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

Oliver Grainge[1], Michael Milford[2], Indu Bodala[1], Sarvapali D. Ramchurn[1], Shoaib Ehsan[1,3]

[1]University of Southampton, UK | [2]Queensland University of Technology, Australia | [3]University of Essex, UK

Abstract & Motivation

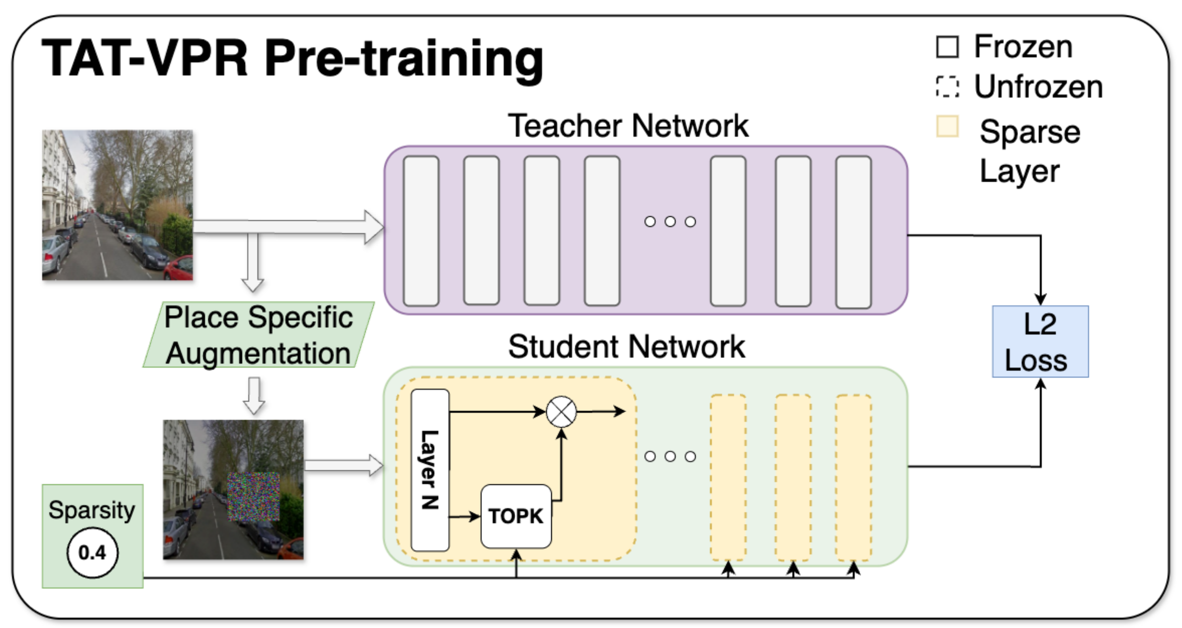
Problem: State-of-the-art Visual Place Recognition (VPR) methods use large Vision Transformers that are too computationally expensive for real-time SLAM on mobile robots and micro-UAVs.

Solution: TAT-VPR delivers dynamic accuracy-efficiency trade-offs through:

- **Ternary weight quantization** ({-1, 0, +1}) for 8× memory reduction
- **Adaptive activation sparsity** for runtime computational control
- **Two-stage distillation** to preserve descriptor quality

Key Results: 40% computation reduction with <1% accuracy loss, enabling deployment on resource-constrained platforms.

Method Overview



[FIGURE 1: TAT-VPR Training Pipeline] *Full-precision DINOv2-BoQ teacher (purple, frozen) provides token-level supervision to ternary student transformer (green). Student applies top-k sparse activation filter during training with distillation loss computed between teacher and student tokens.*

Three-Stage Pipeline

Stage 1: Ternary Quantization

- Convert all weights to ternary values {-1, 0, +1}
- Absolute mean quantization: $\tilde{W} = \mathrm{RoundClip}(W/\gamma, -1, 1)$
- Achieves 8× memory savings vs. 32-bit floating point

Stage 2: Knowledge Distillation

- Full-precision DINOv2-BoQ teacher supervises ternary student
- Token-level MSE loss: $\mathcal{L}_{distill} = \left\|S^l - T^l\right\|_2^2$
- Sparsity sampling from 10% to 60% during training

