

Multimodal Remote Sensing Scene Classification with AI

**Exploring VLMs and Dual-Cross Attention
Networks**

Computer and Information Technology

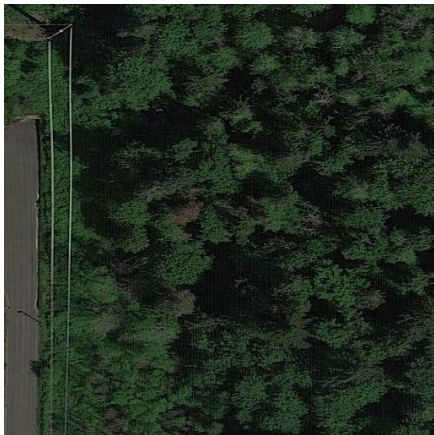
cai379@purdue.edu

10/31/2024

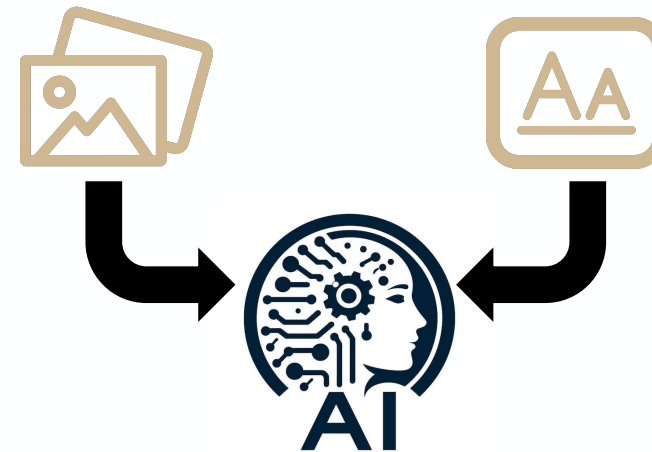


Using AI to enhance scene classification from satellite imagery

Subhead, Franklin Gothic 22



- This project uses AI to interpret complex remote sensing scenes by integrating data from multiple sources. AI enables faster, more accurate classification of scenes, helping make better decisions in industries like agriculture and urban planning.



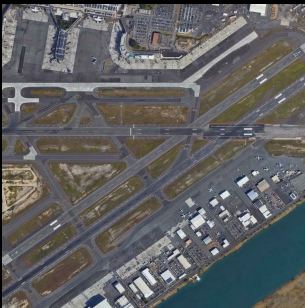
Improves accuracy and speed of classification

Supports better decision-making in critical areas

Makes large-scale, complex data manageable

Challenges in Scene Classification - Image Challenge

High intra-class variance



(a) Runways

High inter-class similarity



(b) Basketball (up) and tennis ball (below) courts

Large variations in the scales



(c) Storage tanks

Coexistence of multiple ground object



(d) Commercial (up) and industrial (below) area

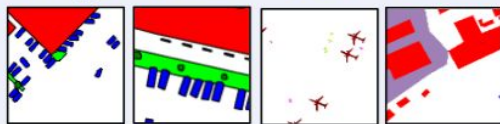
Challenges in Scene Classification - Image Annotation

Previous practice:


- Manual generation of description: labor-intensive, may incomplete
- Automated methods to transform traditional datasets into descriptions
- Example:

Segmentation Datasets (SEG-4)


Vaihingen Potsdam iSAID LoveDA



Mask-to-Box (M2B)



Box-to-Caption (B2C)




A large deal of places with cars are in the middle of the picture .
 There is a place with building in this remote sensing picture .
 4 large vehicles, a small vehicle and 5 planes at the edge of the picture .
 Lots of buildings are located in this remote sensing picture .


Detection Datasets (DET-10)

Satellite Imagery

DOTA DIOR HRRSD RSOD LEVIR HRSC



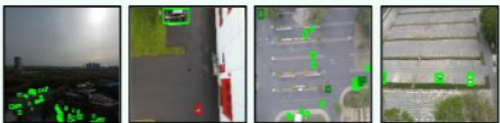
Box-to-Caption (B2C)




Many small-vehicles in the middle of the picture .
 A quantity of airplanes are located in the picture .
 Lots of airplanes are located in the picture .
 There are 15 aircrafts on the ground .
 An airplane in the middle of the picture .
 A ship in the middle of the picture .

