LoCoRe: Image Re-ranking with Long-Context Sequence Modeling

Supplementary Material

A. Implementation details

All training is conducted on 8 NVIDIA A100 PCI-E 40GB GPUs. Training on Google Landmark v2 clean set [23] takes 106 hours on LoCoRe-base for 5 epochs. Models are trained with AdamW optimizer [10], 5e−5 learning rate and weight decay disabled. Global batch size is set to 128 with 4 gradient accumulation steps. We present the configurations of the different LoCoRe variants in Table A. LoCoRe-tiny is initialized from `roberta-tiny-cased`[1] by migrating weights and repeatedly copying absolute position embedding along the sequence dimension[2]. LoCoRe-small is initialized from the first 6 layers of `longformer-base-4096`[3], while LoCoRe-base is initialized from the full `longformer-base-4096`. To accommodate 50 descriptors $\times$ (1 query image + 100 reranking candidates) = 5,050 tokens, the position embeddings in the original models are linearly interpolated to extend their length from 4,096 to 5,120.

When experimenting with local descriptors from DINOv2 [12], we use the same training set as AMES [18], which is approximately half the size of the full GLDv2 clean set, *i.e*. 750k images. We adopt the same global-local ensemble scheme as AMES. The ensemble hyper-parameters are selected based on the best-performing configuration on GLDv2 public validation split and applied to $\mathcal{R}$Oxf and $\mathcal{R}$Par evaluations. For the training of AMES*, we follow the original training process from AMES, changing only the batch size and learning late to 150 and 1e−5, respectively.

For baseline results on CUB-200 [22], Stanford Online Products (SOP) [17] and In-shop [9], we reproduce them using their official code releases and identical training configuration, except for ProxyNCA++ [20], we change the training image size from 256 $\times$ 256 to 224 $\times$ 224 to use the training image size same as the other baselines.

For the performance benchmark in Section 4.3, we use the Deepspeed [15] profiler on a single NVIDIA A100 GPU to measure key performance metrics of the model per 100 reranked gallery images as follows: the number of parameters (# of Params), floating-point operations (FLOPs), throughput in FLOPs per second, latency, and peak memory consumption. All dynamic metrics are reported with 10 warmup steps followed by 10 measurements for reporting the mean and standard deviation. Parameters of visual backbones are excluded from # of Params.

We consider the descriptors to be already extracted and exclude I/O from measuring memory, latency, *etc*. For the geometric verification (GV) method, we run RANSAC in OpenCV [3] with 1,000 iterations on AMD EPYC 9354 CPU and measure the wall-clock time as the latency and the maximum resident set size (Max RSS) as the peak memory consumption. All models are benchmarked with batched input except CVNet Reranker [5]. It is worth noting that CVNet Reranker does not support batched inference since it computes pair-wise multi-scale correlation on raw feature maps (without resizing) from query and gallery images of different sizes. Thus, CVNet Reranker heavily underutilizes the GPU and achieves extremely low throughput and high latency. The FLOP, latency, and peak memory are measured assuming query and gallery images of 512 $\times$ 512 size in CVNet Reranker.

Model Variants	tiny	small	base
Number of Parameters	19.4M	58.7M	111.8M
Number of Layers	4	6	12
Local Attention Window	1024	512	512
Hidden Size	512	768	768
Intermediate Size	2048	3072	3072
Number of Attention Heads	8	12	12
Max Context Length	5120	5120	5120

Table A. **Architectural parameters** of LoCoRe variants.

B. Additional Experimental Results

B.1. Additional comparisons

We present additional experiments with different combinations of global and local features in Table B. We compare with more baseline re-ranking methods, including methods with global, *i.e*. SuperGlobal (SG) Rerank [16], and local, *i.e*. AMES [18], RRT [19], R2Former [24], descriptors. We evaluate the models under different Hard settings, using different global descriptors to generate the shortlist and different backbones for feature extraction. We also test the combination of LoCoRe with other re-ranking schemes.

