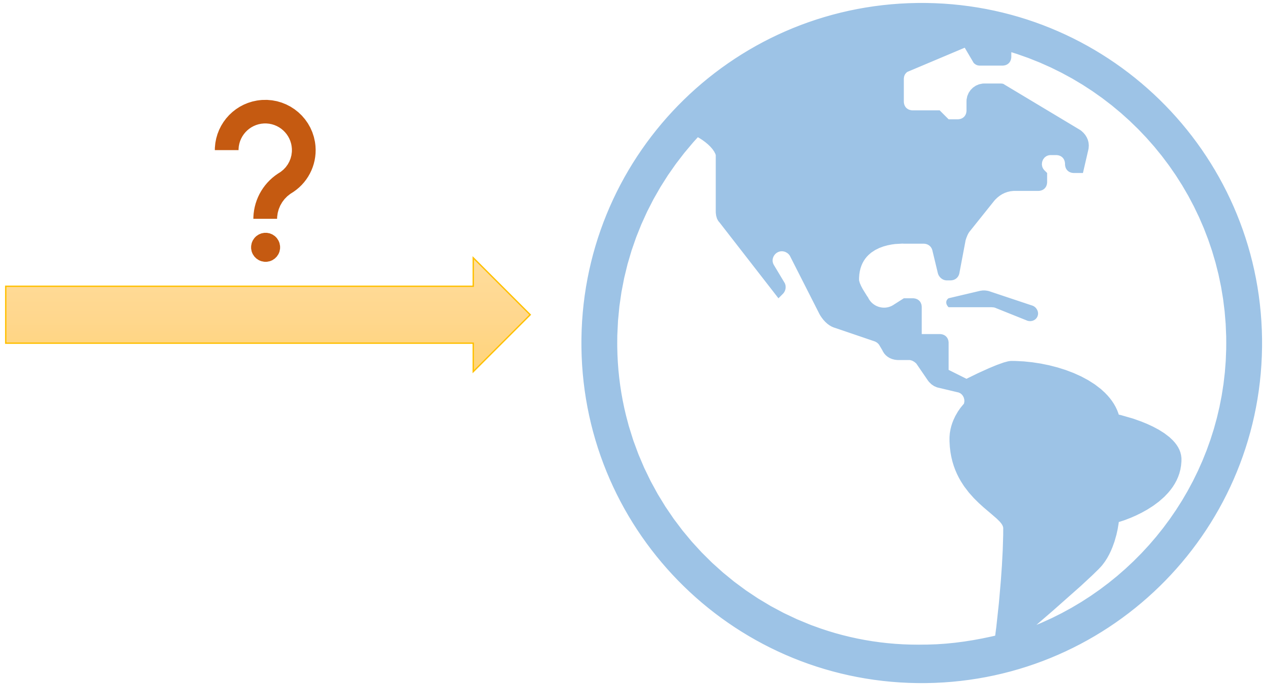
Cross-view Geo-localization via Learning Disentangled Geometric Layout Correspondence

Xiaohan Zhang[1,2]*, Xingyu Li[3]*, Waqas Sultani[4], Yi Zhou[5], Safwan Wshah[1,2]

[1]Department of Computer Science, University of Vermont
[2]Vermont Complex Systems Center, University of Vermont
[3]Shanghai Center for Brain Science and Brain-Inspired Technology
[4]Intelligent Machine Lab, Information Technology University
[5]NEL-BITA, School of Information Science and Technology, University of Science and Technology of China

