EMVP: Embracing Visual Foundation Model for Visual Place Recognition with Centroid-Free Probing

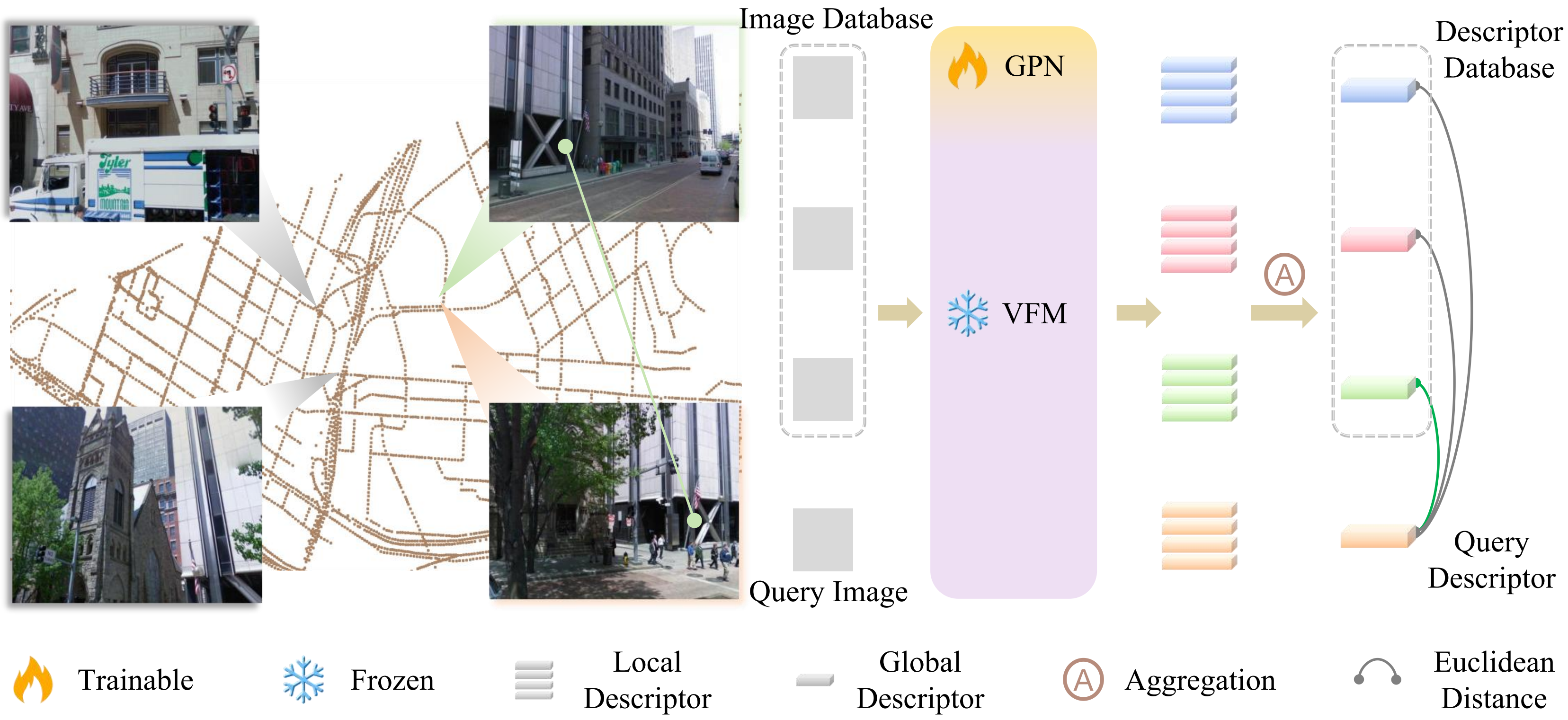
Qibo Qiu[1,2], Shun Zhang[3], Haiming Gao[3],
Honghui Yang[1], Haochao Ying[1], Wenxiao Wang[1], Xiaofei He[1]

