imagery. Its applicability to other remote-sensing modalities, such as multispectral and synthetic aperture radar imagery, remains to be explored.

5 Conclusion

In this paper, we studied feature lifting for open-vocabulary segmentation with frozen prediction components. We showed that directly inserting learned upsampled features into the SegEarth-OV3 pipeline can degrade segmentation performance, even when the feature grid is spatially finer. To address this issue, we proposed CompatLift, which combines a bicubic reference with an image-guided residual update and is trained with same-encoder high-resolution supervision. A simple score calibration step further compensates for the score shift caused by feature replacement. Experiments on eight remote-sensing benchmarks show that CompatLift achieves 54.2 mean mIoU, compared with 53.4 for native SegEarth-OV3 and 52.8 for bicubic feature resizing. It improves over the native SegEarth-OV3 on seven of the eight datasets. These results demonstrate the potential of anchored residual lifting to improve spatial detail while retaining the behavior required by the evaluated frozen SAM 3-based segmentation pipeline.

REFERENCES

Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3828–3837. IEEE, 2024.

Wenxiao Cai, Ke Jin, Jinyan Hou, Cong Guo, Letian Wu, and Wankou Yang. Vdd: Varied drone dataset for semantic segmentation. *Journal of Visual Communication and Image Representation*, 109:104429, 2025.

Qinglong Cao, Yuntian Chen, Chao Ma, and Xiaokang Yang. Open-vocabulary remote sensing image semantic segmentation. *arXiv preprint arXiv:2409.07683*, 2024.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 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.

Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungyong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4113–4123, 2024.

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.

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.

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European conference on computer vision*, pp. 540–557. Springer, 2022.

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

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *European Conference on Computer Vision*, pp. 143–160. Springer, 2024a.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *European Conference on Computer Vision*, pp. 70–88. Springer, 2024b.

Bingyu Li, Haocheng Dong, Da Zhang, Zhiyuan Zhao, Hao Sun, and Junyu Gao. Exploring efficient open-vocabulary segmentation in the remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5982–5991, 2026.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022.

Kaiyu Li, Xiangyong Cao, and Deyu Meng. A new learning paradigm for foundation model-based remote-sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024.

Kaiyu Li, Ruixun Liu, Xiangyong Cao, Xueru Bai, Feng Zhou, Deyu Meng, and Zhi Wang. Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10545–10556, 2025a.

Kaiyu Li, Shengqi Zhang, Yujie Wang, Yupeng Deng, Zhi Wang, Deyu Meng, and Xiangyong Cao. Segearth-ov3: Exploring sam 3 for open-vocabulary semantic segmentation in remote sensing images. *arXiv preprint arXiv:2512.08730*, 2025b.

Yu-Jie Liang, Zihan Cao, Shangqi Deng, Hong-Xia Dou, and Liang-Jian Deng. Fourier-enhanced implicit neural fusion network for multispectral and hyperspectral image fusion. *Advances in neural information processing systems*, 37:63441–63465, 2024.

Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3491–3500, 2024.

Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and Million-AID. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:4205–4230, 2021. doi: 10.1109/JSTARS.2021.3070368.

Hao Lu, Wenze Liu, Zixuan Ye, Hongtao Fu, Yuliang Liu, and Zhiguo Cao. Sapa: Similarity-aware point affiliation for feature upsampling. *Advances in Neural Information Processing Systems*, 35: 20889–20901, 2022.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Saksham Suri, Matthew Walmer, Kamal Gupta, and Abhinav Shrivastava. Lift: A surprisingly simple lightweight feature transform for dense vit descriptors. In *European Conference on Computer Vision*, pp. 110–128. Springer, 2024.

Datao Tang, Xiangyong Cao, Xuan Wu, Jialin Li, Jing Yao, Xueru Bai, Dongsheng Jiang, Yin Li, and Deyu Meng. Aerogen: Enhancing remote sensing object detection with diffusion-driven data generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3614–3624. IEEE, 2025.

Krispin Wandel, Jingchuan Wang, and Hesheng Wang. Vit-up: Faithful feature upsampling for vision transformers. *arXiv preprint arXiv:2606.14024*, 2026.

Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *European conference on computer vision*, pp. 315–332. Springer, 2024.

Jiaqi Wang, Kai Chen, Rui Xu, Ziwei Liu, Chen Change Loy, and Dahua Lin. Carafe: Content-aware reassembly of features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3007–3016. IEEE, 2019.

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.

Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2945–2954, 2023.

