# Foundation Models for Visual Place Recognition

**Name**: Avneesh Mishra
**Roll. No**: 2021701032
**Guide**: Prof. K Madhava Krishna (RRC)

Robotics Research Center

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

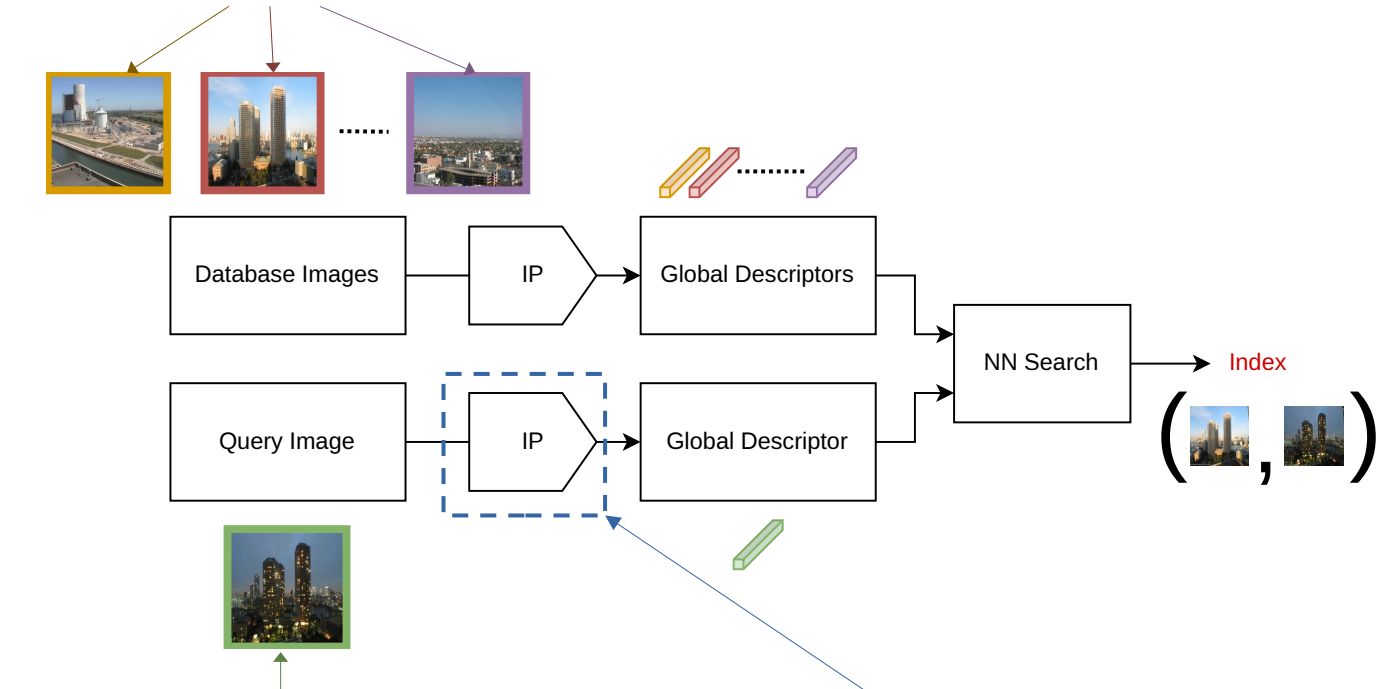
# What is it?

- A **Universal VPR** solution that can be used off-the-shelf (no retraining or fine-tuning)
- Using conventional feature aggregation techniques from VPR on latent features from **Foundation Models**.

VPR + Foundation Models = AnyLoc

# Image Retrieval (IR)

Given a geostamped collection of images (database)

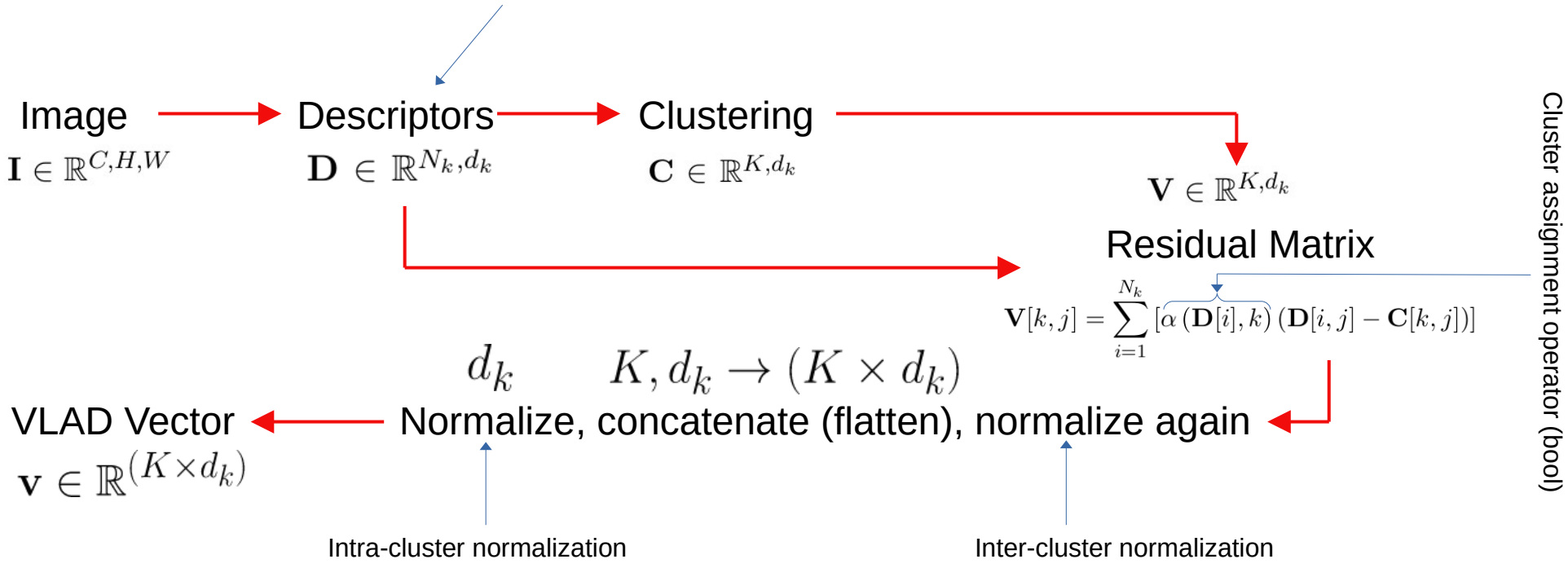



Where was this image taken?