UAV Imagery

VisDrone AU-AIR S-Drone CAPRK



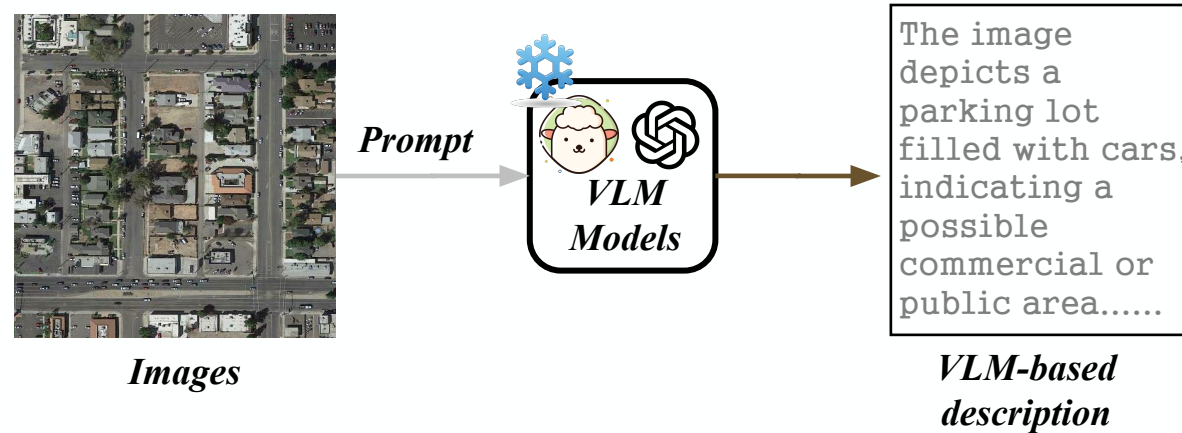
Box-to-Caption (B2C)



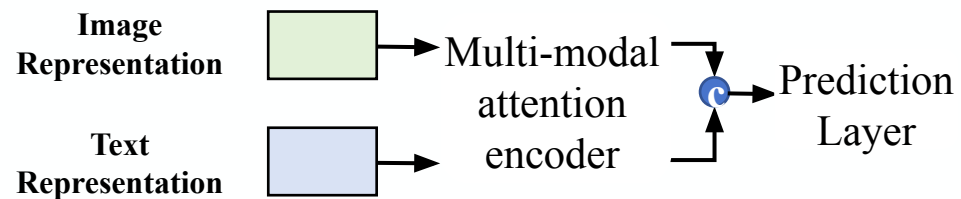
A crowd of cars are located in this remote sensing picture .
 A human and a car at the edge of the picture .
 There are both pedestrians and bikers in the picture .
 There are 3 cars in the picture .

