自监督学习及视觉感知应用

山世光

中国科学院计算技术研究所

sgshan@ict.ac.cn; http://vipl.ict.ac.cn

中国科学院计算技术研究所
Institute of Computing Technology,Chinese Academy of Sciences

目录

- What
- **Why**
- How
- Our related work
- Discussion and future

What is Self-Supervised Learning?

- Yann LeCun的蛋糕“理论”
  - 强调无监督/自监督学习的重要性

“Pure” Reinforcement Learning (cherry)

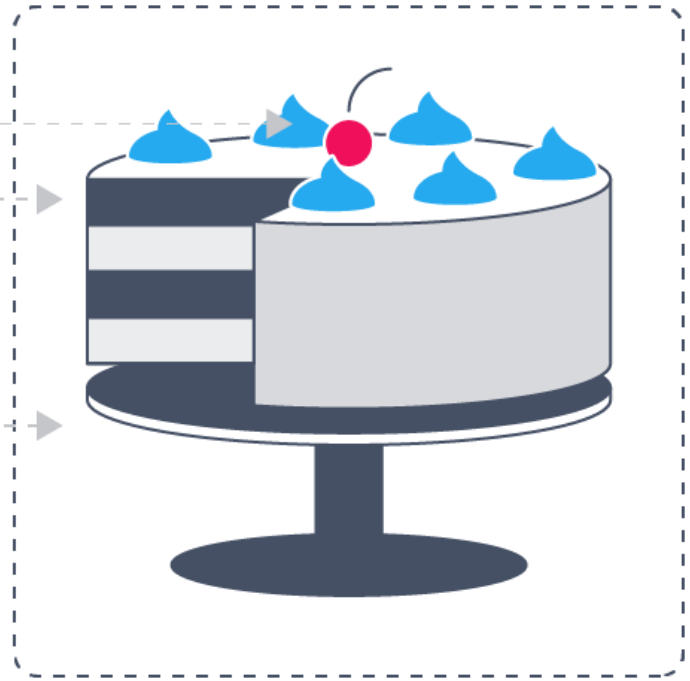
- The machine predicts a scalar reward given once in a while
- **A few bits for some samples**

Supervised Learning (icing)

- The machine predicts a category of a few numbers for each input
- Predicting human-supplied data
- **10 → 10,000 bits per sample**

Self-Supervised Learning (cake génoise)

- The machine predicts any part of its input for any observed part
- Predicts future frames in videos
- **Millions of bits per sample**



What is Self-Supervised Learning?

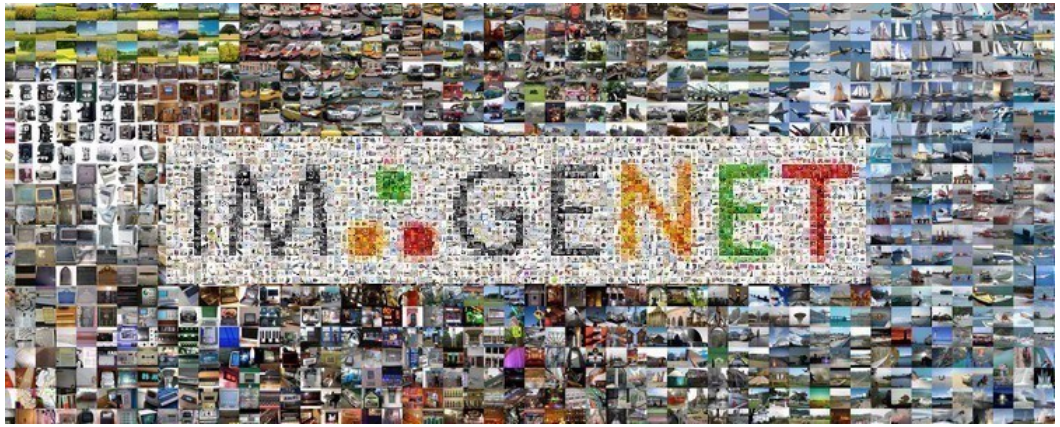
- 无监督学习的一种方式
  - **无标注**的数据，但经常进行的是**有监督的学习**
- 标注从哪里来?
  - **自带干粮**：标签是可以免费获得的，无需额外标注
  - 标签是自定义的代理任务Proxy/pretext task的标签
    - 可以自动生成标签
    - 更基础的任务：e.g.图像变换后内容不变，彩色➔黑白
  - 标签是当前任务的标签
    - 其他渠道的任务（e.g.其他模态，比如ASR vs. VSR）

Why？主要从CV角度

- 角度1：更充分的利用数据和标签
  - 标注困难/昂贵
  - 无标注数据几乎是取之不尽用之不竭的
  - 数据可以自带标签
    - 对人是不言而喻的
    - 对(已有的)算法/模型一脸懵逼
  - 强标注蕴含着大量的弱标签（配合知识）
    - e.g. 标签dog蕴含：眼睛，鼻子，嘴巴，毛发，腿，尾巴...
    - Unexplored topic

Why？主要从CV角度

- 角度2：更好的预训练模型
  - 分类任务➔其他(大样本)任务
  - **问题：真的可以很好的迁移吗?**



ImageNet, 1000 classes

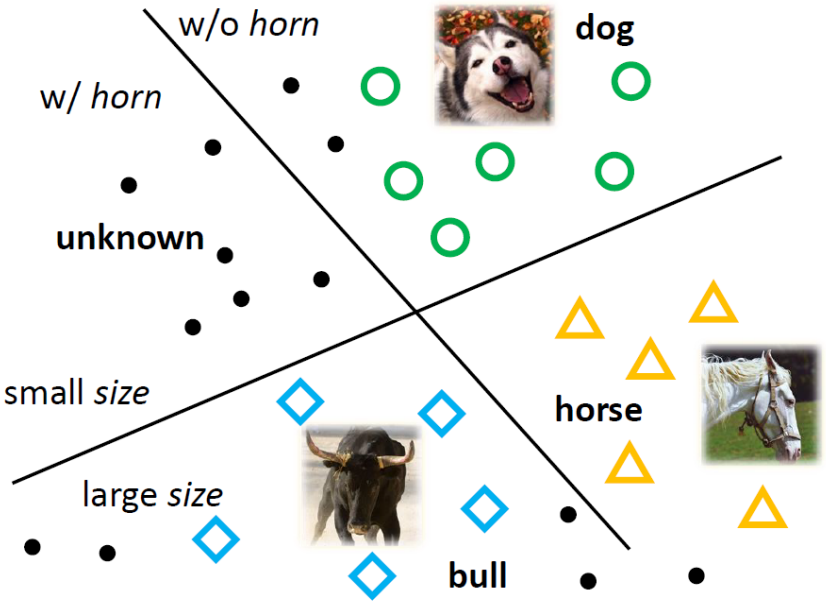
1000类分类任务训练的模型（Pretrained）

初始化，fine-tuning

其他任务

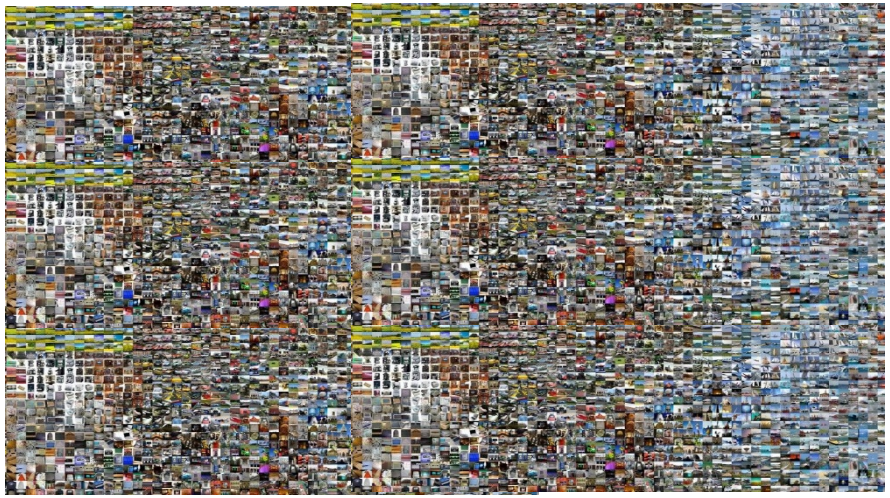
Why？主要从CV角度

- 角度2：更好的预训练模型
  - 分类任务➔其他(大样本)任务
  - **问题：真的可以很好的迁移吗？ No!**
  - 分类loss导致过早丢掉了太多特征
    - 极端例子
      - 分类：区分horse，bull和dog
      - 2个特征就够了：horn和size
      - 丢了其他特征！
    - Unknow unaware ➔ Unknown aware

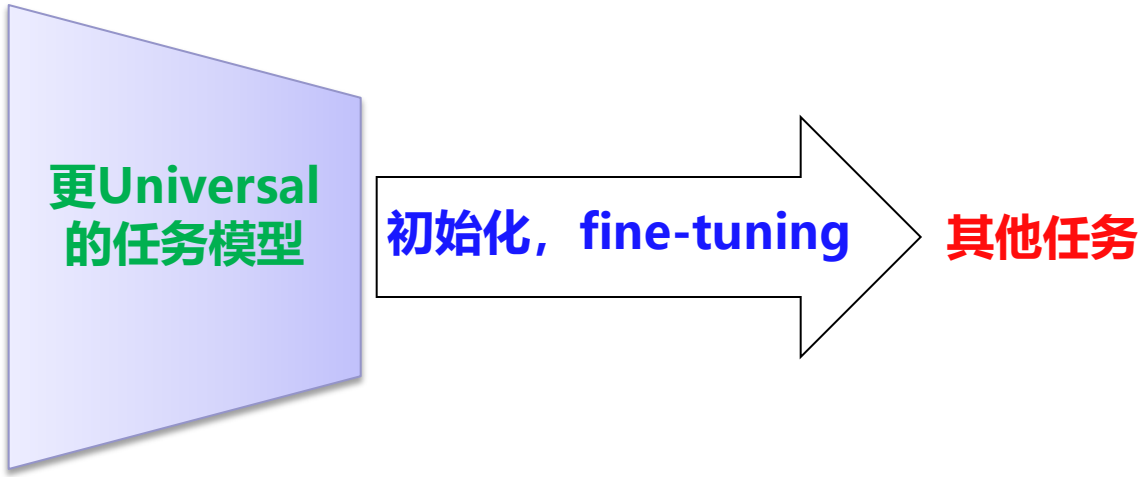


Why？主要从CV角度

- 角度2：更好的预训练模型
  - 分类任务➔其他(大样本)任务
  - 通用/上游任务➔**其他(小样本)下游任务**(downstream tasks)
    - 通用/上游？任务无关？是什么？



更大规模的数据



Why？主要从CV角度

- 角度3：常识表示——**自然语言常识**
  - NLP领域的大杀器：BERT，GPT等
  - 最近创新工场的王咏刚采用人机协同方式，写出一篇短篇科幻小说《迷路》——惊人的能力！！
  - **不会基本写出符合不语法的句子** ➡ **基本不会写出不符合语法的句子**

CV领域类似GPT-3的大杀器是什么？

Why？主要从CV角度

- 角度3：常识表示——**视觉常识表示**
  - **视觉是语言**！回想bag of visual word（VW）时代…
    - 图像的**词法**（lexical model）模型是什么?
      - 词袋模型时代：局部特征的向量量化（VQ）或聚类（clustering）得到视觉单词
      - CNN建模了视觉单词吗？每个Neuron就是一个visual word检测器
    - 图像的**句法**（syntactic model）模型是什么?
      - Visual word之间的连接方式，VW的空间结构/布局
        - 愚蠢的bag of (visual) word——扔掉了结构信息
      - CNN有句法吗?
        - 同层VW：平滑连接；自相似性；远距离合理性(对称性)
        - 跨层VW：边缘➔角点➔motif➔部件➔物体➔场景
        - 类比语言：单词➔短语➔从句➔句子➔段落➔篇章

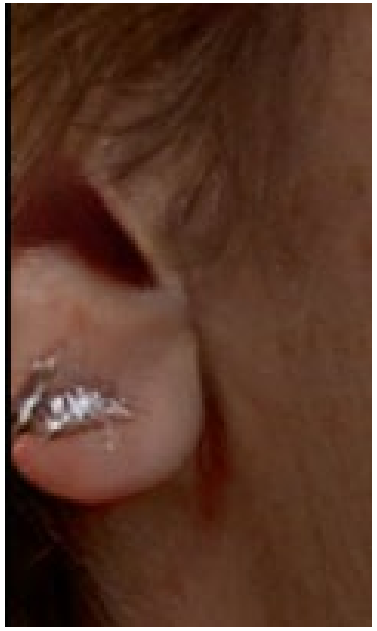
Why？主要从CV角度

- 角度3：常识表示——**视觉常识表示**
  - **视觉是语言**！回想bag of visual word（VW）时代...
  - 合理的语言➔合理的图像/视频?
  - 对人而言，至少Low-level/mid-level的合理性是不言而喻的

http://thispersondoesnotexist.com/

Why？主要从CV角度

- 角度3：常识表示——**视觉常识表示**
  - **视觉是语言**！回想bag of visual word（VW）时代…
  - 合理的语言➔合理的图像/视频?
  - 对人而言，至少Low-level/mid-level的合理性是不言而喻的



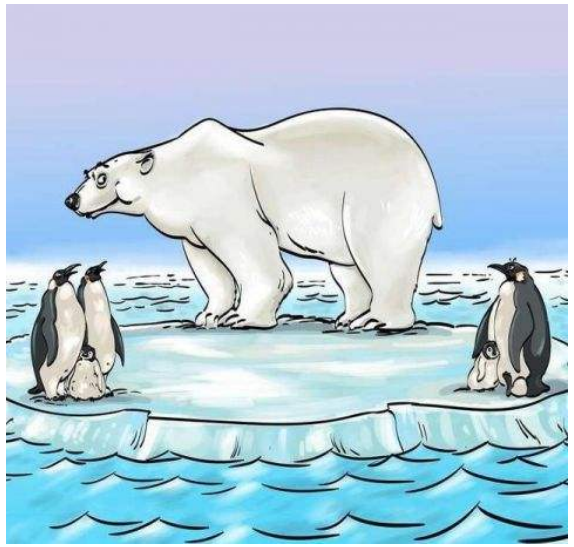
http://thispersondoesnotexist.com/

Why？主要从CV角度

- 角度3：常识表示——**视觉常识表示**
  - **视觉是语言**！回想bag of visual word（VW）时代…
  - 合理的语言➔合理的图像/视频?
  - 对人而言，至少Low-level/mid-level的合理性是不言而喻的

Low-level的不合理

Mid-level的不合理



High-level的不合理

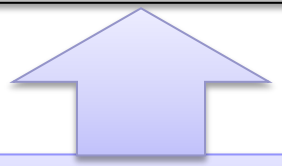
Why？主要从CV角度

- 角度3：**视觉常识表示——为什么对AI至关重要？！**
  - “三岁看大”是胡说八道吗？
  - 婴幼儿期(1-3岁)的脑发育是在发育什么？
    - 语言、概念、符号、逻辑甚至自我意识等等都还很弱（从无到有）
    - 清醒状态下视听感觉系统时时刻刻在工作中。在干嘛？
      - 神经细胞数量在此期间达到顶峰，神经细胞之间的连接远多于成年人
  - 但是我们对这个时期似乎没有记忆，没用？
    - 其实只是缺失需要用语言描述的“外显记忆”
    - “内隐记忆”都在，包括大量“只可意会不可言传”的**常识或直觉**
    - 这类学习或可称为“直觉学习” 或“非符号非逻辑的常识学习”

Why？主要从CV角度

- 角度3：**视觉常识表示——为什么对AI至关重要？！**
  - 三岁看大——计算视角
    - 学习和推理的特点：无监督/弱监督，非符号
    - 表示空间：连续向量空间，分布式表示

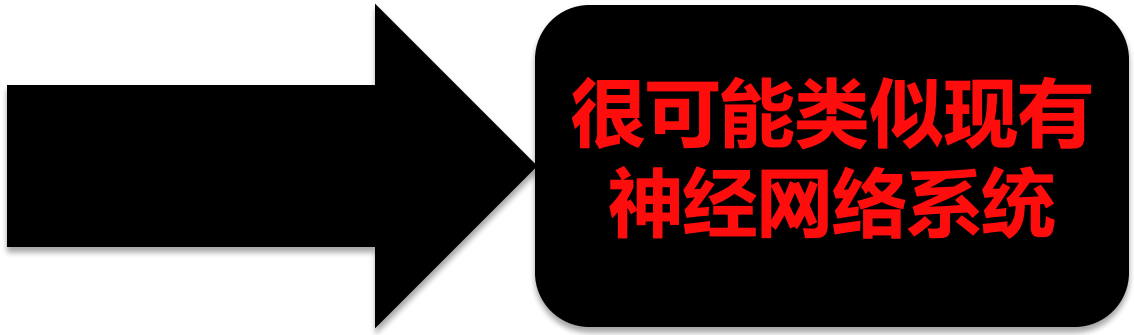
连续向量空间中的非符号表示
（直觉/常识/系统1/快系统）



无监督/弱监督学习
非符号、非逻辑层面的推理

Why？主要从CV角度

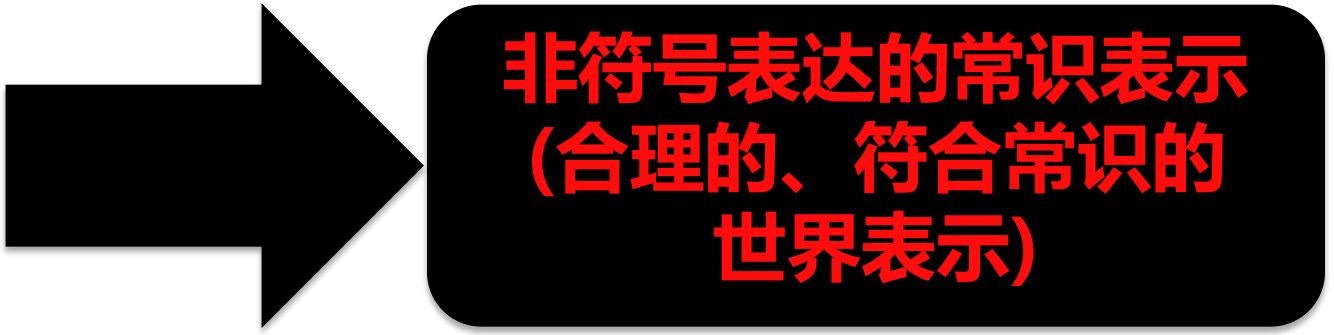
- 角度3：**视觉常识表示——为什么对AI至关重要？！**
  - 三岁看大——计算视角
  - **系统1*——至少是感知层面的系统1(比如人脸识别...)**
    - 非符号、非语言
    - 直觉推理（非逻辑）
    - 主要是无意识的
    - 难以解释
    - 通常是快速的



*系统1和系统2的概念最早是由心理学家Keith Stanovich和Richard West提出

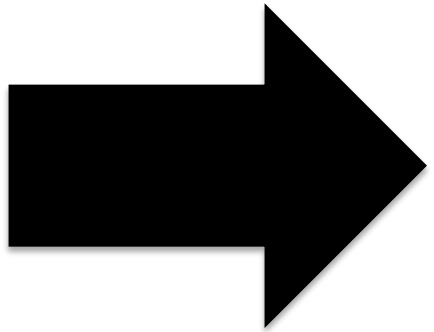
Why？主要从CV角度

- 角度3：**视觉常识表示——为什么对AI至关重要？！**
  - 三岁看大——计算视角
  - 系统1*——至少是感知层面的系统1(比如人脸识别...)
- **需要研究基于超大规模无监督数据的自/弱监督学习**
  - BERT, GPT-3
  - CPL，SimCLR，MoCo
  - 多模态协同
  - 具身智能的配合
  - ???



