



中科院计算所  
INSTITUTE OF COMPUTING TECHNOLOGY, CAS

# Feature Representation in Person Re-identification

Hong Chang

Institute of Computing Technology  
Chinese Academy of Sciences

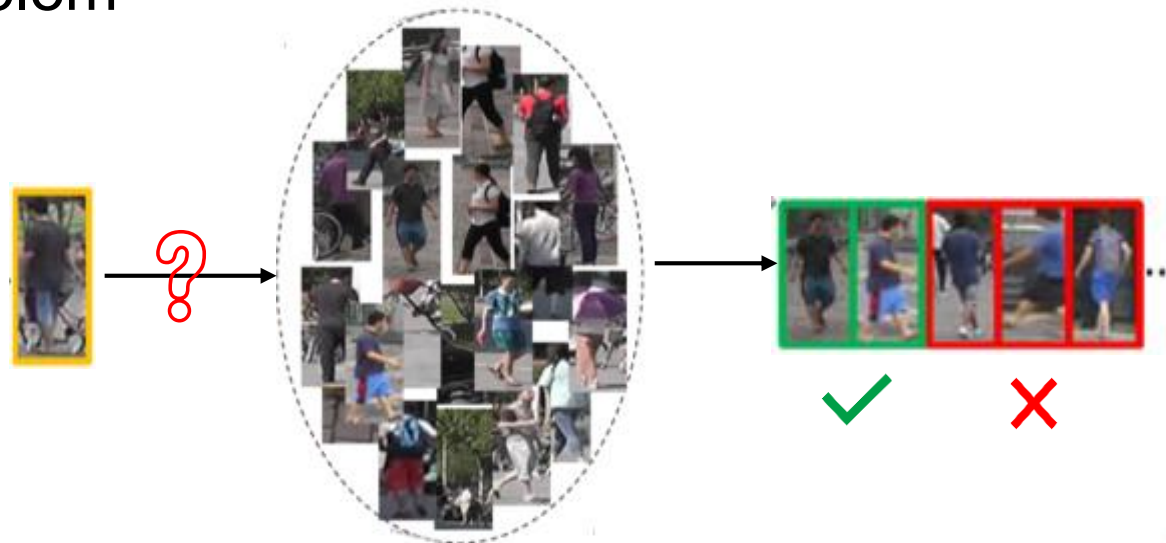
2020.1

# Contents

- Feature representation in person Re-ID
  - Related recent works
- Learning features with
  - High robustness
  - High discriminativeness
  - Low information loss/redundancy
- Discussions

# Person Re-identification

- The problem



- Main challenges



pose



scale



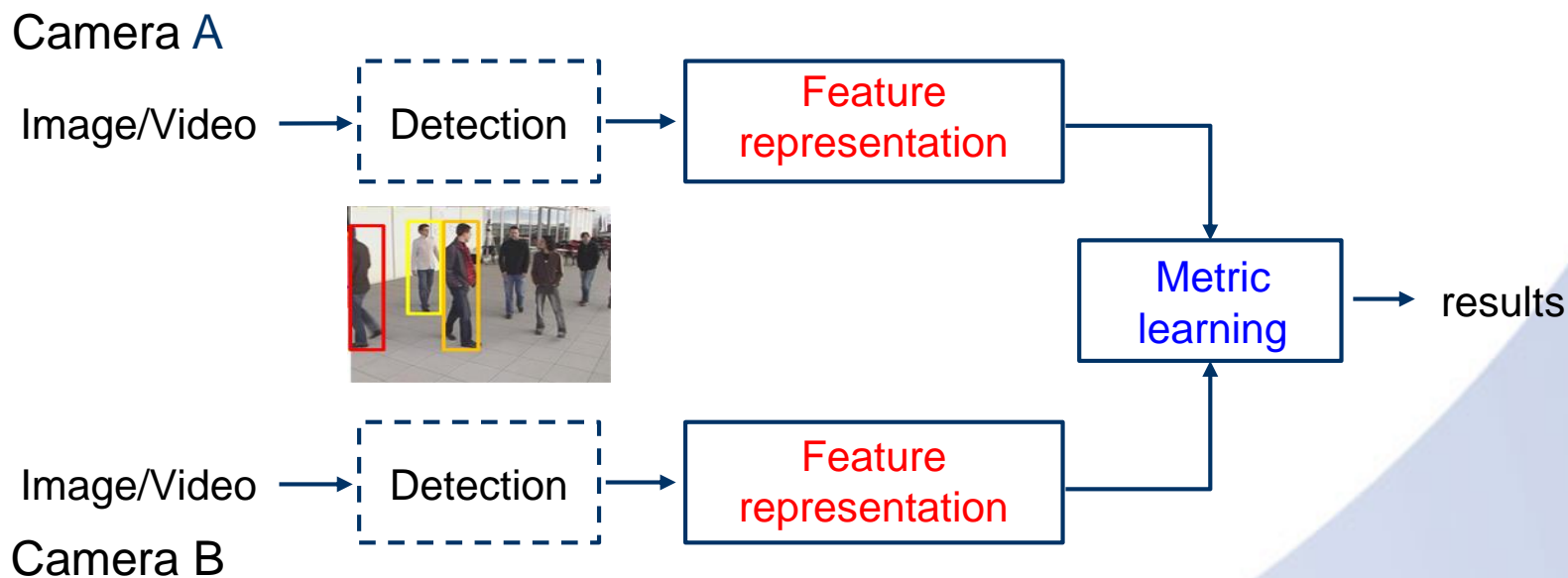
occlusion



illumination

# Feature Representation & Metric Learning

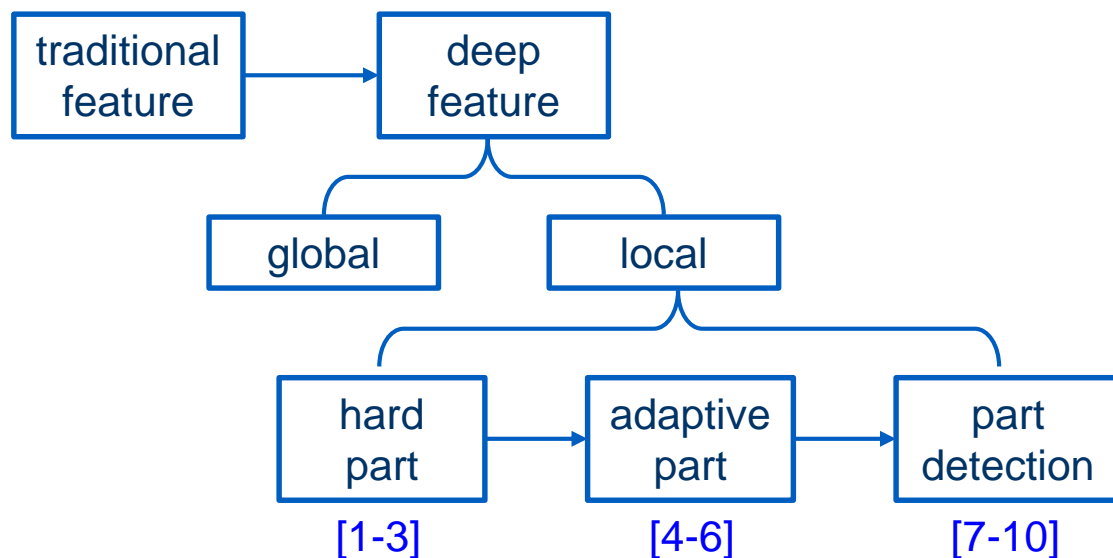
- The work flow of person Re-ID



- Two key components
  - Feature representation
  - Metric learning

# Recent Works in Feature Representation

- For images:



(a)

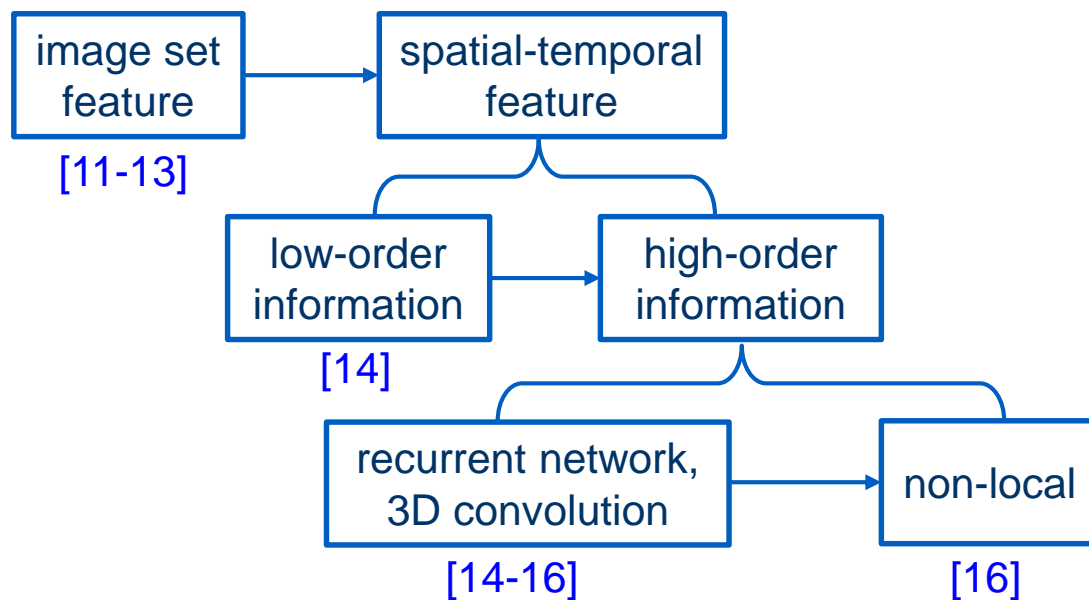


(b)

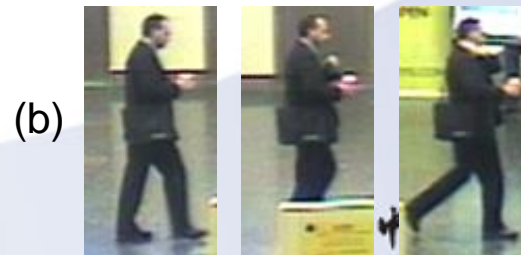
- Better person part alignment
- Weaknesses: part detection loss, extra computation, etc.
- Unsolved problems: (a) discriminative region? (b) occlusion?