[1] Zhejiang University, Hangzhou, China

[2] China Mobile (Zhejiang) Research & Innovation Institute, Hangzhou, China

[3] Zhejiang Lab, Hangzhou, China

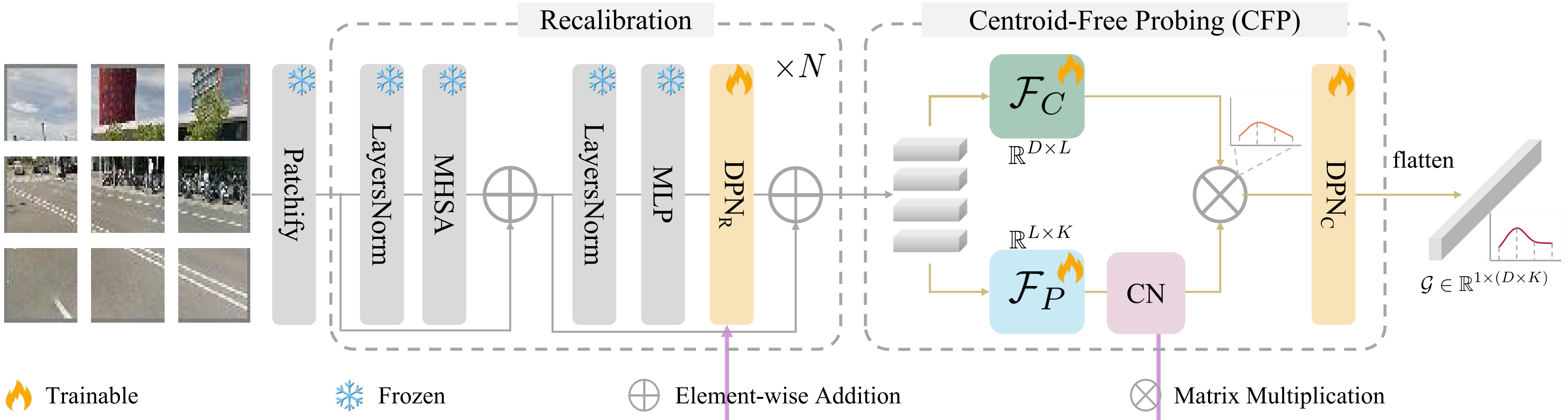
1. Background & Motivation

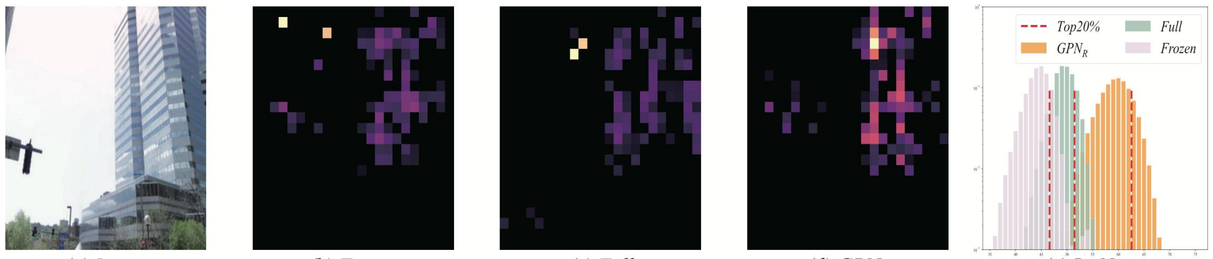


1. The Visual Foundation Model (VFM) has significantly enhanced the performance of Visual Place Recognition (VPR), avoiding training a model from scratch on **environment-specific** data.

