北京大学
PEKING UNIVERSITY

PKU_WICT at TRECVID 2022: Disaster Scene Description and Indexing Task

Yanzhe Chen, HsiaoYuan Hsu, James Ye, Zhiwen Yang, Zishuo Wang,
Xiangteng He, and Yuxin Peng*

Wangxuan Institute of Computer Technology, Peking University

Beijing, China

{pengyuxin@pku.edu.cn}

Outline

北京大学
PEKING UNIVERSITY

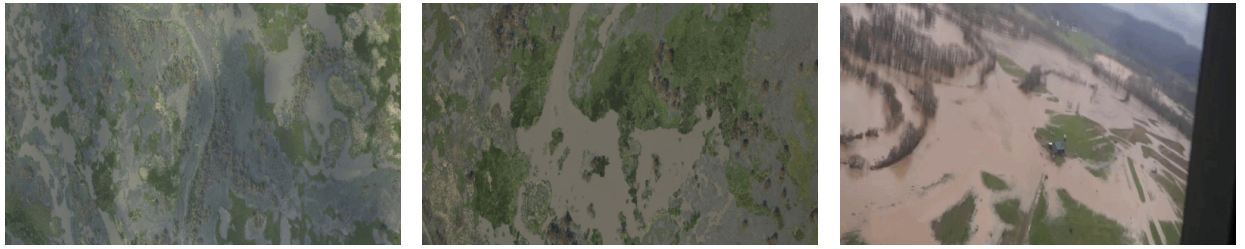
- Introduction
- Our approach
- Results and conclusions
- Our related works

Introduction

北京大学 PEKING UNIVERSITY

- **D**isaster **S**cene **D**escription and **I**ndexing (DSDI) task
  - —Given disaster-related features, retrieve videos containing each of them
  - —Development set:
    - ☐LADI (Low Altitude Disaster Imagery) dataset: in subtask “L”
    - ☐LADI + Others: in subtask “O”

