

Scaling Image Geo-Localization to Continent Level

Supplementary material

In this appendix, we first discuss the societal impact of our work and later present additional ablation studies that motivate the design decisions of our approach, shedding light into its inner workings. We then evaluate the image embeddings learned by our model, study the data distribution of our dataset, and show extensive qualitative examples.

A Societal Impact

This work addresses the research question of geolocating images over very large geographical areas (countries, continents) without the need for rough priors, like GPS. A solution to this problem will undoubtedly raise concerns about privacy, surveillance, discrimination, and personal safety. While these concerns apply to any other works in the field, which we build and improve upon, they grow larger by scaling up the size of the database. As the potential of misuse is significant, we offer this as a proof-of-concept only, and will refrain from releasing model weights to the public. We note, however, that the same risks apply to any current visual place recognition (VPR) systems, which also typically perform best (albeit at a prohibitive cost).

Another potential misuse is in the training set, which covers a vast area of public places (streets, houses), including humans, animals and cars. Our data is anonymized, blurring faces and license plates in order to prevent leaking this information to the model.

On the other hand, we highlight the potential capabilities of such a system, which could enable novel applications in autonomous systems and augmented or virtual reality. It could also enable the creation of much larger 3D vision datasets by helping pose arbitrary images (in conjunction with more traditional solutions like Structure-from-Motion), a process that is currently very time-consuming and typically rejects a very large fraction of images. It also helps push the envelope on the understanding of geospatial patterns from multiple modalities (ground and aerial images). Finally, it offers very significant compute savings over retrieval-based systems (VPR), which are the state of the art in visual geolocalization.

B Additional Evaluations

B.1 Ablations

Ablation—Losses (Table 7). We study the impact of the loss terms under different evaluation settings: (a) ground-to-aerial cross-view retrieval, (b) cell classification, and (c) our hybrid cell prototypes. All terms significantly contribute to the accuracy of our approach. Removing the edges between ground and aerial embeddings harms cross-view localization performance most because they get only indirectly constrained through the prototypes. The most important edge is between the ground images and cell prototypes, behaving as a global, spatial memory. However, this memory is limited by the actual density of samples in the cell, which acts as a bottleneck. Aligning ground images jointly to aerial embeddings and prototypes yields large improvements, especially on prototype retrieval. Empirically, we observed that this reduces overfitting between ground images and prototypes, which

Table 7: **Loss terms.** We study the impact of the loss components between ground-level (G), aerial (A) embeddings, and cell prototypes (P), on recall@ K @200m on BEDENL, under different evaluation settings (a-c). We highlight the **best** and **second best**, per column. The bottom row is our final model.

Terms			(a) Cross-view			(b) Prototypes			(c) Hybrid		
G-A	G-P	A-P	$K=1$	$K=5$	$K=100$	$K=1$	$K=5$	$K=100$	$K=1$	$K=5$	$K=100$
✓			39.2	53.6	74.1	N/A	N/A	N/A	39.2	53.6	74.1
	✓		N/A	N/A	N/A	47.2	59.3	74.5	47.2	59.3	74.5
✓	✓		42.3	56.0	75.2	<u>56.4</u>	<u>67.8</u>	<u>81.0</u>	<u>58.3</u>	<u>70.0</u>	<u>84.6</u>
✓		✓	<u>46.4</u>	<u>59.8</u>	<u>78.6</u>	40.7	54.6	73.9	47.0	60.4	79.6
	✓	✓	15.8	26.4	51.6	47.7	60.5	77.0	47.7	60.5	77.0
✓	✓	✓	49.7	63.3	81.0	57.1	68.6	81.8	60.3	71.6	85.6

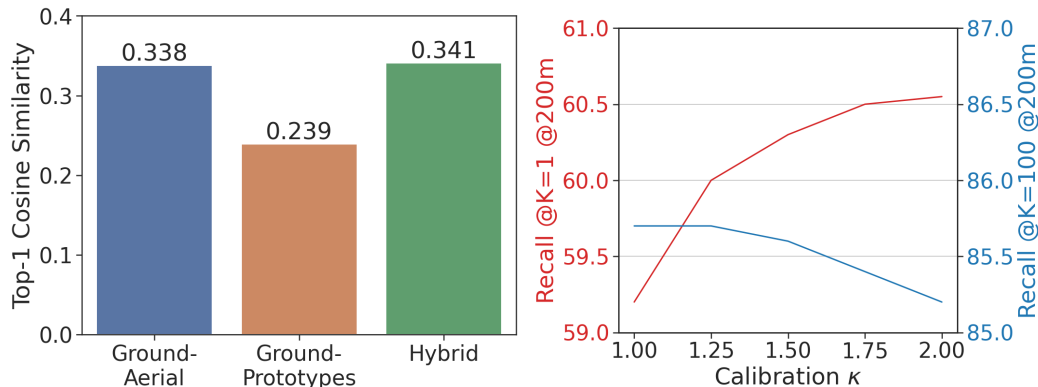


Figure 6: **Ablation of the calibration factor.** **Left:** Average top-1 cosine similarities. **Right:** Impact of the calibration factor κ on top-1 (red) and top-100 (blue) recall at 200m.

Table 8: **Generalization to gaps in the map.** We expand the database with aerial-only embeddings on cells not covered by ground-level images. We report recall both (a) in areas where training data, and thus cell prototypes, are available, and (b) where cell prototypes are not available, a common failure case of approaches that rely only on ground-level images, which we bridge via aerial embeddings. Our hybrid method is able to generalize well to unseen areas in the map.

Method	(a) BEDENL			(b) BEDENL gaps		
	Recall @ K @ 200m			Recall @ K @ 200m		
	$K=1$	$K=5$	$K=100$	$K=1$	$K=5$	$K=100$
SALAD [23]-Aerial	34.2	47.1	67.2	25.8	36.9	57.0
Ours (full)	55.1	67.5	83.1	35.6	46.9	64.9

398 is more common in cells where the data is sparser. The aerial embeddings smooth the feature space
399 and thus reduce the dependency on sampling density, countering overfitting.

400 If the edge between aerial embeddings and cell prototypes is missing, this introduces an asymmetry
401 whereby global constraints come only from the ground-level embeddings, significantly harming
402 performance in cross-view retrieval. The model benefits from regularized aerial embeddings which
403 better constrain the space, serving as a proxy for hard negative mining between ground and aerial
404 images. Combining all loss terms strikes a strong balance between cross-view and prototype retrieval
405 performance. One downside of the full model is its requirement to compute the full similarity matrix
406 to all cell prototypes twice (once for the ground images and once for aerial) during training.

