Learning Disentangled Geometric Layout Correspondence for Cross-View Geo-localization

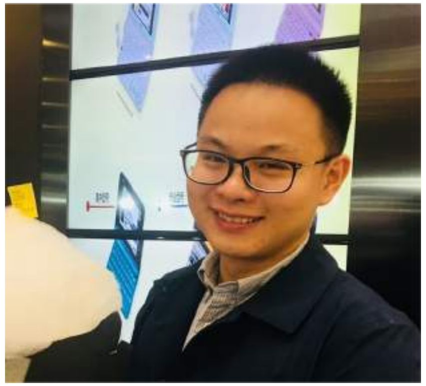
Waqas Sultani
Information Technology University, Lahore

Our team

Xiaohan Zhang
Vermont Complex Systems Center
University of Vermont

Safwan Wshah
Vermont Complex Systems Center
University of Vermont

University of Vermont, USA



Xingyu Li
Shanghai Center for Brain Science and Brain-Inspired Technology

Waqas Sultani
Intelligent Machine Lab
Information Technology University

Information Technology University, Pakistan

Image geo-localization

Query Image

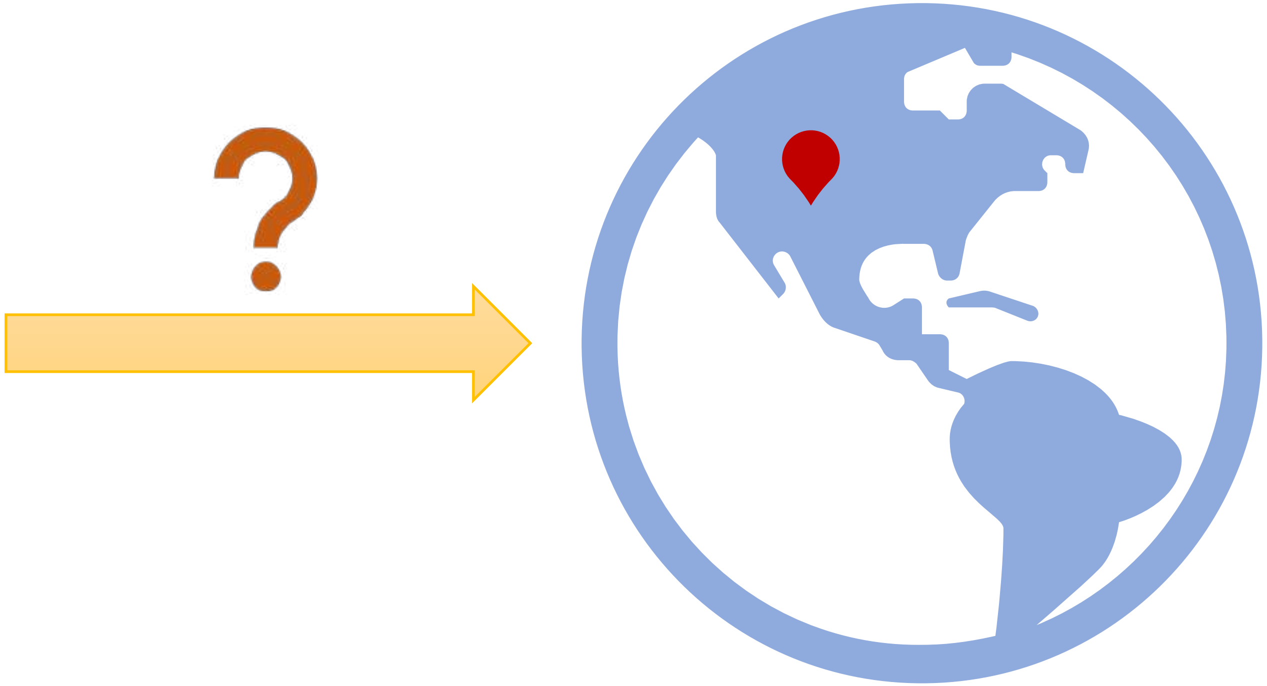


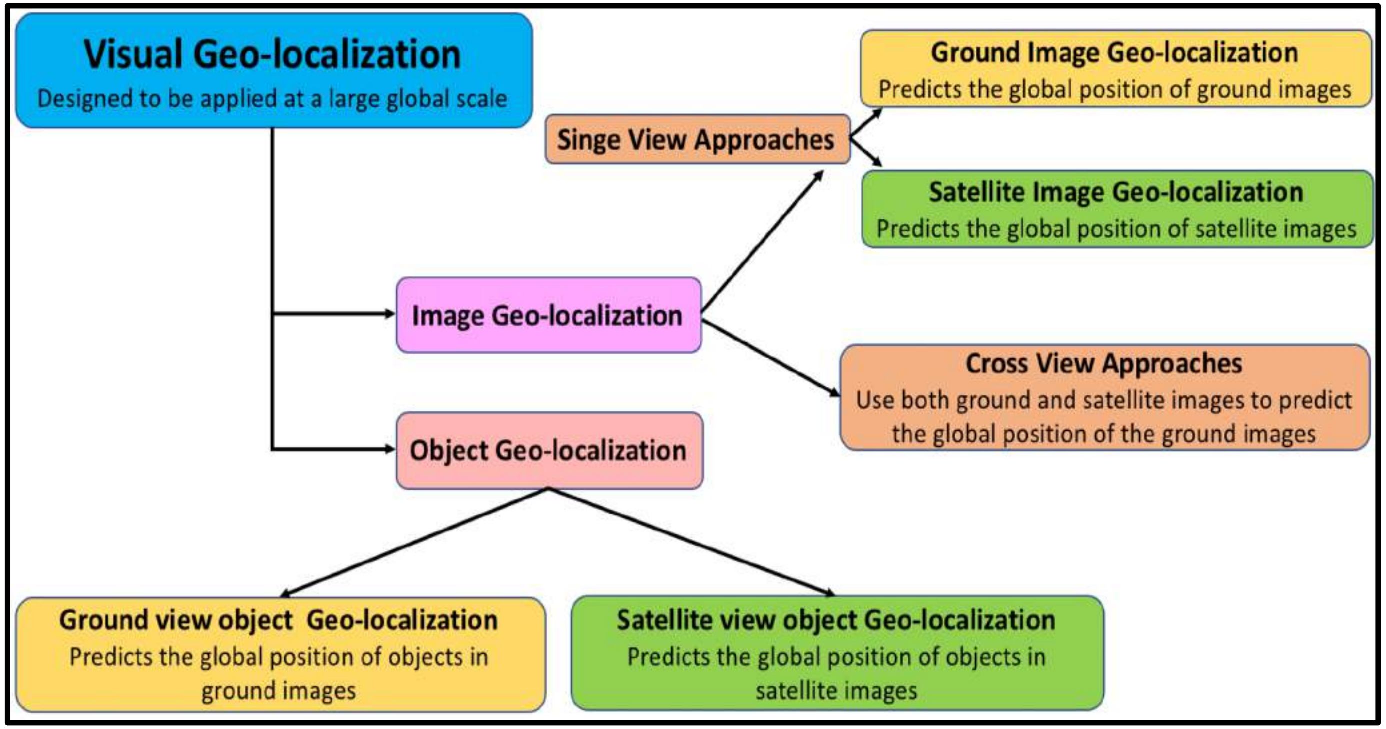
Image and Object Geo-localization

The University of Vermont | ITU Information Technology University

- **Visual Geo-localization**
  Designed to be applied at a large global scale
  - **Image Geo-localization**
    - **Singe View Approaches**
      - **Ground Image Geo-localization**
        Predicts the global position of ground images
      - **Satellite Image Geo-localization**
        Predicts the global position of satellite images
    - **Cross View Approaches**
      Use both ground and satellite images to predict the global position of the ground images
  - **Object Geo-localization**
    - **Ground view object Geo-localization**
      Predicts the global position of objects in ground images
    - **Satellite view object Geo-localization**
      Predicts the global position of objects in satellite images

Daniel Wilson, Xiaohan Zhang, **Waqas Sultani**, Safwan Wshah, "**Visual and Object Geo-localization: A Comprehensive Survey**", arXiv preprint arXiv:2112.15202

Image and Object Geo-localization



Daniel Wilson, Xiaohan Zhang, **Waqas Sultani**, Safwan Wshah, "**Visual and Object Geo-localization: A Comprehensive Survey**", arXiv preprint arXiv:2112.15202

Visual localization

Designed to be applied at a small scale

Object Localization

Predicts position of the object within the image

Camera Pose Estimation

Predicts pose of camera locally

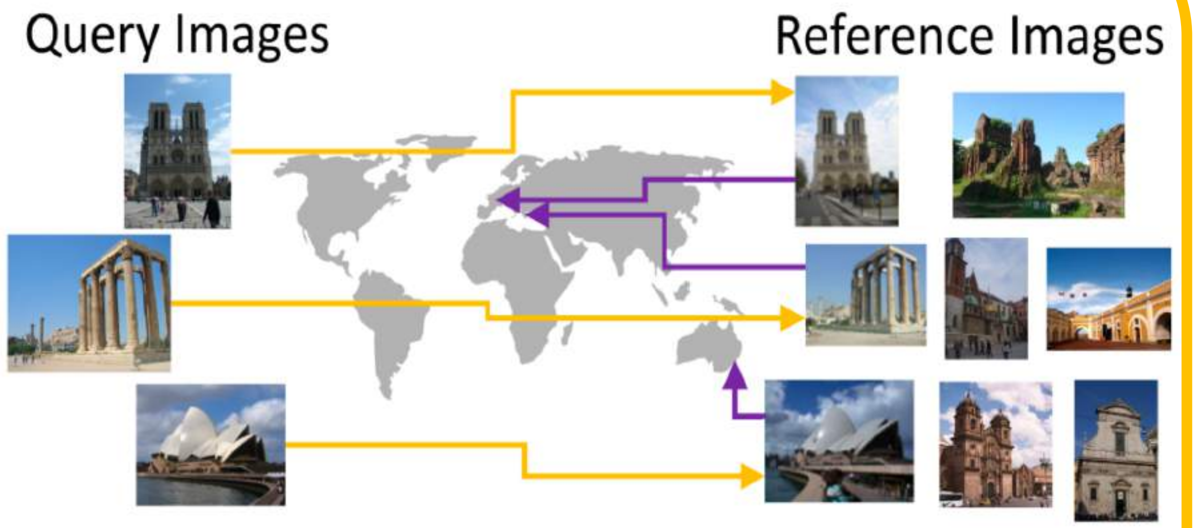
SLAM

Constructs a map of the surrounding environment, typically in real time

Image Geo-localization

Ground-view Geo-localization

Cross-view Geo-localization



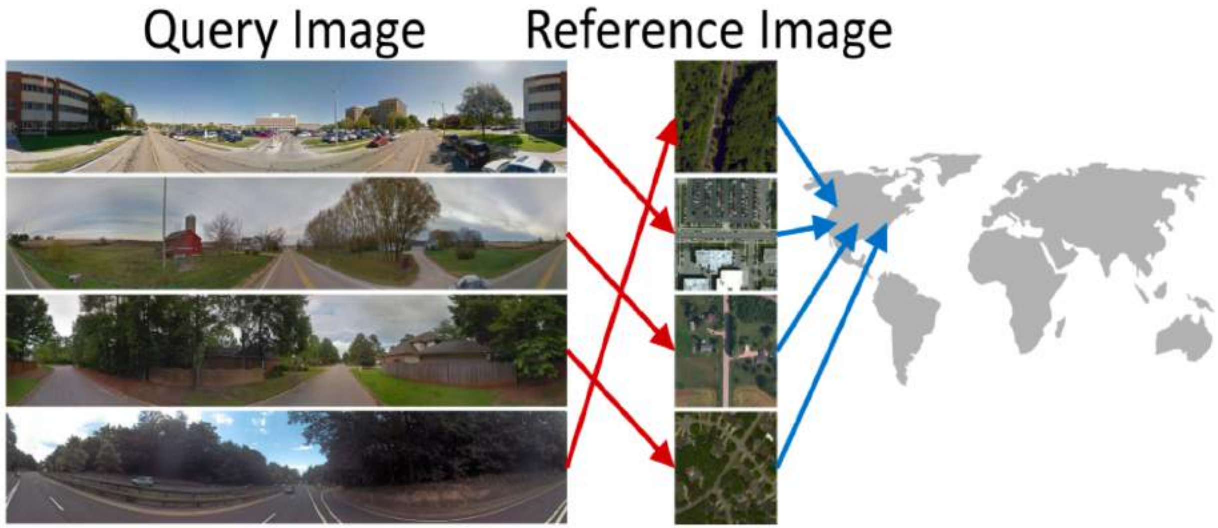
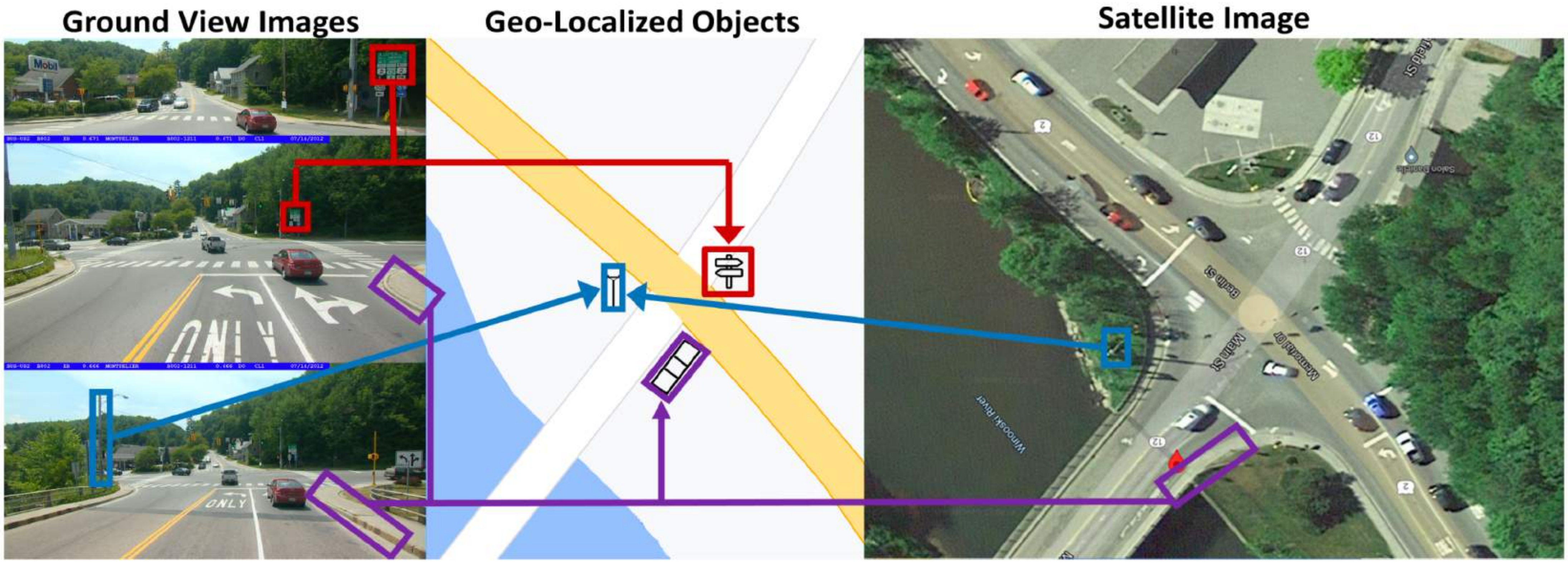