Why？主要从CV角度

- 角度3：**视觉常识表示——为什么对AI至关重要？！**
  - 三岁看大——计算视角
  - 系统1*——至少是感知层面的系统1(比如人脸识别...)
- **需要研究基于超大规模无监督数据的自/弱监督学习**
  - BERT, GPT-3
  - CPL，SimCLR，MoCo
  - 多模态协同
  - 具身智能的配合
  - ???



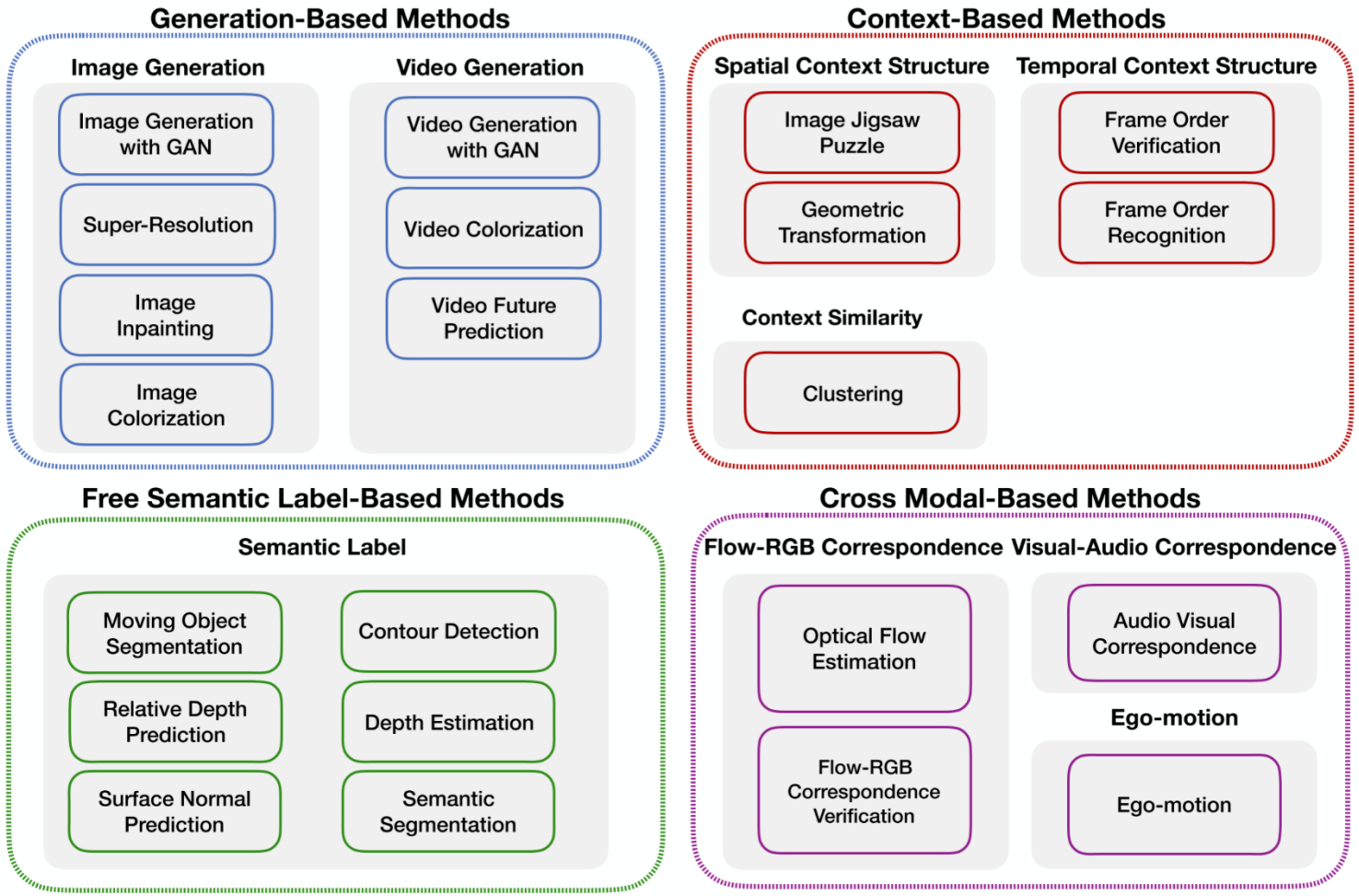
不会基本写出符合不语法的句子

目录

- What?
- Why?
- **How?**
- Our related work
- Discussion and future

How?

- Longlong Jing, Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey IEEE T PAMI, 2020
- 综述内容很全面，但对Contrastive learning方法的介绍偏少



How?

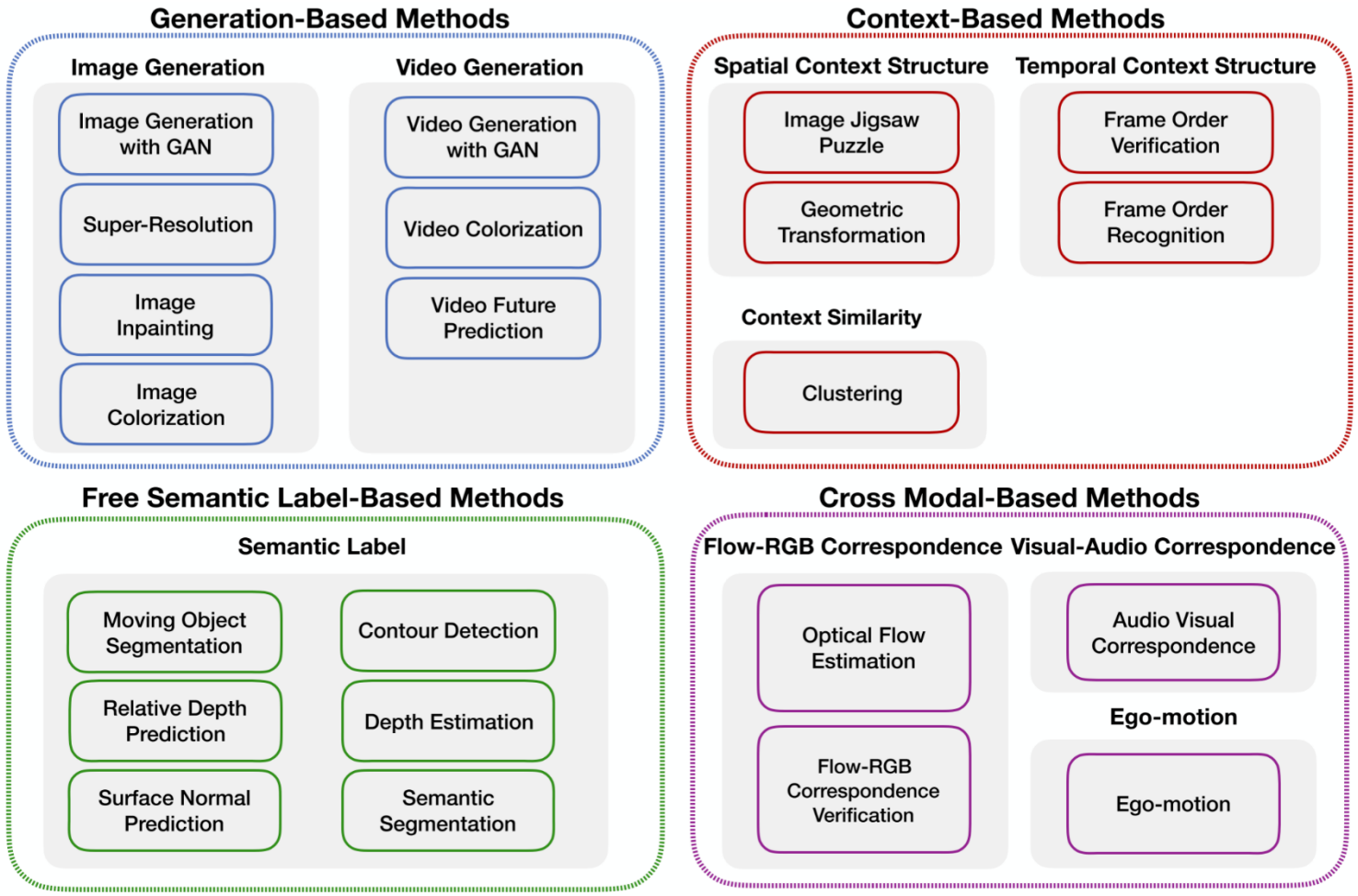
- 核心是定义Pretext/proxy task并自动产生标签
  - 任务应该足够基础，足够通用
    - 图像：建模视觉语言的常识（词法/句法等不同层面）
    - 视频：外加时序维度（Dynamics）
  - 标签能够自动产生
    - 最好是对人**不言而喻**的标签【但模型/算法并未建模的】
    - 通过简单规则可以产生，可以低成本的**大量**生成

How?

- 标签如何自动生成？——**挖掘对人不言而喻的自生标签**
  - **无中挖有**
    - 类似数据增广
    - 标签对图像变换的不变性或同变性
      - 各类变换：几何变换(仿射变换，裁剪，镜像），光照，色彩…
  - **断臂再生**
    - 彩色➔灰度；高分辨率➔低分辨率；全局➔局部
    - 擦除再补(inpainting)【类似自然语言中的用法】
    - 打乱重排（时间或空间）
  - **它山之石**
    - 其他模态：听觉，触觉…

How?

- Longlong Jing, Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey IEEE T PAMI, 2020
- 综述内容很全面，但对Contrastive learning方法的介绍偏少



补图法

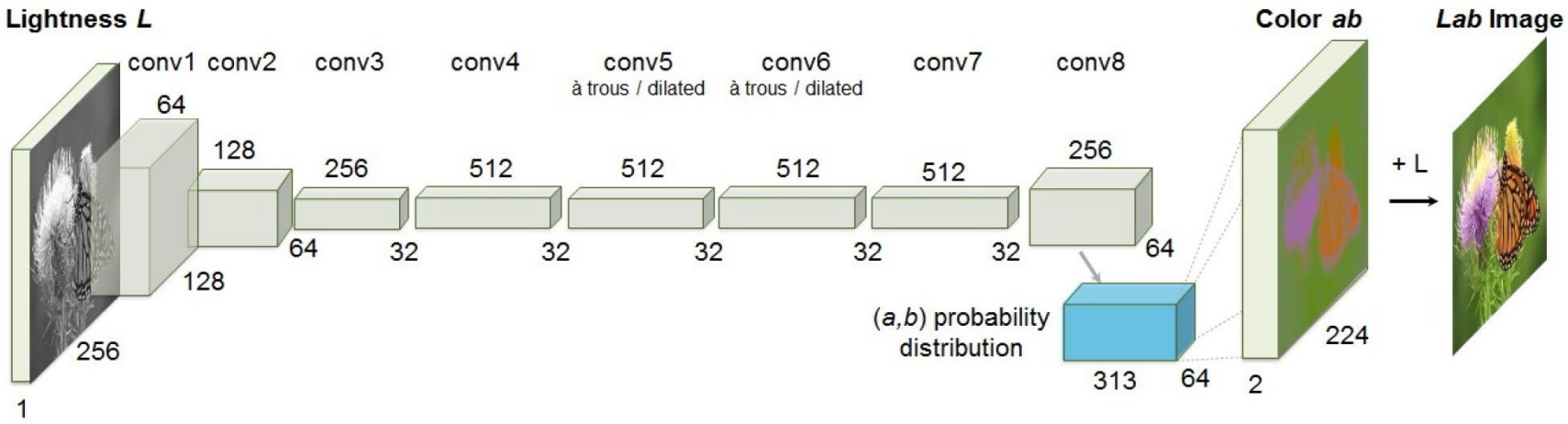
- 用其他区域填补删除区域
  - ConvNet+对抗loss

(a) Input Context　(b) Human Artist　(c) Network Prediction

D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” CVPR 2016

上色法

- Image Generation with Colorization
  - Fully convolution neural network which consists of an encoder for feature extraction and a decoder for the color hallucination to colorization.
  - The network can be optimized with L2 loss



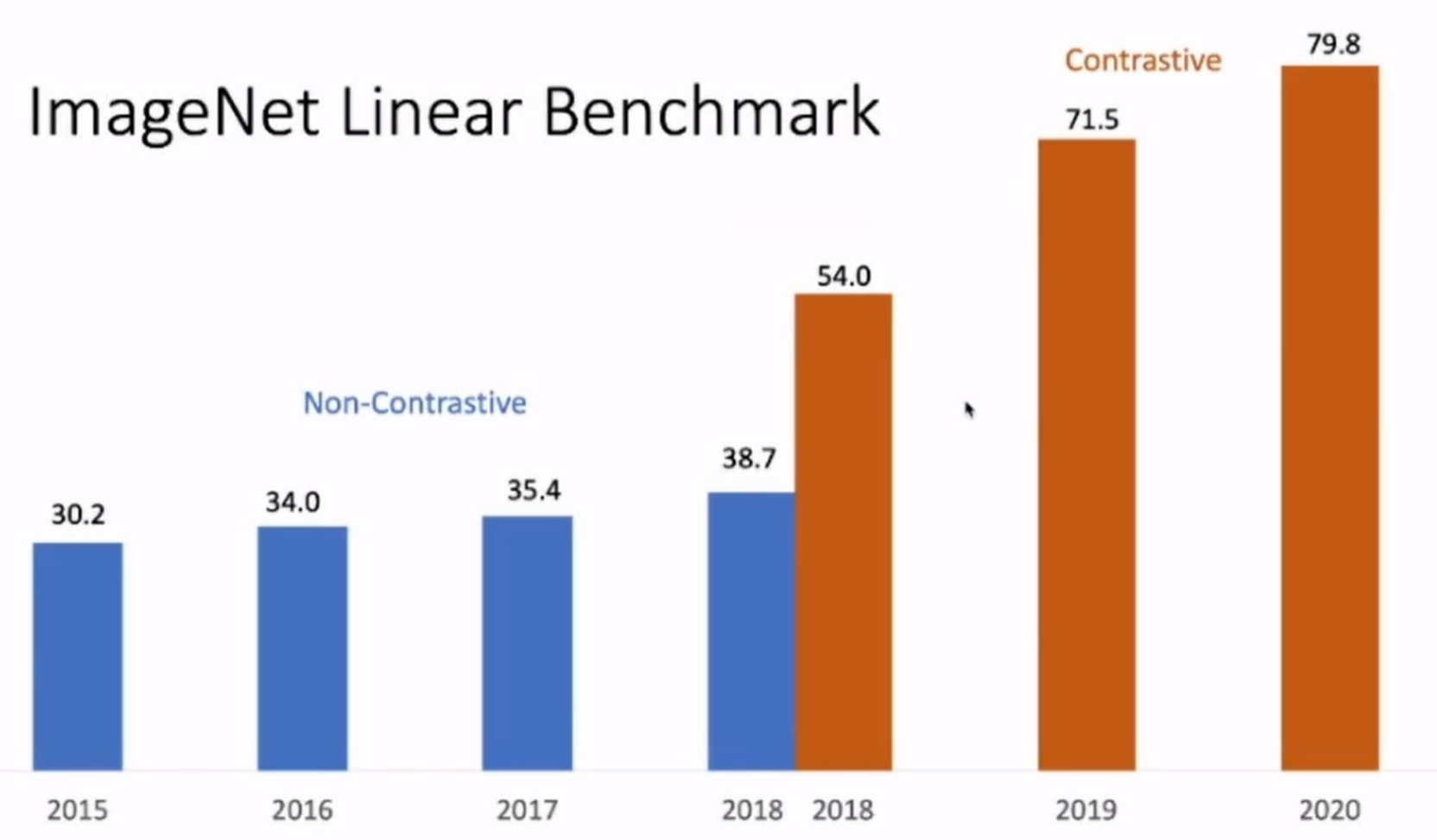
R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in ECCV, pp. 649–666, Springer, 2016

重排图像Patch

- Jigsaw Image Puzzle
  - The shuffled image patches are fed to the network
    - Trained to recognize the correct spatial locations of the input patches by learning spatial context structures of images such as object color, structure, and high level semantic information

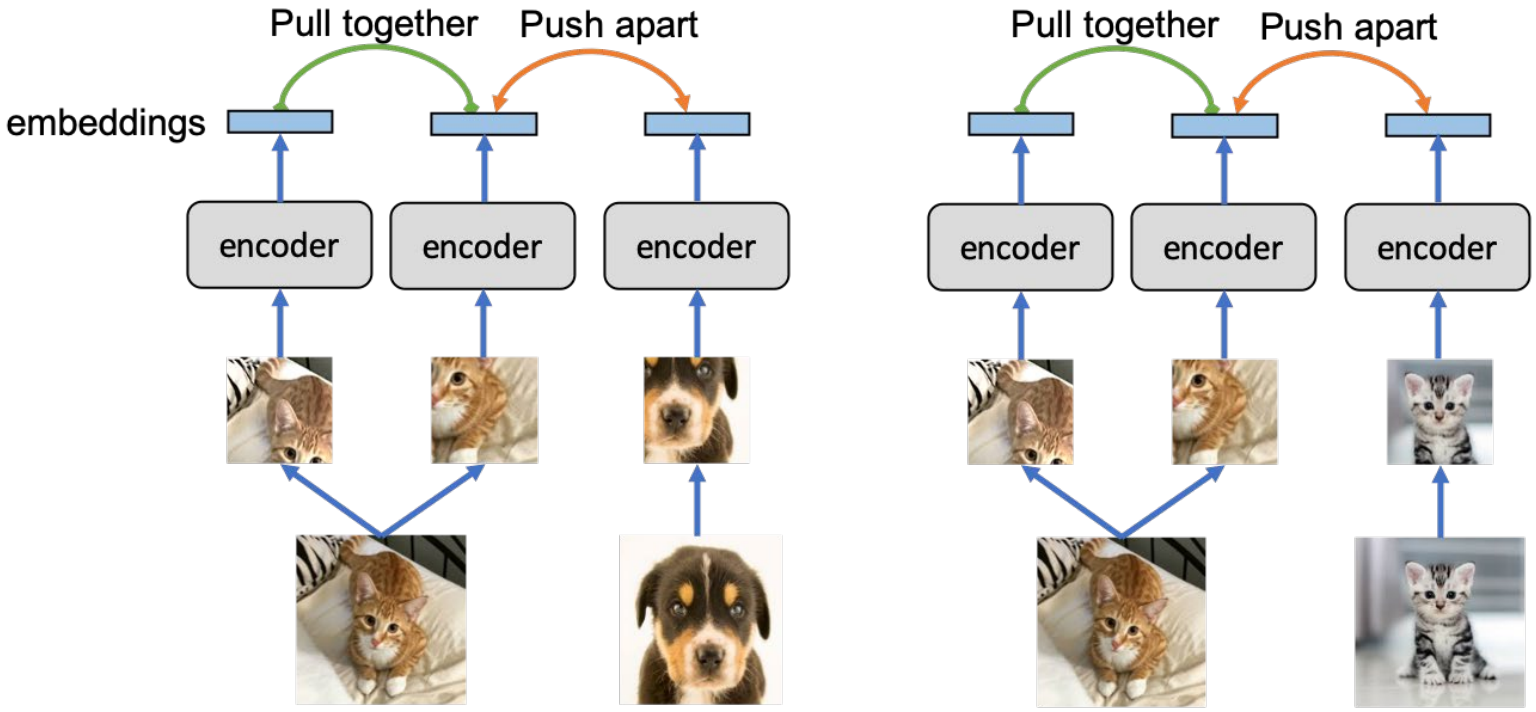
M. Noroozi and P. Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” in ECCV, 2016

Contrastive Learning来袭...



