Sensing, Understanding, and Synthesizing Humans in an Open World

Ziwei Liu

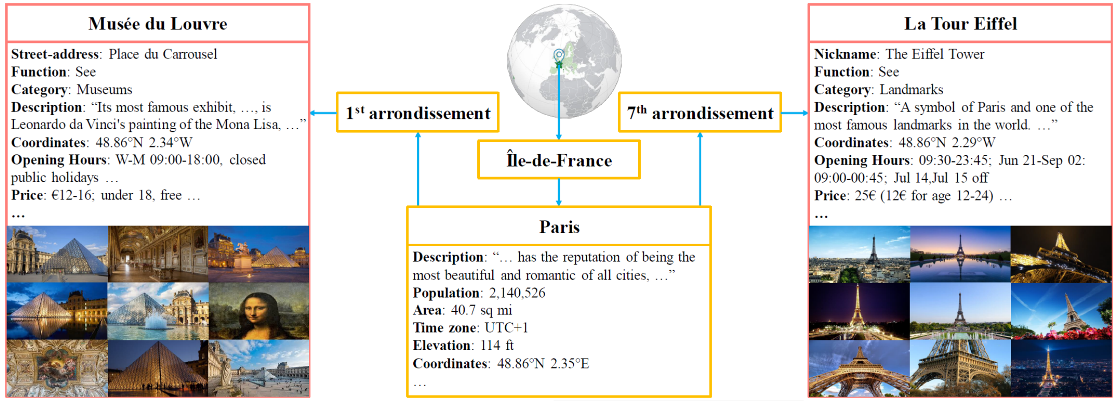
Nanyang Technological University

Human-Centric AI

Human-Centric AI

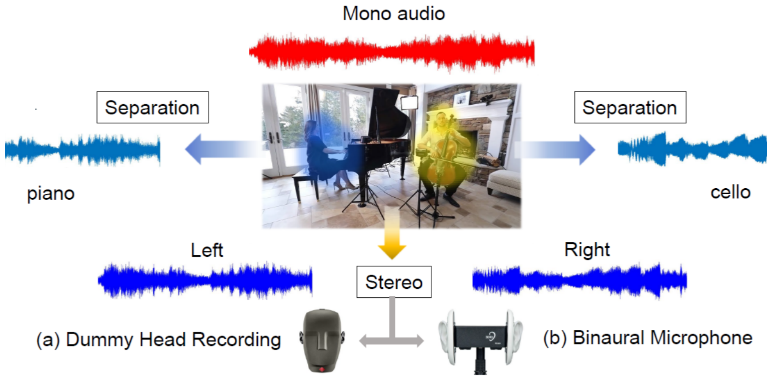
Robust **Sensing**

CelebA-Spoof



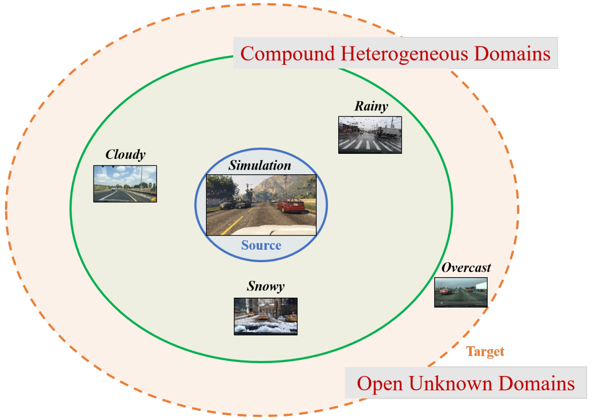
Understanding beyond Recognition

Placepedia



Synthesizing across Modalities

Sep-Stereo



Open World Learning

Open Compound Domain Adaptation

Robust Sensing

CelebA-Spoof

Understanding beyond Recognition

Placepedia

Synthesizing across Modalities

Sep-Stereo

Open World Learning

Open Compound Domain Adaptation

CelebA-Spoof: Large-Scale Face Anti-Spoofing Dataset With Rich Annotations

Yuanhan Zhang*[1,2] Zhenfei Yin*[2] Yidong Li[1] Guojun Yin[2] Junjie Yan[2] Jing Shao[2] Ziwei Liu[3]

[1]Beijing Jiaotong University [2]SenseTime Research

[3]The Chinese University of Hong Kong

*Equal contribution

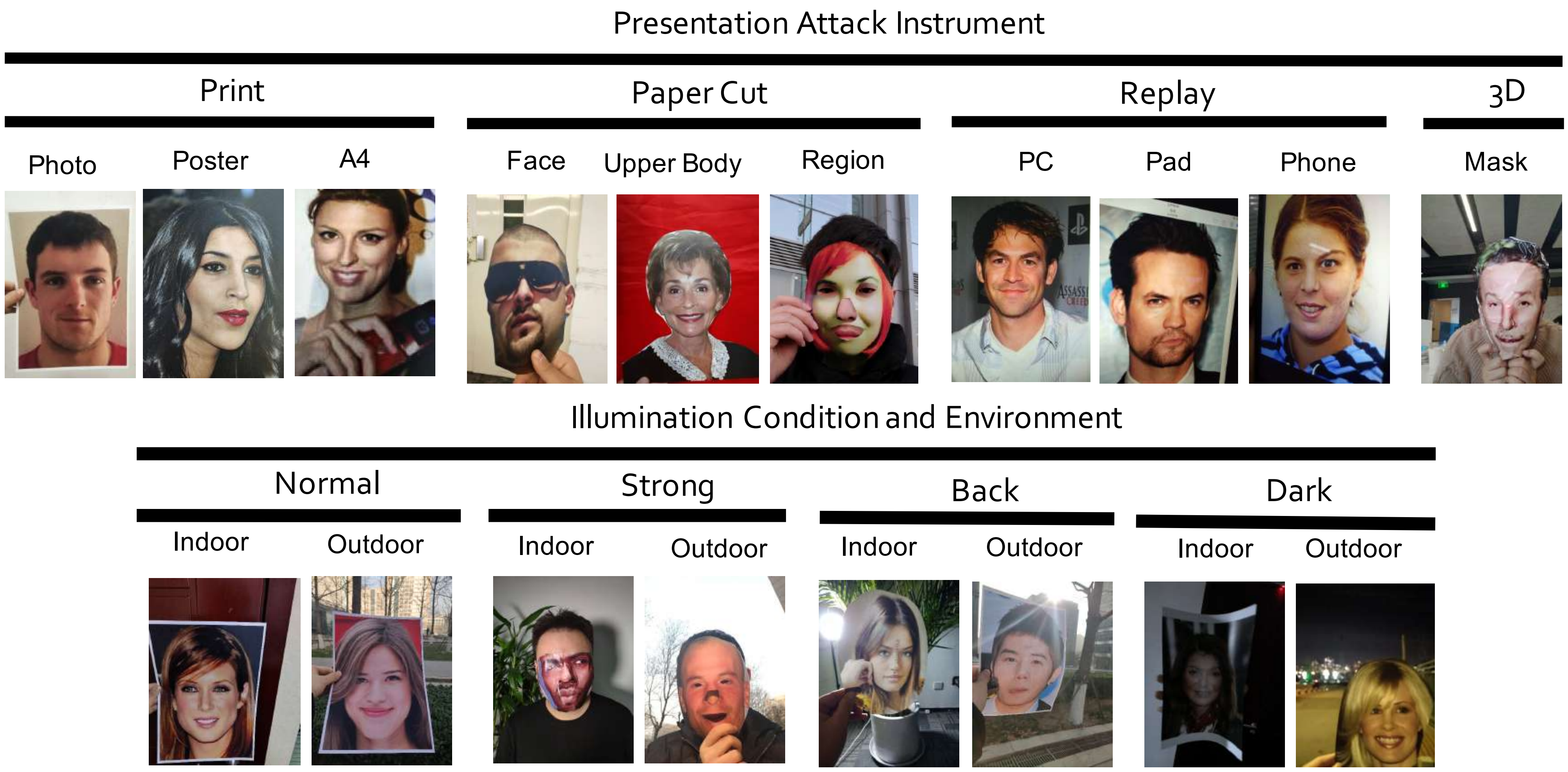
Introduction

CelebA-Spoof Dataset

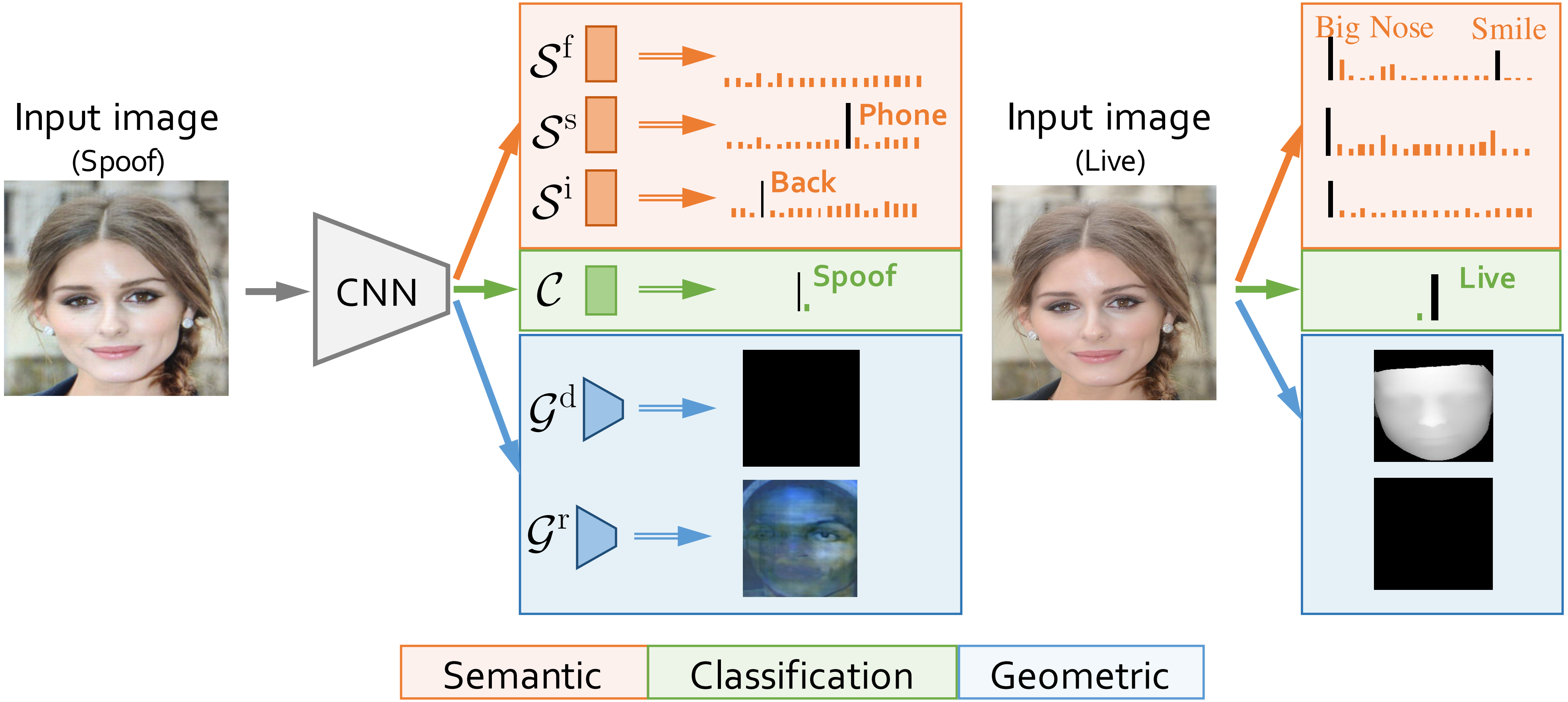
1. Lack of Diversity 2. Lack of Annotations

Dataset	Year	#Subjects	#Data(V/I)	#Annotation
Replay-Attack	2012	50	1,200 (V)	1
CASIA-MFSD	2012	50	600 (V)	
3DMAD	2014	14	255 (V)	
MSU-MFSD	2015	35	440 (V)	
Msspoof	2015	21	4,704 (V)	
HKBU-MARs V2	2016	12	1,008 (V)	
MSU-USSA	2016	1,140	10,260 (I)	
Oulu-NPU	2017	55	5,940 (V)	
SiW	2018	165	4,620 (V)	
CASIA-SURF	2018	1,000	21,000 (V)	
CSMAD	2018	14	260 (V), 17 (I)	
HKBU-MARs V1 +	2018	12	180 (V)	
SiW-M	2019	493	1,628 (V)	
CelebA-Spoof	2020	10,177	625,537 (I)	43

CelebA-Spoof Dataset



Auxiliary Information Embedding Network (AENet)



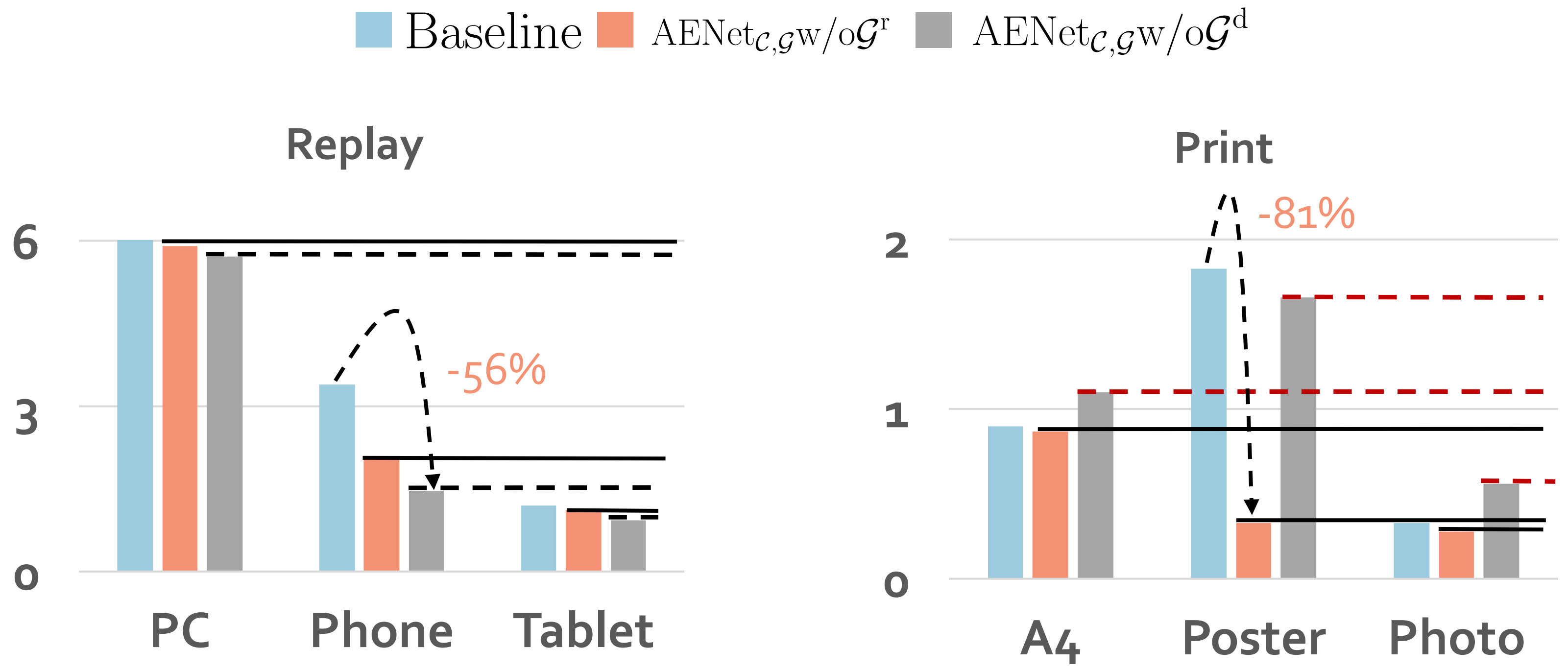
Observation

Depth Maps are More Versatile

Smaller $\mathrm{APCER}_{\mathcal{S}^{\mathrm{s}}}$ is better

- - - - - Reflection map as auxiliary supervision

——— Depth map as auxiliary supervision



Depth Maps are More Versatile

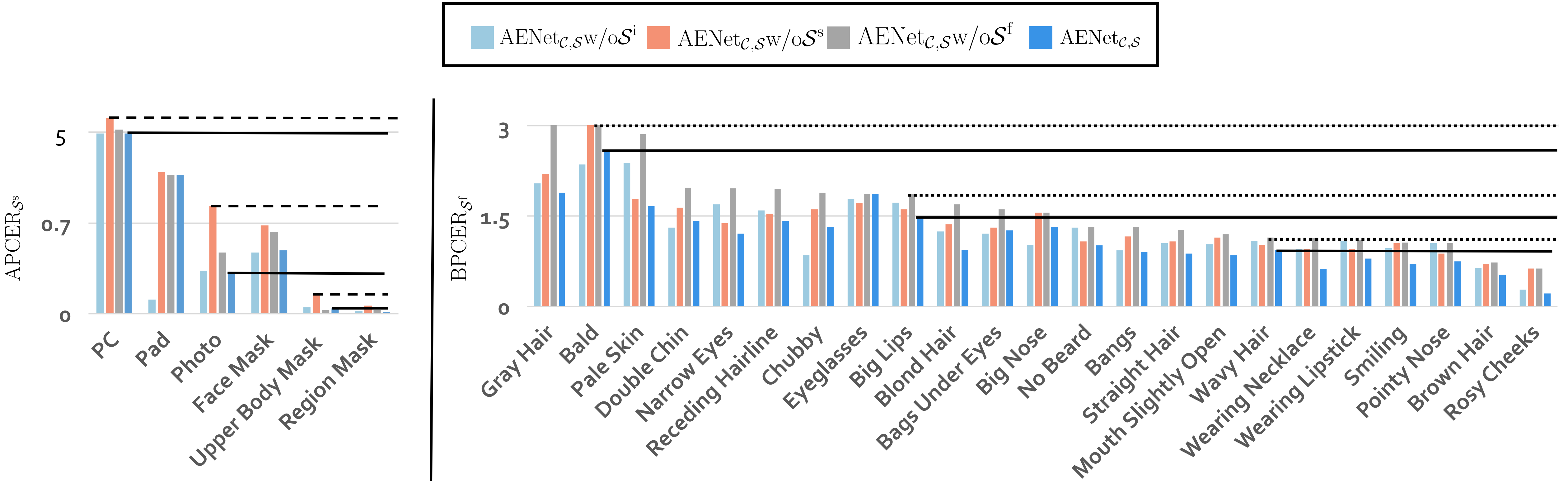
Semantic Attribute Matters

Smaller $\mathrm{APCER}_{\mathcal{S}^{\mathrm{s}}}$ and $\mathrm{BPCER}_{\mathcal{S}^{\mathrm{f}}}$ are better

$\mathrm{AENet}_{\mathcal{C},\mathcal{S}}$

$\mathrm{AENet}_{\mathcal{C},\mathcal{S}}\mathrm{w/o}\mathcal{S}^{\mathrm{s}}$

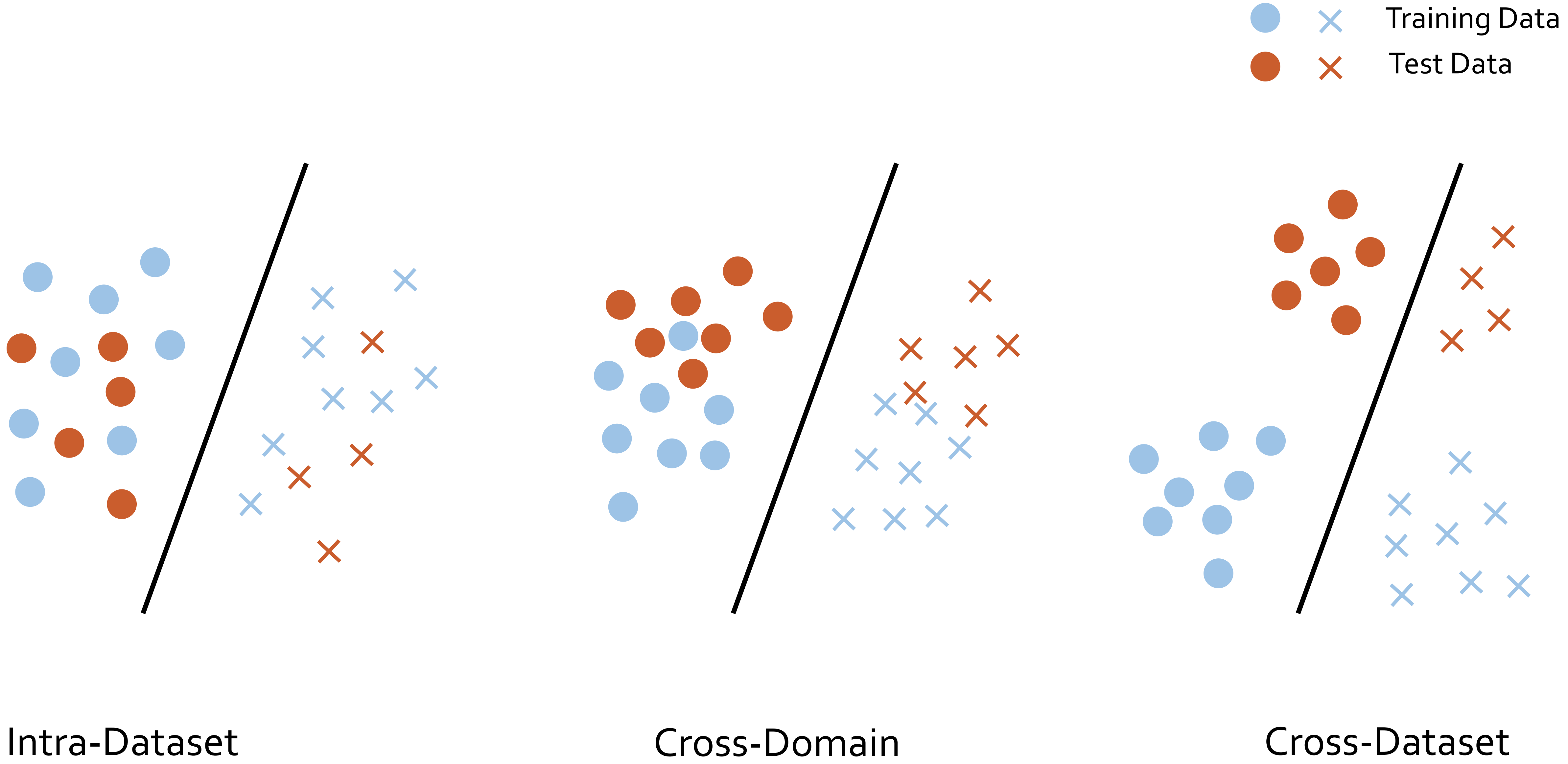
$\mathrm{AENet}_{\mathcal{C},\mathcal{S}}\mathrm{w/o}\mathcal{S}^{\mathrm{f}}$



Semantic Attribute Matters

Benchmark

Benchmarks


Training Data
Test Data
Intra-Dataset
Cross-Domain
Cross-Dataset

Benchmarks

Intra-dataset Benchmark

Model	Backbone	Parm. (MB)	Recall (%)↑			AUC↑	EER (%)↓	APCER (%)↓	BPCER (%)↓	ACER (%)↓
			FPR = 1%	FPR = 0.5%	FPR = 0.1%					
Auxiliary* [23]	-	22.1	97.3	95.2	83.2	0.9972	1.2	5.71	1.41	3.56
BASN [16]	VGG16	569.7	98.9	**97.8**	**90.9**	**0.9991**	1.1	4.0	1.1	2.6
AENet$_{\mathcal{C},\mathcal{S},\mathcal{G}}$	ResNet-18	42.7	**98.9**	97.3	87.3	0.9989	**0.9**	**2.29**	**0.96**	**1.63**

Cross-Domain Benchmark

Protocol	Model	Recall (%) ↑			AUC↑	EER (%)↓	APCER (%)↓	BPCER (%)↓	ACER (%)↓
		FPR = 1%	FPR = 0.5%	FPR = 0.1%					
	Baseline	93.7	86.9	69.6	**0.996**	2.5	5.7	2.52	4.11
1	AENet$_{\mathcal{C},\mathcal{G}}$	93.3	88.6	**74.0**	0.994	2.5	5.28	2.41	3.85
	AENet$_{\mathcal{C},\mathcal{S}}$	93.4	89.3	71.3	**0.996**	2.4	5.63	2.42	4.04
	AENet$_{\mathcal{C},\mathcal{S},\mathcal{G}}$	**95.0**	**91.4**	73.6	0.995	**2.1**	**4.09**	**2.09**	**3.09**
	Baseline	#	#	#	0.998 ±0.002	1.5±0.8	8.53±2.6	1.56±0.81	5.05±1.42
2	AENet$_{\mathcal{C},\mathcal{G}}$	#	#	#	0.995±0.003	1.6±4.5	8.95±1.07	1.67±0.9	5.31±0.95
	AENet$_{\mathcal{C},\mathcal{S}}$	#	#	#	0.997±0.002	1.2±0.7	**4.01±2.9**	1.24±0.67	3.96±1.79
	AENet$_{\mathcal{C},\mathcal{S},\mathcal{G}}$	#	#	#	**0.998±0.002**	**1.3±0.7**	4.94±3.42	**1.24±0.73**	**3.09±2.08**

