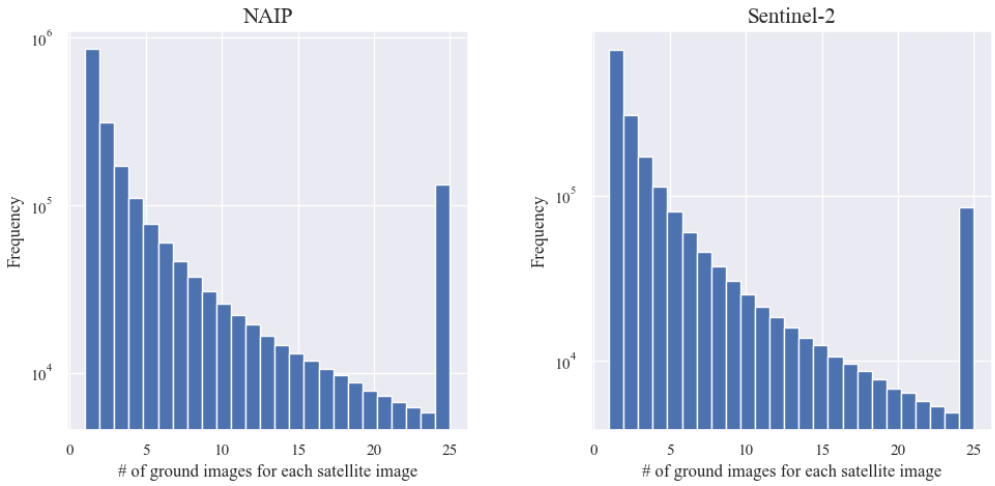


Figure 7: Distribution of the number of images corresponding to a single satellite images in our NAIP (Left) and Sentinel-2 (Right) datasets.

# A Training Dataset Details

The dataset at NAIP resolution contains 10.2 million pairs of satellite-ground image pairs. It contains 2.0 million unique satellite images and 7.9 million unique ground images. Note that while the overlap between satellite images is minimal they are not completely non-overlapping. As a result, the same ground image can be a part of two different satellite images in our training data. Fig. 7 (left) shows the distribution of ground images for each NAIP image in our dataset. The distribution of the number of ground images per satellite image is a decreasing function. The number of satellite images with 25 images is much higher than the rest because we subsample 25 images whenever a satellite image contains more than 25 ground images.

**NAIP Preprocessing:** We download the RGB bands of NAIP images. Since NAIP imagery is already stored in a normalized fashion we do not need to take anymore additional steps.

The dataset at Sentinel-2 resolution contains 8.7 million pairs of satellite-ground image pairs. It contains 1.9 million unique satellite images and 7.6 million unique ground images. Fig. 7 (right) shows the distribution of ground images for each Sentinel-2 image which follows the same trend as NAIP dataset. Fig. 8 shows example of paired data from our NAIP and Sentinel-2 dataset.

**Sentinel-2 Preprocessing:** We download the B4, B3, B2 (corresponding to RGB) bands of Sentinel-2 images. Note that the data from Sentinel-2 is not coming from an RGB camera sensor. The values in B4, B3, and B2 capture the reflectance for a specific wavelength close to Red, Green, and Blue. We downscale these intensities by 3000 in order to get images that approximate RGB camera sensors (see Fig. 8 (bottom)). This is a standard practice followed by other works as well(Mall et al., 2023; Manas et al., 2021).

## A.1 Spatial and Temporal Alignment

**Temporal alignment:** The temporal revisit for Sentinel-2 is about a week. For every Flickr image, we collect the temporally closest cloudless satellite image (see Sec. 3.2). The Flickr images in our dataset go back to 2014 while the earliest sentinel images we can get are from 2017. For the first three years of Flickr data, we sample the closest seasonally aligned images. On average the temporal gap between a Flickr image and a Sentinel image is around 21 days. So while it is difficult to connect ground images with satellite images of the same day, we can still align seasonality.


Figure 8: Examples of pairs from our NAIP and Sentinel-2 dataset. Each satellite image can contains multiple ground images. The ground images are shown at their corresponding location within the satellite images. NAIP (top) and Sentinel-2 (bottom).

**Spatial alignment:** As stated in Sec. 3.2, the Flickr API also provides accuracy in the geotags along with the geotags. We only use the images with the highest level of geotag accuracy (street level). However, it is not unreasonable to think that some of these geotags can still be incorrect. There are two reasons why we believe our model is robust to such noise. First, prior work on image-text contrastive learning such as OpenCLIP (Cherti et al., 2023) has shown that the training loss and architectures for such vision language models are robust even at a lower signal-to-noise ratio. Second, the patches we contrast image features with are big enough (14m for NAIP and 140m for Sentinel-2) that our model is invariant to a small amount of noise in geotagging.