Image Geo-localization

Object Geo-localization



Cross-view image geo-localization

Query Image

Reference database

Challenges in cross-view geo-localization:

- Drastic view changes
- Different capturing time
- Different object resolution

Cross-view image geo-localization

Query Image

Reference database

Challenges in cross-view geo-localization:

- Drastic view changes
- Different capturing time
- Different object resolution

Limitations

- The performance of cross-view geo-localization methods **degrades on cross-area benchmarks**.
- Lack of ability to extract the spatial configuration of visual feature layout.
- Models overfit the low-level details from the training set.

Key Idea

- ✓ **Explicitly disentangle geometric information from the raw features**
- ✓ **Learn the spatial correlations among visual features from aerial and ground pairs**

Xiaohan Zhang, Xingyu Li, **Waqas Sultani**, Yi Zhou, Safwan Wshah, “**Learning Disentangled Geometric Layout Correspondence for Cross-View Geo-localization”,** AAAI 2023 (Oral).

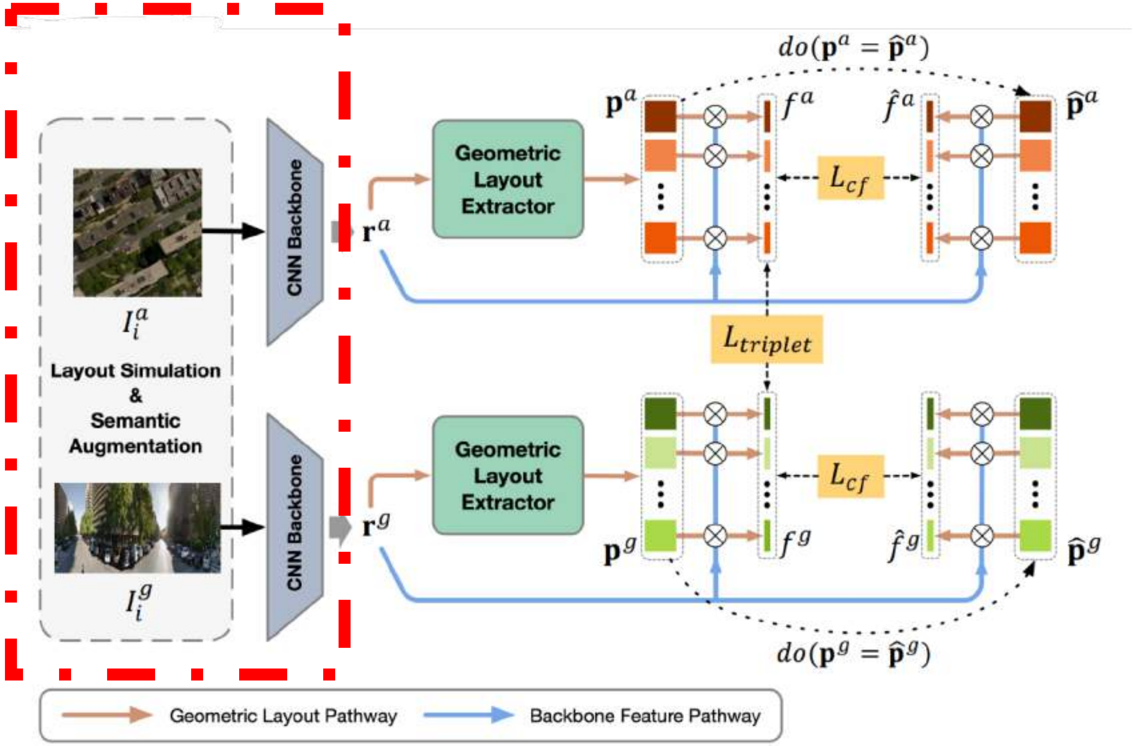
Overview

- ✓ GeoDTR module generates a set of geometric layout descriptors which produce a high quality latent representations.

- ✓ Analysis the effect of data augmentation for improved cross-area cross-view geo-localization performance.

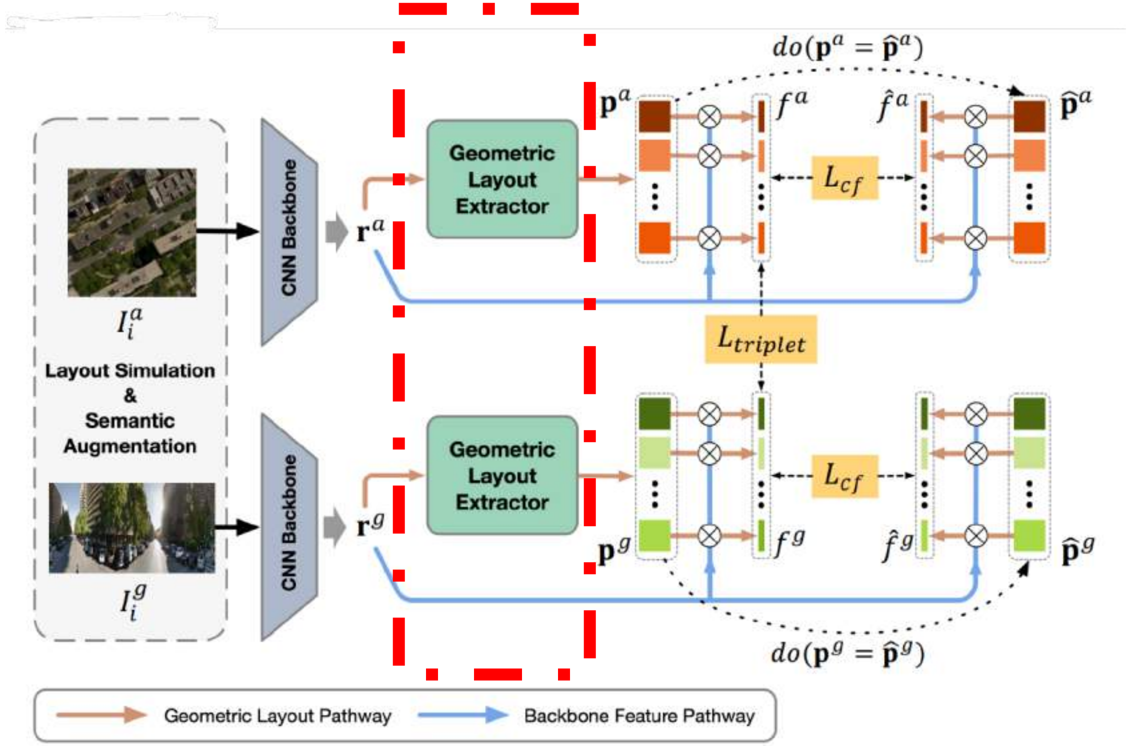
- ✓ To help geometric layout descriptor in exploring spatial information, we propose to employ counterfactual-based learning process.

GeoDTR Overview



1. CNN backbones extract raw features $r^{a(g)}$ from input images $I_i^{a(g)}$ augmented by Layout simulation and Semantic augmentation (LS).

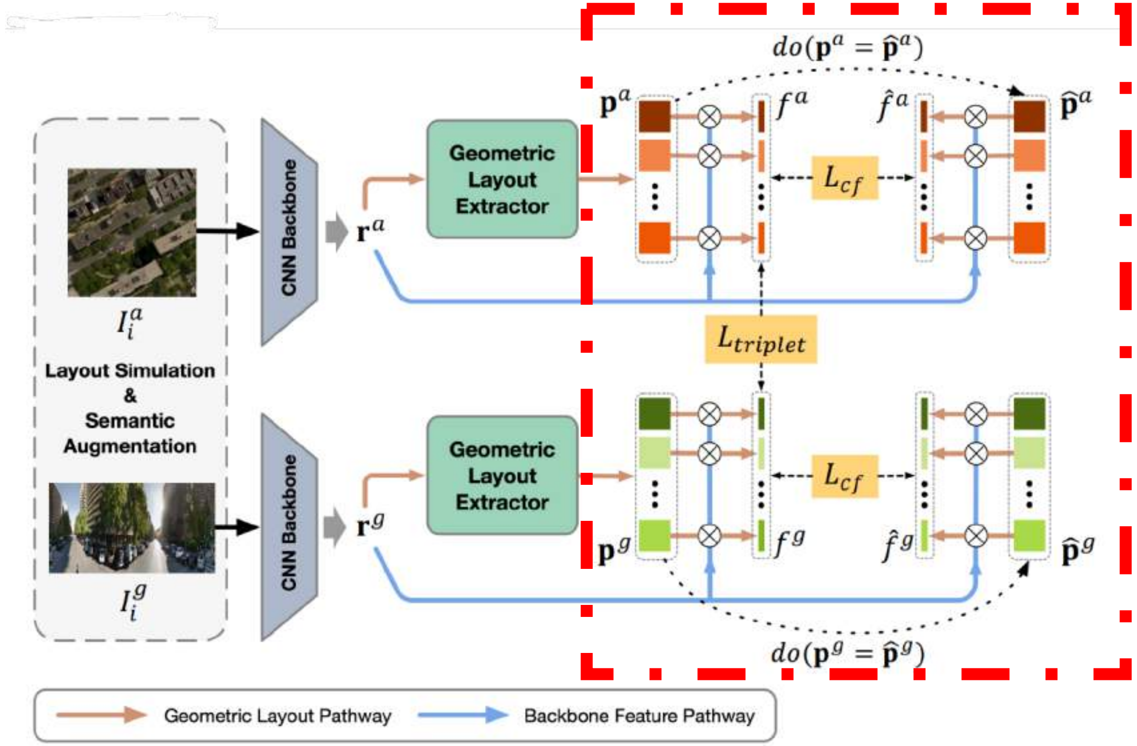
GeoDTR Overview



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.

GeoDTR Overview

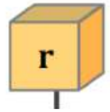


1. CNN backbones extract raw features $\mathrm{r}^{a(g)}$ from input images $I_i^{a(g)}$ augmented by Layout simulation and Semantic augmentation (LS).

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.

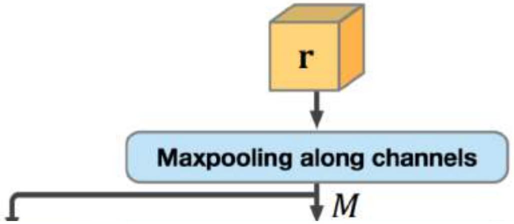
3. A Counterfactual learning paradigm is adopted to generate a counterfactual descriptors $\widehat{\mathrm{P}}^{a(g)}$.

Geometric layout extractor



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

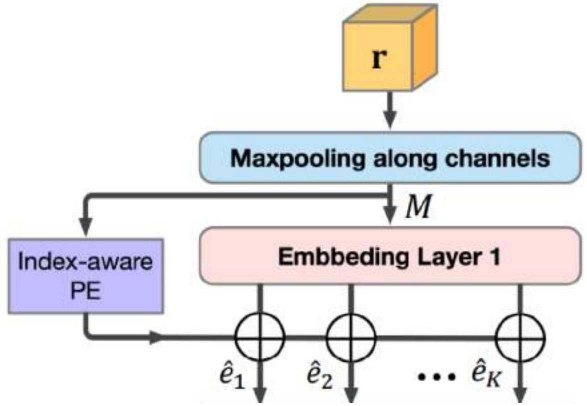
Geometric layout extractor



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

Geometric layout extractor

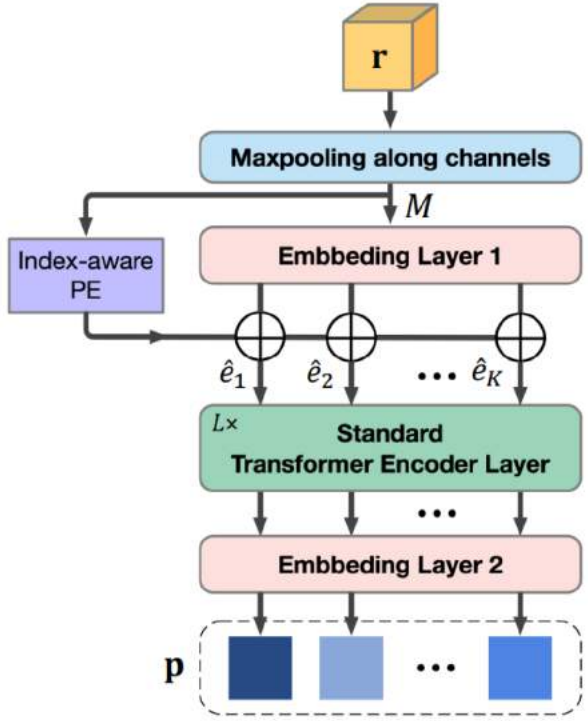


Geometric Layout Extractor takes raw feature r extracted by backbone as input.