407 **Ablation—Calibration prototypes and aerial embeddings (Figure 6).** One important hyperpa-
408 rameter in our study is the calibration factor we use when combining the aerial embeddings with
409 the prototypes, *i.e.*, for our ‘hybrid’ model, at inference time. One key insight here is that the actual
410 similarity scales are different: queries show about 1.5 times larger similarity to the aerial embeddings,
411 both on the training and test sets. The left panel in Fig. 6 illustrates this observation. We partially
412 attribute this to the different granularity between aerials (L16) and cell prototypes (L15), as the
413 coarser granularity of the prototypes means that they need to average over larger areas and thus more
414 visual content, yielding lower similarity scores to each query. On the right panel we show recall
415 metrics for different values of the calibration factor κ —recall@200m for both the top-1 and top-100
416 candidates. The best trade-off in recall is observed at approximately $\kappa = 1.5$, which corresponds to
417 the offset factor between the similarities. This supports our design choice to select κ based on this
418 delta between the two similarities. Overall, recall performance demonstrates robustness to changes in
419 the calibration factor.

420 **Ablation—Holes in training data (Table 8).** A benefit of aerial embeddings over prototypes is their
421 ability to generalize to unseen cells. In Table 3 (in the main paper) we discussed how this helps
422 the model generalize to different countries. In practice, we are more interested in the generalization
423 capabilities to *gaps* in-between prototypes, *i.e.*, smaller regions or ‘holes’ in our database where we
424 might have aerial coverage but not enough ground-level images to build cell prototypes.

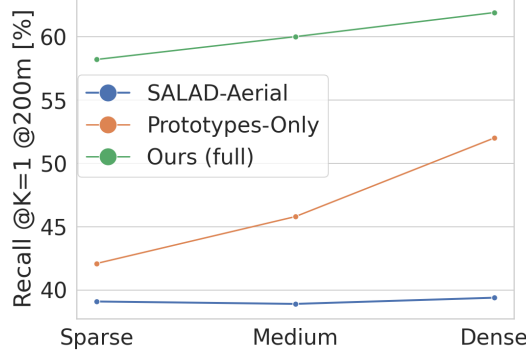


Figure 7: **Ablation of localization performance by data density on BEDENL.** We group cells by the amount of panos contained, from ‘sparse’ to ‘dense’. Cross-View Retrieval methods are robust to data density, while prototypes ‘remember’ places seen during training, and are thus inherently less robust.

In the main paper, we evaluate only on cells where we can train the model, omitting cells for which we have enough *test* data (from 2023) but not enough *training* data (from other years) — this allows us to provide a fair comparison between cross-view and prototype-based retrieval, as they use the same subset of the data. In this experiment we aim to increase the test coverage *beyond* that of the training set. We collect 166k test images that are at least 200m from their closest prototype center in BEDENL, and from the year reserved for the test set (2023). We extend the aerial database to contain these areas, increasing the database size from 4.8M to 8.5M $L=16$ cells, and evaluate our best model against SALAD [23]-Aerial. The results in Table 8 show that our method can localize images in these areas, although with a significant performance drop. This captures the use-case where we are missing cell prototypes and must fall back to pure cross-view retrieval. Note that compared to the results in Table 1 (in the main paper), the in-domain performance (*i.e.*, for cells with a prototype) also drops slightly because of *almost doubling* the size of the database.

Ablation—Impact of data density (Figure 7). Cross-view retrieval is more robust to variations in data density, as it is less impacted by a lack of covisibility. Yet prototypes are superior in very densely sampled cells, e.g. in city centers. Our hybrid method is able to combine these properties by utilizing both bird’s-eye and ground-level cues.

B.2 Cross-Area Visual Place Recognition

Setup: We perform an additional study on classic image retrieval. We evaluate our model trained on BEDENL in a country not included in our training set, Portugal. This dataset consists of 18.8M images spaced 40 meters apart, similar to the distribution of the BEDENL training split. We evaluate on 197k test images from a different year. This benchmark evaluates the strength of learned image embeddings to large viewpoint and seasonal changes. **Baselines:** Unlike in the paper, we here use the *official weights of SALAD* [23]. This model is trained on the smaller Street-View dataset GSV-Cities [79], which contains images from major metros around the globe (including Lisbon, which is part of this test set). We further add our own cross-view retrieval baselines to this benchmark (database size 1.2M aerial images). Note that the features produced by SALAD are 4x larger than ours — too large in fact to run over UK+IE, which we used in the main paper. **Results:** We report the benchmark results in Table 9. Notably our learned embeddings outperform SALAD [76] trained on GSV-Cities [79]. This can be explained by the extensive amount of rural images in the test set, a domain not covered by GSV-cities [79], which pose a major challenge in country-wide geo-localization. The image embeddings learned from prototypes only generalize equally well to new domains, which is in contrast to the evaluation in-domain (*i.e.*, on BEDENL). Overall, image-based retrieval, despite the large viewpoint changes, still generalizes much better than cross-view retrieval to new areas. Notably, the gap between VPR and cross-view retrieval is significantly larger than in training areas (*i.e.*, for BEDENL, Table 1), as the model has to overcome both spatial, temporal, and viewpoint domain gaps.

Table 9: **Cross-Area Visual Place Recognition in Portugal.** We perform an additional experiment that compares visual place recognition between ground view-images to cross view retrieval, both with a large spatial domain gap, in a country not covered by the training set, Portugal. Visual place recognition generalizes significantly better. Our model outperforms the popular VPR baseline SALAD [23], also trained on StreetView imagery.

Evaluation	Method	Training	Portugal			dim.
			Recall@ K @200m			
			$K=1$	$K=5$	$K=100$	
ground retrieval	SALAD – official weights [23]	GSV-Cities	27.3	36.2	53.9	8448
	Ours (prototypes-only)	BEDENL	47.6	58.7	74.8	2176
	Ours (full)	BEDENL	50.3	62.2	78.8	2176
aerial retrieval hybrid	SALAD [23]-Aerial	BEDENL	7.4	13.5	32.8	2176
	Ours (full)	BEDENL	10.8	18.5	39.6	2176

Table 10: **Fine-grained localization.** We report recall for our best model on the two main datasets used in the paper, BEDENL and EuropeWest, at a *finer 100 m threshold* rather than 200 m, which closely aligned to the finer-grained cells ($L=16$) used for cross-view retrieval. This confirms our landmark achievement of 50.2% top-1 recall at 100m over a large subset of Europe (see: abstract). Notably, our prototypes are coarser at $L=15$, with an approximate size of 200×200 m. Despite this, combining coarse prototypes with finer aerial embeddings greatly boosts recall even at finer thresholds.