# B PROMPTING

Like CLIP, we test our models using template prompts such as `a photo of a {label}`. For classification metrics, with labels $l_j \in \mathbf{L}$, an image $x$ can be classified as label $l_{j^*}$,

$$j^* = \arg\max_j f_S^I(x).f_T(\texttt{a photo of a } \{l_j\}) \tag{4}$$

In practice, we use multiple prompts and use the average text representation from those. For GRAFT we use the following prompts: `A photo of a {label}`, `A photo taken from inside a {label}`, `I took a photo from a {label}`. For baselines, we use the prompts prescribed in CLIP: `A centered satellite photo of {label}`, `A centered satellite photo of a {label}`, `A centered satellite photo of the {label}`.

Note that since our model is trained with internet images as an intermediary, it performs better with prompts describing internet imagery such as: `a photo of a {label}`. Whereas CLIP performs better with prompts describing satellite images such as `a satellite image of a {label}`.

# C ADDITIONAL IMPLEMENTATION DETAILS

## C.1 TRAINING DETAILS

Regardless of the training datasets (NAIP or Sentinel-2), we train all models for 10 epochs using AdamW with weight decay set to 1e-2. For image-level model, we linearly ramp up the learning rate from 0 to 1e-5 and then decrease the learning rate using a cosine schedule. For pixel-level model, we linearly ramp up the learning rate from 0 to to 5e-5 and then decrease the learning rate to zero using a cosine schedule. All models are initialized using CLIP's weights and the temperature hyperparameter is set to $\tau = 0.07$.

## C.2 INTERNAL VALIDATION DATASET

For the development of the image-level models, we collected an internal validation set using OpenStreetMap (OSM contributors, 2023). This validation set contains 14 categories with a total of 2632 single-label images. We measured the performance of our image-level models on this dataset using mean average precision. For developing pixel-level model, we use a subset of NAIP-OSM we collected (see Appendix C.3). This subset contains 32 categories with 17k images. We select the best performing pixel-level models based on the zero-shot accuracy of patch prediction.

## C.3 IMAGE-LEVEL UNDERSTANDING EVALUATION DETAILS.

We measured our image-level model on 5 different datasets. The low-resolution Sentinel-2 model is evaluated on EuroSat and BigEarthNet Datasets.

**Eurosat:** We use the validation set of EuroSAT as our test set (the real test set is not public). It contains 5400 64×64 images from Sentinel-2 imagery from Europe. Each image is labeled with a single label out of 10 classes.

Table 5: List of 33 categories in the proposed NAIP-OSM dataset.

| | | | | |
|---|---|---|---|---|
| airport | football field | baseball field | beach | bridge |
| cemetery | commercial area | dam | equestrian facility | farmland |
| forest | garden | golf course | highway | marina |
| parking garage | park | parking lot | pond/lake | railroad |
| residential area | river | roundabout | sand area | school building |
| shooting range | soccer field | supermarket | swimming pool | tennis court |
| university building | warehouse | wetland | | |

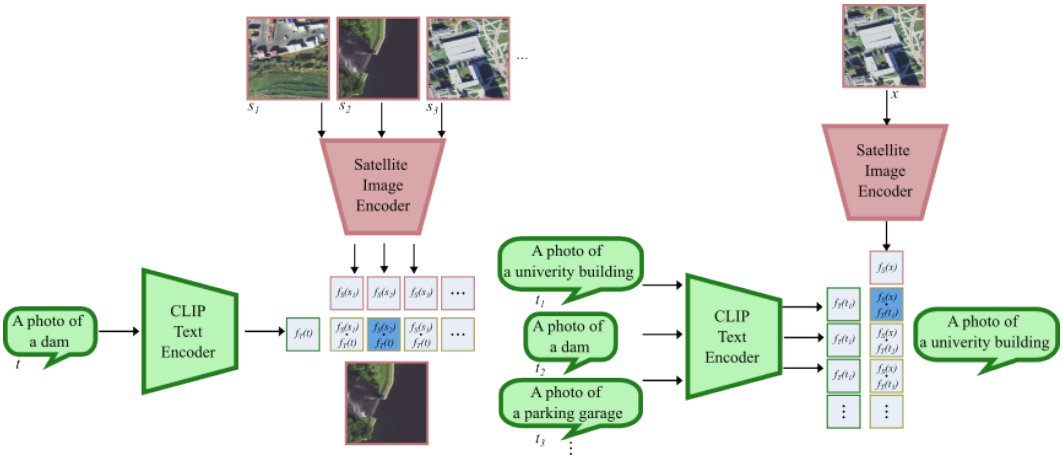


Figure 9: Inferring for text-to-image retrieval (left) and zero-shot classification (right) tasks with GRAFT.