Image geo-localization

Query Image



Cross-view image geo-localization

Query Image

Reference database

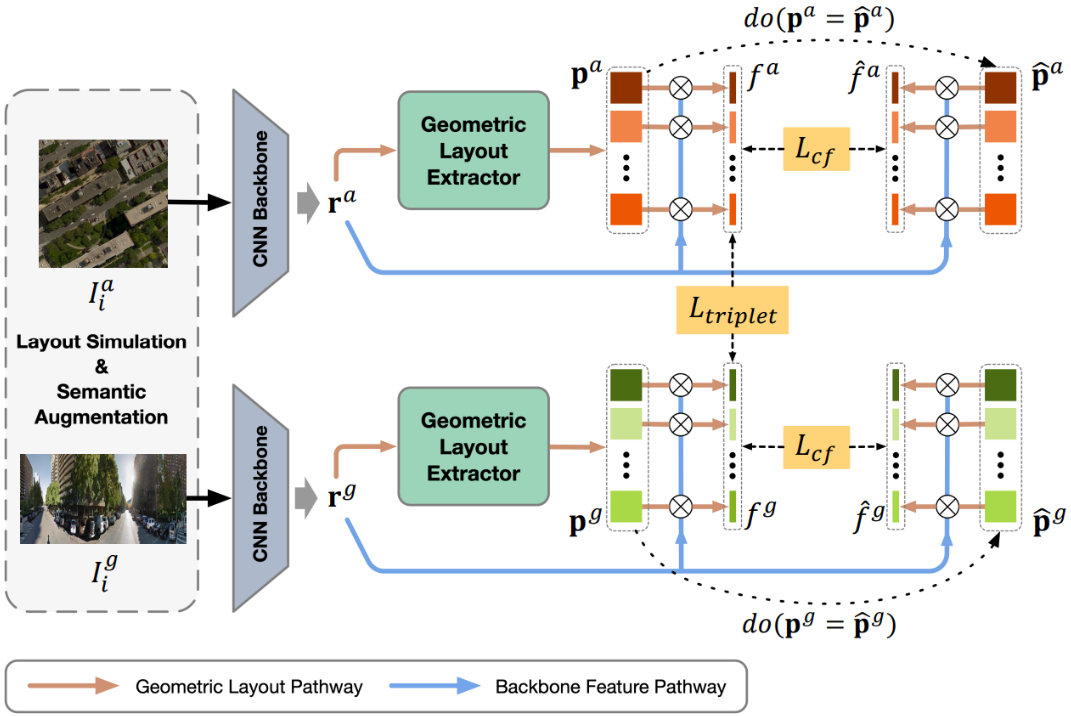
Challenges in cross-view image geo-localization

1. Existing CNN-based methods are limited by the nature of CNNs which explore the local correlation among pixels.

2. Recent researches explore applying transformers in cross-view geo-localization. However, they only implicitly model the spatial information.

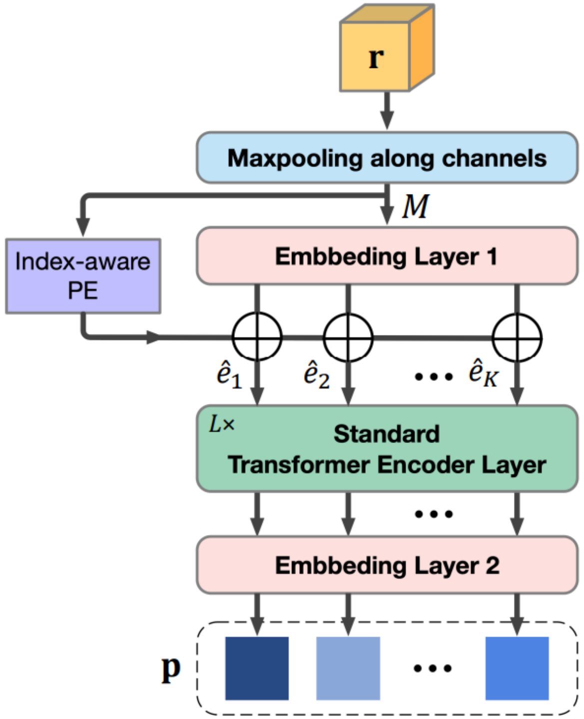
3. The performance of cross-view geo-localization methods degrades on cross-area benchmarks.

GeoDTR Overview



1. CNN backbones extract raw features $\mathrm{r}^{a(g)}$ from input images $I_i^{a(g)}$ augmented by Layout simulation and Semantic augmentation (LS).
2. $\mathrm{r}^{a(g)}$ are then passed to Geometric Layout Pathway to get layout descriptors $\mathrm{P}^{a(g)}$ and Backbone Feature Pathway to produce latent feature $f^{a(g)}$ by Frobenius product.
3. A Counterfactual learning paradigm is adopted to generate a counterfactual descriptors $\widehat{\mathrm{P}}^{a(g)}$.

