A Study of Zero-shot Learning for Visual Search on Satellite and Aerial Images

A. Chuong Dang
Automated Mapping Platform
Woven Alpha, Inc.
Tokyo, Japan
dan.anh.chuong@woven-planet.global

Ion-George Todoran
Automated Mapping Platform
Woven Alpha, Inc.
Tokyo, Japan
george.todoran@woven-planet.global

Srushti Rashmi Shirish
Automated Mapping Platform
Woven Alpha, Inc.
Tokyo, Japan
srushti.rashmishirish@woven-planet.global

***Abstract*—In this article, we present a visual search system based on the latest Deep Learning techniques, which enables users to find images containing similar content as a query image. As many applications battle the dual challenge of limited labelled data that does not cover all the possible classes, we propose to mitigate this issue with an approach we call zero shot learning. We prove the potential of this approach by extensively experimenting on 3 of all time popular aerial imagery datasets. In addition, we show that pre-training the model on top-down imagery improves the final performance of the visual search system.**

***Keywords*—visual search; zero-shot learning; data indexing; deep learning**

I. Introduction

At present there are more than 150 operational satellites equipped with sensors gathering petabytes of data each year for a variety of Earth observations tasks [1]. Usually, this data is ingested and indexed in huge databases considering information like the date of the image, the polygon of the covered area, the number of spectral bands, the image resolution, the number of bits per pixel, the vendor name, and so on. Then it is fairly easy to extract the desired data by employing a query composed of the aforementioned information, e.g., give me all the Maxar WorldView-3 RGB orthorectified imagery acquired after 2017 covering Paris area.

Unfortunately, this type of queries cannot be used for more complex tasks like give me all the images in Africa that contain power plants. In order to enable such a query we need to extract the semantic content of each image. With the Deep Learning revolution, Machine Learning (ML) models based on Convolutional Neural Networks (CNN) proved to be efficient in extracting the semantic content of an image [2]–[4].

Open Street Map (OSM) contains 130 object classes and for many practical applications we could consider they cover all of the use cases. Nevertheless, we identify two practical limitations of this approach. The first one is that the semantic content of many images is too complex to be tagged as belonging to a single class. Labeling such images with more than one classes is difficult and prone to significant errors. In order to exemplify this, consider the image taken from the SpaceNet5 challenge [5] and shown in Figure 1. In this image, one might identify various semantic contents of interest: buildings, beach, road, cars, pools, tennis court and so on. The second limitation is related to the closed-world assumption. In practice we often need to add one or more classes that were not included in the training dataset. In order to solve the above-mentioned problems, researchers considered weak supervision and **zero-shot learning strategies**, i.e., to recognize the objects whose instances were never seen during the training. Since our model has never learnt these additional classes, neither has learnt to classify complex real world scenarios explicitly, we call this approach a zero shot learning.

Figure 1. Example of a satellite image with complex visual content

When the semantic content of an image is complex it is easier to provide a query image and ask for similar images. This is the definition of visual search and it includes the construction of a model that transforms an image into a rich, semantic vector representation, called an embedding. Thus, the model is not trying to directly extract the content of the images but to learn an embedding representation that pulls similar images closer in the embedding space and pushes dissimilar images apart.

Figure 2 shows a visual search service consisting of two parts: the first one extracts an embedding for each image based

