

# RANGE: Retrieval Augmented Neural Fields for Multi-Resolution Geo-Embeddings

Aayush Dhakal<sup>1</sup> Srikumar Sastry<sup>1</sup> Subash Khanal<sup>1</sup> Adeel Ahmad<sup>1,2</sup> Eric Xing<sup>1</sup> Nathan Jacobs<sup>1</sup>

Washington University in St. Louis<sup>1</sup> Taylor Geospatial Institute<sup>2</sup>

## Abstract

The choice of representation for geographic location significantly impacts the accuracy of models for a broad range of geospatial tasks, including fine-grained species classification, population density estimation, and biome classification. Recent works like SatCLIP and GeoCLIP learn such representations by contrastively aligning geolocation with co-located images. While these methods work exceptionally well, in this paper, we posit that the current training strategies fail to fully capture the important visual features. We provide an information theoretic perspective on why the resulting embeddings from these methods discard crucial visual information that is important for many downstream tasks. To solve this problem, we propose a novel retrieval-augmented strategy called RANGE. We build our method on the intuition that the visual features of a location can be estimated by combining the visual features from multiple similar-looking locations. We evaluate our method across a wide variety of tasks. Our results show that RANGE outperforms the existing state-of-the-art models with significant margins in most tasks. We show gains of up to 13.1% on classification tasks and 0.145  $R^2$  on regression tasks. All our code and models will be made available at: <https://github.com/mvrl/RANGE>.

## 1. Introduction

Several machine learning tasks require the use of geolocation as an input feature. Multiple works have shown the benefits of using geolocation to solve vital ecological, geographical, and geological tasks. SINR [5] used location as input to predict species distribution, Lang *et al.* [20] used location with satellite image to learn a global canopy height model, and SatCLIP [18] used location to solve several geospatial tasks such as biome classification, population density estimation, housing price prediction, etc. Since location is a ubiquitous input variable in all geospatial settings, finding good representations for it is fundamental when solving geospatial tasks.

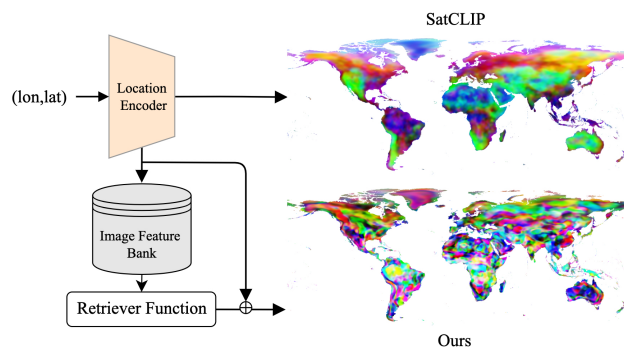


Figure 1. Adding explicit visual features allows us to generate more high-resolution location embeddings. We use a retrieval strategy that allows us to generate multi-resolution retrieval-augmented neural field of geo-embeddings (RANGE).

Previous work [33] has shown that using location directly (represented as two floating point numbers of latitude and longitude) as an input to a machine learning model yields poor performance. Therefore, several works have attempted to find better ways to represent location data. Some of these approaches [10, 25] use a non-parametric method to encode location. However, recent works [18, 27, 39] have shown that it is possible to learn a general-purpose location representation by learning a parametric model (often a neural network) in a self-supervised setting. There are several strategies that can be used to train the parametric model. A common approach used by state-of-the-art models [18, 27, 39] involves learning visual cues associated with a given geolocation. Two recent state-of-the-art models, GeoCLIP [39] and SatCLIP [18], use contrastive training to align geo-embeddings with co-located images; the former aligns them with ground-level images while the latter aligns them with satellite images. Images of a location capture dense information about the area, such as the landcover, landuse, etc. Therefore, learning location representations that capture this dense information is beneficial for solving geospatial tasks. Prior work [18] has shown that rep-

representations from such models are extremely powerful and are capable of solving diverse geospatial tasks both on their own and in combination with images. In this work, we argue that the current approaches to learning geo-embeddings using images, while powerful, still fail to preserve all the relevant visual information. The location representations from these models only capture low-resolution shared information between location and image and ignore crucial high-resolution information that is unique to the image. In section 2, we define this problem more formally from the perspective of multi-view redundancy.

To address this issue, we propose a retrieval-augmented strategy that approximates the unique visual information present in a location. There are three key ideas to our solution. First, it is possible to approximate the visual features of a given location by strategically combining the visual features of other similar-looking locations. Second, contrastively trained models like SatCLIP [18] and GeoCLIP [39], while lacking in preserving modality-specific information, are excellent in learning modality-alignment [13]. Therefore, it is possible to retrieve semantically aligned images, with alignment scores, for a given location using these models. Third, we note that the variance of semantic information visible in satellite images across the Earth is relatively low (as opposed to the wider diversity in consumer photographs). Hence, it is possible to capture a large percentage of unique aerial semantics with a limited number of satellite images. By combining these three key ideas, we propose our retrieval-augmented framework, RANGE (Retrieval-Augmented Neural Fields for Multi-Resolution Geo-Embeddings). Our retriever function uses both semantic and spatial alignment to approximate the visual features of a given location using an auxiliary database of image features. Our function is robust to the database size and works well even in limited sample settings. The hyperparameter in the retriever function can be adjusted to generate geo-embeddings at different frequencies. In summary, these are the main contributions of our paper:

- We show why existing image-location contrastive methods result in sub-optimal representations from the perspective of multi-view redundancy.
- We propose a novel retrieval-based strategy, RANGE, to generate geo-embeddings that preserve both shared and unique modality information.
- We propose a retriever function that estimates the visual embedding of a location using both semantic and spatial alignment. Our retriever function is robust to different database sizes and can generate geo-embeddings at multiple frequencies by enforcing spatial smoothness.

