Aerial Scene Parsing

From Tile-level Scene Classification to Pixel-wise Semantic Labeling

Gui-Song Xia

School of Computer Science, Wuhan University
Institute of Artificial Intelligence, Wuhan University
State Key Lab. LIESMARS, Wuhan University

Advanced RS Technology

RS technology has significantly improved the Earth observation ability.

LandSat-1 LandSat-4 Spot-1 IKONOS QuickBird WorldView-1 GeoEye-1 WorldView-3

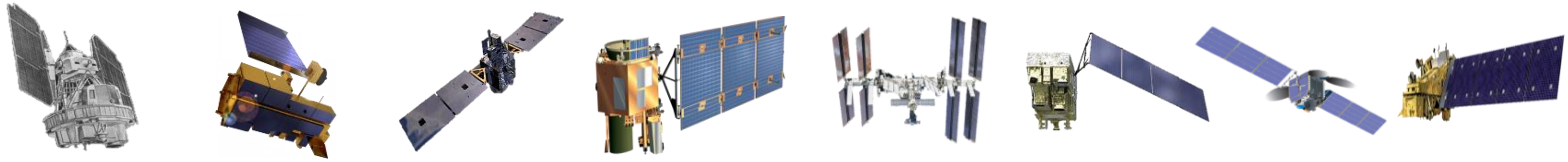
1972	1975	1978	1982	1984	1986	1990	1993	1998	1999	2001
LandSat-1	LandSat-2	LandSat-3	LandSat-4	LandSat-5	Spot-1	Spot-2	Spot-3	Spot-4	IKONOS	QuickBird
78m	78m	78m	30m	30m	10m	10m	10m	10m	1m	0.61m
18d	18d	18d	16d	16d	26d	26d	26d	26d	3d	1-6d

2002	2003	2007	2008	2009	2011	2012	2014	2016	2021	-->
Spot-5	OrbView-3	WorldView-1	GeoEye-1	WorldView-2	Pleiades-1A	Spot-6	WorldView-3	WorldView-4	Legion	SCOUT
2.5m	1m	0.5m	0.41m	0.5m	0.5m	1.5m	0.31m	0.31m	0.29m	0.29m
26d	3d	1.7d	2-3d	1.1d	1d	1d	<1d	<1d	15t/d	24t/d

The characterization of features on the earth surface.

Advanced RS Technology

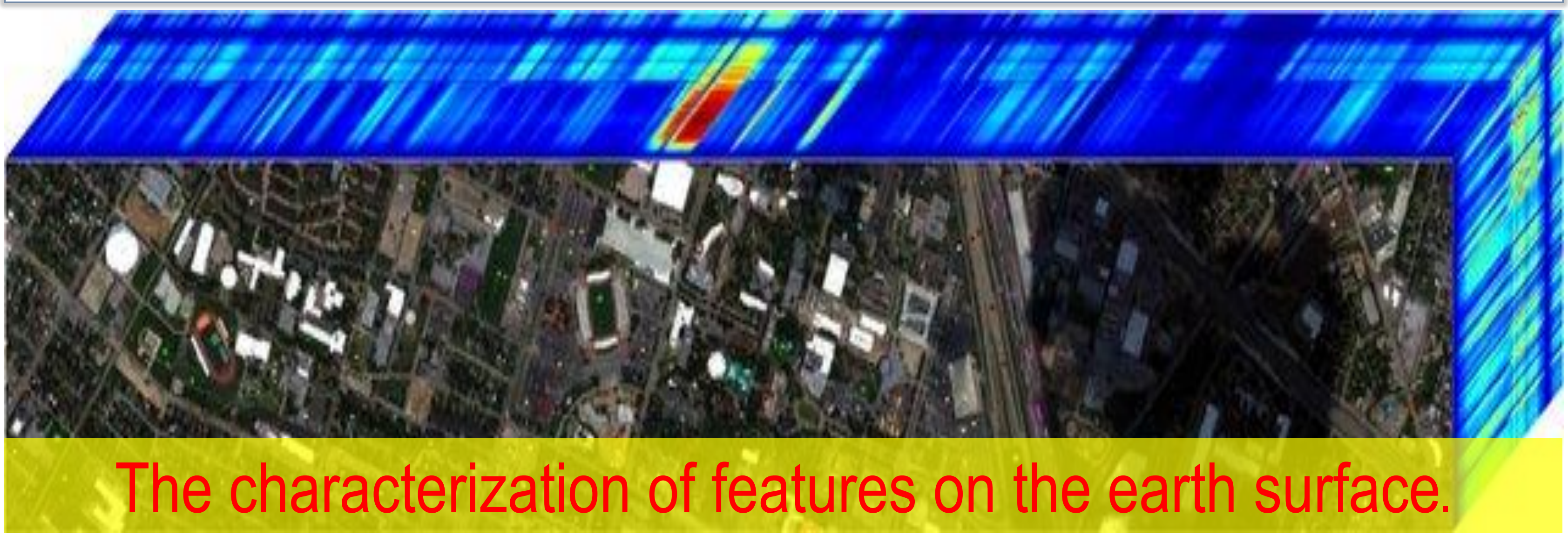
RS technology has significantly improved the Earth observation ability.



LandSat-1 | EOS AM-1 | MightySat II | EO-1 | ISS DESIS | GF-5 | TEMPO | JPSS-2

1972	1986	1999	2000	2000	2002	2008	2008	2018	2018	2018
LandSat-1	NOAA-10	EOS AM-1	MightySat2.1	EO-1	ADEOS-2	IMS-1	HJ-1A	ISS	HysIS	GF5
8	20	36	145	220	115	64	115	235	316	330
0.475-12.6	0.669-14.5	0.62-14.385	0.45-1.05	0.4-25	0.45-0.95	0.4-0.95	0.45-0.95	0.4-1.0	0.4-2.4	0.4-2.5

2019	2020	2021	2022	2022	2022	2024	2024	2024	2024
PRISMA	HyperScout2	GISAT-1	EnMAP	TEMPO	JPSS-2	FLEX	MTG-S1	METOP-SG A1	METOP-SG A1
240	45	150	232	666	1305	300	1720	3936	16921
0.4-2.5	0.4-14.0	0.9-2.5	0.42-2.45	0.29-0.74	3.9-15.3	0.5-0.78	4.6-14.3	0.270-2.385	3.62-15.5



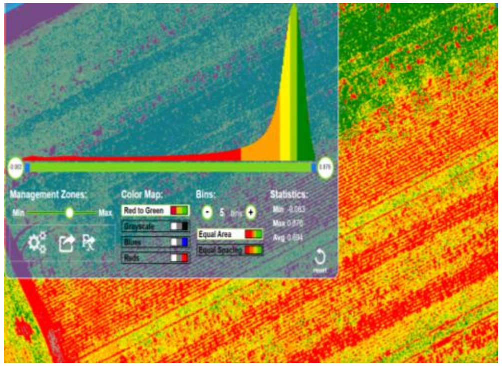
Applications of RS Images

CAPTAIN
COMPUTATIONAL AND PHOTOGRAMMETRIC VISION

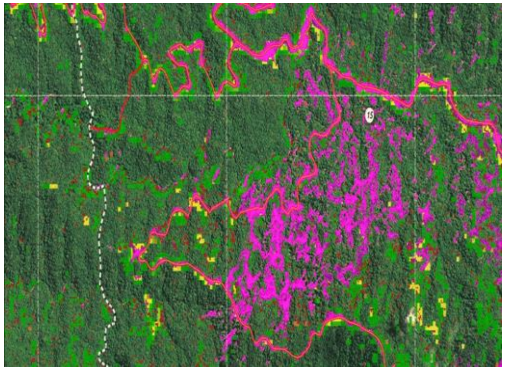
Interpretation of RS images plays important roles in many real-world applications

Information investigation

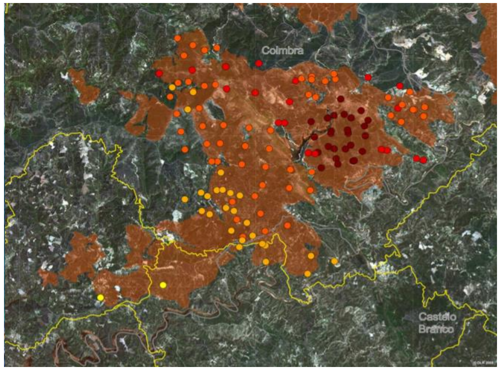
Smart city



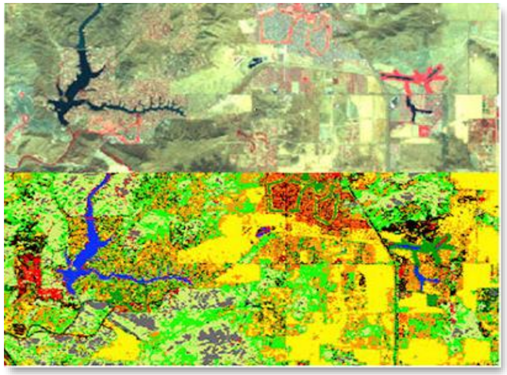
Precision agriculture



Environ. monitoring



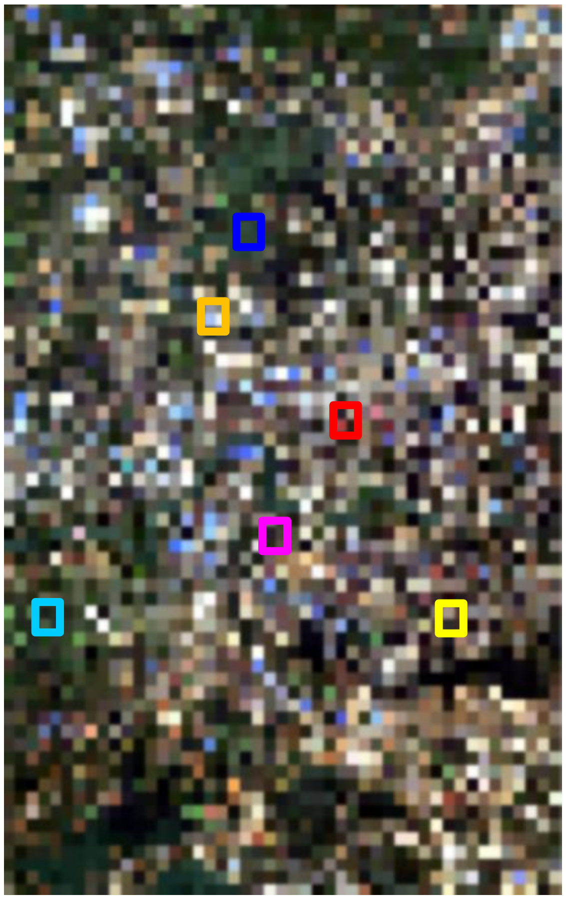
Disaster assessment

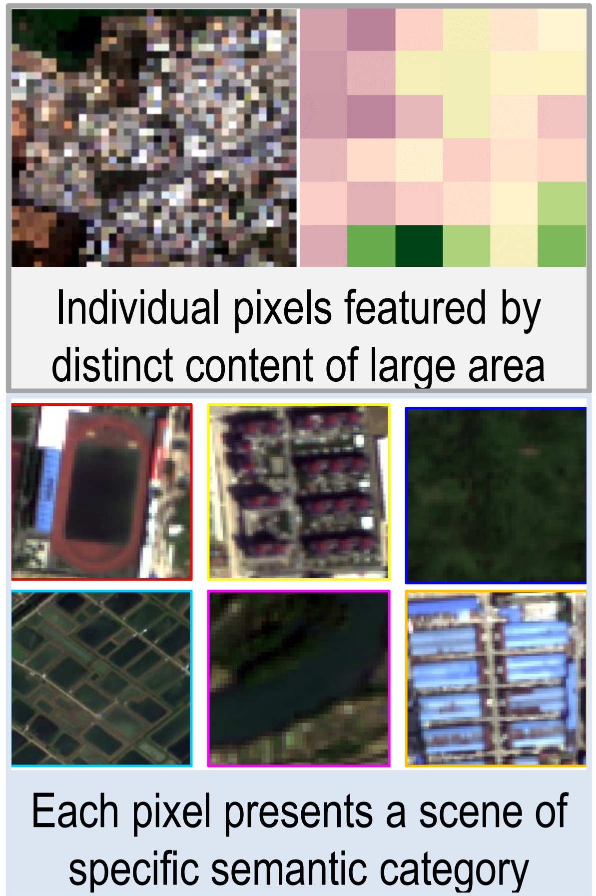


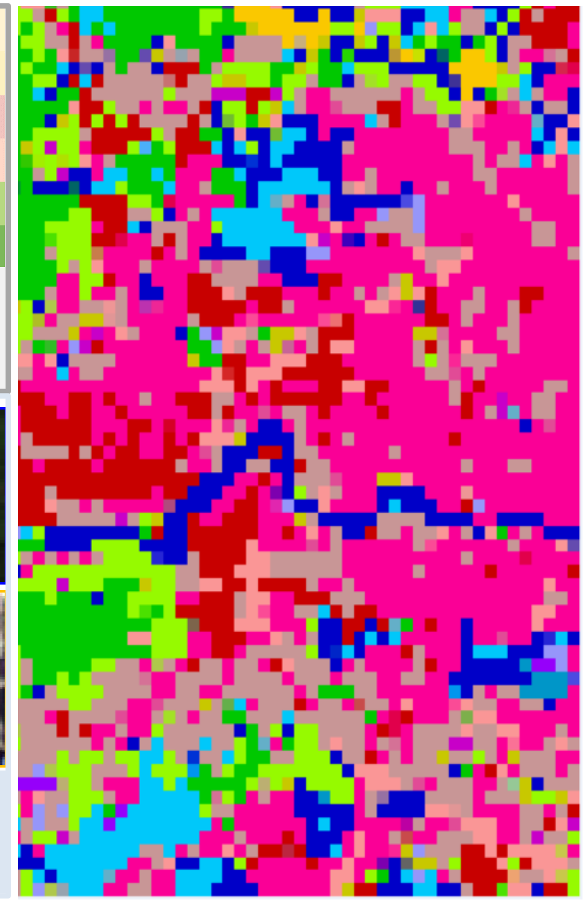
Land cover mapping

Pixel-wise Classification

Pixel-wise classification for low resolution aerial image classification



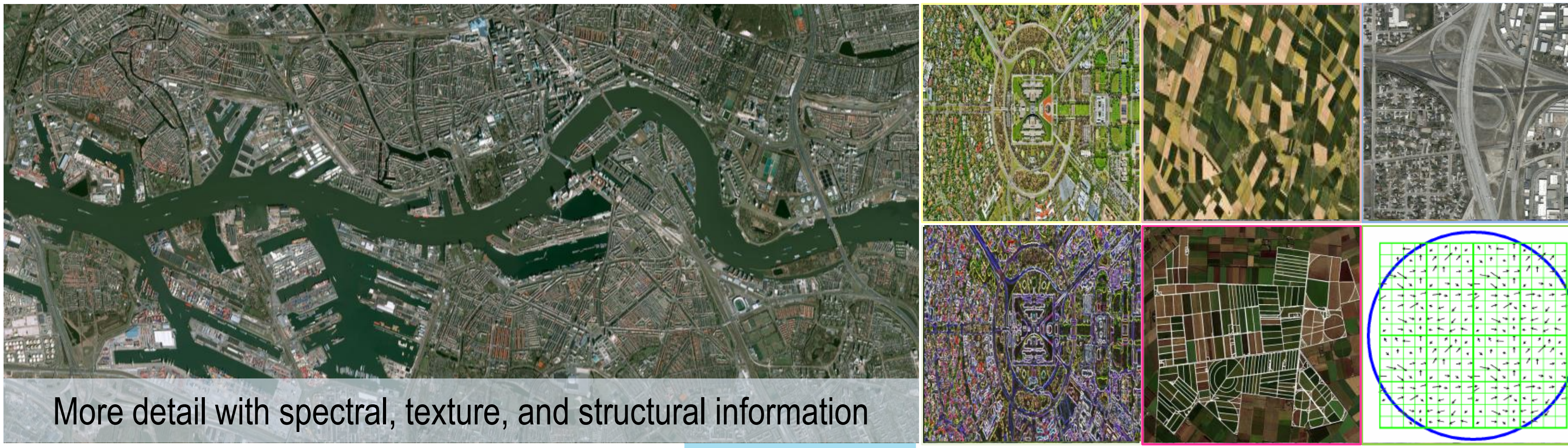




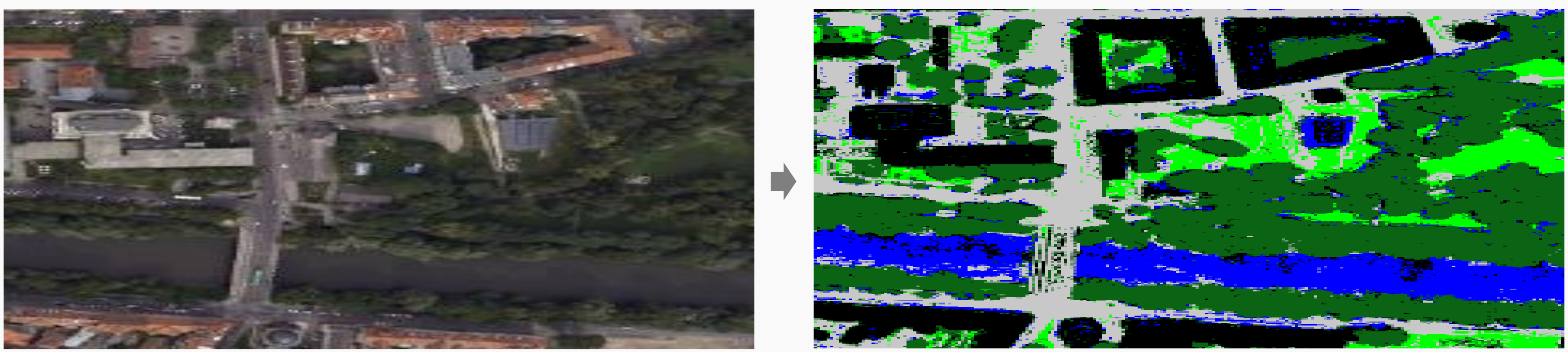
Low resolution image → Pixel-based classification