Cross-Dataset Benchmark

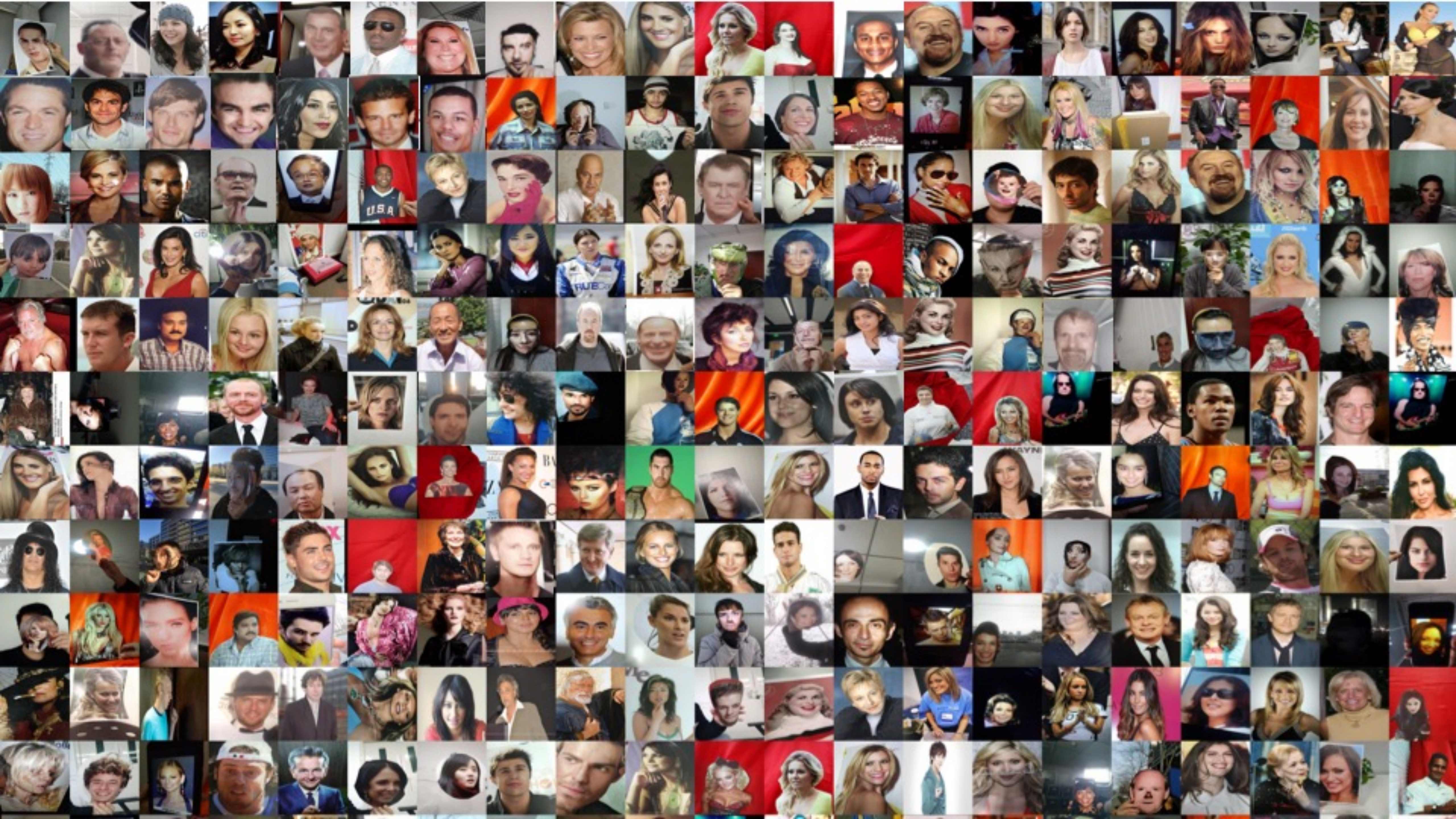
Model	Training	Testing	HTER (%) ↓
FAS-TD-SF [34]	SiW	CASIA-MFSD	39.4
FAS-TD-SF [34]	CASIA-SURF	CASIA-MFSD	37.3
AENet$_{\mathcal{C},\mathcal{S},\mathcal{G}}$	SiW	CASIA-MFSD	27.6
Baseline	CelebA-Spoof	CASIA-MFSD	14.3
AENet$_{\mathcal{C},\mathcal{G}}$	CelebA-Spoof	CASIA-MFSD	14.1
AENet$_{\mathcal{C},\mathcal{S}}$	CelebA-Spoof	CASIA-MFSD	12.1
AENet$_{\mathcal{C},\mathcal{S},\mathcal{G}}$	CelebA-Spoof	CASIA-MFSD	**11.9**

CelebA-Spoof

Large-Scale


U.S.A.
citi


U.S.A
citi



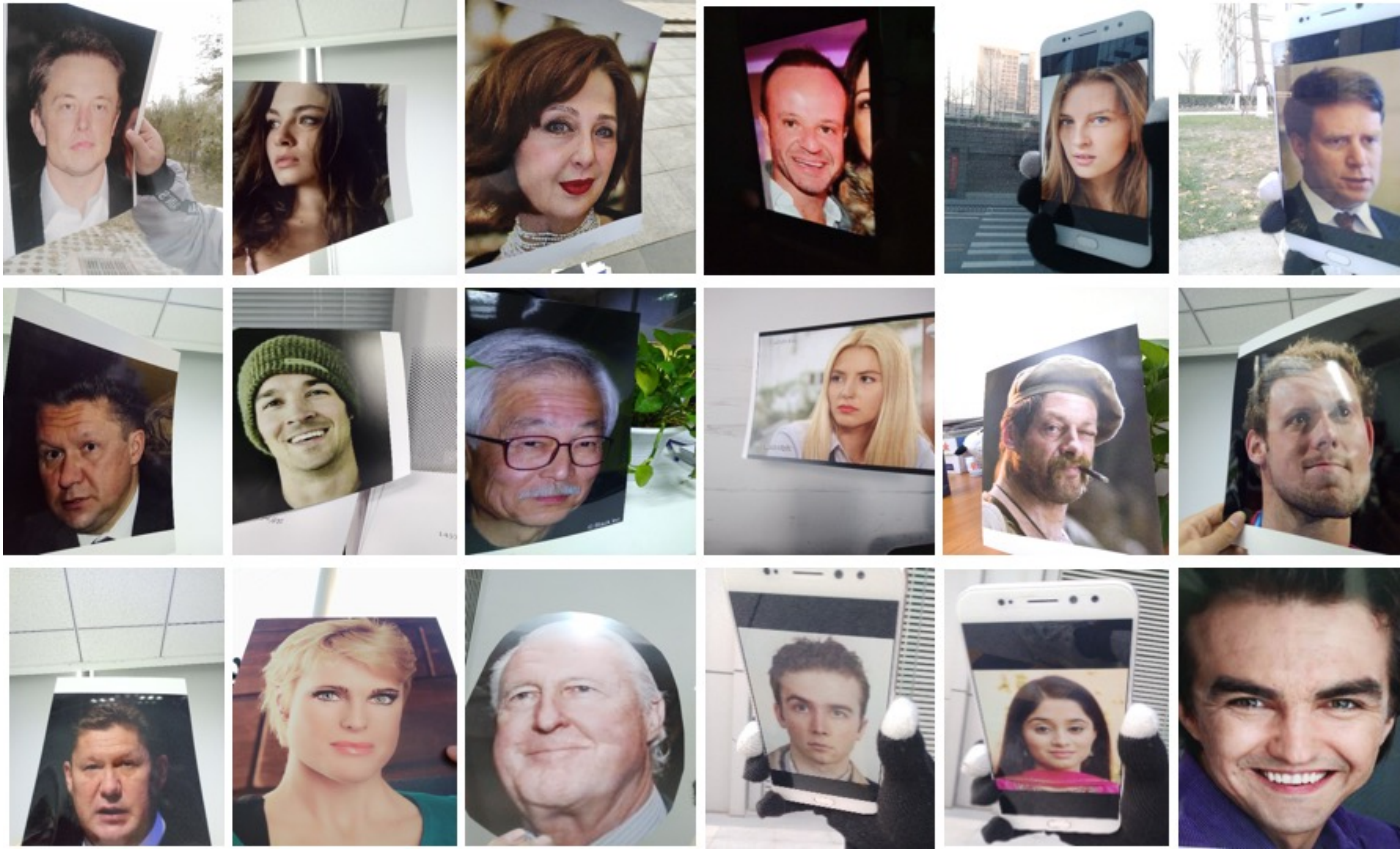
Collection Dimensions

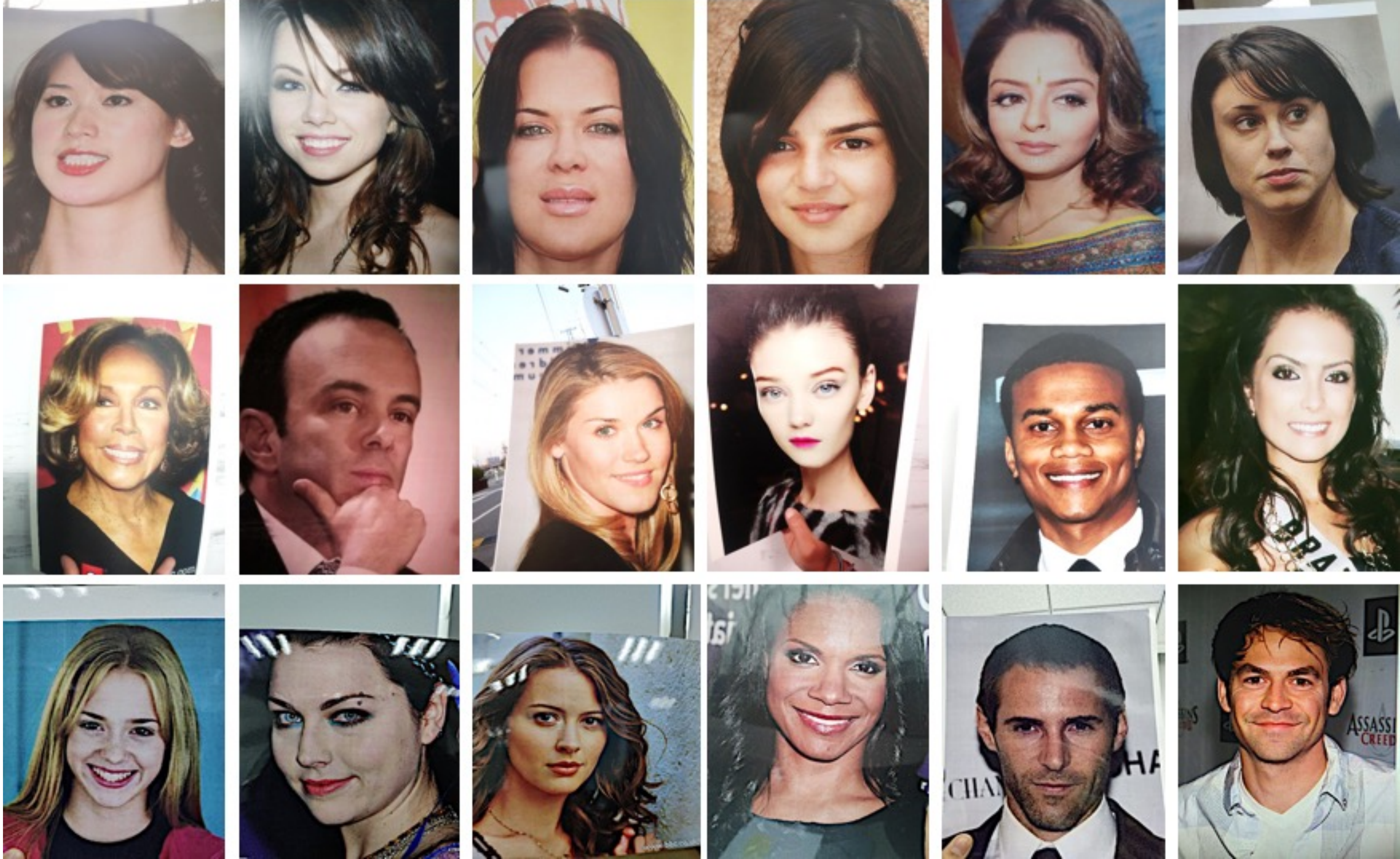
Collection Dimensions

Angle

Input Sensor

Shape

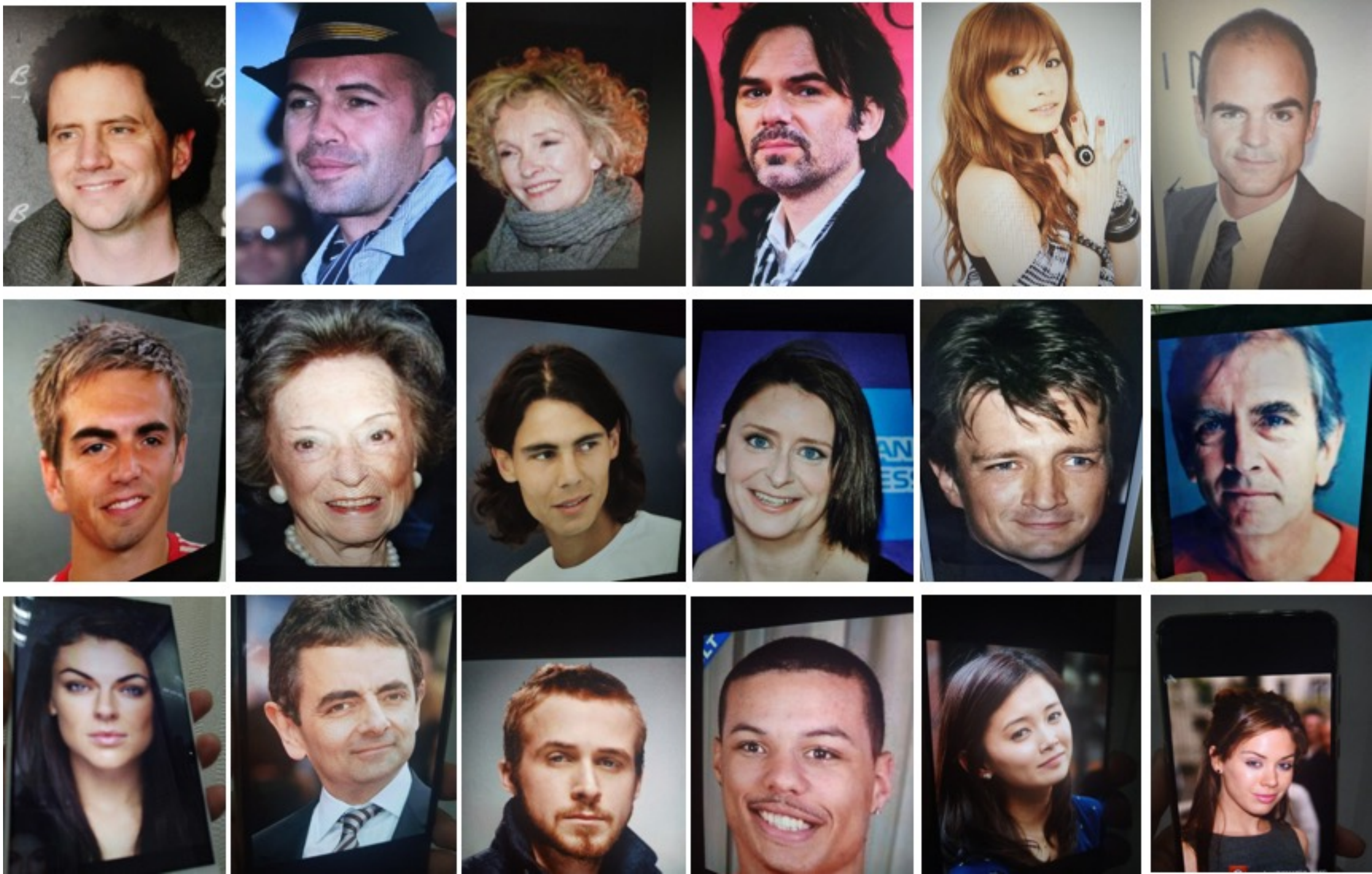




Rich Annotations

Rich Annotations

Illumination Condition		Spoof Type				Environment
Normal	Strong	Photo	Poster		A4	Indoor
		Face Mask	Upper Body Mask		Region Mask	
Back	Dark	PC	Tablet	Phone	3D Mask	Outdoor



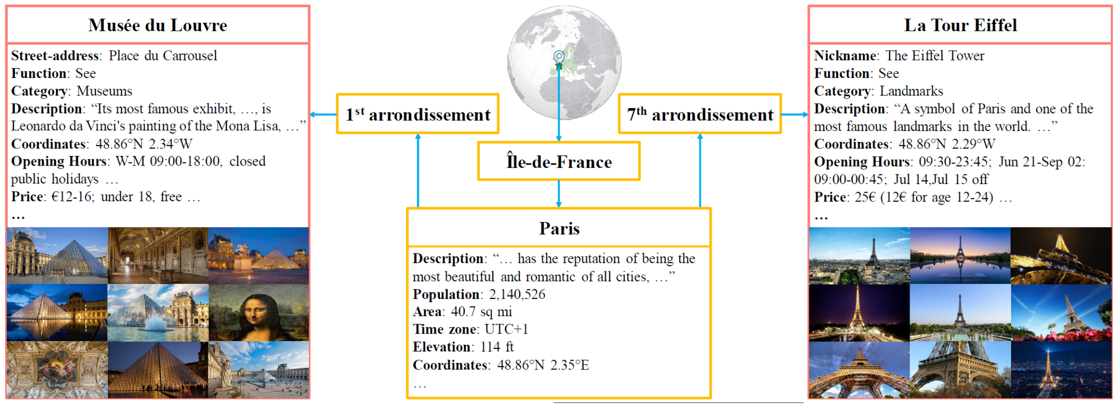
Github Page: https://github.com/Davidzhangyuanhan/CelebA-Spoof

Robust Sensing

CelebA-Spoof

Synthesizing across Modalities

Sep-Stereo



Understanding beyond Recognition

Placepedia

Open World Learning

Open Compound Domain Adaptation

Placepedia: Comprehensive Place Understanding with Multi-Faceted Annotations

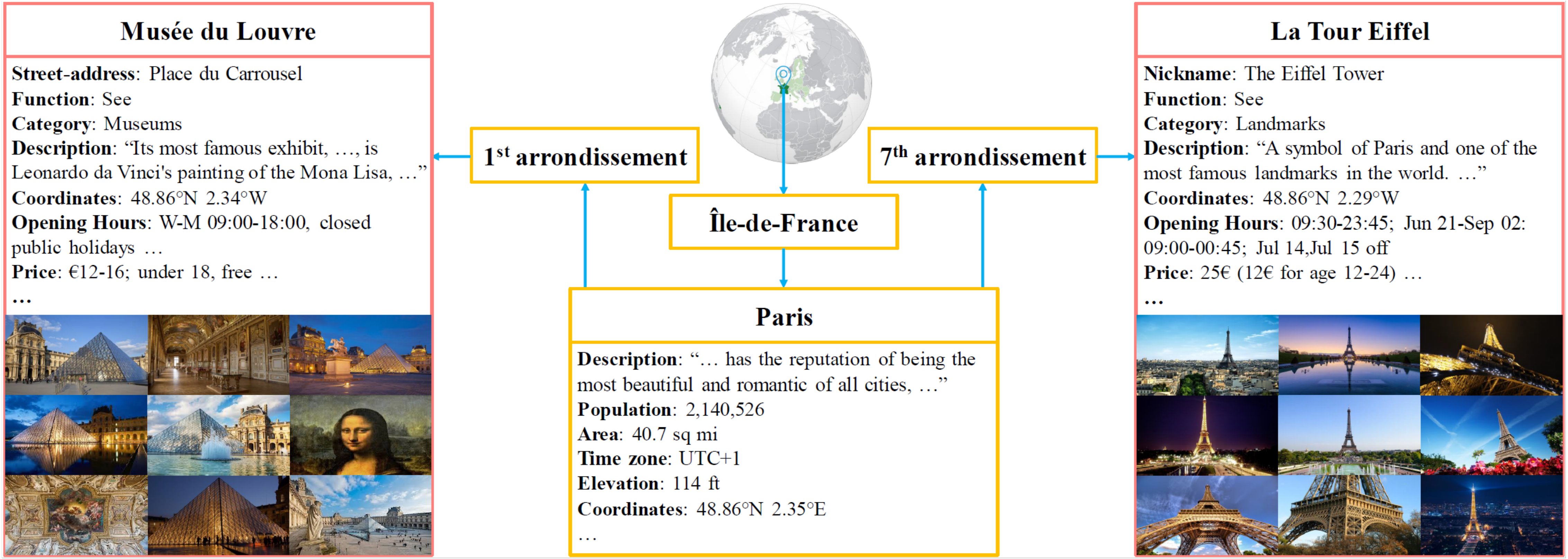
Huaiyi Huang , Yuqi Zhang, Qingqiu Huang, Zhengkui Guo, Ziwei Liu, Dahua Lin

The Chinese University of Hong Kong

Dataset Overview

- Place is an important element in visual understanding.
  - Functionality
  - Cultural style
  - Economic type
- Comprehensive place understanding
  - Far beyond categorizing a place with an image
  - Requires information of multiple aspects

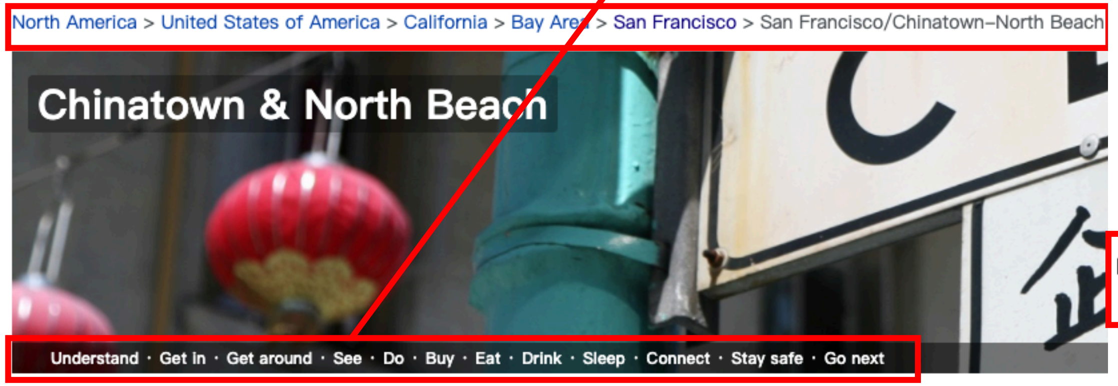
Dataset Overview

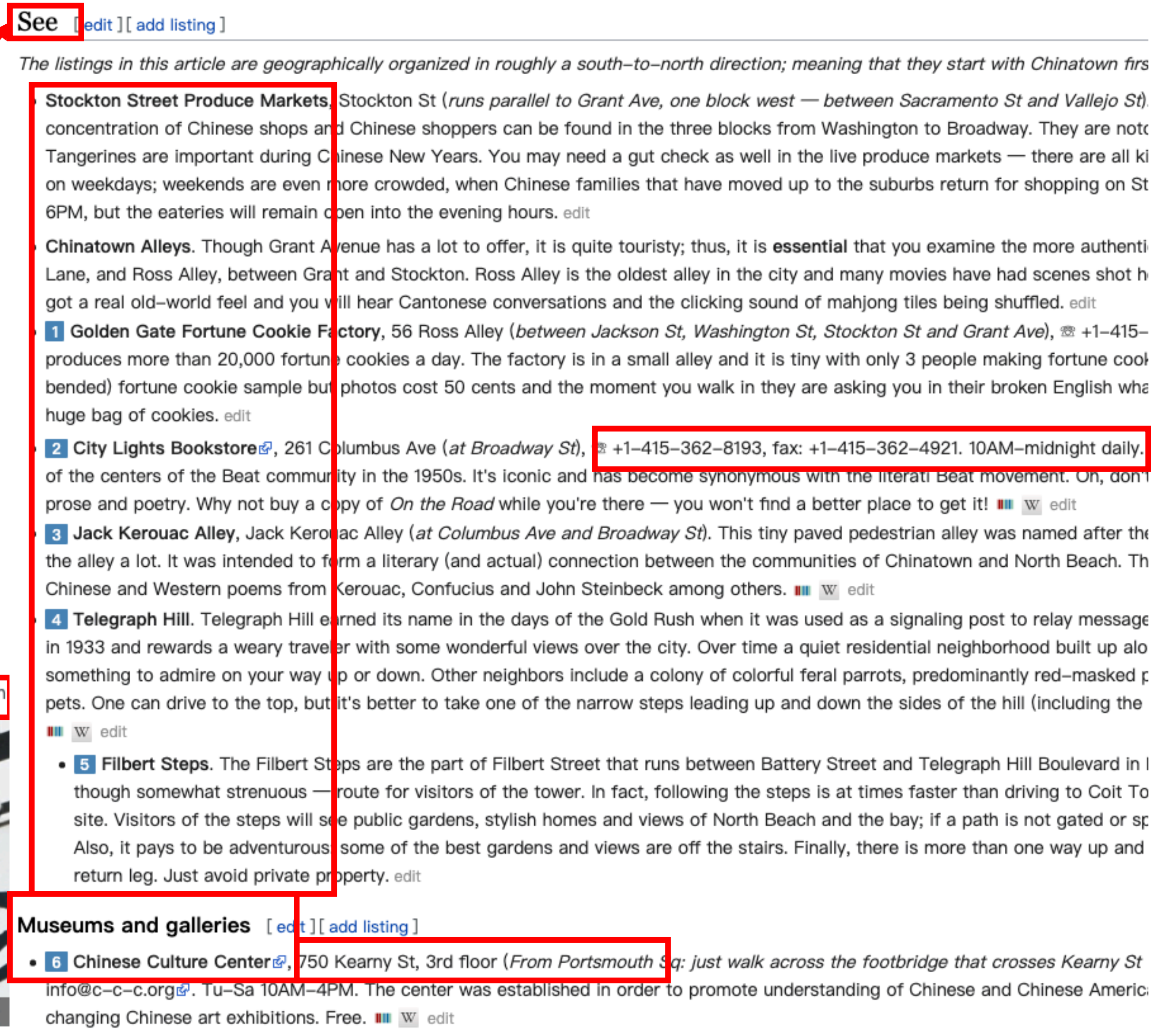


Get Very Rich Data of Place

- Wikivoyage!