## 2. Aligning Location and Image is a Multi-View Non-Redundancy Problem

Let  $(L, H, T)$  represent location modality, image modality, and geospatial task, respectively. Models like SatCLIP and GeoCLIP align co-located images and locations by minimizing the infoNCE objective. Solving the multimodal infoNCE objective maximizes the lower bound on the mutual information, preserving the shared information within the learned representations [29]. As such, the final learned representations of these models contain only information that is shared across both location and image. This property is useful under the multi-view redundancy assumption which states that the information shared between the two modalities is exactly relevant to solve the underlying task. In other words, the multi-view redundancy setting assumes that any additional unique information from the image or location would not be useful in solving the task  $T$  [37]. However, this assumption does not hold in the location-image learning setting as there is high unique information contained in the image modality that is useful for solving downstream tasks. Formally, we argue that location-image learning falls under the multi-view non-redundancy setting: there is *unique* information in the image modality that is not shared with the location modality, and that information is relevant for solving our geospatial task  $T$ . In this case, contrastive training discards valuable task-relevant information from the image, leading to poor downstream performance [23].

Empirically, there is abundant unique information present in the image features that are relevant for solving downstream tasks. The location representations from contrastively trained models like SatCLIP and GeoCLIP only capture the low-resolution shared information, ignoring all the valuable task-relevant information that is unique to the image. To demonstrate this property concretely, we quantitatively show that there are tasks that can be solved more accurately by adding the image representation to the shared image-location representation from SatCLIP [18]. First, we choose 4 tasks: Biome, Ecoregions, Elevation, and Population Density estimation. We download a corresponding Sentinel-2 image for every data point. For each task, we train one model using the SatCLIP features and another model using the SatCLIP features combined with the SatMAE [6] features of the satellite image [6]. The results in Table 1 show that adding image features improves the performance across all tasks. The increase in performance demonstrates that the images contain unique information relevant to solving the task that is lost during contrastive training. Hence, “location and image” does not follow the assumption of multi-view redundancy.

Although adding corresponding satellite image information to a location embedding improves performance, this is not feasible in practice. For many geospatial tasks, we

	Biome	Ecoregion	Elevation	Population
SatMAE	58.8	28.4	0.388	0.600
SatCLIP	68.9	69.3	0.666	0.684
loc $\oplus$ img	74.9	73.5	0.749	0.765
gain (%)	+8.71	+6.06	+12.46	+11.84

Table 1. We show that using the image-location shared information from SatCLIP embeddings is sub-optimal to solve some geospatial problems. Adding image features to the embeddings provides useful visual information that improves the accuracy of the task. We show the accuracy for the classification tasks and  $R^2$  value for the regression tasks. The results indicate that there are valuable visual features that are not captured by the SatCLIP embeddings.

wish to make predictions across the globe with millions of points. In such cases, using the “image+location” framework would require us to retrieve/store/process millions of images. We propose an efficient method for approximating the visual information using a compact database. Our results suggest that this approximation can outperform the use of true visual features in some cases.

### 3. Related Works

#### 3.1. Learning Representations for Geolocation

Learning effective representations for geolocation is valuable for diverse downstream geospatial tasks, such as geolocalization [39], canopy height prediction [20], fine-grained species classification [25, 34], species distribution modeling [5, 34], geographical question answering [26], soundscape mapping [16], etc. Geolocation in these tasks can be represented using either non-parametric [10, 33] or parametric [18, 27, 39] methods. Rußwurm *et al.* [33] represents geolocation using spherical harmonic basis functions, and Aodha *et al.* [25] uses a sine and cosine encoding of geolocation that removes discontinuity at dateline. In recent years, learned representations for geolocation, typically parameterized by neural networks, have been popularized. Some of these methods learn geolocation representations by using self-supervised pre-training strategies that distill visual information from either co-located overhead [18, 27] or ground-level images [27, 39]. The location encoders trained by these methods have demonstrated utility across various downstream geospatial tasks. We use the representations from these location encoders as baselines for our model.

#### 3.2. Multimodal Contrastive Learning

Contrastive learning is an efficient and scalable pre-training strategy for learning a shared embedding space across multiple modalities [7, 9, 11, 15, 16, 22, 31, 41–43]. As demonstrated by CLIP [31] and more recent works [22, 41, 42] in

vision-language learning, the shared embedding space can be utilized for zero-shot classification and leveraged across a variety of downstream tasks. Contrastively trained multimodal embedding spaces between geolocation and satellite imagery [18, 27] or ground-level imagery [39], have also demonstrated impressive transferability of both location and visual representations to various geospatial tasks.

Despite its success, recent works highlight the limitations of multimodal contrastive learning. One such study [23] shows that contrastively trained representations capture the shared information between the two modalities and thus only work best under certain assumptions. Similarly, another work [13] demonstrates that the final layers of encoders in the contrastive learning framework preserve only the information necessary for solving the alignment objective, discarding other modality-specific details. We argue that some of the underlying assumptions of multimodal contrastive training do not hold true in the image-location setting. Therefore, the resulting location representations from these methods are sub-optimal for downstream tasks. To address this issue, we propose a method to approximate the image-specific information for a given location.

#### 3.3. Retrieval Augmented Methods

There has been a recent surge in retrieval-augmented generation (RAG) methods, primarily focused on generative tasks [1, 2, 4, 17, 19, 21, 35, 36, 40]. These methods generally consist of three main components: a retriever, an external database, and a generator. For a given query, the retriever selects a set of documents from the database and provides them, along with the original query, to the generator to produce the desired output. RAG has been successfully employed to improve text generation for large language models (LLMs) [2, 21, 36], text-to-image generation [1, 4, 17], text-to-3D generation [35], audio captioning [19], and more. Within the RAG framework, one could also replace the generator with another task-specific component, such as a classifier [24], to enhance performance by leveraging the rich information from the retrieved documents. Inspired by these retrieval-augmented methods, we design our framework, RANGE, which retrieves rich visual information from our database for a given geolocation query and utilizes it for various downstream geospatial tasks.