How to build good global descriptors?

# VLAD

- Build *global features* from *local features*



# VLAD

- Build *global features* from *local features*

Image $\mathbf{I} \in \mathbb{R}^{C,H,W}$ → Descriptors $\mathbf{D} \in \mathbb{R}^{N_k,d_k}$ → Clustering $\mathbf{C} \in \mathbb{R}^{K,d_k}$ → $\mathbf{V} \in \mathbb{R}^{K,d_k}$

Residual Matrix

$$\mathbf{V}[k,j] = \sum_{i=1}^{N_k} [\alpha(\mathbf{D}[i],k)(\mathbf{D}[i,j] - \mathbf{C}[k,j])]$$

$$d_k \qquad K, d_k \rightarrow (K \times d_k)$$

Normalize, concatenate (flatten), normalize again → VLAD Vector $\mathbf{v} \in \mathbb{R}^{(K \times d_k)}$

---

# GeM Pooling

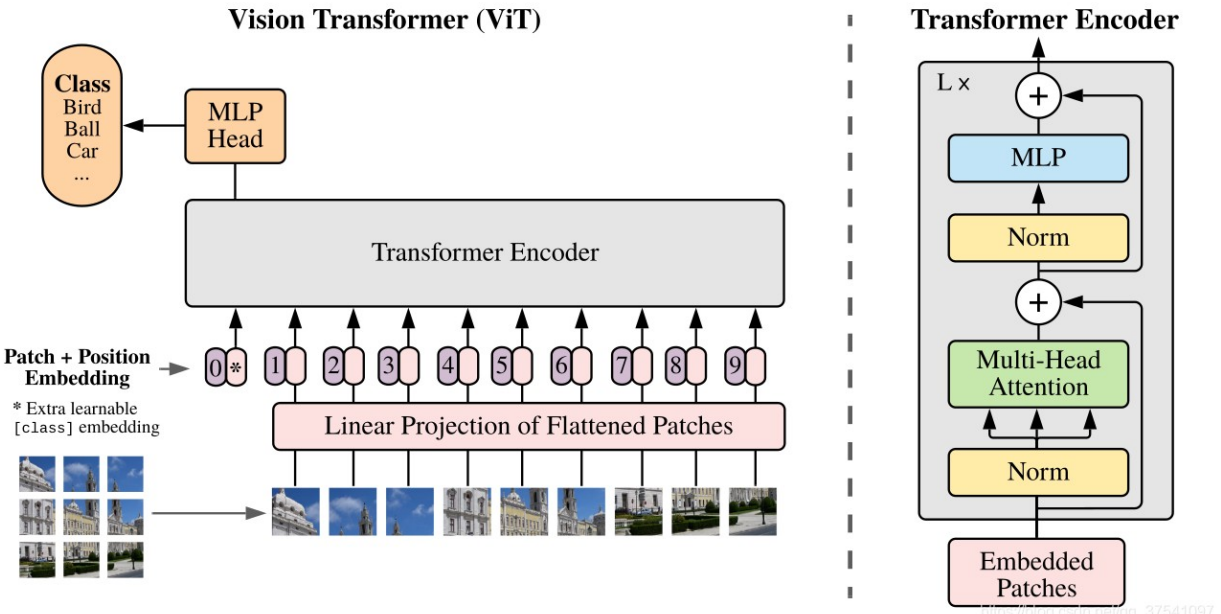
Image $\mathbf{I} \in \mathbb{R}^{C,H,W}$ → Descriptors $\mathbf{f} \in \mathbb{R}^{N_k,d_k}$ → $\mathbf{f}_G = \left( \frac{1}{n} \sum_{i=1}^{n} (\mathbf{f}[i,:])^p \right)^{\frac{1}{p}} \in \mathbb{R}^{d_k}$

Generalized Mean Pooling

P = 3 works!