Variations for Hard setup. As mentioned in the main paper, there can be two approaches regarding how to handle *easy* images in the hard setup: (i) **Hard**: *easy* images are used to re-rank and removed before the evaluation (typically used in the literature [16]), and (ii) **Hard***: *easy* images are completely removed from the database. While the two choices (Hard and Hard*) are equivalent for pair-wise

[1] https://huggingface.co/haisongzhang/roberta-tiny-cased

[2] https://github.com/allenai/longformer/blob/master/scripts/convert_model_to_long.ipynb

[3] https://huggingface.co/allenai/longformer-base-4096

Global	Local	Re-rank	$\mathcal{R}$Oxf+1M			$\mathcal{R}$Par+1M		
			Medium	Hard	Hard*	Medium	Hard	Hard*
SG	N/A	N/A†	78.8	61.9		83.9	69.1	
		N/A	78.5	61.4		83.6	68.4	
		SG-Rerank†	84.4	71.1	N/A	84.9	71.4	N/A
		SG-Rerank	84.0	69.4	63.9	85.2	72.3	75.7
	CVNet	R2Former	79.9	63.7		83.8	69.7	
		RRT	79.3	62.7		83.6	69.1	
		AMES	80.7	65.7		84.6	71.8	
		LoCoRe	81.9	68.6	64.9	84.6	71.4	70.7
		SG + LoCoRe	84.7	71.5	65.6	86.2	74.8	76.1
	DINOv2	R2Former*	81.0	66.2		84.9	72.1	
		RRT*	81.0	66.1		85.5	73.3	
		AMES*	81.3	67.3		85.8	74.3	
		LoCoRe	85.8	75.8	73.2	86.8	75.9	76.5
		SG + LoCoRe	86.5	76.3	73.7	87.2	76.9	78.2
DINOv2	N/A	N/A	59.6	35.2		77.0	58.9	
		SG-Rerank	62.2	40.5	31.2	79.8	60.5	65.8
	DINOv2	R2Former*	67.8	44.6		78.6	61.3	
		RRT*	68.8	46.0		79.6	64.0	
		AMES*	68.9	46.8		79.9	64.7	
		LoCoRe	73.4	54.9	52.5	80.9	66.4	66.7
		SG + LoCoRe	71.2	54.4	48.7	81.9	68.7	69.5

Table B. **Additional results** with re-ranking top-400 candidates. Hard*: *easy* images are completely removed from the database. Hard: *easy* images are used to re-rank and removed before the evaluation. †: results in the SuperGlobal paper [16]. LoCoRe is reported with the base variant. SG + LoCoRe: re-ranking with SG first and then with LoCoRe. * indicates models trained with 768 hidden size, serving as a fair comparison with LOCORE. N/A: not available.

re-ranking methods, this is not the case when interactions between database images are considered (*i.e*. LoCoRe, SG-rerank). In Table B, it is evident that the two setup lead to significantly different results. In most cases, mAP considerably drops in $\mathcal{R}$Oxf, comparing results in Hard and Hard*; whereas, mAP increases in $\mathcal{R}$Par.

Performance with other backbones. First, we benchmark all models when the shortlist is generated based on DINOv2 global descriptors. It is noteworthy that DINOv2 global descriptors are significantly worse than SG ones. In this setup, LoCoRe outperforms all other re-ranking schemes by a vast margin.

Second, we evaluate LoCoRe using local descriptors extracted from CVNet backbones. CVNet local descriptors have a higher dimension than that of DINOv2, *i.e*. 1024 vs 768; hence, we used a learnable linear projector to match the embedding dimensionality of the transformer. LoCoRe achieves competitive performances compared with the pair-wise re-rankers, with only AMES outperforming it in a few cases. Yet, all local-based re-rankers are outperformed by SG-Rerank. Nevertheless, LoCoRe with DINOv2 outperforms SG-Rerank.

