



# Farewell to Mutual Information: Variational Distillation for Cross-Modal Person Re-identification

主讲人姓名：田旭东

主讲人简介：华东师范大学 2021 级博士研究生





## ➤ 研究团队概况



马利庄  
教授/博导  
国家杰青



王长波  
教授/博导  
学院书记



谢源  
教授/博导  
香江学者

## 华东师范大学多媒体与视觉实验室

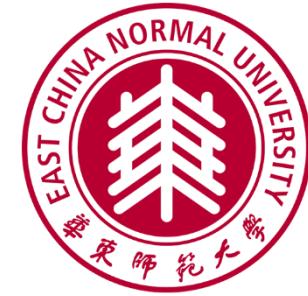
- 专任教师: 10+; 博士研究生: 10; 硕士研究生: 30+
- 团队研究方向: 1) 低层视觉; 2) 2D感知 (图像语义分割和跨模态行人重识别, 神经网络压缩和加速等); 3) 3D感知 (点云分类和分割); 4) 计算机图形学; 5) 模型压缩
- AI和CV顶会: 年均4~8篇; 顶刊: 年均3~5篇

形成了一系列**自有知识产权**的国际领先的科研成果



# 目 录

- 1 论文概述**
- 2 算法模型剖析**
- 3 代码复现**





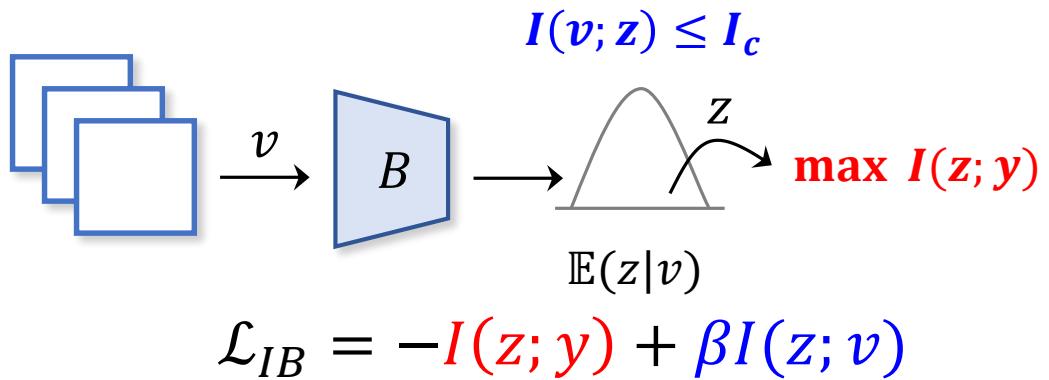
## 一、论文概述

- 研究背景
- 研究方法与成果



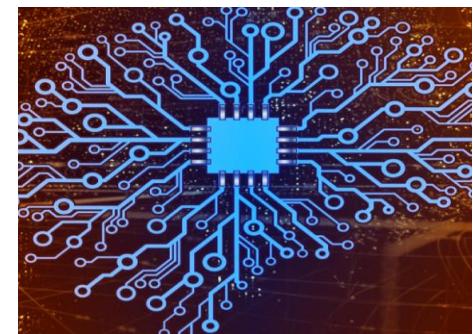
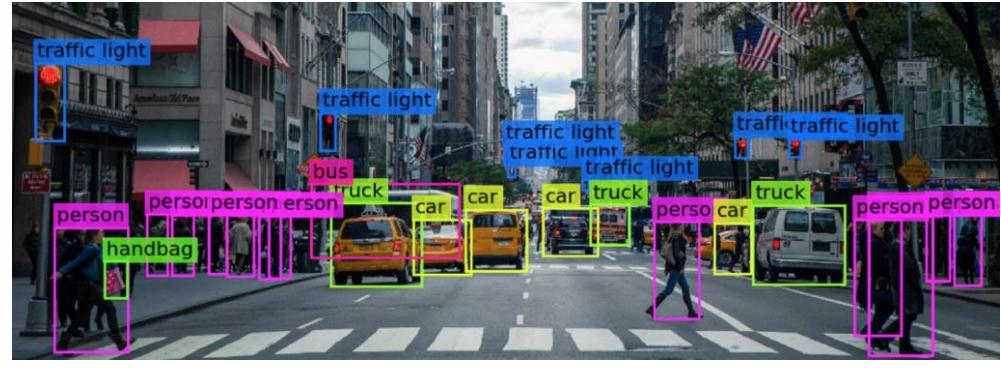
## What is information bottleneck principle?

- 核心目标：训练编码器以最大化保留对任务有帮助的信息，同时尽可能去除冗余信息[1]。



## Import applications

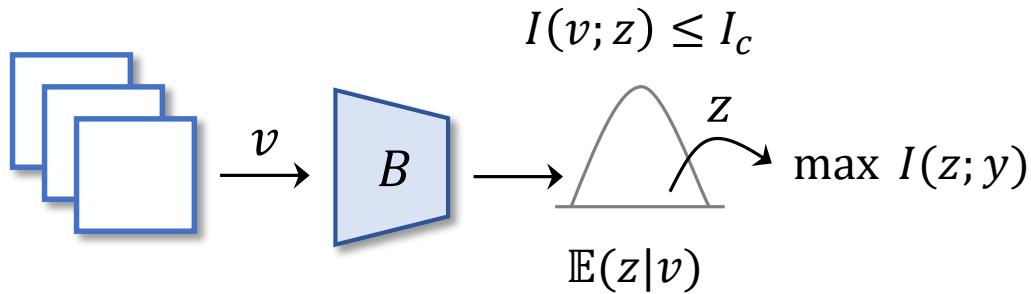
- 计算机视觉 [2]
- 自然语言处理 [3]
- 神经科学 [4]...



## The information bottleneck method (IB)

➤ 贡献:

- 提供了一种信息论指导下的表征学习方法



➤ 不足:

- 其有效性严重依赖互信息估算精度

$$\mathcal{L}_{IB} = -I(z; y) + \beta I(z; v)$$

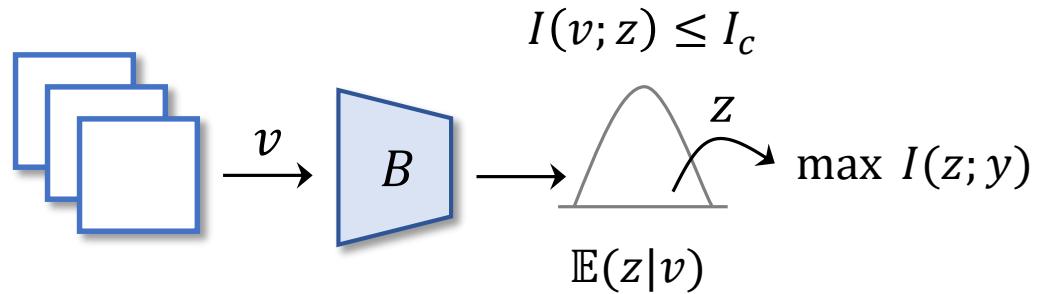
$$I(z; v) = \iint p(z|v) \tilde{p}(v) \log \frac{p(z|v)}{p(z)} dz dv$$

underlying distribution    latent variable space

## The information bottleneck method (IB)

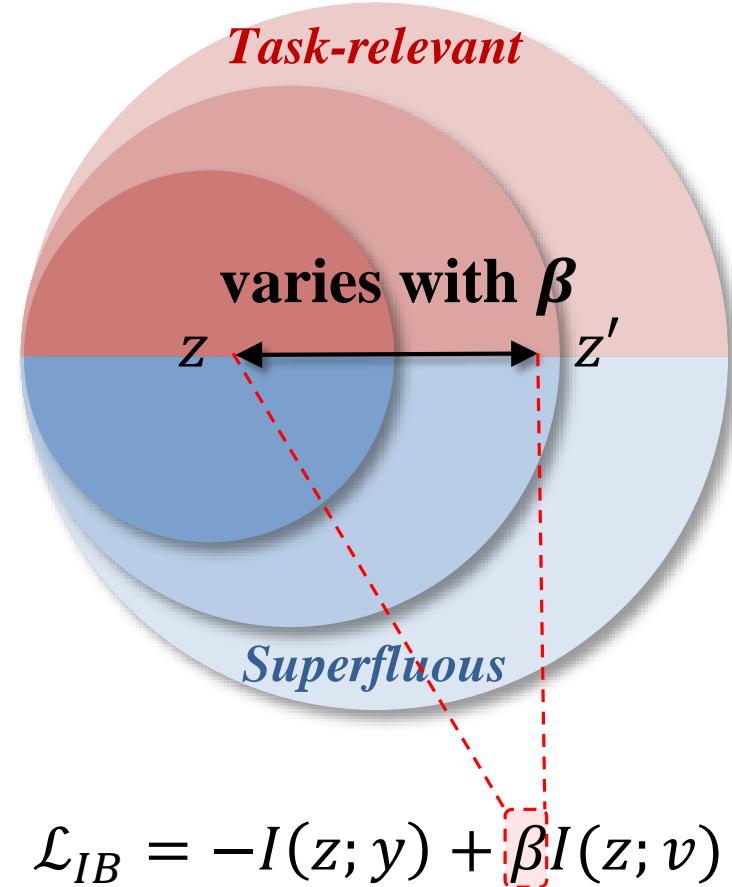
➤ 贡献:

- 提供了一种信息论指导下的表征学习方法



➤ 不足:

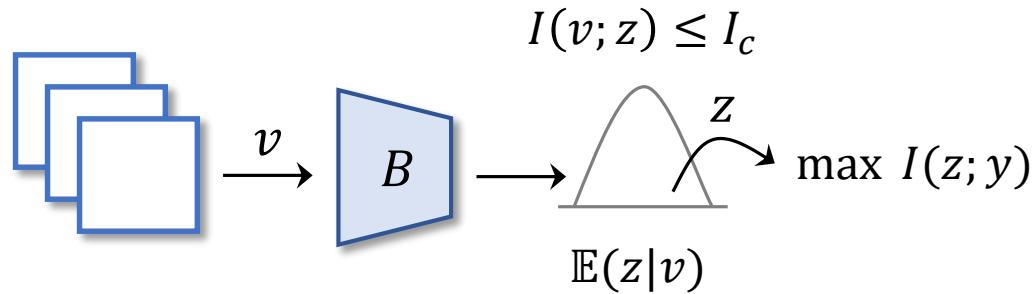
- 其有效性严重依赖互信息估算精度
- 预测性能与简洁性之间难以权衡



## The information bottleneck method (IB)

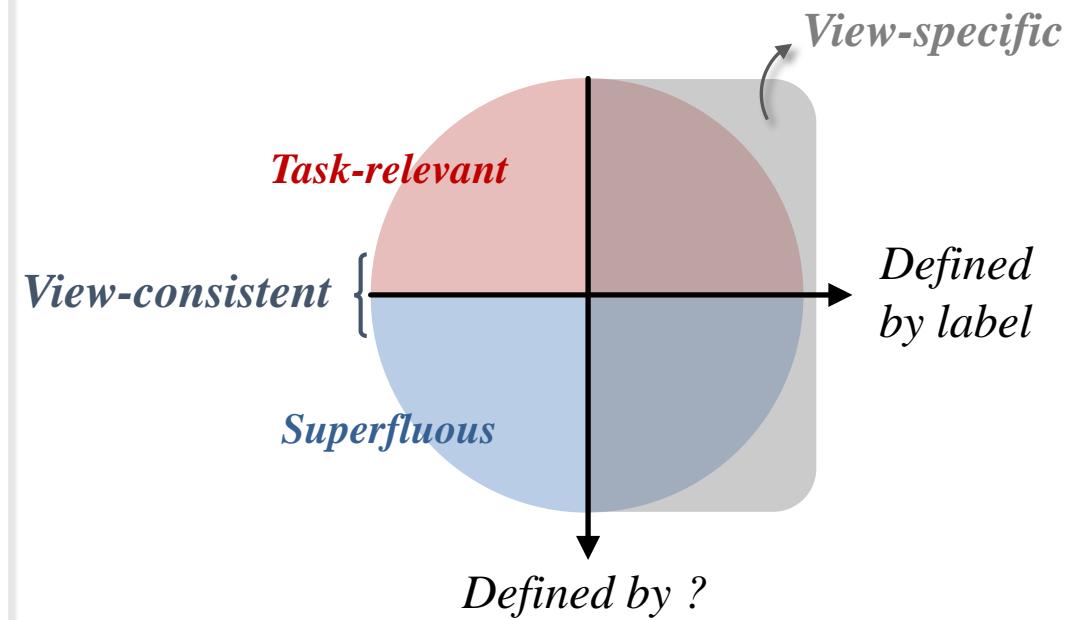
➤ 贡献:

- 提供了一种信息论指导下的表征学习方法



➤ 不足:

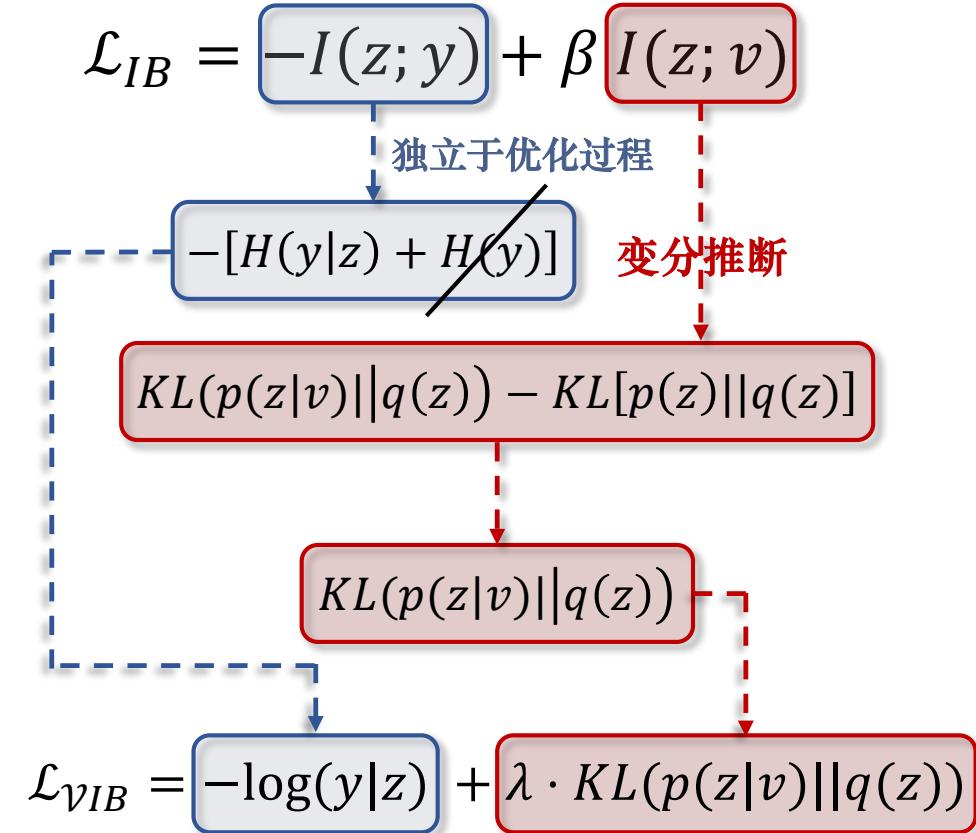
- 其有效性严重依赖互信息估算精度
- 预测性能与简洁性之间难以权衡
- 对多视图问题乏力



## Deep Variational Information Bottleneck (VIB)

➤ 贡献:

- 引入变分推断的思想，将互信息的优化转为熵的计算



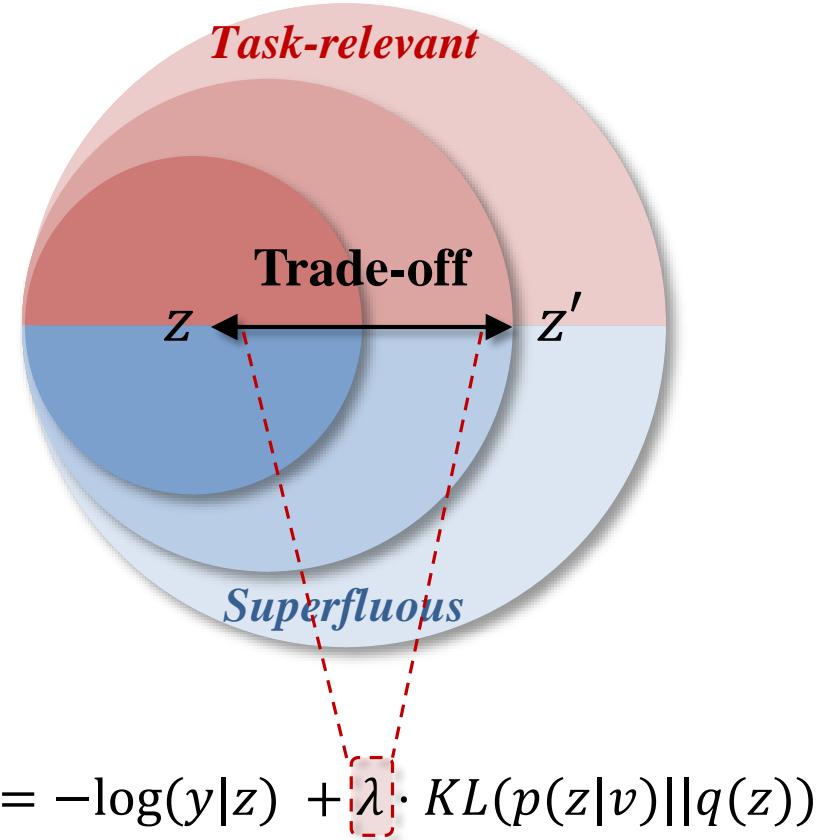
## Deep Variational Information Bottleneck (VIB)

➤ 贡献:

- 通过变分推断将  $I(z; v)$  转化为熵的计算

➤ 不足:

