AMES: Asymmetric and Memory-Efficient Similarity Estimation for Instance-level Retrieval -Supplementary material-

Pavel Suma[1], Giorgos Kordopatis-Zilos[1], Ahmet Iscen[2], and Giorgos Tolias[1]

[1]VRG, FEE, Czech Technical University in Prague [2]Google DeepMind

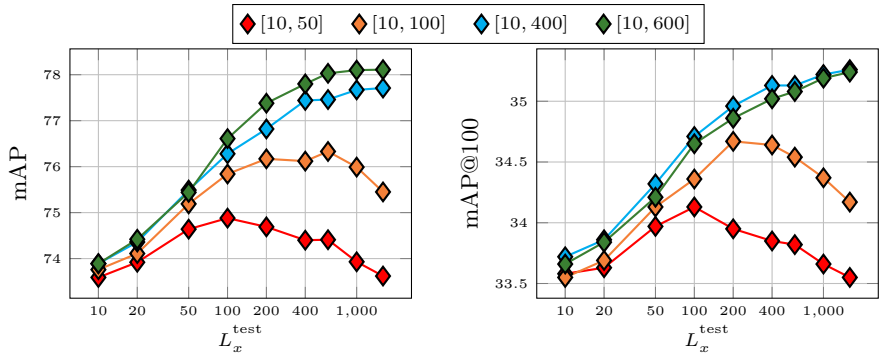


Fig. A: Impact of L_x^{train}, L_q^{train} sampling range. AMES is trained by limiting the descriptor set size to a particular range during training while testing for all set sizes, both within and outside the range used for training. Performance evaluated on $\mathcal{R}$OP+1M (left) and GLDv2 (right) for AMES. All runs use global descriptors with PQ8 for initial ranking and ensemble similarity to re-rank m=1600 images.

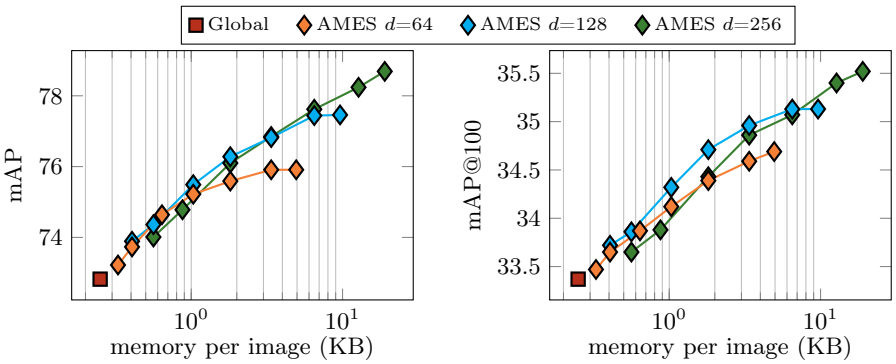


Fig. B: Impact of dimensionality d. Performance evaluated on $\mathcal{R}$OP+1M (left) and GLDv2 (right) for AMES. All runs use global descriptors with PQ8 for initial ranking and ensemble similarity to re-rank m=1600 images.

ℓ_2 loss	β	$\mathcal{R}$OP+1M	GLDv2
–	0	75.5	34.3
Z_N	1	76.3	34.6
	10	76.4	34.7
	100	76.3	34.5
S	1	75.7	34.4
	10	75.9	34.4
	100	75.9	34.4

Table A: Impact of distillation. Z_N stands for distillation with ℓ_2 loss on the output tokens (proposed method). S stands for distillation with ℓ_2 loss on the final local similarity scores. Z_N and $\beta = 10$ are the default choices for AMES.

N	$\mathcal{R}$OP+1M	GLDv2
0	72.8	33.3
1	73.0	33.3
3	75.1	34.1
5	75.5	34.3
7	75.6	34.3

Table B: Impact of network depth. $N = 0$ stands for the performance of the global-only similarity for reference. $N = 5$ is the default choice for AMES.

ITQ	FT	$\mathcal{R}$OP+1M	GLDv2
–	–	74.6	34.1
–	✓	74.8	34.0
✓	–	75.1	34.3
✓	✓	75.5	34.3

Table C: Impact of binarization settings. **ITQ** stands for initialization with ITQ [1]. **FT** stands for fine-tuning W, as opposed to keeping it frozen. ITQ and FT are the defaults for AMES.

A Additional results

In this section, we provide additional experimental results with different model settings and training hyper-parameters. Unless specified otherwise, $L_q^{\text{test}} = 600$ and $L_x^{\text{test}} = 50$, global descriptors with PQ8, $m = 1600$ and AMES binary without distillation are used.

