Sample4Geo: Hard Negative Sampling For Cross-View Geo-Localisation

Fabian Deuser Konrad Habel Norbert Oswald

Institute for Distributed Intelligent Systems
University of the Bundeswehr Munich
Munich, Germany
{fabian.deuser, konrad.habel, norbert.oswald}@unibw.de

Abstract

Cross-View Geo-Localisation is still a challenging task where additional modules, specific pre-processing or zooming strategies are necessary to determine accurate positions of images. Since different views have different geometries, pre-processing like polar transformation helps to merge them. However, this results in distorted images which then have to be rectified. Adding hard negatives to the training batch could improve the overall performance but with the default loss functions in geo-localisation it is difficult to include them. In this work, we present a simplified but effective architecture based on contrastive learning with symmetric InfoNCE loss that outperforms current state-of-the-art results. Our framework consists of a narrow training pipeline that eliminates the need of using aggregation modules, avoids further pre-processing steps and even increases the capacity of generalisation of the model to unknown regions. We introduce two types of sampling strategies for hard negatives. The first explicitly exploits geographically neighboring locations to provide a good starting point. The second leverages the visual similarity between the image embeddings in order to mine hard negative samples. Our work shows excellent performance on common cross-view datasets like CVUSA, CVACT, University-1652 and VIGOR. A comparison between cross-area and same-area settings demonstrate the good generalisation capability of our model.

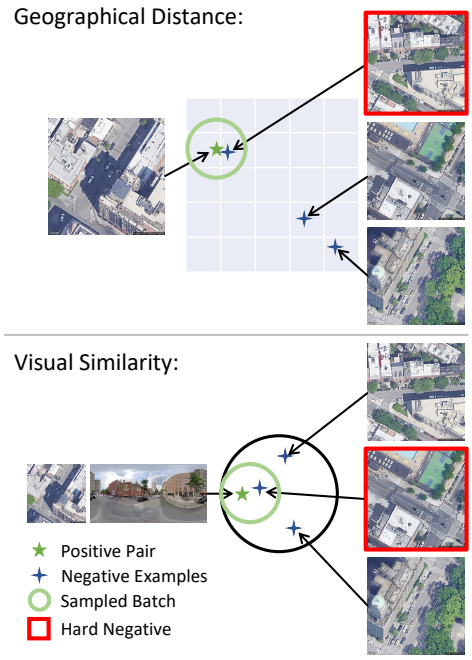

Figure 1: Our two sampling strategies. The first is based on the geographical distance between the satellite images. The second uses the cosine similarity between the street-view and satellite view embeddings to find hard negatives within a margin.

1. Introduction

Determining a geo-location from images without metadata is one of the puzzles yet to be solved in the computer vision community. Solving this problem can help in areas such as agriculture and automotive. For example, a robotic agent in agriculture for spraying fertiliser requires a high precision location. This can be achieved with a real-time kinematic (RTK) GPS, but these sensors are expensive and short-time signal outages can obstruct workflows. Therefore aerial image based localisation in these highly repetitive environments can further enhance positioning [4]. Another challenge in cities is the so-called urban canyon effect, which blocks signals such as GPS or reduces their accuracy. Especially in large cities, GPS signals are noisy due to tall buildings. The evaluation of 250,000 driving hours in New York City traffic by Brosh et al. showed an error of 10 me-

ters in 40% of GPS signals. Therefore a computer vision solution based on image retrieval was proposed [2] to correct these signals.

While classic approaches sought to do this with visual clues such as sun position and the resulting shadows [7, 19, 17] or weather [16, 15], current approaches are increasingly focusing on image retrieval based on deep learning [1, 20, 21]. In cross-view image retrieval [37, 38, 44, 47], different views of an image, e.g. a ground image and a satellite image, have to match to determine the searched position. The ground image queries are compared with a database of satellite images with known geographical positions.

Previous approaches mostly used CNN based methods [38, 13, 3, 30, 36, 35, 28, 29], while present research is mainly focused on the Transformer or MLP Mixer architecture as a backbone for geo-localisation [40, 43, 46, 48]. In the latter case, reasonable performance can only be achieved if the weights between the two view-specific encoders are not shared, resulting in larger models. Furthermore, the polar transformation is often utilised to bridge the geometrical gap between the views [28]. Triplet loss as a standard loss in cross-view geo-localisation uses only one negative example per batch and is prone to model collapsing when hard negatives are used [39].