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 $\mathrm{E} = [e_1, e_2, \dots e_K]$.

Geometric layout extractor



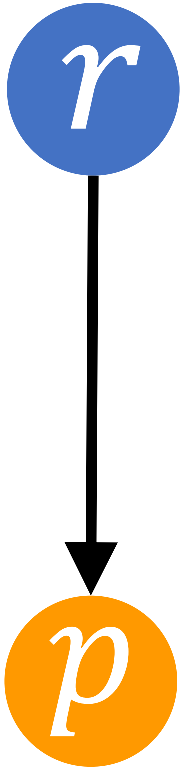
Geometric Layout Extractor takes raw feature r extracted by backbone as input.

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 $\mathrm{E} = [e_1, e_2, \dots e_K]$.

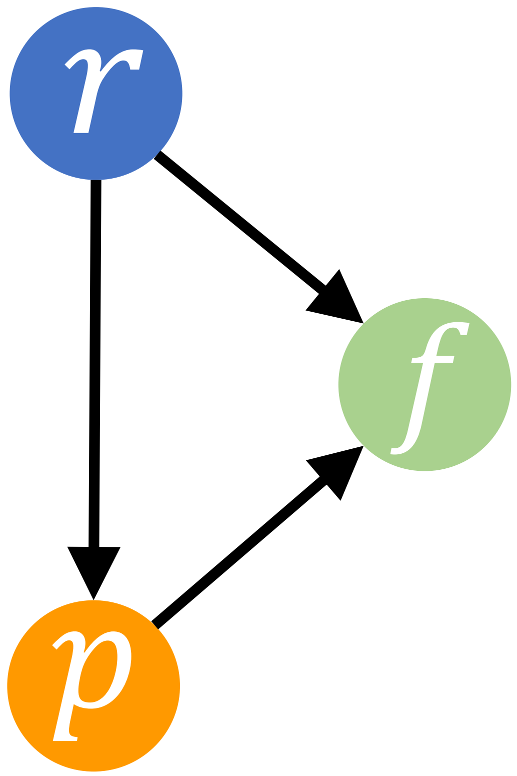
Finally, a transformer is applied to explore correlations in E. After the transformer, another embedding layer produces geometric layout descriptors P.

Counterfactual-based learning process



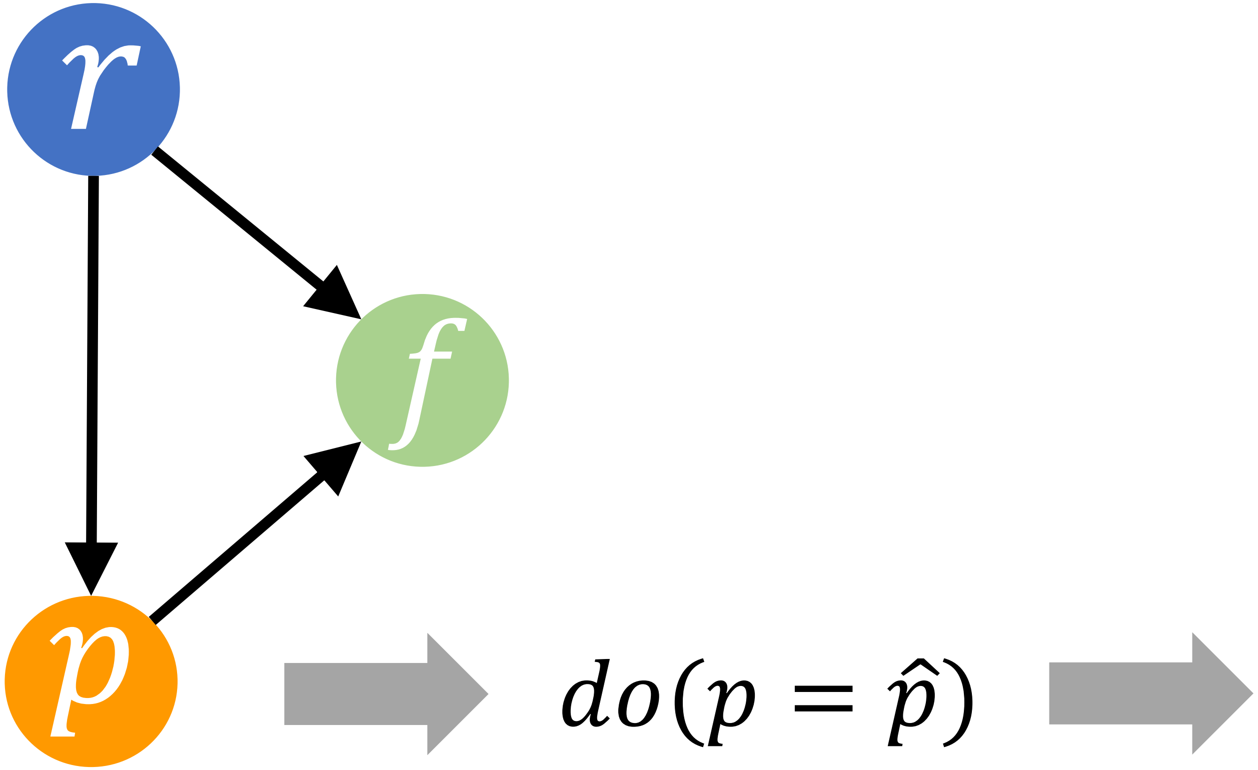
$\boldsymbol{p}$ is obtained from $\boldsymbol{r}$ by geometric layout extractor

Counterfactual-based learning process



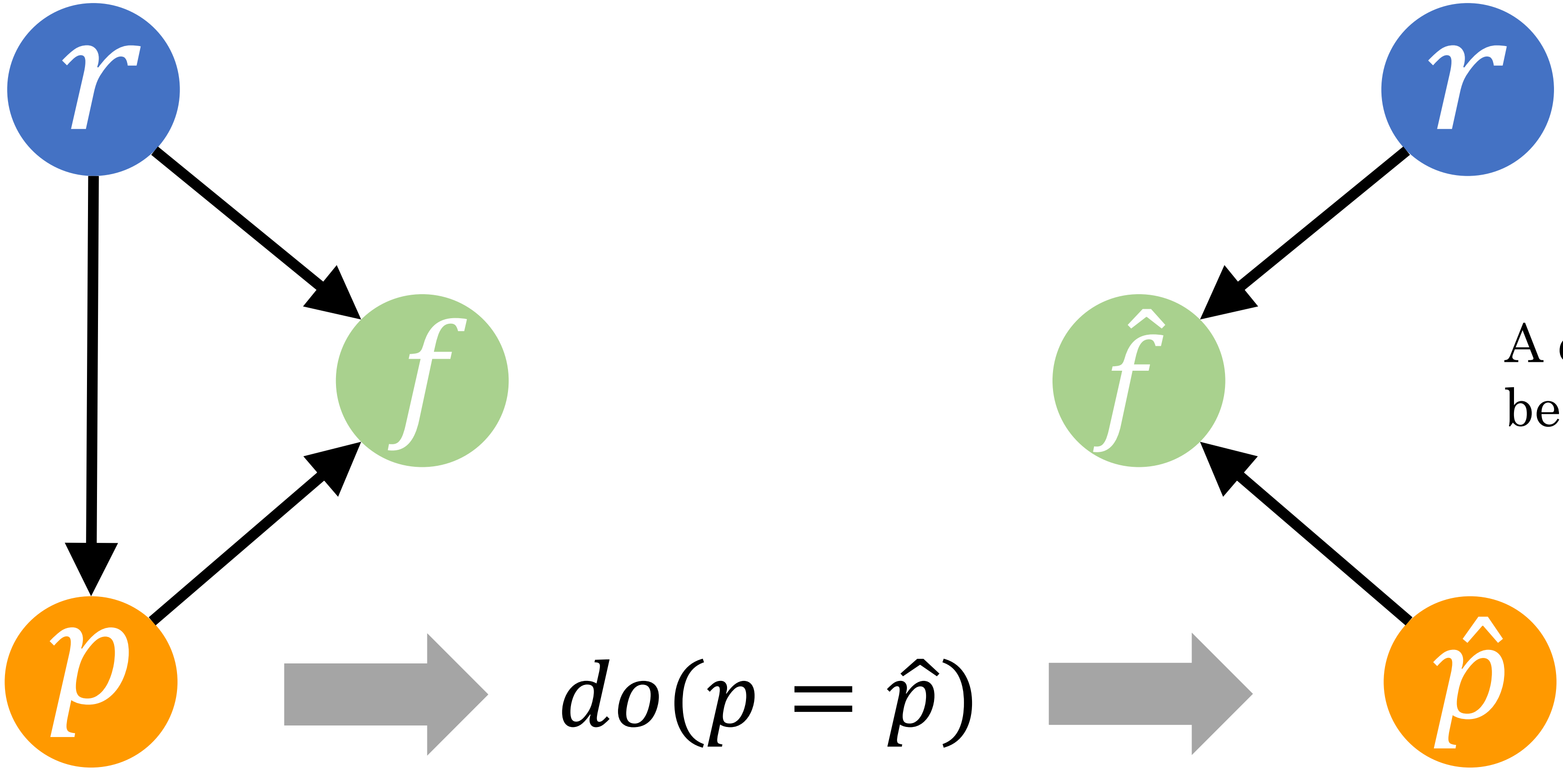
$\boldsymbol{f}$ is the Frobenius product of $\boldsymbol{r}$ and $\boldsymbol{p}$.

Counterfactual-based learning process



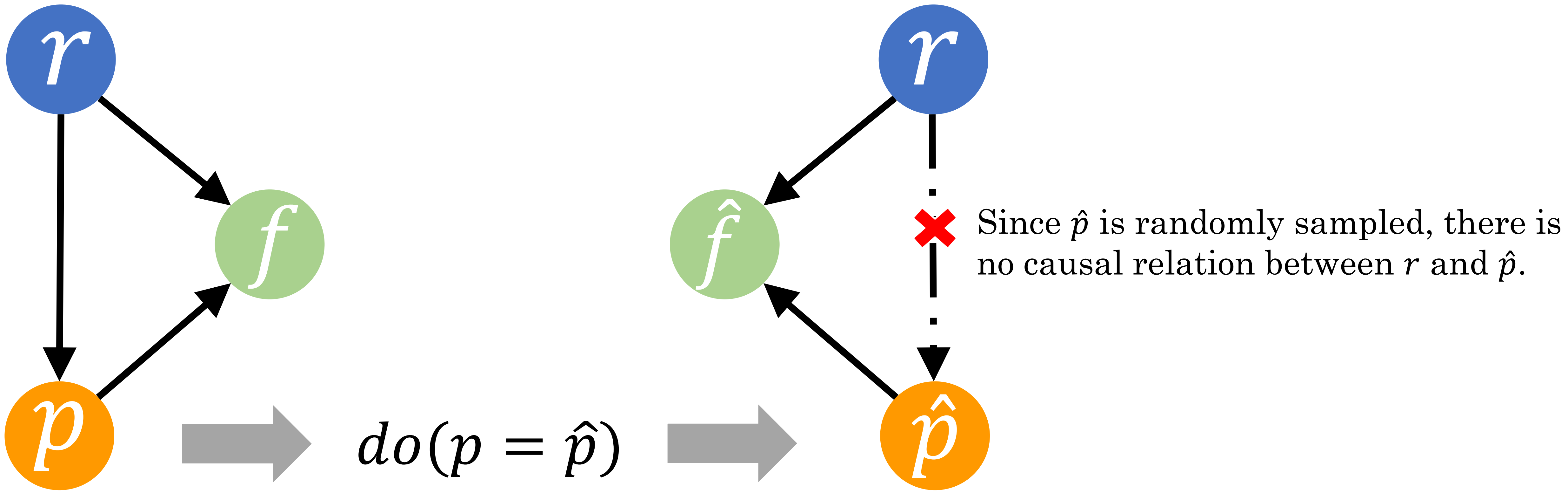
An intervention $\boldsymbol{do}(\boldsymbol{p} = \widehat{\boldsymbol{p}})$ is applied on $\boldsymbol{p}$ which replaces $\boldsymbol{p}$ into randomly sampled vectors $\widehat{\boldsymbol{p}}$.

