Learning Diverse Human Representation in the Wild

Ziwei Liu

The Chinese University of Hong Kong

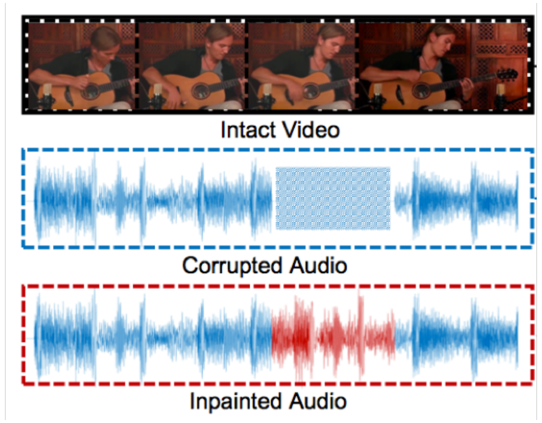
Human-Centric AI

Human-Centric AI

Face Representations

Human Representations


NATURE REPUBLIC
PC방
DVD
영화관
노래방
고객이 즐거워 한다면 이 정도 쯤이야
고객이 즐거워 한다면 이 정도 쯤이야
6월 6일 대개봉
인류 기원의 충격적 비밀이 밝혀진다!
프로메테우스
미고운 성형외과 피부과
Tel.774-8050
ZIOZIA
COFFEE
DEC.32ND
Nail Olive
Nail Olive
DEC.32ND
DEC.32ND
FLOATI MINI
RAINBOW PC
노정식 성형외과
778-7933
Roem
JUNO HAIR
준오헤어
삼성부동산
THE FACESHOP
THE FACESHOP
LLOYD
THE GIFT



Diverse Modalities

Visual-Audio Representation

Diverse Poses & Textures

Colorful 3D Human Representation

Diverse Categories & Relations

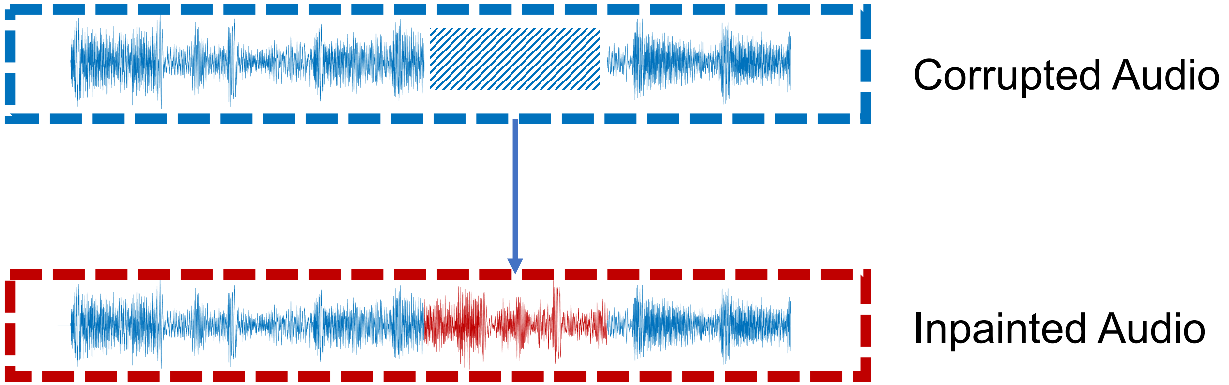
Fashion Collocation Representation

Diverse Modalities

Vision-Infused Deep Audio Inpainting,
ICCV 2019

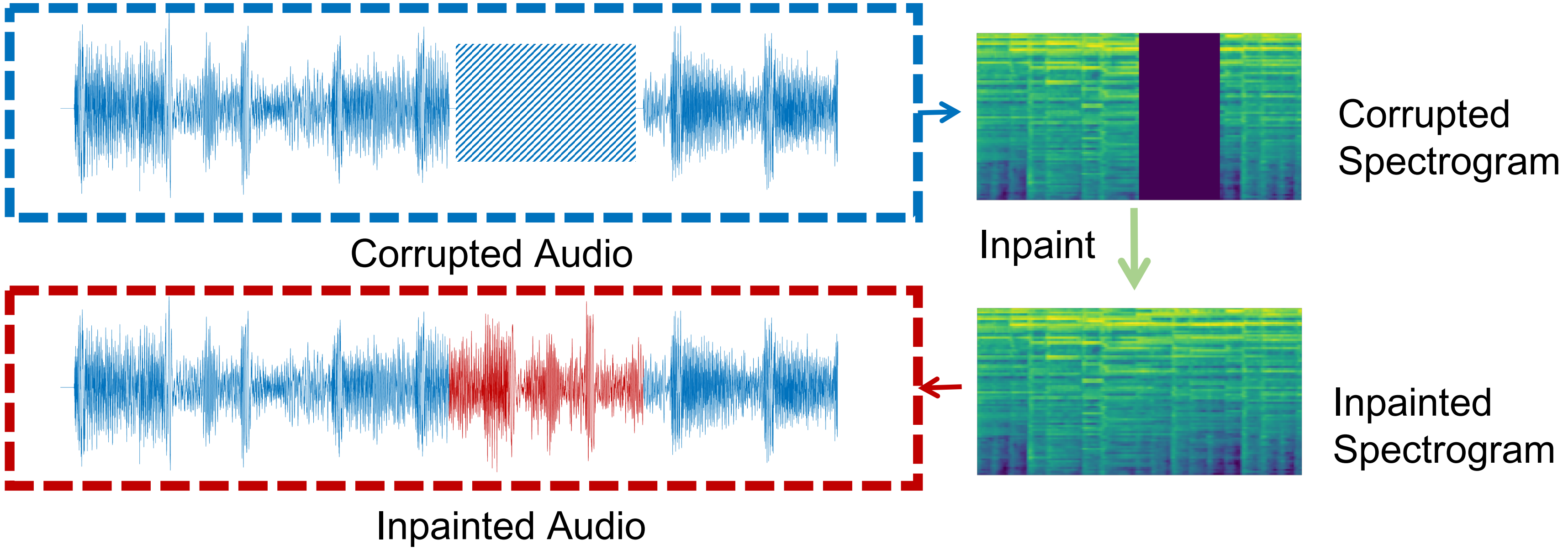
Motivation

- Audio signals often suffer from local distortions where the intervals are corrupted.
- Audio Inpainting: To fill the corrupted information with newly generated samples.



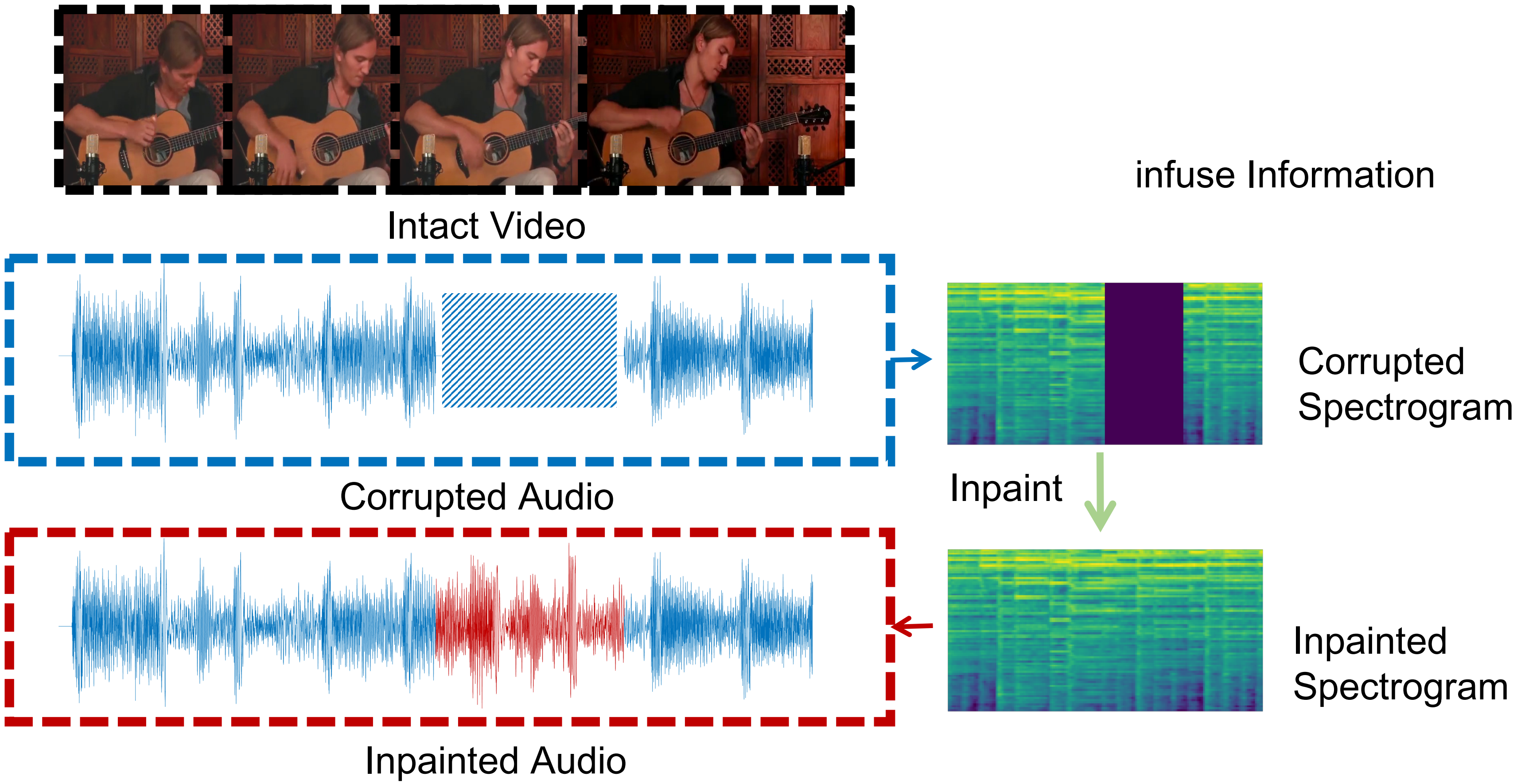
Core Idea

- Formulate audio inpainting into spectrogram inpainting.



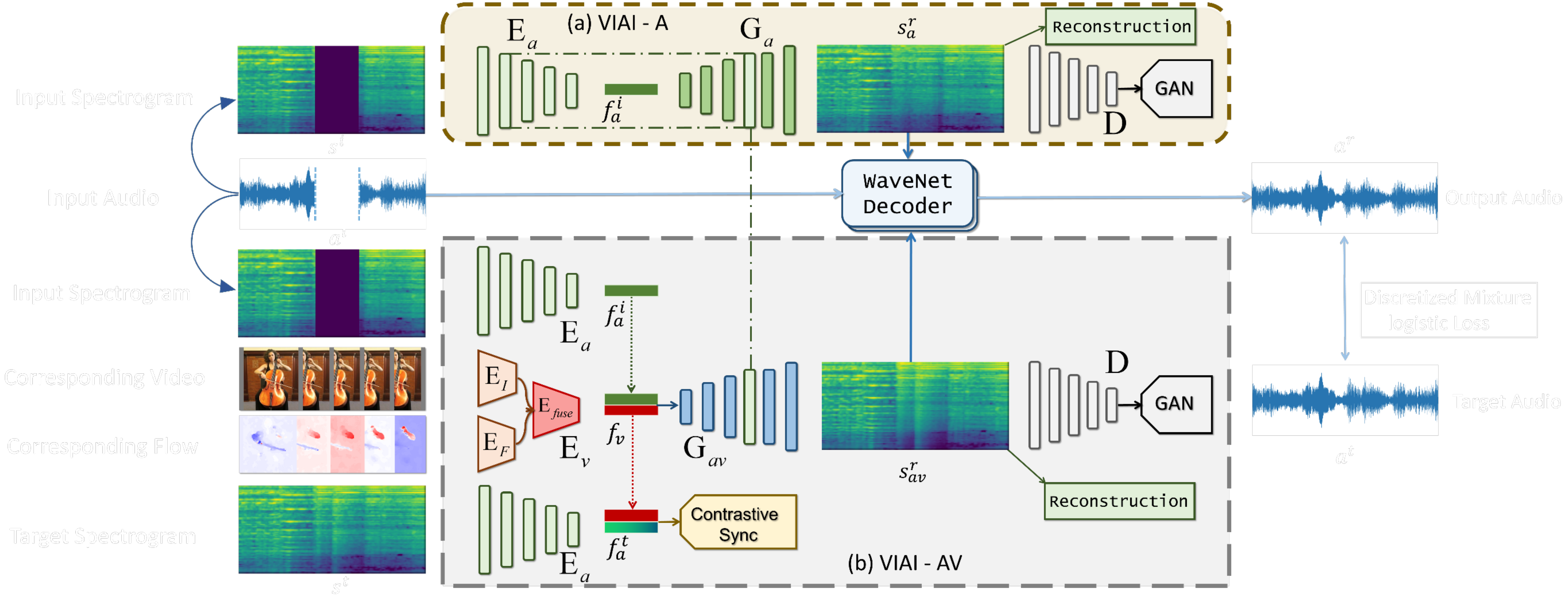
Core Idea

- Utilize intact video to guide audio inpainting.

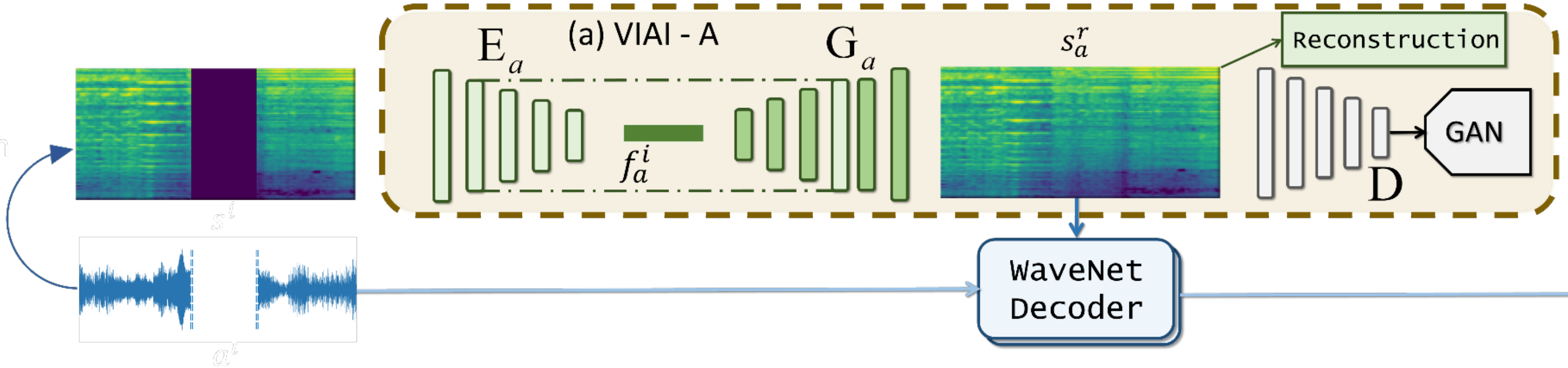


Approach

- Overview: Vision-Infused Audio Inpainter (VIAI)



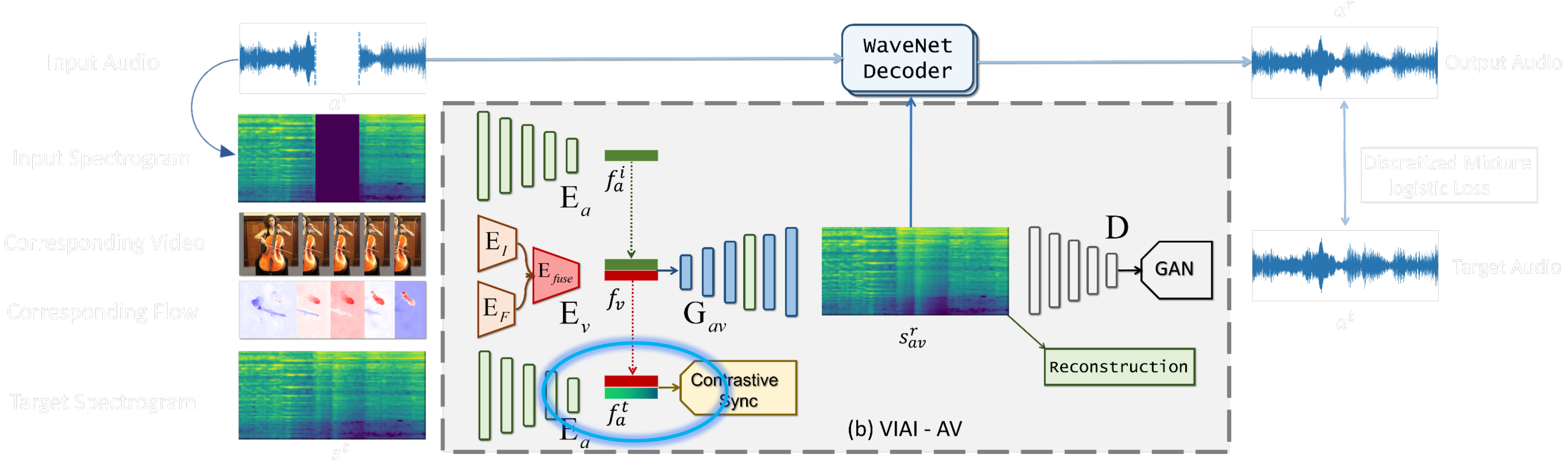
VIAI–Audio Branch (VIAI-A)



- Using the 2D Time-Frequency representation of Mel-Spectrogram for audios.
- Formulating the problem into inpainting spectrogram with Generative Adversarial Networks

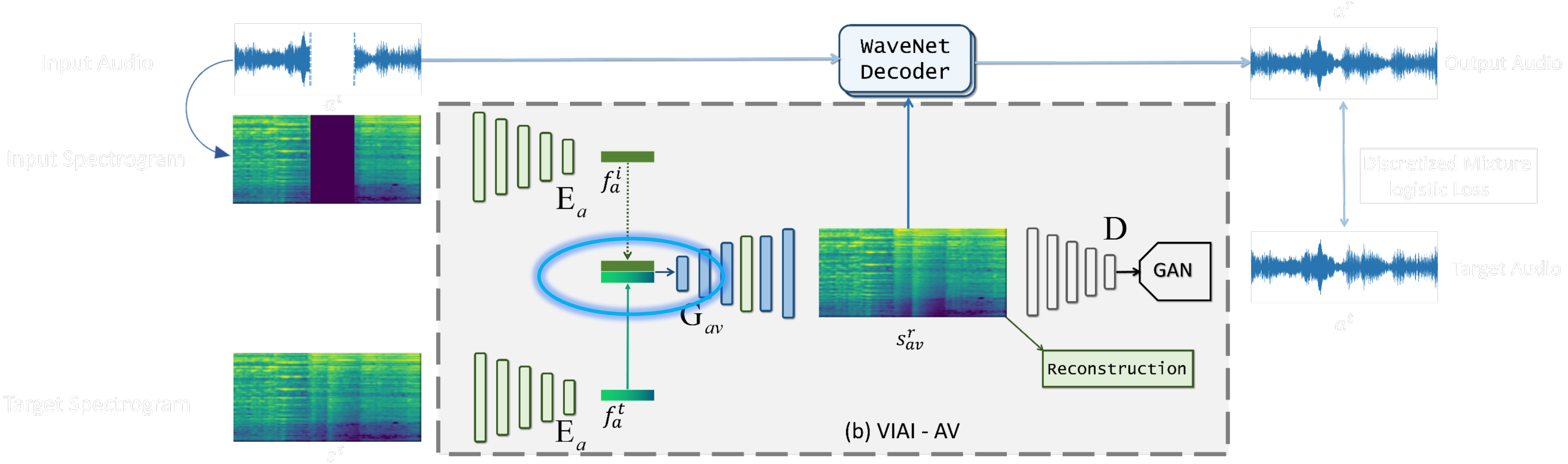
VIAI–Audio-Visual Branch (VIAI-AV)

- Learning synchronization between intact video and audio.
- Concatenate the synchronized features for reconstruction.