### 4. Method

#### 4.1. Problem Setup

We consider a dataset with paired geolocations and co-located satellite images  $\{g_i, s_i\}$ . The goal is to learn a meaningful representation for  $g_i$  by aligning it with  $s_i$ , which can be done by minimizing the clip objective [18, 31, 39]. We have a trainable geolocation encoder  $E$ , a frozen pretrained image encoder  $I$ , and a projection layer

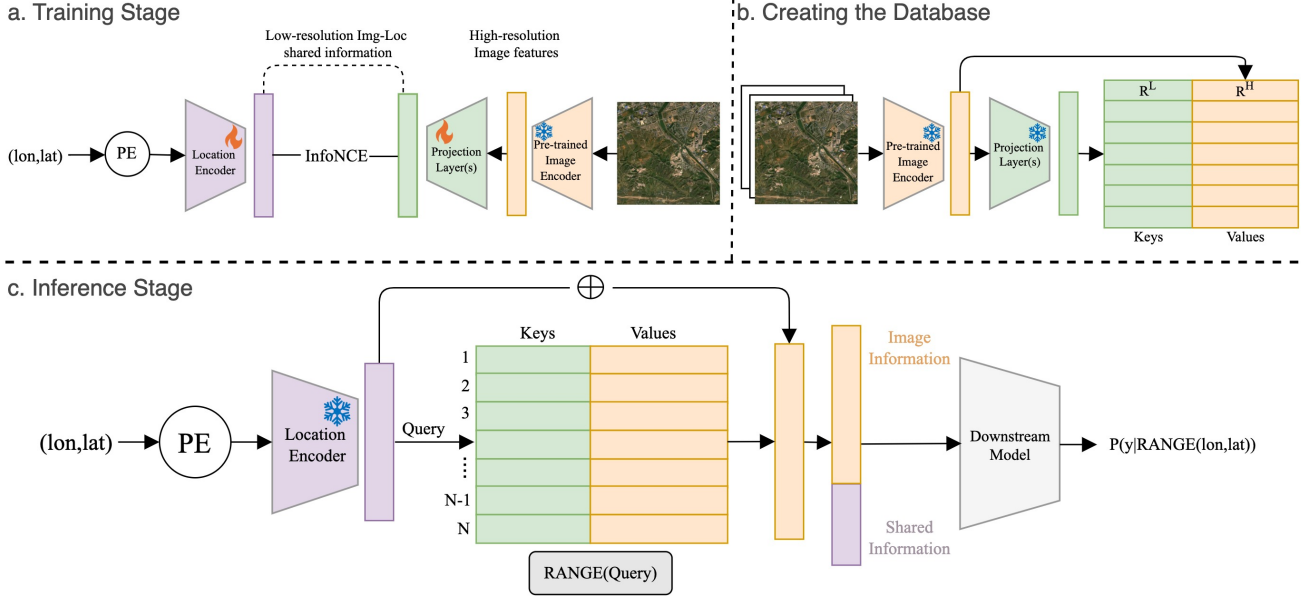


Figure 2. Framework of RANGE. (a) In the training stage, a shared embedding space is learned between locations and images. (b) We create a database of low-resolution and high-resolution image embeddings using the trained projection layer and a powerful pretrained image model, respectively. (c) During inference, we use a location as the query, low-resolution image embeddings as keys, and high-resolution image embeddings as values. Using our retriever function, we compute the approximate high-resolution embeddings for the query. We concatenate ( $\oplus$ ) the approximated visual feature with our query embedding.

$P$ . Let  $G_i = E(g_i)$  represent the location embedding,  $S_i = I(s_i)$  represent the frozen image embedding, and  $P_i = P(I(s_i)) = P(S_i)$  represent the image embedding obtained from the projection layer for the  $i^{th}$  sample. We minimize the objective:

$$L_i^{loc} = \frac{-1}{k} \sum_{i=1}^k \log \frac{\exp(G_i \cdot P_i / \tau)}{\sum_{j=1}^k \exp(G_i \cdot P_j / \tau)} \quad (1)$$

$$L_i^{img} = \frac{-1}{k} \sum_{i=1}^k \log \frac{\exp(G_i \cdot P_i / \tau)}{\sum_{j=1}^k \exp(G_j \cdot P_i / \tau)} \quad (2)$$

$$L_i = (L_i^{loc} + L_i^{img}) / 2 \quad (3)$$

The training process is shown in Figure 2-(a). We contrastively align the geolocations and satellite images by minimizing their CLIP objective. This formulation is identical to SatCLIP [18], so we use the pretrained SatCLIP image and location encoders for simplicity. For query location  $g_i$ ,  $S_i$  encodes the high-resolution visual features since  $I$  is a powerful pre-trained image encoder. Both  $P_i$  and  $G_i$  encode low-resolution shared information between  $g_i$  and  $s_i$ . The resulting geo-embedding,  $G_i$ , is sub-optimal as it loses unique visual information crucial for many downstream tasks. To solve this, we propose a retriever function  $R_\tau$  that allows us to approximate the high-resolution information  $S_i$  given  $g_i$ . Let  $\epsilon$  denote the deviation from the true

visual embedding. We can then explicitly add the approximate visual information  $\hat{S}_i$  to our location embedding  $G_i$ :

$$R_\tau(g_i) = \hat{S}_i = S_i + \epsilon \quad (4)$$

$$RANGE(g_i) = \hat{S}_i \oplus G_i \quad (5)$$

## 4.2. Retrieval Augmented Neural Fields

**Creating the Retrieval Database:** Once the contrastive training is done, we create a retrieval database as shown in Figure 2-(b). We take  $\{g_i, s_i\}$  sampled uniformly across the globe using the SatCLIP [18] dataset. For every location  $g_i$  in our data, we compute: a) the shared-information image embedding  $R_i^L = P(I(s_i))$ , and b) the high-resolution image embedding  $R_i^H = \bar{I}(s_i)$ ; we refer to these sets of embeddings as low-resolution embeddings and high-resolution embeddings, respectively, in the rest of the paper. In practice,  $\bar{I}$  can differ from  $I$ ; the only condition is that  $\bar{I}$  is a pretrained image feature extractor. For our purposes, we use SatMAE [6] as  $\bar{I}$ .