Impact of $L_x^{\text{train}}, L_q^{\text{train}}$ sampling range. Figure A displays the results of AMES trained with four different sampling ranges. In the smaller sampling ranges, performance increases until L_x^{test} is close to the maximum range used during training. After that point, performance almost consistently drops or remains the same. This behavior aligns with the observation in prior work in different research field [11], where transformer-based models are trained and tested with different input sequence lengths. In the larger sampling ranges, performance steadily improves as L_x^{test} increases. It saturates for the larger descriptor set sizes, *i.e.* performance gains are marginal for L_x^{test} greater than 600 with significant memory and computation overhead. The two larger sampling ranges report very close results; hence, we select $[10, 400]$ in our default settings for better efficiency in training time and memory allocation. Nevertheless, training with an even larger sampling range could potentially yield even better performance.

Impact of dimensionality d. To further investigate the memory footprint and performance trade-off, we explore two additional dimensionalities of the binary local descriptors in Figure B. In low memory regimes, using more local descriptors with lower dimensions is advantageous. On the other hand, it is

					Medium				Hard				GLDv2
L_q^{test}	L_x^{test}	bin	dist	global	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	
600	600	✓	✓	PQ8	88.3±0.4	83.5±0.4	93.2±0.2	86.9±0.1	77.0±0.7	69.6±0.5	86.1±0.4	75.7±0.1	35.6±0.1
600	50	✓	✓	PQ8	87.0±0.3	81.1±0.2	92.6±0.0	85.4±0.0	75.0±0.5	66.0±0.3	84.7±0.0	72.9±0.1	34.7±0.1
50	50	✓	✓	PQ8	86.9±0.2	80.9±0.3	92.4±0.1	85.0±0.1	74.7±0.4	65.4±0.4	84.3±0.1	72.0±0.2	34.3±0.1
600	600	✓	–	PQ8	87.7±0.4	82.4±0.3	92.7±0.3	86.2±0.4	75.6±0.7	67.3±0.5	85.1±0.6	74.0±0.8	35.1±0.0
600	50	✓	–	PQ8	86.7±0.5	80.6±0.3	92.2±0.2	84.7±0.2	74.4±0.9	65.2±0.5	83.9±0.4	71.4±0.5	34.3±0.1
50	50	✓	–	PQ8	86.5±0.2	80.2±0.3	92.0±0.2	84.2±0.2	73.9±0.3	64.4±0.3	83.4±0.3	70.5±0.4	34.0±0.0
600	600	–	–	PQ8	89.3±0.2	84.7±0.4	93.3±0.0	87.2±0.2	78.1±1.0	71.5±0.7	86.6±0.0	76.5±0.4	35.8±0.3
600	50	–	–	PQ8	87.2±0.2	81.5±0.2	92.6±0.1	85.4±0.1	75.5±0.8	66.9±0.7	84.8±0.3	73.1±0.4	34.8±0.2
50	50	–	–	PQ8	86.8±0.2	80.7±0.1	92.3±0.1	85.0±0.1	74.4±0.7	65.1±0.4	84.1±0.2	71.9±0.2	34.4±0.1
600	600	–	–	full	89.1±0.1	84.4±0.4	93.2±0.1	87.1±0.2	78.3±0.3	71.2±0.5	86.3±0.2	76.2±0.3	35.9±0.2
600	50	–	–	full	87.3±0.3	81.3±0.2	93.0±0.1	85.6±0.2	75.8±0.8	66.2±0.7	85.6±0.1	73.3±0.2	34.7±0.2
50	50	–	–	full	87.3±0.2	80.9±0.0	92.6±0.0	85.0±0.1	75.5±0.6	65.5±0.2	84.7±0.2	72.0±0.2	34.4±0.1

Table D: AMES performance for different settings reported separately per dataset with standard deviations across three experiments using a different seed. mAP used for $\mathcal{R}$Oxf and $\mathcal{R}$Par. mAP@100 used for GLDv2.

preferable to use fewer higher-dimensional descriptors instead of including all available ones in the high memory regime. A subset of the descriptors carries most of the necessary information, and adding more introduces redundancy or noise in the matching.

Impact of distillation. In Table A, we evaluate the impact of the hyper-parameter β, and we compare our distillation scheme with another alternative that applies distillation on the output similarity scores, commonly used in the literature [4,8]. Different values of β do not significantly affect performance. Our distillation scheme considerably outperforms the similarity-based approach. This is expected considering that the latter may oppose the supervision loss, whereas ours is complementary.

Impact of network depth. Table B reports the performance of AMES implemented with various numbers of blocks N. The performance saturates after using more than three transformer blocks.

Impact of binarization settings. In Table C, we assess our implementation choices for our binarization scheme, *i.e.* the initialization with ITQ and the fine-tuning of the trainable parameters of the matrix W. Both choices are necessary to achieve the final performance, with the proper initialization having a significant impact since the model does not achieve similar performance while learning W from scratch.

Performance mean and standard deviation. For future reference, we provide the numeric results of our AMES model on the different $\mathcal{R}$Oxford and $\mathcal{R}$Paris settings, as well as the private split of GLDv2, in Table D. Results of AMES in Table 1 from the main paper are repeated in Table E with their standard deviation. The mean and standard deviation of three training runs with different seeds are reported.