Contrastive learning

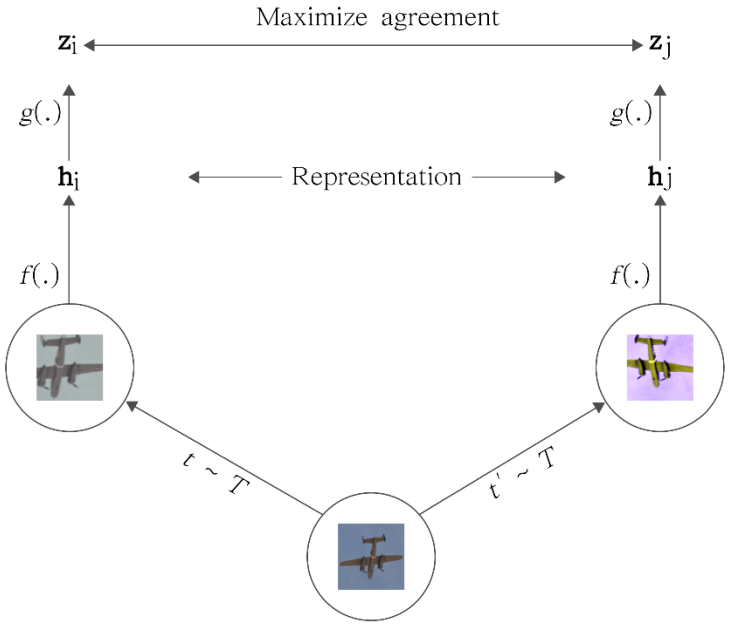
- 学习图像的上下文结构
  - 拉近同图patch（Positive），推远非同图patch（negative）

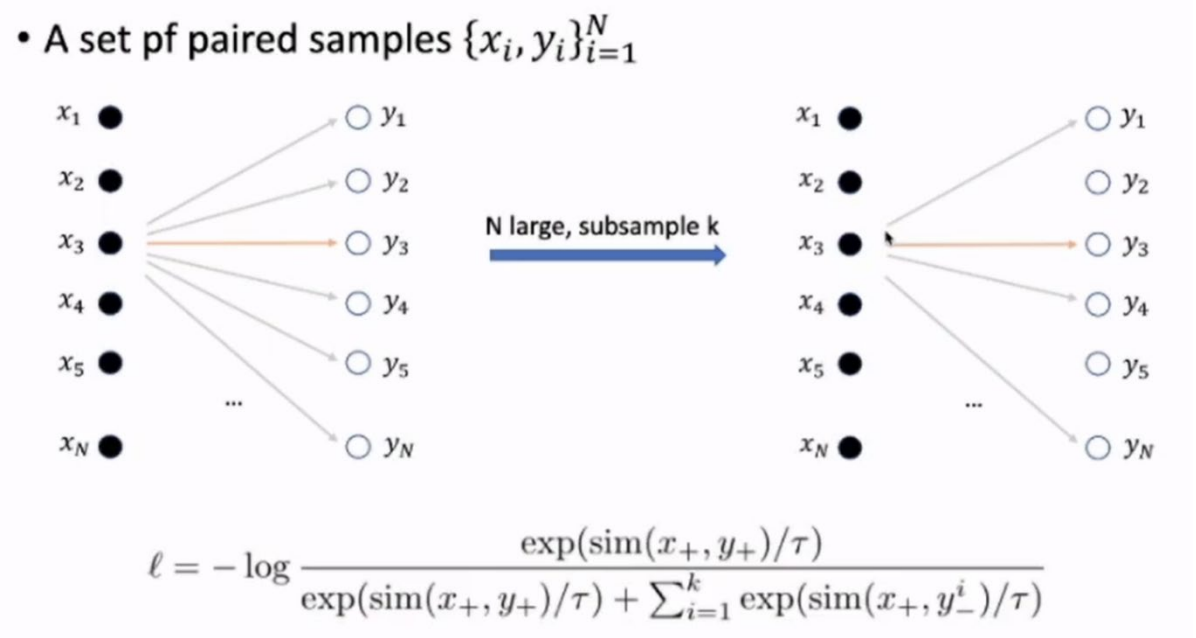


R Hadsell, S Chopra, Y LeCun. Dimensionality reduction by learning an invariant mapping, CVPR2006
K Sohn, Improved deep metric learning with multi-class n-pair loss objective, NIPS2016

Contrastive learning

- 学习图像的上下文结构
  - 拉近同图patch（Positive），推远非同图patch（negative）
  - 采样negative samples

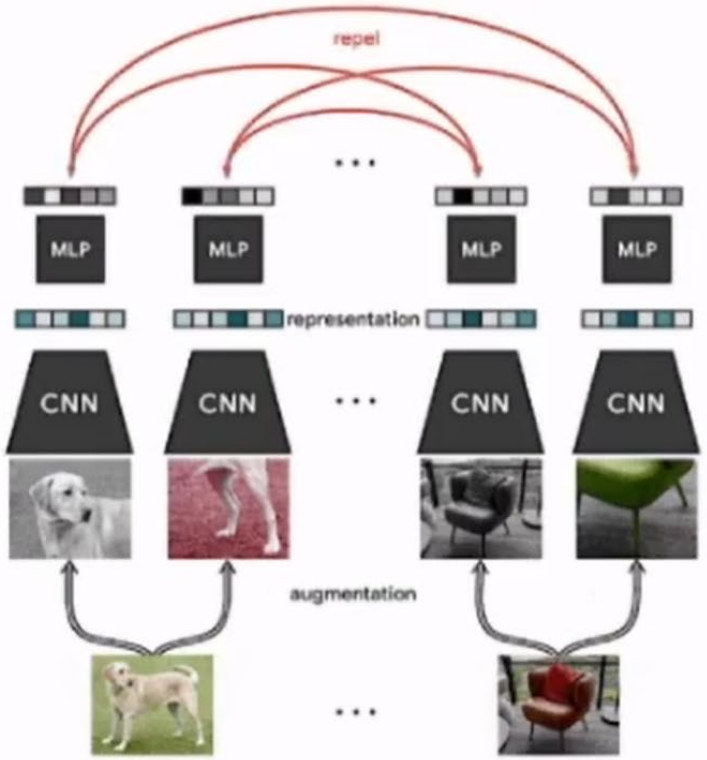




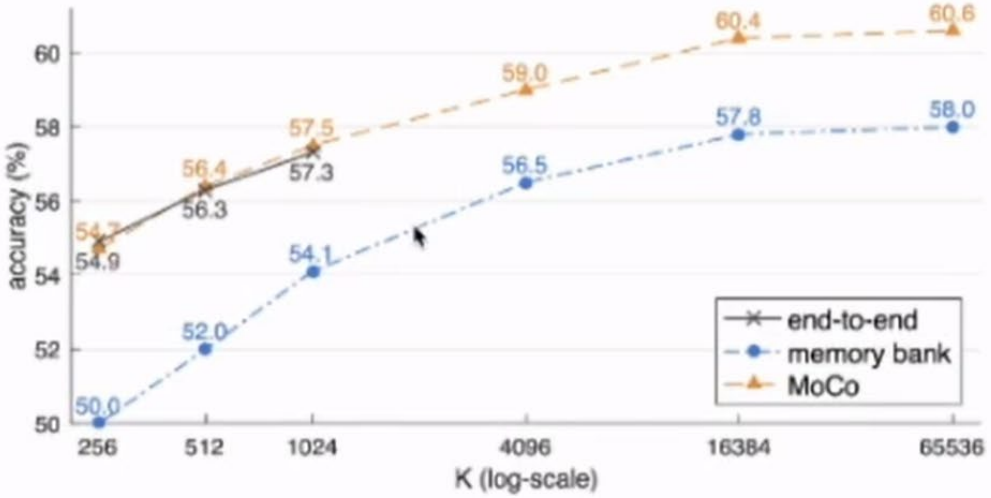
R Hadsell, S Chopra, Y LeCun. Dimensionality reduction by learning an invariant mapping, CVPR2006
K Sohn, Improved deep metric learning with multi-class n-pair loss objective, NIPS2016

Contrastive learning

(1) Augmentations as (X, Y)



(2) Large number of negatives



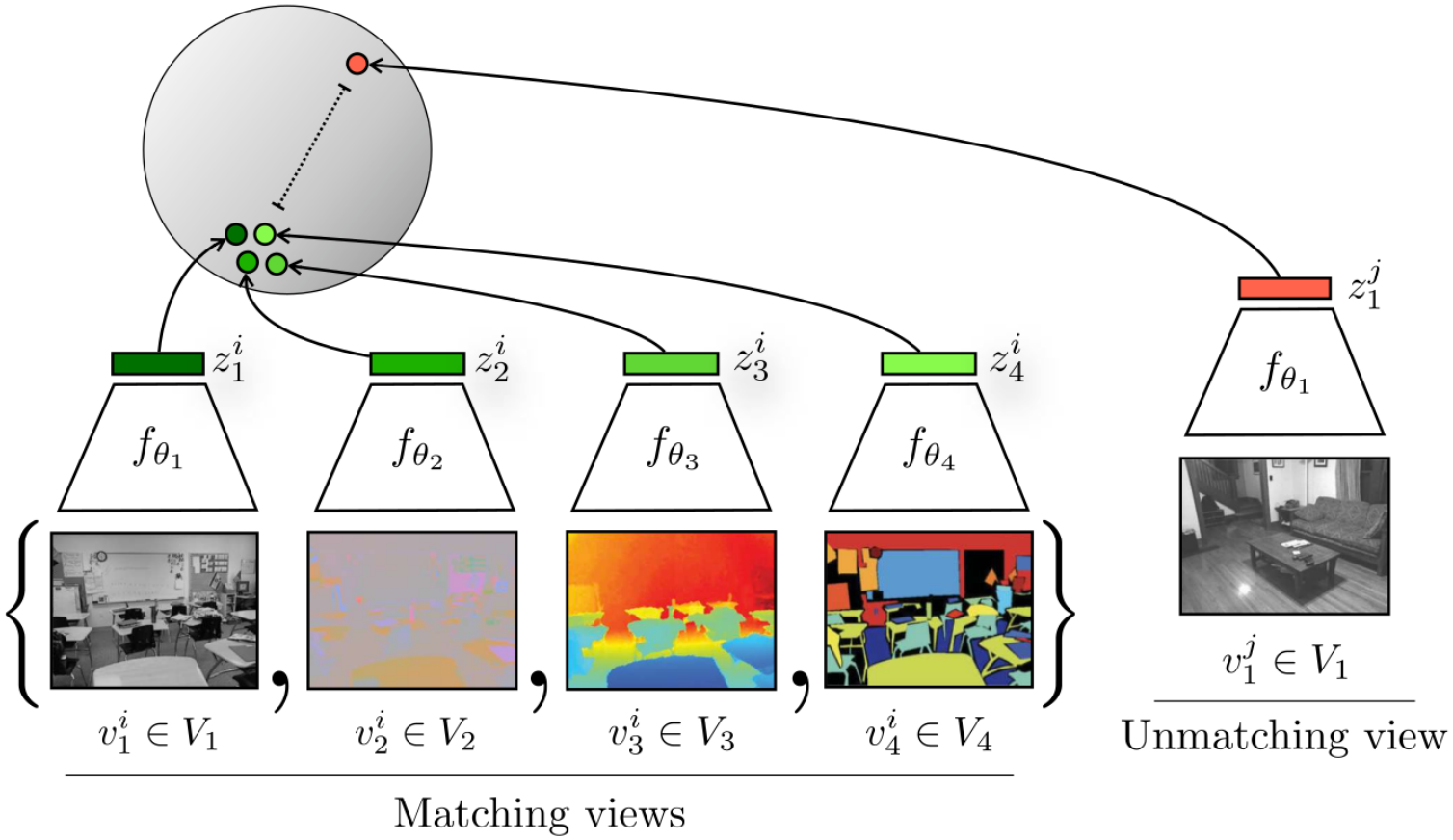
- Memory bank (InstDis)
- Momentum Encoder (MoCo)
- Large Batch (SimCLR)

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. CVPR2020

T Chen, S Kornblith, M Norouzi, G Hinton. A simple framework for contrastive learning of visual representations. arXiv 2020

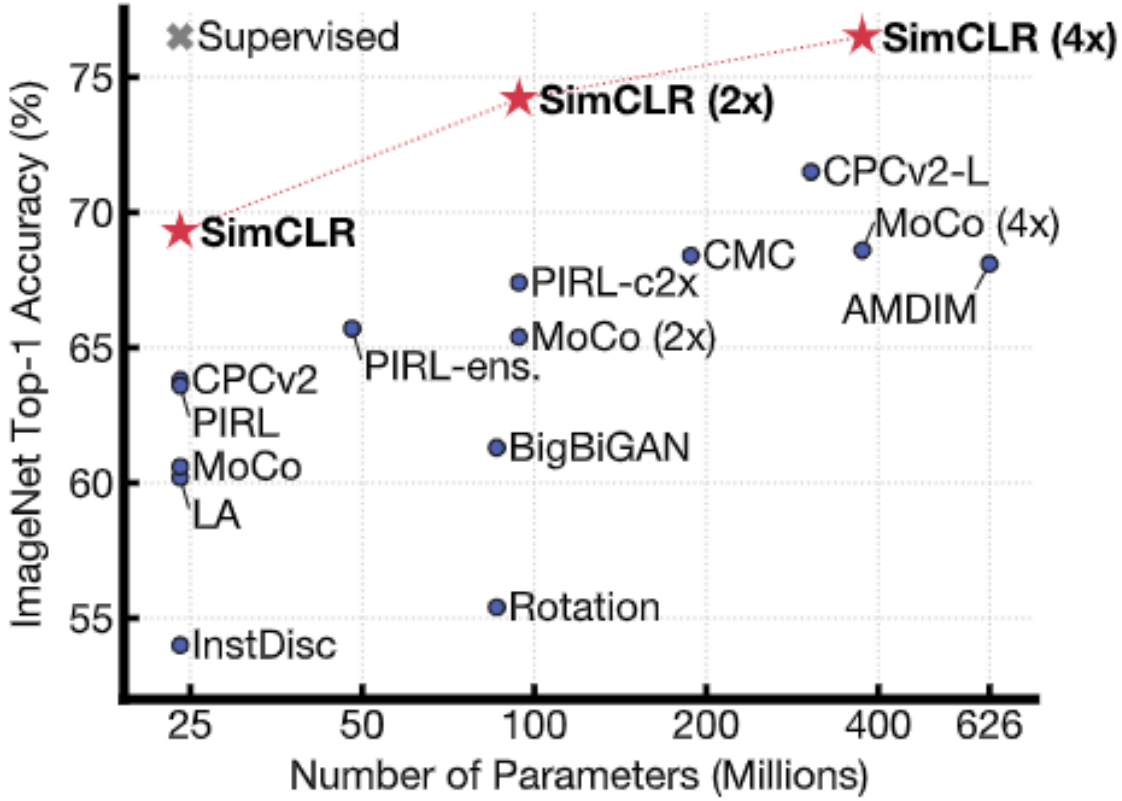
Contrastive Multi-view Learning

- Yonglong Tian et al., MIT, arXiv 2020



CL改进系列的性能

- 已经逼近监督方法



https://space.bilibili.com/562085182/ 【Ting Che, Yonglong Tian, Weidi Xie亲讲】

目录

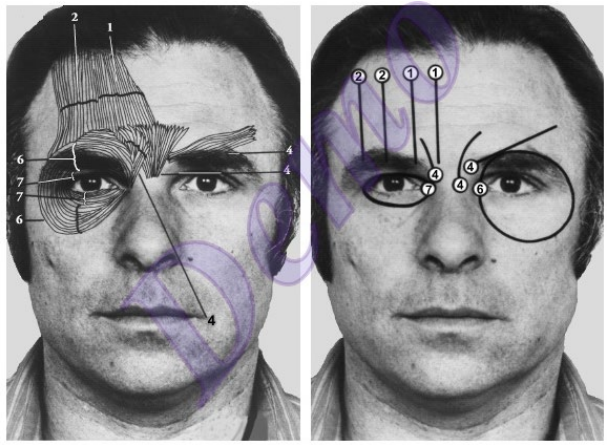
- What?
- Why?
- How?
- **Our related work**
- Discussion and future

Yong Li, Jiabei Zeng, Shiguang Shan, Xilin Chen. Twin-Cycle Autoencoder: Self-supervised Representation Learning from Entangled Movement for Facial Action Unit Detection.
IEEE/CVF CVPR2019，IEEE T-PAMI

基于自监督表示学习的AU建模与检测

工作1：面向面部动作检测的无监督学习

- 面部动作检测
- AU
  - Action Unit



上半脸动作单元AU

AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
内眉上扬	外眉上扬	眉毛下压	上眼睑上扬	脸颊抬起	眼睑收紧
*AU 41	*AU 42	*AU 43	AU 44	AU 45	AU 46
眼睑下垂	眯眼	闭眼	半眯眼	眨眼	半眨眼