# Foundation Models

## Model

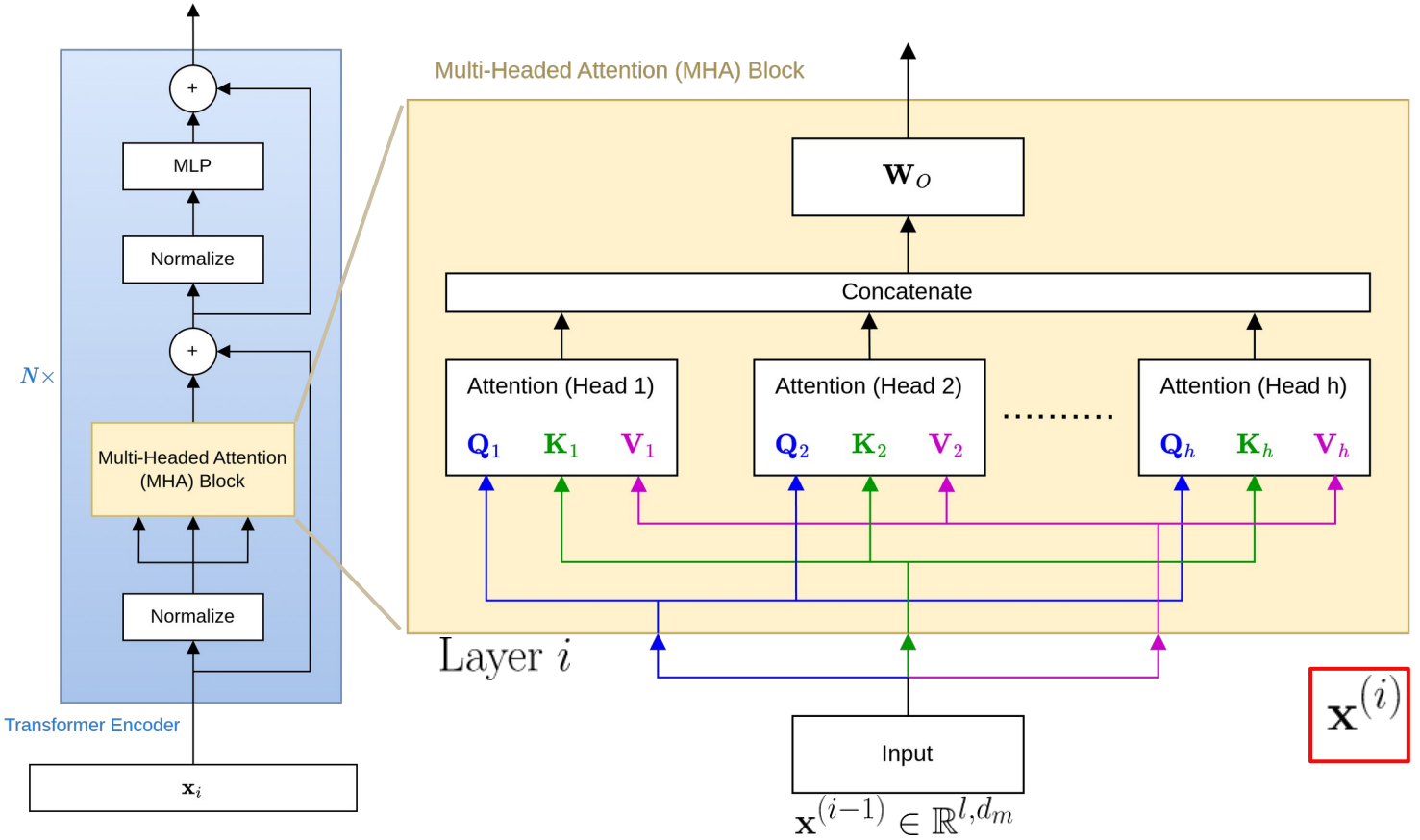



+ Big Datasets

+ SSL Losses

+ Scalable Optimizers

# ViT Features




$$\tilde{\mathbf{x}} = \mathrm{LN}(\mathbf{x})$$

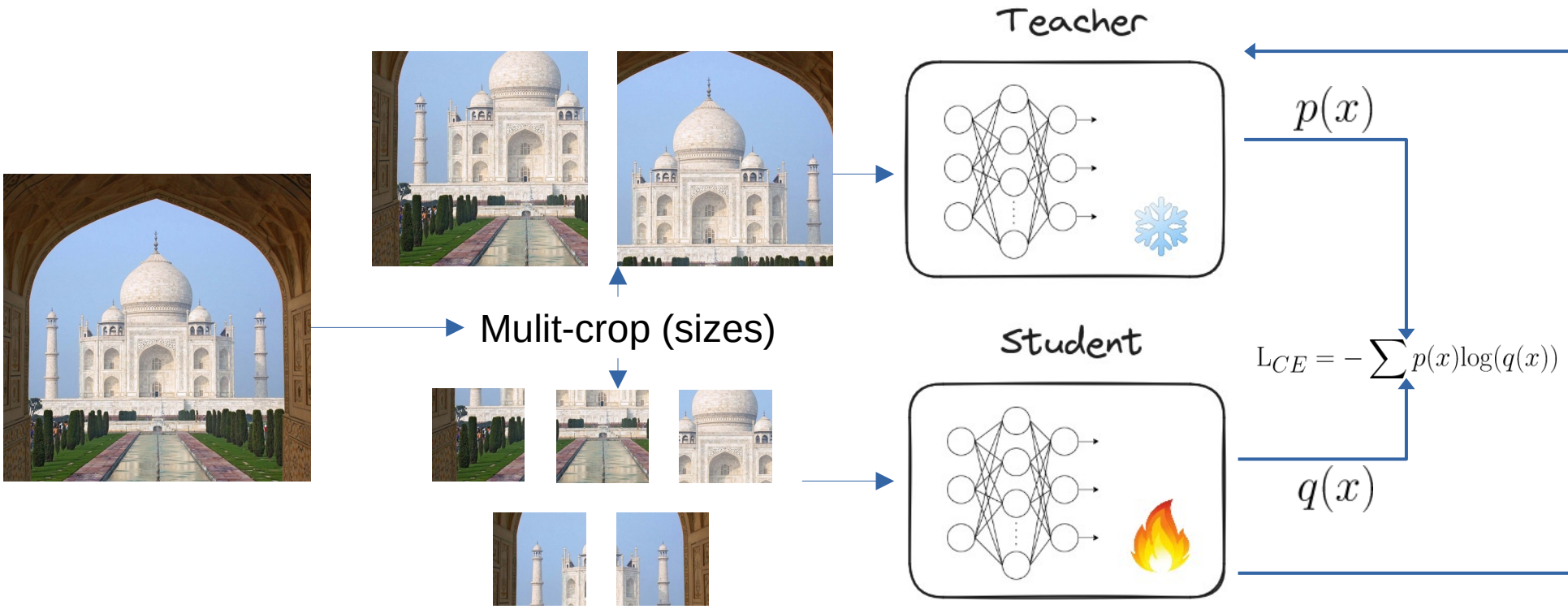
$$\mathbf{Q}_h = \tilde{\mathbf{x}}\mathbf{W}_{Q_h}$$
$$\mathbf{K}_h = \tilde{\mathbf{x}}\mathbf{W}_{K_h}$$
$$\mathbf{V}_h = \tilde{\mathbf{x}}\mathbf{W}_{V_h}$$

$$\mathbf{A}_h = \mathrm{softmax}\left(\frac{\mathbf{Q}_h\mathbf{K}_h^\top}{\sqrt{d_k}}\right)\mathbf{V}_h$$

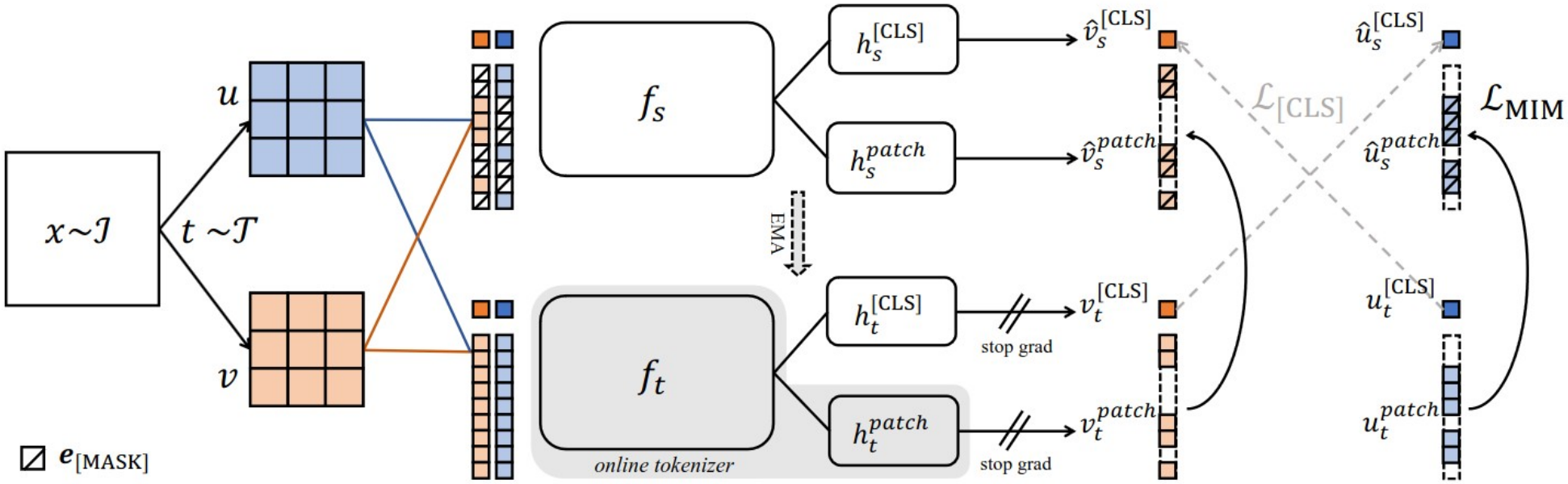
$$\mathbf{x}^{(i)} = [\mathbf{A}_1 \oplus \cdots \oplus \mathbf{A}_h]\,\mathbf{W}_{O_i}$$

