EPM: Episodic Predictive Memory as a Latent World Model for Season-Invariant Visual Place Recognition

Jiaxin Li[1] Chen Sun[2*]

Abstract

A robot following a familiar outdoor route faces two opposite difficulties: the same place can look very different across snow, rain and so on, while distant road segments may look nearly identical. Single-image VPR descriptors can struggle because appearance changes or repeated route geometry create different kinds of ambiguity. We propose EPM (Episodic Predictive Memory), a lightweight route memory model for few-shot cross-season VPR. It combines local VLAD features, recurrent route context, and FAISS episodic retrieval. On Boreas cross-season pairs, EPM reaches 89.9% SR@25m at 538 FPS with 12.7M parameters, outperforming CosPlace by 7.7 pp and reducing per-pair variance by 6x. Frozen-representation probing shows that EPM states are more predictable under egomotion than strong visual baselines. The results suggest that EPM learns a structured latent route state, offering a lightweight bridge between episodic route memory and latent world-model ideas.

1. Introduction

A robot that re-traverses a long outdoor loop across four seasons must solve two failure modes at once: *the same place looks different* across snow, leaf fall, and illumination shifts [1–4], and *different places along the same route can look similar* (repeated lane geometry, similar buildings). A single-image descriptor can struggle with this trade-off because it lacks route-level temporal context.

Recent VPR methods either train descriptors on large geo-tagged datasets [5–7] or reuse frozen foundation features [8–11]. On the 12 cross-season pairs of Boreas [12], off-the-shelf CLIP and DINOv2 reach only 25.8% and 42.8% SR@25m, and the strongest baseline, CosPlace, reaches 82.2%.

To address this, we propose **EPM** (Episodic Predictive Memory), a 12.7M-parameter system that specializes to the deployment route rather than chasing universal features. Our contributions are:

(1) EPM itself, pretrained on Nordland [13] and few-shot fine-tuned on the target Boreas [12] route. EPM reaches 89.9% SR@25m at 538 FPS, +7.7 pp over CosPlace.

(2) *Latent-space prediction* a for season-invariant route representations. EPM predicts future states in latent space rather than raw images or explicit positions, encouraging the learned state vector to capture route-level properties that persist across seasonal changes. This opens a path toward bringing world-model-style representation learning [14–17] into visual place recognition.

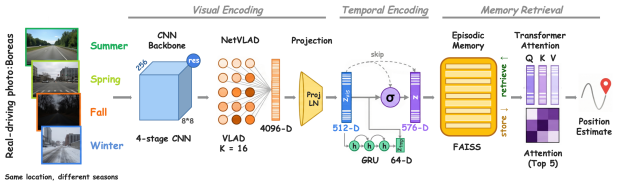

Figure 1. EPM architecture. CNN+VLAD (K=16) encodes 512D visual states; GRU temporal encoder produces 64D features via gated fusion yielding 576D state vectors; FAISS episodic memory stores states for Transformer attention-based retrieval. GPS used only in training.

2. EPM: Episodic Predictive Memory

EPM (Fig. 1) treats cross-season VPR as learning a compact route state for episodic retrieval. Each frame is compressed into a state $\mathbf{z}_{\text{state}} \in \mathbb{R}^{576}$ that fuses a visual latent with a temporal context vector. To keep this state season-stable and disambiguate aliased frames, we adopt two design choices. (i) Cross-season change mainly affects global color and texture, so we aggregate features locally with VLAD. (ii) Single frames are often perceptually aliased, so we add a recurrent encoder over consecutive frames.

Visual encoder. A 4-stage stride-2 CNN with residual connections extracts 256-channel features from a 128×128 RGB image. A differentiable VLAD layer with K=16 centers aggregates the spatial locations into per-cluster

*Corresponding author. [1]Beijing University of Posts and Telecommunications. [2]The University of Hong Kong.

residuals, projected and LayerNorm-normalized to $\mathbf{z}_{\text{vis}} \in \mathbb{R}^{512}$.The local aggregation encourages the representation to rely on route-stable structures such as road boundaries, building outlines, and skyline geometry.