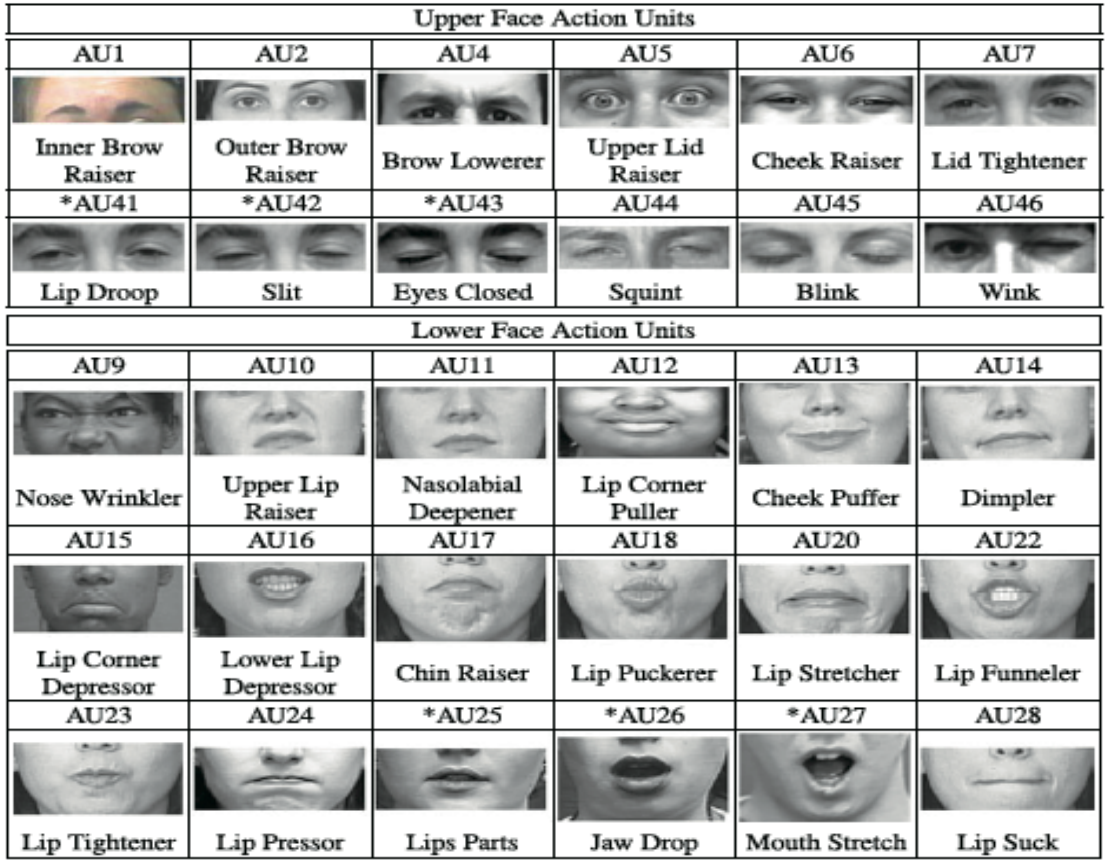
下半脸动作单元AU

AU 9	AU 10	AU 11	AU 12	AU 13	AU 14
鼻子褶皱	上唇上扬	人中抬起	嘴角上扬	腮颊鼓起	收紧嘴角
AU 15	AU 16	AU 17	AU 18	AU 20	AU 22
嘴角下拉	下嘴唇下坠	下巴缩紧	噘嘴	嘴唇舒展	龇牙
AU 23	AU 24	*AU 25	*AU 26	*AU 27	AU 28
嘴唇收缩	嘴唇压紧	上下嘴唇分开	下颚下拉	张嘴	抿嘴

工作1：面向面部动作检测的无监督学习

- **动机**：AU数据匮乏，**可否利用**无监督视频？

面部动作单元标注代价昂贵：标注一分钟视频需要专家花费30分钟以上

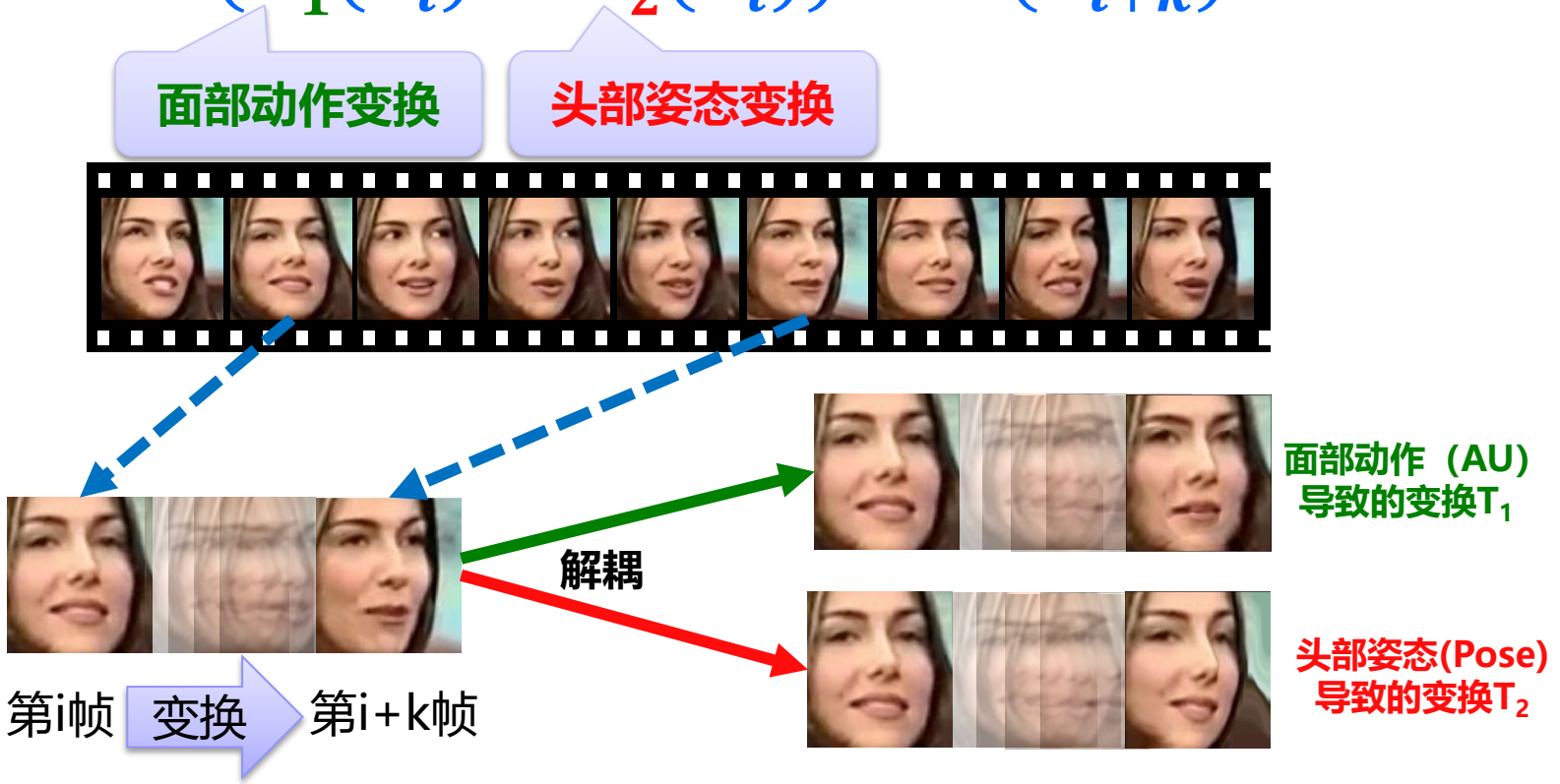


工作1：面向面部动作检测的无监督学习

- **方法**：以视频中的面部动作为**自监督Pretext task**

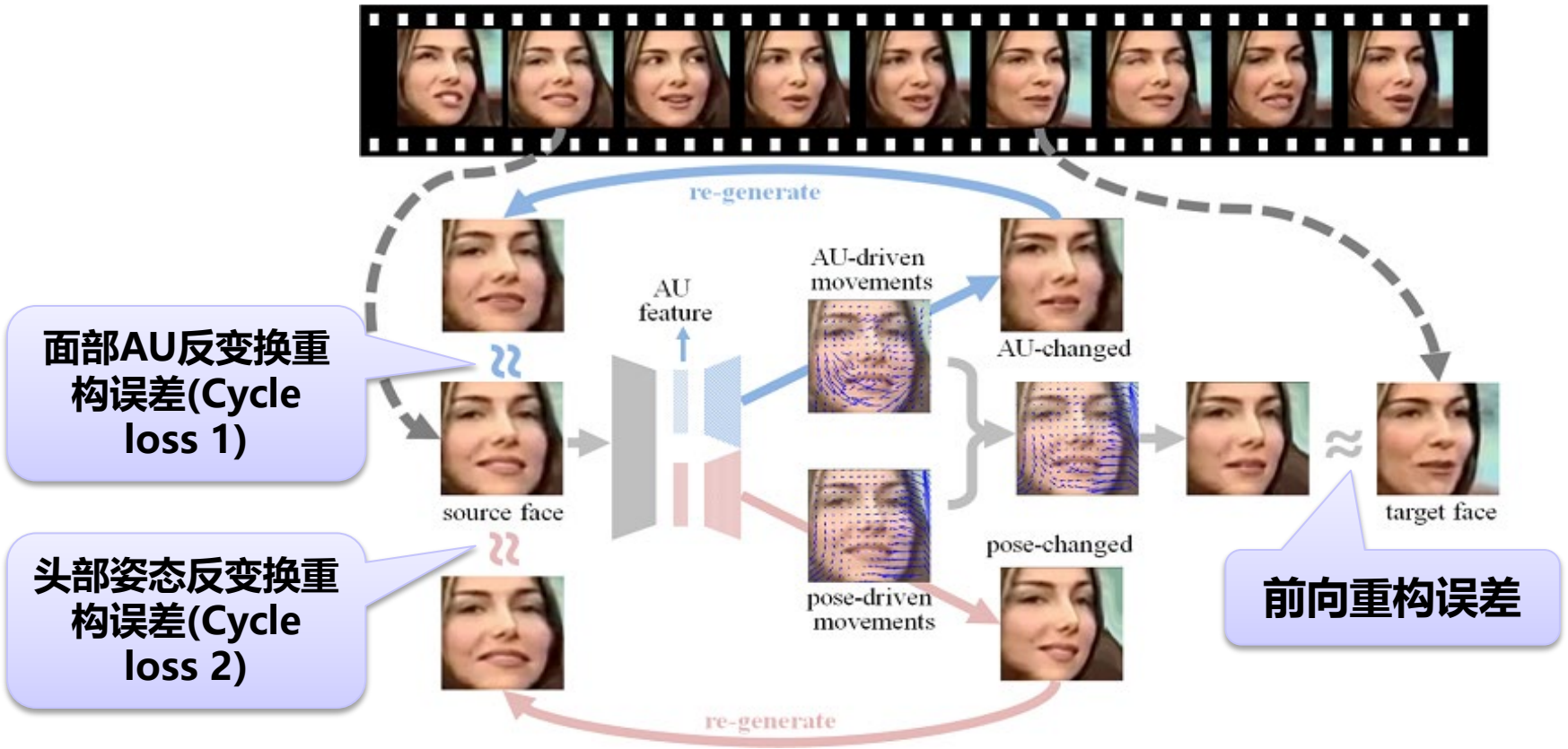
$$F\left(T(x_i)\right) = F(x_{i+k})$$

$$F(T_1(x_i) + T_2(x_i)) = F(x_{i+k})$$



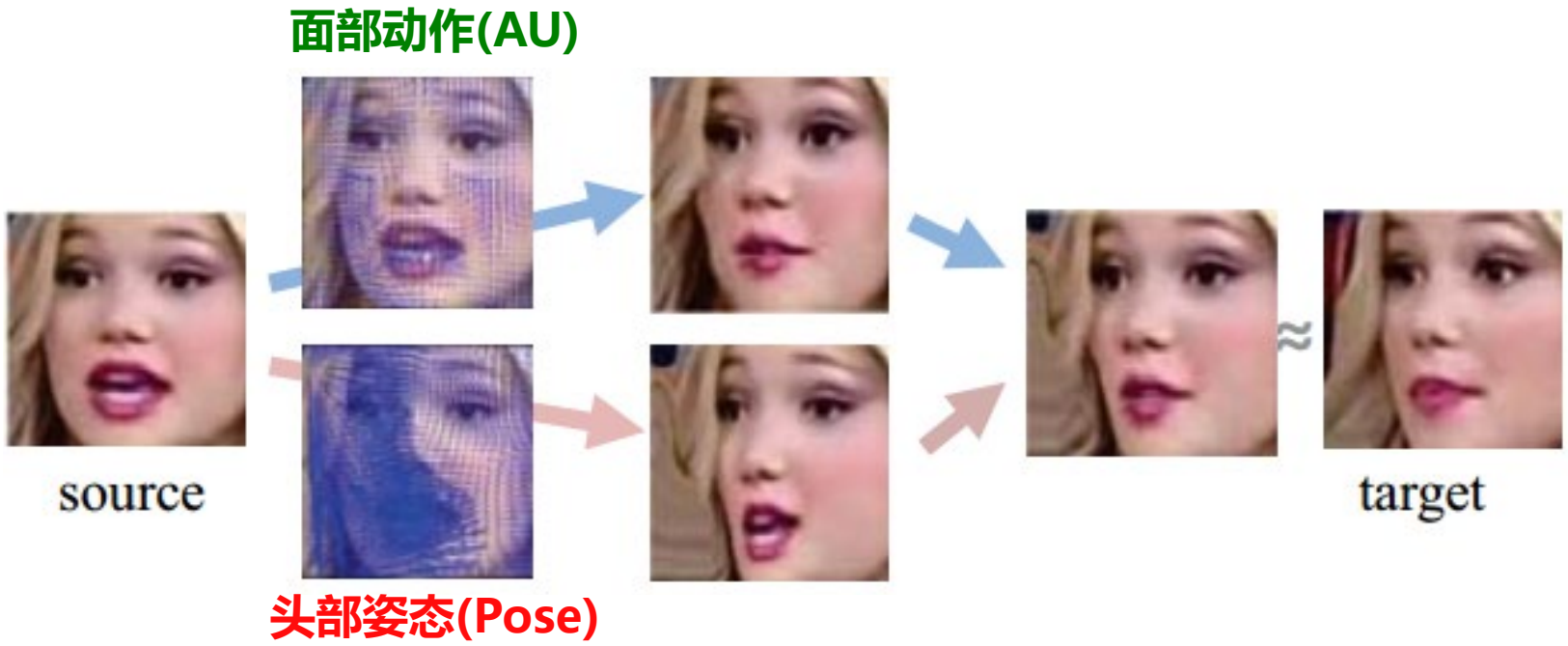
工作1：面向面部动作检测的无监督学习

- **方法**：以视频中的面部动作为自监督Pretext task
- **损失函数**：双反变换重构损失(Cycle loss)+前向重构误差



工作1：面向面部动作检测的无监督学习

- **方法**：以视频中的面部动作为自监督Pretext task
- **损失函数**：双反变换重构损失(Cycle loss)+前向重构误差
- **效果**：解耦效果示例



工作1：面向面部动作检测的无监督学习

- **方法**：以视频中的面部动作为自监督Pretext task
- **损失函数**：双反变换重构损失(Cycle loss)+前向重构误差
- **效果**：8层轻量级网络(无监督)+单FC层分类**媲美复杂方法**

BP4D数据集 F1 score

	Methods/AU	1	2	4	6	7	10	12	14	15	17	23	24	ave
Descriptor	Handcrafted [21]*	43.4	40.7	43.3	59.2	61.3	62.1	68.5	52.5	36.7	54.3	39.5	37.8	50.0
	ResNet-80 face	39.3	40.6	38.5	64.2	67.5	71.0	65.3	57.2	37.8	51.3	35.1	32.6	49.9
	VGG emotion	46.4	36.3	49.6	76.0	77.6	80.2	87.8	60.8	40.4	59.1	43.7	48.2	58.8
Supervised	AlexNet [52]*	40.3	39.0	41.7	62.8	54.2	75.1	78.1	44.7	32.9	47.3	27.3	40.1	48.6
	DRML [2]*	36.4	41.8	43.0	55.0	67.0	66.3	65.8	54.1	33.2	48.0	31.7	30.0	48.3
	EAC-Net [3]*	39.0	35.2	48.6	76.1	72.9	81.9	86.2	58.8	37.5	59.1	35.9	35.8	55.9
	ROI [4]*	36.2	31.6	43.4	77.1	73.7	85.0	87.0	62.6	45.7	58.0	38.3	37.4	56.4
	JAA-Net [5]*	47.2	44.0	54.9	77.5	74.6	84.0	86.9	61.9	43.6	60.3	42.7	41.9	60.0
Self-supervised	SplitBrain [26]	39.0	32.0	39.7	72.9	70.6	78.2	83.7	57.8	37.3	53.6	32.3	45.1	53.5
	DeformAE [47]	39.5	34.5	40.8	70.5	68.4	76.3	82.9	60.7	23.1	54.1	34.3	43.1	52.3
	Fab-Net [24]	43.3	35.7	41.6	72.9	63.0	75.9	83.5	57.7	26.5	48.2	33.6	42.4	52.0
	TAE (w/o L_1, w/o att.)	40.6	38.2	44.7	71.9	67.5	76.0	81.3	61.6	34.0	54.7	39.4	46.7	54.7
	TAE (w/ L_1, w/o att.) [53]	43.1	32.2	44.4	**75.1**	70.5	80.8	85.5	61.8	34.7	58.5	37.2	**48.7**	56.1
	TAE (w/ L_1 & att., $K = 2$)	42.0	41.9	44.5	74.2	73.5	80.6	85.9	60.4	41.4	61.3	44.1	46.0	57.9
	TAE (w/ L_1 & att., $K = 4$)	40.7	42.1	**52.7**	71.0	**73.7**	79.7	85.0	62.1	44.0	**62.9**	43.7	48.1	58.8
	***TAE** (w/ L_1 & att., $K = 8$)*	**47.0**	**45.9**	50.9	74.7	72.0	**82.4**	**85.6**	62.3	**48.1**	62.3	**45.9**	46.3	**60.3**
	TAE (w/ L_1 & att., $K = 16$)	44.2	43.5	50.1	73.8	73.5	81.5	85.1	**62.8**	44.4	60.5	45.3	45.9	59.2

媲美监督方法

工作1：面向面部动作检测的无监督学习

- **方法**：以视频中的面部动作为自监督Pretext task
- **损失函数**：双反变换重构损失(Cycle loss)+前向重构误差
- **效果**：8层轻量级网络(无监督)+单FC层分类**媲美复杂方法**

GFT数据集，F1 score

	Methods/AU	1	2	4	6	10	12	14	15	23	24	ave
Descriptor	Handcrafted [14]*	38	32	13	67	64	78	15	29	49	44	42.9
	ResNet-50 face	24.3	50.7	18.2	39.9	44.7	41.6	17.4	27.8	31.0	25.9	32.2
	VGG emotion	23.9	40.6	26.4	73.6	69.3	74.4	21.1	24.9	26.0	20.2	40.0
Supervised	AlexNet [14]*	44	46	2	73	72	82	5	19	43	42	42.8
	ResNet-50	23.5	37.8	3.5	79.1	70.1	82.1	20.9	11.7	49.1	40.3	41.8
Self-supervised	SplitBrain [26]	19.0	40.6	8.7	60.2	66.6	75.4	5.6	26.7	22.9	32.3	35.8
	DeformAE [47]	17.3	40.1	4.8	64.1	69.1	72.1	7.8	3.9	8.0	25.2	31.2
	Fab-Net [24]	44.4	42.3	9.4	60.6	68.7	70.4	8.7	1.7	5.5	20.8	33.3
	TAE (w/o L_1, w/o att.)	43.7	41.0	5.9	71.9	72.3	78.4	11.8	23.5	24.5	35.1	40.8
	TAE (w/ L_1, w/o att.) [53]	43.9	49.5	6.3	71.0	76.2	79.5	10.7	28.5	34.5	41.7	44.2
	TAE (w/ L_1 & att., $K = 2$)	38.6	51.1	8.9	75.9	72.4	80.3	14.8	43.1	35.5	42.9	46.3
	TAE (w/ L_1 & att., $K = 4$)	33.8	45.0	5.9	**77.8**	74.3	82.6	12.2	43.7	41.1	45.1	46.1
	***TAE** (w/ L_1 & att., $K = 8$)*	**46.3**	**48.8**	13.4	76.7	74.8	81.8	19.9	**42.3**	**50.6**	**50.0**	**50.5**
	TAE (w/ L_1 & att., $K = 16$)	45.8	54.1	**14.6**	78.3	**76.3**	**83.1**	**20.1**	41.6	37.4	43.6	49.5