 ISBN: 978-1-61208-983-6

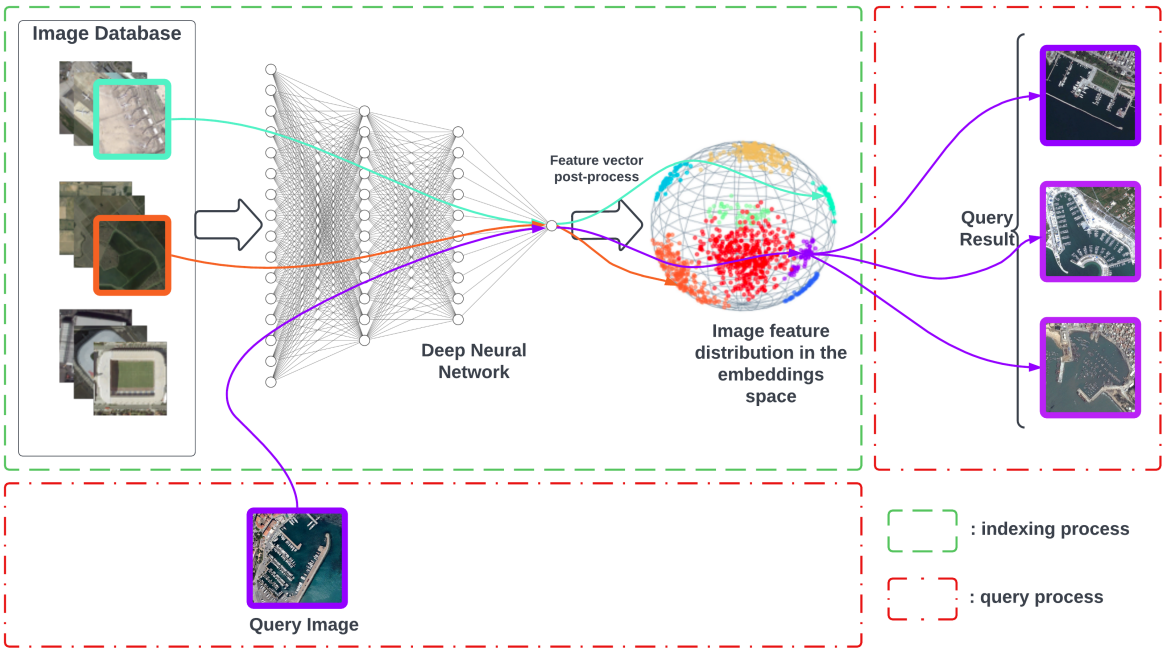

Figure 2. High-level architecture for a visual search system using a deep neural network for learning an image feature vector representation.

on its contents, and the second one deals with the query search in the embedding space. In this paper we mainly focus on the first part and recommend the work published in [6] for the search part.

A visual search service opens new opportunities for improving an existing application or for developing new applications by making use of the right data (inspired from [4]):

- Constructing the training and testing datasets for a new ML-based application, e.g., train an ML model for detecting houses with a pool - for this we need to extract images containing houses with and without pools.
- Better visualisation and understanding of large-scale satellite imagery - it is now possible to easily search and discover similar objects at planetary scale.
- Fast prototyping of a new application - visual search enables us to only extract and use the data that the user really needs for proving the feasibility of a new idea.

This paper is organized as follows. In Section II, we present the prior work related to constructing a visual search system. In Section III, we introduce our solution for constructing a performing visual search system along with the datasets used for doing our experiments. Then, in Section IV, we present the results of various experiences using classic and recent DNN architectures on 3 popular datasets. In Section V, we conduct an ablation study and discuss the importance of feature dimensionality reduction. Finally, we conclude with a summary of future research directions in Section VI.

II. Prior Work

The *classic* solution for extracting embedding vectors for a visual search systems was to aggregate hand-crafted features [7]–[9]. Even though ingenious, these techniques were very difficult to be used on complex imagery at large scale, as it is the case for satellite or aerial imagery.

In the seminal work [10] the authors proved that the representations produced by a deep learning algorithm could be used for image retrieval tasks. Since then, all major work in the domain of visual search focused on extracting the embedding vectors using a deep learning model. We now briefly present the deep learning work for visual search from two aspects:

- *CNN for image retrieval*: Training a CNN model in a supervised manner proved to deliver good results, either by extracting local features (corresponding to some objects of interest) [11] [12] or by extracting global features (corresponding to the entire image). [4]). The main drawback of this technique is the need for huge amounts of manually annotated images with increasing levels of annotation detail.
- *Deep metric learning*: This technique tries to learn the similarity between two images using for example a siamese network [13] or a triplet loss technique [14]. Such models take as input positive samples (corresponding to similar images) and negative samples (not alike images) and it will learn a similarity metric.

