PKU_ICST at TRECVID 2017: Instance Search Task

Yuxin Peng, Xin Huang, Jinwei Qi, Junchao Zhang, Junjie Zhao, Mingkuan Yuan, Yunkan Zhuo, Jingze Chi, and Yuxin Yuan

Institute of Computer Science and Technology,
Peking University, Beijing 100871, China
{pengyuxin@pku.edu.cn}

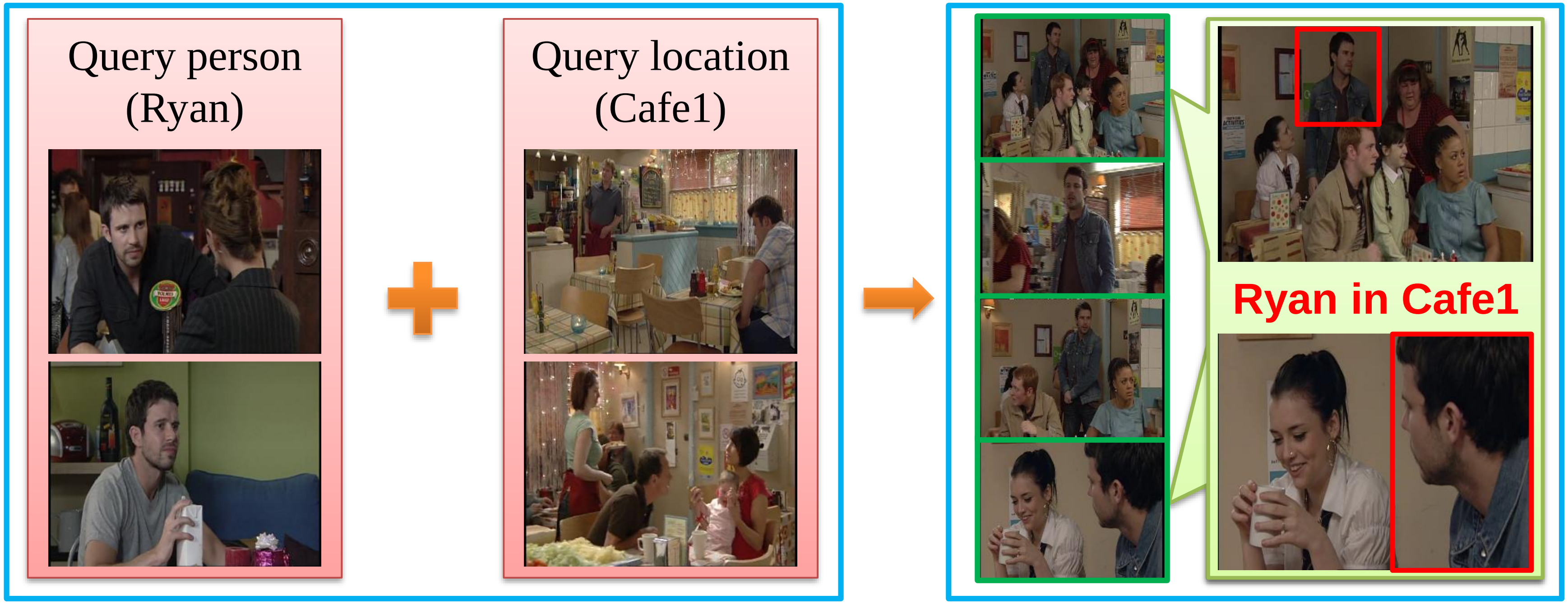
北京大學 PEKING UNIVERSITY

Outline

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

Introduction

- **Instance search (INS) task**
  - Provided: separate person and location examples
  - Topic: combination of a person and a location
  - Target: retrieve specific persons in specific locations



Outline

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

Our approach

- **Overview**

Location-specific search

Similarity computing stage

Result re-ranking stage

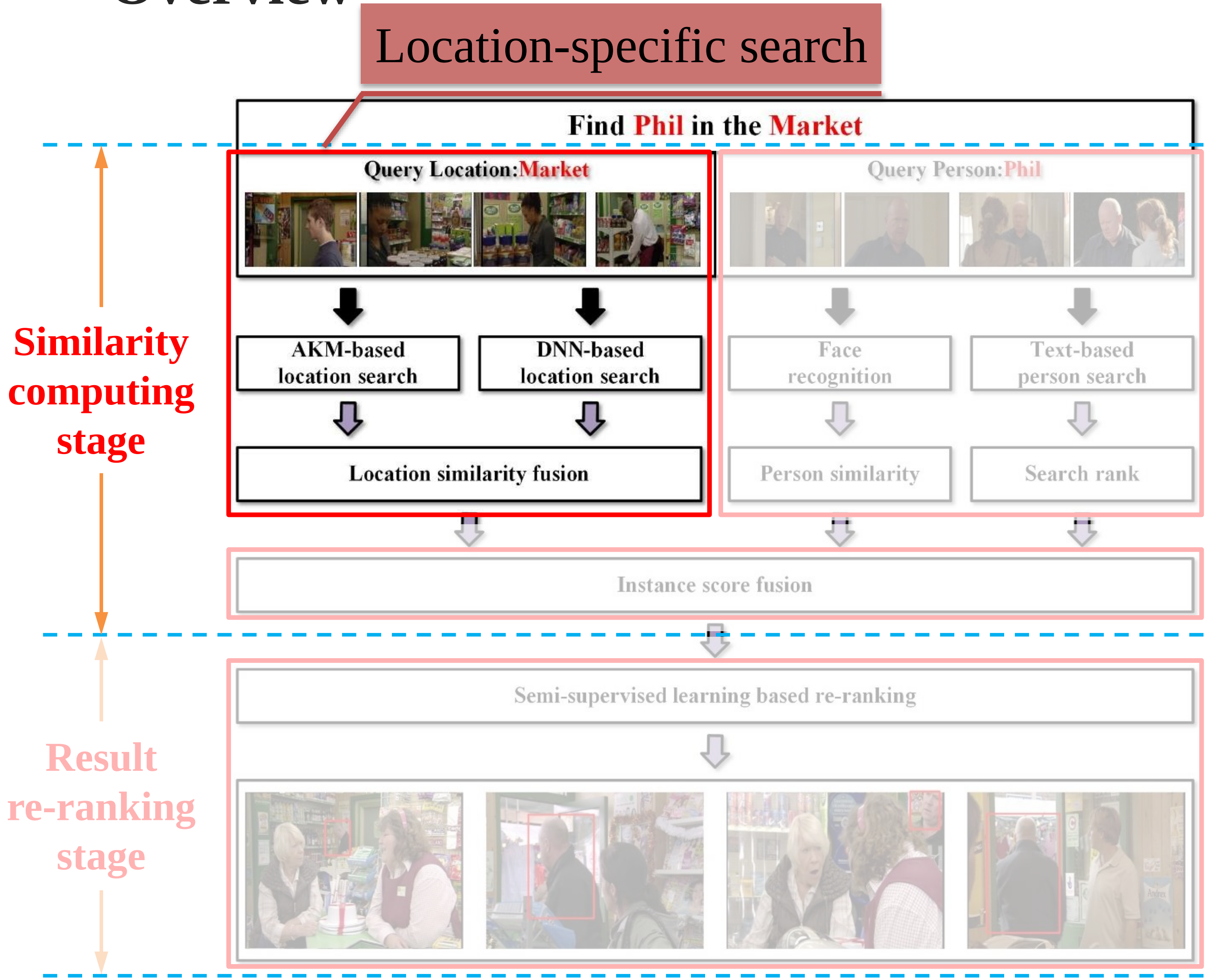
Person-specific search

Fusion

Semi-supervised re-ranking

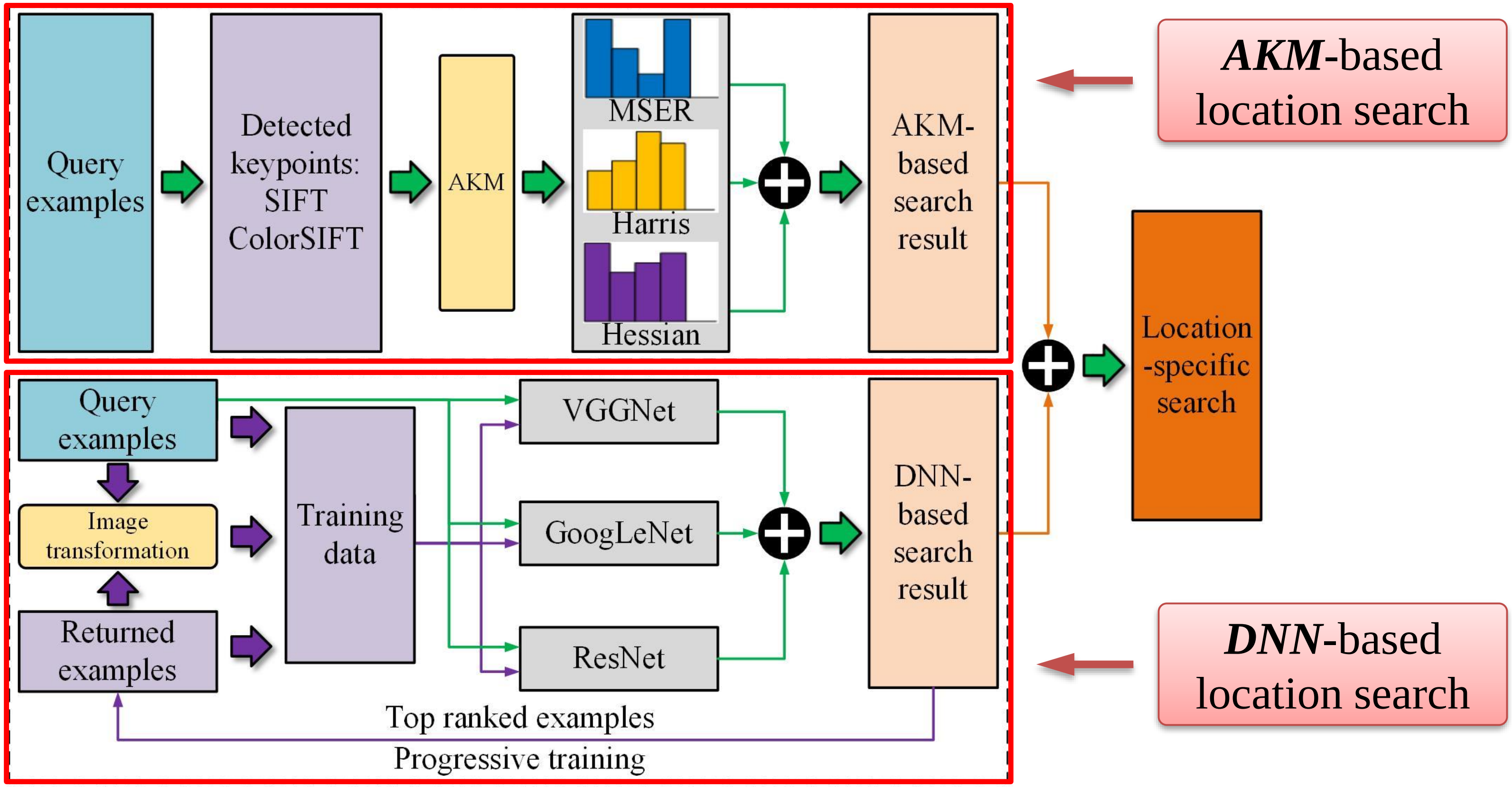
Our approach

- **Overview**



Our approach

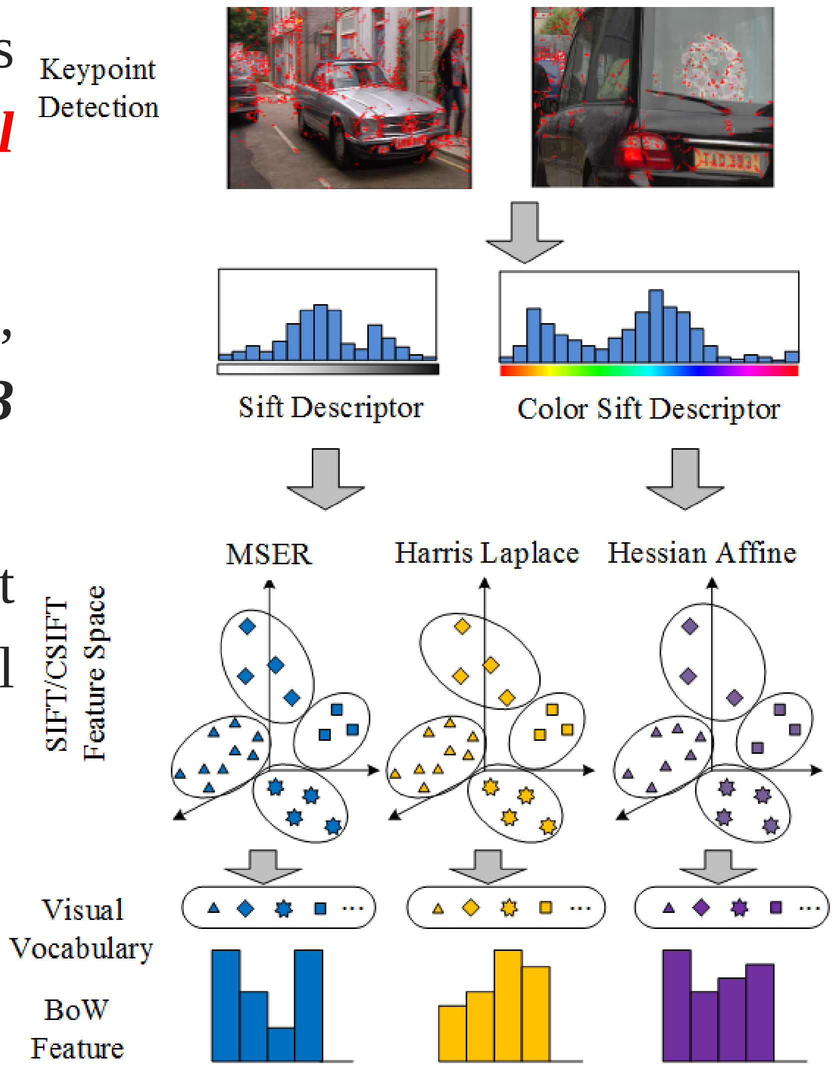
- **Location-specific search**
  - Integrates ***handcrafted*** and ***deep*** features
  - Similarity score: $sim_{location} = w_1 \cdot AKM + w_2 \cdot DNN$



Location-specific search

- **AKM-based location search**
  - Keypoint-based BoW features are applied to capture ***local details***
  - Total 6 kinds of BoW features, which are combinations of ***3 detectors*** and ***2 descriptors***
  - AKM algorithm is used to get ***one-million*** dimensional visual words
- Similarity score:

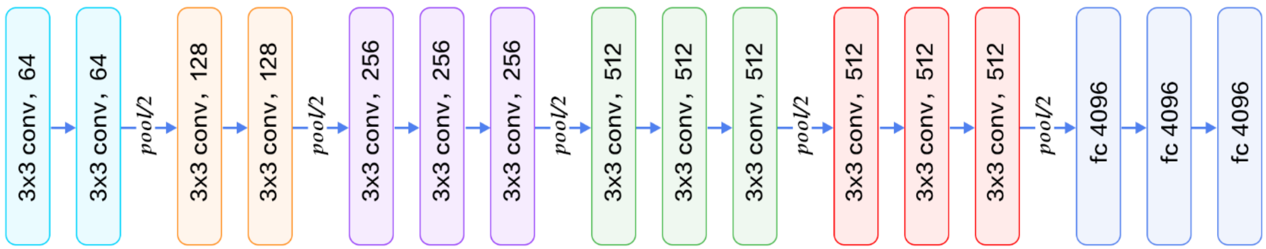
$$AKM = \frac{1}{N}\sum_{k} BOW^{(k)}$$



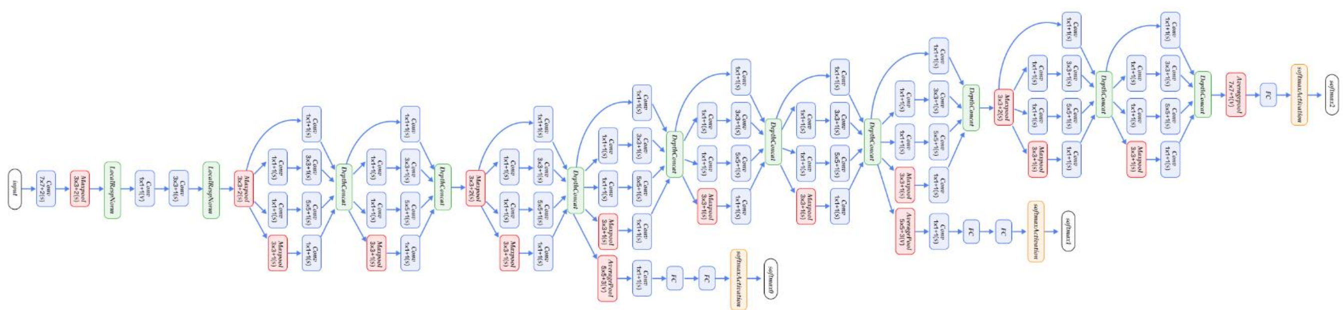
Location-specific search

- **DNN-based location search**
  - DNN features are used to capture ***semantic information***
  - Ensemble of 3 CNN models