Top-m re-ranking. Figure C shows the retrieval performance for re-ranking using local similarity only or with the global-local ensemble. Local similarity does not work by itself; it harms the performance compared to the global similarity

global desc.	local desc.	re-ranking			Medium				Hard				GLDv2
		bin	dist	top-m	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	$\mathcal{R}$Oxf	+1M	$\mathcal{R}$Par	+1M	
CVNet [5]	**CVNet** [5]	–	–	100	84.9±0.2	78.6±0.2	90.6±0.0	81.3±0.0	71.1±0.5	62.4±0.7	81.6±0.0	66.5±0.1	35.3±0.2
		✓	✓	100	84.6±0.7	78.4±0.6	90.6±0.0	81.3±0.0	70.8±1.4	61.9±1.0	81.5±0.1	66.3±0.2	35.0±0.2
		–	–	400	86.8±0.3	81.3±0.2	91.9±0.0	84.1±0.1	74.2±0.6	66.3±0.5	84.1±0.1	71.5±0.1	35.5±0.2
		✓	✓	400	86.3±1.0	80.9±0.8	91.9±0.1	84.0±0.0	73.7±1.6	65.5±1.2	84.1±0.3	71.2±0.3	35.1±0.2
SG [9]	**CVNet** [5]	–	–	1600	89.1±0.1	84.4±0.4	93.2±0.1	87.1±0.2	78.3±0.3	71.2±0.5	86.3±0.2	76.2±0.3	35.9±0.2
		✓	✓	1600	88.5±0.4	83.6±0.4	93.2±0.2	87.0±0.1	77.2±0.9	69.8±0.6	86.3±0.4	75.9±0.3	35.5±0.2
	DINOv2 [7]	–	–	1600	92.4±0.9	87.1±0.5	95.2±0.1	89.8±0.1	83.1±1.1	76.1±0.7	90.2±0.4	81.0±0.3	38.3±0.2
		✓	✓	1600	90.7±0.3	85.1±0.2	94.9±0.1	89.3±0.2	80.0±0.9	72.6±0.8	89.7±0.2	80.0±0.4	37.8±0.0
		with SG re-rank [9]											
	CVNet [5]	–	–	1600	91.1±0.2	86.6±0.4	94.3±0.1	88.8±0.1	80.4±0.7	74.1±0.6	88.6±0.3	79.9±0.2	36.0±0.2
		✓	✓	1600	90.7±0.3	85.9±0.3	94.3±0.1	88.9±0.1	79.4±0.7	72.9±0.6	88.7±0.4	79.8±0.2	35.8±0.1
	DINOv2 [7]	–	–	1600	93.6±0.6	88.2±0.4	95.3±0.1	90.1±0.1	84.8±0.9	77.7±0.8	90.7±0.4	82.0±0.2	38.5±0.2
		✓	✓	1600	92.7±0.3	86.7±0.1	95.2±0.2	89.8±0.2	83.0±0.6	75.4±0.5	90.4±0.3	81.4±0.5	38.0±0.1

Table E: AMES performance for different backbones reported separately per dataset with standard deviations across three experiments using a different seed. mAP used for $\mathcal{R}$Oxf and $\mathcal{R}$Par. mAP@100 used for GLDv2. Global descriptors are in full precision.

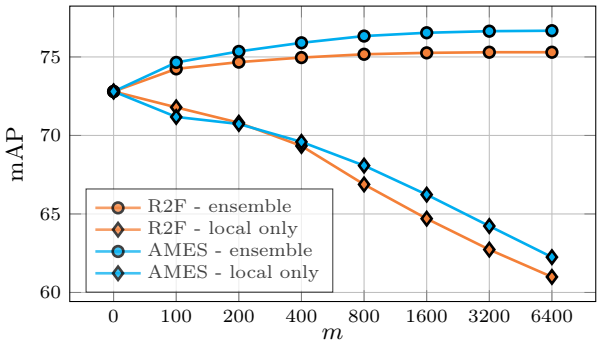


Fig. C: Impact of re-ranking the top-m images obtained by global similarity. The global and local similarity ensemble is effective, but local similarity fails by itself. Results on $\mathcal{R}$OP+1M.

ranking. By contrast, the ensemble similarity consistently increases performance as the number of re-ranked images increases. The same holds both for our method and R^2Former. This observation is likely a byproduct of the high-performing global descriptors used in our work and is not fully aligned with the original findings by other methods using older descriptors, such as RRT using DELG.

Transformer architecture comparison. In our experiments with RRT, we observe a small performance decrease by including the global descriptor in the input token set, which is part of the original RRT architecture. We attribute this to the difficulty of mapping both descriptors in the same space and the insignificant value of one extra token compared to the many tokens of the local descriptors. The results of Figure 3 in the main paper are obtained with the original RRT setup. In addition to that, we use the RRT model architecture and train with the same input token set as AMES, *i.e.* excluding the global descriptor. Results are presented for the binarized variants of both models in Figure D, where AMES seems to consistently outperform RRT even after our fix to it.