Stage 3: Fine-tuning

- Supervised training on GSV-Cities dataset
- Multiple aggregation heads: BoQ, SALAD, MixVPR, CLS
- Only head + last 2 layers updated to avoid overfitting

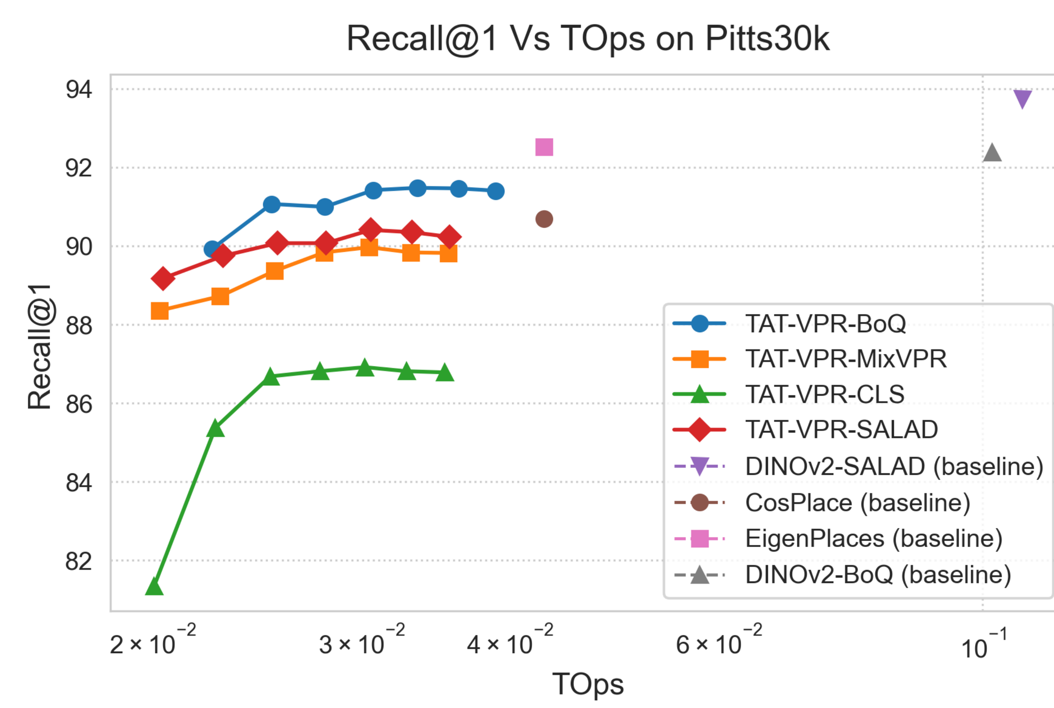
Key Technical Innovations

Ternary Weight Quantization
Memory Footprint: 32-bit → 2-bit (8× reduction) Quantization:
$W \in \mathbb{R} \rightarrow \tilde{W} \in \{-1, 0, +1\}$