**Retriever Function:** Let  $R^L = \{R_1^L, R_2^L, \dots, R_N^L\}$ , and  $R^H = \{R_1^H, R_2^H, \dots, R_N^H\}$  represent the set of all low-resolution and high-resolution image embeddings in our database. After creating the database, we need a method to sample appropriate high-resolution image information for a query location. Naively, we can set this as a lookup operation, where  $G_i$  is the query,  $R^L$  is the set of all keys, and  $R^H$

is the set of all values. We have a function  $\text{sim}(G_i, R_j^L)$ , which gives the alignment scores between  $i$  and  $j$ . In our setting, this similarity function is a simple cosine similarity. For a query  $G_i$ , we find  $R_j^L$  with the highest alignment score and return the key  $R_j^H$ . In practice,  $R_j^H$  is noisy. The semantically closest image in the database can contain additional information that is irrelevant to the location. Therefore, naively adding  $R_j^H$  has the potential of introducing incorrect information to our geo-embeddings.

We instead use a soft selection criteria. First, we compute the query’s alignment score with each key in the database using cosine similarity. We use a softmax function to convert the alignment scores to probabilities. We also use a temperature parameter  $\tau$  to slightly reshape the probability distribution;  $\tau$  is extremely robust as it does not need to be fine-tuned for different tasks or for databases with different distributions. We use the resulting probabilities to compute a weighted average across all values, i.e., high-resolution image embeddings in the database. The resulting embedding is an approximation of  $S_i$ , where the information contributed by each image is weighted by its alignment score with the query. We concatenate this approximate high-resolution image embedding with our original location embedding  $G_i$  to obtain the multi-resolution RANGE embedding  $\text{RANGE}_i$ . A high-level view of this process is shown in Figure 2-(c).

$$\begin{aligned} \text{RANGE}_i &= R_\tau(G_i) \oplus G_i \\ &= \frac{1}{N} \sum_{k=1}^N \frac{e^{\text{sim}(G_i, R_k^L)/\tau}}{\sum_{j=1}^N e^{\text{sim}(G_i, R_j^L)/\tau}} * R_k^H \oplus G_i \\ &\text{where } G_i \in \mathbb{R}^M, R_k^L \in \mathbb{R}^M, R_k^H \in \mathbb{R}^N. \end{aligned} \quad (6)$$

**Adding smoothness constraints:** We also propose another version of our model called  $\text{RANGE}^+$ . Here, we additionally use geodesic similarity to approximate the visual features of the location. We convert our query location to 3D cartesian coordinates and use angular distance to find the spatially closest image in the database. This spatial retrieval allows us to impose spatial smoothness on our embeddings since locations close to each other are forced to have similar retrieval, irrespective of the semantics. Similar to  $\text{RANGE}$ , we compute a weighted average across all values in the database using the angular similarity between the query and keys. However, we explicitly mask samples with angular similarity lower than a threshold, i.e., significantly distant samples have no contribution. The spatially retrieved embedding is added to the semantically retrieved embedding and weighted using a  $\beta$  parameter. This  $\beta$  parameter controls the level of spatial smoothness enforced on the  $\text{RANGE}$  embeddings, as shown in Figure 5. Let  $g_i^{3D}$  be the query location,  $l$  be the set of all database locations

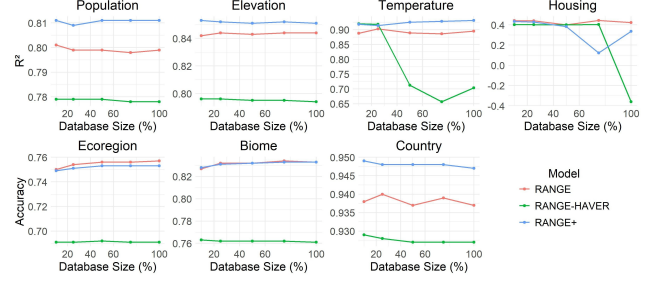


Figure 3. Performance of our model with respect to the database size. The results show that compared to RANGE-HAVER, both RANGE and RANGE<sup>+</sup> are very robust to changes in database size. We can maintain the same performance even when only using 10% of the samples in the database.

in 3D cartesian coordinates, and  $\text{sim}(g_i^{3D}, l_k)$  be their angular similarity:

$$\begin{aligned} \text{RANGE}_i^+ &= \frac{1}{N} (\beta * \sum_{k=1}^N \frac{e^{\text{sim}(G_i, R_k^L)/\tau_1}}{\sum_{j=1}^N e^{\text{sim}(G_i, R_j^L)/\tau_1}} * R_k^H + \\ &(1 - \beta) \sum_{k=1}^N \frac{e^{\text{sim}(g_i^{3D}, l_k)/\tau_2}}{\sum_{j=1}^N e^{\text{sim}(g_i^{3D}, l_j)/\tau_2}} * R_k^H) \oplus G_i \\ &\text{where } G_i \in \mathbb{R}^M, g_i^{3D} \in \mathbb{R}^3, l_k \in \mathbb{R}^3 \end{aligned} \quad (7)$$

When  $\beta$  is set to 0, only spatially relevant features are used, and we refer to this setting as RANGE-HAVER.