# Recent Works in Feature Representation

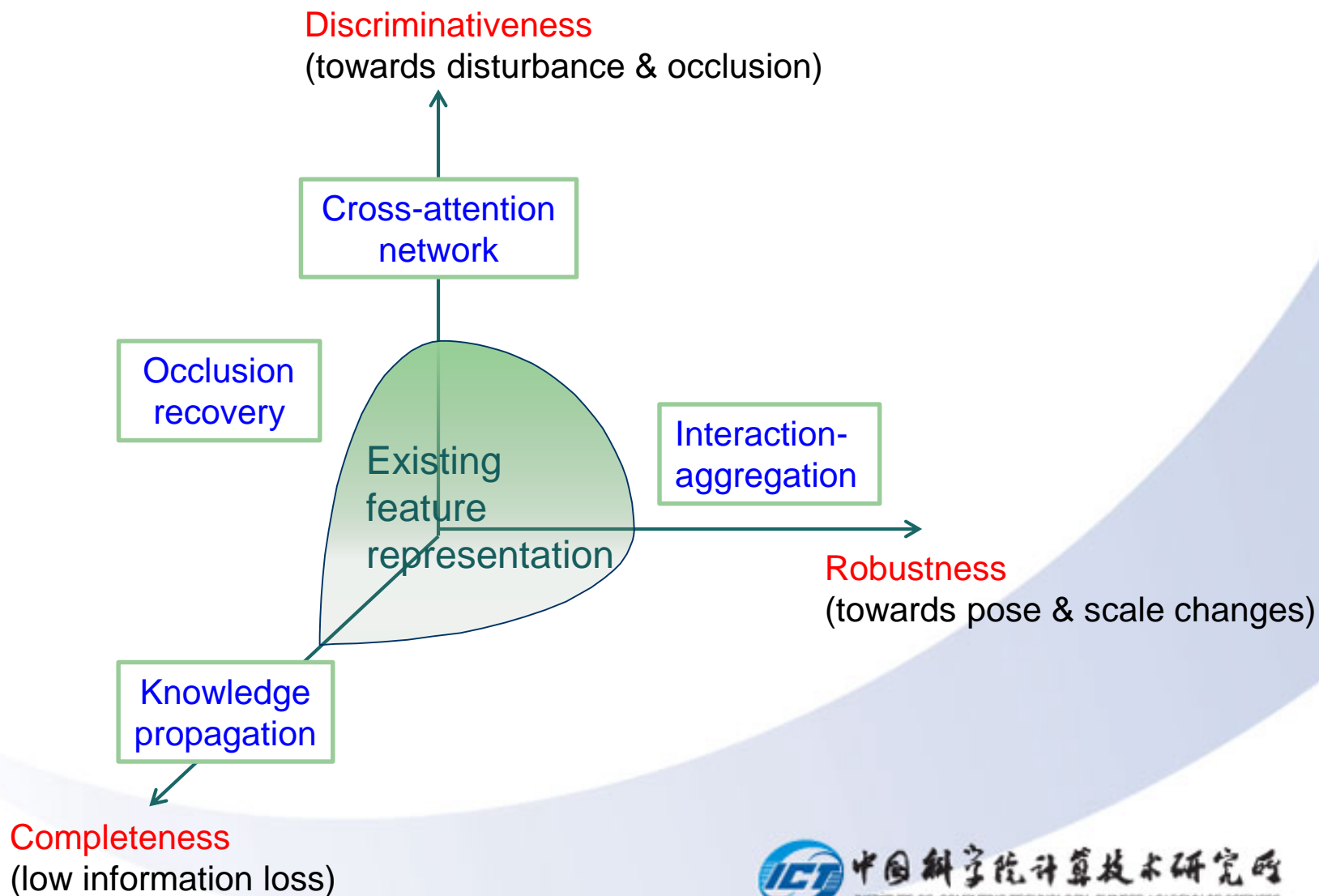
- For videos:



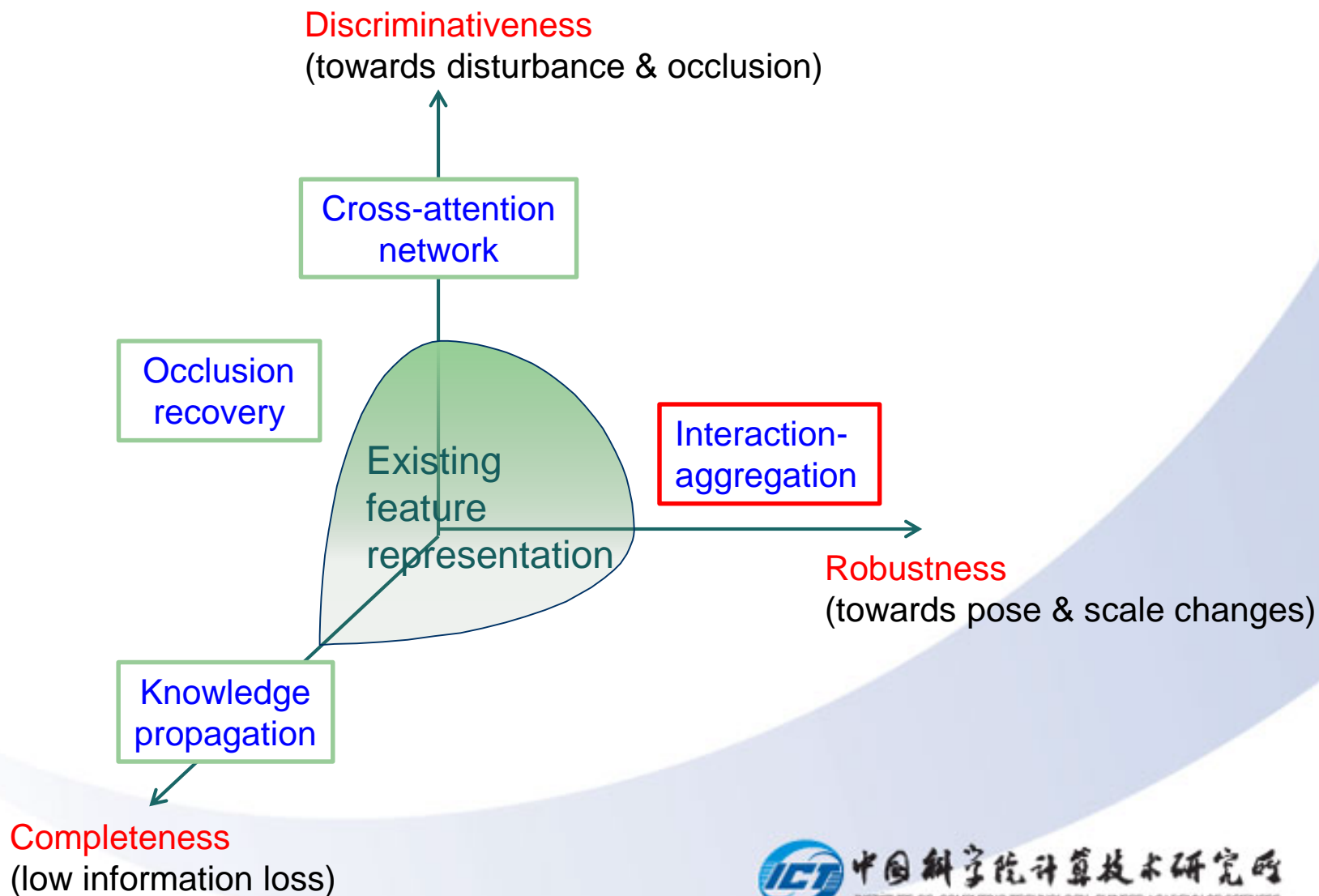
– Unsolved problems: (a) disturbance? (b) occlusion?