Counterfactual-based learning process

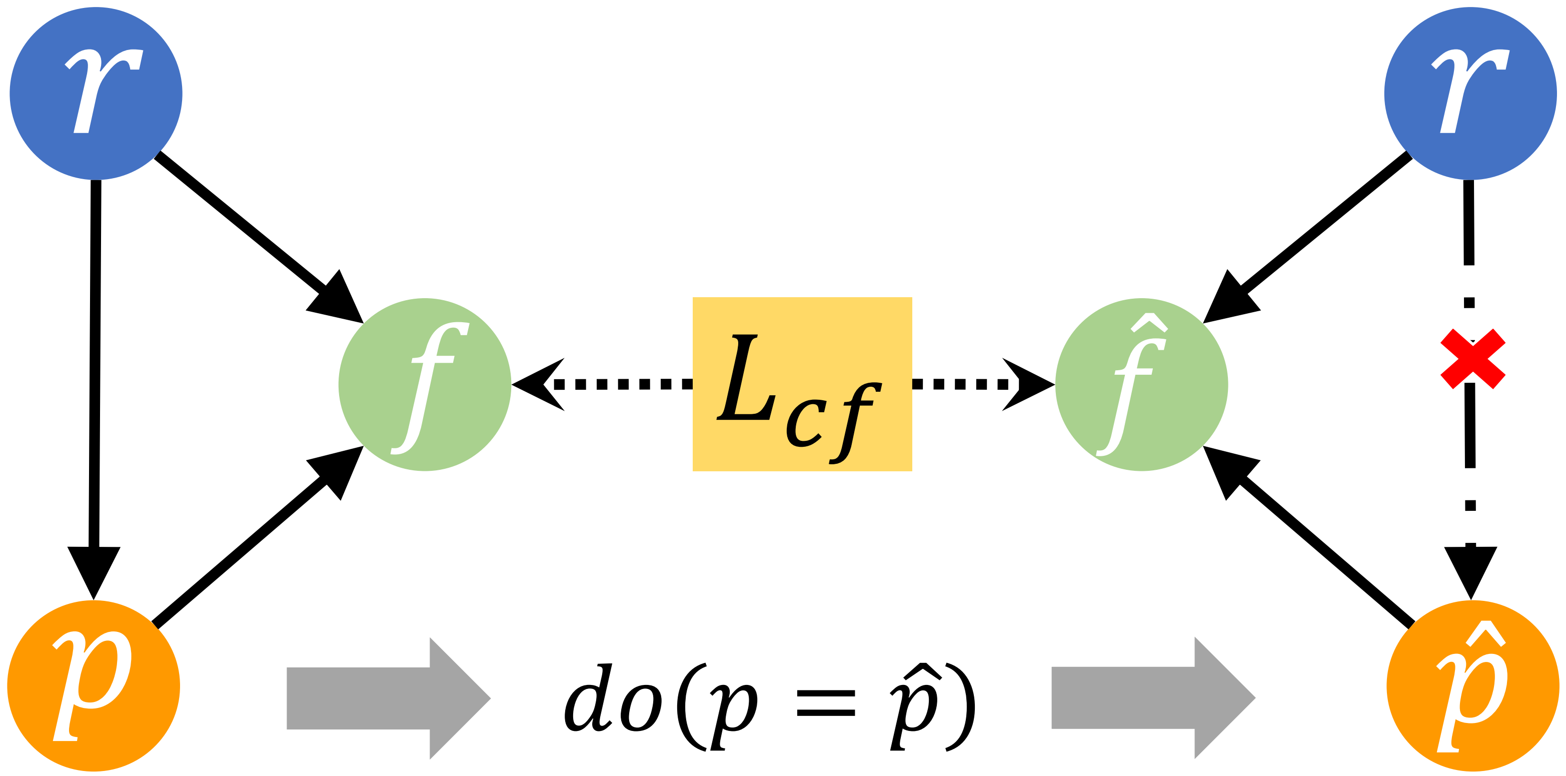


A counterfactual latent feature $\hat{f}$ can be produced via r and $\hat{p}$.

Counterfactual-based learning process



Counterfactual-based learning process



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

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

Data augmentation

- Usually break the correspondence between aerial-ground pairs and incapable to provide diverse layout.

Aerial image

Ground image

Data augmentation

- Usually break the correspondence between aerial-ground pairs and incapable to provide diverse layout.

- No sufficient attention on the low-level details.

Aerial image

Ground image (Cropped)

Data augmentation

- ✓ Layout simulation
- ✓ Semantic augmentation

LS techniques

Layout simulation

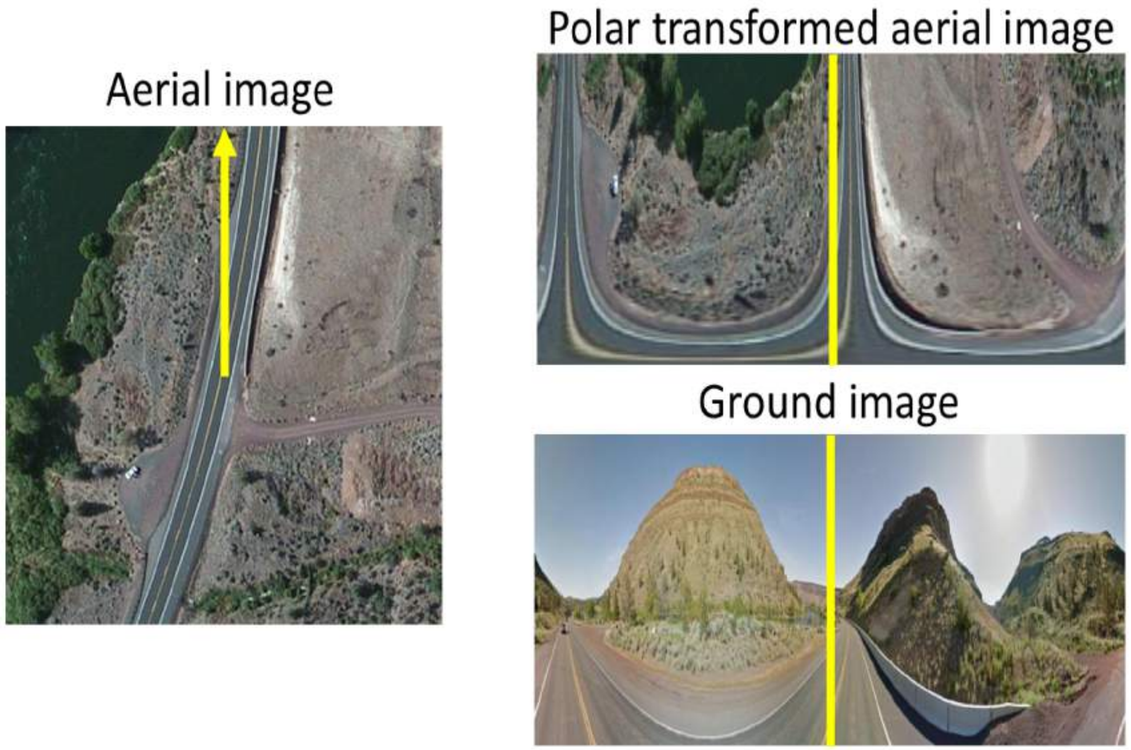
The University of Vermont | ITU INFORMATION TECHNOLOGY UNIVERSITY

- Layout simulation aims to generate aerial-ground pairs **with unseen layouts** by using geometric transformations that satisfy the following requirements:

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

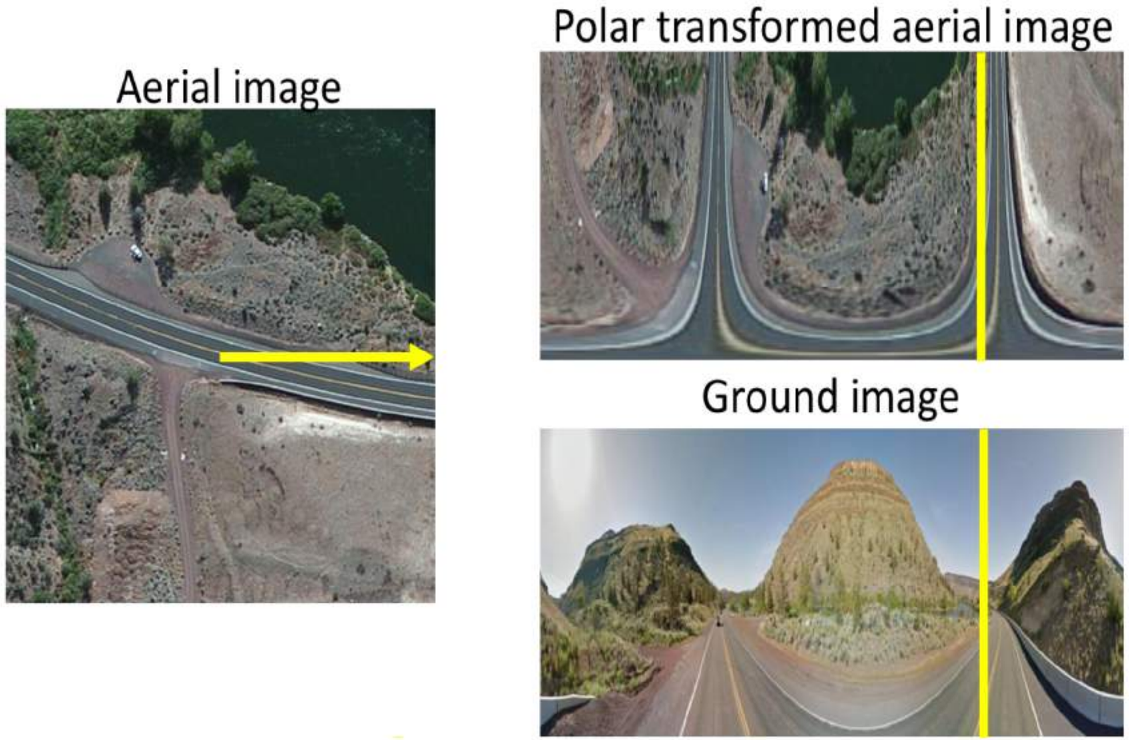
II. The generation process must **maintain the low-level details**.

Layout simulation



Rotation

Layout simulation



Rotation

Layout simulation

Aerial image

Polar transformed aerial image

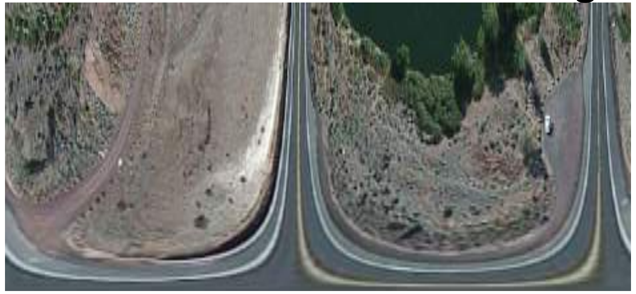
Ground image

FLIP

Layout simulation

Aerial image

Polar transformed aerial image

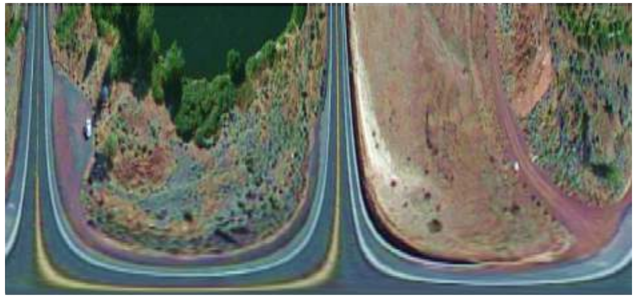


Ground image

FLIP

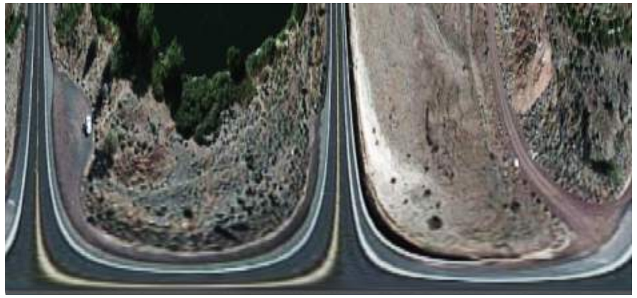
Semantic augmentation

- 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



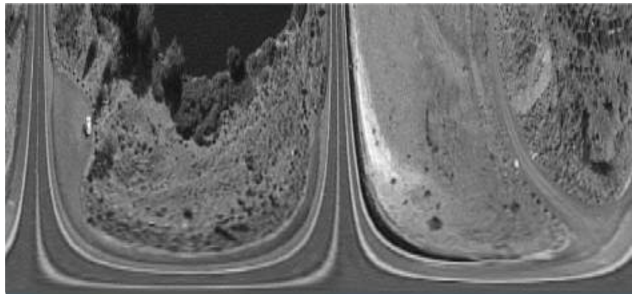
Semantic augmentation

- 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



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 descriptors 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.

We use $R@1$, $R@5$, $R@10$, and $R@1\%$.

Experiment – CVUSA same-area

The University of Vermont

INFORMATION TECHNOLOGY UNIVERSITY

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%**

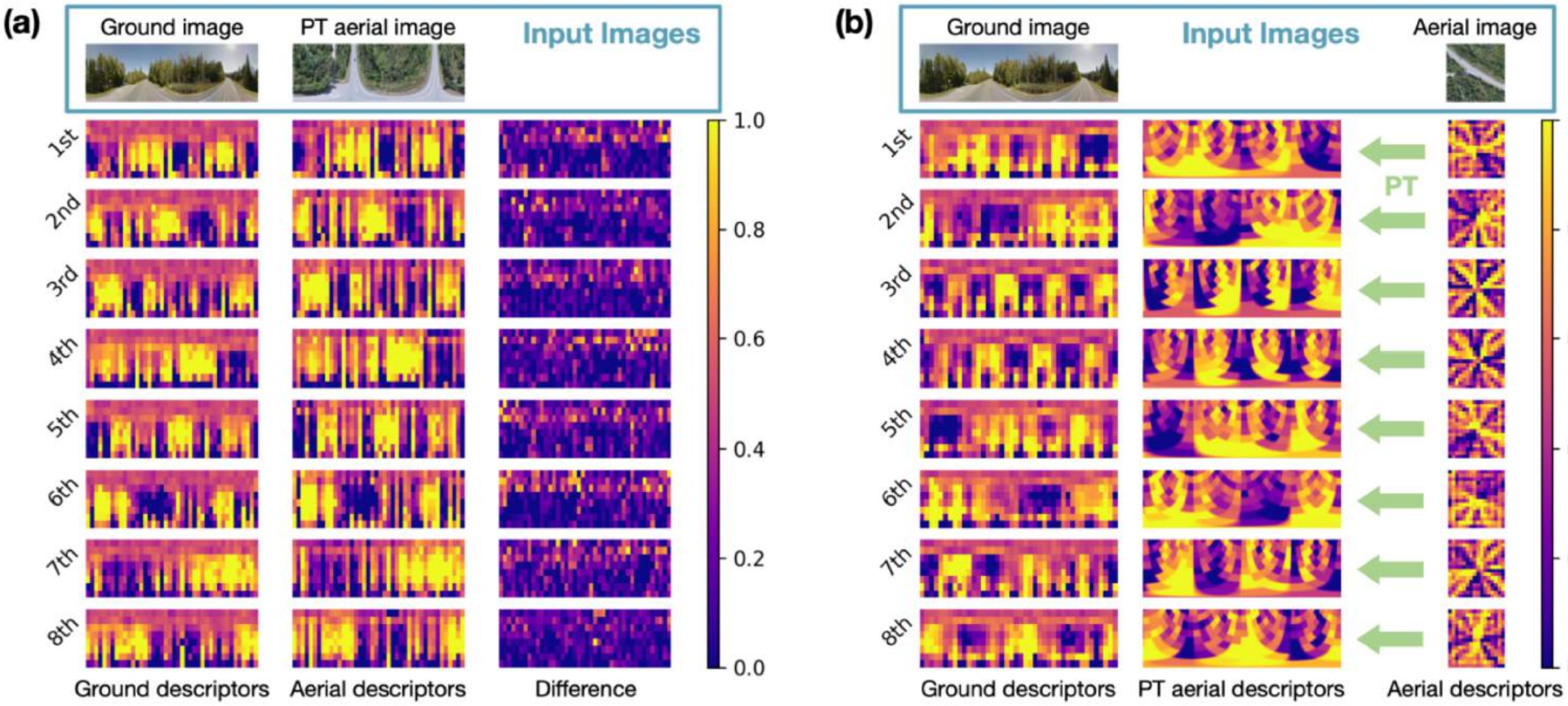
Experiment – Cross-area

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	Configuration	Same-area R@1	Same-area R@5	Same-area R@10	Same-area R@1%	Cross-area R@1	Cross-area R@5	Cross-area R@10	Cross-area 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

