

TAT-VPR: Ternary Adaptive Transformer for Dynamic and Efficient Visual Place Recognition

Oliver Grainge¹, Michael Milford², Indu Bodala¹, Sarvapali D. Ramchurn¹ and Shoaib Ehsan^{1,3}

Abstract—TAT-VPR is a ternary-quantized transformer that brings dynamic accuracy–efficiency trade-offs to *visual SLAM* loop-closure. By fusing ternary weights with a learned activation-sparsity gate, the model can control computation by up to 40% at run-time without degrading performance (Recall@1). The proposed two-stage distillation pipeline preserves descriptor quality, letting it run on micro-UAV and embedded SLAM stacks while matching state-of-the-art localization accuracy.

Index Terms—Visual Place Recognition, SLAM, Quantization

I. INTRODUCTION & BACKGROUND

Visual Place Recognition (VPR) is often formulated as an image-retrieval task, matching a query image to a geotagged image database. State-of-the-art methods use foundation-scale Vision Transformer (ViT) global descriptors [1]–[3], which are robust to viewpoint, lighting, and seasonal changes. However, their high computational and memory demands limit their use on low-power mobile robots, especially for real-time SLAM loop closure [4], [5]. Consequently, many lightweight SLAM systems still rely on hand-crafted or aggregated point features, sacrificing the robustness of modern transformers [6], [7].

To mitigate this, the field has turned to neural network compression. Binary and low-bit quantization can push precision below 8-bit with minimal accuracy loss [4], [8], [9], while pruning and sparsity reduce redundancy without degrading performance [5]. In NLP, methods like Q-Sparse [10] and BitNet-4.8a [11] show that sub-4-bit quantization and activation sparsity are compatible. Task-aware knowledge distillation further enhances efficiency by transferring capacity from large ViT teachers to smaller students [12]. Yet, these approaches are typically fixed to one accuracy-efficiency tradeoff: once deployed, they can’t adapt to changing conditions such as low power or high speeds in cluttered environments.

We propose TAT-VPR, an end-to-end extreme-quantization pipeline (Figure 1) that combines ternary weight quantization, adaptive activation gating, and teacher–student distillation.

This work was supported by the UK Engineering and Physical Sciences Research Council through grants EP/Y009800/1 and EP/V00784X/1

¹O. Grainge, I. Bodala, S. D. Ramchurn and S. Ehsan are with the School of Electronics and Computer Science, University of Southampton, United Kingdom (email: oeg1n18@soton.ac.uk; i.p.bodala@soton.ac.uk; sdrl@soton.ac.uk; s.ehsan@soton.ac.uk).

²M. Milford is with the School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, QLD 4000, Australia (email: michael.milford@qut.edu.au).

³S. Ehsan is also with the School of Computer Science and Electronic Engineering, University of Essex, United Kingdom, (email: sehsan@essex.ac.uk).

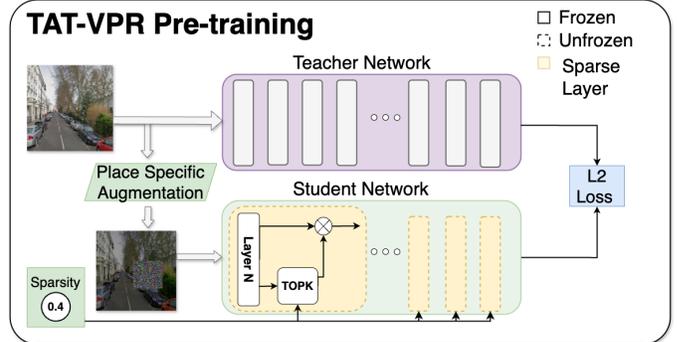


Fig. 1. Overview of the TAT-VPR pre-training pipeline. A full-precision DINOv2-BoQ teacher [3] (purple, frozen) provides token-level supervision to a ternary student transformer (green). During training, the student applies a top-k sparse activation filter. A distillation loss is computed between teacher and student tokens to guide compression-aware representation learning.

The result is a model that can dynamically control inference cost by sparsifying activations on demand. On standard VPR benchmarks, TAT-VPR achieves under 1% Recall@1 drop compared to the dense model while dynamically cutting inference operations by 40% and shrinking model size 5×, offering a practical, adaptive solution for resource-aware visual localization.

II. METHOD

Our model leverages a ViT-Base architecture for global image–descriptor extraction, modified with ternary weights and controllable sparse activations. The core components are outlined below.

A. Ternary Quantized ViT Backbone

Every weight tensor is quantized to the ternary set $\{-1, 0, +1\}$ with absolute mean quantization:

$$\widetilde{\mathbf{W}} = \text{RoundClip}\left(\frac{\mathbf{W}}{\gamma + \varepsilon}, -1, 1\right) \quad (1)$$

where $\gamma = \frac{1}{MD} \|\mathbf{W}\|_1$ is the absolute mean of the tensor and $\varepsilon = 10^{-6}$ prevents division by zero. RoundClip rounds to the nearest integer then clips to the interval $[-1, 1]$. Resulting in a 8× memory saving compared to 32-bit floating point.