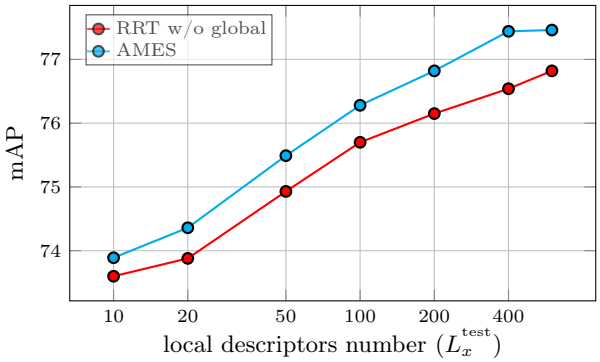


Fig. D: AMES *vs.* RRT for varying L_x^{test}, and $L_q^{\text{test}} = 600$ on $\mathcal{R}$OP+1M. RRT is modified to remove the global descriptor from its input tokens.

In-depth analysis of global-local similarity Using an ensemble of global and local models to compute the final similarity for each pair of images is clearly beneficial based on quantitative results. In Figure E, we present the global and local similarities, estimated by the corresponding models separately, for six selected queries. In several cases, the two types of similarities are not linearly correlated, making it easier to separate matching and non-matching pairs.

B Additional implementation details

Backbone architecture and training process. Our network is trained for 15 epochs with 300 batch size. Every batch consists of 100 triplets, in which one training sample acts as an anchor. All training samples are used as an anchor exactly once per epoch. The rest of the triplet consists of a matching sample from the same class as the anchor and a non-matching sample. Both are randomly drawn from the anchor's 300 nearest neighbors with a distribution proportional to the cube of their global similarities. The network is optimized with AdamW [6] and cosine learning rate scheduler with the initial value of $2 \cdot 10^{-4}$. The variance δ for the binarization layer is set to 10^{-3}, following [4]. Figure F shows the distillation process between the teacher and the student models.

In the experiments with CVNet, we follow the extraction pipeline from the original paper. We feed the input images in 3 resolutions and extract the local descriptors from the penultimate layer feature map. When using SuperGlobal, we skip Scale-GeM and ReLUP to reproduce performance close to the reported one. SG is reported to achieve 73.4 and 33.4 on $\mathcal{R}$OP+1M and GLDv2, respectively, without re-ranking and 81.2 and 35.0 with re-ranking $m=1600$ images; see Table 1 for our reproduction. For DINOv2 [7] experiments, we use its ViT-B/14 variant with registers to extract patch tokens and CLS tokens as our local and global descriptors. We resize the input images such that the larger image side equals 518 pixels and pad the rest of the image to a square.