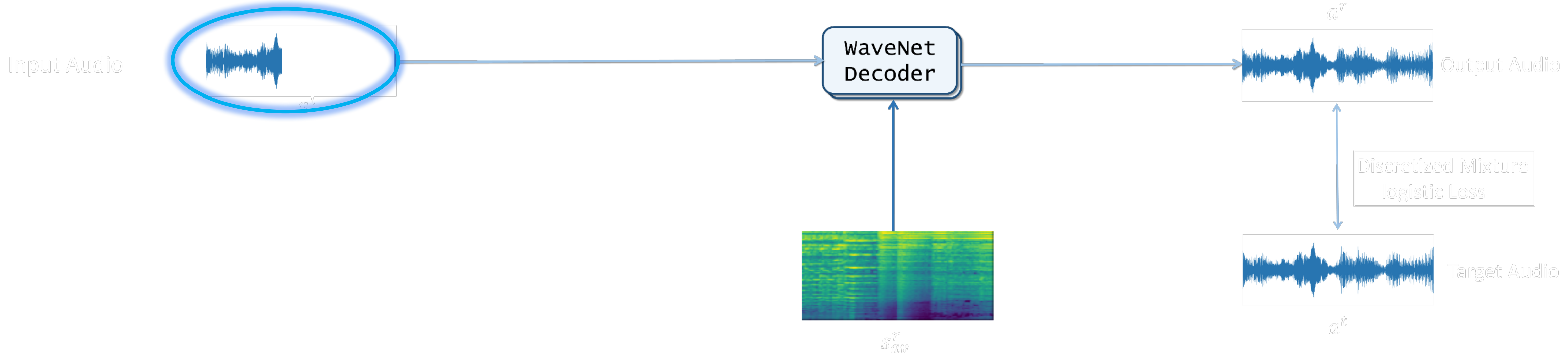
VIAI–Audio-Visual Branch (VIAI-AV)

- Probe loss of using intact audio for reconstruction (VIAI-AA’).
- Forcing the network to learn from bottleneck features.



WaveNet Decoder

- WaveNet is used to convert Mel-spectrogram back to raw audio.
- Utilizing the given audio for better restoration.



Experiments

Score \ Approach	SampleRNN [33]	Visual2Sound [56]	bi-SampleRNN	bi-Visual2Sound	VIAI-A	**VIAI-AV**	**VIAI-AA'**
PSNR	9.1	10.2	12.8	13.6	22.2	**23.2**	**26.6**
SSIM	0.33	0.35	0.38	0.41	0.61	**0.64**	**0.75**
SDR	4.89	3.70	4.20	4.72	6.54	**6.63**	**6.89**
OPS	51.1	51.3	51.2	52.2	52.4	**56.3**	**56.7**

ICCV 2019
Seoul, Korea

Vision-Infused Deep Audio Inpainting

Hang Zhou[1] Ziwei Liu[1] Xudong Xu[1] Ping Luo[2] Xiaogang Wang[1]

1. The Chinese University of Hong Kong
2. The University of Hong Kong

Conclusions

- Discriminative representation’s is capable of distilling and disentangling information from both modalities.

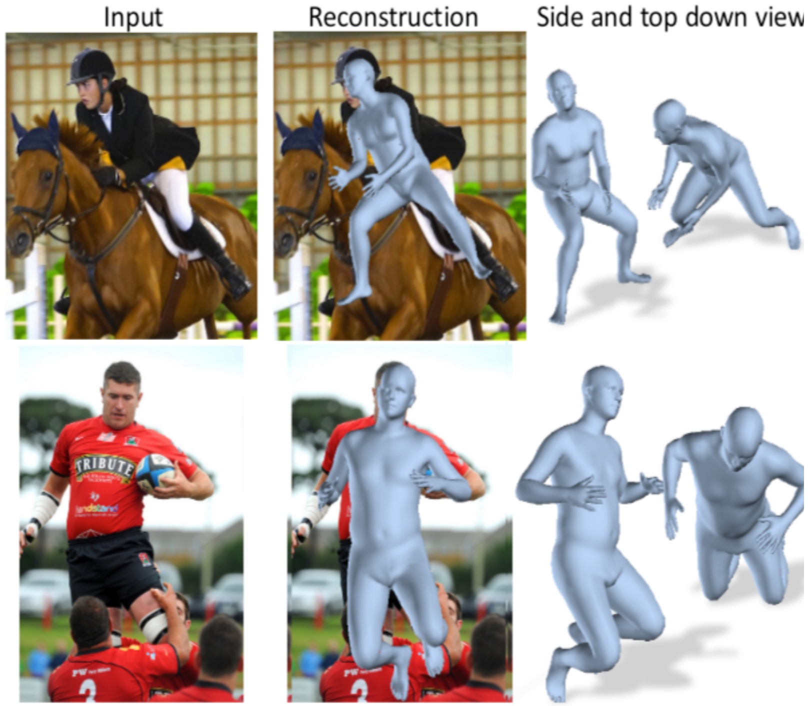
- Audio problems can be easier solved by operating on spectrograms using vision techniques for image processing.

- Synchronization between audio and visual information is the fundamental self-supervision which is crucial for various tasks.

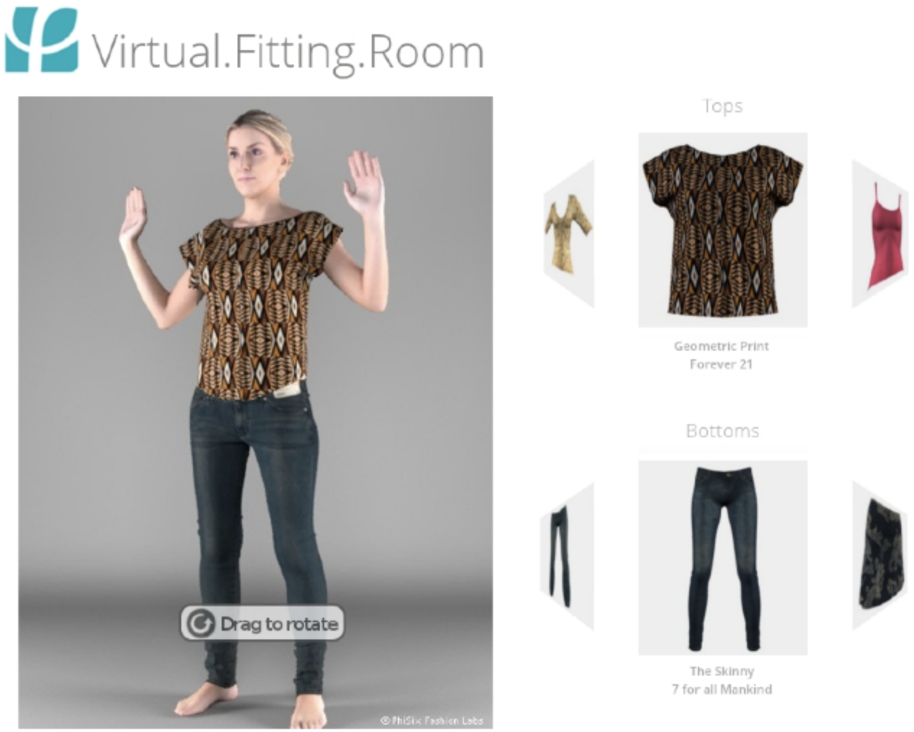
Diverse Poses

Delving Deep into Hybrid Annotations for 3D Human Recovery in the wild, ICCV 2019

Background (I)



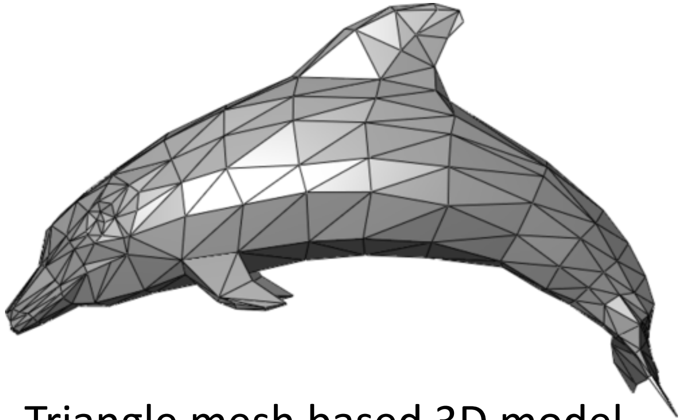
3D Human Reconstruction



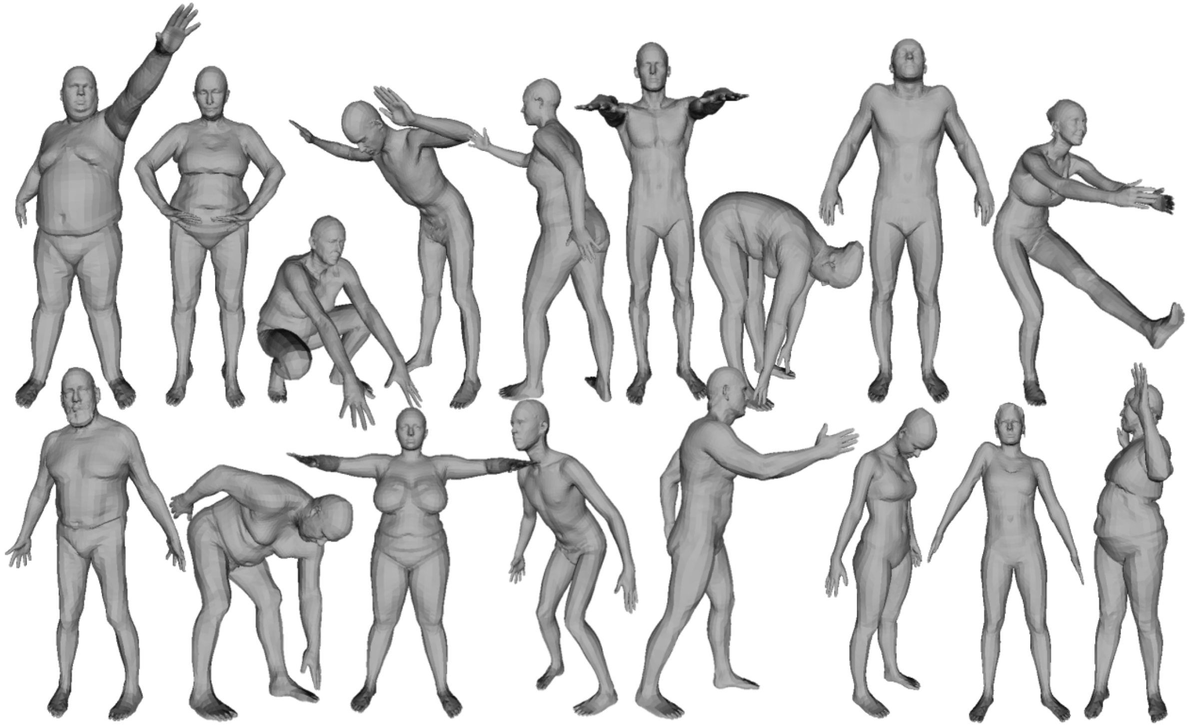
Virtual Try-on

- 3D Human Reconstruction means acquire 3D human representation from given images or videos.
- It can facilitate many technologies such as augmented reality and virtual try-on.

Background (II)



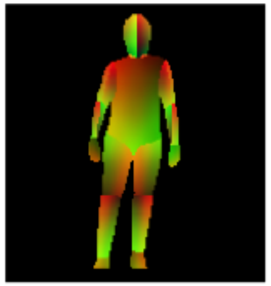
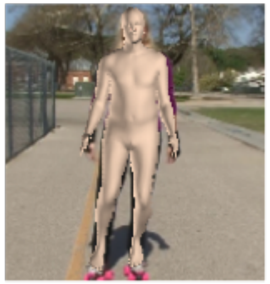
Triangle mesh based 3D model



SMPL

- We use SMPL, a parametric triangle mesh based 3D model to represent 3D human.
- SMPL is parameterized by two parameters: **pose parameters** $\theta \in \mathbb{R}^{72}$ and **shape parameters** $\beta \in \mathbb{R}^{10}$.
- To estimate 3D human representation, we only need to predict the pose and shape parameters.

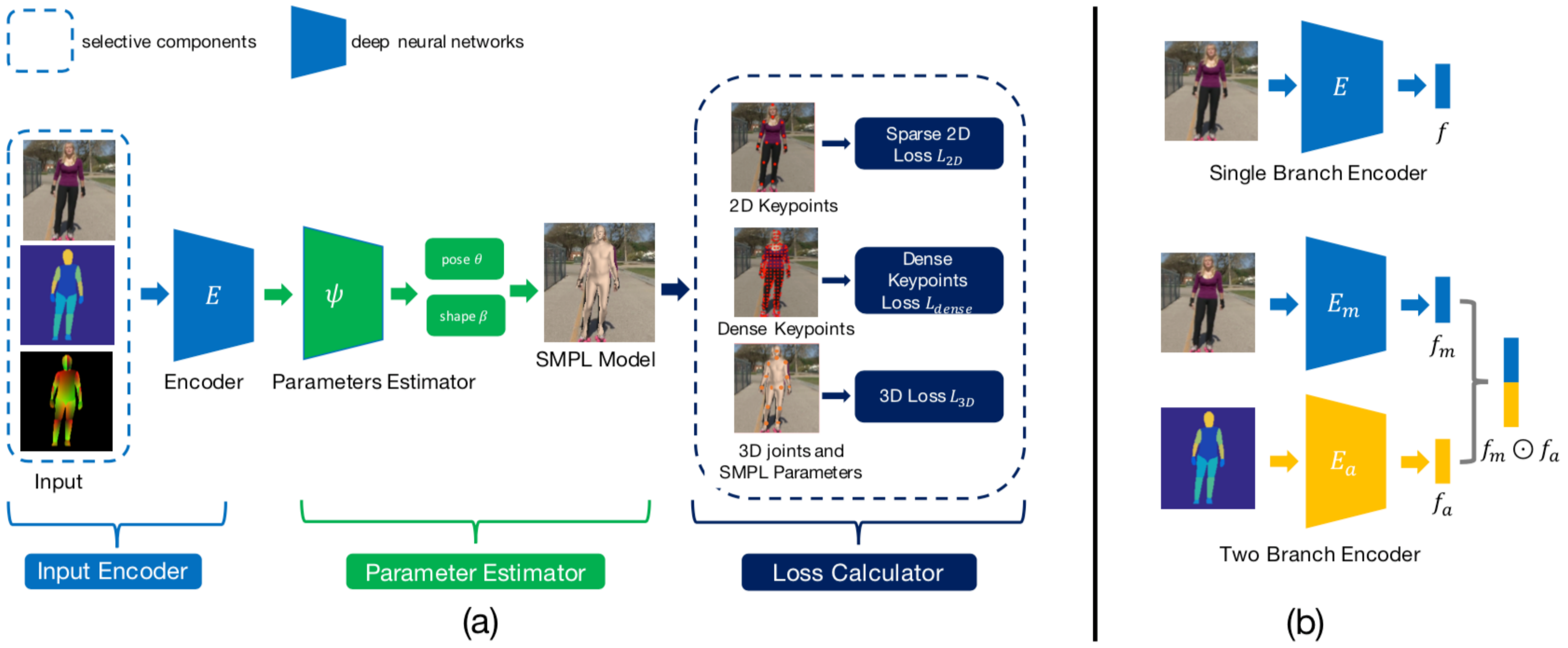
Motivation

Annotation	Sparse 2D	Dense Labeling	Dense Correspondence	In-the-wild 3D
Examples				
Annotation Cost	$	$$	$$$	$$$$$

- In the experiment, we first study the efficiency of different annotations.
- We study the efficiency of those annotations when serving as input and serving as supervision.
- We use per-vertex distance (PVE) as the evaluation metric.
- The experiments are conducted on COCO-DensePose, UP-3D and 3DPW.

$$PVE = \sum_{i=1}^{O} ||P_i - \bar{P}_i||_2^2$$

Framework



- The overall framework is composed of three parts:
  - Input Encoder
  - Parameter Estimator
  - Loss Calculator

Learning Strategy (I)

3D Loss L_{3D}

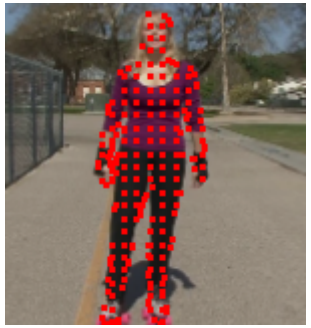
$$L_{3D\_joints} = \sum_{i=1}^{M} ||(J_i^{3D} - \hat{J}_i^{3D})||_1,$$
$$L_{SMPL} = \sum_{i=1}^{O} ||R(\theta_i) - R(\hat{\theta}_i)||_1 + ||\beta_i - \hat{\beta}_i||_1$$
$$L_{3D} = L_{3D\_joints} + L_{SMPL}$$

Hard to acquire !

Sparse2D Loss L_{2D}

$$L_{2D} = \sum_{i=1}^{S} ||(J_i^{2D} - \hat{J}_i^{2D})||_1$$

Too Sparse !