## 5. Experiments and Results

### 5.1. Quantitative Evaluation on Downstream Tasks

We evaluate the efficacy of RANGE and RANGE<sup>+</sup> embeddings on a wide variety of downstream applications. We create our database using the SatCLIP [18] dataset with 82k locations. For RANGE<sup>+</sup>, we set  $\beta$  to 0.5. We choose 3 classification tasks and 4 regression tasks for this experiment as used by prior work [18]. The classification tasks are biome classification [8], ecoregion classification [8], and country classification [18]. The regression tasks are air temperature prediction [14], elevation prediction [32], population density prediction [32], and housing price prediction [30]. Results on climate data from ERA5 are shown in Section 10 of the supplementary material. The objective is to learn a linear predictive model that solves the underlying task. For fairness, we use the same hyperparameters for all tasks, i.e., the same  $\tau$  and  $\beta$  parameters for each task. The details of the hyperparameters are listed in the supplementary material. We compare our method with state-of-the-art parametric [5, 18, 27, 28, 39] and non-parametric [10, 25] representations. The results of this experiment are shown in Table 2.

The results show that our method outperforms all state-of-the-art models in the classification tasks with significant

	Biome	EcoRegions	Country	Temperature	Elevation	Population	Cali-Housing
Direct	29.1	0.6	66.9	0.381	0.025	0.053	0.238
Cartesian_3D	30.2	1.8	66.9	0.362	0.030	0.162	0.240
Wrap [25]	34.4	1.1	69.7	0.861	0.085	0.328	0.239
Theory [10]	33.5	1.0	72.5	0.849	0.093	0.330	0.254
SphereM [28]	36.4	27.3	72.7	0.629	0.139	0.302	0.423
SphereM <sup>+</sup> [28]	58.7	50.1	76.1	0.886	0.294	0.421	0.543
SphereC [28]	36.3	52.9	72.9	0.461	0.185	0.335	0.496
SphereC <sup>+</sup> [28]	53.2	61.6	73.6	0.842	0.260	0.392	<u>0.544</u>
CSP-INat [27]	61.1	57.1	75.9	0.717	0.388	0.554	0.462
CSP-FMoW [27]	61.4	58.0	81.3	0.865	0.399	0.580	0.541
SINR [5]	67.9	54.9	88.3	<b>0.942</b>	0.644	0.726	0.420
GeoCLIP [39]	<u>70.2</u>	71.6	81.3	0.916	0.604	0.698	<b>0.708</b>
SatCLIP [18]	68.9	69.3	82.8	0.825	0.666	0.684	0.400
RANGE	<b>83.3</b>	<b>75.7</b>	<u>93.7</u>	0.895	<u>0.844</u>	<u>0.799</u>	0.422
RANGE <sup>+</sup>	<b>83.3</b>	<u>75.3</u>	<b>94.7</b>	<u>0.931</u>	<b>0.851</b>	<b>0.811</b>	0.336

Table 2. Our model shows improvements over a variety of tasks compared to state-of-the-art models. The first 3 columns are classification tasks with accuracy metrics and the last 4 columns are regression tasks with  $R^2$  metric. We show improvements with significant margins across many of the tasks. Our model underperforms in the Cali-housing dataset. We further analyze this behavior in our discussions.

	type	temp.	Biome	Eco	Country	Temperature	Elevation	Population	Cali-Housing
SatCLIP	base-model	-	68.9	69.3	82.8	0.825	0.666	0.684	0.400
RANGE	top-1	-	75.6	65.2	85.6	0.817	0.766	0.742	0.444
	top-k	-	82.8	76.8	90.6	0.884	0.810	0.771	<b>0.619</b>
	soft selection	fixed	83.3	75.7	93.7	0.895	0.844	0.799	0.422
	soft selection	per task	<u>83.5</u>	<u>75.8</u>	94.5	0.922	<b>0.857</b>	0.809	<u>0.465</u>
	soft selection	per task	<u>83.5</u>	75.3	<u>94.7</u>	<u>0.931</u>	0.851	<u>0.811</u>	0.336
RANGE <sup>+</sup>	soft selection	per task	<b>83.7</b>	<b>75.9</b>	<b>94.9</b>	<b>0.932</b>	<u>0.855</u>	<b>0.813</b>	0.460
gain (%)			+21.5	+9.5	+14.6	+12.9	+28.37	+18.8	+15.0

Table 3. Ablation of different versions of our model compared with SatCLIP as the base model. Our soft selection using weighted averages over the database performs better than top-k selection. We also find that a single temperature works well for most tasks. Fine-tuning the temperature per task is generally not required as the gain is marginal. The last row shows the difference between the base model and RANGE<sup>+</sup>.

	top-1	top-3	top-5	top-10
Img	66.1	83.3	88.0	92.2
Img+CSP	72.9	87.9	91.6	94.8
Img+GeoCLIP	72.9	88.2	91.9	<u>95.2</u>
Img+CSP_INat*	74.4	88.8	92.2	94.9
Img+SatCLIP	75.1	88.7	91.9	94.5
Img+RANGE	<b>75.2</b>	<b>89.6</b>	<b>92.9</b>	<b>95.5</b>
Img+RANGE <sup>+</sup>	<u>75.1</u>	<u>89.5</u>	<u>92.8</u>	<b>95.5</b>

Table 4. **Top-k classification accuracy on INat-2018 test split:** Location information acts as a strong prior in fine-grained species classification. Our method shows competitive top-k results against state-of-the-art models.

margins. In biome, ecoregion, and country classification, we advance the state-of-the-art by 13.1%, 4.1%, and 6.4%, respectively. Among the regression tasks, we achieve SoTA

performance for elevation and population density prediction with significant gains and narrowly come second in the air temperature prediction task. Our method, however, underperforms in the housing price prediction task. We suspect this is because the visual features that define housing change dramatically over time. The standard California housing dataset is based on a 1990 census, while our retrieved visual features are extracted from 2020 Sentinel data. The results also demonstrate that the hyperparameters for our model are robust and do not need to be tuned for specific tasks to achieve good performance. Furthermore, RANGE<sup>+</sup>, on average, outperforms RANGE by a small margin, which can be attributed to the spatial smoothness constraint.