- 表征判别性能与简洁性之间的 trade-off 没得到解决



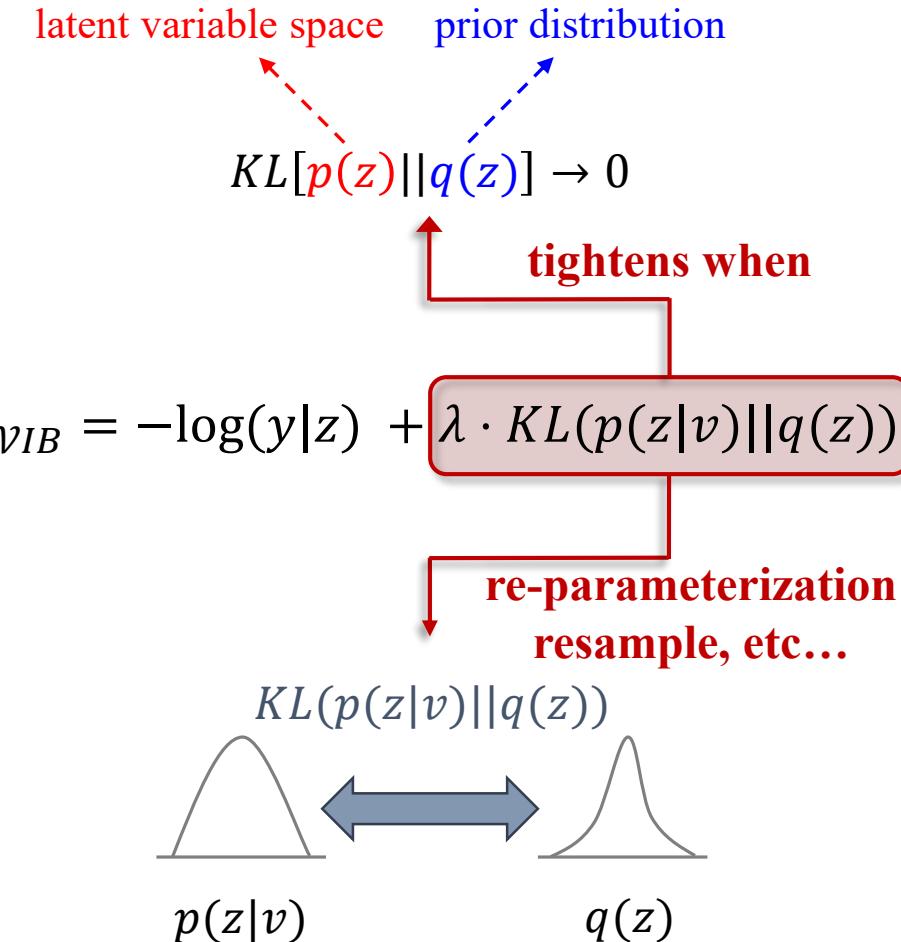
## Deep Variational Information Bottleneck (VIB)

➤ 贡献:

- 通过变分推断将  $I(z; v)$  转化为熵的计算

➤ 不足:

- 表征判别性能与简洁性之间的 trade-off 没得到解决
- 无法保证变分上界的有效性
- 涉及重参数、重采样等复杂操作





## 一、论文概述

- 研究背景
- 研究方法



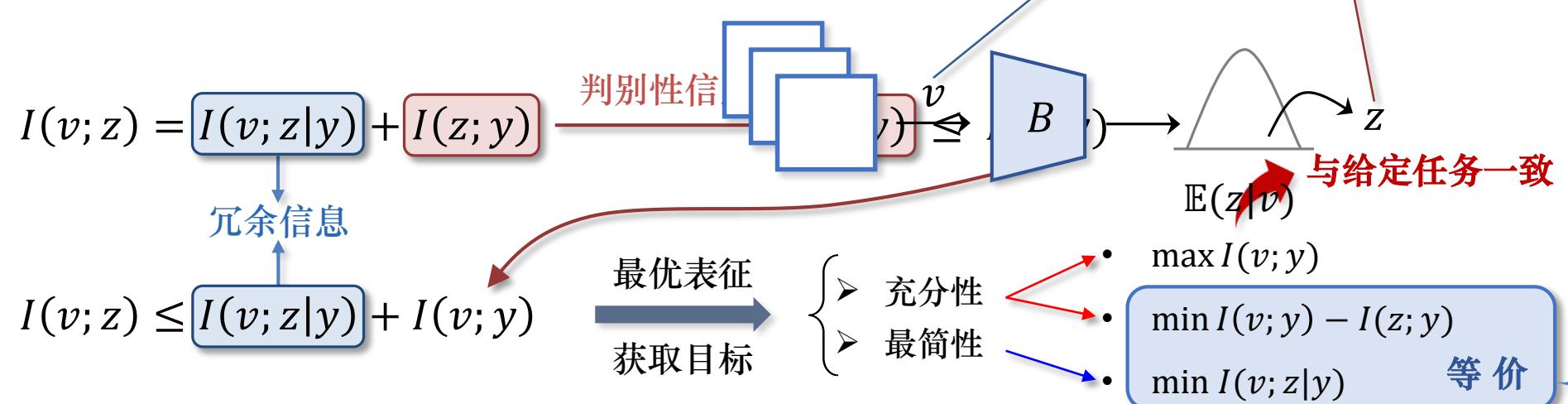
## 充分性

➤  $z$  包含所有关于  $y$  的判别性信息

## 定义

观察量  $v$  的表征  $z$  对于任务目标  $y$  是充分的，当且仅当  $I(v; y) = I(z; y)$ .

再无权衡难题



# 研究方法

## 定理一

- 最小化  $I(v; y) - I(z; y)$  等价于最小化  $v, z$  关于任务目标  $y$  条件熵的差值，即：

$$\min I(v; y) - I(z; y) \Leftrightarrow \min H(y|z) - H(y|v),$$

其中条件熵定义为  $H(y|z) := -\int p(z)dz \int p(y|z) \log p(y|z)dy.$

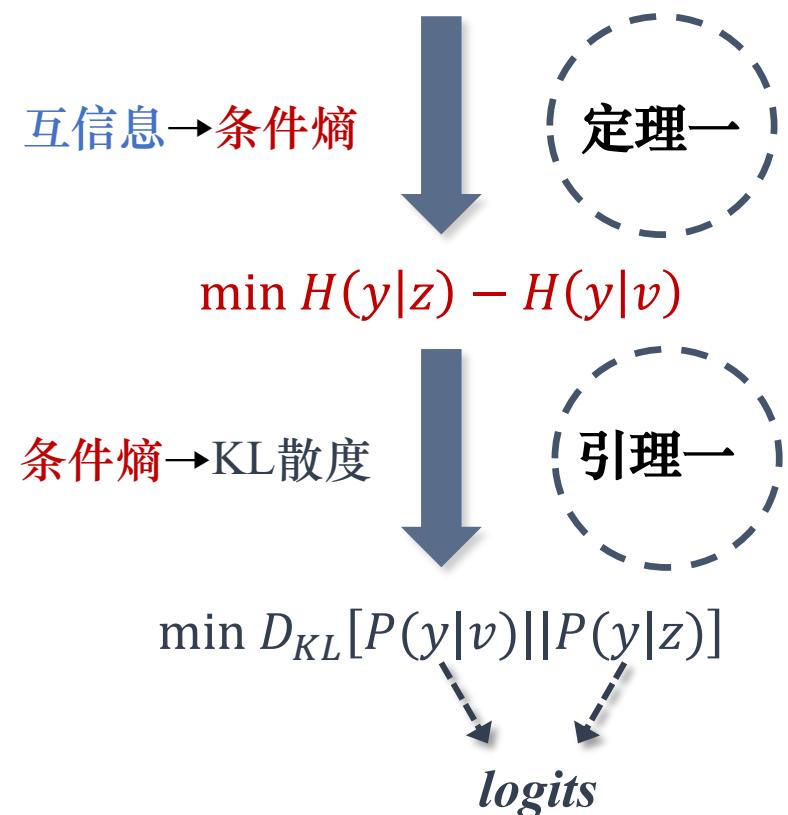
## 引理一

- 当表征  $z$  对任务目标  $y$  做出的预测与其观察量  $v$  的相同时，表征  $z$  对于任务目标  $y$  具备充分性，即：

$$D_{KL}[\mathbb{P}_v || \mathbb{P}_z] = 0 \Rightarrow H(y|z) - H(y|v),$$

其中  $\mathbb{P}_v = p(y|v)$  和  $\mathbb{P}_z = p(y|z)$  代表预测分布，且  $D_{KL}$  表示 KL 散度。

- $\min I(v; y) - I(z; y)$
- $\min I(v; z|y)$  等价



## 定理一

- 最小化  $I(v; y) - I(z; y)$  等价于最小化  $v, z$  关于任务目标  $y$  条件熵的差值，即：

$$\min I(v; y) - I(z; y) \Leftrightarrow \min H(y|z) - H(y|v),$$

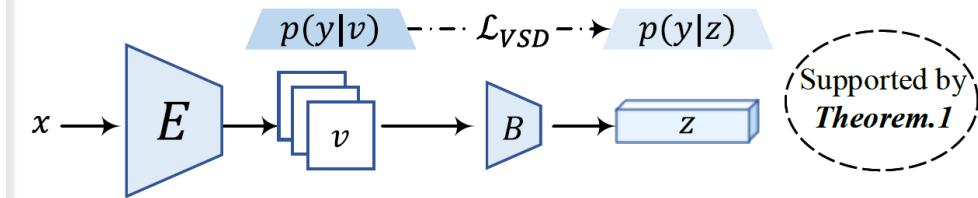
其中条件熵定义为  $H(y|z) := -\int p(z)dz \int p(y|z) \log p(y|z)dy$ .