“flooding” →



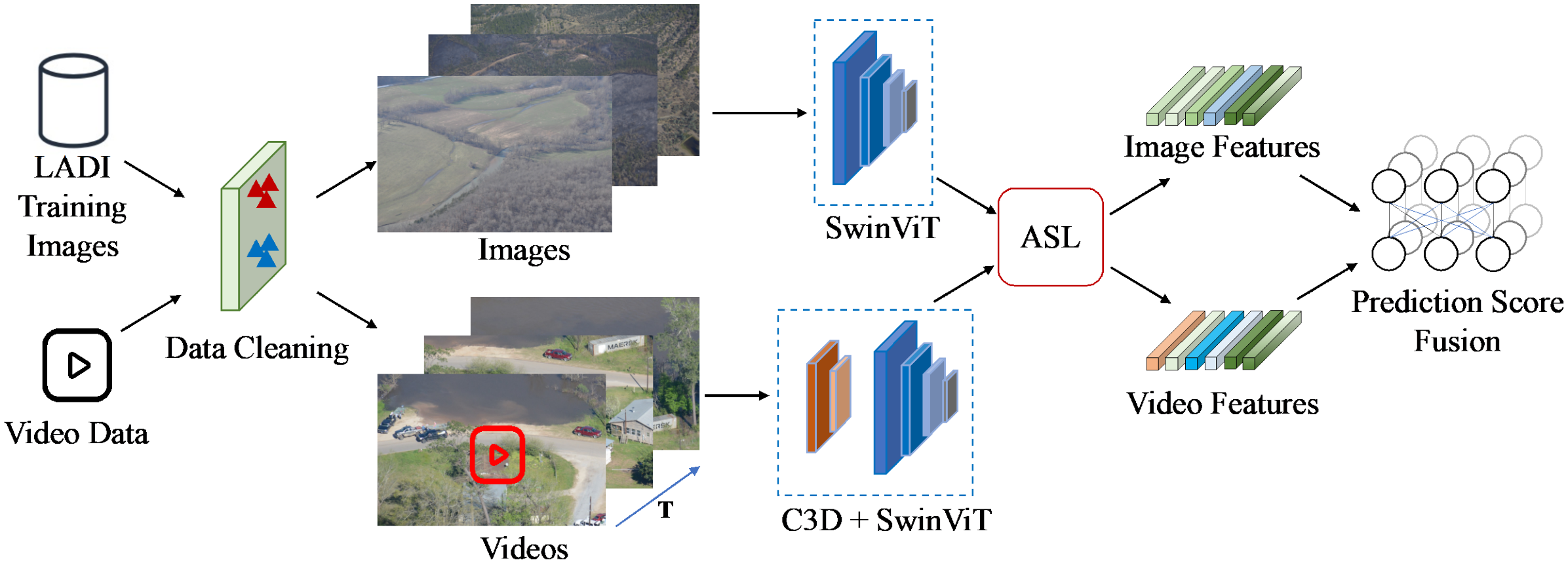
“smoke / fire” →

Outline

北京大学 PEKING UNIVERSITY

- Introduction
- Our approach
- Results and conclusions
- Our related works

Our approach



Data Cleaning
(on LADI, image-level or video-level labels)

0 ⇄ 1 p%

Revise
p% of skeptically wrong labels

Discard
(100-p)% of skeptically wrong labels

Visual Features Extraction

SwinViT

Image Backbone

Video Pre-fusion

Long-tail Distribution handling

W

Asymmetric Loss

Video Reconstruction

Data cleaning

北京大学 PEKING UNIVERSITY

- **Motivation:** Alleviate noise of the annotations in the LADI dataset

***None* of them were annotated!**

Data cleaning

北京大学 PEKING UNIVERSITY

- **Method:** Confident learning
- **Strategies:**
  - —**Revise:** samples with low confidence in the training set are revised
  - —**Discard:** samples with low confidence in the training set are discarded directly
  - —**Hybrid:** a portion of samples with the lowest confidence are revised, while the rest of them are directly discarded

Curtis Northcutt, Lu Jiang, Isaac Chuang, "**Confident learning**: Estimating uncertainty in dataset labels", Journal of Artificial Intelligence Research (JAIR), 2021, 70: 1373-1411.

Image Feature Extraction

北京大学 PEKING UNIVERSITY

- The backbone model plays an important role in the DSDI task
- We trained three backbones with the LADI dataset and evaluated on the DSDI-2021 testing set

Backbone	mAP
EfficientNet-B5	23.62
ViT	25.49
SwinViT	27.97

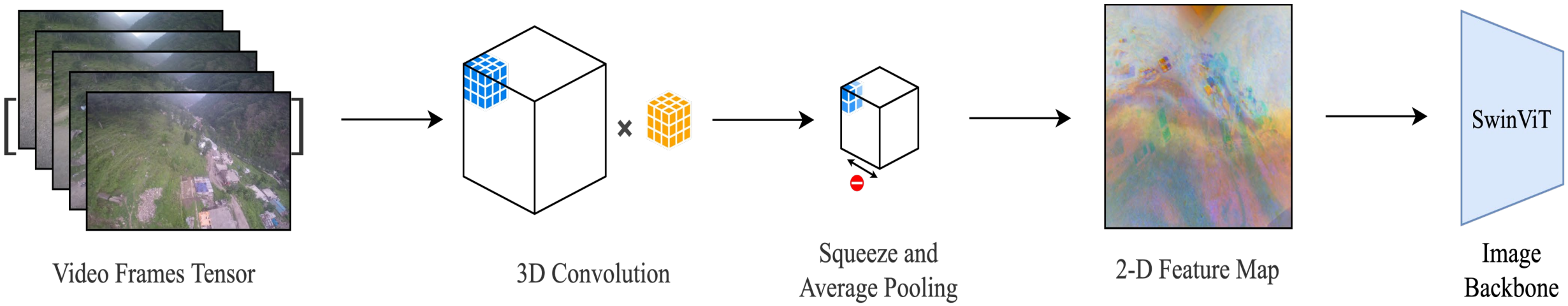
EfficientNet-B5:Efficientnet: Rethinking model scaling for convolutional neural networks, International conference on machine learning (ICML), 2019

ViT: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv, 2020.