# Feature Representation for Person Re-ID



# Feature Representation for Person Re-ID



# Interaction-Aggregation Feature Representation

- To deal with pose and scale changes



pose



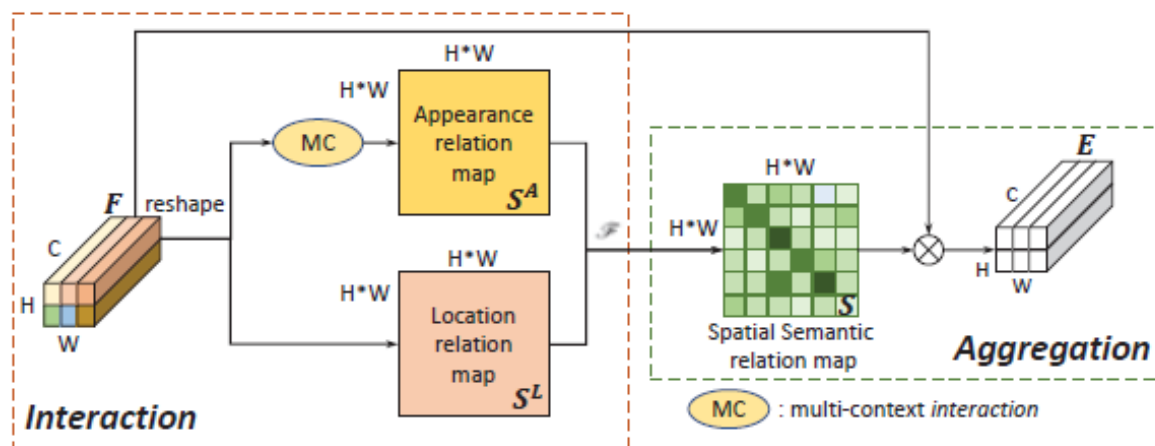
scale

- Main idea:
  - Unsupervised, Light weight
  - Semantic similarity

# Interaction-Aggregation Feature Representation

## ● Spatial IA

- adaptively determines the receptive fields according to the input person pose and scale



- **Interaction**: models the relations between spatial features to generate a semantic relation map  $S$ .

$$(S_K^A)_{ij} = \frac{\exp\left(\sum_{k=1}^{K \times K} (p_{i,k}^T p_{j,k})\right)}{\sum_{t=1}^{H \times W} \exp\left(\sum_{k=1}^{K \times K} (p_{i,k}^T p_{t,k})\right)}$$

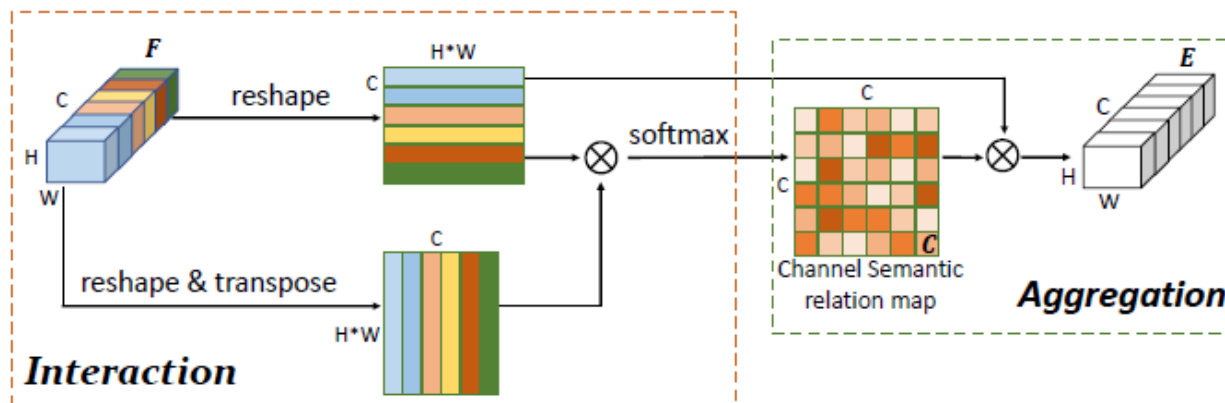
$$(S^L)_{ij} = \left\| \frac{1}{2\pi\sigma_1\sigma_2} \exp\left[-\frac{1}{2}\left(\frac{(x_j - x_i)^2}{\sigma_1^2} + \frac{(y_j - y_i)^2}{\sigma_2^2}\right)\right] \right\|_1$$

- **Aggregation**: aggregates semantically related features across different positions based on  $S$ .

# Interaction-Aggregation Feature Representation

## ● Channel IA

- selectively aggregates channel features to enhance the feature representation, especially for small scale visual cues



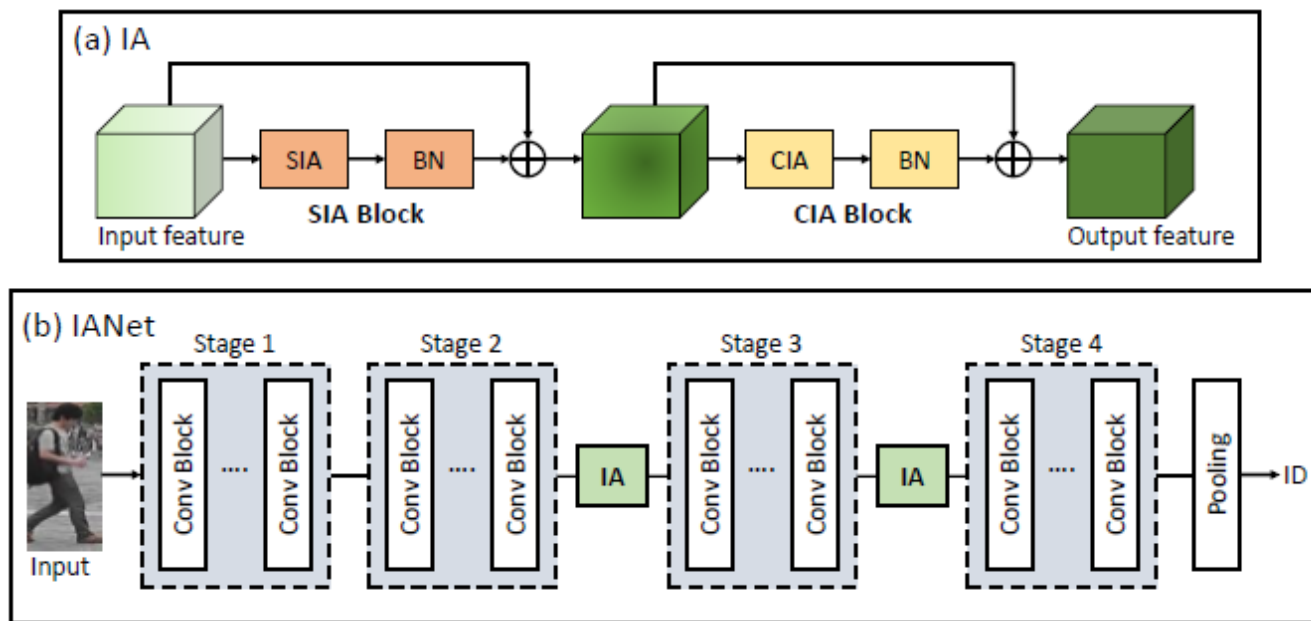
- **Interaction**: models the relations between channel features to generate a semantic relation map  $C$ .

$$C_{mn} = \frac{\exp(f_m^T f_n)}{\sum_{l=1}^C \exp(f_m^T f_l)}$$

- **Aggregation** based on relation map  $C$

# Interaction-Aggregation Feature Representation

- Overall model

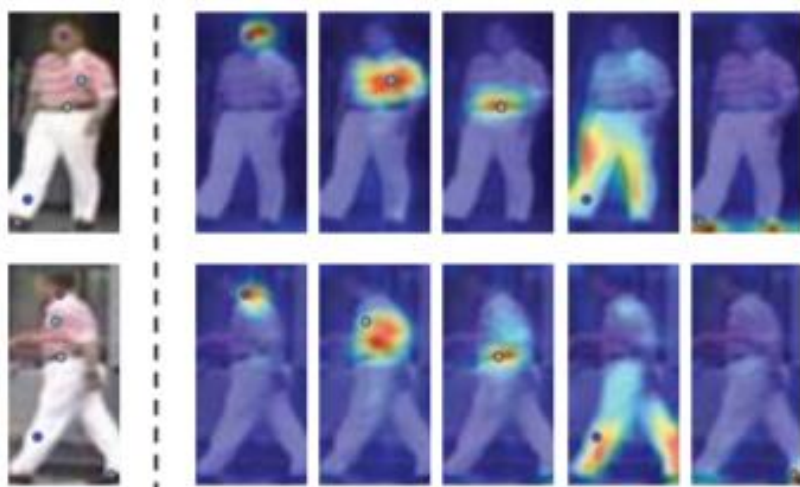


- IANet: CNN with IA modules
- Extension: spatial-temporal context IA

# Interaction-Aggregation Feature Representation

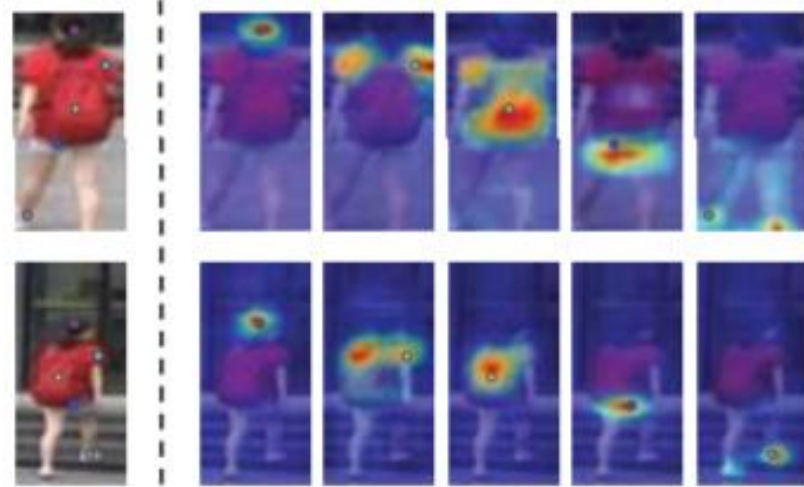
## ● Visualization results

- receptive fields: sub-relation maps with high relation values
- SIA can **adaptively localize** the body parts and visual attributes under various poses and scales.



