ICASSP 2021 TORONTO Canada
June 6-11, 2021
Metro Toronto Convention Centre

Hierarchical Attention Fusion for Geo-Localization

Liqi Yan, Yiming Cui, Yingjie Chen, Dongfang Liu

June. 2021

Outline

1. Introduction
2. Method
3. Experiments
4. Conclusion

1 Introduction

1 Introduction

Related Works for Geo-localization Task

- **Problem: Landmarks with medium or small sizes are difficult to be recognized.** (because CNNs intend to down-sample the spatial resolution of the input image by a significant margin [4,7,8])

- **Reason: Only using features from one semantic level.** (The feature maps from a single semantic level fail to fully explore rich visual clues from landmarks of different scales.)

[4] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5297–5307.

[7] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm, "Learned contextual feature reweighting for image geolocalization," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2136–2145.

[8] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich, "Superpoint: Self-supervised interest point detection and description," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 224–236.

1 Introduction

Related Works for Geo-localization Task

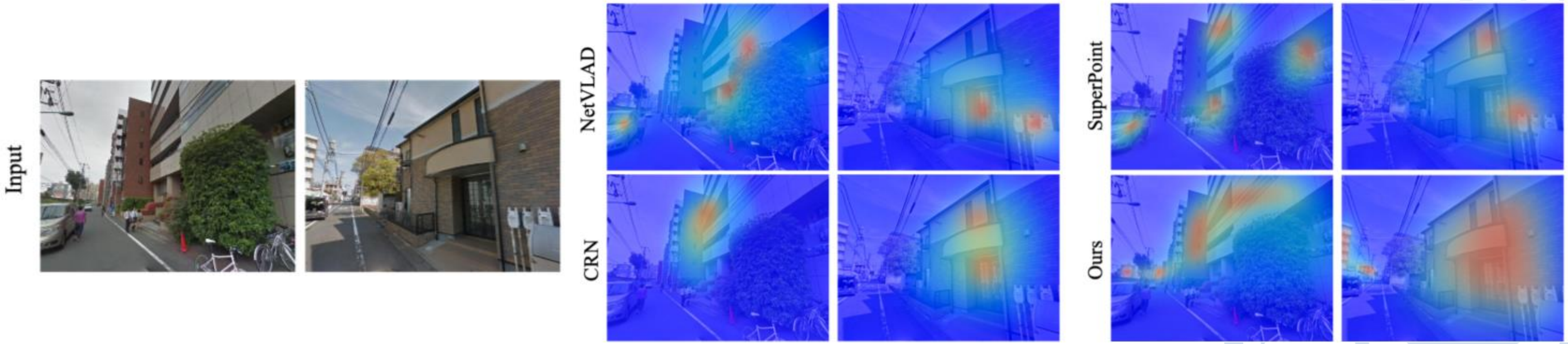


Fig. 1: Comparison of feature emphasis. Compared to conventional methods [4,7,8], our method exploits the multiscale features for hierarchical attention to depict image representation of landmarks with different scales and distance.

[4] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5297–5307.

[7] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm, "Learned contextual feature reweighting for image geolocalization," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2136–2145.

[8] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich, "Superpoint: Self-supervised interest point detection and description," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 224–236.

1 Introduction

ICASSP 2021 TORONTO Canada June 6-11, 2021 Metro Toronto Convention Centre

Principal Contributions

- A **hierarchical attention fusion network**, a novel algorithm for geo-localization.
- A **self-supervised loss function** to captures pairwise image relationships in training.
- Experimental results demonstrate that the proposed method sets **a new state-of-the-art** on several geo-localization benchmarks.