Temporal encoder and latent state. A single-layer GRU (128D hidden) over visual features produces $\mathbf{z}_{\text{tmp}} \in \mathbb{R}^{64}$, fused with $\mathbf{z}_{\text{vis}}$ through a sigmoid gate g into $\mathbf{z}_{\text{state}} = [\mathbf{z}_{\text{vis}} \,\|\, g \odot \mathbf{z}_{\text{tmp}}] \in \mathbb{R}^{576}$. This state is purely image-derived.

Training. We train with an episodic protocol that mixes weather conditions on the same route segment. A small GRU dynamics head f_ϕ rolls the state forward k=3 steps under egomotion actions $\mathbf{a}_{t:t+k}$ (GPS displacements, training-only). A cosine latent-prediction loss $\mathcal{L}_{\text{pred}} = \frac{1}{k}\sum_i 1 - \cos(\hat{\mathbf{z}}_{t+i}, \mathbf{z}_{t+i})$ makes the encoder self-predictive without an image decoder [14, 15]. InfoNCE [18] aligns same-place states across weather, and VICReg [19] prevents collapse.

Episodic memory and retrieval. During deployment, the 576D states of the reference traversal are stored in a FAISS episodic memory. For a query frame, EPM retrieves the top-5 nearest reference states and re-weights them with a 4-head Transformer attention module to estimate position. In *Sequential* mode, retrieval is restricted to a forward sliding window consistent with route progress.

3. Experiments

We evaluate on Boreas [12], an 8 km urban loop in Toronto. The dataset provides 18 traversals across four seasons ($\sim$ 180K frames). We use a geographic train/test split and form 12 directed cross-season evaluation pairs on held-out route segments. EPM is first pretrained on 27K Nordland [13] images, then few-shot fine-tuned on Boreas. Each fine-tuning episode samples three support and one query weather conditions on the same route segment. We report SR@25m, with PCA-192 whitening on all descriptors before retrieval.

Main results. Table 1 reports few-shot cross-season retrieval. EPM is the only model that stays above 80% on every season pair, and its worst case (spring→winter) is comparable to CosPlace's best. DINOv2 and CLIP fall below 50% despite far larger pretraining, suggesting that generic visual semantics alone do not provide reliable metric place identity under severe appearance change. CosPlace closes most of the gap with place-specific supervision but remains brittle on snow-dominated pairs. Ablations match the design rationale in Sec. 2: removing VLAD hurts most, followed by InfoNCE, then the recurrent encoder. As a sanity check that retrieved positions are metrically usable, we replay some of the 12 traversal pairs through a Pure-Pursuit controller fed only EPM's per-frame retrieved positions (Figure 2).

Probing frozen representations. We freeze each model's encoder and attach an identical 1.7M-parameter GRU+MLP probe conditioned on GPS displacement [21, 22], plus four auxiliary linear probes (Table 2).

Model	Params / Data	SR@25m ↑	MAD ↓
EPM (ours)	12.7M / 27K	**89.9**	**1.9**
CosPlace [5]	11.2M / 40M	82.2	11.4
NetVLAD [20]	17.1M / Pitts30k+	43.9	—
DINOv2 [10]	22M / 142M	42.8	—
CLIP [11]	151M / 400M	25.8	—

Table 1. **Cross-season VPR on Boreas** (12 pairs, PCA-192, Sequential). MAD: mean absolute deviation across pairs.

Probe	**EPM**	CosPlace	DINOv2	CLIP	NetVLAD
FD loss ↓	**.016**	.057	.029	.072	.116
Imag@5m ↑	**87.5**	81.0	78.2	81.6	81.8
Noise σ=5 ↑	40.4	36.0	**94.4**	79.7	75.2
Season ↓	**89.5**	96.5	99.3	99.8	97.9
Place gap ↑	**+.212**	-.007	+.019	+.078	-.030

Table 2. **Probing frozen representations.** *FD*: forward-dynamics MSE. *Imag@5m*: 5-step latent imagination Hit@5 m (%). *Noise*: Gaussian-noise robustness (%). *Season*: linear season-classification accuracy (%). *Place gap*: cross-season same-vs-different-place embedding gap.

