

Sat2Cap: Mapping Fine-Grained Textual Descriptions from Satellite Images

Supplementary Material

1. Implementations Details

We use a ViT-32B as the CLIP image encoder. This image encoder is kept frozen throughout our training procedure. We use a ViT-32B architecture as the backbone for our Sat2Cap model. The Sat2Cap backbone is initialized using CLIP weights. Following [4] we use an AdamW optimizer [3] with a learning rate of $1e - 05$ with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. We also use a learnable temperature parameter which is initialized at $\tau = 0.07$. We use Cosine Annealing with Warm Restarts [2] as the learning rate scheduler.

We augment the overhead images using RandomResizedCrop and RandAugment [1]. 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.

2. 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 1 shows the retrieval results at two different time settings (11:00 p.m. and 08:00 a.m.). Figure 2 shows the captions generated for the same location over long (order of months) and short (order of hours) term variations. Finally, Figure 3 shows how our model can generate maps that adapt to temporal variations for a given prompt.

3. 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[4] text embedding and compute its similarity with all images in the test set. Figure 4 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 4, 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 looks like trails nearby residential areas. Hence, Sat2Cap embeddings demonstrate a good understanding of fine-grained variations of textual concepts.

4. 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 5.

5. Geolocating 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 6 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.

6. Dataset

We introduced a cross-view dataset with overhead images and co-located ground-level images taken from the YFCC100M [5] dataset. Figure 7 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.