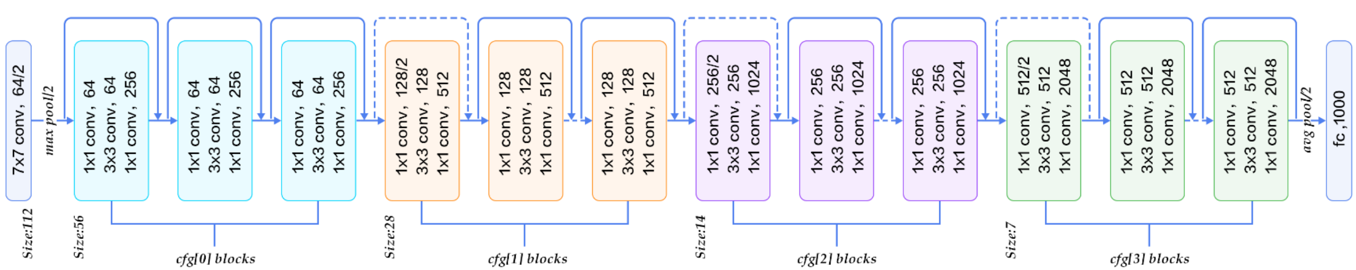
VGGNet



GoogLeNet

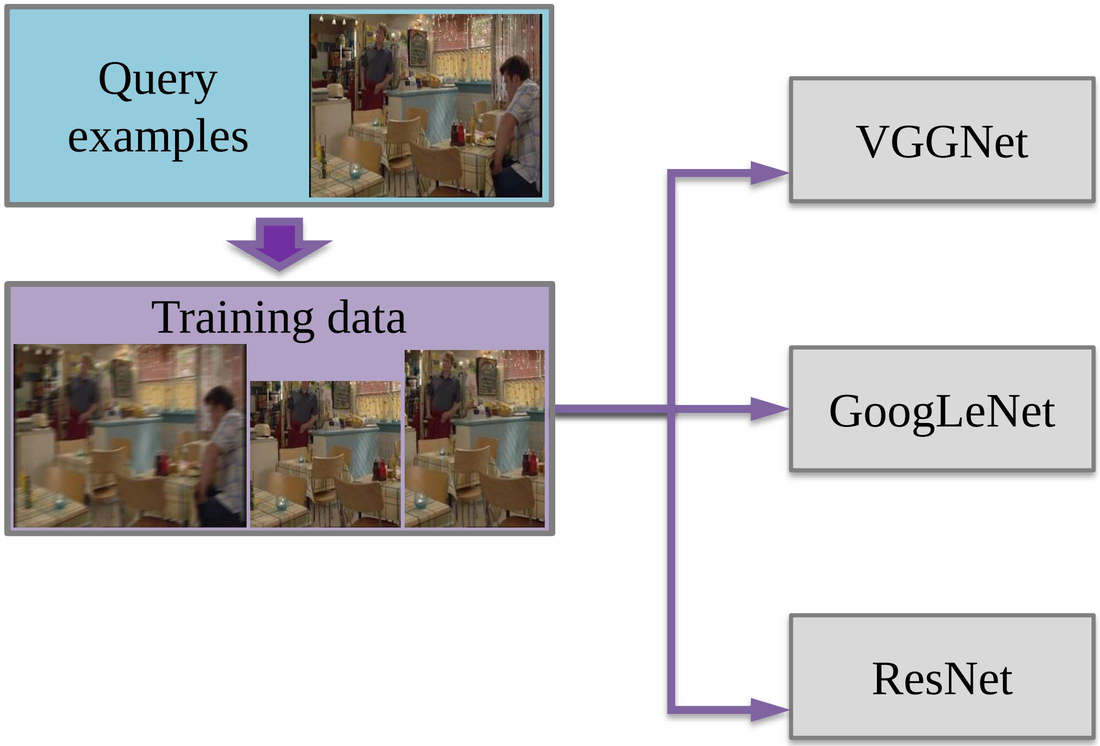


ResNet



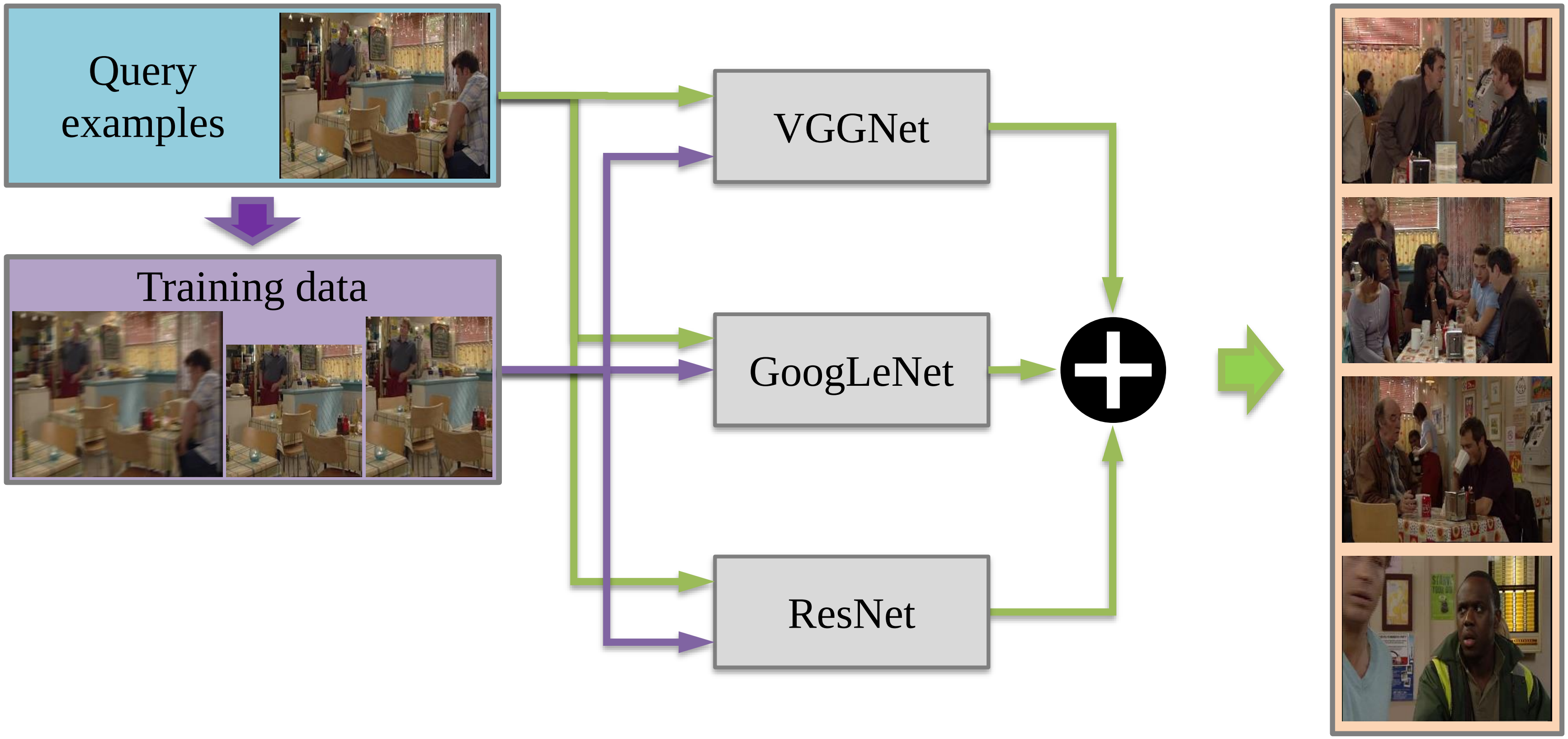
Location-specific search

- **DNN-based location search**
  - All 3 CNNs are trained with ***progressive training*** strategy
- **Progressive training**



Location-specific search

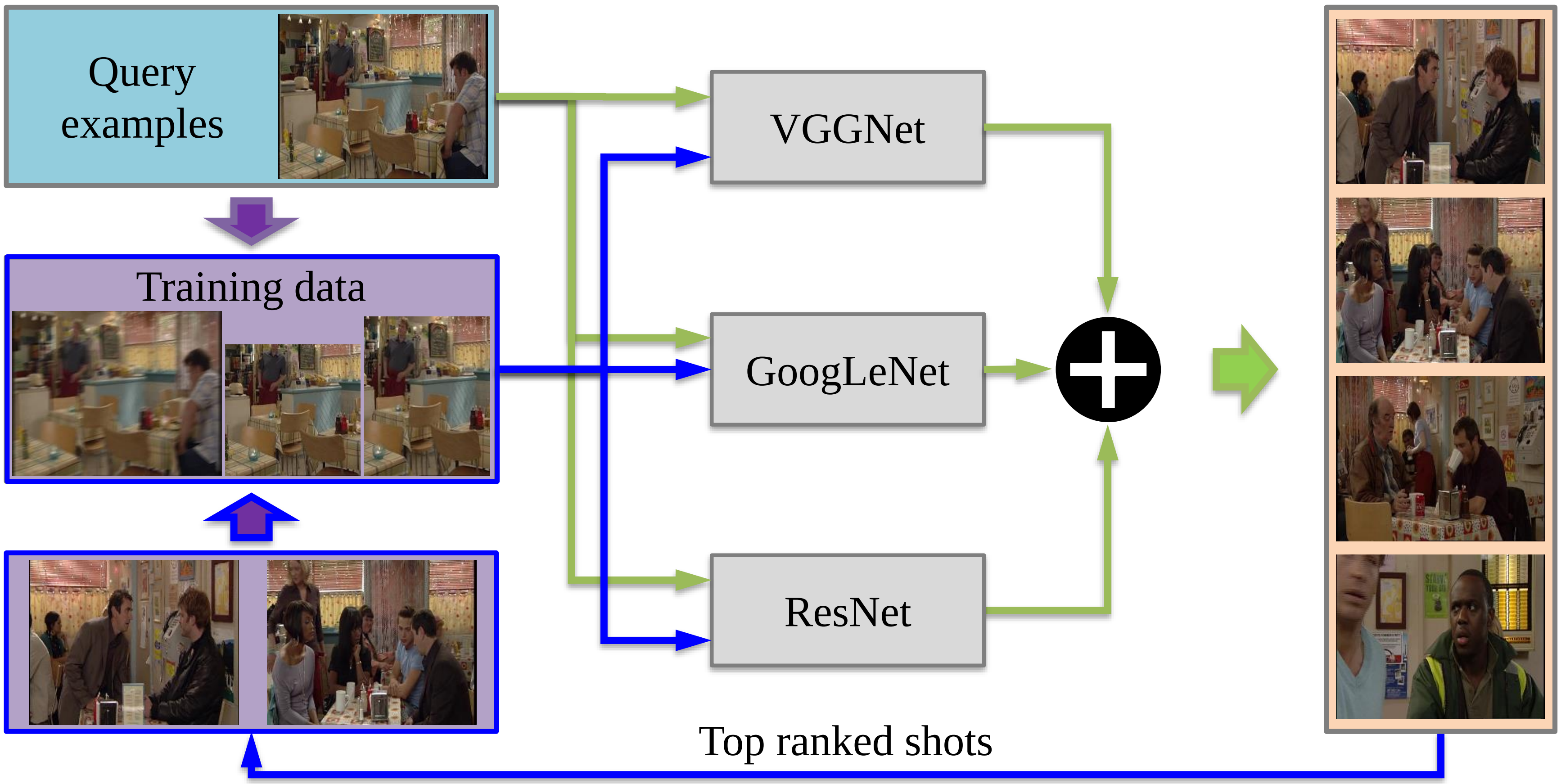
- **DNN-based location search**
  - All 3 CNNs are trained with ***progressive training*** strategy
- **Progressive training**



PEKING UNIVERSITY

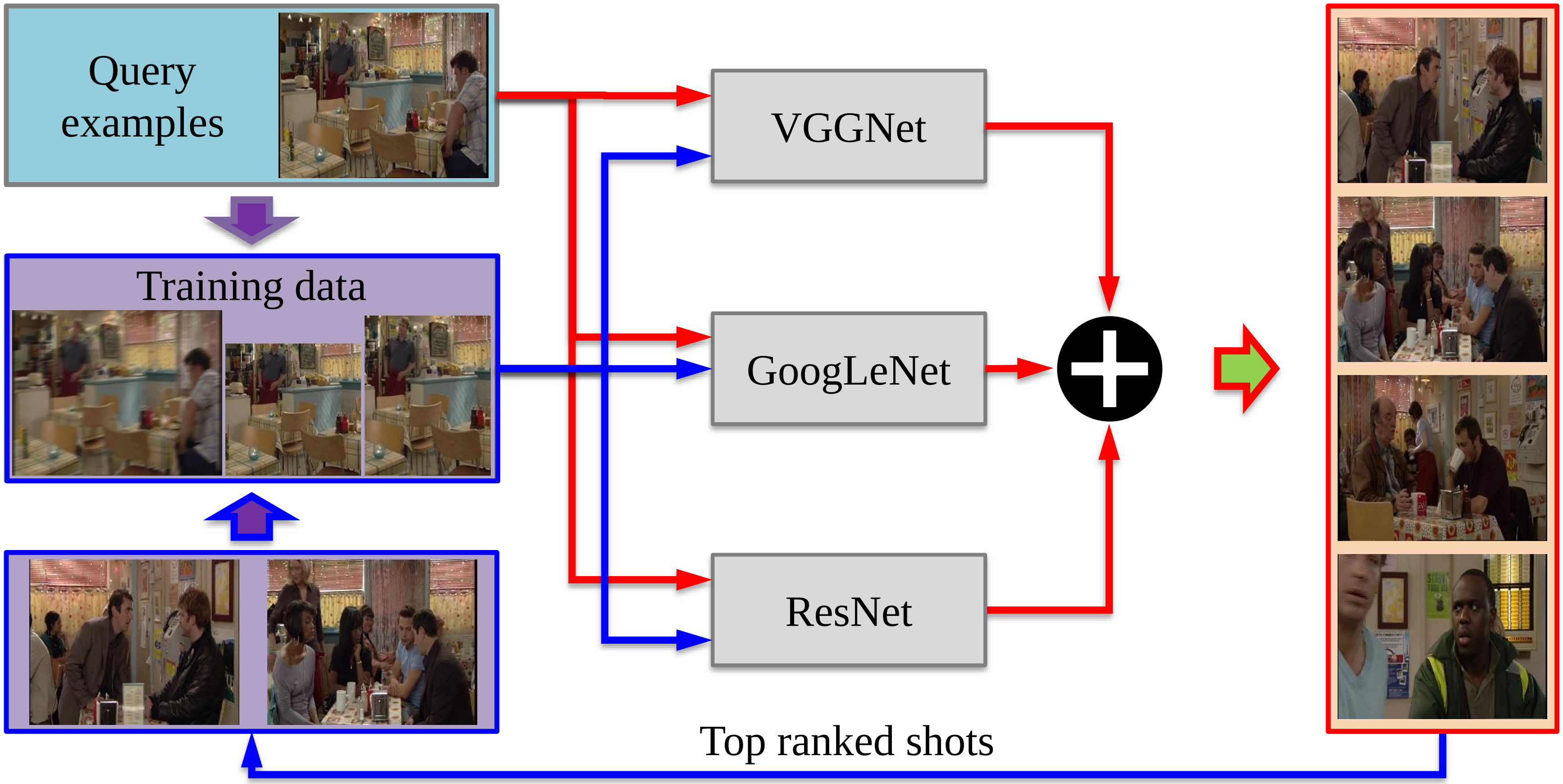
Location-specific search

- **DNN-based location search**
  - All 3 CNNs are trained with ***progressive training*** strategy
- **Progressive training**