- 360K places of 25K cities
- Lots of meta data





Get Very Rich Data of Place

- Some are not place. Refine the place list!
  - It has the attribute GPS coordinates or address
  - It is identified as a location by Google Entity Recognition [1] or Stanford Entity Recognition [2]

- 1 **Grant Plaza Hotel**, 465 Grant Ave (*at Pine St*), ☎ +1-415-434-3883, fax: +1-41
  2 night stay. Rates are reasonable. Make sure you ask for one of the outside rooms

→ Is a place.

- **Chinese New Year Festivities**. Jan or Feb. Celebrated for over 5,000 years,
  firecrackers, "lucky-money" envelopes, colorful banners, over 100 ornately the

→ Is not a place.

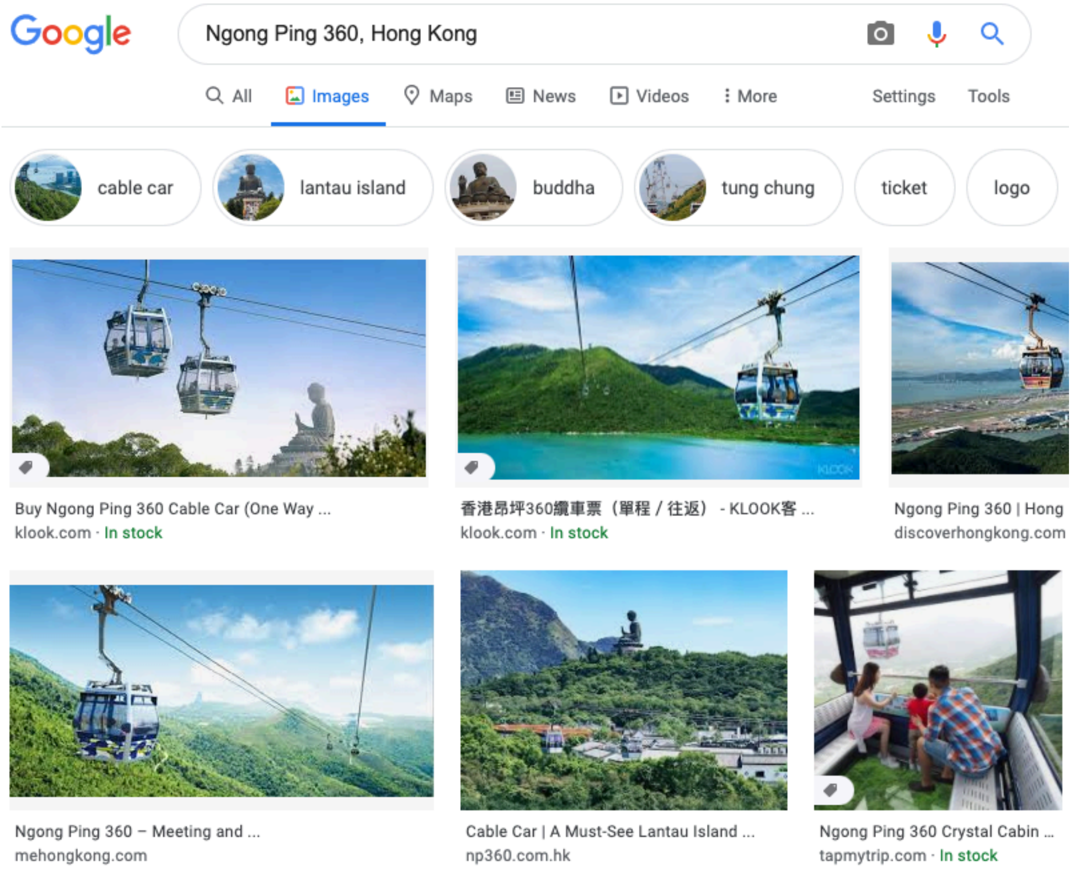
- Remains 320K places

[1] https://cloud.google.com/natural-language

[2] https://nlp.stanford.edu/software/CRF-NER.html

Get Very Rich Data of Place

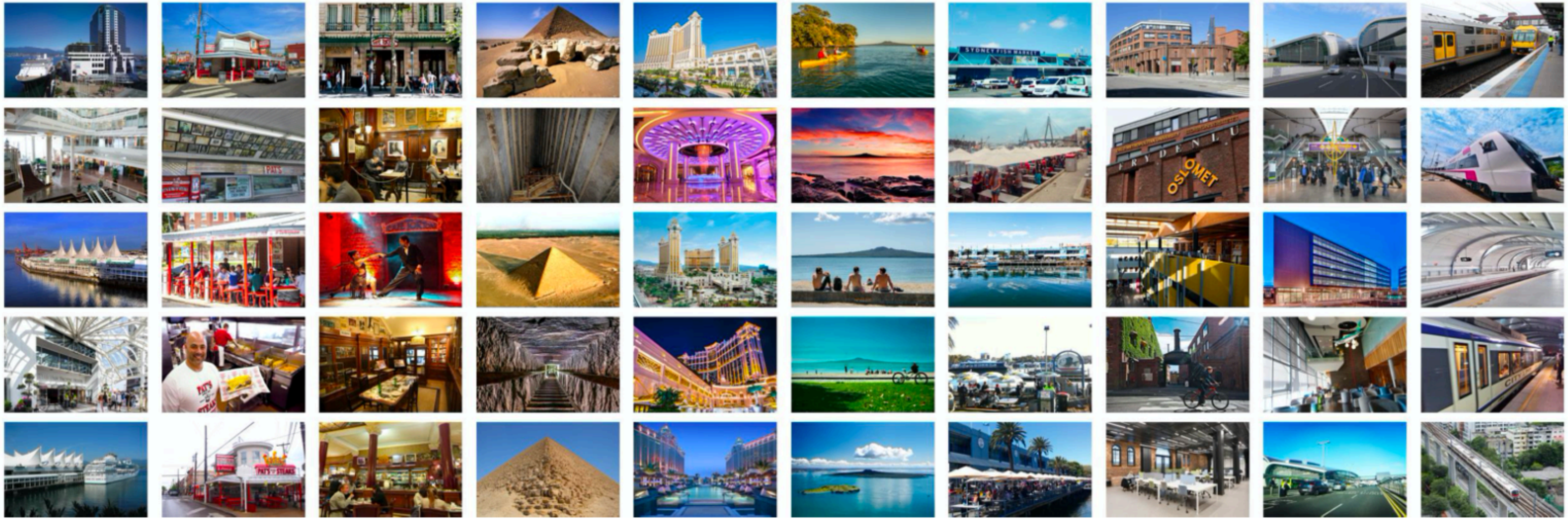
- Google Image open source



- 240K places
- 35M images

Challenges

Pan Pacific	Pat's King of Steaks	Cafe Tortoni	Red Pyramid	Galaxy Macau	Rangitoto Island	Sydney Fish Market	Oslo Metropolitan University	Dublin Airport	Airport Rail Link
Vancouver	Philadelphia Region	Buenos Aires	Cairo	Macau	Auckland	Sydney	Oslo	Dublin	Bangkok
Canada	United States of America	Argentina	Egypt	Macau	New Zealand	Australia	Norway	Ireland	Thailand
Splurge_Hotels	Restaurants	Café	Landmarks	Gambling	Islands	Markets	Universities	Airports	Public_transportations
Sleep	Eat	Drink	See	Do	See & Do	Buy	Learn	Get in	Get around



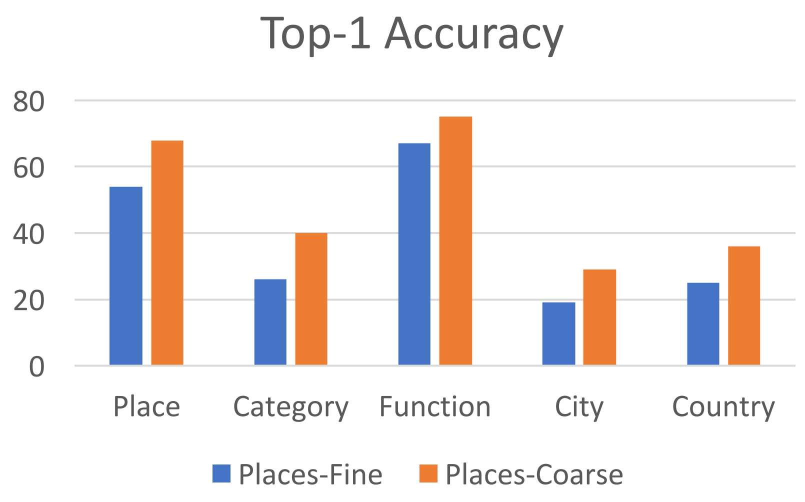
1) Daytime to nighttime   2) Different angles   3) Inside and Outside

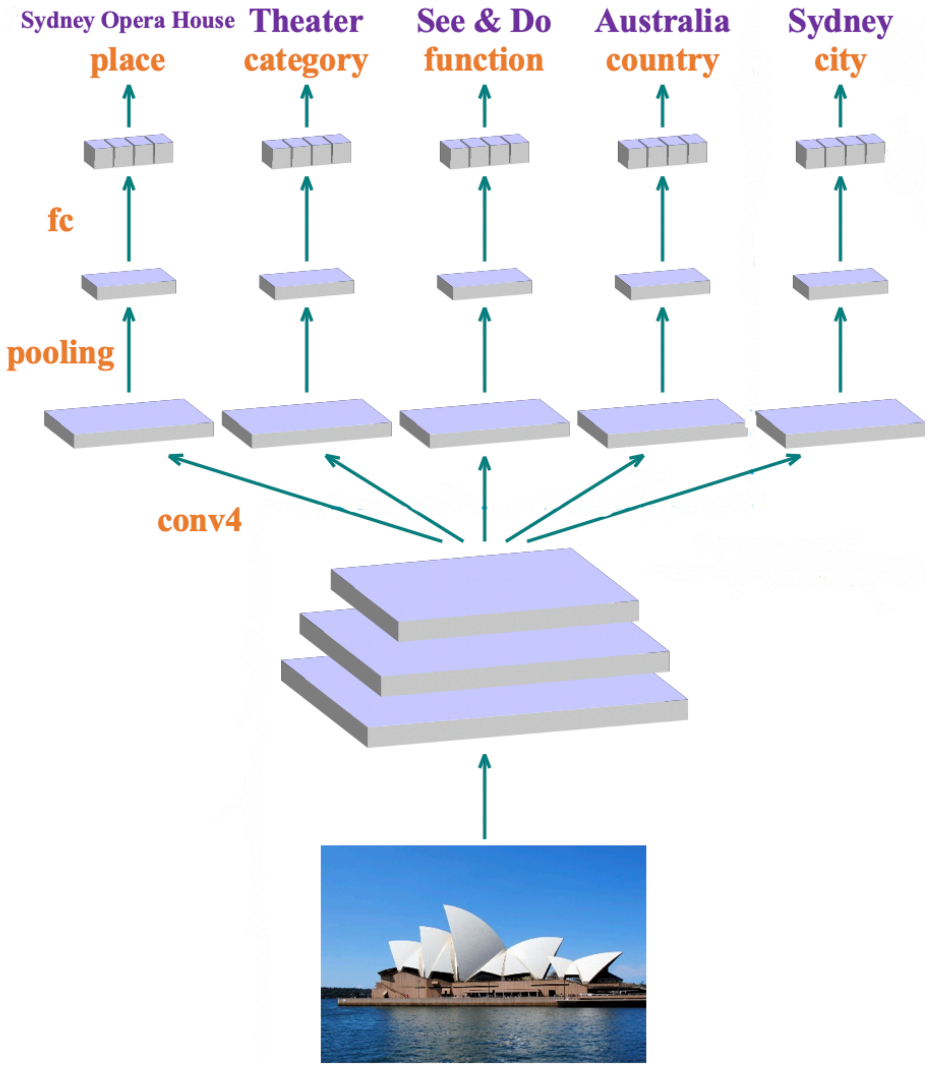
Comprehensive Place Understanding

- Benchmarks
  - Datasets:
    - Places-Coarse: 26K places
    - Places-Fine: 1K places
  - Tasks:
    - Place Retrieval (determine if two images belong the same place)
    - Place Categorization (classify places into categories like museums, parks, churches, and temples)
    - Function Categorization (classify places by their functionality such as eat, sleep, see, buy, and so on)
    - City/Country Recognition (classify places into their cities or countries)

A unified Framework to Predict All Tasks

- Duplicate the last convolution/pooling/fc layers of ResNet50 to five branches





Findings

Ground Truth

Parks
Franklin Park

Parks
Franklin Park

Parks
Parque México

Parks
Parque Ecológico do Tietê

Category Recognition
Place Retrieval

Tombs

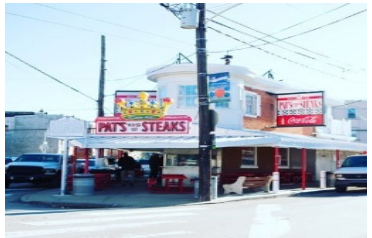
Museums

Museums

Pubs

Category Recognition

Eat

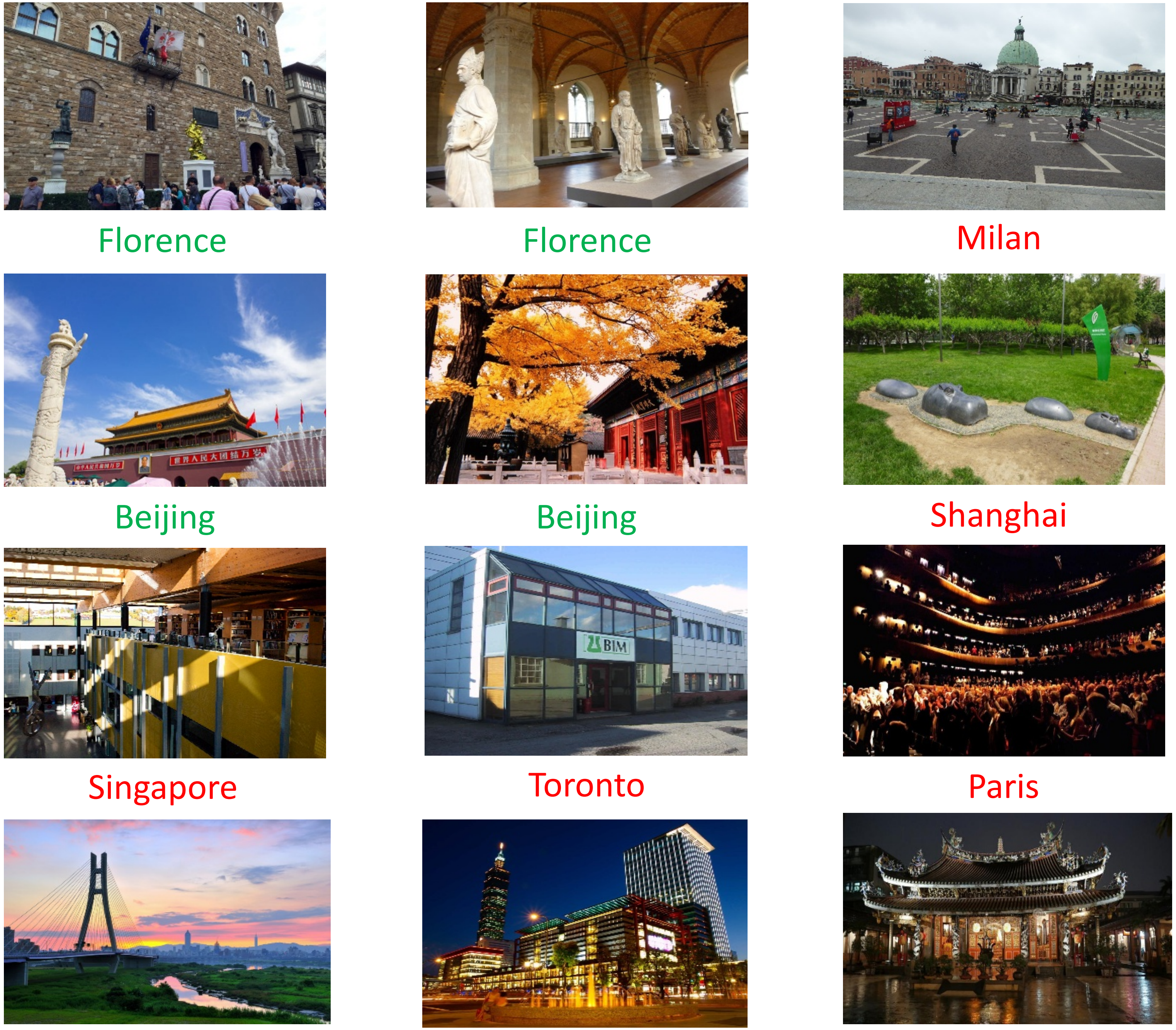


Buy

Buy

Drink

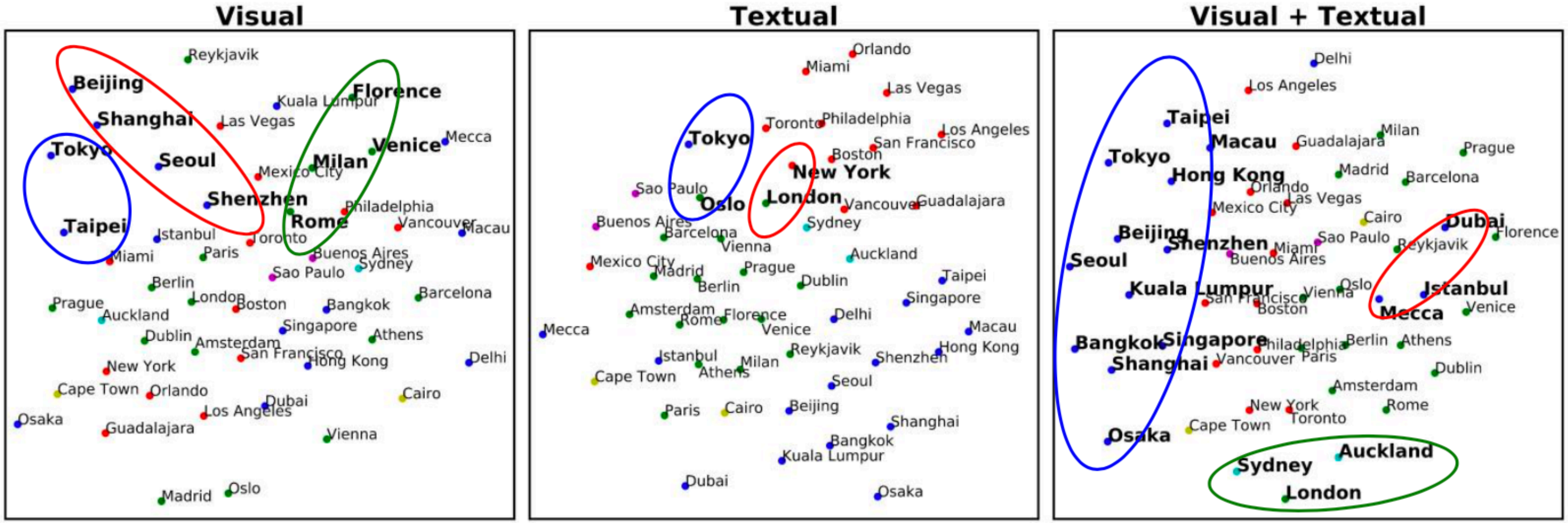
Function Categorization


Ground Truth
City Recognition
Florence
Beijing
Oslo
Taipei
Florence
Florence
Milan
Beijing
Beijing
Shanghai
Singapore
Toronto
Paris
Taipei
Tokyo
Seoul
BIM

City Embedding

- For vision:
  - Using place images
  - Extract the feature from the city recognition model
- For text:
  - Using city descriptions
  - Embed the content of texts into numeric space based on Bert pre-defined model
- For both vision & text: concatenate two vectors above

City Embedding Visualization



City Embedding

- City description
- Calculate the weights
  - Economic
  - Cultural
  - Political

  as in [1]
- Pearson correlation
- Compare with neuron

[1] Son, J.S., Thill, J.C.: Is your city economic, cultural, or political? recognition of city image based on multidimensional scaling of quantified web pages. In: Spatial Analysis and Location Modeling in Urban and Regional Systems, pp. 63–95. Springer (2018)

Conclusion

- A large-scale place dataset
  - comprehensively annotated with multiple aspects
- Explore place understanding
  - Build several benchmarks and study a unified model to recognize places
  - Remains lots of challenges
- Learn city embedding representations
  - Learning from both visual & textual domains can better characterize a city
  - Economic/cultural/political elements could be expressed in different types of images

Placepedia: Comprehensive Place Understanding with Multi-Faceted Annotations

Project page: https://hahehi.github.io/placepedia.html

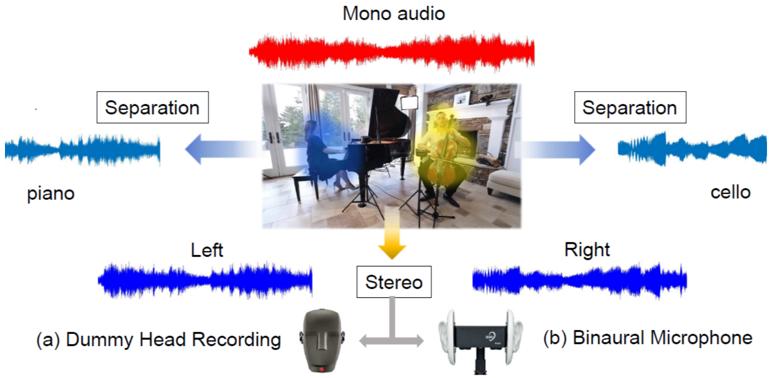
Code and models: https://github.com/hahehi/placepedia

Robust Sensing

CelebA-Spoof

Understanding beyond Recognition

Placepedia



Synthesizing across Modalities

Sep-Stereo

Open World Learning

Open Compound Domain Adaptation

Sep-Stereo: Visually Guided Stereophonic Audio Generation by Associating Source Separation

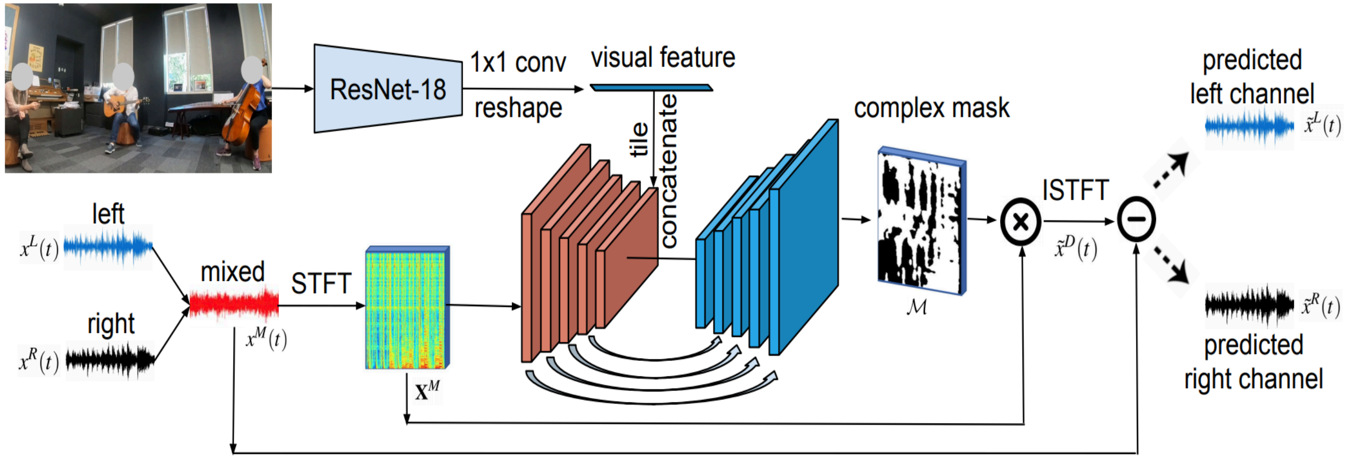



Hang Zhou*, Xudong Xu*, Dahua Lin, Xiaogang Wang, Ziwei Liu

CUHK – SenseTime Joint Lab, The Chinese University of Hong Kong

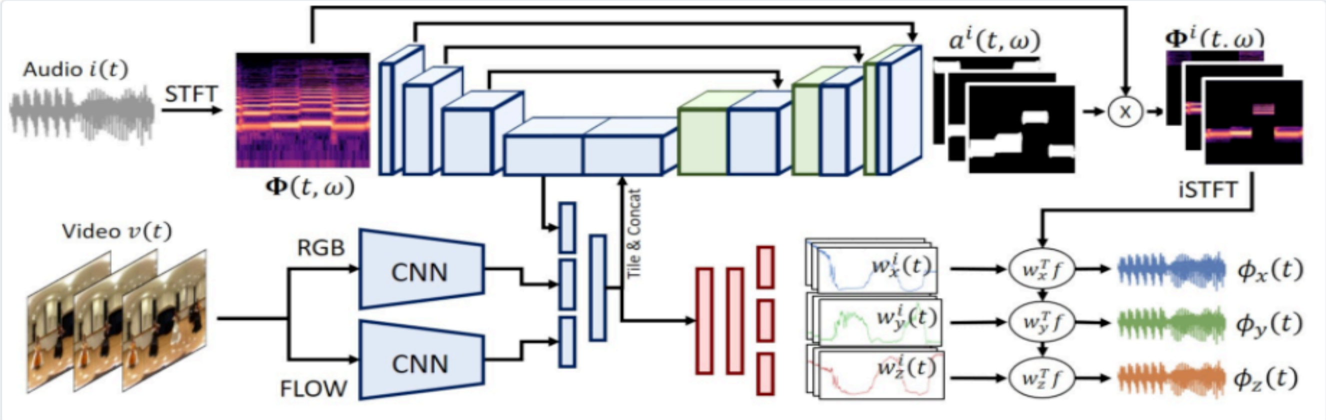
Prior Works: Fully Supervised Training

2.5d Visual Sound (Gao et al, CVPR 2019)



Mono2Binaural relies on self-collected dataset, FAIR-Play, to train the network.

