Hierarchical Attention Fusion for Geo-Localization

Liqi Yan[1], Yiming Cui[2], Yingjie Chen[3], Dongfang Liu[3]*

（[1]Fudan University, China）

（[2]University of Florida , USA）

（[3]Purdue University, USA）

*liu2538@purdue.edu

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

INTRODUCTION

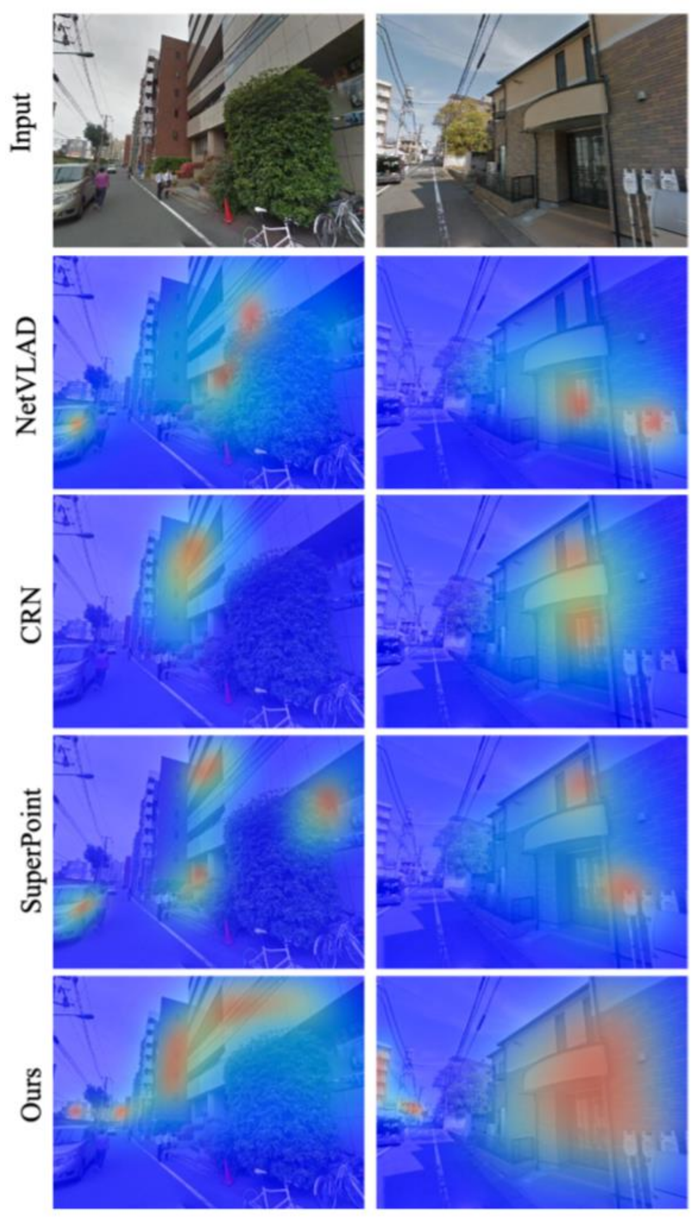

Fig.1. Comparison of feature emphasis.

Related Works

- **Problem:** Landmarks with medium or small sizes are difficult to be recognized.
- **Reason:** Concurrent methods [1-3] only use features from one semantic level.
- **Our method:** Exploiting the multiscale features for hierarchical attention to depict image representation of landmarks with different scales and distance, as shown in Fig. 1.

Contribution:

- A hierarchical attention fusion network.
- A self-supervised loss function.
- A new state-of-the-art on several geo-localization benchmarks.