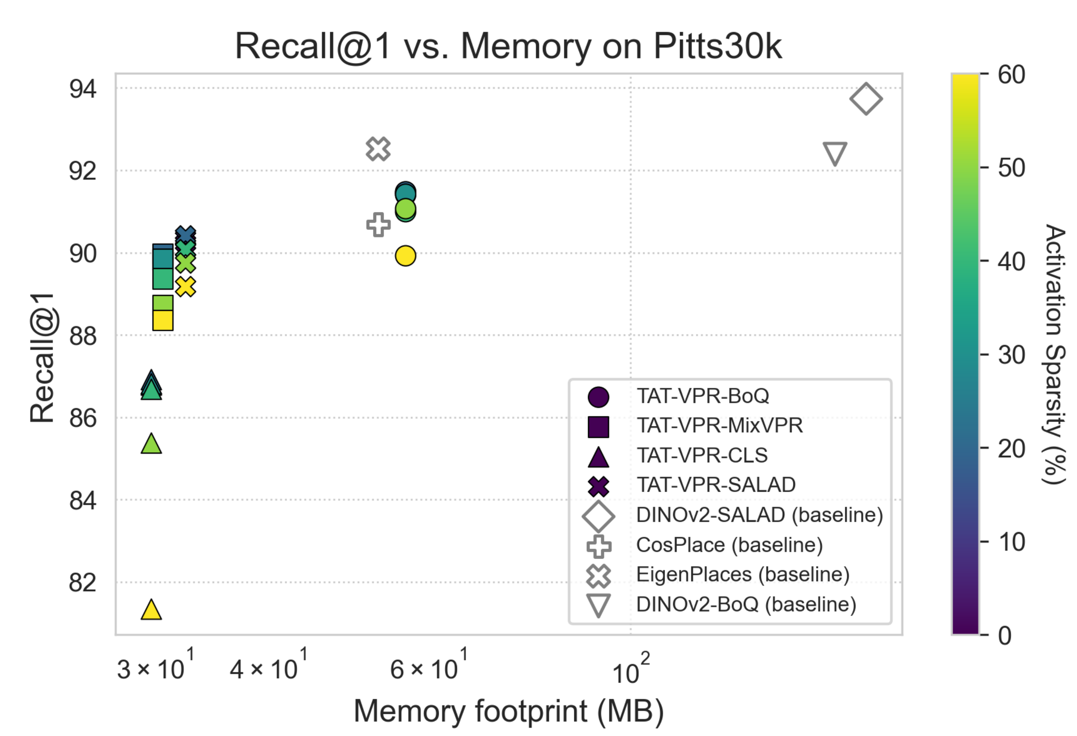
Dynamic Activation Sparsity
Runtime Control: Keep top-k% activations Computation Savings: Up to 40% TOPs reduction Implementation: $M = \mathrm{TopK}(|X|, k)$, $Y = (X \odot M)\tilde{W}^T$

Teacher-Student Distillation
Teacher: Full-precision DINOv2-BoQ (frozen) Student: Ternary transformer
Loss: Token-level supervision

Experimental Results



[FIGURE 2A: Accuracy vs. Computational Cost] Show Image *TAT-VPR enables dynamic accuracy-efficiency trade-offs. Curves show different activation sparsity levels (0-60%). Up to 40% TOPs reduction achievable with <1% Recall@1 loss.*



[FIGURE 2B: Accuracy vs. Memory Footprint] Show Image *TAT-VPR models with ternary weights achieve 5× memory reduction compared to full-precision baselines while maintaining competitive accuracy on Pitts30k dataset.*

Impact & Applications

Micro-UAV SLAM

- Real-time loop closure detection
- Extended flight time through power savings

Mobile Robotics

- Resource-aware navigation
- Adaptive computation based on battery/processing load

Edge Computing

- Dynamic scaling based on available resources
- Practical deployment on resource-limited platforms

Conclusion

TAT-VPR bridges the gap between state-of-the-art VPR accuracy and practical deployment constraints.

✅ **Dynamic scalability:** Single model adapts computation at runtime
✅ **Extreme efficiency:** 5× memory reduction, 40% computation savings
✅ **Preserved quality:** <1% accuracy drop vs. dense models
✅ **Real-world ready:** Enables VPR on micro-UAVs and embedded SLAM

Future work: Hardware acceleration for ternary operations, extended evaluation on physical robotic platforms.

Acknowledgments

Funding: This work was supported by UK EPSRC grants EP/V009600/1 and EP/V007343/1.

Institutions: We thank our collaborating institutions - University of Southampton (UK), Queensland University of Technology (Australia), and University of Essex (UK) - for their support and resources.

Correspondence: Oliver Grainge (oeg1n18@soton.ac.uk)

References:

[1] Berton et al., "Deep Visual Geo-localization Benchmark," CVPR 2022 (GSV-Cities)
[2] Zaffar et al., "CosPlace: Cross-modal Place Recognition," ICCV 2021
[3] Berton et al., "EigenPlaces: Training Viewpoint Robust Models," ICCV 2023
[4] Keetha et al., "BoQ: A Place is Worth a Bag of Learnable Queries," CVPR 2023
[5] Ali-bey et al., "MixVPR: Feature Mixing for Visual Place Recognition," WACV 2023
[6] Torii et al., "24/7 Place Recognition by View Synthesis," TPAMI 2018 (Pitts30k)
[7] Oquab et al., "DINOv2: Learning Robust Visual Features," TICL 2024
[8] Dosovitskiy et al., "An Image is Worth 16x16 Words," ICLR 2021 (ViT)

RSS Unifying Visual SLAM Workshop 2025: https://visual-slam-lab.github.io/unifying-visual-slam