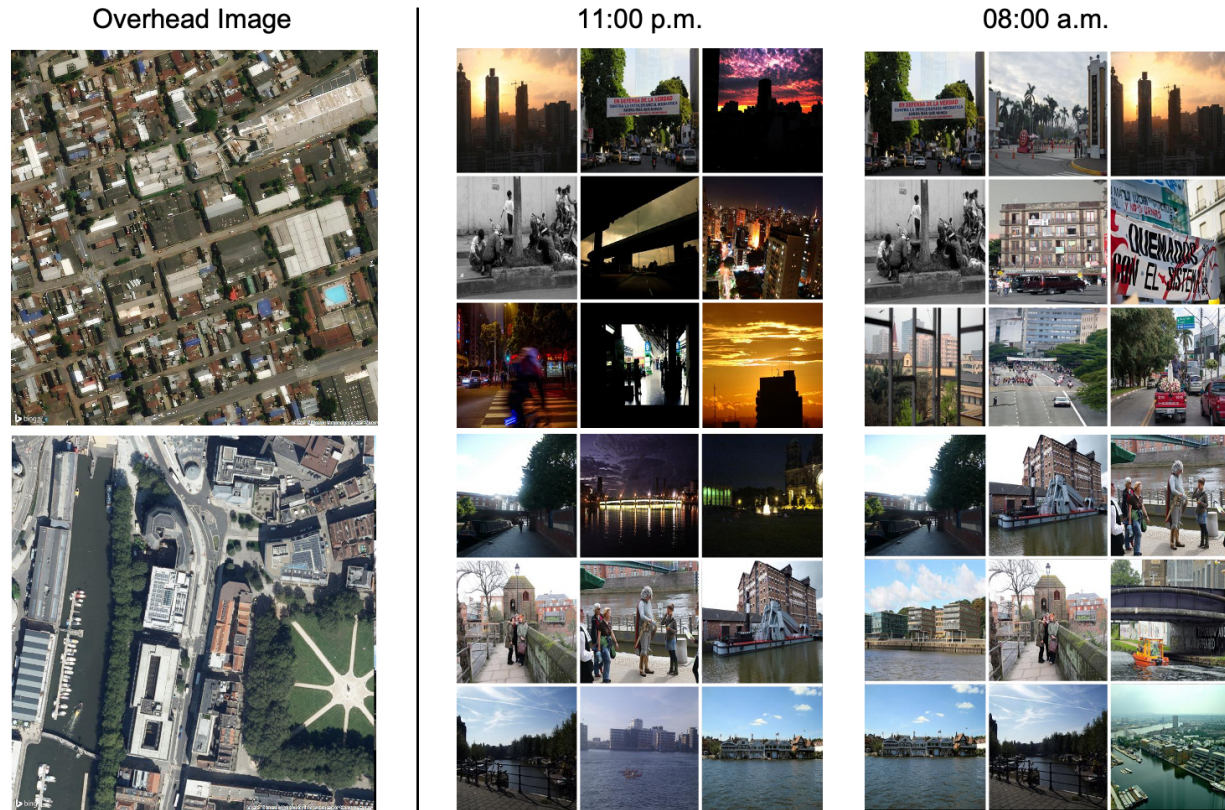


Figure 1. **Top-9 overhead-to-ground image retrieval with temporal manipulation:** We show the 9 closest ground-level images for a query overhead image at two different time settings (11:00 p.m. and 08:00 a.m.).

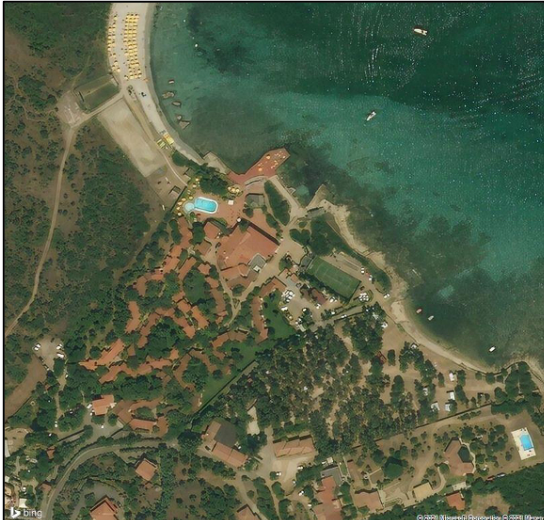
	Model	Date/Time	Description
	CLIP	-	<i>"aerial view of a beach"</i>
	Ours	May 20 08:00 am	<i>"property image sea facing apartment with swimming pool, terrace in a quiet residential area."</i>