Combination with SG-Rerank. It is straightforward to combine local and global-based re-ranking. To this end, we combine LoCoRe with SG-Rerank by applying global re-ranking first, followed by local re-ranking. This combination achieves the best performance when SG is used as global descriptor. However, this combination hurts LoCoRe performance on $\mathcal{R}$Oxf when DINOv2 is used as global.

Performance per query. To highlight the advantages of our proposed list-wise re-ranking over pair-wise re-ranking, we present several scatter plots in Figure A, showing the average precision of each sample in $\mathcal{R}$Oxf+1M Hard before and after re-ranking with different re-ranking paradigms. We compare our model with AMES [18], which is considered the state-of-the-art solution for pair-wise re-ranking. In the first two plots, we observe that most data points are concentrated in the upper-left half and above the red reference line, indicating that both re-ranking paradigms improve the ranking accuracy for the majority of query images. However, the list-wise re-ranking method driven by LoCoRe has barely any sample points below the red reference line, meaning

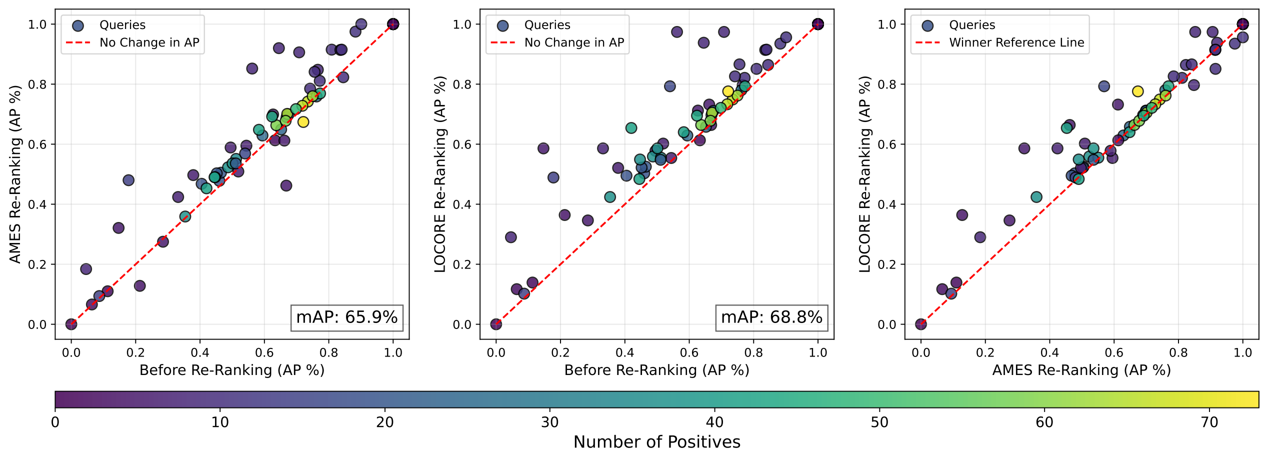

Figure A. **Average precision per query scatter plot** on $\mathcal{R}$Oxf+1M Hard for global-only vs. AMES (*Left*), global-only vs. LoCoRe-small (*Middle*) and AMES vs. LoCoRe-small (*Right*). Global descriptors are from RN101-Superglobal, which by itself achieves mAP=61.4%. Re-ranking is performed for top-100 candidates, and the color bar indicates the number of positive images in the shortlist for each query.

Global	Local	L	K	$\mathcal{R}$Oxf+1M Medium	$\mathcal{R}$Oxf+1M Hard*	$\mathcal{R}$Par+1M Medium	$\mathcal{R}$Par+1M Hard*
SG	DINOv2	50	100	85.8	73.2	86.8	76.5
		100	50	83.9	68.6	85.2	72.5
		25	200	84.5	72.1	85.6	75.1

Table C. Additional results for LoCoRe-base with different combinations of the number of local descriptors L and the number of re-ranking candidates K on $N = 400$ candidates.