SwinViT:Swin transformer: Hierarchical vision transformer using shifted windows. International Conference on Computer Vision (ICCV). 2021

Video Feature Extraction

- **Motivation:** Correlation between frames and temporal information can help understand videos
- **Method:** 3D CNN + 2D image backbone (fine-tuned on LADI)



Loss Function

北京大学 PEKING UNIVERSITY

- ASL loss is applied since the LADI dataset shows a long-tail distribution:

$$L_+ = (1-p)^{\gamma_+} \times \log(p)$$

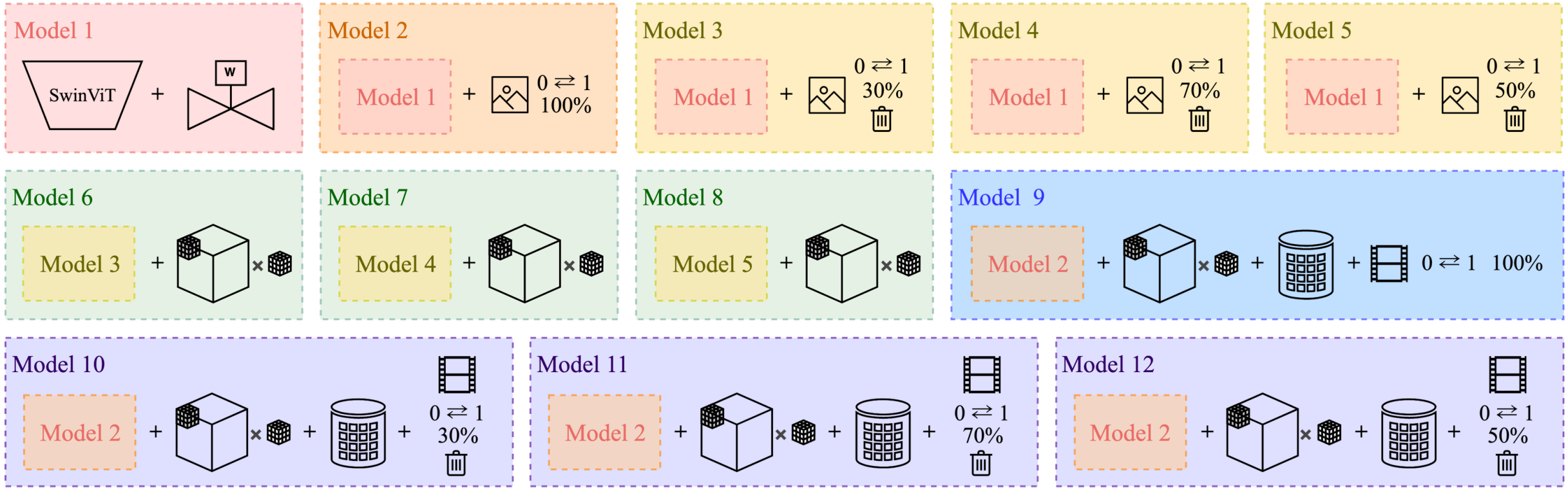
$$L_- = p_m^{\gamma_-} \times \log(1-p_m)$$

ASL loss: Asymmetric loss for multi-label classification. International Conference on Computer Vision, 2021.