## 引理一

- 当表征  $z$  对任务目标  $y$  做出的预测与其观察量  $v$  的相同时，表征  $z$  对于任务目标  $y$  具备充分性，即：

$$D_{KL}[\mathbb{P}_v || \mathbb{P}_z] = 0 \Rightarrow H(y|z) - H(y|v),$$

其中  $\mathbb{P}_v = p(y|v)$  和  $\mathbb{P}_z = p(y|z)$  代表预测分布，且  $D_{KL}$  表示 KL 散度.



$$\mathcal{L}_{VSD} = \min_{\theta, \phi} \mathbb{E}_{v \sim E_\theta(v|x)} [\mathbb{E}_{z \sim E_\phi(z|v)} [D_{KL}[\mathbb{P}_v || \mathbb{P}_z]]]$$

- 变分自蒸馏 (Variational Self-Distillation, VSD):
  - 无需进行互信息估算且更精确地拟合
  - 解决优化时的权衡难题
  - 不涉及重参数、采样等繁琐操作

## Consistency

- 仅保存判别性且满足视图间一致性的信息，以增强表征对于视图变化的鲁棒性

### 定义

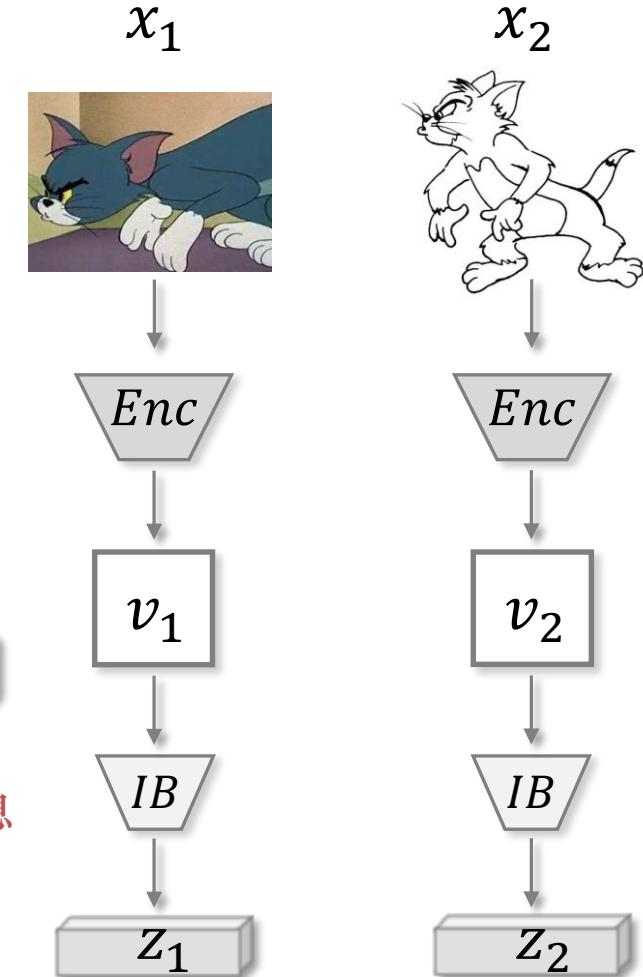
表征  $z_1, z_2$  满足视图间一致性，当且仅当  $I(z_1; y) = I(v_1 v_2; y) = I(z_2; y)$ .

### 定理二

给定两个满足充分性的观察量  $v_1, v_2$ ，其对应的表征  $z_1$  和  $z_2$  满足视图间一致性，当且仅当满足下面条件：

$$I(v_1; z_1 | v_2) + I(v_2; z_2 | v_1) \leq 0 \text{ and } I(v_2; v_1 | y) + I(v_1; v_2 | y) \leq 0$$

最小化                    最大化



## 消除视图特异性信息

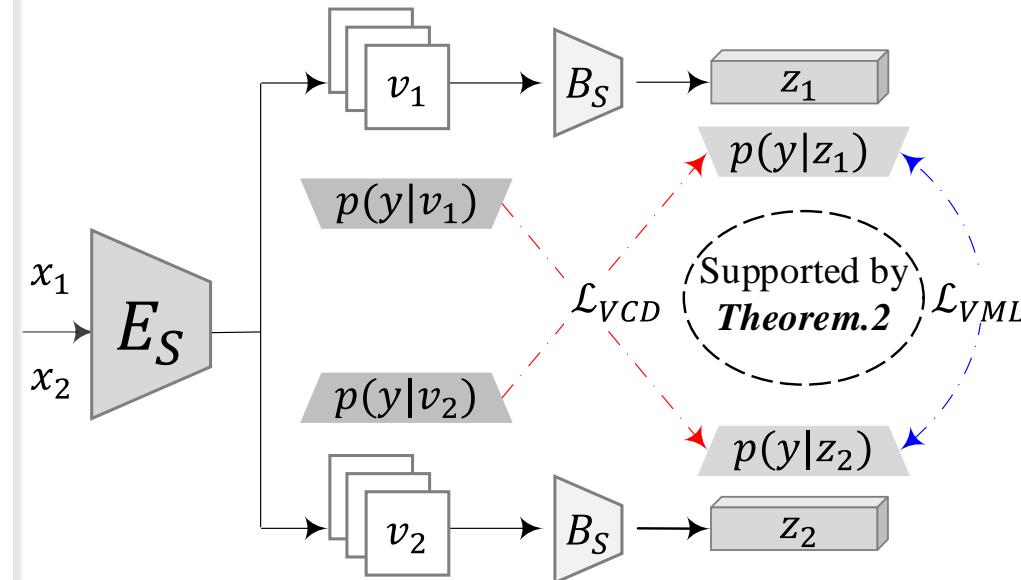
- 变分互学习 (*Variational Mutual Learning, VML*) : 最小化  $z_1, z_2$  预测分布之间的JS散度以消除其所包含的视图特异性信息, 具体目标如下,

$$\mathcal{L}_{VML} = \min_{\theta, \phi} \mathbb{E}_{v_1, v_2 \sim E_\theta(v|x)} \mathbb{E}_{z_1, z_2 \sim E_\phi(z|v)} [D_{JS}[\mathbb{P}_{z_1} || \mathbb{P}_{z_2}]]$$

## 消除冗余信息

- 变分交叉蒸馏 (*Variational Cross-Distillation, VCD*) : 在留存的视图一致性信息中, 通过交叉地优化观察量与不同视图表征之间的KL散度提纯判别性信息, 同时剔除冗余信息, 具体目标如下 ( $v_1$  与  $z_1$  同理) :

$$\mathcal{L}_{VCD} = \min_{\theta, \phi} \mathbb{E}_{v_1, v_2 \sim E_\theta(v|x)} \mathbb{E}_{z_1, z_2 \sim E_\phi(z|v)} [D_{KL}[\mathbb{P}_{v_2} || \mathbb{P}_{z_1}]]$$



## 消除视图特异性信息

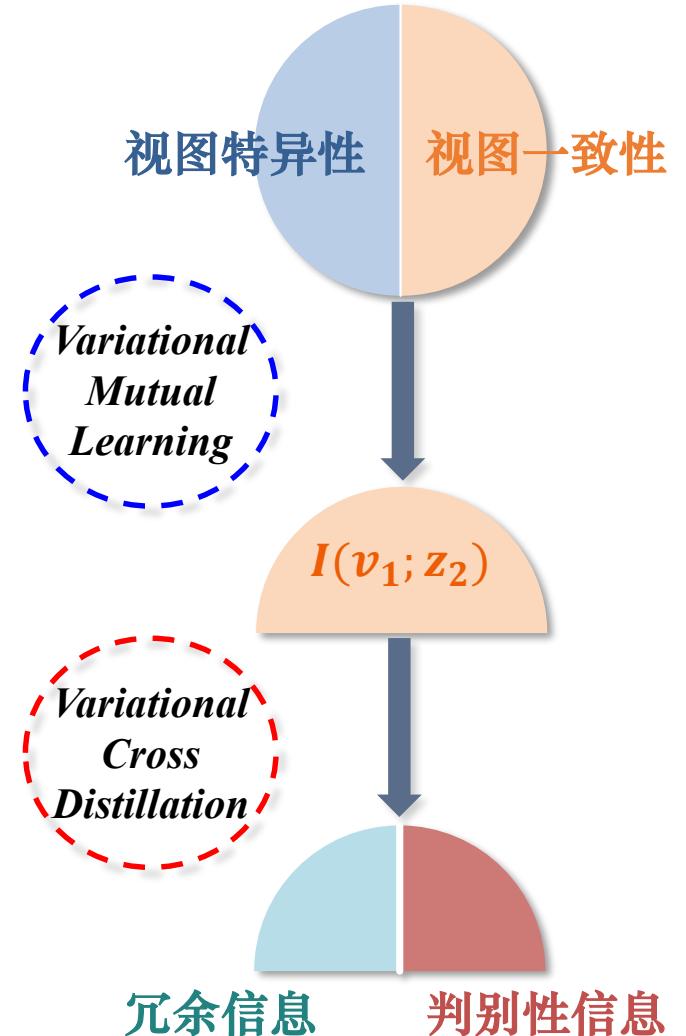
- 变分互学习 (*Variational Mutual Learning, VML*) : 最小化  $z_1, z_2$  预测分布之间的JS散度以消除其所包含的视图特异性信息, 具体目标如下,

$$\mathcal{L}_{VML} = \min_{\theta, \phi} \mathbb{E}_{v_1, v_2 \sim E_\theta(v|x)} \mathbb{E}_{z_1, z_2 \sim E_\phi(z|v)} [D_{JS}[\mathbb{P}_{z_1} || \mathbb{P}_{z_2}]]$$

## 消除冗余信息

- 变分交叉蒸馏 (*Variational Cross-Distillation, VCD*) : 在留存的视图一致性信息中, 通过交叉地优化观察量与不同视图表征之间的KL散度提纯判别性信息, 同时剔除冗余信息, 具体目标如下 ( $v_1$  与  $z_1$  同理) :

$$\mathcal{L}_{VCD} = \min_{\theta, \phi} \mathbb{E}_{v_1, v_2 \sim E_\theta(v|x)} \mathbb{E}_{z_1, z_2 \sim E_\phi(z|v)} [D_{KL}[\mathbb{P}_{v_2} || \mathbb{P}_{z_1}]]$$