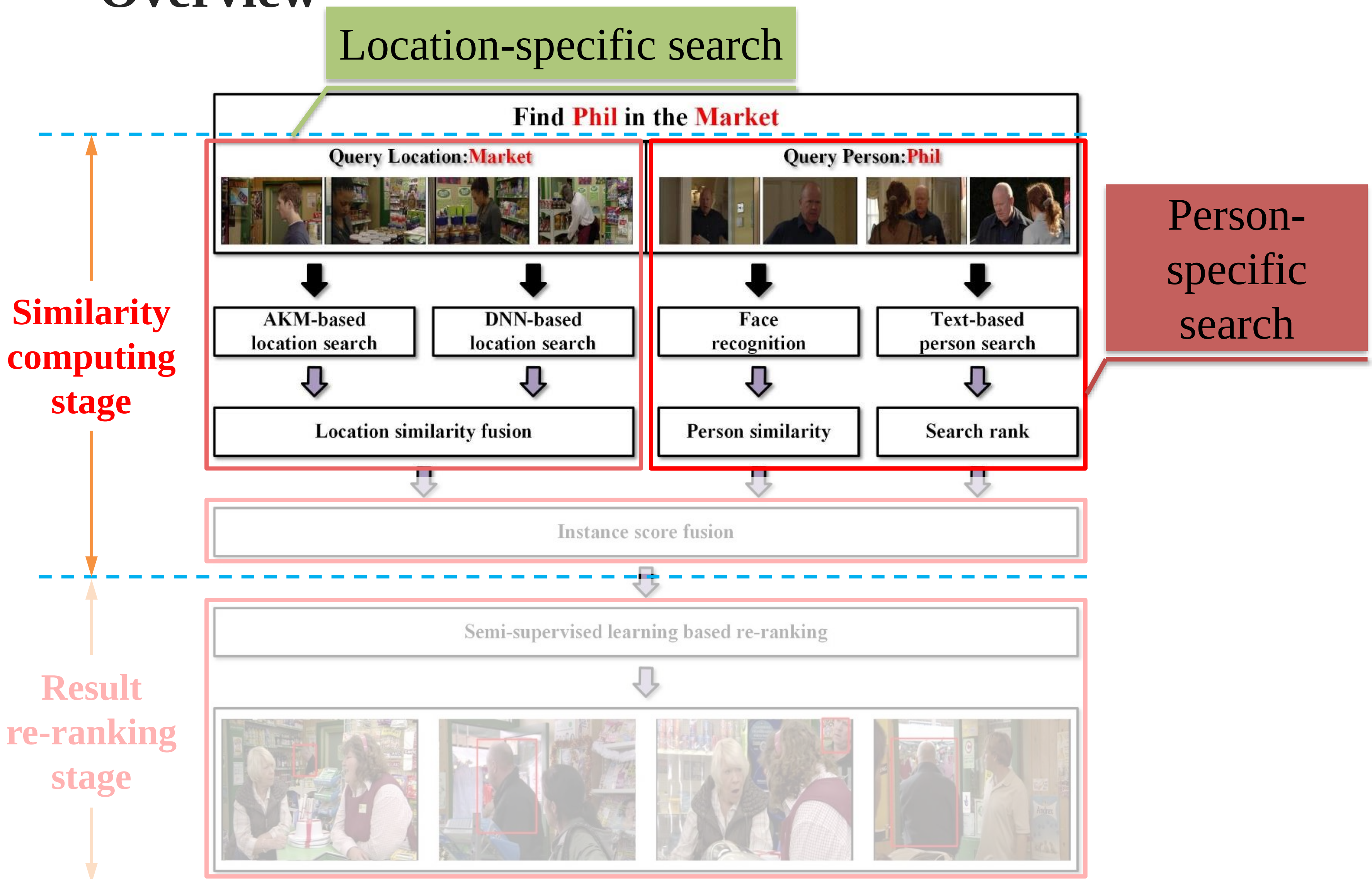
Location-specific search

- **DNN-based location search**
  - All 3 CNNs are trained with ***progressive training*** strategy
- **Progressive training**



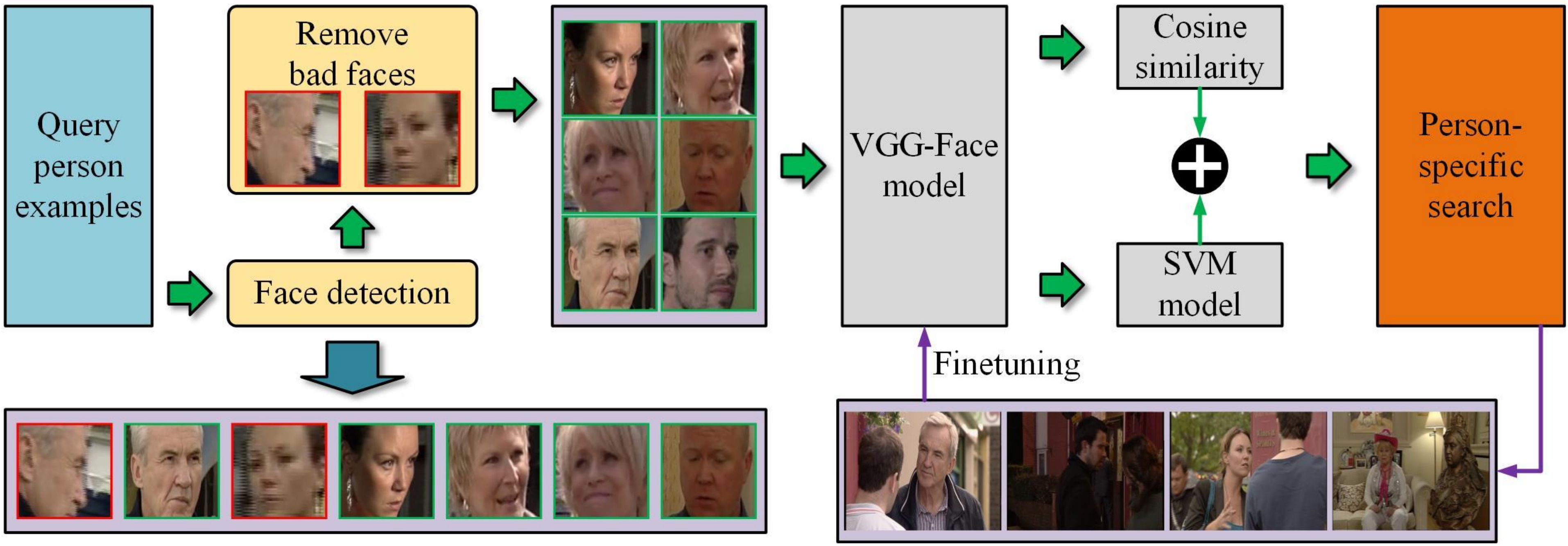
Our approach

- **Overview**



Our approach

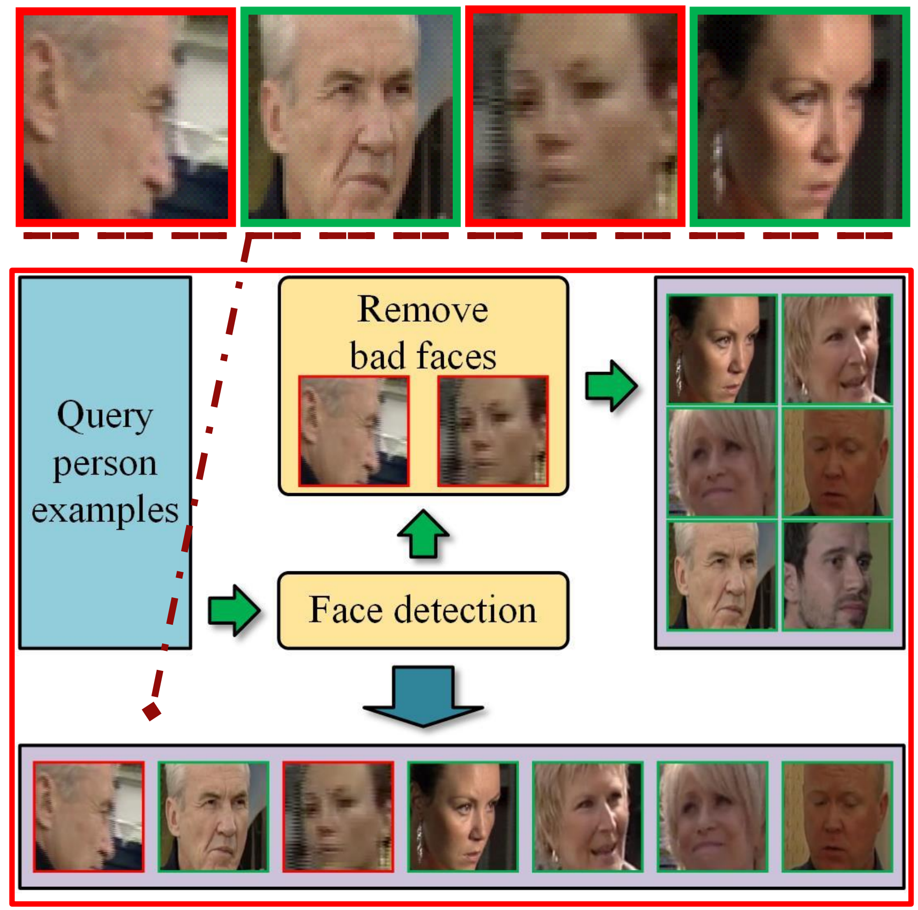
- **Person-specific search**
  - We apply ***face recognition*** technique based on deep model



  - We also conduct ***text-based person search,*** where persons' auxiliary information is minded from the provided video transcripts

Person-specific search

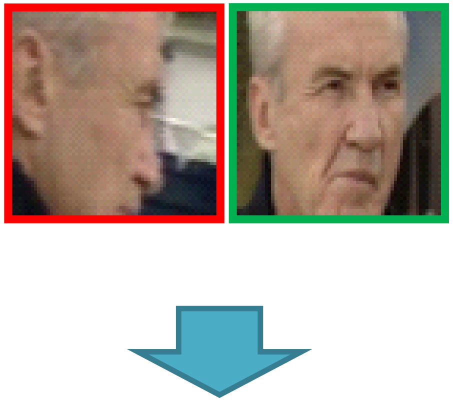
- **Face recognition based person search**
  - Face detection

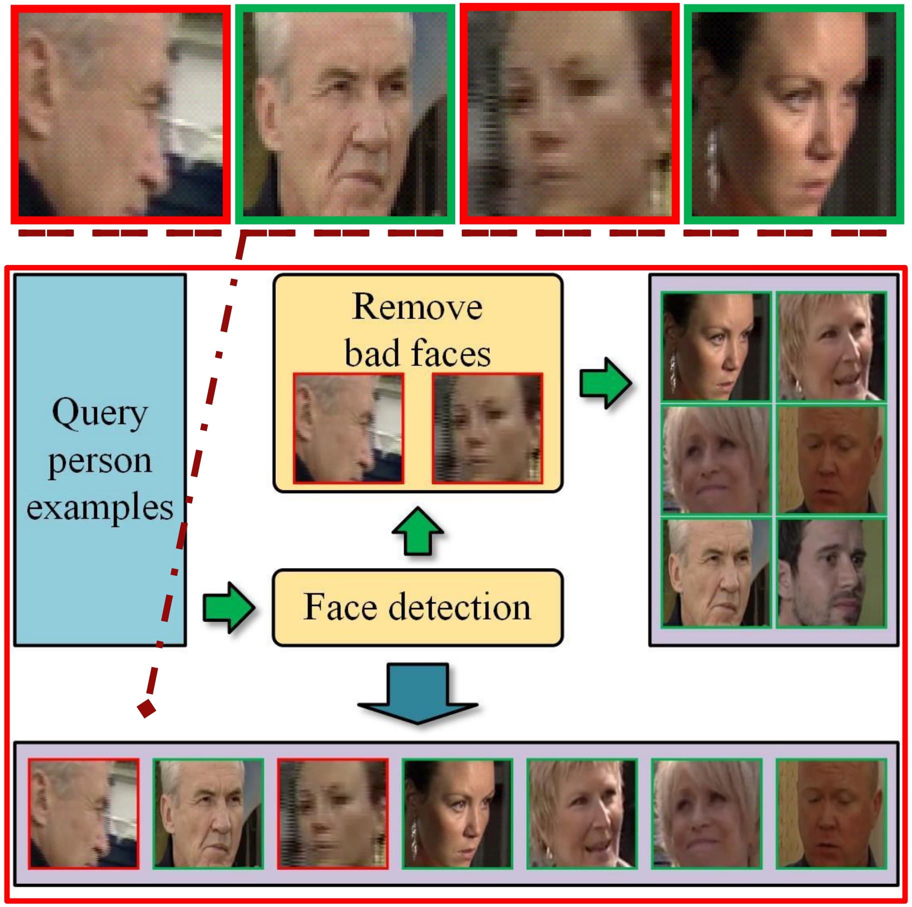


Person-specific search

- **Face recognition based person search**
  - Face detection
  - ***Remove "bad" faces*** automatically: hard to distingush

Before removal of bad faces:

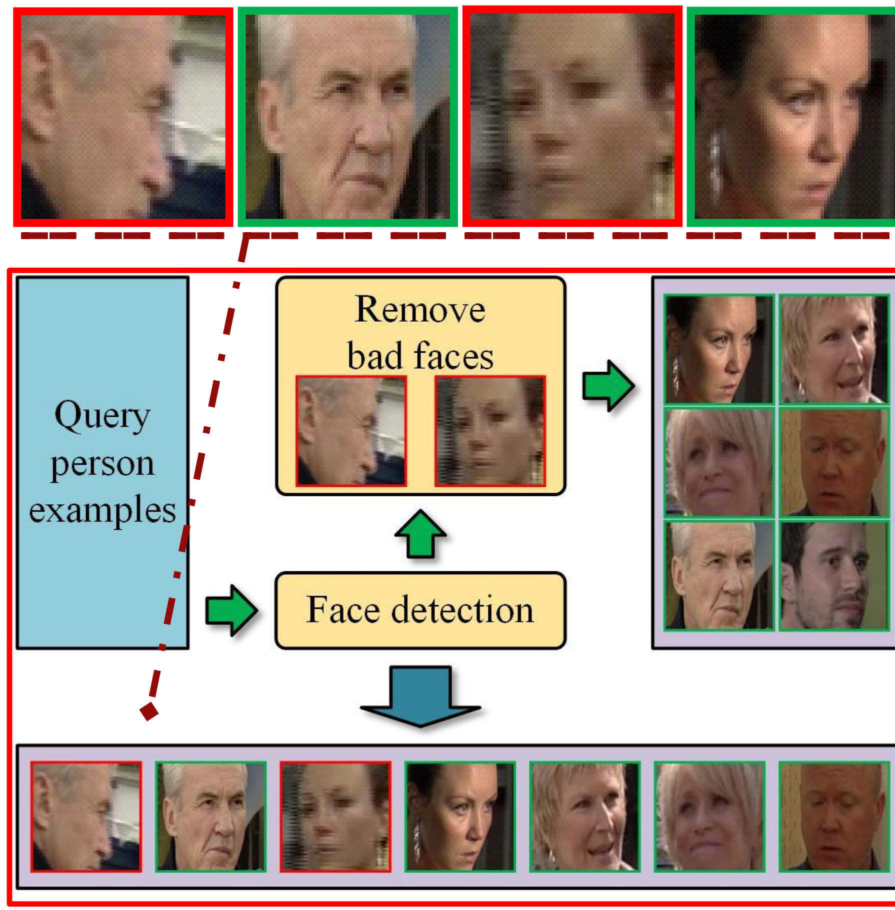




Person-specific search

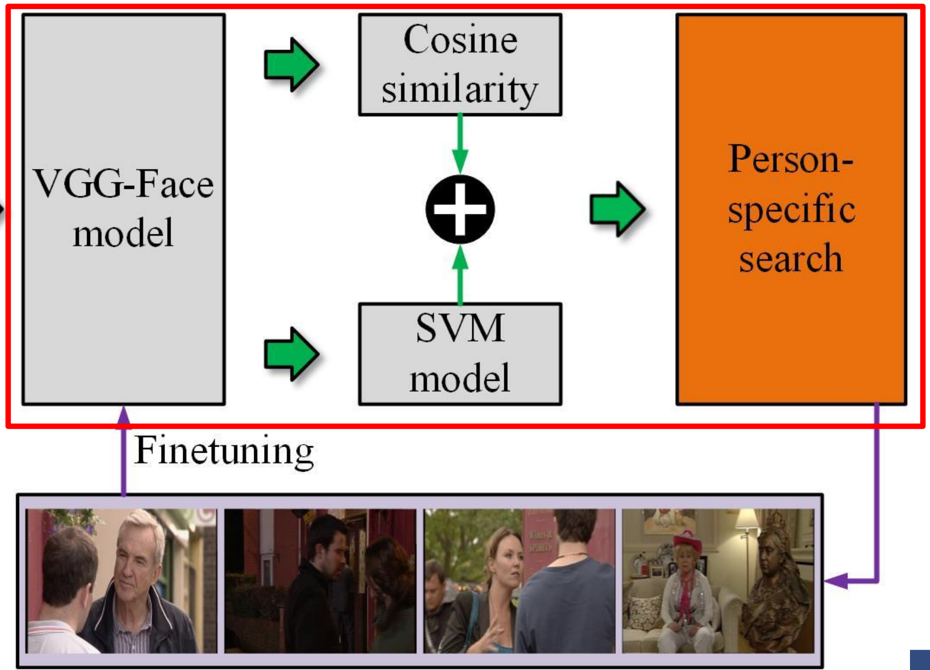
- **Face recognition based person search**
  - Face detection
  - ***Remove "bad" faces*** automatically: hard to distingush

Before removal of bad faces:



Person-specific search

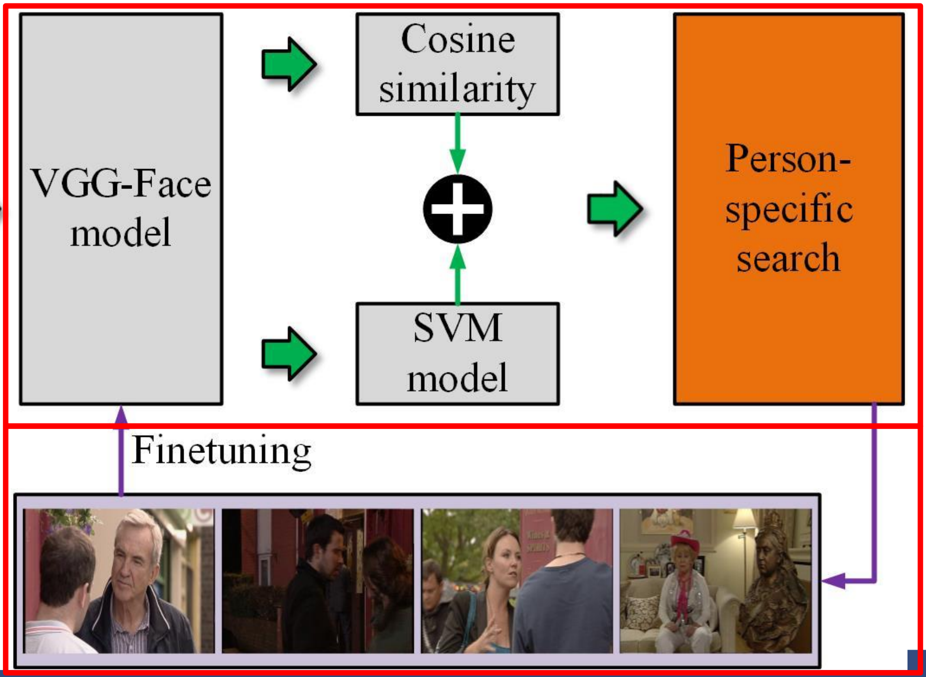
- **Face recognition based person search**
  - We use VGG-Face model to extract face features
  - We integrate ***cosine similarity*** and ***SVM prediction*** scores to get the person similarity scores.



$$sim_{person} = w_1 \cdot COS + w_2 \cdot SVM$$

Person-specific search

- **Face recognition based person search**
  - We use VGG-Face model to extract face features
  - We integrate ***cosine similarity*** and ***SVM prediction*** scores to get the person similarity scores.
  - We adopt similar progressive training strategy to finetune the VGG-Face model

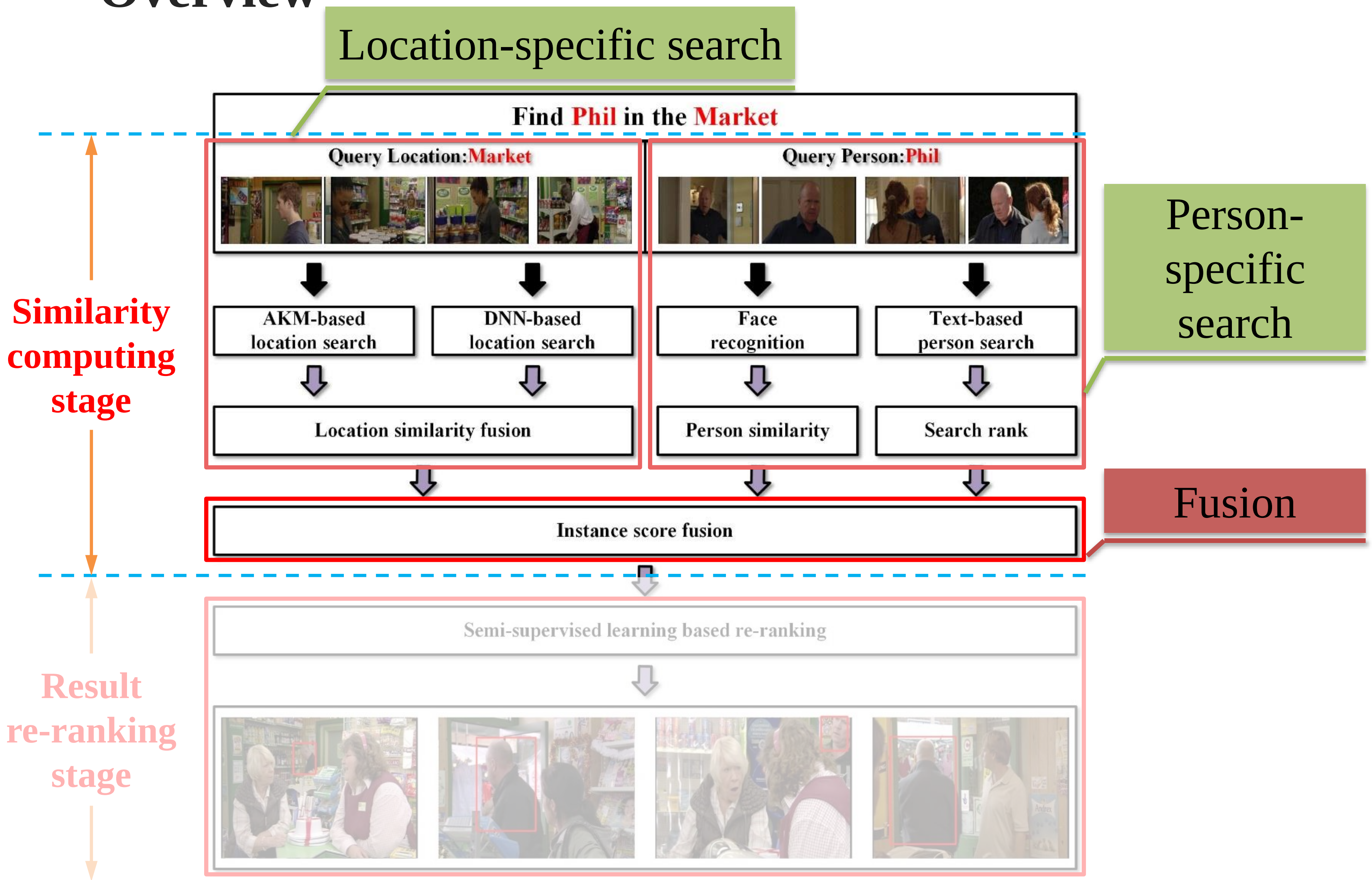


$$sim_{person} = w_1 \cdot COS + w_2 \cdot SVM$$

Progressive training

Our approach

- **Overview**

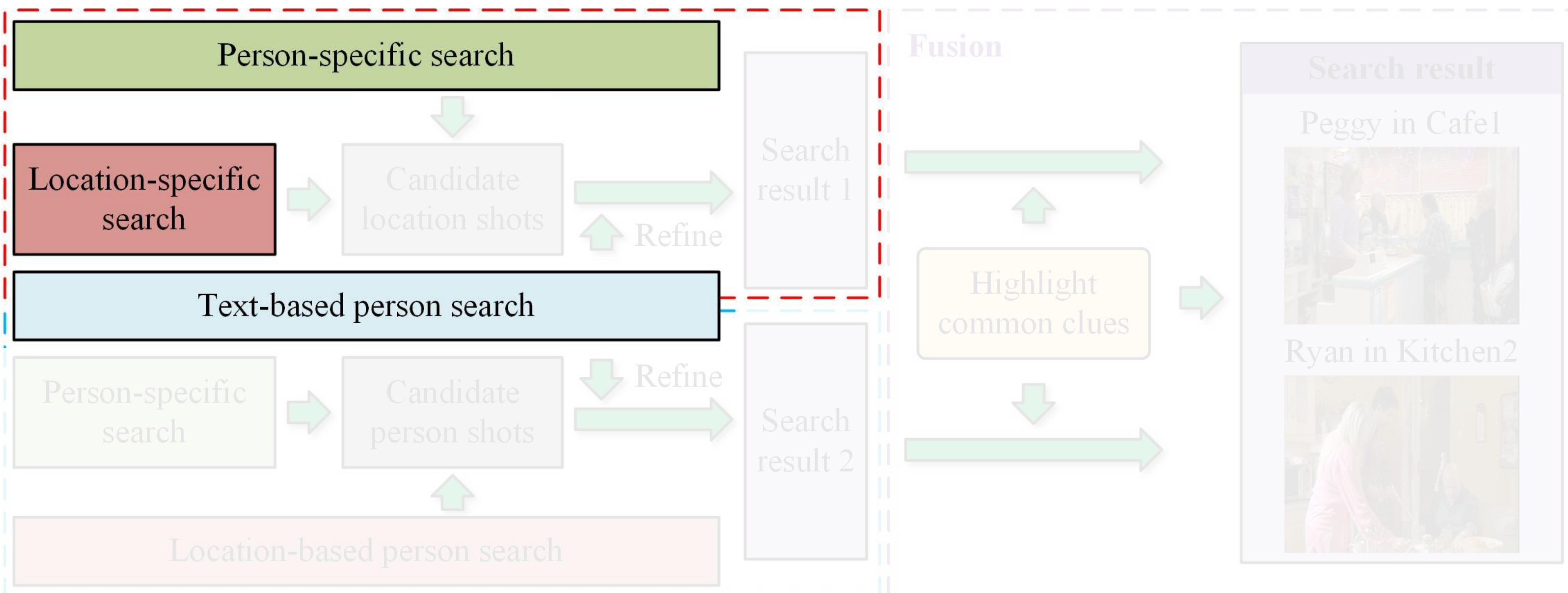


Our approach

- **Instance score fusion**
  - Direction 1, we ***search person in specific location***

$$s_1 = \mu \cdot sim_{person}$$

  - μ is a bonus parameter based on text-based person search

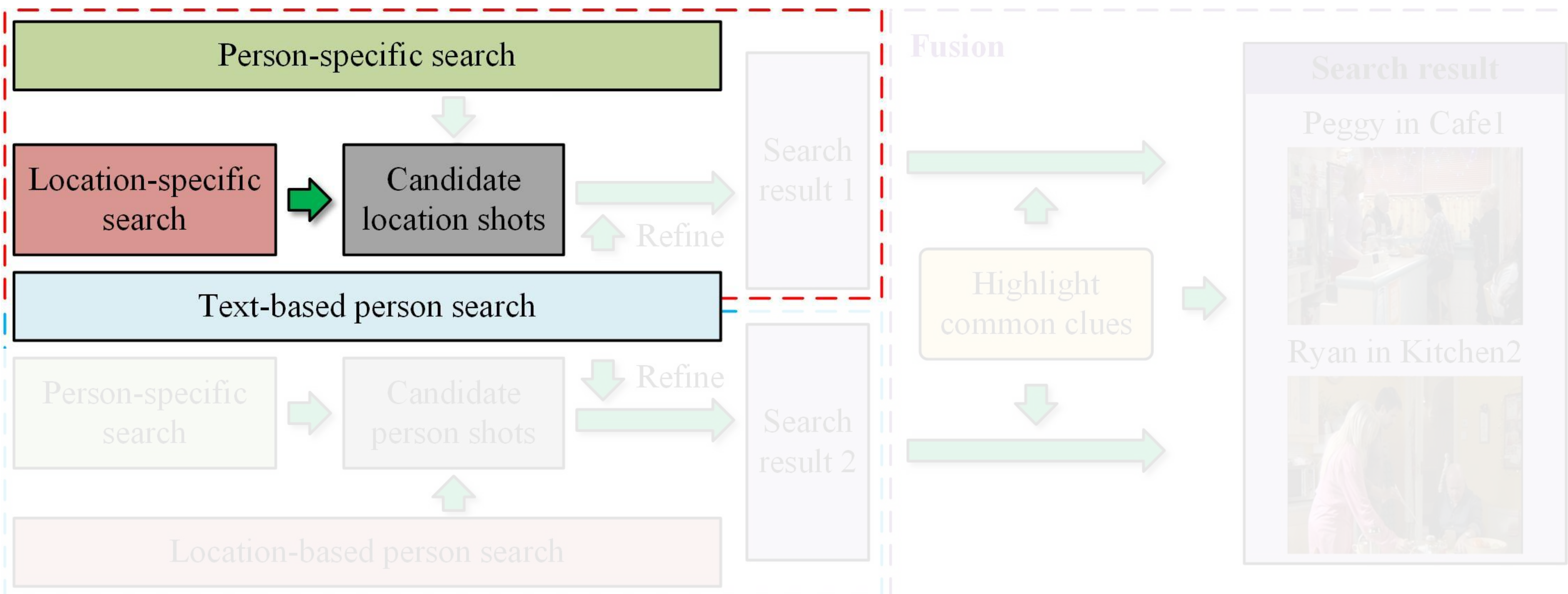


Our approach

- **Instance score fusion**
  - Direction 1, we ***search person in specific location***

$$s_1 = \mu \cdot sim_{person}$$

  - μ is a bonus parameter based on text-based person search

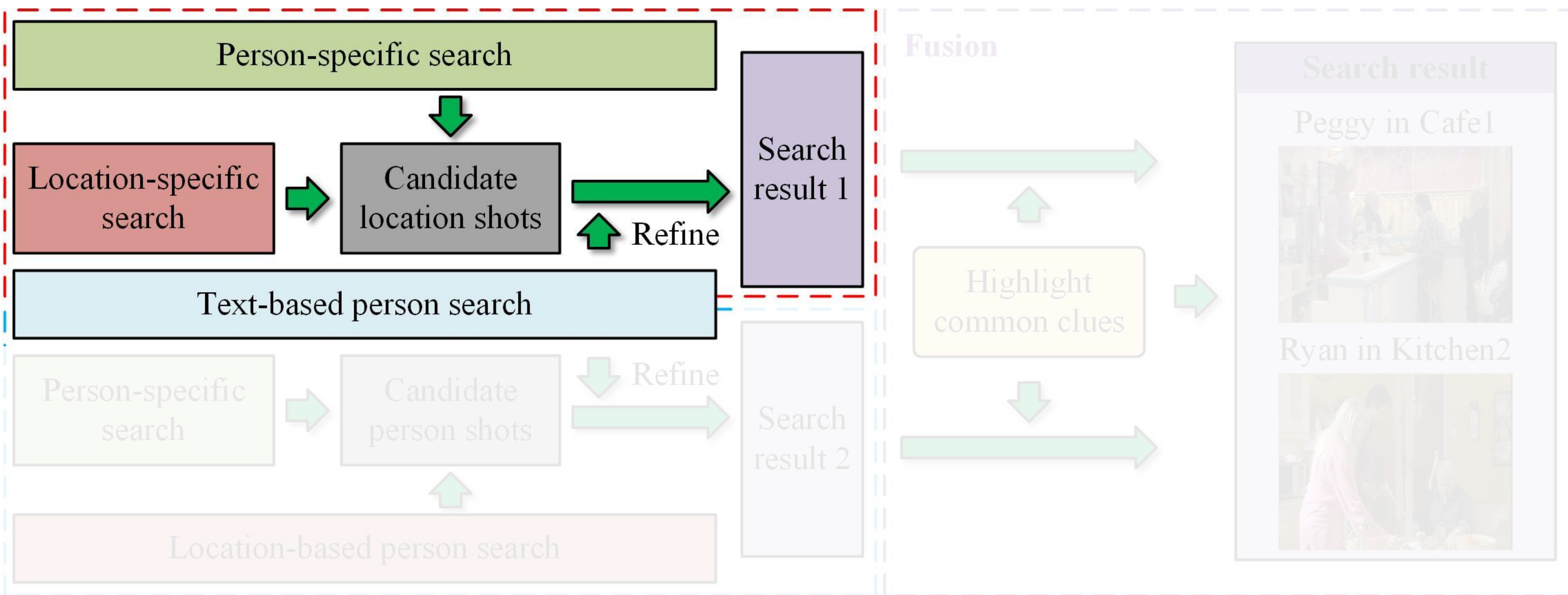


Our approach

- **Instance score fusion**
  - Direction 1, we ***search person in specific location***

$$s_1 = \mu \cdot sim_{person}$$

  - μ is a bonus parameter based on text-based person search

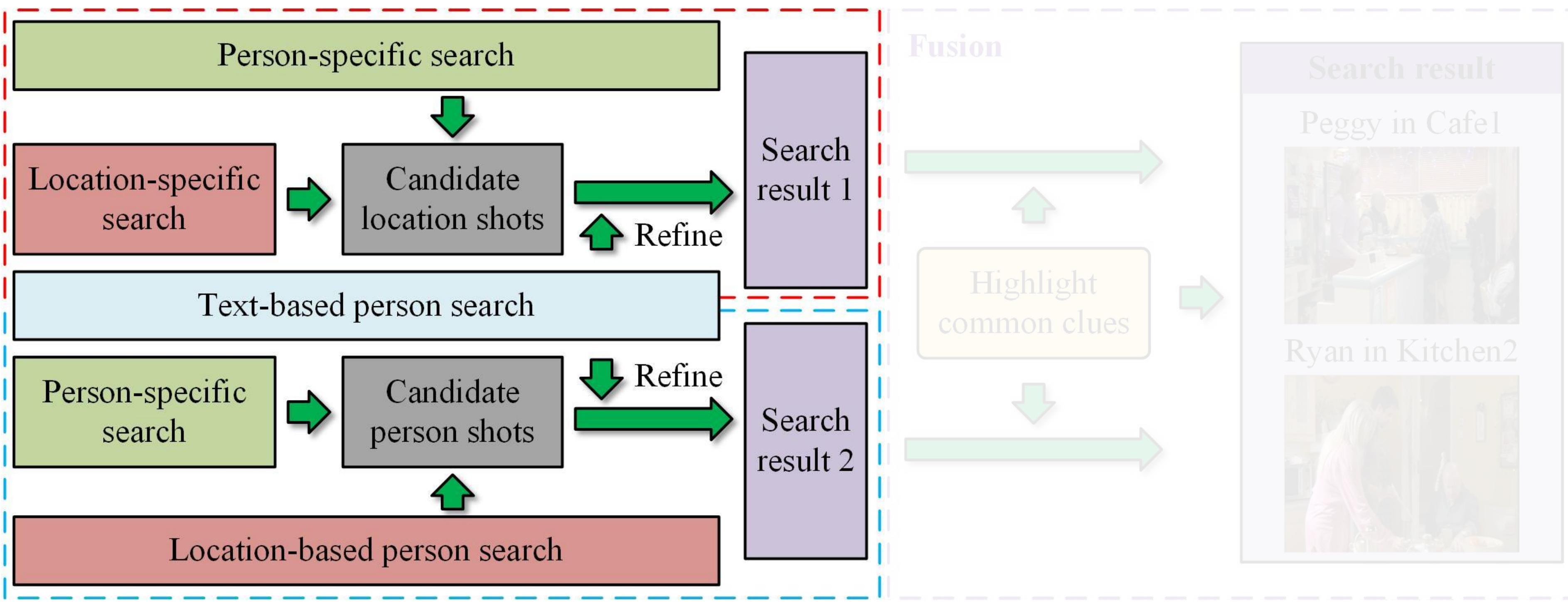


Our approach

- **Instance score fusion**
  - Direction 2, we ***search location containing specific person***