Spatial AudioGen (Morgado et al, NeurIPS 2018)



Spatial AudioGen exploits the stereophonic data on YouTube.

Limitation: Stereophonic Data Collection

Data collection equipment in:
2.5d Visual Sound

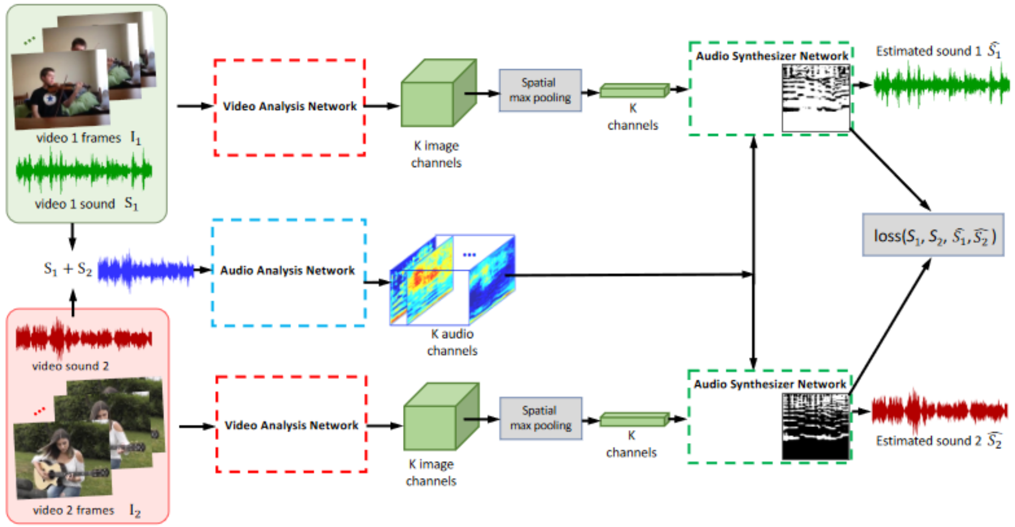
(a) Dummy Head Recording 

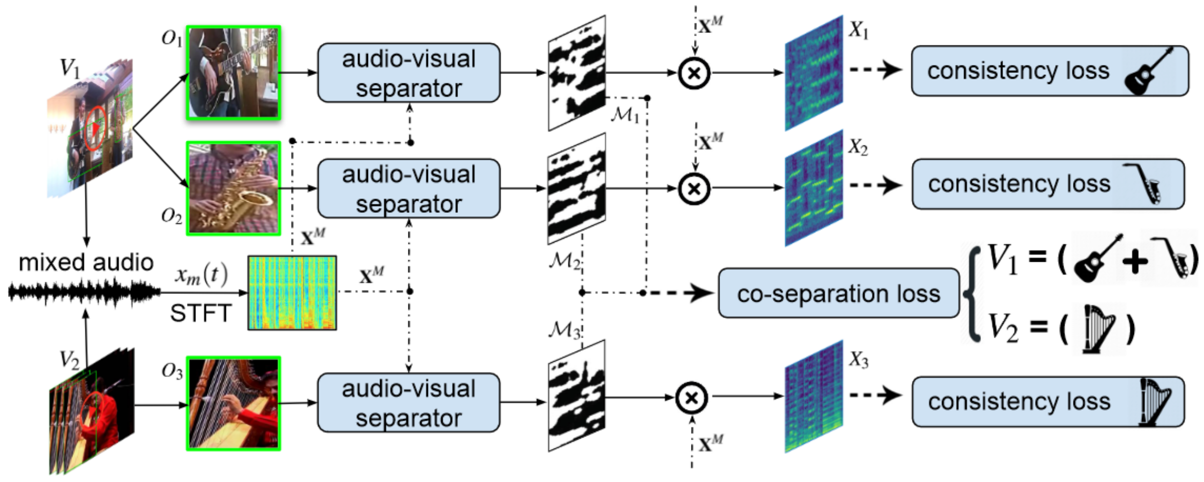
 (b) Binaural Microphone

Mono Audios in Source Separation

The Sound of Pixels
(ECCV 2018)

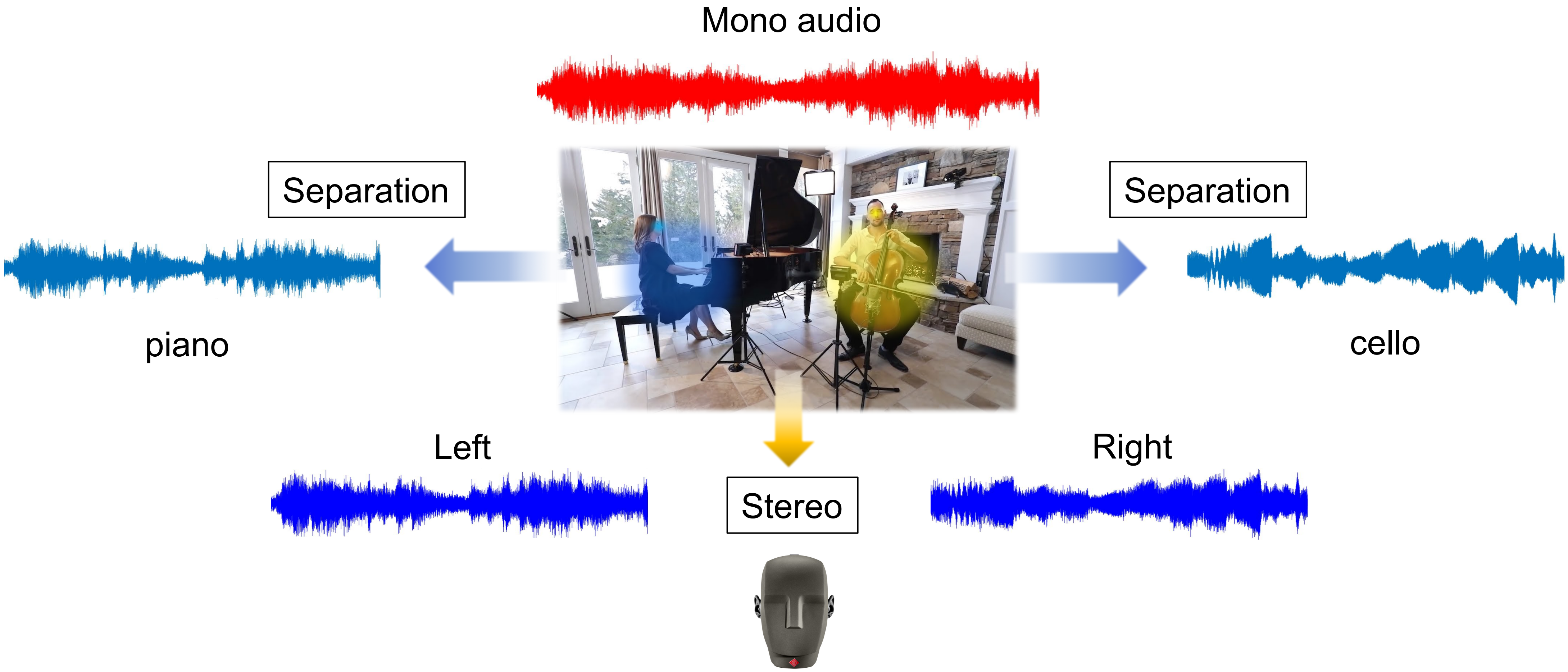
Co-Separating Sounds of Visual Objects
(ICCV 2019)





- Massively availed mono audios have been successfully used in the field of source separation.

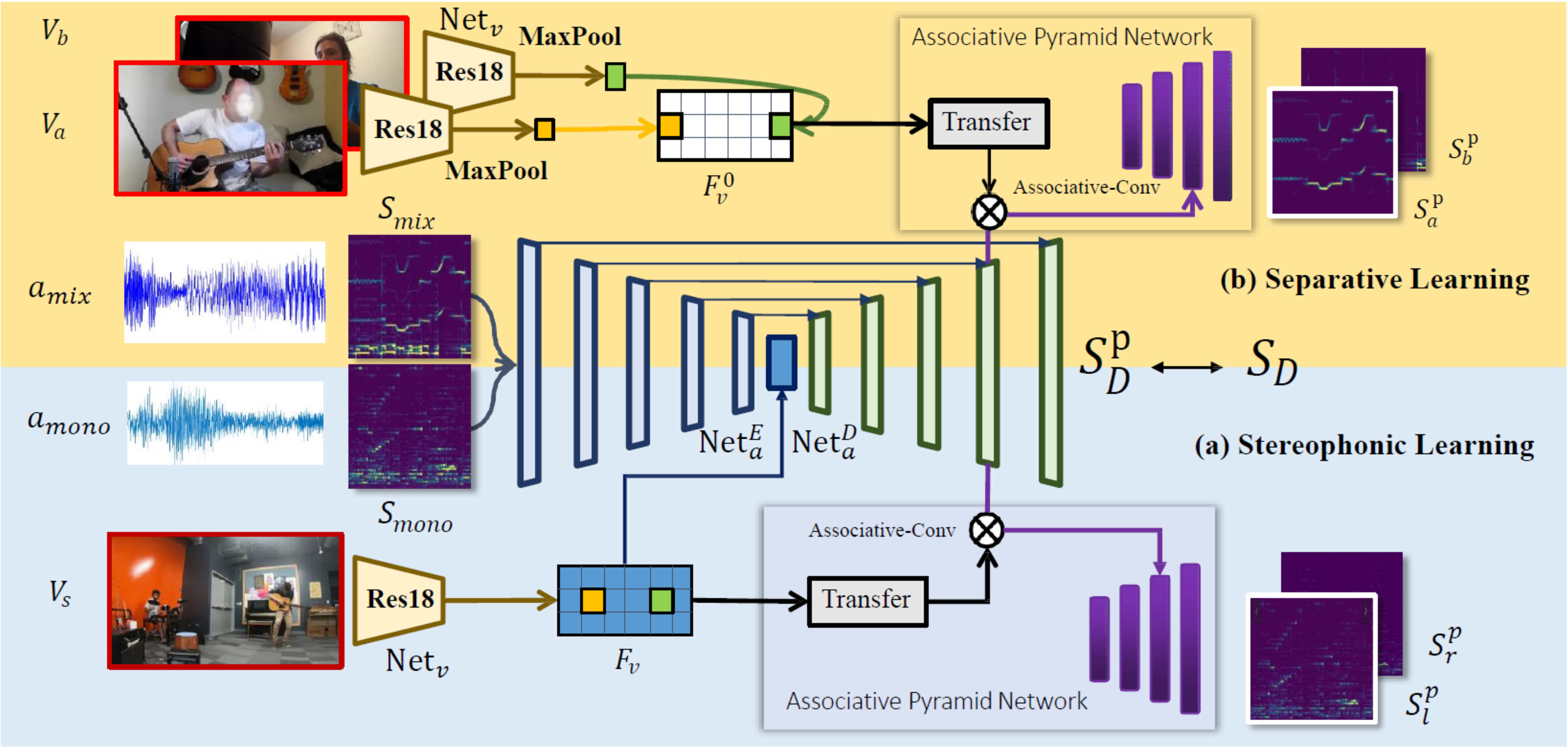
Key Insight: Regard the problem of separating two audios as an extreme case of creating binaural audio



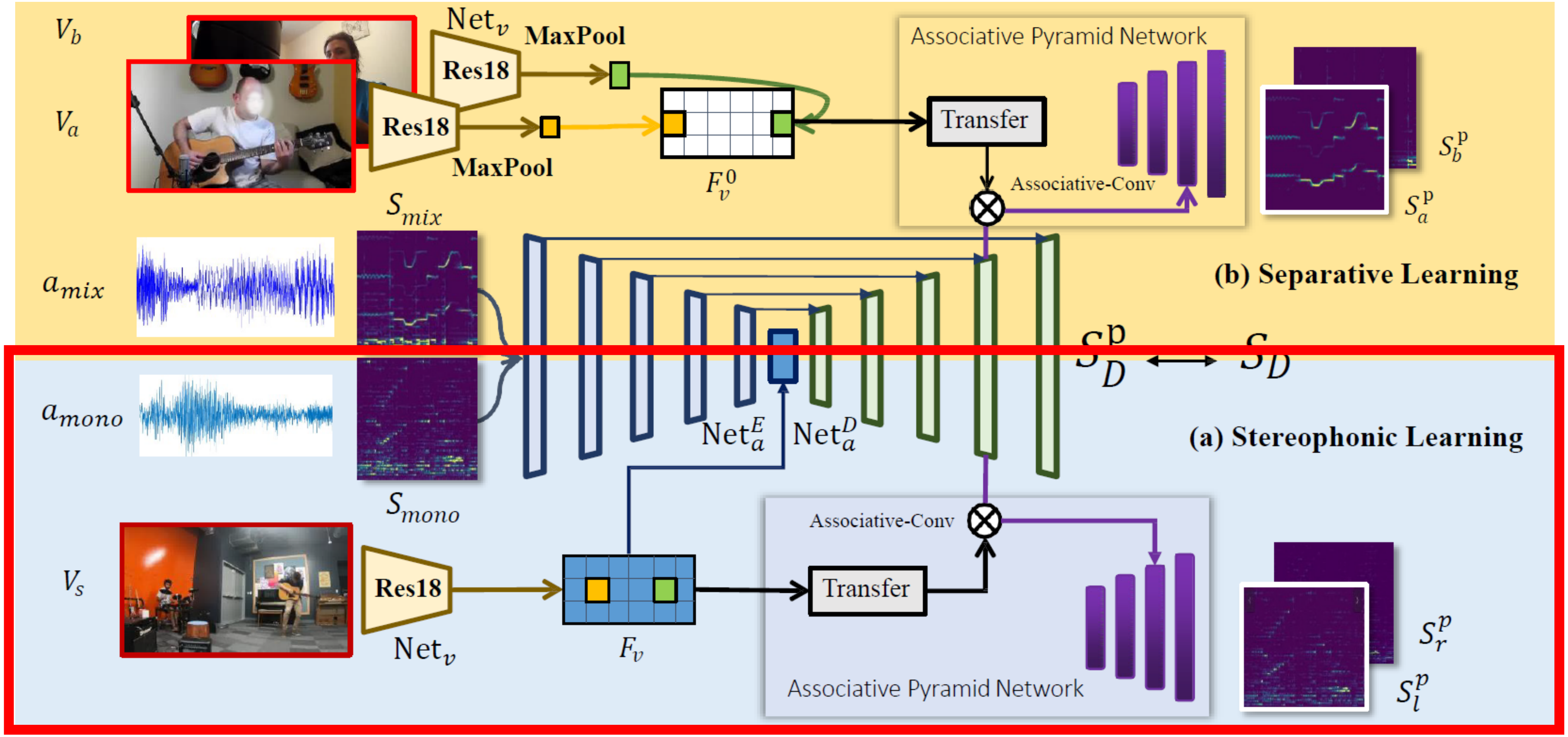
Key Insight: Regard the problem of separating two audios as an extreme case of creating binaural audio

- Mono audios can be used to facilitate the generation of stereo audio.

A Unified Framework: Sep-Stereo



Stereophonic Learning

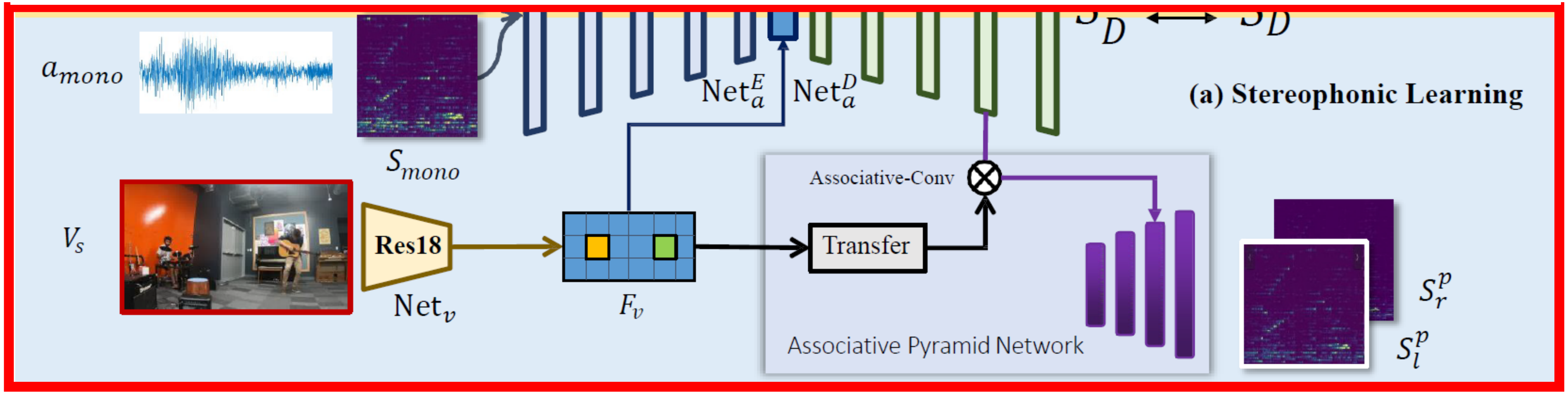


Stereophonic Learning

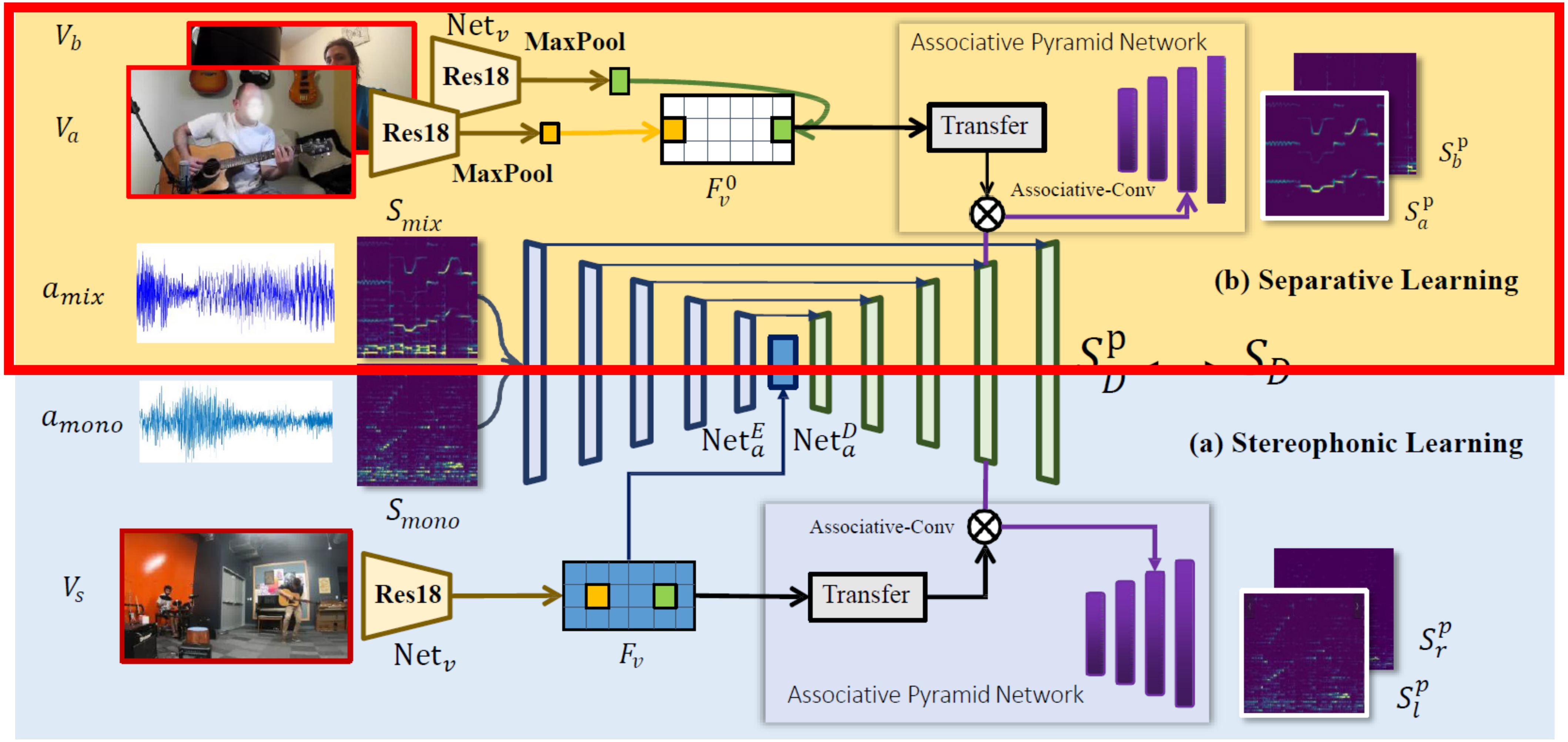
Training Data Settings: Same as 2.5D Visual Sound.

Base Network: Same as 2.5D Visual Sound.

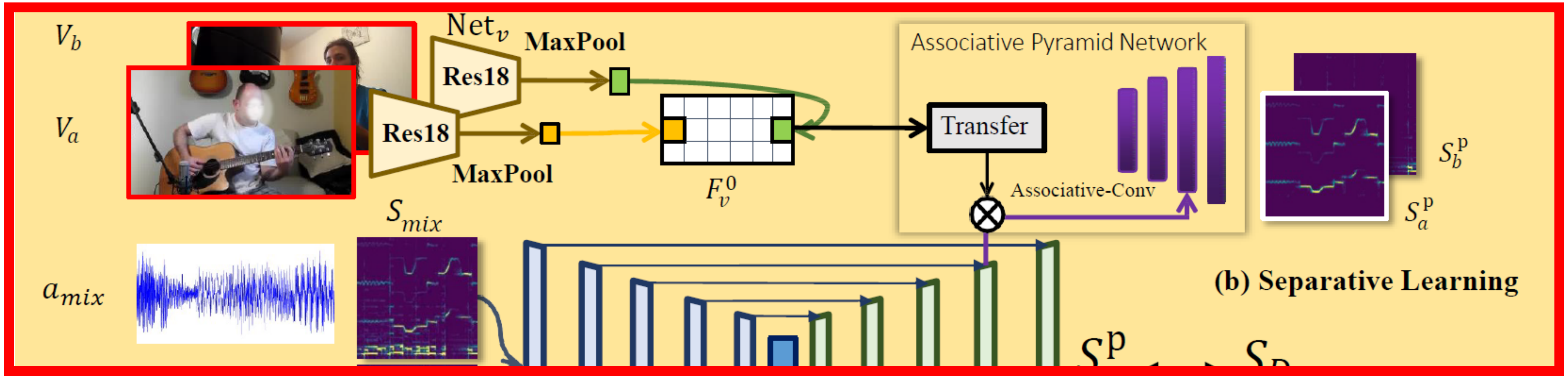
Visual Feature: Kept as 14x7 feature map.