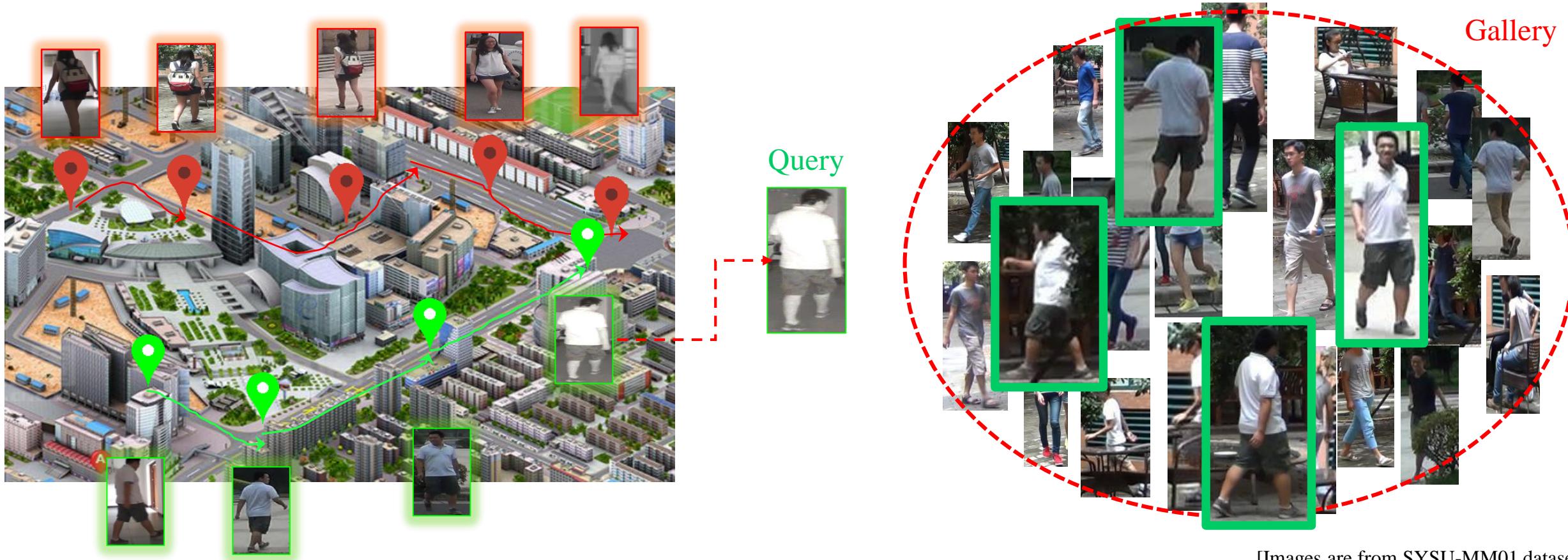
## 二、算法模型剖析

- 框架总览
- 实验标准
- 结果分析



## 跨模态行人重识别

- 给定红外或可见光照下的人像，跨模态行人重识别旨在找出相同目标在另一种模态（即光照）下的图像。



# 框架总览

## 模型结构总览

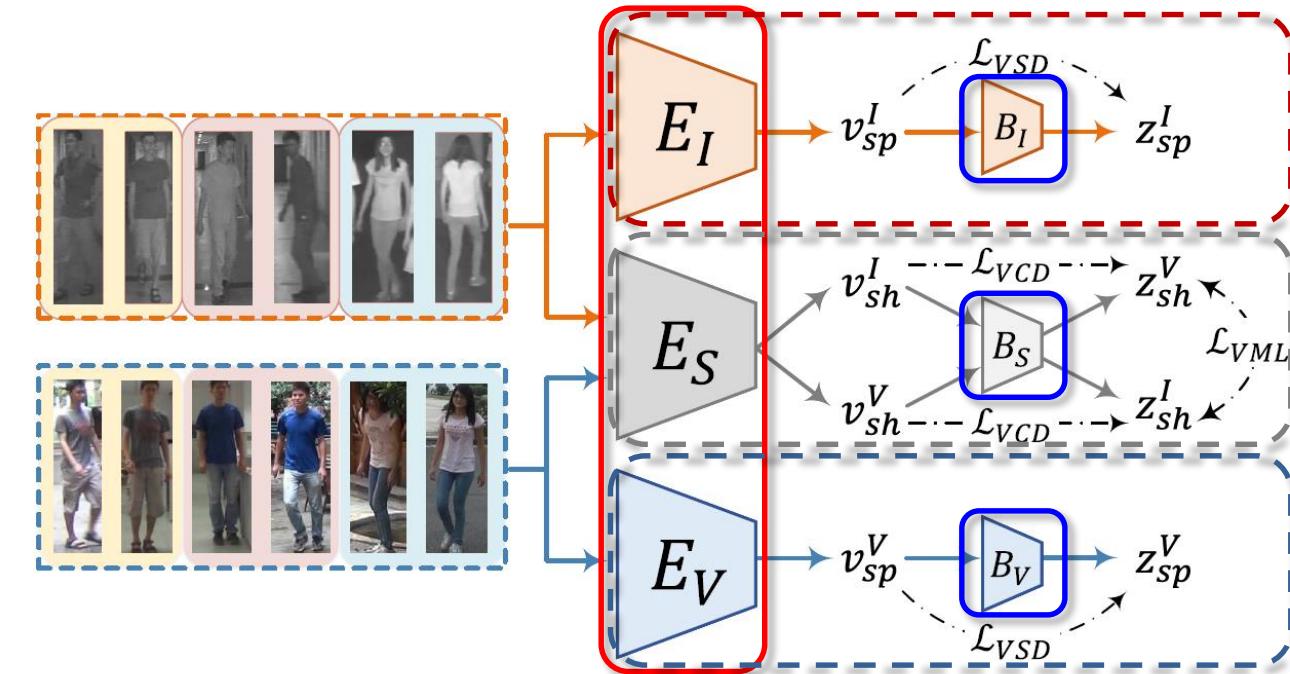
- 模型共包含三条独立分支，且每条分支仅包含一个编码器以及一个信息瓶颈。

## 损失函数总览

- 损失函数由两部分构成，即论文中提出的变分蒸馏，以及 Re-ID 最常用的训练约束。注意 VSD 只约束单模态分支，而 VCD 协同 VML 一起约束跨模态分支。

$$\mathcal{L}_{train} = \mathcal{L}_{ReID} + \beta \cdot (\mathcal{L}_{VSD} + \mathcal{L}_{VCD} + \mathcal{L}_{VML})$$

$$\mathcal{L}_{ReID} = \mathcal{L}_{cls} + \mathcal{L}_{metric} + \alpha \cdot \mathcal{L}_{DML}$$



$E_{I \setminus S \setminus V}$ : implemented with **ResNet-50**.

$B_{I \setminus S \setminus V}$ : implemented with **multi-layer perceptron** of 2 hidden ReLU units of size 1024 and 512 respectively.



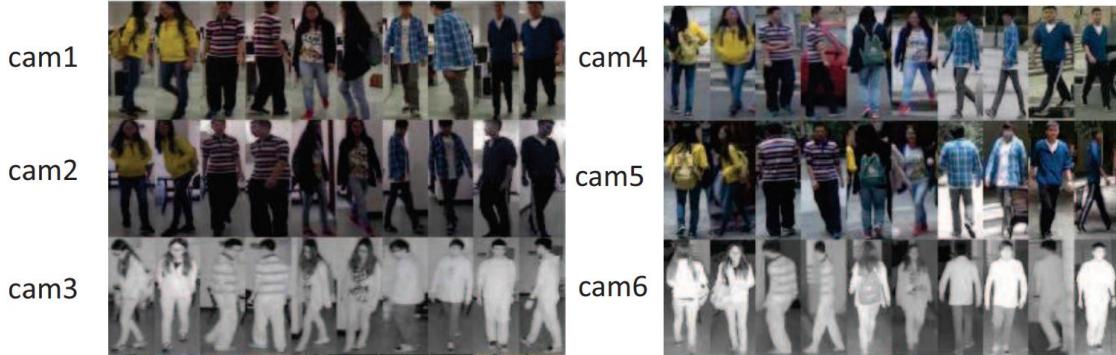
## 二、算法模型剖析

- 框架总览
- 实验标准
- 结果分析



## SYSU-MM01

- 数据集共包括 491 个目标的 287,628 张可见光图像以及 15,792 张红外光图像。每个目标的图像都来源于 6 个不重叠摄像头分别在室内和户外进行拍摄的拍摄结果。
- 评测标准包含全场景查询（*all-search*）和室内查询（*indoor-search*）。论文中所有实验结果都采用标准评测准则。



## RegDB

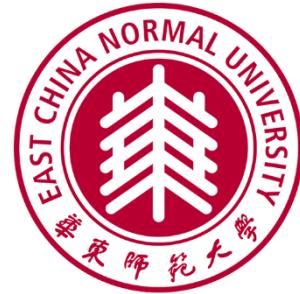
- 数据集共包括 412 个目标，且每个目标对应十张在同一时刻拍摄的可见光图像以及红外光图像。
- 评测标准包括可见光搜红外（*visible-to-infrared*）以及红外搜可见光（*infrared-to-visible*）。最终评测结果为十次实验的平均精度，且每次实验都开展于随机划分的评估集。