III. Method

A. Visual Search Architecture

The central purpose of our approach is to learn an embedding function $\mathbf{e} = f_{\boldsymbol{\theta}}(x)$ where $f_{\boldsymbol{\theta}}$ represents a deep neural network (DNN) with parameters $\boldsymbol{\theta}$, mapping an image x to a feature vector $\mathbf{e}$.

In the embedding space, the distance metric $||f_{\boldsymbol{\theta}}(x_i) - f_{\boldsymbol{\theta}}(x_j)||$ gets the particular meaning of the similarity between the 2 images x_i and x_j. Therefore, a good embedding function should map visually similar images closer to each other in the N dimensional space, where N is the size of the feature vector.

In Figure 2 we illustrate a high-level architecture for a visual search system that uses a DNN for learning a feature vector representation for each image in the training dataset.

B. Indexing and query process

In order to train a DNN for learning the embedding function $\mathbf{e}$, we first need to carefully construct a training image dataset making sure it covers all the semantic content of interest. The visual content of these images could then be labeled for supervised training or the unlabeled data could be directly used for unsupervised training. At the end of the training, the learned embedding function will then be applied to all the images we have. If the dimension of the resulting feature vectors is considered too high they could be passed through a dimensionality reduction post-processing step and then stored in a database. We call this the **indexing process**.

With the indexing process finished, it is now possible to take a query image as input, pass it through the same DNN in order to extract its feature vector and then search for similar images in the embedding space (**query process**).

C. Data

For our experiments we used 3 public aerial imagery datasets, having the main properties summarized in Table I. The heterogeneity of these datasets in terms of image resolution, image dimension, and the labeling strategy, allows us to validate that our method works in the general case of any RGB overhead imagery. For example the UC Merced Land Use Dataset [17] corresponds to aerial orthoimagery from USGS National Map of 20 US regions, having a 30 cm resolution.

TABLE I
AERIAL AND SATELLITE IMAGERY DATASETS USED FOR EXPERIMENTS

Dataset	# images	resolution	# classes	image size
UC Merced [17]	2.1k	0.3 m	21	256x256
AID [18]	10k	8 to 0.5 m	30	600x600
RESISC45 [19]	31.5 k	30 to 0.2 m	45	256x256

In the next section, we present multiple experiments using state-of-the-art DNN models, trained using supervised and unsupervised strategies on the 3 aerial imagery datasets.

IV. EXPERIMENTS AND RESULTS

We strongly believe that the key component to have a robust zero-shot visual search system is to improve the process of extracting feature vector representation. This gives us a better latent space for the query process. To verify this hypothesis, we carried out several experiments seeking the optimal method of pre-training DNN, which is then used to extract embedding vectors for aerial/ satellite images. Our experiments range from utilizing DNNs pre-trained on photographic datasets, to investigating the capability of unsupervised pre-training methods, as well as employing more state-of-the-art DNNs' architecture.

For evaluating quantitative performance of the visual search system we use mean Average Precision (mAP) metric as defined in [15] and Recall@K (R@K) metric as introduced in [16]. Experiments and results are detailed in the following paragraphs.

A. Supervised pre-trained DNNs using large photographic imagery datasets

We first investigated the extraction of embeddings for satellite and aerial images using CNN(s) that have been trained on large scale photographic object imagery (e.g., ImageNet1k [20]) and indoor scene datasets (e.g., Places365 [21]). The interesting conclusion was that even though the CNN model trained on photographic images has little prior knowledge of aerial top-down images, it performed relatively well in modeling the feature space for top-down viewed images. To have a fair comparison across all runs, we only adopted ResNet50 [22] as the backbone CNN architecture in this experiment. The results in term of mAP and R@1 for the 3 datasets described in Section III-C are reported in Table II.