Separative Learning



Separative Learning

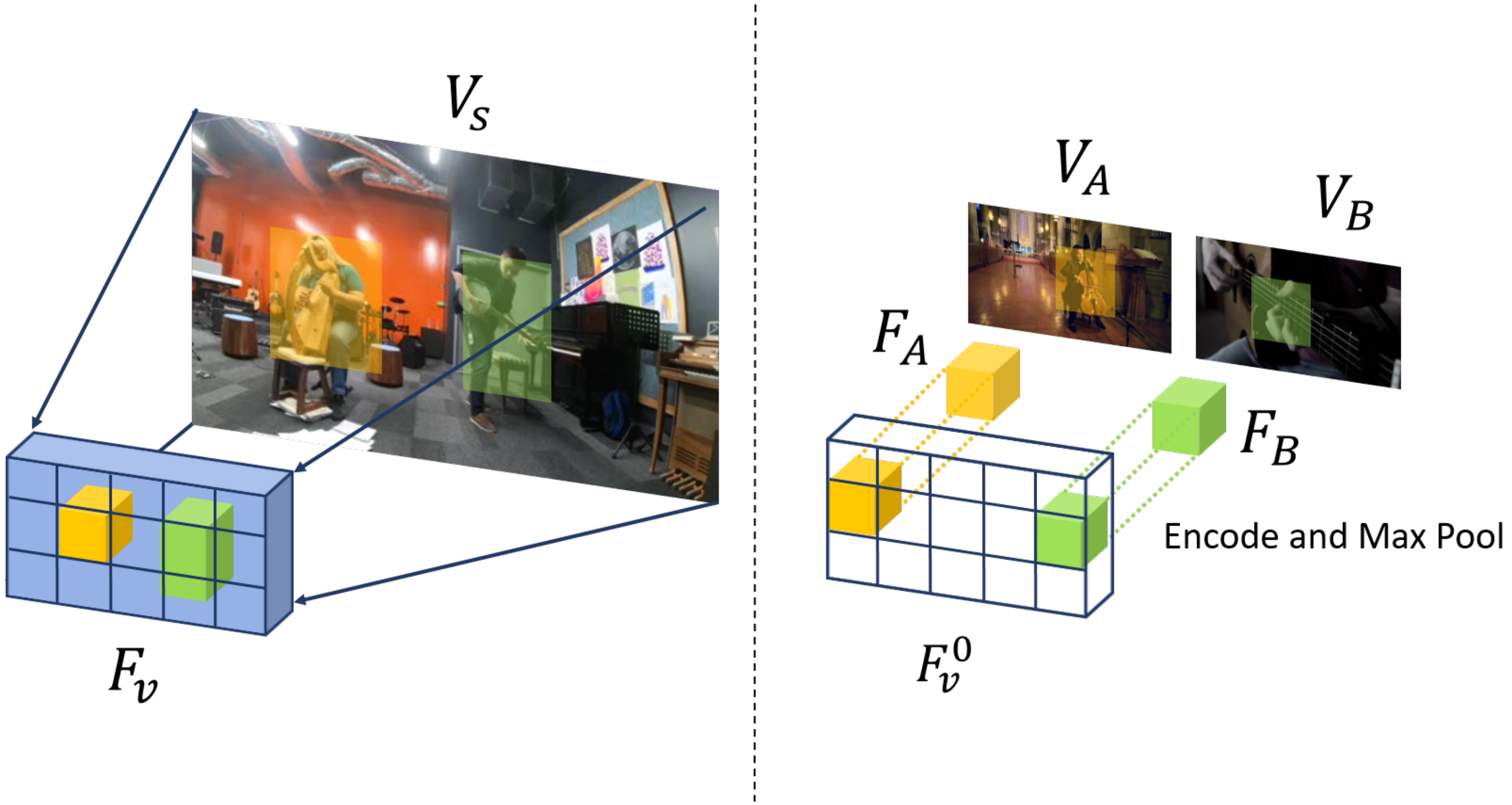


Training Data Settings: Same as Sound of Pixels.

Base Network: Same as 2.5D Visual Sound.

Visual Feature: Max-pooled and rearranged.

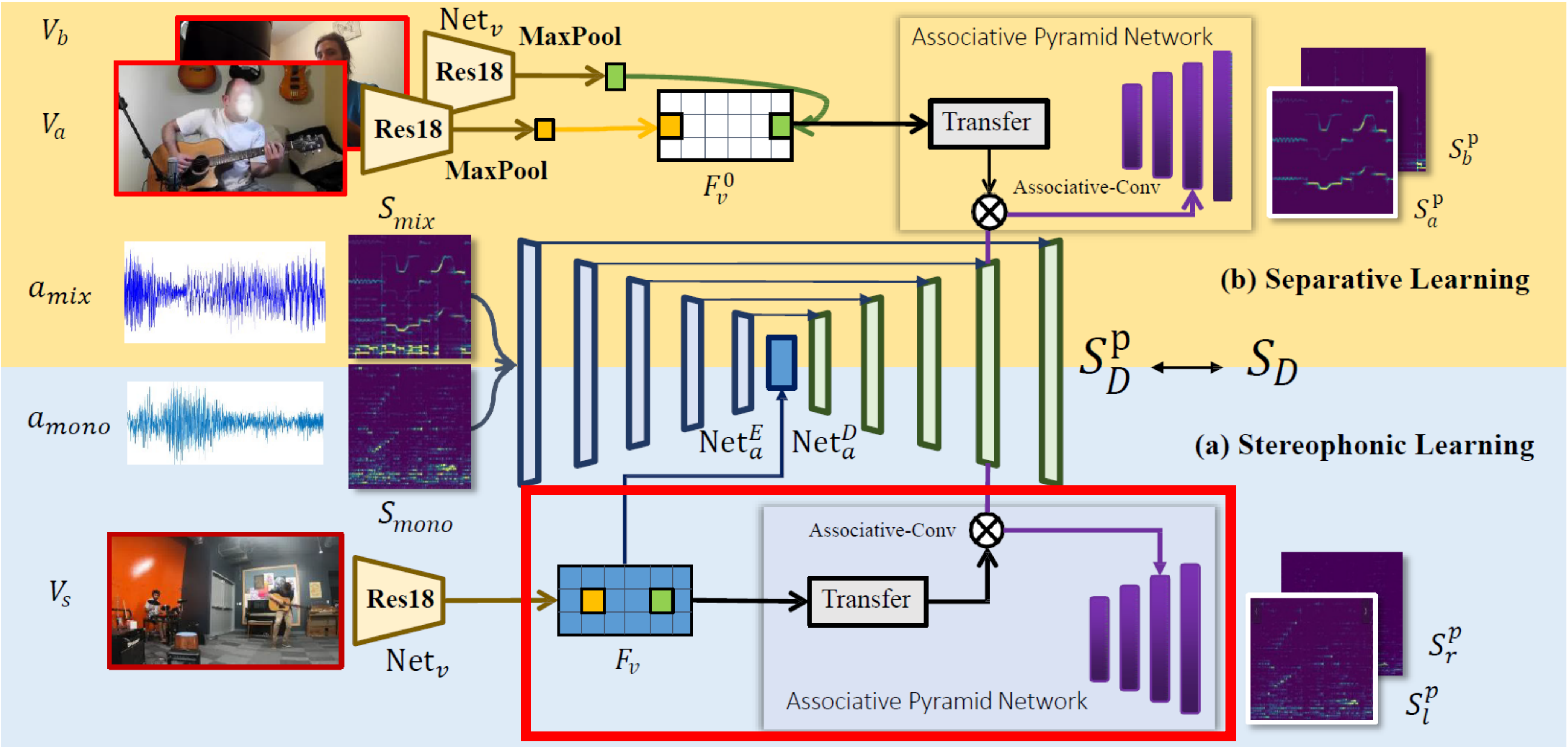
Difference: Visual Feature Rearrangement



Stereophonic Learning

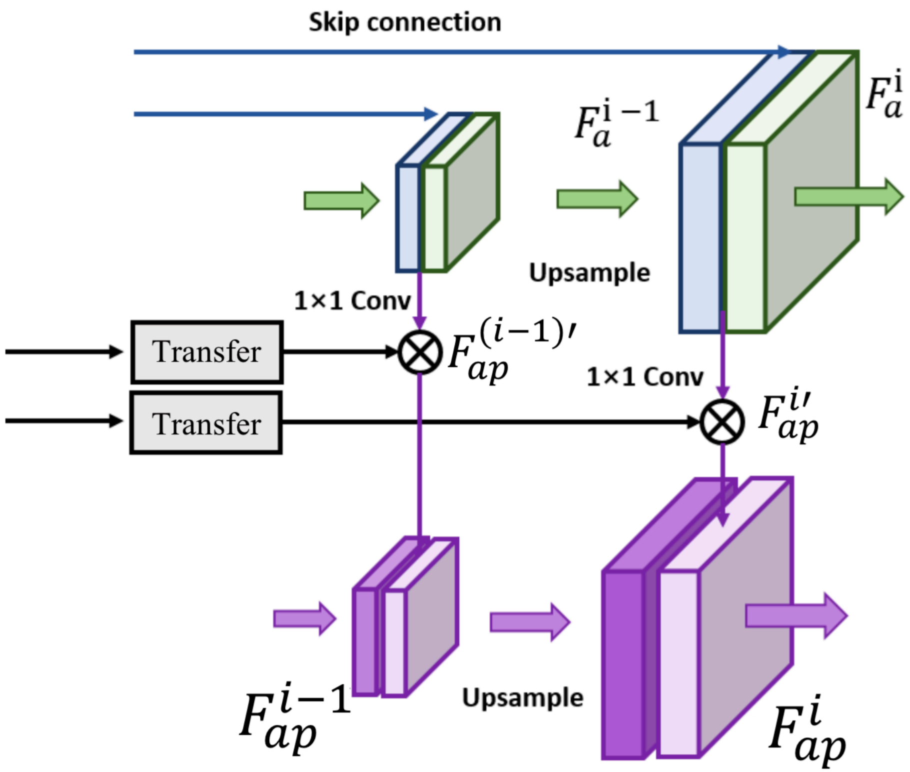
Separative Learning

Associative Pyramid Network



Associative Pyramid Network

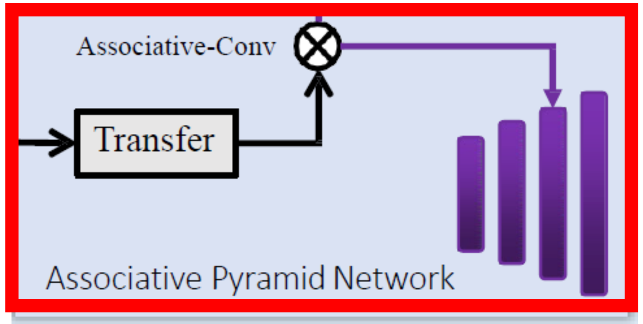
- Associative Pyramid Network better associates the visual features and the audio features with a learned Associative-Conv operation.



(a) APNet Structure

$$F_{ap}^{i'} = \underset{K_v^i}{\mathrm{Conv2d}}(F_a^i)$$

$$F_{ap}^{i} = \mathrm{Cat}([\mathrm{DeConv}(F_{ap}^{i-1}), F_{ap}^{i'}])$$



Mono audio helps stereo audio generation

- Extensive experiments demonstrate that Sep-Stereo can achieve better performance on the task of stereo audio generation with the help of mono audio data.

	Training Data		FAIR-Play		YT-Music	
Method	Stereo	Mono	$STFT_D$	ENV_D	$STFT_D$	ENV_D
Mono2Binaural	✓	✗	0.959	0.141	1.346	0.179
Baseline (MUSIC)	✓	✓	0.930	0.139	1.308	0.175
Assoicative-Conv	✓	✗	0.893	0.137	1.147	0.150
APNet	✓	✗	0.889	0.136	1.070	0.148
Sep-Stereo (Ours)	✓	✓	**0.879**	**0.135**	**1.051**	**0.145**

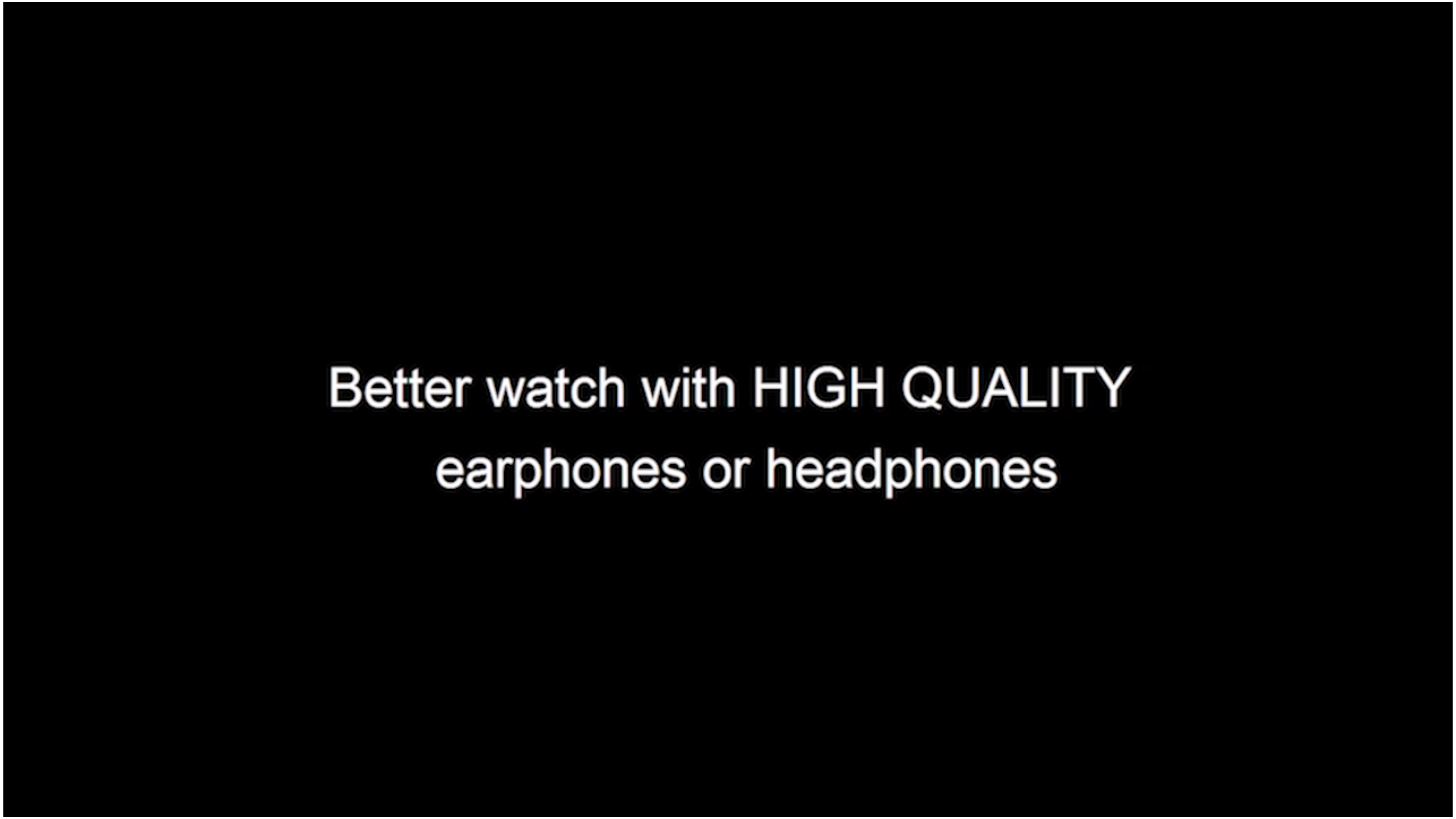
Demo Results

For **stereo**, we compare our results with Mono2Binaural (model of 2.5D Visual Sound), we show the mono input, results from two models and the ground truth.

For **separation**, we compare with PixelPlayer (Sound of Pixels).

Results on MUSIC duets demonstrate the generalization of our method.

Demo Results



Future Work

- Separation:
  - Universal separation, tackling music and speech, even general sound with one model.
  - Adopt the ideas from state-of-the-art audio source separation for the pursuit of models with more capacity.
  - Exploring the task of visually guided audio generation and separation together.
- Stereo:
  - The problem of overfitting still remains unsolved.
  - How to incorporate the setting of the room into the generation of stereo.

Project page: https://hangz-nju-cuhk.github.io/projects/Sep-Stereo

Code and models: https://github.com/SheldonTsui/SepStereo_ECCV2020

Robust Sensing

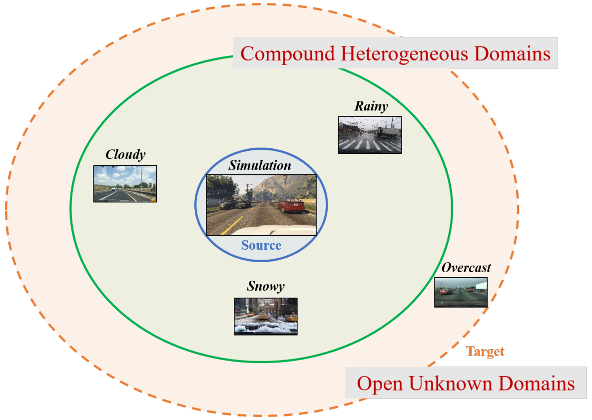
CelebA-Spoof

Understanding beyond Recognition

Placepedia

Synthesizing across Modalities

Sep-Stereo



Open World Learning

Open Compound Domain Adaptation

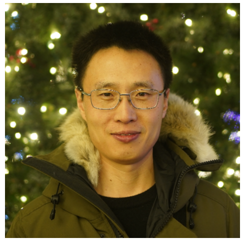
Open Compound Domain Adaptation

Ziwei Liu* Zhongqi Miao* Xingang Pan Xiaohang Zhan Dahua Lin Stella X. Yu Boqing Gong



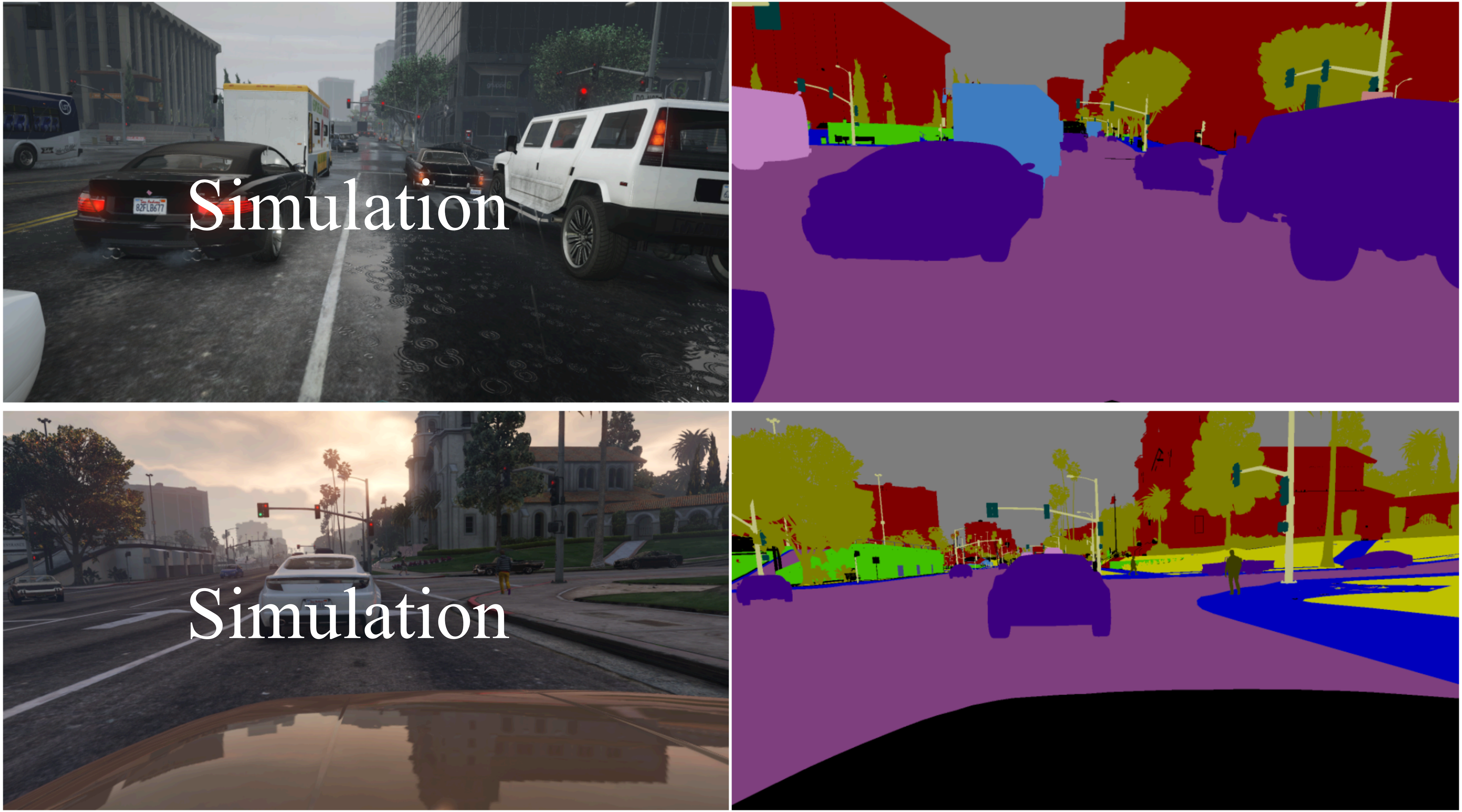






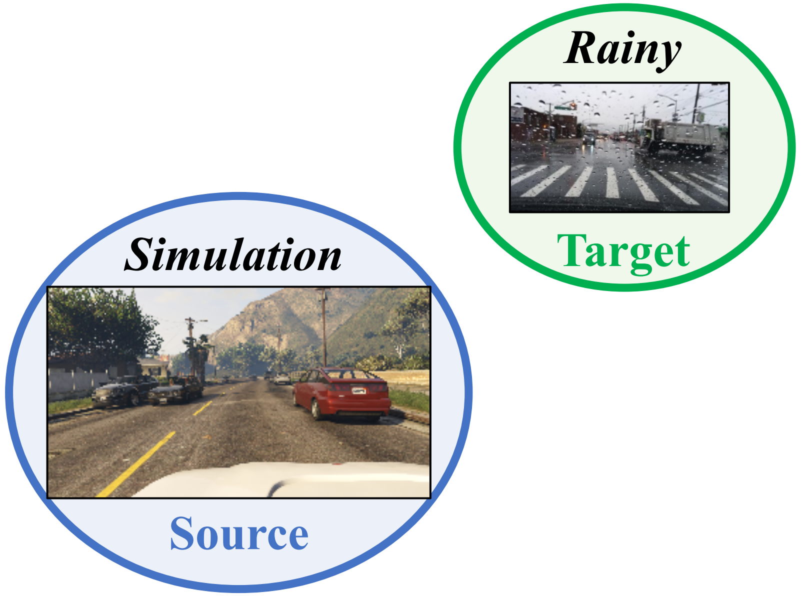
The Chinese University of Hong Kong UC Berkeley / ICSI Google Inc.