## 5.2. Using Location Embeddings as Geo-prior

We evaluate our model on the task of fine-grained species classification. We use the iNaturalist 2018 dataset [38],

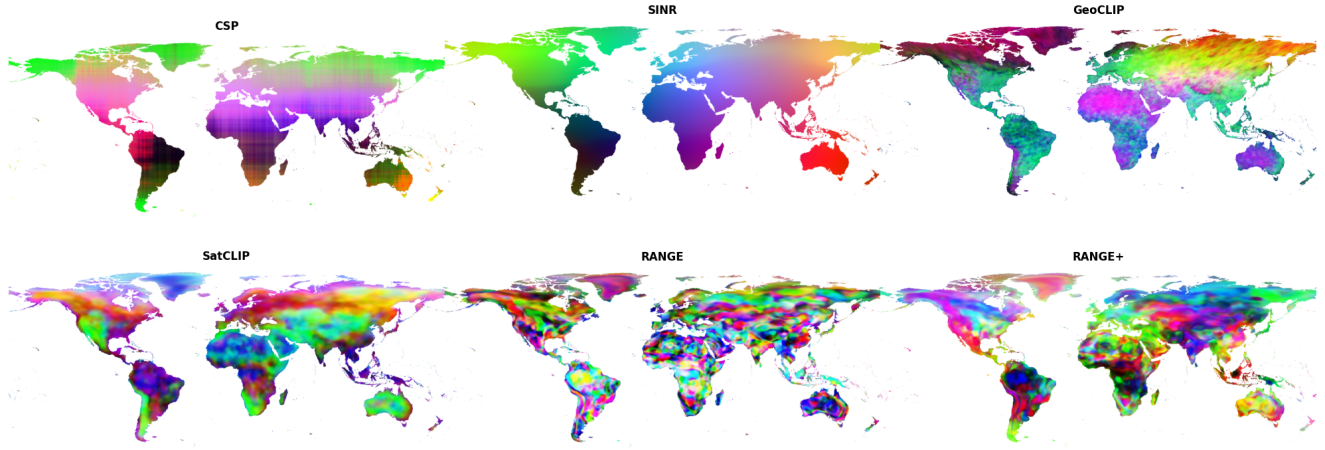


Figure 4. We visualize the geo-embeddings from different models by projecting them into a 3-dimensional vector using Independent Component Analysis (ICA). The results suggest that by explicitly adding visual features, our method learns more high-frequency information compared to the existing models.

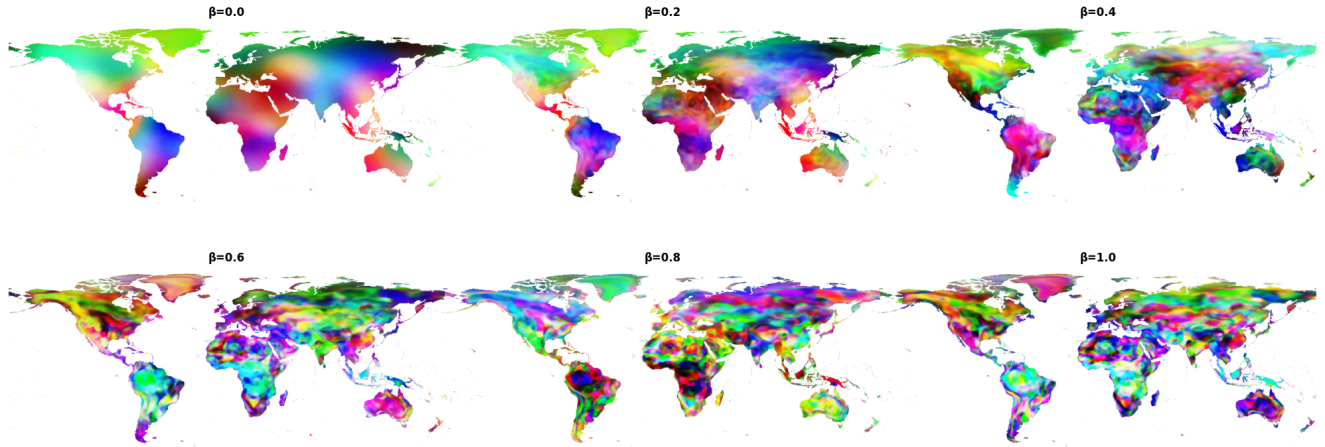


Figure 5. Interpolating the  $\beta$  parameter in  $\text{RANGE}^+$  allows us to control the spatial smoothness of our embeddings. The results show that  $\text{RANGE}^+$  can be used to generate neural fields of geo-embeddings at multiple frequencies.

which contains data from 8,142 species. Prior works [18, 25, 33] have shown that adding location information can improve the performance of pretrained image classification models on iNat as knowing the location of an image can provide a strong prior about its content. For our experiment, we follow a similar setup as [33]. We take a pretrained image classification model, which gives us  $P(y | I)$ . We then train a linear model that takes as input the pre-trained geo-embeddings and predicts the species categories, i.e.,  $P(y | G)$ . The final distribution is computed as the product of the two distributions, i.e.,  $P(y | I, G) = P(y | I) * P(y | G)$ . This experiment also aims to show that location representations obtained using self-supervised training can be eas-

ily generalized towards tasks like iNat. Hence, we exclude SINR [5] from this comparison as it is trained in a supervised manner on the iNaturalist data itself.

The results of this experiment are shown in Table 4. The performance of the pretrained image classifier is 66.1%. The addition of location information improves the accuracy across all models. We see that RANGE and  $\text{RANGE}^+$  achieve competitive scores, narrowly outperforming state-of-the-art models. The results from this experiment highlight the benefits of using good location representations as geo-prior to solve fine-grained image classification tasks.