Use these as features

# DINO


Teacher
Student
Mulit-crop (sizes)
$p(x)$
$q(x)$
$L_{CE} = -\sum p(x)\log(q(x))$
Exponential Moving Average

# iBOT




$$\mathrm{L}_{MIM}^{u_t \to \hat{u}_s} = -\sum_{i=1}^{N} m_i \mathrm{P}_{\theta'}^{t} \left(\mathbf{u}_t^{patch}\right)^{\top} \log\left(\mathrm{P}_{\theta}^{s}\left(\hat{\mathbf{u}}_s^{patch}\right)\right)$$

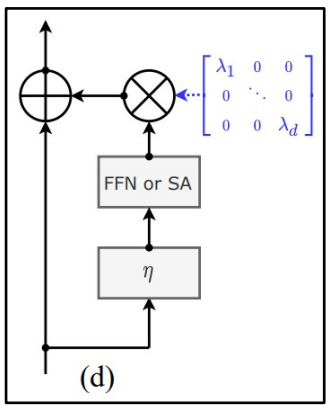
$$\mathrm{L}_{MIM}^{v_t \to \hat{v}_s} = -\sum_{i=1}^{N} m_i \mathrm{P}_{\theta'}^{t} \left(\mathbf{v}_t^{patch}\right)^{\top} \log\left(\mathrm{P}_{\theta}^{s}\left(\hat{\mathbf{v}}_s^{patch}\right)\right)$$

$$\mathrm{L}_{[CLS]}^{u_t \to \hat{v}_s} = -\mathrm{P}_{\theta'}^{t}\left(\mathbf{u}_t^{[CLS]}\right) \log\left(\mathrm{P}_{\theta}^{s}\left(\hat{\mathbf{v}}_s^{[CLS]}\right)\right)$$

$$\mathrm{L}_{[CLS]}^{v_t \to \hat{u}_s} = -\mathrm{P}_{\theta'}^{t}\left(\mathbf{v}_t^{[CLS]}\right) \log\left(\mathrm{P}_{\theta}^{s}\left(\hat{\mathbf{u}}_s^{[CLS]}\right)\right)$$

# DINOv2

**= iBOT + LayerScale + Stochastic Depth + SwiGLU Activation**




Shrink the depth of network during training by randomly dropping entire ResBlocks

$$\mathrm{SwiGLU}(x, W, b, c, \beta) = \underbrace{\mathrm{Swish}_\beta(xW + b)}_{} \otimes \underbrace{(xV + c)}_{}$$

Gated Linear Unit

$$\mathrm{Swish}_\beta(x) = x \times \frac{1}{1 + e^{-\beta x}}$$

FairScale

xFormers

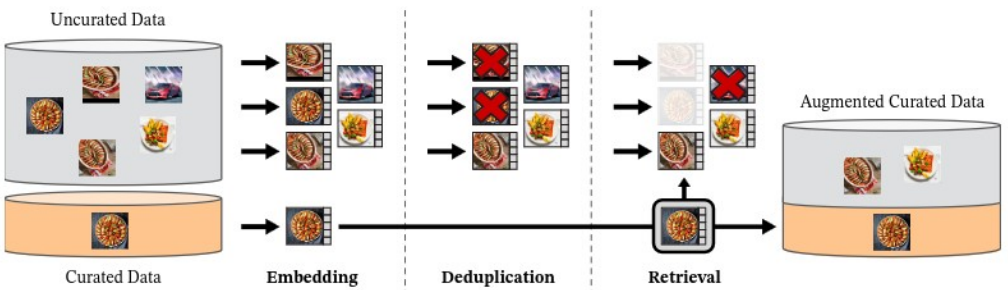
timm

+ Better regularization

+ Better Teacher centering

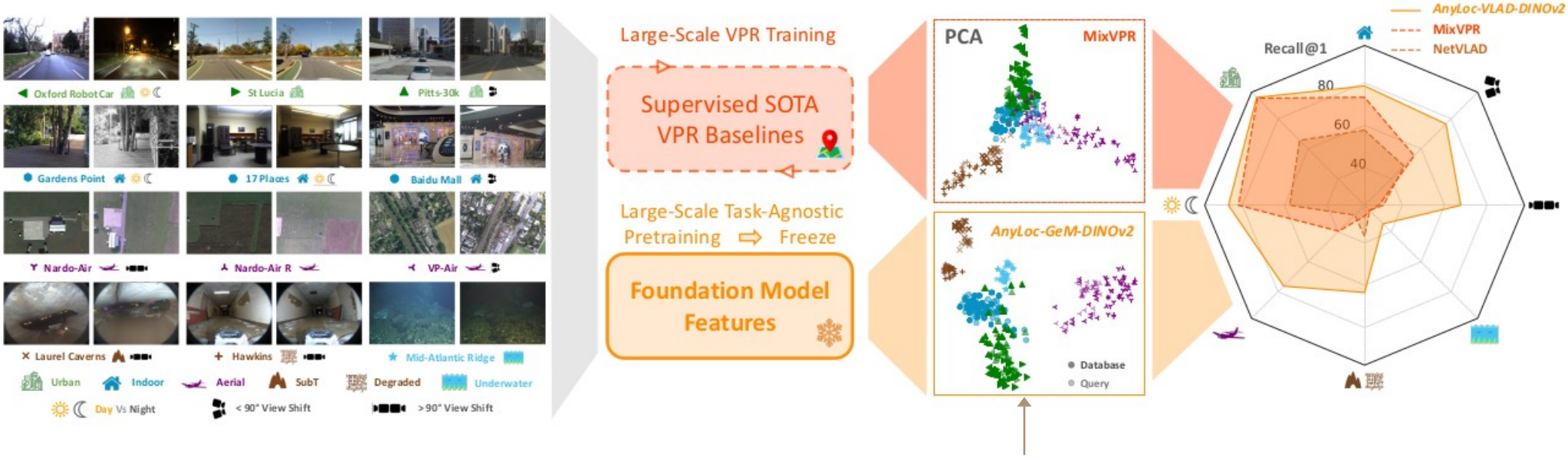
+ Distillation from larger to smaller models

