

Sat2Cap: Mapping Fine-Grained Textual Descriptions from Satellite Images

Aayush Dhakal¹, Adeel Ahmad^{1,2}, Subash Khanal¹, Sri Kumar Sastry¹, Hannah Kerner³, Nathan Jacobs¹

¹Washington University in St. Louis, ²Taylor Geospatial Institute, ³Arizona State University

{a.dhakal, aadeel, k.subash, s.sastry, jacobsn}@wustl.edu, hkerner@asu.edu,

Abstract

We propose a weakly supervised approach for creating maps using free-form textual descriptions. We refer to this work of creating textual maps as zero-shot mapping. Prior works have approached mapping tasks by developing models that predict a fixed set of attributes using overhead imagery. However, these models are very restrictive as they can only solve highly specific tasks for which they were trained. Mapping text, on the other hand, allows us to solve a large variety of mapping problems with minimal restrictions. To achieve this, we train a contrastive learning framework called Sat2Cap on a new large-scale dataset with 6.1M pairs of overhead and ground-level images. For a given location and overhead image, our model predicts the expected CLIP embeddings of the ground-level scenery. The predicted CLIP embeddings are then used to learn about the textual space associated with that location. Sat2Cap is also conditioned on date-time information, allowing it to model temporally varying concepts over a location. Our experimental results demonstrate that our models successfully capture ground-level concepts and allow large-scale mapping of fine-grained textual queries. Our approach does not require any text-labeled data, making the training easily scalable. The code, dataset, and models will be made publicly available.

1. Introduction

Maps are a fundamental data product for a wide variety of domains. Traditional map-making involved extensive ground-based surveys. However, such methods are extremely time-consuming, expensive, and labor-intensive. As a result, overhead remote-sensing imagery has emerged as an important data modality for map creation. Machine learning methods have enabled scalable, accurate mapping of attributes using overhead imagery. The current paradigm of prior methods is to learn models that leverage the visual cues from overhead images to predict specific pre-defined attributes (e.g., a fixed set of land cover classes). Persello et al [21] describe the numerous applications of

deep learning in addressing the Sustainable Development Goals (SDGs), including crop monitoring, deforestation mapping, wildfire monitoring, and more. Salem et al. [24] used overhead images to map transient attributes [13] and scene categories [39] across large regions, while Streletsov et al. [26] predicted residential building energy consumption using overhead imagery. Similarly, Bency et al. [3] also used satellite images to map housing prices.

All these prior methods focused on learning some specific pre-defined attributes. These attribute-specific models are highly restrictive as they cannot map anything beyond their preset list of variables. To overcome this limitation, we introduce a framework that enables the mapping of fine-grained textual descriptions of concepts that are observable only on the ground. Our approach allows the mapping of any concept that can be expressed in natural language and, thus, serves as a general framework for zero-shot mapping. For example: using our model, one can create a map of concepts like “harvesting crops” or “busy streets” without training any task-specific models.

Modeling the relationship between text and images is a well-studied problem in deep learning. Numerous methods [14, 22, 35] have been proposed to learn the relationship between these two modalities. Models such as CLIP [22] and ALBEF [14] are trained on a large database of captioned images to learn a multimodal embedding space that unifies vision and text. However, we observe that directly using these models on overhead imagery leads to the collapse of representations to a few coarse textual concepts like city, beach, island, etc. Overhead images capture a broad perspective of a geolocation but offer limited insight into the intricate concepts and dynamics within the location. Hence, directly using overhead images with such models for text-driven mapping would only allow us to map coarse-level textual concepts. On the other hand, ground-level images and their respective CLIP embeddings provide detailed and fine-grained concepts of a location. Yet, several challenges hinder the direct utilization of ground-level imagery for mapping tasks. Firstly, ground-level images are sparsely available, i.e., obtaining a ground-level image for every location on Earth is not feasible. Secondly, the coverage and

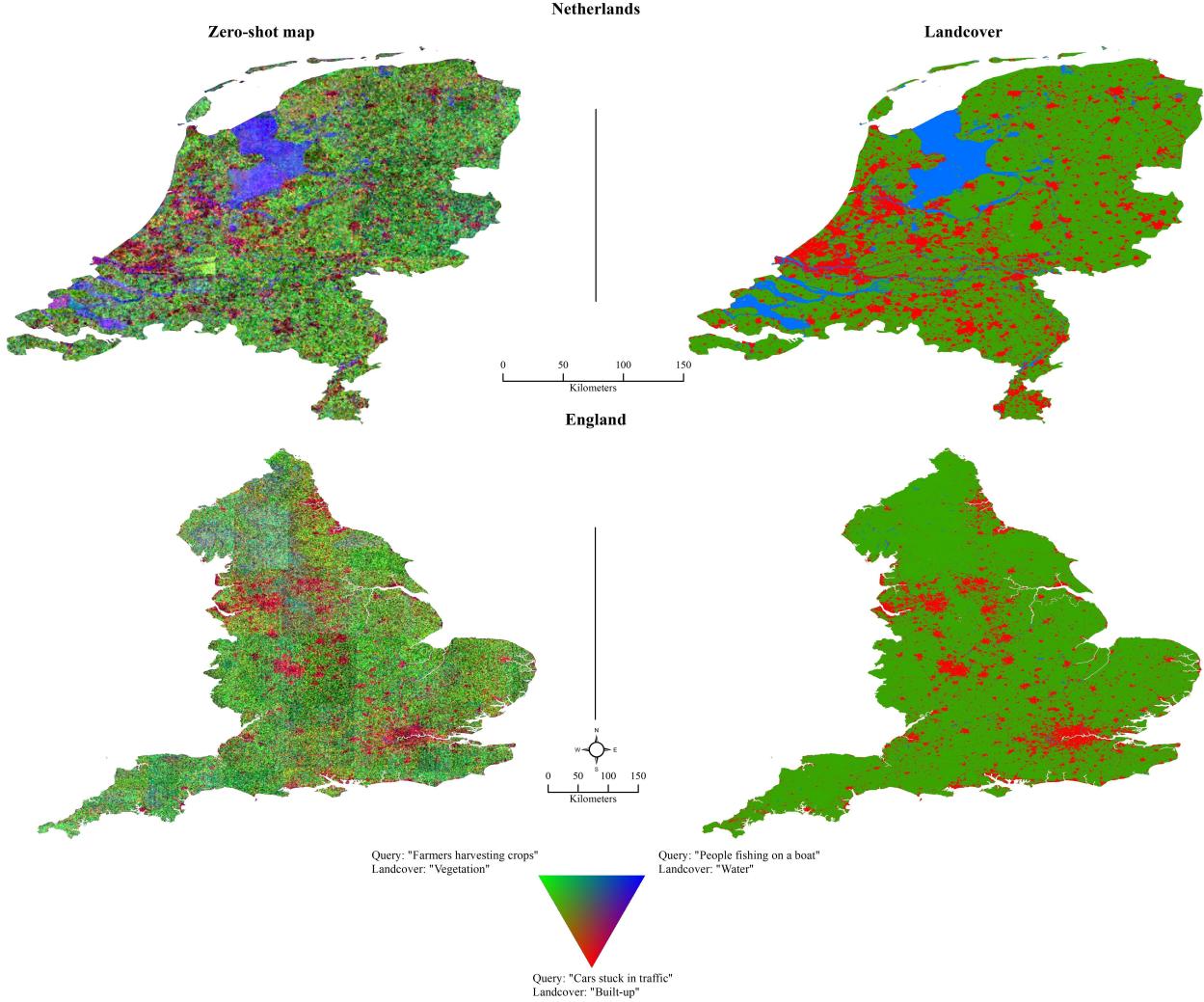


Figure 1. Country-level maps of textual descriptions: (Col 1-2) shows the country-level maps created using Sat2Cap for three prompts: “Cars stuck in traffic”, “People fishing on a boat,” and “Farmers harvesting crops.” We compare the predicted zero-shot maps with landcover maps of the region.

quality of a ground-level image for the same location can have large variations, which could introduce unwanted variations during inference.