Our Approach: Multimodal Scene Classification with the help of AI VLM

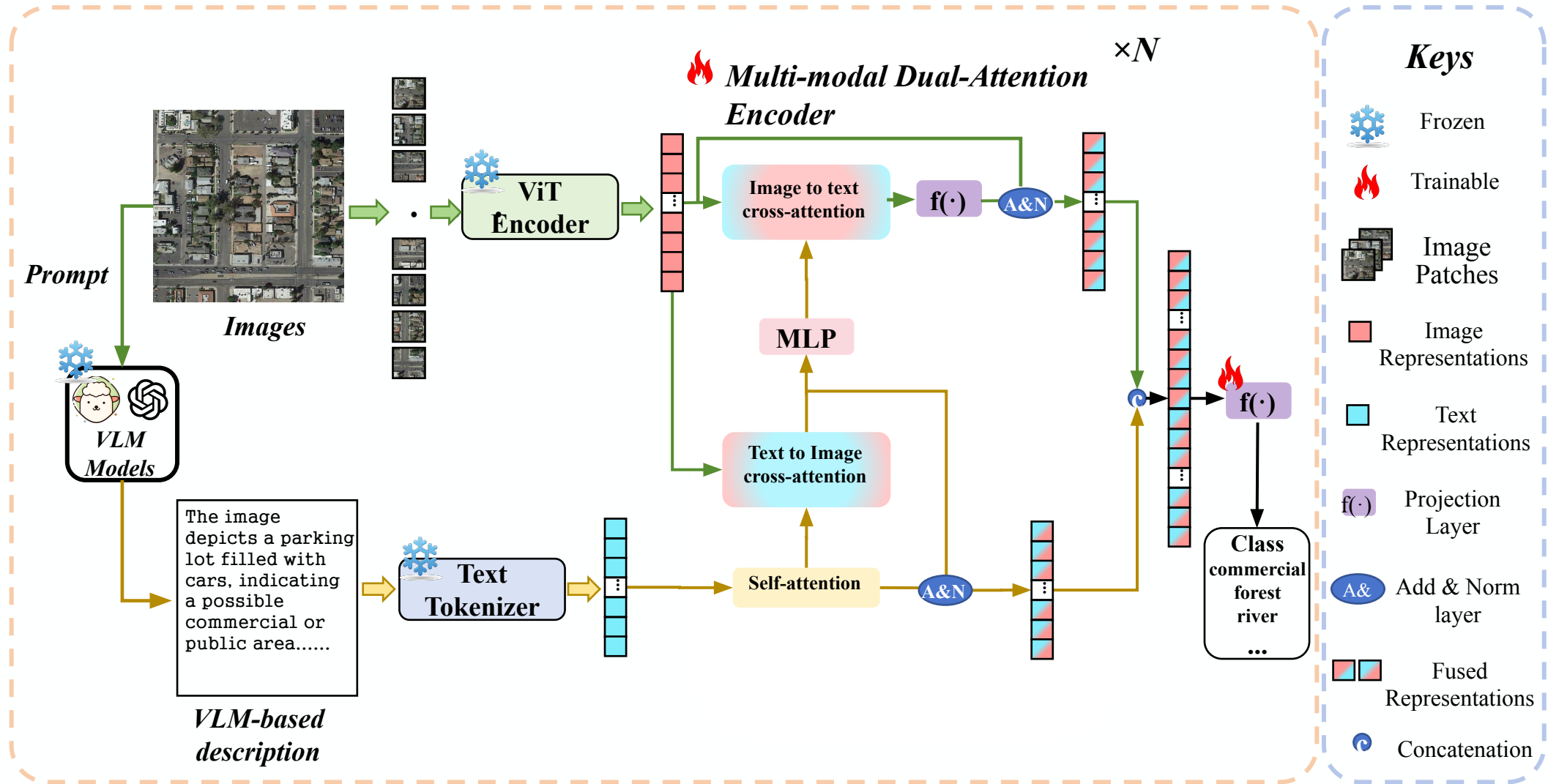
Key Technology 1:
Vision Language Models (VLMs)



Key Technology 2:
Dual-Cross Attention Networks



Project Workflow

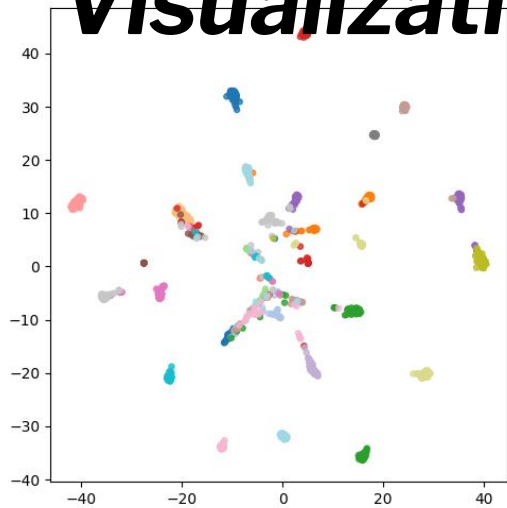


Key Results and Benefits: Quantitative Analysis

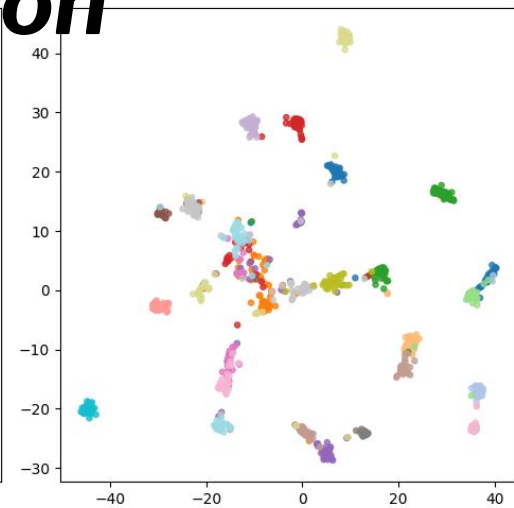
Table 1. Summary of experimental results in terms of average and standard deviations of Overall Accuracy (OA%), Average Accuracy (AA%), and *Kappa* coefficient (*Kappa*%).