$$s_2 = \mu \cdot sim_{location}$$

  - μ is a bonus parameter based on text-based person search

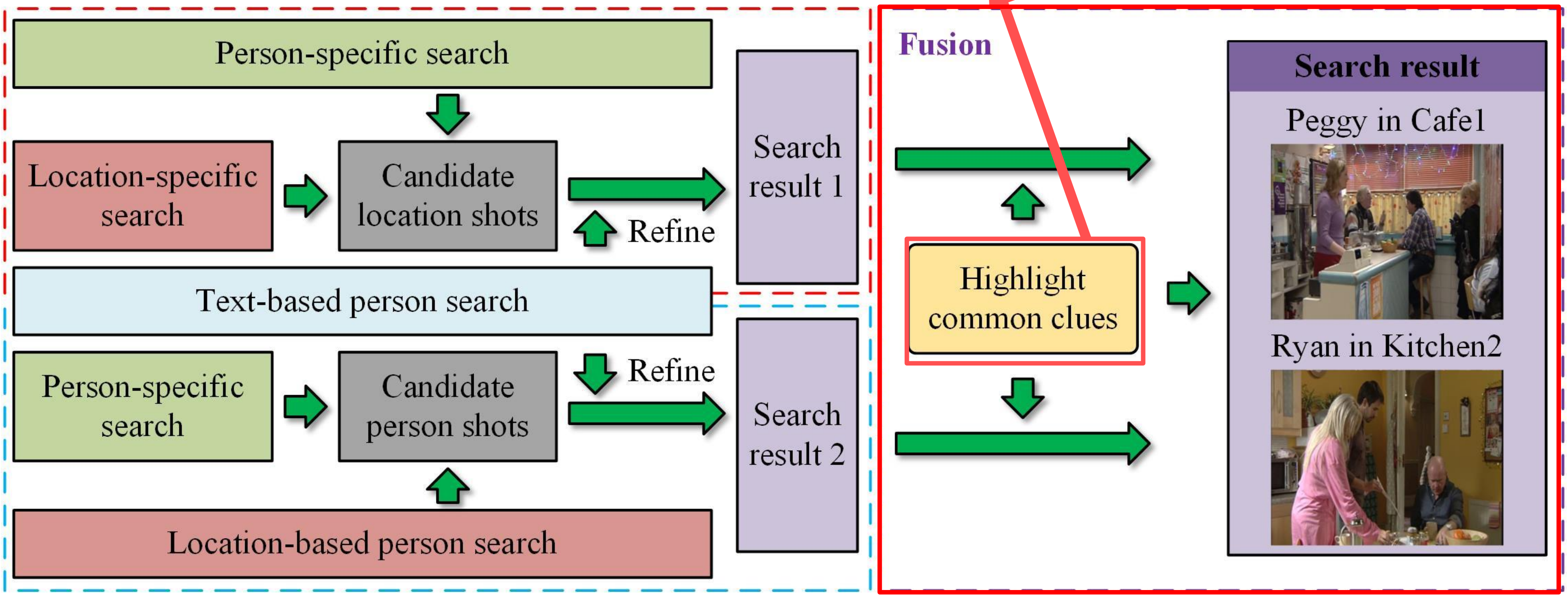


Our approach

- **Instance score fusion**
  - Combine scores of above two directions:

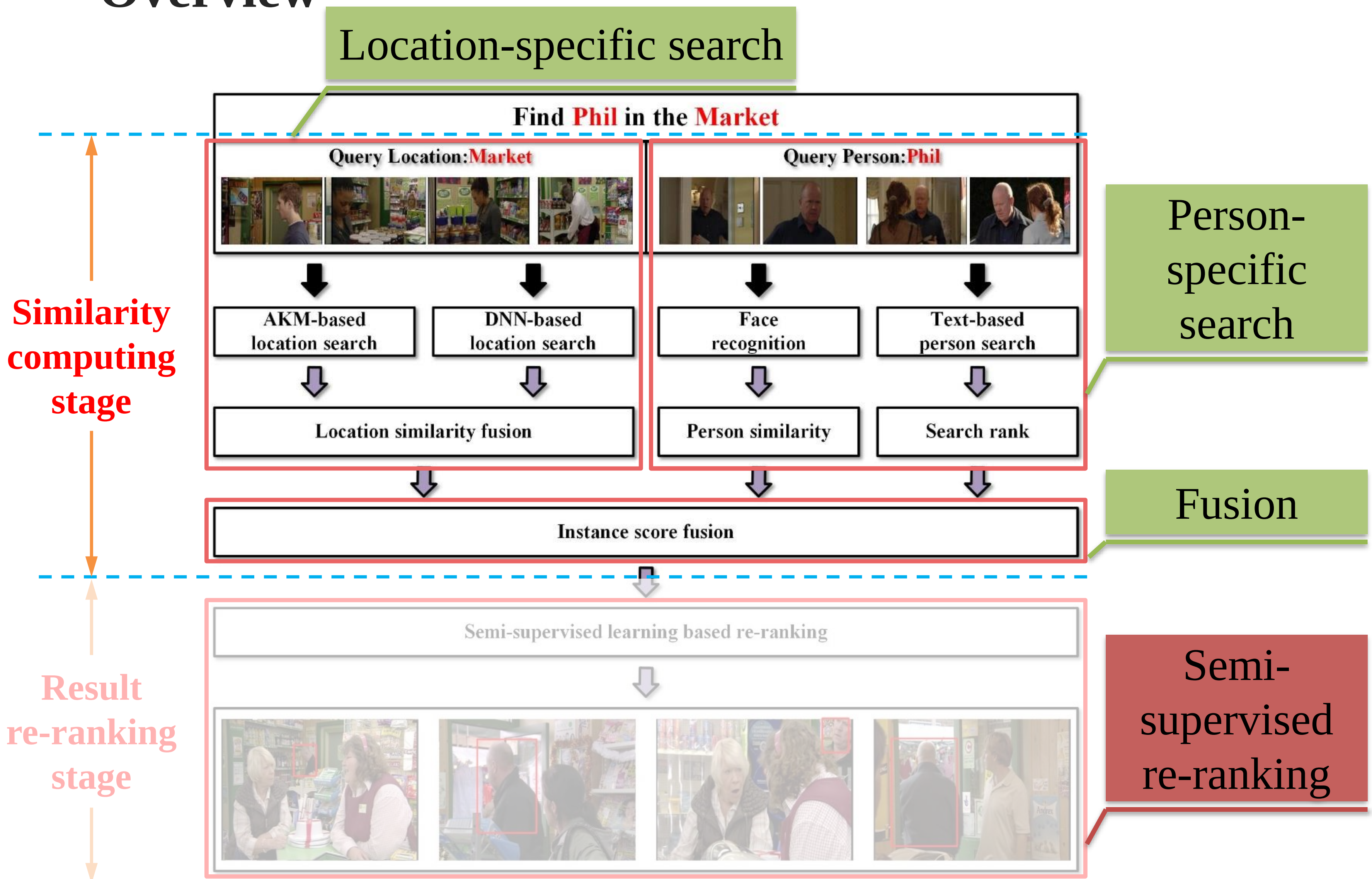
$$s_f = \omega \cdot (\alpha \cdot s_1 + \beta \cdot s_2)$$

  - $\boldsymbol{\omega}$ indicates whether the shot is ***simultaneously*** included in candidate location shots and candidate person shots



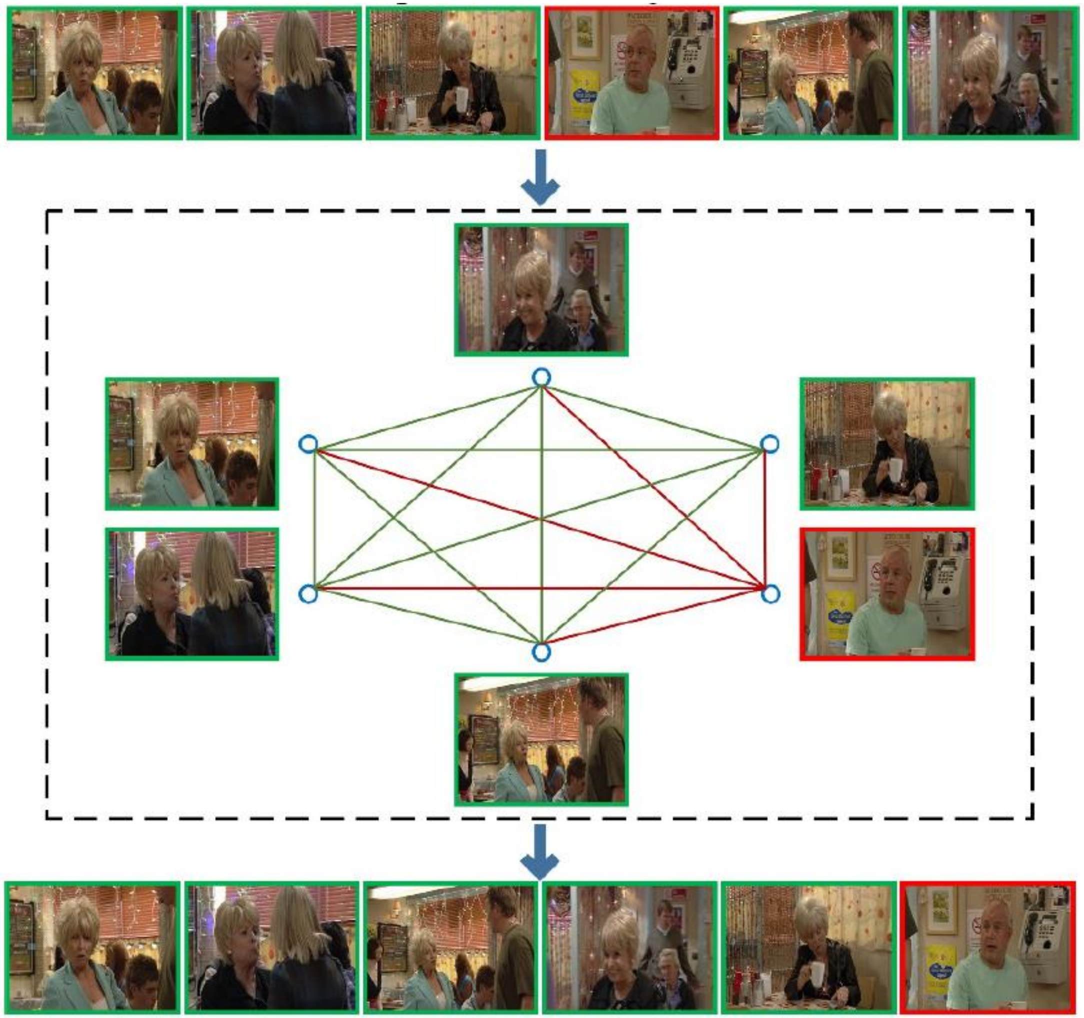
Our approach

- **Overview**



Our approach

- **Re-ranking**
  - Most of the top ranked shots are correct and look similar
  - Noisy shots with ***large dissimilarity*** can be filtered using similarity scores among top ranked shots
  - A ***semi-supervised re-ranking method*** is proposed to refine the result



Re-ranking

- **Semi-supervised re-ranking algorithm**
  - Obtain affinity matrix *W* of top-ranked shots *F*:
  - $$W_{ij} = \begin{cases} \frac{F_i^T \cdot F_j}{|F_i| \cdot |F_j|}, i \neq j \\ 0, \quad i = j \end{cases}, \qquad i,j = \{1,2,\cdots,n\}$$
  - Update *W* according to *k-NN* graph:

$$W_{ij} = \begin{cases} W_{ij}, F_i \in KNN(F_j) \\ 0, \quad otherwise \end{cases}, \qquad i,j = \{1,2,\cdots,n\}$$

  - Construct the matrix:

$$S = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$

  - Re-rank search result:

$$G_{t+1} = \alpha S G_t + (1-\alpha) Y$$

where *Y* is the ranked list obtained by above ***fusion*** step

Outline

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

Results and Conclusions

- **Results**
  - We submitted 7 runs, and ranked ***1st*** in ***both automatic and interactive search***
  - Interactive run is performed based on RUN2 with expanding positive examples as queries

Type	ID	MAP	Brief description
Automatic	RUN1_A	0.448	AKM+DNN+Face
	RUN1_E	0.471	AKM+DNN+Face
	RUN2_A	0.531	RUN1+Text
	RUN2_E	0.549	RUN1+Text
	RUN3_A	0.528	RUN2+Re-rank
	RUN3_E	**0.549**	RUN2+Re-rank
Interactive	RUN4	**0.677**	RUN2+Human feedback

PEKING UNIVERSITY

Results and Conclusions

- **Conclusions**
  - Video examples are helpful for accuracy improvement
  - Automatic removal of “bad faces” is important
  - Fusion of location and person similarity is a key factor of the instance search

Type	ID	MAP	Brief description
Automatic	RUN1_A	0.448	AKM+DNN+Face
	RUN1_E	0.471	AKM+DNN+Face
	RUN2_A	0.531	RUN1+Text
	RUN2_E	0.549	RUN1+Text
	RUN3_A	0.528	RUN2+Re-rank
	RUN3_E	**0.549**	RUN2+Re-rank
Interactive	RUN4	**0.677**	RUN2+Human feedback

北京大學 PEKING UNIVERSITY

Outline

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

1. Video concept recognition (1/2)

- **Video concept recognition**
  - Learn semantics from video content and classify videos into pre-defined categories automatically.
  - For examples: human action recognition and multimedia event detection, etc.