		Dec 20 10:00 am	<i>"Sailboat on the sea in winter"</i>
		Dec 20 05:00 pm	<i>"Person on the beach at night"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

Figure 2. **Dynamic Caption Generation:** Our Sat2Cap embeddings dynamically adapt to temporal manipulations, facilitating dynamic caption generation.

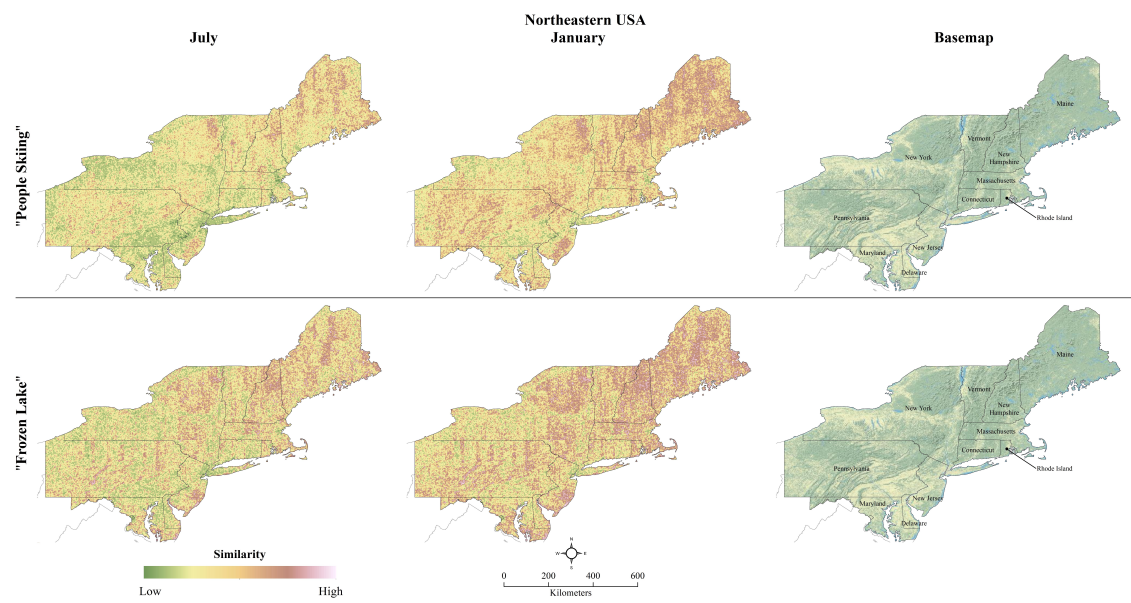
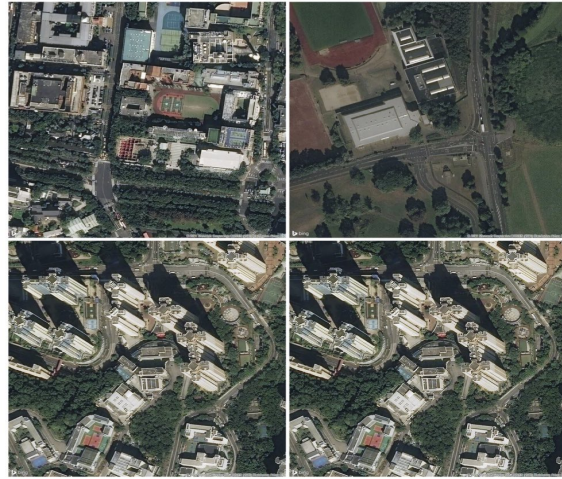