Dataset Baselines	AID			PatternNet			Million-AID-2			Million-AID-3			DeepGlobe		
	OA/top-5	AA	<i>Kappa</i>	OA/top-5	AA	<i>Kappa</i>	OA/top-5	AA	<i>Kappa</i>	OA/top-5	AA	<i>Kappa</i>	OA/top-3	AA	<i>Kappa</i>
<i>IO</i>	95.6/99.8 (0.4/0.1)	94.3 (0.5)	94.8 (0.3)	98.0/99.8 (0.5/0.1)	97.5 (0.4)	97.0 (0.3)	88.3/96.1 (0.2/0.2)	87.8 (0.3)	87.3 (0.3)	85.9/96.9 (0.4/0.2)	85.5 (0.4)	85.0 (0.4)	69.9/85.2 (0.5/0.3)	69.0 (0.5)	68.4 (0.4)
<i>TO</i>	89.5/97.0 (1.1/0.1)	88.0 (1.0)	88.5 (0.9)	92.1/97.3 (0.8/0.2)	91.5 (0.9)	91.0 (0.7)	84.5/95.5 (0.4/0.3)	83.4 (0.4)	83.6 (0.4)	80.6/95.2 (0.5/0.4)	78.5 (0.4)	79.0 (0.4)	63.3/89.7 (0.9/0.8)	62.5 (1.0)	62.0 (0.8)
<i>EF</i>	95.5/99.8 (0.8/0.1)	94.0 (0.7)	94.5 (0.6)	97.2/99.7 (0.5/0.1)	96.5 (0.5)	96.0 (0.5)	88.5/96.5 (0.5/0.4)	88.2 (0.5)	88.4 (0.5)	85.6/97.6 (0.4/0.4)	83.5 (0.4)	85.1 (0.3)	79.7/94.2 (0.3/0.3)	79.0 (0.4)	78.5 (0.3)
<i>LF</i>	95.9/99.8 (0.7/0.1)	94.5 (0.6)	94.9 (0.5)	97.8/99.6 (0.5/0.0)	97.0 (0.6)	96.5 (0.4)	88.3/97.0 (0.7/0.4)	87.9 (0.5)	88.0 (0.5)	85.2/96.8 (0.7/0.5)	83.8 (0.3)	85.2 (0.3)	79.5/94.0 (0.2/0.2)	79.5 (0.2)	78.2 (0.3)
<i>no CAtt</i>	97.0/99.5 (0.7/0.1)	96.0 (0.6)	96.5 (0.5)	98.5/99.8 (0.2/0.1)	98.5 (0.2)	98.4 (0.3)	92.5/99.2 (0.8/0.6)	92.0 (0.5)	92.5 (0.5)	90.2/97.8 (0.4/0.3)	88.5 (0.3)	88.9 (0.4)	88.3/99.7 (0.4/0.2)	88.3 (0.4)	88.0 (0.2)
<i>ICAtt</i>	97.5/99.8 (0.4/0.1)	96.5 (0.3)	97.0 (0.2)	99.2/100.0 (0.2/0.1)	99.2 (0.3)	99.1 (0.2)	94.6/99.5 (0.8/0.4)	94.1 (0.7)	94.3 (0.6)	92.5/99.5 (0.5/0.4)	91.6 (0.3)	92.0 (0.3)	85.3/99.7 (0.2/0.1)	85.3 (0.2)	84.8 (0.3)
<i>TCAtt</i>	97.2/99.9 (0.6/0.0)	96.2 (0.5)	96.7 (0.4)	99.0/100.0 (0.5/0.1)	99.0 (0.5)	98.9 (0.4)	93.7/99.5 (0.6/0.5)	93.2 (0.6)	93.5 (0.6)	92.0/99.6 (0.5/0.4)	91.0 (0.3)	92.0 (0.3)	89.7/99.9 (0.5/0.2)	89.7 (0.5)	87.6 (0.4)
Ours	98.9/100.0 (0.8/0.0)	98.1 (0.3)	97.0 (0.2)	99.4/100.0 (0.3/0.0)	99.4 (0.4)	98.5 (0.3)	97.4/99.4 (0.5/0.1)	97.1 (0.5)	97.0 (0.4)	95.6/99.8 (0.8/0.2)	95.0 (0.7)	94.5 (0.6)	91.3/99.9 (1.2/0.5)	90.5 (1.1)	90.0 (1.0)

