Cross-view Image Geo-localization

August 25, 2022
Academic talk at Qualcomm

Dr. Chen Chen
Assistant Professor
Center for Research in Computer Vision
University of Central Florida
https://www.crcv.ucf.edu/chenchen/
chen.chen@crcv.ucf.edu
Outline

- Introduction (image geo-localization)

- Cross-view image geo-localization
  - Orientational alignment in image geo-localization
  - Spatial alignment in image geo-localization
  - Vision transformer for image geo-localization

- Future work
What is image geo-localization?

What’s the localization of the place?

Input

Visual Information (Images)

Output

Location in terms of Longitude and Latitude

40.4419, -79.9986
What is image geo-localization?

Query street-view image

GPS location?
(Latitude, Longitude) = (40.441426, -80.003586)

Geo-tagged reference database

Find match

Street-view images (i.e., same view)

Top-1 match (ranked by similarity)

(Latitude, Longitude) = (40.441426, -80.003586)
What is image geo-localization? 

Query street-view image 

Geo-tagged reference database 

GPS location? 

(Angle, Longitude) = (40.441426, 80.003586) 

Find match 

Aerial images (i.e., cross view)
How to define similarity?

Similar? ✔

Similar? ✗

Similar? ✗
Learn similarity via metric learning

Query → Model → Embedding Space

Positive → Model

Negative → Model

Triple Loss
Why is image geo-localization important?

• Accurate Visual Localization for Automotive Applications

Why is image geo-localization important?

• Cross-View Policy Learning for Street Navigation

Why is image geo-localization important?

• UAV Pose Estimation using Cross-view Geo-localization

Cross-view image geo-localization

- Only small number of cities in the world are covered by ground-level imagery
- A more complete coverage for overhead reference data such as satellite/aerial imagery

GPS location?

(Latitude, Longitude) = (40.441426, -80.003586)
Outline

• Introduction (image geo-localization)

• Cross-view image geo-localization
  • Orientational alignment in image geo-localization
  • Spatial alignment in image geo-localization
  • Vision transformer for image geo-localization

• Future work
Revisiting Street-to-Aerial View Image Geo-localization and Orientation Estimation


High Accuracy Depends on Assumptions

• Orientation Alignment

In real-world applications, images may not be geometrically aligned.
Research problems

• How would the alignment information affect the retrieval model in terms of performance?

• Without assuming the inference image pairs are aligned, how to effectively improve the retrieval performance?

• Is it possible to estimate the alignment information when no explicit supervision is given?
Research problems

• How would the alignment information affect the retrieval model in terms of performance?

• Without assuming the inference image pairs are aligned, how to effectively improve the retrieval performance?

• Is it possible to estimate the alignment information when no explicit supervision is given?
Impact of orientation alignment

Top-1 recall accuracy with different alignment settings

<table>
<thead>
<tr>
<th>Validation</th>
<th>Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aligned</td>
<td>60.1%</td>
</tr>
<tr>
<td>Rotate</td>
<td>13.5%</td>
</tr>
<tr>
<td>Aligned</td>
<td>43.7%</td>
</tr>
<tr>
<td>Rotate</td>
<td>44.2%</td>
</tr>
</tbody>
</table>
Research problems

• How would the alignment information affect the retrieval model in terms of performance?

• Without assuming the inference image pairs are aligned, how to effectively improve the retrieval performance?

• Is it possible to estimate the alignment information when no explicit supervision is given?
Overall Framework (network architecture)

Metric learning techniques are independent of the alignment assumption
Matching Loss

Triplet loss function

\[
L = \frac{1}{N} \sum_{i}^{N} \max(0, d_i^p - d_i^n + m)
\]

\(d_i^p\) and \(d_i^n\) denote the distance between the i-th anchor and its positive and negative samples

\(N\): N triplets in a batch

\(m\): a positive margin parameter

Weighted soft-margin loss function

\[
L = \frac{1}{N} \sum_{i}^{N} \sigma(\alpha(d_i^p - d_i^n)), \quad \alpha > 0
\]

soft-margin function \(\sigma(d) = \log(1 + \exp(d))\)

Matching Loss

Binomial deviance loss (Yi et al.)

\[ L = \frac{1}{N_p} \sum_{i}^{N_p} \sigma(-\alpha(s_i^p - m)) + \frac{1}{N_n} \sum_{i}^{N_n} \sigma(\alpha(s_i^n - m)) \]