## 二、算法模型剖析

- 框架总览
- 实验方法
- 结果分析



Settings			All Search				Indoor Search			
Type	Method	Venue	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
Network Design	Zero-Pad [41]	ICCV'17	14.80	54.12	71.33	15.95	20.58	68.38	85.79	26.92
Metric Design	TONE [43]	AAAI'18	12.52	50.72	68.60	14.42	20.82	68.86	84.46	26.38
Metric Design	HCML [43]	AAAI'18	14.32	53.16	69.17	16.16	24.52	73.25	86.73	30.08
Metric Design	BDTR [45]	IJCAI'18	17.01	55.43	71.96	19.66	-	-	-	-
Network Design	cmGAN [4]	IJCAI'18	26.97	67.51	80.56	31.49	31.63	77.23	89.18	42.19
Metric Design	D-HSME [11]	AAAI'18	20.68	32.74	77.95	23.12	-	-	-	-
Generative	D <sup>2</sup> LR [40]	CVPR'19	28.9	70.6	82.4	29.2	-	-	-	-
Metric Design	MAC [42]	MM'19	33.26	79.04	90.09	36.22	36.43	62.36	71.63	37.03
Generative	AlignGAN [36]	ICCV'19	42.4	85.0	93.7	40.7	45.9	87.6	94.4	54.3
Generative	X-modal [17]	AAAI'20	49.92	89.79	95.96	50.73	-	-	-	-
Generative	JSIA-ReID [37]	AAAI'20	38.1	80.7	89.9	36.9	52.9	43.8	86.2	94.2
Network Design	cm-SSFT [19]	CVPR'20	52.4	-	-	52.1	-	-	-	-
Network Design	Hi-CMD [3]	CVPR'20	34.94	77.58	-	35.94	-	-	-	-
Network Design	DDAG [44]	ECCV'20	<b>54.75</b>	<b>90.39</b>	<b>95.81</b>	<b>53.02</b>	<b>61.02</b>	<b>94.06</b>	<b>98.41</b>	<b>67.98</b>
Representation	ours	-	<b>60.02</b>	<b>94.18</b>	<b>98.14</b>	<b>58.80</b>	<b>66.05</b>	<b>96.59</b>	<b>99.38</b>	<b>72.98</b>

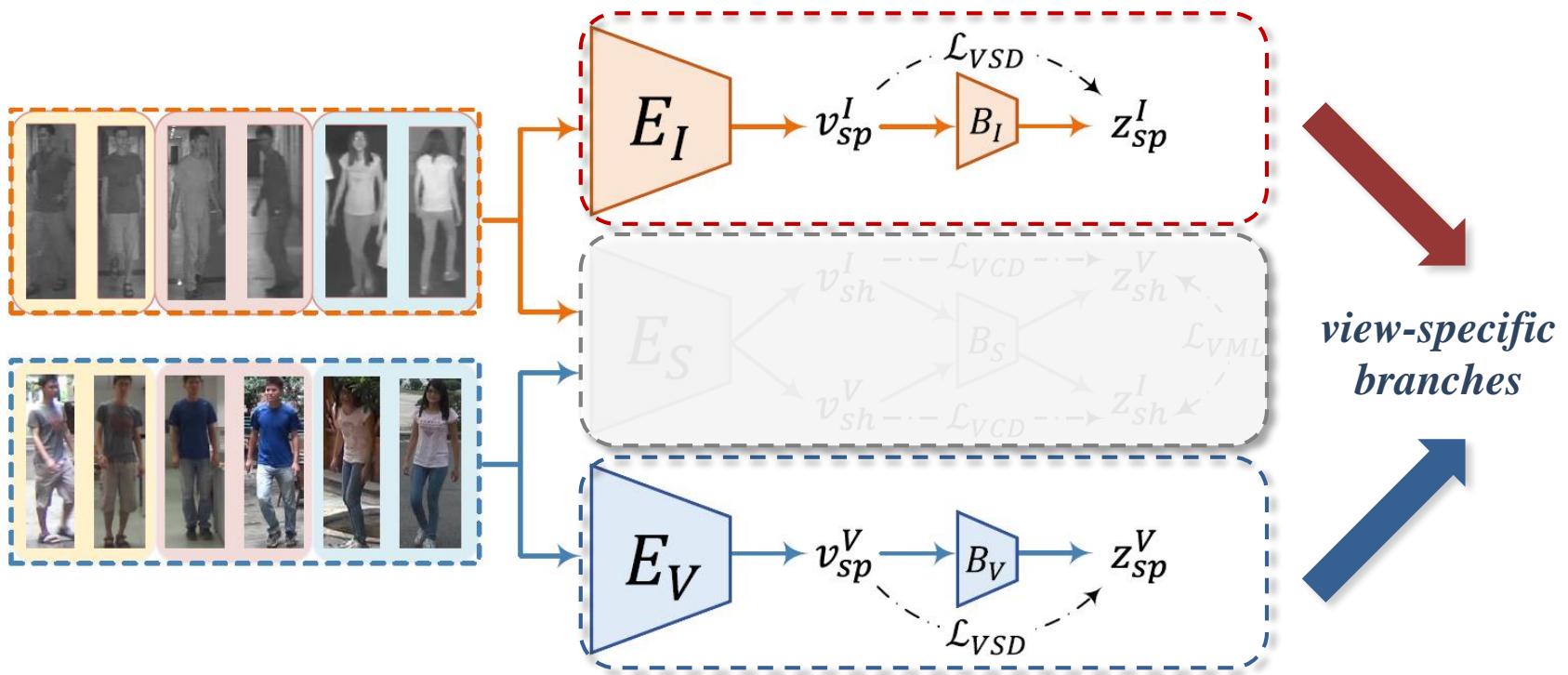
- Performance of the proposed method compared with the state-of-the-arts. Note all methods are measured by CMC and mAP on **SYSU-MM01** under **single-shot** mode.

Settings		Visible to Thermal		Thermal to Visible	
Method	Venue	Rank-1	mAP	Rank-1	mAP
Zero-Pad [41]	ICCV'17	17.8	18.9	16.6	17.8
HCML [43]	AAAI'18	24.4	20.1	21.7	22.2
BDTR [45]	IJCAI'18	33.6	32.8	32.9	32.0
D-HSME [11]	AAAI'18	50.8	47.0	50.2	46.2
D <sup>2</sup> LR [40]	CVPR'19	43.4	44.1	-	-
MAC [42]	MM'19	36.4	37.0	36.2	36.6
AlignGAN [36]	ICCV'19	57.9	53.6	56.3	53.4
X-modal [17]	AAAI'20	62.2	60.2	-	-
JSIA-ReID [37]	AAAI'20	48.5	49.3	48.1	48.9
cm-SSFT [19]	CVPR'20	62.2	63.0	-	-
Hi-CMD [3]	CVPR'20	<b>70.9</b>	<b>66.0</b>	-	-
DDAG [44]	ECCV'20	69.3	63.5	<b>68.1</b>	<b>61.8</b>
ours	-	<b>73.2</b>	<b>71.6</b>	<b>71.8</b>	<b>70.1</b>

- Comparison with the state-of-the-arts on *RegDB* under *visible-thermal* and *thermal-visible* settings.