To address these issues, we present a weakly-supervised cross-view approach for learning fine-grained textual concepts for geographic locations. In our work, “fine-grained” refers to concepts that are observable in ground-level images but are hard to infer from the low-resolution view of satellite images. To do this, we first create a large-scale dataset with paired overhead and ground-level images. Our dataset uses a subset of the YFCC100M [27]. More details about the dataset are presented in Section 3.1. Using this paired dataset, we learn the CLIP embeddings of the ground-level images at a given location. CLIP embed-

dings of ground-level images can describe detailed textual concepts of that location. Using the overhead image, our Sat2Cap model learns to predict the expected CLIP embeddings of the ground-level scene. Compared to the CLIP embeddings, Sat2Cap embeddings tend to capture more fine-grained textual concepts for a given geolocation.

To account for the temporal associations between various concepts and locations, our model is conditioned on temporal data, specifically, the date and time stamps from the Flickr imagery. This allows our model to learn concepts that can be dynamically adapted to different date and time settings. Our method is also weakly-supervised and thus does not require any text labels. To summarize, these are the primary contributions of our work:

- A weakly-supervised approach for learning fine-grained textual concepts of geographic locations
- A zero-shot approach for creating large-scale maps from textual queries as seen in Figure 1
- A new large-scale cross-view dataset

2. Related Works

2.1. Deep Learning Based Mapping

The ability to map attributes of interest is a fundamental task in Remote Sensing and has wide-ranging implications for achieving SDGs. Deep Learning methods have been used extensively [2, 8, 12, 19, 21, 40] in recent years to make mapping efficient and scalable. Alhassan et al. [1] finetuned imagenet pretrained models to make landcover predictions. Similarly [7, 11] leveraged large-scale annotated data from different sensors to improve land use and landcover classification using deep learning methods. Other works specifically focus on mapping visual attributes. For instance, [32] used features from both overhead and ground-level imagery and introduced a cross-view approach to map scenic-ness. Later works focused on creating dynamic maps: [24, 31] conditioned their model on temporal information along with overhead images to learn dynamic concepts for a given location. However, across this huge research area, the prevailing paradigm is to create task-specific models over a fixed set of attributes. We attempt to generalize the mapping process to *any* attribute by introducing a framework to create maps of free-form textual prompts.

2.2. Vision-Language Pretraining

Vision-Language (VL) models have shown great promise in their ability to model complex relationships between the vision and text space. ConVIRT [38] and VirTex [6] both introduced methods that used image-text pairs to learn rich visual representations. CLIP [22] demonstrated the results of VL pretraining on a large-scale dataset (400M pairs) and validated the efficacy of large-scale VL pretraining for several downstream tasks. Florence [36] and ALIGN [10] further increased the scale of data by training on 900M and 1.8B pairs, respectively. Other works [14, 33–35] have since focused on learning better VL embedding space. With the existence of these powerful pretrained VL models, many researchers have utilized their embedding spaces to solve specific downstream tasks. CLIPCap [17] and [4] used CLIP space to generate image captions. Models like [18, 23, 29] utilized the CLIP space for text-to-image generation.

Recently, there has been work in creating image-text datasets of overhead images and captions for VL pertaining of geospatial models. ChatEarthNet [37] created a dataset with paired image and text captions using ChatGPT. Similarly, SkyScript [30] used OpenStreetMap (OSM) data to

create overhead image and text-paired datasets for VL pre-training. However, both approaches only gather coarse textual information for a given location using either the low-resolution visual cues from an overhead image or fixed pre-set tags from OSM data.

Our approach fundamentally differs from prior work as we try to capture the more intricate concepts occurring at the ground-level of a given location by leveraging the corresponding crowdsourced images. Utilizing the ground-level images uploaded from a location allows us to access extremely fine-grained information of that location that is not distillable solely from satellite images or OSM data.

3. Method

Our objective is to learn an embedding space that describes the expected ground-level scene given a geographic location and an overhead image. We have ground-level images $\{g_1, g_2, \dots, g_n\}$, corresponding overhead images $\{o_1, o_2, \dots, o_n\}$, and respective metadata for the ground-level images $\{e_1, e_2, \dots, e_n\}$. Each e_i contains the latitude and longitude information of the sample, as well as the date and time for when the ground-level image was captured. We also have a CLIP image encoder f_θ that generates CLIP embeddings for a given ground-level image.

3.1. Dataset

We created a large-scale cross-view dataset to train our model. The ground-level images in the dataset are taken from the YFCC100M [27] dataset. The YFCC100M dataset contains 99.3 million images, which are collected from Flickr. Our cross-view dataset uses a smaller sample from this collection which excludes all US imagery. We filter out all images that are not geotagged. The resulting dataset contains 6.1M images. Each of these images has a geolocation, timestamp, and other meta-information. For each image, we download an overhead image centered at its location. Specifically, we use the Bing Maps API to download 800x800 patch satellite images at 0.6m/px resolution. We randomly sampled 100k images to be used for testing.

3.2. Approach

We initialize our Sat2Cap image encoder g_θ with the weights of f_θ . A batch of ground-level images $\{g_1, g_2, \dots, g_k\}$ is passed through the CLIP encoder to get the ground-level CLIP embeddings. These embeddings serve as the target for alignment. A batch of corresponding overhead images $\{o_1, o_2, \dots, o_k\}$ is passed through the Sat2Cap image encoder to obtain the respective embeddings:

$$G_i = f_\theta(g_i) \quad (1)$$

$$O_i = g_\theta(o_i) \quad (2)$$

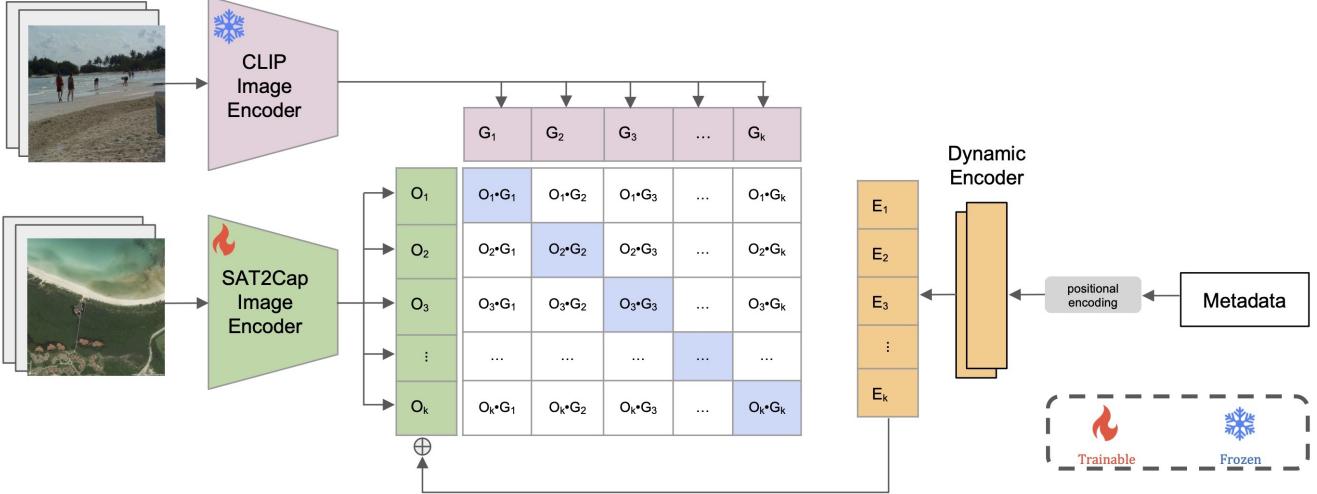


Figure 2. Sat2Cap Framework: The frozen CLIP Image Encoder takes as input the ground-level images and generates their CLIP embeddings. The trainable Sat2Cap Image Encoder takes as input the overhead images, and the Dynamic Encoder takes as input the date, time, and location information. The respective overhead image embeddings and meta-information embeddings are added element-wise, and the resulting embeddings are contrastively trained with the CLIP embeddings of the ground-level images.