超越监督方法

工作1：面向面部动作检测的无监督学习

- 提出了一种自监督学习方法
  - 图像变换T的变与不变：pretext task
  - 特征解耦/解卷(Disentangling)：交叉组合再验证
- 启示
  - Pretext task/auxiliary task/surrogate tasks的挖掘和定义是无监督学习的关键；Pretext task或许可以考虑设定为基础性的、已经解决的比较好的任务
  - 视频中的运动信息对无监督学习大有可为

    VALSE B站视频：https://space.bilibili.com/562085182/

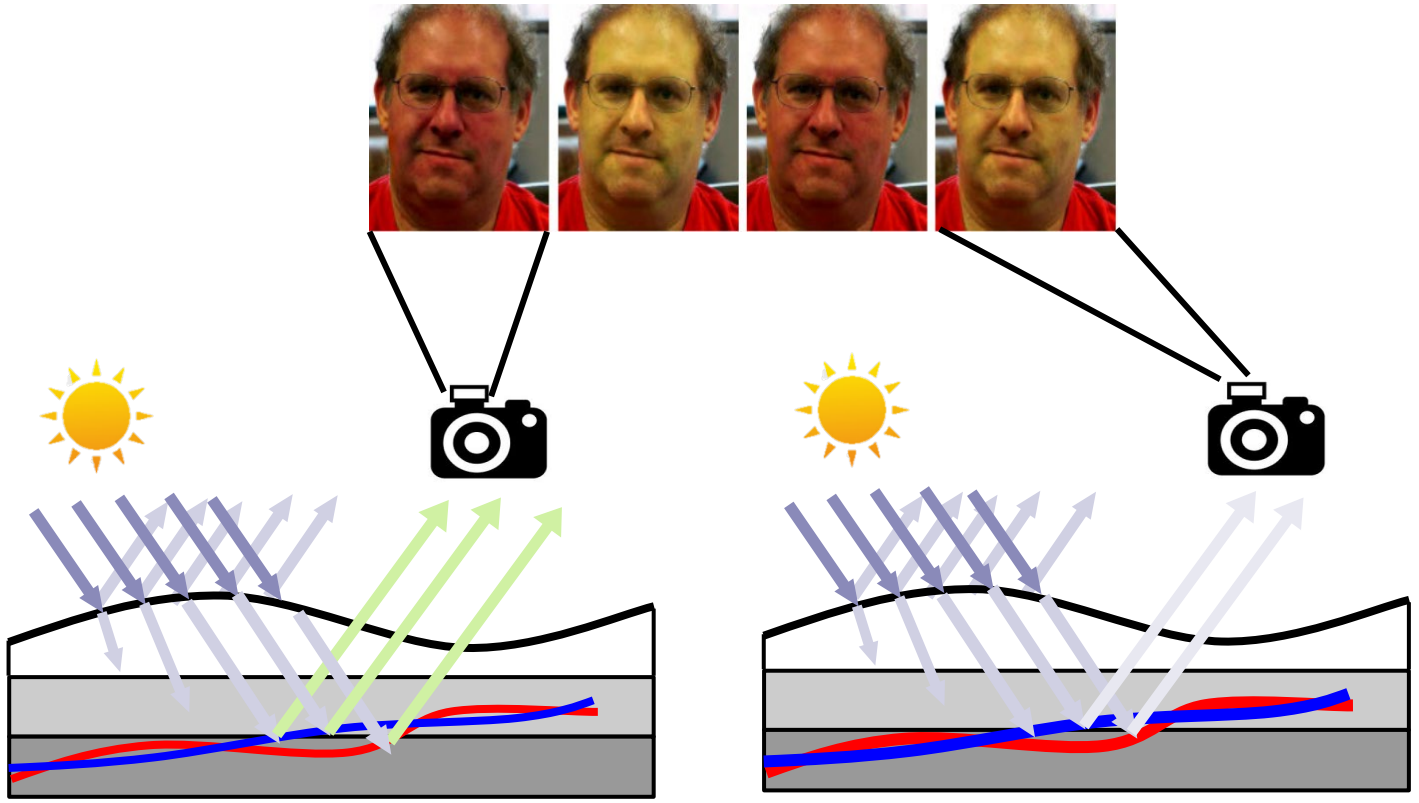
Xuesong Niu, Shiguang Shan, Hu Han, Xilin Chen. RhythmNet: End-to-end Heart Rate Estimation from Face via Spatial-temporal Representation. IEEE Transactions on Image Processing. vol. 29, pp. 2409-2423, 2020

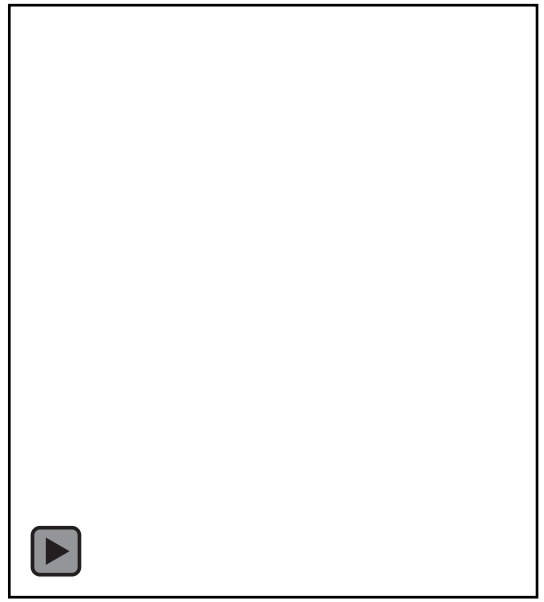
Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, Guoying Zhao, "Video-based Remote Physiological Measurement via Cross-verified Feature Disentangling," European Conference on Computer Vision, Aug. 23-28, 2020.

工作2：基于交叉验证解耦的生理信号估计

基于视频的远距、无接触生理信号测量

- 原理





Magnified skin color changes due to heartbeat using the EVM algorithm[1]

[1] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, “Eulerian video magnification for revealing subtle changes in the world,” ACM Trans. Graph, vol. 31, no. 4, pp. 1–8, 2012.

挑战

- The physiological signal in face video is **very weak**, and it can be easily affected by head movements, lighting conditions, and sensor diversities.
- **Datasets are usually small**, could not provide enough data of various less-constrained scenarios.

One possible way is to use pseudo data for regularization and augmentation.

核心思想——正确的解耦信号和噪声

- 输入视频x，解耦其中的生理信号p和其他信号n，然后交叉合并，之后再解耦，验证解耦的正确性
- 给定x_1和x_2
  - 解耦x_1和x_2

$$x_1 \rightarrow p_{x1} + n_{x1}, \quad x_2 \rightarrow p_{x2} + nx_2$$

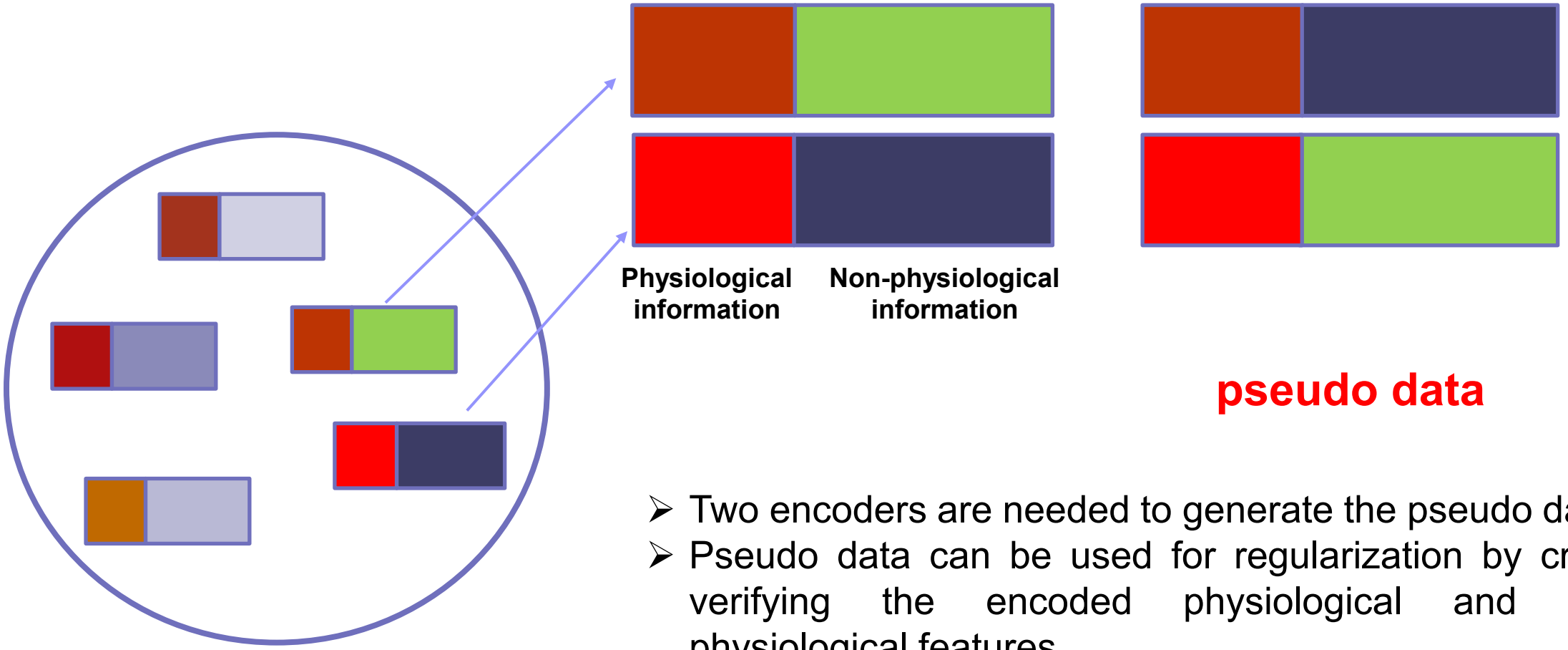
  - 交换n1和n2

$$y_1 = px_1 + n_{x2}, \quad y_2 = p_{x2} + nx_1$$

  - 解耦y1和y2

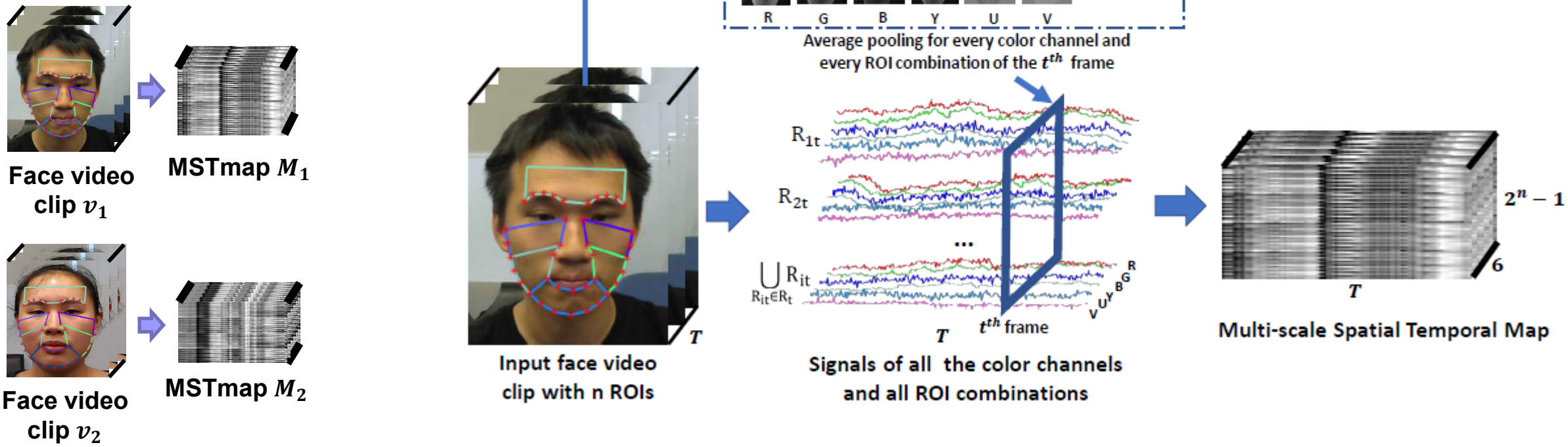
$$y_1 \rightarrow p'_{x1} + n'_{x2}, \quad y_2 \rightarrow p'_{x2} + n'_{x1}$$

核心思想——伪数据

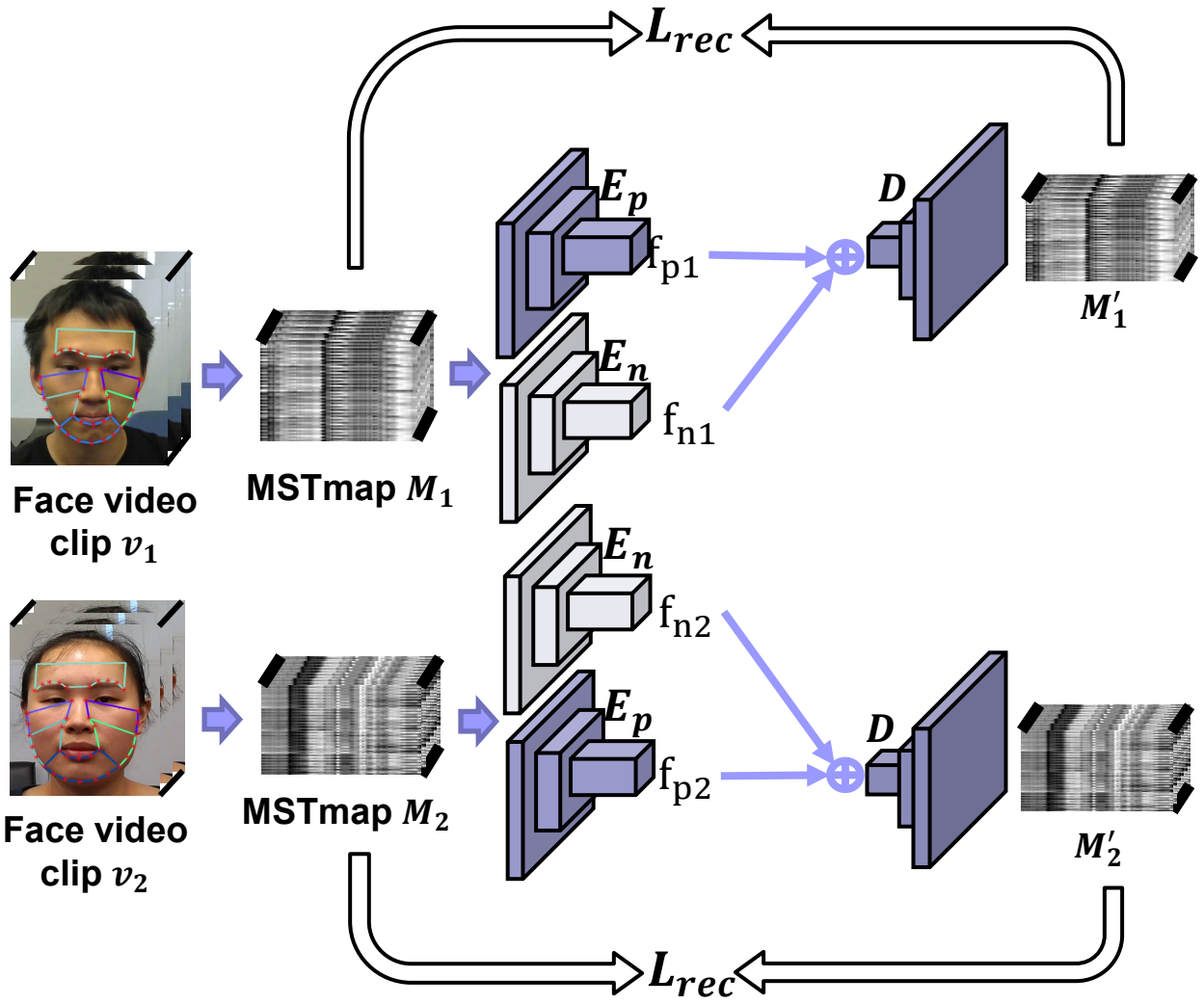


- Two encoders are needed to generate the pseudo data.
- Pseudo data can be used for regularization by cross-verifying the encoded physiological and non-physiological features.
- Pseudo data can be used as the augmentation to train the physiological estimator.

信号提取


Face video clip v_1
MSTmap M_1
Face video clip v_2
MSTmap M_2
t^{th} frame
All ROI combinations
Color channels
R
G
B
Y
U
V
Average pooling for every color channel and every ROI combination of the t^{th} frame
Input face video clip with n ROIs
t^{th} frame
Signals of all the color channels and all ROI combinations
$2^n - 1$
6
Multi-scale Spatial Temporal Map