In this paper, we present a weight-sharing Siamese CNN that learns class-agnostic embeddings based on the InfoNCE loss [25, 11, 5] and demonstrate the advantage of the CNN architecture for our approach. We use a technique from multimodal pre-training [26] to calculate the loss symmetrically to further aid the understanding of the different domains in the viewpoints. Hard negative examples, i.e. examples that are hard to distinguish for the model from positives examples, in deep metric learning are a key ingredient to achieve superior performance, thus we present two sampling methods. In the early epochs, we leverage GPS coordinates to contrast close geographic neighbours as a good initialisation for our second sampling. In later epochs based on a similarity metric, such as cosine similarity, we collect visually similar samples for the batch construction, to focus on hard negatives examples. To summarise our contribution in this work:

- We show the superior performance of our contrastive training with the symmetric InfoNCE loss whilst using a single encoder for both views.
- We propose GPS-Sampling as a task-specific sampling technique for hard negatives at the beginning of the training.
- We present Dynamic Similarity Sampling (DSS) to select hard negatives during training based on the cosine similarity between street and satellite views.
- Our framework consists of a simple training pipeline that eliminates the need of special aggregation modules or complex pre-processing steps, whilst outperforming current approaches in performance and generalisation ability.

2. Related Work

One of the first works by Workman et al. [38] showed that CNN extracted features are far superior to hand-crafted features. In their work, they also presented CVUSA which is nowadays one of the main benchmarks for geo-localisation. For training their two-stream CNN as a Siamese Network [6, 31] a simple L2 target was used to reduce the distance between view specific features. In following work Zhai et al. [41] used noisy semantic labels for the aerial view and minimised the distance between these labels and a learned transformation for the street-view. In particular, this should better reflect the common semantic layout within the views. Vo et al. [34] introduced the soft-margin triplet loss as a standard for cross-view geo-localisation. The Triplet loss decreases the distance of a positive example to an anchor and increases the distance to a negative example. Subsequent work from Hu et al. [13] included the NetVLAD-layer [1] in their architecture to further enhance the global description generation. The outputed feature maps are aggregated based on a differential clustering instead of a simple average or max pooling. Because of the slow convergence of the soft-margin triplet loss they proposed a weighted soft-margin ranking loss where the distance between a positive and a negative sample is additionally scaled.

Initially attention was also paid to the orientation within the images, as in datasets such as CVUSA [38] the alignment is known a priori and can serve as an additional feature, as Liu et al. [21] showed. To help the network better exploit directions, the number of input channels was expanded and colour-coded orientations, from top to bottom and left to right, were used for both street-view and satellite-view. Furthermore, they introduced CVACT as a dataset.

In several subsequent publications [34, 3, 29, 14], the orientation was further used to create an auxiliary objective. The street-view image is shifted as an augmentation and the degree of shift has to be predicted.

Since the geometry between the satellite image and the street-view differs, Shi et al. [28] applied polar transformation to the satellite image. The resulting stretched image is similar to the domain of the street-view image, this is still a common pre-processing technique today. In addition, they extended their network with a Spatial-Aware Feature Aggregation (SAFA) module to pool important features from the feature maps in a learnable fashion.

The polar transformation has the disadvantage that the distortion injects disturbances into the image. Toker et

al. [32] developed an approach that learns to remove these disturbances with the help of a GANs [10]. The street-view image was used as ground truth for the discriminator of the GAN. At inference time, the generator and discriminator were no longer used and only the latent representation is used to find similarities to the street-view.

Yang et al. [40] introduced the use of a ResNet Backbone in combination with a Transformer for that task. The feature maps output by the CNN are provided with a positional encoding and then entered into the Transformer.

Because the previous mentioned benchmarks are slowly saturating, a more challenging dataset named VIGOR was created by Zhu et al. [47]. While CVUSA, CVACT and VIGOR only use street-view to satellite images Zheng et al. created University-1652 [44] with UAV images instead of street-views.

Previous approaches commit to a single step of prediction which can not be corrected, therefore TransGeo [46] and SIRNet [24] have emphasised additional refinement. In SIRNet, additional refinement modules are added to the CNN backbone. A Softmax-based decision process uses a variable number of modules, up to four. With TransGeo, an additional zoom step is performed using the attention map in its Transformer architecture. This results in smaller objects being viewed at a higher resolution. To increase the generalisation of their approach, TransGeo also uses Adaptive Sharpness-Aware Minimisation (ASAM) [18] to smooth the loss surface. ASAM significantly slows down the training process, because for all training data an additional forward pass through the network is needed. Since the geometric perspective is strongly shifted by the different viewing angles, GeoDTR [42] tries to extract additional geometric properties with a Transformer-based extractor. First, features are generated independently using a CNN view. However, since ground truth labels for these geometric features are missing, an additional counterfactual loss is used to contrast the output features of the CNN with those of the Geometric Extractor to ensure dissimilarity. A triplet loss is then calculated between the features of the view-dependent geometric extractors. An alternative architecture, namely SAIG-D, by Zhu et al. [48] utilises a MLP-Mixer [33]. They replaced the patch stem with a convolutional stem to support local feature learning. In addition they propose a feature aggregation module of its own, based on two fully connected layers with a GELU activation function and a following down projection. Similar to TransGeo they utilised Sharpness-aware minimisation (SAM) [9] to further smooth the loss surface.