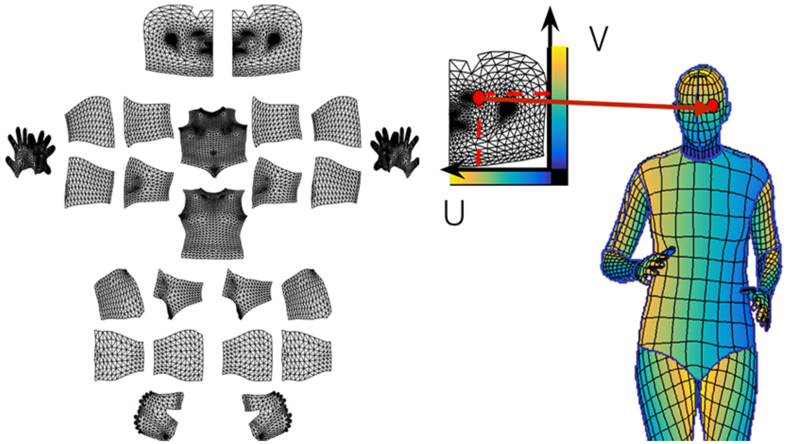


Dense Keypoints Loss L_{dense}

$$[v_{i1}, v_{i2}, v_{i3}], [b_{i1}, b_{i2}, b_{i3}] = \phi(D_i),$$
$$\hat{X}_i = \sum_{j=1}^{3} \hat{P}_i^{2D}[v_{ij}] \times b_{ij},$$
$$L_{dense} = \sum_{i=1}^{T} ||(X_i - \hat{X}_i)||_1,$$

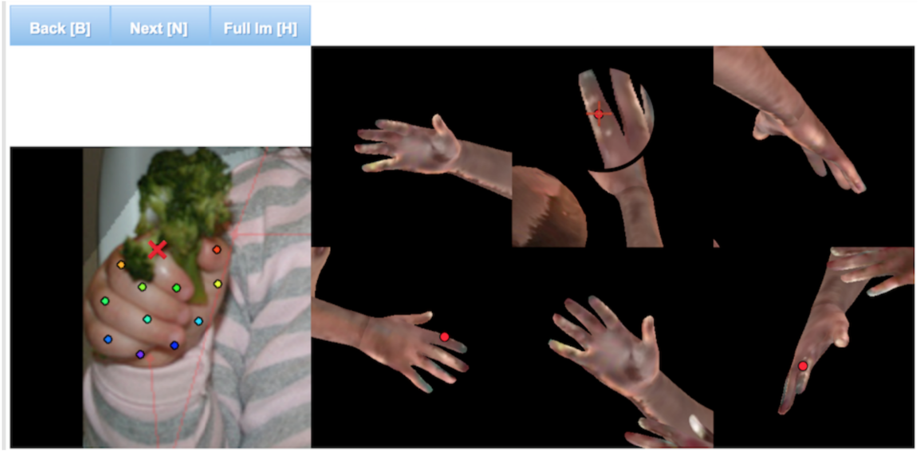
- Previous works mainly use 3D annotations and sparse 2D annotations in training.
- Sparse 2D keypoints are too sparse to provide enough guidance.
- 3D annotations are hard to acquire.
- We propose to use dense keypoints in recovering 3D human model.

Learning Strategy (II)



DensePose Model

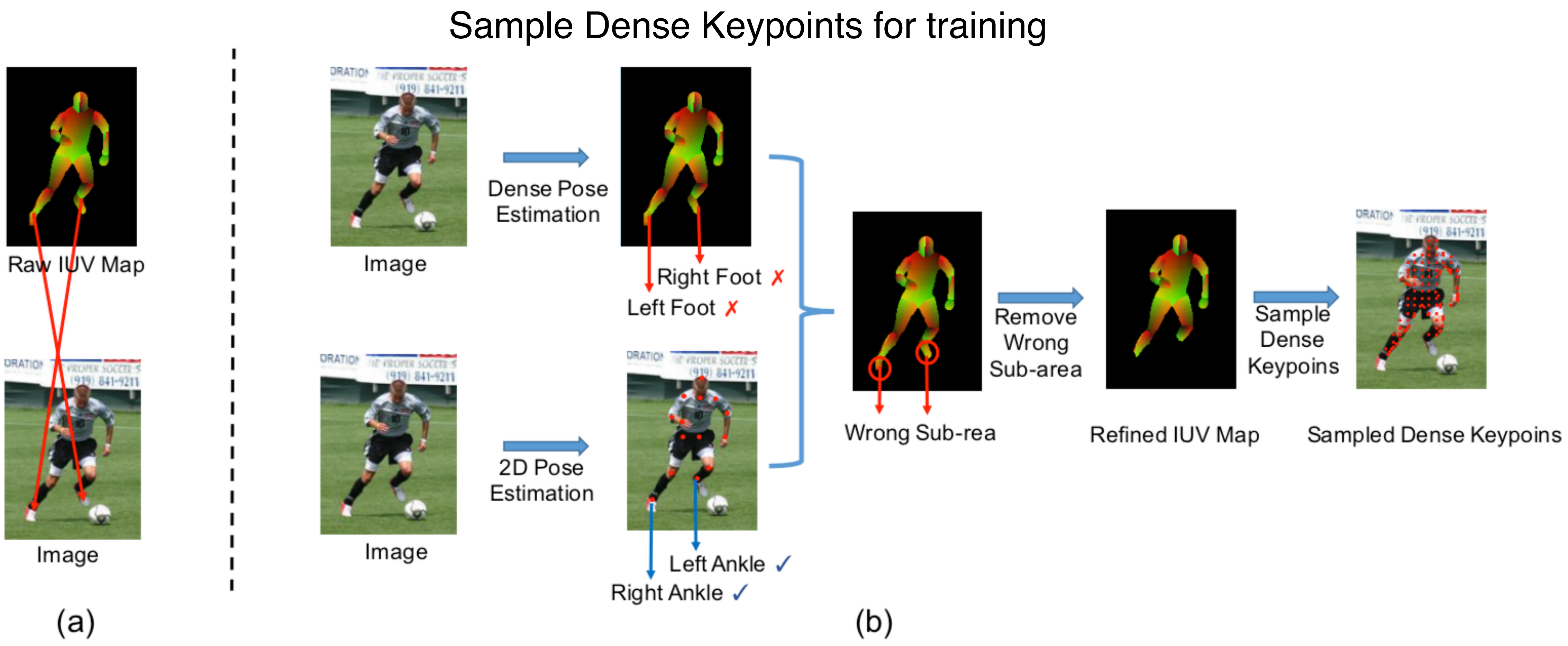
IUV Maps generated
by DensePose



Annotating
Dense Keypoints

- DensePose build dense correspondence between 2D images and human body surface.
- For each dense keypoints, the annotations include (I, U, V). I indicates which body part this point belongs to. (U, V) indicates the precise position.
- Dense keypoints could be annotated by human annotators without using auxiliary equipements.

Learning Strategy (III)



- We use the predicted IUV maps from DensePose model and sample dense keypoints from them.
- We conduct refinement using the accurate sparse 2D keypoints to remove erroneous IUV maps.

Experiments

Table 3. **Influence of different annotations.** The evaluation metrics are PVE, MPJPE and PVE-T, separately. For all metrics, lower is better. "3D" refers to paired in-the-wild 3D annotations. "20% 3D" refers to 20% randomly selected 3D annotations. "Sparse 2D" refers to sparse 2D keypoints. "Dense" refers to dense correspondence, namely, IUV maps generated by DensePose [1, 19].

Supervision → Input ↓	3D & Dense & Sparse 2D	20% 3D & Dense & Sparse 2D	3D & Sparse 2D	Dense & Sparse 2D	Sparse 2D Only
IUV Only	**120.0** / **103.1** / 31.8	125.0 / 107.2 / 32.6	125.2 / 106.4 / 32.1	138.7 / 121.2 / 54.7	204.3 / 177.0 / 92.1
Segment Only	123.0 / 105.1 / 32.7	126.7 / 110.0 / 33.2	124.8/ 107.8 / 31.7	147.4 / 130.1 / 55.9	203.8 / 176.7 / 93.3
Image Only	123.7 / 105.9 / 30.9	127.5 / 110.6 / 32.2	127.4 / 108.5 / 30.7	137.7 / 120.3 / 51.7	203.2 / 178.5 / 106.2
Image & IUV	122.4 / 105.1 / **30.2**	125.0 / 107.6 / 32.1	125.5 / 107.3 / 30.7	133.8 / 117.2 / 52.5	197.3 / 172.8 / 107.9
Image & Segment	121.5 / 104.3 / 31.0	126.4 / 107.0 / 31.6	125.8 / 106.8 / 31.5	142.2 / 124.2 / 56.6	201.2 / 177.5 / 101.7

Delving Deep into Hybrid Annotations for 3D Human Recovery

Paper ID 2209

This video is composed of two parts:

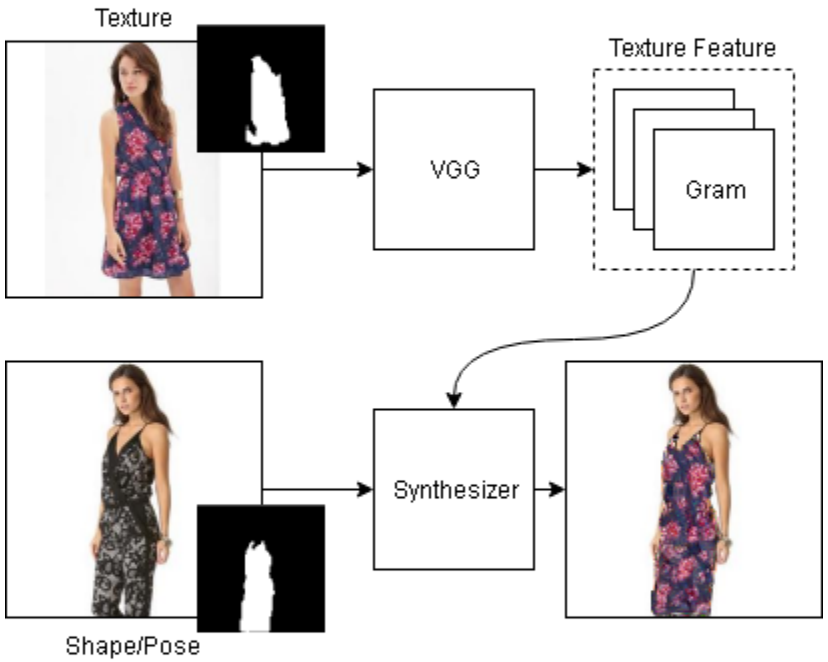
I. Influence of different annotations
II. Comparison with previous state-of-the-arts.

Diverse Textures

Learning to Synthesis Fashion Textures,
(in submission)

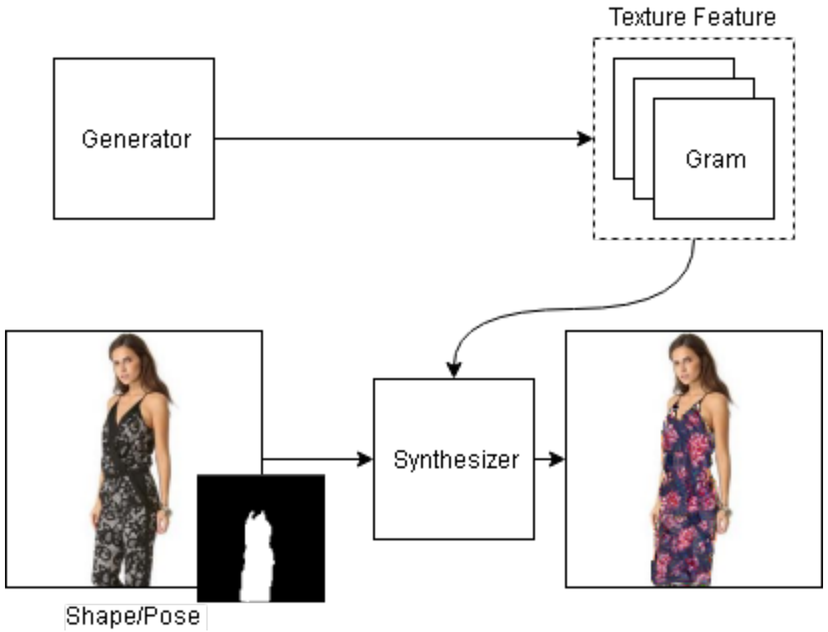
Fashion Texture Synthesis

- Use Gram matrix as texture feature to synthesize images
  - Flexible
  - Visually pleasing



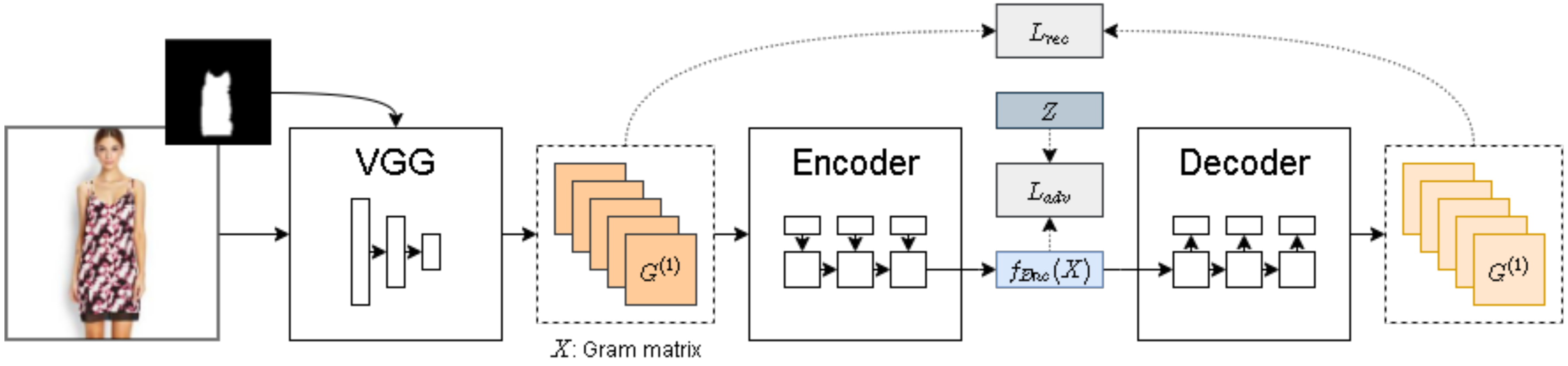
Fashion Texture Synthesis

- Two-step generation



Pose Texture Synthesized

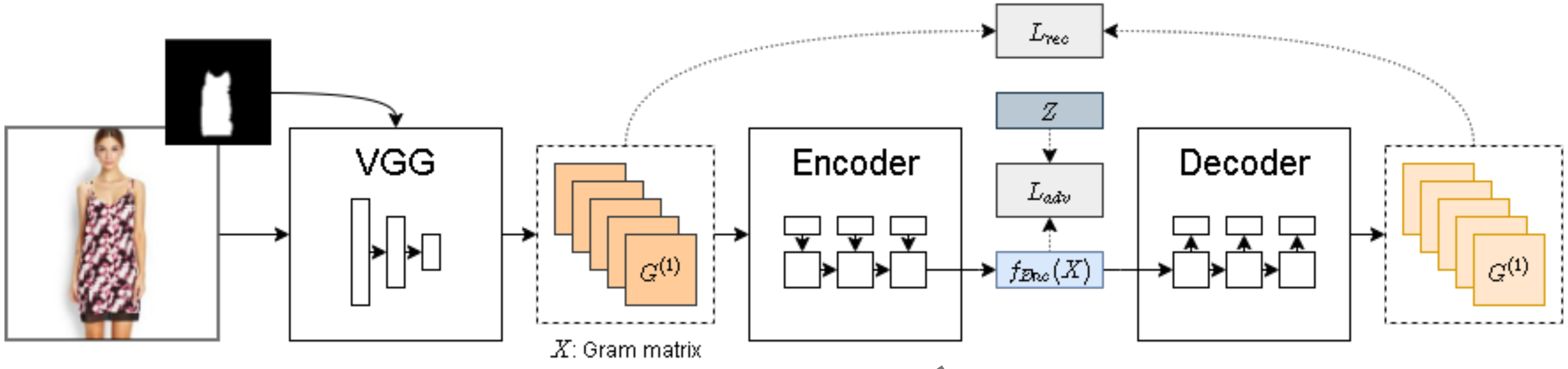
Generative Framework



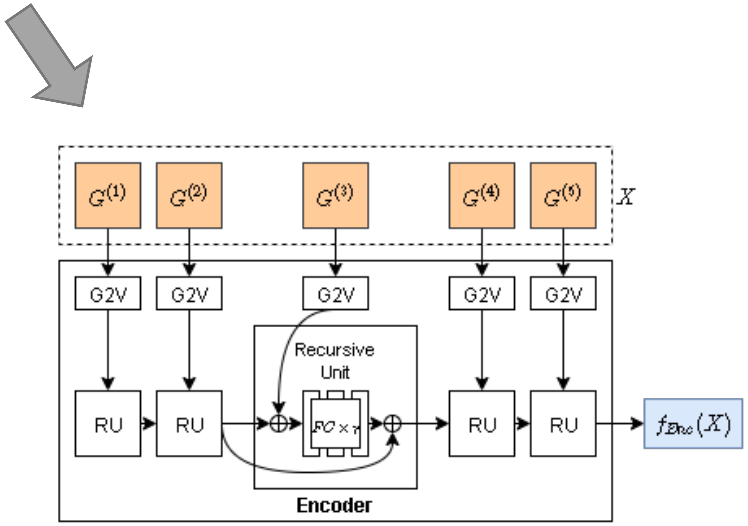
- Training Gram-WAE-GAN
  - Reconstruct the input Gram matrix
  - Match the latent distribution with the prior

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, & Bernhard Schoelkopf. Wasserstein Auto-Encoders. In *ICLR 2018*.