Method

2

2 Method

Architecture

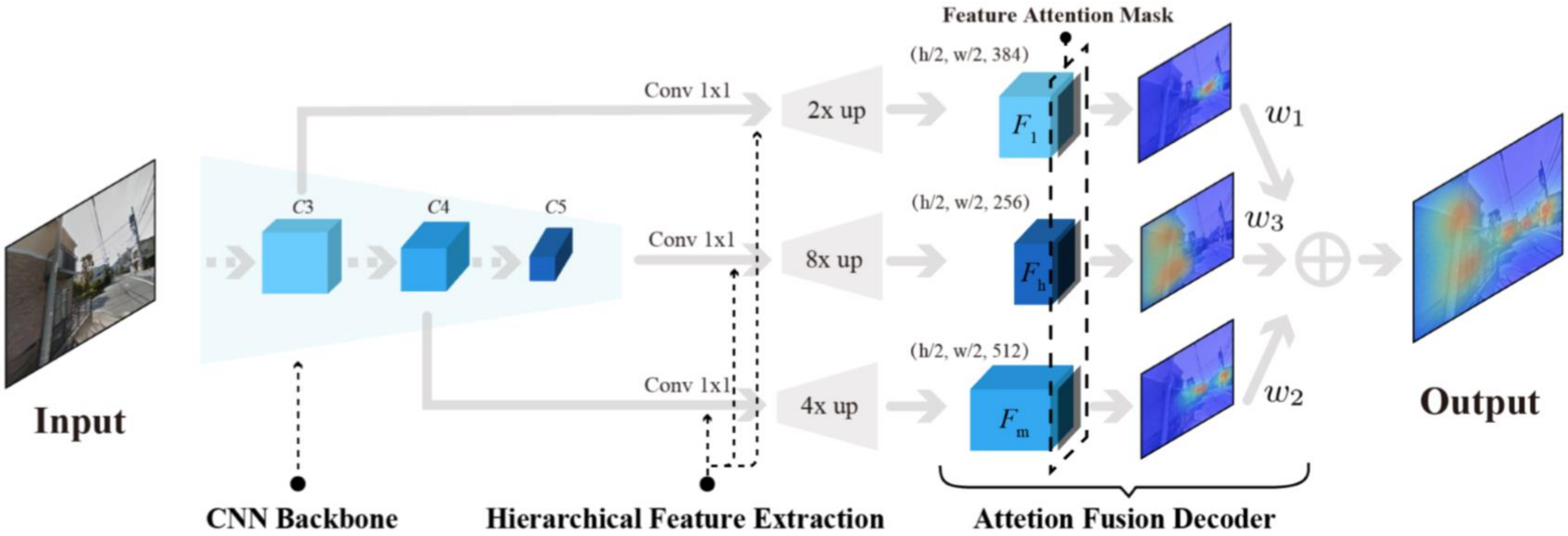
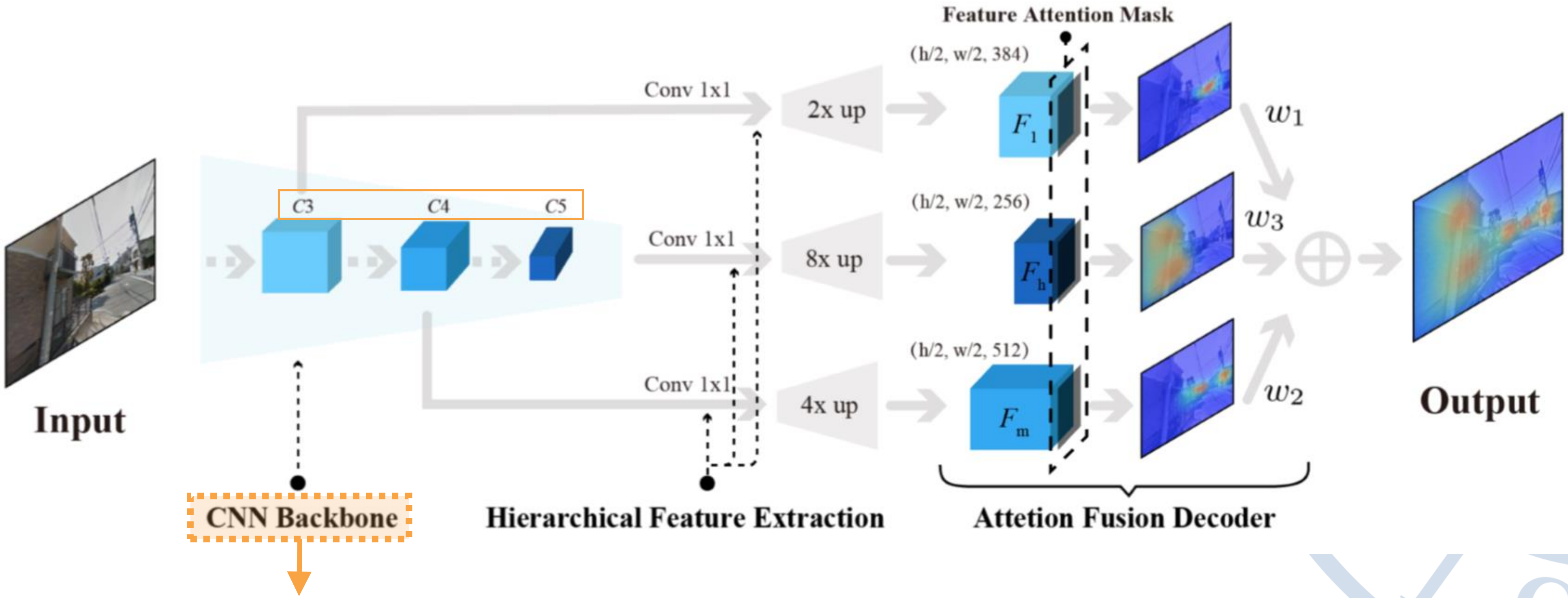