Perception for Autonomous Driving

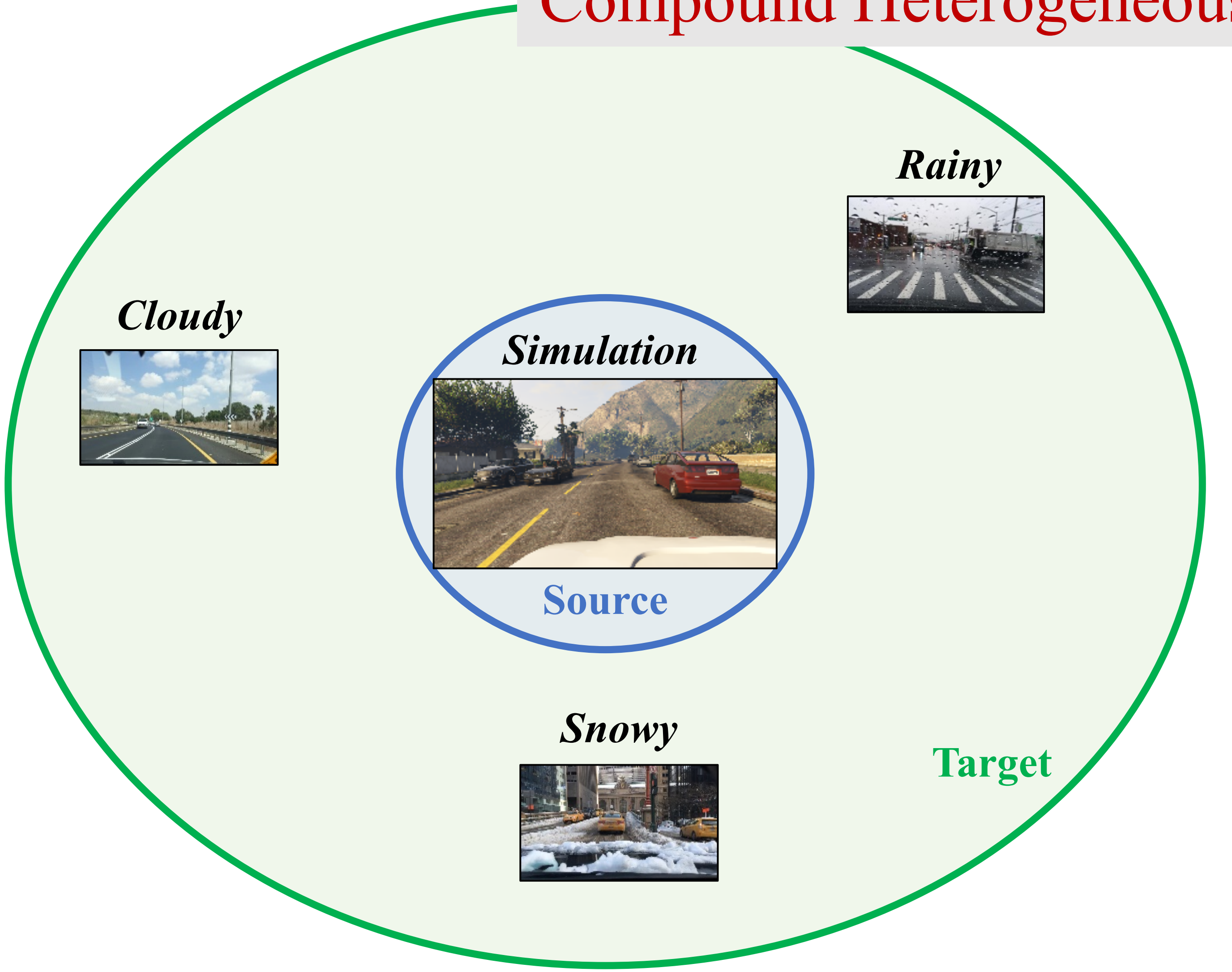

Simulation
Simulation

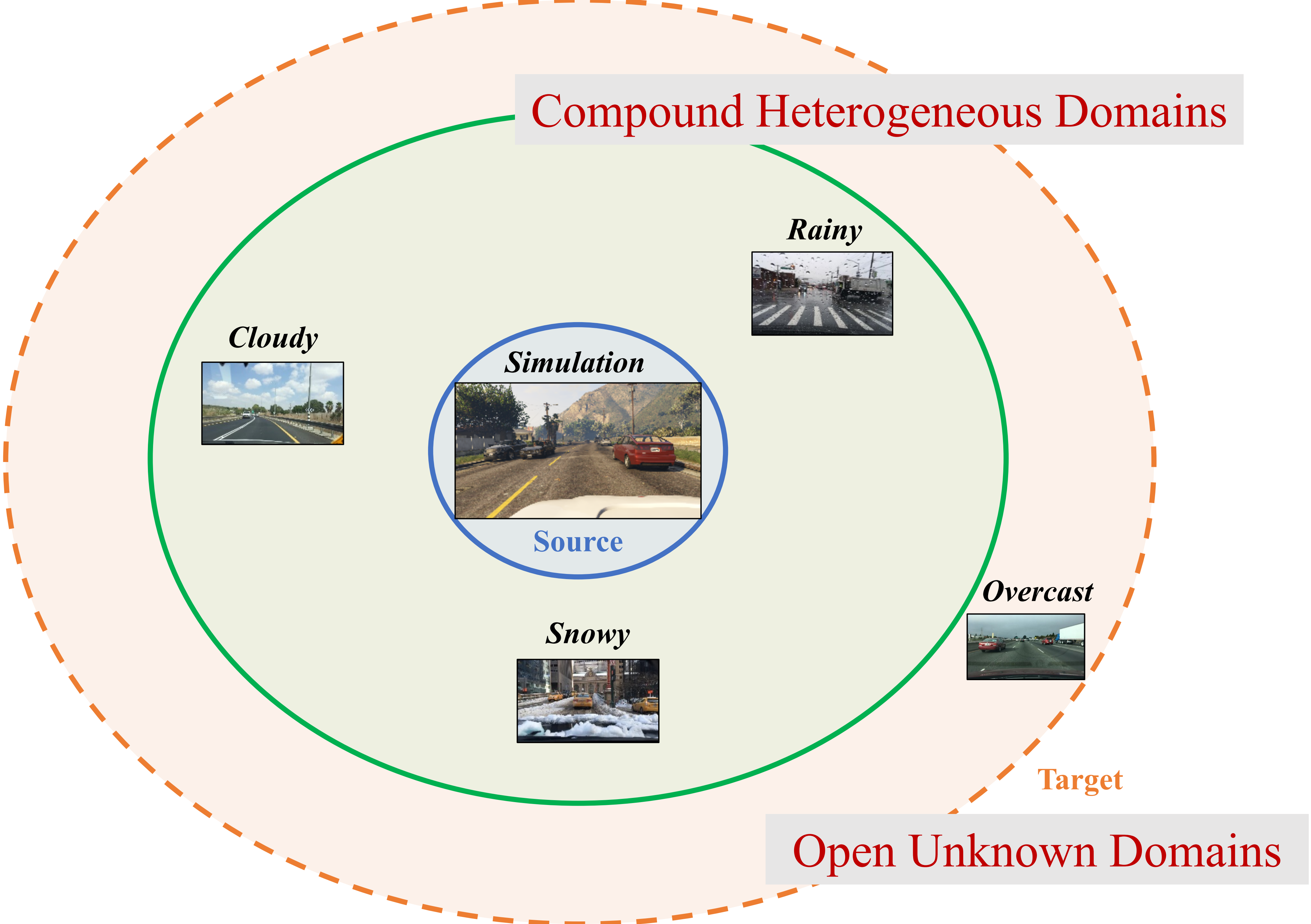
Simulation

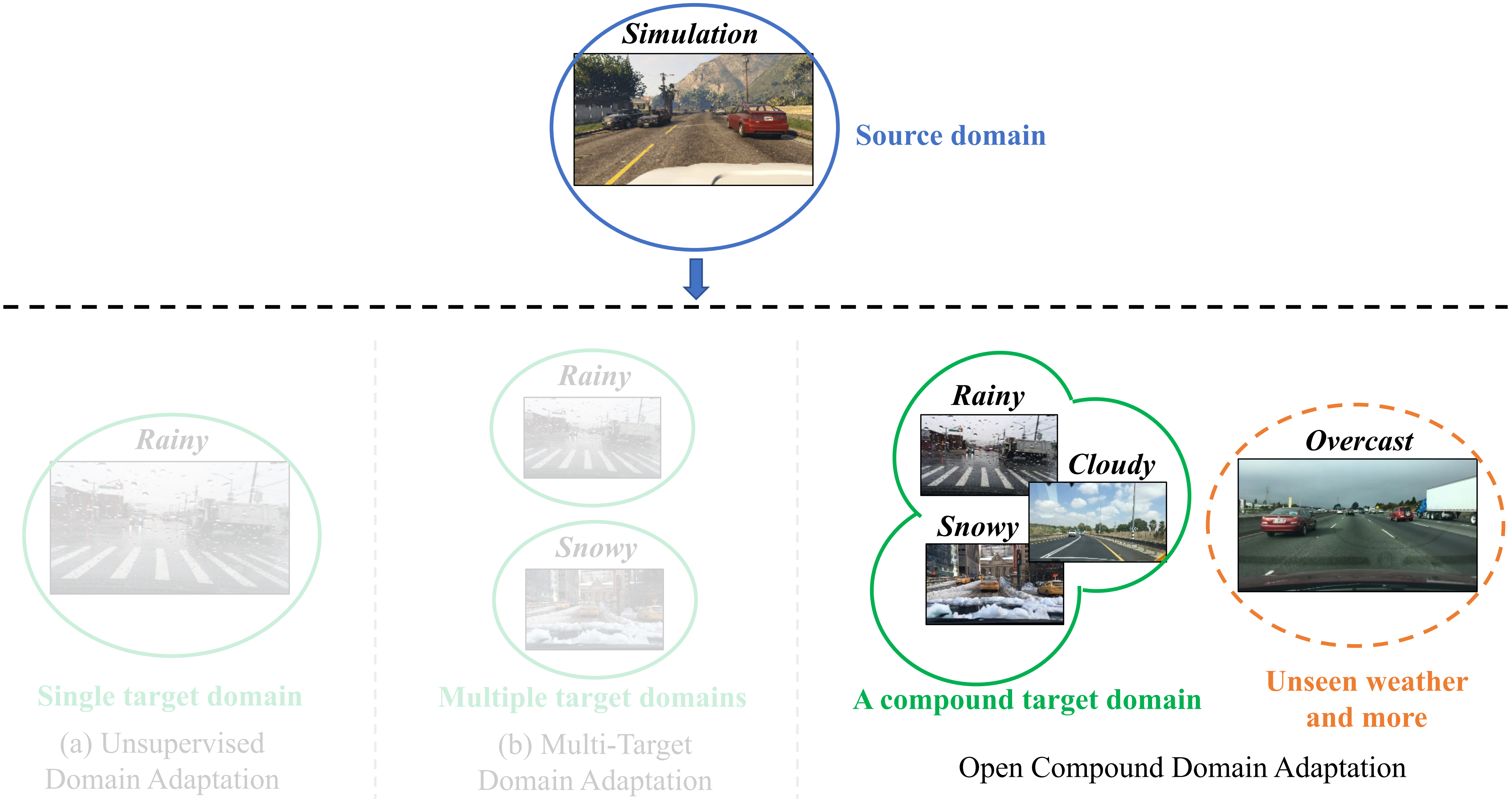
Open World Driving Conditions


Rainy
Target
Simulation
Source

Compound Heterogeneous Domains



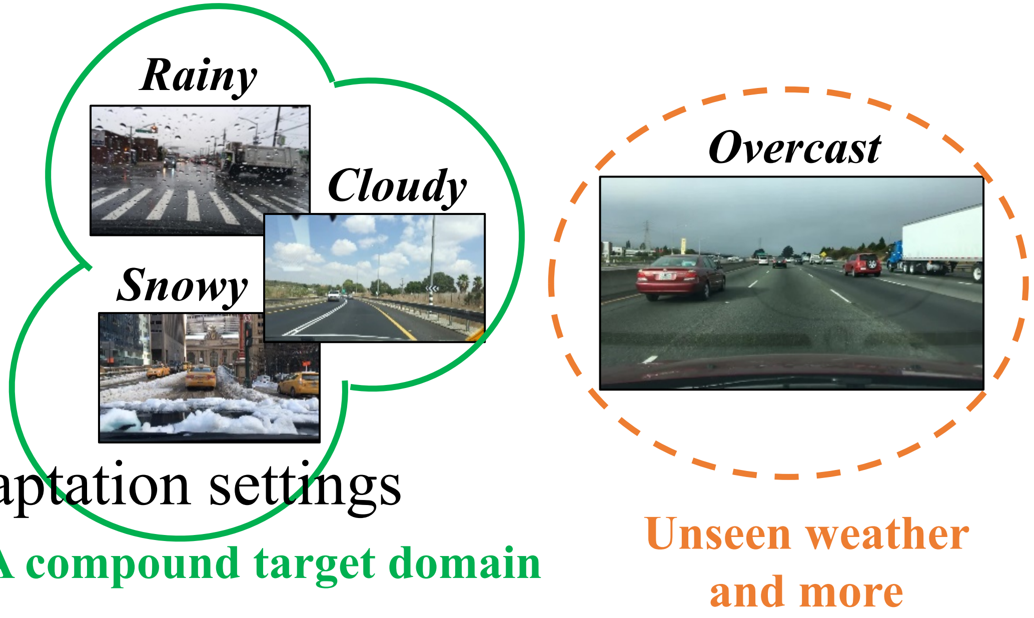

Compound Heterogeneous Domains
Rainy
Cloudy
Simulation
Source
Overcast
Snowy
Target
Open Unknown Domains



Simulation
Source domain
Rainy
Single target domain
(a) Unsupervised Domain Adaptation
Rainy
Snowy
Multiple target domains
(b) Multi-Target Domain Adaptation
Rainy
Cloudy
Snowy
A compound target domain
Overcast
Unseen weather and more
Open Compound Domain Adaptation

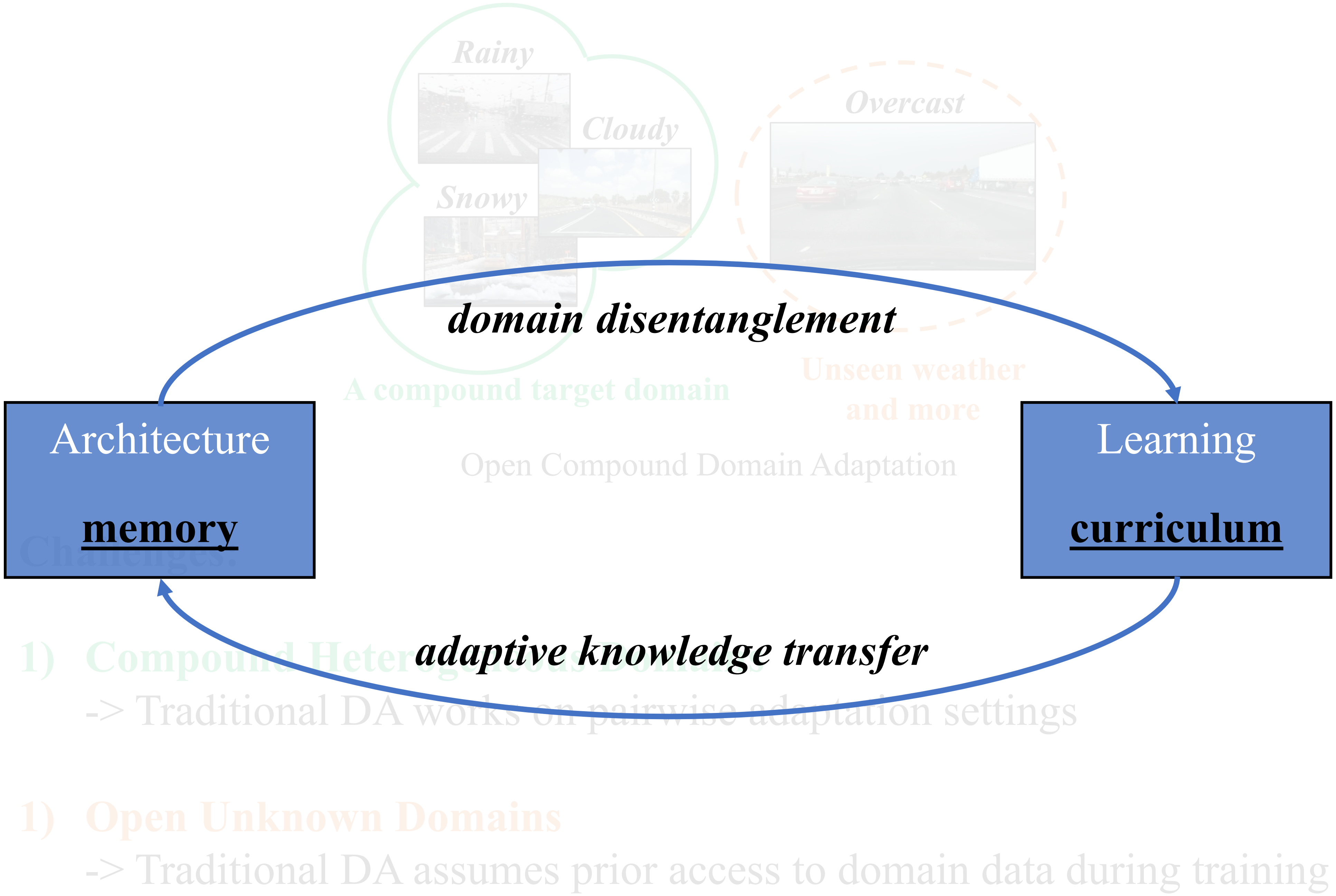
Challenges:

1) **Compound Heterogeneous Domains**
   -> Traditional DA works on pairwise adaptation settings



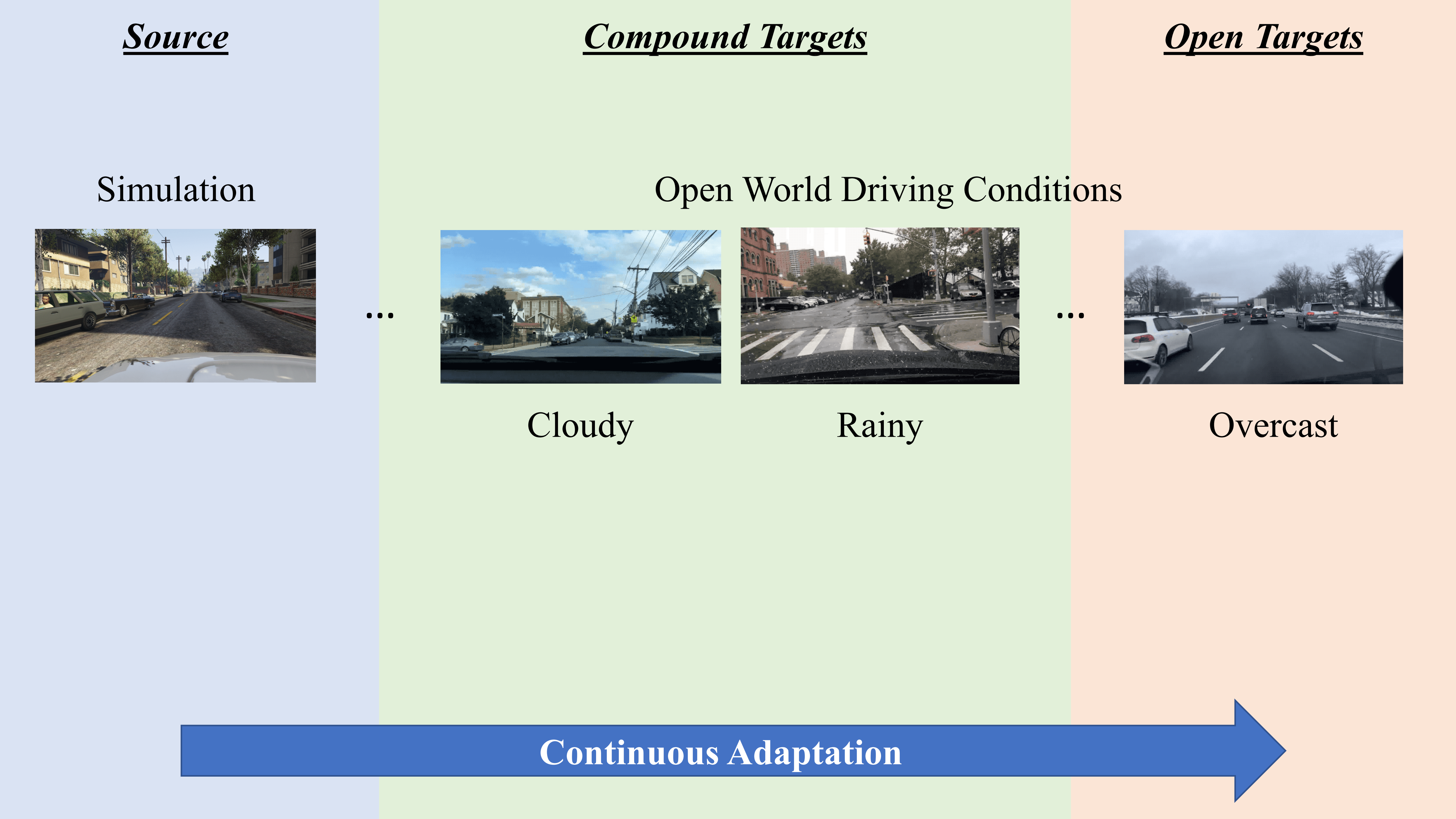
Open Compound Domain Adaptation

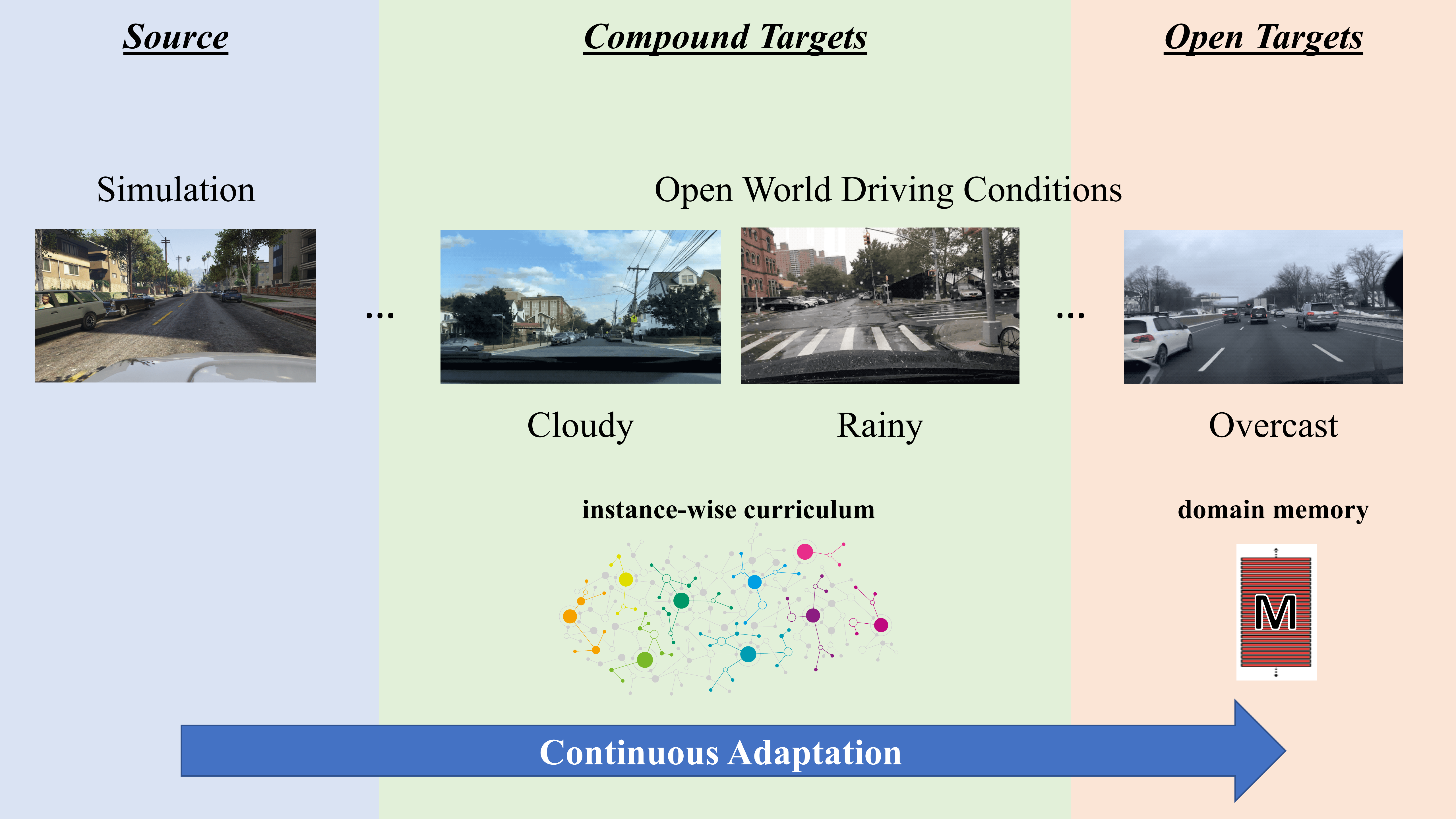
1) **Open Unknown Domains**
   -> Traditional DA assumes prior access to domain data during training

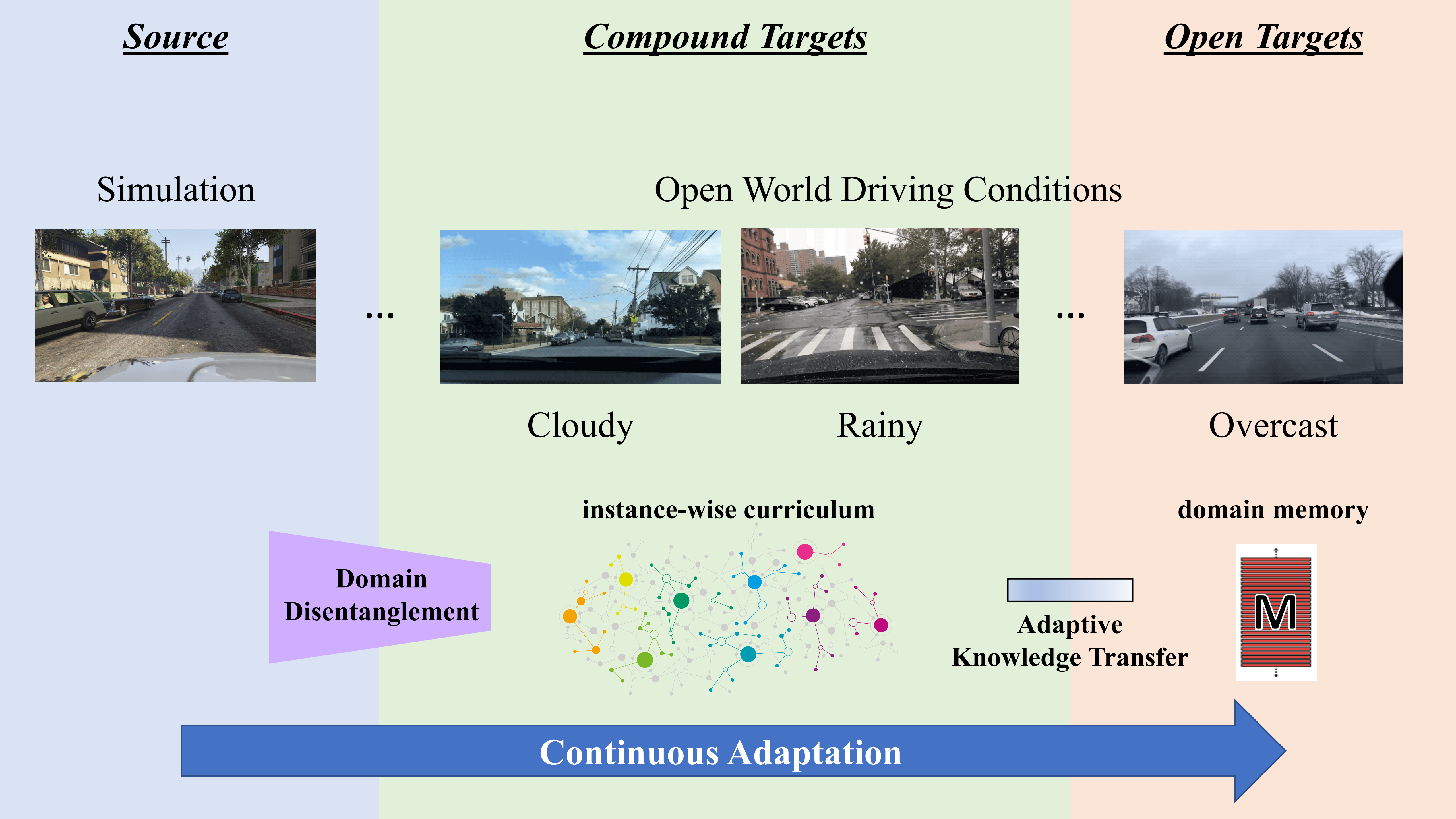

Rainy
Cloudy
Snowy
Overcast
A compound target domain
Unseen weather and more
Open Compound Domain Adaptation
domain disentanglement
Architecture
memory
Learning
curriculum
adaptive knowledge transfer
1) Open Unknown Domains
-> Traditional DA assumes prior access to domain data during training