2. This paper focuses on crucial role of **probing** in effectively adapting a VFM for improved image representation.

2. Solution





(a) Image (b) Frozen (c) Full (d) GPN$_R$ (e) L_2 Norm

$$
\begin{aligned}
\mathcal{G} &= NetVLAD(\mathcal{X}, \mathcal{C}) \\
&= \sum_{i=1}^{L} \left([X_i - C_1; X_i - C_2; ...; X_i - C_K] \odot [\underbrace{p_{i1}, p_{i1}, ...p_{i1}}_{D}; ...; p_{iK}, p_{iK}, ...p_{iK}] \right) \\
&= \sum_{i=1}^{L} X_i^T \times P_i - \underbrace{[C_1; C_2; ...; C_K] \odot \left(\sum_{i=1}^{L} P_i \right).expand(K, D)}_{constant},
\end{aligned}
$$

$$
\mathcal{G} = \sum_{i=1}^{L} X_i^T \times P_i = \mathcal{X}^T \mathcal{P} \approx \mathcal{F}_C(\mathcal{X})^T \mathcal{F}_P(\mathcal{X})
$$

Controlling the **preservation** α of task-specific information for each image, enabling more flexible fine-tuning.

Remove the explicit calculation of semantic centroids $\mathcal{C}$ for the enhanced generalization.

3. Results

Table 1: Comparison with state-of-the-art methods. $^\flat$ denotes models trained on the GSV-Cities dataset. Due to the high quality of annotations in GSV-Cities, results from models marked with $^\flat$ generally outperform those from their corresponding papers. In contrast, results from models without $^\flat$ are reported in their respective papers.

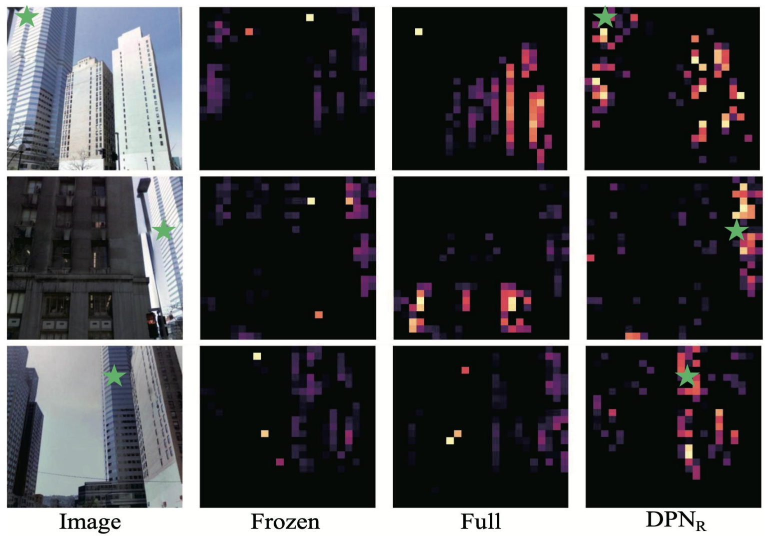
(a) Comparison with single-stage methods.

Method	MSLS Val			NordLand⋆ [52]			Pitts250k-test			SPED		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
SPE-VLAD$^\flat$ [54]	78.2	86.8	88.8	25.5	40.1	46.1	89.9	96.1	97.3	73.1	85.5	88.7
Gated NetVLAD$^\flat$ [55]	82.0	88.9	91.4	34.4	50.4	57.7	89.7	95.9	97.1	75.6	87.1	90.8
NetVLAD$^\flat$ [4]	82.6	89.6	92.0	32.6	47.1	53.3	90.5	96.2	97.4	78.7	88.3	91.4
Conv-AP$^\flat$ [32]	83.4	90.5	92.3	38.2	54.8	61.2	92.4	97.4	98.4	80.1	90.3	93.6
CosPlace$^\flat$ [17]	83.0	89.9	91.8	34.4	49.9	56.5	91.5	96.9	97.9	75.3	85.9	88.6
MixVPR$^\flat$ [16]	88.0	92.7	94.6	58.4	74.6	80.0	94.6	98.3	99.0	85.2	92.1	94.6
EigenPlaces [19]	89.3	93.7	95.0	54.4	68.8	74.1	94.1	98.0	98.7	69.9	82.9	87.6
SALAD$^\flat$ [12]	92.2	96.4	97.0	76.0	89.2	92.0	95.1	98.5	99.1	92.1	96.2	96.5
EMVP-L$^\flat$ (Ours)	**93.9**	**97.3**	**97.6**	**78.4**	**89.7**	**92.4**	**96.5**	**99.1**	**99.5**	**94.6**	**97.5**	**98.4**

(b) Comparison with two-stage methods that include a re-ranking stage, marked with $^\sharp$.