Images

receptive fields



Images

receptive fields

# Interaction-Aggregation Feature Representation

- Visualization for pose and scale robustness
- Quantitative results

Ablation study

Model	Market-1501		DukeMTMC	
	top-1	mAP	top-1	mAP
baseline	90.4	76.2	82.1	66.0
CIA (stage <sub>3</sub> )	91.9	79.3	84.3	68.7
SIA (stage <sub>3</sub> )	94.1	82.5	85.9	72.2
IA (stage <sub>3</sub> )	94.3	82.8	85.9	72.3
IA (stage <sub>2</sub> )	94.4	82.8	86.5	71.8
IA (stage <sub>23</sub> )	<b>94.4</b>	<b>83.1</b>	<b>87.1</b>	<b>73.4</b>

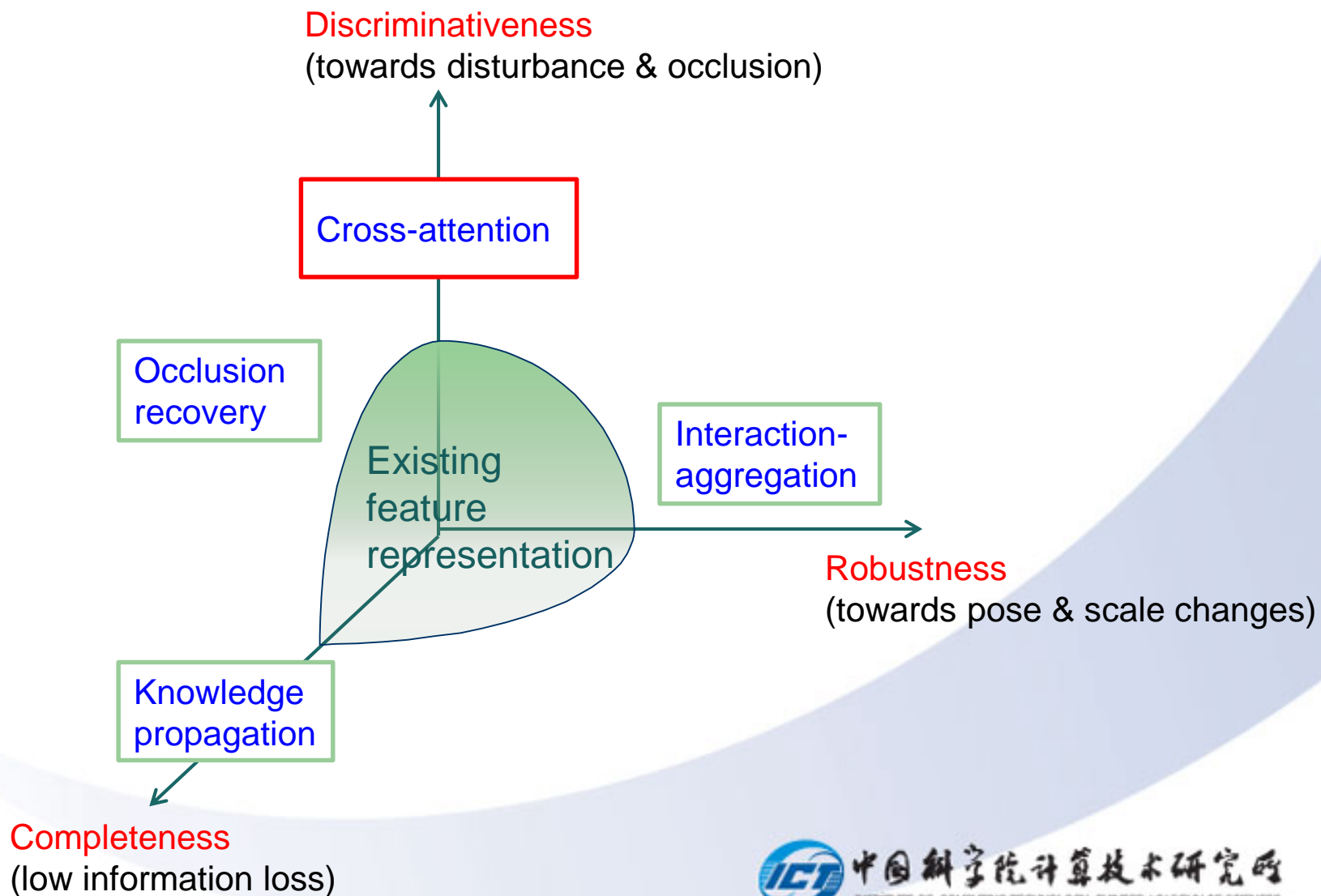
**G**: global feature  
**P**: part feature  
**MS**: multi-scale feature

Market-1501&DukeMTMC

Methods		Market-1501		DukeMTMC	
		top-1	mAP	top-1	mAP
<b>G</b>	SVDNet [41]	82.3	62.1	76.7	56.8
	Dual [9]	91.4	76.6	81.8	64.6
	Mancs [45]	93.1	82.3	84.9	71.8
<b>P</b>	Spindle* [58]	76.9	–	–	–
	SPReID* [20]	92.5	81.3	84.4	70.9
	RPP [42]	93.8	81.6	83.3	69.2
<b>MS</b>	DaRe(R) [49]	86.4	69.3	75.2	57.4
	KPM [38]	90.1	75.3	80.3	63.2
	Group [5]	93.5	81.6	84.9	69.5
	IANet	<b>94.4</b>	<b>83.1</b>	<b>87.1</b>	<b>73.4</b>

[17] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen. Interaction-and-aggregation network for person re-identification, in CVPR, 2019.

# Feature Representation for Person Re-ID



# Cross-Attention Feature Representation

- Motivation: to localize the **relevant regions** and generate **more discriminative** features

- Person re-identification



- Few-shot classification



Meta-Test (curtain)

Existing method

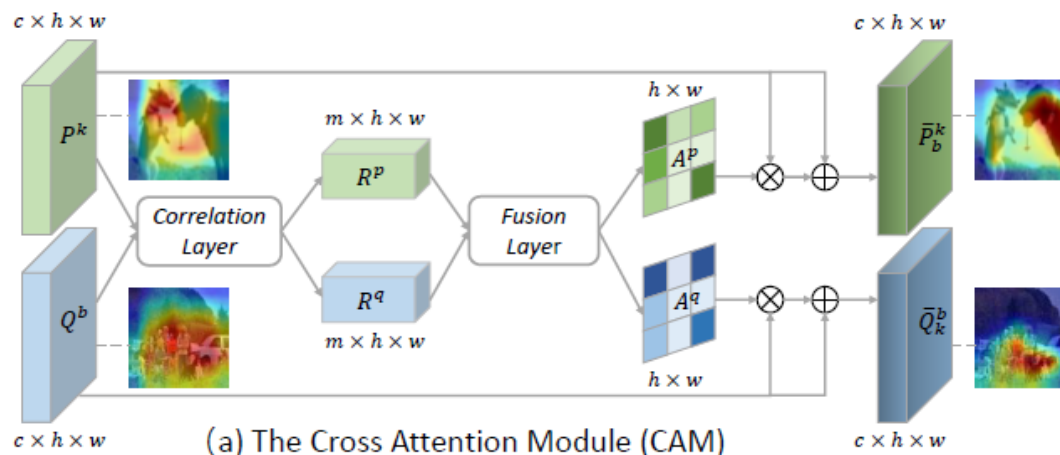
Our CAN

- Main idea: utilizing semantic relations **meta-learns** where to focus on!

# Cross-Attention Feature Representation

## ● Cross-attention module

- highlights the relevant regions and generate more discriminative feature pairs

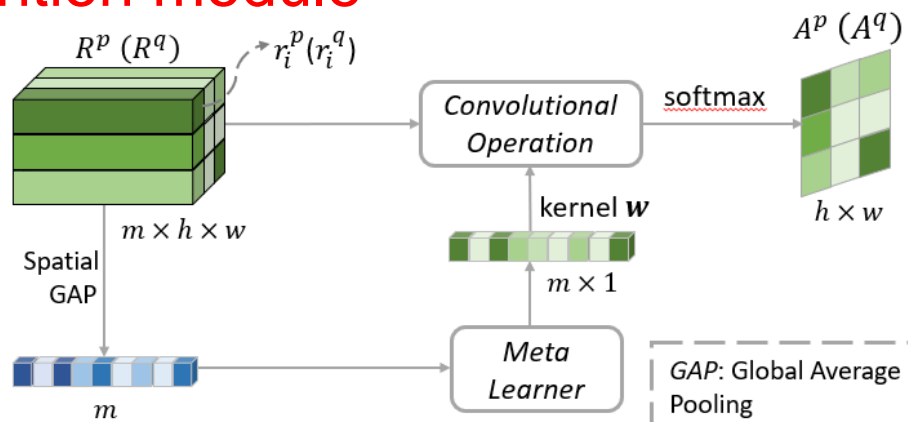


- **Correlation Layer**: calculate a **correlation map**  $R \in \mathbb{R}^{(h \times w) \times (h \times w)}$  between support feature  $P$  and query feature  $Q$ . It denotes the **semantic relevance** between each spatial position of  $P, Q$ .

$$R_{ij} = \left( \frac{p_i}{\|p_i\|_2} \right)^T \left( \frac{q_i}{\|q_i\|_2} \right) \quad R^p = R^T = [r_1^p, r_2^p, \dots, r_{hw}^p] \quad R^q = R = [r_1^q, r_2^q, \dots, r_{hw}^q]$$

# Cross-Attention Feature Representation

## ● Cross-attention module



- **Fusion Layer**: generate the **attention map pairs**  $A^p(A^q) \in \mathbb{R}^{h \times w}$  based on the corresponding correlation maps  $R$ .