To align the overhead image embeddings with the ground-level CLIP embeddings, we contrastively train our model using the InfoNCE [20] loss as follows:

$$L = \frac{1}{k} \sum_{i=0}^k -\log \frac{\exp(O_i \cdot G_i / \tau)}{\sum_{j=0}^k \exp(O_j \cdot G_j / \tau)} \quad (3)$$

We also keep a queue Q and fill it with CLIP embeddings of ground images. Here, $|Q| >> k$ and the embeddings from the queue are used as additional negative samples for our contrastive objective. As in MoCo [9], the queue is continuously updated during the training with the most recent batch. Minimizing this objective minimizes the distance between co-located overhead and ground-level images in the CLIP space. It is worth noting that throughout the training process, the CLIP image encoder is frozen. Hence, with our training procedure, we allow the overhead images to move close to images from their respective ground-level scene in the CLIP space. As a byproduct, the overhead images also move closer to the textual descriptions of ground-level images, which we are ultimately interested in mapping.

3.3. Learning Dynamic Concepts of Places

Many ground-level concepts are temporally dependent. Concepts like “crowded street”, or “snowy place” can dramatically vary based on the exact time we query about them. To model such dynamic concepts, we condition Sat2Cap on the timestamps of the ground-level images.

For each sample, we extract the year, month, day, and hour in which the ground image was taken. We also add

the geolocation information to provide a stronger signal to the model. We encode this meta-information using sin-cos encoding and pass it through a shallow, fully connected network, which we call the Dynamic Encoder (h_θ). The output from h_θ is added element-wise to the output from the Sat2Cap encoder before computing the objective.

$$S_i = O_i + E_i \quad (4)$$

where O_i is the output from the Dynamic Encoder. Now the objective function changes to:

$$L_{dynamic} = \frac{1}{k} \sum_{i=0}^k -\log \frac{\exp(S_i \cdot G_i / \tau)}{\sum_{j=0}^k \exp(S_j \cdot G_j / \tau)} \quad (5)$$

where S_i is the sum of outputs from the Image Encoder and Dynamic Encoder

Our complete framework is shown in Figure 2. To prevent overfitting to the meta-information, we implement random dropout of the Dynamic Encoder during training. Our experiments from Section 4.1 show that this addition significantly improves the performance of our model.

4. Experiments and Results

Our model learns a powerful geo-text embedding space that can be used for a variety of applications. We describe 4 experiments to demonstrate the efficacy of our model.

4.1. Cross-View Image Retrieval

In this experiment, we show that our model learns a strong relationship between co-located overhead images

Method				Overhead2Ground (10K)			Ground2Overhead (10K)		
Model	Meta/Training	Dropout	Meta/Inference	R@5↑	R@10↑	Median-R↓	R@5↑	R@10↑	Median-R↓
CLIP	-	-	-	0.007	0.013	1700	0.108	0.019	2857
ours	✗	✗	✗	0.398	0.493	15	0.356	0.450	11
	✓	✗	✗	0.322	0.413	34	0.254	0.343	20
	✓	✗	✓	0.368	0.467	23	0.298	0.398	13
	✓	✓	✗	0.467	0.564	13.5	0.366	0.462	7
	✓	✓	✓	0.493	0.591	12	0.390	0.482	6

Table 1. **Cross-view retrieval performance of Sat2Cap model:** The table shows that CLIP performs poorly for the task of cross-view retrieval. Moreover, we study the performance of our model under various settings. The **Meta/Training** column ablates the impact of adding meta-information in training. The **Dropout** column ablates the impact of randomly dropping out the Dynamic Encoder in training. The **Meta/Inference** column ablates the impact of adding meta-information in inference. Our experiments show that using all three achieves the best performance.



Figure 3. **Top-9 overhead-to-ground image retrieval:** We use the Sat2Cap embeddings of the overhead images and CLIP embeddings of the ground-level images and show the 9 closest ground-level images retrieved for a query overhead image. The retrieval was performed from a gallery of 10,000 samples.

and ground images in the CLIP space. We randomly sample 10,000 image pairs from the test set for this experiment. First, we predict the Sat2Cap embeddings for all overhead images and the corresponding CLIP embeddings for all the ground-level images in the test set. We then compute top-k (or R@k) and median rank metrics between the Sat2Cap overhead embeddings and the CLIP ground embeddings. The top-k metric measures how often the ground truth falls within the top-k closest images for a given query image. Table 1 shows all of the cross-view retrieval results.

As a baseline, we use the retrieval results using CLIP

overhead embeddings and CLIP ground embeddings. The cross-view retrieval for CLIP model is extremely low with R@10 score of 0.013 and a median rank of 1700. These low scores suggest that the overhead image CLIP embeddings do not contain high-frequency information about the ground-level scene. For our model, we first experiment without using the Dynamic Encoder. Just by contrastively training the Sat2Cap image encoder with ground-level CLIP embeddings, we achieve a high R@10 score of 0.493 and a median rank of 15, as seen in Table 1.

All remaining experiments are conducted on models

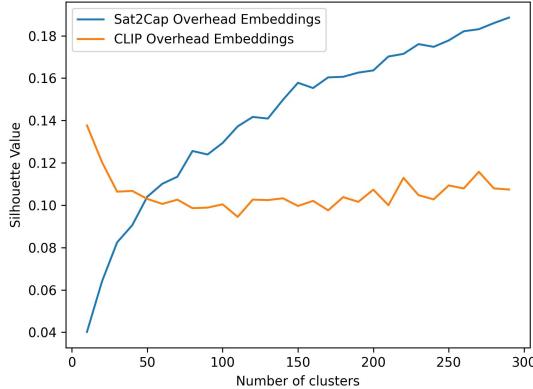


Figure 4. Silhouette value of CLIP embedding vs. Sat2Cap embedding clusters: We use k-means clustering with identical parameters to get the clusters for different values of k . Results show that Sat2Cap embeddings can be well separated into a larger number of clusters than the corresponding CLIP embeddings.

trained using the Dynamic Encoder. Table 1 shows that initially, the retrieval scores drop when using the Dynamic Encoder. This happens because the model starts to overfit on the meta-information, ignoring important cues from the overhead images. We see a further 5.4% drop in R@10 metrics when we use meta-information in training but remove it during inference. To reduce the possibility of overfitting, we randomly drop the Dynamic Encoder during training. Simply adding dropout during training increases the R@10 score by 12.4%. Furthermore, removing meta-information during inference is less severe (2.7%) when using dropout. Hence, our model achieves good cross-view retrieval scores even if meta-information is not provided during inference.

Figure 3 shows the 9 closest images retrieved from a given overhead image. We see that our model is able to retrieve ground-level images by relating concepts rather than direct visual matching. For example, in (d), our model relates farmland with cattle and livestock, which are concepts that semantically match the location but are not visible in the overhead image. Sat2Cap is also capable of encoding temporal information in the embeddings whose results are shown in the supplementary material.

Our results highlight that CLIP space, in itself, is remarkably poor at learning the relationship between a location and the corresponding ground-level scene. The low cross-view retrieval scores imply that we cannot accurately reason about ground-level scene using the CLIP embeddings of co-located overhead images. On the other hand, the high cross-view retrieval scores of Sat2Cap suggest that the Sat2Cap embedding of a location approximates the CLIP embeddings of the images at the ground-level. This ultimately re-

sults in an emergent alignment between location and textual descriptions of the ground level.

