Outline

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

---


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.


1 Introduction

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.
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.

2 Method
Hierarchical Feature Extraction

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

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:

\[
F'_l = \sum_{n=1}^{x} \sum_{r \in R} m^r \cdot f^r_n, \
F'_m = \sum_{n=1}^{y} \sum_{r \in R} m^r \cdot f^r_n, \
F'_h = \sum_{n=1}^{z} \sum_{r \in R} m^r \cdot f^r_n
\]

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

\[
F_l = \{f_1, ..., f_x\} \
F_m = \{f_1, ..., f_y\} \
F_h = \{f_1, ..., f_z\}
\]
2 Method

Attention Fusion Decoder

- Coupled descriptor and detector.

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

\[
K = \sum_{i=1}^{h} \sum_{j=1}^{w} F'_{ij}, K_{ij} \in \mathbb{R}^x
\]  

- $K_{ij}$ is the Euclidean distance of each descriptor between images at each pixel point $(i,j)$.

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

\[
D = \sum_{n=1}^{x} F^{n:n}, D^n \in \mathbb{R}^{h \times w}
\]  

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

\[
S_{ij} = \frac{D_{l}^{(ij)n'}}{\sum_{i'=1}^{h} \sum_{j'=1}^{w} D_{l}^{(ij)n'}}
\]  

most strong detection on the response maps
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:

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

  Thus, the triple ranking loss is defined as:

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

- Our overall loss is:

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

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.
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
3 Experiments

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

Image retrieval benchmarks:

<table>
<thead>
<tr>
<th>Method</th>
<th>Oxford 5K</th>
<th>Paris 6k</th>
<th>Holidays</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>full</td>
<td>crop</td>
<td>full</td>
</tr>
<tr>
<td>Ours</td>
<td>67.81</td>
<td>69.52</td>
<td>75.10</td>
</tr>
<tr>
<td>CRN</td>
<td>63.95</td>
<td>65.52</td>
<td>72.88</td>
</tr>
<tr>
<td>NetVLAD</td>
<td>63.09</td>
<td>65.33</td>
<td>72.53</td>
</tr>
<tr>
<td>SuperPoint</td>
<td>63.14</td>
<td>65.50</td>
<td>72.83</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th>Method</th>
<th>Pitts 250k-test</th>
<th>TokyoTM-val</th>
<th>Tokyo 24/7</th>
<th>Sf-0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_1$</td>
<td>0.1</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>$w_2$</td>
<td>0.4</td>
<td>0.3</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>$w_3$</td>
<td>0.5</td>
<td>0.4</td>
<td>0.5</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Table 2: 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

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