Figure 3. **Dynamic Maps:** We show zero-shot maps of the northeast US at two different temporal settings (July and January)

"Ongoing soccer match"



"Ongoing basketball match"



"People driving cars"



"People riding horses"



"Person on a long hike"



"Person on a long run"



Figure 4. **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.

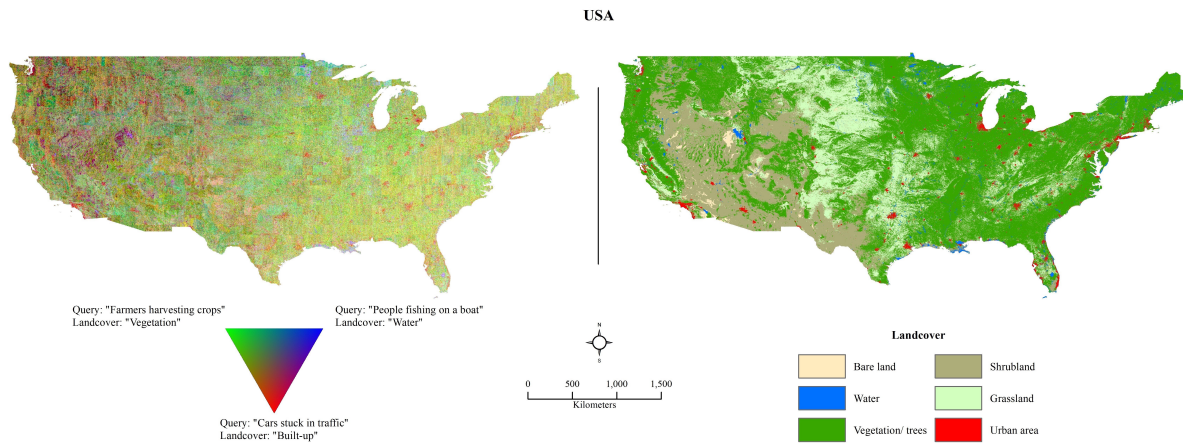


Figure 5. Zero-shot map of the USA

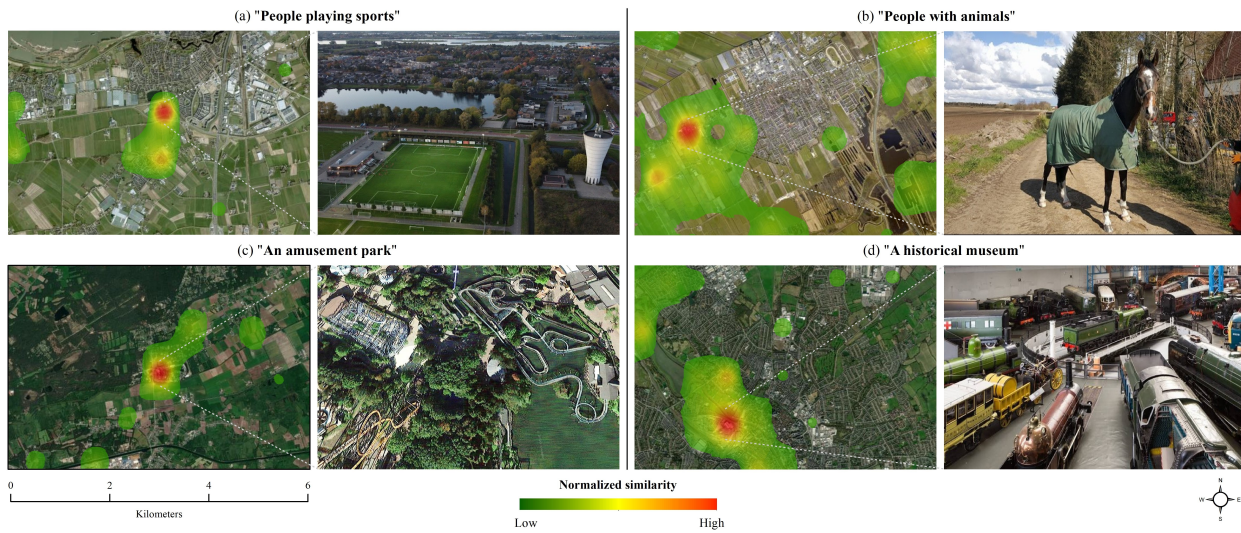


Figure 6. **Localizing textual queries at finer resolution:** For each prompt, the image on left shows the big region which is used for inference. The image on the right shows an image of the ground-level scene at the point with the highest activation, which was taken by entering the location in Google Maps

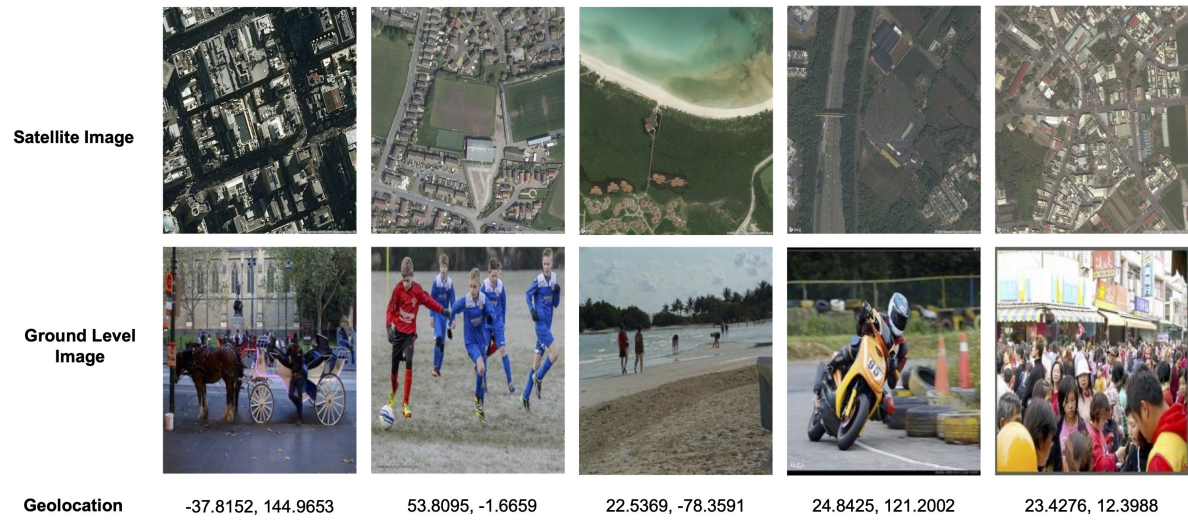


Figure 7. Examples of co-located overhead and ground images in our dataset. The ground-level images describe more detailed concepts of the given locations than their overhead counterparts.