Method	MSLS Val			NordLand⋆⋆ [50]			Pitts30k-test		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
SP-SuperGlue$^\sharp$ [56]	78.1	81.9	84.3	25.8	35.4	38.2	87.2	94.8	96.4
Patch NetVLAD$^\sharp$ [5]	79.5	86.2	87.7	51.6	60.1	62.8	88.7	94.5	95.9
DELG$^\sharp$ [57]	83.2	90.0	91.1	51.3	66.8	69.8	89.9	95.4	96.7
TransVPR$^\sharp$ [10]	86.8	91.2	92.4	58.8	75.0	78.7	89.0	94.9	96.2
R2Former$^\sharp$ [11]	89.7	95.0	96.2	60.6	66.8	68.7	91.1	95.2	96.3
SelaVPR$^\sharp$ [13]	90.8	96.4	97.2	85.2	95.5	98.5	92.8	96.8	97.7
TransVPR w/o re-ranking [10]	70.8	85.1	89.6	15.9	38.6	49.4	73.8	88.1	91.9
SelaVPR (gobal) [13]	87.7	95.8	96.6	72.3	89.4	94.4	90.2	96.1	97.1
EMVP-L$^\flat$ (Ours)	**93.9**	**97.3**	**97.6**	**88.7**	**97.3**	**99.3**	**94.0**	**97.5**	**98.2**

Achieving State-of-the-Art performance with minimal trainable parameters.



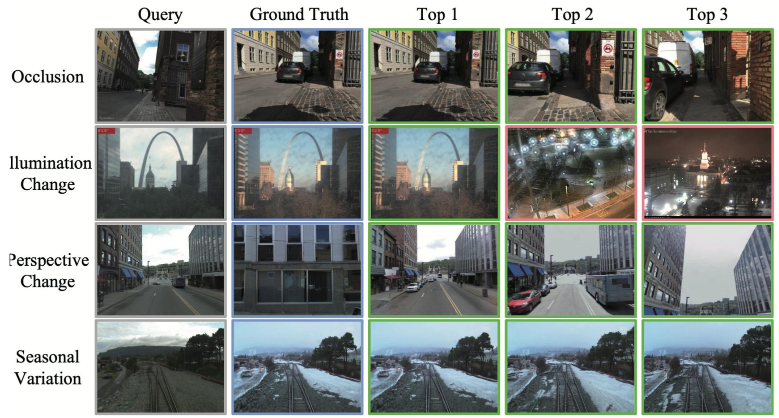


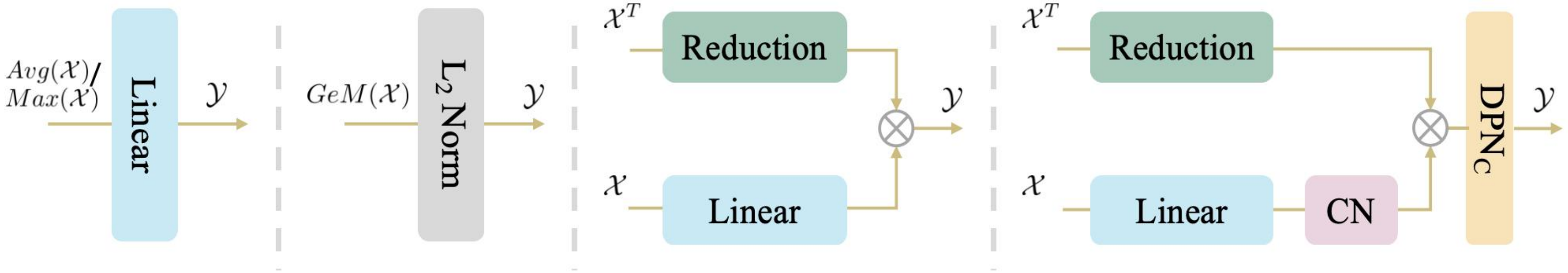
Figure 4: Query (gray) and top 3 retrieved frames (green: successful, red: failed). Moreover, one of the true (blue) matches is displayed for comparison.