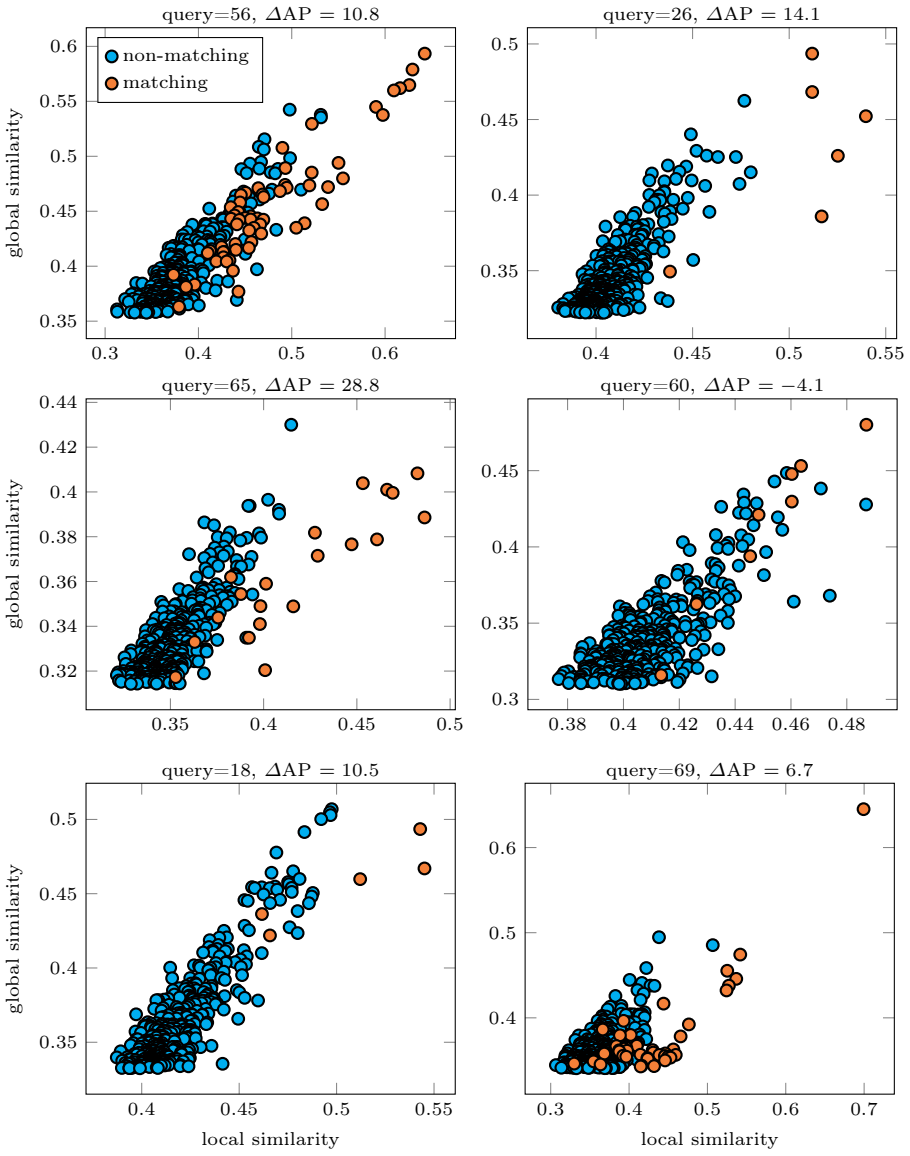


Fig. E: Comparison of global and local similarity. Each figure shows the top 400 retrieved images after the initial ranking for a single query from $\mathcal{R}$Oxf+1M in the hard setting. ΔAP represents the change in average precision for the specific query between using global similarities only and our ensemble similarities. Matching (positive) and non-matching (negative) are according to the query image ground truth.

γ \ λ	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
1e-4	28.8	31.0	31.0	31.0	31.0	31.0	31.0	31.0	31.0	31.0	31.0	31.0	31.0	31.0	31.0	31.0	31.0	31.0	31.0	31.0	31.0
1e-3	28.7	31.9	31.4	31.3	31.2	31.2	31.1	31.0	31.0	31.1	31.0	31.0	31.0	31.0	31.0	31.0	31.0	31.0	31.0	31.0	31.0
1e-2	28.7	32.6	32.9	32.6	32.3	32.1	32.0	31.9	31.8	31.6	31.5	31.4	31.4	31.3	31.2	31.2	31.1	31.0	31.1	31.0	31.0
1e-1	28.7	29.7	30.3	31.1	31.7	32.1	32.4	32.7	32.9	32.9	33.0	32.9	32.7	32.5	32.3	32.2	32.0	31.9	31.5	31.3	31.0
1e0	28.7	28.9	29.1	29.2	29.4	29.7	29.9	30.0	30.2	30.3	30.6	31.0	31.4	31.8	32.2	32.4	32.6	32.8	32.7	32.1	31.0
1e1	27.6	30.3	30.4	30.5	30.6	30.7	30.8	30.8	30.9	30.9	30.9	31.0	31.0	31.1	31.2	31.3	31.5	31.9	32.2	32.2	31.0

Table F: Global-local ensemble tuning via grid search on the validation set (GLDv2 public retrieval benchmark). Performance measured with mAP@100.

The local feature detector architecture is adopted from DOLG [12]. It consists of two 1×1 convolution layers followed by a Softplus activation. There is a BatchNorm [2] and ReLU in between the two convolutions. It is applied on top of a dense 3D activation map to provide a weight per local descriptor. For CVNet, the detector is applied on top of the third ResNet101 block output of 1024 dimensions (3D activation map depth). For DINOv2, the detector is applied on top of the ViT output, *i.e.* the set of patch tokens of 768 dimensions. During training, we use these local descriptor weights to extract a global descriptor by performing weighted average pooling and train with triplet loss. The triplets are sampled in the same manner as in AMES; the margin for the loss is set to 0.9. We train only the detector part on top of a frozen backbone for 20k iterations with a batch size of 10. We use AdamW optimizer and cosine learning rate scheduler with an initial value equal to 10^{-4}. During testing, we use the weights to select the L strongest local descriptors per image. This acts in the form of a classical local feature detector.

Estimation of mAP. We rely on two different implementations for estimating mAP on $\mathcal{R}$Oxford/$\mathcal{R}$Paris or mAP@100 on GLDv2. We choose the corresponding implementations that use the trapezoids for the area under the curve or average precision values, respectively, in our effort to better match the standard practice in the prior work.