3. Methodology

Deep metric learning for cross-view geo-localisation uses Triplet loss by default [27], with various extensions such as soft-margin triplet loss [34] or weighted soft-margin triplet loss [13]. Triplet loss aims to decrease the distance between a positive example and an anchor, while increasing the distance between a negative example and the anchor. To use Triplet loss, suitable triplets must be sampled beforehand. Based on this formulation, only one positive and the anchor is always contrasted with one negative example. Arandjelovic et al. [1] already showed the advantage of using two negatives as quadruple loss instead of a triplet. Without the aforementioned extensions hard negatives in the Triplet loss result in a collapsed model [39]. With this in mind we designed our framework to leverage multiple hard negatives.

3.1. Symmetric InfoNCE Loss

An alternative for contrastive learning that exploits all available negatives in the batch, as opposed to triplet loss, is the so-called InfoNCE [25, 26] or NTXent loss [11, 5], see equation 1.

$$\mathcal{L}(q, R)_{\text{InfoNCE}} = -\log \frac{\exp(q \cdot r_+/\tau))}{\sum_{i=0}^{R} \exp(q \cdot r_i/\tau))} \tag{1}$$

q denotes an encoded street-view, the so-called query, and R is a set of encoded satellite images called references. Only one positive r_i, namely r_+ matches to q. The InfoNCE loss uses the dot-product to calculate the similarity between query and reference images and is low when the query and the positive match are similar, and high when the negative r_i are dissimilar to q. As loss function for the similarity between the views the cross-entropy is calculated. The temperature parameter τ is a hyperparameter that can either be learned [26] or set to a static value. So far, InfoNCE loss has mostly been used in a non-symmetric fashion for unsupervised representation learning for images [11, 5]. A symmetric formulation showed to be useful in multi-modal pre-training [26] to bridge the gap between the modalities. Therefore we utilise this loss function in the same symmetric fashion to leverage the flow of information in both directions: satellite-view to street-view and vice versa. In the InfoNCE loss, a positive example is always contrasted with N-1 negative examples, where N denotes the batch size, thus delimiting many examples at once. But in cases where there are several positive examples, such as University-1652, this requires a custom sampler to prevent multiple positives for the same ground truth label in one batch. We provide an ablation study of the importance of symmetry in the InfoNCE loss and a comparison to the triplet loss in our supplementary material.

3.2. Model Architecture

One of our main goals in this paper is to use an off-the-shelf architecture that does not require any special adaptations to the imagery in order to solve the geo-localisation problem in a meaningful way. Since recent publications

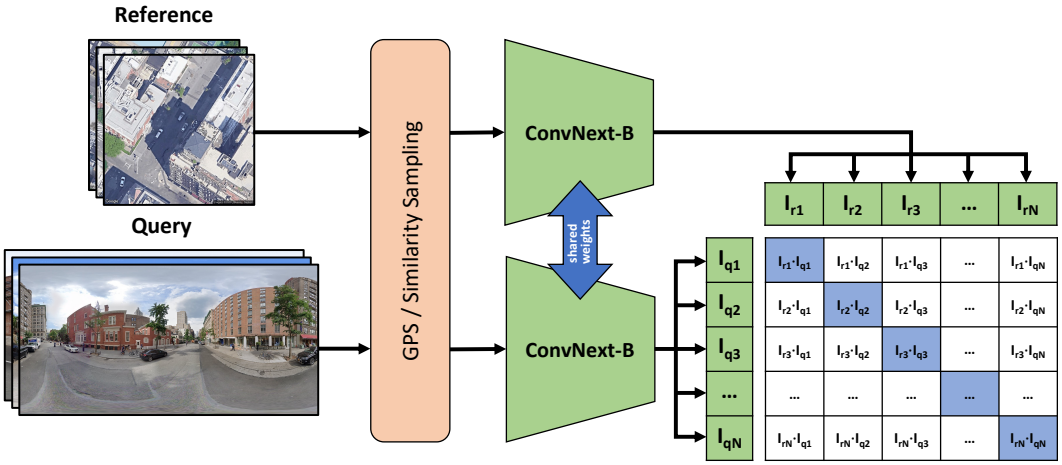

Figure 2: **Architecture overview of our approach.** We use an off-the-shelf ConvNeXt-B and mine hard negatives based on our GPS and visual similarity sampling. The InfoNCE loss is used in a symmetric fashion to learn discriminative features in both view directions.

have increasingly used Vision-Transformer [46, 40, 43] or MLP mixer architectures [48], we compare two architectures. For this purpose we used ConvNeXt [22] as state-of-the-art CNN and a Vision Transformer [8]. We decide in favour of ConvNeXt on the basis of the results in section 4.3.1. In line with previous approaches we use a Siamese Network as depicted in figure 2. We utilise mean-pooled feature vectors without any attention-pooling or modules for refinement. The ConvNeXt is a modernised variant of the ResNet [12] and uses many improvement such as a patchified stem, higher kernel sizes, GELU activation functions, LayerNorm and depthwise convolutions. While current work does not use weight-sharing due to the different view-domains, we emphasise the use of a weight-shared ConvNeXt as a single encoder for both views.