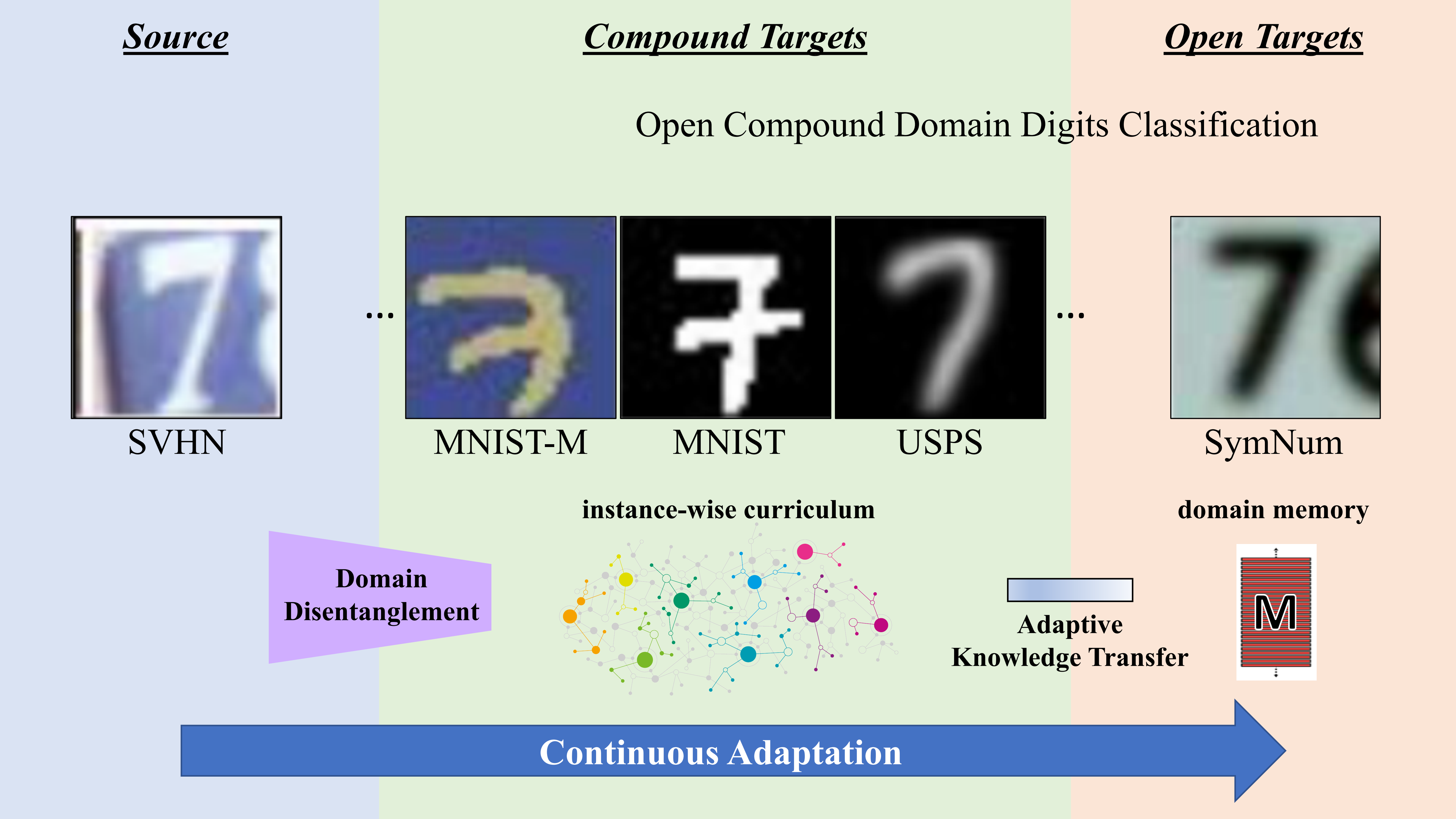
Simulation

Open World Driving Conditions


Source
Compound Targets
Open Targets
Simulation
Open World Driving Conditions
...
...
Cloudy
Rainy
Overcast
Continuous Adaptation


Source
Compound Targets
Open Targets
Simulation
Open World Driving Conditions
...
...
Cloudy
Rainy
Overcast
instance-wise curriculum
domain memory
M
Continuous Adaptation

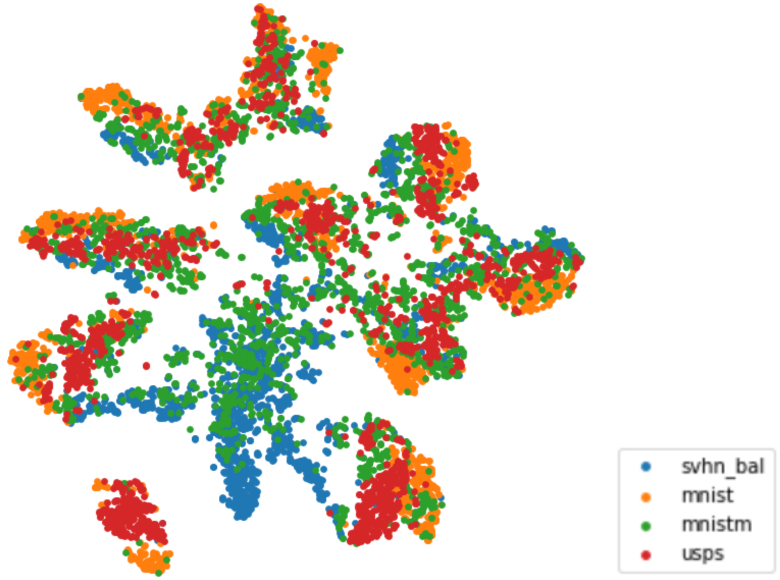

Source
Compound Targets
Open Targets
Simulation
Open World Driving Conditions
...
...
Cloudy
Rainy
Overcast
instance-wise curriculum
domain memory
Domain
Disentanglement
Adaptive
Knowledge Transfer
M
Continuous Adaptation

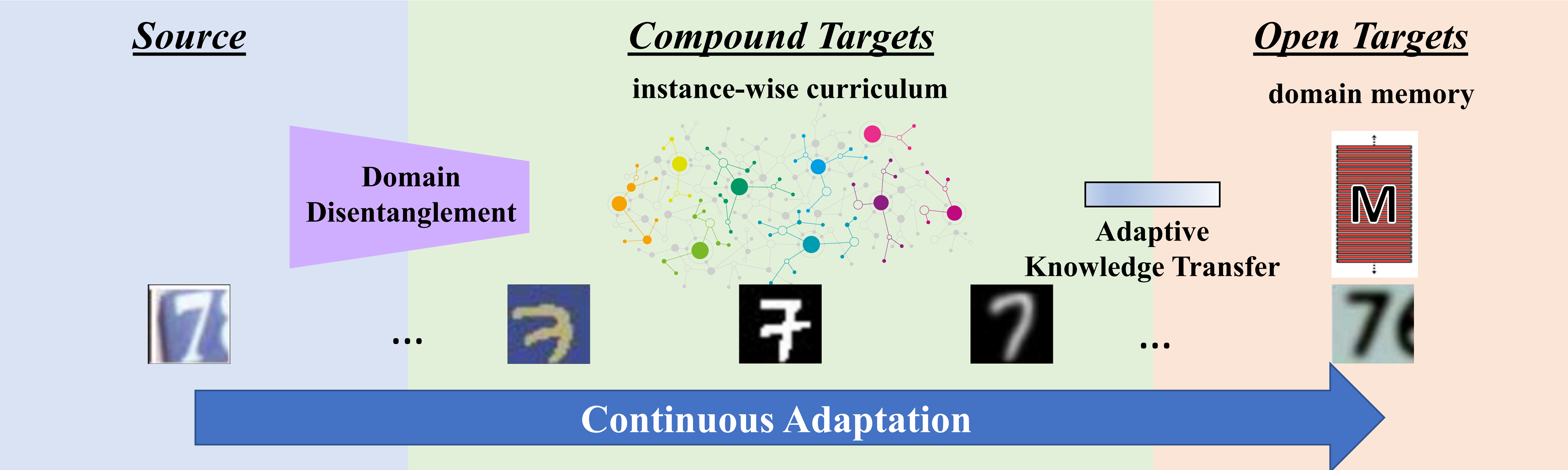

Source
Compound Targets
Open Targets
Open Compound Domain Digits Classification
...
...
SVHN
MNIST-M
MNIST
USPS
SymNum
instance-wise curriculum
domain memory
Domain Disentanglement
Adaptive Knowledge Transfer
M
Continuous Adaptation

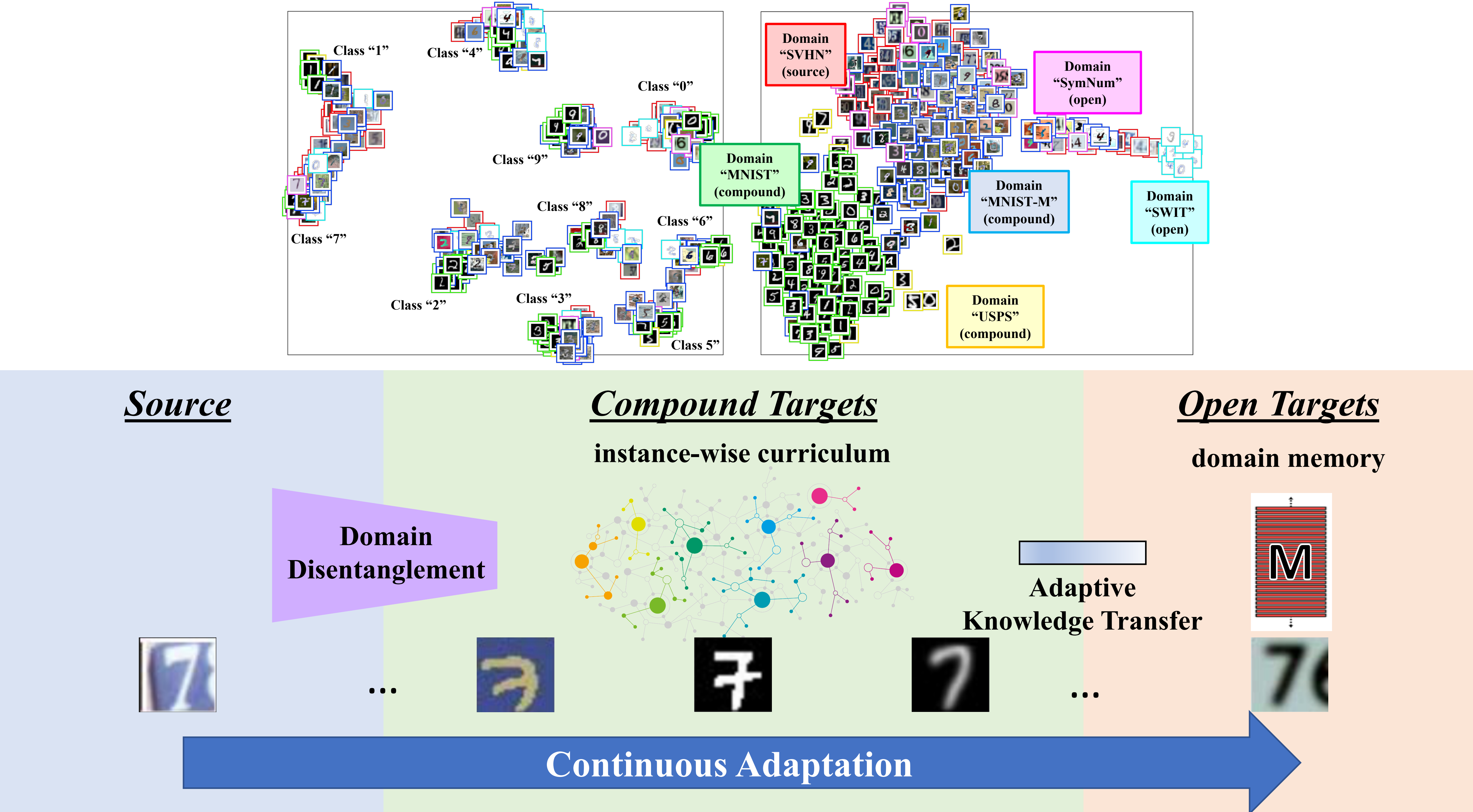
Adversarial Domain Characteristics Disentanglement

$$\min_{E_{domain}} \quad -\sum_i z^i_{random} \log D(E_{domain}(x^i))$$

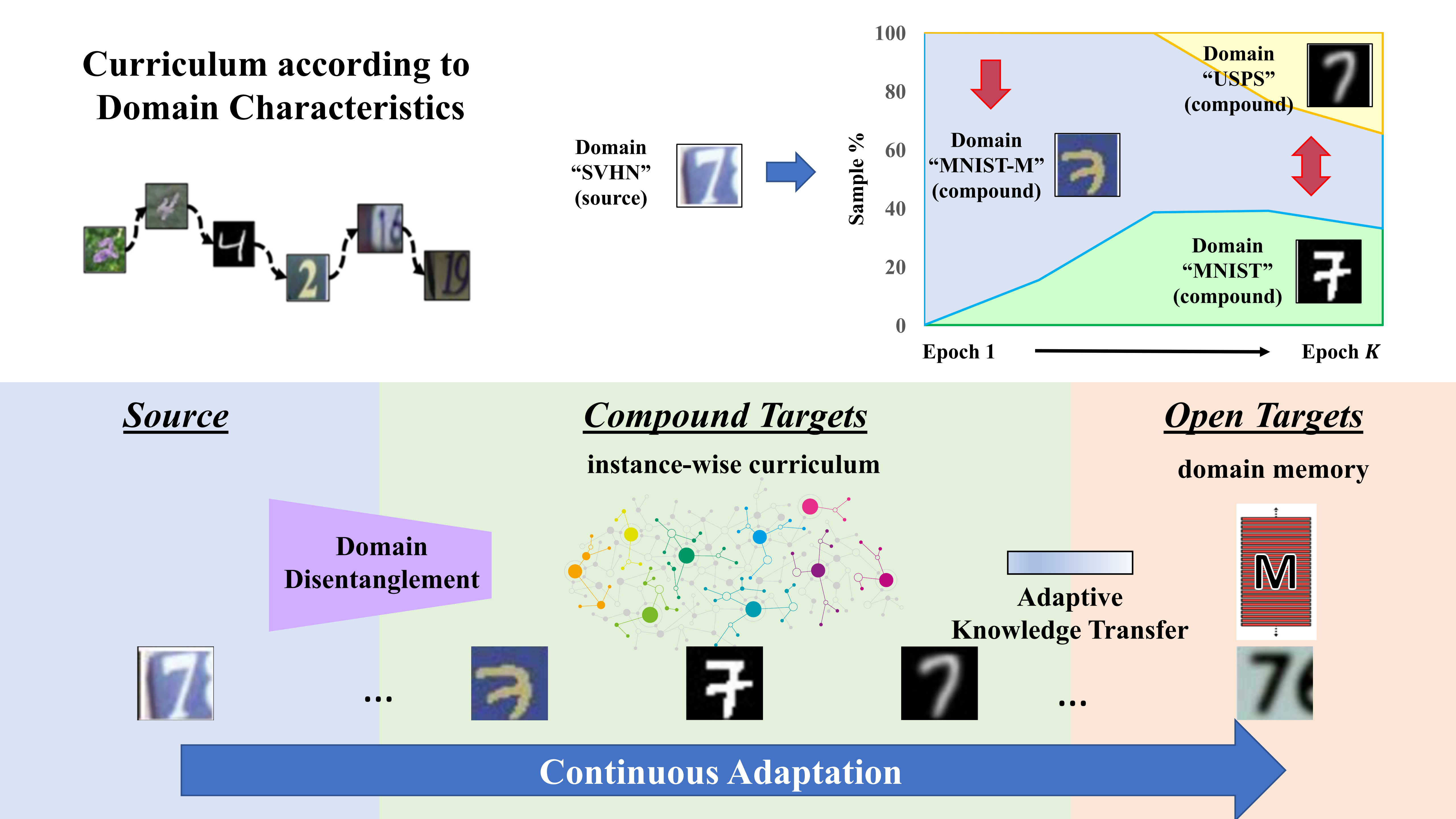
$$\min_{D} \quad -\sum_i y^i \log D(E_{domain}(x^i))$$

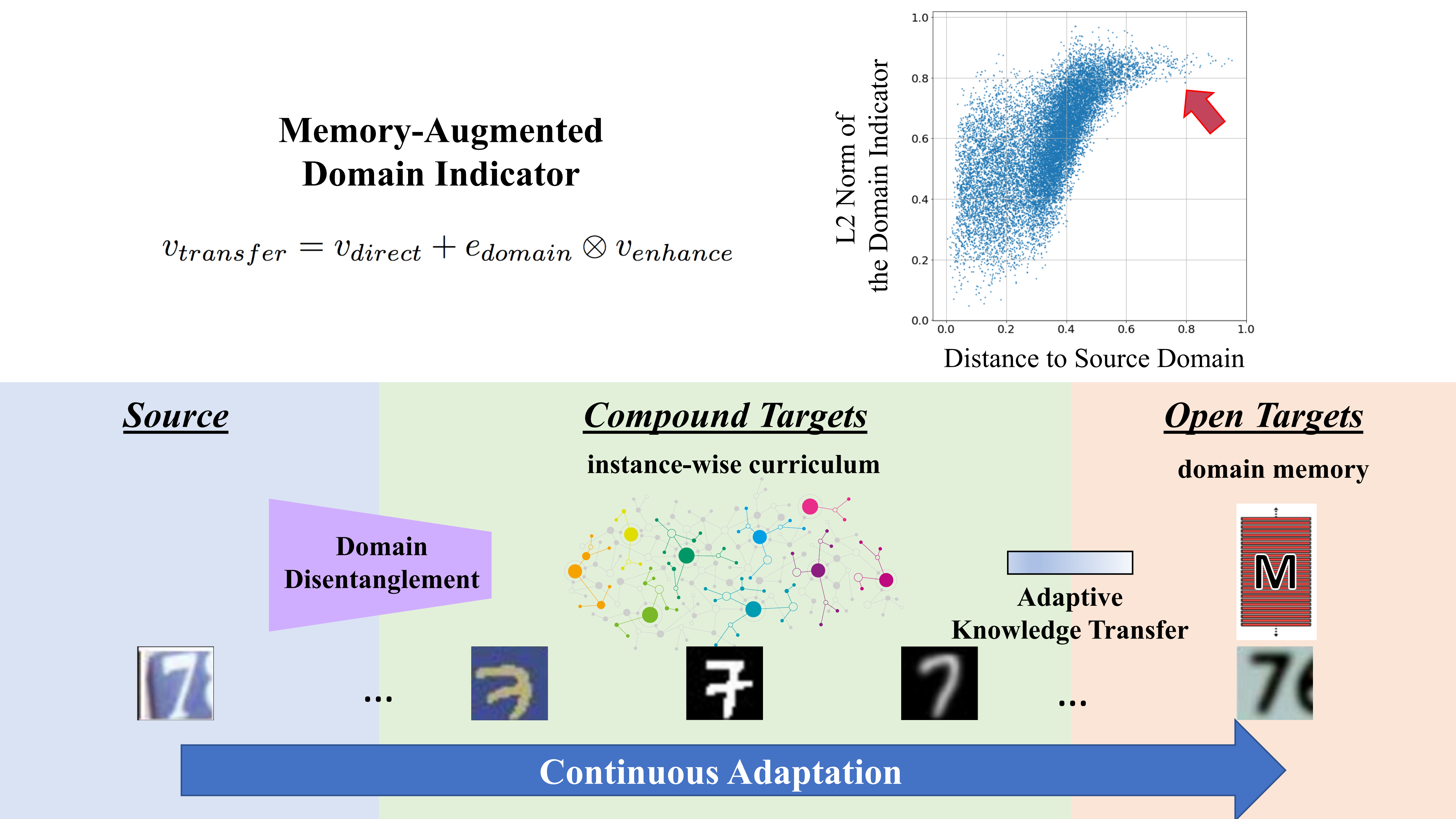






Class "1"
Class "4"
Class "0"
Class "9"
Class "7"
Class "8"
Class "6"
Class "2"
Class "3"
Class 5"
Domain "SVHN" (source)
Domain "SymNum" (open)
Domain "MNIST" (compound)
Domain "MNIST-M" (compound)
Domain "SWIT" (open)
Domain "USPS" (compound)
Source
Compound Targets
instance-wise curriculum
Open Targets
domain memory
Domain Disentanglement
Adaptive Knowledge Transfer
M
Continuous Adaptation


Curriculum according to Domain Characteristics
Domain "SVHN" (source)
Sample %
100
80
60
40
20
0
Domain "USPS" (compound)
Domain "MNIST-M" (compound)
Domain "MNIST" (compound)
Epoch 1
Epoch K
Source
Compound Targets
instance-wise curriculum
Open Targets
domain memory
Domain Disentanglement
Adaptive Knowledge Transfer
M
…
…
Continuous Adaptation

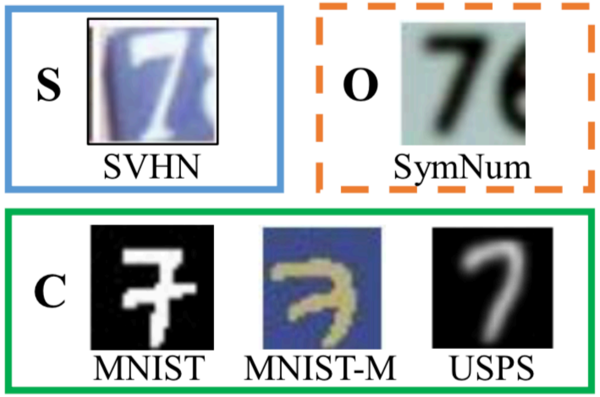

Memory-Augmented
Domain Indicator
$v_{transfer} = v_{direct} + e_{domain} \otimes v_{enhance}$
L2 Norm of
the Domain Indicator
Distance to Source Domain
0.0
0.2
0.4
0.6
0.8
1.0
Source
Compound Targets
Open Targets
instance-wise curriculum
domain memory
Domain
Disentanglement
Adaptive
Knowledge Transfer
M
...
Continuous Adaptation