4.2. Embedding Space Comparison: CLIP vs. Sat2Cap

In Section 4.1, we used retrieval scores to draw conclusions about the Sat2Cap and CLIP space. Here, we highlight the differences in Sat2Cap and CLIP space more explicitly. In our work, the term “fine-grained” refers to the concepts that are easily visible at the ground level but are hard to infer from overhead images. For example, an overhead image over a location might tell us that it is a city, but ground-level images of that location capture detailed information such as how crowded the place is, if there are many restaurants around that area, does the location regularly hosts street festivals, and so on. Our method learns a model that takes an overhead view of a location and predicts a representation of ground-level images at that location in CLIP space.

We hypothesize that, for given overhead images, Sat2Cap learns diverse concepts while CLIP collapses to a few coarse concepts. To examine this, we run k-means clustering for both CLIP and Sat2Cap overhead image embeddings. We then compute the silhouette value for the resulting clusters, which tells us the quality of these clusters. This value ranges from -1 to +1, and it tells us how well the embeddings are separated/clustered for a given value of k ; higher values indicate better clusters. Figure 4 shows the plot of silhouette value for different values of k . CLIP embeddings are well clustered for small values of k but gradually worsen as we increase the number of clusters. This suggests that CLIP embeddings collapse to only a few concepts and are not further separable to reason about more diverse concepts. On the contrary, Sat2Cap embeddings perform poorly for low values of k but outperform CLIP as the value of k increases. The result indicates that the overhead embeddings from the Sat2Cap model are more diverse, allowing them to learn a variety of fine-grained concepts. The Sat2Cap embeddings can be well separated into a large number of clusters (concepts) and do not suffer the same issue of collapse as CLIP, making them suitable for the task of fine-grained mapping.

4.3. Zero-Shot Map of Fine-grained Concepts

We use Sat2Cap embeddings to create zero-shot maps using fine-grained prompts. To do this, we choose a region and download high-resolution satellite images ($0.6m/px$) over the region. We precompute the Sat2Cap embeddings for all the images. At inference time, for a given text query, we predict the CLIP embedding and compute its similarity with the overhead image embeddings. The process of computing the similarities only takes about 4 seconds.

Figure 5 shows the zero-shot map of Amsterdam for two prompts: “Heavy trucks transporting goods”, and “Peo-

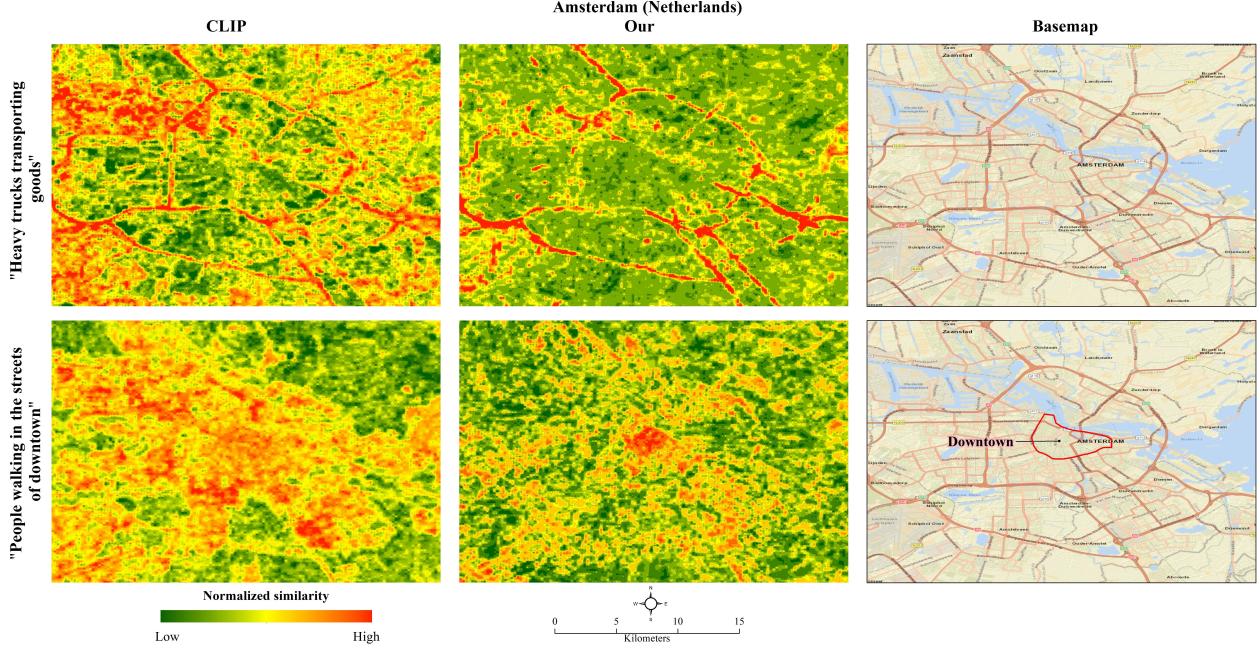


Figure 5. Fine-grained maps CLIP vs Sat2Cap: We create zero-shot maps on a city level using CLIP and Sat2Cap for two prompts: “Heavy trucks transporting goods” and “People walking in the streets of downtown.” Compared to CLIP, Sat2Cap activations are more localized to the appropriate regions for a given prompt. Sat2Cap is better at distinguishing between fine-grained concepts like **highway** and **downtown street**.

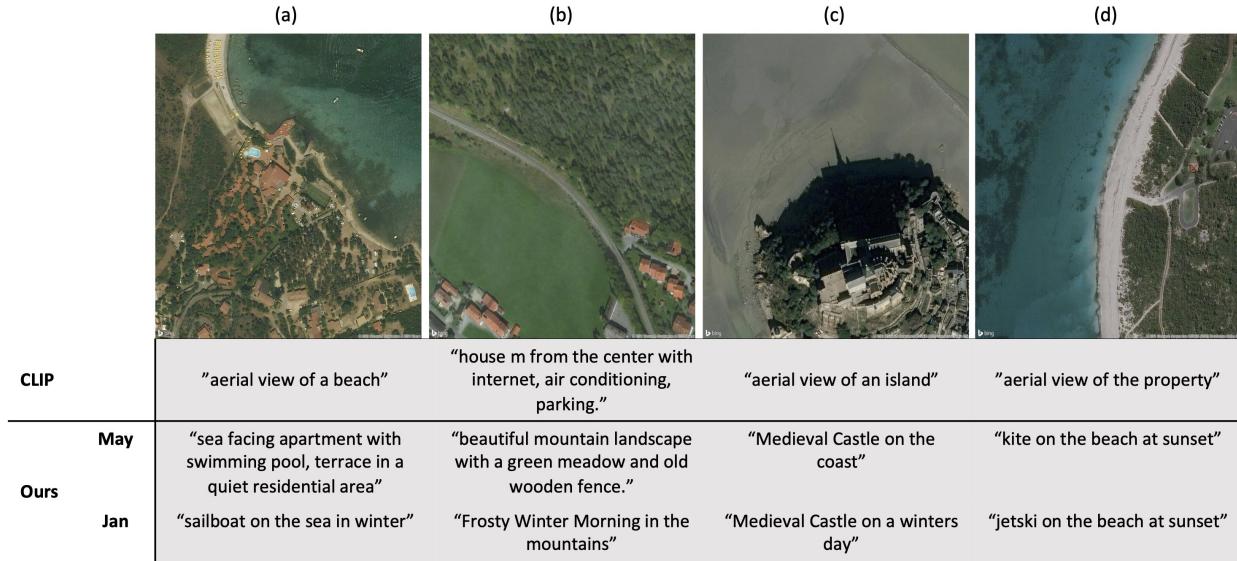


Figure 6. Captions generated by the CLIPCAP model [17] using CLIP embeddings vs. Dynamic Sat2Cap embeddings. (Row-1) shows the results from CLIP embeddings, which produce many generic descriptions. (Row-2 and 3) show the results from our Sat2Cap embeddings for the month of May and January, respectively. The captions generated using Sat2Cap embeddings are more fine-grained and dynamic. While (d) does not add any winter properties for the January query, this behavior is expected as the image is over Australia, where January falls in the middle of summer.