Prediction Score Fusion

北京大学 PEKING UNIVERSITY

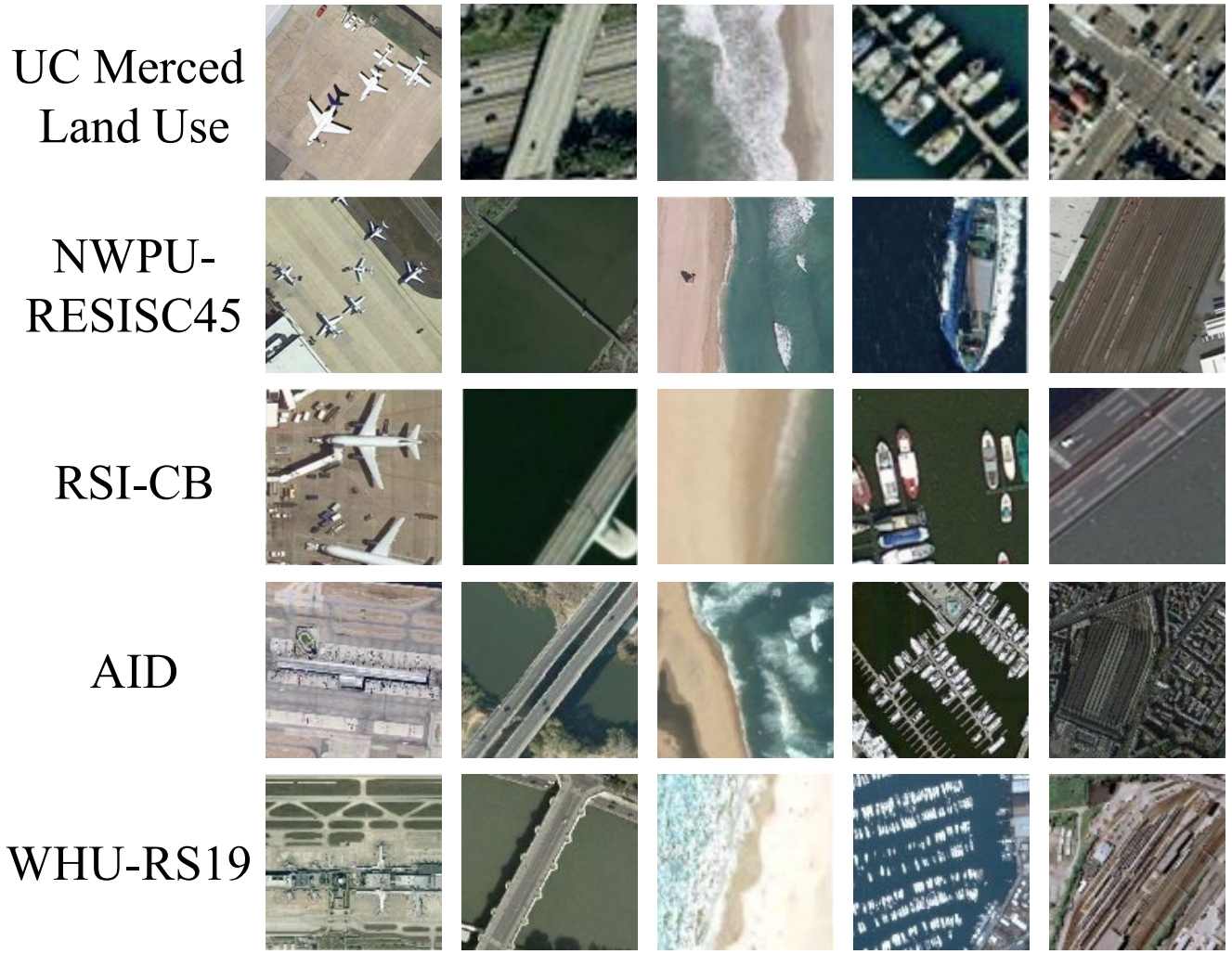
- **Models with Different Settings**
  - —Structure: image-level / video-level
  - —Confident Learning: dropping-bases / flipping-based
  - —Hyper-parameters: learning rates, weight decay
- **Fusion Strategy**
  - —Assigning Weights: {0,1,2}
  - —Normalization
  - —Weighted Average