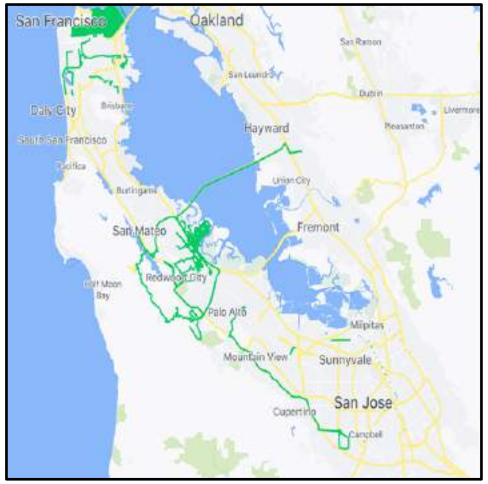


Summary

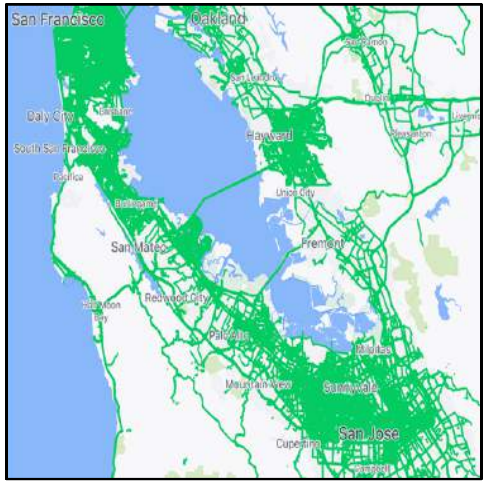
1. **GeoDTR** disentangles geometric information from raw features to better captures the correspondence between aerial and ground images.
2. **Layout simulation and semantic augmentation (LS)** techniques improve the performance of GeoDTR (as well as other existing models) on cross-area experiments.
3. A novel **counterfactual-based learning schema** guides GeoDTR to better grasp the spatial configurations and therefore produce better latent feature representations.

Limitations

- Cross-view image geo-localization heavily rely on panoramic query images.
- Limited field-of-view (FOV) images are more common than panoramas.



(a) Panoramic images coverage



(b) Limited FOV images coverage

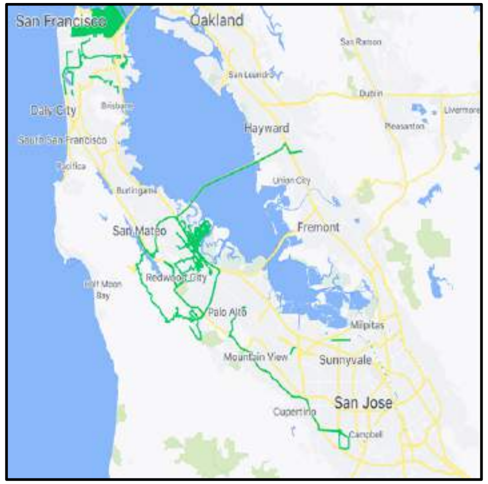
Cross-View Image Sequence Geo-localization



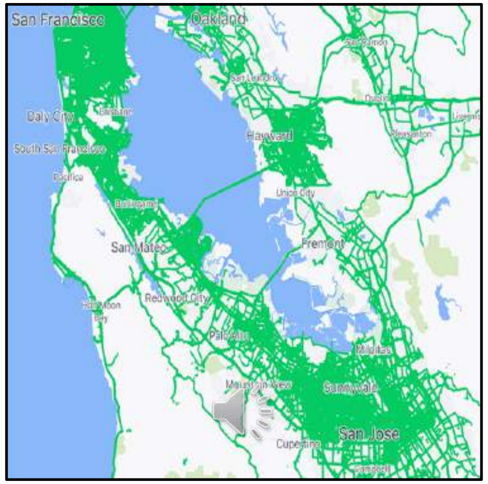
Xiaohan Zhang, **Waqas Sultani**, Safwan Wshah , **Cross-View Image Sequence Geo-localization,** WACV 2023

Motivation

- Cross-view image geo-localization heavily rely on panoramic query images.
- Limited field-of-view (FOV) images are more common than panoramas.
- Sequence of limited FOV images expands the range of visibility of a single limited FOV image.
- We propose to geo-locate sequences of limited FOV images instead of panoramas.



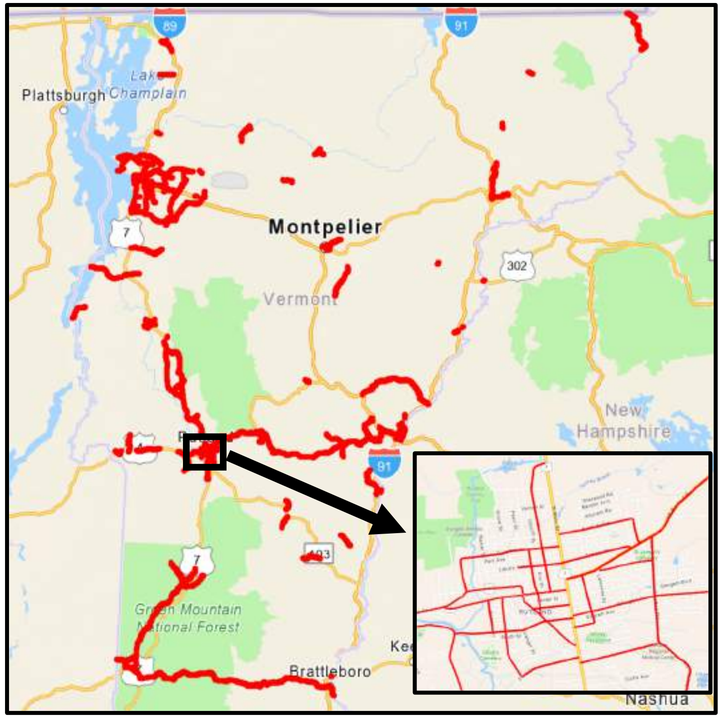
(a) Panoramic images coverage



(b) Limited FOV images coverage

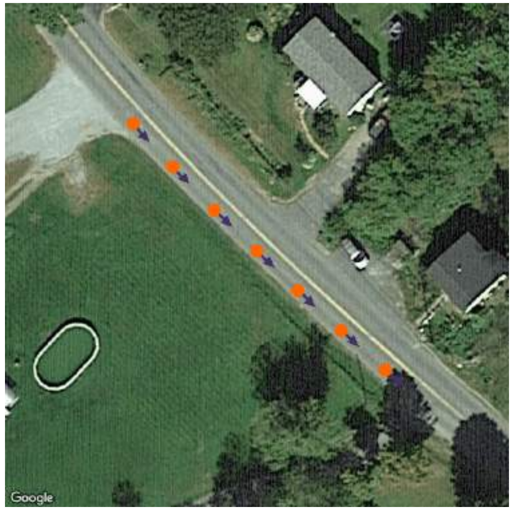
Dataset

- Covers more than *500 km of road* in Vermont, USA.
- Various coverage area, urban, suburban, highway, etc.
- Dataset contains *118,549* ground images and forms *38,863* satellite-sequence pairs.
- The dataset does not contain panoramic images.

