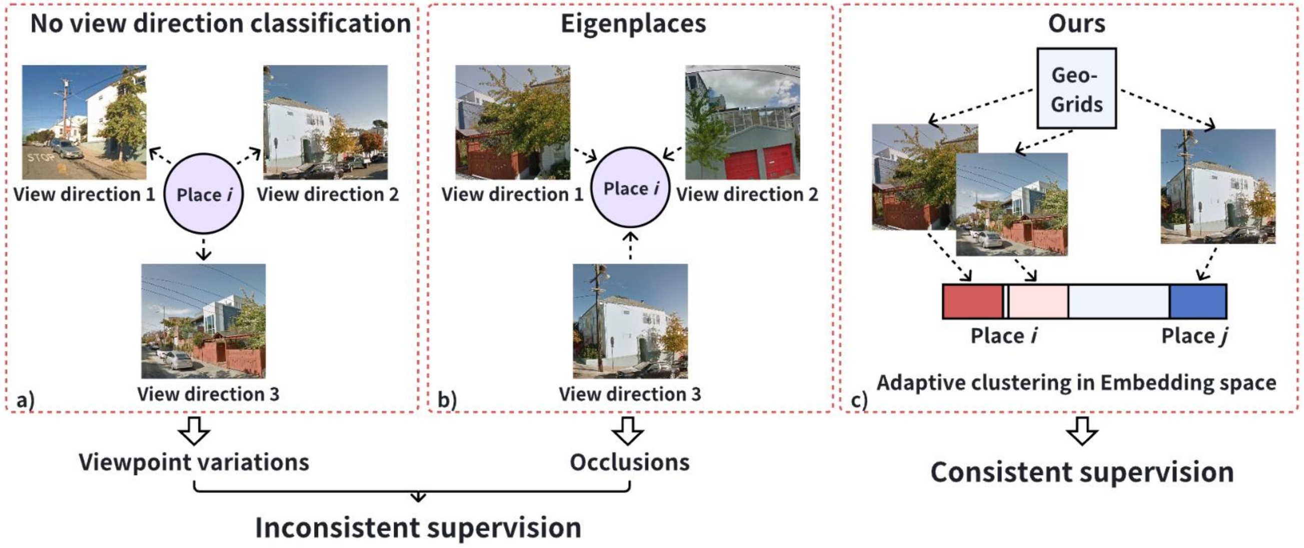
MutualVPR: A Mutual Learning Framework for Resolving Supervision Inconsistencies via Adaptive Clustering

Qiwen Gu, Xufei Wang, Junqiao Zhao, Siyue Tao, Tiantian Feng, Ziqiao Wang, Guang Chen