B. Activation-Sparsity Scheduling

To let the model trade accuracy for efficiency at run time, we apply a *top-k* activation mask. Given activations $\mathbf{X} \in \mathbb{R}^{N \times D}$,

$$\mathbf{M} = \text{TopK}(|\mathbf{X}|, k), \quad \mathbf{Y} = (\mathbf{X} \odot \mathbf{M}) \widetilde{\mathbf{W}}^T, \quad (2)$$

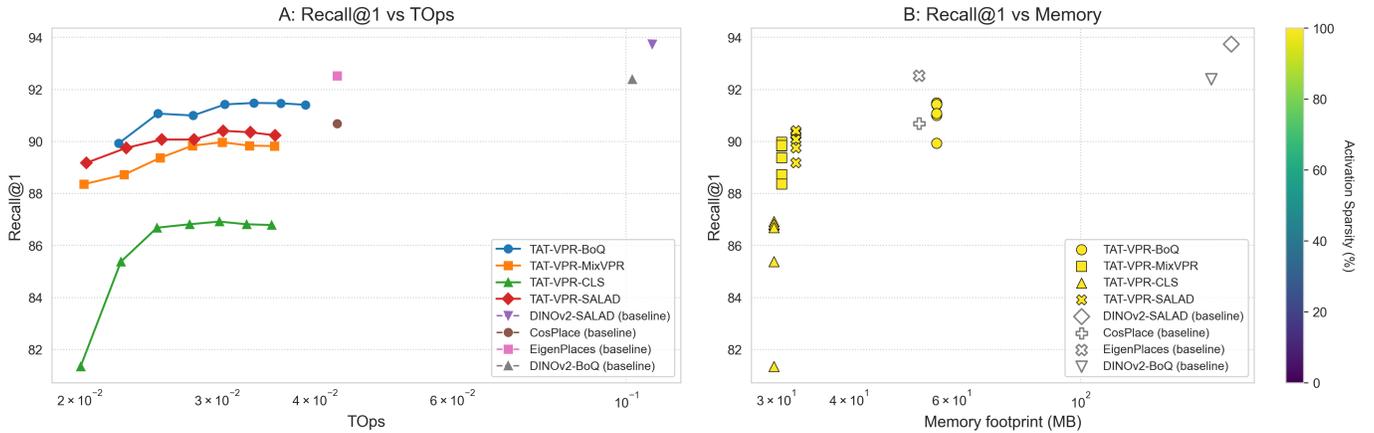


Fig. 2. (A) Recall@1 versus Tera-Operations (TOPs) for a feature-extraction forward pass, showing TAT-VPR curves at activation sparsity levels from 0% up to 60%. (B) Recall@1 versus memory footprint on the Pitts30k dataset, highlighting memory savings from ternary-weight backbones.

where the binary mask $\mathbf{M} \in \{0, 1\}^{N \times D}$ keeps the largest-magnitude $k\%$ entries in \mathbf{X} and \odot denotes the hadamard product. Because zeroed elements of \mathbf{X} can be skipped by sparse matrix kernels, only $k\%$ of the usual multiply-accumulate operations are executed, yielding proportional savings in latency, energy, and TOPs (Tera Operations).

C. Knowledge Distillation with a BoQ Teacher

Extreme quantization and sparsity inevitably limit representational capacity, so we pre-train the student with guided distillation from a full-precision DINOv2-BoQ teacher [3]. We use a single token-level mean-squared error loss on the output tokens:

$$\mathcal{L}_{\text{distill}} = \frac{1}{ND} \|\mathbf{S}^{(L)} - \mathbf{T}^{(L)}\|_2^2, \quad (3)$$

where $\mathbf{T}^{(L)}, \mathbf{S}^{(L)} \in \mathbb{R}^{N \times D}$ are the teacher and student tokens at the final layer L . This objective suffices to recuperate the accuracy lost to ternarisation and sparsity (2). During this stage we linearly raise the sparsity sampling range of k from 10% to 60%, forcing the network to concentrate information in sparse activations whilst avoiding under-training.

D. Fine-tuning

We fine-tune the pre-trained compact backbone on the GSV-CITIES dataset using a supervised multi-similarity retrieval loss [13]. Four aggregation heads are evaluated on top of the frozen sparse ternary backbone: Bag-of-learnable queries, SALAD, MixVPR, and a lightweight classification token head [1], [3], [14]. Only the head and the last two backbone layers are updated to preserve low-precision representations and prevent overfitting.

III. RESULTS AND DISCUSSION

Figure 2-A shows the Recall@1 versus computational cost (in TOPs) curves for TAT-VPR, where computational cost is controlled via runtime activation sparsity. Across all aggregation heads, up to a 40% reduction in TOPs can be achieved with less than a 1% loss in Recall@1 on the Pitts30k dataset.

TABLE I
RECALL@1 / RECALL@1 PER MB ON SVOX CONDITION SPLITS.

