自监督学习的三个基本范式及应用

杨 健

PCA Lab

自监督学习的概念

- 自监督学习属于无监督学习的范畴
- **概念：不提供标签数据**（无），通过构造 **辅助任务**（Proxy tasks）来挖掘数据自身的表征特性提供监督信息（有），从而构造出监督学习问题的求解形式（没有标签却能构造出目标函数）
- 核心目标：为下游任务提供泛化性更强的表示（representation）
- 借助于自监督学习，近年来表示学习取得了长足的进展
- 成功的典范：从自然语言大模型到视觉基础模型

表示学习历久弥新的成功范式

- PCA+LDA (Fisherfaces, 1997)
- Layer-by-layer Pretraining + Fine-tuning（Hinton, Science 2006）
- Self-supervised Pretraining + Fine-tuning (Current)

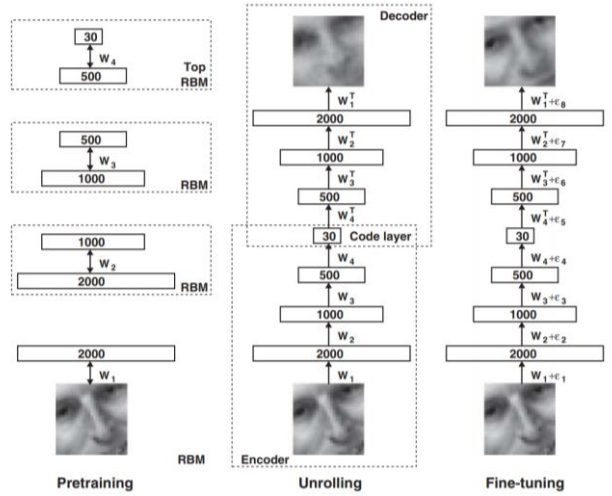
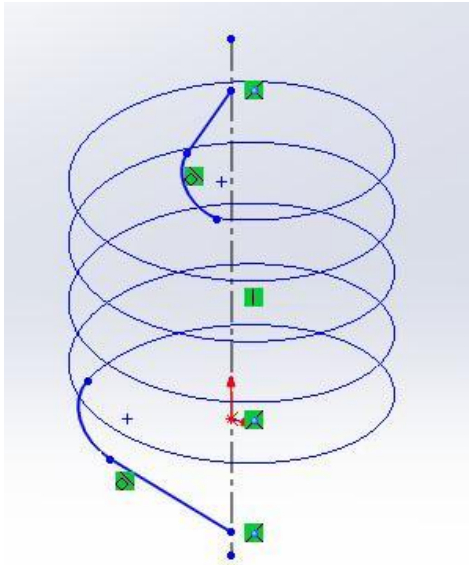


Fig. 1. Pretraining consists of learning a stack of restricted Boltzmann machines (RBMs), each having only one layer of feature detectors. The learned feature activations of one RBM are used as the "data" for training the next RBM in the stack. After the pretraining, the RBMs are "unrolled" to create a deep autoencoder, which is then fine-tuned using backpropagation of error derivatives.



Self-supervised Pretraining + Fine-tuning

- 符合人类的学习规律（自监督为主、监督为辅）
- 为小样本问题提供了可行的途径
- 为多任务问题提供了统一的基础

主要内容

(1) 对比范式

样本之间的相似性与差异性（“求于外”）

(2) 掩码范式

样本内在的上下文信息（“求于内”）

(3) 内隐范式

空间几何关系

对比范式

- 基本思想：“不怕不识货（非监督），只怕货比货（对比）”
- 基本策略：相似性（similarity）、差异性（ dis-similarity）
- 基本手段：正样本对、负样本对

对比学习的损失函数

- Noise-Contrastive Estimation(**NCE**) **loss** (or called the normalized temperature-scaled cross entropy loss)

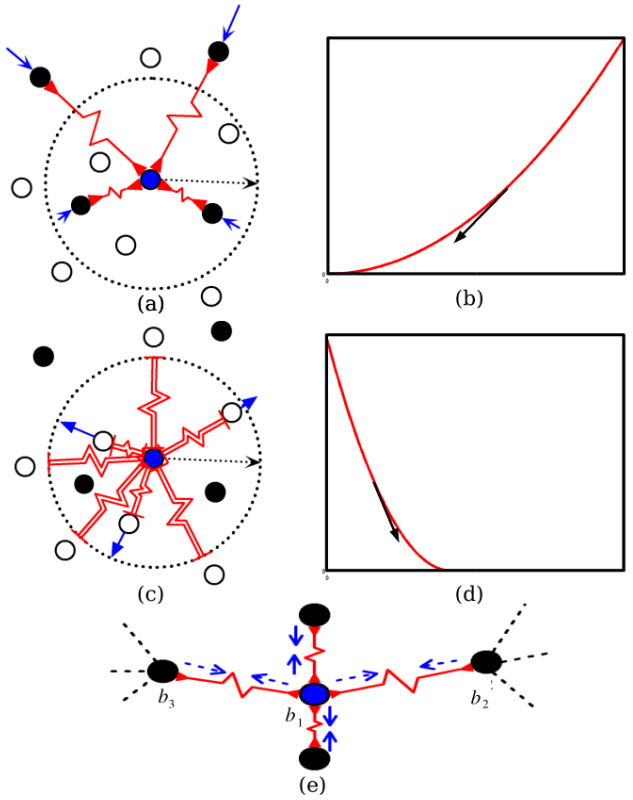
$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

$$\text{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} / \|\boldsymbol{u}\| \|\boldsymbol{v}\|$$

τ denotes a temperature parameter

对比的思想渊源

- **Class Discrimination**
  Fisher Discriminant (supervised)
- **Cluster Discrimination**
  Unsupervised Discriminant Projection (PAMI2007)
- **Dimensionality reduction by Invariant Mapping**, LeCun等首次给出了Contrastive Loss Function（CVPR2006）



Hadsell, R., Chopra, S., & LeCun, Y. Dimensionality reduction by learning an invariant mapping. *CVPR 2006*.

Instance discrimination

- **PCA** (A special case of Instance Discrimination)
- Unsupervised Feature Learning via Non-Parametric **Instance Discrimination** (CVPR2018)

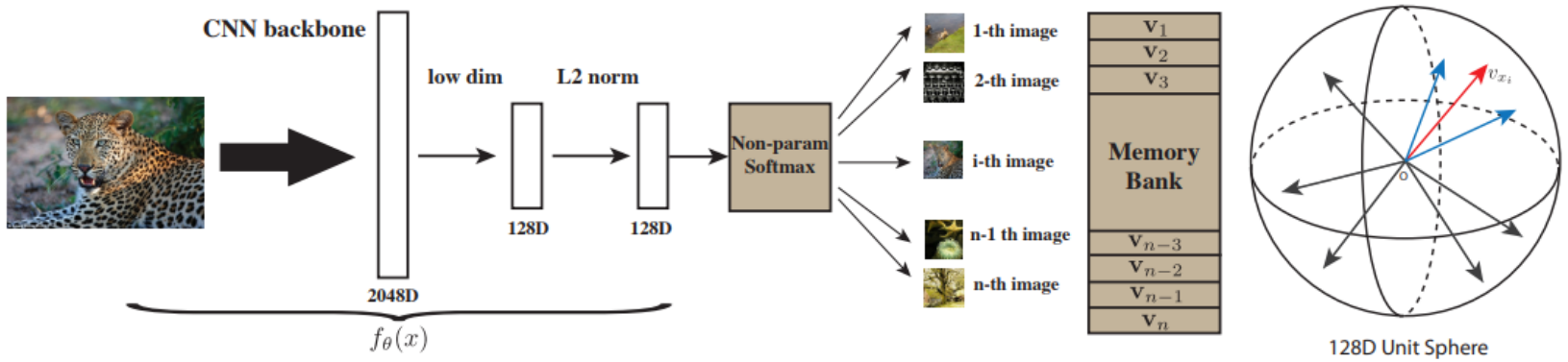


Figure 2: The pipeline of our unsupervised feature learning approach. We use a backbone CNN to encode each image as a feature vector, which is projected to a 128-dimensional space and L2 normalized. The optimal feature embedding is learned via instance-level discrimination, which tries to maximally scatter the features of training samples over the 128-dimensional unit sphere.