Recursive Structure



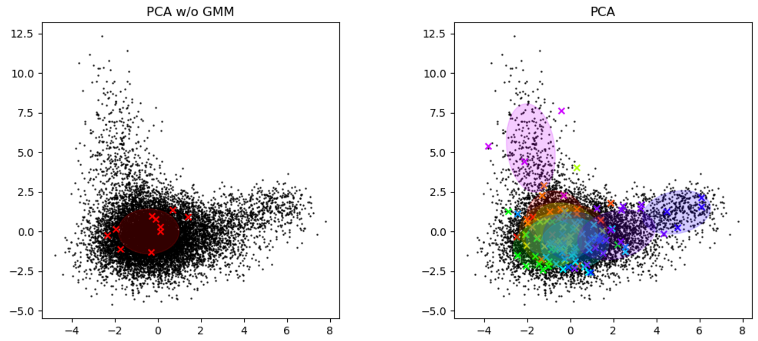
- Model a set of Gram matrices from multi-granularity levels

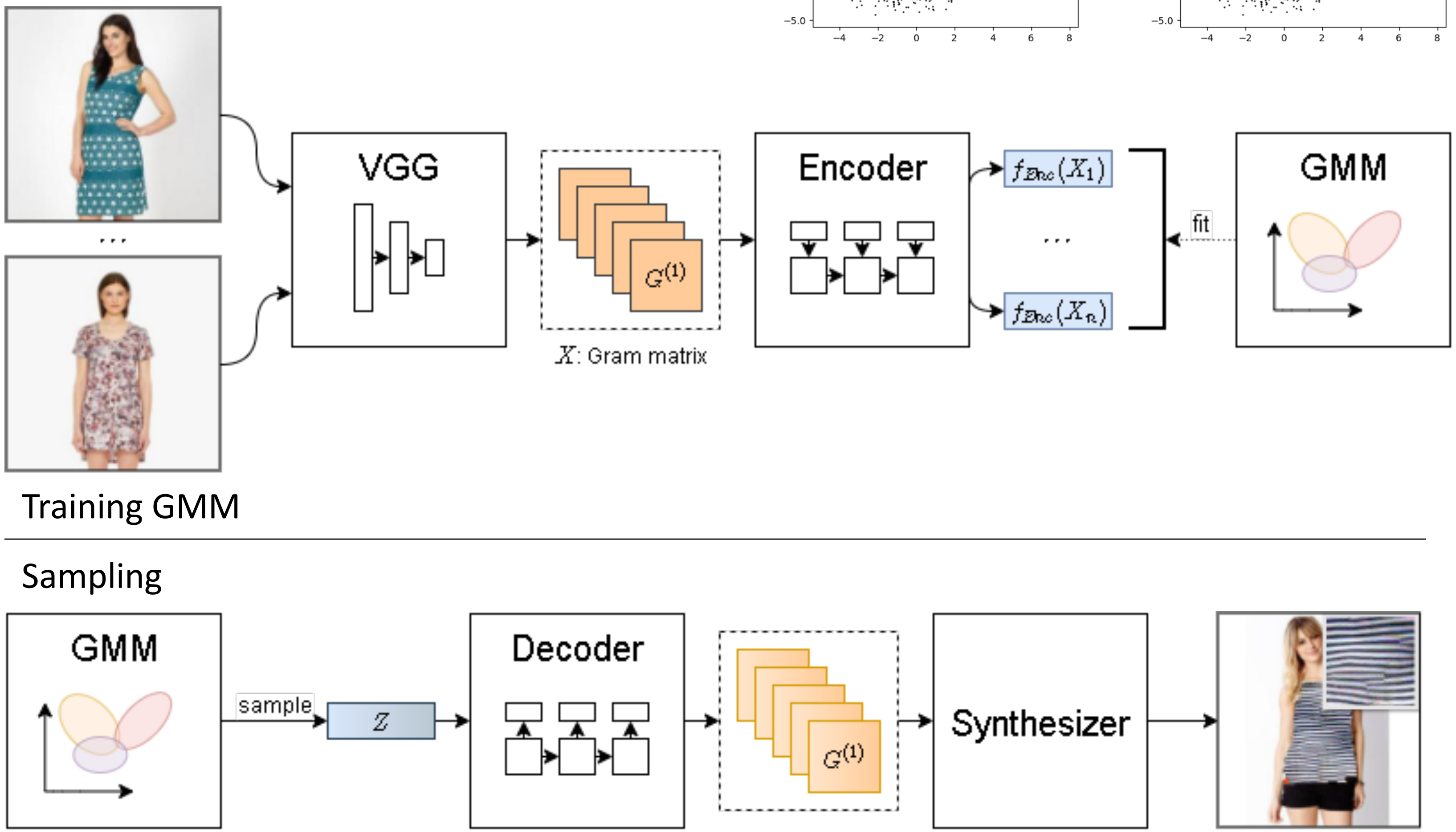


Shallow

Mixed

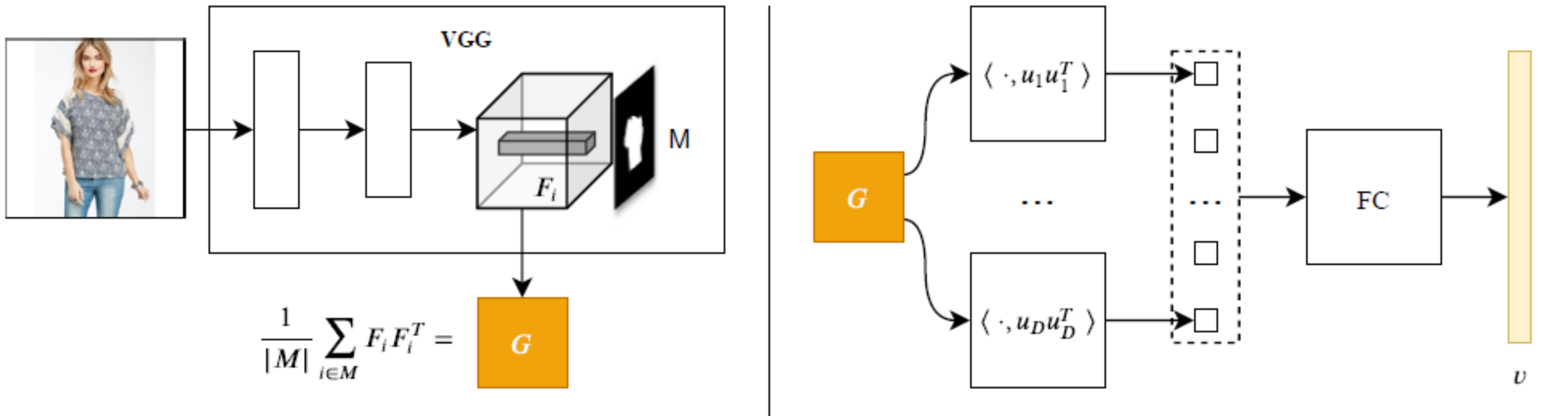
GMM Sampling





L. A. Gatys, A. S. Ecker, & M. Bethge (2015). Texture Synthesis Using Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 28*.

Gram Transformation



(a) Gram matrix

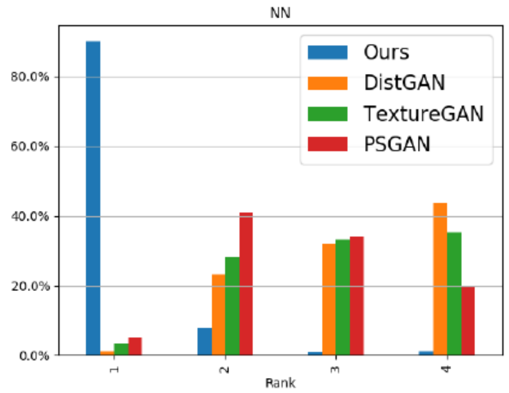
(b) Gram2Vec

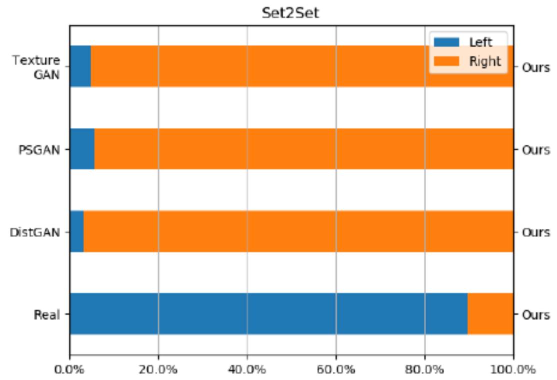
- Transform the Gram matrix to a low dimensional vector
  - Number of parameters: 184M -> 10.8M

Results

	Method	FID
Baseline	DistGAN [87]	41.97
	PSGAN [5]	77.10
	TextureGAN [93]	44.38
Ablation Study	FC transformation	***37.32***
	MLP structure	45.72
	No GMM sampling	40.83
	Ours	**37.74**

Results

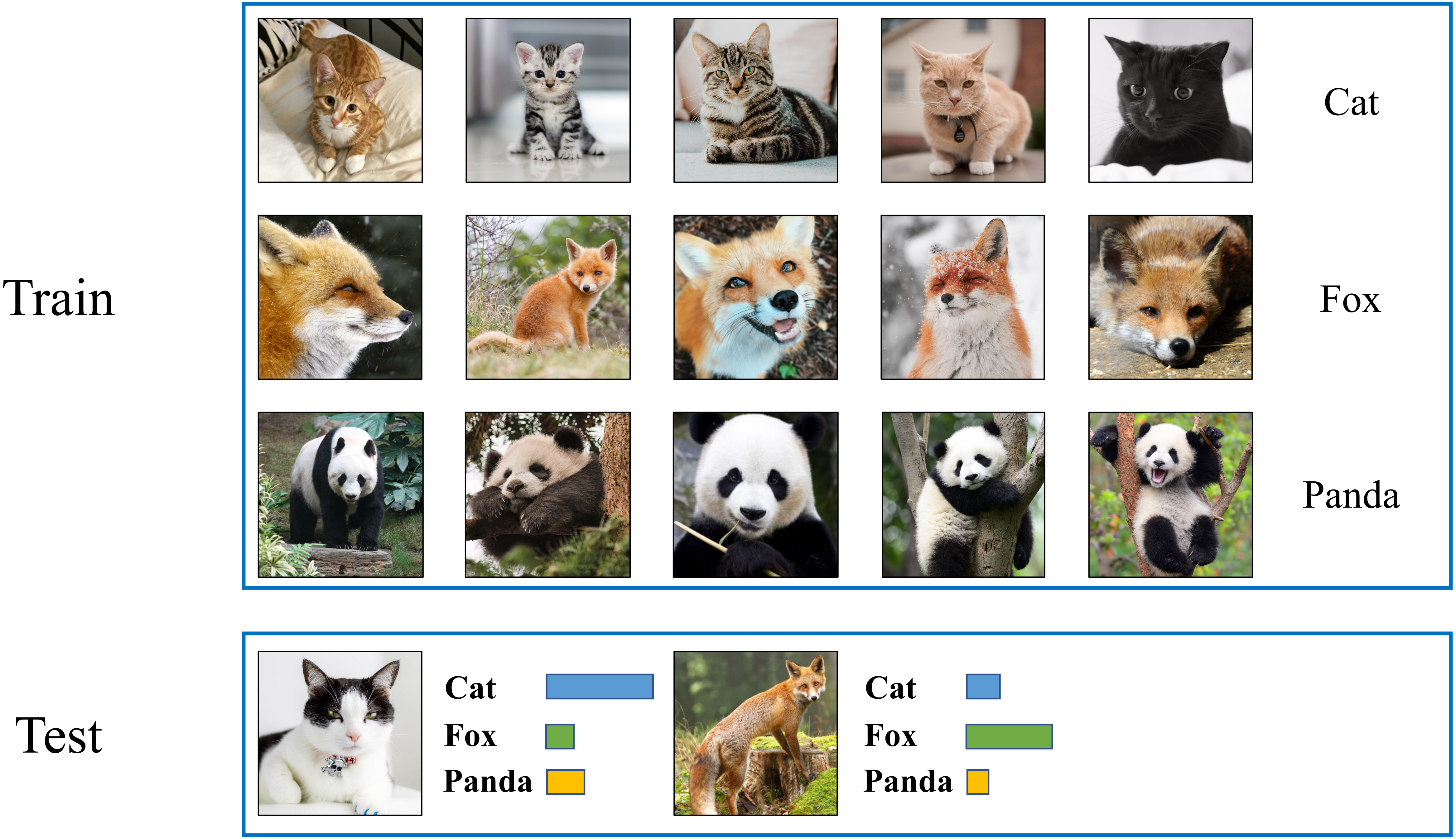

NN
Ours
DistGAN
TextureGAN
PSGAN
80.0%
60.0%
40.0%
20.0%
0.0%
1
2
3
4
Rank

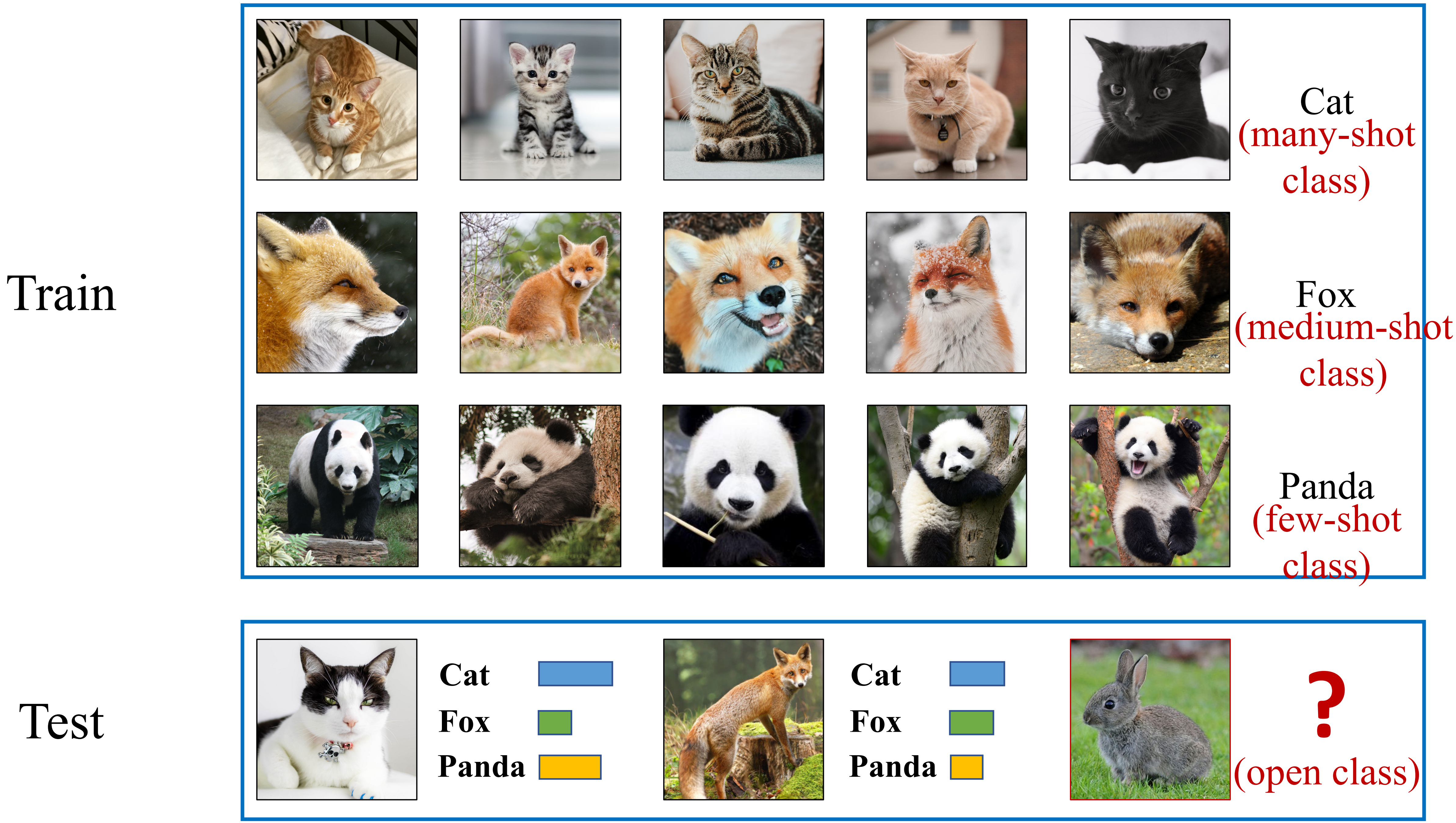

Set2Set
Left
Right
Texture GAN
PSGAN
DistGAN
Real
Ours
Ours
Ours
Ours
0.0%
20.0%
40.0%
60.0%
80.0%
100.0%

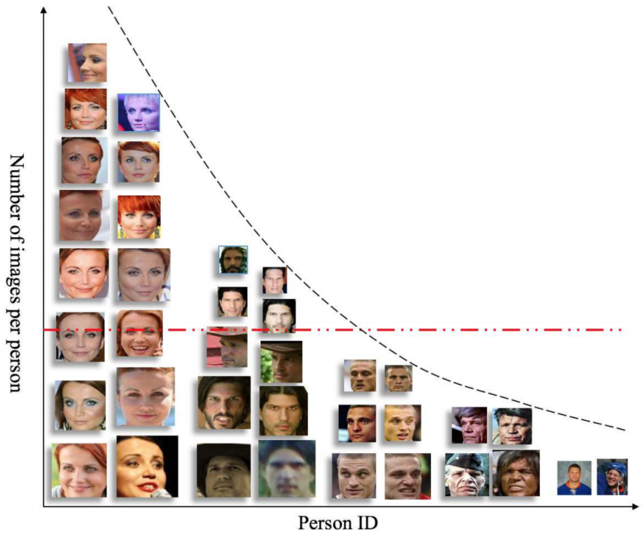

Test Image
PSGAN
Ours
Test Image
PSGAN
Ours

Diverse Categories

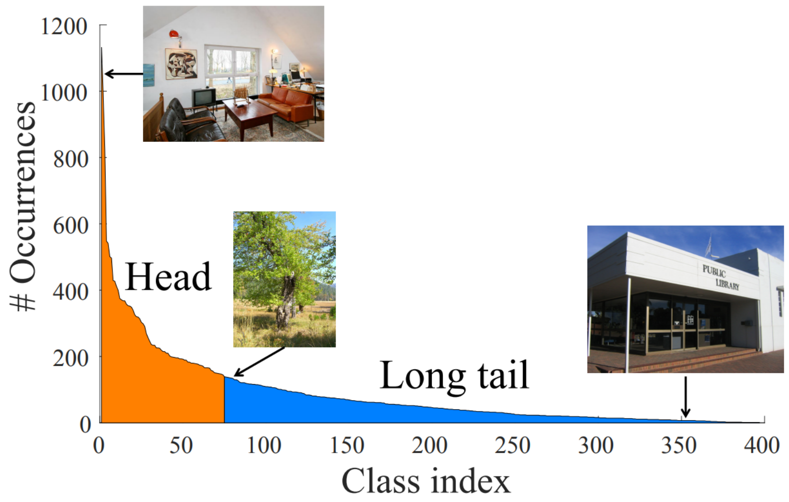
Large-Scale Long-Tailed Recognition in an Open World, CVPR 2019


Train
Cat
Fox
Panda
Test
Cat
Fox
Panda
Cat
Fox
Panda

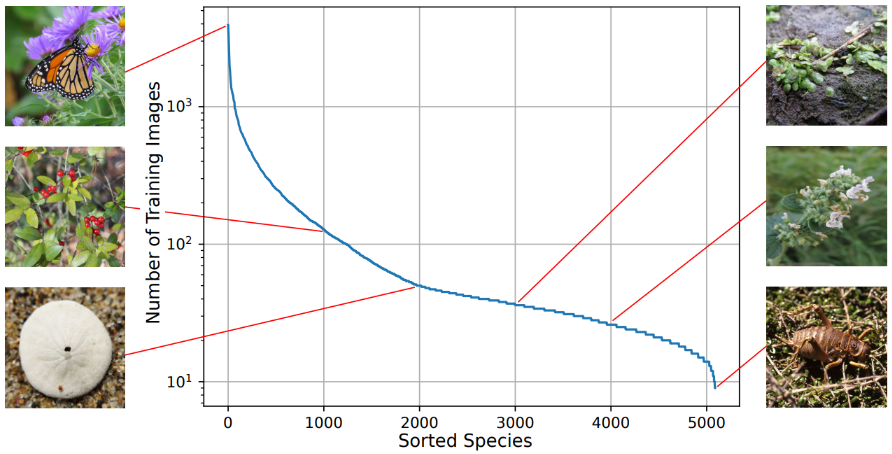

Train
Cat
(many-shot class)
Fox
(medium-shot class)
Panda
(few-shot class)
Test
Cat
Fox
Panda
Cat
Fox
Panda
?
(open class)