Ours (full)	Level	BEDENL			EuropeWest		
		Recall@ K @100m			Recall@ K @100m		
		$K=1$	$K=5$	$K=100$	$K=1$	$K=5$	$K=100$
Cross-view retrieval	L16	45.5	60.1	78.6	40.3	55.8	75.3
Prototype retrieval	L15	24.4	29.6	34.6	23.4	30.2	36.3
Hybrid	L16	54.3	69.6	84.4	50.2	67.5	83.0

B.3 Fine-grained Localization results

We provide an additional table that reports localization results at the finer 100 m threshold in Table 10. Note that our prototypes alone are too coarse for an evaluation at this threshold, usually spanning a region of 200×200 m at $L=15$ (for compute reasons). The aerial embeddings, in contrast, are at a finer threshold ($L=16$), *i.e.*, a quarter of the area covered by the cell prototypes, thus enabling finer-grained localization — given the space limits we reported only results at 200 m in the main paper, which allows us to make direct comparisons for all variants.

Our proposed hybrid evaluation that averages cell prototypes with aerial embeddings yields significant improvements. Note that we bridge the granularity gap by nearest-neighbor interpolation, *i.e.*, four $L=16$ aerial embeddings are paired with the same $L=15$ cell prototype. This evaluation confirms our landmark achievement of 50.2% top-1 recall at 100m over 10 countries in Europe.

B.4 Spatial error distribution

To better understand the improvements of our hybrid retrieval method, we illustrate the localization errors spatially. We conduct this experiment for the cross-view retrieval baseline SALAD [23]-Aerial, our best baseline that does not use aerial images (Ours (prototypes-only)), and our full model. The results are illustrated in Fig. 8. Notably our full model (middle) achieves improvements uniformly in all areas, which are mostly rural cells with low data density. There, the aerials, which are almost unaffected by data density, yield large improvements over cell prototypes, which needs to remember content seen during training. Prototypes, on the other hand, are inherently globally discriminative, and exhibit strong performance in more densely sampled areas of our datasets, which is weakly correlated with population density.

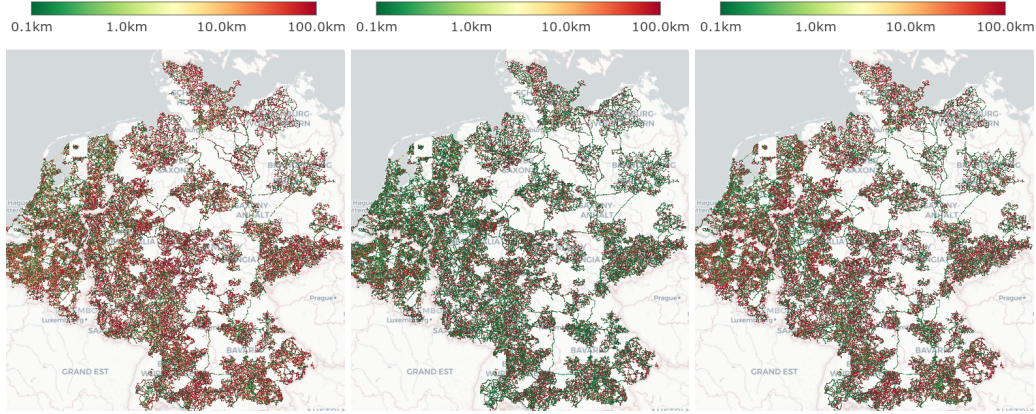


Figure 8: **Localization errors for queries in BEDENL.** Left: SALAD [23]-Aerial, Middle: Ours (full), Right: Ours (Prototypes). Our method mostly improves especially in rural areas, where ground-level training data is sparser.

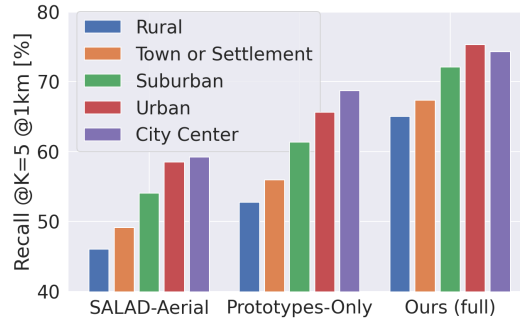


Figure 9: **Ablation of localization performance by population density on EuropeWest.** We report recall@5@1km. Performance improves in more densely populated areas, for all methods.

482 B.5 Evaluation by population density

483 We group test queries of EuropeWest by their urban population density (if available, many cells
 484 on highways and country roads are excluded), and report recall @5 at 1km in Fig. 9. Notably, all
 485 methods achieve higher accuracy on urban cells because of the adaptive sampling.

486 B.6 Performance improvements

487 Training our full model requires computing the similarity between both ground-view and aerial
 488 embeddings to all the prototypes. The performance bottleneck is two-fold: First the actual dot product,
 489 and second the all-to-all transform to gather the sharded similarities to the correct device. Improving
 490 inference speed on the actual similarity computation would require heuristics (e.g. training and
 491 maintaining a shortlist [80]), which would drastically increase the complexity of our method, and
 492 we thus refrain from doing so. Experimentally we found the all-to-all transform to be the actual
 493 bottleneck in our system, as it involves transferring the full similarity vector per batch element, twice.
 494 We alleviate this by first broadcasting and replicating all modality embeddings to each device, and
 495 then compute the similarity to the shard of prototypes on the specific device. We then compute the
 496 loss directly on the device, which improves training speed from 0.7 steps/sec to 1.0 steps/sec, without
 497 any impact on accuracy. However, this still requires maintaining the full gradient to all prototypes,
 498 which is both inefficient and harms accuracy, as many prototypes not visible in the batch get tiny,
 499 noisy updates. Furthermore, one can utilize approximate nearest-neighbor search within each device
 500 to mine hard negatives, and only add these elements to the loss. This improves the training speed to
 501 1.2 steps/sec on BEDENL, while also achieving slightly higher accuracy (+0.9 top-1 recall at 200m).
 502 However, for simplicity we report all results without approximate nearest neighbor search in the
 503 paper. For reference, our prototype-only baseline runs at 1.7 steps/sec, and our cross-view retrieval
 504 baseline (SALAD [23]-Aerial) at 1.5 steps/sec. On our hardware (128 TPUs), localizing a query
 505 image (encoding + retrieval) on EuropeWest takes around 0.4 sec.

C Implementation and baselines

C.1 Baselines

We discuss the (re-)implementations of the major baselines we compare against.