LADI + Other (O)

北京大学 PEKING UNIVERSITY

- **Collecting Images from Other Public Datasets**
  - — Remote Sensing Datasets
    - UC Merced Land Use
    - NWPU-RESISC45
    - RSI-CB
    - AID
    - WHU-RS19
  - — 16 categories in total

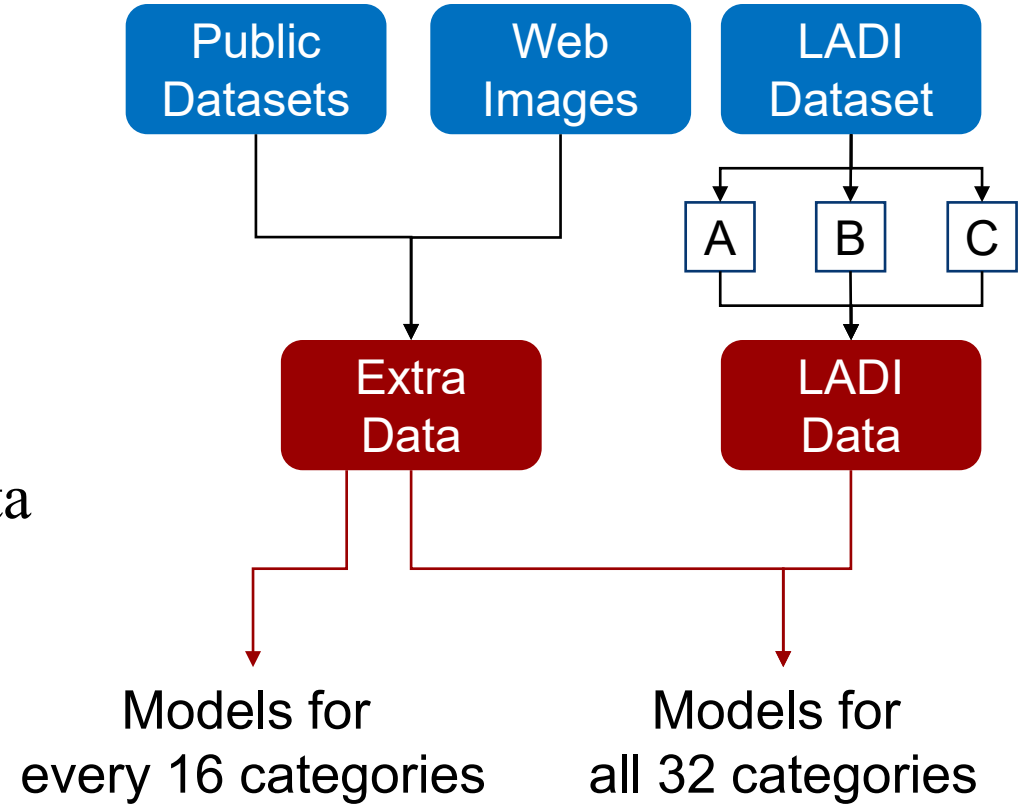


LADI + Other (O)

北京大学 PEKING UNIVERSITY

- **Collecting Images from Web Image Crawling**
  - Categories (Not covered in public datasets)
    - Landslide, washout, rubble, …
- **Extra Data Utilization Strategy**
  - A: all LADI data
  - B: same amount of LADI data
  - C: filtered LADI data

  } + Extra data
- **Fine-tuning O models from L models**
  - Models for every 16 categories
  - Models for all 32 categories