Geometric layout extractor



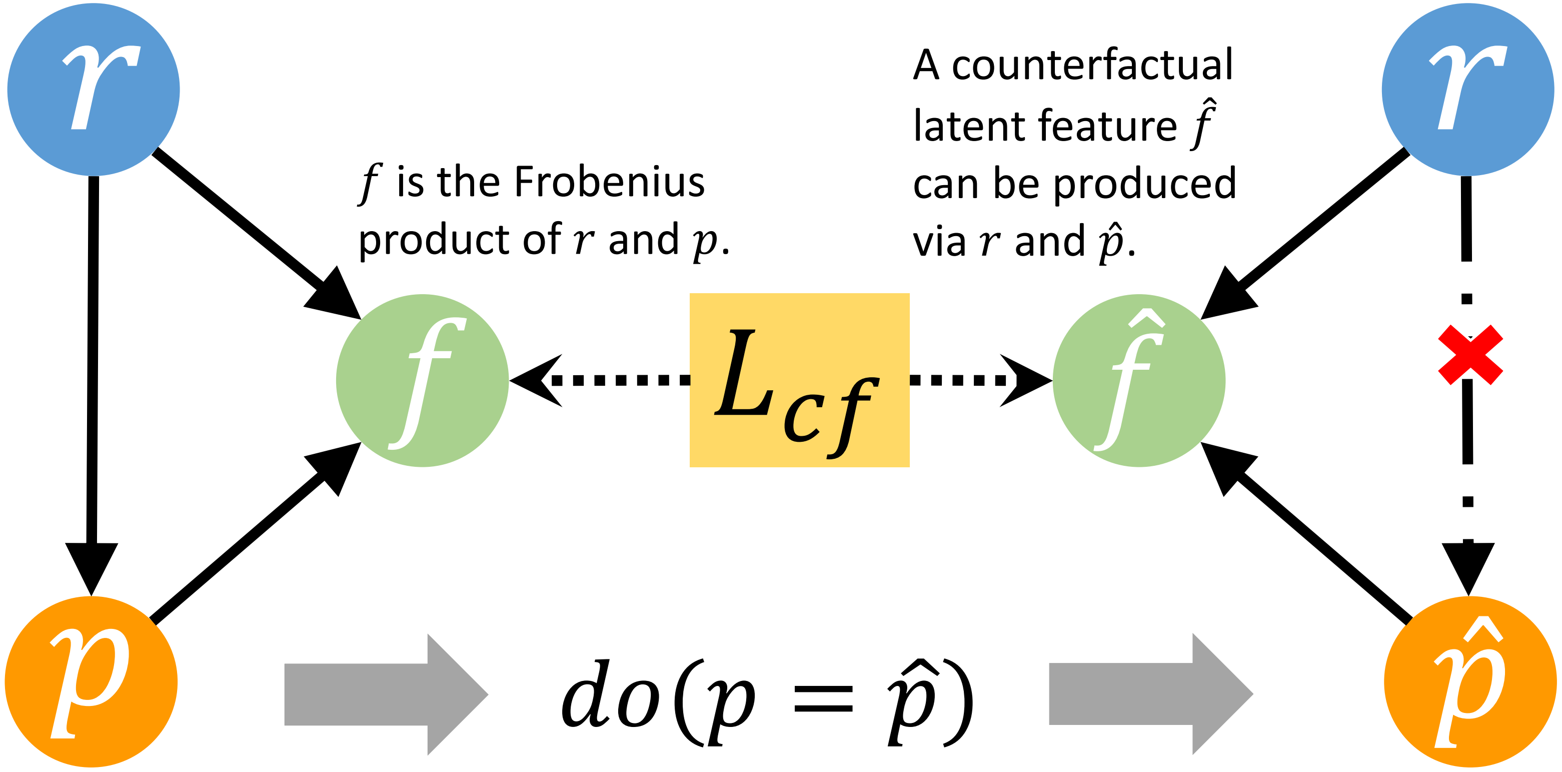
Geometric Layout Extractor takes raw feature r extracted by backbone as input.

A max pooling layer along channel is applied to obtain Saliency feature map M

An Embedding layer projects M into K subspaces. Then combined with index-aware position encoding and K embedding vectors to get $\mathrm{E} = [e_1, e_2, \dots e_K]$.