Pixel-wise Classification

Pixel-wise classification for high resolution aerial image classification



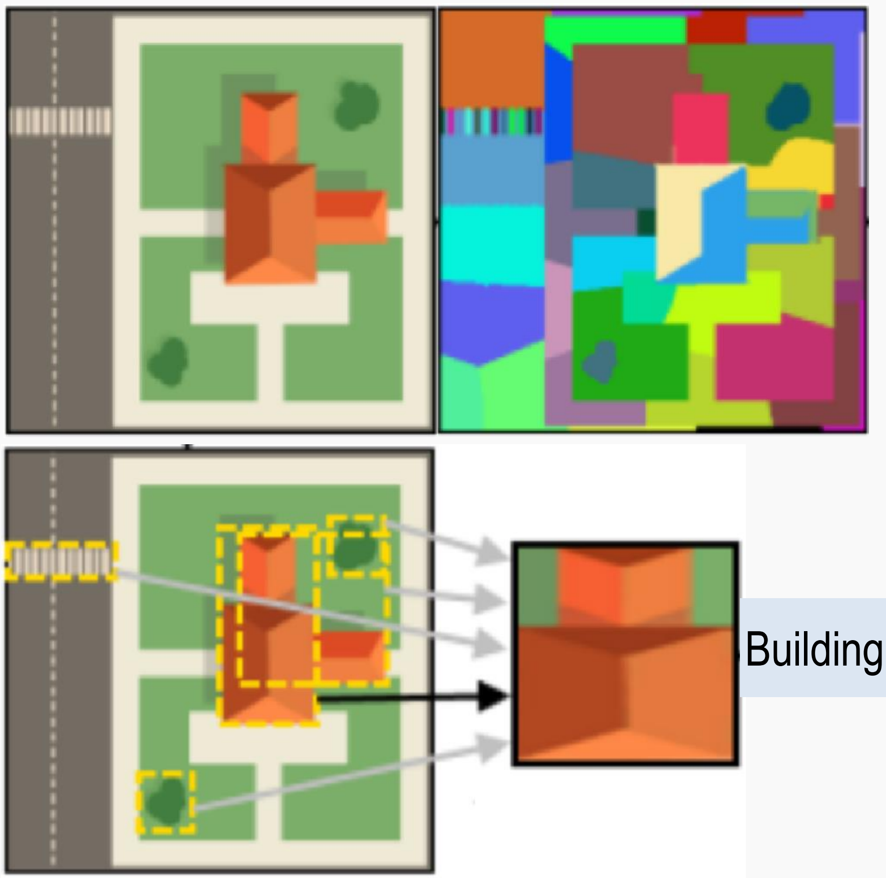
More detail with spectral, texture, and structural information



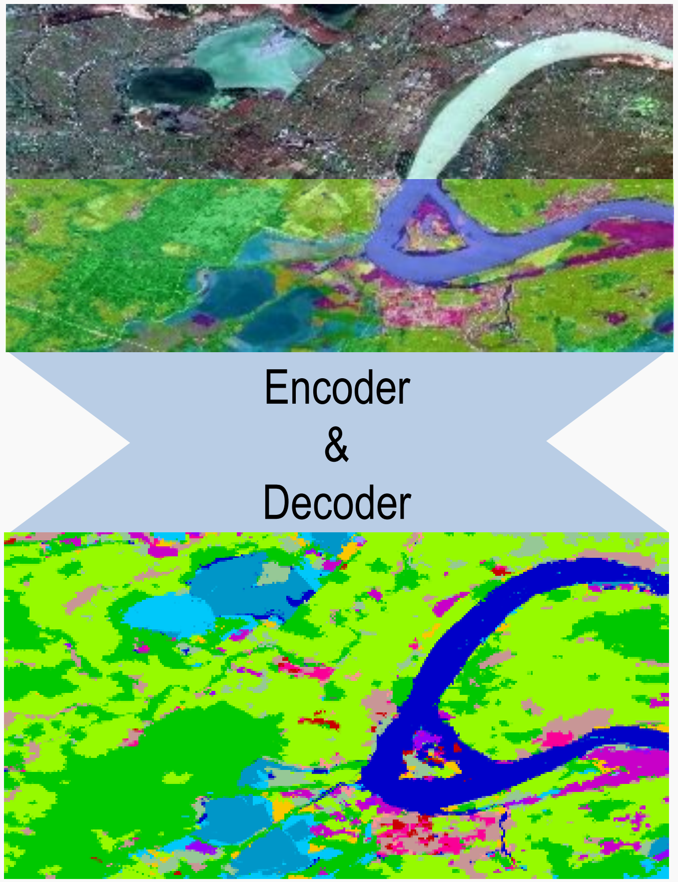
Messy result by pixel-based classification

Segmentation-based Classification

Complex modeling process from pixels to semantics



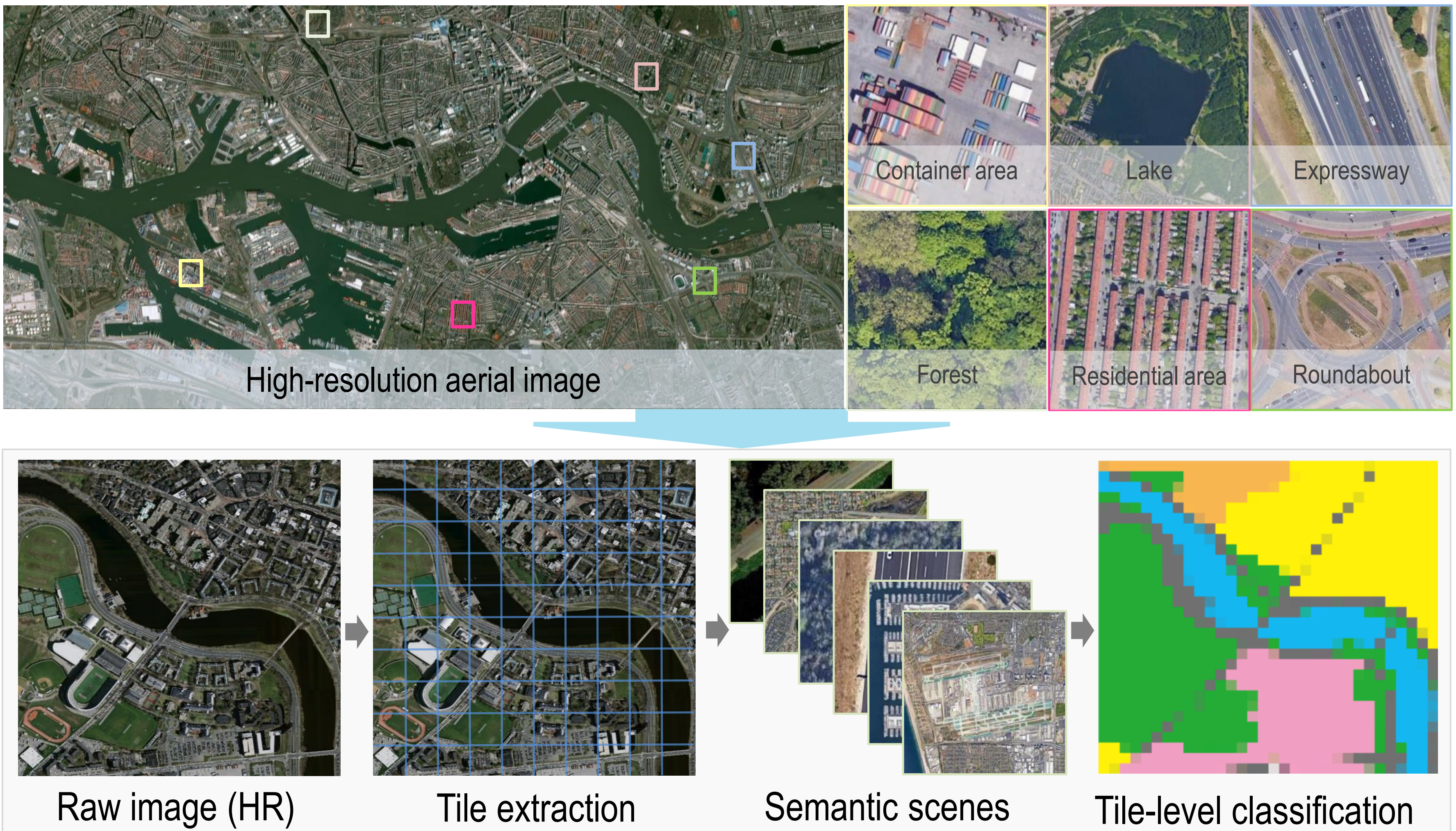
OBIA: relation modeling for over-segmented regions is required for semantic scene recognition



End2end segmentation: large-scale and well-annotated pixel-wise labels

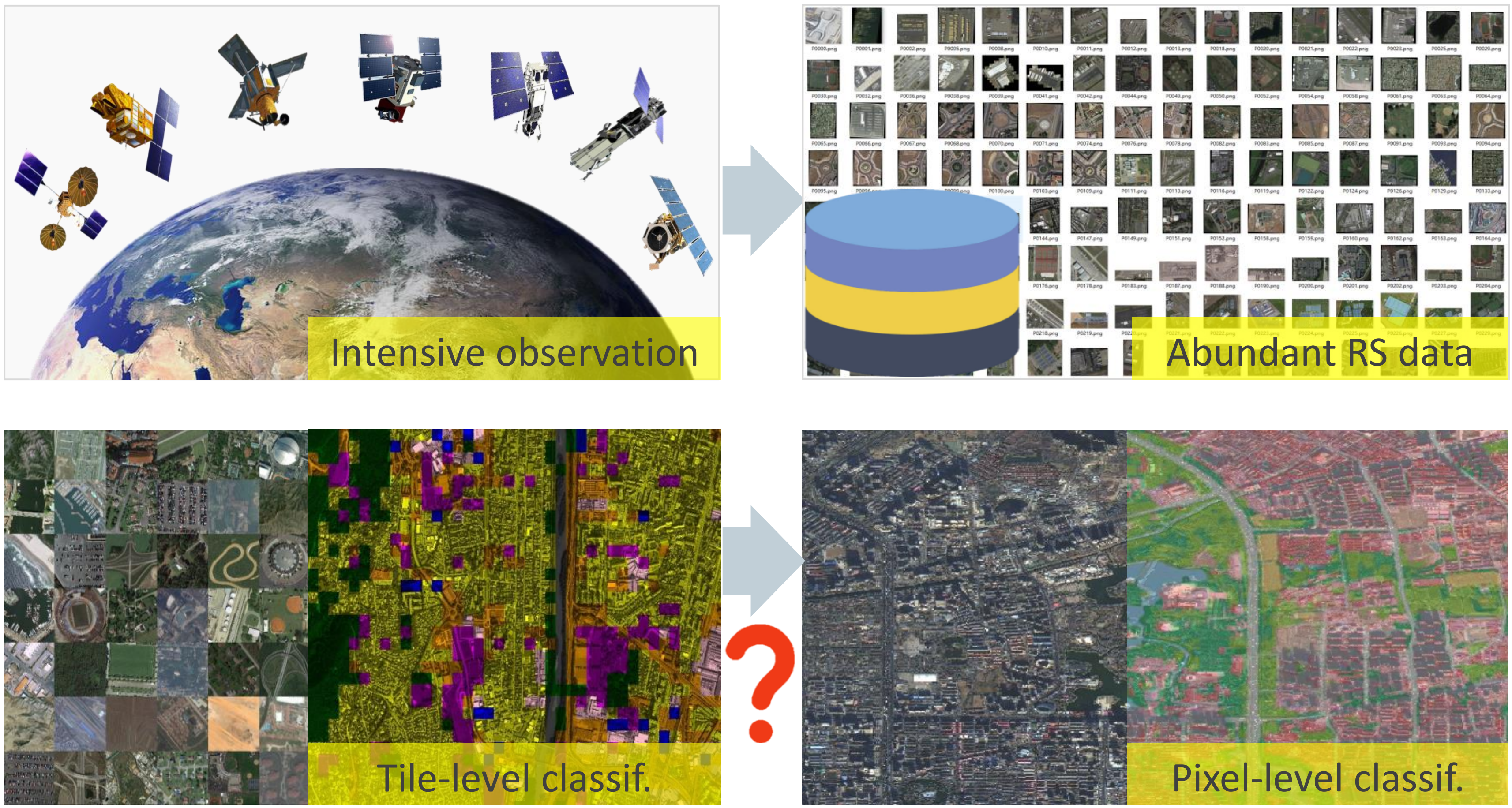
Tile-level Classification

Complicated features and components as a whole of scene



High-quality Classification

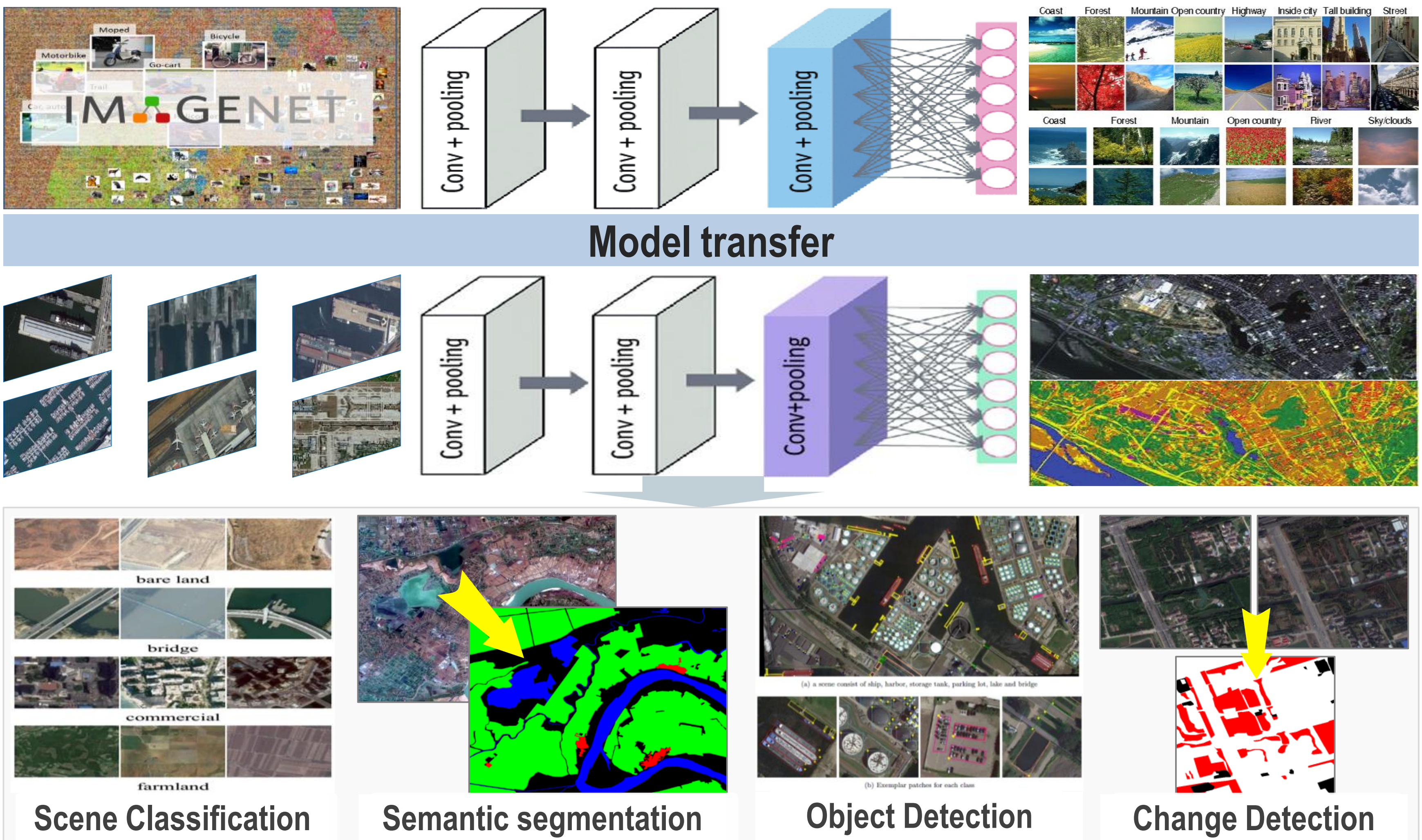
Current situation: Increasing demand for high-quality semantic classification