\+



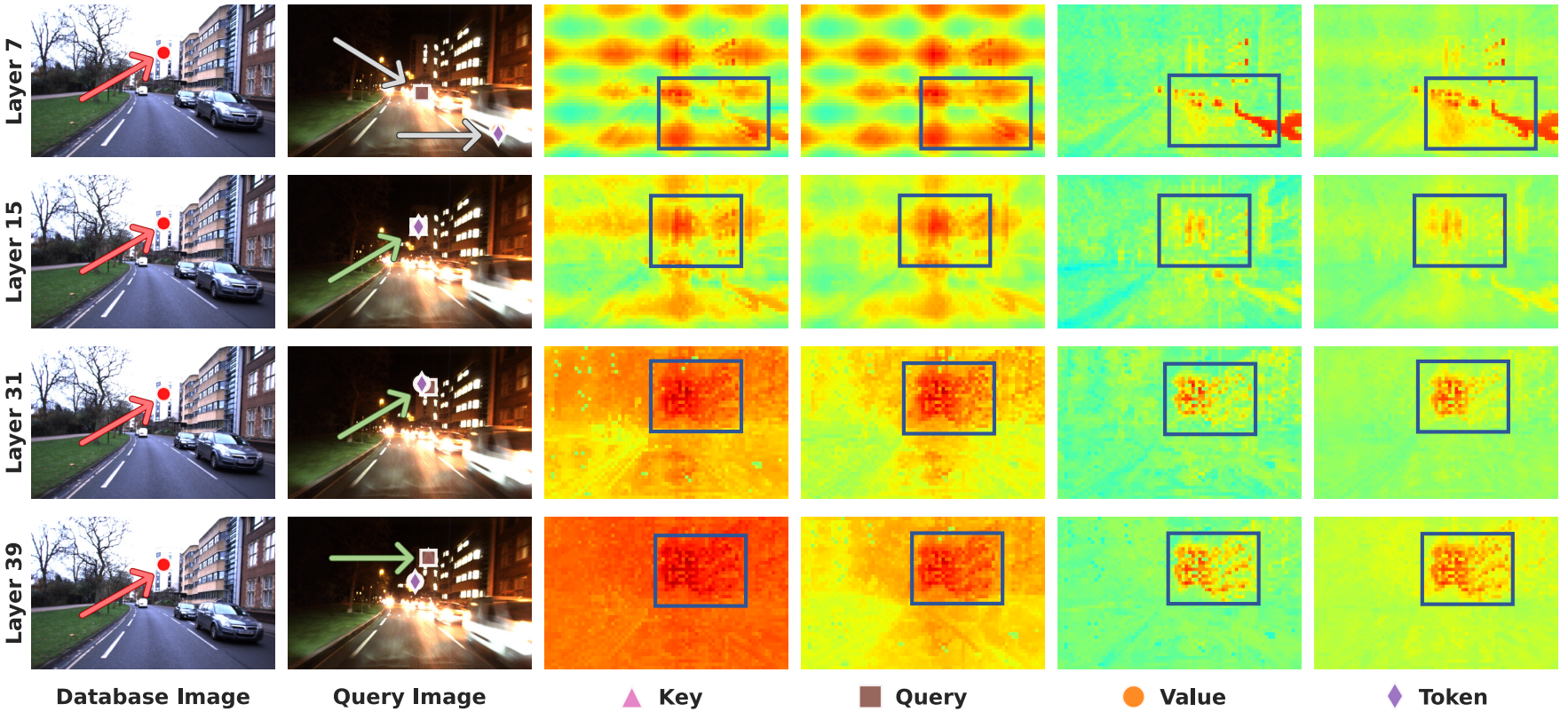
# AnyLoc

1. Take DINO or DINOv2
2. Extract latent (intermediate) features from a facet in a layer
3. Do global pooling over these “patch descriptors”