	Meta Training	Dropout	Meta Inference	Cosine Similarity		BERT Score
				MPNET_Base_v2	E5_Small	
CLIP	-	-	-	0.1560	0.7519	0.7135
ours	\times	\times	\times	0.3334	0.7969	0.7692
	\checkmark	\times	\times	0.2727	0.7806	0.7476
	\checkmark	\times	\checkmark	0.2888	0.7857	0.7572
	\checkmark	\checkmark	\times	0.2755	0.7831	0.7533
	\checkmark	\checkmark	\checkmark	0.2755	0.7833	0.7537

Table 2. **Caption generation alignment:** The table shows the alignment between captions generated from co-located overhead and ground-level images. Captions generated using Sat2Cap have better alignment with the descriptions of co-located ground-level scenery.

ple walking in the streets of downtown.” Column 1 shows the maps generated using CLIP embeddings. For the first prompt, the Sat2Cap activations are localized to the highway network of the city, whereas CLIP has many false positives over general built-up areas. This shows that Sat2Cap can pick the subtle nuances of fine-grained text that CLIP fails to do. Similarly, for the second prompt, CLIP shows high activations throughout the city, while Sat2Cap’s activations are more localized to the downtown area.

We also create country-level zero-shot maps for two countries: the Netherlands and England. We query the images with 3 prompts, each related to a distinct land cover class. Figure 1 shows the comparison of our zero-shot maps with the land cover maps obtained from ESRI. We see that the zero-shot maps highly correlate with the ESRI land cover maps of the respective countries.

4.4. Fine-grained and Dynamic Caption Generation

To generate captions from our embeddings, we use the CLIPCap [17] model. CLIPCap allows us to generate captions by learning a mapping from the CLIP space to the text space. Figure 6 shows qualitative examples of captions generated by passing CLIP embeddings vs Sat2Cap embeddings as input for a given overhead image. We observe that when using CLIP embeddings of the overhead images as input, the captions generated by CLIPCap mostly describe generic concepts of a location like a beach, island, property, etc. In contrast, our Sat2Cap embeddings produce more fine-grained and aesthetically pleasing captions. Furthermore, we use the dynamic encoder to generate captions in two different months, for the same location. Figure 6 shows that Sat2Cap accurately models the seasonal variations and aligns the captions towards respective temporal inputs such as capturing winter concepts for January. However, in Figure (d), we see that the model does not add any winter-specific information for the January input. This is expected behavior since the image is from Australia, where the month of January falls in the middle of summer. The example further demonstrates that Sat2Cap learns a joint model of time and location. Fine-grained captions for a location should be

well-aligned with the captions describing the ground-level scene. We quantitatively evaluate the quality of the captions by looking at their alignment with the text descriptions for respective ground-level images. First, we use the CLIPCap model to generate captions for ground-level images and use that as our ground truth. We then use the CLIP embeddings and Sat2Cap embeddings of the overhead images to generate the respective captions for that location. Table 2 shows the similarity metrics between the ground-truth text and generated text. To compute the cosine similarity, we use two different sentence transformers from HuggingFace, *MPNET_Base_v2* [25] and *E5_Small* [28]. The captions generated using Sat2Cap embeddings demonstrate significantly better alignment with the ground-image captions than their CLIP counterparts. The results show that Sat2Cap captions better describe the ground-level characteristics of a given location. Table 2 also shows that the model trained without metadata has the highest cosine similarity and BERTScore with ground-level descriptions. We suspect this happens because of the uncertainty introduced by the use of a pretrained caption generator, i.e., small noise in the metadata introduces big deviations in the generated text.

5. Conclusion

We presented a weakly supervised framework for learning a semantically rich embedding space between geolocation and fine-grained text. For this task, we introduced a new large-scale cross-view dataset with 6.1M samples. Our approach does not depend on text supervision, and learns textual representations of a location using only ground-level images and the CLIP embedding space. In addition to higher retrieval performance, we demonstrated that our framework can efficiently generate high-quality zero-shot maps from fine-grained text prompts. This ability to create maps from fine-grained text prompts offer greater flexibility when compared to traditional methods. Finally, we demonstrated that Sat2Cap embeddings can be used to generate dynamic captions that align with the ground-level scene.

Sat2Cap: Mapping Fine-Grained Textual Descriptions from Satellite Images

Supplementary Material

6. Implementations Details

We use a ViT-32B as the CLIP image encoder. This image encoder is kept frozen throughout the training. We use a ViT-32B architecture as the backbone for our Sat2Cap model. The Sat2Cap backbone is initialized using CLIP weights. Following [22] we use an AdamW optimizer [16] with a learning rate of $1e-05$ with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. We use a learnable temperature parameter initialized at $\tau = 0.07$. We use Cosine Annealing with Warm Restarts [15] as the learning rate scheduler. We augment the overhead images using RandomResizedCrop and RandAugment [5]. The overhead images are normalized using the mean and standard deviation of the training set. The training was carried out using Nvidia A100 40GB GPU. Since a larger number of negative samples is beneficial for contrastive learning, we simulate a large batch size using a memory bank approach. We initialize a queue of size 9600 and fill it with precomputed ground-level image CLIP embeddings which are used as negative samples for computing the loss.

7. Text to Overhead Image Retrieval

Our framework uses ground-level images as pseudo-labels to learn the textual concepts of geolocation. Although Sat2Cap does not require any text labels during training, it effectively learns an embedding space where geolocation and their fine-grained descriptions are well aligned. To show this, we randomly selected 1000 overhead images from our training set, and compute their Sat2Cap embeddings. For a given text query, we generate the CLIP[22] text embedding and compute its similarity with all images in the test set. Figure 7 shows examples of 4 closest overhead images retrieved for a given query.

We experiment with small perturbations of prompts to analyze how our retrieval results change with minute variations of query. We see in Figure 7, the prompt “people driving cars” retrieves city or residential areas. However, replacing the phrase “driving cars” with “riding horses” retrieves locations with farmland. Similarly, the prompt “person on a long hike” exclusively retrieves mountainous regions, while the prompt “person on a long run” retrieves images that look like trails nearby residential areas. Hence, Sat2Cap embeddings demonstrate a good understanding of fine-grained variations of textual concepts.

8. Learning Dynamic Concepts

The dynamic encoder allows our model to learn temporally varying concepts over a location. Here we show more

qualitative results showcasing the dynamic properties of our model. Figure 8 shows the retrieval results at two different time settings (11:00 p.m. and 08:00 a.m.). Figure 9 shows the captions generated for the same location over long (order of months) and short (order of hours) term variations. Finally, Figure 10 shows how our model can generate maps that adapt to temporal variations for a given prompt.

9. Country Level Map of US

We also create a zero-shot map of the US. However, due to the massively large area, we had to downsample our overhead image acquisition by 10x. Although we are predicting at a much coarser resolution, we still achieve a reasonable zero-shot map as seen in Figure 11.

10. Geolocalizing Textual Queries

Our model can be used to localize textual queries at a finer resolution. For this experiment, we draw a $24km^2$ bounding box over a region. We compute the Sat2Cap similarity for all the overhead images in that box with a given text query. We, then, normalize the similarities between 0 and 1 and clip the values below 0.5. Figure 12 shows the results of this experiment. The red spot indicates the location with the highest activations for the given query. For each query, the left figure shows the total area of inference, and the right figure shows a fine-grained image at the location with the highest activation, obtained from Google. We see that our model makes reasonable localization for the given queries. For example: in (a) our model activates over a soccer stadium. Figure (b) shows that when we compose the concept of people with animals, our model shows very high activation in farm-like areas which is where these two concepts would most likely co-occur. These results show that our model can reasonably localize the most plausible point within a given area, where one might observe a given query. This property can be beneficial in solving visual search problems in the geospatial domain.