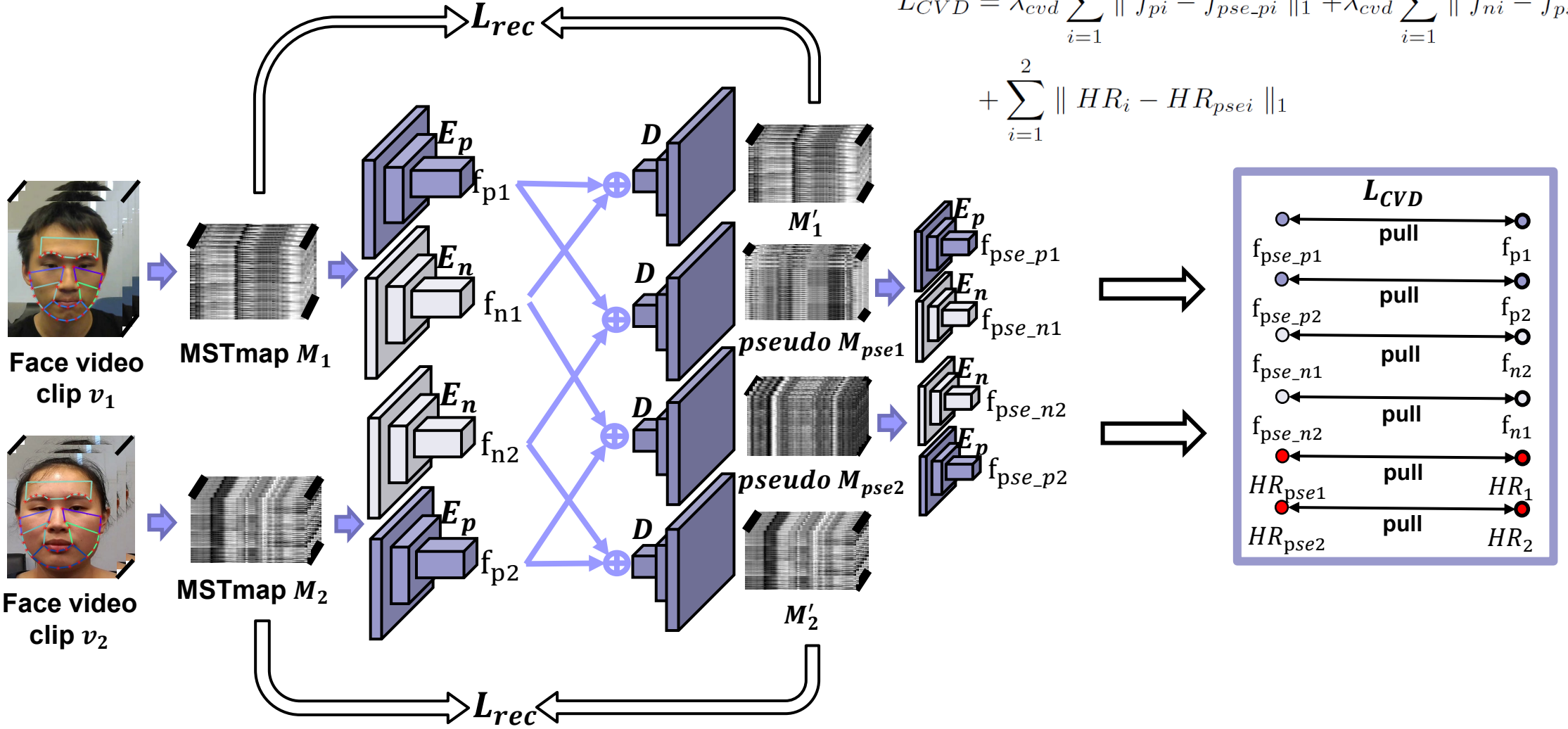
Proposed Method



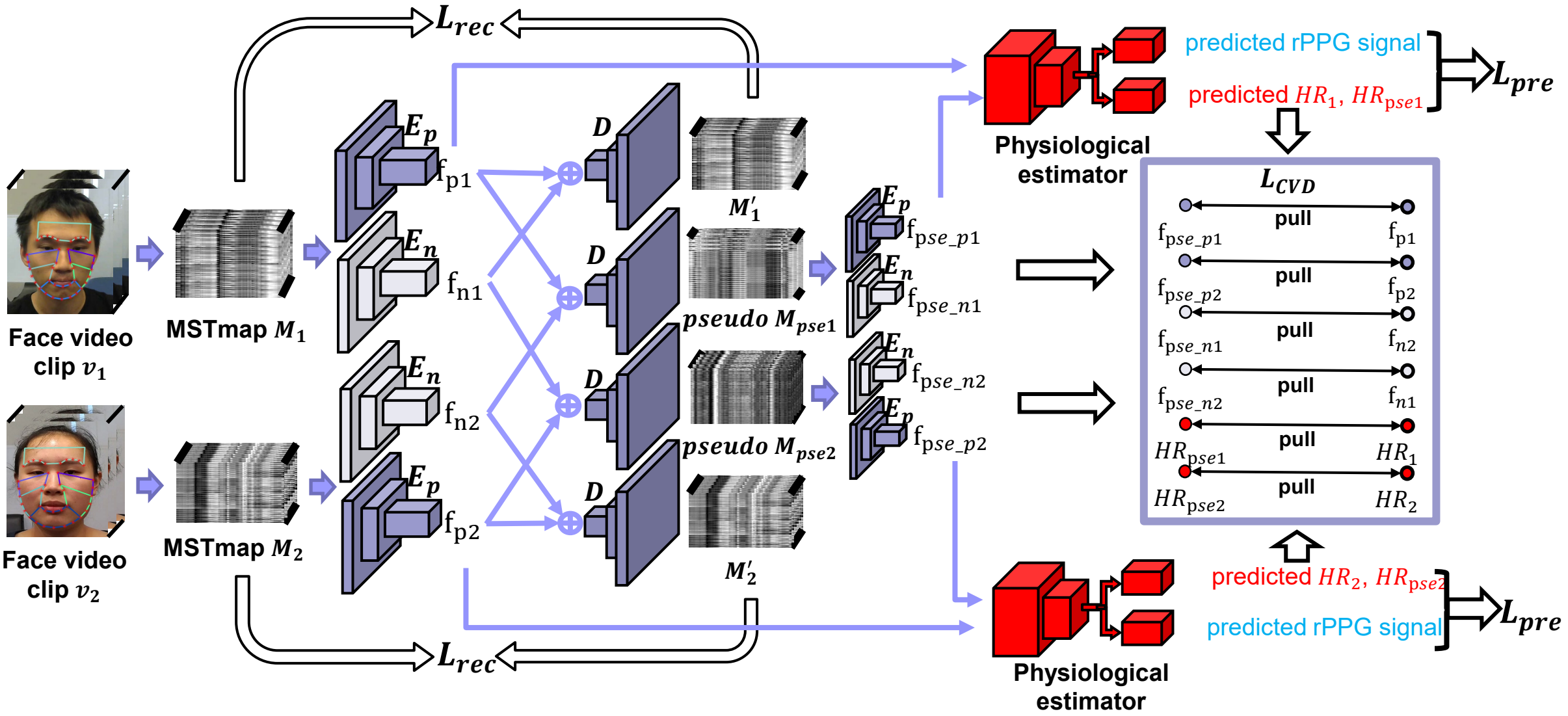
$$L_{rec} = \lambda_{rec} \sum_{i=1}^{2} \| M_i - M_i^{'} \|_1$$

Proposed Method

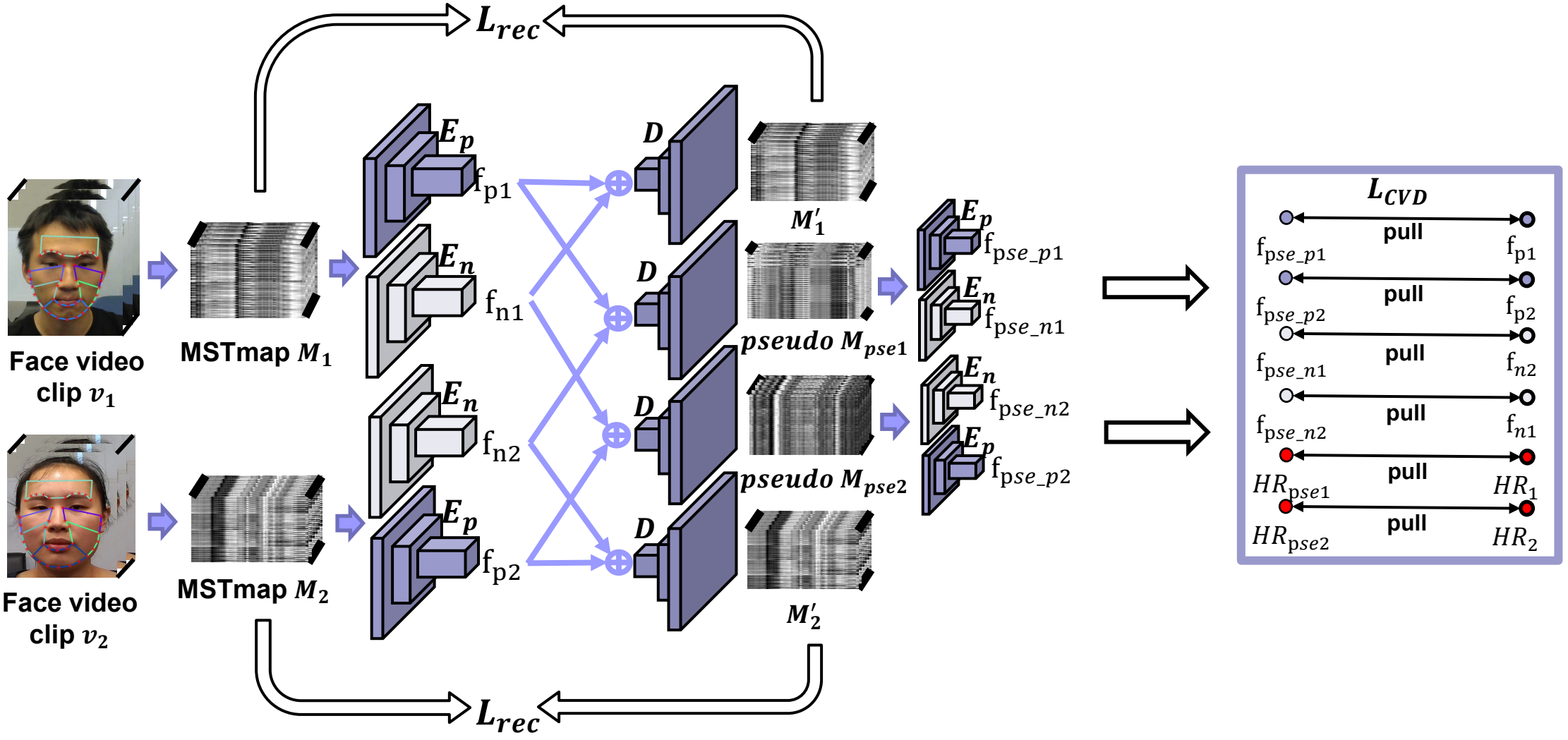
$$L_{CVD} = \lambda_{cvd} \sum_{i=1}^{2} \| f_{pi} - f_{pse\_pi} \|_1 + \lambda_{cvd} \sum_{i=1}^{2} \| f_{ni} - f_{pse\_n(3-i)} \|_1 + \sum_{i=1}^{2} \| HR_i - HR_{psei} \|_1$$



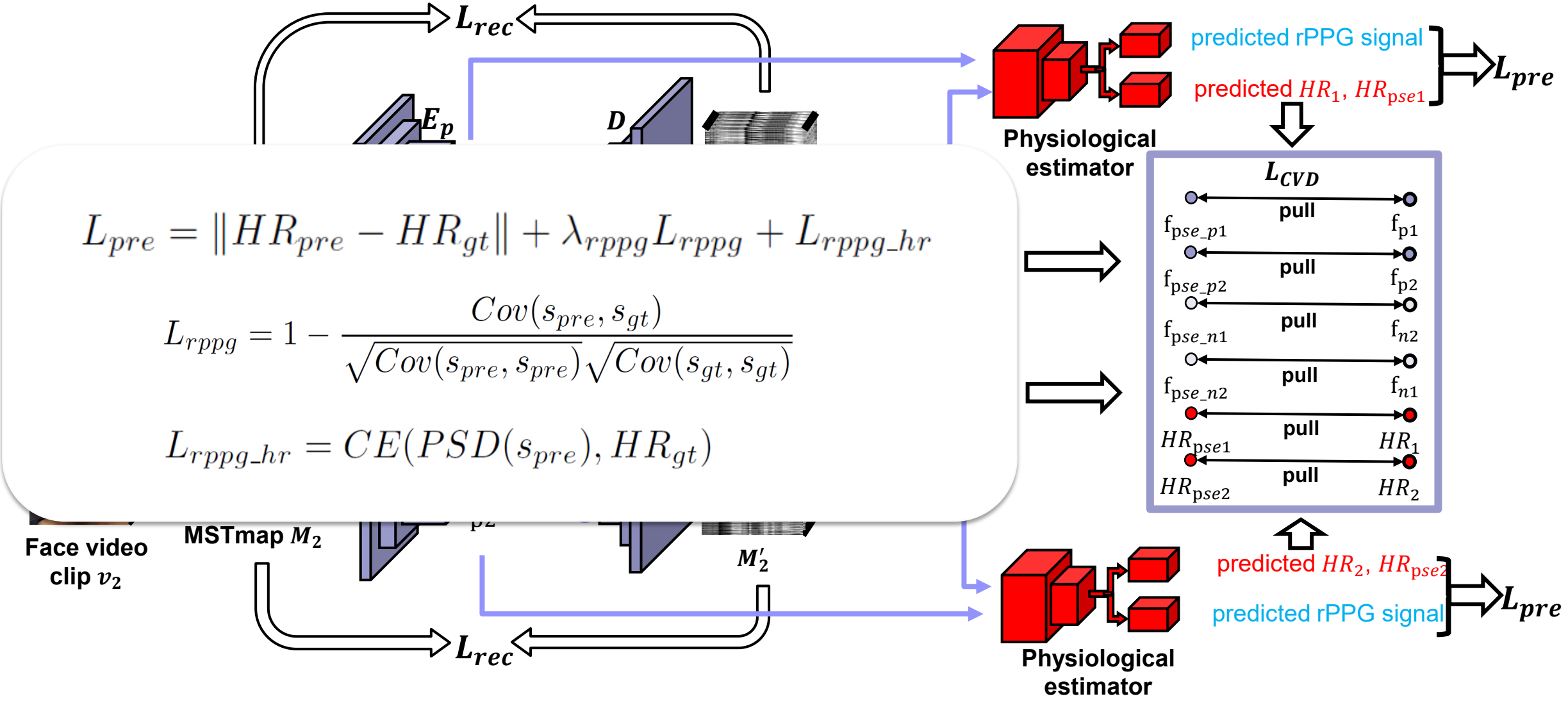
Proposed Method



Proposed Method



Proposed Method



$$L_{pre} = \|HR_{pre} - HR_{gt}\| + \lambda_{rppg} L_{rppg} + L_{rppg\_hr}$$

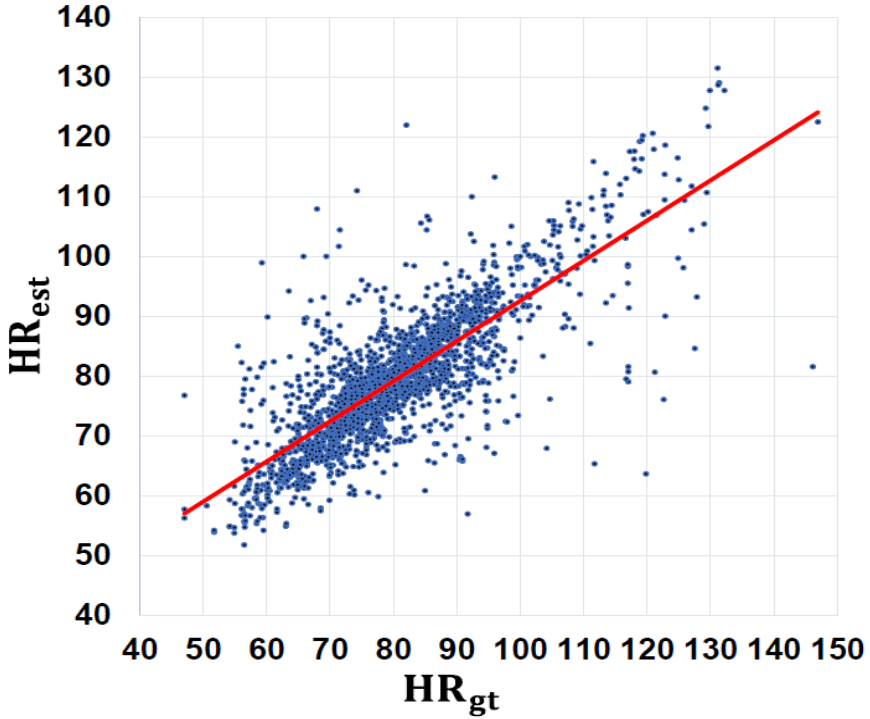
$$L_{rppg} = 1 - \frac{Cov(s_{pre}, s_{gt})}{\sqrt{Cov(s_{pre}, s_{pre})}\sqrt{Cov(s_{gt}, s_{gt})}}$$

$$L_{rppg\_hr} = CE(PSD(s_{pre}), HR_{gt})$$

Results on Average HR Estimation

- 5-fold subject-exclusive intra-database testing on VIPL-HR

Method	Std↓ (bpm)	MAE↓ (bpm)	RMSE↓ (bpm)	r ↑
SAMC	18.0	15.9	21.0	0.11
POS	15.3	11.5	17.2	0.30
CHROM	15.1	11.4	16.9	0.28
I3D	15.9	12.0	15.9	0.07
DeepPhy	13.6	11.0	13.8	0.11
RhythmNet	8.11	5.30	8.14	0.76
Proposed	**7.92**	**5.02**	**7.97**	**0.79**



Results on Average HR Estimation

- 10-fold subject-exclusive intra-database testing on OBF

Method	Std↓ (bpm)	RMSE↓ (bpm)	r ↑
ROI_{green}	2.159	2.162	0.99
CHROM	2.73	2.733	0.98
POS	1.899	1.906	0.991
rPPGNet	1.758	1.8	0.992
Proposed	**1.257**	**1.26**	**0.996**

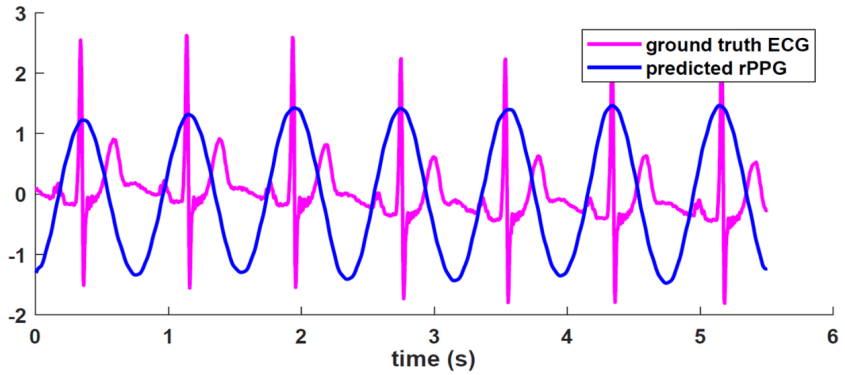
- Cross-database testing on MMSE-HR (training on VIPL-HR)

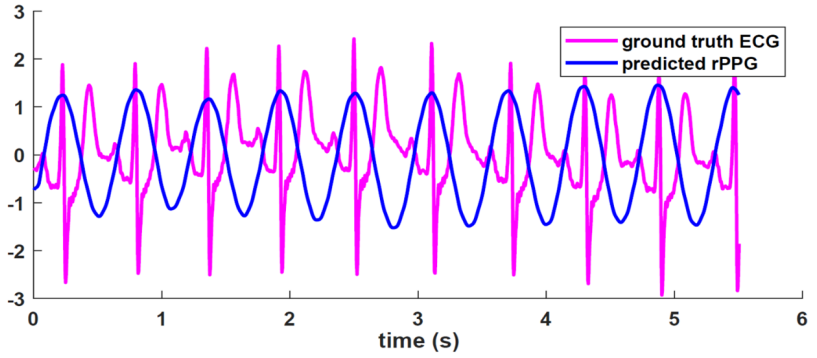
Method	Std↓ (bpm)	RMSE↓ (bpm)	r ↑
Li2014	20.02	19.95	0.38
CHROM	14.08	13.97	0.55
SAMC	12.24	11.37	0.71
RhythmNet	6.98	7.33	0.78
Proposed	**6.06**	**6.04**	**0.84**

Results on RF Measurement and HRV Analysis

- 10-fold subject-exclusive intra-database testing on OBF

Method	RF(Hz)			LF(u.n)			HF(u.n)			LF/HF		
	Std	RMSE	r	Std	RMSE	r	Std	RMSE	r	Std	RMSE	r
ROI_{green}	0.078	0.084	0.321	0.22	0.24	0.573	0.22	0.24	0.573	0.819	0.832	0.571
CHROM	0.081	0.081	0.224	0.199	0.206	0.524	0.199	0.206	0.524	0.83	0.863	0.459
POS	0.07	0.07	0.44	0.155	0.158	0.727	0.155	0.158	0.727	0.663	0.679	0.687
rPPGNet	0.064	0.064	0.53	0.133	0.135	0.804	0.133	0.135	0.804	0.58	0.589	0.773
Proposed	**0.058**	**0.058**	**0.606**	**0.09**	**0.09**	**0.914**	**0.09**	**0.09**	**0.914**	**0.453**	**0.453**	**0.877**