- **Fervers *et al.* [29]:** We adopt the same multi-head attention head, and the decoupled, bidirectional InfoNCE loss with label smoothing factor 0.1. For a fair comparison with our method, and in contrast to the original paper, we replace the ConvNext [81] backbone with a ViT [39], similar to all other baselines. We use a temperature $\tau = \frac{1}{36}$ as in the original paper. The original paper used a spacing of $5m$ between training images, which proved infeasible for us to run at this scale. We therefore equalize the data for a fair comparison. Similarly, the original paper adopted a cell size of 30×30 meters, which would increase size of the database by a factor of 10. We thus evaluate their method on the same resolution as ours (100×100 meter), and adopt the offset accordingly. One core insight of Fervers *et al.* [29] is that an image pyramid per cell yields substantial improvements. However, when experimenting we found this to be a major performance bottleneck, reducing throughput from 1.5 steps/sec to < 0.5 steps/sec. Furthermore, this insight is orthogonal to our method and would improve every approach that relies on aerial images. We therefore use a single level, *i.e.*, images of 256×256 px and a resolution of $0.6 \frac{m}{px}$, for both the baseline and for our model. We train the network for an equal amount of steps as our method. The authors also propose a look-ahead hard example mining (HEM) strategy, yet the authors note that this is not required with large batch sizes (our setup uses a batch size of 8192), and it is a performance bottleneck which requires an additional forward pass per batch. Instead, we try to strengthen the baseline by performing an offline hard-negative mining (on a trained model) using the aerial embeddings. For each cell, we encode its aerial image and find the top-k most similar features from other cells. During training, we then load per element images from 64 of its neighboring cells, and run training for 2 full epochs.
- **SALAD [23]-Aerial:** This baseline uses the same architecture as the original SALAD [23], both for the aerial and ground-level encoder (weights are not shared). In contrast to the original model, we initialize from iBOT [40] weights and finetune the entire network to account for the large domain gap between ground and aerial images. Similar to our work, we adopt the Multi-Similarity Loss [33], but without online negative mining. Instead, we contrast to all other elements in the batch. Similar to Fervers *et al.* [7], we use a bidirectional loss to contrast ground-level to aerial images and vice-versa. The remaining hyperparameters are identical to our full implementation.
- **Haversine loss [1]:** We adopt the haversine loss from PIGEON [1], which is a form of spatial label smoothing. We change the haversine temperature from $\tau = 75$ km in the original paper, tuned for coarser localization, to $\tau = 200$ m, which we empirically found to provide a nice trade-off between robustness and accuracy. The architecture and head are identical to our network, and we use l2-normalized embeddings with a learned temperature initialized to $\tau = 0.01$.
- **Hierarchical loss [2]:** OSV-5M [2] has demonstrated that a simple hierarchical loss on a quad-tree yields substantial improvements. We adopt this baseline in our evaluation, and adapt it with a tree of height $h = 4$ and using every second level in the loss, *i.e.*, we supervise the sum of probabilities at $L=15$, $L=13$, $L=11$, and $L=9$, and use a learned temperature initialized to $\tau = 0.01$. We use the same architecture and training setup as our main method.

C.2 Critical hyperparameters

We found that the most critical hyperparameters besides the learning rate are in the loss function. Of these, the base parameter $\lambda = 0.2$ has the largest impact on performance, controlling the push on the similarities. The parameters $\beta = 100$ and $\alpha = 2$ control the balancing between positive and negative examples. During training, the network tends to first push the prototypes apart to be close to orthogonal, and then increase the similarity to the respective ground-level and aerial embeddings in their region.

Table 11: **Comparison with existing public datasets.** They generally exhibit neither sufficiently dense coverage nor the spatial extent as large as our dataset. Some datasets do not include aerial imagery or only forward-facing ground-level imagery.

dataset	training			evaluation			aerial images	ground images
	images	spacing	countries	images	split	size		
EuropeWest	470M	40m	10	4.5M	temporal	continent	✓	random
Fervers <i>et al.</i> [7]	72M	5m	2	11M	cross-area	state	✓	forward-facing
OSV-5M [2]	5M	1km	225	210K	spatial	global		forward-facing



Figure 10: **Sampling of ground-level training images.** We illustrate the sampling within 4 $L=15$ cells for 3 examples. Each dot represents a panorama and each color corresponds to a different year in 2017–2024. We sample panoramas such that they are at least 40 m apart, which is a good trade-off between density and coverage. Left: intersection in a rural area, Middle: rural town, Right: a major city in BEDENL.

557 C.3 Efficient experimentation

558 As training these networks from scratch is expensive, we found that pre-training networks for cross-
559 view retrieval, and then fine-tuning the network with prototypes yields equal performance at a fraction
560 of the time. There, we initialize the backbones from the pretrained weights, and randomly initialize
561 the weights of the heads and prototypes. In these experiments, we also found it beneficial to increase
562 the learning rate of the head to 0.01, similarly as for the cell codes.

563 D Data

564 **Comparison to public datasets.** We provide a qualitative comparison to public datasets in Table 11.
565 Notably, no existing dataset has sufficient spatial extent and density for fine-grained, continental-scale
566 geo-localization. The dataset introduced by Fervers *et al.* [29] is most similar to ours, yet their actual
567 spatial distribution of training and test images is less dense and biased by the viewpoint (the authors
568 acknowledge that the “frontal street-view perspective is heavily overrepresented” in their dataset [29]).
569 Instead, we rely on random crops from 360° panoramas to model arbitrary viewpoints.

570 **Pinhole rendering:** We render 224×224 px pinhole images from stitched panoramas with height
571 768px (thus minimizing aliasing). To gain robustness to different intrinsics and viewpoints, we
572 randomly sample a roll in range $[-10^\circ, 10^\circ]$, a pitch in $[-5^\circ, 15^\circ]$, and a field of view in range
573 $[45^\circ, 75^\circ]$. For the yaw, we sample a random offset per panorama, stratify the yaw, (for example, four
574 yaw at 90° increments), and randomly perturb each yaw with a uniform random offset.

575 **Local sampling:** As discussed in a previous section, we employ spatio-temporal farthest point
576 sampling to maximize coverage in both axes. We illustrate our sampling in Fig. 10. Our approach
577 avoids oversampling cells that are only crossed by a few streets, while yielding dense coverage in
578 metros. This strikes a nice balance between accuracy and efficiency, thereby enabling the large scale
579 experiments conducted in this work. However, we would like to point out that this inherent adaptive
580 density might still not be sufficient in metros, as it 1) does not account for increased occlusions
581 (from traffic and denser settlements) and more frequent temporal changes (e.g. construction sites) in