Fig. 2: The architecture of the proposed method. Our method uses hierarchical features to close the semantic gap in feature learning. We perform the attention fusion over the obtained features to produce strong image representation for landmarks with different scales.

2 Method

Hierarchical Feature Extraction

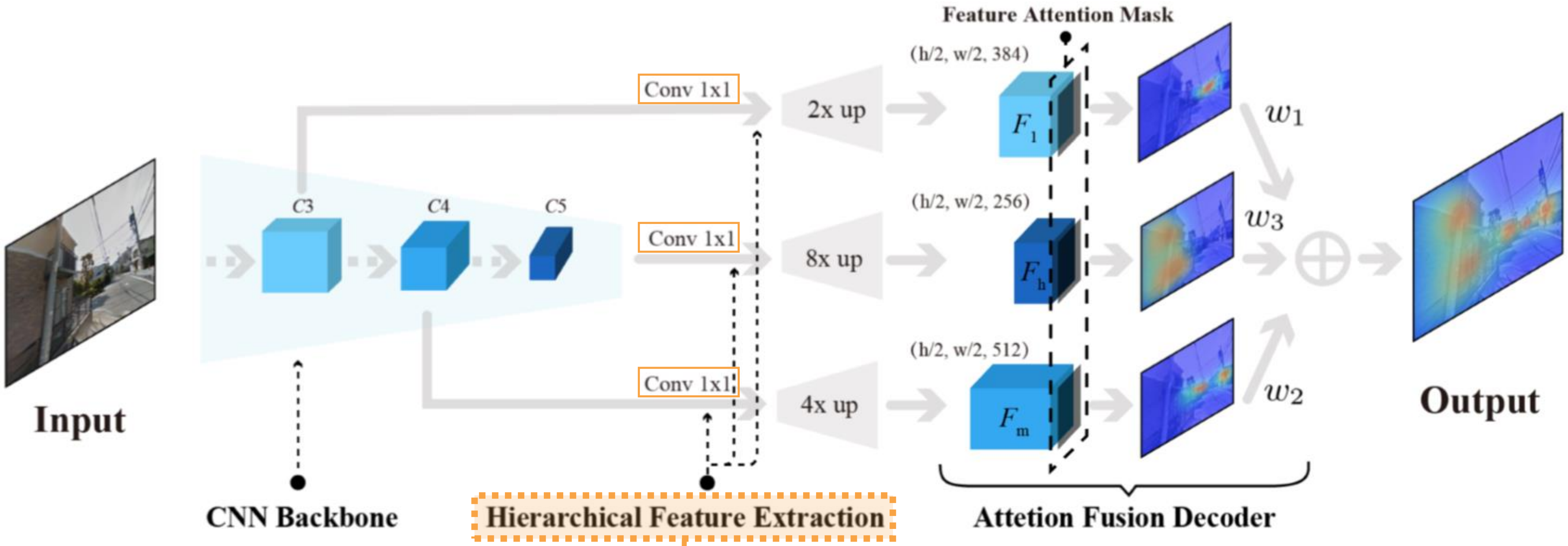


- We use **VGG16** [9] as the backbone network for feature extraction. We extract hierarchical features from **Con3_2**, **Con4_3**,and **Con5_3** respectively.