Ablation Module	$\mathcal{R}$@1	$\mathcal{R}$@10	mAP@R
Global descriptors	80.8	92.1	65.1
LoCoRe-tiny	82.4	93.1	68.0
LoCoRe-small	83.3	92.7	69.4
LoCoRe-base	83.8	92.9	71.0
LoCoRe-RWKV	81.4	92.3	66.7
LoCoRe-Mamba	80.6	92.1	66.4

Table D. Ablation studies for LoCoRe recurrent models on the SOP dataset. Re-ranking is performed with the top 100 candidates.

the re-ranking only improves the retrieval on the individual query level. The distinction between the two models is most prominent in the final plot, where the number of sample points above the winner reference line far exceeds those below, demonstrating that LoCoRe outperforms AMES on more query samples. We also observed that the list-wise re-ranking method is relatively robust in terms of the number of positive samples included in the shortlist, as the color distribution of the sample points does not exhibit any discernible pattern. This indicates the general versatility of LoCoRe.

B.2. Additional ablations

Number of images vs number of descriptors. We explore the relationship between the number of local descriptors and the number of image candidates within a given context window in Table C. Specifically, we set the context window to 5,120 and examine three configurations of LoCoRe: (i) using 100 gallery images with 50 local descriptors per image, *i.e.* the default setup, (ii) using 200 gallery images with 25 local descriptors per image, *i.e.* more candidate images but fewer descriptors per image, and (iii) using 50 database images with 100 local descriptors per image, *i.e.* more descriptors per image but fewer candidate images. The LoCoRe in the default settings reports the best results.

Comparison with other recurrent models. Other model architectures with no restrictions on context length that could be employed instead of LongFormer are the recently proposed recurrent models Mamba [2] and RWKV [13]. As the causal nature of the recurrence-based model does not align well with our re-ranking motivation and is strictly less expressive than bi-directional encoders [4, 14], we follow the common practice in recurrent visual encoder community [1, 6, 8] to build a bi-directional variant that serves as an efficient sequence encoder. To ensure recurrent models can still handle long-range interactions and alleviate the inherent information bottleneck in the design of recurrent models, we devise a mechanism that resembles the query global attention in Section 3.2 by interleaving recurrent blocks with uni-directional transformer blocks [21]. These transformer blocks compute attention scores between intermediate hidden states of query image tokens and intermediate hidden states of gallery image tokens and produce fused intermediate representations for the following layers to process. The uni-directional attention guarantees that every gallery image has similar difficulty accessing the query image, irrespective of its position in the sequence relative to the query. Although we find that these recurrence-based models could slightly outperform the base global retrieval model, they do not surpass our transformer-based results, as shown in Table D.