\( s_i^p \) and \( s_i^n \) denote the cosine similarity between the i-th anchor and its positive and negative samples

\( N_p \) and \( N_n \) represent the number of positive and negative pairs

\( m \) : a positive margin parameter

\[ \sigma(d) = \log(1 + \exp(d)) \]

Matching Loss

Our new loss function

\[
L = \sum_{i}^{N_p} \sigma \left( -\alpha_p \left( \frac{s_i^p - m_p}{\alpha_p N_p} \right) \right) + \sum_{i}^{N_n} \sigma \left( \frac{\alpha_n \left( s_i^n - m_n \right)}{\alpha_n N_n} \right)
\]

When positive samples are much fewer than negative samples, as in cross-view geo-localization with only one positive match, it would be easier to pulling the only matched sample close to the anchor rather than pushing all negative samples away (i.e., assign a much smaller value to \( \alpha_p \) than \( \alpha_n \)).
Geo-localization Results

<table>
<thead>
<tr>
<th>Method</th>
<th>CVUSA</th>
<th>Vo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1%</td>
<td>Top-1</td>
</tr>
<tr>
<td>Scott [22] (ICCV’15)</td>
<td>34.3%</td>
<td>-</td>
</tr>
<tr>
<td>Zhai [26] (CVPR’17)</td>
<td>43.2%</td>
<td>-</td>
</tr>
<tr>
<td>Vo [21] (ECCV’16)</td>
<td>63.7%</td>
<td>-</td>
</tr>
<tr>
<td>CVMNet [8] (CVPR’18)</td>
<td>93.6%</td>
<td>22.5%</td>
</tr>
<tr>
<td>Lending [13] (CVPR’19)</td>
<td>93.19%</td>
<td>31.71%</td>
</tr>
<tr>
<td>Reweight [3] (ICCV’19)</td>
<td>98.3%</td>
<td>46.0%</td>
</tr>
<tr>
<td>GAN [14] (ICCV’19)</td>
<td>95.98%</td>
<td>48.75%</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>97.7%</strong></td>
<td><strong>54.5%</strong></td>
</tr>
</tbody>
</table>

Table 2: Top-1 and top-1% recall accuracy comparison on CVUSA and Vo datasets.

“R@k”: If the ground-truth reference image appears in the top k retrieved images, it is considered as correct
Ablation study – effect of alignment

<table>
<thead>
<tr>
<th>Method</th>
<th>w/ alignment</th>
<th>w/o alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1%</td>
<td>Top-1</td>
</tr>
<tr>
<td><strong>Baseline</strong> (soft-margin loss)</td>
<td>98.8%</td>
<td>60.1%</td>
</tr>
<tr>
<td><strong>Ours</strong> (proposed loss)</td>
<td>99.1%</td>
<td>70.4%</td>
</tr>
</tbody>
</table>

w/o alignment: images are randomly rotated in the training and test sets

- The improvements of the proposed technique are consistent across both settings
Geo-localization Examples
Geo-localization Examples
Geo-localization Examples

A failure case
Visual Explanation of the Matching Results

• Visual explanation using Grad-CAM
What is Grad-CAM?

- Gradient-weighted Class Activation Mapping (Grad-CAM)

Visual Explanation of the Matching Results

The most activated regions are likely to be the same objects.
Visual Explanation of the Matching Results

Street view id: 49288

Aerial view positive, id: 49288

Similarity: 0.74
Research problems

• How would the alignment information affect the retrieval model in terms of performance?

• Without assuming the inference image pairs are aligned, how to effectively improve the retrieval performance?

• Is it possible to estimate the alignment information when no explicit supervision is given?
Orientation Estimation with Grad-CAM

We find the Grad-CAM activation maps have the rotation-invariant property!
The angle distributions of activated pixels from two views would be similar if the image pair is well aligned. Find the angle $\phi$ so that $p_{aerial}(\theta + \phi)$ best matches $p_{street}(\theta)$. 
Orientation Estimation Example

126°

Circular Convolution

125.2°
Conf = 0.95
Summary

• Ablation study and visual explanation lead to a key observation – the orientation alignment has a great impact on the retrieval performance (overlooked by prior work)

• We show that improvements on metric learning techniques boost the retrieval performance

• We discover that the orientation information between cross-view images can be estimated when the alignment is unknown
More on visual explanation

Outline

• Introduction (image geo-localization)

• Cross-view image geo-localization
  • Orientational alignment in image geo-localization
  • Spatial alignment in image geo-localization
  • Vision transformer for image geo-localization

• Future work
VIGOR: Cross-View Image Geo-localization beyond One-to-one Retrieval

Spatial alignment

• Existing works simply assume that each query ground-view image has one corresponding reference aerial-view image whose center is exactly aligned at the location of the query image.