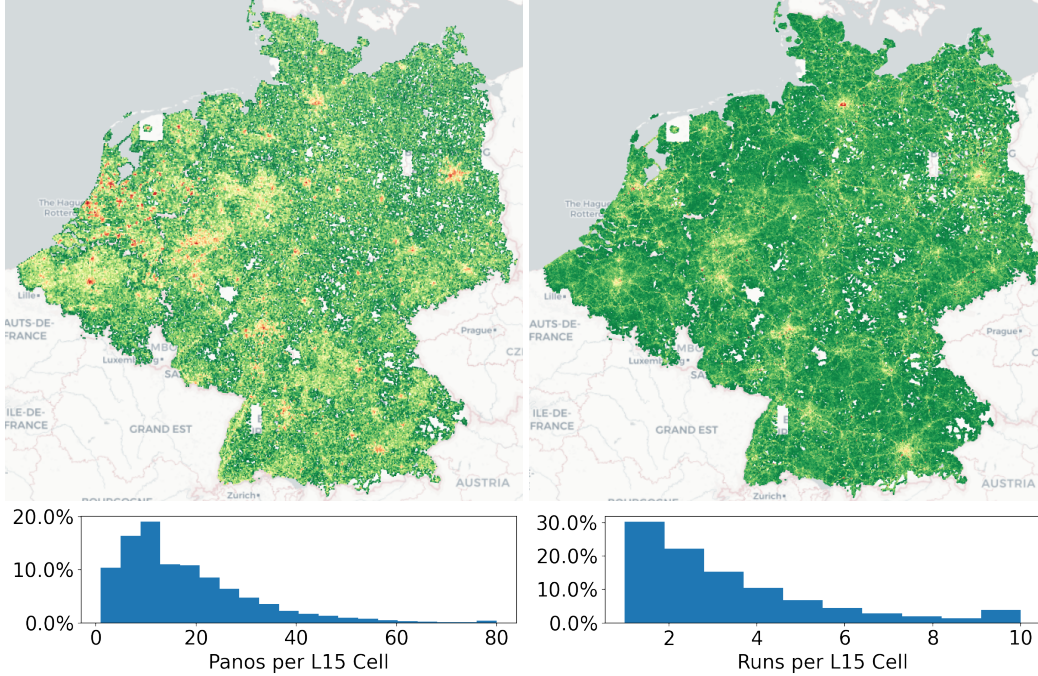


Figure 11: **Data density.** The training data densely covers BEDENL. The density of panoramas (left column) is weakly proportional the the population density, while temporal diversity from runs (right) is largest in urban centers and on highways. On average, a cell in our dataset has 20 panos (4 sampled views per pano) from 3 unique runs. Both have a direct impact on the final localization accuracy, as shown by Fig. 3.

Table 12: **Number of training and evaluation images per country (millions).**

	PT	ES	IT	AT	CH	DE	FR	BE	NL	CZ
Training	19.0	52.3	76.9	16.4	8.3	115.5	149.5	13.5	17.0	14.5
Test	0.2	0.5	0.7	0.2	0.1	1.2	1.4	0.2	0.2	0.1

urban areas, and 2) is only weakly proportional to the actual population density. This motivated us to perform additional experiments on BEDENL+, a dataset that mixes additional urban training samples to the existing dataset, BEDENL (BEDENL+ is a superset of BEDENL).

Spatial coverage and density: We illustrate the spatial and temporal coverage in Fig. 11. Most cells in our dataset have low data variety, averaging at around 20 panos per cell. The density is significantly higher in urban and suburban areas. Contrary to the density in panoramas, the amount of unique runs these panoramas were captured in shows a significantly different behaviour: We have the largest temporal diversity on highways. This also supports our qualitative observation of increased accuracy on highways (see e.g. Fig. 4), which at first sight is counter-intuitive because of the lack of visual landmarks there.

Holes in dataset: Our dataset exhibits a few rectangular holes in the dataset. In these areas, our dataset lacks aerial image coverage from the same sensor, and we thus exclude this from training. During evaluation, however, as can be seen in Fig. 4, we utilize satellite imagery of lower quality there. While this does impact the accuracy, our method is still able to correctly localize a substantial part of images there. We therefore suspect that mixing aerial and satellite imagery would further increase the robustness of our localizations system.

Statistics by country: In Table 12 we provide detailed statistics about the number of train and test images per country. Notably, our test set is sampled uniformly over space. While it could be argued that this biases results towards sparse, rural areas, which usually have harder queries, we believe this accurately reflects how one would assess a true global localization system, which should exhibit spatially uniform performance.

E Visualizations

We qualitatively show query examples in the EuropeWest dataset, labelled by their top-1 localization error (in km), see Fig. 12. We compare retrieval to the strongest cross-view (G-A) and classification-only (G-P) baselines on this dataset. Notable, cross-view retrieval methods achieve higher accuracy in rural areas with few bird’s-eye occlusion obstacles, while cell prototypes excel in urban areas with large, vertical facades. Our method combines the best of both worlds by relying on the factor that best describes the respective area.

The last couple of rows show failure cases. Notably, low-texture regions such as the border wall of highways or zoomed-in images, large dynamic occlusions (trucks and cars), and repetitive vegetation are limitations of our approach. Overall, we qualitatively made the observation that the network tends to localize images very well if 1) the road is visible and ahead, which coincides with 2) the image has large depth, and therefore can observe distant features. We believe that these distant landmarks are beneficial as they are easier to observe from different angles, and can help to coarsely remember locations, especially in rural areas.