- The kernel  $w$  fuses the correlation vector into an attention scalar.

$$A_i^p = \text{softmax} \left( (w^T r_i^p) / \tau \right)$$

- The kernel  $w$  should draw attention to the target object.
- A **meta fusion layer** is designed to generate the kernel  $w$ .

$$w = W_2 \left( \sigma \left( W_1 (GAP(R^P)) \right) \right)$$

# Cross-Attention Feature Representation

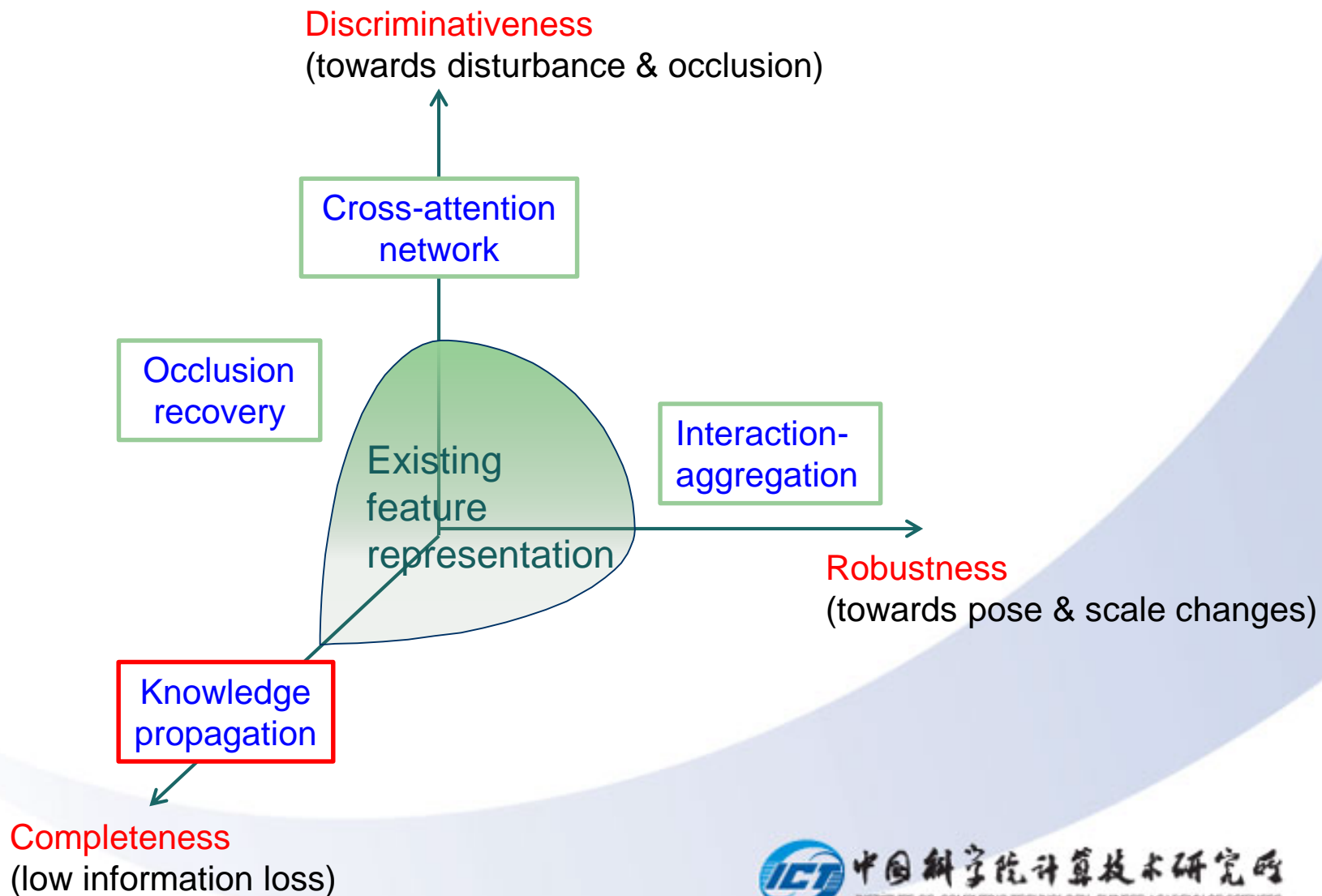
- Experiments on few-shot classification
  - state-of-the-art on miniImageNet and tieredImageNet datasets

	model	Embedding	miniImageNet		tieredImageNet	
			1-shot	5-shot	1-shot	5-shot
<b>O</b>	MAML [7]	ConvNet	48.70	55.31	51.67	70.30
	MTL [34]	ResNet-12	61.20	75.50	-	-
	LEO [31]	WRN-28	61.76	77.59	66.33	81.44
	MetaOpt [12]	ResNet-12	62.64	78.63	65.99	81.56
<b>P</b>	MetaNet [18]	ConvNet	49.21	-	-	-
	MM-Net [3]	ConvNet	53.37	66.97	-	-
	adaNet [19]	ResNet-12	56.88	71.94	-	-
<b>M</b>	MN [38]	ConvNet	43.44	60.60	-	-
	PN [33]	ConvNet	49.42	68.20	53.31	72.69
	RN [35]	ConvNet	50.44	65.32	54.48	71.32
	DN4 [13]	ConvNet	51.24	71.02	-	-
	TADAM [24]	ResNet-12	58.50	76.70	-	-
	<b>Our CAN</b>	ResNet-12	<b>63.85</b>	<b>79.44</b>	<b>69.89</b>	<b>84.23</b>
<b>T</b>	TPN [15]	ResNet-12	59.46	75.65	-	-
	<b>Our CAN+T</b>	ResNet-12	<b>67.19</b>	<b>80.64</b>	<b>73.21</b>	<b>84.93</b>

**O**: Optimization-based  
**P**: Parameter-generating  
**M**: Metric-learning  
**T**: Transductive

[18] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen. Cross Attention Network for Few-shot Classification. In NeurIPS, 2019.

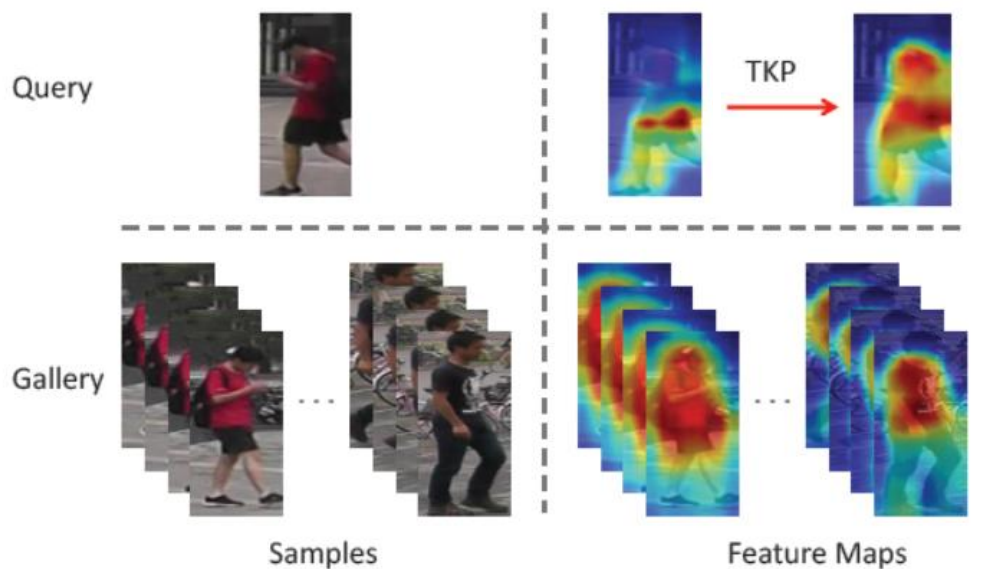
# Feature Representation for Person Re-ID



# Temporal Knowledge Propagation

- Image-to-video Re-ID

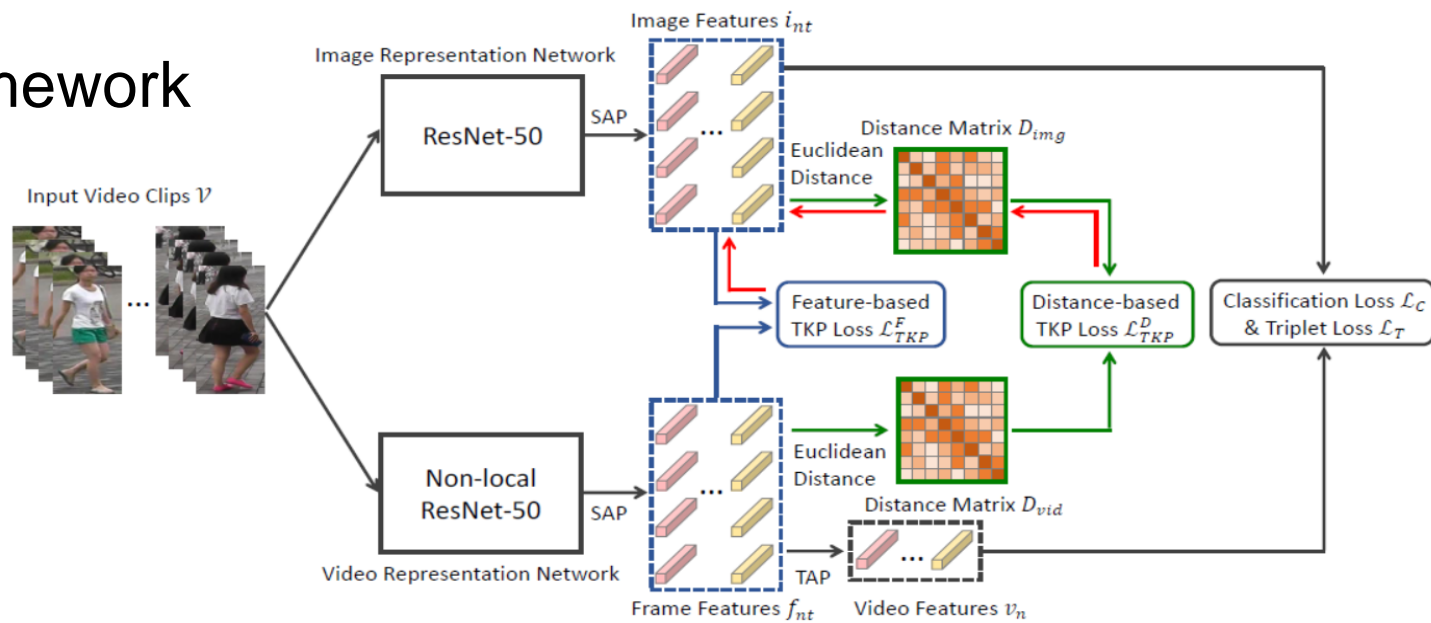
- Image lacks temporal information
- Information asymmetry increases matching difficulty