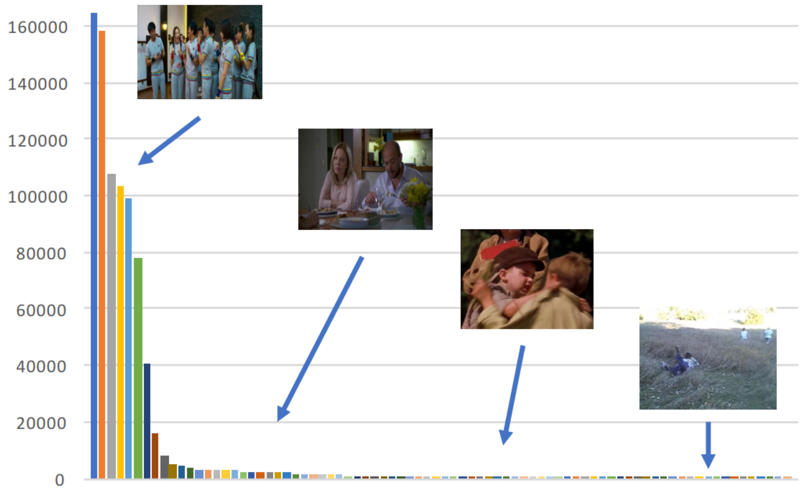
Faces [Zhang et al. 2017]



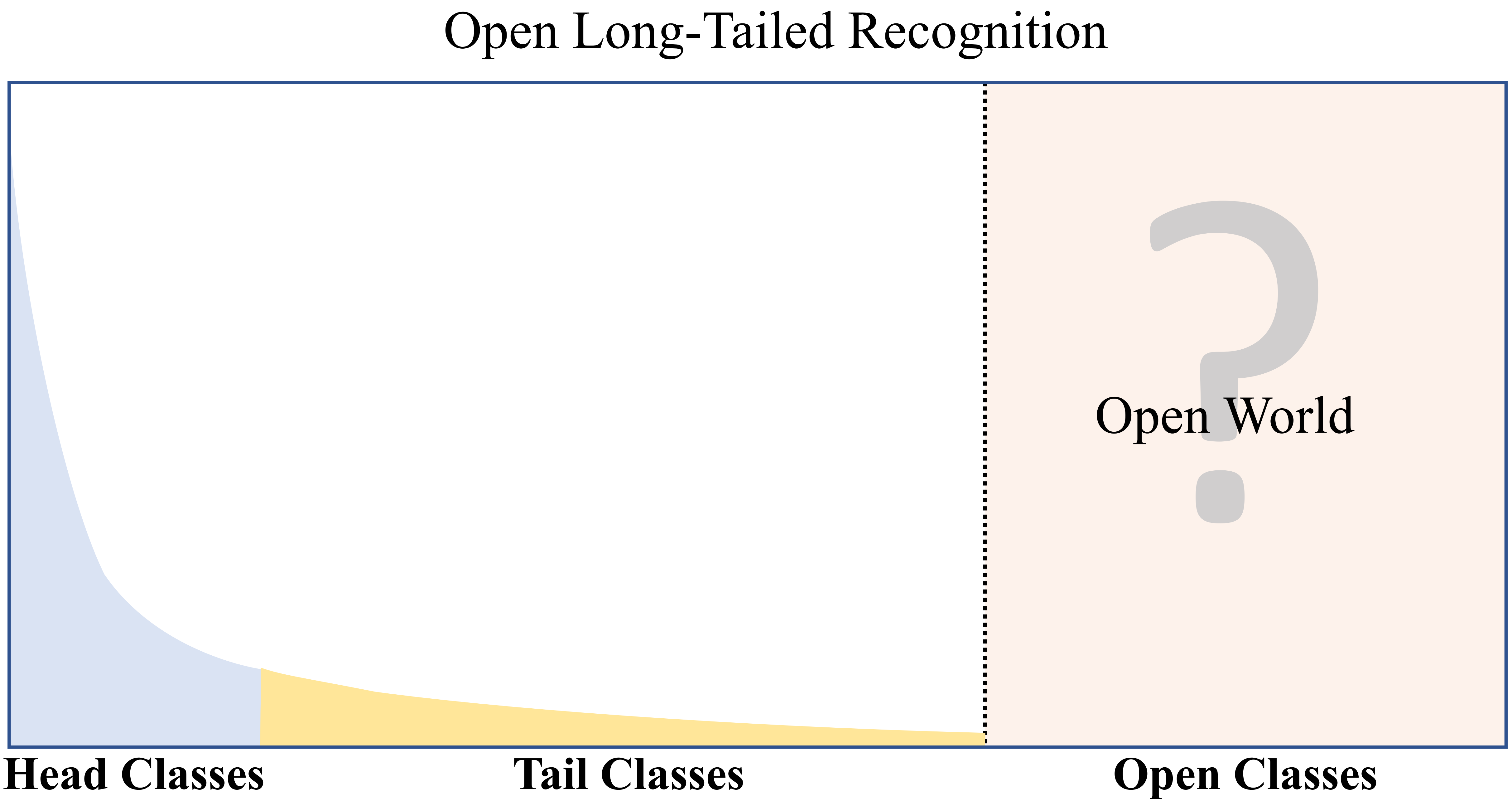
Places [Wang et al. 2017]



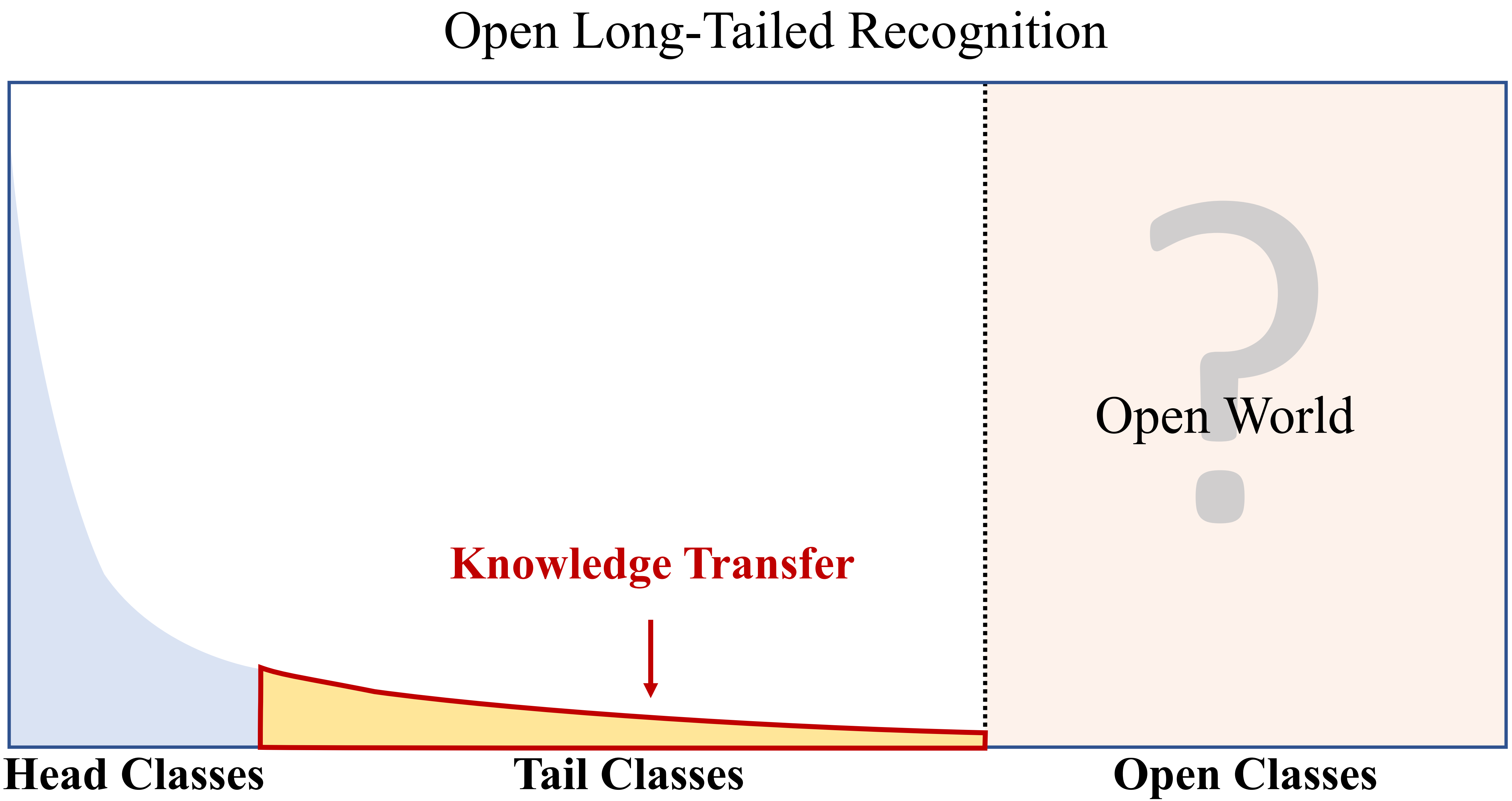
Species [Van Horn et al. 2019]

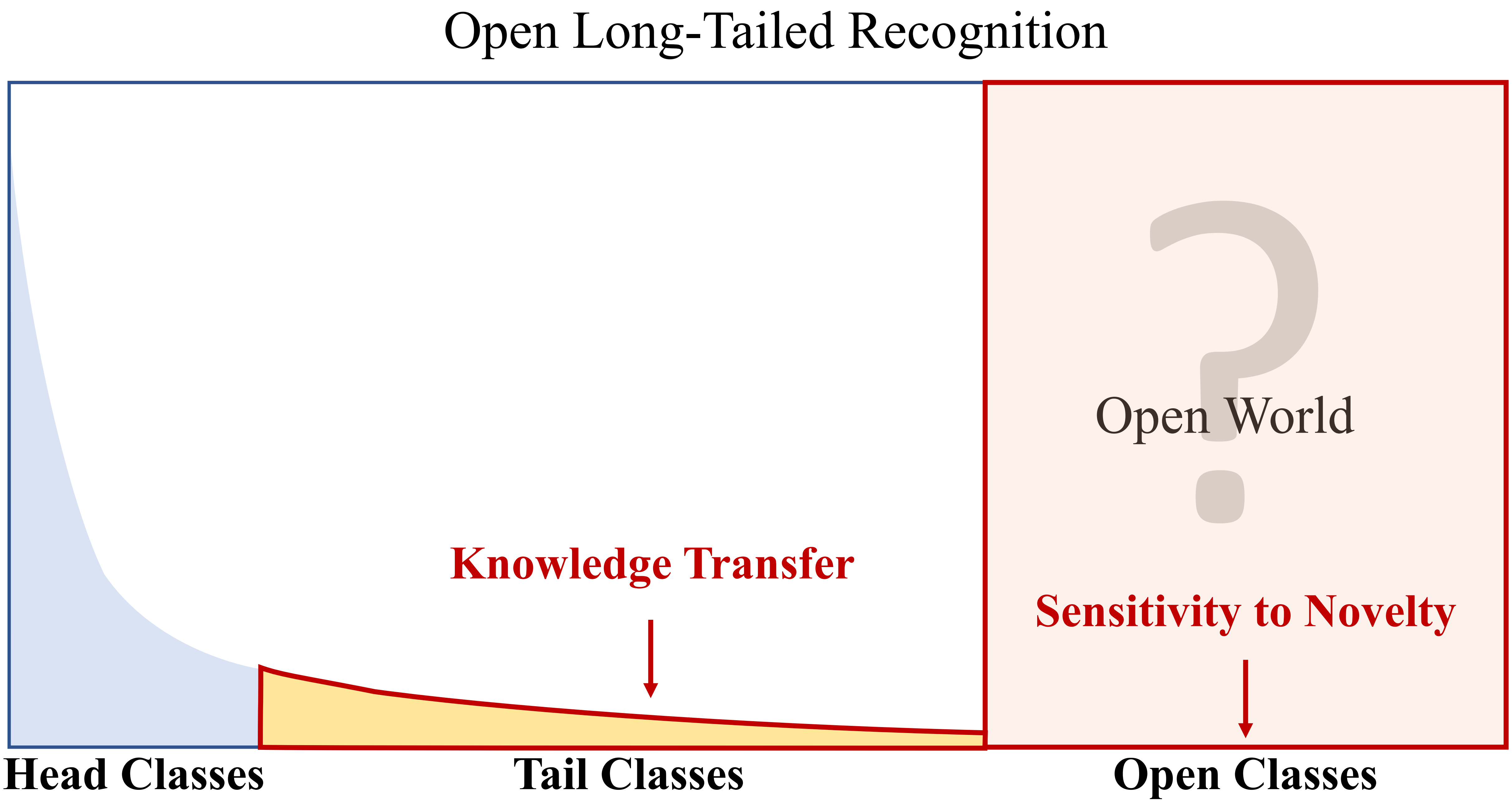


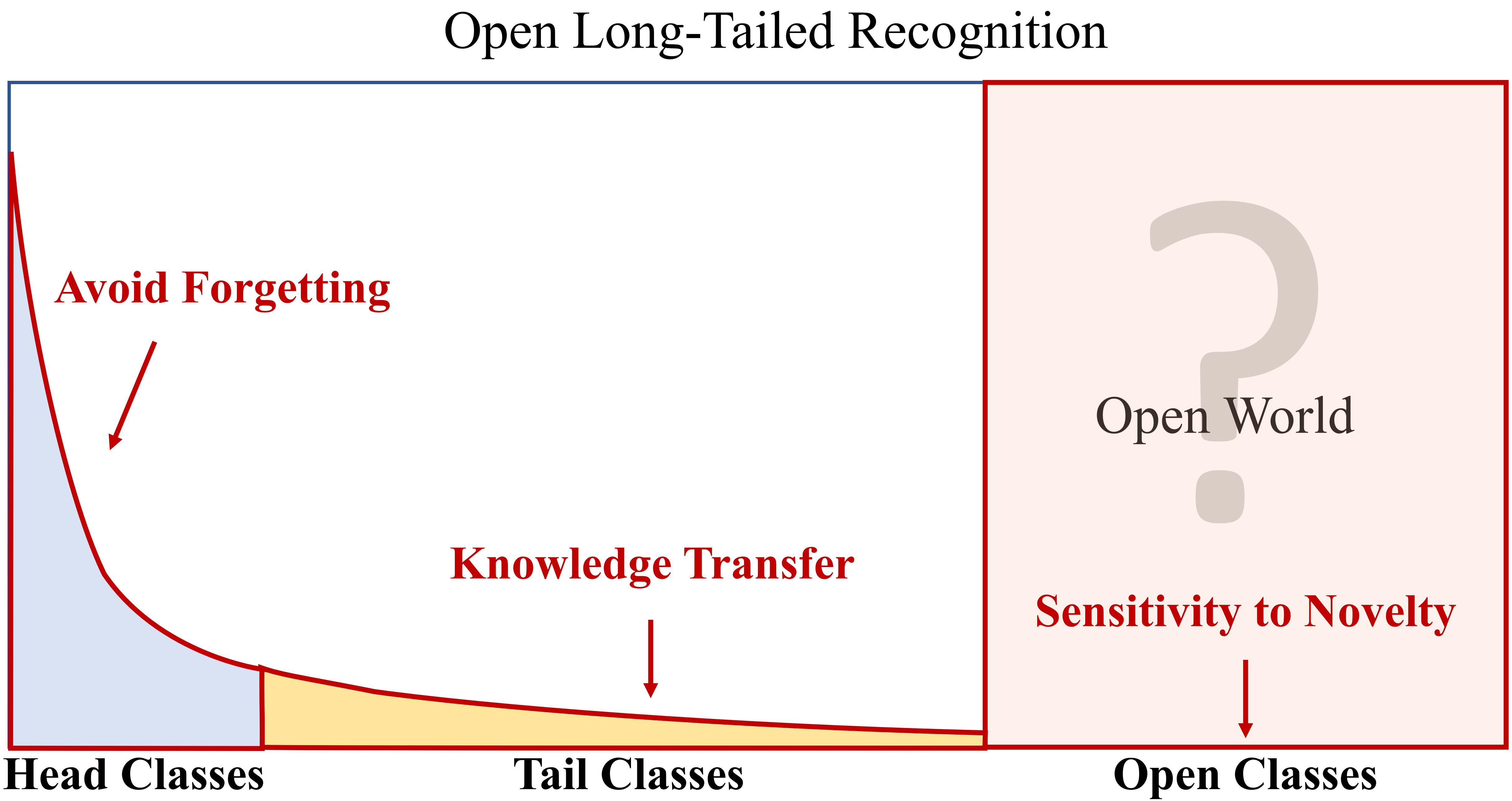
Actions [Zhang et al. 2019]