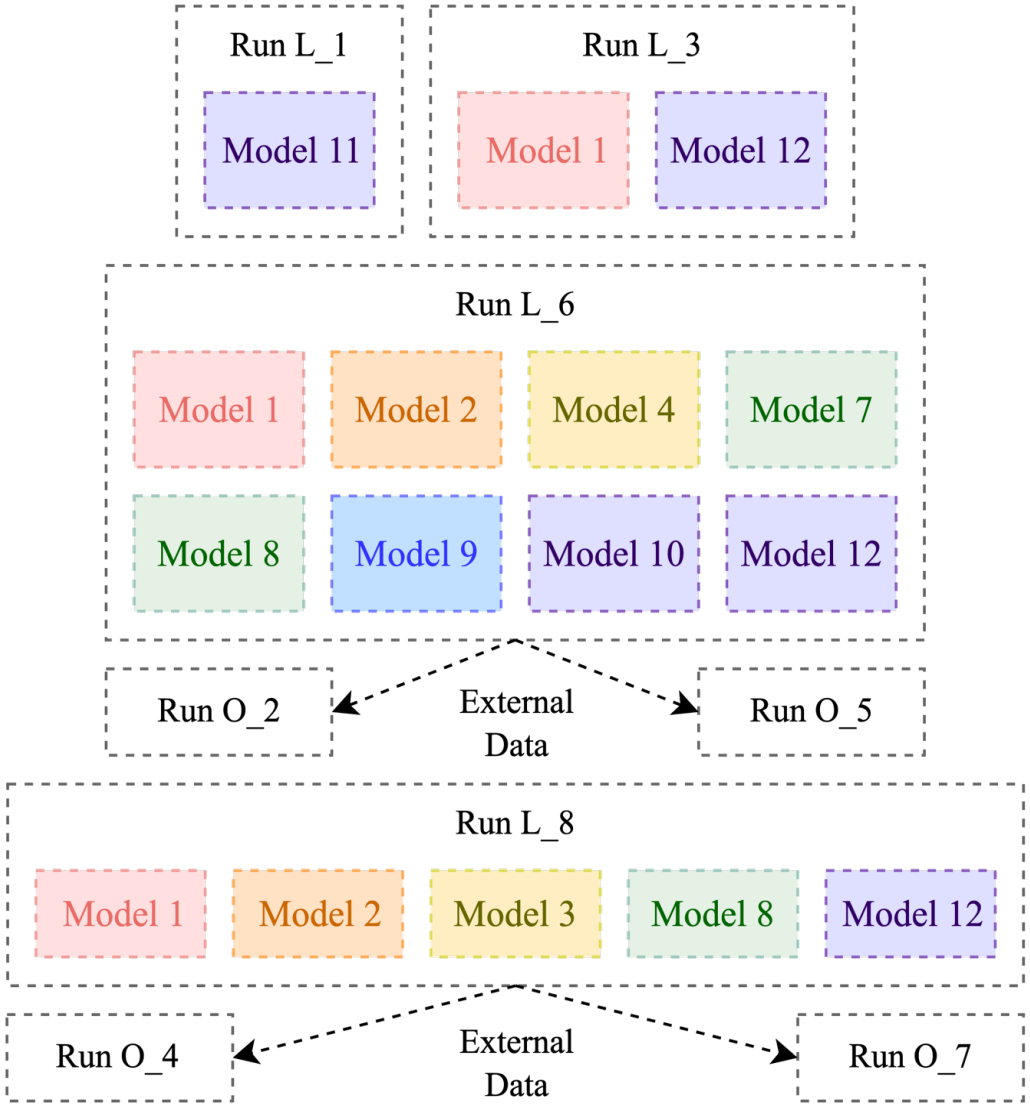
Outline

北京大学 PEKING UNIVERSITY

- Introduction
- Our approach
- Results and conclusions
- Our related works

Results and conclusions

Type	ID	MAP
L	L_PKU_WICT_1	0.4653
	L_PKU_WICT_3	0.4678
	L_PKU_WICT_6	0.4680
	L_PKU_WICT_8	0.4227
O	O_PKU_WICT_2	0.4995
	O_PKU_WICT_4	0.4819
	O_PKU_WICT_5	0.4287
	O_PKU_WICT_7	0.5006