3.3. Near Neighbour Sampling Based on GPS

In most of the cases hard examples can not be selected before training, since the model is not fitted to the domain of the training data. In order to bypass this drawback we leverage the geographical nature of cross-view geo-localisation. Images from the same area, or even the same street in the VIGOR dataset, share common properties like vegetation, street signs, or housing type. Since these are easy features that contrast a rural from a urban scenery, they will not contribute much to the objective function. Therefore we propose a simple GPS-based sampling strategy for the negative mining initialisation before the training even started. For CVUSA and VIGOR GPS locations are present in the dataset and near neighbours are selected based on the haversine distance. In the CVACT dataset the locations are in the UTM coordinate system, which is why we use the euclidean distance to determine the neighbours. The GPS coordinate is only used during training for the sampling strategy in the early epochs. For the University-1652 dataset this initialisation is not possible, therefore a simple shuffling is used.

3.4. Dynamic Similarity Sampling

After a certain amount of epochs, GPS-based sampling is replaced by our Dynamic Similarity Sampling (DSS). During one full inference epoch on the training data the visual distances between all samples are calculated using the cosine similarity. In order to sample future batches the top K nearest neighbours for every query image are selected. We set K smaller or equal than the batch size, to have multiple regions or cities in one batch. These K neighbours are sorted based on their similarity and then $\frac{k}{2}$ nearest samples are selected for the batch, the remaining $\frac{k}{2}$ samples are randomly selected from the remaining samples in K. The random selection process ensures enough diversity for the hard negatives since we only calculate new distances every e epochs to shorten the training process. We set our hyperparameters as follows: $k = 64$, $K = 128$ and $e = 4$. An ablation over the best choice of k can be found in our supplementary material. Before we add our k samples to the batch, a lookup is proceeded to avoid double entries within an epoch. For different settings of k we did not observe the collapsing model problem described in [39] even when $k = K$.

4. Evaluation

4.1. Experiments and Results

In our evaluation we conducted experiments on four standard benchmarks, namely CVUSA, CVACT, University-1652 and VIGOR. In the subsequent tables we

compare our approach with previous work.

4.1.1 CVUSA & CVACT

CVUSA As one of the first cross-view datasets, CVUSA [38, 41] consists of 35,532 view pairs for training and 8,884 for evaluation and is one of the standard benchmarks. The satellite with a resolution of 750×750 and street-view images with a resolution of 224×1232 are aligned based on the camera extrinsics. This alignment ensures that the geographical north is located in the upper centre of the satellite image, and the street-view image is warped accordingly. The task here is a 1 to 1 mapping, every satellite-view has one corresponding street-view image.

CVACT In the dataset from Liu et al. [21] the training and validation splits are the same size as CVUSA, but they provide an additional test split with 92802 images. Unlike CVUSA, Canberra (Australia) is used as the region in the dataset and mostly urban scenery is included. Also a higher image resolution is provided with 1200×1200 for satellite and 832×1664 for street-views. The task is still a 1 to 1 mapping and the images are aligned like in the CVUSA dataset. For both CVACT and CVUSA, the metrics are Recall@k (R@k) with $k \in \{1, 5, 10\}$ and R@1%.

Results As shown in table 1 our approach outperforms all previous work in the R@1 metric. Since the CVUSA and CVACT Val are saturated as a benchmark we also compare our work on the CVACT test set to show the generalisation ability of our model. The test split of the CVACT dataset is more difficult, because streetview images were sampled at a much higher density thus resulting in many semi-positive matches in the R@1 compared to R@5. It should also be noted that we do not use polar transformation. Approaches that do use the polar transformation usually experience a slight improvement on the benchmarks. As input image size for both datasets we use 140×768 for the street-view and 384×384 for the satellite-view.

4.1.2 University-1652

While the previous two datasets matches $360°$ street-views with satellite data, the dataset proposed by Zheng et al. [44] includes multiple drone images from a building. The training images consists of over $50k$ drone views from 701 university buildings and the task is to match the drone views to the according satellite image and vice versa. Another peculiarity of this dataset is that several buildings in the test set represent only adversary classes and there are no drone views referencing them, thus additional to the recall also the average precision (AP) is indicated. In our comparison 2 we show the capability of your approach to different domains since we have a $1 : n$ mapping. As input image size we downsample from 512×512 to 384×384 for both the UAV and the satellite-view.

4.1.3 VIGOR

For VIGOR, Zhu et al. [47] collect 90,618 satellite reference images and 105,214 street-view query images. The four cities, New York, Seattle, San Francisco and Chicago in this benchmark are used for two different split settings: same-area and cross-area. In the SAME setting, images of all cities are used in training and validation and in CROSS setting, training carried out on New York and Seattle and evaluation is done on San Francisco and Chicago to test the generalisation of the approach. Another novelty is the introduction of so-called semi-positive images. For each positive pair, there are three semi-positive sat-view neighbours which also cover regions of the street-view image. The position from which they were taken is not in the centre region of the image. Due to these three semi-positive images for each satellite image, it is difficult to achieve a high R@1 score. To determine the performance without semi-positive images, the hit rate is calculated. The hit rate is understood here as the R@1 with masked semi-positives, as these serve as a distraction. The images are provided with a resolution of 640×640 for satellite and 1024×2048 for street-views. We use 384×768 for the street-view and 384×384 for the satellite-view as input size for our model.