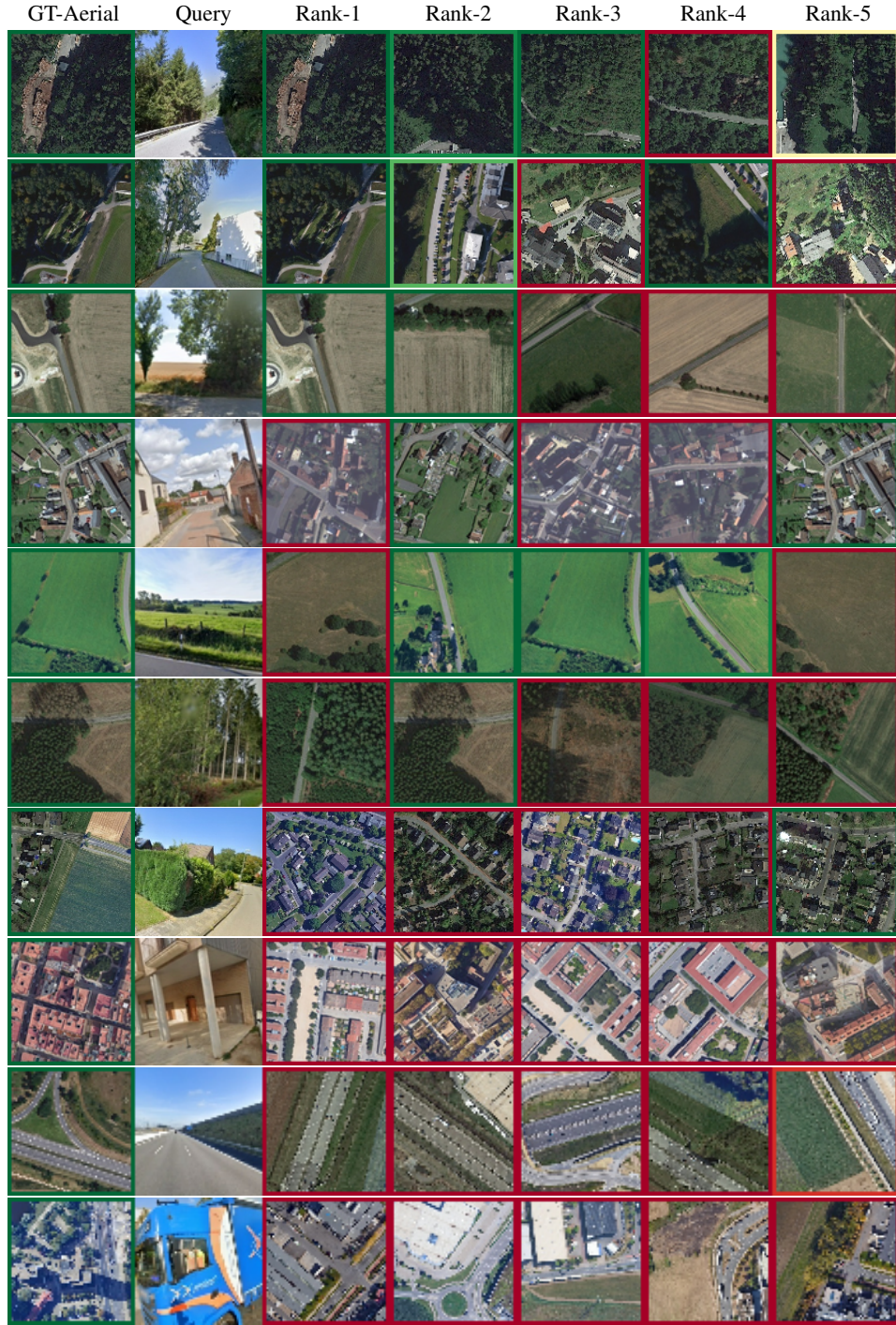
Fig. 13 shows the top-5 retrieved aerial images of our method. Notably, the network is able to utilize patterns in the vegetation, such as the spacing between trees (row 3), or features at large depth (row 5) to robustly find the correct area. Failure cases are typically close-up images (third-last row), or ambiguous queries such as empty highways, or very large occlusions.

We finally analyze the PCA visualizations of our network and one major ablation in Fig. 14. Supervising just prototypes yields smooth and very informative prototypes that clearly encode semantic and geological entities. Empirically, we observe PCA features to be more informative the coarser the cells are (*i.e.*, the stronger the information bottleneck).

Finally, we show self-similarity patterns between cell prototypes in Fig. 15. While the prototypes are almost fully orthogonal to each other, which helps localization, they are locally similar, hinting that the network indeed uses coarser geospatial patterns for localization.



Figure 12: **Localization errors of queries in EuropeWest.** a) Database aerial image for the ground truth location, b) Query image and error for our prototype-only variant, c) Query image and error for the baseline SALAD [23]-Aerial, d) Query image and error for our full hybrid approach. Our approach is able to correctly localize ambiguous rural cells by combining ground- and aerial cues, and is robust in many scenarios. Failure occurs in scenarios with occlusion by transient objects (cars, trucks) and queries with narrow field of view. In general, the method performs significantly better the less the view is obscured.



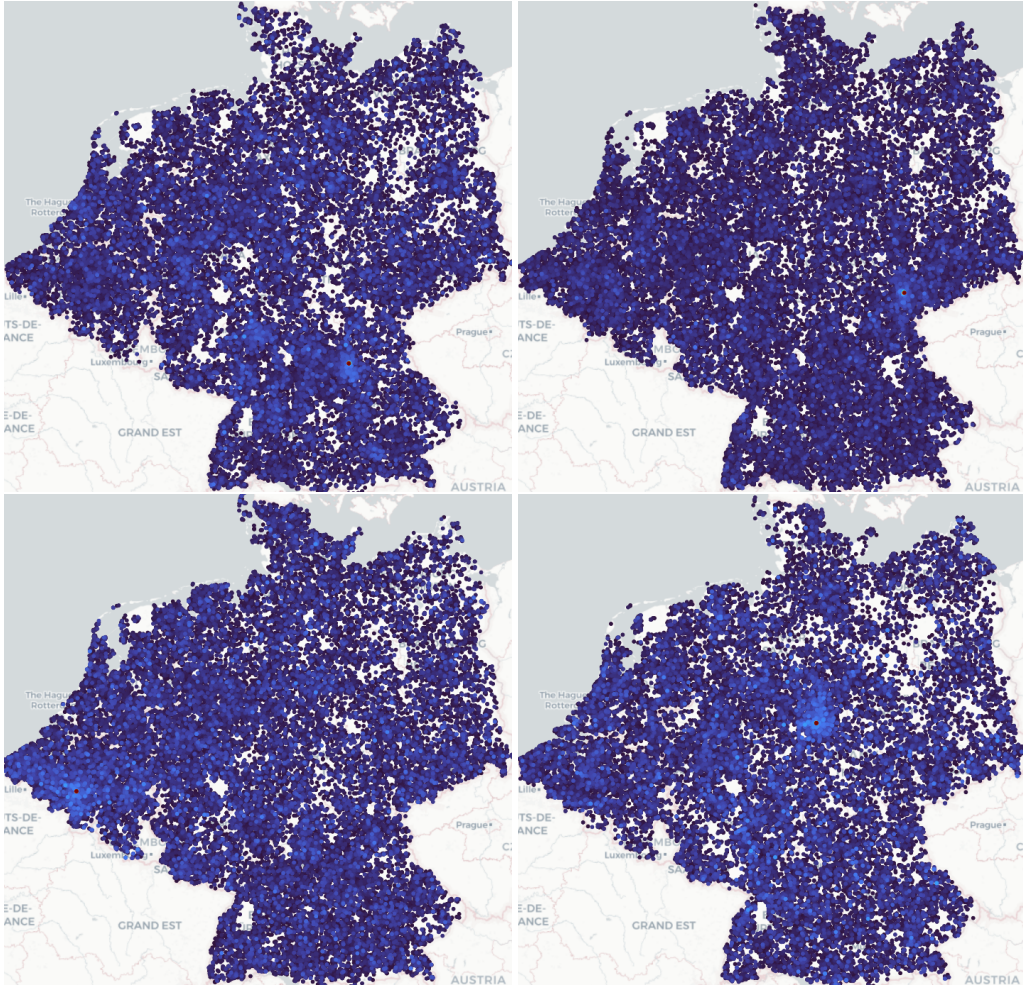


Figure 15: **Self-similarities between prototypes in BEDENL.** We show the self-similarities of 4 prototypes (red dots) to their top 50k neighbors. Red and blue correspond to a high and low similarities, respectively. The prototypes are almost fully orthogonal, yet locally smooth.