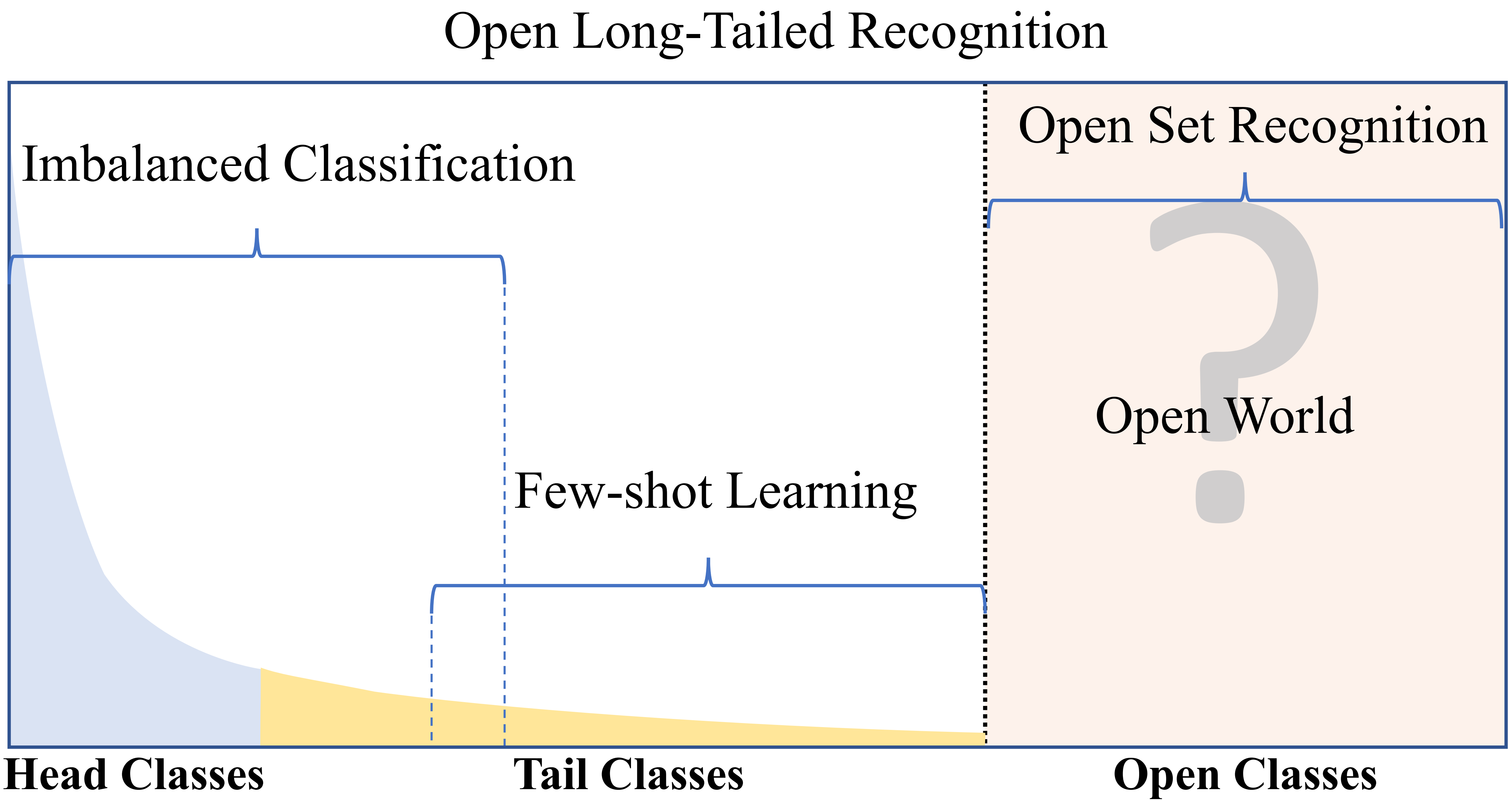


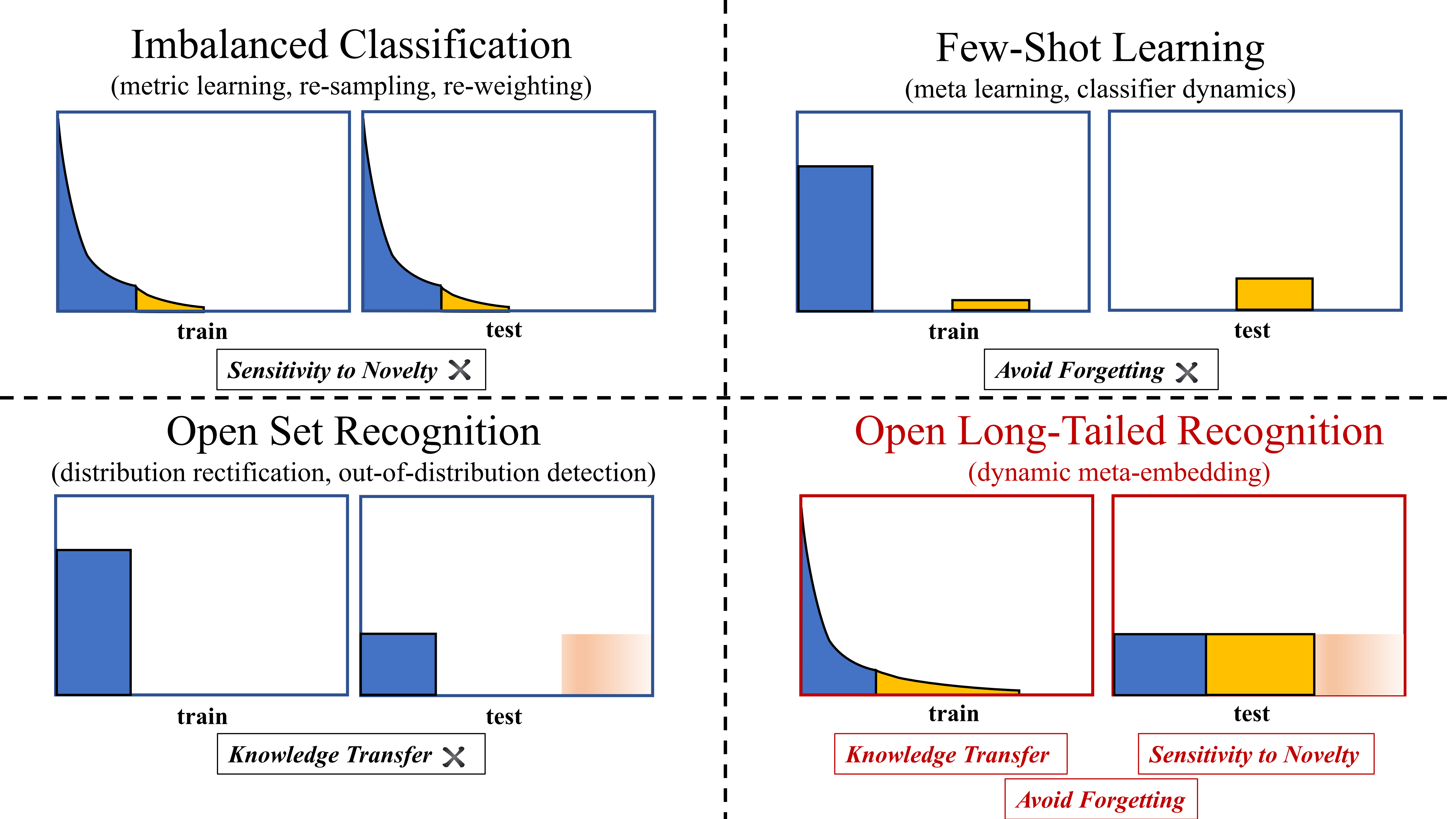
Open Long-Tailed Recognition
Open World
Head Classes
Tail Classes
Open Classes


Open Long-Tailed Recognition
Open World
Knowledge Transfer
Head Classes
Tail Classes
Open Classes


Open Long-Tailed Recognition
Open World
Knowledge Transfer
Sensitivity to Novelty
Head Classes
Tail Classes
Open Classes


Open Long-Tailed Recognition
Avoid Forgetting
Open World
Knowledge Transfer
Sensitivity to Novelty
Head Classes
Tail Classes
Open Classes

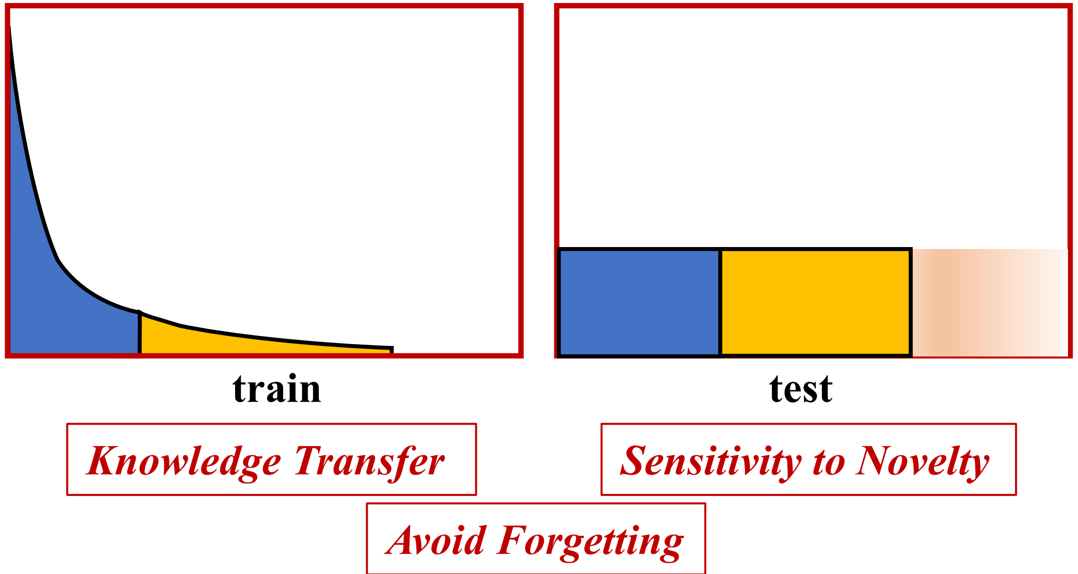

Open Long-Tailed Recognition
Open Set Recognition
Imbalanced Classification
Open World
Few-shot Learning
Head Classes
Tail Classes
Open Classes

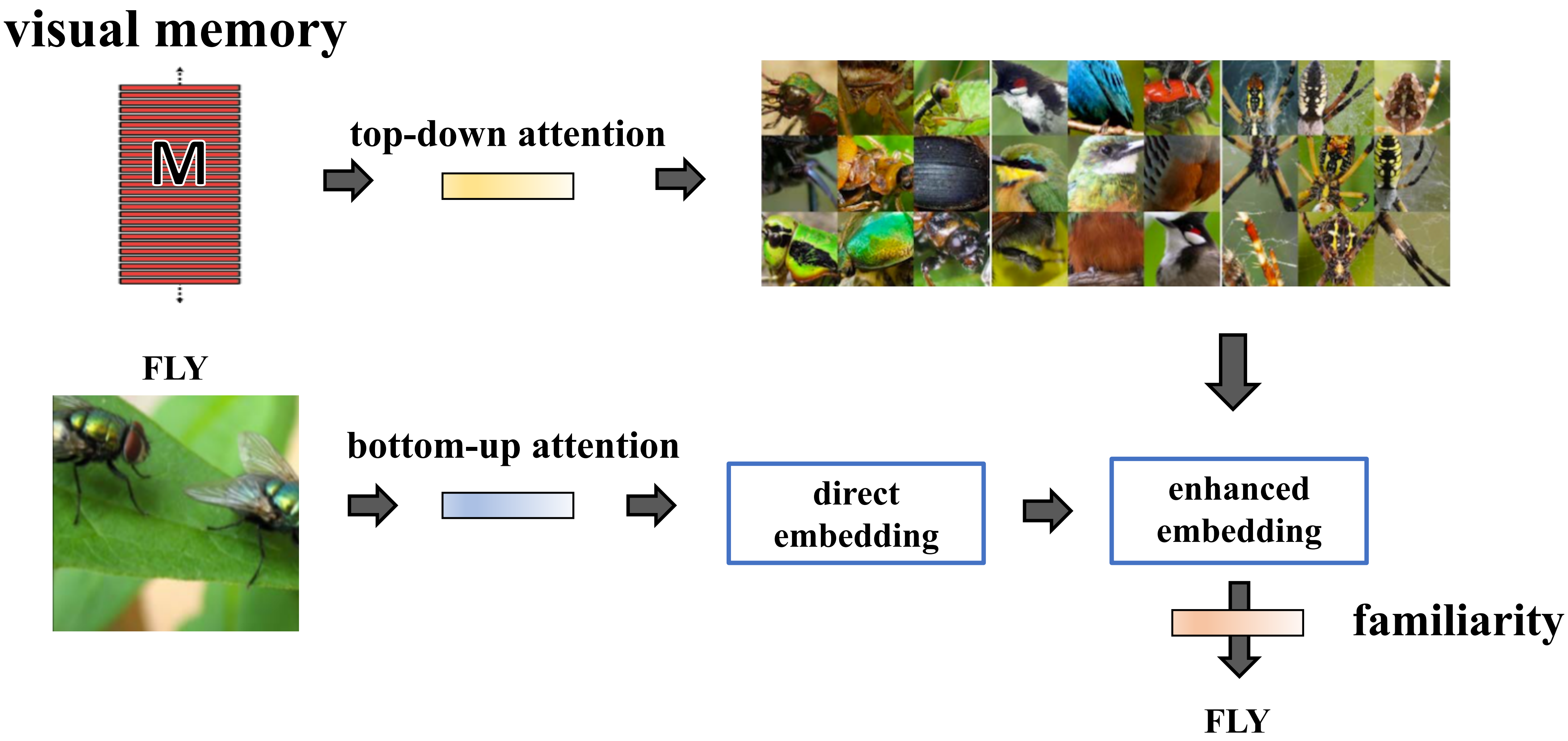


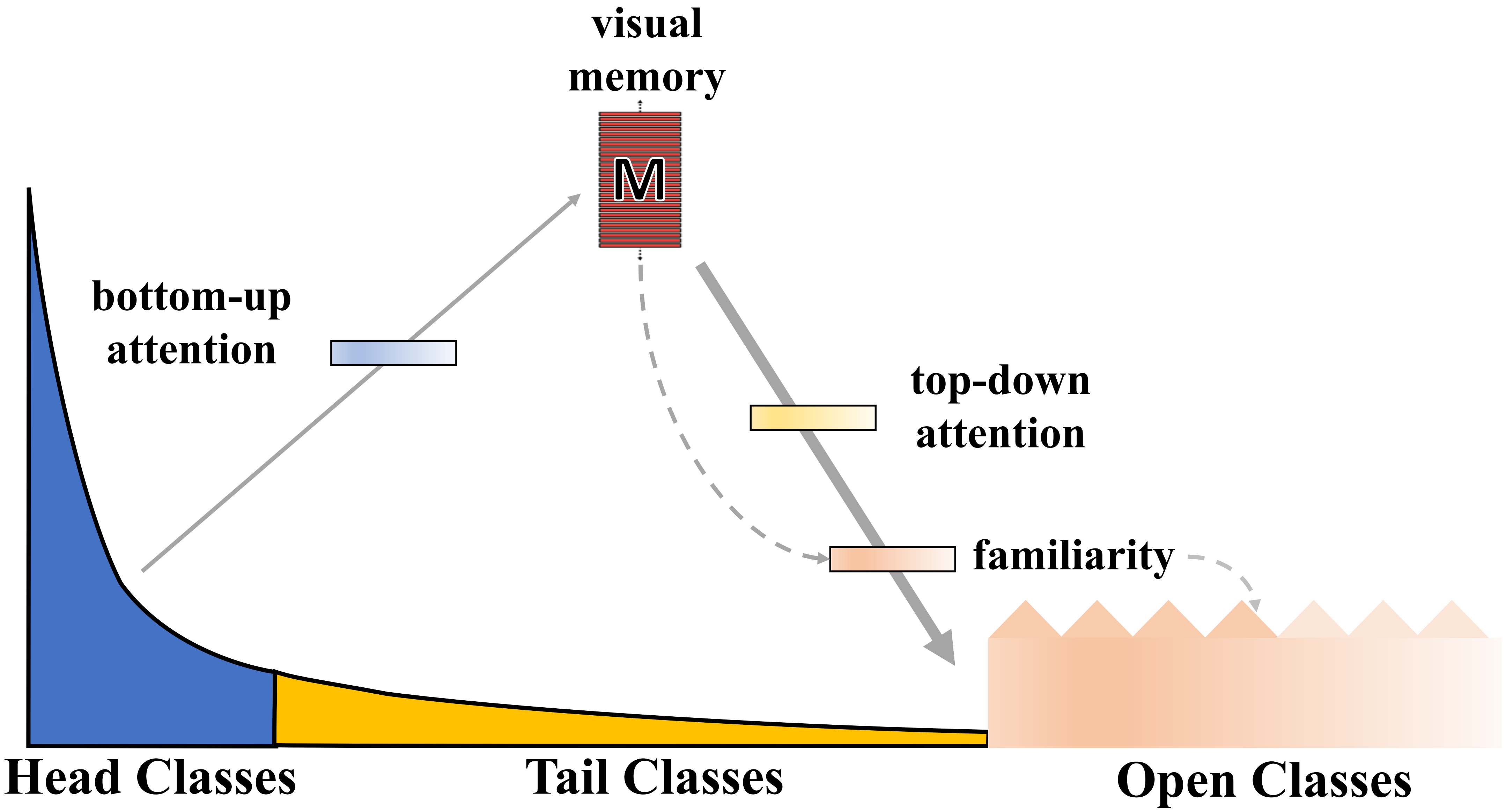
Imbalanced Classification
(metric learning, re-sampling, re-weighting)
train
test
Sensitivity to Novelty
Few-Shot Learning
(meta learning, classifier dynamics)
train
test
Avoid Forgetting
Open Set Recognition
(distribution rectification, out-of-distribution detection)
train
test
Knowledge Transfer
Open Long-Tailed Recognition
(dynamic meta-embedding)
train
test
Knowledge Transfer
Sensitivity to Novelty
Avoid Forgetting