11. Dataset

We introduced a cross-view dataset with overhead images and co-located ground-level images taken from the YFCC100M [27] dataset. Figure 13 shows a few samples from our dataset. The ground-level images provide us with detailed fine-grained concepts of a location that cannot be directly inferred when looking at the overhead imagery.

"Ongoing soccer match"



"Ongoing basketball match"



"People driving cars"



"People riding horses"



"Person on a long hike"



"Person on a long run"

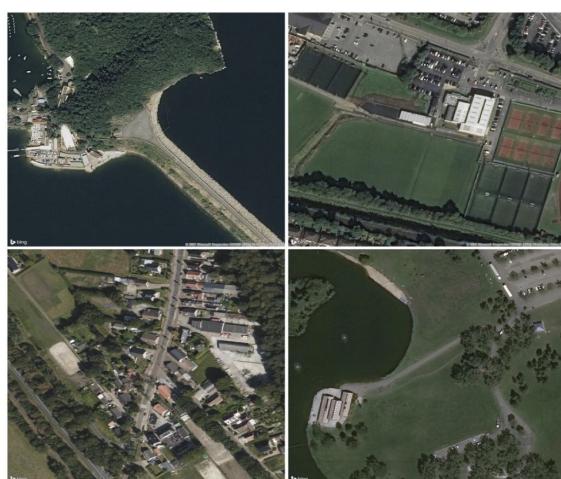


Figure 7. **Top-4 text-to-overhead retrieval:** We retrieve the top-4 closest overhead image from a given text prompt. Our results show that Sat2Cap embeddings can accurately relate geolocations with fine-grained textual prompts.

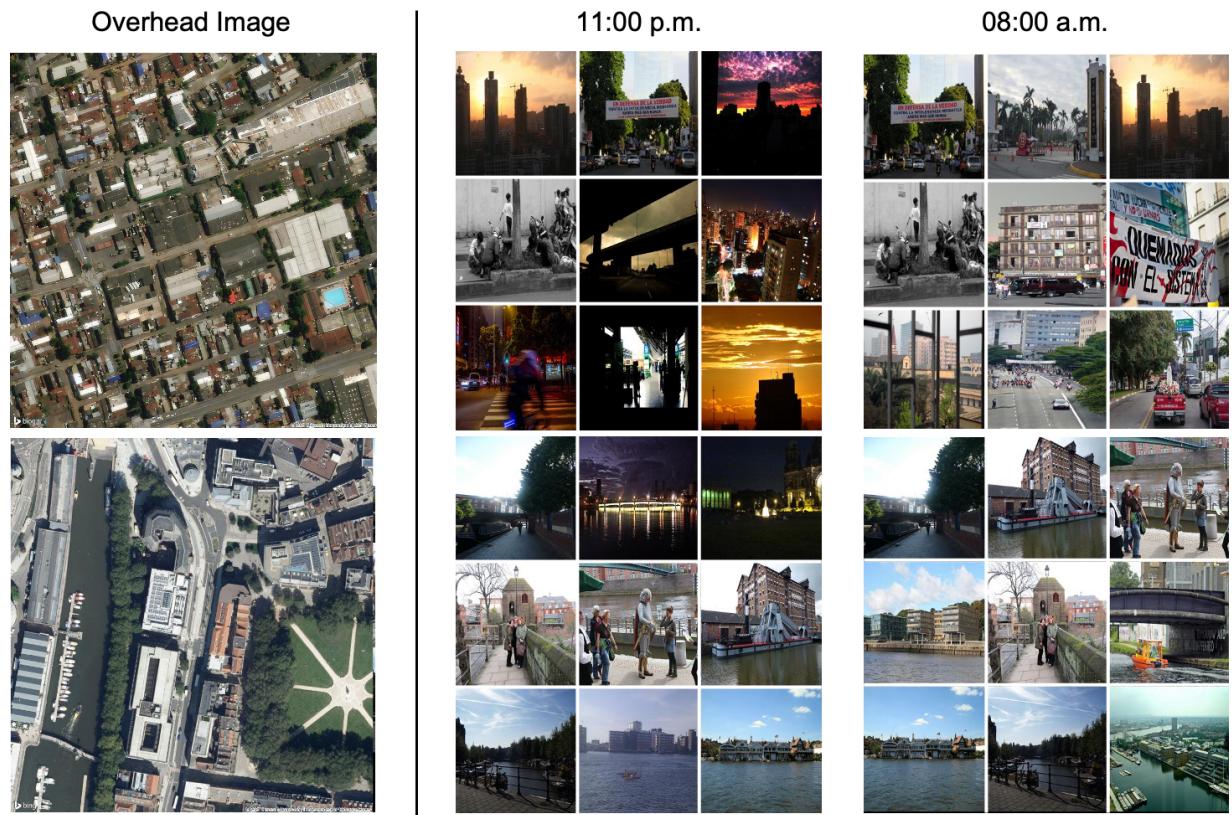


Figure 8. **Top-9 overhead-to-ground image retrieval with temporal manipulation:** We show the 9 closest ground-level images for a query overhead image at two different time settings (11:00 p.m. and 08:00 a.m.).

Model	Date/Time	Description
CLIP	-	<i>"aerial view of a beach"</i>
Ours	May 20 08:00 am	<i>"property image sea facing apartment with swimming pool, terrace in a quiet residential area."</i>
	Dec 20 10:00 am	<i>"Sailboat on the sea in winter"</i>
	Dec 20 05:00 pm	<i>"Person on the beach at night"</i>
	Dec 20 11:00 pm	<i>"Nightime on the beach"</i>

Figure 9. **Dynamic Caption Generation:** Our Sat2Cap embeddings dynamically adapt to temporal manipulations, facilitating dynamic caption generation.

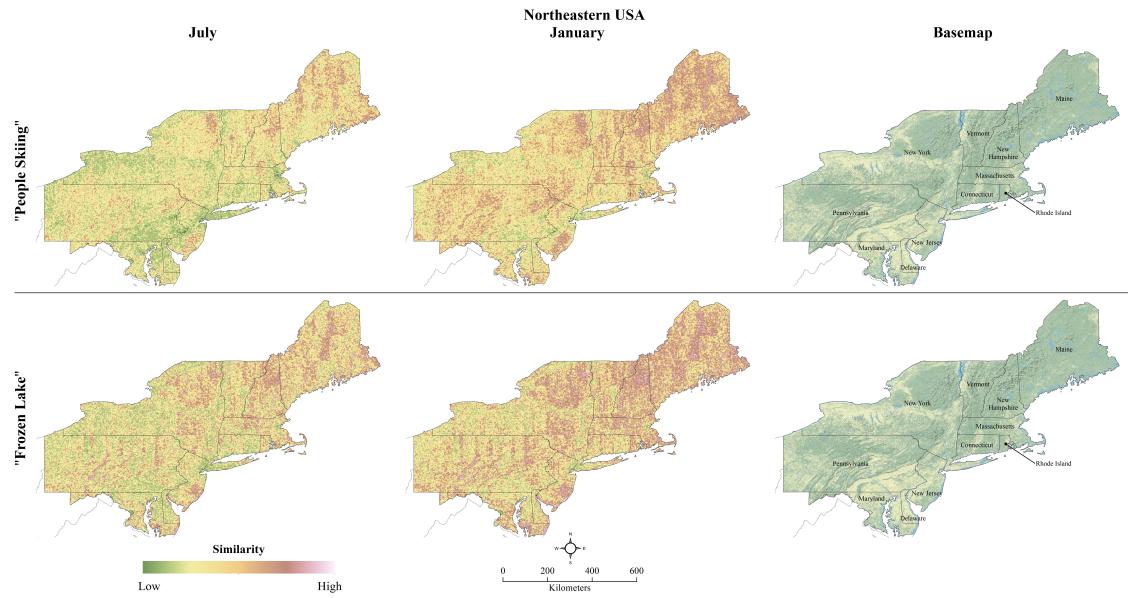


Figure 10. **Dynamic Maps:** We show zero-shot maps of the northeast US at two different temporal settings (July and January)