In the same-area setting, table 3, we already outperform current work by far and show the ability of our approach even in the harder to handle CROSS setting. The higher hit rate and the performance at Recall@5 also shows that all approaches naturally have problems to distinguish between the semi-positives and the ground truth image as they are very close to each other. A visualisation of which features in an image make it difficult to distinguish between different ground truth and semi-positive images is provided in section 4.3.4. While transfer to new, unknown cities and regions has been difficult in previous work, our model also shows outstanding performance in the cross-area setting.

4.2. Implementation Details

In our experiments the architecture is not altered and the ConvNeXt-B with $88M$ parameters is used. An ablation of different architecture sizes and types is provided in section 4.3.1. To minimise the overfitting on the training set we use label smoothing of 0.1 in the InfoNCE loss and the temperature parameter τ is a learnable parameter. As Zhang et al. [42] showed, synchronous augmentations are important in order not to disturb the positions and orientations encoded in the image. Therefore, we used synchronous horizontal flipping as well as rotation on the satellite images and shift the street-view images accordingly to preserve the

Approach	CVUSA				CVACT Val				CVACT Test			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
LPN [35]	85.79	95.38	96.98	99.41	79.99	90.63	92.56	-	-	-	-	-
SAFA† [28]	89.84	96.93	98.14	99.64	81.03	92.80	94.84	-	-	-	-	-
TransGeo [46]	94.08	98.36	99.04	99.77	84.95	94.14	95.78	98.37	-	-	-	-
GeoDTR [42]	93.76	98.47	99.22	99.85	85.43	94.81	96.11	98.26	62.96	87.35	90.70	98.61
GeoDTR†	95.43	98.86	99.34	99.86	86.21	95.44	96.72	98.77	64.52	88.59	91.96	**98.74**
SAIG-D [48]	96.08	98.72	99.22	99.86	89.21	96.07	97.04	98.74	-	-	-	-
SAIG-D† [48]	96.34	99.10	99.50	99.86	89.06	96.11	97.08	**98.89**	67.49	89.39	92.30	96.80
Ours	**98.68**	**99.68**	**99.78**	**99.87**	**90.81**	**96.74**	**97.48**	98.77	**71.51**	**92.42**	**94.45**	98.70

Table 1: **Quantitative comparison between our approach and state-of-the-art approaches on CVUSA [41] and CVACT [21].** † denotes which models are using the polar transformation for their satellite input as a pre-processing technique.

Approach	Drone2Sat		Sat2Drone	
	R@1	AP	R@1	AP
LPN [35]	75.93	79.14	86.45	74.79
SAIG-D [48]	78.85	81.62	86.45	78.48
DWDR [36]	86.41	88.41	91.30	86.02
MBF [45]	89.05	90.61	92.15	84.45
Ours	**92.65**	**93.81**	**95.14**	**91.39**

Table 2: **Quantitative comparison between our approach and state-of-the-art approaches on University-1652 [44].**

Approach	R@1	R@5	R@10	R@1%	Hit Rate
SAME					
SAFA† [28]	33.93	58.42	68.12	98.24	36.87
TransGeo [46]	61.48	87.54	91.88	99.56	73.09
SAIG-D [48]	65.23	88.08	-	**99.68**	74.11
Ours	**77.86**	**95.66**	**97.21**	99.61	**89.82**
CROSS					
SAFA† [28]	8.20	19.59	26.36	77.61	8.85
TransGeo [46]	18.99	38.24	46.91	88.94	21.21
SAIG-D [48]	33.05	55.94	-	94.64	36.71
Ours	**61.70**	**83.50**	**88.00**	**98.17**	**69.87**

Table 3: **Quantitative comparison between our approach and the current state-of-the-art on VIGOR [47].** The same split contains images from all four cities in the dataset in training and validation. The cross split contains exclusively two cities and the model is then evaluated on the unknown other two cities. Our result show a much better generalisation on unknown regions than previous approaches. † denotes which models are using the polar transformation for their satellite input as a pre-processing technique.

orientation of the satellite images. In order not to focus completely on certain regions in some images, we also use grid and coarse dropout and to further increase the generalisation we use colour jitter. Since University-1652 is a $1:n$ or $n:1$ task we use a custom sampler to prevent the occurrence of multiple images from the same class within the batch. The InfoNCE Loss treats all other examples as negatives, thus without a sampling strategy positive examples would be treated as negatives too. Each experiment is trained for 40 epochs with a batch size of 128 using the AdamW optimiser [23] with a initial learning rate of 0.001 and a cosine learning rate scheduler with a 1 epoch warm-up period.