Finally, a transformer is applied to explore correlations in E. After the transformer, another embedding layer produces geometric layout descriptors P.

Counterfactual-based learning process



f is the Frobenius product of r and p.

A counterfactual latent feature $\hat{f}$ can be produced via r and $\hat{p}$.

A counterfactual loss is applied on f and $\hat{f}$ to maximize the distance as follow,

$$L_{cf} = \log(1 + e^{-\beta[|f,\hat{f}|_2]})$$

Since $\hat{p}$ is randomly sampled, there is no causal relation between r and $\hat{p}$.

p is obtained from r by geometric layout extractor

An intervention $do(p = \hat{p})$ is applied on p which replaces p into randomly sampled vectors $\hat{p}$.

Layout simulation

Layout simulation aims to generate aerial-ground pairs and satisfy the following requirement:

I. The generated aerial-ground pairs should **keep the correspondence**.

II. The generation process **varies the layouts** (e.g., the relative position of a house near a road) of the scene in the aerial-ground pairs.

III. The generation process must **maintain the low-level details**.

Layout simulation

Aerial image

Polar transformed aerial image

Ground image

Rotation

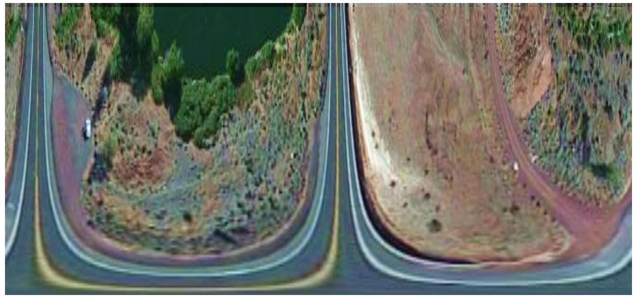
Flip

Semantic augmentation

Semantic augmentation modifies the low-level features in aerial and ground images

separately by randomly adjusting or applying:

- Brightness
- Contrast
- Saturation
- Gaussian blur
- Image grayscale
- Image posterizing



Training objectives

1. Counterfactual loss :

$$L_{cf}^{a(g)} = \log(1 + e^{-\beta[|f^{a(g)}, \hat{f}^{a(g)}|_2]})$$

2. Soft margin triplet loss :

$$L_{triplet} = \log(1 + e^{\alpha[|f_i^g, f_i^a|_2 - |f_i^g, f_j^a|_2]})$$

3. Total loss :

$$L = L_{triplet} + L_{cf}^{a(g)}$$

Implementation details

- A ResNet-34 is employed as backbone.
- α and β are set to 10 and 5 respectively.
- The model is trained on a single Nvidia V100 GPU for 200 epochs with an AdamW optimizer.
- The number of descriptor K is set to 8.
- Our code can is open-sourced at *https://gitlab.com/vail-uvm/geodtr*

Experiments Setup

CVUSA:

- 35,532 training pairs
- 8,884 testing pairs.

CVACT :

- 35,532 training pairs
- 8,884 validation pairs (CVACT_val).
- 92,802 testing pairs (CVACT_test).

Evaluation Metrics:

Similar to existing methods, we choose to use recall accuracy at top K ($R@K$) for evaluation purposes. In the following experiments, we adopt $R@1$, $R@5$, $R@10$, and $R@1\%$.

Experiment – CVUSA same-area

Method	R@1	R@5	R@10	R@1%
FusionGAN	48.75%	-	81.27%	95.98%
CVFT	61.43%	84.69%	90.49%	99.02%
SAFA	81.15%	94.23%	96.85%	99.49%
SAFA†	89.84%	96.93%	98.14%	99.64%
DSM†	91.93%	97.50%	98.54%	99.67%
CDE†	92.56%	97.55%	98.33%	99.57%
L2LTR	91.99%	97.68%	98.65%	99.75%
L2LTR†	94.05%	98.27%	98.99%	99.67%
TransGeo	94.08%	98.36%	99.04%	99.77%
SEH†	**95.11%**	98.45%	99.00%	99.78%
Ours w/ LS	93.76%	**98.47%**	**99.22%**	**99.85%**
Ours w/ LS†	**95.43%**	**98.86%**	**99.34%**	**99.86%**