EMVP-B successfully finds the closest match in challenging scenarios.

4. Comparison with Different Probing Methods

Table 2: Comparing different backbones and probings. LP, MP, CFP, CN, and DPN_C indicate linear probing, moment probing, centroid-free probing, constant normalization, and dynamic power normalization in probing, respectively. For fairness, results produced by ViT-based models are obtained by fully fine-tuning the last 4 blocks. *Baseline* refers to the simplified NetVLAD adapted by SALAD. The best and the second best results are **bolded** and underlined, respectively.

ID	Method	Fea. Dim	Backbone	MSLS Val			NordLand			Pitts250k-test			SPED		
				R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
1	NetVLAD [4]	32768	ResNet50	82.6	89.6	92.0	32.6	47.1	53.3	90.5	96.2	97.4	78.7	88.3	91.4
2	NetVLAD	8192	ViT-B	90.1	95.4	96.8	70.1	86.5	90.2	95.4	98.4	99.1	90.6	95.4	96.7
3	NetVLAD	24576	ViT-B	92.4	95.9	**96.9**	71.8	86.5	90.1	**95.6**	**98.7**	**99.3**	90.8	95.7	96.7
4	LP	768+256	ViT-B	85.3	93.5	95.4	38.1	55.3	61.8	91.3	96.9	98.1	83.0	92.3	94.0
5	MP	2048	ViT-B	87.3	94.5	96.4	42.6	62.6	70.0	92.5	97.3	98.5	85.2	92.6	94.6
6	GeM	4096	ViT-B	85.4	93.9	95.0	35.4	52.5	59.6	89.5	96.5	98.0	83.0	92.1	93.9
7	Baseline	8192+256	ViT-B	90.3	95.7	96.1	56.5	73.0	78.6	94.4	98.4	99.1	88.0	94.7	95.6
8	+ CN=Softmax	-	-	91.3	95.7	96.4	68.0	82.0	86.2	94.9	98.3	99.0	89.3	94.9	96.4
9	+ CN= ℓ_2 norm.	-	-	90.8	95.9	96.6	66.4	80.9	84.5	94.5	98.1	99.0	89.0	93.9	95.7
10	+ DPN_C (*i.e.*, CFP)	-	-	**92.6**	**96.2**	96.8	**74.6**	**87.6**	**91.3**	95.2	**98.7**	**99.3**	**92.1**	**95.9**	**97.2**



(a) LP (b) GeM Pooling (c) Baseline (d) CFP (Ours)