4.3. Ablation Study

In our ablation study we dive into the design choices made in this work like the model architecture or the effectivity of our sampling. To provide additional insights we extract activation heatmaps of the satellite and street-view images of our model to visualise important regions in both views.

4.3.1 Architecture Evaluation

For a long time, within the research community of cross-view geo-localisation, CNNs were the most important building block for learning useful representations [28, 13, 41, 32, 42]. More recent publications explore other architectures such as transformers [46, 40, 43] or the MLP mixer [48]. In order to determine a proper selection for our approach, we compare two standard architectures: a Vision Transformer (ViT) [8] and a current state-of-the-art CNN [22] in table 4.

The advantage of a CNN is that it can handle different input image sizes. Vision Transformers are not capable of dealing with multiple input sizes due to their fixed size positional encoding. Another constraint of the Transformer

Approach	#Params (M)	Shared W.	R@1	R@5	R@1%	Hit Rate
SAFA [28]	$14.7 * 2$	X	33.93	58.42	98.24	36.87
TransGeo [46]	$22.4 * 2$	X	61.48	87.54	99.56	73.09
SAIG-D [48]	$15.6 * 2$	X	65.23	88.08	99.68	74.11
Ours (ViT)	$86.8 * 2$	X	69.89	92.81	99.43	83.49
Ours (Base)	$88.5 * 2$	X	75.81	94.86	99.61	88.08
Ours (Nano)	15.5	✓	70.70	91.17	94.19	99.40
Ours (Tiny)	28.5	✓	73.90	93.60	99.47	84.77
Ours (Base)	88.5	✓	77.86	95.66	99.61	89.82
Ours (Large)	197.7	✓	78.27	95.98	99.66	90.50

Table 4: **Parameter efficiency of our models compared to the current state-of-the-art.** Increasing the number of parameters brings only a marginal advantage and shows that our sampling and the symmetric loss is more important. Results are reported on the VIGOR SAME split.

architecture is its scalability in terms of quadratic memory consumption due to the attention mechanism. Previous approaches propose to use separate encoders for satellite and street-view images. Therefore we also tested ConvNeXt without weight sharing, but we achieve better results when using the same encoder for both views. In table 4, we compare models with two separate encoders for satellite and street-view with our approach that requires only one encoder for both views.

Since SAIG-D, one of the best approaches so far on datasets like VIGOR, CVUSA and CVACT, has a much smaller number of parameters a comparison is necessary. The question is how well our model performs when using the similar number of parameters. As shown in table 4 the results from our work is very consistent, a higher amount of parameters has only a marginal impact on the performance. Instead of providing two individual backbones per view we use a single encoder and show it outperforms two separate encoders, even in settings with similar model size.

4.3.2 Effectiveness of Sampling Strategies

Without any sampling strategy we still achieve competitive results when using contrastive training leveraging the InfoNCE loss. But a good sampling strategy drastically improves the performance, as shown in table 5. We also tested the sampling strategies for CVUSA and CVACT, and the results are included in our supplementary material.

Due to the presence of geographical neighbours in the VIGOR SAME split, the model benefits from this method during the whole training course. However, since locations from all four cities can occur in the SAME split, an ablation is also interesting to see whether GPS sampling has an effect on the cross area split. Those nearby locations are not present in the test set and an increase in performance with the help of GPS sampling on the cross-area split would indicate the generalisation of this sampling. As we successfully show the results are consistent in our same-area and cross-area experiments. Even if only GPS sampling is used during the training we achieve almost the same R@1 as with visual similarity sampling. When using GPS-based sampling without DSS, it can be assumed, that areas within short geographical distances are also quite similar and harder to distinguish based on visual features. The performance with a combination of both strategies is larger in the cross-area setting in comparison to the same-area setting. Based on this analysis we decided to use both sampling strategies for the reported CVUSA and CVACT results.

Sampling	R@1	R@5	R@10	R@1%	Hit Rate
SAME					
Random	65.23	91.62	95.85	99.85	78.77
GPS	74.69	92.21	94.66	99.34	83.49
DSS	76.94	95.50	97.12	**99.66**	89.31
GPS + DSS	**77.86**	**95.66**	**97.21**	99.61	**89.82**
CROSS					
Random	36.38	63.96	72.43	97.18	43.66
GPS	57.17	78.44	83.79	97.22	62.81
DSS	58.59	81.45	86.44	97.98	66.63
GPS + DSS	**61.70**	**83.50**	**88.00**	**98.17**	**69.87**

Table 5: **Quantitative comparison between our proposed sampling techniques for the VIGOR SAME split.**

Approach	R@1	R@5	R@10	R@1%
CVUSA → CVACT				
L2LTR [40]	47.55	70.58	-	91.39
L2LTR† [40]	52.58	75.81	-	93.51
GeoDTR [42]	47.79	70.52	-	92.20
GeoDTR†	53.16	75.62	-	93.80
Ours	**56.62**	**77.79**	**87.02**	**94.69**
CVACT → CVUSA				
L2LTR [40]	33.00	51.87	-	84.79
L2LTR† [40]	37.69	57.78	-	89.63
GeoDTR [42]	29.13	47.86	-	81.09
GeoDTR†	44.07	**64.66**	-	90.09
Ours	**44.95**	64.36	**72.10**	**90.65**

Table 6: **Generalisation ablation when trained on the CVUSA dataset and evaluated on CVACT and vice versa.** † denotes approaches that used the polar transformation.