Z. Wu, et al., Unsupervised feature learning via non-parametric instance discrimination. CVPR 2018.

动量对比：Momentum Contrast（Moco）

- MoCo基于Instance discrimination的思想，引入动量更新encoder的机制，解决memory bank过大时的工程问题
- **揭示了在下游的物体检测/分割等任务中，自监督的预训练可以超越监督的预训练的性能**

pre-train	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
random init.	31.0	49.5	33.2	28.5	46.8	30.4
super. IN-1M	38.9	59.6	42.7	35.4	56.5	38.1
MoCo IN-1M	38.5 (−0.4)	58.9 (−0.7)	42.0 (−0.7)	35.1 (−0.3)	55.9 (−0.6)	37.7 (−0.4)
MoCo IG-1B	38.9 (0.0)	59.4 (−0.2)	42.3 (−0.4)	35.4 (0.0)	56.5 (0.0)	37.9 (−0.2)

(a) Mask R-CNN, R50-**FPN**, **1**× schedule

AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
36.7	56.7	40.0	33.7	53.8	35.9
40.6	61.3	44.4	36.8	58.1	39.5
40.8 (+0.2)	61.6 (+0.3)	44.7 (+0.3)	36.9 (+0.1)	58.4 (+0.3)	39.7 (+0.2)
41.1 (+0.5)	61.8 (+0.5)	45.1 (+0.7)	37.4 (+0.6)	59.1 (+1.0)	40.2 (+0.7)

(b) Mask R-CNN, R50-**FPN**, **2**× schedule

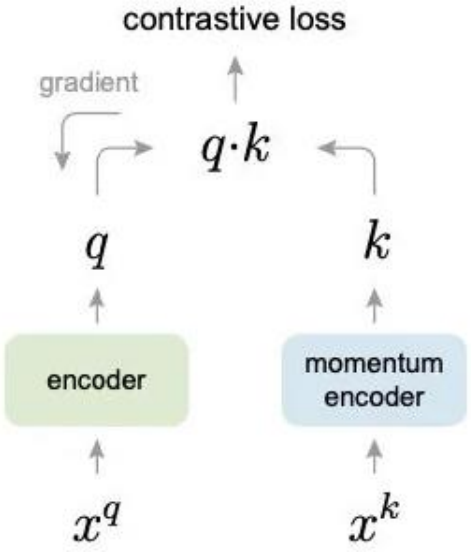
pre-train	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
random init.	26.4	44.0	27.8	29.3	46.9	30.8
super. IN-1M	38.2	58.2	41.2	33.3	54.7	35.2
MoCo IN-1M	38.5 (+0.3)	58.3 (+0.1)	41.6 (+0.4)	33.6 (+0.3)	54.8 (+0.1)	35.6 (+0.4)
MoCo IG-1B	39.1 (+0.9)	58.7 (+0.5)	42.2 (+1.0)	34.1 (+0.8)	55.4 (+0.7)	36.4 (+1.2)

(c) Mask R-CNN, R50-**C4**, **1**× schedule

AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
35.6	54.6	38.2	31.4	51.5	33.5
40.0	59.9	43.1	34.7	56.5	36.9
40.7 (+0.7)	60.5 (+0.6)	44.1 (+1.0)	35.4 (+0.7)	57.3 (+0.8)	37.6 (+0.7)
41.1 (+1.1)	60.7 (+0.8)	44.8 (+1.7)	35.6 (+0.9)	57.4 (+0.9)	38.1 (+1.2)

(d) Mask R-CNN, R50-**C4**, **2**× schedule

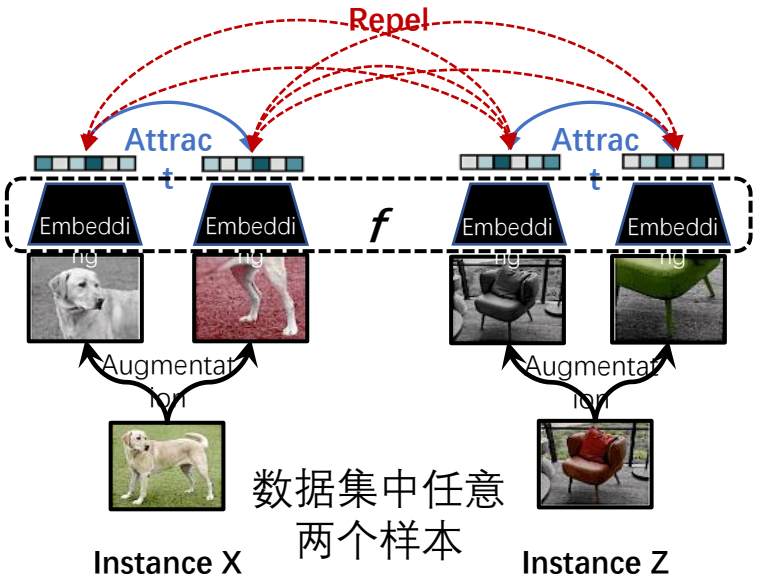
Table 5. **Object detection and instance segmentation fine-tuned on COCO**: bounding-box AP (AP^{bb}) and mask AP (AP^{mk}) evaluated on `val2017`. In the brackets are the gaps to the ImageNet supervised pre-training counterpart. In green are the gaps of at least +0.5 point.

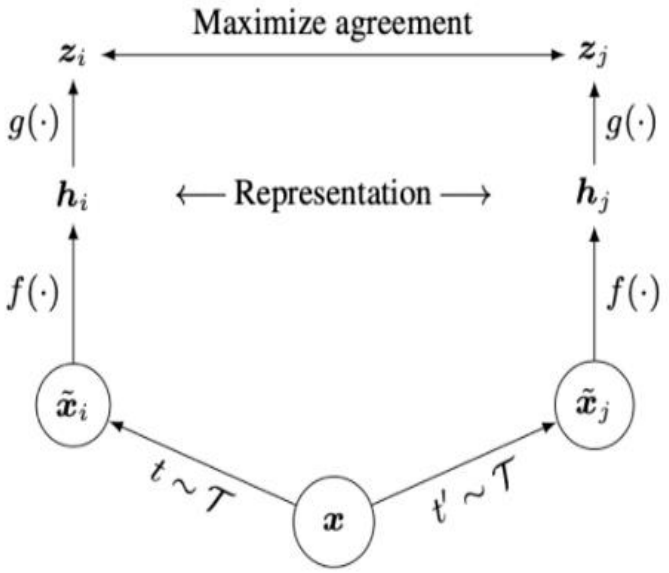


He, Kaiming, et al. "Momentum contrast for unsupervised visual representation learning." *CVPR* 2020.

SimCLR: 一个简单有效的对比学习框架

- 把数据集中任意一个样本与自身的增广样本视作“正样本对”，把数据集中任意两个样本视作“负样本对”，通过学习表征来使得正样本对的特征间距离尽量小，并使得负样本对的特征间距离尽量大

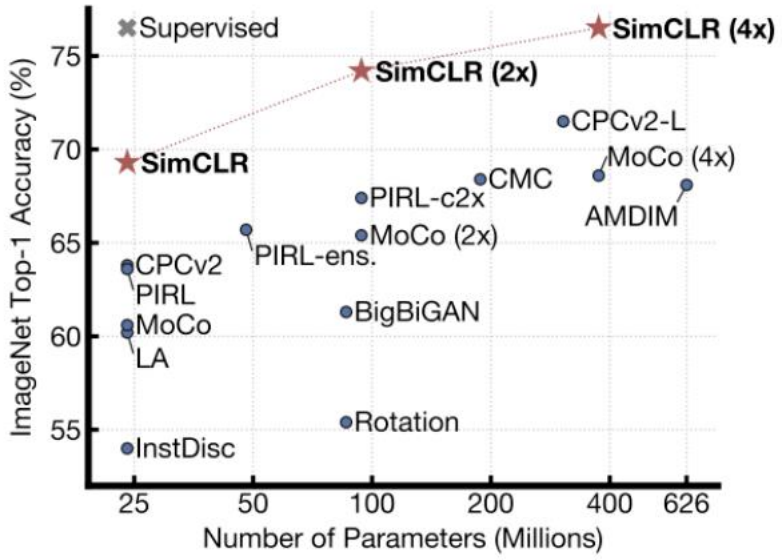




T. Chen, et al., A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020.

SimCLR: Results

- SimCLR：在ImageNet上取得了与监督训练可比较的结果
- SimCLRv2：Unsupervised pretraining using a large ResNet + fine-tuning with 10% of labels，较大幅度超越了全监督训练的结果



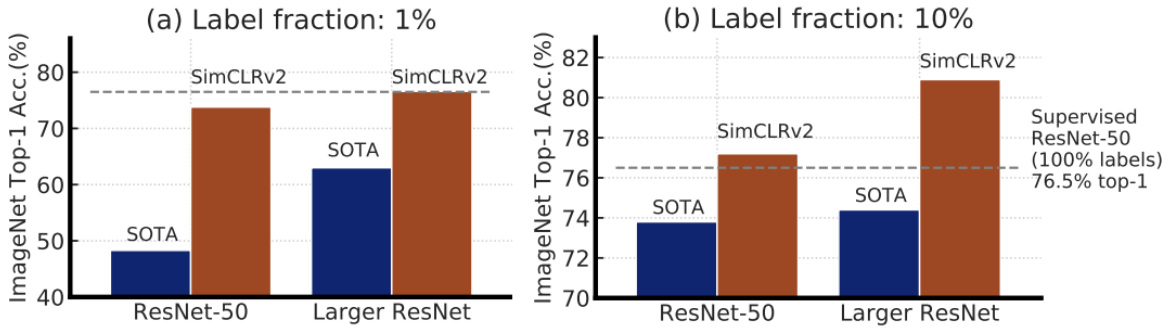


Figure 2: Top-1 accuracy of previous state-of-the-art (SOTA) methods [1, 2] and our method (SimCLRv2) on ImageNet using only 1% or 10% of the labels. Dashed line denotes fully supervised ResNet-50 trained with 100% of labels. Full comparisons in Table 3.

T. Chen, et al., A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020.
Chen T, Kornblith S, Swersky K, et al. Big self-supervised models are strong semi-supervised learners[J]. NIPS, 2020

问题：对比学习的有效性

问题：现有对比学习把数据集中任意两个样本视作负样本对彼此拉远，显然这把同类别的样本对也拉远了，这样构造的自监督信息感觉是不合理的。为何对比学习有效呢?