Coarse result by tile-level classification and high computational cost for pixel-wise classification

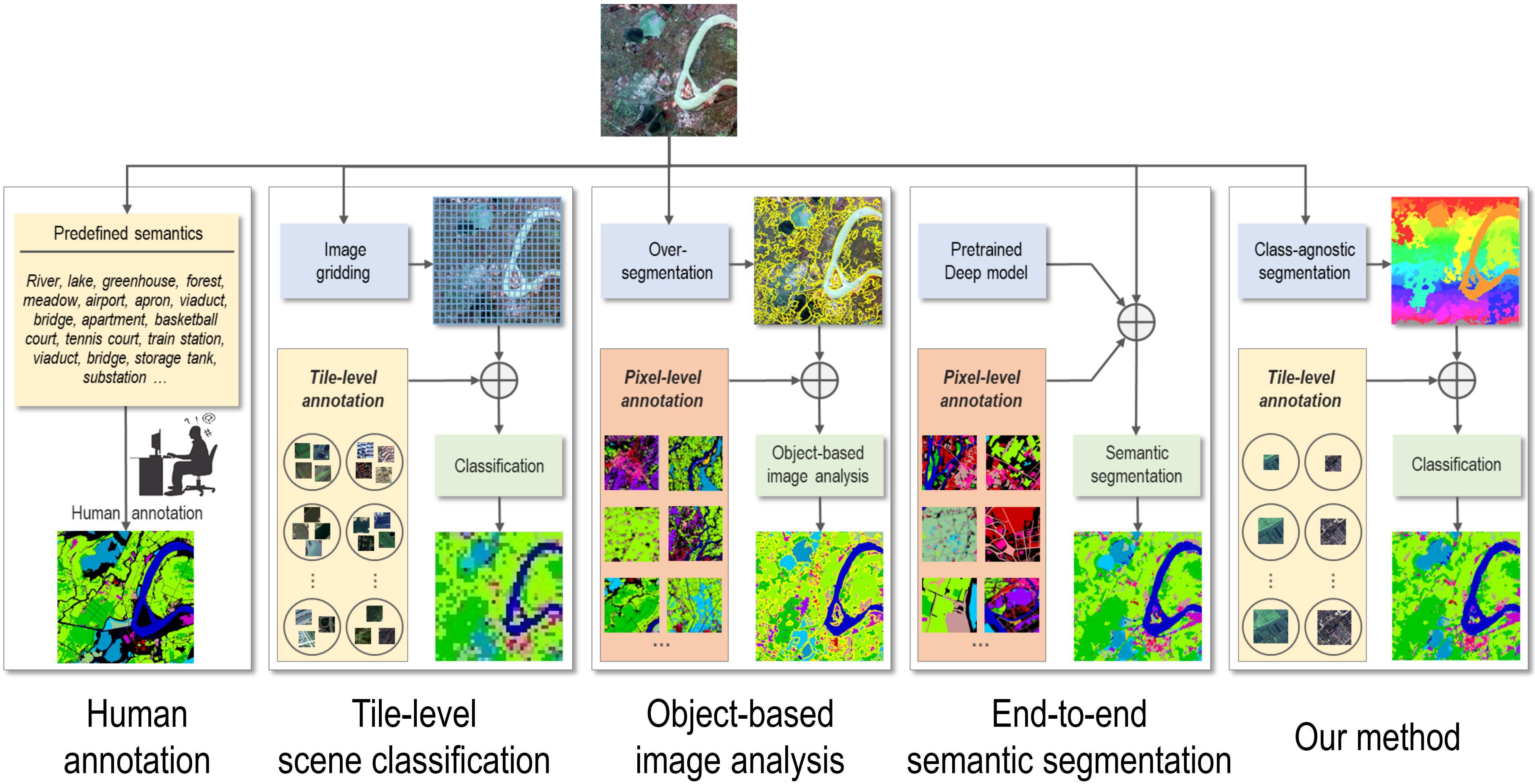
Model Adaption

Model optimization: parameters from natural images transferred for RS images



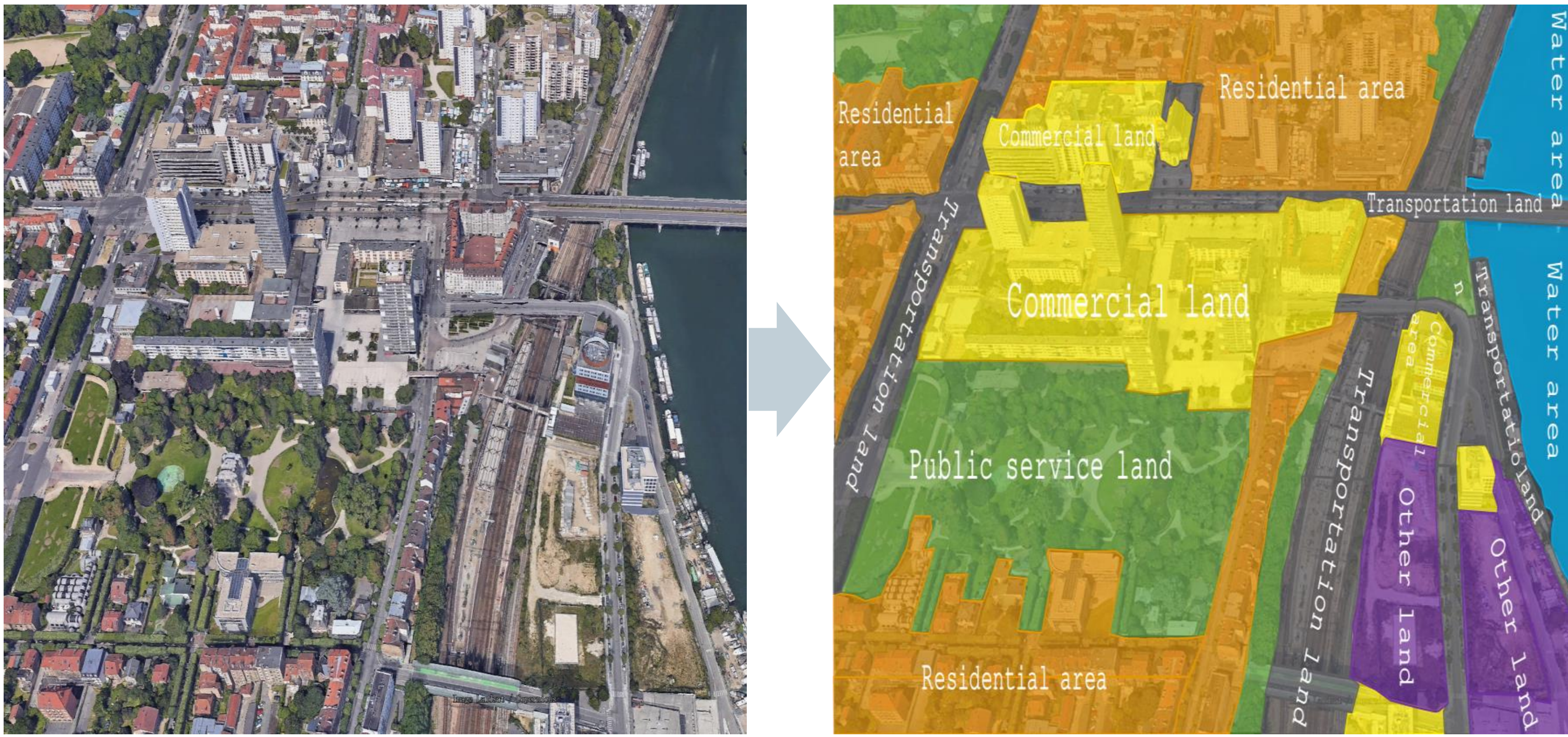
Aerial Scene Parsing

Target: A full-scene semantic structure interpretation of the aerial image content



Interpretation of RS Images

Image classification: transfer raw imagery data into semantic information



Motivation

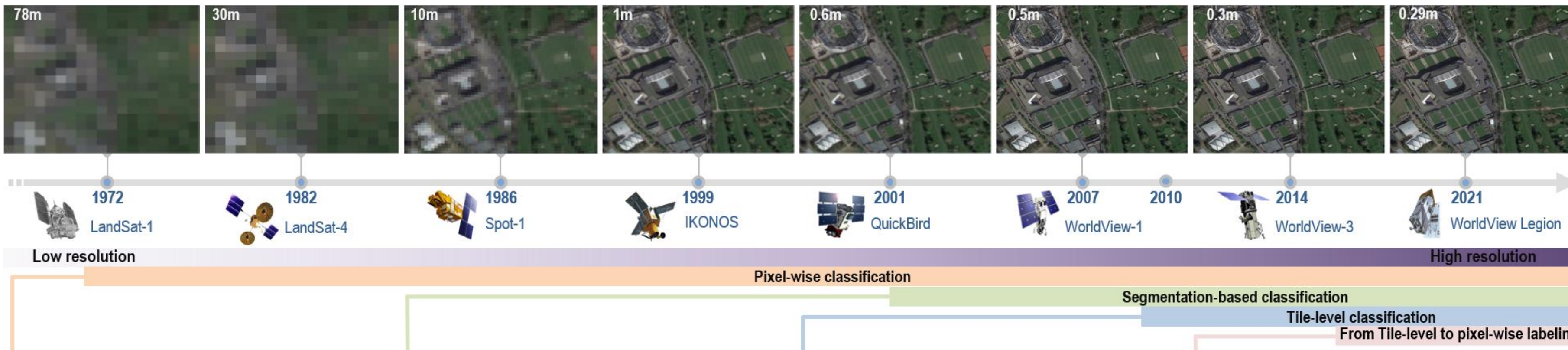
- **Bridging tile-level classification toward pixel-wise semantic labeling**
  - **Unification of tile-level scene classification and OBIA** for image interpretation
  - Emphasis on **tile-level interpretation** with high-level semantics while neglecting their **homogeneous components in pixel level**.
  - **Pixels** are no longer isolated units, of which semantics are highly related to their **contextual information** in high-resolution aerial images.
- **Weak generalization ability of interpretation methods**
  - **Potential of data-driven interpretation methods** remains to be further liberated and evaluated on large-scale available datasets.
  - **Insufficiency in learning and utilizing domain knowledge** from the relevant interpretation data and tasks.

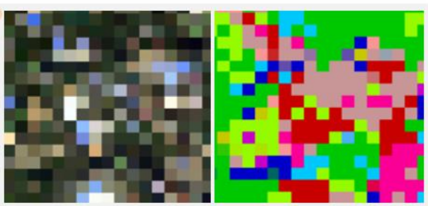
Outline

- Background
- Revisiting Aerial Image Interpretation
- Introduction to Million-AID
- Aerial Scene Classification: A New Benchmark
- Knowledge Transfer: From Tile-level to Pixel-level
- Conclusions

Road Map

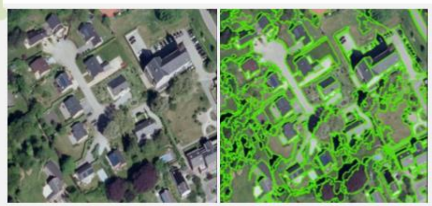
Interpretation prototypes develop with the improvement of aerial image quality





Spectral and textural attributes of pixels are mainly employed for semantic classification.

Spectral and textural description: @Haralick et al., 1973; @Swain et al., 1991; @Manju-nath et al., 1996; @Ojala et al., 2002; @Xia et al., 2010.
Statistical analysis: @Bruzzone et al., 1999; @Chen et al., 2008; @Li et al., 2010; @Li et al., 2011. @Camps-Valls et al., 2013; @Zhao et al., 2016.
Learning classifiers: @Kavzoglu et al., 2003; @Lee et al., 2007; Review @Lu et al., 2007; @Mountrakis et al., 2011; @Belgiu et al., 2016; @Xia et al., 2018.
Subpixel classification: @Wang, 1990; @Atkinson, 1997; Liu et al., 2005; @Somers et al., 2011;@Wang et al., 2017; @He et al., 2020; @Yu et al., 2021.

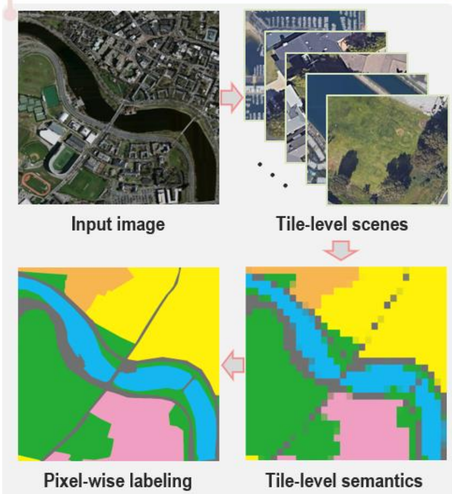


High-resolution image with richer spectral, textural, and structural detail for homogenous segmentation.

OBIA paradigm: @Blaschke et al., 2001; @Blaschke et al., 2014; Review @Hossain et al., 2019.
Spectral-spatial segmentation: @Cheng et al., 2001; @Kaur et al., 2011; @Martha et al., 2011; @Han et al., 2018; @ Tang et al., 2020; @ Shang et al., 2021.
Morphological methods: @Zhang et al., 2014; @Liu et al., 2015; @Yang et al., 2017; @Su, 2019. @Su et al., 2020; Review @Hossain et al., 2019; @Niu et al., 2021.
Deep learning segmentation: @Long et al., 20115; @Maggio et al., 2016; Review @Zhu et al., 2017; @Li et al., 2019; @Audebert et al., 2019; OCNN @Zhang et al., 2018; @Zhang et al., 2020; @Martins et al., 2020.

Complicated visual features and components as a whole for semantic scene recognition.

Hierarchical scene parsing: @Zhu et al., 2010.
Low-level and mid-level features: @Yang et al.,2010; @Chen et al., 2015; @Zhong et al., 2015; Review @Cheng et al., 2017; @Xia et al., 2017.
Deep learning schemes: Transfer learning @Hu et al., 2015; Feature fusion @Li et al., 2017; Chaib et al., 2017; Metric learning @Cheng et al., 2018; Attention mechanism @Wang et al., 2019; Architecture search @Ma et al., 2021; Patch classification @Sharma et al., 2017; @Paoletti et al., 2018; @Sharma et al., 2018; @Liu et al., 2020; Few-shot @Cheng et al., 2021; @Li et al., 2021; Datasets @Long et al., 2021.

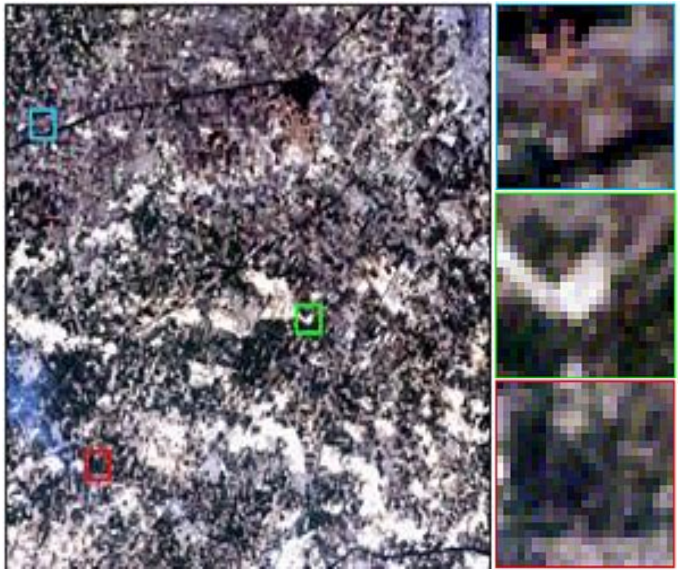


Full-scene semantic structure interpretation that bridges tile-level scene classification toward pixel-wise semantic labeling for high-resolution aerial images.