4.3.3 Generalisation Capabilities

A very interesting question about different approaches is the generalisation capability, i.e. the ability of models trained on one region or scenery and then used for another region. As we have already shown in table 3 in particular, our approach generalises to unknown cities very well, but

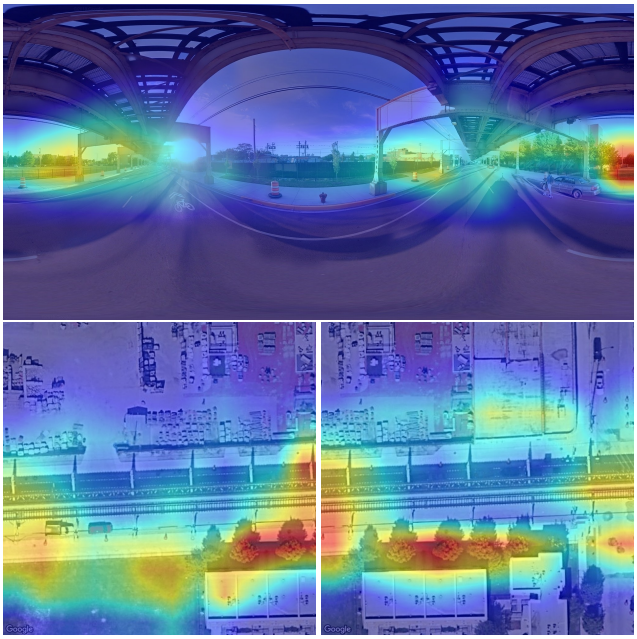

Figure 3: **Heatmap for a semi-positive hit on the VIGOR dataset.** The upper picture shows the street-view query, the bottom left side the prediction and on the right side the actual ground truth is shown. Features like trees and buildings are fundamental for the prediction and in both similar features are relevant.

not yet with nearly the same performance as in the same-area setting. For our ablation, we additionally evaluated the models trained on CVUSA and CVACT to show the transfer between these datasets. As can be seen from table 6, our approach scores well, especially when compared to approaches without polar transformation. In our supplementary material we provide further visualisations of the behaviour of our model for this setting. This shows that incorrectly recognised satellite images are usually very similar in terms of the course of the road. The model is biased towards the road course in the image and encodes it in the feature vector. This is one of the reasons why previous approaches benefit from polar transformation, which leads to a better alignment of street-view and distorted satellite scenery.

4.3.4 Visualisation

Human reasoning for cross-view geo-localisation is usually based on powerful features such as the course of a road or distinctive buildings. Therefore, it is particularly interesting to understand which regions are important for the model. In our supplementary material we present multiple examples from CVUSA. These examples show that in CVUSA streets are one of the most important features in an image. For the VIGOR dataset this is not the case. In figure 3 a semi-positive hit can be observed and activations from the surrounding landscape are more important for the prediction. However, the VIGOR dataset contains multiple semi-positives for each position. A clear prediction is harder since multiple important features, like the trees in figure 3 are present in both satellite images.

5. Conclusion

In our work, we present a simple but effective approach to solve the geo-localisation task. Our proposed model consists of a single image encoder for both satellite and ground view based on a modern CNN. This lightweight approach leverages contrastive learning by using the InfoNCE Loss as training objective. We further demonstrate, that an effective sampling strategy leads to superior results on the VIGOR, CVUSA, CVACT and University-1652 datasets. Compared to other state-of-the-art approaches, our model does not need complex pre-processing steps or any task specific aggregation modules, nor the use of loss surface smoothing like SAM or ASAM to achieve outstanding generalization performance in cross-area settings.

6. Discussion

During our work we identified several problems as well as future challenges. While the traditional datasets such as CVUSA, CVACT and VIGOR only contain 360° street-view images, University-1652 is one of the few datasets that contain a limited field of view of the drone images. However, all images are taken close to the building. This simplifies the geo-localisation task enormously in both cases as non-orientation bound features have to be matched. Another reason for the outstanding performance on CVUSA and CVACT is the perfect alignment of the position of the Street View image to the centre of the satellite image. This is fixed in VIGOR, but as well here the benchmark becomes progressively saturated. To improve the applicability, additional datasets with explicitly not aligned central positions and unknown geographical orientation between satellite and street-view image are required. In addition, the FoV should be limited to below 120 degrees, as this reflects the FoV from a current standard smartphone.