PlayingGitar

HorseRiding

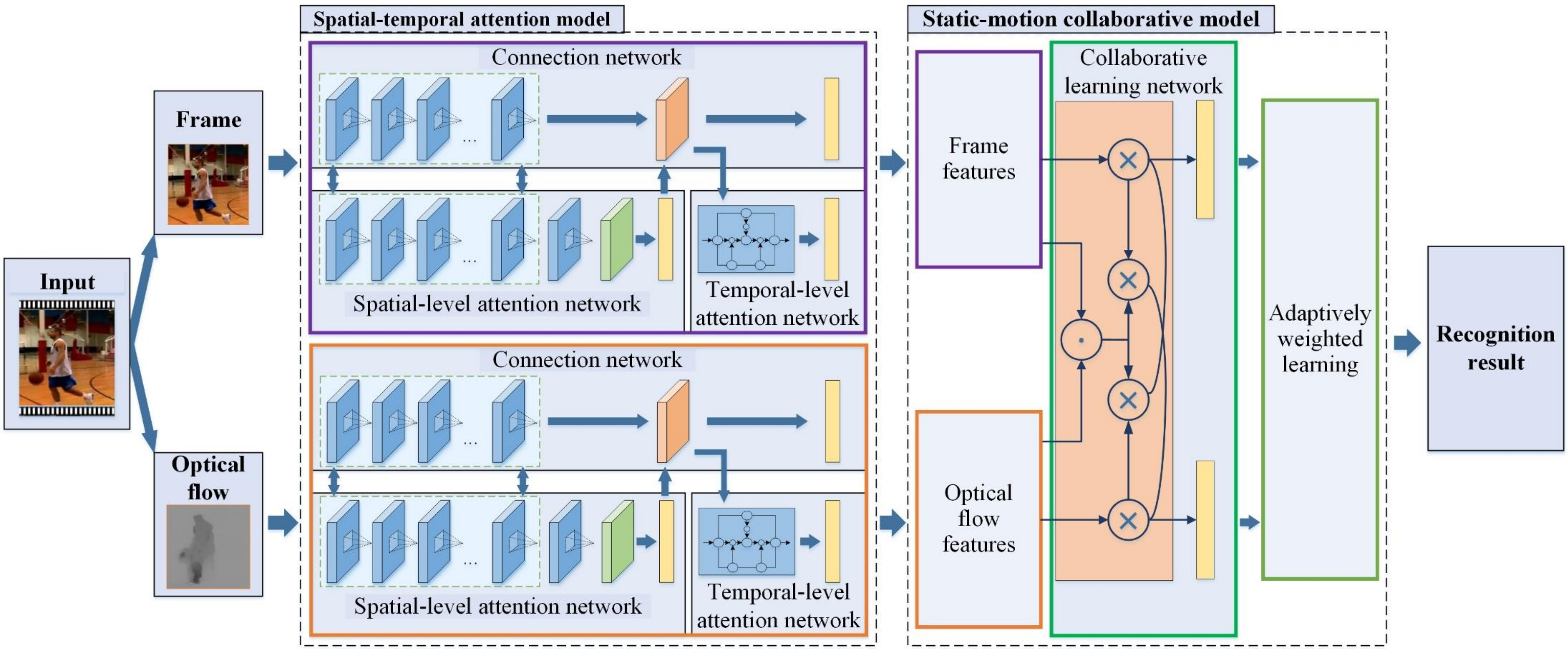
Birthday Celebration

Parade

PEKING UNIVERSITY

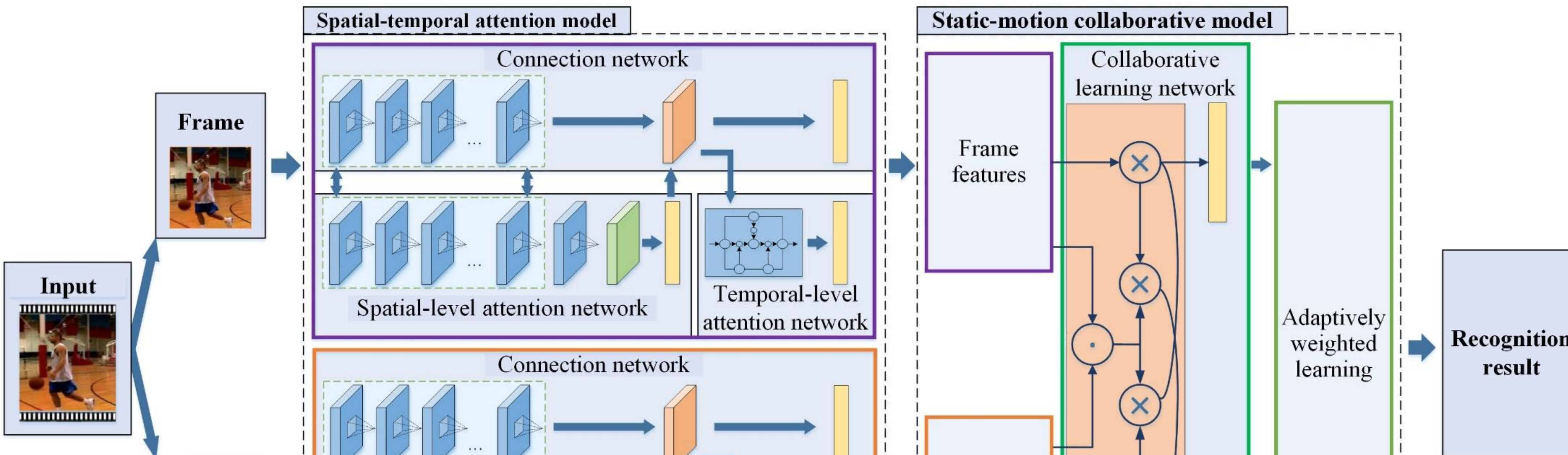
1. Video concept recognition (2/2)

- We propose **two-stream collaborative learning with spatial-temporal attention**
  - **spatial-temporal attention model**: jointly capture the video evolutions both in spatial and temporal domains
  - **static-motion collaborative model**: adopt collaborative guidance between static and motion information to promote feature learning



1. Video concept recognition (2/2)

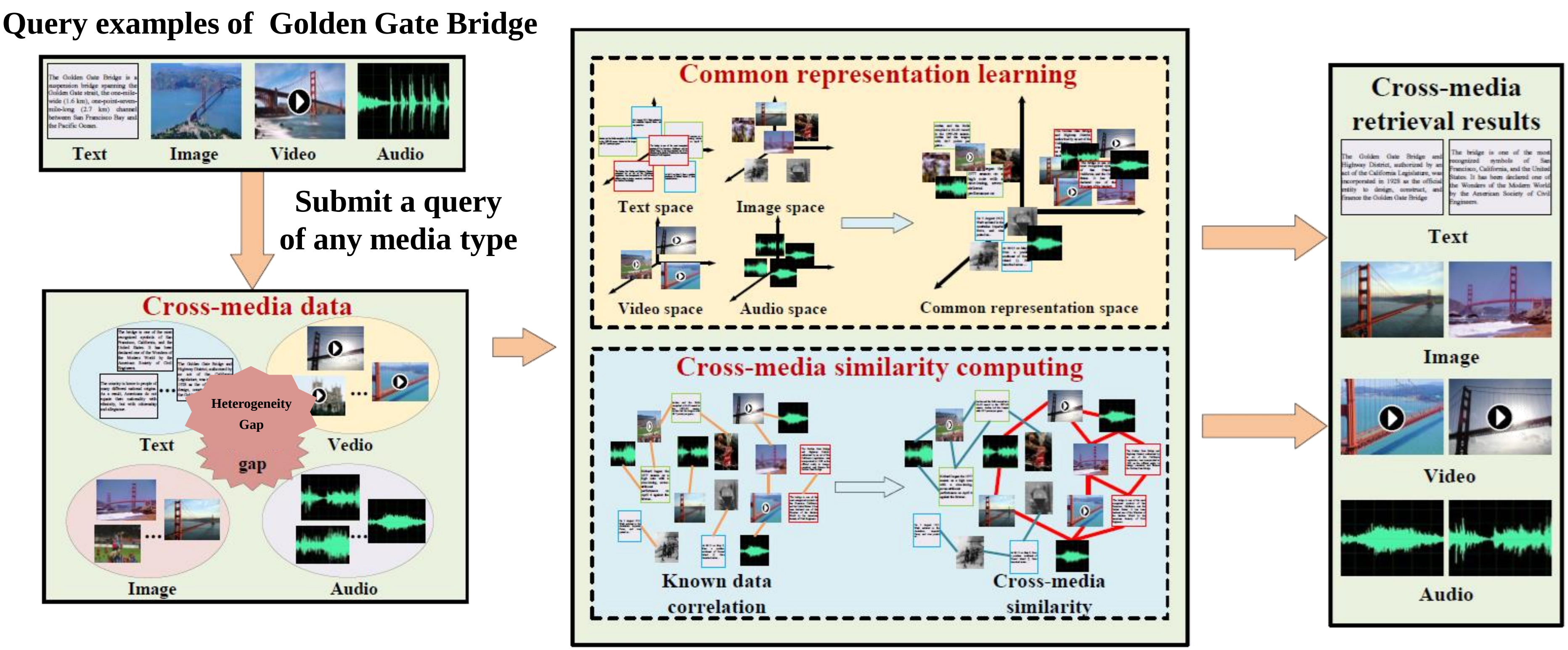
- We propose **two-stream collaborative learning with spatial-temporal attention**
  - **spatial-temporal attention model**: jointly capture the video evolutions both in spatial and temporal domains
  - **static-motion collaborative model**: adopt collaborative guidance between static and motion information to promote feature learning



Yuxin Peng, Yunzhen Zhao, and Junchao Zhang, "Two-stream Collaborative Learning with Spatial-Temporal Attention for Video Classification", ***IEEE TCSVT 2017 (after minor revision) arXiv: 1704.01740***

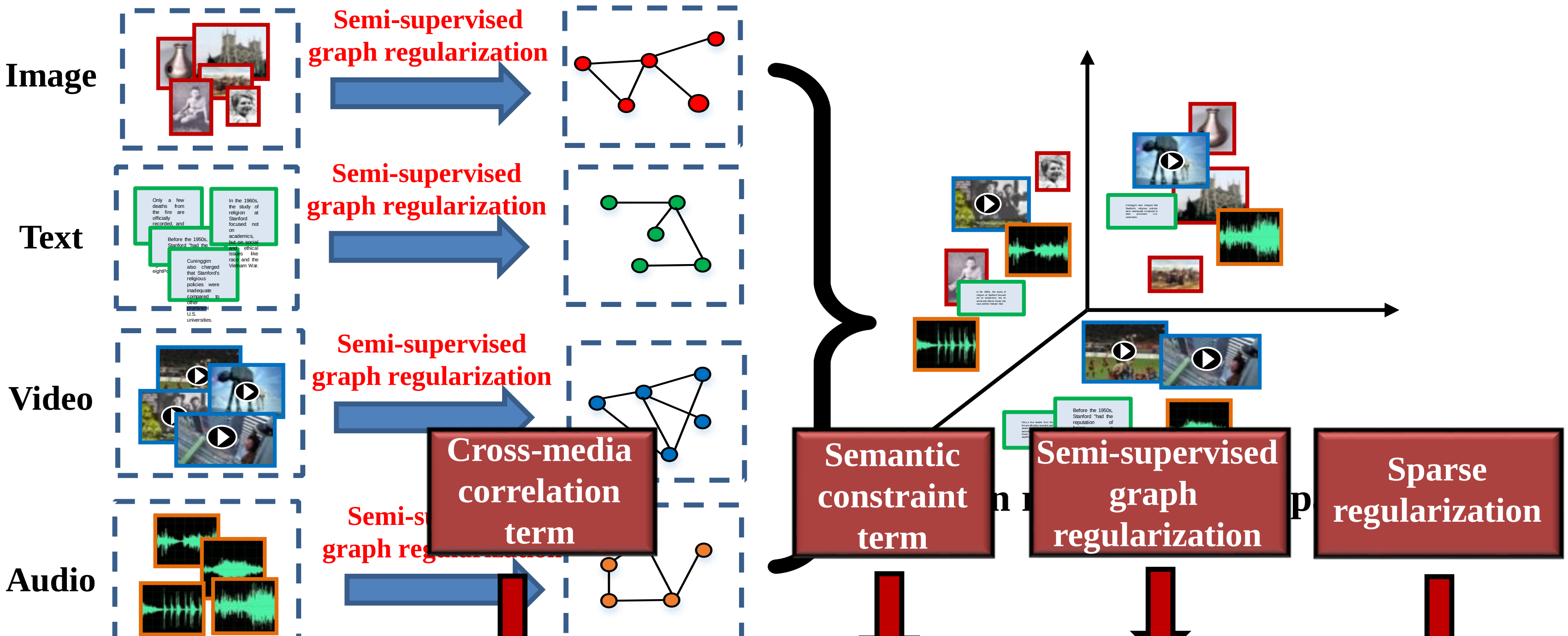
2. Cross-media Retrieval (1/5)

- **Cross-media retrieval:**
  - Perform retrieval among different media types, such as image, text, audio and video
- **Challenge:**
  - **Heterogeneity gap**: Different media types have inconsistent representations



2. Cross-media Retrieval (2/5)

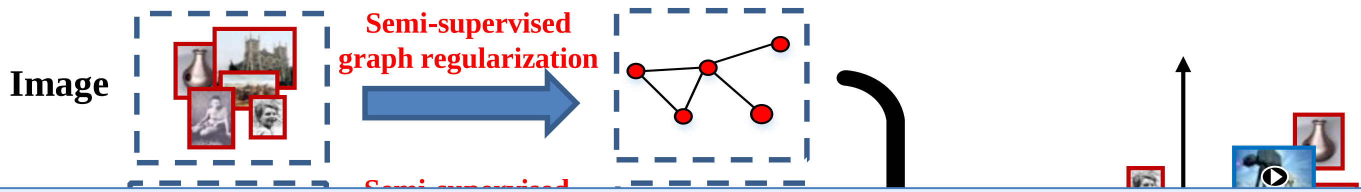
- We propose **common representation learning based on sparse and semi-supervised regularization**, which models correlation and high-level semantics in a **unified framework**, and exploits complementary information among multiple media types to reduce noise



$$\underset{P^{(1)},\dots,P^{(s)}}{\operatorname{argmin}} \sum_{i=1}^{s}\sum_{j=i+1}^{s} \boxed{\left\| P^{(i)^T} X_{mij}^{(i)} - P^{(j)^T} X_{mij}^{(j)} \right\|_F^2} + \sum_{i=1}^{s}\left(\boxed{\left\| P^{(i)^T} X^{(i)} - Y^{(i)} \right\|_F^2} + \boxed{\lambda(\Omega(P^{(i)}))} + \boxed{\left\| P^{(i)} \right\|_{2,1}} \right)$$

2. Cross-media Retrieval (2/5)

- We propose **common representation learning based on sparse and semi-supervised regularization**, which models correlation and high-level semantics in a **unified framework**, and exploits complementary information among multiple media types to reduce noise

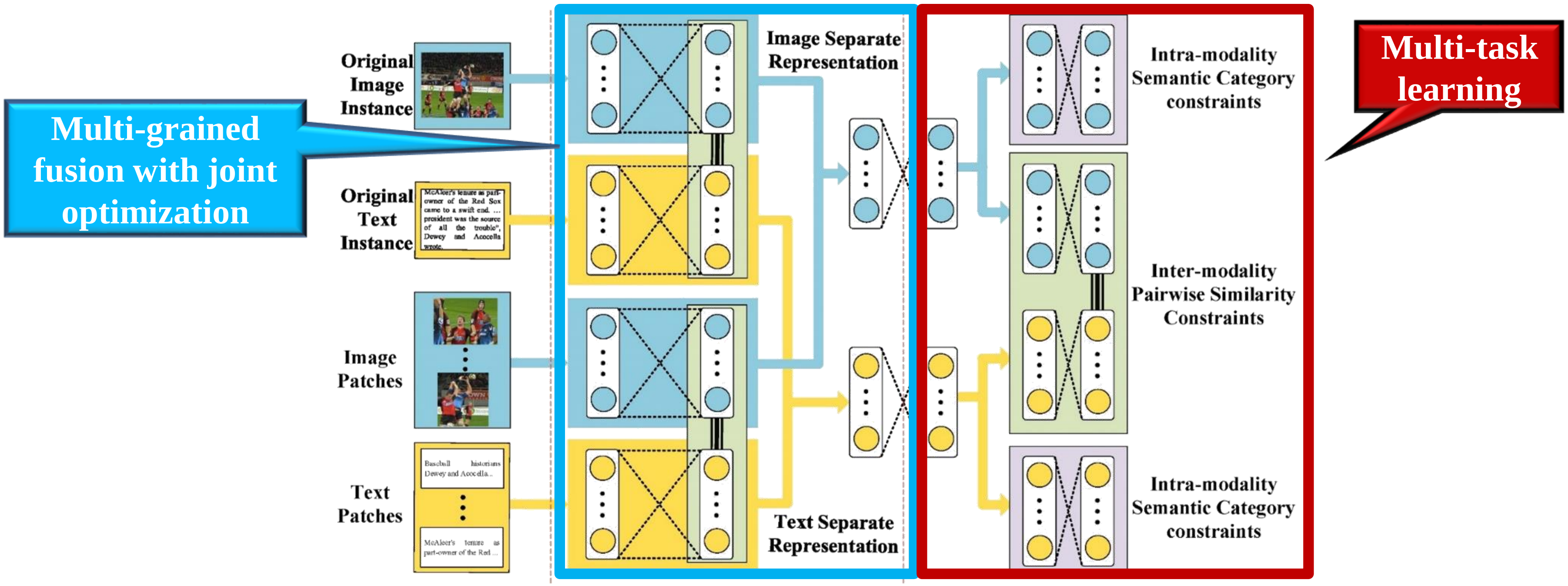


Comment from Reviewers of TCSVT: "the proposed method is **quite novel**.", and "**jointly represents several media** for cross-media retrieval, while the previous works usually deal with two different media"

- Yuxin Peng, Xiaohua Zhai, Yunzhen Zhao, and Xin Huang, "Semi-Supervised Cross-Media Feature Learning with Unified Patch Graph Regularization", ***IEEE TCSVT*** **2016**
- Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao, "Learning Cross-Media Joint Representation with Sparse and Semisupervised Regularization", ***IEEE TCSVT 2014***