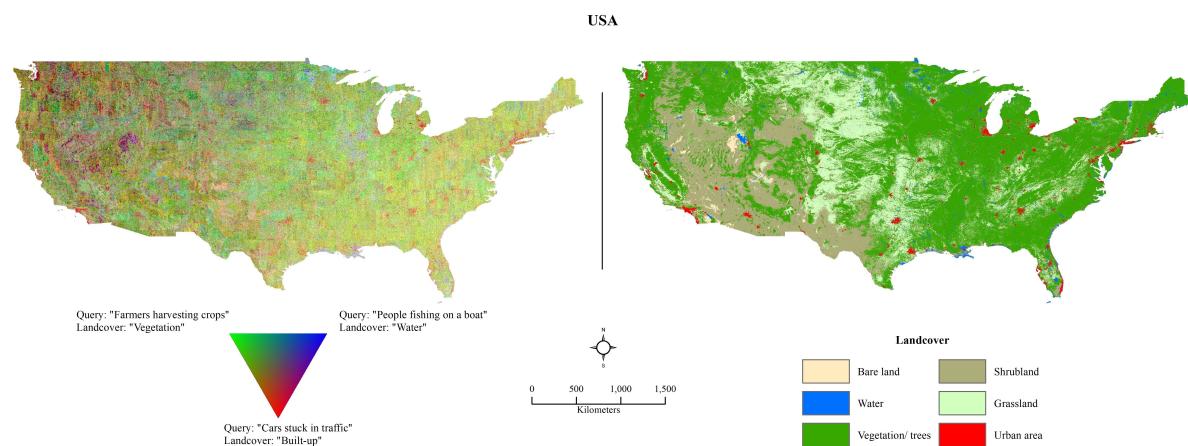


Figure 11. Zero-shot map of the USA

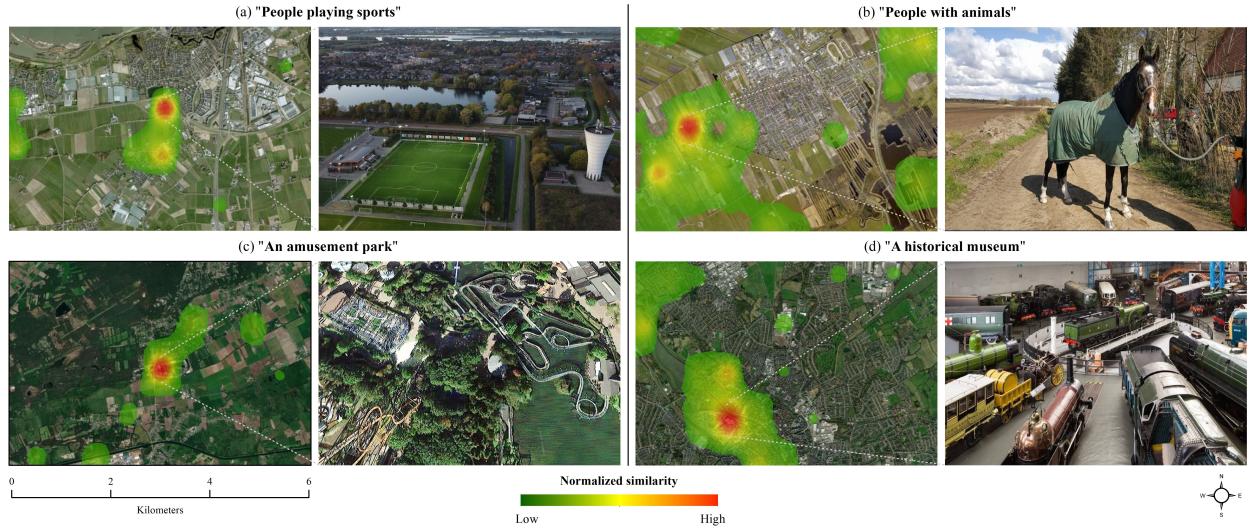


Figure 12. **Localizing textual queries at finer resolution:** For each prompt, the image on left shows the big region which is used for inference. The image on the right shows an image of the ground-level scene at the point with the highest activation, which was taken by entering the location in Google Maps

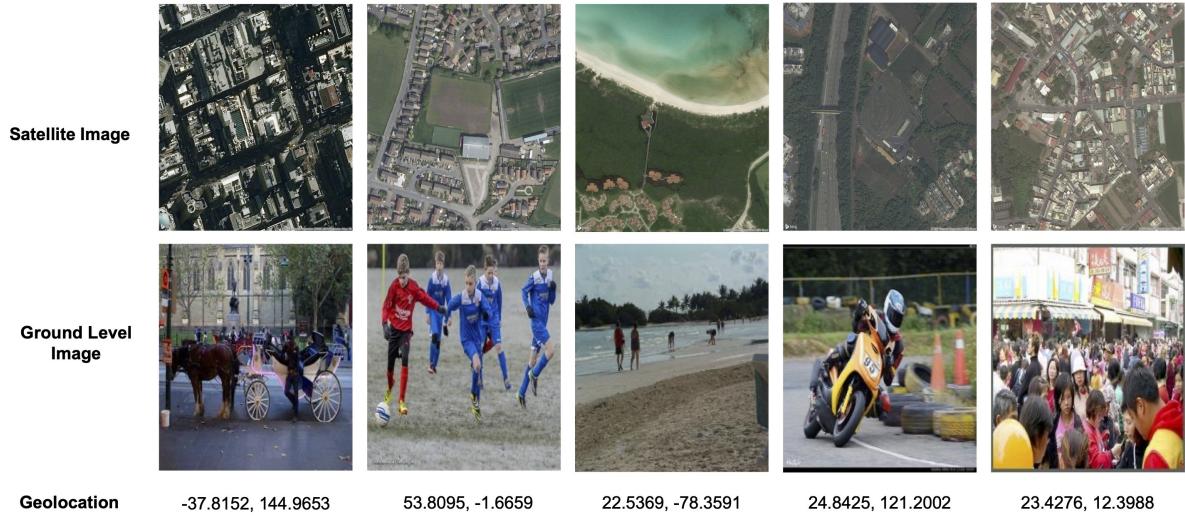


Figure 13. Examples of co-located overhead and ground images in our dataset. The ground-level images describe more detailed concepts of the given locations than their overhead counterparts.