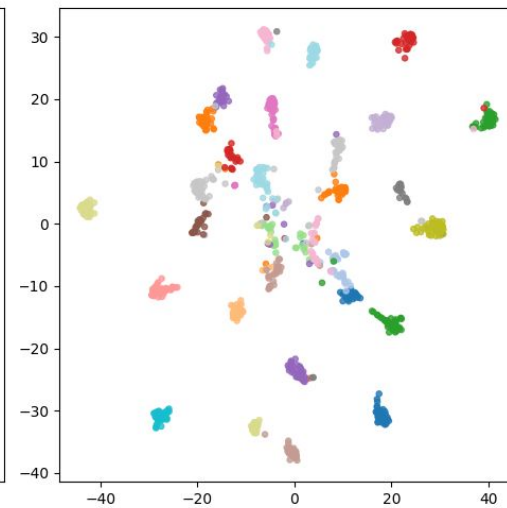
Key Results and Benefits : t-SNE Classification Visualization



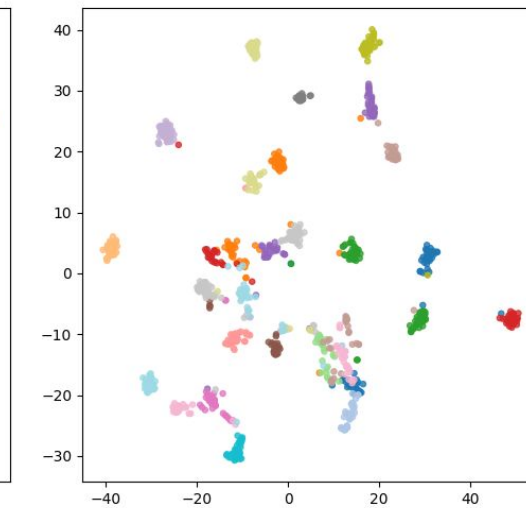
(a) Image Only



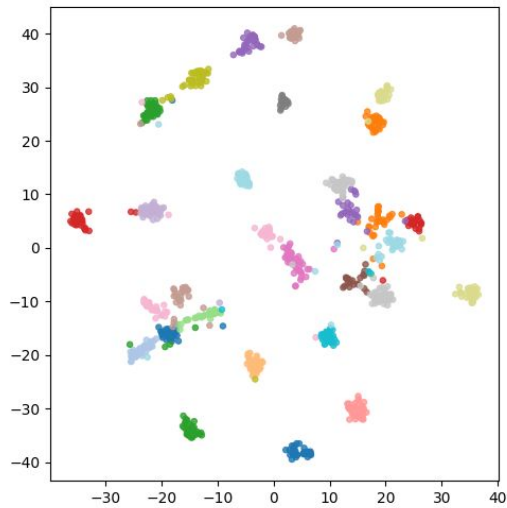
(b) Text Only



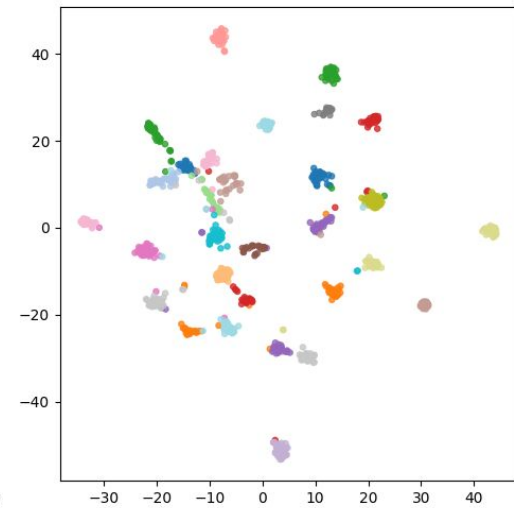
(c) Early Fusion



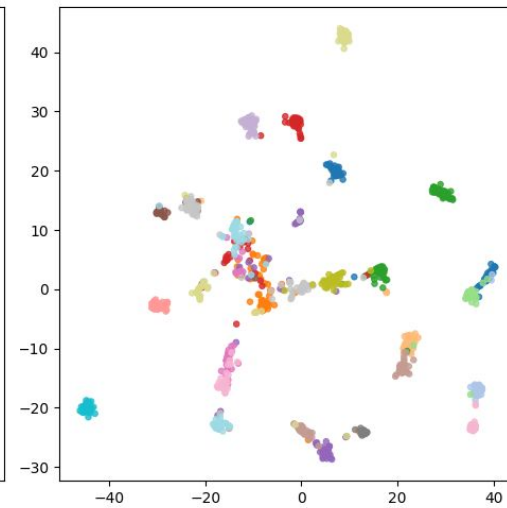
(d) Late Fusion



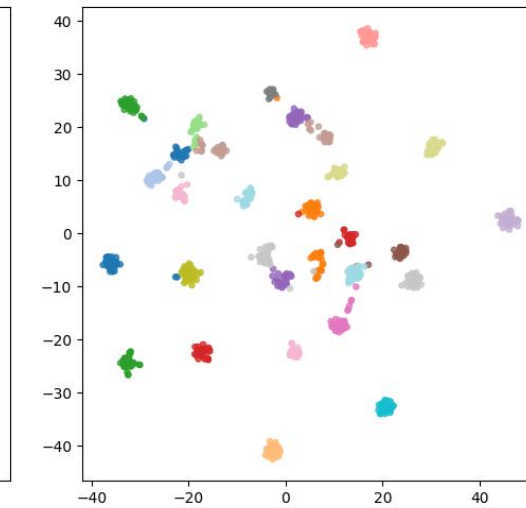
(e) no CAtt



(f) ICAtt



(g) TCAtt

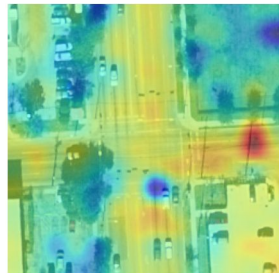
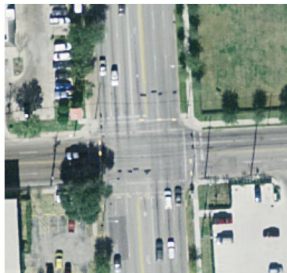


(h) Ours