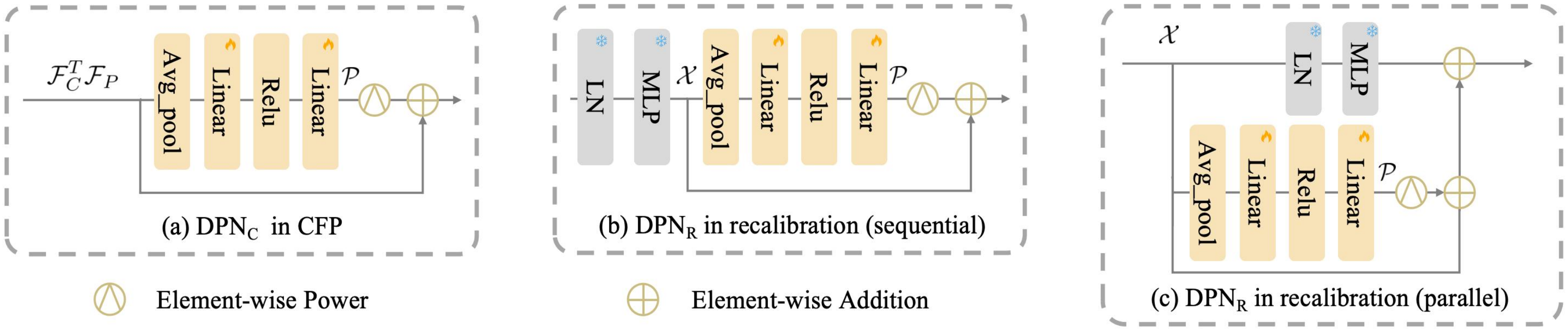
Takeaways

- **First-order** methods are inferior to CFP, due to information loss.
- The **second-order** MP is inferior to CFP, lacking of leveraging the priors provided by semantic centroids.
- Directly **removing** centroids using bilinear pooling leads to a performance drop.
- **Increasing** the feature dimension of NetVLAD can significantly enhance the performance. However, it is costly when dealing with the storage of sizable global descriptors.
- **CN** makes this reinterpretation operation empirically more robust. And the improvement brought by CN is dependent on its specific implementation.

5. Comparison with Different Adapters

Table 3: Comparing different fine-tuning methods. $\mathrm{DPN_C}$ and $\mathrm{DPN_R}$ indicate DPN in CFP and recalibration, respectively. Results of both parallel and sequential versions of $\mathrm{DPN_R}$ are reported. For fairness, only the last 4 blocks can be fine-tuned, and all methods employ the same backbone, *i.e.*, ViT-B. The best and the second best results are **bolded** and underlined, respectively.

Method	Fine-tuning Type	Params. (M)	MSLS Val			NordLand			Pitts250k-test			SPED		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
AnyLoc	Zero-shot	-	68.7	78.2	81.8	16.1	25.4	30.4	87.2	94.4	96.5	85.3	94.4	95.4
SALAD	Full	27.1	92.2	96.4	97.0	76.0	**89.2**	92.0	95.1	98.5	99.1	**92.1**	96.2	96.5
CFP	+PSRP	0.14	92.7	96.6	96.9	73.0	86.5	89.5	95.3	98.6	99.2	91.3	95.9	96.9
	+ $\mathrm{DPN_R}$(para.)	**0.05**	92.4	96.5	96.8	71.8	85.2	88.9	95.4	98.5	99.1	91.3	96.2	96.7
(*i.e.*, EMVP-B)	+ $\mathrm{DPN_R}$(seq.)	**0.05**	**93.2**	**96.9**	**97.2**	**76.4**	88.8	**92.1**	**95.7**	**98.9**	**99.3**	91.8	**96.5**	**97.4**



(a) $\mathrm{DPN_C}$ in CFP

(b) $\mathrm{DPN_R}$ in recalibration (sequential)

(c) $\mathrm{DPN_R}$ in recalibration (parallel)

Element-wise Power

Element-wise Addition

Takeaways

- Current VFMs (i.e., DINOv2) **lack sufficient zero-shot** capabilities for diverse data in the VPR domain. SALAD achieves high performance by fully fine-tuning on DINOv2.
- VPR models are typically deployed on mobile robots, and full-parameter update approach **imposes the higher demands on communication**.
- The **sequential** $\mathbf{DPN_R}$ performs better. This is primarily because the sequential method recalibrates the backbone features more thoroughly.
- Compared with SALAD and PSRP, $\mathrm{DPN_R}$ outperforms them by achieving the best performance while saving **64.3%** of trainable parameters (0.14M vs 0.05M).

Welcome to visit our poster

Thank you !

Follow the OR code for more details