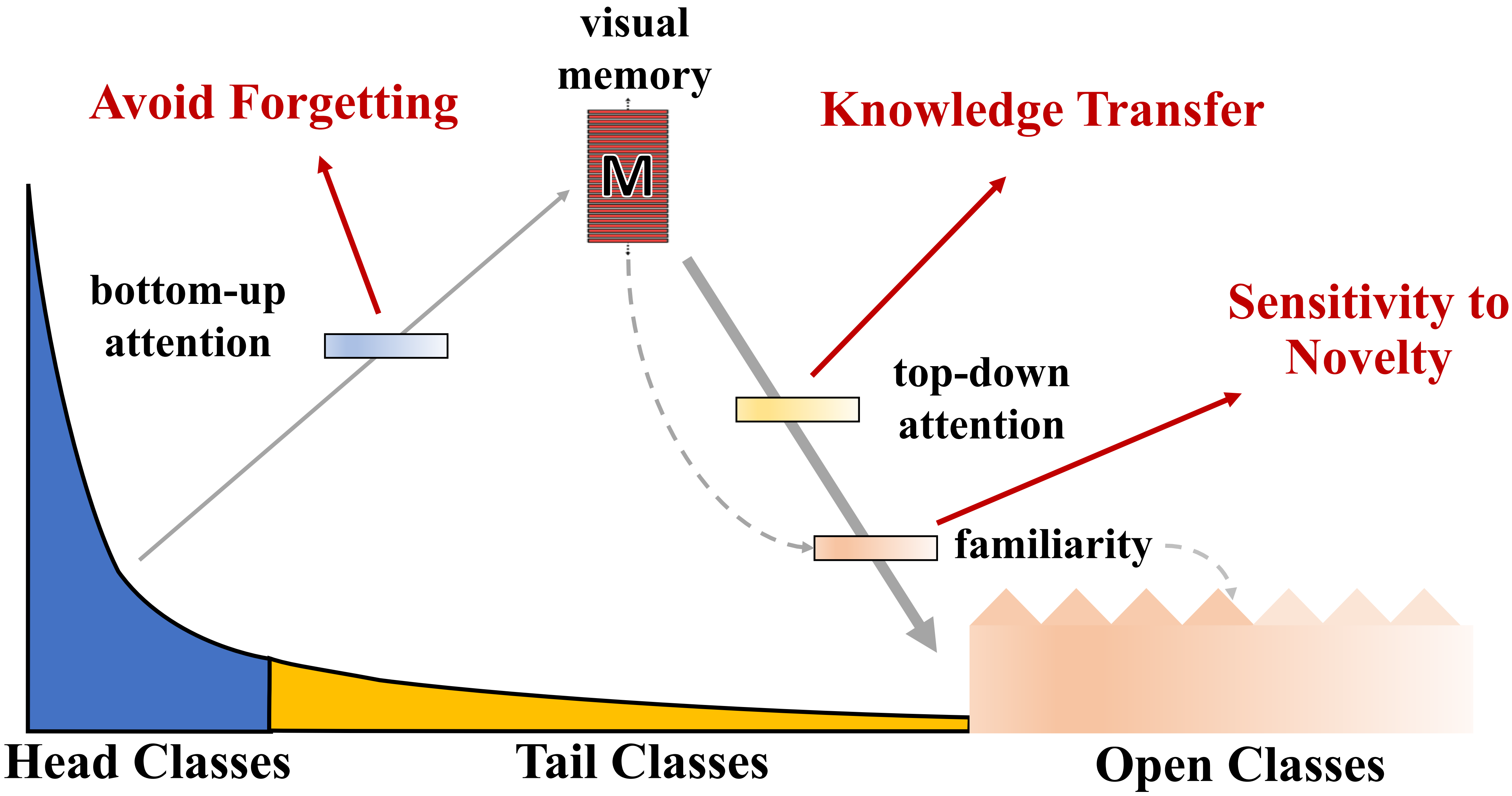
Open Long-Tailed Recognition

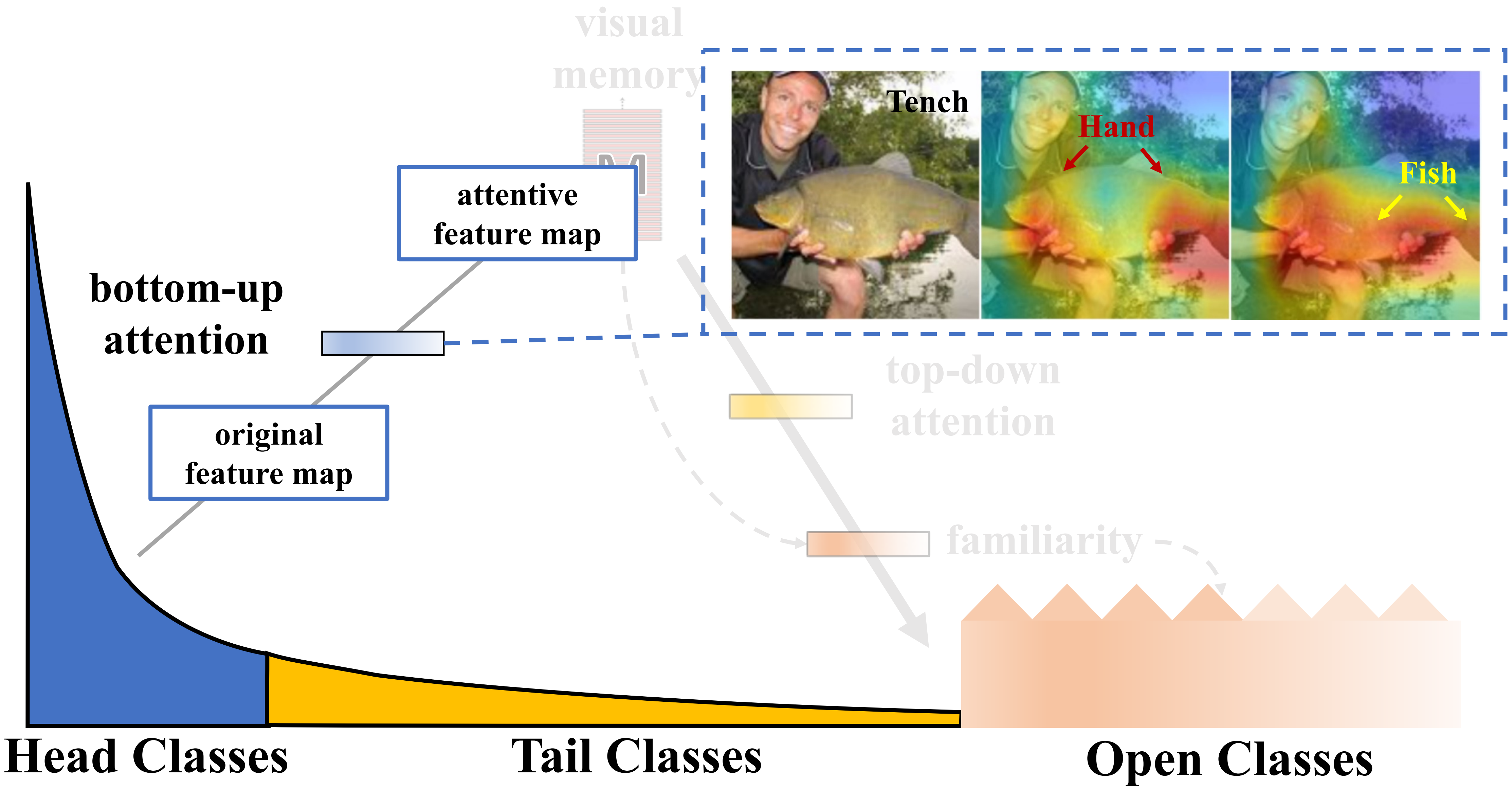
(dynamic meta-embedding)

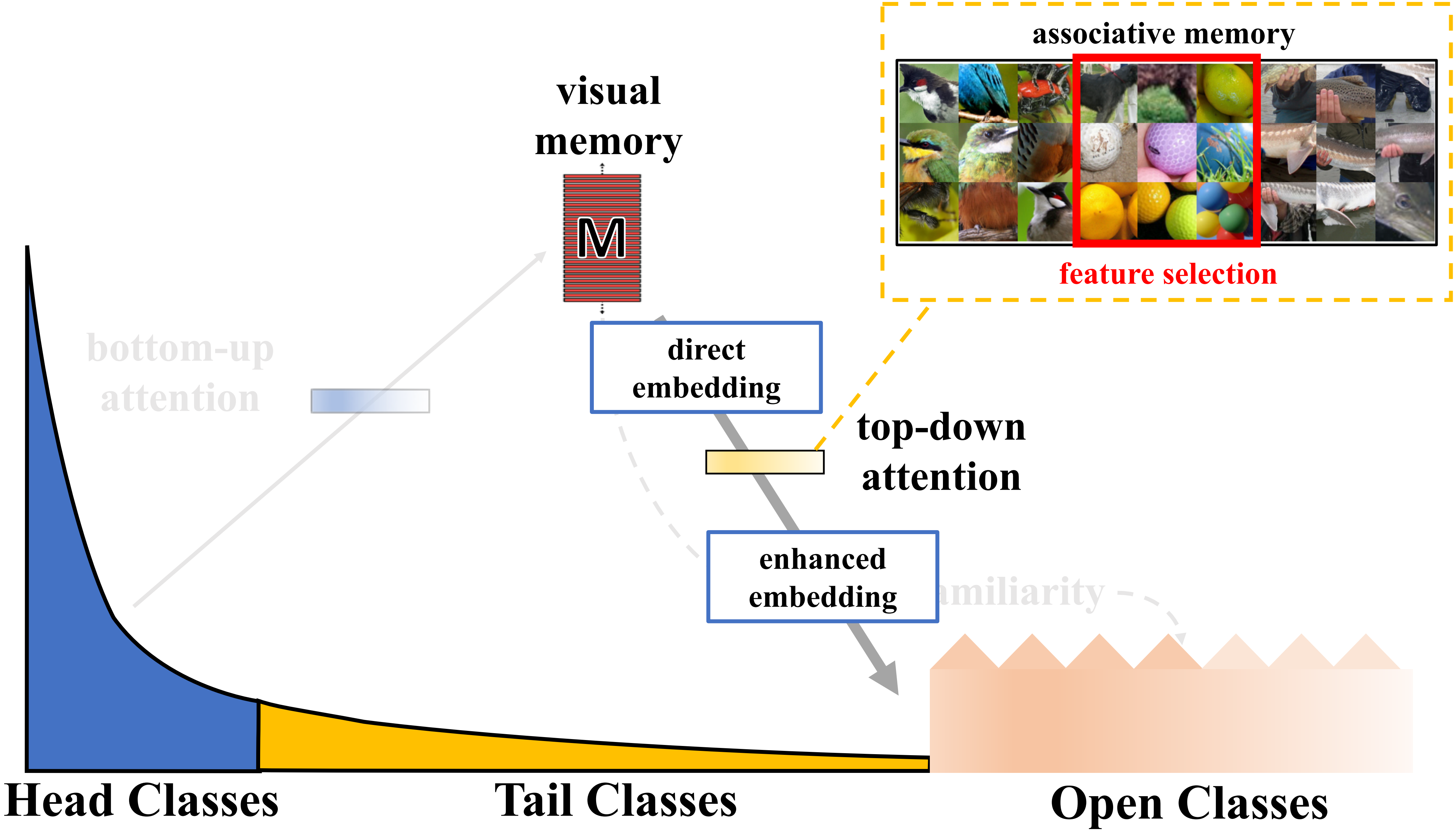


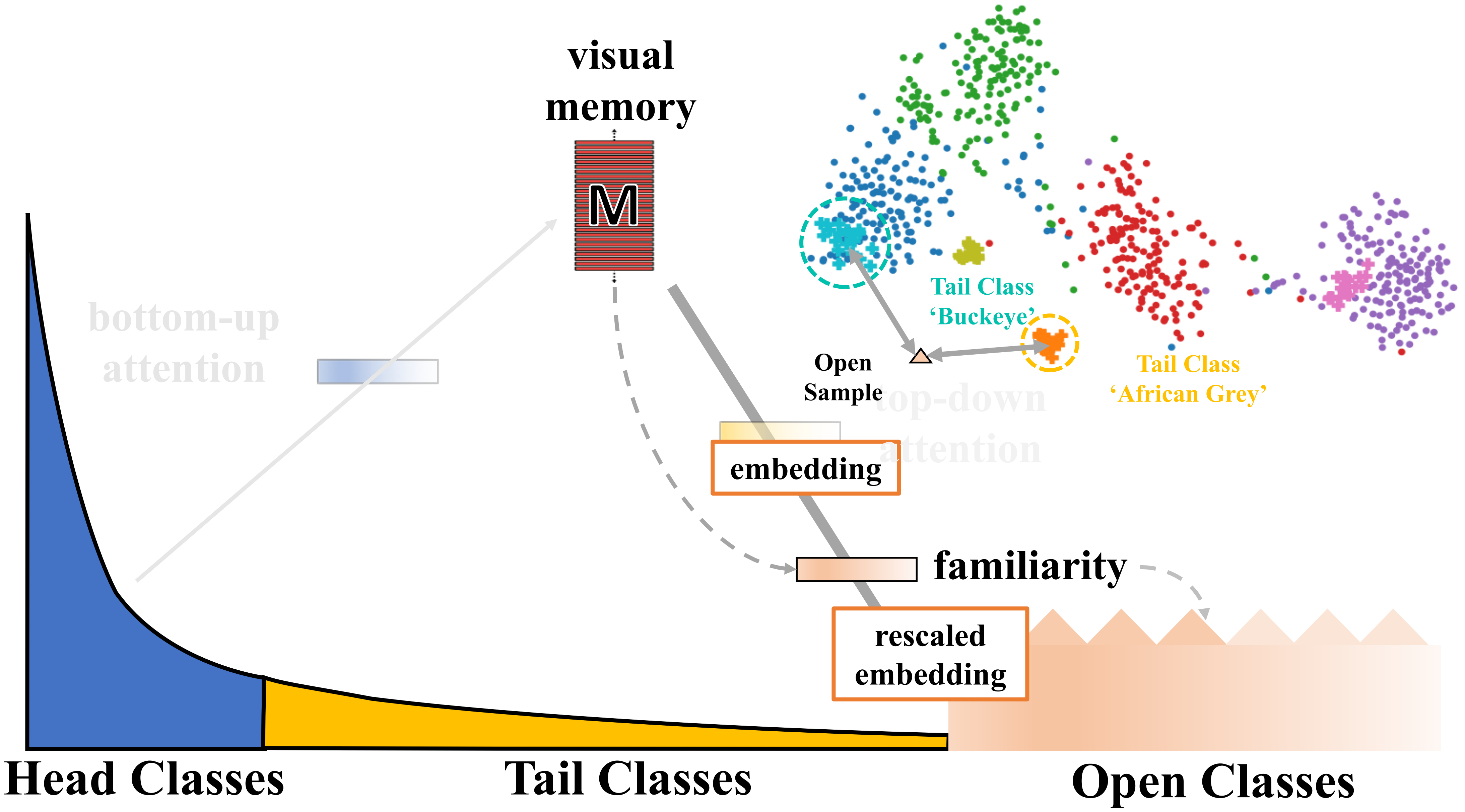

visual memory
M
top-down attention
FLY
bottom-up attention
direct embedding
enhanced embedding
familiarity
FLY


visual memory
M
bottom-up attention
top-down attention
familiarity
Head Classes
Tail Classes
Open Classes


visual memory
M
Avoid Forgetting
Knowledge Transfer
Sensitivity to Novelty
bottom-up attention
top-down attention
familiarity
Head Classes
Tail Classes
Open Classes

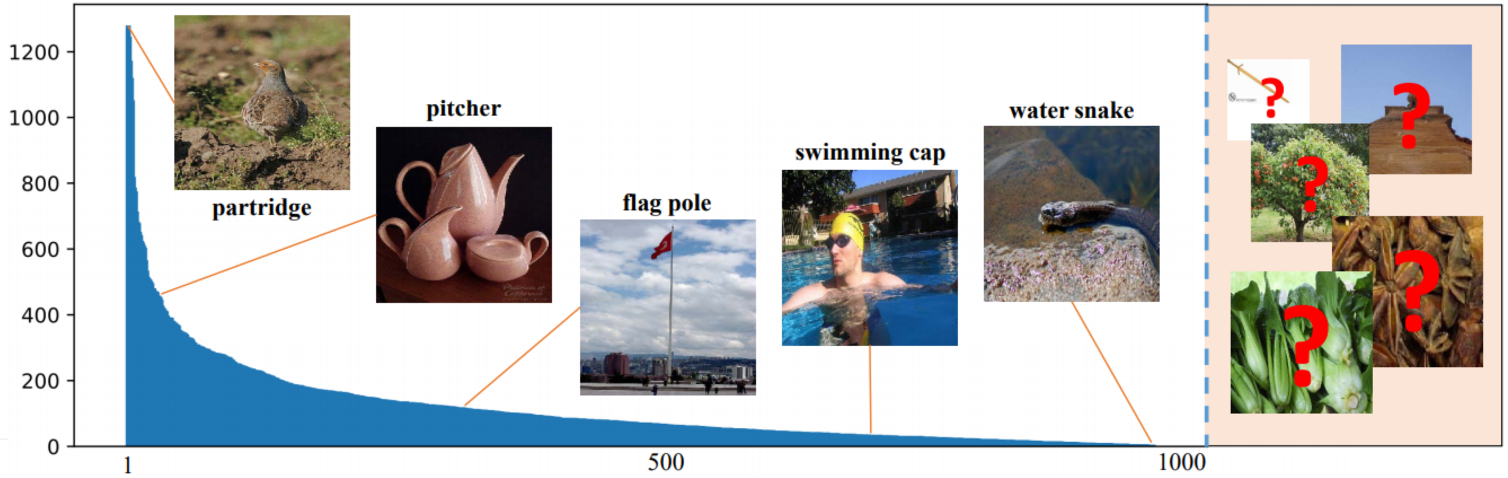

visual
memory
Tench
Hand
Fish
attentive
feature map
bottom-up
attention
top-down
attention
original
feature map
familiarity
Head Classes
Tail Classes
Open Classes


associative memory
feature selection
visual memory
M
bottom-up attention
direct embedding
top-down attention
enhanced embedding
familiarity
Head Classes
Tail Classes
Open Classes


visual memory
M
bottom-up attention
Tail Class 'Buckeye'
Open Sample
Tail Class 'African Grey'
top-down attention
embedding
familiarity
rescaled embedding
Head Classes
Tail Classes
Open Classes

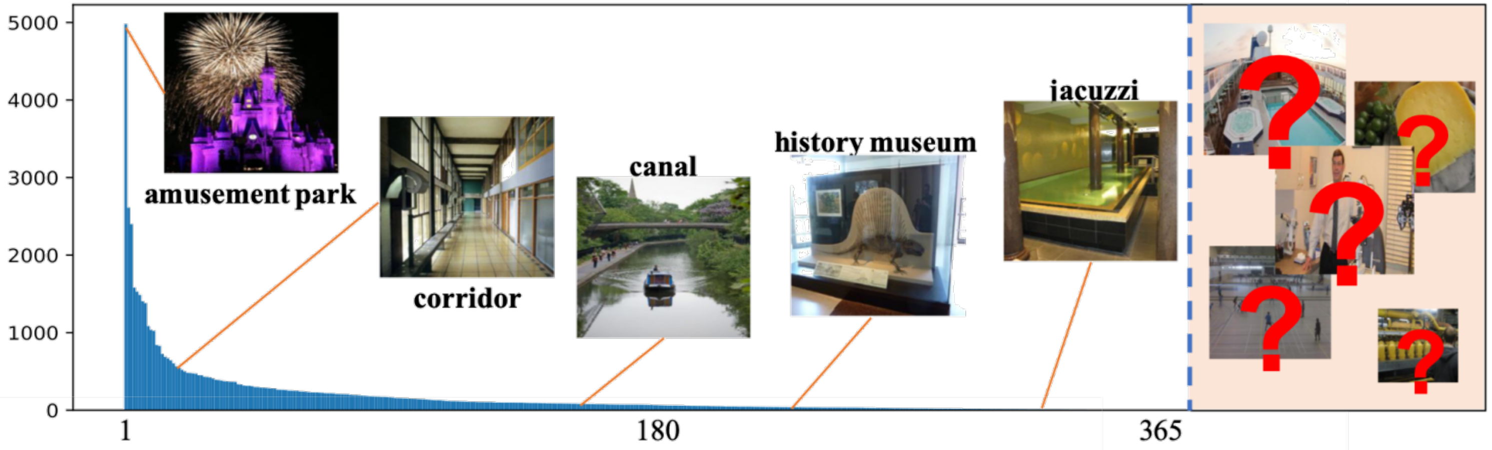
ImageNet-LT Benchmark

Absolute Performance Gain: ~20%



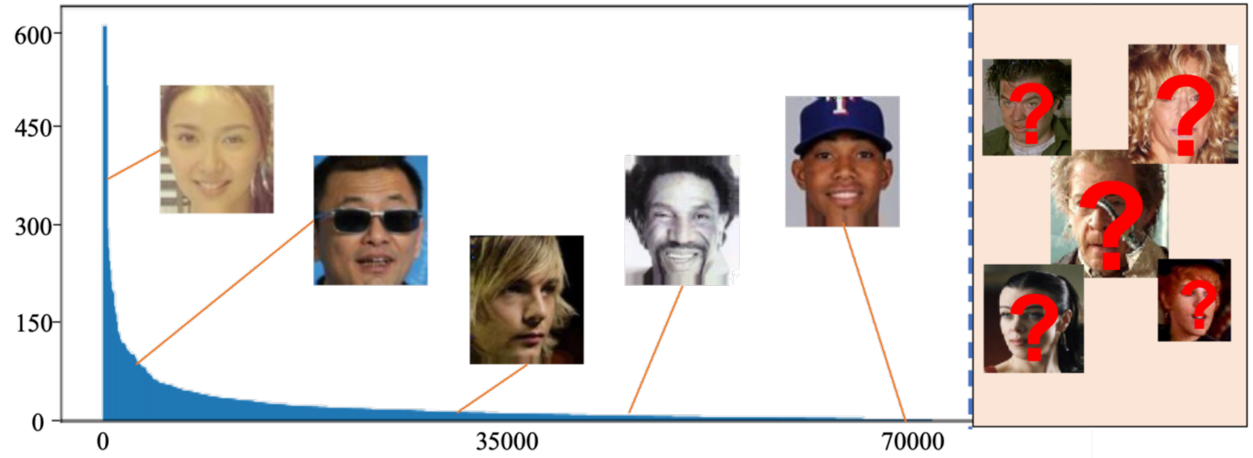
Places-LT Benchmark

Absolute Performance Gain: ~10%



MS1M-LT Benchmark

Absolute Performance Gain: ~2%



Overall F1 Score on **ImageNet-LT**, **Places-LT** and **MS1M-LT** Benchmarks

Methods	**ImageNet-LT**	**Places-LT**	**MS1M-LT**
Plain Model	0.295	0.366	0.738
Sample Re-weighting (Focal Loss)	0.371	0.453	-
Metric Learning (Range Loss)	0.373	0.457	0.722
Open Set Recognition (OpenMax)	0.368	0.458	-
Few-shot Learning (FSLwF)	0.347	0.375	-
Dynamic Meta-Embedding	**0.474**	**0.464**	**0.745**