[9] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

2 Method

Hierarchical Feature Extraction

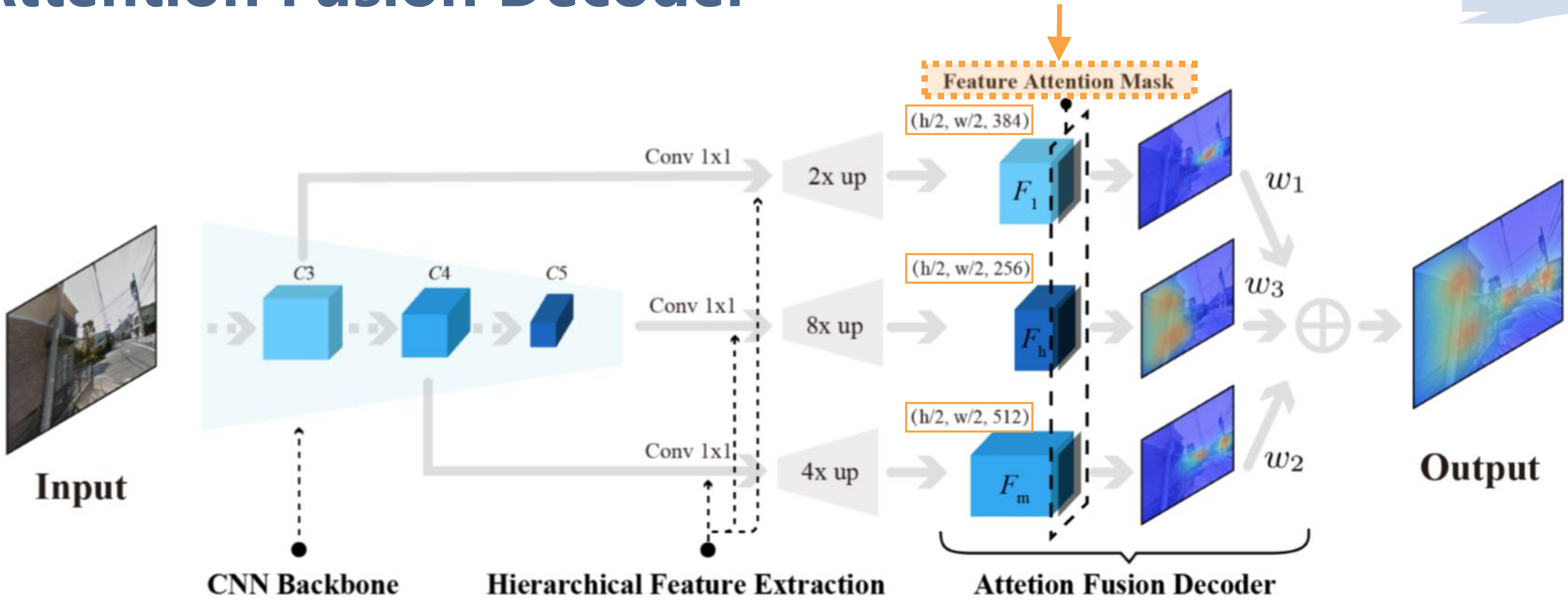


- The obtained hierarchical feature maps are then processed by a modified SuperPoint structure [8].

[8] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich, "Superpoint: Self-supervised interest point detection and description," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 224–236.

2 Method

Attention Fusion Decoder



- **Feature attention mask.** We implement three learnable feature attention masks $\{m_1, m_2, m_3\}$ which are appended to $\{F_l, F_m, F_h\}$ separately. We define the **attention-weighted features** as:

$$F_l' = \sum_{n=1}^{x} \sum_{r \in R} m^r \cdot f_n^r, F_m' = \sum_{n=1}^{y} \sum_{r \in R} m^r \cdot f_n^r, F_h' = \sum_{n=1}^{z} \sum_{r \in R} m^r \cdot f_n^r \quad \begin{aligned} F_l &= \{f_1, \dots, f_x\} \\ F_m &= \{f_1, \dots, f_y\} \\ F_h &= \{f_1, \dots, f_z\} \end{aligned} \qquad (1)$$

where R denotes a set of spatial regions on the feature map.