Tongji University, Shanghai, China

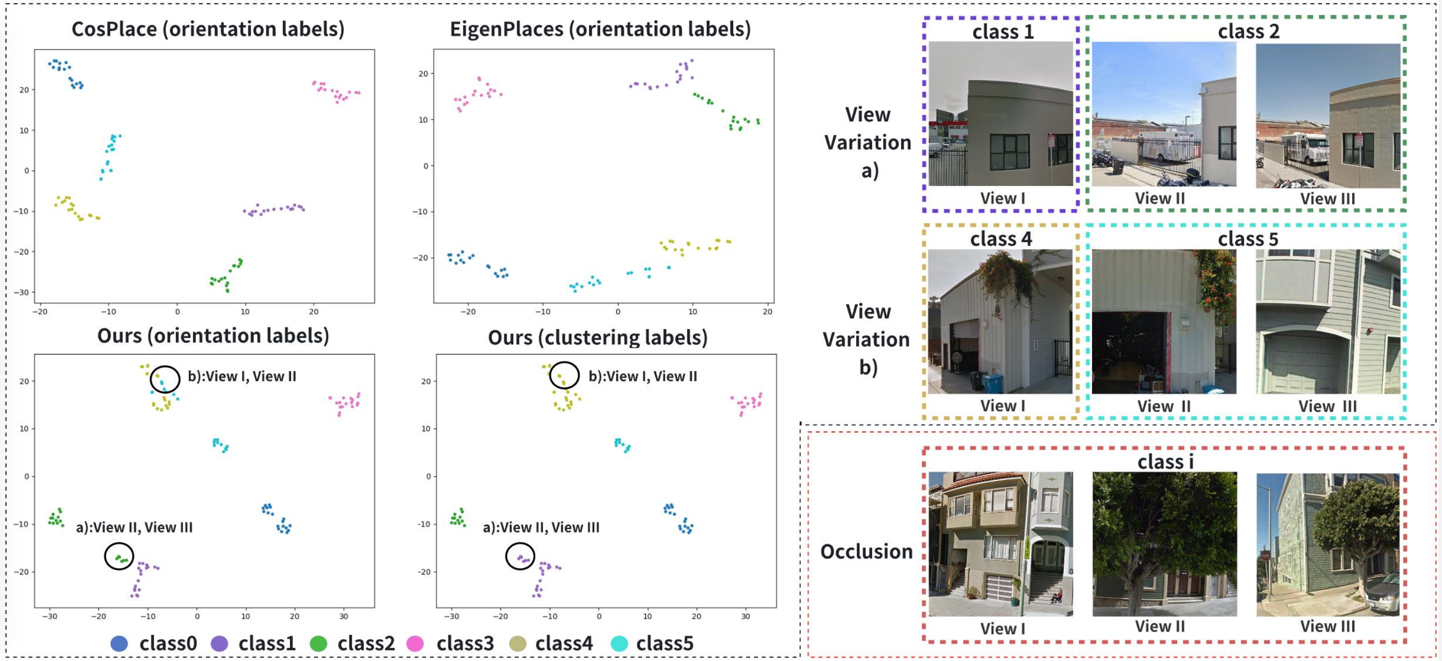
Introduction & Motivation

- What is VPR?
  - Visual Place Recognition (VPR) is a key component for localization in autonomous systems.
  - It enables a robot or vehicle to determine its location by matching a query image against a database of geo-tagged images.
- How it Fails :
  - (a) Viewpoint Variations: Simple geo-labels incorrectly group visually distinct scenes.
  - (b) Occlusions: Fixed splits fail when occlusions are present, again grouping dissimilar images.
- The Core Problem: Inconsistent Supervision
  - Classification-based VPR is scalable, but suffers from Inconsistent Supervision: rigid labels misalign with true visual similarity.



In-Depth Analysis: Why Prior Methods Fail

- t-SNE Plots Confirm the Problem:
  - Prior work (like CosPlace, EigenPlaces) uses "orientation labels" which create flawed feature spaces.



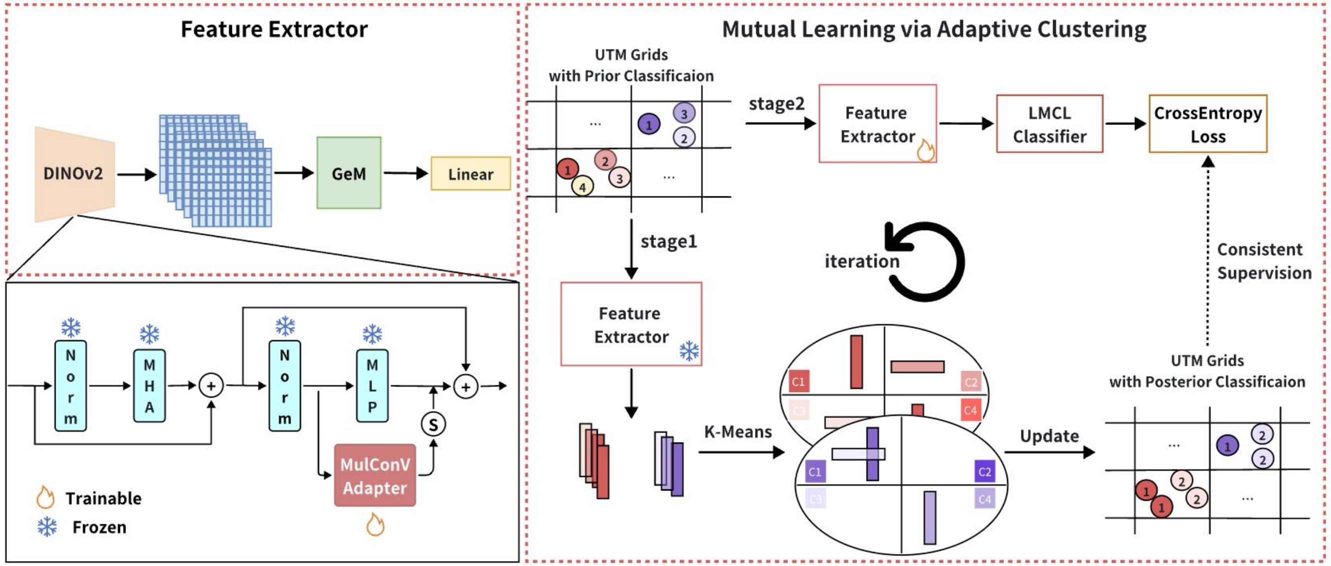
- This leads to two key failures:
  - Splitting: Visually similar views are split into different classes.
  - Merging: Visually distinct views are merged into the same class.
- Our Goal (Bottom-Right Plot):
  - Our "clustering labels" (bottom-left plot) demonstrate the correct approach: grouping images by their actual visual similarity.

Our Contribution: MutualVPR

- We propose MutualVPR: A mutual learning framework that resolves supervision inconsistencies via adaptive clustering.
- Our Core Contributions:
  - A mutual learning framework where feature learning and clustering co-evolve, effectively mitigating supervision inconsistency.
  - An adaptive clustering strategy that dynamically refines pseudo-labels based on visual semantics.
  - This handles view directions and occlusions without needing any orientation labels.
  - The model achieves SOTA performance on challenging VPR benchmarks, especially in occluded scenes.