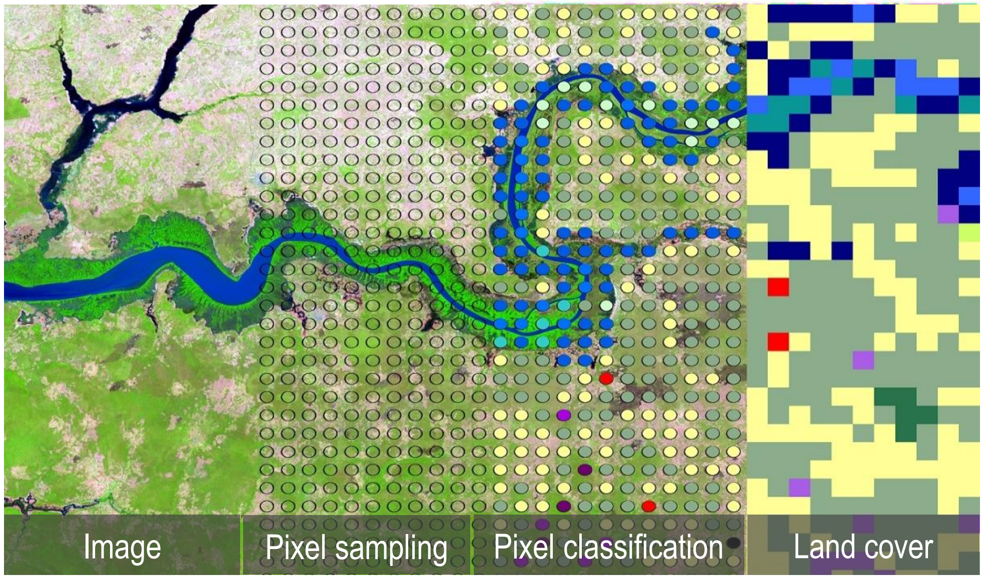
Pixel-wise Classification

Aerial images with low resolution

- Sizes of objects are smaller than the image resolution
- Spectral and texture attributes are mainly employed
- Pixel sampling and statistical analysis with content attributes



Low resolution image



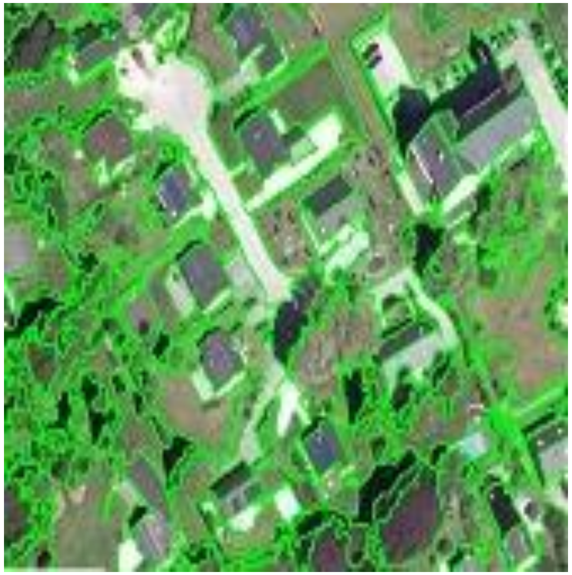
Pixel-wise image classification

Segmentation-based Analysis

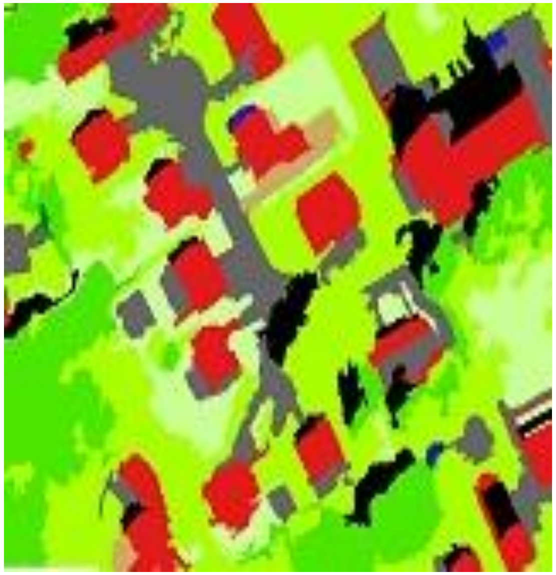
Object-based analysis

- Ground objects as basic units for semantic information identification
- Homogeneous segmentation by spectral, texture, and structural attributes
- lack semantic description, object relation modeling, scale challenge

Image with rich detail



Homogenous segments



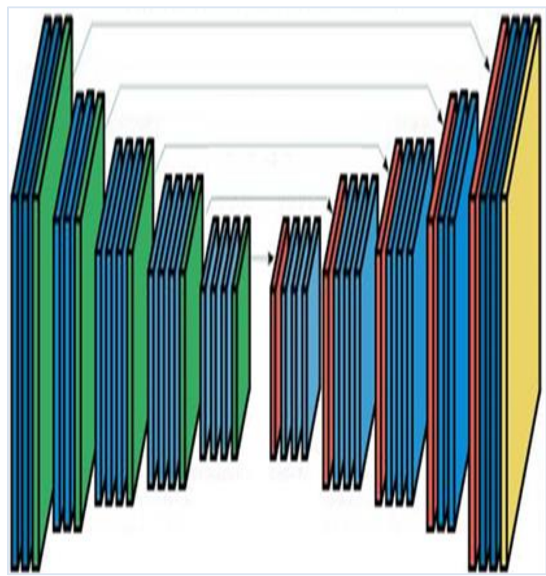
Object classification

Segmentation-based Analysis

End2end segmentation

- Simultaneously produce homogeneous segments and semantic classes
- Improved architectures and feature integration to advance accuracy
- Optimization with massive pixel labels, computational burden, generalization

High-resolution aerial image



Convolutional encoder-decoder

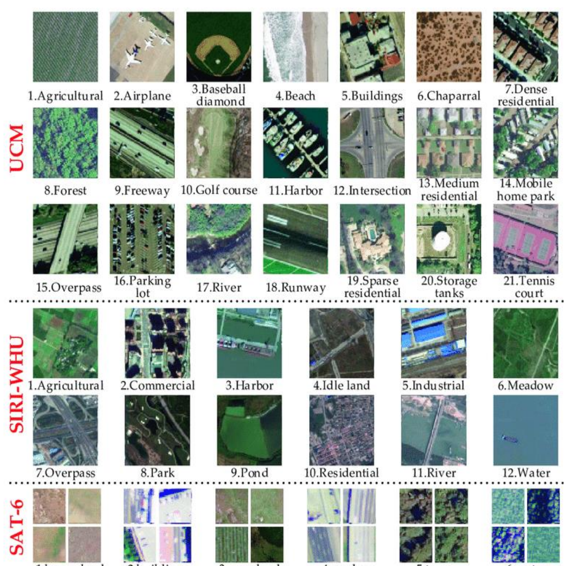
Semantic segmentation result

Tile-level Understanding

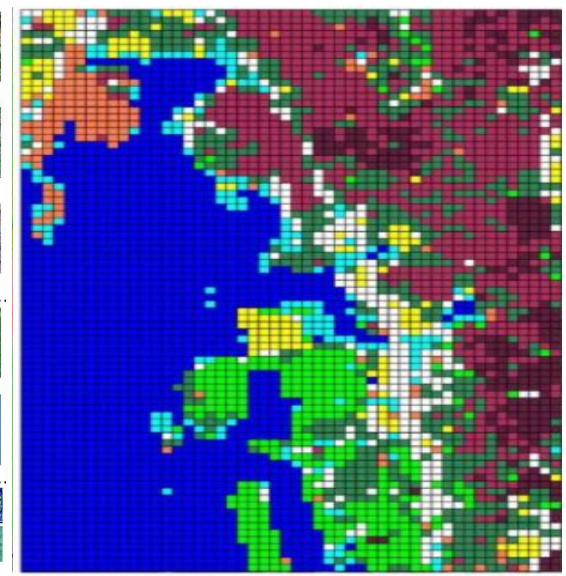
Scene recognition within local area

- Complicated features and content as a whole with high-level knowledge
- Scene representation from handcrafted to deep learning features
- Coarse interpretation result, accuracy saturation of existing datasets

Real-world complex content



Limited classes and scale



Course classification result

Analysis

Pixels are highly related to their neighbors in high resolution aerial images

Low resolution:
Isolated pixels as basic semantic units

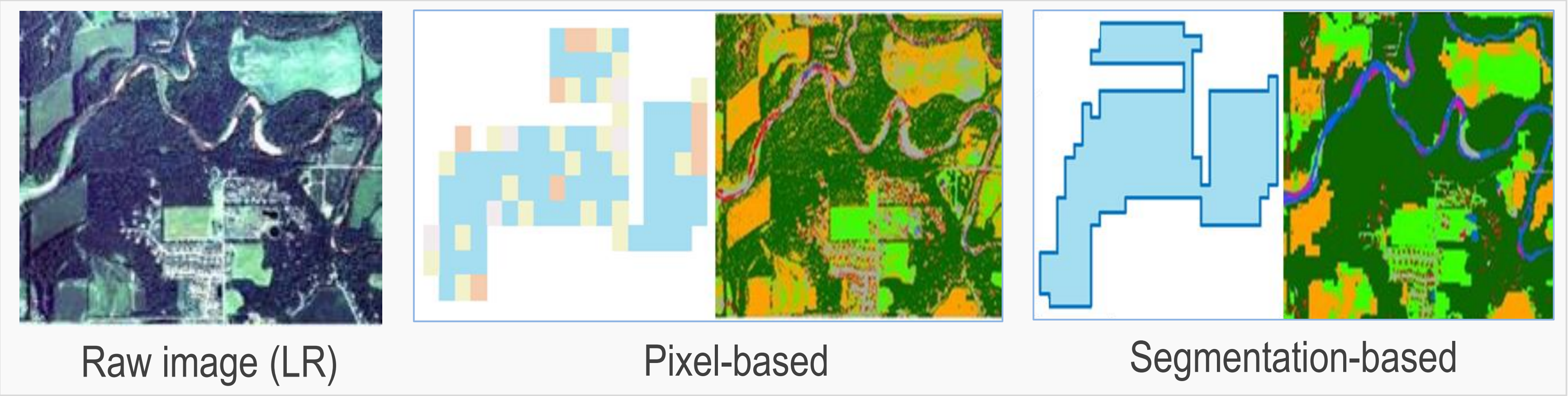
High resolution:
Pixels must be considered with contextual information

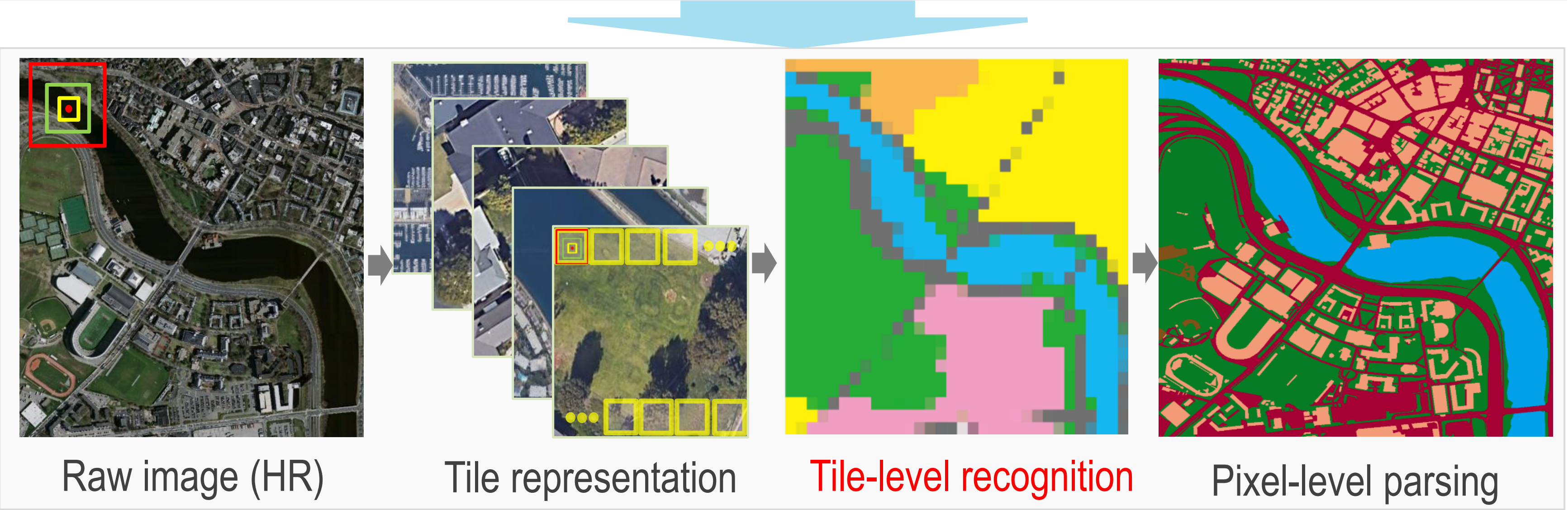
Enlarged areas of homogeneity

More rich detail with Noisy information

Analysis

Tile-level representation for high resolution aerial image classification





Outline

- Background
- Revisiting Aerial Image Interpretation
- **Introduction to Million-AID**
- Aerial Scene Classification: A New Benchmark
- Knowledge Transfer: From Tile-level to Pixel-level
- Conclusions

Tile-level Datasets

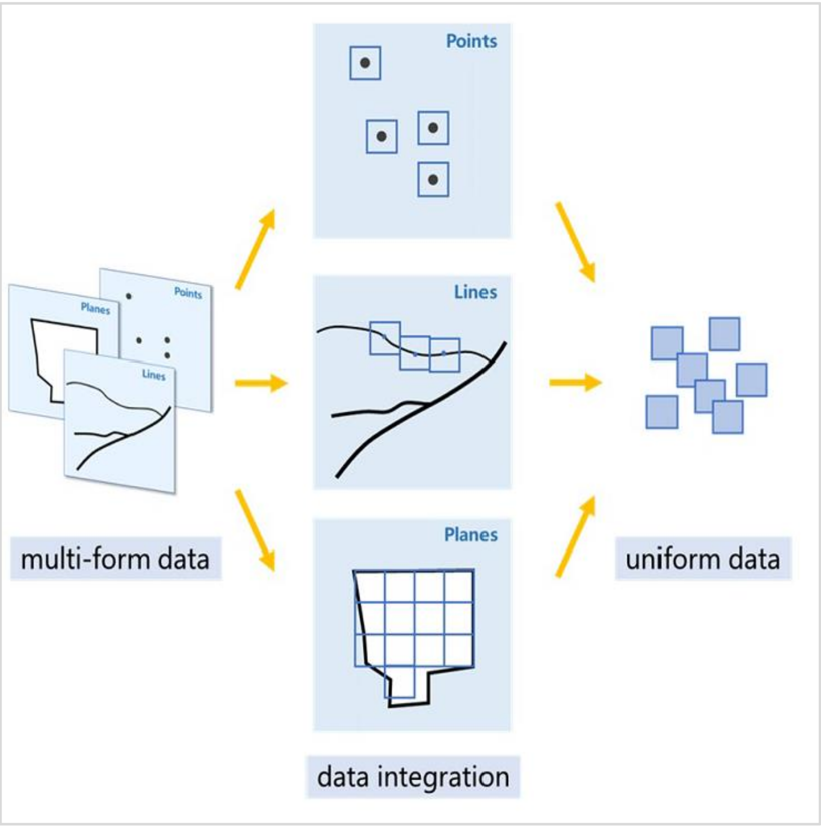
Aerial scene datasets

- **Small scale and poor diversity:** small number of categories and instances
- **Accuracy saturation:** lack standard evaluation benchmarks