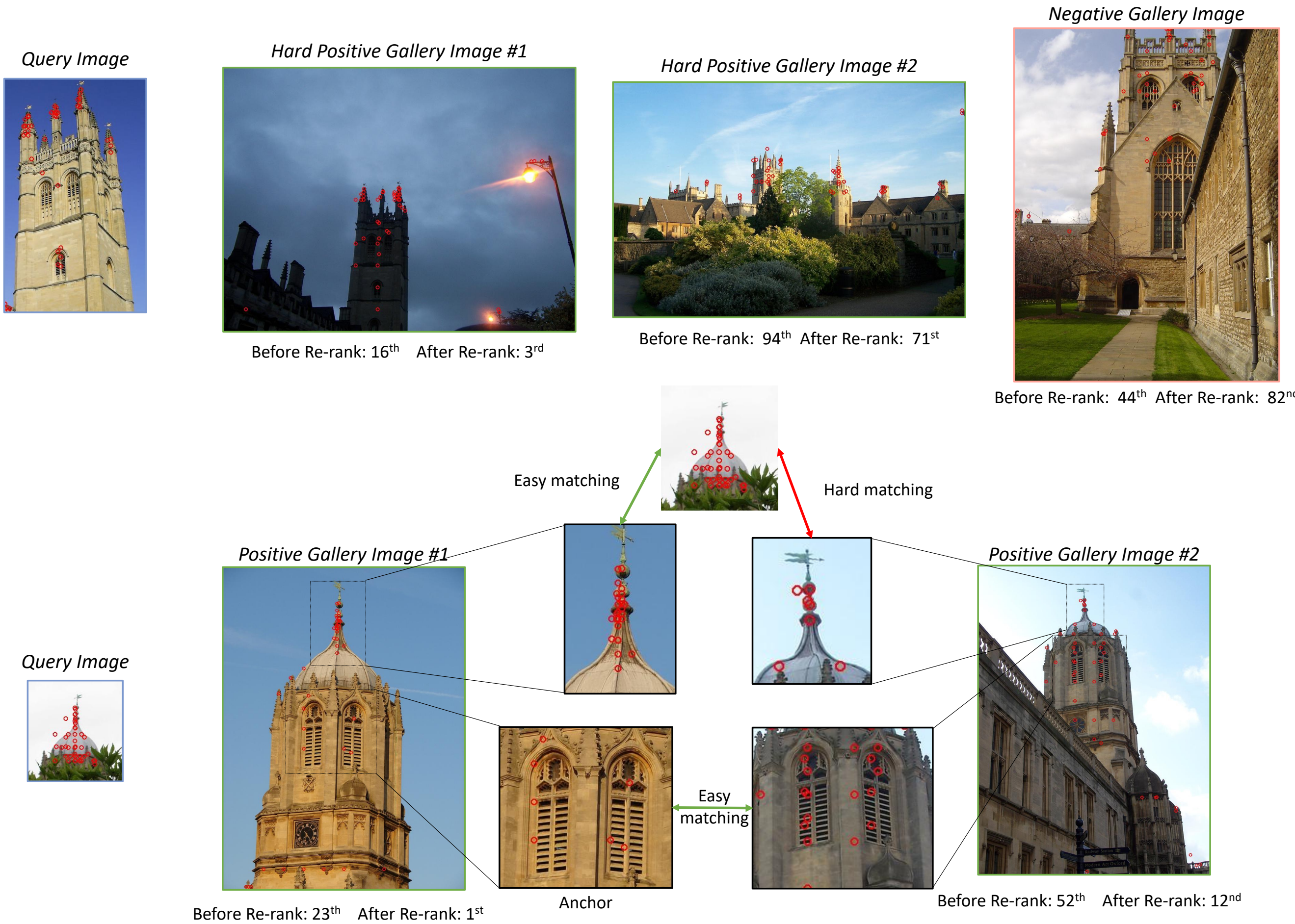

Figure B. **Qualitative analysis** on $\mathcal{R}$Oxford dataset of LoCoRe-base on RN50-DELG descriptors. **Upper:** two hard positive gallery images get assigned with higher ranks while a negative gallery image is put in lower ranks after re-ranking. **Lower:** the first gallery image can be easily identified as positive due to its dense matching with the query image; it can also serve as a perfect anchor image for refining the ranking of the second gallery image due to their transitive relationship.

B.3. Qualitative Results

We illustrate the re-ranking performance of LoCoRe in Figure B as qualitative results. The upper example underscores the superior performance of our method, demonstrated by its success in elevating the ranking of two hard positive images and lowering that of the negative gallery image. We also show in the lower example that our model is able to capture the transitive relationship between query and gallery images. The transitive relationship is based on the assumption that generally, if two gallery images are similar and one of them is predicted as positive, then the other should be calibrated with higher confidence. In the lower example, the correspondence between the query image and the first gallery image is easy to catch as the common geometric features are evident, resulting in Easy matching in the figure. However, although the global retriever returns the second gallery image as re-ranking candidates, the sparse local features focused on the top of the tower make it hard for pair-wise re-ranker to assign this gallery image a high confidence score. This misalignment is calibrated by our list-wise re-ranking paradigm since the windows in both gallery images can serve as the anchor to propagate the positive prediction from the easy candidate to the hard one.