Global-local ensemble parameter tuning. The optimal parameters for each (L_q^{test},L_x^{test}) setting are tuned independently. We conduct a hyper-parameter tuning via grid search on the validation set to find the values for λ and γ from the global-local ensemble and local similarity, respectively. We use the public split of the GLDv2 as the validation set for tuning and re-rank $m = 400$ images. Table F illustrates grid search results of an example run with L_q^{test}=600 and L_x^{test}=600. Higher values of λ are usually paired with higher values of γ. Consequently, even when λ is large, its most confident predictions are accounted for in the final ensemble. The differences for $\lambda = 0$ among different rows are due to the ties in the ranking of database images; many local similarities are either 0 or 1 with very large γ.

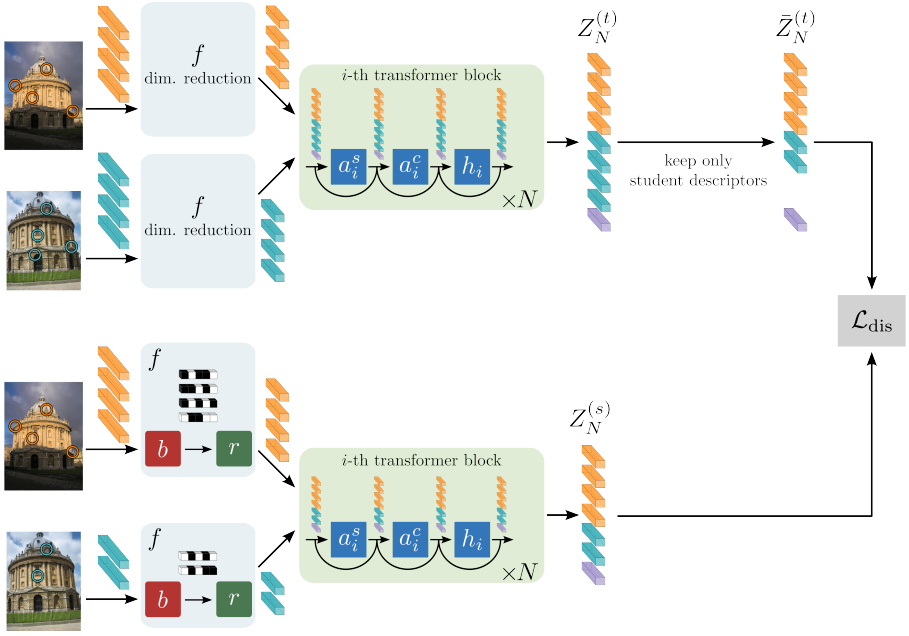


Fig. F: The distillation process of AMES. The teacher model (top) distills its output tokens to the student model (bottom) via the root mean squared error loss. The student operates in an asymmetric way on a subset of the local descriptors used by the teacher for the database image. The loss is applied to the intersection of the two sets. Function f refers to a different function for the teacher (dimensionality reduction by a linear layer) and the student (dimensionality reduction, binarization, and re-mapping to the real coordinate space). The distillation loss is combined with the binary cross-entropy loss.

Competing methods. We use the publicly available implementations for all the competitors and employed methods, *i.e.* ASMK[1], RRT[2], R^2Former[3], CVNet[4], and SuperGlobal[5]. For a fair comparison in our trade-off evaluation, we follow a similar tuning process for all competing approaches.

ASMK includes an internal quantization process that aggregates vectors per cell; therefore, the effective number of local descriptors that need to be stored is typically lower than the input number. Also, a visual word *id* needs to be stored for each effective descriptor, which amounts to 2 bytes per descriptor if

[1] `github.com/jenicek/asmk`
[2] `github.com/uvavision/RerankingTransformer`
[3] `github.com/bytedance/R2Former`
[4] `github.com/sungonce/CVNet`
[5] `github.com/ShihaoShao-GH/SuperGlobal`

stored as an unsigned integer, using delta coding. The total memory footprint is derived from the sum of the effective descriptor vectors and their word *ids*. The binary and simplified variant [10] is used, while local similarity is estimated for all the database images as this is more of an indexing than a re-ranking approach. As the input to ASMK, we perform dimensionality reduction down to 128 dimensions by PCA whitening [3] learned on the training set, while its internal representation is comprised of 128-bit vectors.

RRT and R^2Former are trained with our implementation framework with varying descriptor set sizes and the same projection function f at their input, as in our fp variant. In both cases, we use $d = 128$ for the reduction of local descriptors. We use a similar reduction layer for RRT on the global descriptors to map them to the same dimension space as the local ones. In this way, we train those on exactly the same input local descriptors, training dataset, and training process as AMES.

CVNet's memory footprint of the re-ranker is computed based on the quantized variant of the approach, as reported in Lee *et al*. [5] since it demonstrates a marginal performance drop compared to the full precision variant.

C Qualitative results

In Figure G, we show examples of hard positive images whose ranking is significantly improved using AMES for re-ranking. The most prominent examples include small objects among severe background clutter. We conclude that local similarity is essential in handling clutter, but current benchmarks only include a small number of such cases.