## 消融实验：

- 单一模态分支条件下，变分蒸馏 vs 信息瓶颈



Variational  
Self-Distillation

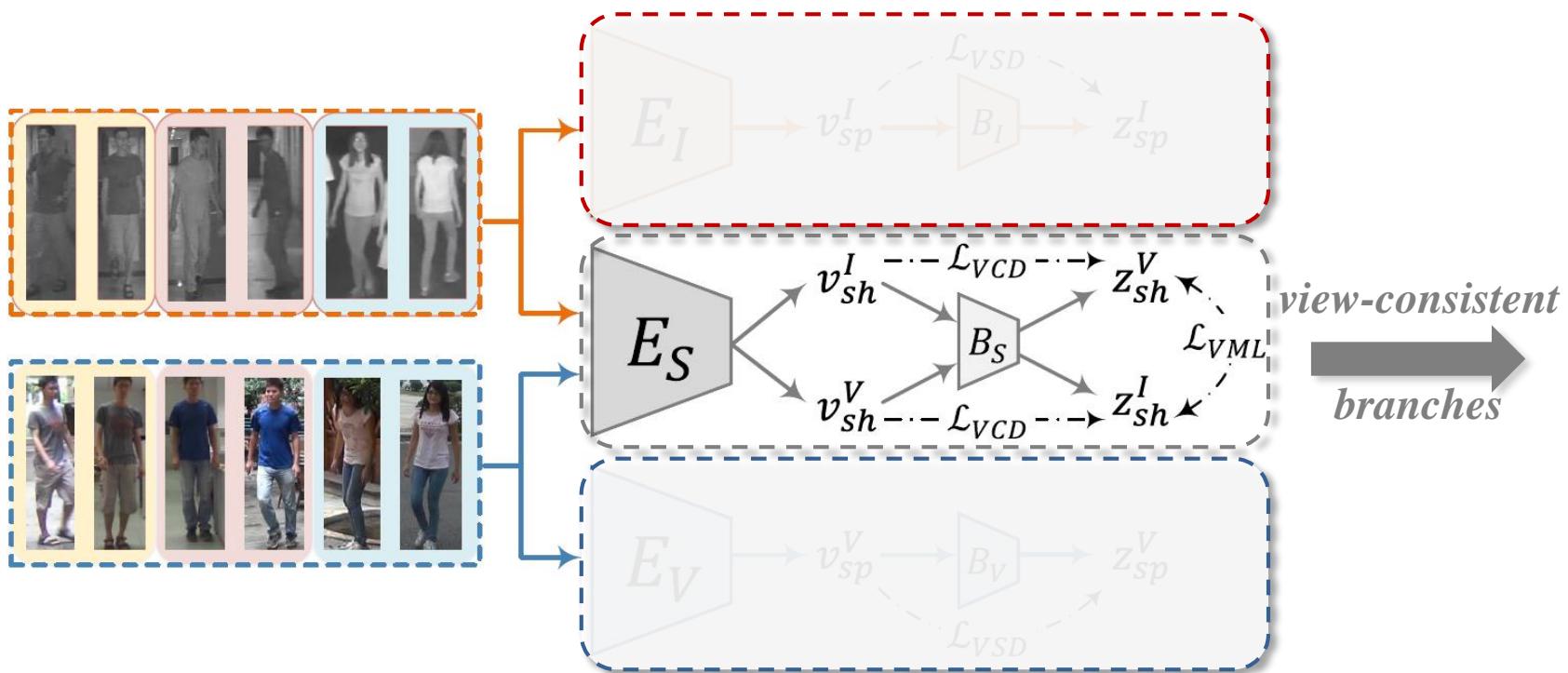
Rank-1	mAP
<b>59.62</b>	<b>57.99</b>

Information  
Bottleneck

Rank-1	mAP
<b>28.69</b>	<b>32.42</b>

## 消融实验:

- 多模态分支条件下，变分蒸馏 vs 信息瓶颈



**Variational Cross-Distillation**

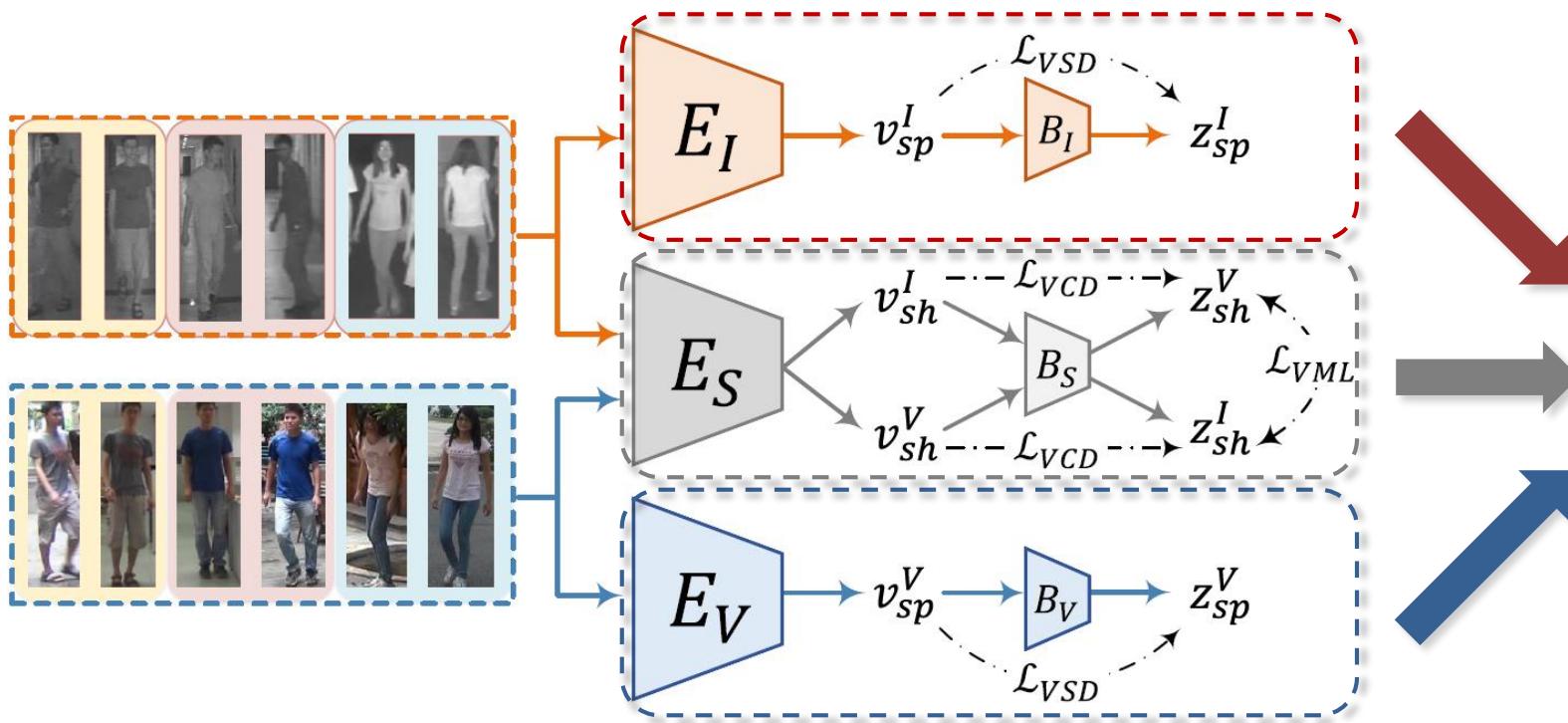
Rank-1	mAP
<b>49.22</b>	<b>48.74</b>

**Information Bottleneck**

Rank-1	mAP
<b>24.34</b>	<b>28.01</b>

## 消融实验：

- 三分支条件下，变分蒸馏 vs 信息瓶颈



Variational  
Distillation

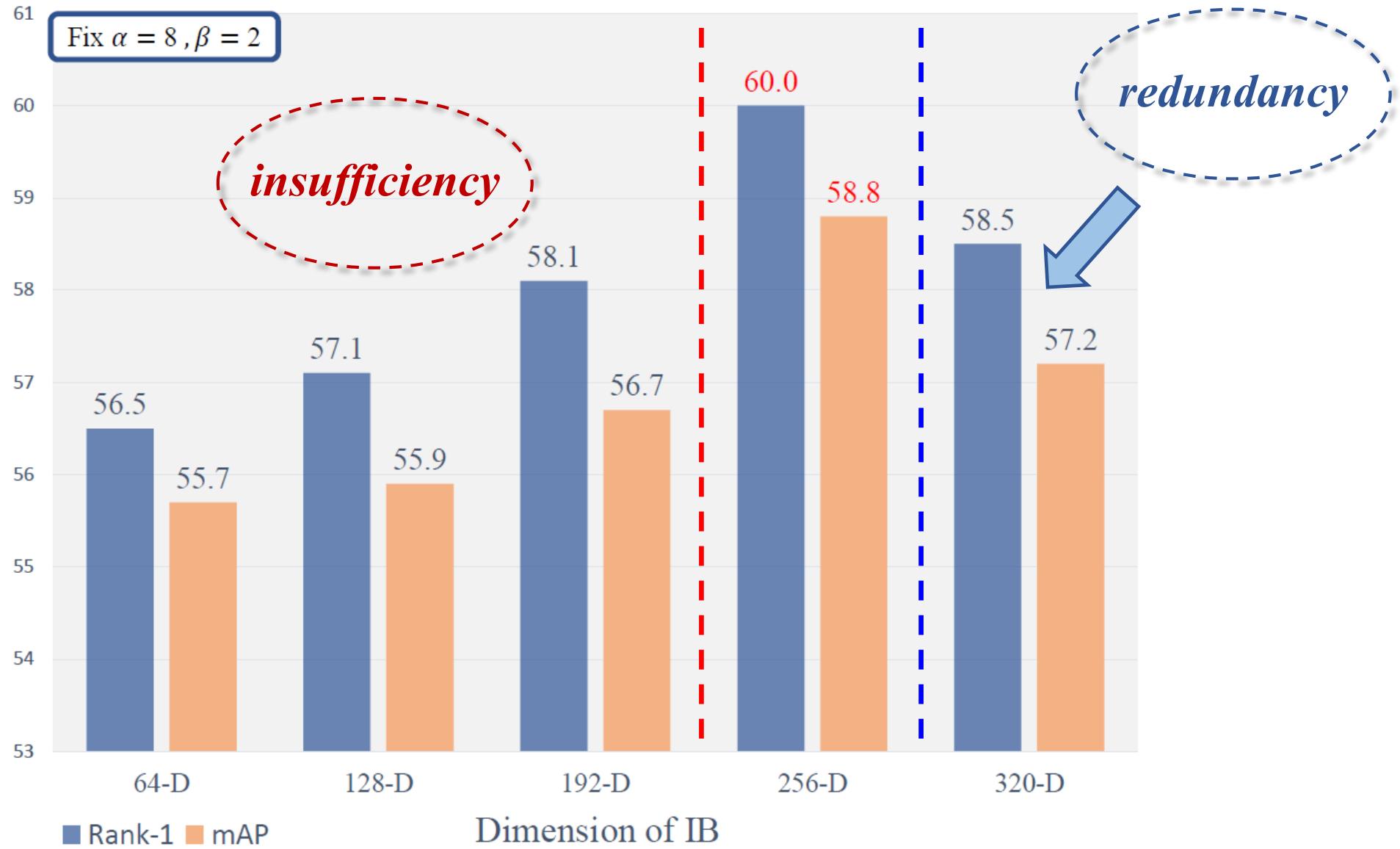
Rank-1	mAP