工作2：小结

- 基于解耦的rPPG信号提取
  - 自监督：交叉验证解耦（类比：汉语➔英语➔汉语）
  - 深度编码器，解耦，验证

Code

https://github.com/nxsEdson/CVD-Physiological-Measurement

Download VIPL-HR

https://vipl.ict.ac.cn/view_database.php?id=15

Download VIPL-HR-V2

https://vipl.ict.ac.cn/view_database.php?id=17

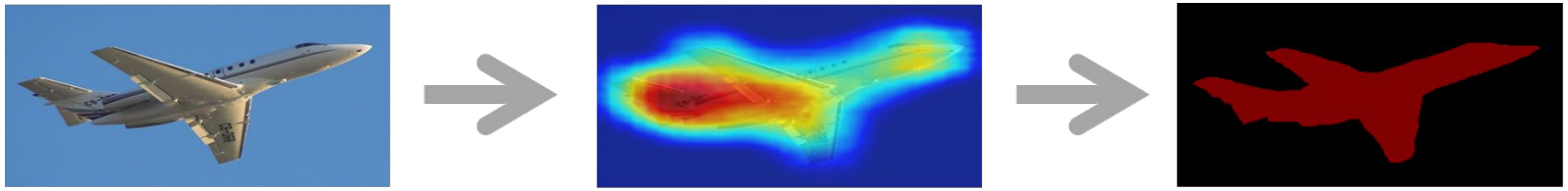
工作3：基于同变性自监督的弱监督分割方法

Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, Xilin Chen. Self-supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation. IEEE/CVF International Conference on Computer Vision and Pattern Recognition. Jun. 2020

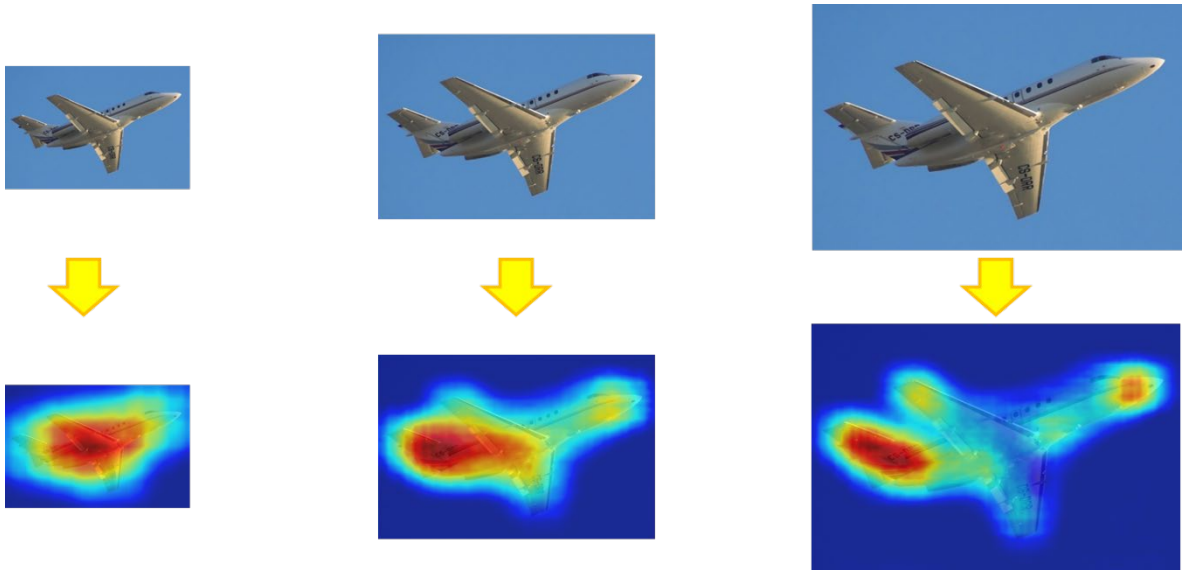
基于同变性的弱监督分割方法

- 思路：以不同变换图像的分割图一致性设计Loss

$$F(T1(x)) = F(T2(x)) = F(T3(x))$$



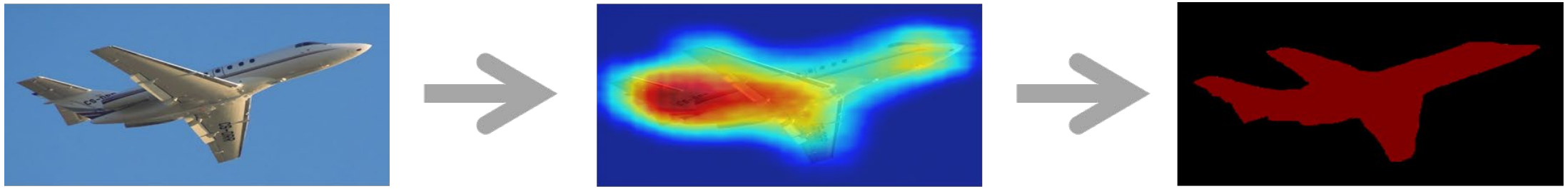
输入图像　　类别响应图　　像素级别伪标签



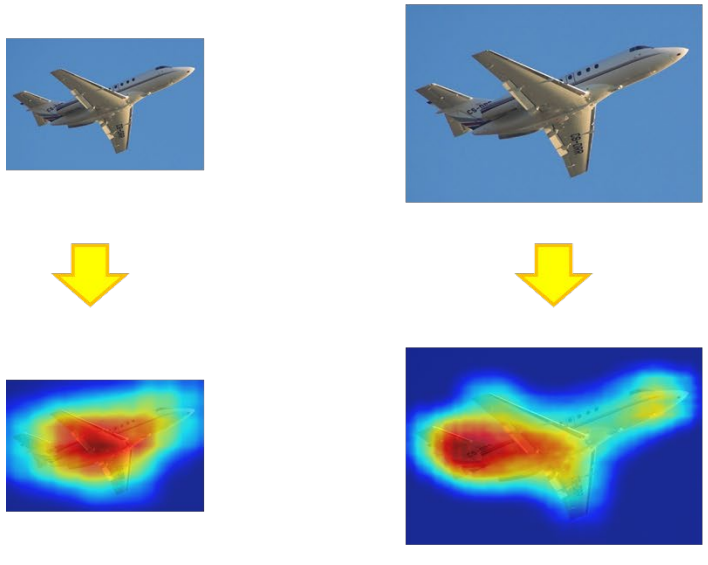
问题：输入图像的尺度不一致，会产生不一致的CAM

基于同变性的弱监督分割方法

- 由类别标签训练得到的类别响应图（CAM）常作为弱监督语义分割的中间结果，并以此为基础进行调整获得像素级伪标签



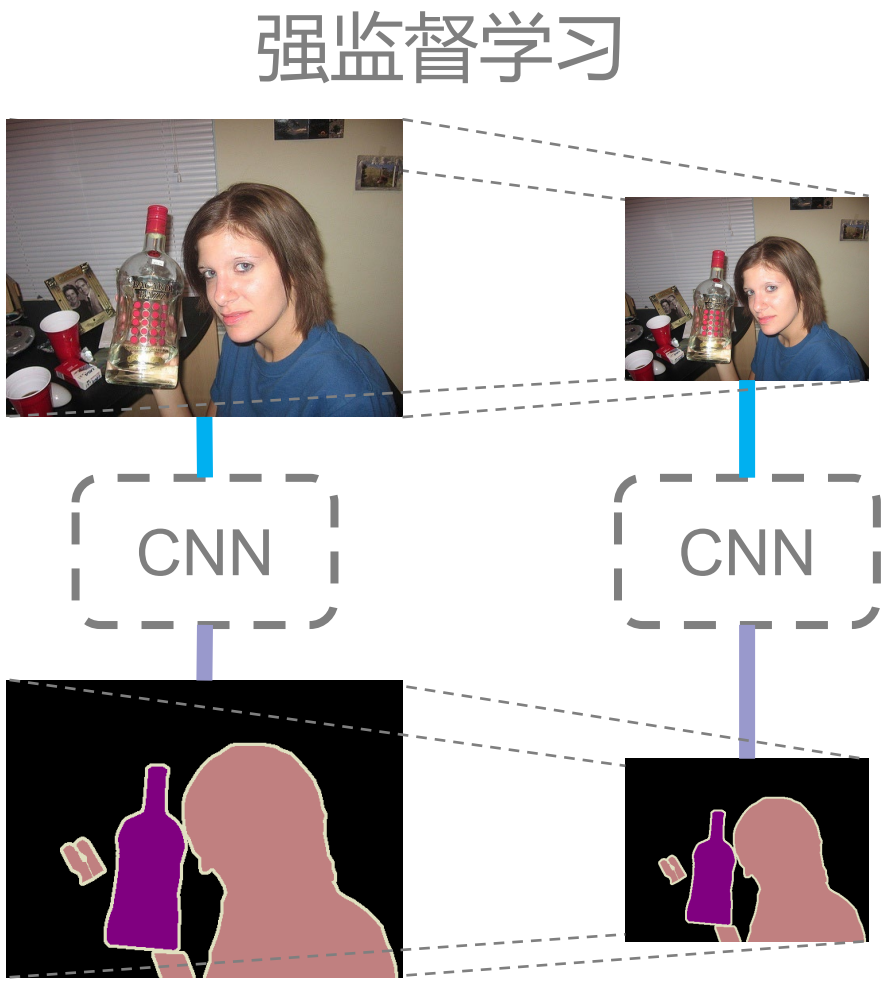
输入图像　　类别响应图　　像素级别伪标签

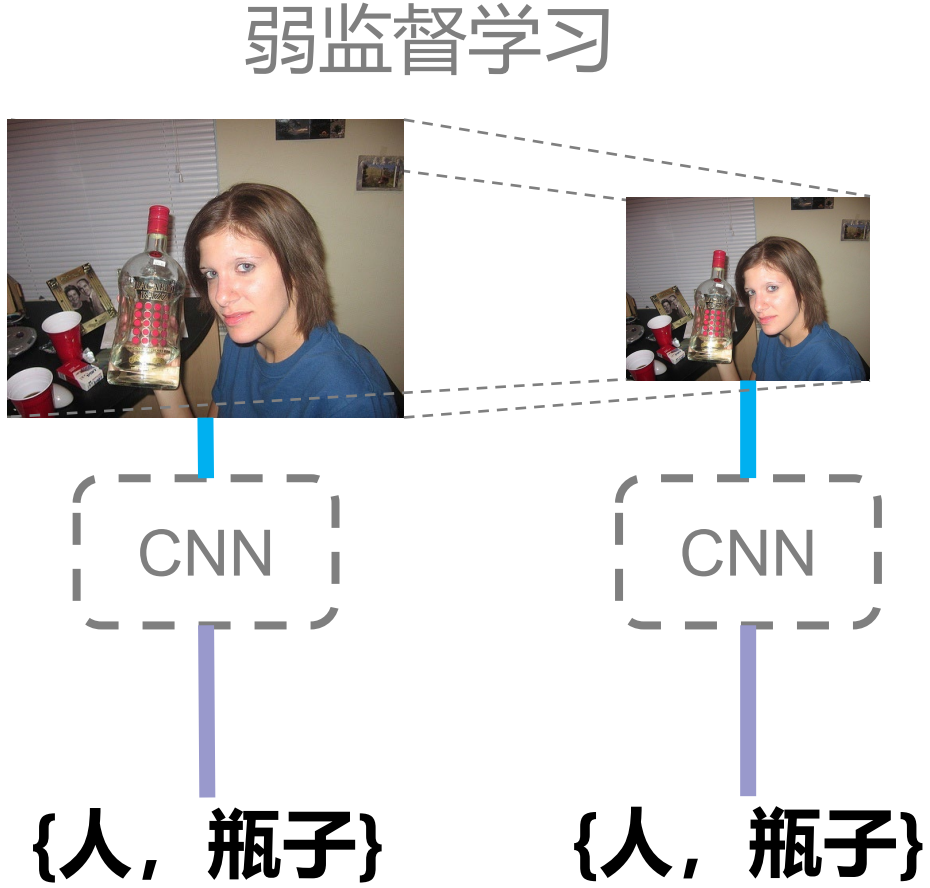


输入图像的尺度不一致
会产生不一致的CAM

基于同变性的弱监督分割方法

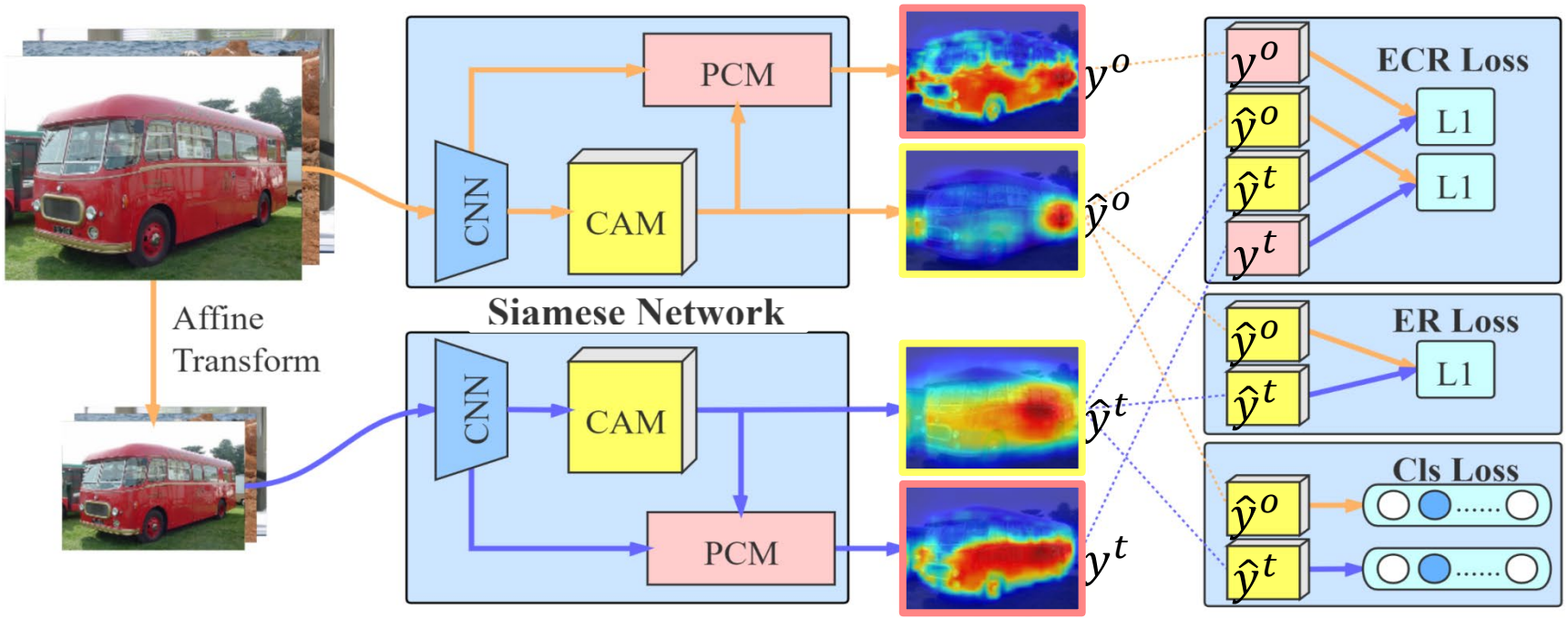
- CAM的不一致性来源于数据增广时的**同变性约束缺失**





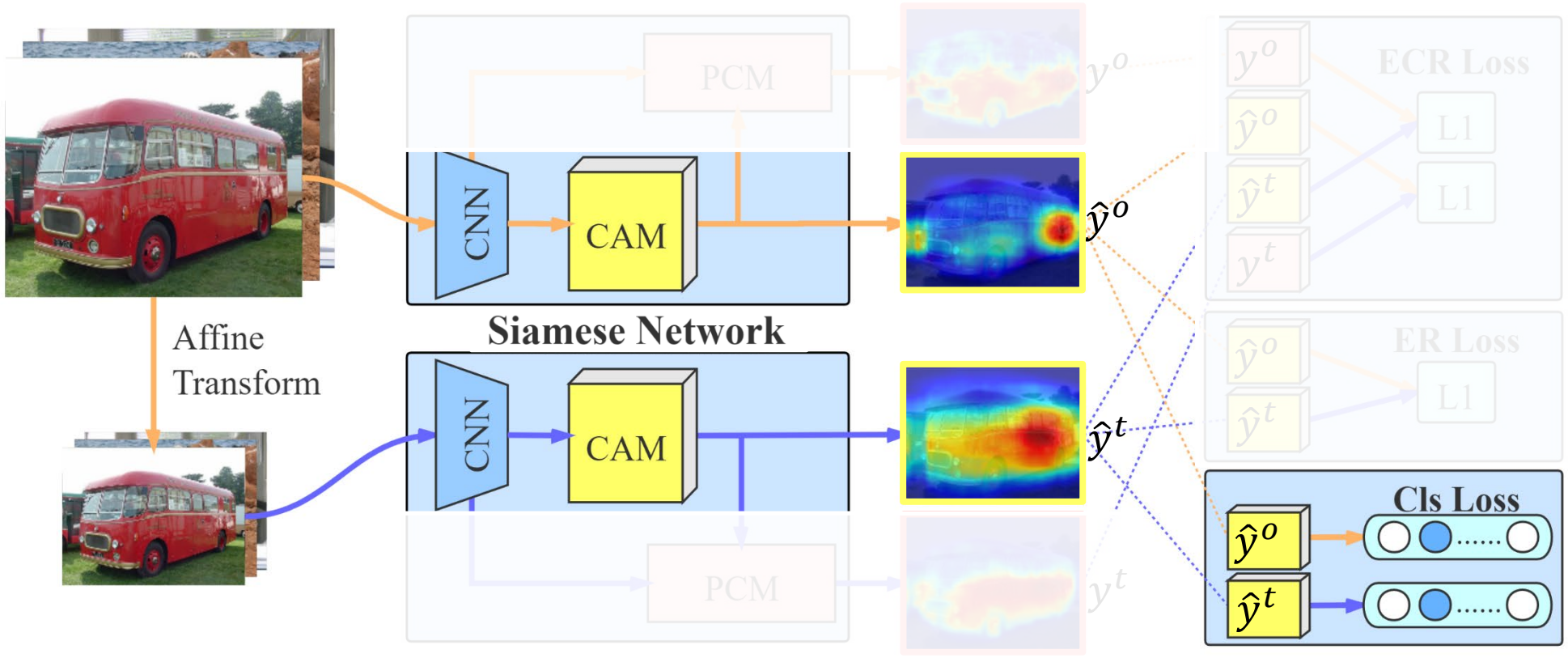
基于同变性的弱监督分割方法

- 自监督同变注意力机制（SEAM）
  - 利用**孪生网络**结构生成不同尺度的CAM
  - 为了增强网络对于同变性的学习能力提出**PCM**对CAM进行调整
  - 利用**同变性交错损失**来保证调整后的CAM能够收敛到合理结果