Experiment – CVACT same-area

Method	CVACT_val				CVACT_test			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
CVFT	61.05%	81.33%	86.52%	95.93%	26.12%	45.33%	53.80%	71.69%
SAFA	78.28%	91.60%	93.79%	98.15%	-	-	-	-
SAFA†	81.03%	92.80%	94.84%	98.17%	55.50%	79.94%	85.08%	94.49%
DSM†	82.49%	92.44%	93.99%	97.32%	35.63%	60.07%	69.10%	84.75%
CDE†	83.28%	93.57%	95.42%	98.22%	61.29%	85.13%	89.14%	98.32%
L2LTR	83.14%	93.84%	95.51%	**98.40%**	58.33%	84.23%	88.60%	95.83%
L2LTR†	84.89%	94.59%	95.96%	98.37%	60.72%	85.85%	89.88%	96.12%
TransGeo	84.95%	94.14%	95.78%	98.37%	-	-	-	-
SEH†	84.75%	93.97%	95.46%	98.11%	-	-	-	-
Ours w/ LS	**85.43%**	**94.81%**	**96.11%**	98.26%	**62.96%**	**87.35%**	**90.70%**	**98.61%**
Ours w/ LS†	**86.21%**	**95.44%**	**96.72%**	**98.77%**	**64.52%**	**88.59%**	**91.96%**	**98.74%**