Large-Margin Contrastive Learning

- **对比学习的有效性分析**
- **提出了Large-Margin的对比学习方法**

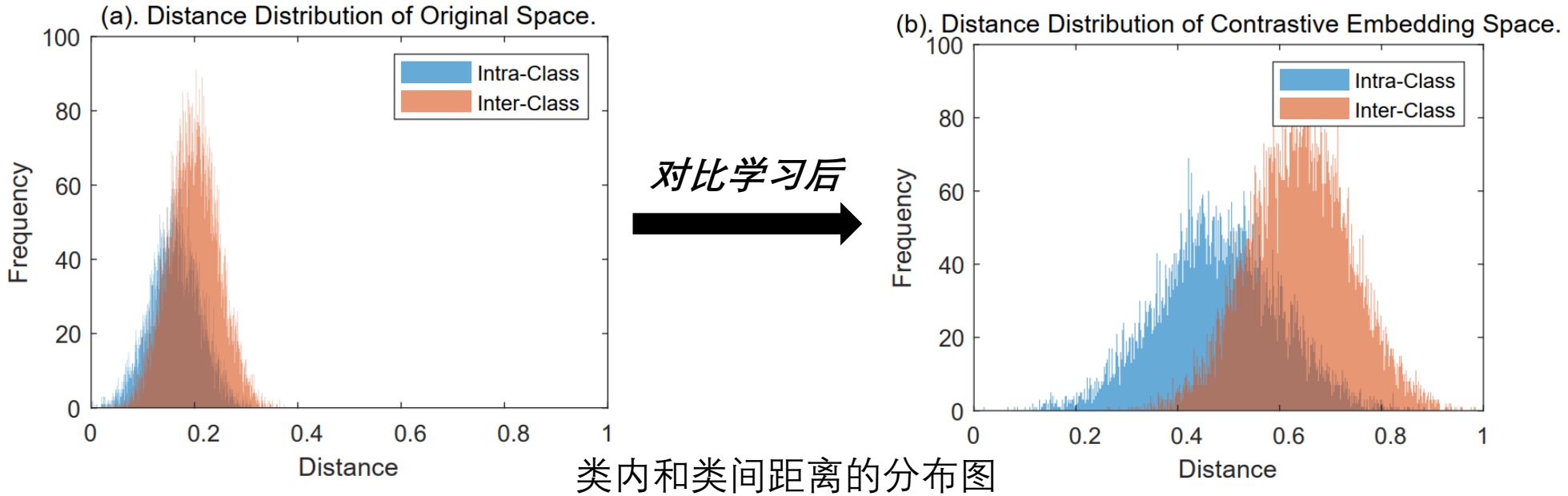
Shuo Chen, Gang Niu, Chen Gong, Jun Li, Jian Yang*, Masashi Sugiyama, Large-Margin Contrastive Learning with Distance Polarization Regularizer, ICML 2021.

对比学习的有效性分析

- 理论分析：对比学习使得样本间距离**连续地分布**在归一化距离空间，即(0, 1)之间，从而自然地产生较小距离和较大距离，从而产生了鉴别性

Theorem 1. *Assume that the optimal feature embedding* $\widehat{\varphi} \in \arg\min_{\varphi \in \mathcal{H}} \mathcal{L}_{\mathrm{NCE}}(\varphi)$ *and the corresponding distance value* $\mathcal{D}_{ij}^{\widehat{\varphi}} = (1 - \widehat{\varphi}(\boldsymbol{x}_i)^\top \widehat{\varphi}(\boldsymbol{x}_j))/2$. *Then for any given* $\mu \in [0, 1]$ *and* $\epsilon > 0$, *there exists sufficiently large* N *such that* $\min_{1 \le i < j \le N} \{|\mathcal{D}_{ij}^{\widehat{\varphi}} - \mu|\} < \epsilon$.

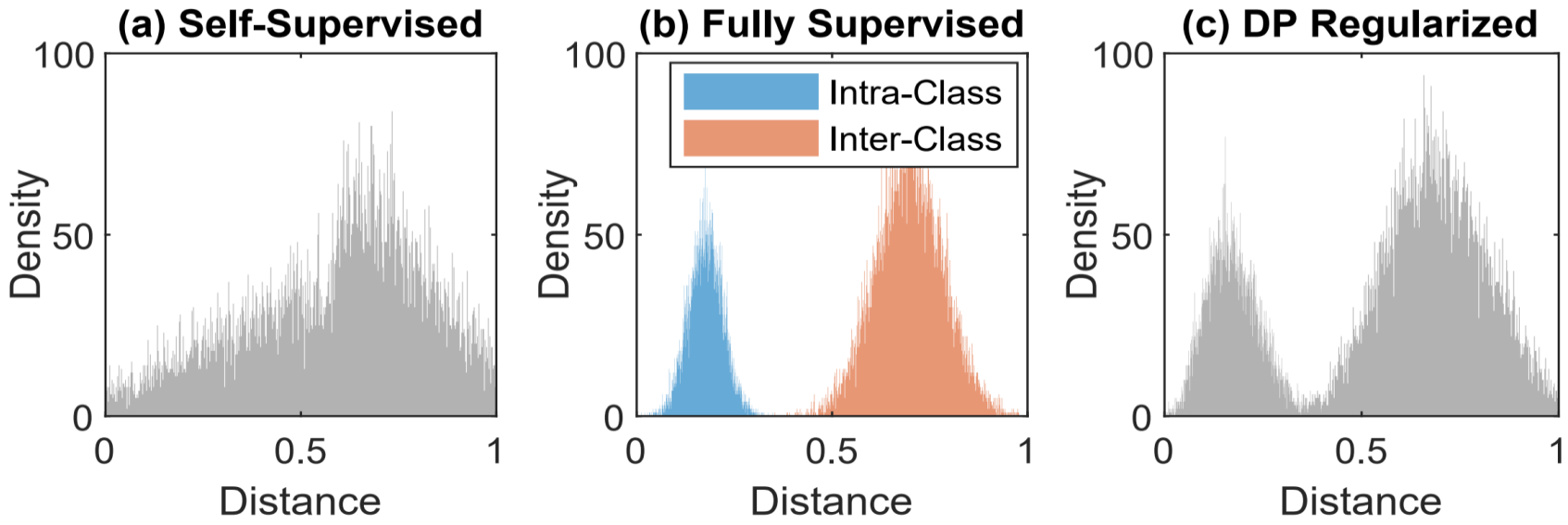
对比学习在拉大两两样本间距离的同时，能够有效凸显类间与类内的距离差异，从而能够自动学得具有鉴别性的特征