Another common characteristic of the previous datasets is the focus on urban environments. CVUSA offers a somewhat broader spectrum here, but the ground view images are obviously all taken on streets. This does not reflect the variety of scenes in the wild and as shown in our supplementary our approach focuses most times on the streets and intersections to achieve matching, especially on the CVUSA dataset. Future datasets should also include ground views not taken exclusively from roads to increase diversity, usefulness and practicability.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016.

[2] Eli Brosh, Matan Friedmann, Ilan Kadar, Lev Yitzhak Lavy, Elad Levi, Shmuel Rippa, Yair Lempert, Bruno Fernandez-Ruiz, Roei Herzig, and Trevor Darrell. Accurate visual localization for automotive applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[3] Sudong Cai, Yulan Guo, Salman Khan, Jiwei Hu, and Gongjian Wen. Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8390–8399, 2019.

[4] Nived Chebrolu, Philipp Lottes, Thomas Läbe, and Cyrill Stachniss. Robot localization based on aerial images for precision agriculture tasks in crop fields. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 1787–1793, 2019.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[6] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1, 2005.

[7] F. Cozman and E. Krotkov. Robot localization using a computer vision sextant. In *Proceedings of 1995 IEEE International Conference on Robotics and Automation*, volume 1, pages 106–111 vol.1, 1995.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[11] 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.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Sixing Hu, Mengdan Feng, Rang M. H. Nguyen, and Gim Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7258–7267, 2018.

[14] Wenmiao Hu, Yichen Zhang, Yuxuan Liang, Yifang Yin, Andrei Georgescu, An Tran, Hannes Kruppa, See-Kiong Ng, and Roger Zimmermann. Beyond geo-localization: Fine-grained orientation of street-view images by cross-view matching with satellite imagery. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 6155–6164, New York, NY, USA, 2022. Association for Computing Machinery.

[15] Nathan Jacobs, Kylia Miskell, and Robert Pless. Webcam geo-localization using aggregate light levels. In *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 132–138, 2011.

[16] Nathan Jacobs, Nathaniel Roman, and Robert Pless. Toward fully automatic geo-location and geo-orientation of static outdoor cameras. In *2008 IEEE Workshop on Applications of Computer Vision*, pages 1–6, 2008.

[17] Imran N. Junejo and Hassan Foroosh. Gps coordinates estimation and camera calibration from solar shadows. *Computer Vision and Image Understanding*, 114(9):991–1003, 2010.

[18] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pages 5905–5914. PMLR, 2021.

[19] Jean-François Lalonde, Srinivasa Narasimhan, and Alexei Efros. What do the sun and the sky tell us about the camera? *International Journal of Computer Vision*, 88:24–51, 05 2010.

[20] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5007–5015, 2015.

[21] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5624–5633, 2019.

[22] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[24] Xiufan Lu, Siqi Luo, and Yingying Zhu. It's okay to be wrong: Cross-view geo-localization with step-adaptive iterative refinement. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

[25] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[26] 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.

[27] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[28] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. *Spatial-Aware Feature Aggregation for Cross-View Image Based Geo-Localization*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[29] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4064–4072, 2020.

[30] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal feature transport for cross-view image geo-localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11990–11997, 2020.

[31] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[32] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2021.

[33] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

[34] Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 494–509. Springer, 2016.

[35] Tingyu Wang, Zhedong Zheng, Chenggang Yan, Jiyong Zhang, Yaoqi Sun, Bolun Zheng, and Yi Yang. Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2):867–879, 2021.

[36] Ting Wang, Zhedong Zheng, Zunjie Zhu, Yu-Fei Gao, Yi Yang, and Chenggang Yan. Learning cross-view geo-localization embeddings via dynamic weighted decorrelation regularization. *ArXiv*, abs/2211.05296, 2022.

[37] Scott Workman and Nathan Jacobs. On the location dependence of convolutional neural network features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2015.

[38] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3961–3969, 2015.

[39] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE international conference on computer vision*, pages 2840–2848, 2017.

[40] Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29009–29020. Curran Associates, Inc., 2021.

[41] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 867–875, 2017.

[42] Xiaohan Zhang, Xingyu Li, Waqas Sultani, Yi Zhou, and Safwan Wshah. Cross-view geo-localization via learning disentangled geometric layout correspondence. *arXiv preprint arXiv:2212.04074*, 2022.

[43] Jianwei Zhao, Qiang Zhai, Rui Huang, and Hong Cheng. Mutual generative transformer learning for cross-view geo-localization. *arXiv preprint arXiv:2203.09135*, 2022.

[44] Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM international conference on Multimedia*, pages 1395–1403, 2020.

[45] Runzhe Zhu, Mingze Yang, Ling Yin, Fei Wu, and Yuncheng Yang. Uav's status is worth considering: A fusion representations matching method for geo-localization. *Sensors*, 23(2):720, Jan 2023.

[46] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1162–1171, 2022.

[47] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021.

[48] Yingying Zhu, Hongji Yang, Yuxin Lu, and Qiang Huang. Simple, effective and general: A new backbone for cross-view image geo-localization. *arXiv preprint arXiv:2302.01572*, 2023.