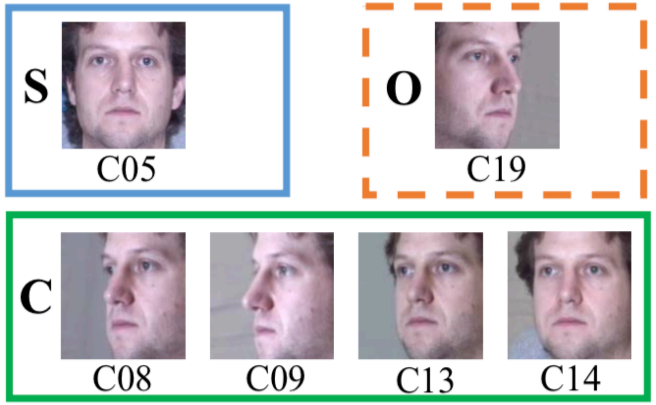
C-Digits Benchmark

Absolute Performance Gain: ~5%

C-Faces Benchmark

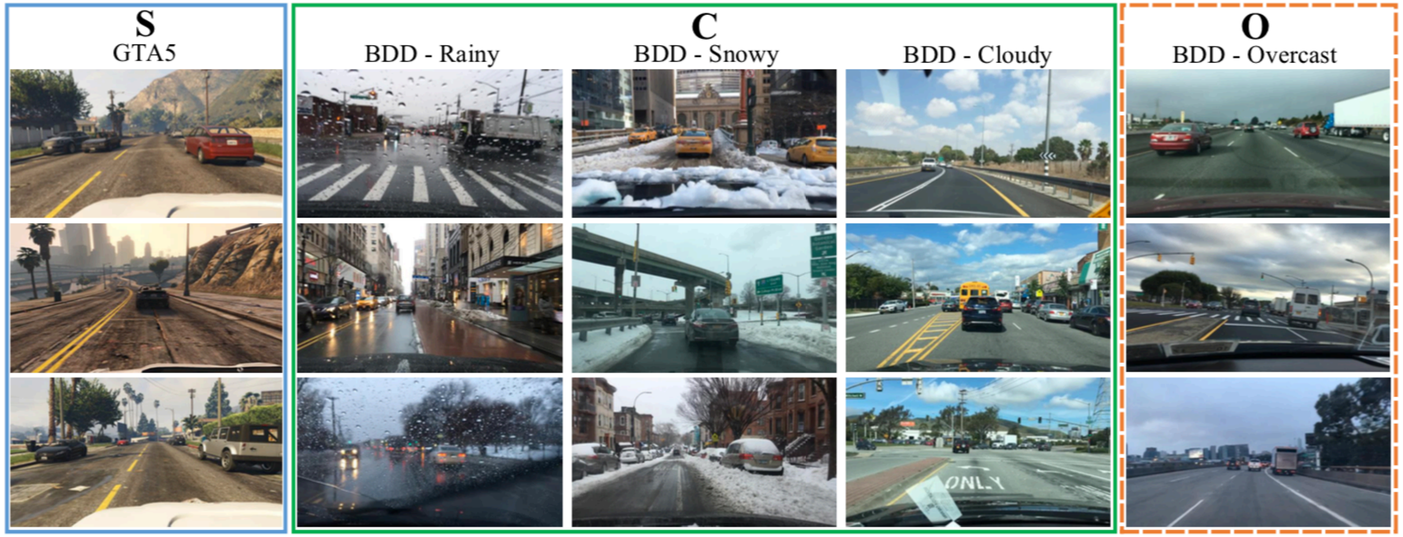
Absolute Performance Gain: ~10%





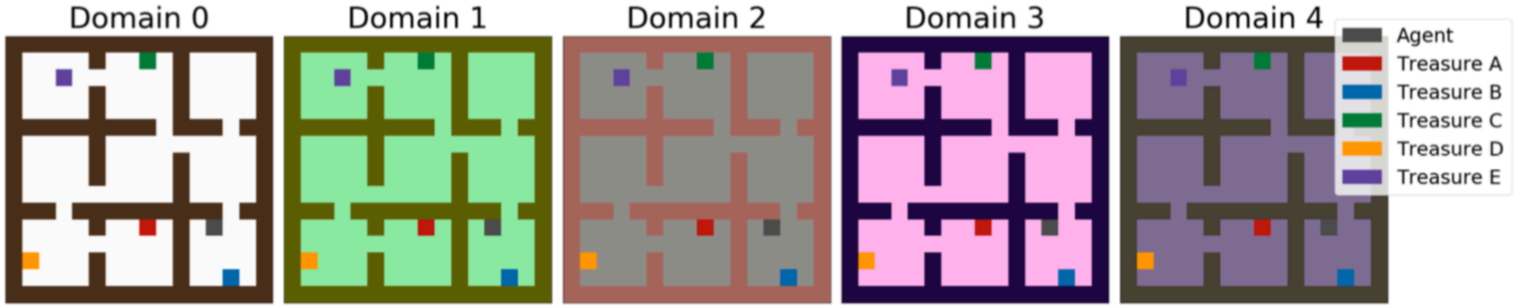
C-Driving Benchmark

Absolute Performance Gain: ~2%

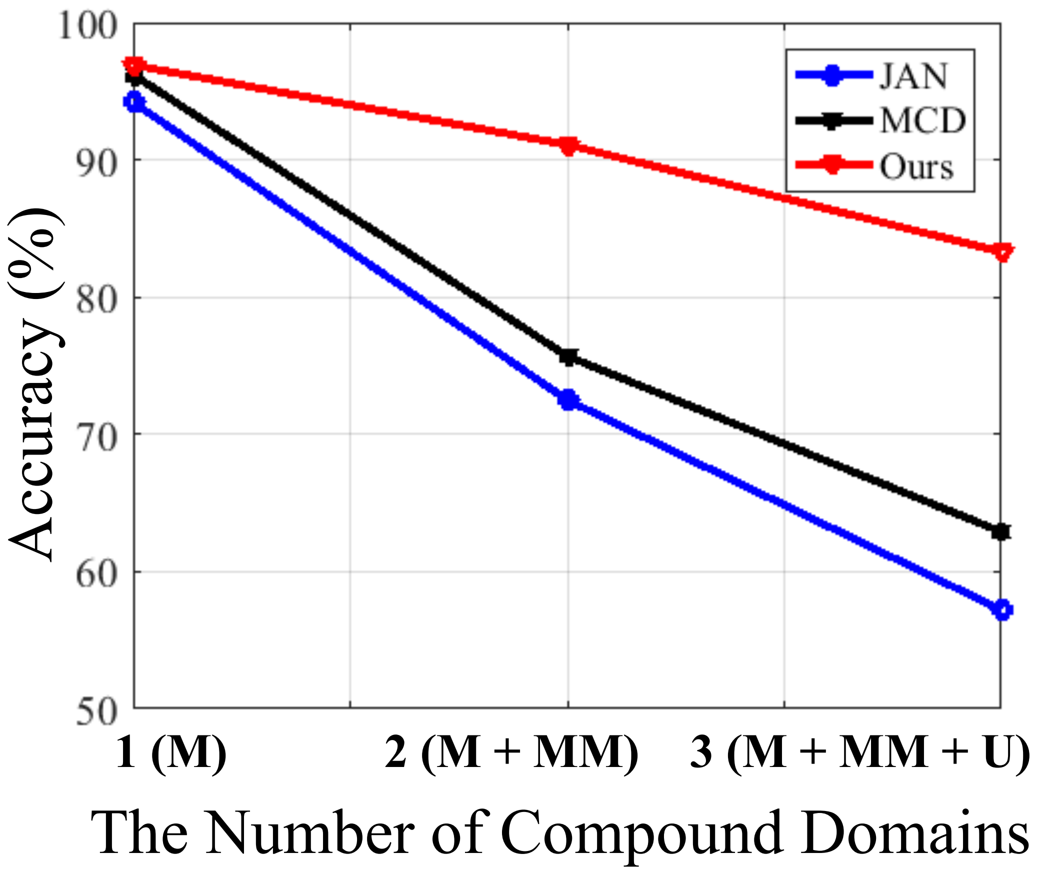


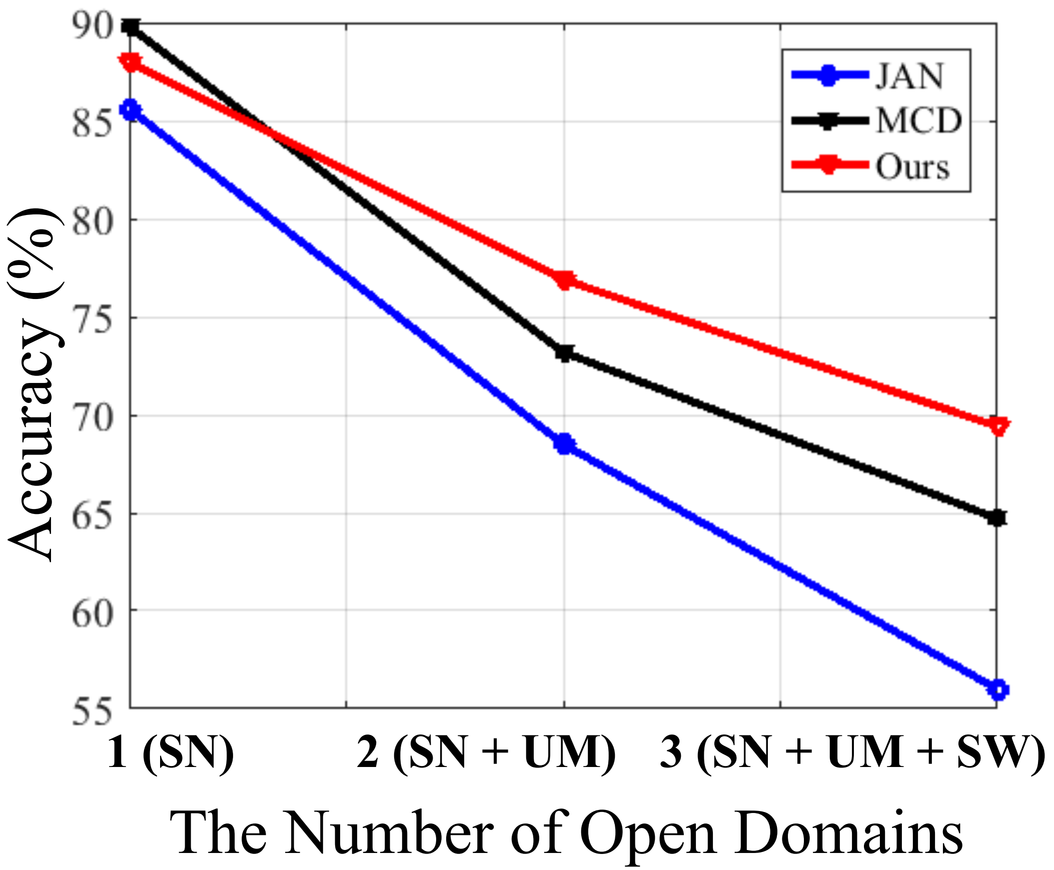
C-Mazes Benchmark

Absolute Performance Gain: ~30%



Robustness to the complexity of compound domains and open domains





Adaptation Results on C-Driving

(semantic segmentation)

Source Domain (Simulation)

Source Only

Ours

Compound Target Domain (Rainy)

Source Only

Ours

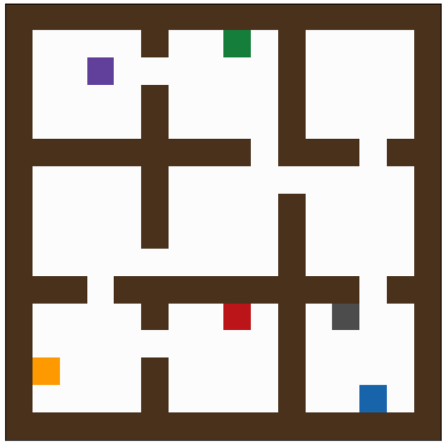
Open Target Domain (Overcast)

Source Only

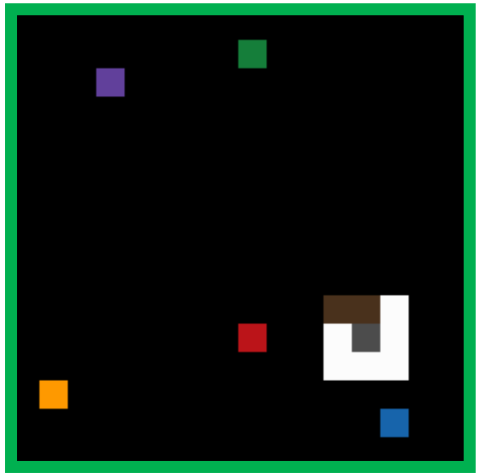
Ours

Adaptation Results on C-Mazes

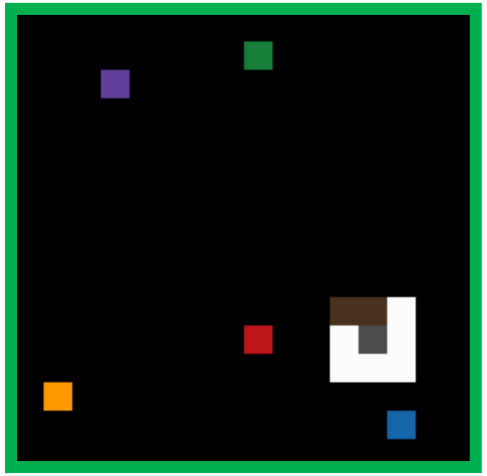
(reinforcement learning)



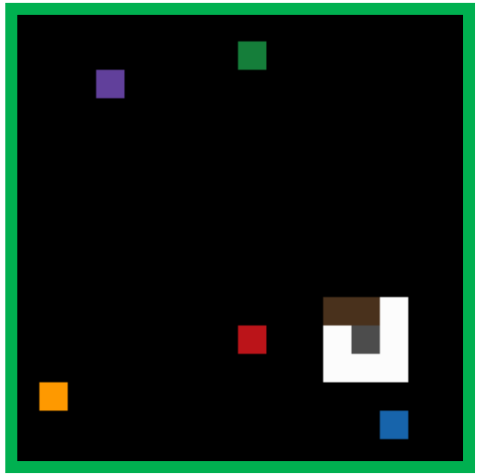
Source Domain



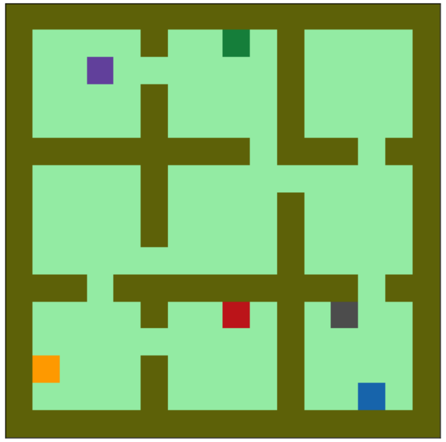
MTL
(succeed)



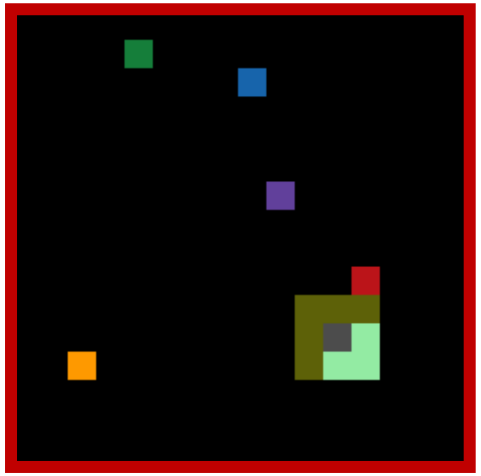
SynPo
(succeed)



Ours
(succeed)

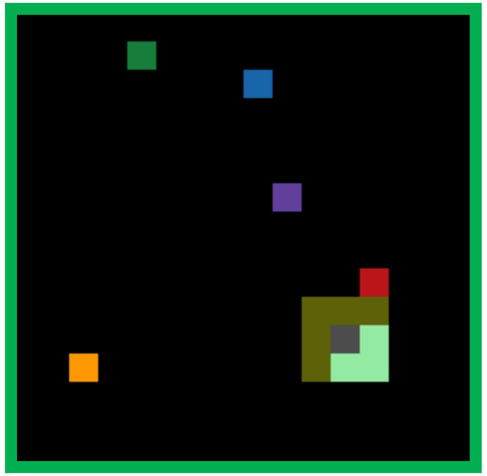


Open Target Domain 1



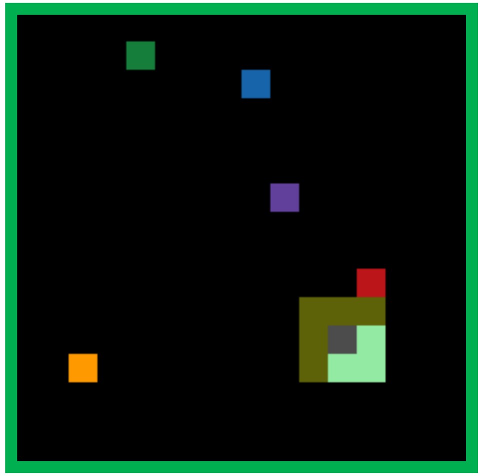
MTL

(fail)



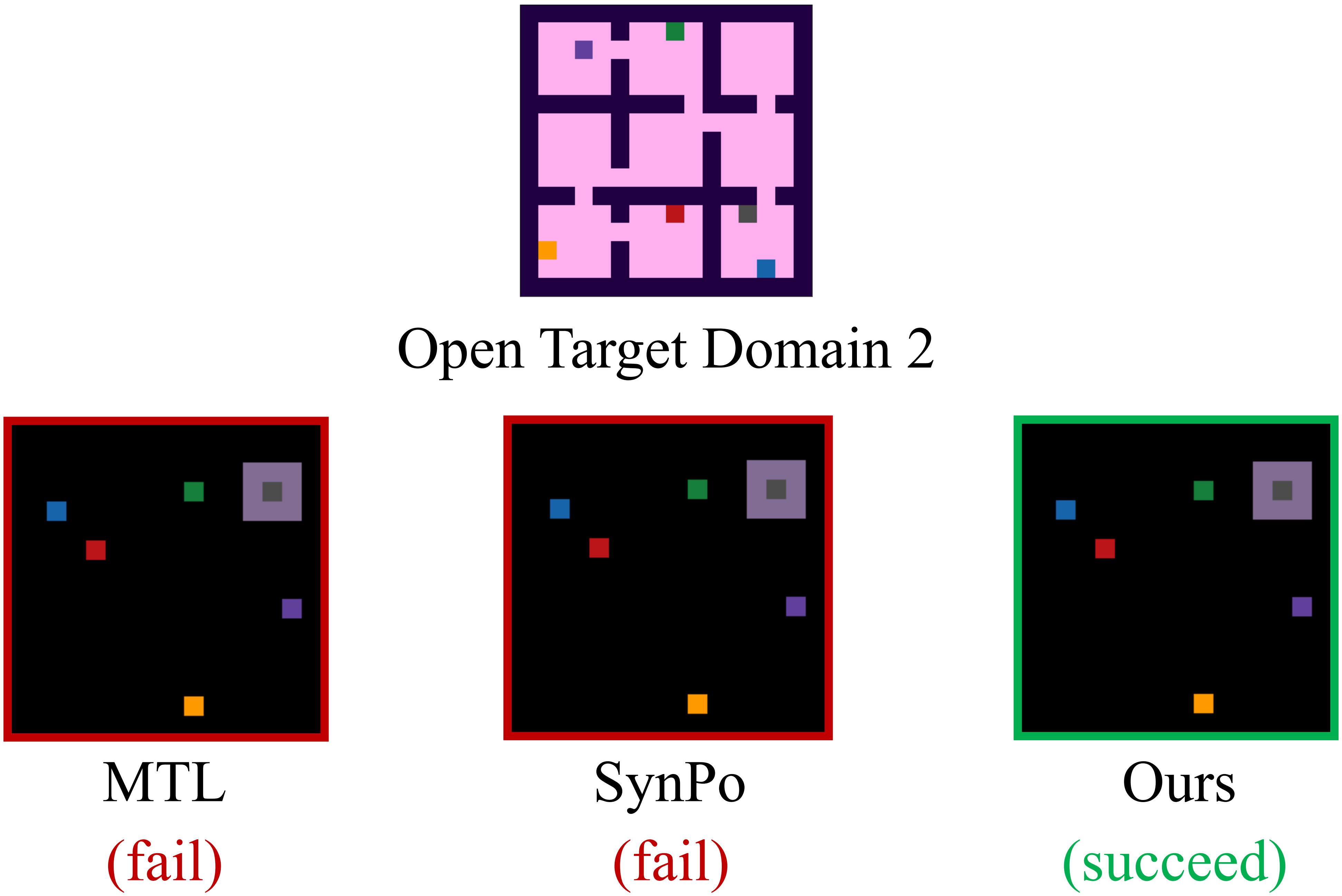
SynPo

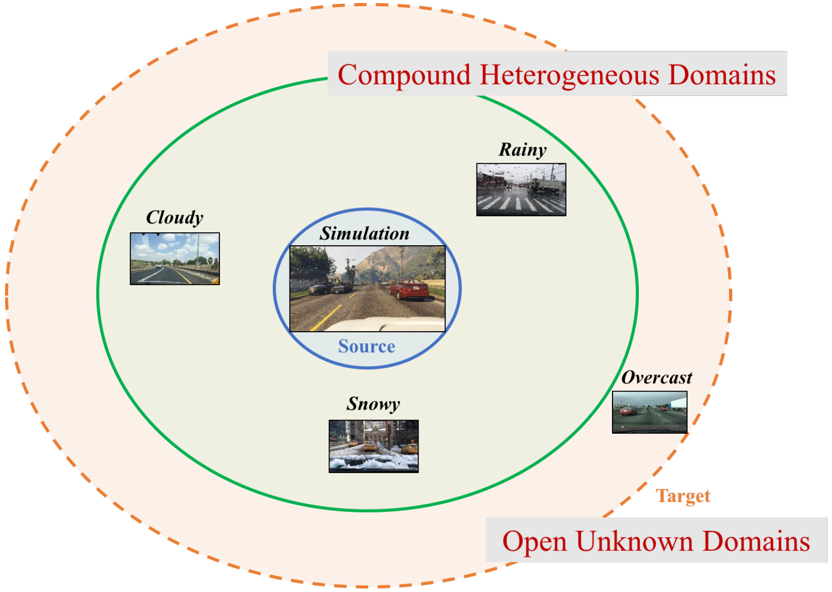
(succeed)



Ours

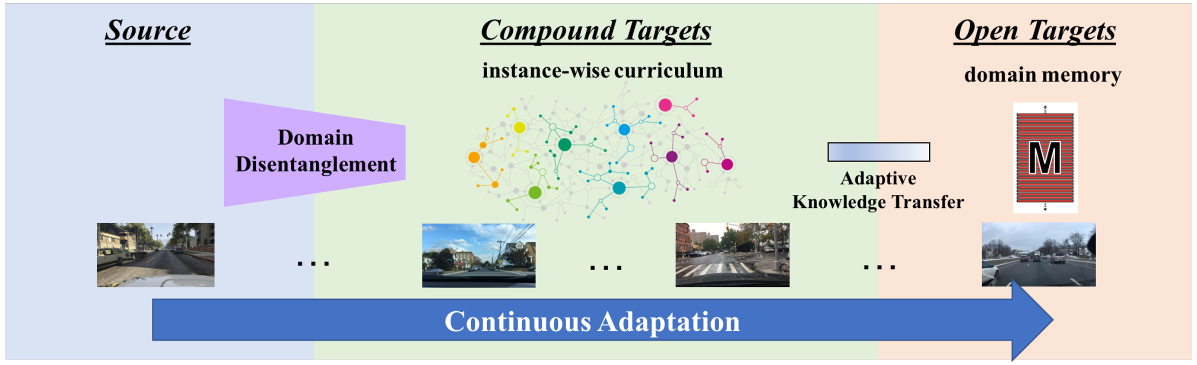
(succeed)


Open Target Domain 2
MTL
(fail)
SynPo
(fail)
Ours
(succeed)



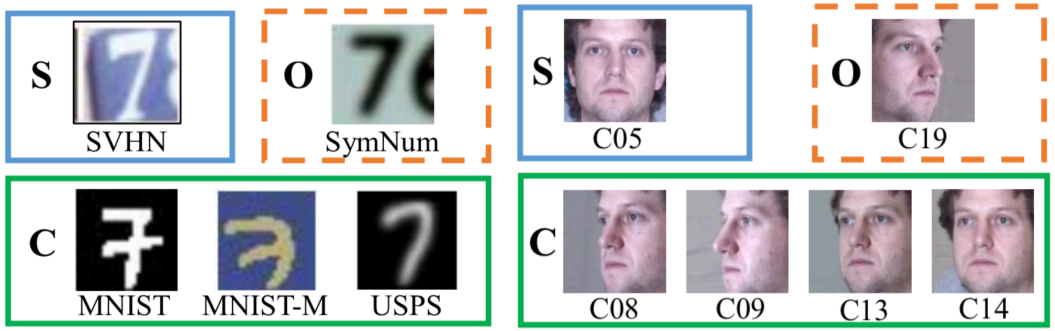
New Task

Open Compound Domain Adaptation(OCDA)



New Approach

Instance-wise Curriculum + Domain Memory



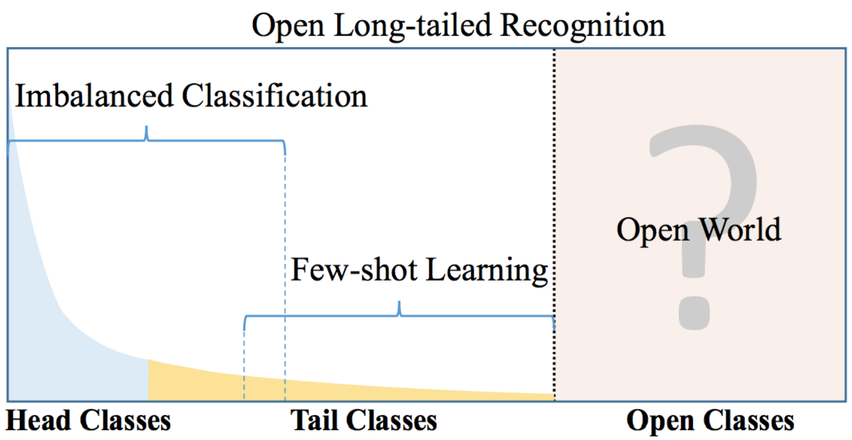
New Benchmarks

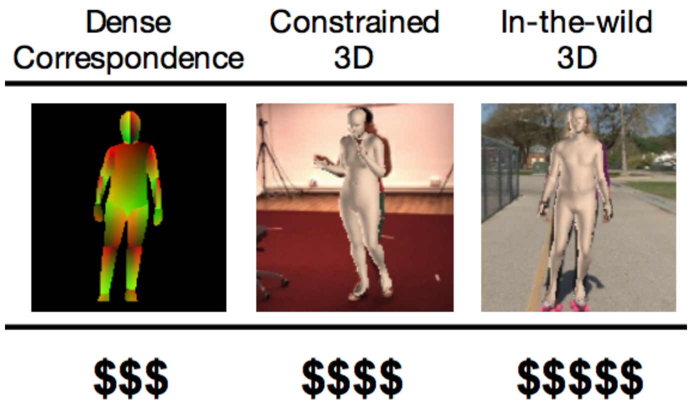
C-Digits, C-Faces, C-Driving, and C-Mazes

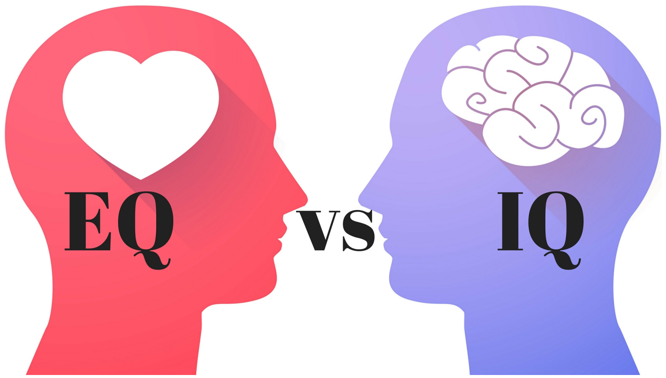
Code, models and benchmarks are available at

Project Page: https://liuziwei7.github.io/projects/CompoundDomain.html

What's Next







"Devils are in the Tails"

"Blessing of Dimension"

"Ghost in the Shell"

Thanks!

Science is what we understand well enough to explain to a computer. Art is everything else we do.

Homepage: https://liuziwei7.github.io/