GeM pooling over DINOv2’s L31, value -> Notice clusters

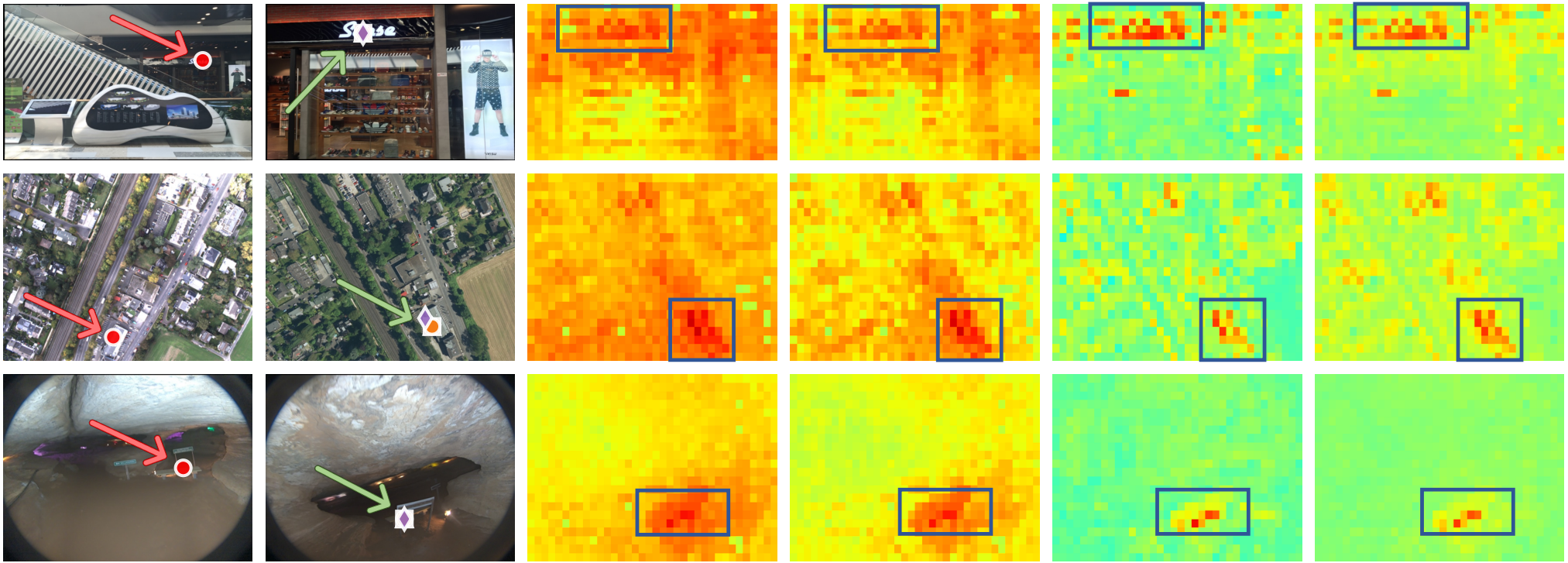
# Layer Selection




Layer 31 detects the building well

ViT-G/14 – DINOv2 – Layer selection: Notice L31, value and token facet

# Facet Selection




ViT-G/14 – DINOv2 – Facet selection (layer 31 chosen): value facet has best contrast

# Models

DINOv2: ViT-G/14, Layer 31, Value Facet → AnyLoc-GeM-DINOv2

32 clusters → AnyLoc-VLAD-DINOv2

1536 x 32 = 49152 to 512 (PCA) → AnyLoc-VLAD-DINOv2-PCA

**96x reduction in PCA!**

DINO: ViT-S/8, Layer 9, Key Facet + 128 clusters → AnyLoc-VLAD-DINO

384 x 128 = 49152 to 512 (PCA) → AnyLoc-VLAD-DINO-PCA

**ours**

---

**baselines**

NetVLAD: Soft VLAD

CosPlace

MixVPR

CLIP: OpenCLIP ViT-BigG/14

# Choice of VLAD Database

1. **Global**: All datasets
2. **Structured**: Datasets in left
3. **Unstructured**: Datasets in left
4. **Map**: Only the dataset being tested
5. **Domain**: Datasets of the same type (symbol)

Datasets used

| Structured | | | | | Unstructured | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Dataset | $N_{Db}$ | $N_q$ | Loc. | Type | Dataset | $N_{Db}$ | $N_q$ | Loc. | Type |
| Baidu [127] | 689 | 2292 | 10 m |  | Hawkins [28] | 65 | 101 | 8 m |  |
| Gardens [56, 150] | 200 | 200 | 5 fr |  | Laurel [28] | 141 | 112 | 8 m |  |
| 17 Places [140] | 406 | 406 | 5 fr |  | Nardo-Air [14] | 102 | 71 | 60 m |  |
| Pitts-30k [161] | 10k | 6816 | 25 m |  | VP-Air [45] | 2.7k | 2.7k | 3 fr |  |
| St. Lucia [177] | 1549 | 1464 | 25 m |  | Mid-Atlantic [8] | 65 | 101 | 0.3 m |  |
| Oxford [121] | 191 | 191 | 25 m |  | | | | | |

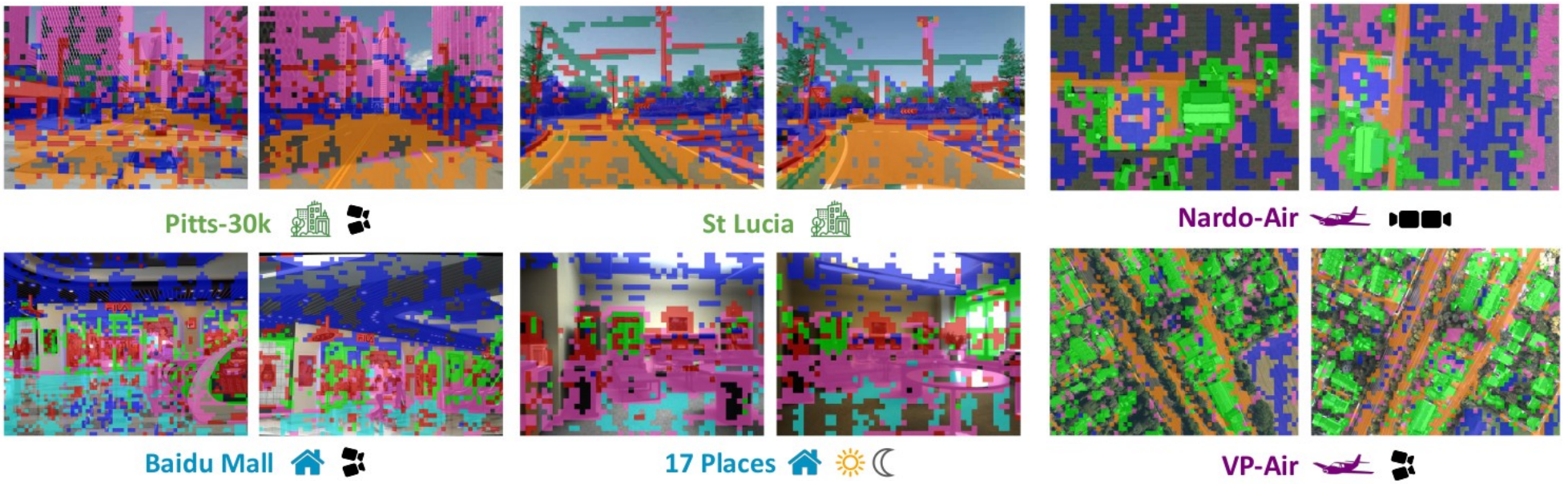
# Best Vocabulary to Use: Domain

| Vocabulary Type | Indoor | Outdoor | Aerial |
|---|---|---|---|
| Global | 77.0 | 93.9 | 57.1 |
| Structured | 77.0 | 93.3 | 56.4 |
| Unstructured | 74.8 | 89.0 | 75.8 |
| Map-Specific | 78.0 | 92.3 | 62.9 |
| Domain-Specific | **78.6** | **94.4** | **76.2** |

**AnyLoc-VLAD-DINOv2** average on different vocabularies

Use **domain-specific** for portability (unknown dataset but known *setting*)!