类内和类间距离的分布图

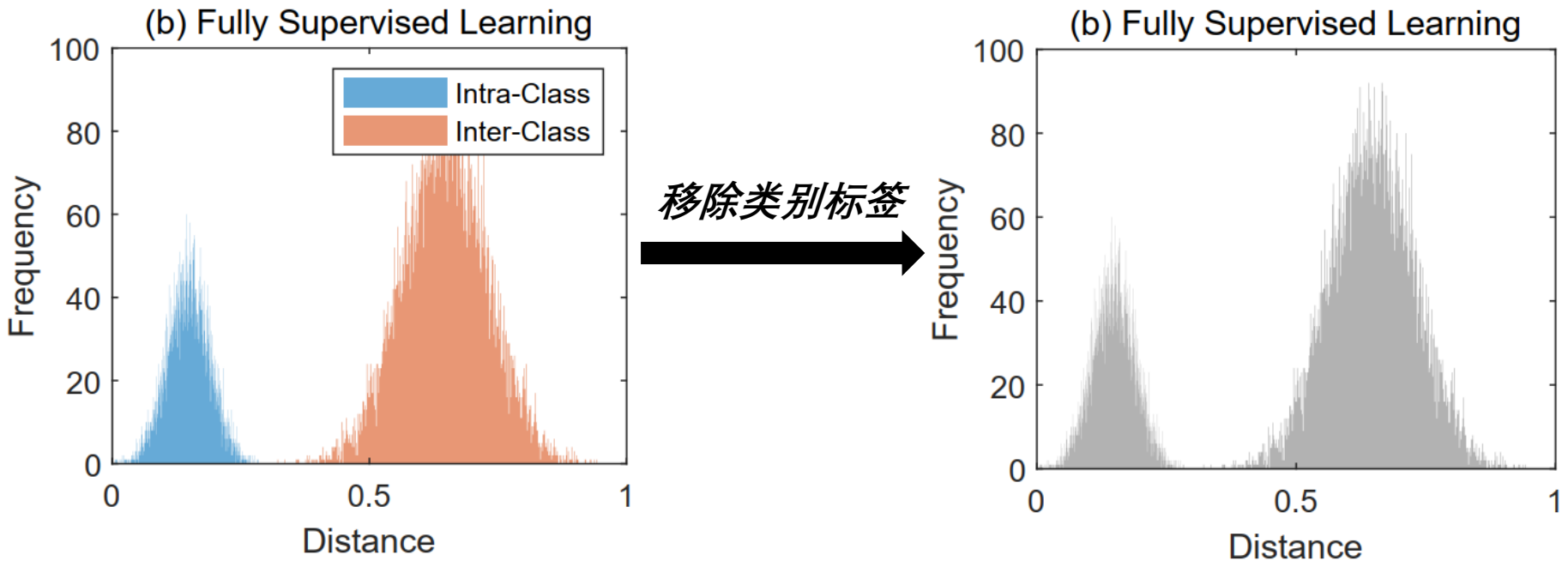
研究动机：大间隔距离

- CIFA-10数据集上原始对比学习、全监督学习、正则化的对比学习的距离分布



模型算法：借鉴度量学习的距离特性

- 全监督的度量学习下的距离分布具有“极化”特点

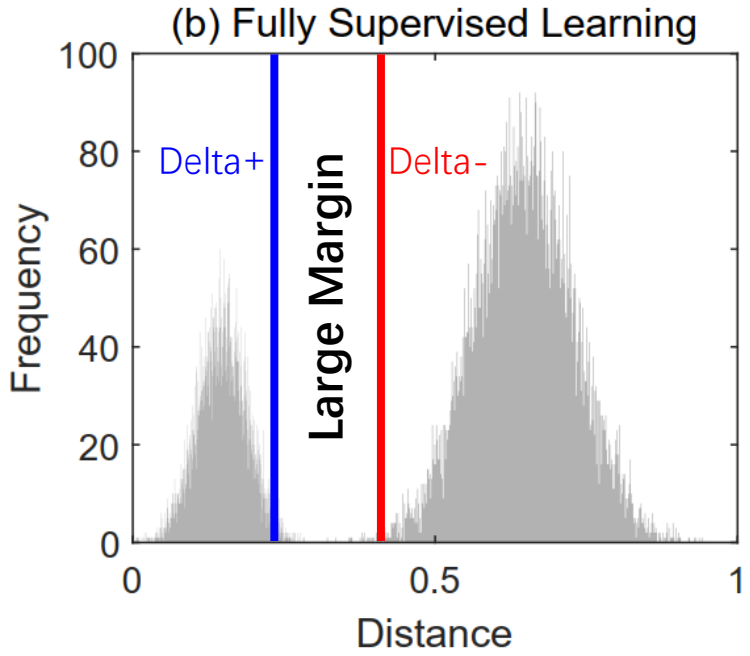


“度量学习” — “监督信息” = “距离极化”

模型算法：构建距离极化正则项

- 引入距离极化正则:

$$\mathcal{R}_0(\boldsymbol{\varphi}) = \| \min[(\boldsymbol{D}^{\boldsymbol{\varphi}} - \boldsymbol{\Delta}^{+}) \odot (\boldsymbol{D}^{\boldsymbol{\varphi}} - \boldsymbol{\Delta}^{-}), 0] \|_0$$



- 正则化后的对比学习目标:

$$\min_{\boldsymbol{\varphi} \in \mathcal{H}} \{\mathcal{F}(\boldsymbol{\varphi}) = \mathcal{L}_{\mathrm{NCE}}(\boldsymbol{\varphi}) + \lambda \mathcal{R}_1(\boldsymbol{\varphi})\}$$

原始NCE损失函数

L1范数松弛后的正则

理论分析：距离度量的泛化误差界

我们证明了，**产生错误距离度量（类间小于类内距离）的数学期望值具有上界**，且具有以下性质：

1）随着正则项参数λ的增大，即距离极化正则项的权重提升，能够有效降低模型的泛化误差界；

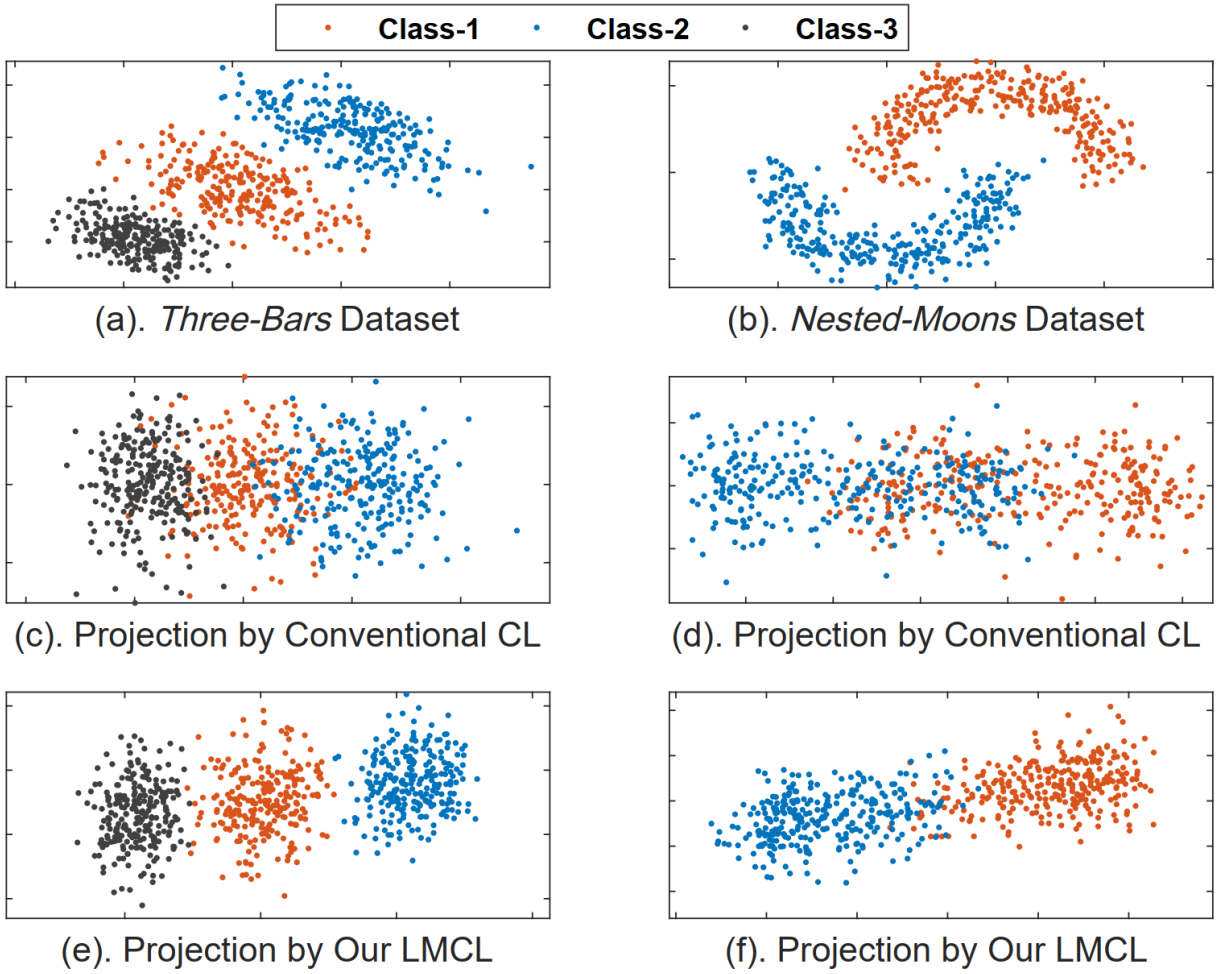
（2）随着类别数C的增加，泛化误差界同样会逐渐缩小，这说明了更多的类别数，也就是更丰富多样的数据，有助于提升模型的泛化性

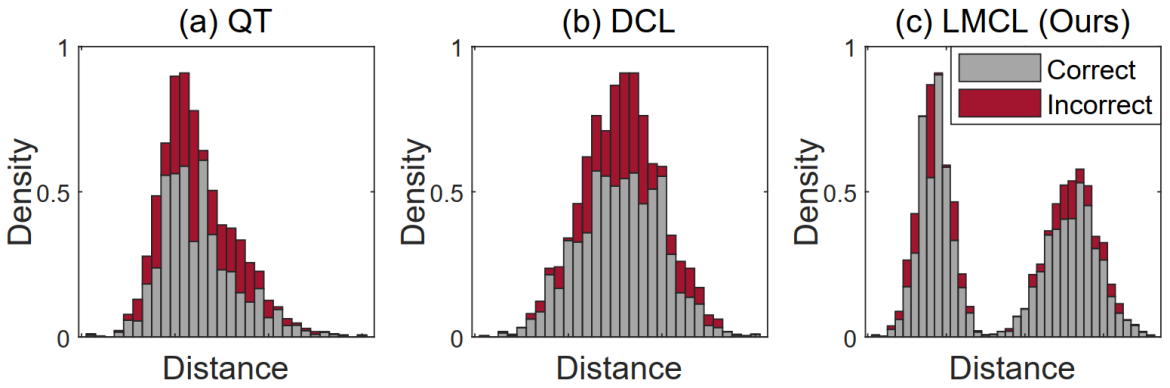
Theorem 3. *Assume that* $\boldsymbol{\varphi}^* \in \arg\min_{\boldsymbol{\varphi}\in\mathcal{H}} \mathcal{L}_{\mathrm{NCE}}(\boldsymbol{\varphi}) + \lambda\mathcal{R}_1(\boldsymbol{\varphi})$, *and the underling class labels of training data* $\{\boldsymbol{x}_i\}_{i=1}^N$ *are* $\{y_i\}_{i=1}^N$. *Then we have that*

$$\begin{aligned}
&\mathbb{E}_{y_i\neq y_j}[\max(\delta_\mu^- - \mathcal{D}_{ij}^{\boldsymbol{\varphi}^*}, 0)] + \mathbb{E}_{y_k=y_l}[\max(\mathcal{D}_{kl}^{\boldsymbol{\varphi}^*} - \delta_\mu^+, 0)] \\
&\leq (\delta^- - \delta^+)\mathcal{R}_1(\boldsymbol{\varphi}^*) + (K_{\max}/K_{\min})/C \\
&\leq 4(\delta^- - \delta^+)/\lambda + (K_{\max}/K_{\min})/C, \qquad (13)
\end{aligned}$$

where the constants $\delta_\mu^- = \delta^- - \mu$, $\delta_\mu^+ = \delta^+ + \mu$, $\mu \in (0, \delta^- - \delta^+)$, $K_{\min} = \min_{1\leq k\leq C} \|\boldsymbol{y} - k\cdot\mathbf{1}_{N\times 1}\|_0$, *and* $K_{\max} = \max_{1\leq k\leq C} \|\boldsymbol{y} - k\cdot\mathbf{1}_{N\times 1}\|_0$.