• This is not practical for real-world applications, because the query image may be generated at arbitrary locations in the area of interest and the reference images should be captured before the queries emerge.
Spatial alignment

Query

Reference

Query images can be at arbitrary locations.
May not be at the center of any reference aerial image.
VIGOR dataset

Dataset Setting: given an area of interest (AOI), the reference aerial images are densely sampled to achieve a seamless coverage of the AOI and the street-view queries are captured at arbitrary locations.
VIGOR dataset

Dataset Setting: beyond one-to-one correspondence (one to many)
Data Distribution

Manhattan  Chicago  San Francisco  Seattle
Datasets Comparison

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Satellite images</td>
<td>~ 450,000</td>
<td>128,334</td>
<td>44,416</td>
<td>90,618</td>
</tr>
<tr>
<td>Panoramas in total</td>
<td>~ 450,000</td>
<td>128,334</td>
<td>44,416</td>
<td>238,696</td>
</tr>
<tr>
<td>Panoramas after balancing</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>105,214</td>
</tr>
<tr>
<td>Street-view GPS locations</td>
<td>Aligned</td>
<td>Aligned</td>
<td>Aligned</td>
<td>Arbitrary</td>
</tr>
<tr>
<td>Full panorama</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Multiple cities</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Orientation information</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Evaluation in terms of meters</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Seamless coverage on area of interest</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Number of references covering each query</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>
Coarse-to-fine Cross-view Localization

Beyond One-to-one

How to make use of the semi-positive images?

Directly considering semi-positive as positive results in a low accuracy.

We force the ratio of the similarities in the embedding space to be close to the ratio of IOUs.

IOU-based semi-positive assignment loss

$$\mathcal{L}_{IOU} = \left( \frac{S_{semi}}{S_{pos}} - \frac{IOU_{semi}}{IOU_{pos}} \right)^2$$

<table>
<thead>
<tr>
<th>Semi-positive Assignment</th>
<th>Same-Area</th>
<th></th>
<th>Cross-Area</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1</td>
<td>Top-5</td>
<td>Top-1%</td>
<td>Hit Rate</td>
</tr>
<tr>
<td>No semi-positive (i.e. baseline, (\mathcal{L}_{triplet}))</td>
<td>38.0</td>
<td>62.9</td>
<td>97.6</td>
<td>41.8</td>
</tr>
<tr>
<td>Positive ((\mathcal{L}_{triplet}))</td>
<td>20.3</td>
<td>45.7</td>
<td>97.9</td>
<td>25.4</td>
</tr>
<tr>
<td>IOU ((\mathcal{L}<em>{triplet}+\mathcal{L}</em>{IOU}))</td>
<td><strong>41.1</strong></td>
<td><strong>65.9</strong></td>
<td><strong>98.3</strong></td>
<td><strong>44.8</strong></td>
</tr>
</tbody>
</table>
Beyond Retrieval

Offset prediction within the retrieved top-1 aerial image

$$L_{offset} = (lat - lat^*)^2 + (lon - lon^*)^2$$

$lat$ and $lon$ denote the predicted offset of the query GPS location relative to the reference GPS

$lat^*$ and $lon^*$ denote the ground-truth offset
Comparison with State-of-the-art

• Retrieval Performance

<table>
<thead>
<tr>
<th></th>
<th>Same-Area</th>
<th></th>
<th>Cross-Area</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1</td>
<td>Top-5</td>
<td>Top-1%</td>
<td>Hit Rate</td>
</tr>
<tr>
<td>Siamese-VGG ($\mathcal{L}_{triplet}$)</td>
<td>18.1</td>
<td>42.5</td>
<td>97.5</td>
<td>21.2</td>
</tr>
<tr>
<td>SAFA ($\mathcal{L}_{triplet}$)</td>
<td>33.9</td>
<td>58.4</td>
<td>98.2</td>
<td>36.9</td>
</tr>
<tr>
<td>SAFA+Mining (baseline, $\mathcal{L}_{triplet}$)</td>
<td>38.0</td>
<td>62.9</td>
<td>97.6</td>
<td>41.8</td>
</tr>
<tr>
<td>Ours ($\mathcal{L}_{hybrid}$)</td>
<td><strong>41.1</strong></td>
<td><strong>65.8</strong></td>
<td><strong>98.4</strong></td>
<td><strong>44.7</strong></td>
</tr>
</tbody>
</table>