In Figure H, we illustrate several matching image pairs, *i.e.* a query and a database image, and visualize the locations and importance of the local descriptors for the local similarity estimation with our AMES model. We experiment with two different values for L_x^{test}, corresponding to symmetric and asymmetric matching. We measure the importance of the local descriptors based on the dot product similarity between the output matching token t_N and the output tokens X_N and Q_N of the two images. The network has learned to ignore the background descriptors that do not have a matching pair in both images. This is prevalent mainly in the limited memory settings, but it is also noticeable in the settings with more descriptors. Note that on the side of the query image, the importance of descriptors also changes between the symmetric and asymmetric settings, which reflects the availability of matching descriptors on the side of the database image. All images are from the GLDv2 test set.

D Limitations

We discuss limitations of AMES and possible improvements of those. (i) AMES relies on external sources/models for image descriptors, for both global and local. Hence, the quality of the employed descriptors is an important factor for the final performance, regardless of the performance gains AMES introduces.

Nevertheless, AMES is agnostic to the type of local descriptors and applicable in an off-the-shelf way. Training both AMES and the backbone representation in an end-to-end manner is possible, but we do not pursue this direction in this work. (ii) AMES lacks of modeling image geometry. It does not contain any mechanism that encodes the spatial structure of the images. We consider it a promising direction for future work for better generalization of the proposed approach. Our preliminary trials using conventional positional encodings do not bring any performance improvements.

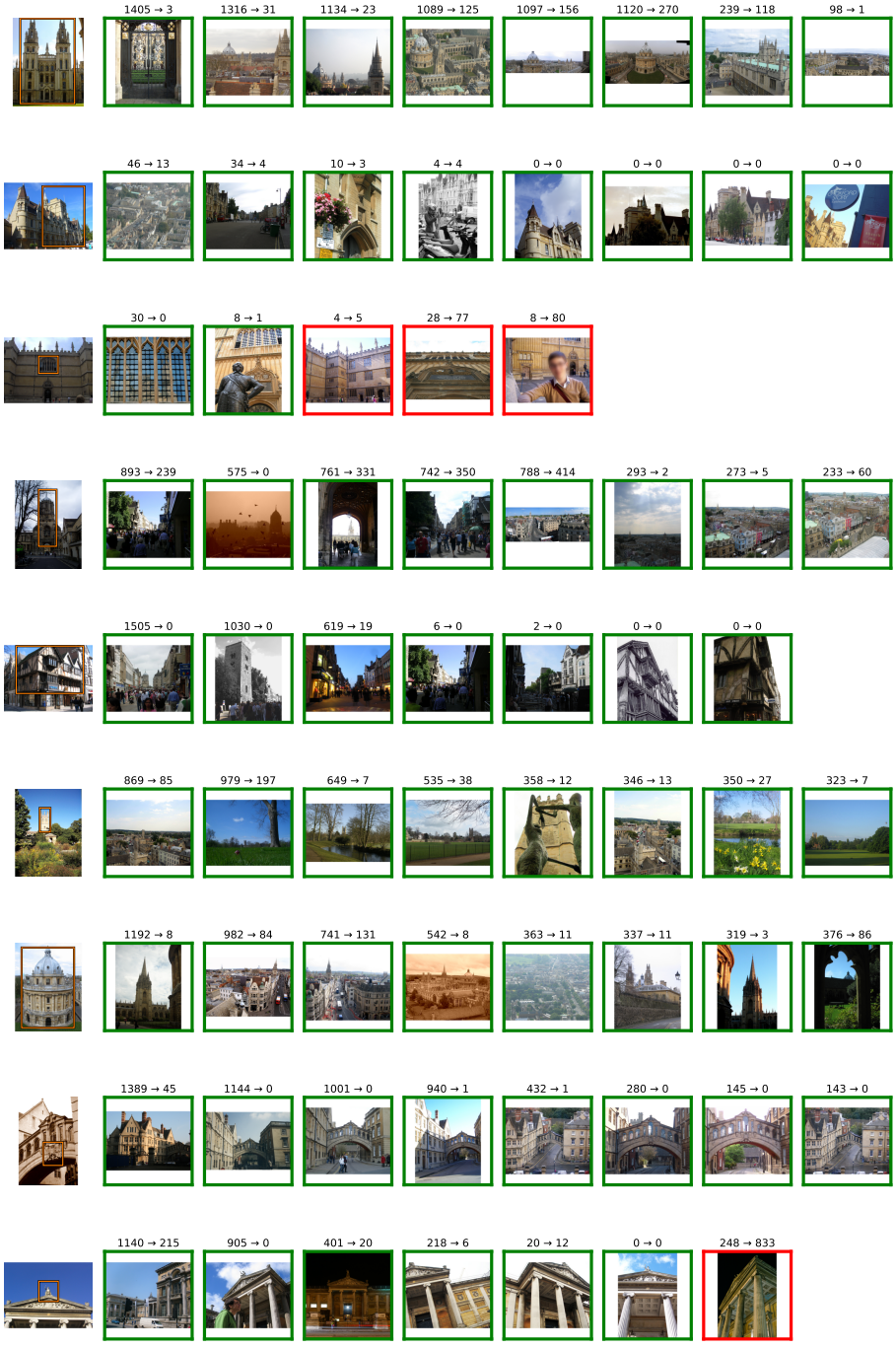
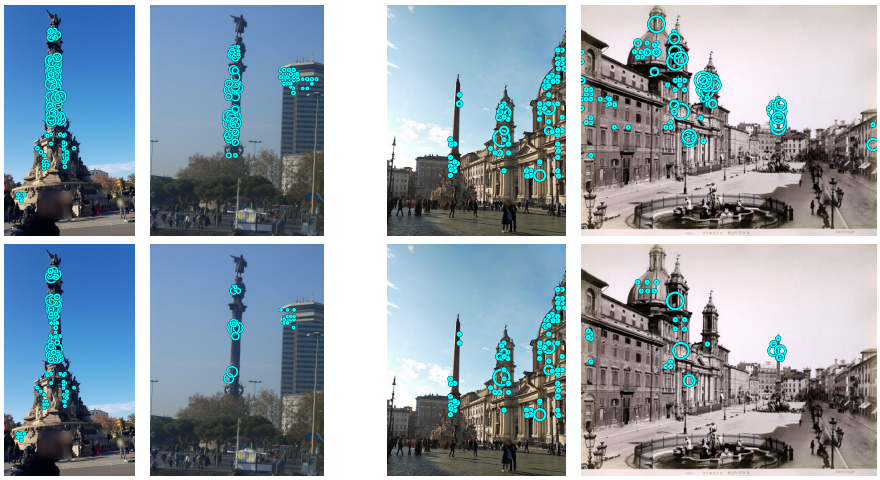
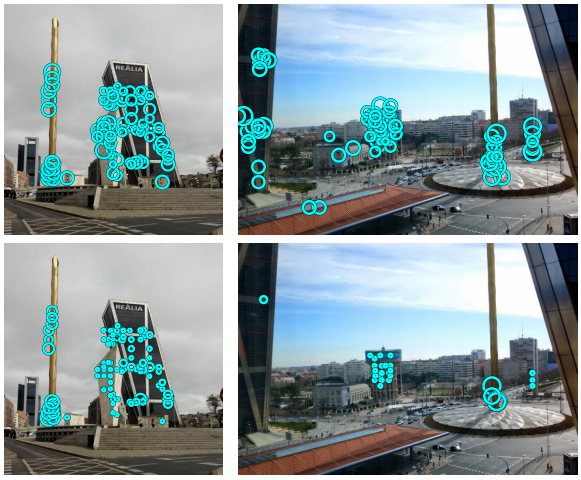