**BigEarthNet (BEN):** BigEarthNet (Sumbul et al., 2019) is a 19-class multilabel classification dataset. The test set contains 104k 120×120 image from Sentinel-2.

The high-resolution NAIP model is evaluated on 3 datasets.

**Sat-4 and Sat-6:** The test set of the SAT-4 and SAT-6 datasets contains 64k and 81k 28×28 images respectively. SAT-4 and SAT-6 are labeled with 4 and 6 categories respectively on NAIP images. The fourth category of SAT-4 dataset is 'None', which is not suitable for zero-shot evaluation, so we evaluate with the remaining three categories only.

**NAIP-OSM:** Finally, since there are not a lot of evaluation benchmarks on NAIP images. We leverage OpenStreetMap data to create a large-scale multi-label classification and retrieval dataset for NAIP. This dataset contains 33 category labels and 1.7 million 224×224 NAIP images. Tab. 5 shows the list of categories in this dataset.

Since for many of these benchmarks the images are smaller than the resolution our models and baselines support. We try a variety of image pre-processing and report the best number for each method. We try zero padding, reflection padding, and resizing transforms and report the best numbers for each method. CLIP and CLIP-RSICD work best with resizing to 224×224 images. For GRAFT ViT-B/16 we use zero-padding as it results in the best performance. For RemoteCLIP, CACo, and SeCo, we use the pre-processing proposed by them. For Satlas, reflection padding works the best. When evaluating GRAFT ViT-B/32 on EuroSat, we had to pad in images in a way that the 64×64 images align with the input patches of ViT-B/32.

Fig. 9 illustrates how we use our model for text-to-image retrieval and zero-shot classification on these 5 benchmarks.

For multiclass classification (EuroSAT), we report the top1 accuracy. For the multi-label classification task (BigEarthNet), we report mean average precision (area under precision-recall curve). For retrieval tasks, we report the standard metric used for evaluating ranking: mAP@100 and mAP@20.

### C.4 SEGMENTATION

Our segmentation results are evaluated on the newly proposed SatlasPretrain (Bastani et al., 2023) segmentation task. SatlasPretrain contains 2 test sets for NAIP and Sentinel-2 images. For NAIP, we use the test set along with the 2020 NAIP imagery. This results in a total of about 39k 512×512 NAIP images. The segmentation labels consist of 11 landcover classes (3 of which never occur in the test set). We evaluate the per-class accuracy on the remaining 8 classes. While the NAIP images are of size 512×512, the landcover labels are at a lower resolution of 32×32. With a ViT-B/16 model, this allows us to evaluate patch prediction without having to upscale either using interpolation or other bottom-up segmentation models such as SAM. For evaluation, we divide the 512×512 images into 4 224×224 and pass them through our models and baselines.

For Sentinel-2, we use the test set along with the *sentinel-2-small* imagery. This results in a total of about 2183 512×512 Sentinel-2 images. Same as before we evaluate with 8 out of 11 classes. Similiar to NAIP, for evaluation, we divide the 512×512 images into 4 224×224 and pass them through our models and baselines.

However, for Sentinel-2, the landcover masks are also of size 512×512. Therefore we evaluate GRAFT in two ways. Firstly, we evaluate the performance by upscaling the logits (bicubic interpolation) and using the maximum logit value as a prediction. This gives us the performance of the standalone GRAFT model. Alternatively, we also combine our model with SAM. In order to perform semantic segmentation, we use SAM's segment everything mode to get bottom-up segments. Then we assign each segment the class corresponding to the majority of the pixels in it. Our experiments in Sec. 4 suggest that using SAM in leads to a minor improvement in performance, suggesting the standalone GRAFT can also perform good segmentation.

### C.5 VQA

As stated in the main paper, we conducted VQA evaluation with 500 unique questions from the high-resolution RSVQA testset. Note that where there are 500 unique questions, the number of image-question-answer tuples is much larger since the same question can be asked on multiple images. In total, there are 2985 image-question-answer tuples in the evaluation set (presence: 576, Area: 1161, Comparison: 590, Count: 658).