Results and conclusions

北京大学 PEKING UNIVERSITY

- **Conclusions**
  - Data cleaning is helpful to improve the accuracy of retrieval
  - Combination of image and video feature extraction is a key factor for the DSDI task
  - More attention to the fine-grained classification and the combination of more well-designed loss functions may be helpful

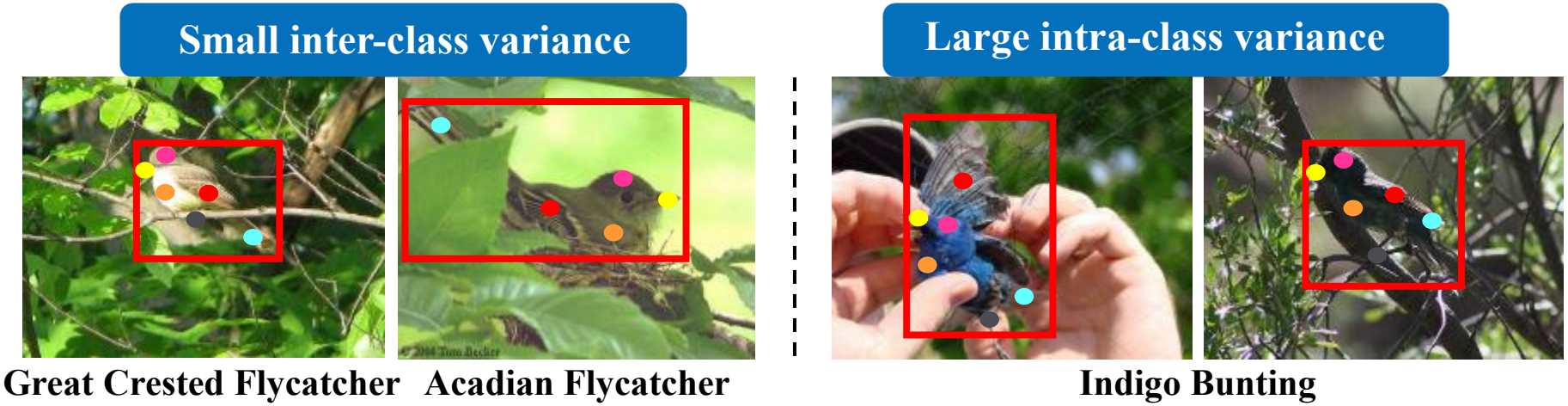
Outline

北京大学 PEKING UNIVERSITY

- Introduction
- Our approach
- Results and conclusions
- Our related works

Fine-grained Image Classification

- Fine-grained Image Classification: Recognize **fine-grained** categories of **coarse-grained** categories (e.g., recognize birds as Great Crested Flycatcher or Acadian Flycatcher)
- One of the most challenging tasks: Birds are easily disturbed by **deformation, occlusion, background** and other complex factors
  - –**Small inter-class variance, and Large intra-class variance**
  - –Dataset contains **200** categories, each with **less than 30** training images
  - –Can be extended for other complex objects like airplane



Fine-grained Image Classification

北京大学 PEKING UNIVERSITY

- Existing methods neglect discriminative regions' interdependencies and composed holistic object structure.
- We propose the Structure Information Modeling Transformer (SIM-Trans) to incorporate object structure information into transformer for enhancing discriminative representation learning to contain both the appearance information and structure information *(SIM-Trans: Structure Information Modeling Transformer for Fine-grained Visual Categorization, **ACM MM 2022**)*