- Our solution: **temporal knowledge propagation**

# Temporal Knowledge Propagation

## ● The framework



- Propagation via **features**:

$$\mathcal{L}_{TKP}^F = \frac{1}{NT} \sum_{n=1}^N \sum_{t=1}^T \|i_{nt} - f_{nt}\|_2^2$$

- **Integrated Triplet Loss**:

$$\mathcal{L}_T = \mathcal{L}_{I2V} + \mathcal{L}_{V2I} + \mathcal{L}_{I2I} + \mathcal{L}_{V2V}$$

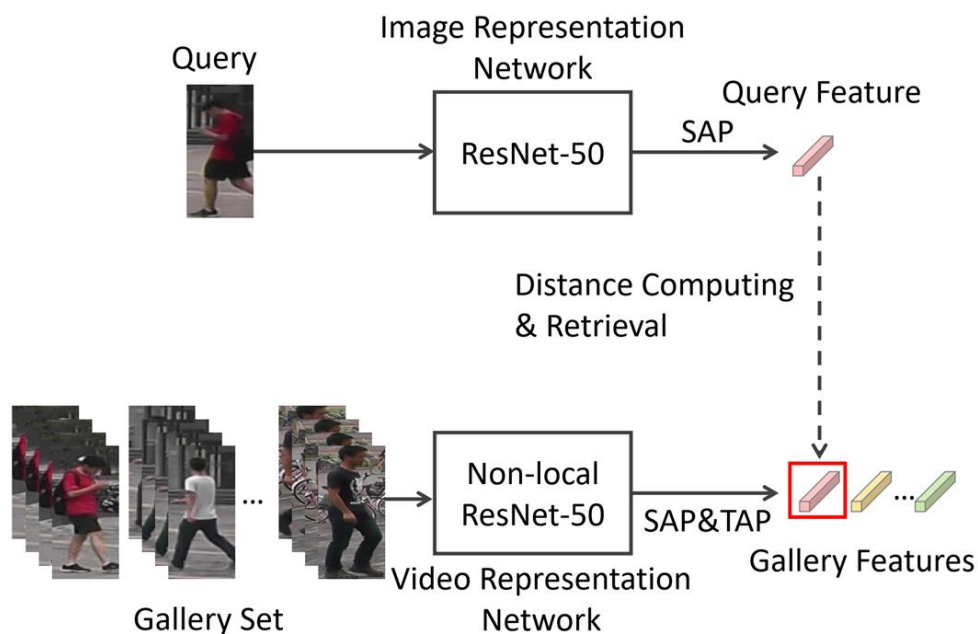
- Propagation via **cross sample distances**:

$$\mathcal{L}_{TKP}^D = \frac{1}{NT} \|D_{img} - D_{vid}\|_F^2$$

# Temporal Knowledge Propagation

- Testing pipeline of I2V Re\_ID

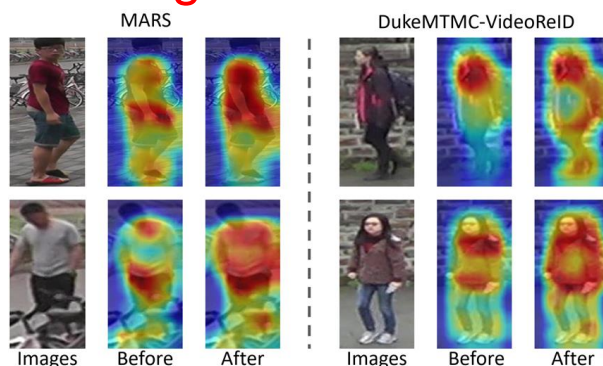
- SAT: spatial average pooling
- TAP: temporal average pooling



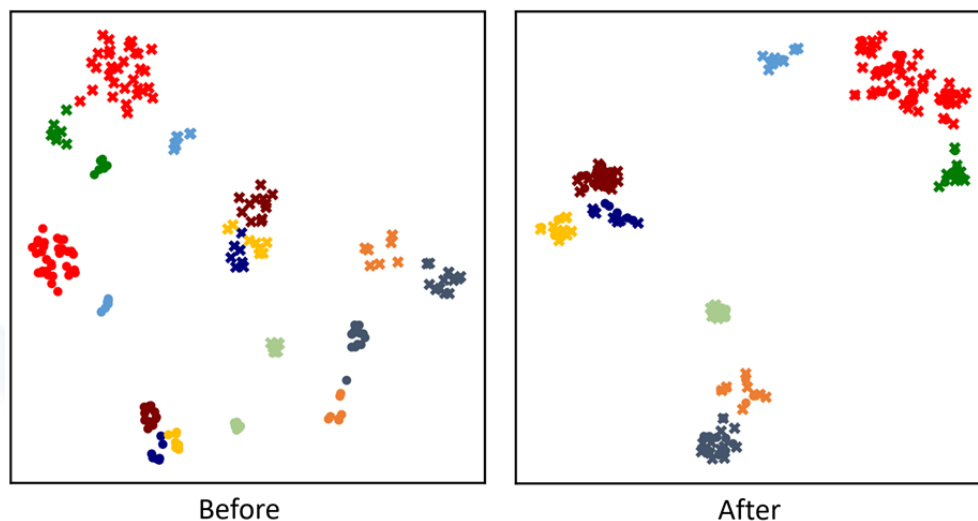
# Temporal Knowledge Propagation

## ● Visualization

- The learned **image features** focus on more foreground



- More consistent **feature distributions** of two modalities



# Temporal Knowledge Propagation

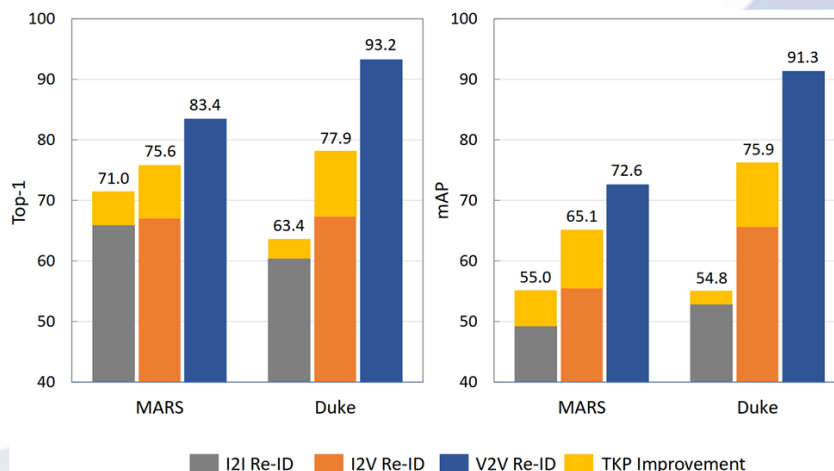
## ● Experimental results

I2V ReID on iLIDS-VID				
Models	top-1	top-5	top-10	top-20
PSDML [43]	13.5	33.8	45.6	56.3
LERM [14]	15.3	37.1	49.7	62.0
XQDA [21]	16.8	38.6	52.3	63.6
KISSME [18]	17.6	41.7	55.3	68.7
PHDL [44]	28.2	50.4	65.9	80.4
TMSL [37]	39.5	66.9	79.6	86.6
P2SNet [31]	40.0	68.5	78.1	90.0
<b>TKP</b>	<b>54.6</b>	<b>79.4</b>	<b>86.9</b>	<b>93.5</b>

I2V ReID on MARS				
Models	top-1	top-5	top-10	mAP
P2SNet [31]	55.3	72.9	78.7	-
Res50 [9]+XQDA [21]	67.2	81.9	86.1	54.9
<b>TKP</b>	<b>75.6</b>	<b>87.6</b>	<b>90.9</b>	<b>65.1</b>