### 5.3. Ablation Study

**Robustness to Database Size:** The RANGE database that we used for our model in Section 5.1 was created using the SatCLIP dataset with 82k locations. However, our method is extremely robust and can be used with much smaller database sizes. We conducted an experiment where we used smaller samples of the SatCLIP data to create our RANGE database. We created a database with 75%, 50%, 25%, and 10% of the total data. We then trained a linear model for RANGE and RANGE<sup>+</sup> using each of these subsets as our retrieval database. We also trained another version of our method where we only use geodesic distance for retrieval, i.e.  $\beta = 0$ , which we call RANGE-HAVER. We used the same tasks described in Section 5.1 to evaluate these models.

Our results from figure 3 show that we can obtain similar performance while using only 10% of the original database size. In terms of storage, this database only occupies 85 Megabytes in disk, which is very efficient. The results also support our initial claim that there is low variance in satellite images; thus, a small number of satellite images can capture many important semantics across the globe. We see that the performance for both RANGE and RANGE<sup>+</sup> is stable across different settings and different tasks. The results also show that RANGE-HAVER, which only relies on spatially retrieved information, can be unstable, as seen in figure 3. These results highlight the pitfalls of depending completely on spatial similarity for retrieval. Locations that are spatially close do not necessarily share similar semantics. Therefore, as you change the sample distribution in the database, you can introduce noisy retrievals, which amplify the noise in the estimated visual embedding. Semantic-based retrieval strategy mitigates such issues.

**Impact of architectural components:** Table 3 shows the impact of different components of our method compared to SatCLIP as the base model. We also compare our method with two other methods of retrieval: top-1 and top-k (where k was set to 100). The results show that our retriever function generally yields better results than the top-k selection approach. This can be attributed to the fact that top-k averaging does not consider the measure of similarity between query location and keys. We also performed a linear search to find the optimal temperature parameters of RANGE for each task separately. Our results show we are able to make marginal improvements by tuning the temperature parameter for individual tasks. The results indicate the temperature parameters of RANGE are robust and do not need to be fine-tuned for individual tasks in most cases.

### 5.4. Qualitative Evaluation of Geo-embeddings

**Visual comparison of geo-embeddings:** We qualitatively compare the RANGE embeddings with other location em-

beddings. We use Independent Component Analysis (ICA) to project the embeddings to a 3-dimensional vector and use them as RGB channels to visualize them over the globe. Figure 4 shows the visualizations for each model. Qualitatively, we see that models like CSP and SINR are extremely smooth, suggesting that they predominantly capture low-frequency information. GeoCLIP and SatCLIP capture relatively high-frequency information. The visualizations further suggest that RANGE and RANGE<sup>+</sup> embeddings are able to capture even higher frequency information.

**Visualizing the impact of  $\beta$  parameter:** Secondly, we visualize the impact of the  $\beta$  parameter in equation (7). A quantitative evaluation of the  $\beta$  parameter is presented in Section 9 of the supplementary material. Setting  $\beta = 1$  gives us the RANGE embeddings, setting  $\beta = 0$  gives us RANGE-HAVER embeddings, and setting  $\beta = 0.5$  gives the RANGE<sup>+</sup> embeddings. This parameter controls the contribution of spatially-retrieved visual information to the semantically-retrieved visual information. This additional information acts as a spatial smoothness constraint on the RANGE embeddings. At  $\beta = 0$ , we add the maximum constraint, which gives us low-frequency embeddings similar to CSP and SINR. As we increase the value of  $\beta$ , we increase the frequency of our embeddings, i.e. locations that are spatially close to each other but are semantically different have different representations. Thus, the  $\beta$  parameter allows us to explicitly control the smoothness of the RANGE embeddings and, therefore, allows generating location embeddings at multiple frequencies.

## 6. Discussion and Conclusion

Self-supervised learning has been a key enabler of the rapid recent progress in computer vision and natural language processing. Representations from models like SimCLR [3], and CLIP [31] have facilitated the creation of many powerful vision and language models. Recent works like SatCLIP [18], GeoCLIP [39], and CSP [27] showed that the same self-supervised techniques can be used to learn powerful representations for geographic locations, which are useful for a broad range of tasks.

In this paper, we introduced a simple, yet effective strategy for improving such geographic embeddings, moving past previous assumptions of multiview redundancy. We proposed a retrieval-augmented strategy for estimating visual features from an auxiliary database of visual embeddings. Our approach results in significant improvements for a variety of tasks as compared to purely parametric embedding strategies with only a modest increase in storage requirements. Our method is efficient, robust, and multi-scale. We hope our insights and results substantiate our recommendation of using RANGE as a general-purpose location encoder for geospatial tasks.

## 7. Acknowledgments

This research used the TGI RAILs advanced compute and data resource which is supported by the National Science Foundation (award OAC-2232860) and the Taylor Geospatial Institute.