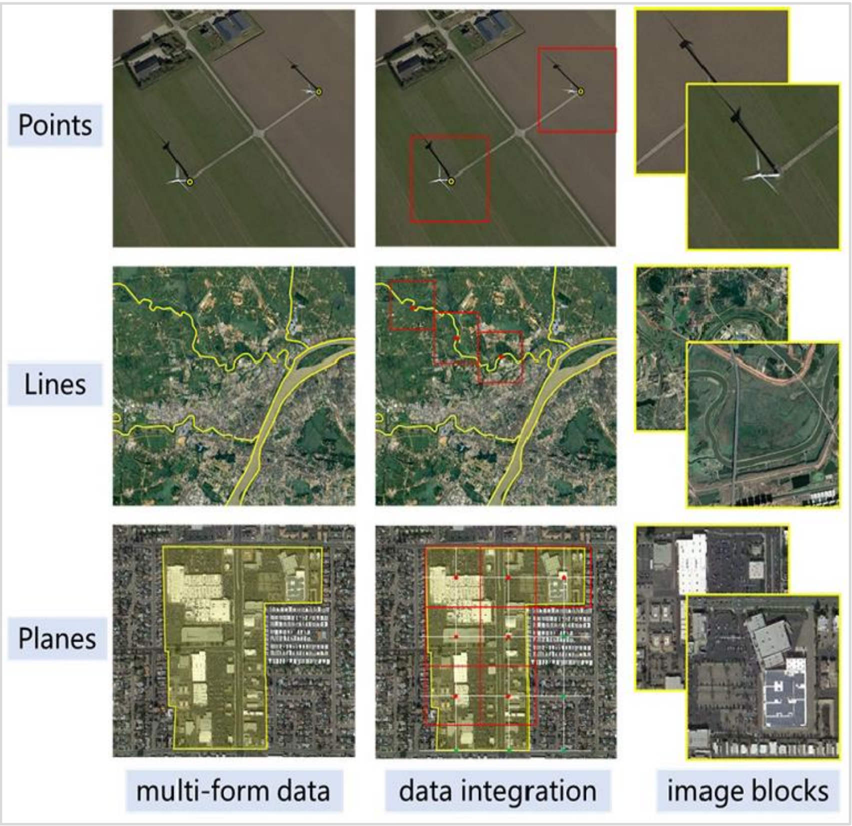
Dataset	#Cat.	#Images per cat.	#Images	Resolution (m)	Image size	GL/IT/SP	Year
UC-Merced	21	100	2,100	0.3	256×256	✗✗✗	2010
WHU-RS19	19	50 to 61	1,013	up to 0.5	600×600	✗✗✗	2012
RSSCN7	7	400	2,800	--	400×400	✗✗✗	2015
SAT-4	4	89,963 to 178,034	500,000	1 to 6	28×28	✗✗✗	2015
SAT-6	6	10,262 to 150,400	405,000	1 to 6	28×28	✗✗✗	2015
BCS	2	1,438	2,876	--	600×600	✗✗✗	2015
RSC11	11	~100	1,232	~0.2	512×512	✗✗✗	2016
SIRI-WHU	12	200	2,400	2	200×200	✗✗✗	2016
NWPU-RESISC45	45	700	31,500	0.2 to 30	256×256	✗✗✗	2016
AID	30	220 to 420	10,000	0.5 to 8	600×600	✗✗✗	2017
RSI-CB128	45	173 to 1,550	36,000	0.3 to 3	128×128	✗✗✗	2017
RSI-CB256	35	198 to 1,331	24,000	0.3 to 3	256×256	✗✗✗	2017
Planet-UAS	17	--	40,408	3 to 5	256×256	✓✓✓	2017
RSD46-WHU	46	500 to 3,000	117,000	0.5 to 2	256×256	✗✗✗	2017
MASATI	7	304 to 1,789	7,389	--	512×512	✗✗✗	2018
EuroSAT	10	2,000 to 3,000	27,000	10	64×64	✓✓✓	2018
PatternNet	38	800	30,400	0.06 to 4.7	256×256	✗✗✗	2018
fMoW	62	--	132,716	0.5	74×58 to 16184×16288	✓✓✓	2018
WiDS Datathon 2019	2	--	20,000	3	256×256	✗✗✗	2019
Optimal-31	31	60	1,860	--	256×256	✗✗✗	2019
BigEarthNet	43	328 to 217,119	590,326	10,20,60	20×20,60×60,120×120	✓✓✓	2019
CLRS	25	600	15,000	0.26 to 8.85	256×256	✗✗✗	2020
MLRSN	46	1,500 to 3,000	109,161	0.1 to 10	256×256	✗✗✗	2020

Dataset Construction

- Semi-automatic scene image collection: integration of public geographical features



Geographical point, line, and plane features



Scene image interpretation

Semantic Categories

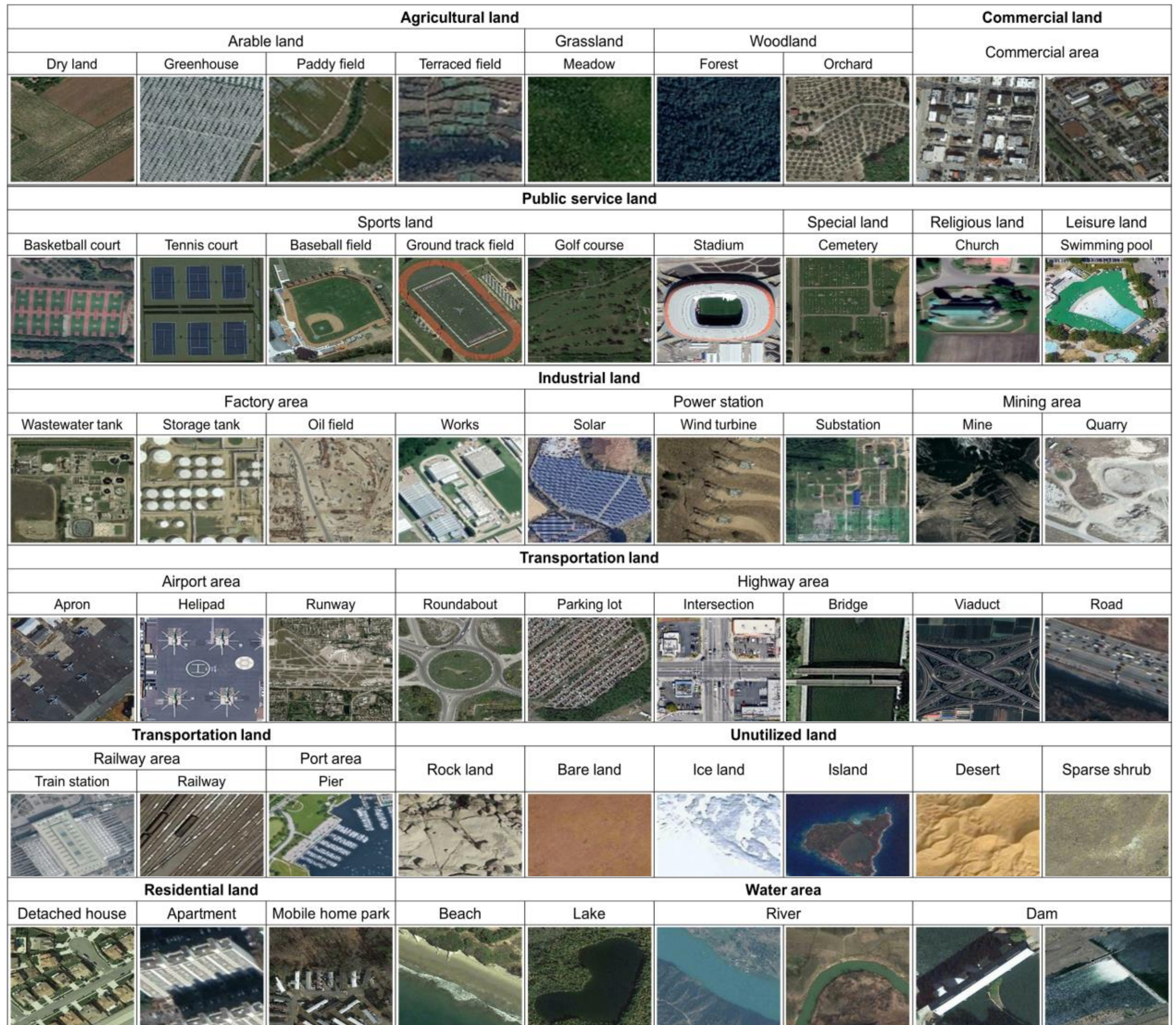
CAPTAIN
COMPUTATIONAL AND PHOTOGRAMMETRIC VISION

- Hierarchical category organization with land use standard

- First-level:
  8 categories
- Second-level:
  28 categories
- Third-level:
  37 categories

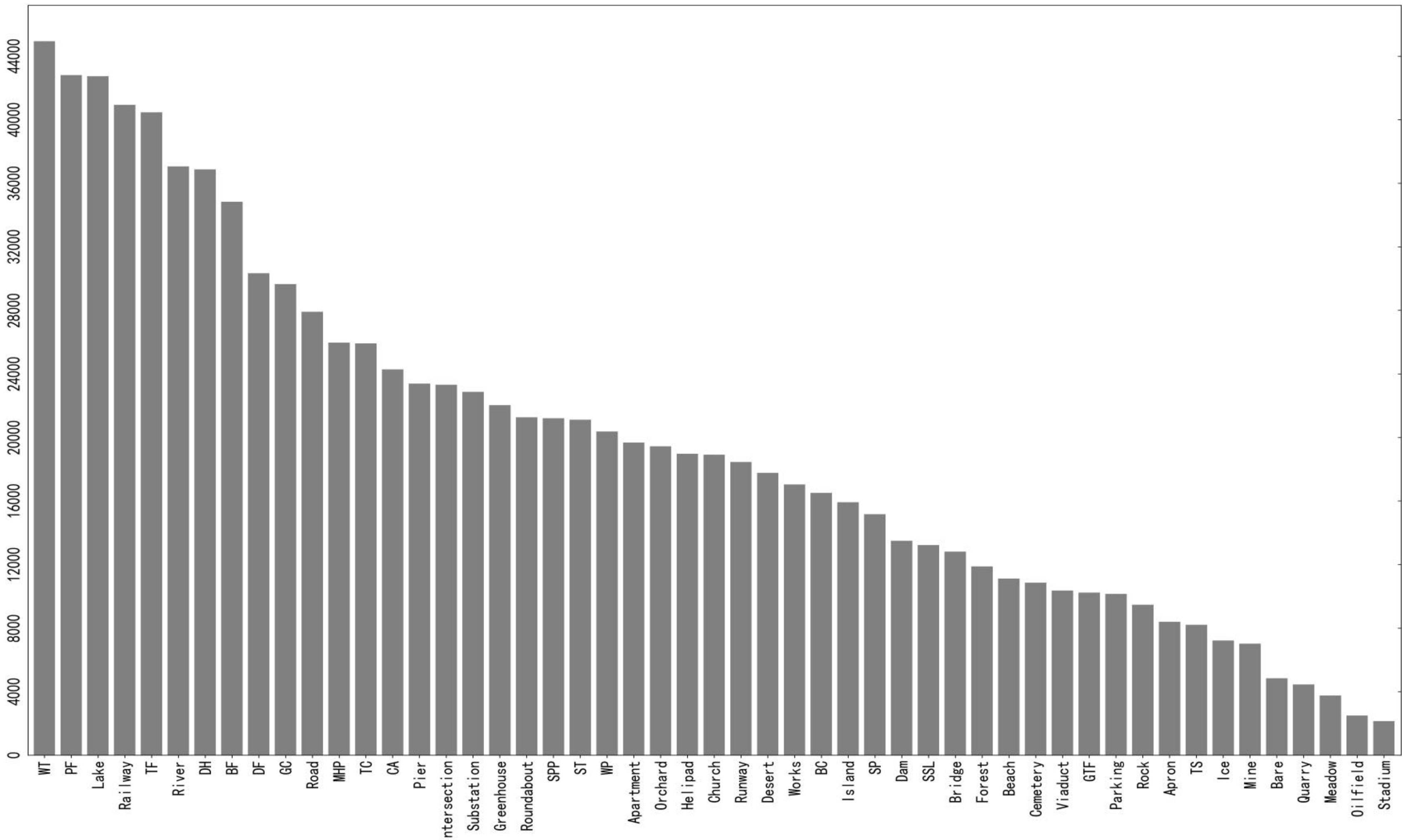
Multi-class classification:
51 fine-grained classes

Multi-label classification:
73 hierarchical classes



Dataset Scale

- Over 1M instances with unbalanced distribution: 2k to 45k samples in each category



Dataset Diversity

- Over 1M instances with unbalanced distribution: 2k to 45k samples in each category

Scene diversity

Apron

Baseball field

Similarity

Bridge

Viaduct

Scale variation

Storage tank

Wind turbine

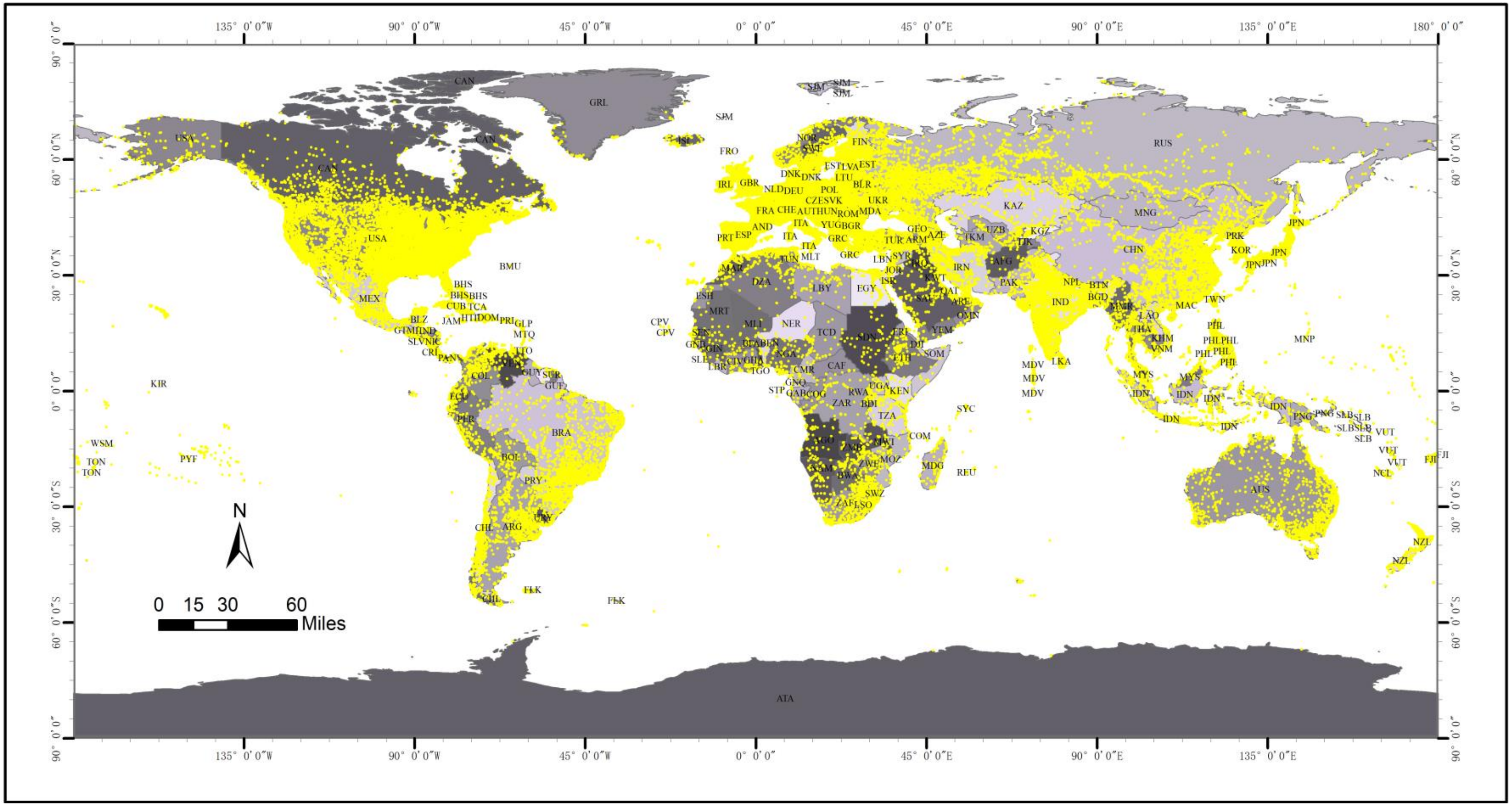
Complex.