METHODS

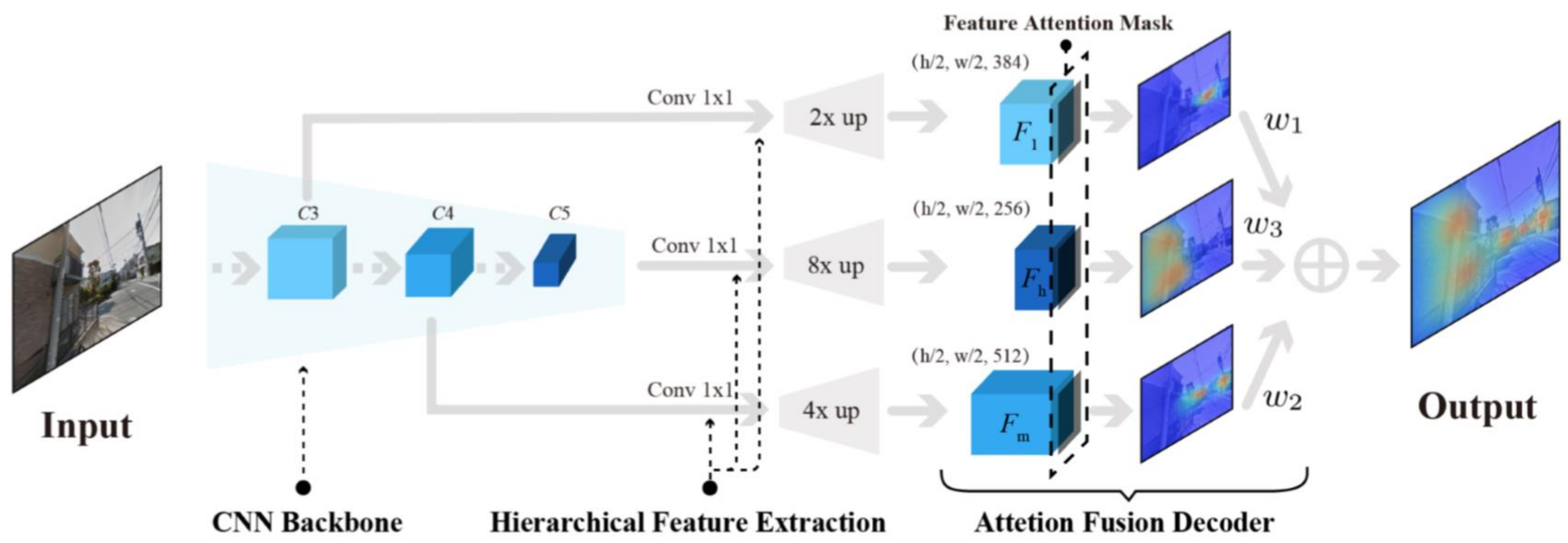

Fig.2. The architecture of the proposed method

Network Architecture: As Fig. 2 shows, we perform the attention fusion over the obtained features to produce strong image representation for landmarks with different scales.

Training Objective: For a pair of image (I_q, I_r), we include a detection term to compute their differences:

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

Thus, the triple ranking loss is defined as:

$$\mathcal{L}(I_q, I_r^+, I_r^-) = max(M + \Delta\mathcal{D}(I_q, I_r^+) - \Delta\mathcal{D}(I_q, I_r^-), 0)$$

RESULTS

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).

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 which produces the best recall@5 for each benchmarks.