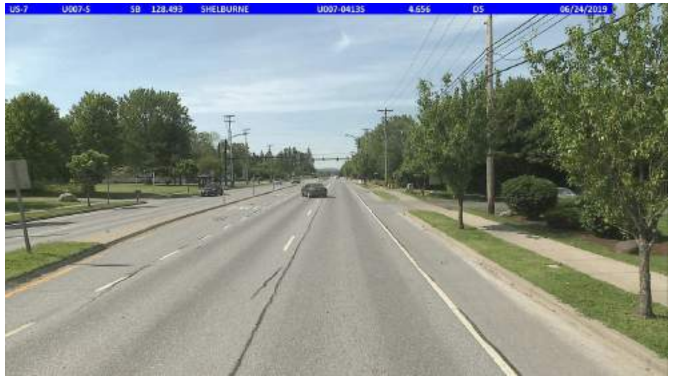


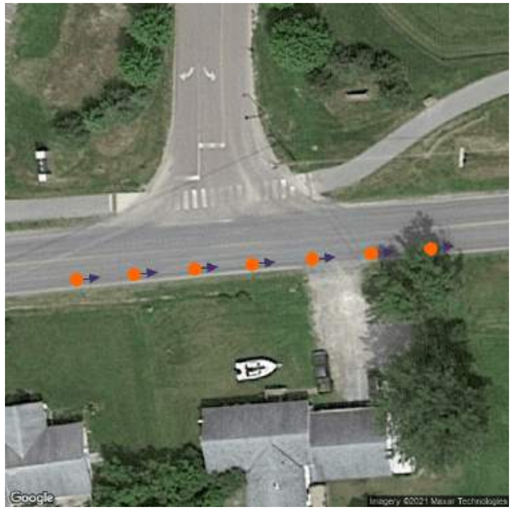
Dataset coverage area

Samples from our dataset



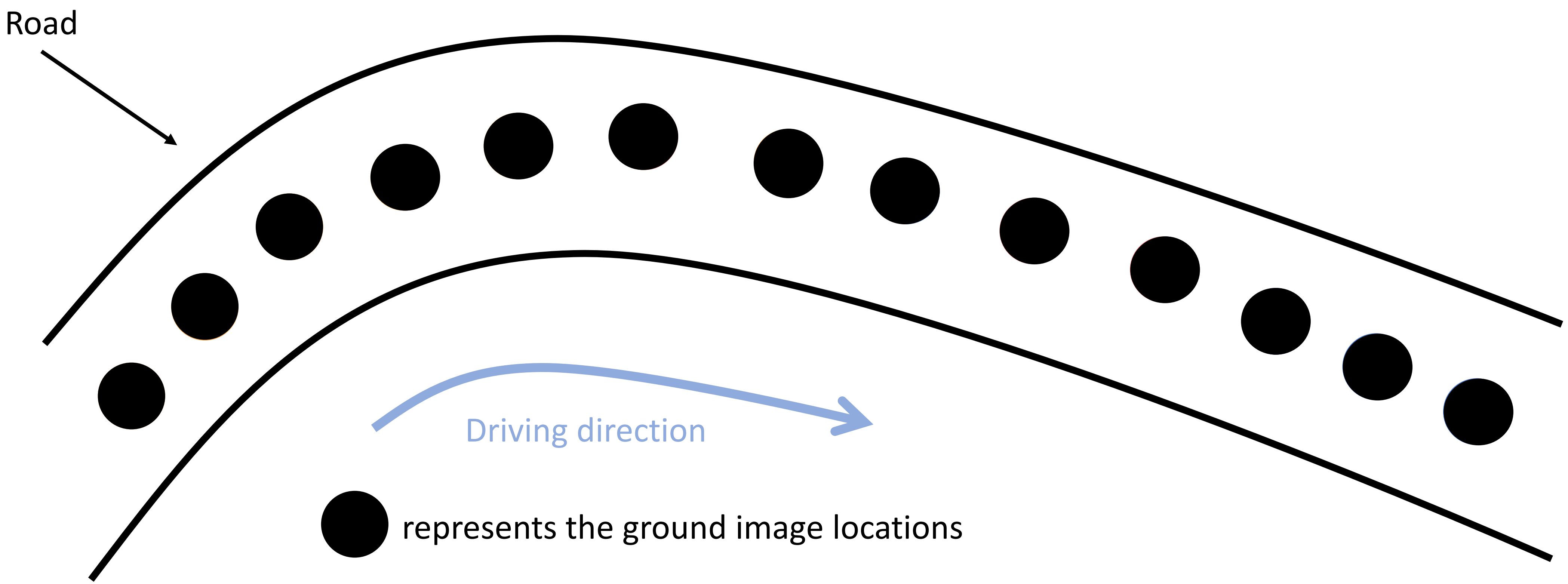
Samples from our dataset



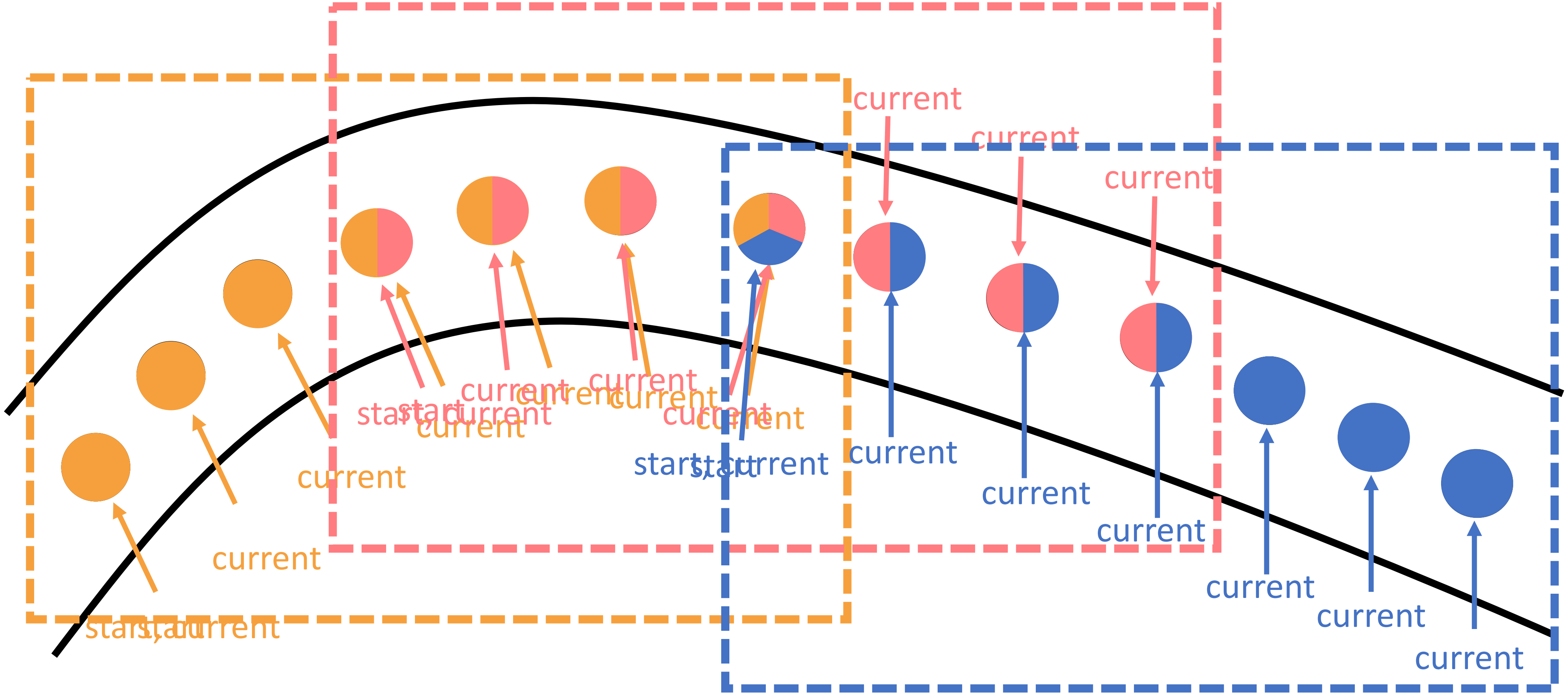


The University of Vermont
ITU INFORMATION TECHNOLOGY UNIVERSITY

Sequence Formation

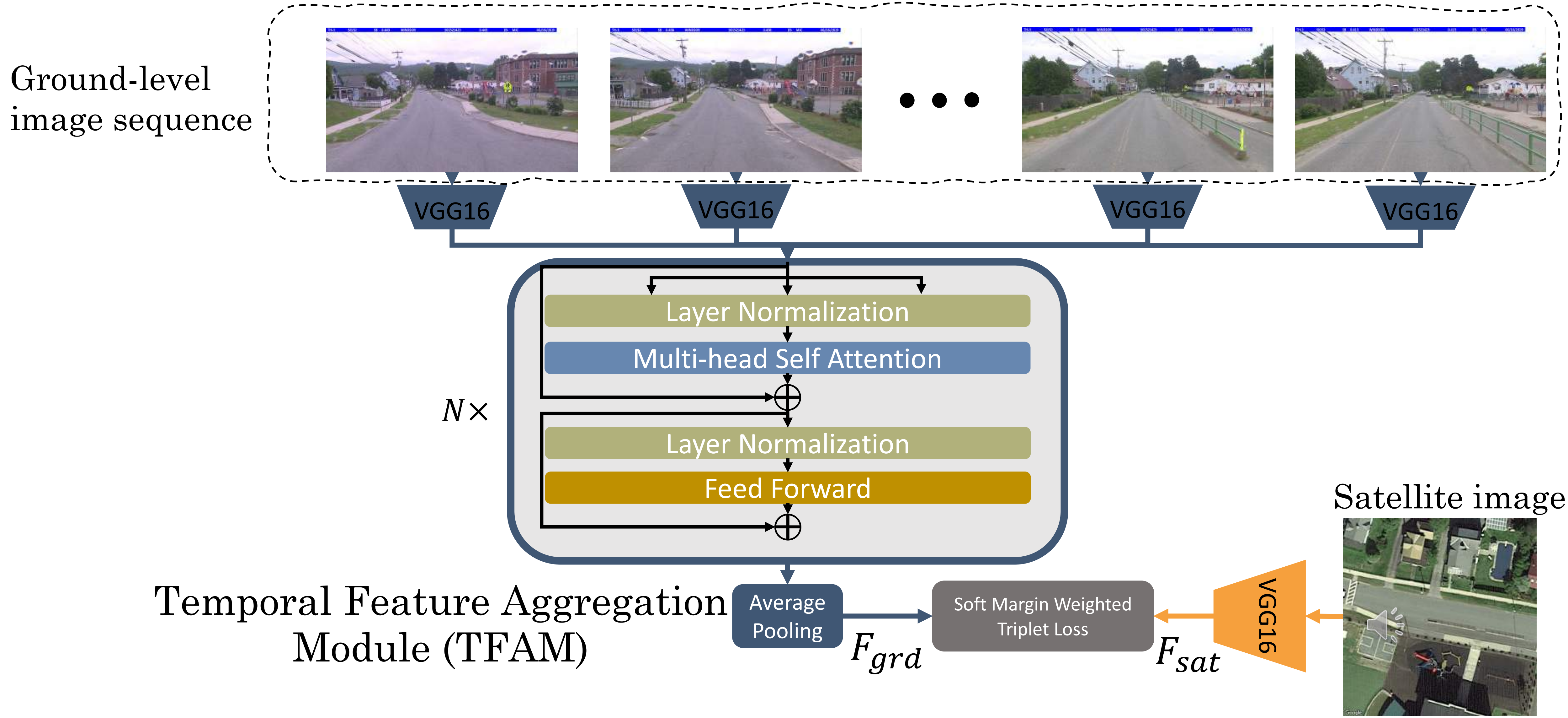


Sequence Formation

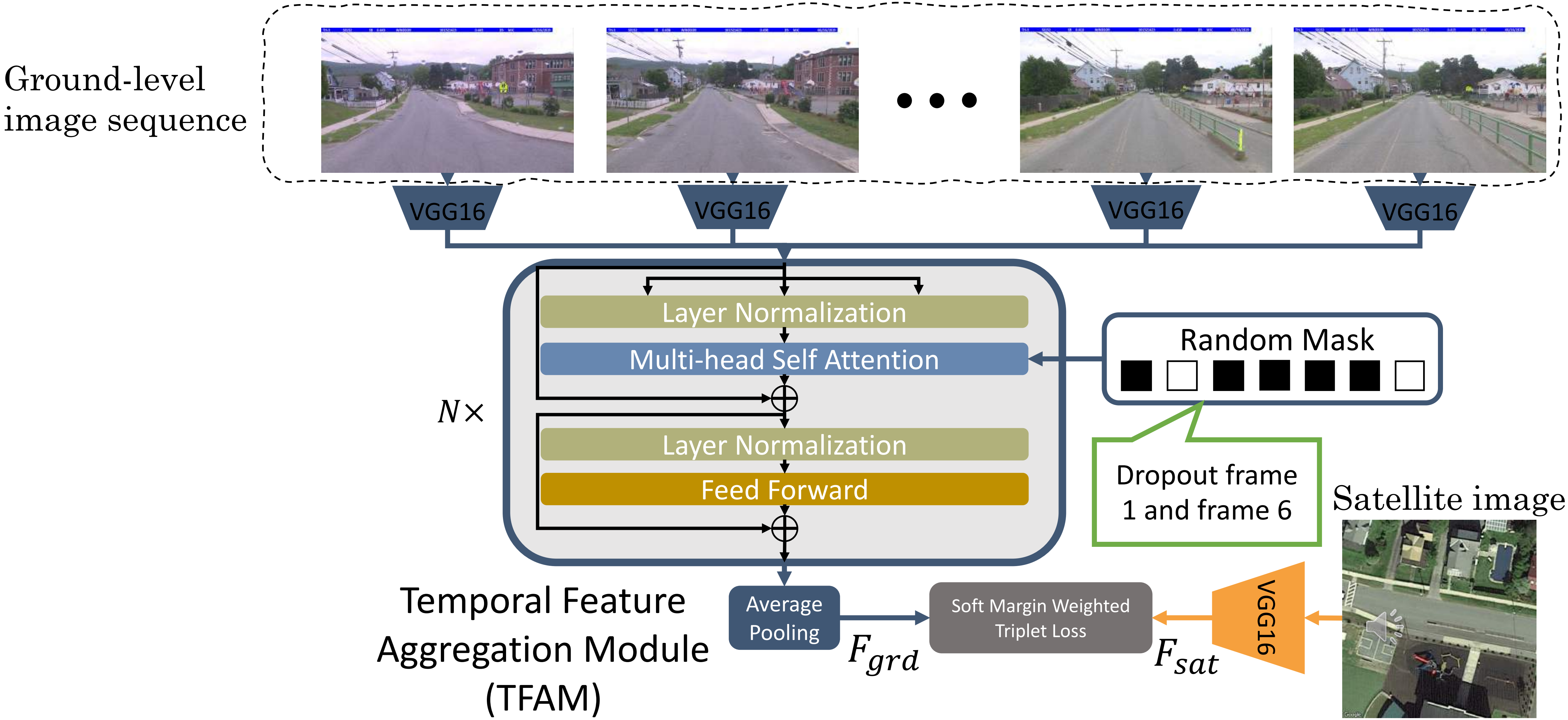


$$distance(start, current) > \Delta$$

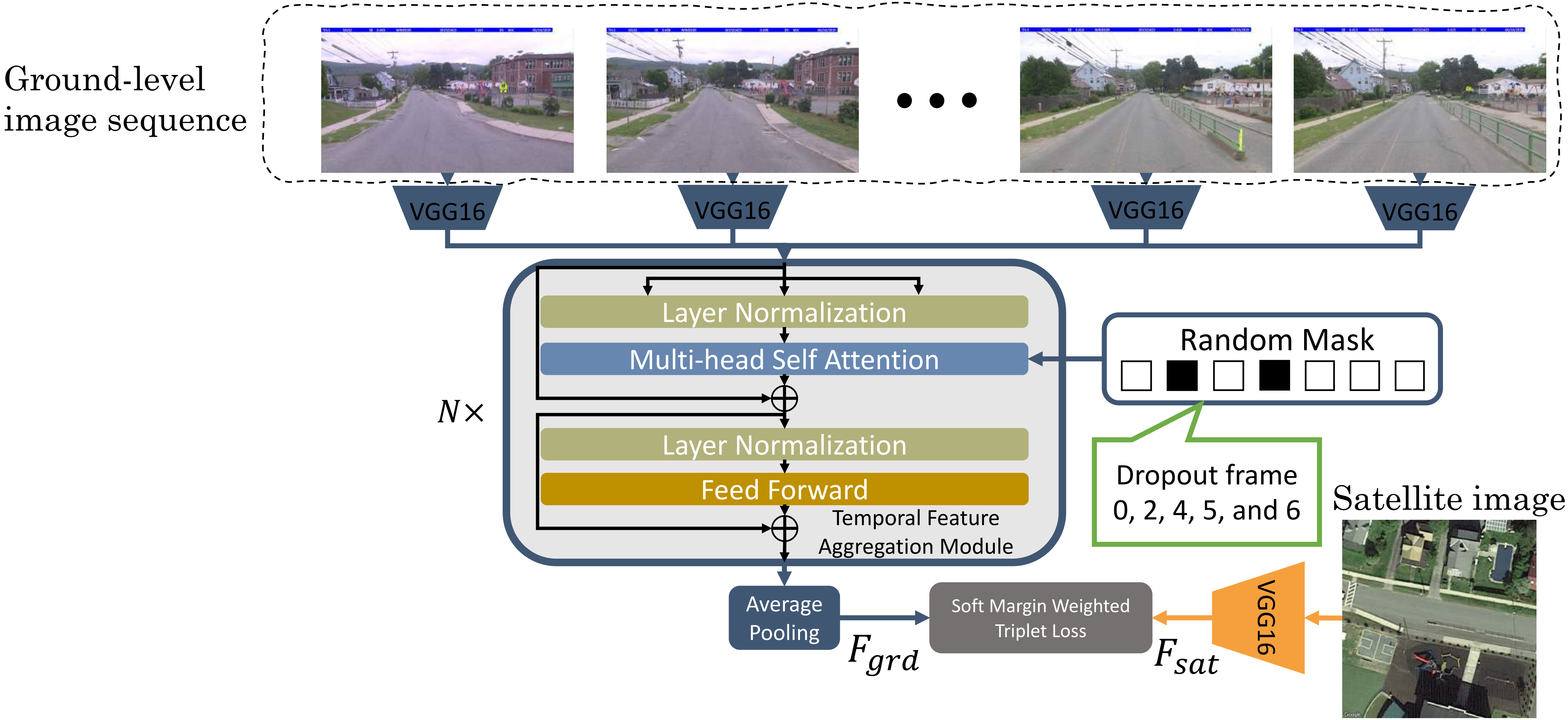
Proposed Model



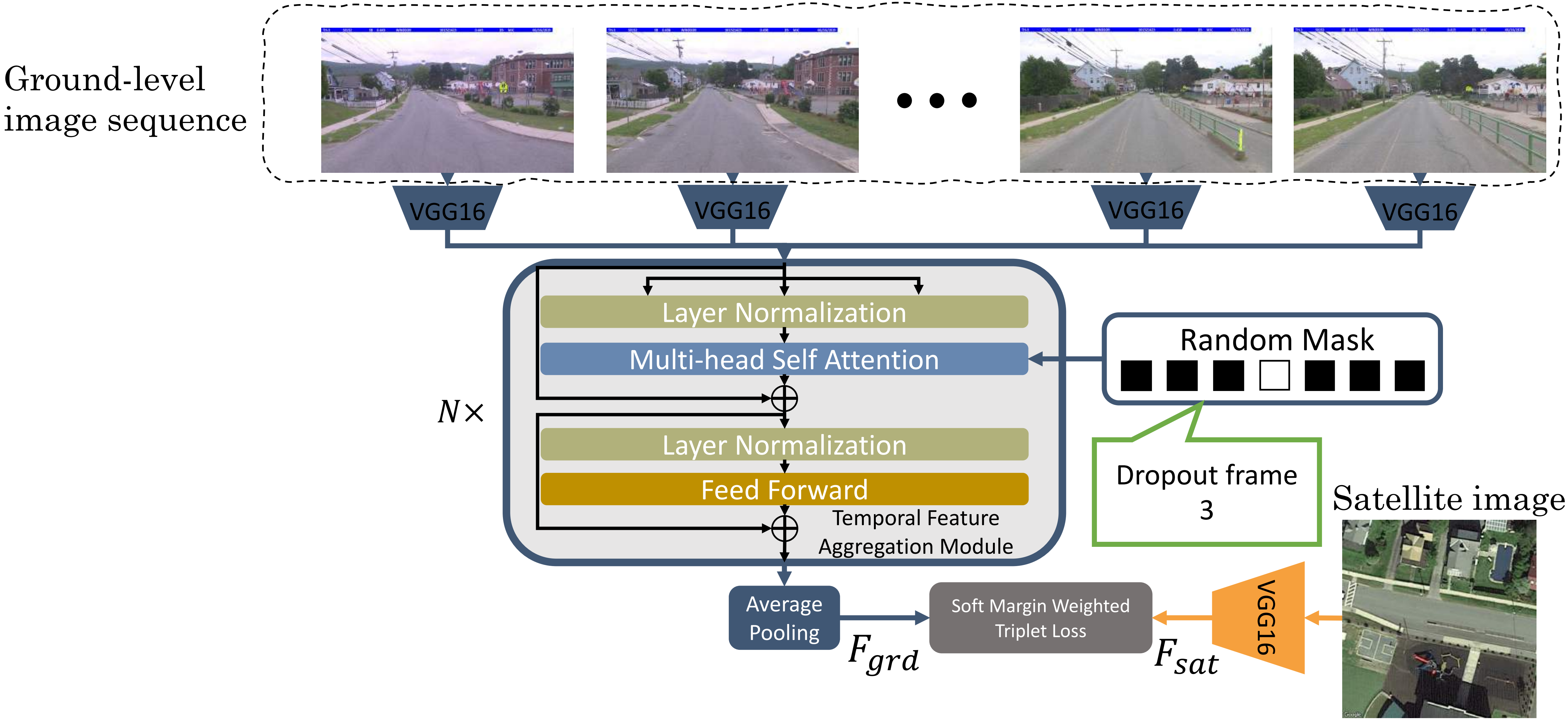
Sequential Dropout



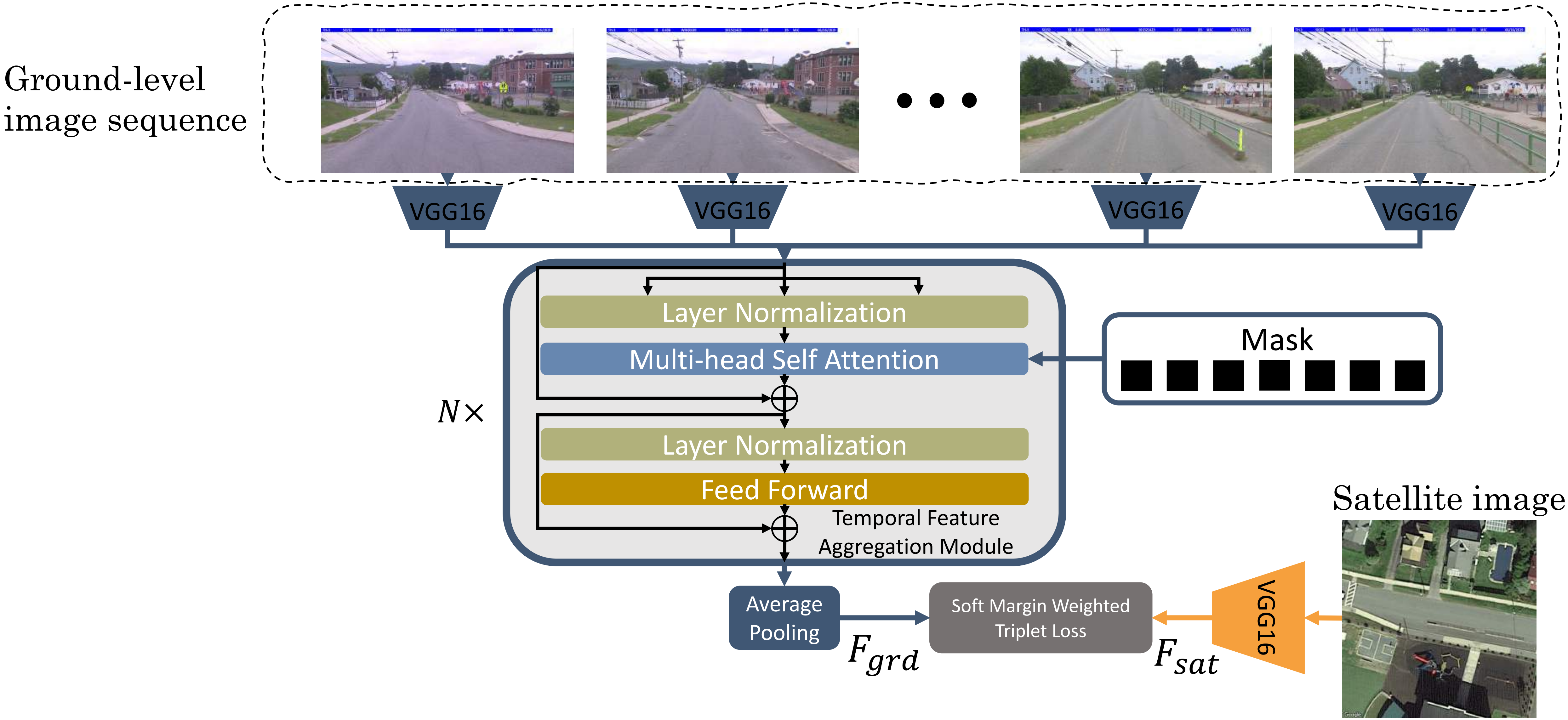
Sequential Dropout



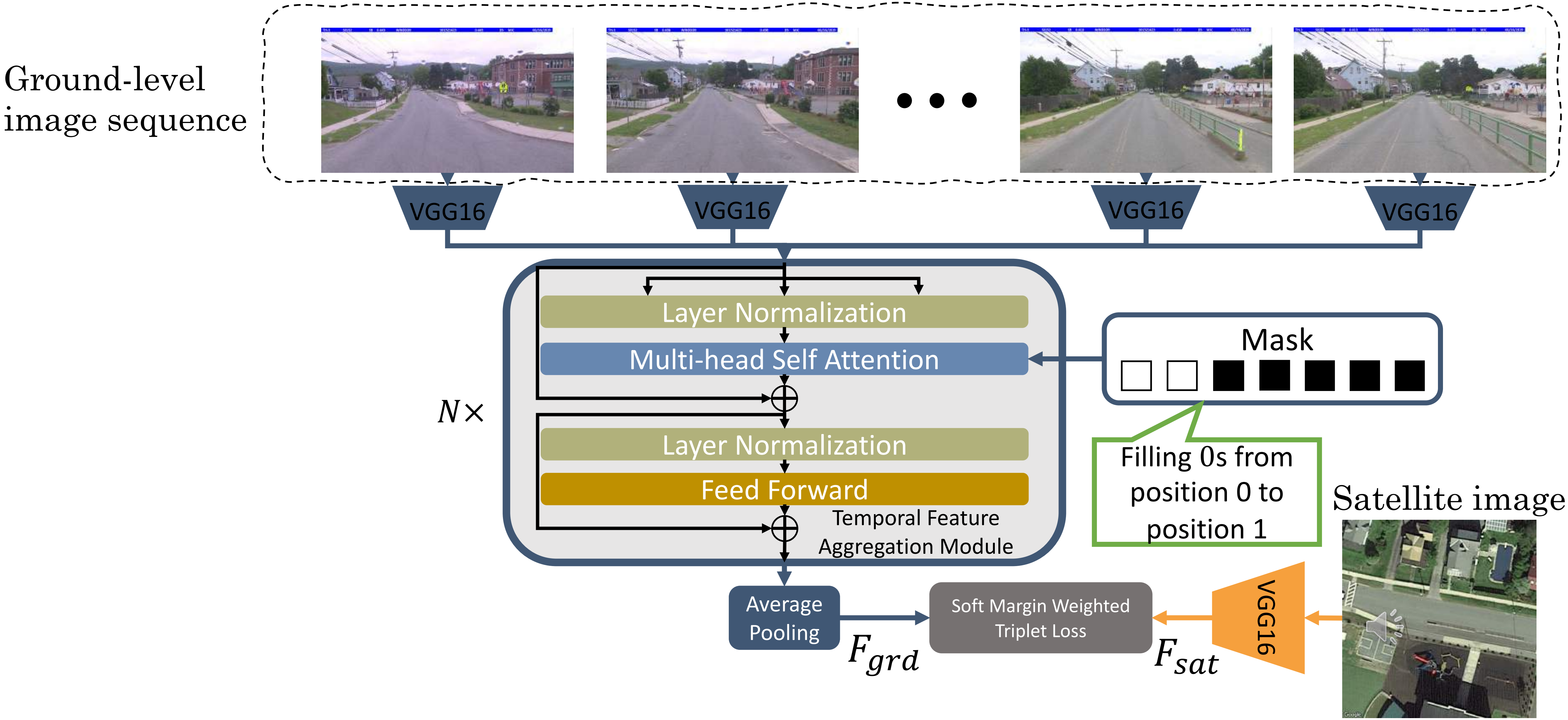
Sequential Dropout



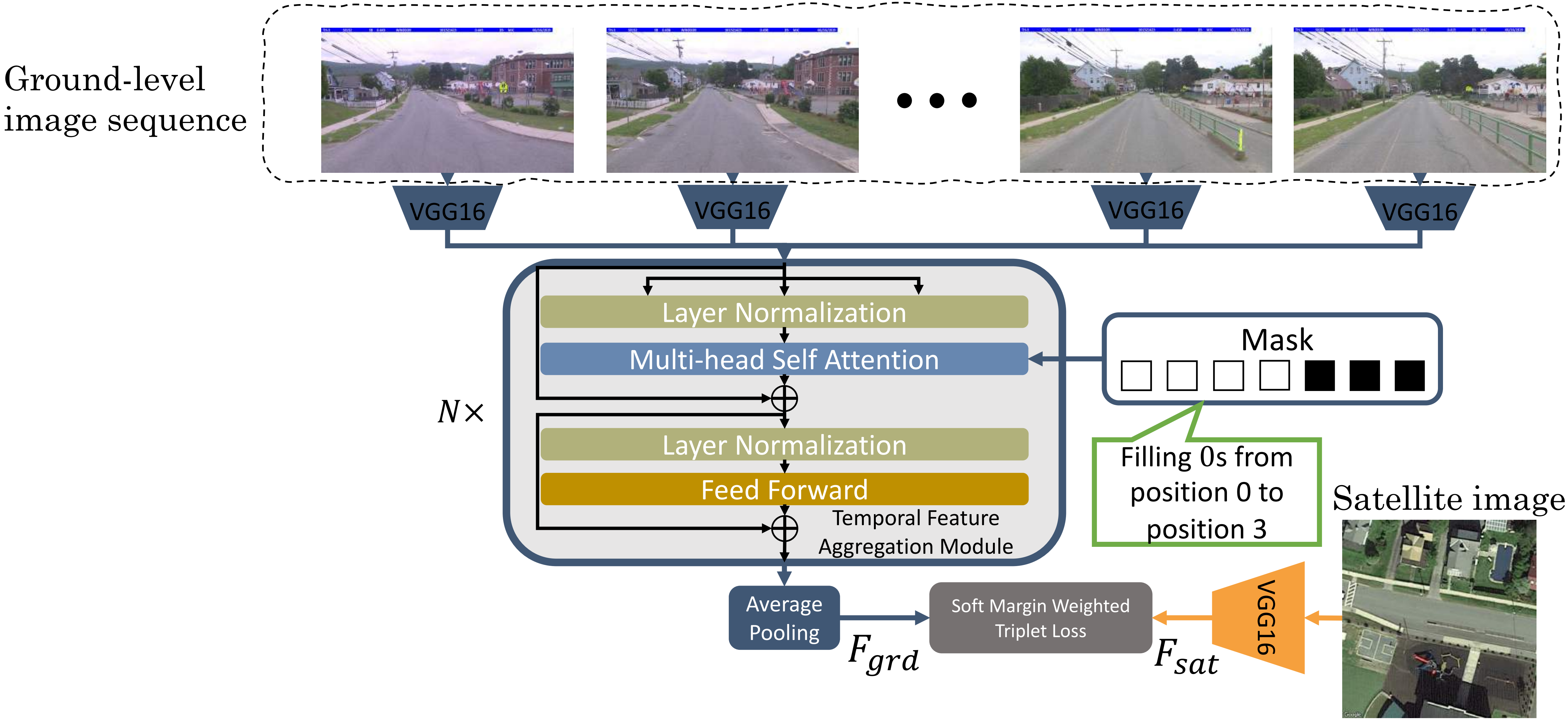
Testing – using the full sequence



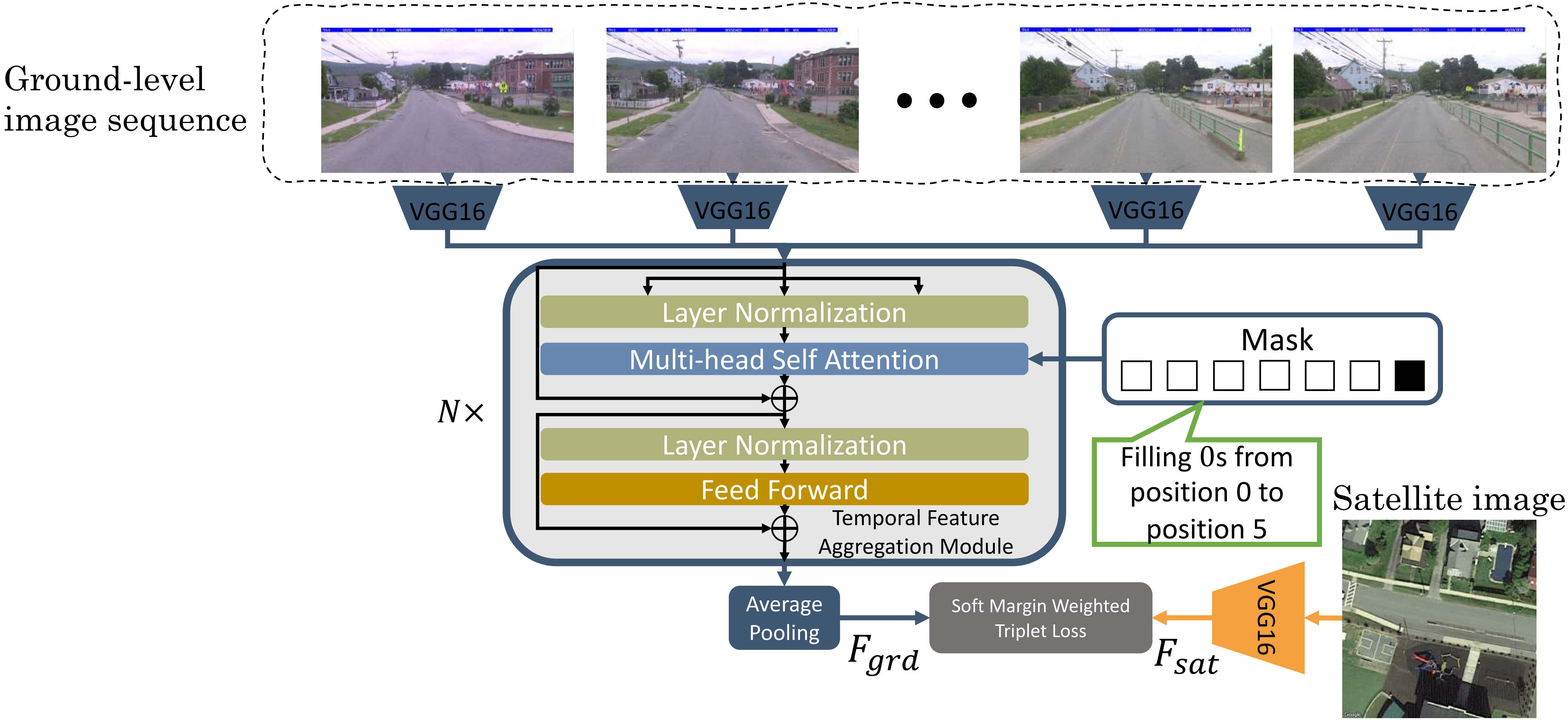
Inferencing



Testing – simulating sequence length=3



Testing – simulating sequence length=1



Experiments

Baseline Methods:

- SAFA (center) : Training on using center image as query and testing on query center image only.