2 Method

Attention Fusion Decoder

- **Coupled descriptor and detector.**

- Using the attention-weighted features F', we define **descriptor** as a set of vectors K:

$$K = \sum_{i=1}^{h} \sum_{j=1}^{w} F'^{ij:}, K^{ij} \in \mathbb{R}^{x} \tag{2}$$

➢ Kij is the Euclidean distance of each descriptor between images at each pixel point (i, j).

- Thus the **detectors** D can be denoted as:

$$D = \sum_{n=1}^{x} F'^{::n}, D^{n} \in \mathbb{R}^{h \times w} \tag{3}$$

- We then perform an image-wise normalization of the detection to obtain the detection score at a pixel (i, j):

most strong detection on the response maps

$$s_{ij} = \frac{\boxed{D_l^{(ij)n'}}}{\sum_{i'=1}^{h} \sum_{j'=1}^{w} D_l^{(ij)n'}} \tag{4}$$

2 Method

Training Objective

- For a pair of image (I_q, I_r):

- We include a detection term to compute their **differences** in feature space:

$$\Delta\mathcal{D}(I_q, I_r) = \sum_{c\in\mathcal{C}} \frac{s_q^{c'} s_r^{c'}}{\sum_{c'\in\mathcal{C}} s_q^{c'} s_r^{c'}} \underbrace{\left\| K_q^c - K_r^c \right\|_2}_{\text{descriptor distance}} \quad (5)$$

- Thus, the **triple ranking loss** is defined as:

$$\mathcal{L}(I_q, I_r^+, I_r^-) = \max\left(M + \Delta\mathcal{D}\left(I_q, \underbrace{I_r^+}_{\text{positive reference}}\right) - \Delta\mathcal{D}\left(I_q, \underbrace{I_r^-}_{\text{negative reference}}\right), 0\right) \quad (6)$$

- Our **overall loss** is:

$$\mathcal{L}_{total} = w_1 \cdot \mathcal{L}_l + w_2 \cdot \mathcal{L}_m + w_3 \cdot \mathcal{L}_h, \qquad (w_1 + w_2 + w_3 = 1) \quad (7)$$

Notes: $\mathcal{C}$ indicates all the corresponding feature points between the two images. s is the detection scores in (4). $\mathcal{L}_l$, $\mathcal{L}_m$ and $\mathcal{L}_h$ are individual loss for each hierarchical attention.

3

Experiments

3 Experiments

Implementation Setup

Optimizer:

- 30 **epochs**, **learning rate** 0.0001 which is **halved** in every 5 epochs,
- **Momentum** 0.9, **weight decay** 0.001, and a **batch size** of 4 triplets.

Loss function:

- $w_1 = 0.1$, $w_2 = 0.4$, and $w_3 = 0.5$.

Inference:

- The trained models which yield the **best recall@5** on the validation set is used for testing.

3 Experiments

Evaluation Datasets and Metrics

Two types of Benchmarks:

- **Image retrieval datasets:**
  - Oxford5k
  - Paris6k
  - Holidays

  Evaluated by: mean-Average-Precision (mAP)

- **Geo-localization datasets:**
  - Pitts250k-test
  - Tokyo 24/7
  - Tokyo TM val
  - Sf-0

  Evaluated by: Precision-Recall curve

Oxford5k Dataset

Pitts250k Dataset

3 Experiments

Empirical Results

We compare our method with the state-of-the-art methods, NetLAVD, CRN, and SuperPoint

Image retrieval benchmarks:

Method	Oxford 5K		Paris 6k		Holidays	
	full	crop	full	crop	orig	rot
Ours	**67.81**	**69.52**	**75.10**	**78.29**	**84.82**	**88.41**
CRN	63.95	65.52	72.88	75.85	83.19	87.30
NetVLAD	63.09	65.33	72.53	75.67	82.67	86.83
SuperPoint	63.14	65.50	72.83	75.10	82.92	86.90

Table1: Results for compact image representations (256-D).

On all metrics,
our margins consistently exceed the mAP of other methods by **1 to 5%↑**.