$L_{hybrid} = L_{triplet} + L_{IOU} + L_{offset}$
Comparison with State-of-the-art

• Localization in terms of meters

![Graph 1: Same-area Localization](image1)

![Graph 2: Cross-area Localization](image2)
The Effect of Offset Prediction

- Localization in terms of meters

Figure 6. Same-area (left) and cross-area (right) meter-level localization accuracy of different offset prediction methods.
The Effect of Offset Prediction

Figure 8. Case study on meter-level refinement within the retrieved aerial image. Red square, green circle and blue diamond denote the final prediction with regression, ground-truth, and center (i.e. the prediction with only retrieval), respectively.
Noisy GPS Refinement

- Retrieval in a searching scope

<table>
<thead>
<tr>
<th>Search Scope</th>
<th>Same-Area</th>
<th></th>
<th>Cross-Area</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1</td>
<td>Top-5</td>
<td>Top-1</td>
<td>Top-5</td>
</tr>
<tr>
<td>All</td>
<td>41.1</td>
<td>65.8</td>
<td>11.0</td>
<td>23.6</td>
</tr>
<tr>
<td>1000 m</td>
<td>49.2</td>
<td>76.7</td>
<td>19.9</td>
<td>41.5</td>
</tr>
<tr>
<td>500 m</td>
<td>54.1</td>
<td>82.6</td>
<td>26.4</td>
<td>53.3</td>
</tr>
<tr>
<td>200 m</td>
<td>60.9</td>
<td>90.6</td>
<td>37.7</td>
<td>72.0</td>
</tr>
</tbody>
</table>
Summary

• We propose a new benchmark for cross-view image geo-localization beyond one-to-one retrieval, which is a more realistic setting for real-world applications.

• The proposed method significantly improves 10-meter-level accuracy:
  11.4% → 25.5% for same-area evaluation
  2.8% → 6.2% for cross-area evaluation

• Code and dataset are available at
  https://github.com/Jeff-Zilence/VIGOR
Outline

• Introduction (image geo-localization)

• Cross-view image geo-localization
  • Orientational alignment in image geo-localization
  • Spatial alignment in image geo-localization
  • Vision transformer for image geo-localization

• Future work
TransGeo: Transformer Is All You Need for Cross-view Image Geo-localization

Cross-view Image geo-localization

Domain Gap
## Predominant CNN-based Methods on CVUSA

<table>
<thead>
<tr>
<th>Method</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>R@1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workman [30]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>34.30</td>
</tr>
<tr>
<td>Zhai [34]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>43.20</td>
</tr>
<tr>
<td>CVM-Net [10]</td>
<td>22.47</td>
<td>49.98</td>
<td>63.18</td>
<td>93.62</td>
</tr>
<tr>
<td>Liu [14]</td>
<td>40.79</td>
<td>66.82</td>
<td>76.36</td>
<td>96.12</td>
</tr>
<tr>
<td>Reweight [3]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>98.30</td>
</tr>
<tr>
<td>Regmi [19]</td>
<td>48.75</td>
<td>-</td>
<td>81.27</td>
<td>95.98</td>
</tr>
<tr>
<td>Revisit [35]</td>
<td>70.40</td>
<td>-</td>
<td>-</td>
<td>99.10</td>
</tr>
<tr>
<td>SAFA [21]</td>
<td>81.15</td>
<td>94.23</td>
<td>96.85</td>
<td>99.49</td>
</tr>
<tr>
<td>†SAFA [21]</td>
<td>89.84</td>
<td>96.93</td>
<td>98.14</td>
<td>99.64</td>
</tr>
<tr>
<td>†Shi [22]</td>
<td>91.96</td>
<td>97.50</td>
<td>98.54</td>
<td>99.67</td>
</tr>
<tr>
<td>†Toker [26]</td>
<td>92.56</td>
<td>97.55</td>
<td>98.33</td>
<td>99.57</td>
</tr>
</tbody>
</table>

Suffer from Domain Gap

Polar Transform
Polar Transform works well on CVUSA dataset

Polar Transform doesn’t Work on VIGOR dataset

<table>
<thead>
<tr>
<th>VIGOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAFA [21]</td>
</tr>
<tr>
<td>SAFA+Polar</td>
</tr>
</tbody>
</table>
How to Bridge the Domain Gap?