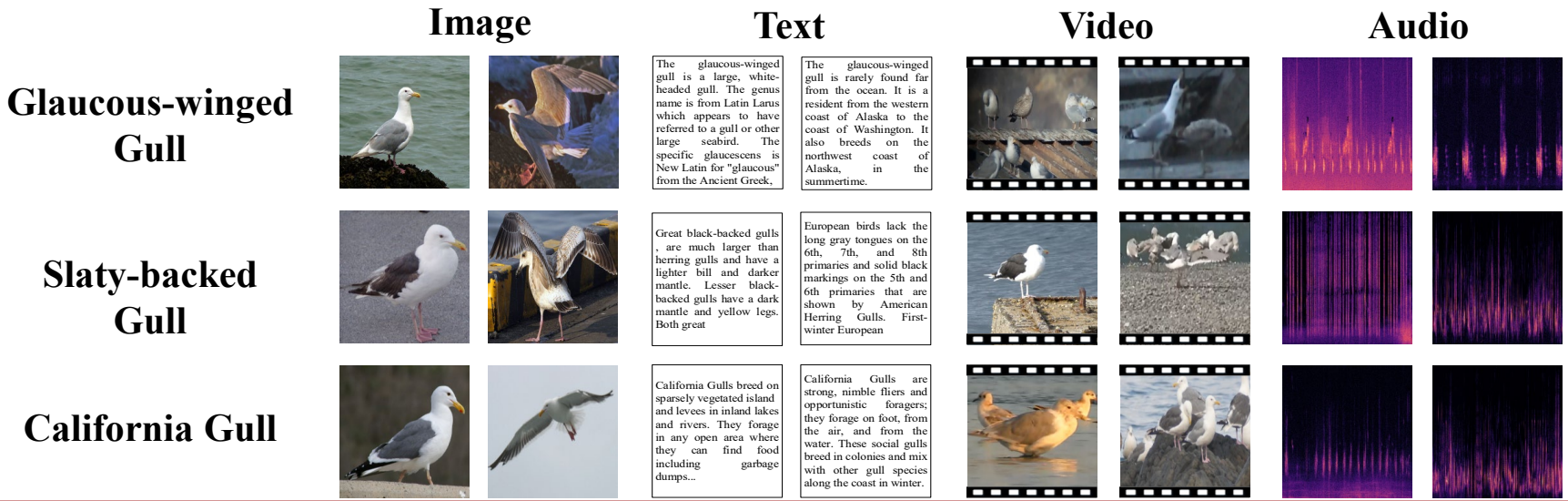
A series of our related researches

- **Fine-grained Visual-textual Representation Learning**[*TCSVT* 2020]
- **Multi-scale and multi-granularity deep reinforcement learning**[*IJCV* 2019]
- **Fine-grained cross-media retrieval** [*ACM MM* 2019]
- **Data augmentation based on selection and generation**[*ACM MM* 2018]
- **Fine-grained visual-textual representation learning**[*CVPR* 2017]
- **Saliency-guided fine-grained discriminative localization** [*ACM MM* 2017]

Our datasets

北京大学 PEKING UNIVERSITY

- **A new dataset and benchmark (PKU FG-XMedia) for fine-grained cross-media retrieval is constructed**
  - The first dataset and benchmark with **4 media types (image, text, video and audio)** for fine-grained cross-media retrieval
  - Consists of **200** fine-grained subcategories of the "Bird"



Download URL：https://github.com/PKU-ICST-MIPL/FGCrossNet_ACMMM2019

Xiangteng He, Yuxin Peng and Liu Xie, "A New Benchmark and Approach for Fine-grained Cross-media Retrieval", ***ACM MM***, 2019.

Contacts

北京大学 PEKING UNIVERSITY

Contact

- Email: pengyuxin@pku.edu.cn
- Lab Website:
  - http://www.wict.pku.edu.cn/mipl
- Github:
  - https://github.com/pku-icst-mipl

多媒体信息处理研究室

Multimedia Information Processing Lab (MIPL)

MIPL GitHub

MIPL WeChat Official Account