TABLE II
PERFORMANCE COMPARISON OF RESNET50 PRE-TRAINED ON PHOTOGRAPHIC DATASETS USING SUPERVISED TRAINING METHOD

Test dataset	Pre-trained dataset	mAP	R@1
UC Merced Land Use	ImageNet1k	***58.9***	***92.9***
	Places365	54.3	90.2
	ImageNet1k & Places365	57.5	92.6
AID	ImageNet1k	***44.6***	***85.4***
	Places365	42.3	83.3
	ImageNet1k & Places365	44.4	84.0
RESISC45	ImageNet1k	34.0	78.7
	Places365	33.2	77.9
	ImageNet1k & Places365	***35.0***	***80.3***

We clearly observed the difference in performance of models pre-trained on different type of datasets. Specifically, the results were better when the model was pre-trained on an object imagery dataset namely ImageNet1k (denoted as bold italic ***numbers***), compared to Places365, an indoor scene dataset. From our perspective, this result is very appealing since it is intuitive to believe that the indoor scenes images may be semantically closer to top-down satellite ones as they both feature complex real world scenes. This observation makes us wonder what kind of performance we would be able to get with a model backbone pre-trained directly on bird eye view imagery. We present our findings in the following section.

B. Supervised pre-trained DNNs using aerial and satellite imagery datasets

Following the setting of previous experiment, we only deployed ResNet50 as the CNN backbone for this experiment. In order to investigate the effectiveness of using bird eye view images, the factor we changed this time was to pre-train the backbone CNN directly on the airborne imagery

 ISBN: 978-1-61208-983-6

datasets. We iteratively pre-trained the extractor backbone on one of the satellite imagery datasets, mentioned in section section III-C, then tested its performance on the remaining datasets. However, as the UC Merced Land Use dataset is too small, we skipped using this dataset for pre-training and only adopted it for testing instead. We present the performance comparisons in tables III to V. The best performance metrics are denoted as bold italic ***numbers*** while underline numbers are used for baseline metrics from previous experiments.

TABLE III
PERFORMANCE COMPARISON OF RESNET50 BACKBONE PRE-TRAINED USING SUPERVISED METHOD AND AERIAL IMAGERY DATASETS ON UC MERCED LAND USE DATASET

Pre-trained dataset	mAP	R@1
ImageNet1k	58.9	92.9
AID (x224)	60.0	91.0
AID (x320)	62.7	90.9
RESISC45	***78.6***	***95.9***

TABLE IV
PERFORMANCE COMPARISON OF RESNET50 BACKBONE PRE-TRAINED USING SUPERVISED METHOD AND AERIAL IMAGERY DATASETS ON AID DATASET

Pre-trained dataset	mAP	R@1
ImageNet1k	44.6	85.4
RESISC45	***69.3***	***89.2***

TABLE V
PERFORMANCE COMPARISON OF RESNET50 BACKBONE PRE-TRAINED USING SUPERVISED METHOD AND AERIAL IMAGERY DATASETS ON RESISC45 DATASET

Pre-trained dataset	mAP	R@1
ImageNet1k	34.0	78.7
AID (x224)	***44.0***	***80.9***
AID (x320)	43.4	79.6

(x224): input image size as 224x224.
(x320): input image size as 320x320.

Looking at the performance tables, it is evident that pre-training the extractor backbone directly on top-down airborne images has a positive affect on the visual search system performance.

C. Unsupervised pre-trained DNNs using aerial satellite imagery datasets

Recently, unsupervised training has become the approach of choice to pre-train DNNs. It is noticeable that many proposed unsupervised pre-training methods [28]–[33] improve the performance of the model in the down-stream tasks with a considerable margin. The technique enables us to train the neural networks on a huge amount of data without having to lable it. This inspired us to carry out the next batch of experiments, in which we investigated the capability of unsupervised pre-training methods.