- Polar Transform + CNN
- Vision transformer
Vision Transformer (ViT)

• Explicit positional information
• Global attention

TransGeo

Stage 1

Embedding

MLP Head

Street-view Transformer Encoder

Triplet Loss

Embedding

MLP Head

Aerial-view Transformer Encoder

Linear Projection

Linear Projection

Street-view image

Aerial-view image

Position Embedding

Class Token

Patch Embedding
Stage 2 - Attend and Zoom-in

Stage 1
- Embedding
- MLP Head
- Street-view Transformer Encoder
  - Linear Projection
  - Street-view image

Stage 2
- Embedding
- MLP Head
- Aerial-view Transformer Encoder
  - Linear Projection
  - Aerial-view image

Share Weights
- Triplet Loss

Attention
- Non-uniform Cropping
- Zoom In
  - (higher resolution aerial image)
Non-uniform Cropping

MLP Head
Aerial-view Transformer Encoder
Linear Projection

Attention
Resize and Binarize

Zoom-in
Non-uniform Cropping

Aerial-view image
## Retrieval Performance on VIGOR

<table>
<thead>
<tr>
<th></th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>R@1%</th>
<th>Hit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Siamese-VGG [35]</td>
<td>18.69</td>
<td>43.64</td>
<td>55.36</td>
<td>97.55</td>
<td>21.90</td>
</tr>
<tr>
<td>SAFA [21]</td>
<td>33.93</td>
<td>58.42</td>
<td>68.12</td>
<td>98.24</td>
<td>36.87</td>
</tr>
<tr>
<td>SAFA+Mining [36]</td>
<td>38.02</td>
<td>62.87</td>
<td>71.12</td>
<td>97.63</td>
<td>41.81</td>
</tr>
<tr>
<td>VIGOR [36]</td>
<td>41.07</td>
<td>65.81</td>
<td>74.05</td>
<td>98.37</td>
<td>44.71</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>61.48</strong></td>
<td><strong>87.54</strong></td>
<td><strong>91.88</strong></td>
<td><strong>99.56</strong></td>
<td><strong>73.09</strong></td>
</tr>
</tbody>
</table>
Meter-level Evaluation

[Graph showing accuracy as a function of threshold for various methods: Siamese-VGG, SAFA, SAFA+Mining, VIGOR w/o Offset, VIGOR, and Ours.]

Threshold (m) on the x-axis and Accuracy (%) on the y-axis.
Performance on CVUSA

- Query is exactly at the center of reference aerial image.

<table>
<thead>
<tr>
<th>Method</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>R@1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workman [30]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>34.30</td>
</tr>
<tr>
<td>Zhai [34]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>43.20</td>
</tr>
<tr>
<td>CVM-Net [10]</td>
<td>22.47</td>
<td>49.98</td>
<td>63.18</td>
<td>93.62</td>
</tr>
<tr>
<td>Liu [14]</td>
<td>40.79</td>
<td>66.82</td>
<td>76.36</td>
<td>96.12</td>
</tr>
<tr>
<td>Reweight [3]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>98.30</td>
</tr>
<tr>
<td>Regmi [19]</td>
<td>48.75</td>
<td>-</td>
<td>81.27</td>
<td>95.98</td>
</tr>
<tr>
<td>Revisit [35]</td>
<td>70.40</td>
<td>-</td>
<td>-</td>
<td>99.10</td>
</tr>
<tr>
<td>SAFA [21]</td>
<td>81.15</td>
<td>94.23</td>
<td>96.85</td>
<td>99.49</td>
</tr>
<tr>
<td>EgoTR [31] (arXiv)</td>
<td>91.99</td>
<td>97.68</td>
<td>98.65</td>
<td>99.75</td>
</tr>
<tr>
<td>†SAFA [21]</td>
<td>89.84</td>
<td>96.93</td>
<td>98.14</td>
<td>99.64</td>
</tr>
<tr>
<td>†Shi [22]</td>
<td>91.96</td>
<td>97.50</td>
<td>98.54</td>
<td>99.67</td>
</tr>
<tr>
<td>†Toker [26]</td>
<td>92.56</td>
<td>97.55</td>
<td>98.33</td>
<td>99.57</td>
</tr>
<tr>
<td>†EgoTR [31] (arXiv)</td>
<td>94.05</td>
<td>98.27</td>
<td>98.99</td>
<td>99.67</td>
</tr>
</tbody>
</table>