References

- [1] Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. PIGEON: Predicting Image Geolocations. In *CVPR*, 2024. 1, 2, 3, 5, 6, 7, 9, 15
- [2] Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronssohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, et al. OpenStreetView-5M: The Many Roads to Global Visual Geolocation. In *CVPR*, 2024. 1, 2, 5, 6, 7, 9, 15, 16
- [3] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Improving image-based localization by active correspondence search. In *ECCV*, 2012. 1
- [4] Bernhard Zeisl, Torsten Sattler, and Marc Pollefeys. Camera pose voting for large-scale image-based localization. In *ICCV*, 2015. 1, 9
- [5] Simon Lynen, Bernhard Zeisl, Dror Aiger, Michael Bosse, Joel Hesch, Marc Pollefeys, Roland Siegwart, and Torsten Sattler. Large-scale, real-time visual-inertial localization revisited. *IJRR*, 39(9):1061–1084, 2020. 1
- [6] Paul-Edouard Sarlin, Daniel DeTone, Tsun-Yi Yang, Armen Avetisyan, Julian Straub, Tomasz Malisiewicz, Samuel Rota Buló, Richard Newcombe, Peter Kotschieder, and Vasileios Balntas. OrienterNet: Visual Localization in 2D Public Maps with Neural Matching. In *CVPR*, 2023. 1, 9
- [7] Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. Uncertainty-aware Vision-based Metric Cross-view Geolocation. In *CVPR*, 2023. 1, 15, 16
- [8] Paul-Edouard Sarlin, Eduard Trulls, Marc Pollefeys, Jan Hosang, and Simon Lynen. SNAP: Self-Supervised Neural Maps for Visual Positioning and Semantic Understanding. In *NeurIPS*, 2023. 1, 9
- [9] Tobias Weyand, Ilya Kostrikov, and James Philbin. PlaNet - Photo Geolocation with Convolutional Neural Networks. In *ECCV*, 2016. 2, 7, 9
- [10] Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. CPlaNet: Enhancing Image Geolocation by Combinatorial Partitioning of Maps. In *ECCV*, 2018. 2, 9
- [11] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization. *NeurIPS*, 2023. 2, 9
- [12] Brandon Clark, Alec Kerrigan, Parth Parag Kulkarni, Vicente Vivanco Cepeda, and Mubarak Shah. Where We Are and What We’re Looking At: Query Based Worldwide Image Geolocation Using Hierarchies and Scenes. In *CVPR*, 2023. 2, 9
- [13] Pengyue Jia, Yiding Liu, Xiaopeng Li, Xiangyu Zhao, Yuhao Wang, Yantong Du, Xiao Han, Xuetao Wei, Shuaiqiang Wang, and Dawei Yin. G3: An Effective and Adaptive Framework for Worldwide Geolocation Using Large Multi-Modality Models. *NeurIPS*, 2024. 2
- [14] Parth Parag Kulkarni, Gaurav Kumar Nayak, and Mubarak Shah. CityGuessr: City-Level Video Geo-Localization on a Global Scale. In *ECCV*, 2024. 2
- [15] Nicolas Dufour, David Picard, Vicky Kalogeiton, and Loic Landrieu. Around the World in 80 Timesteps: A Generative Approach to Global Visual Geolocation. *arXiv:2412.06781*, 2024. 2, 9
- [16] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *CVPR*, 2016. 2
- [17] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking Visual Geo-localization for Large-Scale Applications. In *CVPR*, 2022. 2, 3, 8
- [18] Gabriele Berton, Gabriele Trivigno, Barbara Caputo, and Carlo Masone. EigenPlaces: Training Viewpoint Robust Models for Visual Place Recognition. In *CVPR*, 2023. 2, 8

- [19] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *RA-L*, 9(2):1286–1293, 2023. 2, 9
- [20] Tong Wei, Philipp Lindenberger, Jiří Matas, and Daniel Barath. Breaking the Frame: Visual Place Recognition by Overlap Prediction. In *WACV*, 2025. 2, 9
- [21] Gabriele Berton and Carlo Masone. MegaLoc: One Retrieval to Place Them All. *arXiv:2502.17237*, 2025. 2, 3, 9
- [22] Sergio Izquierdo and Javier Civera. Close, But Not There: Boosting Geographic Distance Sensitivity in Visual Place Recognition. In *ECCV*, 2024. 2, 9
- [23] Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *CVPR*, pages 17658–17668, 2024. 2, 3, 5, 6, 8, 9, 11, 12, 13, 14, 15, 19
- [24] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-Aware Feature Aggregation for Image based Cross-View Geo-Localization. In *NeurIPS*, 2019. 2, 9
- [25] Qingwang Zhang and Yingying Zhu. Benchmarking the Robustness of Cross-View Geo-Localization Models. In *ECCV*, 2024. 2, 9
- [26] Junyan Ye, Zhutao Lv, Weijia Li, Jinhua Yu, Haote Yang, Huaping Zhong, and Conghui He. Cross-view image geo-localization with Panorama-BEV Co-Retrieval Network. In *ECCV*, 2024. 2, 9
- [27] Li Mi, Chang Xu, Javiera Castillo-Navarro, Syrielle Montariol, Wen Yang, Antoine Bosselut, and Devis Tuia. ConGeo: Robust Cross-view Geo-localization across Ground View Variations. In *ECCV*, 2024. 2, 9
- [28] Shixiong Xu, Chenghao Zhang, Lubin Fan, Gaofeng Meng, Shiming Xiang, and Jieping Ye. AddressCLIP: Empowering Vision-Language Models for City-wide Image Address Localization. In *ECCV*, 2024. 2
- [29] Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. Statewide Visual Geolocalization in the Wild. In *ECCV*, 2024. 2, 5, 6, 9, 15, 16
- [30] Bingyi Cao, Andre Araujo, and Jack Sim. Unifying Deep Local and Global Features for Image Search. In *ECCV*, 2020. 2, 8
- [31] Dan Larkin-York, Google Inc., Koordinates Limited, Mike Playle, and Tiago Brito. S2 Geometry Library. <https://github.com/google/s2geometry>, 2015. [Online; accessed 13-May-2025]. 3, 5, 9
- [32] Jonas Theiner, Eric Müller-Budack, and Ralph Ewerth. Interpretable Semantic Photo Geolocation. In *WACV*, 2022. 3
- [33] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-Similarity Loss with General Pair Weighting for Deep Metric Learning. In *CVPR*, 2019. 4, 6, 9, 15
- [34] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4
- [35] Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. In *ECCV*, 2022. 4
- [36] Google Inc. TPU v2 documentation. <https://cloud.google.com/tpu/docs/v2>, 2025. 4
- [37] Danny Cheung. Mapping stories with a new Street View Trekker. <https://blog.google/products/maps/mapping-stories-new-street-view-trekker/>, 2018. 5
- [38] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980*, 2014. 8