3 Experiments

Empirical Results

We compare our method with the state-of-the-art methods, NetLAVD, CRN, and SuperPoint

Geo-localization benchmarks:

- ✓ Effectively exploit multi-scale features.
- ✓ The capacity of having hierarchical attentions on landmarks with different scales and distances.
- ✓ Focusing on the distinctive details of buildings.
- ✓ Avoiding confusing objects such as pedestrians, vegetation, or vehicles which are hard for feature repeatability.

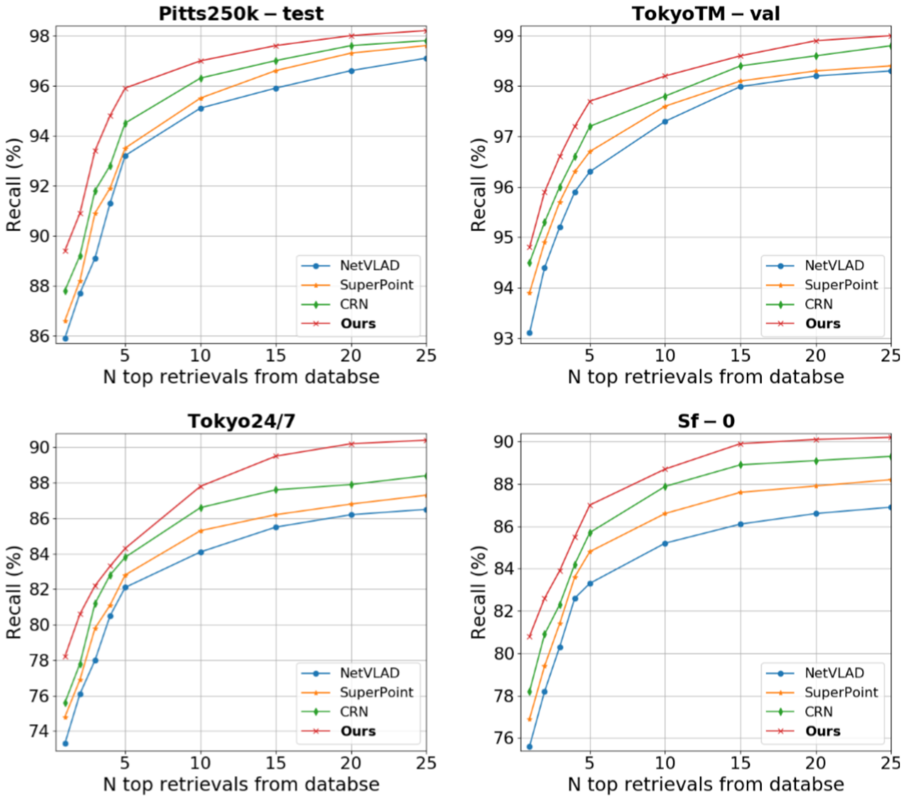


Fig. 3: Comparison of recalls at N top retrievals with the state-of-the-arts methods.

3 Experiments

Empirical Results

Adaptive Weight Analysis:

Method	Pitts 250k-test	TokyoTM-val	Tokyo 24/7	Sf-0
w_1	0.1	0.3	0.2	0.1
w_2	0.4	0.3	0.3	0.1
w_3	0.5	0.4	0.5	0.8

Table2: Best adaptive weights for each benchmarks.

w_1: lower-level features (small scale), w_2: mid-level features (middle scale), w_3: higher-level features (large scale)

- **Pitts 250k-test** focuses on **middle** and **large-scale** buildings.
- **TokyoTM** generally includes small-, middle-, and large-scale buildings.
- **Tokyo 24/7** includes a lot of **landmark details** such as billboards, city lights, or traffic signs by the road.
- **Sf-0** has a dominant w_3 as it mainly focuses on buildings with a **large scale**.

4

Conclusion

4 Conclusion

Empirical Results

- **A hierarchical attention fusion network** for geo-localization.
- **Approach:** Extracting the **multi-scale feature maps** from a convolutional neural network (CNN) to perform **hierarchical attention fusion** for image representations.
- **Advantage:** Since the hierarchical features are **scale-sensitive**, our method is **robust** to landmarks with **different scales and distances**.
- **Experimental Results:** indicate that our method is **competitive** with the latest state-of-the-art approaches on the image retrieval benchmarks and the large-scale geo-localization benchmarks.

THANKS

It's time for Q&A