- SAFA (sequence) : Training on using center image as query and testing on query sequence by feature averaging.
- VIGOR: Training on a query sequence in which center image is considered as "positive" and other images are "semi-positive".

Evaluation Metric:

We choose to use recall accuracy at top K ($R@K$) for evaluation purpose.

$R@K$ measures the probability of the ground truth aerial image ranking within the first K predictions given a query image. In the experiments,

We evaluate for:

$R@1, R@5, R@10,$ and $R@1\%$.

Experiments

	R@1	R@5	R@10	R@1%
VIGOR	0.54%	2.52%	4.48%	18.55%
SAFA(center image as query)	0.68%	2.92%	5.06%	21.81%
SAFA(sequence as query)	0.63%	2.83%	5.03%	21.51%
Ours w/o Sequential Dropout	1.39%	6.50%	10.45%	32.42%
Ours	**1.80%**	**6.45%**	**10.36%**	**34.38%**

Comparison between our proposed method and SOTA methods on the proposed dataset

Ablation Studies

# of TFAMs	# of Heads	R@1	R@5	R@1%
0	0	0.91%	4.49%	26.69%
2	2	1.45%	6.22%	31.84%
4	2	1.40%	6.34%	32.97%
4	4	1.51%	6.27%	32.93%
6	4	1.59%	6.02%	32.14%
6	8	1.80%	6.45%	34.38%