Fig. G: The impact of re-ranking with AMES. We show a query (region within the orange bounding box) per row and its hard positives from the database. We show the number of negative images ranked before the positive using only global similarity (1) and after re-ranking with AMES (2), by (1)→(2). The positives are ordered based on the difference (1)-(2) in descending order. Green (red) border denotes improved (harmed) re-ranking. Retrieval on $\mathcal{R}$Oxford +1M.



(a) (b)



(c)

Fig. H: Our model estimates the image pair similarity with a low number of local descriptors. Circle size reflects importance of the corresponding local descriptor within the model. Top: 100 (query) *vs*. 100 (database) local descriptors. Bottom: memory efficient asymmetric similarity with 100 *vs*. 30 local descriptors. Descriptors of the common object (other objects) are taken more (less) into account even with the lightweight and asymmetric variant.

References

1. Gong, Y., Lazebnik, S., Gordo, A., Perronnin, F.: Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. PAMI (2012)
2. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015)
3. Jégou, H., Chum, O.: Negative evidences and co-occurences in image retrieval: The benefit of pca and whitening. In: ECCV (2012)
4. Kordopatis-Zilos, G., Tzelepis, C., Papadopoulos, S., Kompatsiaris, I., Patras, I.: DnS: Distill-and-select for efficient and accurate video indexing and retrieval. IJCV (2022)
5. Lee, S., Seong, H., Lee, S., Kim, E.: Correlation verification for image retrieval. In: CVPR (2022)
6. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
7. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. In: arXiv (2024)
8. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: CVPR (2019)
9. Shao, S., Chen, K., Karpur, A., Cui, Q., Araujo, A., Cao, B.: Global features are all you need for image retrieval and reranking. In: ICCV (2023)
10. Tolias, G., Jenicek, T., Chum, O.: Learning and aggregating deep local descriptors for instance-level recognition. In: ECCV (2020)
11. Varis, D., Bojar, O.: Sequence length is a domain: Length-based overfitting in transformer models. In: arXiv (2021)
12. Yang, M., He, D., Fan, M., Shi, B., Xue, X., Li, F., Ding, E., Huang, J.: Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features. In: CVPR (2021)