可视化实验结果



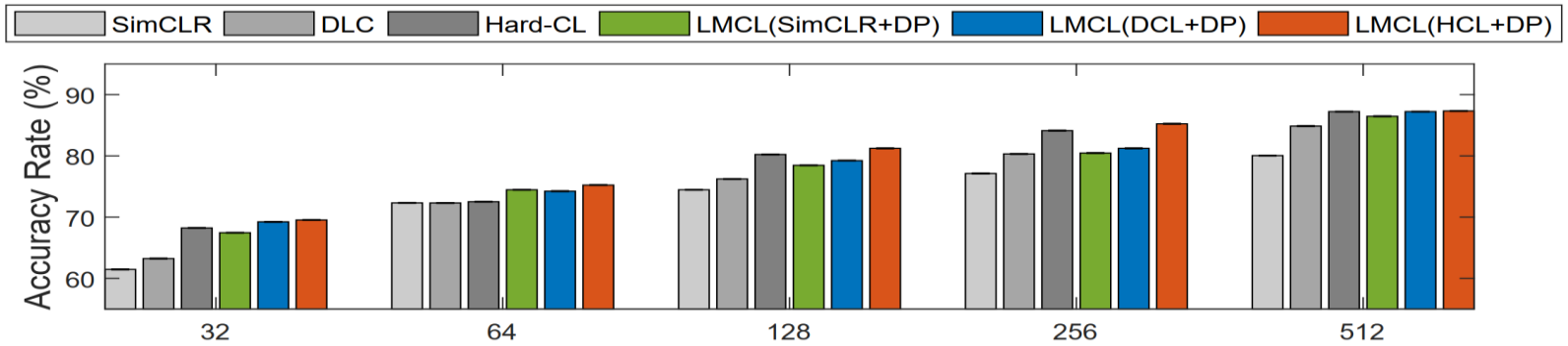


METHOD	*MR*	*CR*	*SUBJ*	*MPQA*	*TREC*	*MSRP*
QT	76.8	81.3	86.6	93.4	89.8	73.6
DCL	76.2	82.9	86.9	93.7	89.1	74.7
HCL	77.4	83.6	86.8	93.4	88.7	73.5
LMCL(QT+DP)	77.3	82.3	86.9	93.7	**90.2**	74.1
LMCL(DCL+DP)	77.2	**83.7**	**87.2**	93.8	90.1	**75.1**
LMCL(HCL+DP)	**78.1**	83.5	**87.2**	**94.0**	89.1	74.2

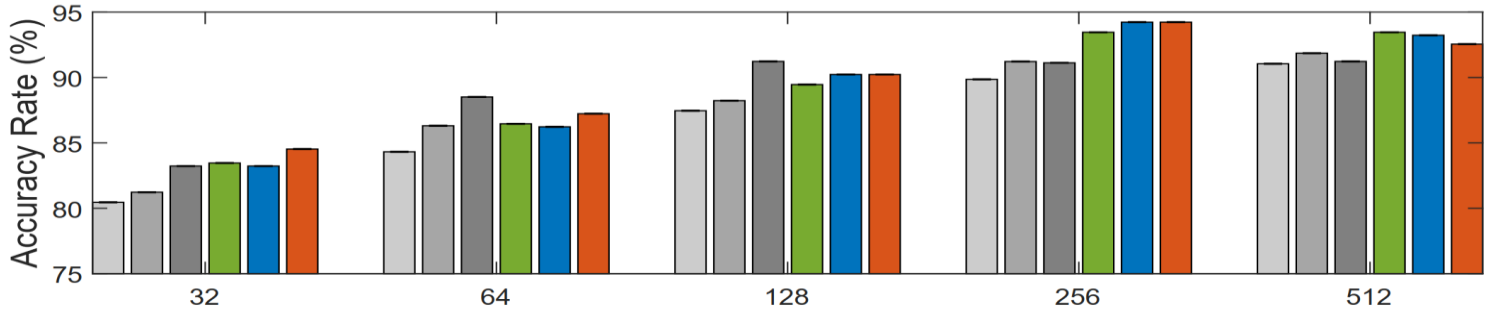
所提出的距离极化正则方法能够使得类内与类间产生大间隔距离差异，从而提升算法性能

分类实验结果

METHOD	1024		4096	
	Top1	Top5	Top1	Top5
CMC	60.23	79.23	73.58	92.06
SwAV	60.93	79.43	75.78	92.86
DCL	61.01	78.99	74.60	92.08
HCL	60.89	79.33	74.66	92.32
LMLC(CMC+DP)	**61.23**	**79.44**	75.67	**93.02**
LMLC(DCL+DP)	61.12	79.20	**75.89**	92.89
LMLC(HCL+DP)	60.92	79.43	74.94	92.39



(a). Classification accuracy of all compared methods on *STL-10* dataset.



(b). Classification accuracy of all compared methods on *CIFAR-10* dataset.

低维空间下的对比学习

- 从理论分析了对比学习中的特征维度对模型泛化性能的影响
- 揭示了维数约简在对比学习中的必要性

S. Chen, C. Gong, J. Li, J. Yang, G. Niu, M. Sugiyama. Learning Contrastive Embedding in Low-Dimensional Space, *NeurIPS 2022.*

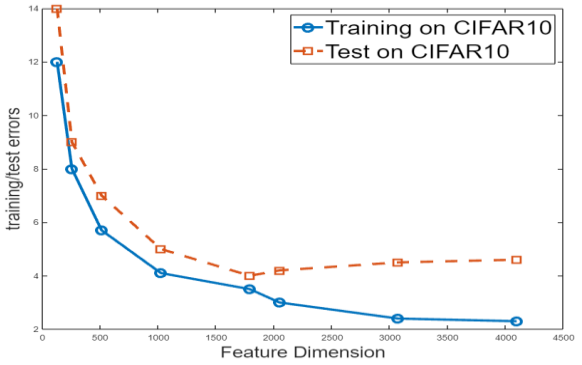
高维特征空间存在过拟合问题

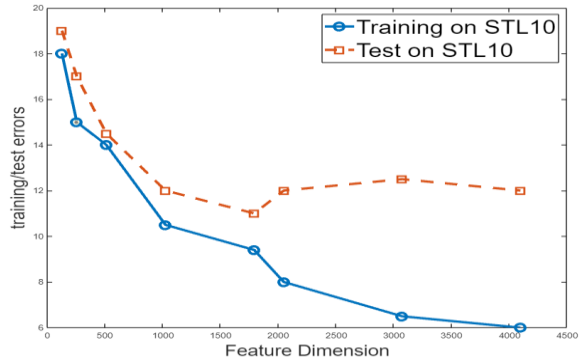
- 随着特征维度不断升高，网络的训练误差一致地下降（NCE loss）
- 但当特征维度超过一定数值时，虽然训练误差在下降，但测试误差反而呈现上升趋势

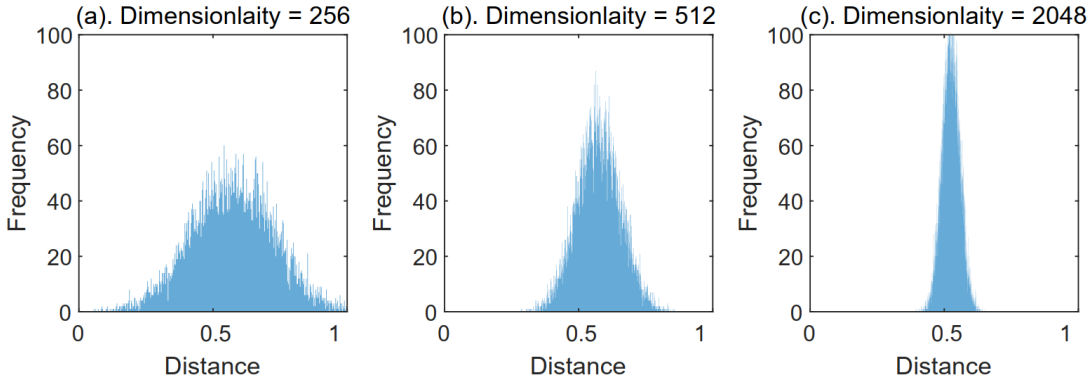
$$\mathcal{L}_{\mathrm{NCE}}(\boldsymbol{\Phi})$$

$$=\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}_j^-\in\mathcal{X}}[-\log(\mathrm{e}^{\boldsymbol{\Phi}(\boldsymbol{x})^\top\boldsymbol{\Phi}(\boldsymbol{x}^+)}/(\mathrm{e}^{\boldsymbol{\Phi}(\boldsymbol{x})^\top\boldsymbol{\Phi}(\boldsymbol{x}^+)}+\sum_{j=1}^{n}\mathrm{e}^{\boldsymbol{\Phi}(\boldsymbol{x})^\top\boldsymbol{\Phi}(\boldsymbol{x}_j^-)}))]$$