Among many available unsupervised training methods for DNNs, we selectively picked DINO (self-**di**stillation with **no** labels) [23], a self-supervised training method, popular for its clarity and effectiveness. Adopting similar settings with earlier experiments, we only used ResNet50 as CNN backbone architecture in this experiment for the sake of fairness in the comparison. Results are reported in tables VI to VIII.

TABLE VI
PERFOMANCES OF RESNET50 BACKBONE PRE-TRAINED USING UNSUPERVISED METHOD AND AERIAL IMAGERY DATASETS ON UC MERCED LAND USE DATASET

Pre-trained dataset	mAP	R@1
ImageNet1k	58.9	***94.7***
AID	55.0	93.1
RESISC45	***63.0***	93.8

TABLE VII
PERFOMANCES OF RESNET50 BACKBONE PRE-TRAINED USING UNSUPERVISED METHOD AND AERIAL IMAGERY DATASETS ON AID DATASET

Pre-trained dataset	mAP	R@1
ImageNet1k	46.7	88.6
RESISC45	***52.1***	***90.1***

TABLE VIII
PERFOMANCES OF RESNET50 BACKBONE PRE-TRAINED USING UNSUPERVISED METHOD AND AERIAL IMAGERY DATASETS ON RESISC45 DATASET

Pre-trained dataset	mAP	R@1
ImageNet1k	***36.6***	***84.6***
AID	36.1	84.0

Our experiment demonstrated that feature space of visual search algorithm is enhanced by pre-training CNN backbones using unsupervised training method with a sufficient to large amount of bird eye view images. This shed light on how to effectively utilize an abundant source of aerial and satellite images without labeling.

D. Vision Transformer as extractor backbone

In order to understand the impact of utilizing recent state-of-the-art DNN architecture as the backbone extractor, we carried on our next experiment. We purposely chose to deploy Vision Transformers (ViT) [27] because it is the best performance DNN in many tasks at the present.

In this experiment, we only changed the DNN's architecture, while the other components were kept the same. Specifically, we deployed the ResNet50 and variants of ViT models pre-trained on ImageNet1k dataset using DINO self-supervised training method as the backbone extractor. We intentionally chose two small variances (having 21M parameters, comparable to 23M in ResNet50) in ViT family, the difference between the two are input patch size (16×16 and 8×8), denoting

as ViT-S/16 and ViT-S/8 respectively. The results in term of performance are reported in tables IX to XI.

TABLE IX
PERFOMANCES OF DIFFERENT PRE-TRAINED DNNS' ARCHITECTURE ON UC MERCED LAND USE DATASET

Backbone architecture	Pre-trained dataset	Pre-trained method	mAP	R@1
ResNet50	ImageNet1k	unsupervised	58.9	94.7
ViT-S/16	ImageNet1k	unsupervised	63.3	***95.7***
ViT-S/8	ImageNet1k	unsupervised	***67.0***	95.4

TABLE X
PERFOMANCES OF DIFFERENT PRE-TRAINED DNNS' ARCHITECTURE ON AID DATASET

Backbone architecture	Pre-trained dataset	Pre-trained method	mAP	R@1
ResNet50	ImageNet1k	unsupervised	46.7	88.6
ViT-S/16	ImageNet1k	unsupervised	49.8	90.2
ViT-S/8	ImageNet1k	unsupervised	***53.7***	***91.7***

TABLE XI
PERFOMANCES OF DIFFERENT PRE-TRAINED DNNS' ARCHITECTURE ON RESISC45 DATASET

Backbone architecture	Pre-trained dataset	Pre-trained method	mAP	R@1
ResNet50	ImageNet1k	unsupervised	36.6	84.6
ViT-S/16	ImageNet1k	unsupervised	39.7	86.9
ViT-S/8	ImageNet1k	unsupervised	***43.0***	***88.8***

ResNet50: Residual Network with depth of 50 layers.
ViT-S/16: Vision Transformer small, 16-patch variance.
ViT-S/8: Vision Transformer small, 8-patch variance.