- [39] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021. 8, 9, 15
- [40] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT Pre-training with Online Tokenizer. In *ICLR*, 2022. 8, 15
- [41] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *TMLR*, 2024. 8
- [42] James Hays and Alexei A Efros. IM2GPS: estimating geographic information from a single image. In *CVPR*, 2008. 8, 9
- [43] Jan Knopp, Josef Sivic, and Tomas Pajdla. Avoiding Confusing Features in Place Recognition. In *ECCV*, 2010. 8
- [44] Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier. Large-Scale Image Retrieval with Compressed Fisher Vectors. In *CVPR*, 2010. 8
- [45] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid. Aggregating Local Image Descriptors into Compact Codes. *IEEE TPAMI*, 34(9):1704–1716, 2011. 8, 9
- [46] James Hays and Alexei A Efros. Large-Scale Image Geolocalization. In *Multimodal Location Estimation of Videos and Images*, 2015. 8
- [47] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. Predicting Good Features for Image Geo-Localization Using Per-Bundle VLAD. In *ICCV*, 2015. 8
- [48] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *CVPR*, 2017. 8, 9
- [49] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 8
- [50] Aude Oliva and Antonio Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *IJCV*, 42:145–175, 2001. 8
- [51] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the Fisher Kernel for Large-Scale Image Classification. In *ECCV*, 2010. 8
- [52] Nam Vo, Nathan Jacobs, and James Hays. Revisiting IM2GPS in the Deep Learning Era. In *CVPR*, 2017. 8
- [53] Ruotong Wang, Yanqing Shen, Weiliang Zuo, Sanping Zhou, and Nanning Zheng. TransVPR: Transformer-Based Place Recognition with Multi-Level Attention Aggregation. In *CVPR*, 2022. 8
- [54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 9
- [55] Amar Ali-Bey, Brahim Chaib-Draa, and Philippe Giguère. MixVPR: Feature Mixing for Visual Place Recognition. In *WACV*, 2023. 9
- [56] Artem Babenko and Victor Lempitsky. Aggregating Deep Convolutional Features for Image Retrieval. In *ICCV*, 2015. 9
- [57] Ali S Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. Visual Instance Retrieval with Deep Convolutional Networks. *ITE Transactions on Media Technology and Applications*, 4(3):251–258, 2016. 9

- [58] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning CNN Image Retrieval with No Human Annotation. *IEEE TPAMI*, 41(7):1655–1668, 2018. 9
- [59] Torsten Sattler, Michal Havlena, Konrad Schindler, and Marc Pollefeys. Large-Scale Location Recognition and the Geometric Burstiness Problem. In *CVPR*, 2016. 9
- [60] Hajime Taira, Ignacio Rocco, Jiri Sedlar, Masatoshi Okutomi, Josef Sivic, Tomas Pajdla, Torsten Sattler, and Akihiko Torii. Is This The Right Place? Geometric-Semantic Pose Verification for Indoor Visual Localization. In *ICCV*, 2019. 9
- [61] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-NetVLAD: Multi-Scale Fusion of Locally-Global Descriptors for Place Recognition. In *CVPR*, 2021. 9
- [62] Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. R2Former: Unified Retrieval and Reranking Transformer for Place Recognition. In *CVPR*, 2023. 9
- [63] Nathan Piasco, Désiré Sidibé, Cédric Demonceaux, and Valérie Gouet-Brunet. A survey on Visual-Based Localization: On the benefit of heterogeneous data. *Pattern Recognition*, 74:90–109, 2018. 9
- [64] Stephen Hausler, Tobias Fischer, and Michael Milford. Unsupervised Complementary-aware Multi-process Fusion for Visual Place Recognition. *arXiv:2112.04701*, 2021. 9
- [65] Frederik Warburg, Martin Jørgensen, Javier Civera, and Søren Hauberg. Bayesian Triplet Loss: Uncertainty Quantification in Image Retrieval. In *ICCV*, 2021. 9
- [66] Kaiwen Cai, Chris Xiaoxuan Lu, and Xiaowei Huang. STUN: Self-Teaching Uncertainty Estimation for Place Recognition. In *IROS*, 2022. 9
- [67] Mubariz Zaffar, Liangliang Nan, and Julian FP Kooij. On the Estimation of Image-matching Uncertainty in Visual Place Recognition. In *CVPR*, 2024. 9
- [68] Gabriele Berton, Lorenz Junglas, Riccardo Zaccone, Thomas Pollok, Barbara Caputo, and Carlo Masone. MeshVPR: Citywide Visual Place Recognition Using 3D Meshes. In *ECCV*, 2024. 9
- [69] Francesco Castaldo, Amir Zamir, Roland Angst, Francesco Palmieri, and Silvio Savarese. Semantic Cross-View Matching. In *ICCV Workshops*, pages 9–17, 2015. 9
- [70] Arsalan Mousavian and Jana Kosecka. Semantic Image Based Geolocation Given a Map. *arXiv:1609.00278*, 2016. 9
- [71] Nam N Vo and James Hays. Localizing and Orienting Street Views Using Overhead Imagery. In *ECCV*, 2016. 9
- [72] Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. CVM-Net: Cross-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization. In *CVPR*, 2018. 9
- [73] Scott Workman and Nathan Jacobs. On the location dependence of convolutional neural network features. In *CVPR Workshops*, 2015. 9
- [74] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-Area Image Geolocation with Aerial Reference Imagery. In *ICCV*, 2015. 9
- [75] Liu Liu and Hongdong Li. Lending Orientation to Neural Networks for Cross-view Geo-localization. In *CVPR*, 2019. 9
- [76] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal Feature Transport for Cross-View Image Geo-Localization. In *AAAI*, 2020. 9, 12
- [77] Shraman Pramanick, Ewa M Nowara, Joshua Gleason, Carlos D Castillo, and Rama Chellappa. Where in the World is this Image? Transformer-based Geo-localization in the Wild. In *ECCV*, 2022. 9
- [78] Lukas Haas, Silas Alberti, and Michal Skreta. Learning Generalized Zero-Shot Learners for Open-Domain Image Geolocation. *arXiv:2302.00275*, 2023. 9

- 810 [79] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. GSV-Cities: Toward Appropriate
811 Supervised Visual Place Recognition. *Neurocomputing*, 513:194–203, 2022. 12
- 812 [80] N. Gupta, P.H. Chen, H-F. Yu, C-J. Hsieh, and I. Dhillon. ELIAS: End-to-end Learning to Index
813 and Search in Large Output Spaces. In *NeurIPS*, 2022. 14
- 814 [81] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining
815 Xie. A ConvNet for the 2020s. *CVPR*, 2022. 15