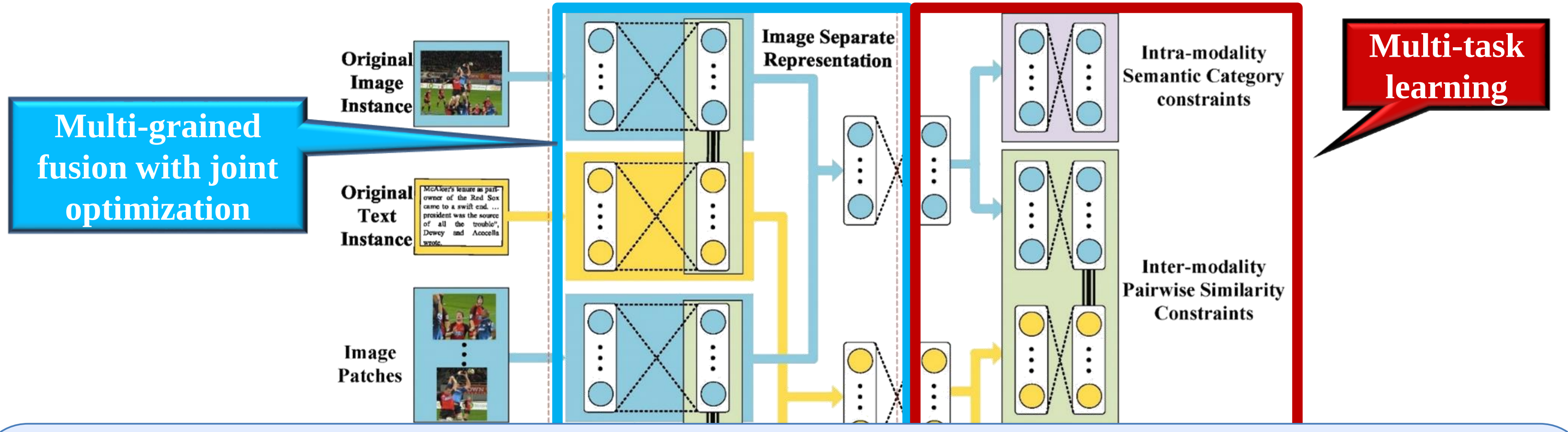
2. Cross-media Retrieval (3/5)

- We propose a **cross-modal correlation learning** approach with **multi-grained fusion** by hierarchical network. It exploits **multi-level association with joint optimization** and adopts **multi-task learning** to preserve intra-modality and inter-modality correlation



2. Cross-media Retrieval (3/5)

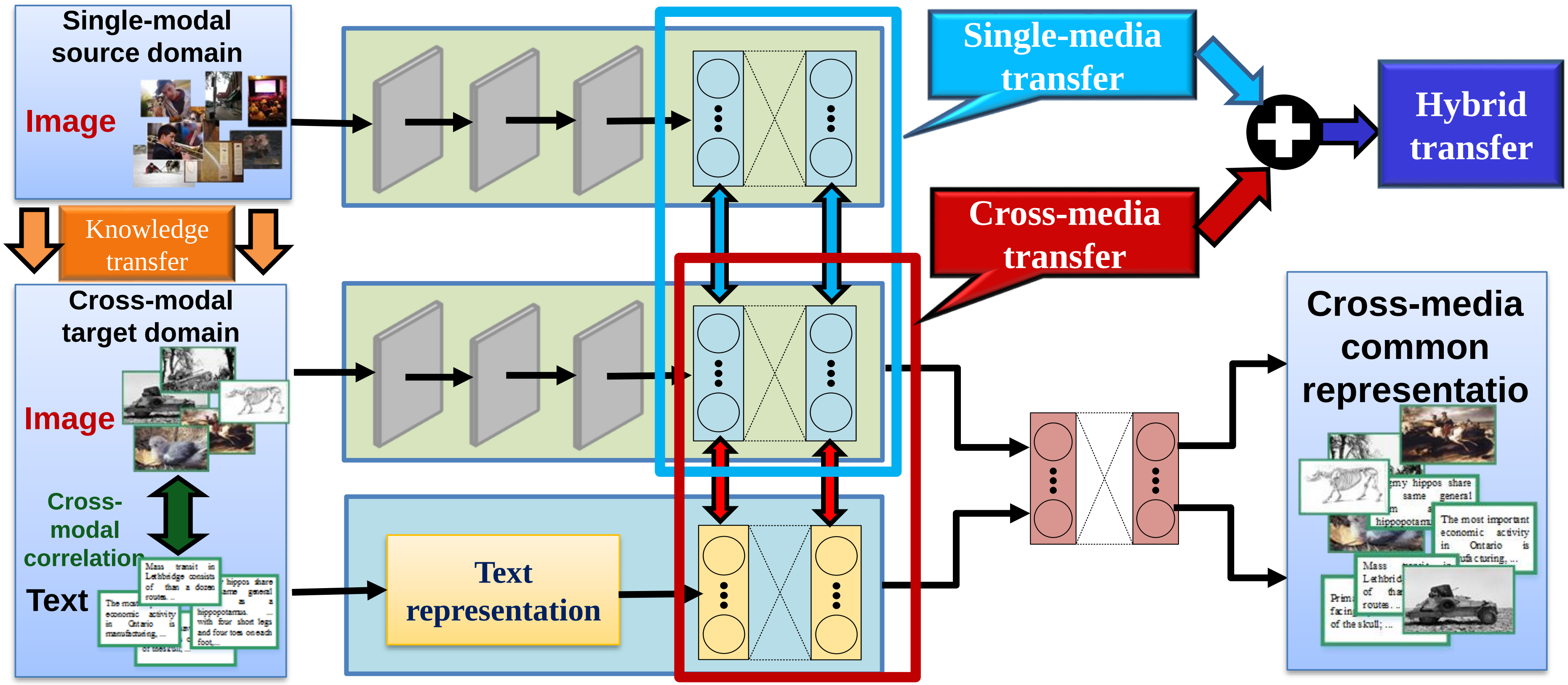
- We propose a **cross-modal correlation learning** approach with **multi-grained fusion** by hierarchical network. It exploits **multi-level association with joint optimization** and adopts **multi-task learning** to preserve intra-modality and inter-modality correlation



- Yuxin Peng, Xin Huang, and Jinwei Qi. "Cross-media Shared Representation by Hierarchical Learning with Multiple Deep Networks". ***IJCAI 2016.***
- Yuxin Peng, Jinwei Qi, Xin Huang, and Yuxin Yuan, "CCL: Cross-modal Correlation Learning with Multi-grained Fusion by Hierarchical Network", ***IEEE TMM* 2017**

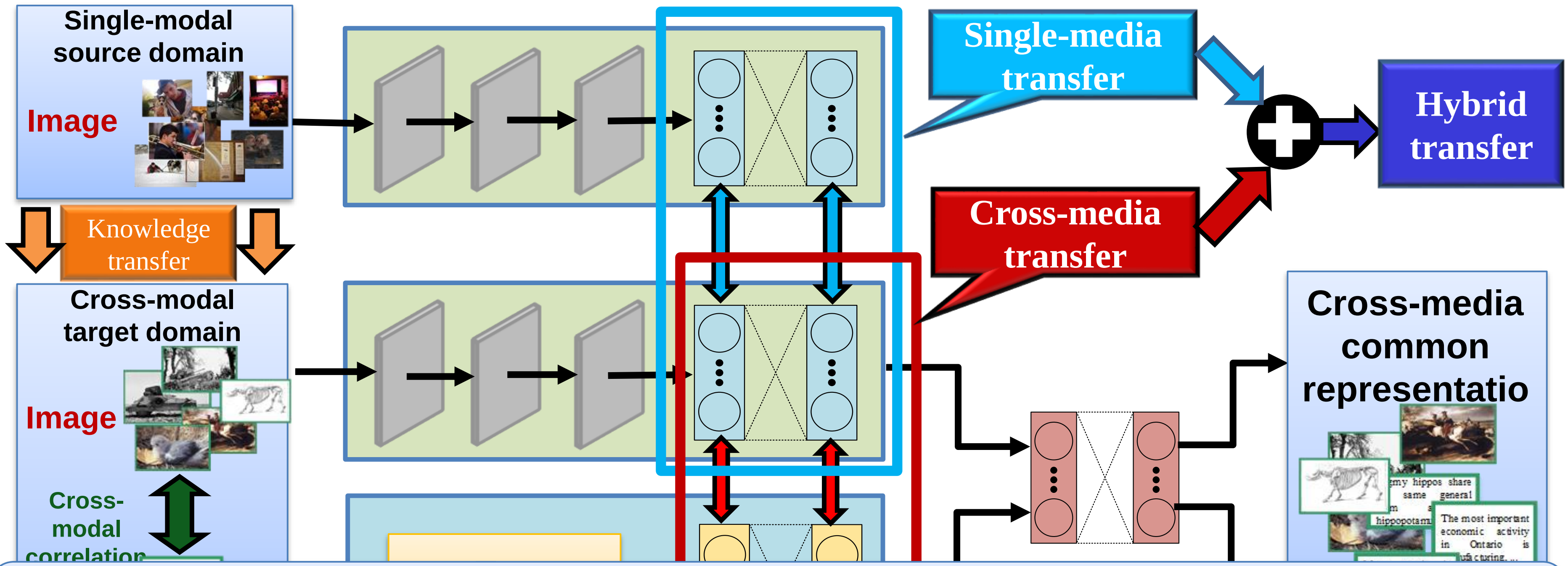
2. Cross-media Retrieval (4/5)

- For addressing the problem of **insufficient training data** in DNN-based cross-media retrieval method, we propose **cross-media hybrid transfer network**, which exploits the semantic information of existing large-scale **single-media datasets** to promote the network training of cross-media common representation learning



2. Cross-media Retrieval (4/5)

- For addressing the problem of **insufficient training data** in DNN-based cross-media retrieval method, we propose **cross-media hybrid transfer network**, which exploits the semantic information of existing large-scale **single-media datasets** to promote the network training of cross-media common representation learning

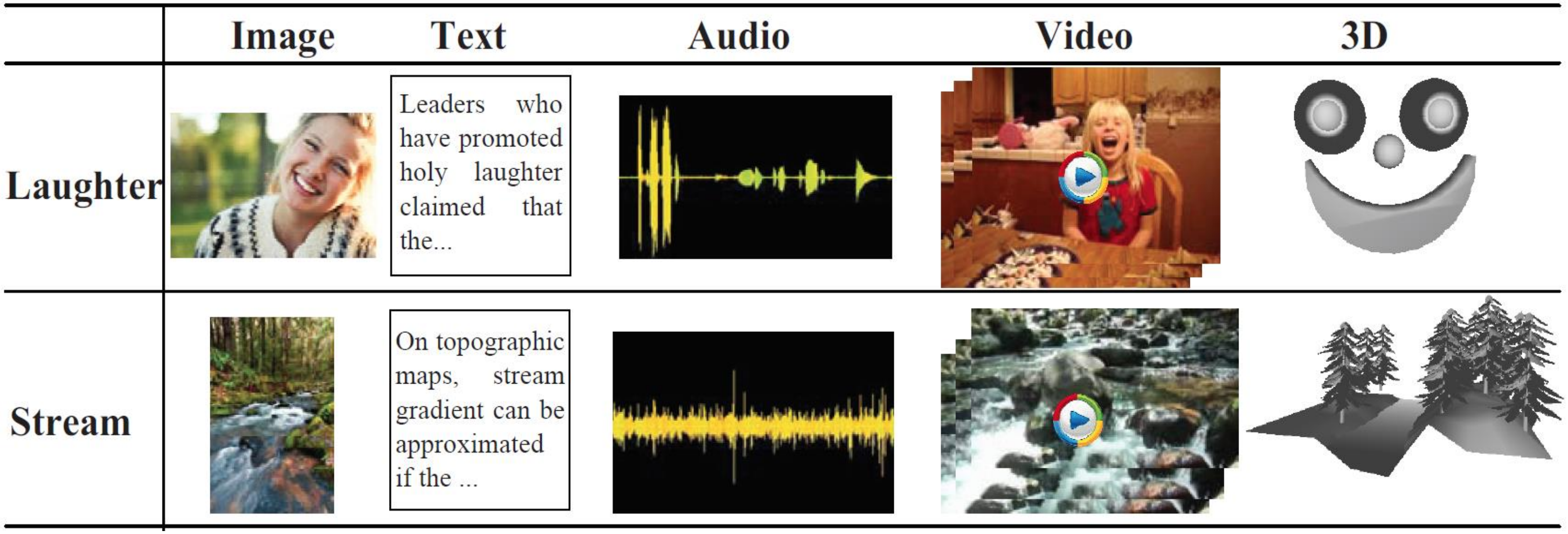


Xin Huang, Yuxin Peng, and Mingkuan Yuan, "Cross-modal Common Representation Learning by Hybrid Transfer Network", ***IJCAI 2017.***

2. Cross-media Retrieval (5/5)

- We have released **PKU-XMedia**、**PKU-XMediaNet** dataset with 5 media types. Datasets and source codes of our related works:

http://www.icst.pku.edu.cn/mipl/xmedia



- Interested in cross-media retrieval? Hope our recent overview is helpful for you

> Yuxin Peng, Xin Huang, and Yunzhen Zhao, "**An Overview of Cross-media Retrieval: Concepts, Methodologies, Benchmarks and Challenges**", IEEE TCSVT, 2017. arXiv: 1704.02223.

3. Fine-grained Image Classification (1/4)

- **Fine-grained Image Classification:**
  - Recognize hundreds of subcategories belonging to the same basic-level category
- **Challenges:**

Large variances in the same subcategory

Black Footed Albatross

Smart fortwo Convertible

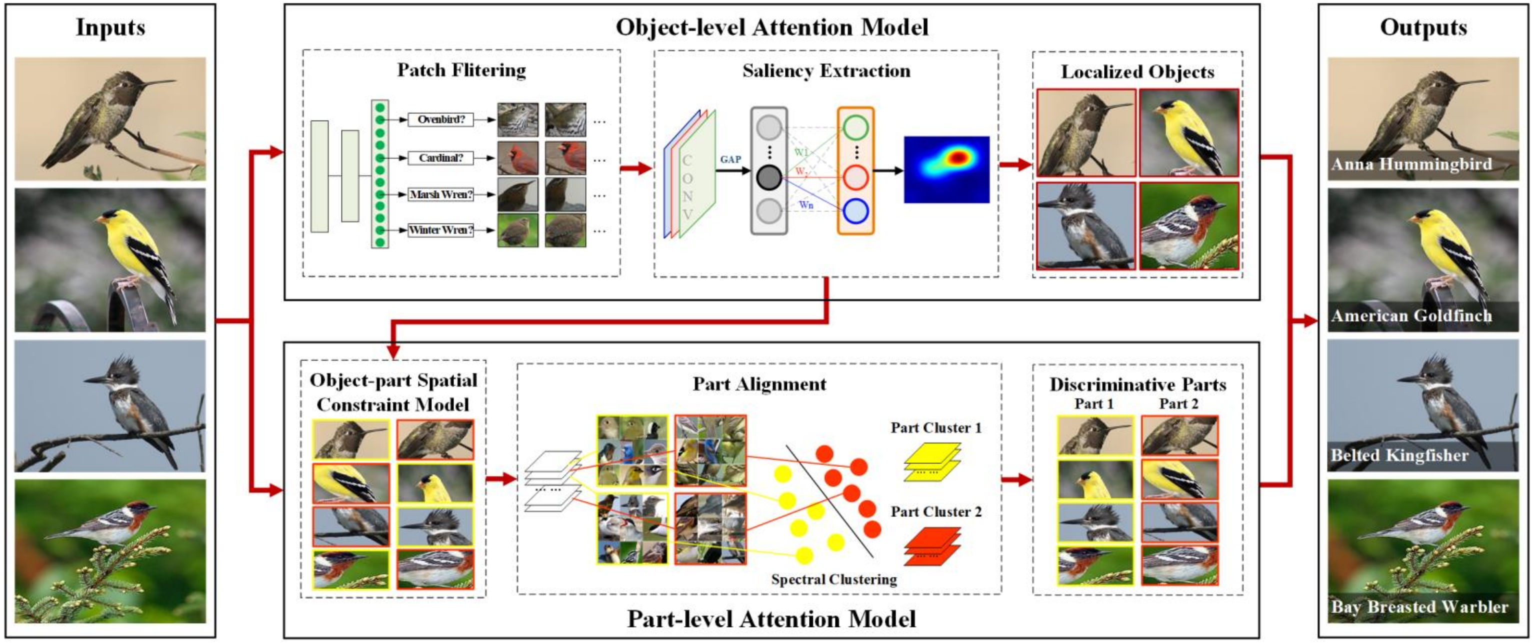
Small variances among different subcategories

Marsh Wren **Rock Wren** **Winter Wren**

BMW 1 **Hyundai Elantra** **Toyota Sequoia**

3. Fine-grained Image Classification (2/4)

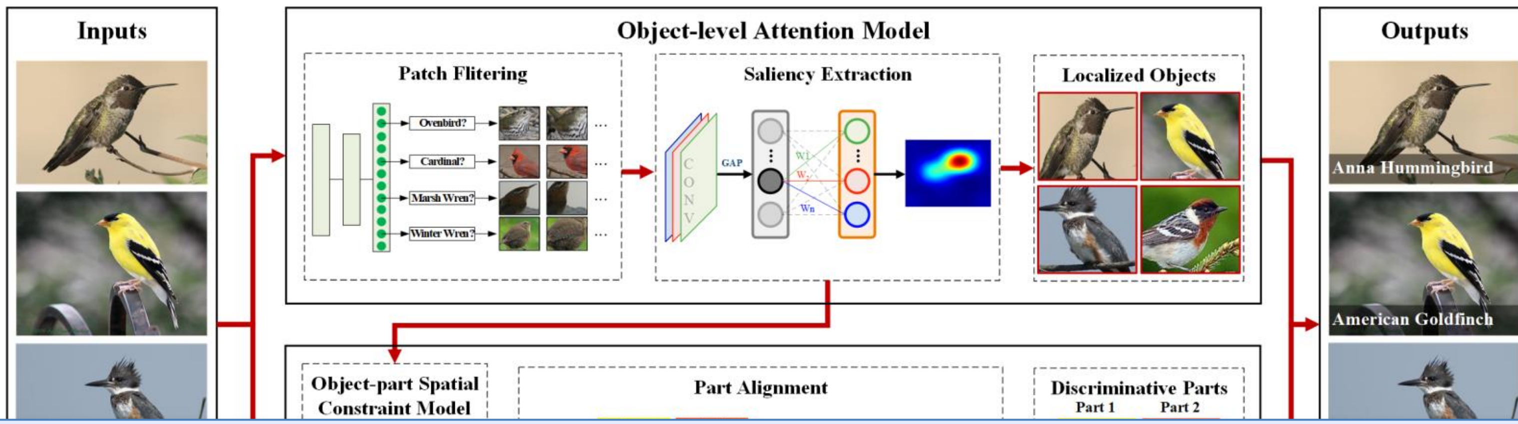
- To address the problem of fine-grained image classification, **object-part attention model** is proposed, which is the **first work** to classify fine-grained images **without using object or parts annotations** in both training and testing phase, but still achieves promising results.