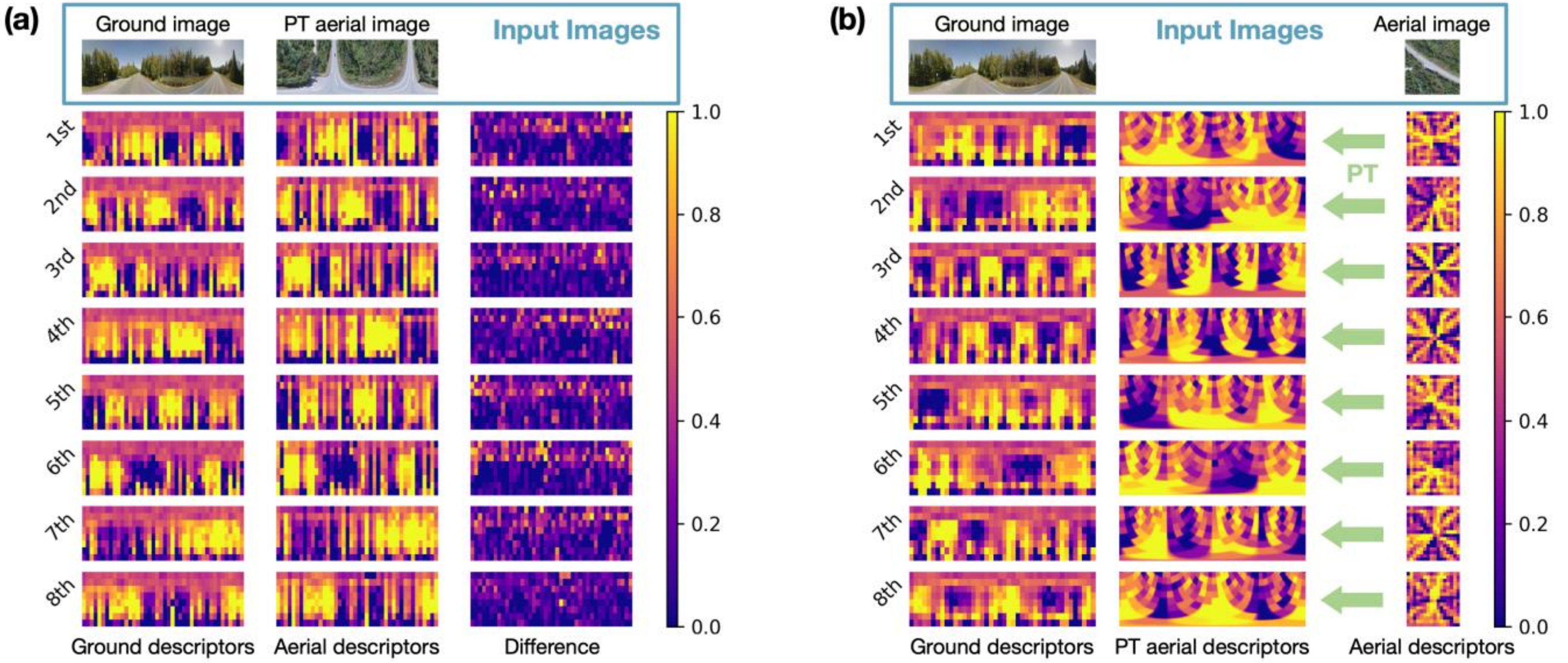
Experiment – Cross-area

Model	Task	R@1	R@5	R@10	R@1%
SAFA†	CVUSA → CVACT	30.40%	52.93%	62.29%	85.82%
DSM†		33.66%	52.17%	59.74%	79.67%
L2LTR†		**47.55%**	**70.58%**	**77.39%**	91.39%
TransGeo		37.81%	61.57%	69.86%	89.14%
Ours w/ LS		43.72%	66.99%	74.61%	**91.83%**
Ours w/ LS†		**53.16%**	**75.62%**	**81.90%**	**93.80%**
SAFA‡	CVACT → CVUSA	21.45%	36.55%	43.79%	69.83%
DSM†		18.47%	34.46%	42.28%	69.01%
L2LTR†		**33.00%**	**51.87%**	**60.63%**	**84.79%**
TransGeo		18.99%	38.24%	46.91%	88.94%
Ours w/ LS		29.85%	49.25%	57.11%	82.47%
Ours w/ LS†		**44.07%**	**64.66%**	**72.08%**	**90.09%**

Experiment – LS on other methods

LS + other methods		Same-area				Cross-area			
	Configuration	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
Trained on CVUSA	SAFA	89.84%	96.93%	98.14%	99.64%	30.40%	52.93%	62.29%	85.82%
	SAFA w/ LS	88.19%	96.48%	98.20%	99.74%	37.15%	60.31%	69.20%	89.15%
	L2LTR	94.05%	98.27%	98.99%	99.67%	47.55%	70.58%	77.52%	91.39%
	L2LTR w/ LS	93.62%	98.46%	99.03%	99.77%	52.58%	**75.81%**	**82.19%**	93.51%
	GeoDTR w/o LS	95.23%	98.71%	99.26%	99.79%	47.79%	70.52%	77.52%	92.20%
	GeoDTR w/ LS	**95.43%**	**98.86%**	**99.34%**	**99.86%**	**53.16%**	75.62%	81.90%	**93.80%**
Trained on CVACT	SAFA	81.03%	92.80%	94.84%	98.17%	21.45%	36.55%	43.79%	69.83%
	SAFA w/ LS	79.88%	92.84%	94.71%	97.96%	25.42%	42.30%	50.36%	76.49%
	L2LTR	84.89%	94.59%	95.96%	98.37%	33.00%	51.87%	60.63%	84.79%
	L2LTR w/ LS	83.49%	94.93%	96.44%	98.68%	37.69%	57.78%	66.22%	89.63%
	GeoDTR w/o LS	**87.42%**	95.37%	96.50%	98.65%	29.13%	47.86%	56.21%	81.09%
	GeoDTR w/ LS	86.21%	**95.44%**	**96.72%**	**98.77%**	**44.07%**	**64.66%**	**72.08%**	**90.09%**

Learned descriptors visualization



Summary

1. In this paper, we propose **GeoDTR** which disentangles geometric information from raw features to better captures the correspondence between aerial and ground images.
2. We propose **layout simulation and semantic augmentation (LS)** techniques that improve the performance of GeoDTR (as well as existing models) on cross-area experiments.
3. We introduce a novel **counterfactual-based learning schema** that guides GeoDTR to better grasp the spatial configurations and therefore produce better latent feature representations.

Related links

Contact Email: Xiaohan.Zhang@uvm.edu

Homepage: https://zxh009123.github.io/

Gitlab link: https://gitlab.com/vail-uvm/geodtr

arXiv link: https://arxiv.org/abs/2212.04074

Thanks for watching !