- 维度升高后，距离的概率分布变得更加集中，从而缺乏类内、类间的鉴别能力







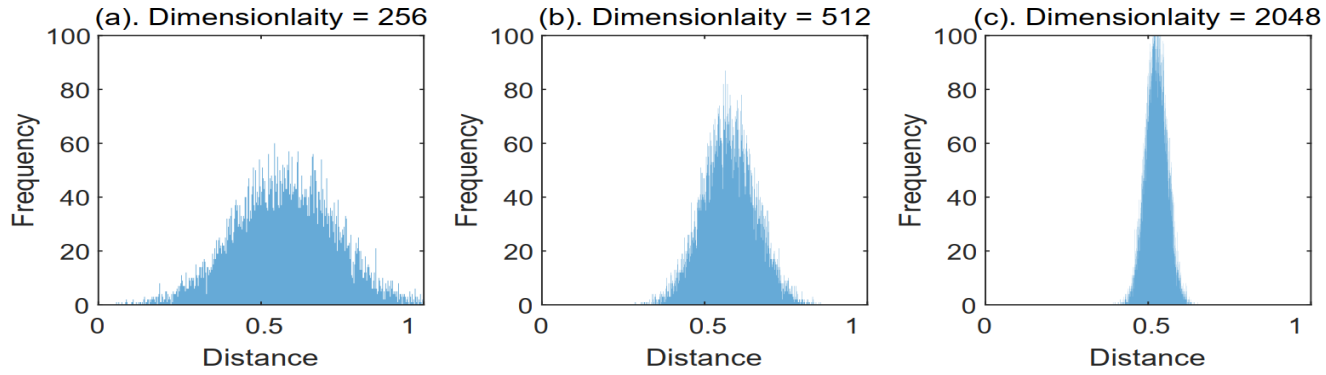
维度灾难

- 维度趋向于无穷大时，样本间距离的最大值和最小值变得近乎相同

Theorem 1. *For any given i.i.d. random data points* $\boldsymbol{x}, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^m$, *we denote* $\mathcal{D}_{\boldsymbol{\Phi}}^{\max}(H) = \max\{\mathcal{D}_{\boldsymbol{\Phi}}(\boldsymbol{x}, \boldsymbol{x}_i) | i = 1, \ldots, n\}$ *and* $\mathcal{D}_{\boldsymbol{\Phi}}^{\min}(H) = \min\{\mathcal{D}_{\boldsymbol{\Phi}}(\boldsymbol{x}, \boldsymbol{x}_i) | i = 1, \ldots, n\}$. *Then we have that* $\lim_{H\to\infty}\{\mathrm{var}[\mathcal{D}_{\boldsymbol{\Phi}}(\boldsymbol{x}, \boldsymbol{x}_i)/\mathbb{E}(\mathcal{D}_{\boldsymbol{\Phi}}(\boldsymbol{x}, \boldsymbol{x}_i))]\} = 0$ *and*

$$\mathcal{P}\left\{\lim_{H\to\infty}(\mathcal{D}_{\boldsymbol{\Phi}}^{\max}(H) - \mathcal{D}_{\boldsymbol{\Phi}}^{\min}(H))/\mathcal{D}_{\boldsymbol{\Phi}}^{\min}(H) = 0\right\} = 1, \tag{5}$$

where the distance function $\mathcal{D}_{\boldsymbol{\Phi}}(\cdot, \cdot)$ *is defined in Eq. (1) and the feature embedding* $\boldsymbol{\Phi}$ *is learned from the training data and independent to the data points* $\boldsymbol{x}, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$.



解决方法：稀疏低维空间重建

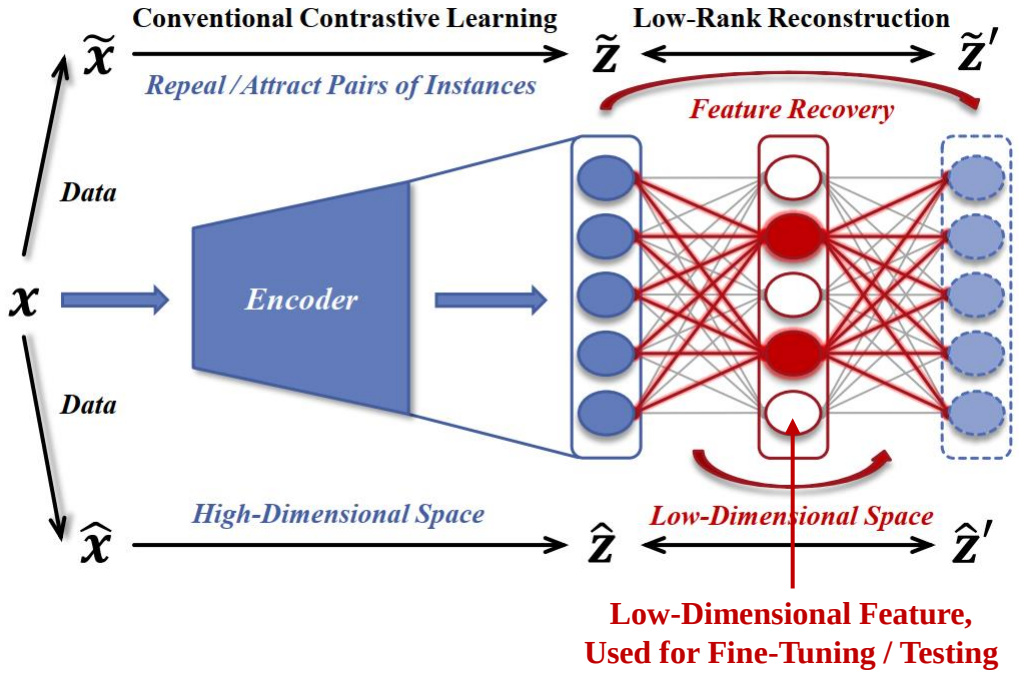
- 现有的对比学习模型基础上，引入额外的稀疏投影层，通过L21范数约束来对高维特征进行自适应的低维重建

$$\mathcal{R}_{2,1}(\boldsymbol{\Phi}, \boldsymbol{L}) = \mathbb{E}_{\boldsymbol{x}\in\mathcal{X}}[\,\|\boldsymbol{L}^\top \boldsymbol{L}\cdot\boldsymbol{\Phi}(\boldsymbol{x}) - \boldsymbol{\Phi}(\boldsymbol{x})\|_2^2\,] + \alpha\|\boldsymbol{L}\|_{2,1}$$

- 正则化后的 Loss function

$$\min_{\boldsymbol{\Phi}\in\mathcal{H},\boldsymbol{L}\in\mathbb{R}^{H\times H}} \{\mathcal{F}(\boldsymbol{\Phi}, \boldsymbol{L}) = \mathcal{L}_{\mathrm{NCE}}(\boldsymbol{\Phi}) + \lambda\mathcal{R}(\boldsymbol{\Phi}, \boldsymbol{L})\}$$

- 从而获得低维具有鉴别力的特征表示



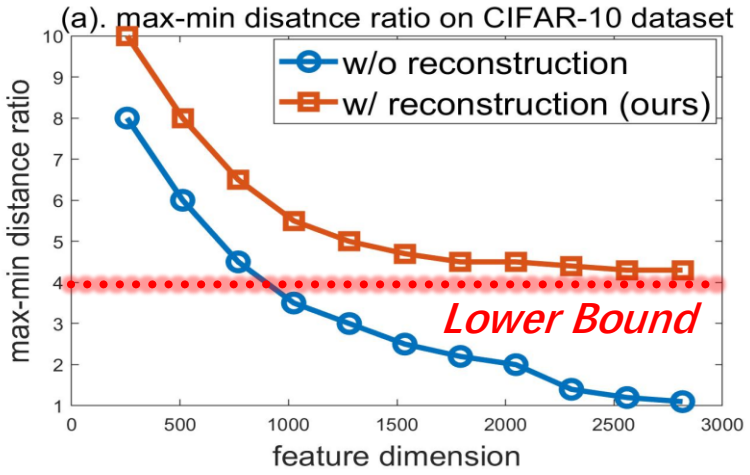
理论分析

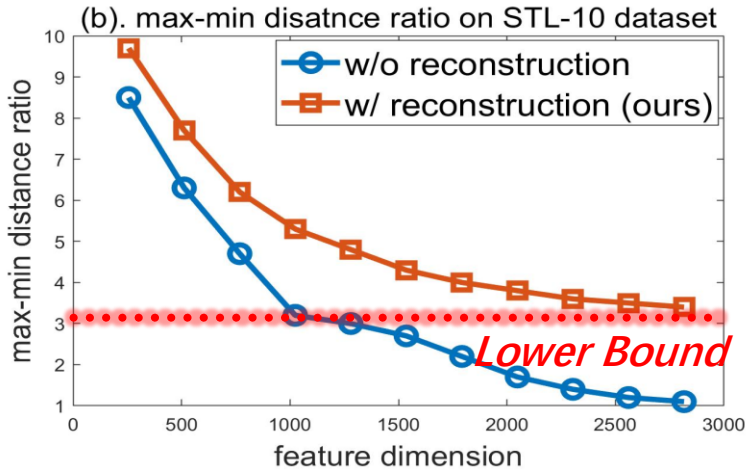
- 引入稀疏低维正则后，能够有效地保证样本对距离的鉴别性，使得鉴别能力具有一定的最坏下界保证

Theorem 3. *For any given $n+1$ i.i.d. random data points $\boldsymbol{x}, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^m$, we denote that $\mathcal{D}^{\max}_{\widehat{\boldsymbol{\Phi}},\widehat{\boldsymbol{L}}} = \max\{\mathcal{D}_{\widehat{\boldsymbol{\Phi}},\widehat{\boldsymbol{L}}}(\boldsymbol{x}, \boldsymbol{x}_i) | i = 1, 2, \ldots, n\}$ and $\mathcal{D}^{\min}_{\widehat{\boldsymbol{\Phi}},\widehat{\boldsymbol{L}}} = \min\{\mathcal{D}_{\widehat{\boldsymbol{\Phi}},\widehat{\boldsymbol{L}}}(\boldsymbol{x}, \boldsymbol{x}_i) | i = 1, 2, \ldots, n\}$, and then we have that*

$$\mathcal{P}\left\{(\mathcal{D}^{\max}_{\widehat{\boldsymbol{\Phi}},\widehat{\boldsymbol{L}}} - \mathcal{D}^{\min}_{\widehat{\boldsymbol{\Phi}},\widehat{\boldsymbol{L}}})/\mathcal{D}^{\min}_{\widehat{\boldsymbol{\Phi}},\widehat{\boldsymbol{L}}} \geq \alpha\lambda C(\mathcal{X})\right\} = 1, \tag{12}$$

where $\mathcal{D}_{\widehat{\boldsymbol{\Phi}},\widehat{\boldsymbol{L}}}(\boldsymbol{x}, \boldsymbol{x}_i) = \|\widehat{\boldsymbol{L}}\widehat{\boldsymbol{\Phi}}(\boldsymbol{x}) - \widehat{\boldsymbol{L}}\widehat{\boldsymbol{\Phi}}(\boldsymbol{x}_i)\|_2 / rank(\widehat{\boldsymbol{L}})$, and $\widehat{\boldsymbol{\Phi}}$ and $\widehat{\boldsymbol{L}}$ are learned from Eq. (8).