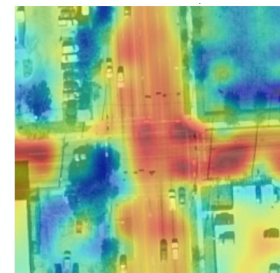


Key Results and Benefits : Attention Heatmap

(a) Intersection

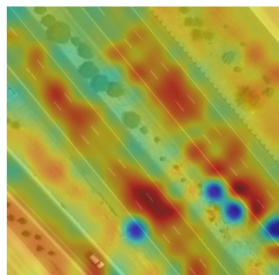
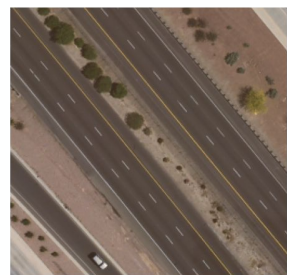


"An intersection only with some houses and plants at the corners."

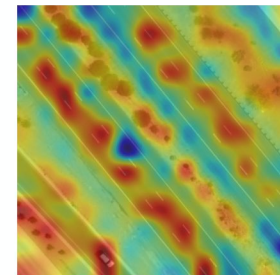


"The aerial image depicts a busy intersection with multiple lanes of traffic both moving and stationary. There are areas of asphalt road, grass, and trees. overall,this scene can be classified as a transportation hub or an intersection."