From the results, we can clearly observe that by simply replacing ResNet50 backbone with a ViT, we can obtain a significant improvement in terms of performance. This implies that one may consider utilizing better DNN architectures available when wanting to enhance the visual search system's performance.

E. Qualitative results

In this section, we would like to present some qualitative search results from our zero-shot visual search system. These search results, shown in Figure 3, are obtained from our system equipped with a ViT-S/8 as the backbone extractor, which had been pre-trained on ImageNet1k dataset using self-supervised method DINO. Visually the results re-confirmed our observations in this study.

V. ABLATION STUDY

Besides the experiments presented above, we also carried out an ablation study on the visual search system, in which we studied the role of feature dimensionality reduction step, e.g., Principal Component Analysis (PCA) [26], Singular Value Decomposition (SVD) [24], k-reciprocal encoding [25]. We discovered an interesting phenomena that needs to be cautiously considered when building a robust and efficient zero-shot visual search system over billions satellite/aerial images.

A. Impact of removing feature dimensionality reduction step

After the feature extraction procedure, one of the post processing steps is to reduce the dimensionality of feature vectors before indexing. There are many methods for feature dimensionality reduction, but for the sake of brevity we only discuss PCA. However, these methods suffer from low scalability since they need to be trained on extracted feature vectors of the gallery images in the database. This process may be trivial for small-medium datasets, yet can be a potential issue when dealing with huge amounts of data, i.e., building a visual search system over billions airborne images. Therefore, in this experiment we examined the influence of removing feature dimensionality reduction step from the system.

TABLE XII
PERFORMANCES OF VISUAL SEARCH SYSTEM WITH & WITHOUT PCA ON AID DATASET

Backbone architecture$^{(**)}$	Dimension Reduction method	mAP	mAP drop ↓	R@1	R@1 drop ↓
ResNet50	PCA	46.7	5.1 ↓	88.6	0.6 ↓
	None	41.6		88.0	
ViT-S/16	PCA	49.8	4.5 ↓	90.2	0.3 ↓
	None	45.3		88.9	
ViT-S/8	PCA	53.7	4.8 ↓	91.7	1.1 ↓
	None	48.9		90.6	

$^{(**)}$: All backbones had been pre-trained on AID dataset using DINO self-supervised method.

From table XII, we can conclude that the performance of the visual search system dropped by a considerable margin in term of mean Average Precision (mAP) and a small margin for Recall at 1 (R@1) when removing dimension reduction method, specifically PCA in our experiment. This demonstrates the necessity of deploying dimension reduction module in the task where high accuracy is required.

VI. CONCLUSIONS AND FUTURE WORK

In this paper we presented two ideas that could help improve the performance of a zero-shot visual search system applied to satellite or aerial imagery: firstly, the importance of pre-training the model using aerial and satellite imagery and secondly, using a new network architecture of deep learning algorithms, i.e., family of ViT, for better feature vector learning.

As future work, we are envisaging to explore more deeply the capacity of ViT for embedding learning in an unsupervised strategy. We believe that applying such strategy to Visual Search will open new opportunities for better using the continuously increasing satellite and aerial data volumes.

REFERENCES

[1] M. Tarasiou and S. Zafeiriou, "DeepSatData: Building large scale datasets of satellite images for training machine learning models," CoRR, abs/2104.13824, 2021.