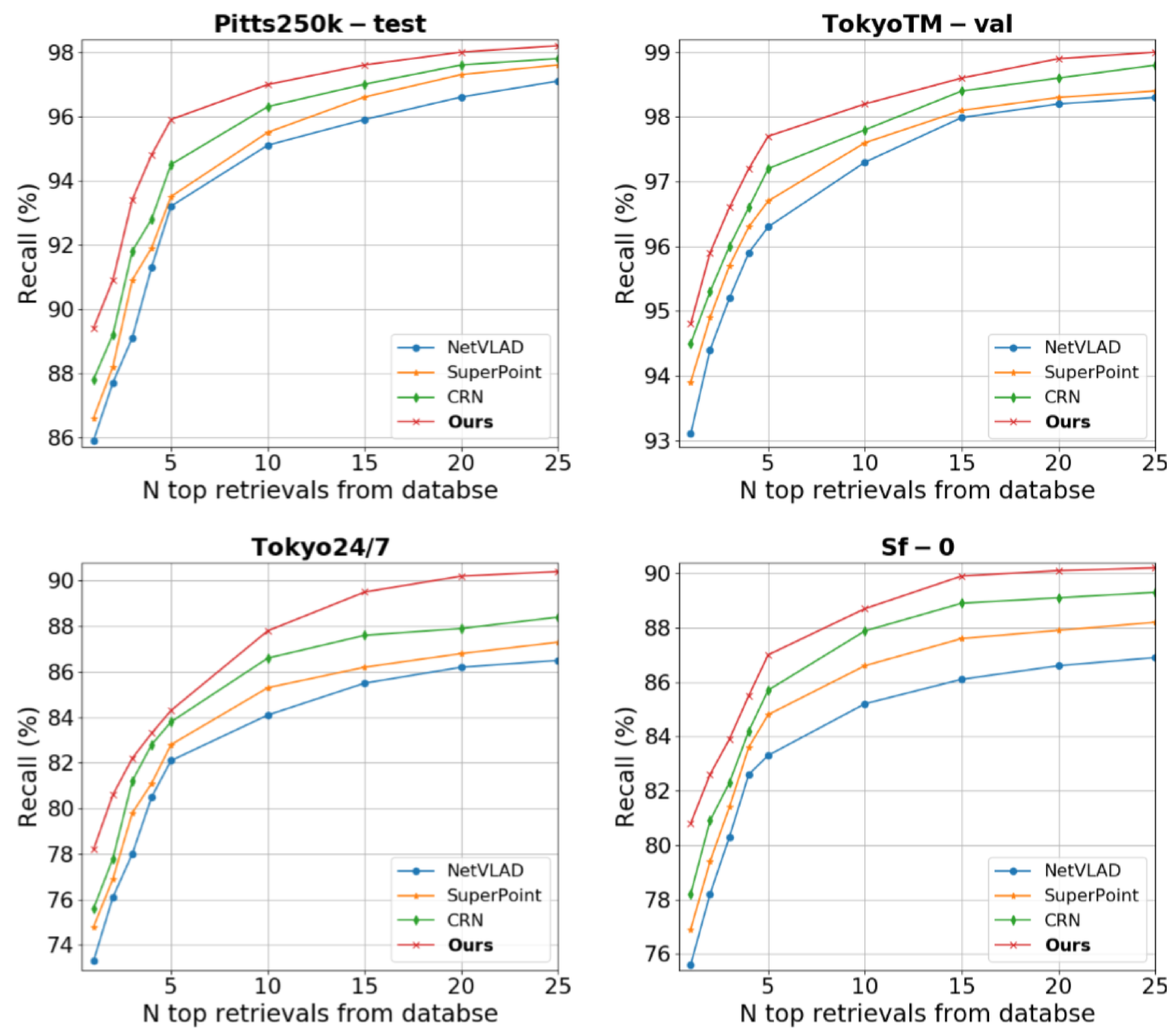

Fig. 3: Comparison of recalls at N top retrievals with the state-of-the-arts methods.

We compare our method with the state-of-the-art methods, NetLAVD [1], CRN [2], and SuperPoint [3].

- **Image retrieval benchmarks:**

The results are displayed in **Table 1**. Our results set the state-of-the-art for compact image representations (256-D) on all three benchmarks. On all metrics, our margins consistently exceed the mAP of other methods by 1 to 5%.

- **Geo-localization benchmarks:**

We report the Precision-Recall plot for each method in **Fig. 3**. Our method outperforms other methods under different recall@n thresholds on all benchmarks.

CONCLUSION

- **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.

REFERENCES

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in Proceedings of CVPR, 2016, pp. 5297–5307.

[2] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm, "Learned contextual feature reweighting for image geolocalization," in Proceedings of CVPR, 2017, pp. 2136–2145.

[3] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich, "Superpoint: Self-supervised interest point detection and description," in Proceedings of CVPR Workshop, 2018, pp. 224–236.