Substation

Wastewater plant

Geographical Distribution

- Scenes around the world: intensively distributed within human inhabited areas



Outline

- Background
- Revisiting Aerial Image Interpretation
- Introduction to Million-AID
- **Aerial Scene Classification: A New Benchmark**
- Knowledge Transfer: From Tile-level to Pixel-level
- Conclusions

Baseline Setup

CAPTAIN
COMPUTATIONAL AND PHOTOGRAMMETRIC VISION

Unified implementation of CNN library

Model	#Layers	#Param.	Acc@1 (%)	Year
AlexNet	8	60M	56.52	2012
VGG16	16	138M	73.36	2014
GoogleNet	22	6.8M	69.78	2014
ResNet101	101	44M	77.37	2015
DenseNet121	121	8M	74.43	2017
DenseNet169	169	14M	75.60	2017

Benchmarking configurations

- **Multi-class scene classification:** 51 fine-grained scene categories
- **Multi-label scene classification:** 73 hierarchical semantic categories

Evaluation metrics

- **Multi-class scene classification:** overall accuracy (OA), average accuracy (AA), Kappa coefficient, mean of intersection-over-union (mIoU)
- **Multi-label scene classification:** per-class precision (CP), recall (CR), F1 (CF) and overall precision (OP), recall (OR), F1 (OF)

Results

- Results of Multi-class scene classification

Performance of Single-label Scene Classification with different CNN models

Metric	AlexNet	VGG16	GoogleNet	ResNet101	DenseNet121	DenseNet169
OA	67.53	77.47	77.37	77.36	79.04	78.99
AA	63.18	74.58	74.86	74.58	76.67	76.67
Kappa	66.61	76.84	76.73	76.73	78.46	78.46

- Results on different datasets with our framework

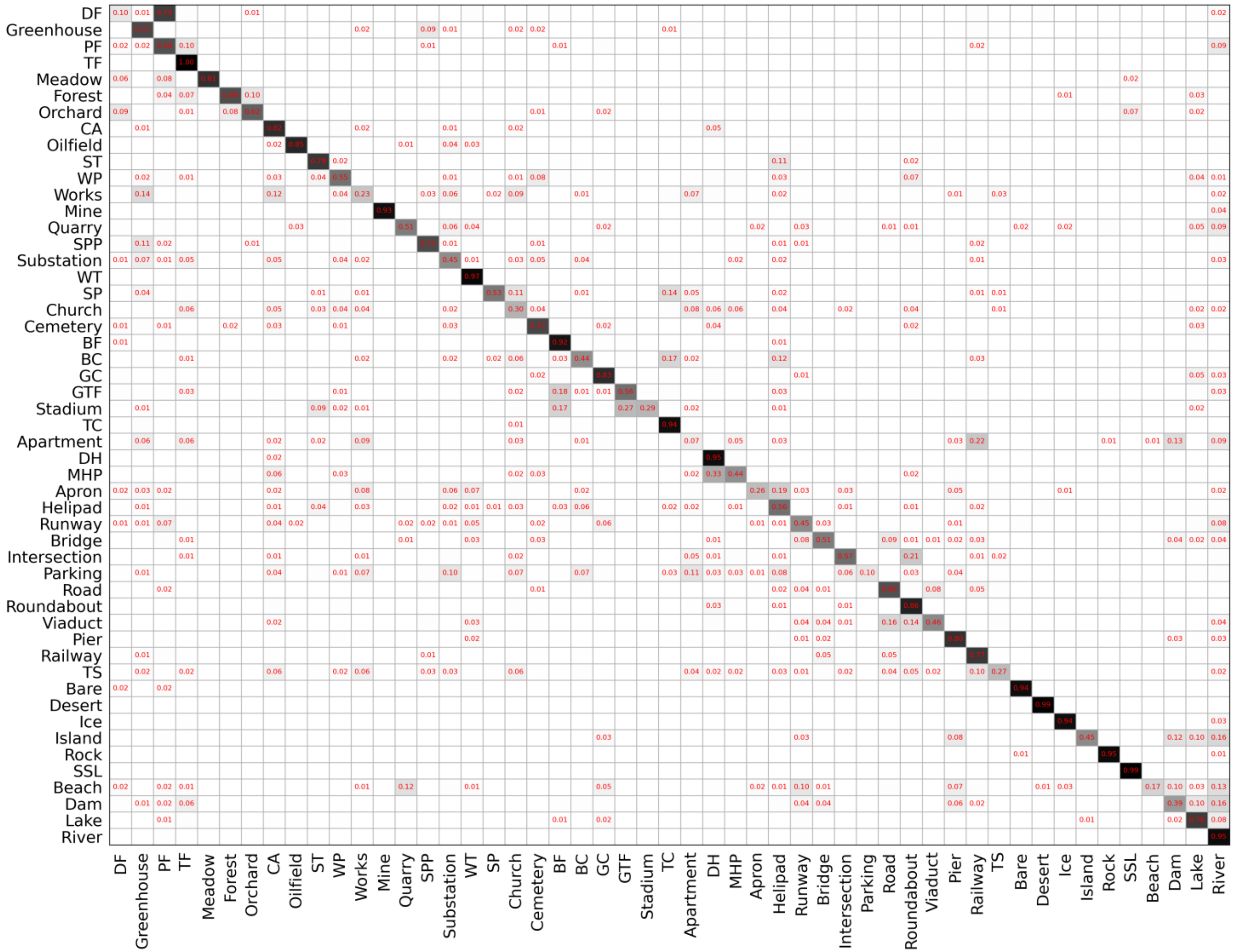
OA Comparison Among Different Datasets

Dataset	AlexNet	VGG16	GoogleNet
AID	86.86	86.59	83.44
AID*	88.79	93.72	92.24
NWPU-RESISC45	85.16	90.36	86.02
NWPU-RESISC45*	87.19	92.76	91.71
Million-AID	67.53	77.47	77.37

** Results using our implemented CNN framework,*

Confusion Matrix

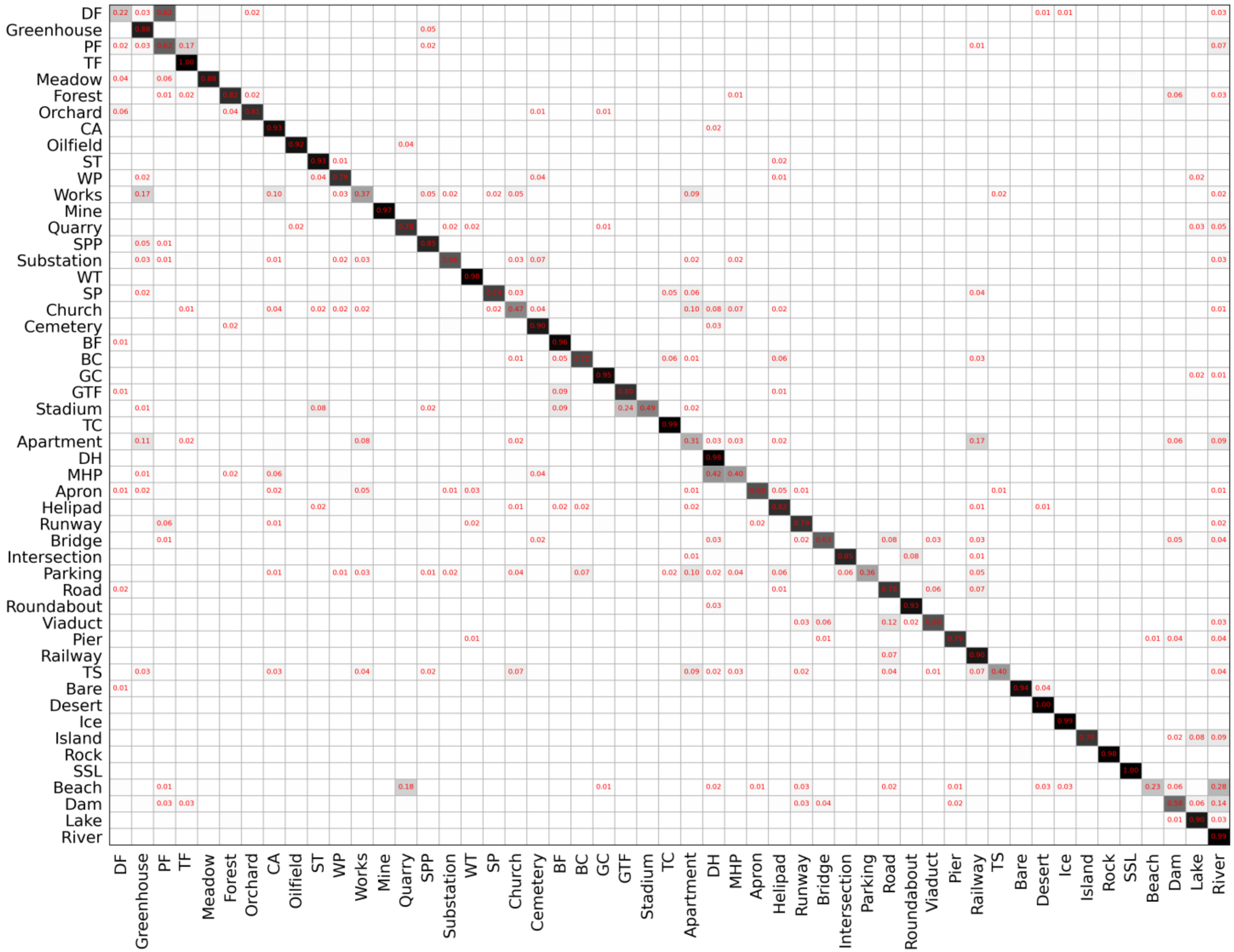
- Shallow network: AlexNet



Confusion Matrix

CAPTAIN
COMPUTATIONAL AND PHOTOGRAMMETRIC VISION

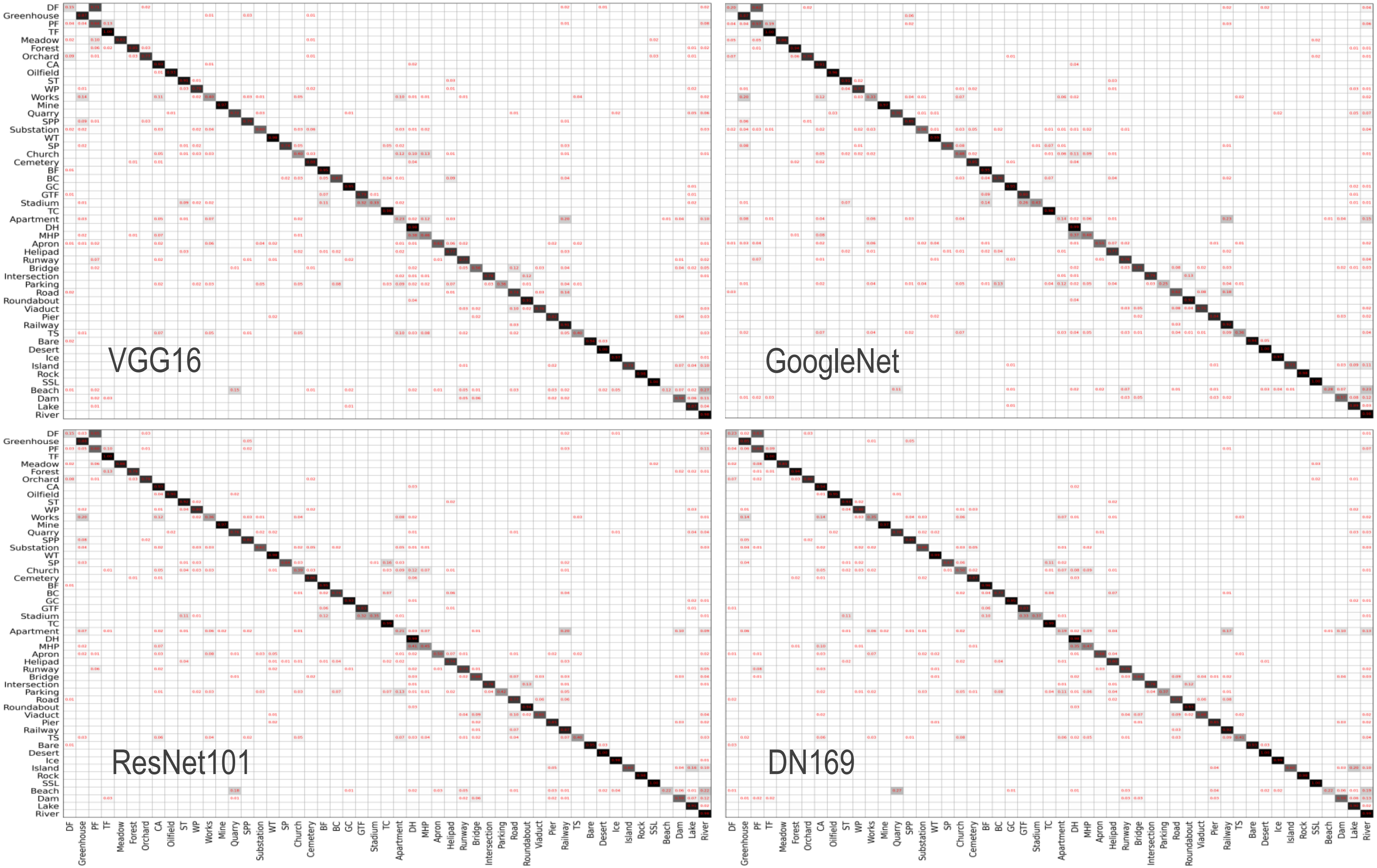
- Deep network: DenseNet121



Confusion Matrix

CAPTAIN
COMPUTATIONAL AND PHOTOGRAMMETRIC VISION

- More results



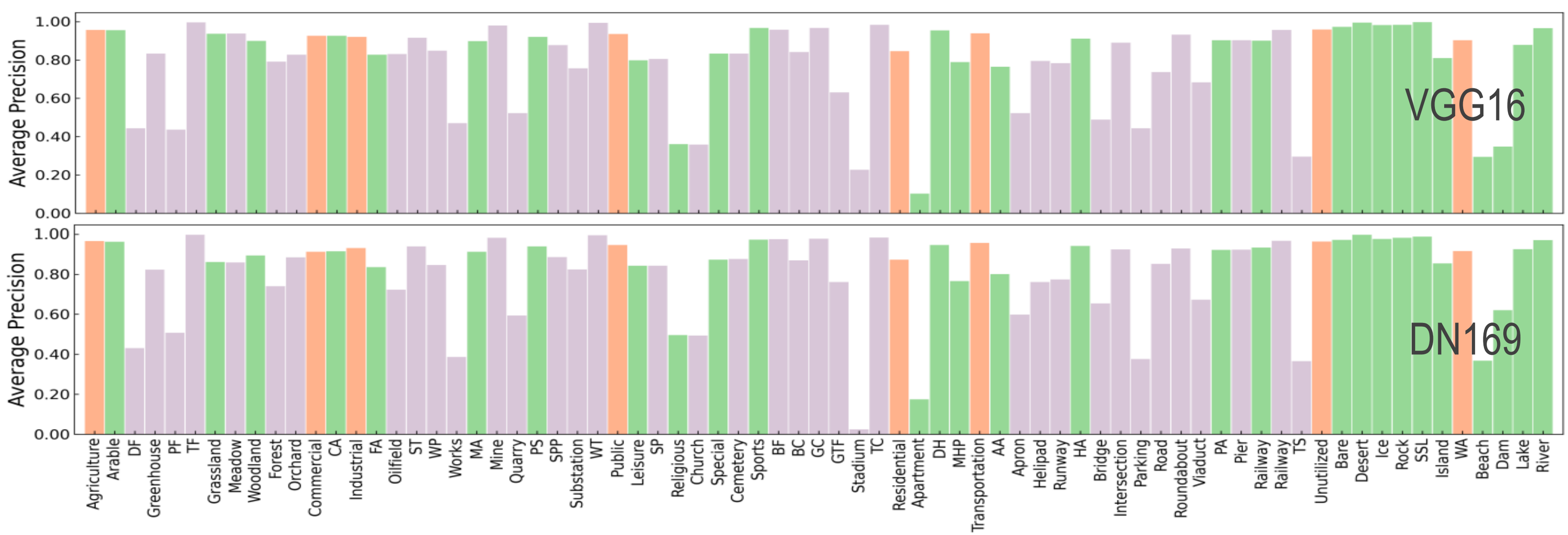
Results

- Results of Multi-label scene classification

Performance of Multi-label Scene Classification with different CNN models

Model	$\tau = 0.5$						$\tau = 0.75$						mAP
	CP	CR	CF1	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1	
AlexNet	71.45	48.19	57.56	76.19	62.84	68.87	78.89	38.51	51.76	85.65	53.03	65.50	61.76
VGG16	82.26	62.20	70.84	86.98	75.31	80.72	84.61	54.29	66.14	91.70	69.37	78.99	79.13
GoogleNet	51.79	33.99	41.04	88.50	59.47	71.14	50.99	23.76	32.42	94.90	47.02	62.89	60.03
ResNet101	79.38	59.67	68.13	88.74	77.31	82.63	76.83	51.56	61.71	93.05	70.93	80.50	80.42
DenseNet121	79.09	56.21	65.71	89.74	75.10	81.77	76.36	47.75	58.76	94.20	67.72	78.79	78.94
DenseNet169	78.54	61.92	69.24	88.50	78.55	83.23	78.52	55.10	64.76	92.66	73.10	81.72	80.99