(b) Freeway

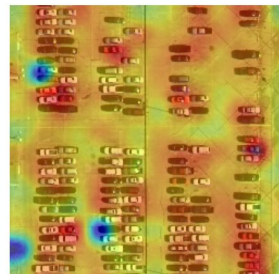
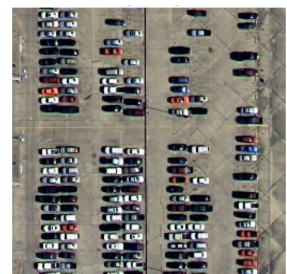


"There are two straight freeways in the desert with no car on them"

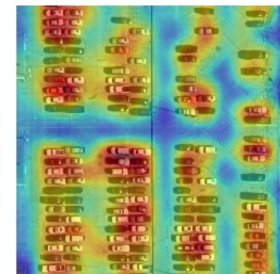


"There are areas of asphalt road,desert vegetation, and some greenery. Overall, this scene can be classified as a desert highway."

(c) Parking Lot

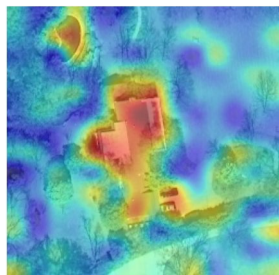


"It is a parking lot with many cars parked neatly and only a few parking spots are free."

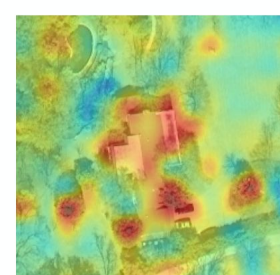


"There are areas of paved parking lot with rows of parked cars,and some empty spaces. Overall,this scene can be classified as a land cover type of urban parking area."

(d) Sparse Residential



"This is a sparse residential area with a villa surrounded by plants and some cars parked there."



"There are areas of green vegetation,a building with a black roof, and a paved area. Overall, this scene can be classified as a residential area with a mix of natural and man-made elements."

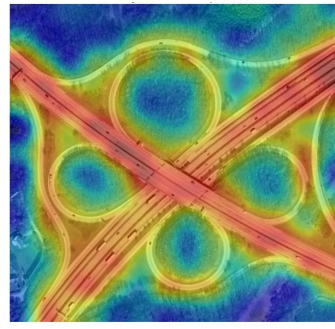
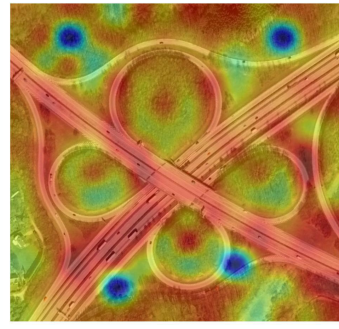
Original

Human-annotated

VLM-generated

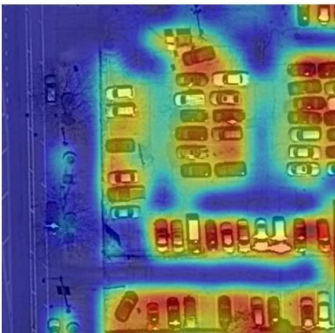
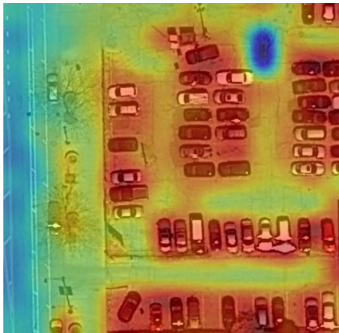
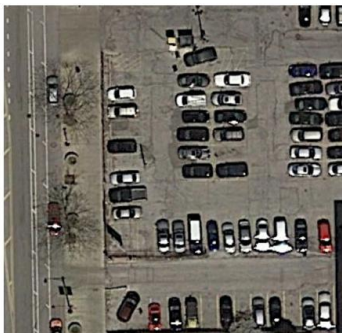
Key Results and Benefits : Attention Heatmap