# VLAD Sanity Check




Visualize cluster assignments (alpha operator) for Reference & Query pairs for each dataset

There are clusters latching onto semantically meaningful objects

# Indoor

| Methods | Baidu Mall | | Gardens Point | | 17 Places | | *Average* | |
|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 |
| NetVLAD [144] | 53.1 | 70.5 | 58.5 | 85.0 | 61.6 | 77.8 | 57.73 | 77.76 |
| CosPlace [33] | 41.6 | 55.0 | 74.0 | 94.5 | 61.1 | 76.1 | 58.9 | 75.2 |
| MixVPR [5] | 64.4 | 80.3 | 91.5 | 96.0 | 63.8 | 78.8 | 73.23 | 85.03 |
| CLIP-CLS [66] | 56.0 | 71.6 | 42.5 | 74.5 | 59.4 | 77.6 | 52.63 | 74.56 |
| DINO-CLS [51] | 48.3 | 65.1 | 78.5 | 95.0 | 61.8 | 76.4 | 62.86 | 78.83 |
| DINOv2-CLS [21] | 49.2 | 64.6 | 71.5 | 96.0 | 61.8 | 78.8 | 60.83 | 79.8 |
| AnyLoc-GeM-DINOv2 | 50.1 | 70.6 | 88.0 | 97.5 | 63.6 | 79.6 | 67.23 | 82.56 |
| AnyLoc-VLAD-DINO | 61.2 | 78.3 | 95.0 | 98.5 | 63.8 | 78.8 | 73.33 | 85.2 |
| AnyLoc-VLAD-DINO-PCA | 62.3 | 81.2 | 91.5 | **99.5** | 63.3 | 78.8 | 72.36 | 86.5 |
| **AnyLoc-VLAD-DINOv2** | **75.2** | 87.6 | 95.5 | **99.5** | **65.0** | 80.5 | **78.56** | 89.2 |
| **AnyLoc-VLAD-DINOv2-PCA** | 74.9 | **89.4** | **96.0** | **99.5** | 64.8 | **81.0** | **78.56** | **89.96** |

# Outdoor

| Methods | Pitts-30k | | St. Lucia | | Oxford | | *Average* | |
|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 |
| NetVLAD [144] | 86.1 | 92.7 | 57.9 | 73.0 | 57.6 | 79.1 | 67.2 | 81.6 |
| CosPlace [33] | 90.4 | **95.7** | 99.6 | 99.9 | 95.3 | 99.5 | **95.1** | 98.36 |
| MixVPR [5] | **91.5** | 95.5 | **99.7** | **100** | 92.7 | 99.5 | 94.63 | 98.33 |
| CLIP-CLS [66] | 55.0 | 77.2 | 62.7 | 80.7 | 46.6 | 60.7 | 54.76 | 72.86 |
| DINO-CLS [51] | 70.1 | 86.4 | 45.2 | 64.0 | 20.4 | 46.6 | 42.23 | 65.66 |
| DINOv2-CLS [21] | 78.3 | 91.1 | 78.6 | 89.7 | 47.1 | 58.1 | 68 | 79.63 |
| AnyLoc-GeM-DINOv2 | 77.0 | 87.3 | 76.9 | 89.3 | 92.2 | 97.9 | 82.03 | 91.5 |
| AnyLoc-VLAD-DINO | 83.4 | 92.0 | 88.5 | 94.9 | 82.2 | 99.0 | 84.7 | 95.3 |
| AnyLoc-VLAD-DINO-PCA | 82.8 | 90.8 | 87.6 | 94.3 | 82.7 | 96.3 | 84.36 | 93.8 |
| **AnyLoc-VLAD-DINOv2** | 87.7 | 94.7 | 96.2 | 98.8 | **99.5** | **100** | 94.46 | 97.83 |
| **AnyLoc-VLAD-DINOv2-PCA** | 86.9 | 93.8 | 96.4 | 99.5 | 96.9 | **100** | 93.4 | 97.76 |

# Aerial

| Methods | Nardo-Air | | Nardo-Air R | | VP-Air | | *Average* | |
|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 |
| NetVLAD [144] | 19.7 | 39.4 | 60.6 | 85.9 | 6.4 | 17.7 | 28.9 | 47.66 |
| CosPlace [33] | 0 | 1.4 | 91.6 | **100** | 8.1 | 14.2 | 33.23 | 38.53 |
| MixVPR [5] | 32.4 | 42.2 | 76.1 | 98.6 | 10.3 | 18.3 | 39.6 | 53.03 |
| CLIP-CLS [66] | 42.2 | 70.4 | 62.0 | 97.2 | 36.6 | 52.8 | 46.93 | 73.46 |
| DINO-CLS [51] | 57.8 | 90.1 | 84.5 | **100** | 24.0 | 38.4 | 55.43 | 76.16 |
| DINOv2-CLS [21] | 73.2 | 88.7 | 71.8 | 91.6 | 45.2 | 59.9 | 63.4 | 80.06 |
| AnyLoc-GeM-DINOv2 | **76.1** | 83.1 | 57.8 | 97.2 | 38.3 | 53.8 | 57.4 | 78.03 |
| AnyLoc-VLAD-DINO | 43.7 | 54.9 | **94.4** | **100** | 17.8 | 28.7 | 51.96 | 61.2 |
| **AnyLoc-VLAD-DINOv2** | **76.1** | **94.4** | 85.9 | **100** | **66.7** | **79.2** | **76.23** | **91.2** |

Too few database images to do PCA!

# Unstructured

| Methods | Hawkins | | Laurel Caverns | | Mid-Atlantic Ridge | |
|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 |
| NetVLAD [144] | 34.8 | 71.2 | 39.3 | 71.4 | 25.7 | 53.5 |
| CosPlace [33] | 31.4 | 59.3 | 24.1 | 47.3 | 20.8 | 40.6 |
| MixVPR [5] | 25.4 | 60.2 | 29.5 | 67.0 | 25.7 | 60.4 |
| CLIP-CLS [66] | 33.0 | 67.0 | 36.6 | 66.1 | 25.7 | 51.5 |
| DINO-CLS [51] | 46.6 | 84.8 | 41.1 | 57.1 | 27.7 | 49.5 |
| DINOv2-CLS [21] | 28.0 | 62.7 | 40.2 | 65.2 | 24.8 | 48.5 |
| AnyLoc-GeM-DINOv2 | 53.4 | 83.9 | 58.9 | 86.6 | 14.8 | 49.5 |
| AnyLoc-VLAD-DINO | 48.3 | 84.8 | 57.1 | 79.5 | **41.6** | **66.3** |
| **AnyLoc-VLAD-DINOv2** | **65.2** | **94.1** | **61.6** | 90.2 | 34.6 | 61.4 |

# Vocabulary Transfer

| Vocabulary Dataset | Evaluation Dataset | Map-Specific R@1 | Vocab-Specific R@1 |
|---|---|---|---|
| Baidu Mall (0.7k) | 17 Places (0.4k) | **64.5** | 63.8 |
| | Gardens Point (0.2k) | **98.0** | 94.5 |
| VP-Air (2.7k) | Nardo-Air (0.1k) | 57.8 | **64.8** |
| | Nardo-Air R (0.1k) | 70.4 | **88.7** |
| Pitts-30k (10k) | Oxford (0.2k) | 94.8 | **99.0** |