Methodology: The MutualVPR Framework

- Our framework jointly refines the feature extractor and cluster assignments in a two-stage loop.
- Feature Extractor:
  - Uses a frozen DINOv2 backbone with a lightweight trainable MulConv adapter.
- Mutual Learning Iteration:
  - Stage 1 (Clustering): Use the current features to run K-Means clustering within each UTM grid. This generates dynamic "Posterior Classification" pseudo-labels based on visual similarity.
  - Stage 2 (Training): Use these updated, consistent pseudo-labels as supervision. Train the feature extractor using an LMCL Classifier.



This loop allows features and supervision (clusters) to co-evolve, creating a mutual learning cycle where they progressively reinforce each other.

Results: Comparison with SOTA Methods

Method	Desc.dim.	Train set	MSLS-val		Pitts30k		Pitts250k		Tokyo24/7		SF-XL-testv1	
			R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
NetVLAD	131072	GSV-Cities	82.8	90.3	87.0	94.3	89.1	94.8	69.5	82.5	-	-
GeM	2048	GSV-Cities	72.5	82.7	84.5	92.8	85.1	93.4	60.3	73.7	25.6	35.5
AnyLoc(ViT-B+GeM)	768	-	32.6	41.6	77.7	88.9	79.3	89.5	71.7	87.6	33.3	45.2
ConvAP	2048	GSV-Cities	81.5	87.5	89.7	95.2	91.2	96.4	74.6	83.2	41.1	53.0
MixVPR	4096	GSV-Cities	87.1	91.4	91.6	95.5	**94.3**	98.1	87.0	93.3	69.2	77.4
CricaVPR*	4096	GSV-Cities	90.0	95.4	94.9	97.3	–	–	93.0	97.5	-	-
CricaVPR$_1$	4096	GSV-Cities	88.5	95.1	91.6	95.7	**94.3**	**98.6**	89.5	94.6	72.8	80.1
CricaVPR$_1$ (PCA)	512	GSV-Cities	87.1	92.6	90.4	94.9	92.5	97.1	87.4	92.9	68.4	77.1
BoQ†	512	GSV-Cities	88.4	93.9	**93.1**	96.1	93.8	97.5	91.9	95.5	79.6	85.9
SALAD†	512+32	GSV-Cities	88.5	94.2	90.6	95.1	92.1	97.0	92.3	95.1	70.2	77.7
SALAD+CM†	512+32	MSLS+GSV-Cities	**90.4**	**96.2**	90.9	95.9	93.2	97.8	**92.8**	96.2	78.4	85.4
EigenPlaces	512	SF-XL	88.1	92.9	92.3	96.1	93.5	97.5	84.8	94.0	**83.8**	**89.6**
CosPlace	512	SF-XL	84.4	90.2	89.6	94.9	90.4	96.6	76.5	89.2	64.8	73.1
MutualVPR (Ours)	512	SF-XL	89.2	95.1	90.9	**96.4**	92.6	97.9	92.4	**96.6**	80.8	86.4

Method	Desc.dim.	SF-XL-Occlusion			
		R@1	R@5	R@10	R@20
GeM	2048	11.8	15.8	17.1	22.4
AnyLoc(ViT-B+GeM)	768	6.6	14.5	19.7	26.3
ConvAP	2048	23.7	26.3	28.9	31.6
MixVPR	4096	30.3	35.5	38.2	44.7
CricaVPR$_1$	4096	40.8	51.3	54.6	59.9
BoQ†	512	38.2	50.0	53.3	59.2
SALAD†	512+32	31.6	42.1	46.1	51.3
SALAD+CM†	512+32	40.8	53.7	58.3	61.3
EigenPlaces	512	36.8	51.8	56.6	59.2
CosPlace	512	32.9	43.4	46.1	48.7
No Classification	512	17.1	25.0	26.3	31.6
MutualVPR (Ours)	512	**47.4**	**65.8**	**71.1**	**73.7**

- Key Finding 1: Balanced SOTA Performance
  - MutualVPR achieves highly competitive performance across standard benchmarks.

- Key Finding 2: Superior Occlusion Robustness
  - On the challenging SF-XL-Occlusion benchmark, MutualVPR achieves 47.4% R@1——best performance！

Limitation & Future Work

- Limitation:
  - The number of clusters (K) is a fixed hyperparameter. This is a limitation because the optimal K is not universal; it depends heavily on the backboneand the dataset's view distribution, which limits the model's adaptability.

- Future Work:
  - We will explore methods to dynamically adjust K based on dataset characteristics. This will enhance the adaptive synergy between the clustering and representation learning, allowing the model to find the optimal grouping granularity by itself.

Thank you!

Code: https://github.com/Gucci233/MutualVPR