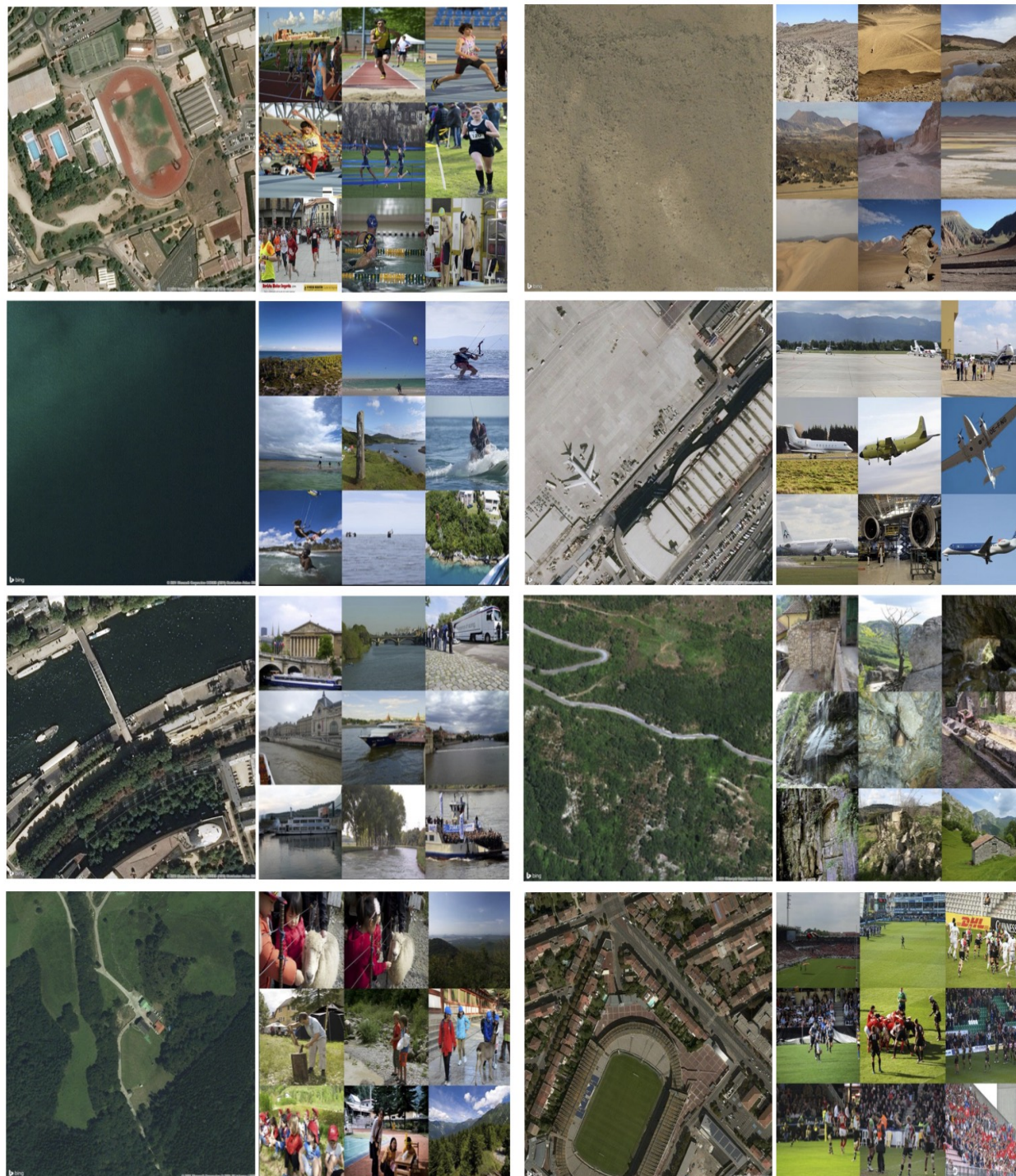


Figure 8. **Top-9 overhead-to-ground retrieval:** Our model can infer fine-grained concepts of ground-level scenes through overhead imagery. Sat2Cap accurately retrieves probable concepts for a given geolocation using an overhead image.

References

- [1] 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](#)
- [2] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. [1](#)
- [3] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. [1](#)
- [4] 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](#)
- [5] 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. [1](#)