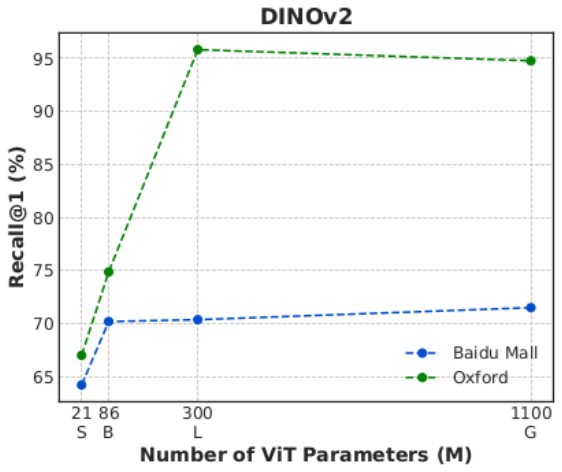
**Vocab-Specific**: Fit clusters on the vocabulary dataset and test on evaluation dataset

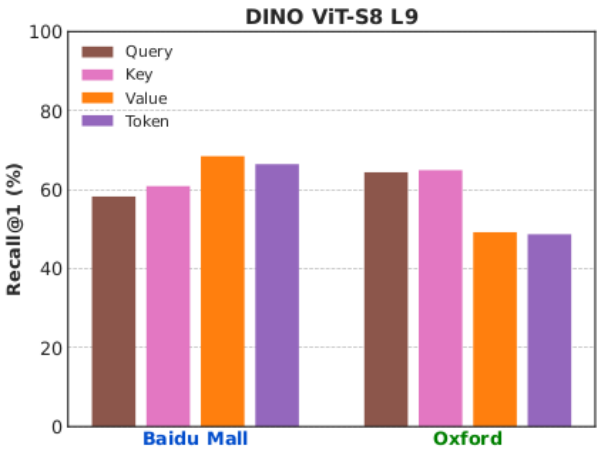
**Map-Specific**: Fit clusters on the evaluation dataset and test on the same dataset

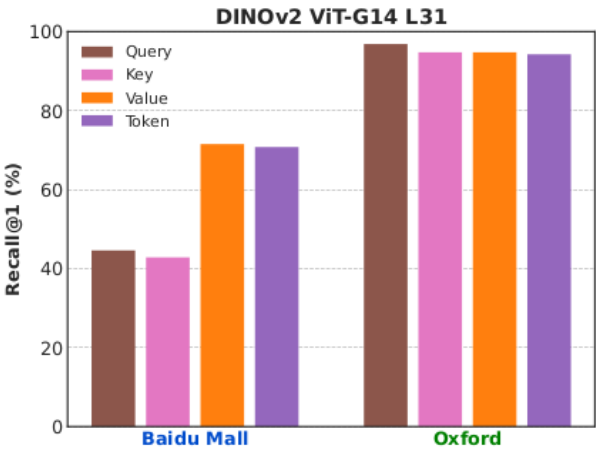
Using a single, large, good quality vocabulary dataset also works

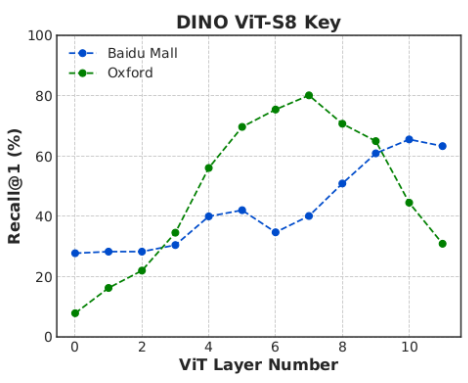
=> You might not need all datasets from the domain

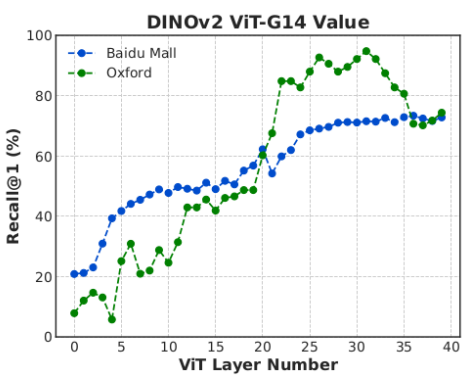
# Ablations




- Larger models are better
- DINO: Intermediate layer & key facet
- DINOv2: Penultimate layer & value facet

# Ablations

| Method | Indoor | Urban | Aerial | SubT & D | Underwater |
|---|---|---|---|---|---|
| ViT-B CosPlace | 62.9 | 80.7 | 26.3 | 26.5 | 18.8 |
| ViT-B CosPlace VLAD | 68.5 | 82.9 | 38.4 | 37.5 | 23.8 |
| ViT-S AnyLoc-VLAD-DINO | 72.9 | 79.6 | 47.8 | 52.7 | **41.6** |
| ViT-B AnyLoc-VLAD-DINOv2 | 77.0 | 82.6 | 53.6 | 60.2 | 35.6 |
| ViT-G AnyLoc-VLAD-DINOv2 | **78.0** | **92.3** | **62.9** | **63.4** | 34.6 |

DINO > CosPlace

| Aggregation Methods | DINO | | | DINOv2 | | |
|---|---|---|---|---|---|---|
| | Baidu ↑ | Oxford ↑ | Dim ↓ | Baidu ↑ | Oxford ↑ | Dim ↓ |
| Global Average Pooling (GAP) | 29.6 | 28.8 | **384** | 41.6 | 78.5 | **1536** |
| Global Max Pooling (GMP) | 34.9 | 38.2 | **384** | 64.4 | 74.9 | **1536** |
| Generalized Mean Pooling (GeM) | 34.7 | 47.6 | **384** | 50.1 | 92.2 | **1536** |
| Soft Assignment VLAD | 33.8 | 28.3 | 49152 | 40.3 | 82.2 | 49152 |
| Hard Assignment VLAD | **60.9** | **64.9** | 49152 | **71.5** | **94.8** | 49152 |

Use VLAD!

# Conclusions

- Good for offline, **too slow for offline**: Make it portable with TensorRT and try distillation with smaller models (ViT-G has 1B+ parameters!)
- Register tokens: remove/hide noise in latent activations
- Other foundation models
- Other forms of aggregation

Thank You