60.02	58.80

Information  
Bottleneck

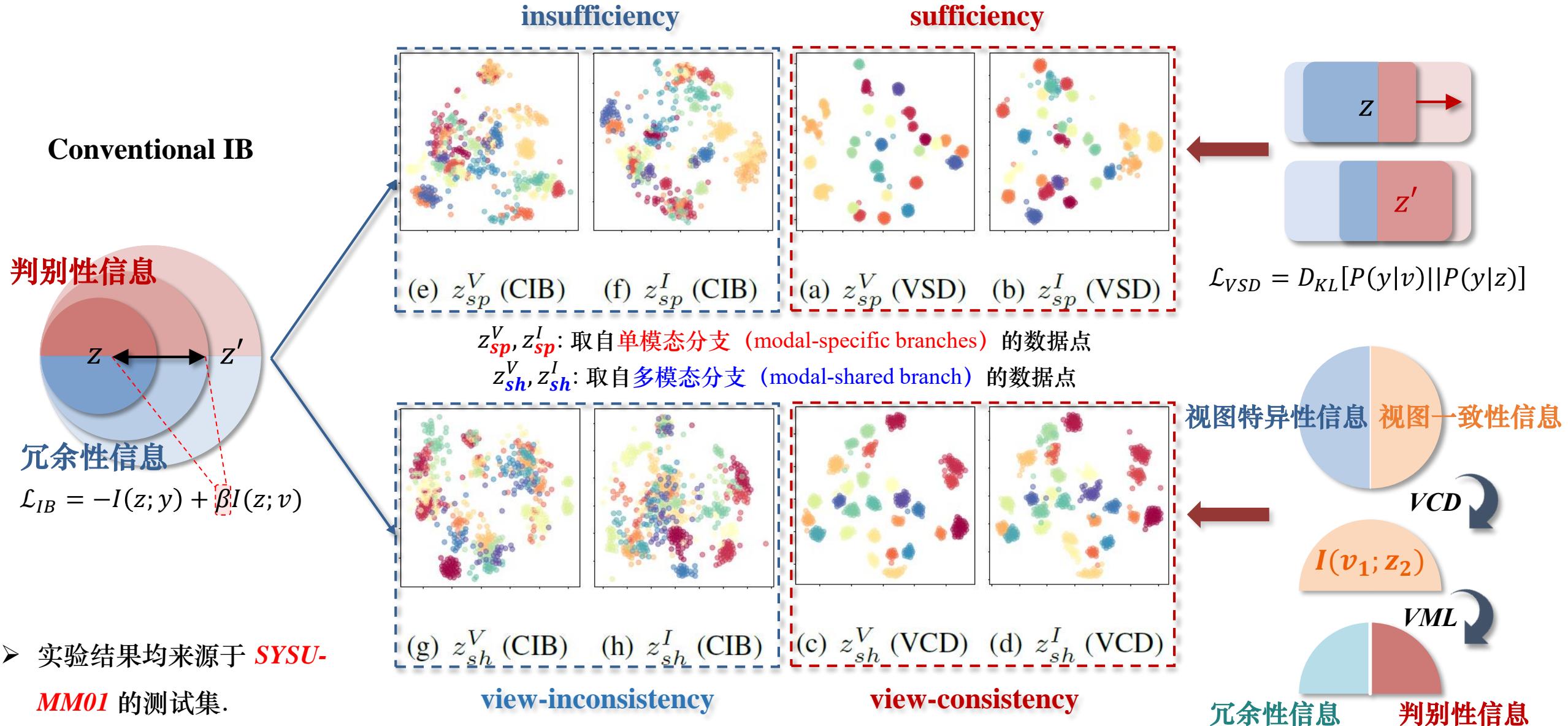
Rank-1	mAP
30.43	33.67

## 消融实验：

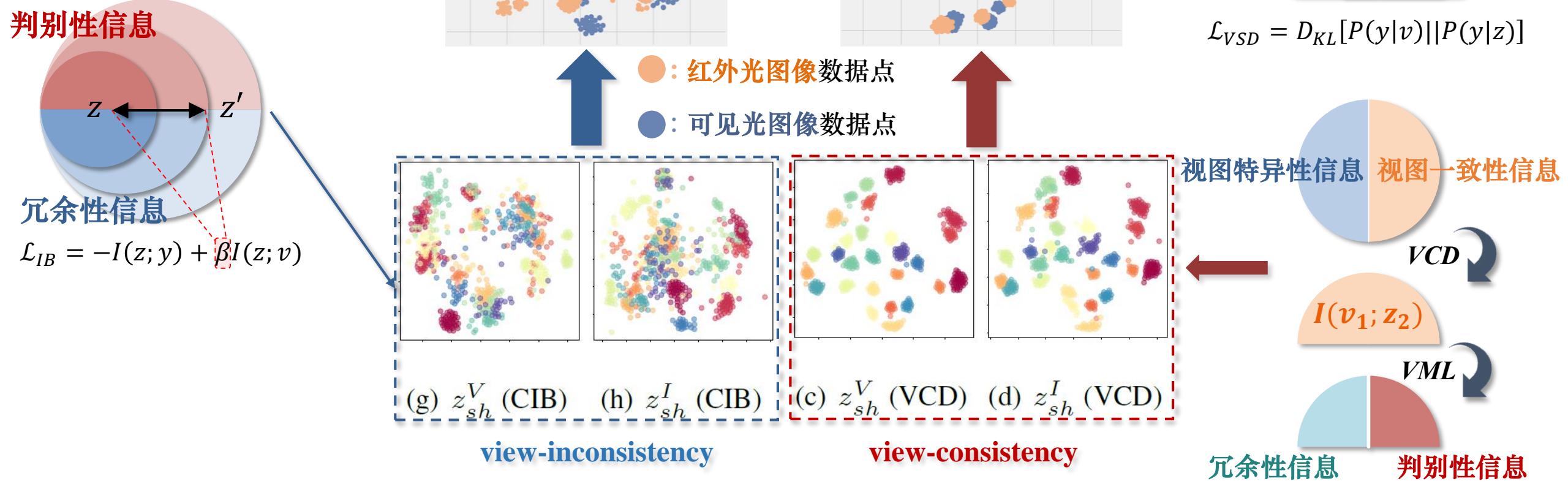
- 不同压缩率的情况下，“充分性”对比



# 结果分析



# 结果分析





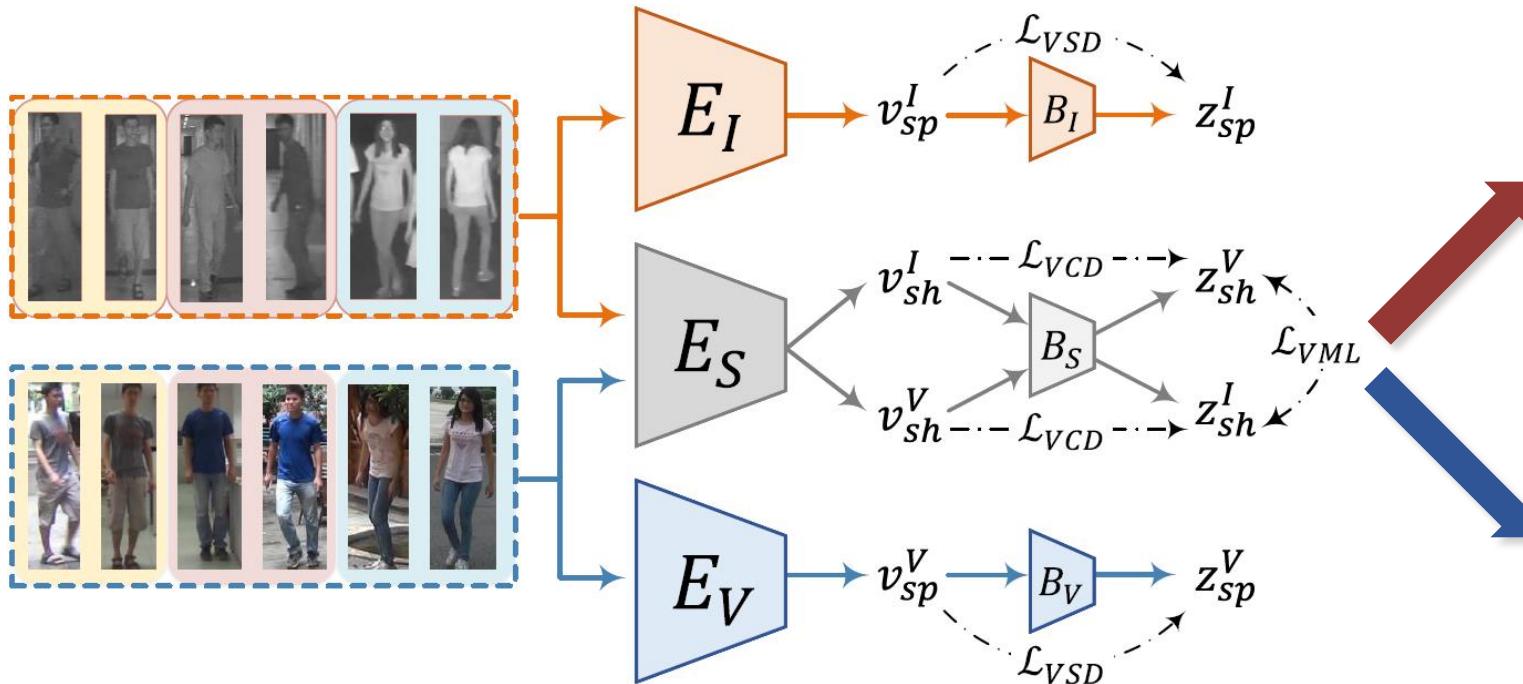
### 三、代码复现

➤ 复现结果



# 复现结果

## ➤ 性能: Pytorch vs MindSpore



Pytorch ver  
baseline (200 epoch)

Rank-1	mAP
54.8	54.0

MindSpore ver  
baseline (100 epoch)

Rank-1	mAP
60.5	56.0

Pytorch ver  
All in (300 epoch)

Rank-1	mAP
60.0	58.8

MindSpore ver  
All in (160 epoch)

Rank-1	mAP
65.0+	60.0+



谢谢聆听

Thank You