- Challenging hierarchical multi-label classification

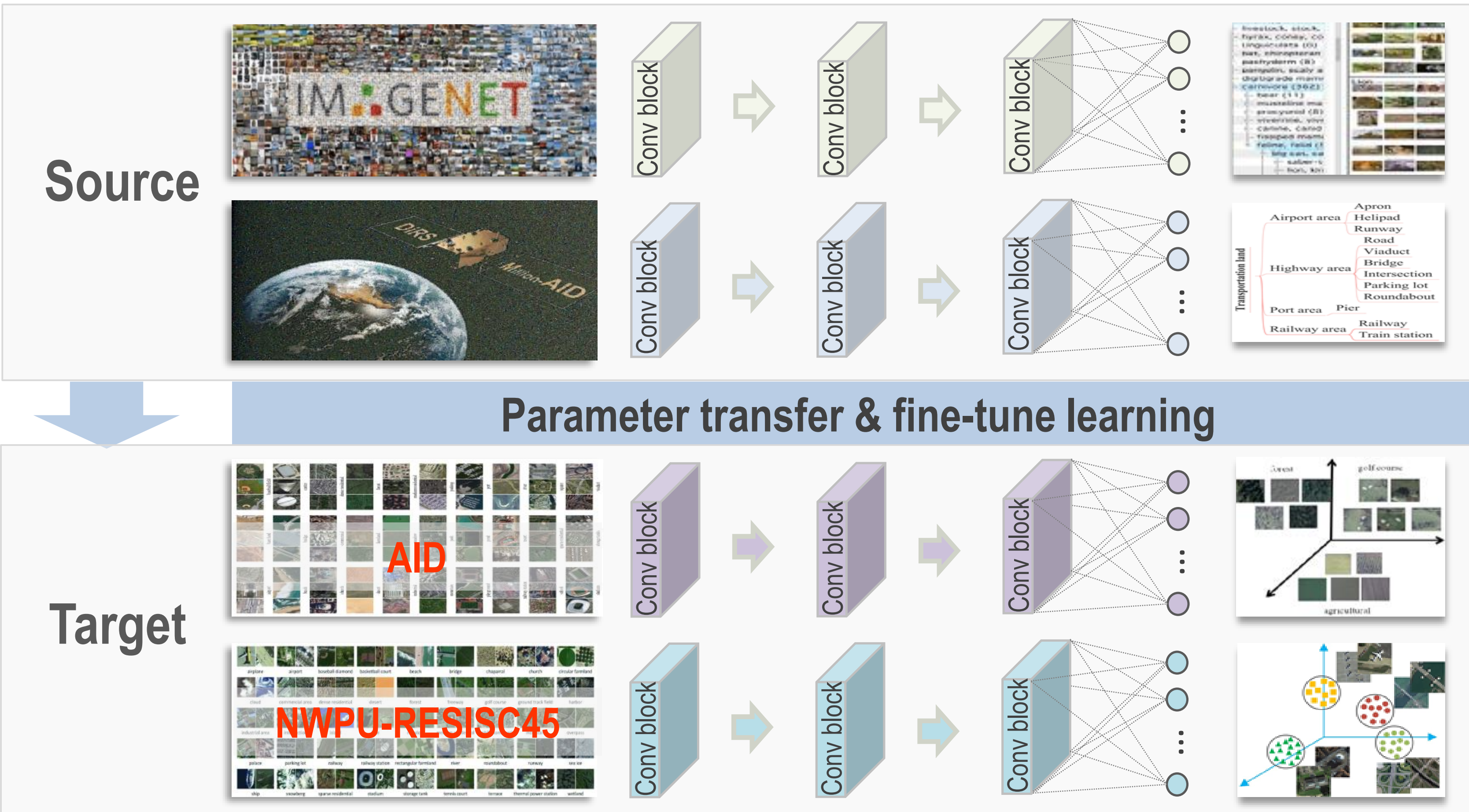


Outline

- Background
- Revisiting Aerial Image Interpretation
- Introduction to Million-AID
- Aerial Scene Classification: A New Benchmark
- **Knowledge Transfer: From Tile-level to Pixel-level**
- Conclusions

Scene Recognition

- Transfer Knowledge from ImageNet and Million-AID for scene recognition



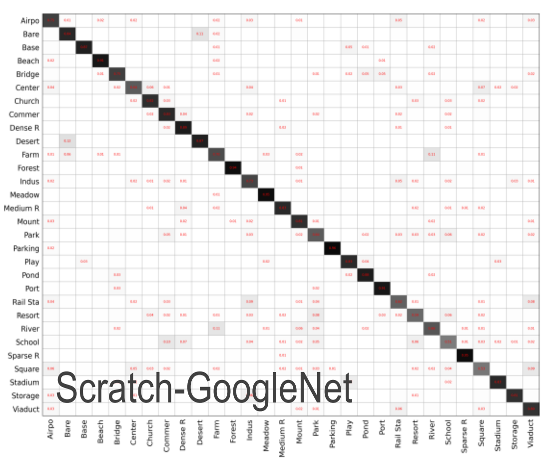
Results on AID

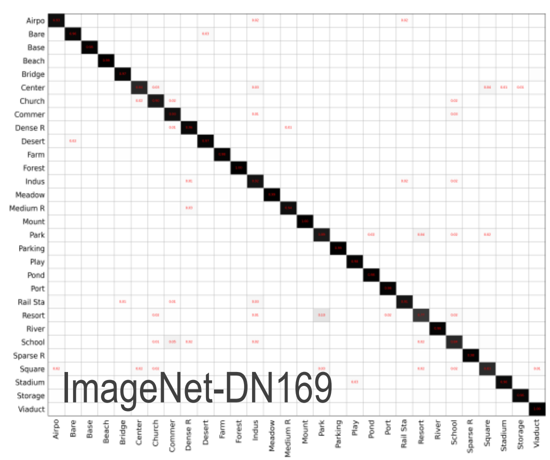
Accuracy comparison

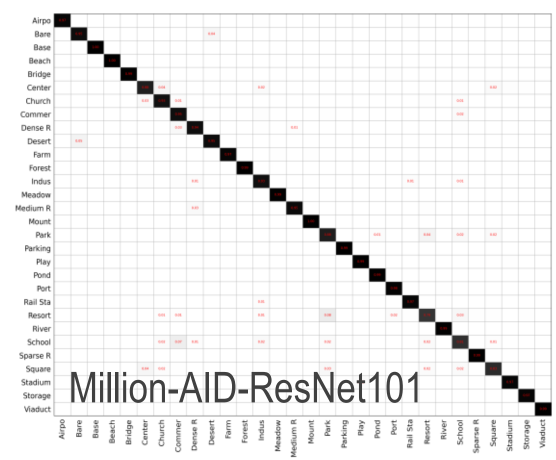
Classification accuracy (%) on AID dataset using different initialization schemes

Metric	Pretrain dataset	AlexNet	VGG16	GoogleNet	ResNet101	DenseNet121	DenseNet169
OA	W/O	33.47 ± 2.15	72.18 ± 0.49	79.05 ± 0.89	49.46 ± 2.07	58.02 ± 0.74	59.16 ± 0.52
	ImageNet	88.79 ± 0.40	93.72 ± 0.21	92.24 ± 0.21	94.52 ± 0.25	94.68 ± 0.19	94.76 ± 0.21
	Million-AID	**90.70 ± 0.43**	**95.33 ± 0.28**	**94.55 ± 0.23**	**95.40 ± 0.19**	**95.22 ± 0.26**	**95.24 ± 0.35**
AA	W/O	33.85 ± 2.35	72.16 ± 0.54	78.88 ± 0.88	49.29 ± 2.06	57.88 ± 0.73	59.04 ± 0.51
	ImageNet	88.52 ± 0.39	93.38 ± 0.22	91.78 ± 0.23	94.18 ± 0.29	94.39 ± 0.21	94.44 ± 0.22
	Million-AID	**90.46 ± 0.45**	**95.14 ± 0.27**	**94.30 ± 0.23**	**95.17 ± 0.19**	**94.97 ± 0.26**	**95.00 ± 0.38**
Kappa	W/O	31.09 ± 2.24	71.19 ± 0.51	78.31 ± 0.92	47.63 ± 2.15	56.50 ± 0.76	57.69 ± 0.53
	ImageNet	88.39 ± 0.42	93.49 ± 0.21	91.96 ± 0.22	94.32 ± 0.26	94.49 ± 0.20	94.57 ± 0.22
	Million-AID	**90.37 ± 0.44**	**95.17 ± 0.29**	**94.35 ± 0.24**	**95.24 ± 0.20**	**95.05 ± 0.27**	**95.07 ± 0.37**

Confusion matrices of different learning schemes







Results on AID

- Example images and predictions

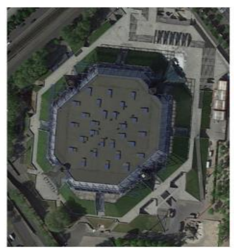

Railway station
Park
Industrial
Railway station


Railway station
Center
Bridge
Railway station


Railway station
Viaduct
Airport
Railway station


Railway station
Mountain
Farmland
Railway station

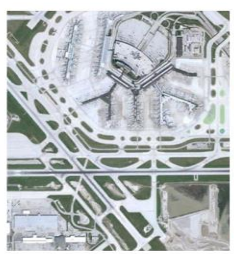

Center
Railway station
Industrial
Center


Center
Square
Stadium
Center


Center
Square
Storage tanks
Center


Center
Airport
Storage tanks
Center


Airport
Railway station
Center
Airport


Airport
Mountain
Square
Airport


Airport
Railway station
Commercial
Airport


Airport
Center
Railway station
Airport

The black labels are the ground truth. The orange labels indicate predictions by GoogleNet trained from scratch, the plum labels the predictions by DenseNet169 pre-trained on ImageNet, and the green labels the predictions by ResNet101 pre-trained on Million-AID.

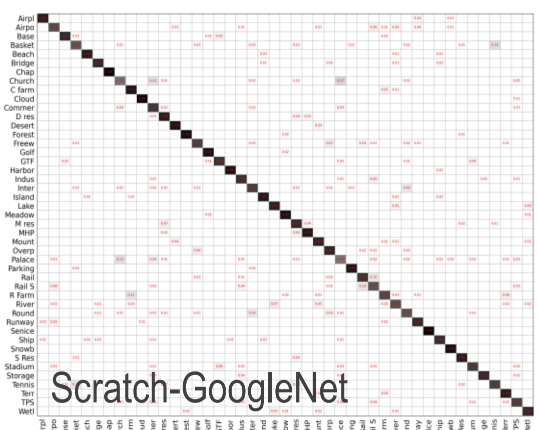
Results on NWPU-RESISC45

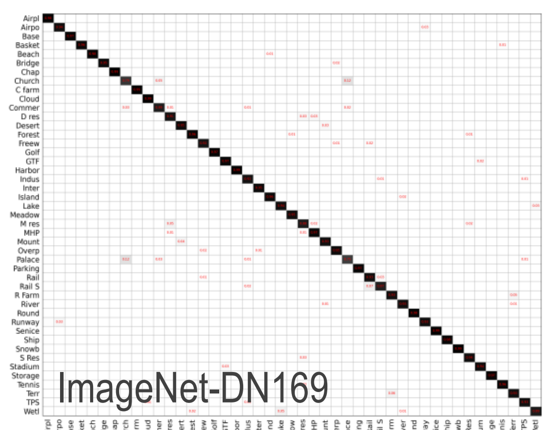
Accuracy comparison

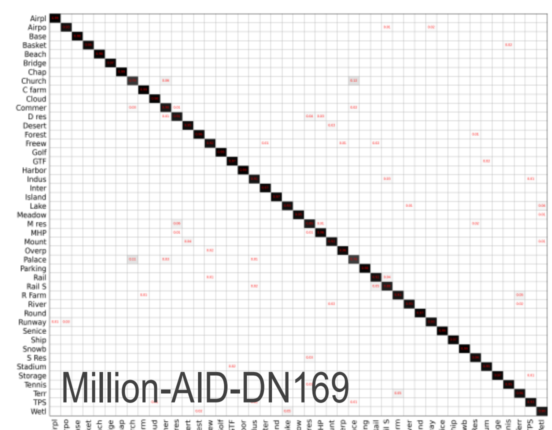
Classification accuracy (%) on NWPU-RESISC45 dataset using different initialization schemes

Metric	Pretrain dataset	AlexNet	VGG16	GoogleNet	ResNet101	DenseNet121	DenseNet169
OA	W/O	37.92 ± 0.70	73.19 ± 0.44	81.77 ± 0.56	58.82 ± 0.74	63.35 ± 0.34	64.51 ± 0.47
	ImageNet	87.19 ± 0.26	92.76 ± 0.18	91.71 ± 0.25	94.06 ± 0.16	93.90 ± 0.19	94.11 ± 0.20
	Million-AID	$\mathbf{88.24 \pm 0.21}$	$\mathbf{93.62 \pm 0.20}$	$\mathbf{93.40 \pm 0.23}$	$\mathbf{94.20 \pm 0.16}$	$\mathbf{94.21 \pm 0.20}$	$\mathbf{94.26 \pm 0.21}$
AA	W/O	37.92 ± 0.70	73.19 ± 0.44	81.77 ± 0.56	58.82 ± 0.74	63.35 ± 0.34	64.51 ± 0.47
	ImageNet	87.19 ± 0.26	92.76 ± 0.18	91.71 ± 0.25	94.06 ± 0.16	93.90 ± 0.19	94.11 ± 0.20
	Million-AID	$\mathbf{88.24 \pm 0.21}$	$\mathbf{93.62 \pm 0.20}$	$\mathbf{93.40 \pm 0.23}$	$\mathbf{94.20 \pm 0.16}$	$\mathbf{94.21 \pm 0.20}$	$\mathbf{94.26 \pm 0.21}$
Kappa	W/O	36.51 ± 0.72	72.59 ± 0.45	81.36 ± 0.58	57.89 ± 0.75	62.51 ± 0.35	63.70 ± 0.48
	ImageNet	86.89 ± 0.21	92.60 ± 0.19	91.52 ± 0.26	93.92 ± 0.17	93.76 ± 0.19	93.98 ± 0.20
	Million-AID	$\mathbf{87.97 \pm 0.21}$	$\mathbf{93.48 \pm 0.20}$	$\mathbf{93.25 \pm 0.24}$	$\mathbf{94.07 \pm 0.16}$	$\mathbf{94.08 \pm 0.20}$	$\mathbf{94.13 \pm 0.21}$