*Ours*  | 94.08| 98.36| 99.04| 99.77|

† means using Polar Transform
The First to Quantitatively Measure Efficiency

<table>
<thead>
<tr>
<th>Method</th>
<th>GFLOPs</th>
<th>GPU Memory</th>
<th>Inference Time per Batch</th>
<th>R@1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAFA</td>
<td>42.24</td>
<td>10.82 GB</td>
<td>111 ms</td>
<td>89.84</td>
</tr>
<tr>
<td>Ours</td>
<td>11.32</td>
<td>9.85 GB</td>
<td>99 ms</td>
<td>94.08</td>
</tr>
</tbody>
</table>
## Unknown Orientation

![Image of a cityscape with a north arrow]

### Table: Performance Metrics

<table>
<thead>
<tr>
<th></th>
<th>Same-Area</th>
<th></th>
<th></th>
<th>Cross-Area</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@5</td>
<td>R@10</td>
<td>R@1%</td>
</tr>
<tr>
<td>VIGOR [36]</td>
<td>19.10</td>
<td>42.13</td>
<td>-</td>
<td>95.12</td>
</tr>
<tr>
<td>TransGeo</td>
<td>47.69</td>
<td>79.77</td>
<td>86.36</td>
<td>99.29</td>
</tr>
</tbody>
</table>
Limited Field of View (FoV)

\[ FoV = 360^\circ \quad FoV = 180^\circ \quad FoV = 90^\circ \]

<table>
<thead>
<tr>
<th></th>
<th>( R@1 )</th>
<th>( R@5 )</th>
<th>( R@10 )</th>
<th>( R@1% )</th>
<th>( R@1 )</th>
<th>( R@5 )</th>
<th>( R@10 )</th>
<th>( R@1% )</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSM [22]</td>
<td>48.53</td>
<td>68.47</td>
<td>75.63</td>
<td>93.02</td>
<td>16.19</td>
<td>31.44</td>
<td>39.85</td>
<td>71.13</td>
</tr>
<tr>
<td>TransGeo</td>
<td>\textbf{58.22}</td>
<td>\textbf{81.33}</td>
<td>\textbf{87.66}</td>
<td>\textbf{98.13}</td>
<td>\textbf{30.12}</td>
<td>\textbf{54.18}</td>
<td>\textbf{63.96}</td>
<td>\textbf{89.18}</td>
</tr>
</tbody>
</table>

Qualitative Results - VIGOR

Street-view Query

Ground-truth

Retrieved Reference Images
Qualitative Results-CVUSA
Summary

• CNN-based methods highly rely on polar transform, but the proposed transformer-based method performs well w/o polar transform for all scenarios, due to the explicit positional embedding.

• Removing a large portion of patches from aerial view does not cause much performance drop, indicating high redundancy in this task. It could be leveraged to reduce computation or improve performance without additional cost.

• Code: https://github.com/Jeff-Zilence/TransGeo2022
Outline

• Introduction (image geo-localization)

• Cross-view image geo-localization
  • Orientational alignment in image geo-localization
  • Spatial alignment in image geo-localization
  • Vision transformer for image geo-localization

• Future work
Future Work – Same-view + Cross-view

- Geo-localization with Multi-view Reference
Geo-localization with Multi-view Reference

• Both street-view and aerial images exist in most cities.

• They are complimentary to each other:
  • Same-view has better performance but does not have full-coverage.
  • Cross-view is easy to collect reference but has low accuracy.

• Two separated fields can be combined.
Cross-view Video Geo-localization

• Query: ground video

Shruti Vyas, Chen Chen, Mubarak Shah. “GAMa: Cross-view Video Geo-localization”, European Conference on Computer Vision (ECCV), 2022 (a new dataset is collected)
References


Thank you