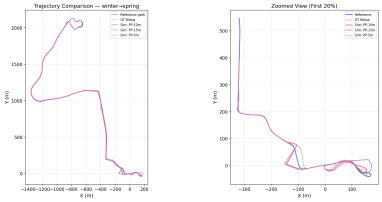

Figure 2. **Route-following replay** (winter→spring). A Pure-Pursuit controller drives a kinematic bicycle model using only EPM's per-frame retrieved positions. Left: full loop; right: first 20% zoom.

(i) Only EPM produces a positive cross-season place gap of substantial magnitude: same-place samples across seasons are, on average, closer than different-place samples in the same season. CosPlace and NetVLAD are slightly negative, suggesting that their embeddings are more affected by appearance changes. This directly explains the retrieval ranking in Table 1. (ii) The forward-dynamics and 5-step imagination probes both favor EPM. The frozen EPM state is therefore not just compact, it is also contain more egomotion-consistent transition structure. This makes it usable as the observation slot of a route-level world model rather than only as a retrieval key. A linear season probe further supports this interpretation: EPM yields the lowest season-classification accuracy among the compared embeddings. This does not mean that season information is absent, but suggests that EPM makes season-specific appearance cues less dominant in the representation.

References

[1] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2016. 1

[2] S. Garg, T. Fischer, and M. Milford, "Where is your place, visual place recognition?," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4416–4425, 2021.

[3] C. Masone and B. Caputo, "A survey on deep visual place recognition," *IEEE Access*, vol. 9, pp. 19516–19547, 2021.

[4] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1643–1649, 2012. 1

[5] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4878–4888, 2022. 1, 2

[6] A. Ali-bey, B. Chaib-draa, and P. Giguère, "MixVPR: Feature mixing for visual place recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2998–3007, 2023.

[7] 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)*, pp. 11080–11090, 2023. 1

[8] 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, 2023. 1

[9] I. Tzachor, B. Lerner, M. Levy, M. Green, T. B. Shalev, G. Habib, D. Samuel, N. K. Zailer, O. Shimshi, N. Darshan, and R. Ben-Ari, "EffoVPR: Effective foundation model utilization for visual place recognition," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[10] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024. 2

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021. 1, 2

[12] K. Burnett, D. J. Yoon, Y. Wu, A. Z. Li, H. Zhang, S. Lu, J. Qian, W.-K. Tseng, A. Lambert, K. Y. Leung, A. P. Schoellig, and T. D. Barfoot, "Boreas: A multi-season autonomous driving dataset," *The International Journal of Robotics Research*, vol. 42, no. 1-2, pp. 33–42, 2023. 1, 2

[13] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? challenging SeqSLAM on a 3000 km journey across all four seasons," in *Proc. Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA)*, 2013. 1, 2

[14] D. Ha and J. Schmidhuber, "World models," *arXiv preprint arXiv:1803.10122*, 2018. 1, 2

[15] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 2

[16] A. Bar, G. Zhou, D. Tran, T. Darrell, and Y. LeCun, "Navigation world models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15791–15801, 2025.

[17] Y. Xu, Y. Pan, and Z. Liu, "Dream to recall: Imagination-guided experience retrieval for memory-persistent vision-and-language navigation," *arXiv preprint arXiv:2510.08553*, 2025. 1

[18] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018. 2

[19] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-invariance-covariance regularization for self-supervised learning," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 2

[20] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5297–5307, 2016. 2

[21] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," in *International Conference on Learning Representations (ICLR) Workshop*, 2017. 2

[22] Y. Belinkov, "Probing classifiers: Promises, shortcomings, and advances," *Computational Linguistics*, vol. 48, no. 1, pp. 207–219, 2022. 2