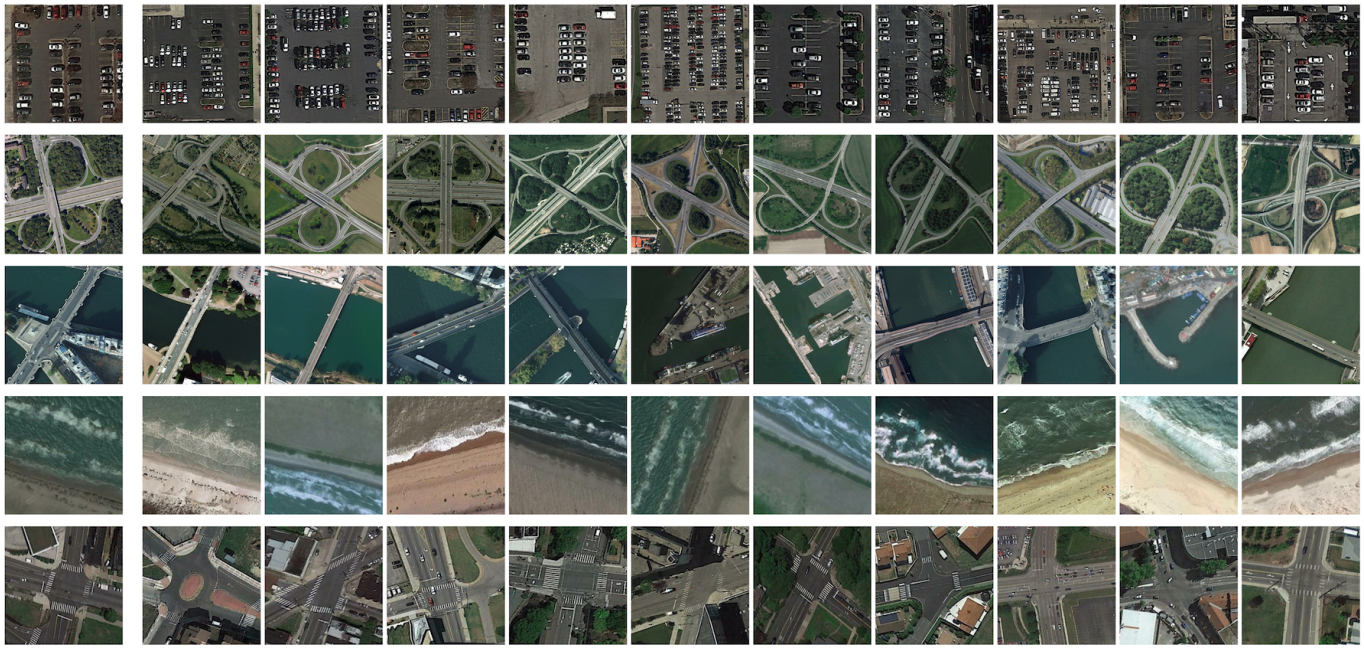

Figure 3. Example of search results. The left most image is the query image, and the rest of the row shows the top ten search results of our system.

[2] S.Bell and K. Bala, “Learning visual similarity for product design with convolutional neural networks,” ACM Trans. Graph., Vol. 34, No. 4, pp 1–10, 2015

[3] K. Sohn, “Improved Deep Metric Learning with Multi-class N-pair Loss Objective,” NIPS, 2016

[4] R. Keisler, S. Skillman, S. Gonnabathula, J. Poehnelt, X. Rudelis, and M. Warren, “Visual search over billions of aerial and satellite images,” Comput. Vis. Image Underst. vol. 187, Issue C, pp 1-6, 2019.

[5] The SpaceNet Partners, “SpaceNet5: Automated Road Network Extraction and Route Travel Time Estimation from Satellite Imagery, ,” https://spacenet.ai/sn5-challenge/, Accessed May 5th 2022

[6] J. Johnson, M. Douze, and H. Jégou, “Billion-scale similarity search with GPUs,” IEEE Transactions on Big Data, vol. 7, No. 3, pp. 535-547, 2019

[7] C. Wengert, M. Douze, and H. Jégou, “Bag-of-colors for improved image search,” In ACM Multimedia, pp. 1437–1440, 2011