## References

- [1] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. In *NeurIPS*, pages 15309–15324, 2022. 3
- [2] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *ICML*, pages 2206–2240, 2022. 3
- [3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 8
- [4] Wenhui Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. In *ICLR*, 2023. 3
- [5] Elijah Cole, Grant Van Horn, Christian Lange, Alexander Shepard, Patrick Leary, Pietro Perona, Scott Loarie, and Oisín Mac Aodha. Spatial implicit neural representations for global-scale species mapping. In *ICML*, 2023. 1, 3, 5, 6, 7
- [6] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In *NeurIPS*, 2022. 2, 4
- [7] Aayush Dhakal, Adeel Ahmad, Subash Khanal, Srikumar Sastry, Hannah Kerner, and Nathan Jacobs. Sat2Cap: Mapping fine-grained textual descriptions from satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 533–542, 2024. 3
- [8] Eric Dinerstein, David Olson, Anup Joshi, Carly Vynne, Neil D Burgess, Eric Wikramanayake, Nathan Hahn, Suzanne Palminteri, Prashant Hedao, Reed Noss, et al. An ecoregion-based approach to protecting half the terrestrial realm. *BioScience*, 67(6):534–545, 2017. 5
- [9] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. 3
- [10] Ruiqi Gao, Jianwen Xie, Song-Chun Zhu, and Ying Nian Wu. Learning grid cells as vector representation of self-position coupled with matrix representation of self-motion. In *ICLR*, 2019. 1, 3, 5, 6
- [11] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 3
- [12] Álvaro González. Measurement of areas on a sphere using fibonacci and latitude–longitude lattices. *Mathematical geosciences*, 42:49–64, 2010. 1
- [13] Kartik Gupta, Thalaiyasingam Ajanthan, Anton van den Hengel, and Stephen Gould. Understanding and improving the role of projection head in self-supervised learning. *arXiv preprint arXiv:2212.11491*, 2022. 2, 3
- [14] Josh Hooker, Gregory Duveiller, and Alessandro Cescatti. A global dataset of air temperature derived from satellite remote sensing and weather stations. *Scientific Data*, 5(1), 2018. 5
- [15] Andy V Huynh, Lauren E Gillespie, Jael Lopez-Saucedo, Claire Tang, Rohan Sikand, and Moisés Expósito-Alonso. Contrastive ground-level image and remote sensing pre-training improves representation learning for natural world imagery. In *ECCV*, 2024. 3
- [16] Subash Khanal, Eric Xing, Srikumar Sastry, Aayush Dhakal, Zhexiong Xiong, Adeel Ahmad, and Nathan Jacobs. Psm: Learning probabilistic embeddings for multi-scale zero-shot soundscape mapping. In *ACM Multimedia*, 2024. 3
- [17] Yuval Kirstain, Omer Levy, and Adam Polyak. X&fuse: Fusing visual information in text-to-image generation. *arXiv e-prints*, pages arXiv–2303, 2023. 3
- [18] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. SatCLIP: Global, general-purpose location embeddings with satellite imagery. In *AAAI*, 2025. 1, 2, 3, 4, 5, 6, 7, 8
- [19] Yuma Koizumi, Yasunori Ohishi, Daisuke Niizumi, Daiki Takeuchi, and Masahiro Yasuda. Audio captioning using pre-trained large-scale language model guided by audio-based similar caption retrieval. *arXiv preprint arXiv:2012.07331*, 2020. 3
- [20] Nico Lang, Walter Jetz, Konrad Schindler, and Jan Dirk Wegner. A high-resolution canopy height model of the earth. *Nature Ecology & Evolution*, 7(11):1778–1789, 2023. 1, 3
- [21] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. 3
- [22] 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. In *NeurIPS*, 2021. 3
- [23] Paul Pu Liang, Zihao Deng, Martin Q Ma, James Y Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. In *NeurIPS*, 2023. 2, 3
- [24] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *CVPR*, 2022. 3

- [25] Oisín Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *ICCV*, 2019. 1, 3, 5, 6, 7
- [26] Gengchen Mai, Krzysztof Janowicz, Ling Cai, Rui Zhu, Blake Regalia, Bo Yan, Meilin Shi, and Ni Lao. Se-kge: A location-aware knowledge graph embedding model for geographic question answering and spatial semantic lifting. *Transactions in GIS*, 24(3):623–655, 2020. 3
- [27] Gengchen Mai, Ni Lao, Yutong He, Jiaming Song, and Stefano Ermon. CSP: Self-supervised contrastive spatial pre-training for geospatial-visual representations. In *ICML*, 2023. 1, 3, 5, 6, 8
- [28] Gengchen Mai, Yao Xuan, Wenyun Zuo, Yutong He, Jiaming Song, Stefano Ermon, Krzysztof Janowicz, and Ni Lao. Sphere2vec: A general-purpose location representation learning over a spherical surface for large-scale geospatial predictions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:439–462, 2023. 5, 6
- [29] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2
- [30] R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997. 5
- [31] 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 *ICML*, pages 8748–8763, 2021. 3, 8
- [32] Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature communications*, 12(1):4392, 2021. 5
- [33] Marc Rußwurm, Konstantin Klemmer, Esther Rolf, Robin Zbinden, and Devis Tuia. Geographic location encoding with spherical harmonics and sinusoidal representation networks. In *ICLR*, 2024. 1, 3, 7
- [34] Srikumar Sastry, Subash Khanal, Aayush Dhakal, Di Huang, and Nathan Jacobs. Birdsat: Cross-view contrastive masked autoencoders for bird species classification and mapping. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7136–7145, 2024. 3
- [35] Junyoung Seo, Susung Hong, Woosok Jang, Inès Hyeonsu Kim, Min-Seop Kwak, Doyup Lee, and Seungryong Kim. Retrieval-augmented score distillation for text-to-3d generation. In *Forty-first International Conference on Machine Learning*, 2024. 3
- [36] Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17, 2023. 3
- [37] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, 2021. 2
- [38] Grant Van Horn, Oisín Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *CVPR*, 2018. 6
- [39] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. GeoCLIP: Clip-inspired alignment between locations and images for effective worldwide geolocalization. In *NeurIPS*, 2024. 1, 2, 3, 5, 6, 8
- [40] Zichao Wang, Weili Nie, Zhuoran Qiao, Chaowei Xiao, Richard Baraniuk, and Anima Anandkumar. Retrieval-based controllable molecule generation. In *The Eleventh International Conference on Learning Representations*. 3
- [41] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun 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
- [42] 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. In *International Conference on Learning Representations*. 3
- [43] Chenhao Zheng, Ayush Shrivastava, and Andrew Owens. Exif as language: Learning cross-modal associations between images and camera metadata. In *CVPR*, 2023. 3