PEKING UNIVERSITY

3. Fine-grained Image Classification (2/4)

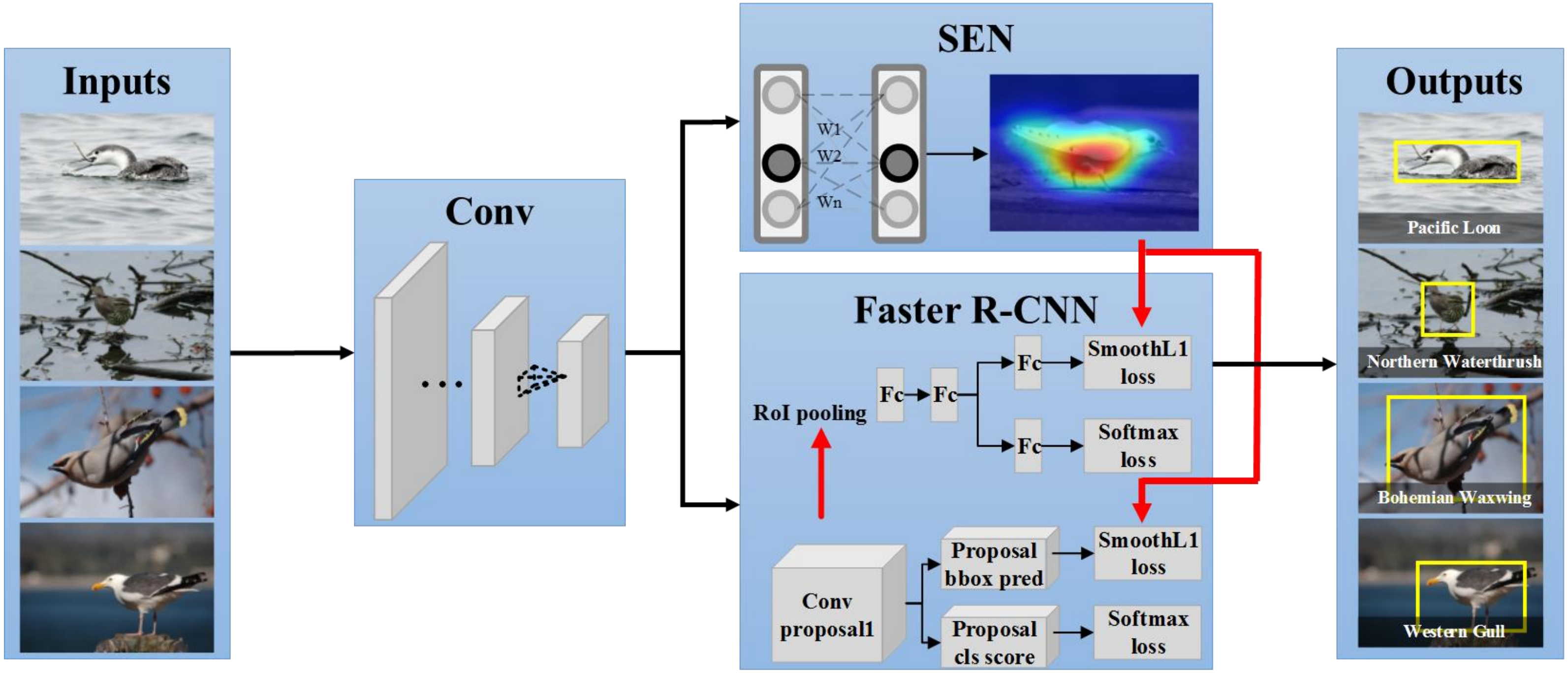
- To address the problem of fine-grained image classification, **object-part attention model** is proposed, which is the **first work** to classify fine-grained images **without using object or parts annotations** in both training and testing phase, but still achieves promising results.



- Yuxin Peng, Xiangteng He, and Junjie Zhao, "Object-Part Attention Model for Fine-grained Image Classification", ***IEEE TIP 2017***
- Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang, "The Application of Two-level Attention Models in Deep Convolutional Neural Network for Fine-grained Image Classification", ***CVPR 2015***

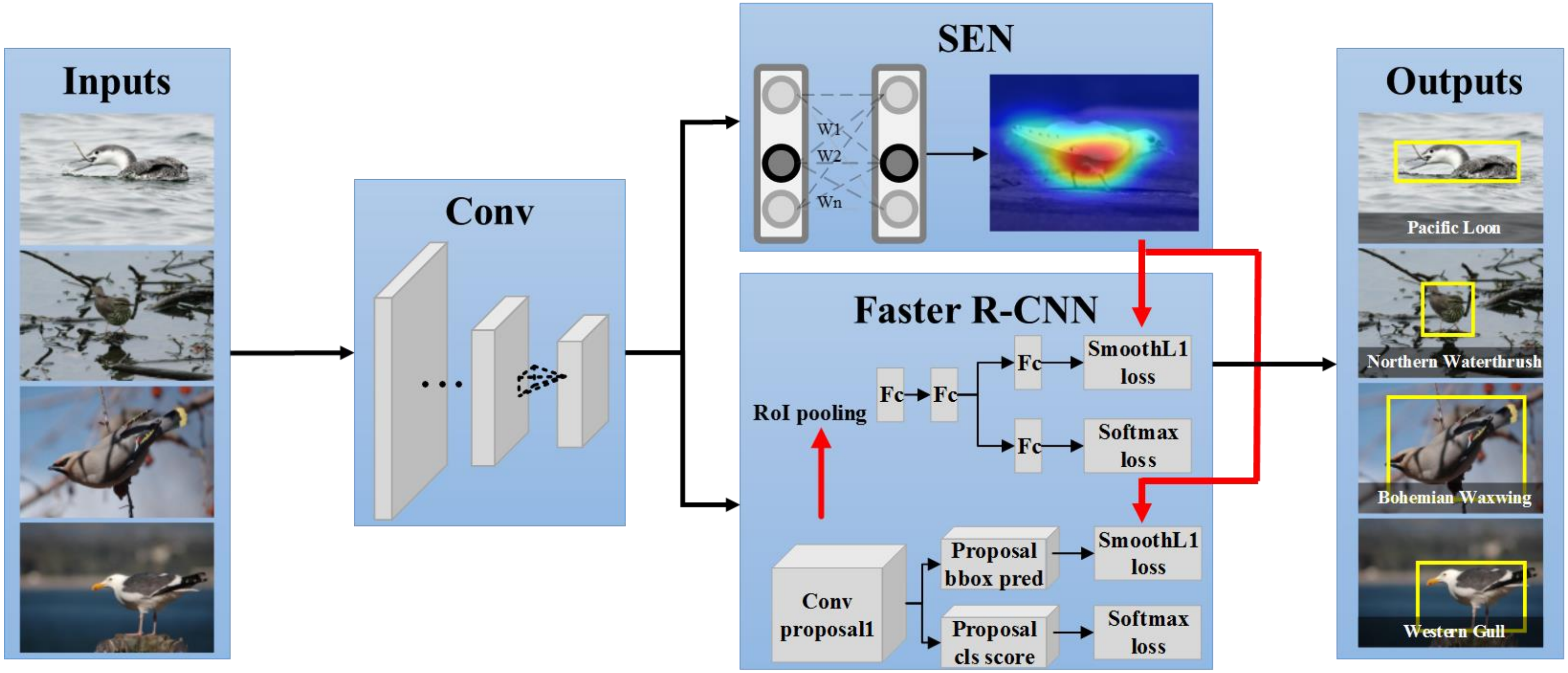
3. Fine-grained Image Classification (3/4)

- To accelerate classification speed, **saliency-guided fine-grained discriminative localization** is proposed, which jointly facilitates fine-grained image classification and discriminative localization



3. Fine-grained Image Classification (3/4)

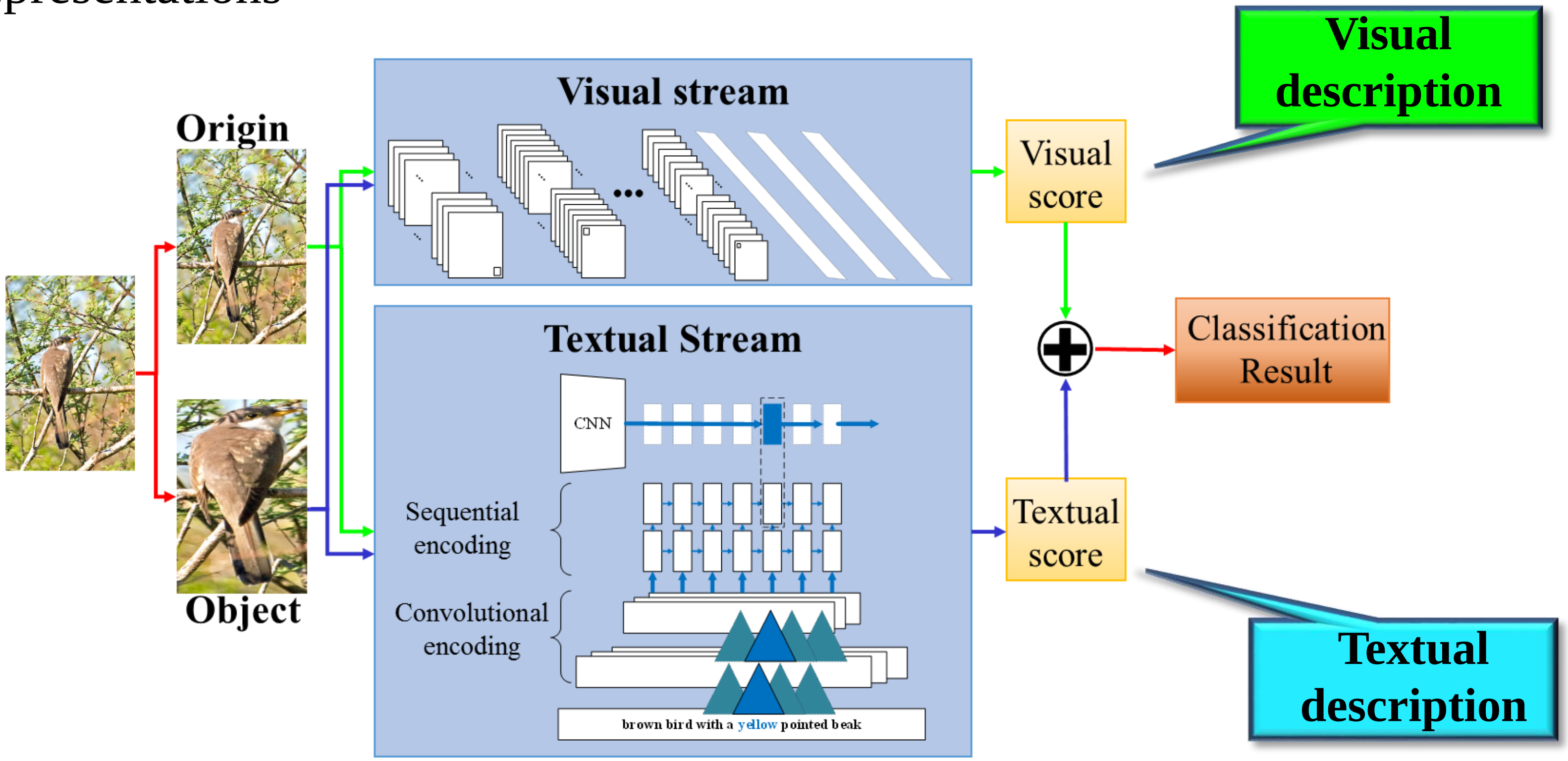
- To accelerate classification speed, **saliency-guided fine-grained discriminative localization** is proposed, which jointly facilitates fine-grained image classification and discriminative localization



Xiangteng He, Yuxin Peng and Junjie Zhao, "Fine-grained Discriminative Localization via Saliency-guided Faster R-CNN", ***ACM MM 2017.***

3. Fine-grained Image Classification (4/4)

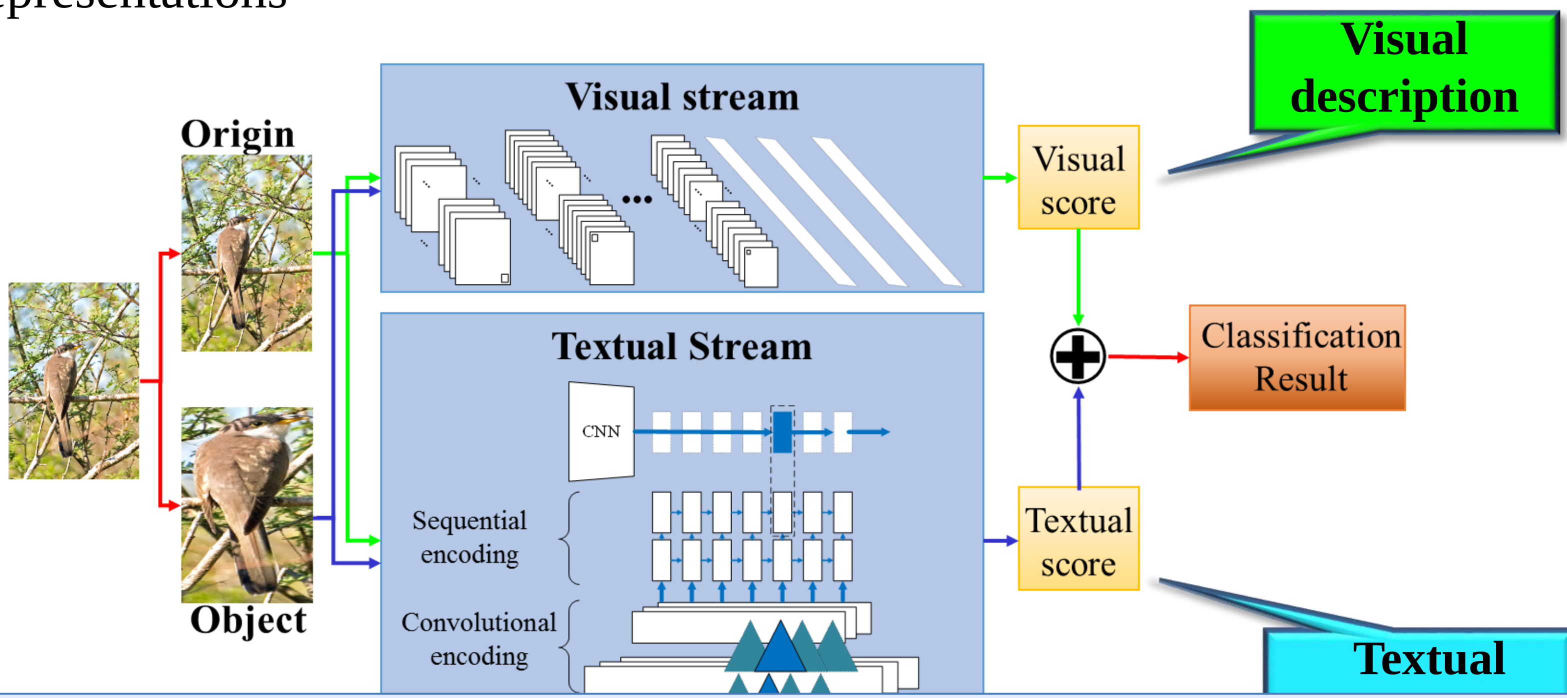
- Considering the complementarity of text, **a two-stream model is proposed to combine vision and language** for learning multi-granularity, multi-view and multi-level representations



PEKING UNIVERSITY

3. Fine-grained Image Classification (4/4)

- Considering the complementarity of text, **a two-stream model is proposed to combine vision and language** for learning multi-granularity, multi-view and multi-level representations



Xiangteng He and Yuxin Peng, “Fine-grained Image Classification via Combining Vision and Language”, ***CVPR 2017***.

Contact：

Email： pengyuxin@pku.edu.cn

Phone： 010-82529699

Lab Website：

http://www.icst.pku.edu.cn/mipl