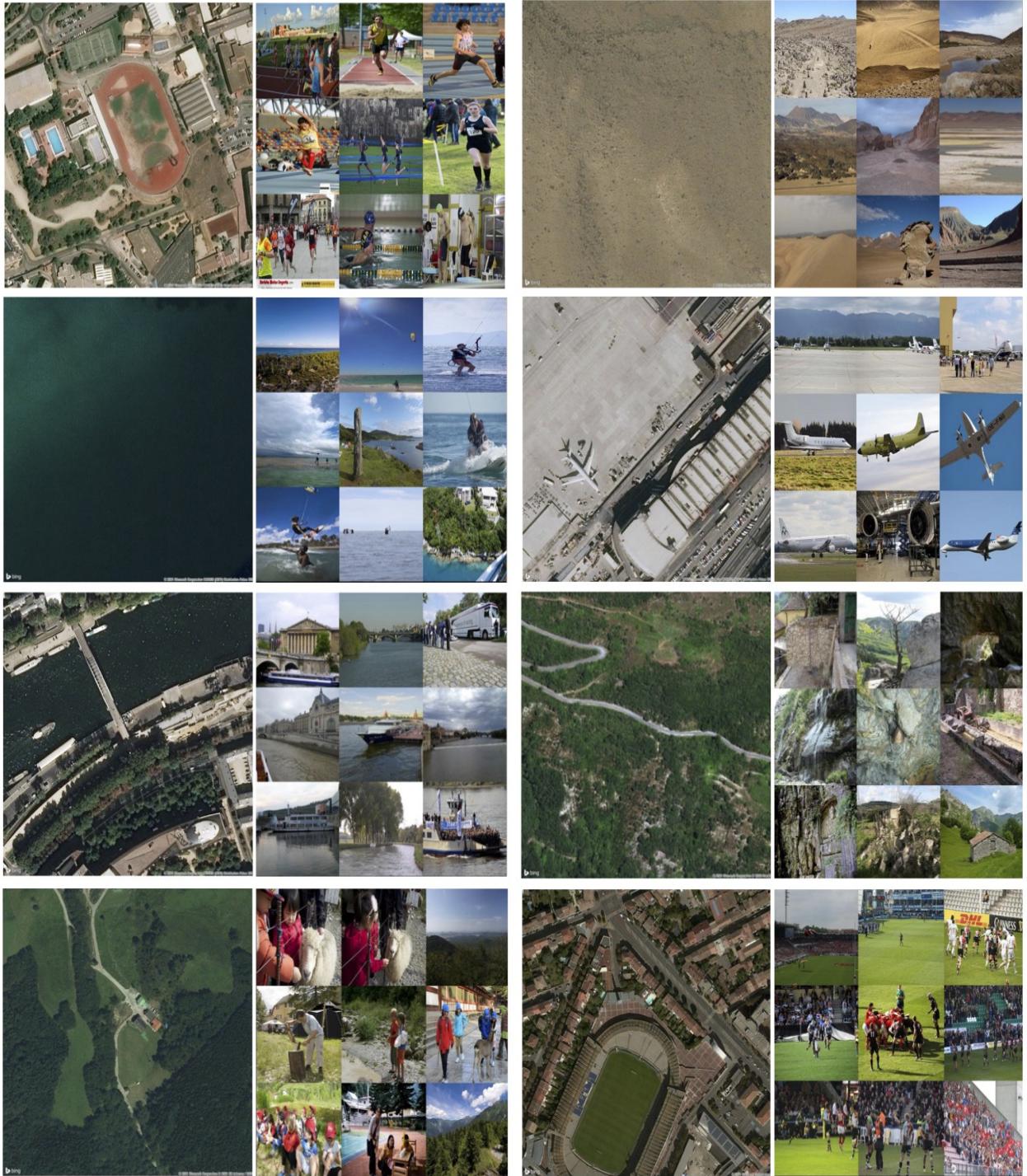


Figure 14. **Top-9 overhead-to-ground retrieval:** Our model can infer fine-grained concepts of ground-level scenes through overhead imagery. Sat2Cap accurately retrieves probable concepts for a given geolocation using an overhead image.

References

- [1] Victor Alhassan, Christopher Henry, Sheela Ramanna, and Christopher Storie. A deep learning framework for land-use/land-cover mapping and analysis using multispectral satellite imagery. *Neural Computing and Applications*, 32: 8529–8544, 2020. 3
- [2] Thorsten Behrens, Karsten Schmidt, Robert A MacMillan, and Raphael A Viscarra Rossel. Multi-scale digital soil mapping with deep learning. *Scientific reports*, 8(1):15244, 2018. 3
- [3] Archith J Bency, Swati Rallapalli, Raghu K Ganti, Mudhakar Srivatsa, and BS Manjunath. Beyond spatial auto-regressive models: Predicting housing prices with satellite imagery. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 320–329. IEEE, 2017. 1
- [4] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dermoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with clip reward. *arXiv preprint arXiv:2205.13115*, 2022. 3
- [5] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 1
- [6] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173, 2021. 3
- [7] Bakhtiar Feizizadeh, Davoud Omarzadeh, Mohammad Kazemi Garajeh, Tobia Lakes, and Thomas Blaschke. Machine learning data-driven approaches for land use/cover mapping and trend analysis using google earth engine. *Journal of Environmental Planning and Management*, 66(3): 665–697, 2023. 3
- [8] Connor Greenwell, Scott Workman, and Nathan Jacobs. What goes where: Predicting object distributions from above. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 4375–4378. IEEE, 2018. 3
- [9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 4
- [10] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 3
- [11] Krishna Karra, Caitlin Kontgis, Zoe Statman-Weil, Joseph C. Mazzariello, Mark Mathis, and Steven P. Brumby. Global land use / land cover with sentinel 2 and deep learning. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 4704–4707, 2021. 3
- [12] Argyro Kavvada, Graciela Metternicht, Flora Kerblat, Naledzani Mudau, Marie Haldorson, Sharthi Laldaparsad, Lawrence Friedl, Alex Held, and Emilio Chuvieco. Towards delivering on the sustainable development goals using earth observations, 2020. 3
- [13] Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on graphics (TOG)*, 33(4):1–11, 2014. 1
- [14] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1, 3
- [15] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 1
- [16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1
- [17] Ron Mokady, Amir Hertz, and Amit H Bermano. Clip-cap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 3, 7, 8
- [18] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3
- [19] Masanori Onishi and Takeshi Ise. Explainable identification and mapping of trees using uav rgb image and deep learning. *Scientific reports*, 11(1):903, 2021. 3
- [20] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4
- [21] C Persello, JD Wegner, R Hänsch, D Tuia, P Ghamisi, M Koeva, and G Camps-Valls. Deep learning and earth observation to support the sustainable development goals: current approaches, open challenges, and future opportunities. *ieee geosci remote sens mag* 10 (2): 172–200, 2022. 1, 3
- [22] 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*, pages 8748–8763. PMLR, 2021. 1, 3

- [23] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3
- [24] Tawfiq Salem, Scott Workman, and Nathan Jacobs. Learning a dynamic map of visual appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12435–12444, 2020. 1, 3
- [25] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020. 8
- [26] Artem Streltsov, Jordan M Malof, Bohao Huang, and Kyle Bradbury. Estimating residential building energy consumption using overhead imagery. *Applied Energy*, 280:116018, 2020. 1
- [27] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 2, 3, 1
- [28] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Dixin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022. 8
- [29] Zihao Wang, Wei Liu, Qian He, Xinglong Wu, and Zili Yi. Clip-gen: Language-free training of a text-to-image generator with clip. *arXiv preprint arXiv:2203.00386*, 2022. 3
- [30] Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. *arXiv preprint arXiv:2312.12856*, 2023. 3
- [31] Scott Workman and Nathan Jacobs. Dynamic traffic modeling from overhead imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12315–12324, 2020. 3
- [32] Scott Workman, Richard Souvenir, and Nathan Jacobs. Understanding and mapping natural beauty. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5589–5598, 2017. 3
- [33] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liquan Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 3
- [34] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19163–19173, 2022.
- [35] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 1, 3
- [36] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 3
- [37] Zhenghang Yuan, Zhitong Xiong, Lichao Mou, and Xiao Xi-ang Zhu. Chatearthnet: A global-scale, high-quality image-text dataset for remote sensing. *arXiv preprint arXiv:2402.11325*, 2024. 3
- [38] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022. 3
- [39] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 1
- [40] Zefang Zong, Jie Feng, Kechun Liu, Hongzhi Shi, and Yong Li. Deepdpm: Dynamic population mapping via deep neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1294–1301, 2019. 3