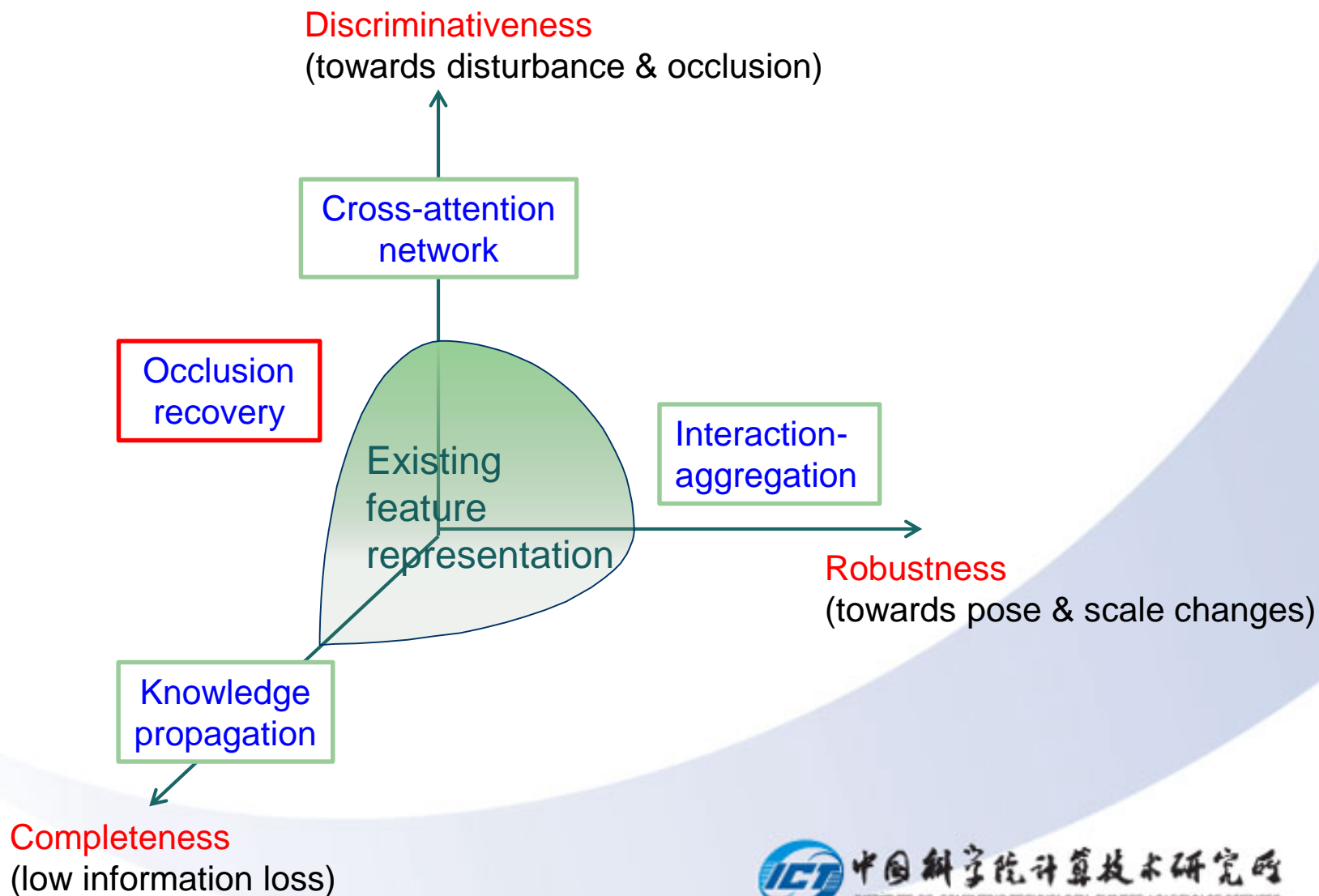
V2V ReID on MARS				
Models	top-1	top-5	top-10	mAP
SDM [38]	71.2	85.7	91.8	-
MGCAM [29]	77.2	-	-	71.2
DuATM [28]	78.7	90.9	-	62.3
multi-snippets [3]	81.2	92.1	-	69.4
DRSA [20]	82.3	-	-	65.8
<b>TKP</b>	<b>84.0</b>	<b>93.7</b>	<b>95.7</b>	<b>73.3</b>

Comparison among I2I, I2V and V2V ReID



[19] X. Gu, B. Ma, H. Chang, S. Shan, X. Chen, Temporal Knowledge Propagation for Image-to-Video Person Re-identification. In ICCV, 2019.

# Feature Representation for Person Re-ID

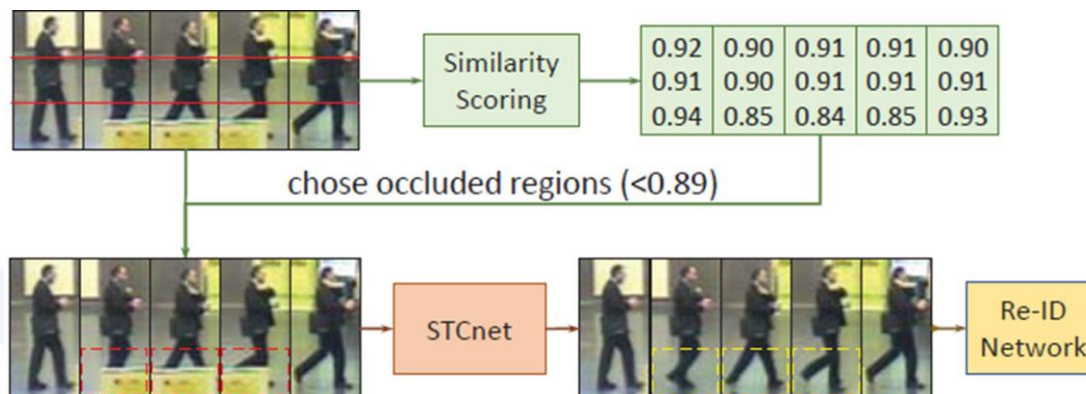


# Occlusion-free Video Re-ID

- Occlusion problem → information loss

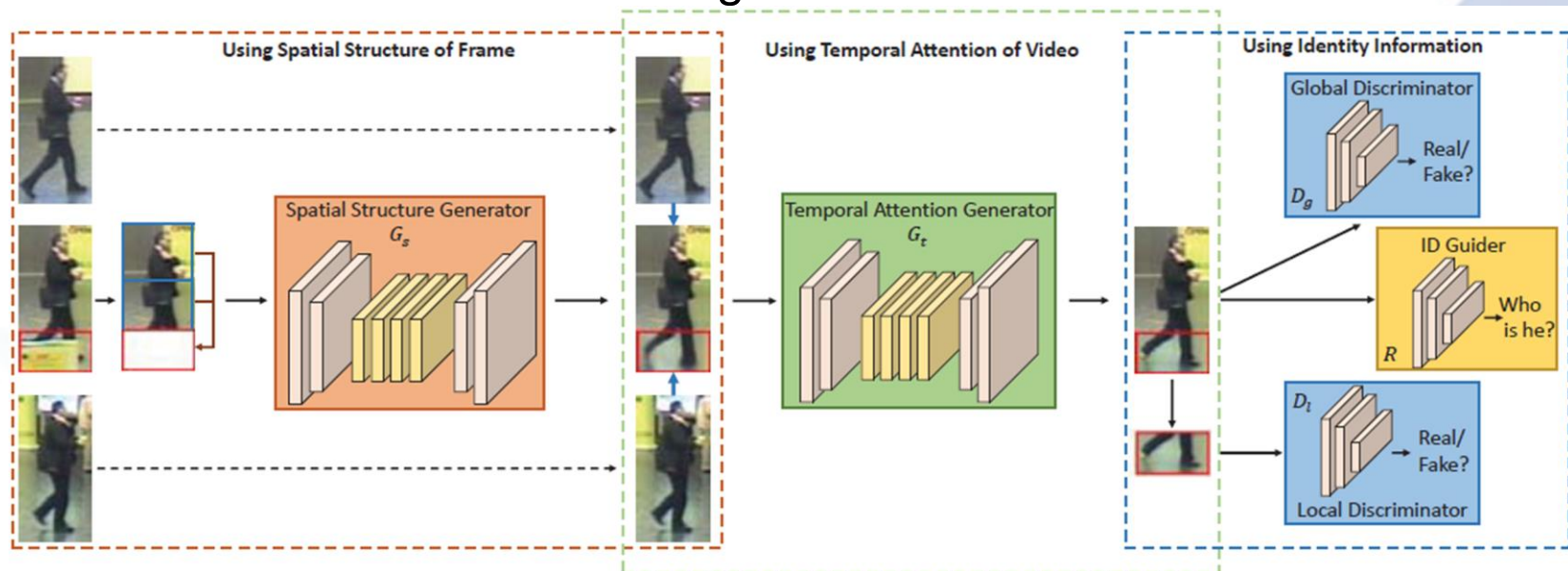


- Our solution: **explicitly recover** the appearance of the occluded parts
- Method overview
  - Similarity scoring mechanism: locate the occluded parts
  - **STCnet**: recover the appearance of the occluded parts



# Occlusion-free Video Re-ID

- Spatial-Temporal Completion network (STCnet)
  - **Spatial Structure Generator**: make a coarse prediction for occluded parts conditioned on the visible parts
  - **Temporal Attention Generator**: refine the occluded contents with temporal information
  - **Discriminator**: real or not?
  - **ID Guider**: classification target



# Occlusion-free Video Re-ID

## ● Visualization results



## ● Quantitative results

Ablation study

Methods	iLIDS	MARS	DukeMTMC
baseline	79.8	84.4 (77.2)	91.4 (90.0)
NL	80.1	86.1 (79.9)	91.8 (91.2)
VRSTC	<b>83.4</b>	<b>88.5 (82.3)</b>	<b>95.0 (93.5)</b>