Method	Snow	Rain	Overcast	Night	Sun
TAT-BoQ	97.0 / 1.71	94.2 / 1.66	97.7 / 1.72	61.5 / 1.08	92.7 / 1.63
TAT-MixVPR	90.6 / 2.95	87.5 / 2.85	93.5 / 3.04	35.2 / 1.14	82.2 / 2.67
TAT-CLS	78.7 / 2.63	75.2 / 2.52	89.7 / 3.00	20.0 / 0.67	66.5 / 2.23
TAT-SALAD	95.3 / 2.93	92.5 / 2.84	96.7 / 2.97	41.6 / 1.28	88.1 / 2.71
DINOv2-SALAD	99.4 / 0.30	98.7 / 0.29	98.5 / 0.29	97.8 / 0.29	97.7 / 0.29
DINOv2-BoQ	98.7 / 0.27	98.5 / 0.27	98.3 / 0.27	95.4 / 0.26	97.1 / 0.27
CosPlace	90.3 / 0.85	85.1 / 0.80	91.4 / 0.86	48.6 / 0.46	76.9 / 0.73
EigenPlaces	91.5 / 0.86	88.0 / 0.83	92.5 / 0.87	59.8 / 0.56	85.2 / 0.80

In contrast, baseline models including [1], [3], [15], [16] rely on a fixed and higher volume of TOPs.

Figure 2-B presents the static memory consumption of TAT models compared to baselines. Despite employing larger backbones in terms of parameter count, most TAT models (excluding TAT-BoQ) consume significantly less memory. This efficiency arises from the use of 2-bit ternary quantized weights in the TAT models.

Table I demonstrates the robustness of TAT models relative to baselines under various appearance change conditions. Due to the distillation of generalizable representations during the pre-training stage, both TAT-BoQ and TAT-SALAD achieve higher recall scores than convolutional baselines such as [15], [16], even when operating at 40% activation sparsity. Moreover, across all datasets, TAT models achieve markedly superior memory efficiency despite the 40% activation sparsity constraint.

IV. CONCLUSION

TAT-VPR brings dynamic scalability to *visual SLAM*: its ternary weights and adaptive activation sparsity let a single network down-shift latency and power on the fly, yet still deliver near-state-of-the-art Recall@1 for loop-closure detection. The 5× memory cut and 40% TOPs savings free headroom for tracking, mapping, and relocalisation threads on micro-UAV and embedded SLAM stacks.

REFERENCES

- [1] S. Izquierdo and J. Civera, "Optimal transport aggregation for visual place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.
- [2] N. Keetha, A. Mishra, J. Karhade, K. M. Jatavallabhula, S. Scherer, M. Krishna, and S. Garg, "Anyloc: Towards universal visual place recognition," *IEEE Robotics and Automation Letters*, vol. 9, no. 2, pp. 1286–1293, 2024.
- [3] A. Ali-Bey, B. Chaib-draa, and P. Giguère, "Boq: A place is worth a bag of learnable queries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 794–17 803.
- [4] O. Grainge, M. Milford, I. Bodala, S. D. Ramchurn, and S. Ehsan, "Design space exploration of low-bit quantized neural networks for visual place recognition," *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 5070–5077, 2024.
- [5] —, "Structured pruning for efficient visual place recognition," *IEEE Robotics and Automation Letters*, vol. 10, no. 2, pp. 2024–2031, 2025.
- [6] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE transactions on robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.
- [7] A. Rosinol, M. Abate, Y. Chang, and L. Carlone, "Kimera: an open-source library for real-time metric-semantic localization and mapping," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 1689–1696.
- [8] B. Ferrarini, M. J. Milford, K. D. McDonald-Maier, and S. Ehsan, "Binary neural networks for memory-efficient and effective visual place recognition in changing environments," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2617–2631, 2022.
- [9] —, "Binary neural networks for memory-efficient and effective visual place recognition in changing environments," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2617–2631, 2022.
- [10] H. Wang, S. Ma, R. Wang, and F. Wei, "Q-sparse: All large language models can be fully sparsely-activated," 2024. [Online]. Available: <https://arxiv.org/abs/2407.10969>
- [11] H. Wang, S. Ma, and F. Wei, "Bitnet a4.8: 4-bit activations for 1-bit llms," 2024. [Online]. Available: <https://arxiv.org/abs/2411.04965>
- [12] G. Peng, Y. Huang, H. Li, Z. Wu, and D. Wang, "Lsdnet: A lightweight self-attentional distillation network for visual place recognition," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 6608–6613.
- [13] A. Ali-bey, B. Chaib-draa, and P. Giguère, "Gsv-cities: Toward appropriate supervised visual place recognition," *Neurocomputing*, vol. 513, pp. 194–203, 2022.
- [14] A. Ali-Bey, B. Chaib-Draa, and P. Giguere, "Mixvpr: Feature mixing for visual place recognition," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 2998–3007.
- [15] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geolocalization for large-scale applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 4878–4888.
- [16] G. Berton, G. Trivigno, B. Caputo, and C. Masone, "Eigenplaces: Training viewpoint robust models for visual place recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 11 080–11 090.