Chengyang Ye, Yunzhi Zhuge, and Pingping Zhang. Towards open-vocabulary remote sensing image semantic segmentation. In *AAAI Conference on Artificial Intelligence*, 2025.

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.

Dengke Zhang, Fagui Liu, and Quan Tang. Corrclip: Reconstructing patch correlations in clip for open-vocabulary semantic segmentation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24677–24687. IEEE, 2025.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European conference on computer vision*, pp. 696–712. Springer, 2022.

Qi Zhu, Jiangwei Lao, Deyi Ji, Junwei Luo, Kang Wu, Yingying Zhang, Lixiang Ru, Jian Wang, Jingdong Chen, Ming Yang, et al. Skysense-o: Towards open-world remote sensing interpretation with vision-centric visual-language modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14733–14744, 2025.

A Implementation Details of Image Guidance and JBU

Given $I_L \in \mathbb{R}^{3\times H\times W}$, we compute its local high-frequency component by subtracting a 3×3 mean-filtered image:

$$I_{\text{low}} = \text{AvgPool}_{3\times 3}\left(\text{ReflectPad}_1(I_L)\right), \qquad \text{HP}(I_L) = I_L - I_{\text{low}}. \tag{12}$$

The horizontal and vertical gradients are computed using one-pixel forward differences:

$$\nabla_x I_L(i,j) = I_L(i,j+1) - I_L(i,j), \qquad \nabla_y I_L(i,j) = I_L(i+1,j) - I_L(i,j). \tag{13}$$

Replication padding is used at the right and bottom boundaries. The four three-channel components are concatenated to form a $12 \times H \times W$ guidance tensor.

The JBU branch first applies bilinear interpolation to F_L and then performs local adaptive aggregation:

$$F_H^{\text{prop}}(x) = \sum_{y\in\mathcal{N}_5(x)} \overline{K}_\theta\big(x, y; G(I_L)\big) F_L^{\uparrow}(y), \tag{14}$$

where $\mathcal{N}_5(x)$ is a radius-5 neighborhood and $\overline{K}_\theta$ is the normalized spatially adaptive kernel predicted from the guidance tensor. Thus, the kernel operates on the bilinearly upsampled feature rather than directly on the original low-resolution feature.

B Additional experiments

B.1 Benchmark

Following SegEarth-OV3, to evaluate the performance of our CompatLift, we compared it with various state-of-the-art open-vocabulary semantic segmentation methods, including SCAN (Liu et al., 2024), SAN (Xu et al., 2023), SED (Xie et al., 2024), Cat-Seg (Cho et al., 2024), OVRS (Cao et al., 2024), GSNet (Ye et al., 2025), RSKT-Seg (Li et al., 2026), SkySense-O (Zhu et al., 2025), CLIP (Radford et al., 2021), MaskCLIP (Zhou et al., 2022), SCLIP (Wang et al., 2024), GEM (Bousselham et al., 2024), ClearCLIP (Lan et al., 2024a), SegEarth-OV (Li et al., 2025a), ProxyCLIP (Lan et al., 2024b), CorrCLIP (Zhang et al., 2025), and SegEarth-OV3 (Li et al., 2025b). Among them, the first eight methods are trained with remote-sensing segmentation data or other domain-specific supervision, while the remaining nine methods are training-free and do not use target-dataset segmentation annotations for training.

B.2 Baseline adaptation

AnyUp. We use the released AnyUp checkpoint with frozen parameters and no additional SAM3 training. SAM3-normalized guidance is converted to RGB and ImageNet-normalized before being fed to AnyUp. AnyUp upsamples the original 256-channel SAM3 FPN level-2 feature by 2×, while the remaining pyramid features are resized with bicubic interpolation.

JAFAR. The SAM 3 adaptation learns from paired 504×504 and 1008×1008 images using the same frozen encoder, with level-2 256-channel features as the student input and teacher target. Its image normalization uses mean and standard deviation 0.5. The recorded training run uses batch size 4, one epoch, and the final beat checkpoint. The training loop does not select this checkpoint through a validation loader. Feature normalization inside the attention keys does not replace the original feature values used to form the output. The teacher-range normalization in its loss rescales the feature error.

ViT-Up. The frozen-backbone SAM 3 variant uses trunk layers $[0, 8, 16, 24, 32]$, four lifting blocks, and five feature decoders. It retains the layer-wise transitions, cross-attention, local FiLM component, and feature losses. LoRA is disabled. Teacher features at image sizes 504, and 1008 supervise all five layers with L2, cosine, and relational KL losses. Training uses Million-AID imagery, batch size 24, AdamW with learning rate 2×10^{-4}, and a configured budget of $20K$ steps. The selected best checkpoint corresponds to the lowest observed validation loss in the saved training trace. The adaptation predicts a dense 1024-channel trunk feature and then applies the frozen neck to construct the FPN, with positional encodings recomputed for that pyramid. We identify it as a SAM 3-adapted system baseline because this intervention differs from replacing only the deepest FPN feature.