## D ADDITIONAL RESULTS

**Open Vocabulary Image Retrieval** Fig. 10 shows more examples of open-world concepts and the understanding abilities of our model. Our method can understand and retrieve several interesting open-world concepts. For example, it can retrieve very niche concepts such as campgrounds or hedge mazes. GRAFT can also be used for tracking concepts related to sustainable development. For example, we can detect solar farms, water treatment facilities, and wind farms (Fig. 2). Such use cases would be very useful to scientists interested in analyzing a region and tracking, for example, renewable sources of energy.

**Multi-spectral Satellite Images** The main contribution of this work is to demonstrate that we can train large-scale vision language models for satellite images without direct text annotations, by using ground images as an intermediary. Therefore, we can also use this technique to train a model for multi-spectral satellite images. To show this, we train another VIT-B/16 model on 100k multi-spectral sentinel images-ground image pairs. To account for the change in input channels, we change the patch embedding input dimension from 3 to 12. The weights for the RGB channels are initialized to CLIP's weights, and the rest are set to zero. On the EuroSat multi-spectral test set, this model leads to a zero-shot classification accuracy of 53.78, which is an improvement of 1.26% on the RGB model (52.52%) trained with the same set of images. This shows that our insight generalizes to multi-spectral datasets.

## E ABLATIONS

We provide additional ablations on GRAFT in this section.


Figure 10: More examples of open-world text-to-image retrieval with GRAFT. Our model can understand various fine-grained concepts such as a water treatment facility or a hedge maze and less concrete concepts such as a university quad.

### E.1 Ablations: Alt-text Alignment

In Sec. 4.4 we show that ground-satellite image alignment results in better-performing models than aligning text to satellite images. We believe the main factor for this is that the ground images can provide much better semantic features to the satellite image encoder for alignment.

We use the captions and tags from Flickr data as the alt-text annotations. Often these captions include less meaningful information such as the filenames ("1819-img.png") automatically uploaded from phones/cameras. Other times the captions are very specific and do not describe the scene in the image, e.g. "A photo from the island where grandma grew up". In both these cases, the captions do not correctly capture the right level of semantic information and therefore the images are more helpful.

The image-text pairs in other datasets such as LAION-5B Schuhmann et al. (2022) also contain such data. However, the methods trained on it can still learn something useful due to the sheer scale of the dataset (5 billion vs 10 million). We posit that, to learn directly from text we might need a dataset with a satellite image-text pair at this scale, which is harder to obtain and train with.

### E.2 Ablations: CLIP Initialization

We initialize our model from CLIP because this helps in better learning and faster convergence. We also tried training a satellite encoder from scratch however this model performed significantly worse than CLIP and Our Best model. The EuroSat zero-shot classification performance for a model trained from scratch is 42.46% vs CLIP (53.59%) and GRAFT (63.76%). Since all the baselines we compare against are also initialized from CLIP and are not trained from scratch this comparison is fair.

### E.3 Ablations: Loss Formulations

In this section, we explore different loss formulations that could be used to create a satellite image representation that aligns with CLIP's representation of the ground images. By default, GRAFT uses Eq. (3) to enforce the intuition that the network should produce a representation of the satellite image that is close to its ground images. A few other losses could achieve similar properties:

$$\mathcal{L}_2^I(\mathcal{B}, f_S^I) = \frac{1}{N_B}\sum_{i=1}^{N_B} -\log\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{\exp(f_S^I(s_i)\cdot f_G(g_i^j)/\tau)}{\sum_{a=1}^{N_B}\sum_{b=1}^{N_i}\exp(f_S^I(s_i)\cdot f_G(g_a^b)/\tau)} \tag{5}$$

$$\mathcal{L}_3^I(\mathcal{B}, f_S^I) = \frac{1}{N_B}\sum_{i=1}^{N_B} -\log\frac{\exp((f_S^I(s_i)\cdot(\bar{z}_i/||\bar{z}_i||))/\tau)}{\sum_{a=1}^{N_B}\exp(f_S^I(s_i)\cdot(\bar{z_a}/||\bar{z_a}||)/\tau)}\,,\ \bar{z}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} f_G(g_i^j) \tag{6}$$

$$\mathcal{L}_4^I(\mathcal{B}, f_S^I) = \frac{1}{N_B}\sum_{i=1}^{N_B}\frac{1}{N_i}\sum_{j=1}^{N_i}||f_S^I(s_i) - f_G(g_i^j)||_2^2 \tag{7}$$

We experiment with all these losses on the NAIP high-resolution dataset and report its performance on the internal validation set (Appendix C.2) in Tab. 6. Even though Khosla et al. (2020) have shown that $\mathcal{L}_2^I$ is suboptimal for their supervised setup, we find the performance to be on par with using the default formulation.