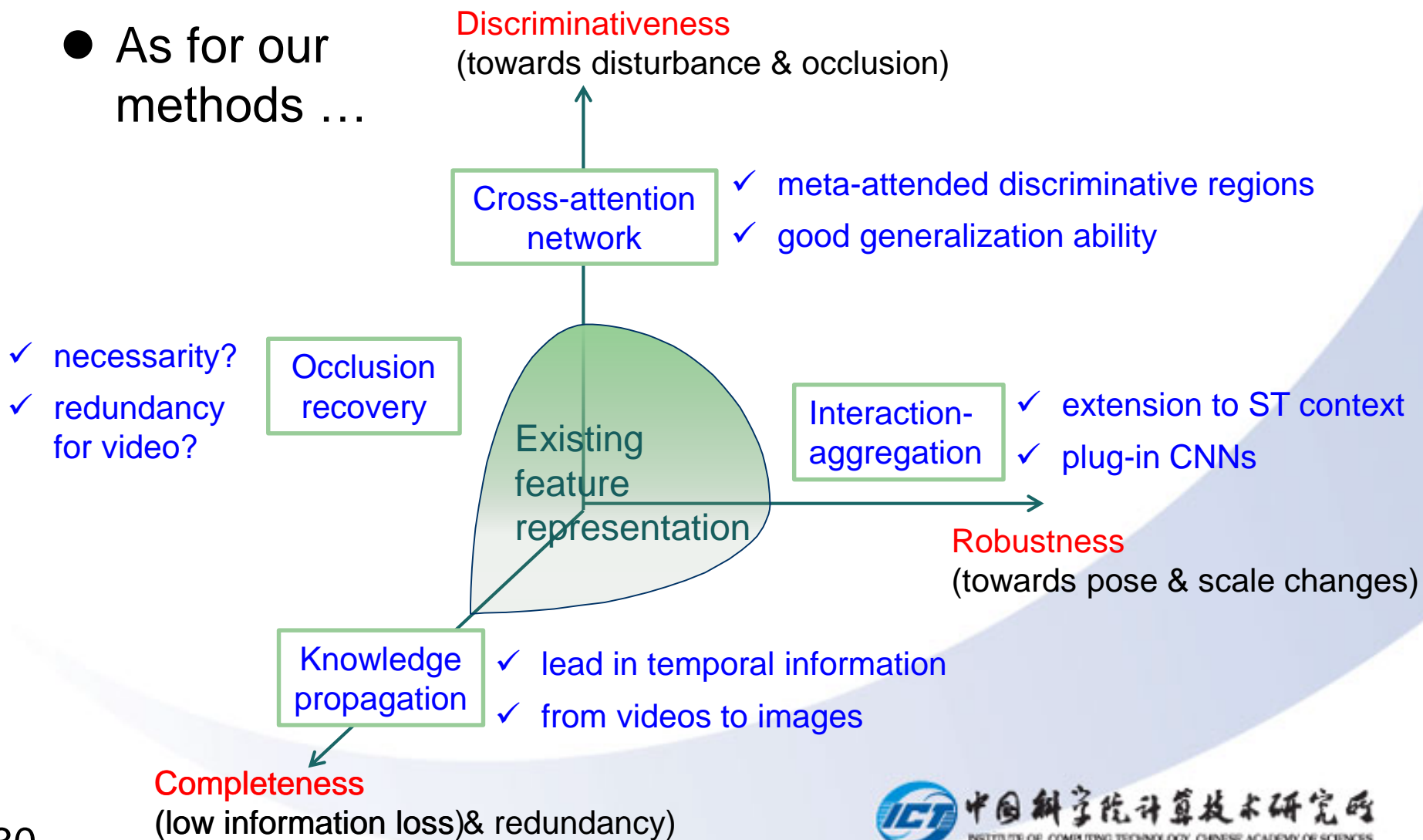
[20] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, VRSTC: Occlusion-free video person re-identification, in CVPR, 2019.

MARS

Methods	MARS			
	rank-1	rank-5	rank-10	mAP
QAN [18]	73.7	84.9	91.6	51.7
K-reciprocal [42]	73.9	-	-	68.5
RQEN [27]	77.8	88.8	94.3	71.7
TriNet [10]	79.8	91.4	-	67.7
EUG [31]	80.8	92.1	96.1	67.4
STAN [15]	82.3	-	-	65.8
Snipped [3]	81.2	92.1	-	69.4
Snippet+OF* [3]	86.3	94.7	<b>98.2</b>	76.1
VRSTC	<b>88.5</b>	<b>96.5</b>	97.4	<b>82.3</b>

# Discussions

- As for our methods ...

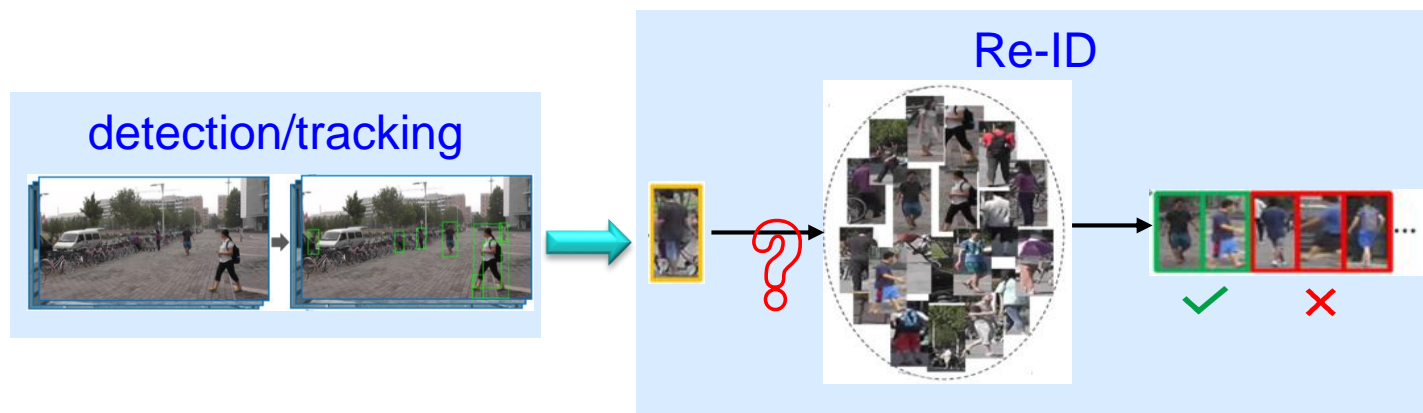


# Discussions

- Limitations in feature representation learning
  - For images, the discriminative ability is **upper bounded**
    - Appearance  $\{x_1, x_2, \dots, x_m\} \rightarrow$  Identity  $y$
    - Large appearance variation & little relation with identity, e.g., the same person with different clothes or accessories
    - Application: short term, restricted regions
  - For videos, **more discriminative spatial temporal features** are required
    - Key: temporal information representation
    - Other information: trajectory, other spatial temporal references
    - Application: more real-world scenarios

# Other Future Works

- Metric learning
  - coordinate with & complement to feature representation
- Person search
  - cooperation of detection/tracking and Re-ID



- Cross-modality person Re-ID
  - Image-to-Video
  - Person Question Answer

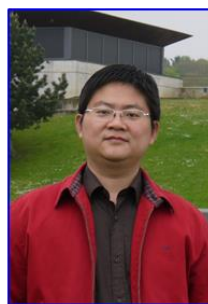
# References

- [1] R. R. Varior, B. Shuai, J. Lu, D. Xu, G. Wang. A siamese long short-term memory architecture for human re-identification. In ECCV, 2016.
- [2] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, J. Sun. Aligned ReID: Surpassing Human-Level Performance in Person Re-Identification. arXiv preprint arXiv:1711.08184.
- [3] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, R. Ji. Pyramidal person re-identification via multi-loss dynamic training. In CVPR, 2019.
- [4] D. Li, X. Chen, Z. Zhang. Learning deep context-aware features over body and latent parts for person re-identification. In CVPR, 2017.
- [5] L. Zhao, X. Li, J. Wang, Y. Zhuang. Deeply-learned part-aligned representations for person re-identification. In ICCV, 2017.
- [6] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), In ECCV, 2018.
- [7] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In CVPR, 2017.
- [8] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian. Glad: globallocal-alignment descriptor for pedestrian retrieval. In ACM, pages 420–428, 2017.
- [9] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, Human semantic parsing for person re-identification. In CVPR, 2018.
- [10] C. Song, Y. Huang, W. Ouyang, and L. Wang. Mask-guided contrastive attention model for person reidentification. In CVPR, pages 1179–1188, 2018.
- [11] Y. Liu, J. Yan, and W. Ouyang. Quality aware network for set to set recognition. In CVPR, 2017.
- [12] S. Li, Slawomir Bak, Peter Carr, Xiaogang Wang. Diversity Regularized Spatiotemporal Attention for Video-based Person Re-identification. In CVPR 18.

# References

- [13] J. Zhang, N. Wang and L. Zhang. Multi-shot Pedestrian Re-identification via Sequential Decision Making. In CVPR, 2018.
- [14] N. McLaughlin, J. M. del Rincon, and P. C. Miller. Recurrent convolutional network for video-based person reidentification. In CVPR, 2016.
- [15] D. Chen, H. Li, T. Xiao, S. Yi, X. Wang. Video Person Re-identification with Competitive Snippet-similarity Aggregation and Co-attentive Snippet Embedding. In CVPR, 2018.
- [16] X. Liao, L. He, Z. Yang. Video-based Person Re-identification via 3D Convolutional Networks and Non-local Attention. In ACCV, 2018.
- [17] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, X. Chen, Interaction-and-Aggregation Network for Person Re-identification. In CVPR, 2019.
- [18] R. Hou, H. Chang, B. Ma, S. Shan, X. Chen, Cross Attention Network for Few-shot Classification. In NeurIPS, 2019.
- [19] X. Gu, B. Ma, H. Chang, S. Shan, X. Chen, Temporal Knowledge Propagation for Image-to-Video Person Re-identification. In ICCV, 2019.
- [20] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, X. Chen, VRSTC: Occlusion-Free Video Person Re-Identification. In CVPR, 2019.

## Co-authors:



# Thanks!

Visual Information Processing and Learning (VIPL)

<http://vipl.ict.ac.cn>