[8] M. Park, J. Jin, and L. Wilson, “Fast content-based image retrieval using quasi-gabor filter and reduction of image feature dimension,“ Fifth IEEE Southwest Symposium on Image Analysis and Interpretation, 2002

[9] R. Arandjelovic and A. Zisserman, “Three things everyone should know to improve object retrieval,“ CVPR, 2012

[10] A. Krizhevsky and G. Hinton, “Using very deep autoencoders for content-based image retrieval,” in Proceedings of the European Symposium of Artifical Neural Networks (ESANN), 2011

[11] J. Ng, F. Yang, and L. Davis, “Exploiting local features from deep networks for image retrieval,” CVPR Workshops, 2015

[12] G. Tolias, T. Jenicek, and O. Chum, “Learning and aggregating deep local descriptors for instance-level recognition,” ECCV, 2020

[13] D. Yi, Z. Lei, S. Liao, and S. Z. Li, ”Deep metric learning for person re-identification,” 22nd International Conference on Pattern Recognition, 2014

[14] W. Ge, W. Huang, D. Dong, and M.R. Scott, “Deep metric learning with hierarchical triplet loss,” ECCV, 2018

[15] M. Everingham, S.M.A. Eslami, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman, “The Pascal visual object classes challenge: A retrospective,” International Journal of Computer Vision, vol. 111, no. 1, pp. 98-136, 2015

[16] H. Jégou, M. Douze, and C. Schmid, “Product quantization for nearest neighbor search,” In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 1, pp. 117-128, 2011

[17] Y.Yang and S. Newsam, “Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification,” ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS), 2010

[18] G.-S. Xia, J.Hu, F. Hu, and B. Shi, “AID: A Benchmark Dataset for Performance Evaluation of Aerial Scene Classification,” In IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 7, pp. 3965-3981, 2017

[19] G. Cheng, J. Han, and X. Lu, “Remote Sensing Image Scene Classification: Benchmark and State of the Art,” In Proceedings of the IEEE, vol. 105, no. 10, pp. 1865-1883, 2017

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. FeiFei, “Imagenet: A large-scale hierarchical image database,” CVPR, 2009

[21] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, ”Places: A 10 million image database for scene recognition,” In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 6, pp. 1452-1464, 2018

[22] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” CVPR, 2016.

[23] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, ”Emerging Properties in Self-Supervised Vision Transformers,” ICCV, 2021.

[24] G. H. Golub and C. Reinsch, ”Singular Value Decomposition and Least Squares Solutions,” In: Bauer, F.L. (eds) Linear Algebra. Handbook for Automatic Computation, vol 2. Springer, 1971, pp.134-151.

[25] Z. Zhong, L. Zheng, D. Cao, and S. Li, ”Re-ranking Person Re-identification with k-reciprocal Encoding,” CVPR, 2017.

[26] K. Pearson, ”LIII. On lines and planes of closest fit to systems of points in space,” The London, Edinburgh, and Dublin philosophical magazine and journal of science, Series 6, vol. 2, Issue 11, pp.559-572, 1901.

[27] A. Dosovitskiy, et al., ”An image is worth 16x16 words: Transformers for image recognition at scale,” ICLR, 2021.

[28] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. ”Momentum Contrast for Unsupervised Visual Representation Learning,” CVPR, 2020.

[29] X. Chen, H. Fan, R. Girshick, and K. He. ”Improved Baselines with Momentum Contrastive Learning,” CoRR abs/2003.04297, 2020.

[30] X. Chen, S. Xie, and K. He. ”An Empirical Study of Training Self-Supervised Vision Transformers,” ICCV, 2021.

[31] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. ”A Simple Framework for Contrastive Learning of Visual Representations,” ICML, 2020.

[32] J.B. Grill, et al., ”Bootstrap your own latent: A new approach to self-supervised Learning,” NeurIPS, 2020.

[33] X. Chen and K. He. ”Exploring Simple Siamese Representation Learning,” CVPR, 2021.

 ISBN: 978-1-61208-983-6