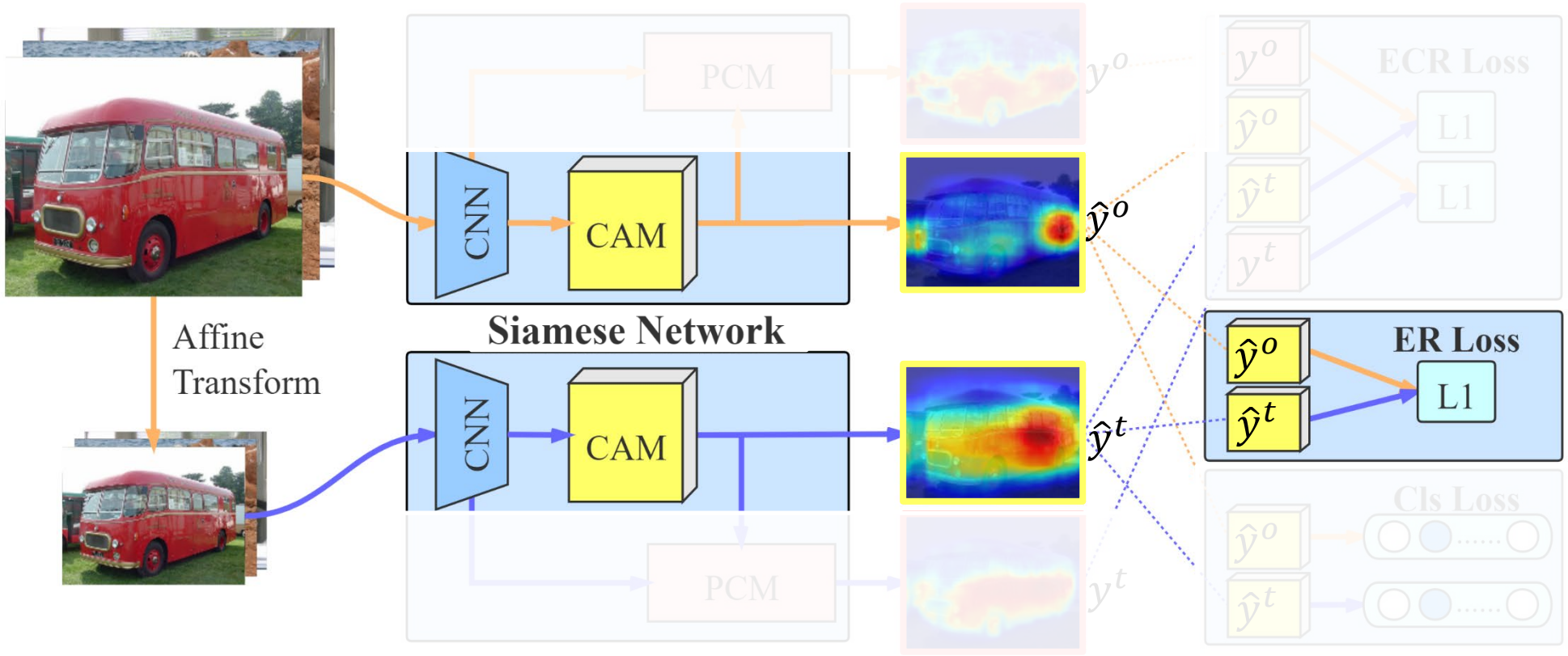
基于同变性的弱监督分割方法

- 自监督同变注意力机制（SEAM）
  - 分类损失（Cls Loss）



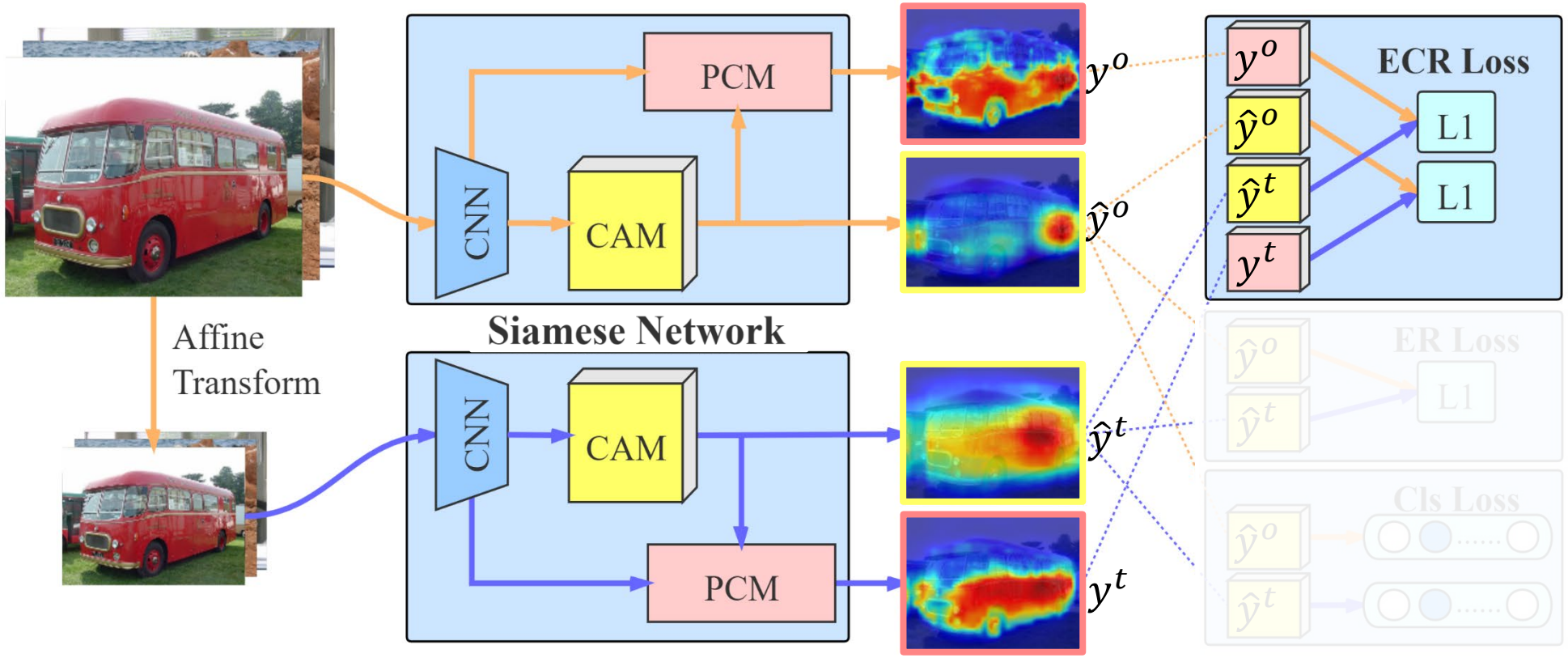
基于同变性的弱监督分割方法

- 自监督同变注意力机制（SEAM）
  - 同变性正则损失（ER Loss）



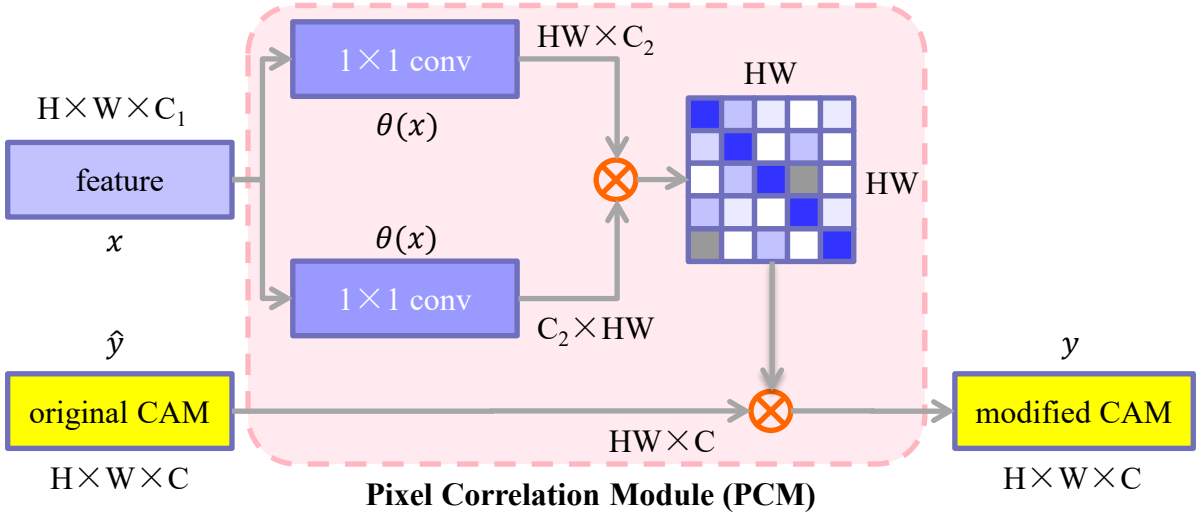
基于同变性的弱监督分割方法

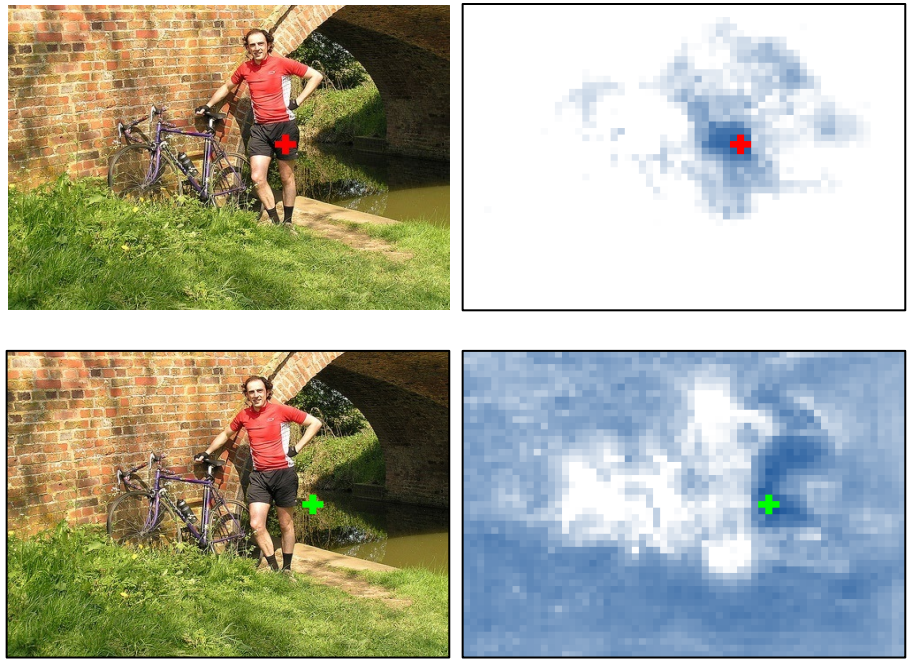
- 自监督同变注意力机制（SEAM）
  - 同变性交错正则损失（ECR Loss）



基于同变性的弱监督分割方法

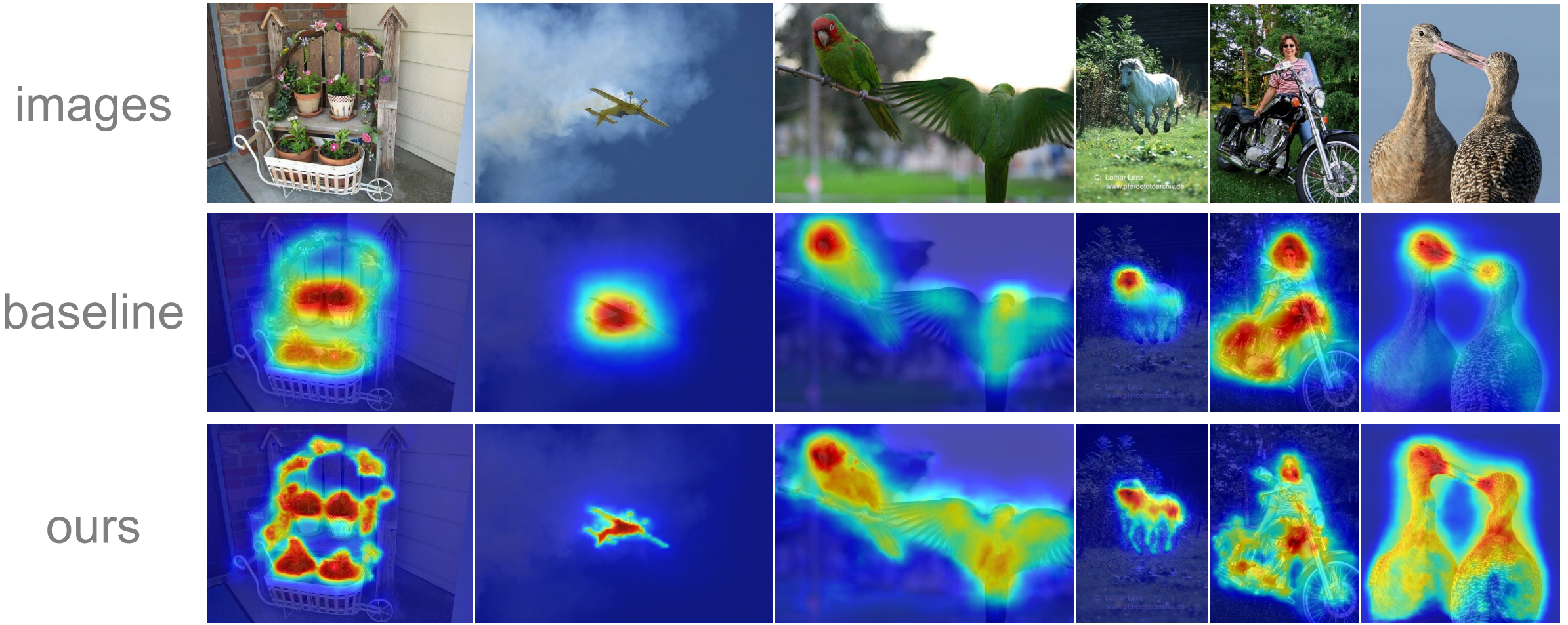
- 像素相关性模块（PCM）源于传统attention结构
- PCM能够利用同变性约束这一自监督信号学习像素间的相似度，并通过像素相似度加权方式对原始CAM进行调优





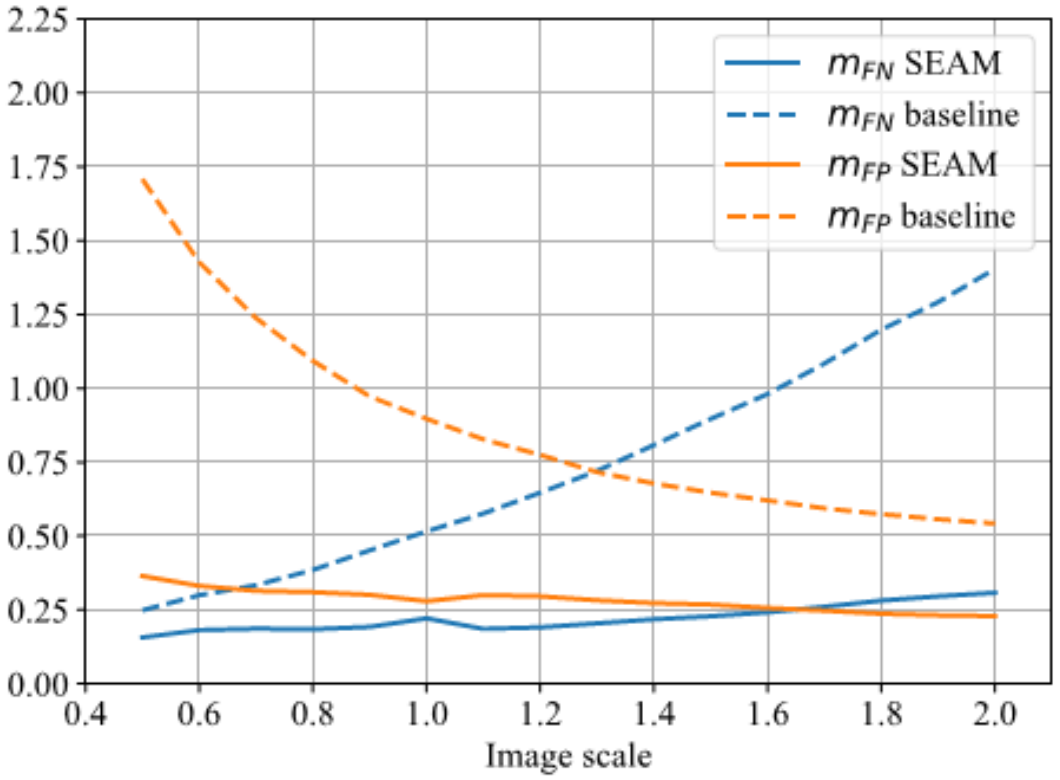
基于同变性的弱监督分割方法

- SEAM能够有效地减少CAM中的**欠激活**与**过激活**



基于同变性的弱监督分割方法

- 统计结果显示SEAM生成的CAM能更全地覆盖整个物体且在输入图像尺度不一致的情况下有更好的一致性



FN_c c类假阴性像素数

FP_c c类假阳性像素数

TP_c c类真阳性像素数

$$m_{FN} = \frac{1}{C-1}\sum_{c=1}^{C-1}\frac{FN_c}{TP_c},$$

$$m_{FP} = \frac{1}{C-1}\sum_{c=1}^{C-1}\frac{FP_c}{TP_c}.$$

基于同变性的弱监督分割方法

- 实验：PASCAL VOC 2012
  - 尺度变换对SEAM性能的提升贡献最大
  - SEAM性能提升与尺度增广范围的增加无关
  - 多尺度融合测试可进一步提升SEAM性能

rescale	flip	rotation	translation	mIoU
				47.43%
√				55.41%
√	√			55.50%
√		√		53.13%
√			√	55.23%

test scale	baseline (mIoU)	ours (mIoU)
[0.5]	40.17%	49.35%
[1.0]	46.10%	51.57%
[1.5]	47.51%	52.25%
[2.0]	46.12%	49.79%
[0.5, 1.0, 1.5, 2.0]	47.43%	55.41%

model	random rescale	mIoU
baseline	[448, 768]	47.43%
baseline	[224, 768]	46.72%
SEAM	[448, 768]	53.47%

methods	backbone	saliency	*val*	*test*
CCNN	VGG16		35.3	35.6
EM-Adapt	VGG16		38.2	39.6
MIL+seg	OverFeat		42.0	43.2
SEC	VGG16		50.7	51.1
STC	VGG16	√	49.8	51.2
AdvErasing	VGG16	√	55.0	55.7
MDC	VGG16	√	60.4	60.8
MCOF	ResNet101	√	60.3	61.2
DCSP	ResNet101	√	60.8	61.9
SeeNet	ResNet101	√	63.1	62.8
DSRG	ResNet101	√	61.4	63.2
AffinityNet	ResNet38		61.7	63.7
CIAN	ResNet101	√	64.1	64.7
IRNet	ResNet50		63.5	64.8
FickleNet	ResNet101	√	64.9	65.3
Our SEAM	ResNet38		64.5	65.7

总结

- 自监督学习任务和标签自动生成

自监督学习目标函数
$F(T_1(x_i) + T_2(x_i)) = F(x_{i+k})$ **视频帧之间关系：解耦+重构**
信号噪声的解耦和重构，交叉验证loss
$F(T_1(x)) = F(T_2(x)) = F(T_3(x))$ 变换同变性

总结与讨论

- 总结
  - What: 无监督学习的一种形式
  - Why: 视觉常识表示学习
  - How: 把不言而喻的标签给算法
- 讨论
  - 大有可为，特别是大规模训练（类别BERT, GPT）
  - 对学术界
    - 挖掘并巧妙的不言而喻的标签生成
    - 对算力需求较小的大规模训练

盲人摸象，任重道远！

谢谢！