Ablation study on # of TFAM and # of attention heads

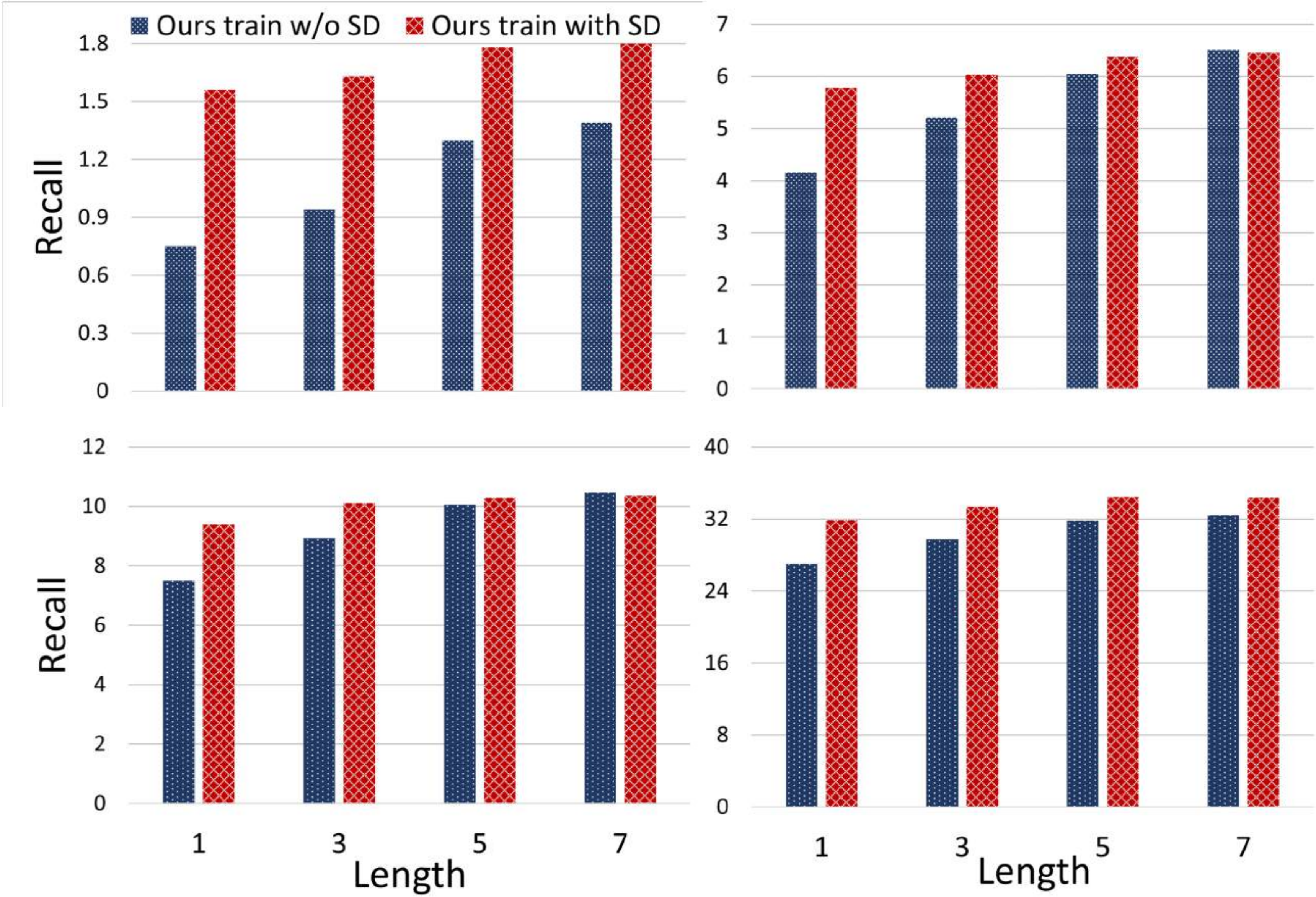
# of dropout images	R@1	R@5	R@1%
1	1.40%	6.08%	31.89%
3	1.51%	6.64%	34.34%
5	1.63%	6.41%	34.40%
6	1.80%	6.45%	34.38%

Ablation study on # of dropout images in the ground sequence

Variant Sequence Lengths

Recall@1

Recall@5



Recall@10

Recall@1%

Qualitative Results

Query sequence

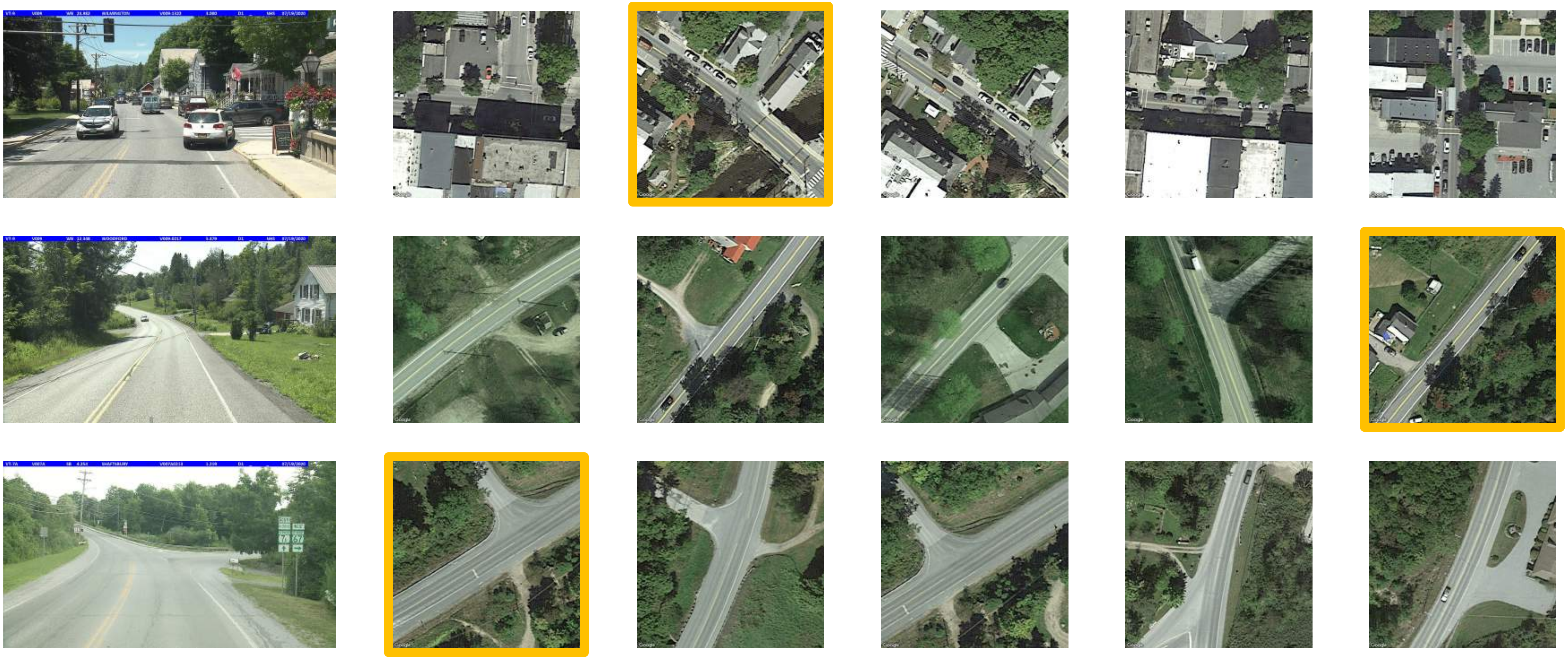
Top-5 predictions

Blue boarder indicates ground truth

Sample results

Query sequence

Top-5 predictions



Blue boarder indicates ground truth

Summary

1. A new end-to-end approach for cross-view image sequence geo-localization.

2. Put forward a novel large-scale cross-view image sequence geo-localization dataset.

3. Propose a new sequential dropout technique to regularize the model to predict coherent features on sequences of different lengths.

Code and dataset are available at https://gitlab.com/vail-uvm/seqgeo

Codes

- https://zxh009123.github.io/
- https://gitlab.com/vail-uvm/geodtr
- https://gitlab.com/vail-uvm/seqgeo

Thanks