Overall F1 Score on **ImageNet-LT**, **Places-LT** and **MS1M-LT** Benchmarks

Methods	**ImageNet-LT**	**Places-LT**	**MS1M-LT**
Plain Model	0.295	0.366	0.738
Sample Re-weighting (Focal Loss)	0.371	0.453	-
Metric Learning (Range Loss)	0.373	0.457	0.722
Open Set Recognition (OpenMax)	0.368	0.458	-
Few-shot Learning (FSLwF)	0.347	0.375	-
Dynamic Meta-Embedding	**0.474**	**0.464**	**0.745**

Overall F1 Score on **ImageNet-LT**, **Places-LT** and **MS1M-LT** Benchmarks

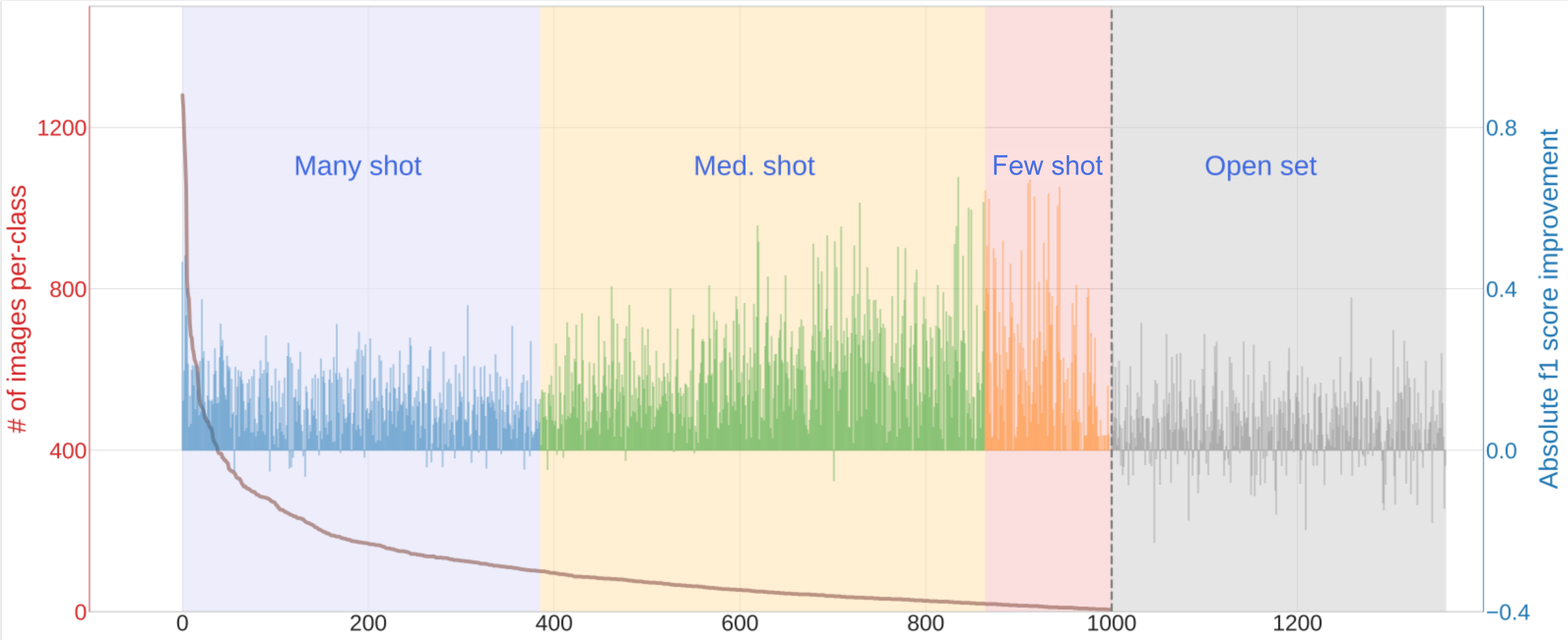
Methods	**ImageNet-LT**	**Places-LT**	**MS1M-LT**
Plain Model	0.295	0.366	0.738
Sample Re-weighting (Focal Loss)	0.371	0.453	-
Metric Learning (Range Loss)	0.373	0.457	0.722
Open Set Recognition (OpenMax)	0.368	0.458	-
Few-shot Learning (FSLwF)	0.347	0.375	-
Dynamic Meta-Embedding	**0.474**	**0.464**	**0.745**

Overall F1 Score on **ImageNet-LT**, **Places-LT** and **MS1M-LT** Benchmarks

Methods	**ImageNet-LT**	**Places-LT**	**MS1M-LT**
Plain Model	0.295	0.366	0.738
Sample Re-weighting (Focal Loss)	0.371	0.453	-
Metric Learning (Range Loss)	0.373	0.457	0.722
Open Set Recognition (OpenMax)	0.368	0.458	-
Few-shot Learning (FSLwF)	0.347	0.375	-
Dynamic Meta-Embedding	**0.474**	**0.464**	**0.745**

Overall F1 Score on **ImageNet-LT**, **Places-LT** and **MS1M-LT** Benchmarks

Methods	**ImageNet-LT**	**Places-LT**	**MS1M-LT**
Plain Model	0.295	0.366	0.738
Sample Re-weighting (Focal Loss)	0.371	0.453	-
Metric Learning (Range Loss)	0.373	0.457	0.722
Open Set Recognition (OpenMax)	0.368	0.458	-
Few-shot Learning (FSLwF)	0.347	0.375	-
Dynamic Meta-Embedding	**0.474**	**0.464**	**0.745**



Many shot
Med. shot
Few shot
Open set
# of images per-class
Absolute f1 score improvement
0
400
800
1200
0
200
400
600
800
1000
1200
−0.4
0.0
0.4
0.8

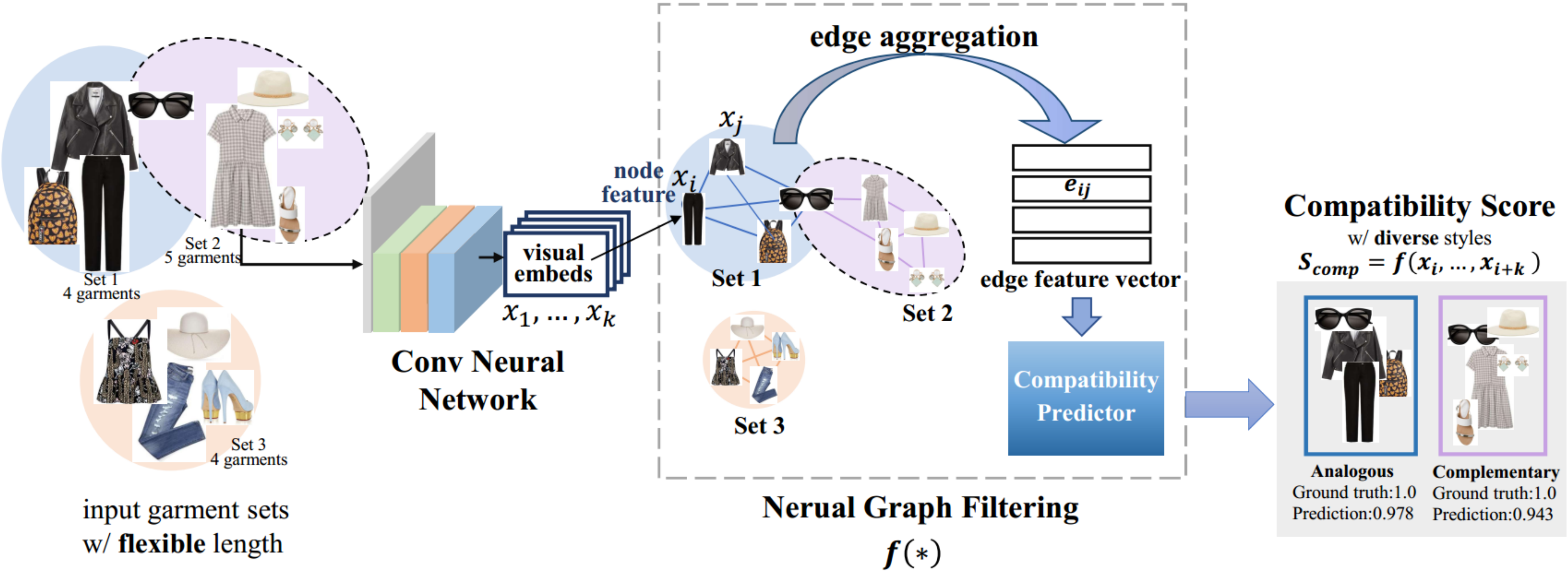
Diverse Relations

Learning Diverse Fashion Collocation by Neural Graph Filtering, (in submission)

Motivation

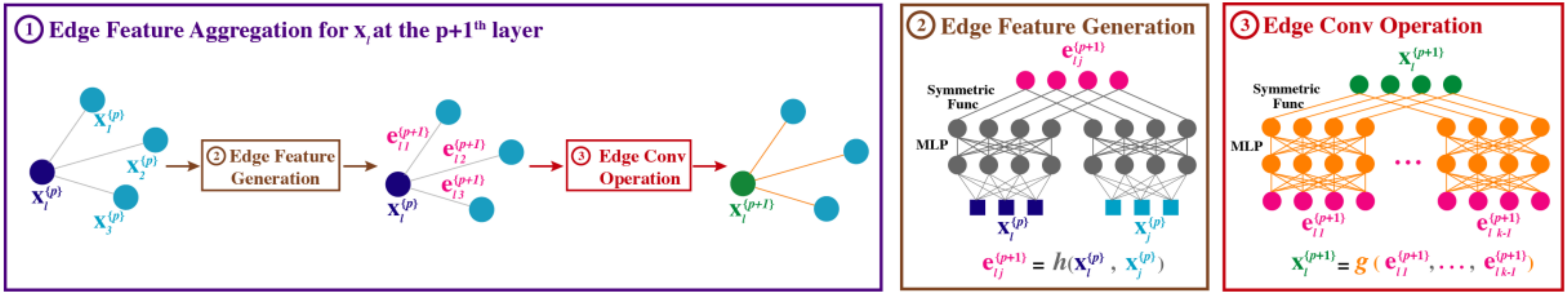
- Increasing demand for intelligent fashion recommendation system
- A successful fashion collocation framework should be featured with two desired properties: **Flexibility** and **Diversity**.
- Existing work can only accept fashion sets with ***fixed length***, e.g., the four-garment set{tops, outerwear, bottoms and shoes} and ***limited categories***, e.g., discarding accessories, bags and hats.

Overall Framework of Diverse Fashion Graph Filtering



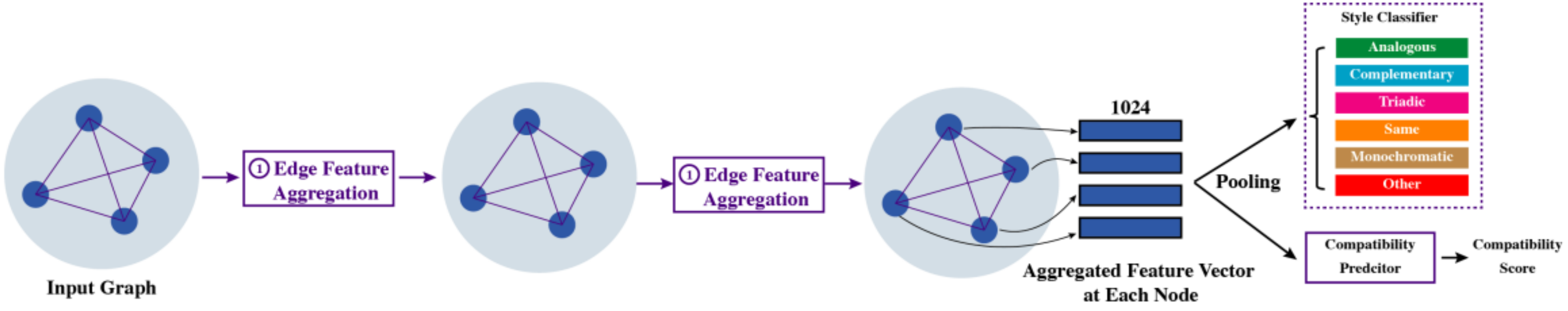
We firstly use the convolutional neural networks to extract the visual embeddings of the input garment sets with **flexible** length, and then consider each visual embedding as a node input to the neural graph network, which not only computes the node features, but also implements edge feature aggregation. Note that one node could appear in several collocations. Afterwards a compatibility predictor calculates the compatibility scores for **diverse** styled garment sets.

Architecture of Neural Graph Filtering



- The graph network architecture constructed using **edge feature aggregation** operations.
- In the last layer, edge information gathered at all the nodes are pooled to compute a compatibility score, and an optional fashion style distribution for a compatible garment set.

Architecture of Neural Graph Filtering



- Graph edge Filtering at **one layer**: aggregates all the edge information connecting to the node under consideration.

Quantitative Evaluation

dataset	Polyvore		Polyvore-D				Polyvore		Polyvore-D	
Metric	AUC	FITB	AUC	FITB	H.(%)		AUC	FITB	AUC	FITB
Bi-LSTM (Han et al. 2017)	0.65	39.7	0.62	39.4	5.0	Euclidean Distance	0.85	54.7	0.82	53.4
CSN (Veit, Belongie, and Karaletsos 2017)	0.83	54.0	0.82	52.5	0	Imbalanced Collocation Handling	0.85	55.1	0.83	54.2
TransNFCM (Xun Yang 2019)	0.75	-	-	-	-	Baseline (Node)	0.92	55.3	0.84	47.8
Wardrobe (Wei-Lin Hsiao 2018)	0.88	-	-	-	7.5	Baseline (Edge Max Pooling)	0.93	57.7	0.87	52.8
Type Aware (Vasileva et al. 2018)	0.86	56.2	0.84	54.9	5.0	Baseline (Edge Avg Pooling)	0.93	58.0	0.86	53.8
Neural Graph Filtering (Ours)	0.94	58.8	0.88	55.1	**82.5**	**Neural Graph Filtering (Ours)**	**0.94**	**58.8**	**0.88**	**55.1**

Fill-in-blank

given a sequence of fashion items, ask for the most compatible one from the four choices

Fashion Compatibility Prediction

score a candidate outfit, higher score means more compatibility

0.805

0. 994

compatible

0. 041 **not compatible**

Diverse Fashion Collocations

Given 1 query item, generate fashion sets of diverse styles and flexible length

Dataset: Polyvore

Diverse Fashion Collocations

Given 1 query item, generate fashion sets of diverse styles and flexible length

Diverse Fashion Collocations

Given 1 query item, generate fashion sets of diverse styles and flexible length

Diverse Fashion Collocations

Dataset: Amazon Fashion

Diverse Fashion Collocations

Dataset: Amazon Fashion

Conclusions

- The concept of **flexible** and **diverse** fashion collocations:
  - support both inputs/outputs with flexible lengths;
  - generate fashion sets with diverse styles
- Novel framework of **neural graph filtering**
  - the graph structure that explores the inter-garment relationship is more suitable for fashion compatibility learning.
- Newly proposed benchmark and evaluation protocols
  - *AmazonFashion* Dataset: comprises of different styles for diversity learning and evaluation

Database and Toolbox

Two New Datasets:

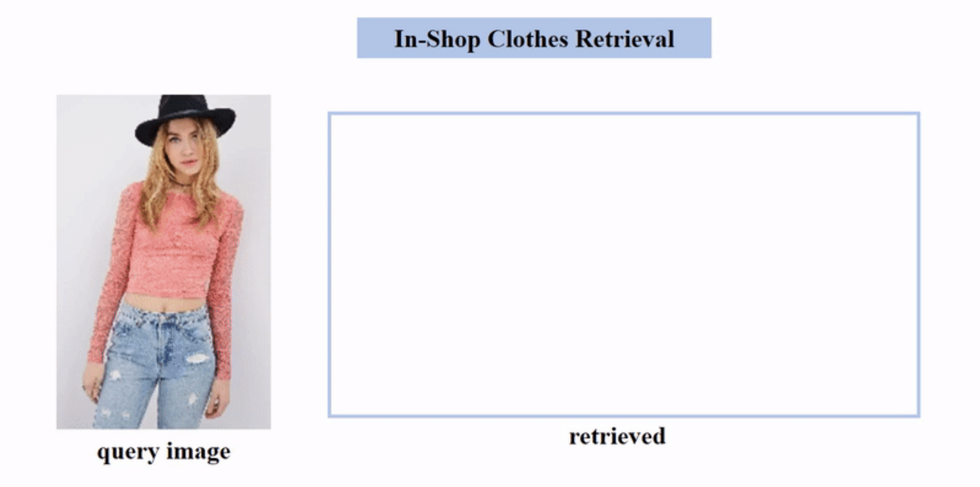
- Fashion Parsing Benchmark
- Fashion Recommendation Benchmark

MMFashion

Open-source toolbox for visual fashion analysis based on PyTorch: https://github.com/open-mmlab/mmfashion

Features

- **Flexible:** modular design and easy to extend
- **Friendly:** off-the-shelf models for layman users
- **Comprehensive:** support a wide spectrum of fashion analysis tasks
  - [x] Fashion Attribute Prediction
  - [x] Fashion Recognition and Retrieval
  - [x] Fashion Landmark Detection
  - [ ] Fashion Parsing and Segmentation
  - [ ] Fashion Compatibility and Recommendation



Thanks!

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

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