如图中所示，蓝线是高维特征的最大最小距离比值，红线是低维重建特征的最大最小距离比值，能够具有一个确定的下界保证

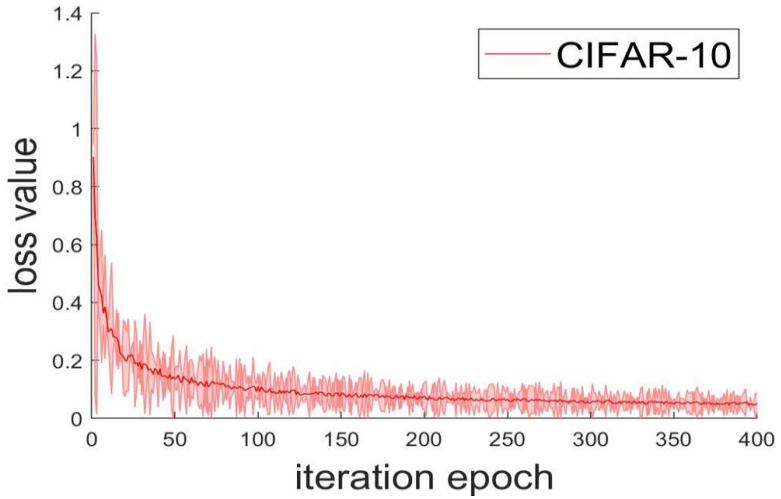
收敛性分析

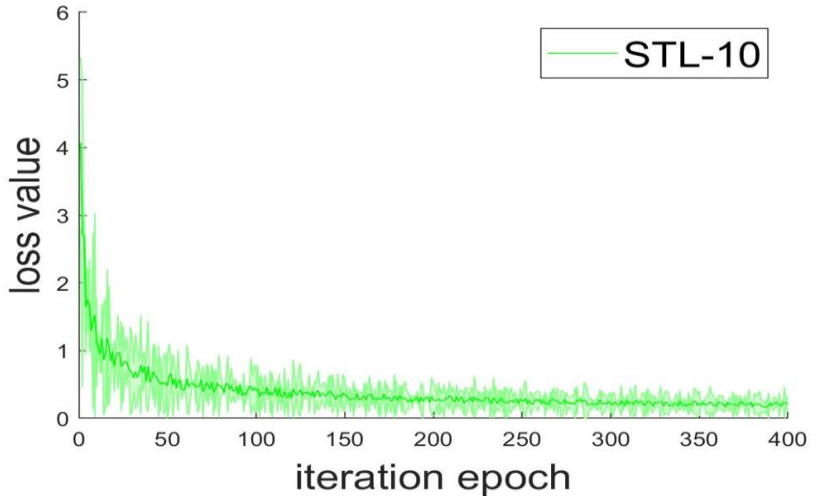
- 设计了随机梯度算法对提出的模型进行优化求解，证明了迭代算法具有根号T的收敛率保证

Theorem 2. *If the function* $\mathcal{F}(\boldsymbol{\Phi}, \boldsymbol{L})$ *has* δ*-bounded gradient* (i.e., $\|\nabla\mathcal{F}(\boldsymbol{\Phi}, \boldsymbol{L})\|_2 < \delta$), *then we let* $\eta = \sqrt{2(\mathcal{F}(\boldsymbol{\Phi}_{(0)}, \boldsymbol{L}_{(0)}) - \mathcal{F}(\boldsymbol{\Phi}^*, \boldsymbol{L}^*))/(S\delta^2 T)}$, *and for the iterations in Algorithm 1 we have that*

$$\min_{0\le t\le T-1} \mathbb{E}[\,\|\nabla\mathcal{F}(\boldsymbol{\Phi}_{(t)}, \boldsymbol{L}_{(t)})\|_2] \le \sqrt{2S(\mathcal{F}(\boldsymbol{\Phi}_{(0)}, \boldsymbol{L}_{(0)}) - \mathcal{F}(\boldsymbol{\Phi}^*, \boldsymbol{L}^*))/T}\delta, \tag{11}$$

where $S > 0$ *is a lipschitz constant such that* $\|\nabla\mathcal{F}(\boldsymbol{\Phi}, \boldsymbol{L}) - \nabla\mathcal{F}(\boldsymbol{\Phi}', \boldsymbol{L}')\|_2 < S\|[\boldsymbol{\Phi}, \boldsymbol{L}] - [\boldsymbol{\Phi}', \boldsymbol{L}']\|_2$.





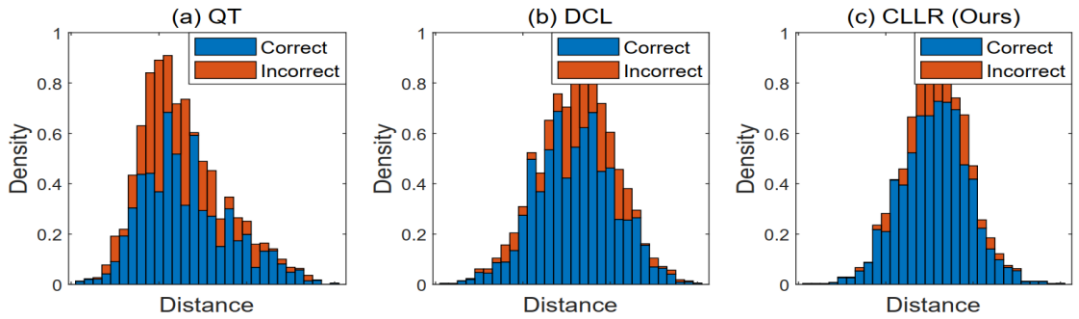
实验结果

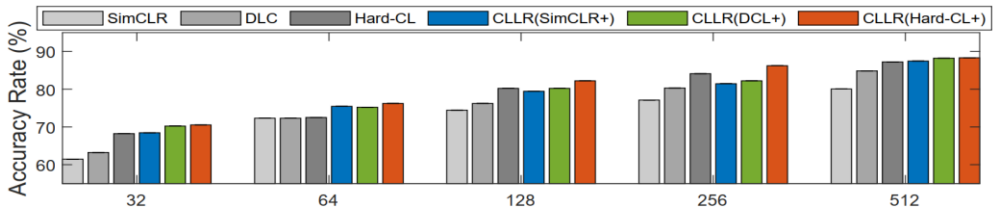
- 消融实验

METHOD	*STL-10*		*CIFAR-10*	
	epochs=100	epochs=400	epochs=100	epochs=400
4096-dim. (w/o $\mathcal{R}(\boldsymbol{\Phi}, \boldsymbol{L})$)	55.1 ± 1.1	75.2 ± 3.1	65.1 ± 1.9	85.4 ± 4.2
3072-dim. (w/o $\mathcal{R}(\boldsymbol{\Phi}, \boldsymbol{L})$)	54.4 ± 3.1	75.2 ± 2.1	67.2 ± 3.5	86.9 ± 6.1
2048-dim. (w/o $\mathcal{R}(\boldsymbol{\Phi}, \boldsymbol{L})$)	56.3 ± 2.1	76.2 ± 1.1	66.3 ± 3.1	89.3 ± 2.1
512-dim. (w/o $\mathcal{R}(\boldsymbol{\Phi}, \boldsymbol{L})$)	56.4 ± 2.5	75.2 ± 0.1	66.4 ± 5.1	90.3 ± 0.6
256-dim. (w/o $\mathcal{R}(\boldsymbol{\Phi}, \boldsymbol{L})$)	55.3 ± 4.1	74.2 ± 2.1	64.3 ± 5.1	88.3 ± 3.1
512-dim. (w/o sparity, $\alpha = 0$)	$\mathbf{56.5 \pm 2.5}$	75.5 ± 0.5	66.2 ± 4.9	90.1 ± 1.2
256-dim. (w/o sparity, $\alpha = 0$)	55.9 ± 2.1	74.1 ± 2.3	64.7 ± 2.1	88.4 ± 2.6
512-dim. (w/ $\ell_{2,1}$-norm)	56.3 ± 8.2 –	78.3 ± 0.5 ✓	$\mathbf{67.5 \pm 0.2}$ –	$\mathbf{92.5 \pm 0.2}$ ✓
512-dim. (w/ nuclear-norm)	56.2 ± 3.2 –	$\mathbf{79.2 \pm 0.2}$ ✓	67.5 ± 2.5 –	92.5 ± 2.3 ✓
256-dim. (w/ $\ell_{2,1}$-norm)	56.2 ± 1.2 –	$\mathbf{79.3 \pm 0.5}$ ✓	65.5 ± 0.5 –	92.3 ± 0.3 ✓
256-dim. (w/ nuclear-norm)	56.3 ± 3.2 –	79.2 ± 0.2 ✓	65.2 ± 5.5 –	$\mathbf{93.1 \pm 1.3}$ ✓