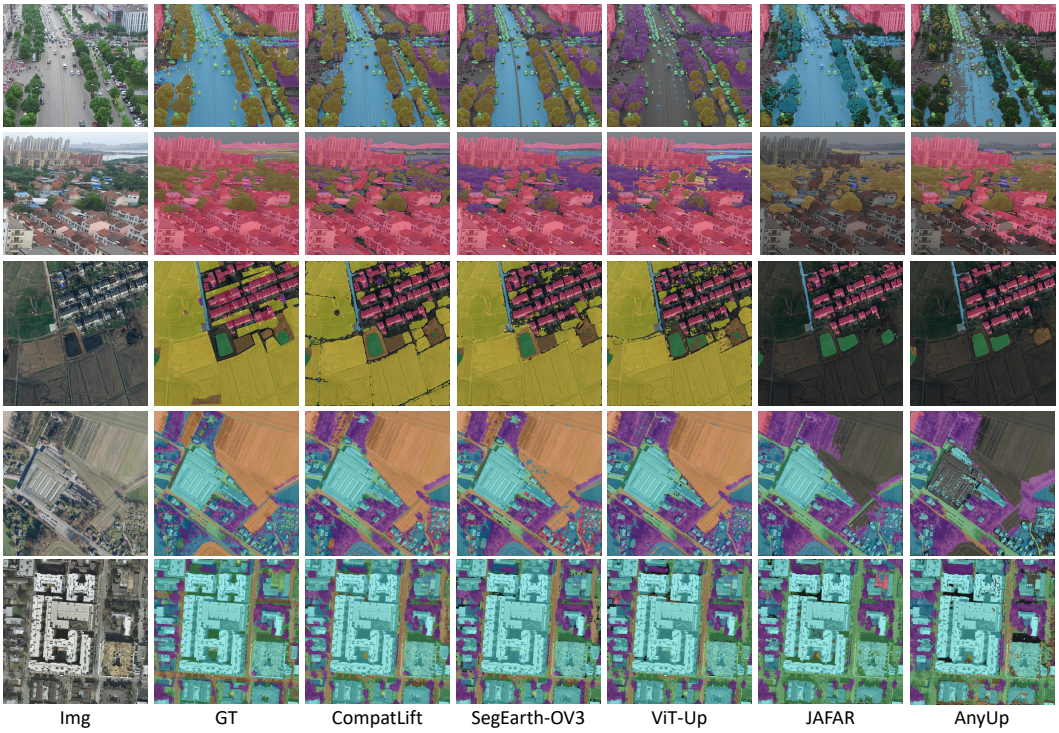


Figure 7: **Additional qualitative comparisons.** Rows 1–2 show UAVid examples, row 3 shows a LoveDA example, and rows 4–5 show OpenEarthMap examples. From left to right: input image (Img), ground truth (GT), CompatLift (ours), SegEarth-OV3, ViT-Up, JAFAR, and AnyUp. Segmentation masks are overlaid on the input images, with colors indicating semantic categories within each dataset.

B.3 Additional Qualitative Results

Figure 7 presents additional segmentation examples from UAVid, LoveDA, and OpenEarthMap. On UAVid (rows 1–2), CompatLift preserves the overall layout of roads, buildings, and vegetation. In the first example, it retains the continuous road region, whereas ViT-Up leaves substantial portions of this region unsegmented. In the second example, JAFAR misses much of the building region, while CompatLift produces more complete building predictions. On LoveDA (row 3), CompatLift captures both the agricultural fields and the adjacent residential area, whereas JAFAR and AnyUp leave large portions of the fields unsegmented. On OpenEarthMap (rows 4–5), CompatLift recovers the major land-cover regions and building footprints, although errors remain around boundaries and small structures.

These examples illustrate that CompatLift can preserve broad semantic regions while retaining local spatial structure. The visual differences are more pronounced relative to JAFAR, Vit-Up, and AnyUp, while the predictions of CompatLift and native SegEarth-OV3 remain broadly better in several scenes.

AI use statement

Generative AI tools were used for language editing and readability improvements. The authors reviewed all suggested revisions to ensure that they preserved the intended technical meaning. No generative AI tools were used for research design, method implementation, data processing, or interpretation of experimental results.