(a) Viadict



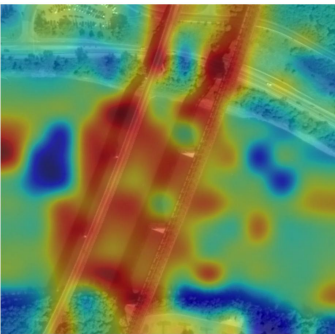
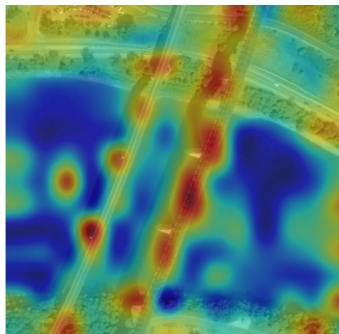
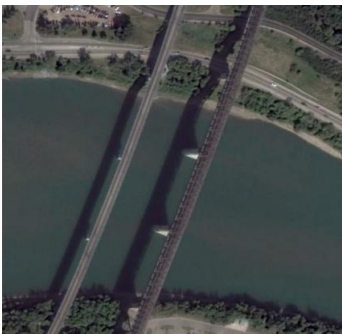
"The image shows a highway intersection with multiple lanes of traffic, surrounded by a mix of greenery, including trees and grass. There are also some buildings and a parking lot visible. Overall, this scene can be classified as a transportation hub with green spaces, possibly indicating a sub-urban or semi-urban area."

(b) Parking



"There are areas of asphalt parking lot with multiple cars, some trees, and a few buildings. Overall, this scene can be classified as an urban land use scene."

(c) Bridge



"There are areas of water, with a large boat moving through it. There is also a bridge spanning the water. Overall, this scene can be classified as a maritime scene with a focus on transportation infrastructure."

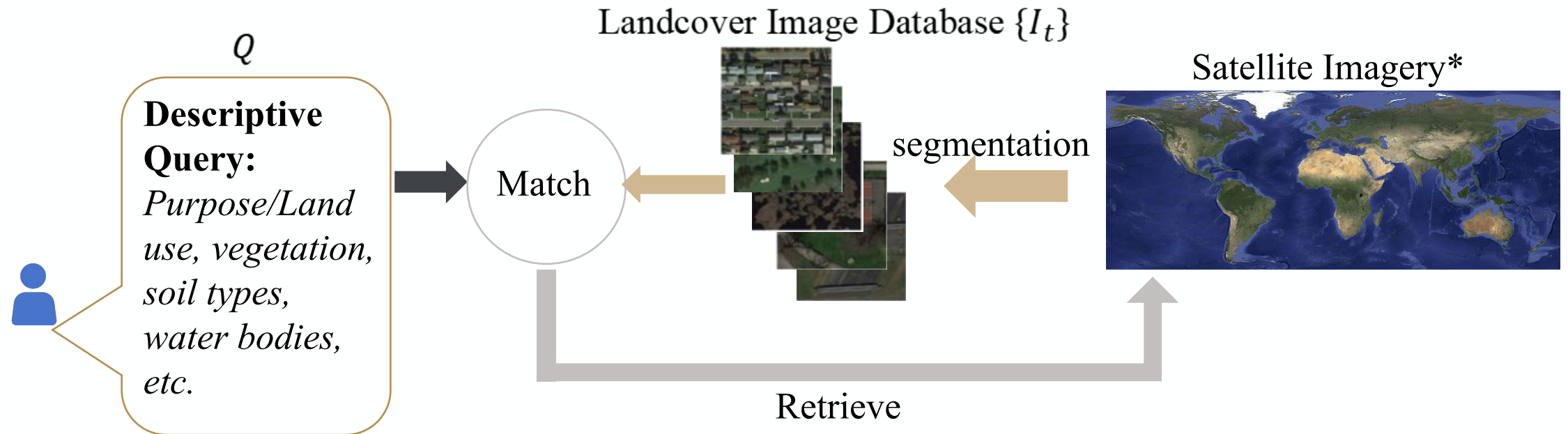
Original

Ablation models

Ours

Corresponding Text




The Future of This AI Project - Retrieval system



Can be vital in:

- Urban Planning and Development,
- Renewable Energy Siting,
- Transportation and Infrastructure,
- Environmental Monitoring and Conservation, etc

Bringing AI Focus to Remote Sensing

- **Effortless Attention on Key Features** 
AI like VLMs can pinpoint essential scene details efficiently, providing valuable focus.
- **Enhanced Scene Understanding** 
Enables models to grasp complex scenes with minimal human intervention, boosting accuracy.
- **Driving Innovation in Geoscience** 
Integrating cutting-edge AI brings fresh insights and fosters advancements in remote sensing.

“
AI enables focused and effective scene understanding, bringing geoscience into a new era of innovation.”

Thank You

Purdue Marketing and Communications, marcom.purdue.edu