However, not all loss formulations are equal. Using the conventional contrastive loss with the average representations of all ground images as positive (Eq. (5)) yields a model that underperforms the GRAFT formulation. Discerning readers might also question whether it is necessary to enforce a satellite image's embedding to stay far from ground images associated with other satellite images. Performance using the l2 loss Eq. (7) suggests that not only it is important that a satellite image's embedding to stay close to all its associated ground images, it is also crucial that its embedding stays away from ground images associated to other satellite images.

### E.4 Additional Ablations: Scaling

GRAFT uses a large amount of satellite-ground image pairs to sidestep the need for textual annotations for training. To understand the behavior of GRAFT, we trained image-level VLMs using

Table 6: Performance of different image-level VLMs trained using different loss formulations to align the satellite image modality with the ground image modality. For the precise definition of the loss functions, please refer to Appendix E.3

| Formulation | mAP |
|---|---|
| $\mathcal{L}_2^I$ | 74.76 |
| $\mathcal{L}_3^I$ | 71.42 |
| $\mathcal{L}_4^I$ | 66.80 |
| Default $\mathcal{L}^I$ | 74.78 |

various amounts of satellite images and reported the performance in Fig. 11. We observe two different behaviors with images with different resolutions. For the model trained on NAIP images (left), the performance platues after $2 \times 10^5$ examples whereas the model trained on low-resolution sentinel-2 images (right) continues to scale with more data. We conjecture that the performance of NAIP models plateus because NAIP only covers the United States and a small amount of satellite images is enough to learn a good embedding.

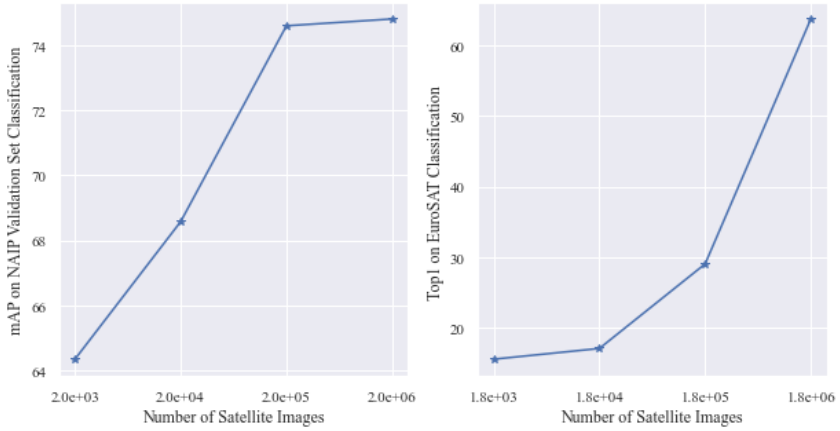


Figure 11: Performance of the Image-level VLMs trained on various amounts of satellite images using GRAFT. Left: a ViT-B/16 VLM trained on high-resolution NAIP images and evaluated on our internal validation set. Right: a ViT-B/16 VLM trained on low-resolution sentinel-2 images and evaluated on EuroSAT classification. The performance for the high-resolution model plateaus after $2 \times 10^5$ examples whereas the low-resolution Sentinel-2 model continues to scale with more data.

## F Future Work

A limitation of GRAFT and other CLIP-like VLMs is that they do not possess text-generation capabilities such as BLIP-2 (Li et al., 2023). Exploring how to use ground images as an intermediary for text generation is an interesting avenue to explore in the future. However, as our results show, we can combine our model, with other frameworks– for example with ViperGPT– to solve tasks such as VQA. In the future, we can also combine GRAFT with other frameworks such as ClipCap (Mokady et al., 2021) to get more powerful VLMs with more capabilities such as captioning.

Another limitation that might arise since we do not fine-tune the text encoder, is that the text encoder might be biased towards concepts of ground images. For example, directional concepts such as "left of" or "top of" might not correspond meaningfully in the satellite images. Despite this aspect, our method performs better than other vision-language models (some of which are supervised with text). This shows that while there might be biases, our model is robust to them for most remote sensing recognition tasks. Nonetheless, it would be interesting for future work to look into this issue and create stronger models. For instance, one possible solution for this could be to fine-tune with a small amount of satellite image-caption data.