Confusion matrices of different learning schemes

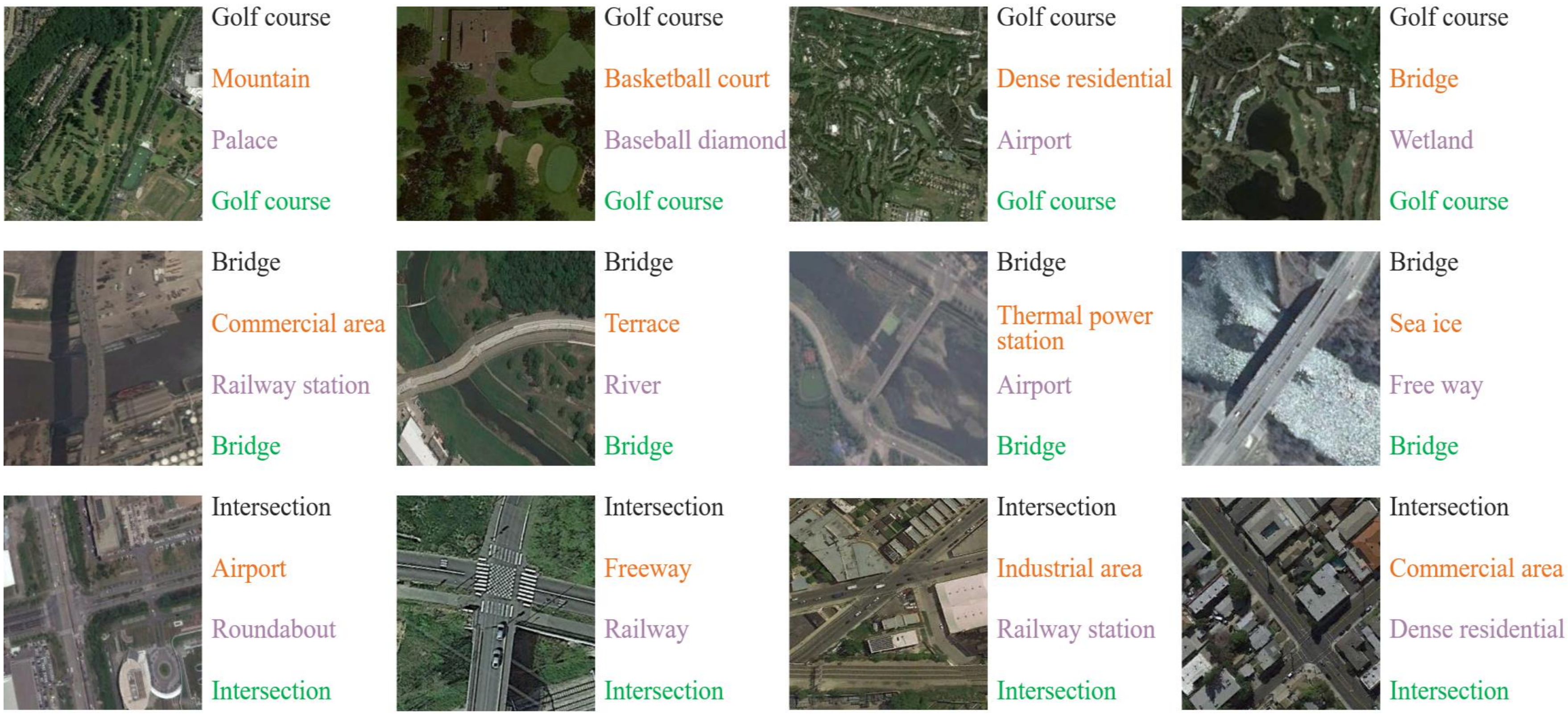

Scratch-GoogleNet


ImageNet-DN169


Million-AID-DN169

Results on NWPU-RESISC45

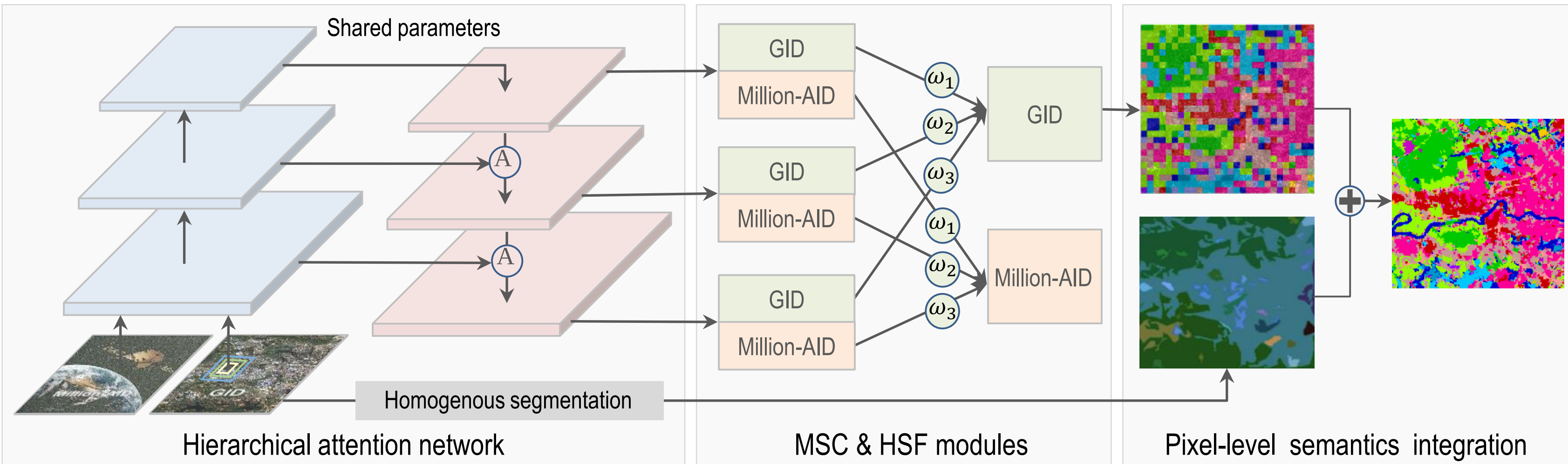
- Example images and predictions



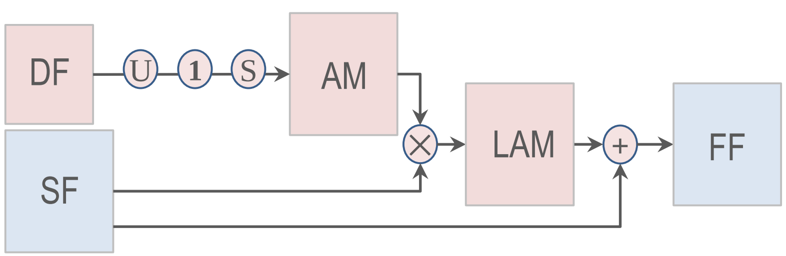
The black labels are the ground truth. The orange labels indicate predictions by GoogleNet trained from scratch, the plum labels the predictions by DenseNet169 pre-trained on ImageNet, and the green labels the predictions by ResNet101 pre-trained on Million-AID.

Semantic Classification

Transfer Knowledge from Million-AID for pixel-level image classification



Attention block



$$Loss^g = \sum_{s=1}^{S} w_s CE_s^g$$

$$Loss^m = \sum_{s=1}^{S} w_s CE_s^m$$

$$Loss = \mu_g Loss^g + \mu_m Loss^m$$

$$\widehat{p}_n(I) = \frac{\sum_{s=1}^{S} w_s p_{s,n}(I)}{\sum_{s=1}^{S} w_s}$$

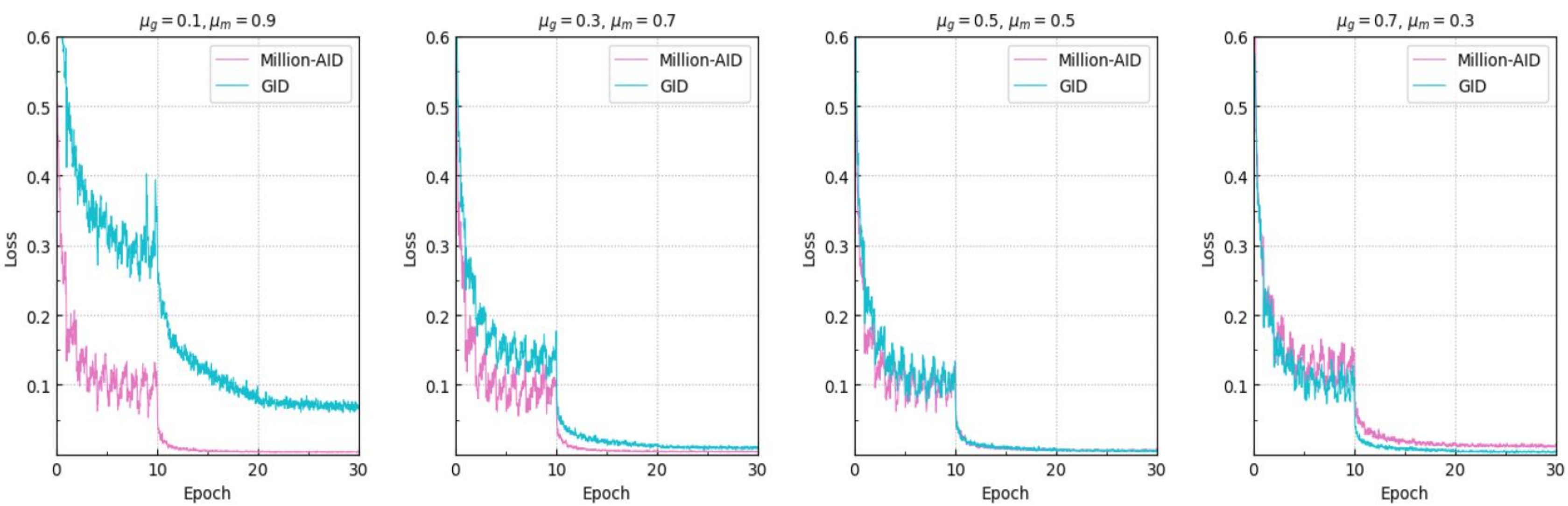
$$l(I) = \underset{n \in [1,2,\ldots,n]}{\arg\max} \; \widehat{p}_n(I)$$

Ablation Study

- Weights influence of different tasks

μ_g	μ_m	GID Kappa (%)	GID OA (%)	GID mIoU (%)	Million-AID Kappa (%)	Million-AID OA (%)	Million-AID AA (%)
0.1	0.9	62.85	69.06	39.88	**90.36**	**90.62**	**89.55**
0.3	0.7	65.15	71.00	41.85	89.44	89.72	88.91
0.5	0.5	**66.65**	**72.38**	**42.71**	**89.67**	**89.94**	**89.14**
0.7	0.3	66.14	72.02	41.75	88.98	89.27	87.84

- Corresponding training loss observation



Comparison

- Quantitative comparison using different Modules

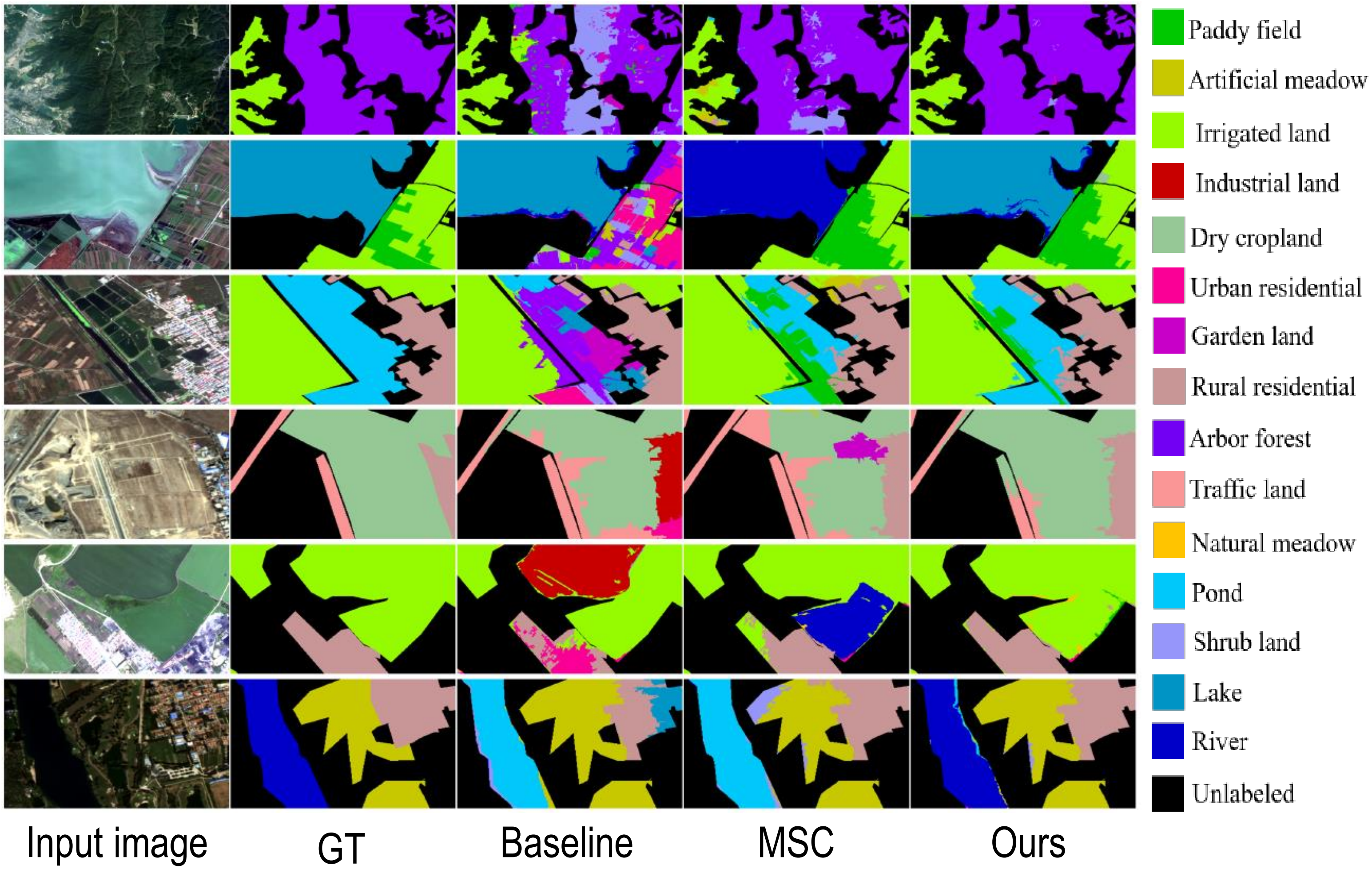
Baseline	MSC	HSR	HSI	Kappa (%)	OA (%)	mIoU (%)
✓				51.59	59.09	30.79
✓	✓			66.65	72.38	42.71
✓	✓	✓		66.79	72.52	43.07
✓	✓	✓	✓	**67.33**	**73.03**	**43.68**

- Quantitative comparison with SOTA methods

Methods	Kappa	OA (%)
MLC + SGDL	0.145	23.61
SVM + SGDL	0.148	23.92
MLP + SGDL	0.199	30.57
RF + SGDL	0.237	33.70
DeepLab V3+ Mobilenet	0.357	54.64
U-Net	0.439	56.59
PSPNet	0.458	60.73
DeepLab V3+	0.478	62.19
DeepLab V3+	0.598	69.16
PT-GID	0.605	70.04
Ours	**0.673**	**73.03**

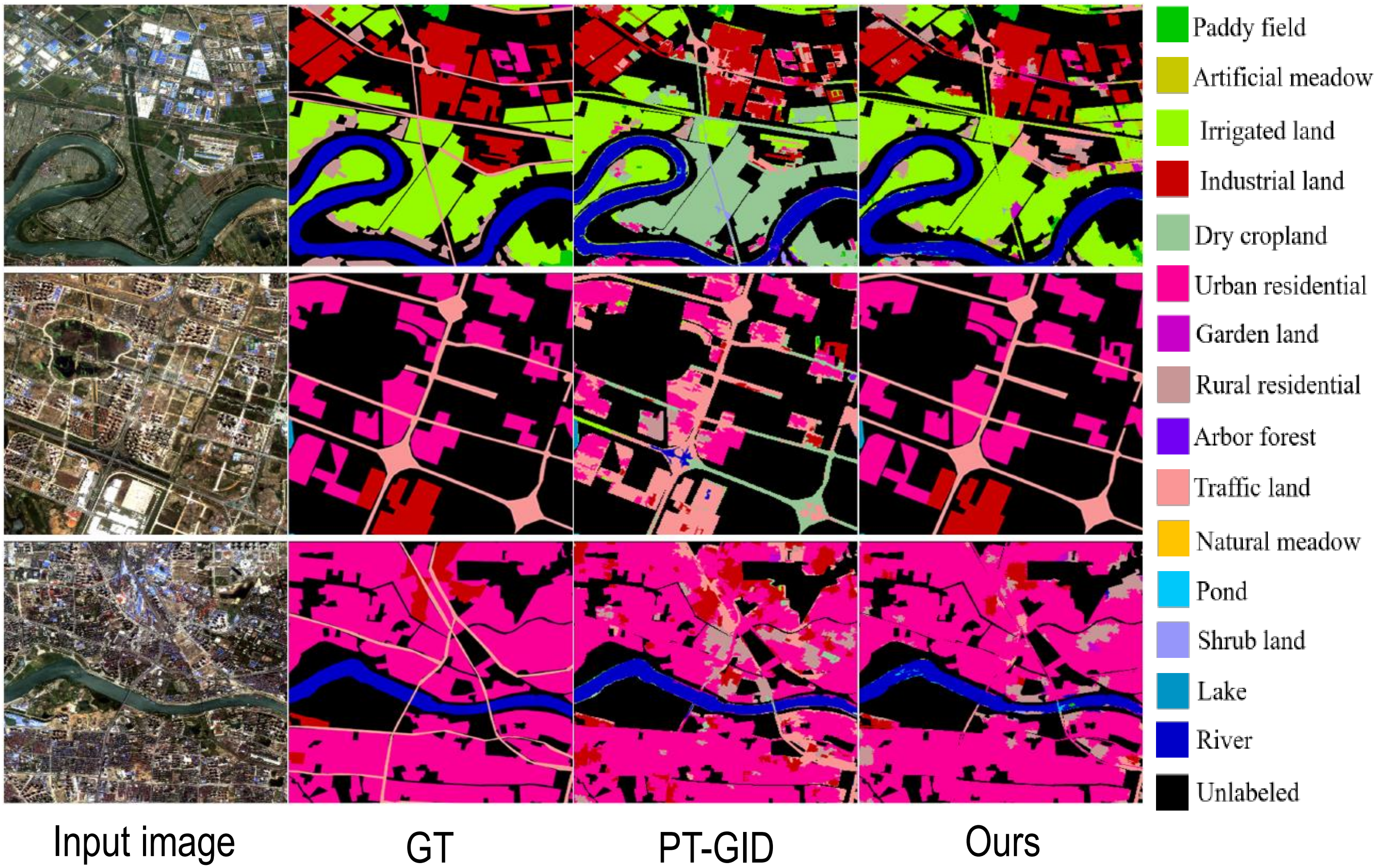
Comparison

- Qualitative comparison with different Modules



Comparison

- Qualitative comparison with SOTA methods



Outline

- Background
- Revisiting Aerial Image Interpretation
- Introduction to Million-AID
- Aerial Scene Classification: A New Benchmark
- Knowledge Transfer: From Tile-level to Pixel-level
- Conclusions

Summary

A review of aerial image interpretation

- Classification prototypes develop with the improvement of image resolution
- Pixel-wise, segmentation-based, and tile-level classification methodologies are established, relying on visual characteristics of images with different resolutions

Tile-level scene classification

- We released a new large-scale dataset, Million-AID, for aerial scene classification
- Million-AID shows better transferability than ImageNet for aerial scene classification

Pixel-wise image parsing

- We verify the tremendous potential of transferring scene knowledge of Million-AID to advance aerial image interpretation from tile-level classification to pixel-wise labeling

CAPTAIN

COMPUTATIONAL AND PHOTOGRAMMETRIC VISION

THANKS

School of Computer Science, Wuhan University

Institute of Artificial Intelligence, Wuhan University

State Key Lab. LIESMARS, Wuhan University

Gui-Song Xia (guisong.xia@whu.edu.cn)