Additionally, in Figure C the easy positive gallery has visual overlap with the query (rooftop). The hard positive gallery has little visual overlap with the query, but larger overlap with the first positive (*e.g.* windows). We wish to answer this question: *Are the local features of the window improving the rank of the hard positive due to a transitive relationship?* We remove local features of the windows (blue crosses), repeat the similarity estimation, and compare the ranks. The decreased similarity score is a sign of LoCoRe capturing transitive relationships.

Figure C. **Visualization of LoCoRe capturing transitive relationships in gallery images.** We prevent LoCoRe from accessing local features of the easy positive corresponding to the windows (blue crosses) and instead randomly sample local features from other negative images. The dropped similarity score indicates LoCoRe relies on the transitivity of local features to calibrate predictions for hard positive gallery images.

C. Limitations and Future Work

Despite the merits in efficiency and re-ranking performance, our model is inherently restricted by the context window of existing encoder-only sequence models. A limited context window limits the number of re-ranking candidates in the gallery and the number of local descriptors that LoCoRe can use. While recurrent models offer more flexibility with the context window size, we find that they could not capture list-wise re-ranking dependencies as well as transformer-based models, resulting in sub-optimal performance. Future work could adopt large-scale decoder-only sequence models which typically have longer context windows and greater capacity for list-wise re-ranking. Additionally, context parallelization techniques (*e.g.*, RingAttention [7], Infini-attention [11]) could help expand the context window of current Transformer encoder models. Lastly, extractive re-ranking as proposed in our work could also be seamlessly adopted for other modalities, *e.g*. document or video re-ranking.

References

[1] Yuchen Duan, Weiyun Wang, Zhe Chen, Xizhou Zhu, Lewei Lu, Tong Lu, Yu Qiao, Hongsheng Li, Jifeng Dai, and Wenhai Wang. Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures, 2024. 3

[2] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv: 2312.00752*, 2023. 3

[3] Itseez. Open source computer vision library. https://github.com/opencv/opencv, 2015. 1

[4] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 4904–4916. PMLR, 2021. 3

[5] Seongwon Lee, Hongje Seong, Suhyeon Lee, and Euntai Kim. Correlation verification for image retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5364–5374. IEEE, 2022. 1

[6] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding, 2024. 3

[7] Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. *arXiv preprint arXiv: 2310.01889*, 2023. 5

[8] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model, 2024. 3

[9] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 1

[11] Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. Leave no context behind: Efficient infinite context transformers with infini-attention. *arXiv preprint arXiv: 2404.07143*, 2024. 5

[12] M. Oquab, Timoth'ee Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, H. Jégou, J. Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2023. 1

[13] Bo Peng, Eric Alcaide, Quentin Gregory Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Nguyen Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Kiran GV, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan S. Wind, Stanisław Woźniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. RWKV: Reinventing RNNs for the transformer era. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 3

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 3

[15] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3505–3506. ACM, 2020. 1

[16] Shihao Shao, Kaifeng Chen, Arjun Karpur, Qinghua Cui, André Araujo, and Bingyi Cao. Global features are all you need for image retrieval and reranking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11036–11046, 2023. 1, 2

[17] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[18] Pavel Suma, Giorgos Kordopatis-Zilos, Ahmet Iscen, and Giorgos Tolias. Ames: Asymmetric and memory-efficient similarity estimation for instance-level retrieval. In *ECCV*, 2024. 1, 2

[19] Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. Instance-level image retrieval using reranking transformers. *IEEE International Conference on Computer Vision*, 2021. 1

[20] Eu Wern Teh, Terrance Devries, and Graham W. Taylor. Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis. *European Conference on Computer Vision*, 2020. 1

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[22] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology, 2011. 1

[23] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2 - a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[24] Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. R2former: Unified retrieval and reranking transformer for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19370–19380, 2023. 1