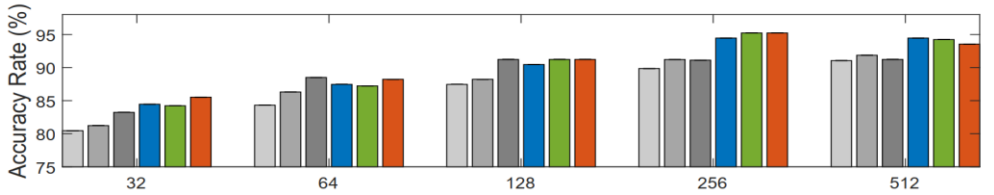
- 图像分类实验（ImageNet）

METHOD	1024		4096	
	Top1	Top5	Top1	Top5
CMC[35]	60.23	79.23	73.58	92.06
SwAV[3]	60.93	79.43	75.78	92.86
DCL[11]	61.01	78.99	74.60	92.08
HCL[30]	60.89	79.33	74.66	92.32
CO2[38]	61.21	79.32	73.96	93.02
CLLR(CMC+$\ell_{2,1}$-norm)	62.03	80.64	75.97	94.22
CLLR(CMC+nuclear-norm)	61.23	80.50	**76.91**	94.03
CLLR(HCL+$\ell_{2,1}$-norm)	61.29	**81.10**	76.88	94.19
CLLR(HCL+nuclear-norm)	**62.43**	80.98	76.89	**94.25**





(a). Classification accuracy of all compared methods on *STL-10* dataset.



(b). Classification accuracy of all compared methods on *CIFAR-10* dataset.

- 文本分类实验

METHOD	*MR*	*CR*	*SUBJ*	*MPQA*	*TREC*	*MSRP*
QT[26]	76.8	81.3	86.6	93.4	89.8	73.6
DCL[11]	76.2	82.9	86.9	93.7	89.1	74.7
HCL[30]	77.4	83.6	86.8	93.4	88.7	73.5
CLLR(DCL+$\ell_{2,1}$-norm)	77.9	83.3	**87.9**	93.7	**91.3**	75.2
CLLR(DCL+nuclear-norm)	**78.2**	**83.7**	87.2	**95.8**	91.2	**75.7**

主要内容

（1）对比范式

Contrastive Learning

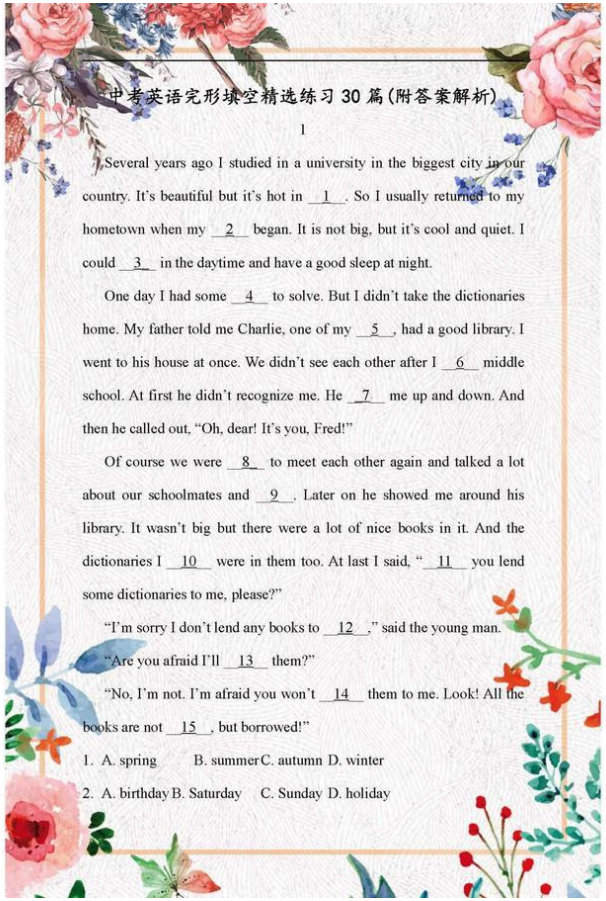
（2）掩码范式

Masked Autoencoders

（3）内隐范式

Masked language modeling (MLM)

- 完形填空是学习自然语言的有效方式
- 人是，机器也是
- 这种预训练的模式非常简单：删去数据的一部分内容然后让机器学习预测删去内容
- Autoregressive language modeling in GPT
- Masked autoencoding in BERT



中考英语完形填空精选练习 30 篇(附答案解析)

1

Several years ago I studied in a university in the biggest city in our country. It's beautiful but it's hot in __1__. So I usually returned to my hometown when my __2__ began. It is not big, but it's cool and quiet. I could __3__ in the daytime and have a good sleep at night.

One day I had some __4__ to solve. But I didn't take the dictionaries home. My father told me Charlie, one of my __5__, had a good library. I went to his house at once. We didn't see each other after I __6__ middle school. At first he didn't recognize me. He __7__ me up and down. And then he called out, "Oh, dear! It's you, Fred!"

Of course we were __8__ to meet each other again and talked a lot about our schoolmates and __9__. Later on he showed me around his library. It wasn't big but there were a lot of nice books in it. And the dictionaries I __10__ were in them too. At last I said, "__11__ you lend some dictionaries to me, please?"

"I'm sorry I don't lend any books to __12__," said the young man.

"Are you afraid I'll __13__ them?"

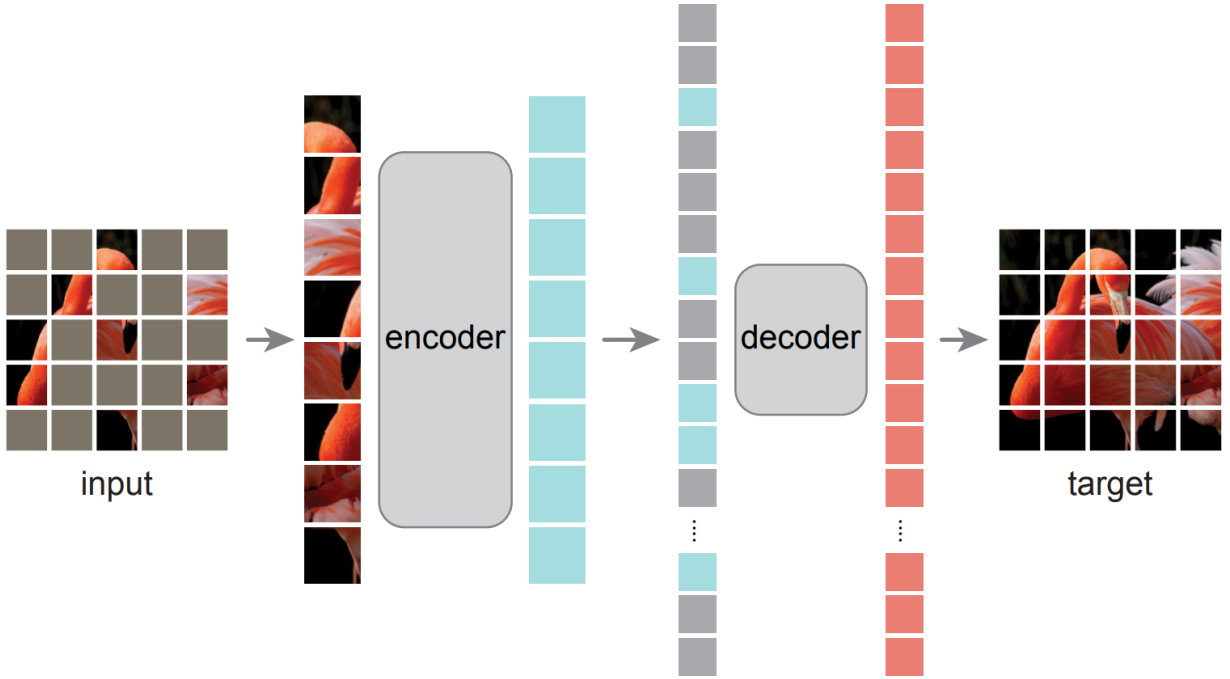
"No, I'm not. I'm afraid you won't __14__ them to me. Look! All the books are not __15__, but borrowed!"

1. A. spring B. summer C. autumn D. winter

2. A. birthday B. Saturday C. Sunday D. holiday

Masked Image Modeling (MIM)

- “填图补缺”是视觉理解的有效方式
- Masked Autoencoders (MAE)



K. He, X. Chen, X. Xie, et al, Masked Autoencoders Are Scalable Vision Learners. *CVPR 2022*.

Z. Xie, Z. Zhang, Y. Cao, et al. Simmim: A Simple Framework for Masked Image Modeling, *CVPR 2022*.

Masked Autoencoders

- Masking

  将图片分割成一系列小块，然后在这些小块中随机、均匀地选取一部分保留，剩下的全部遮蔽。强调要遮蔽大量的像素块（约75%)

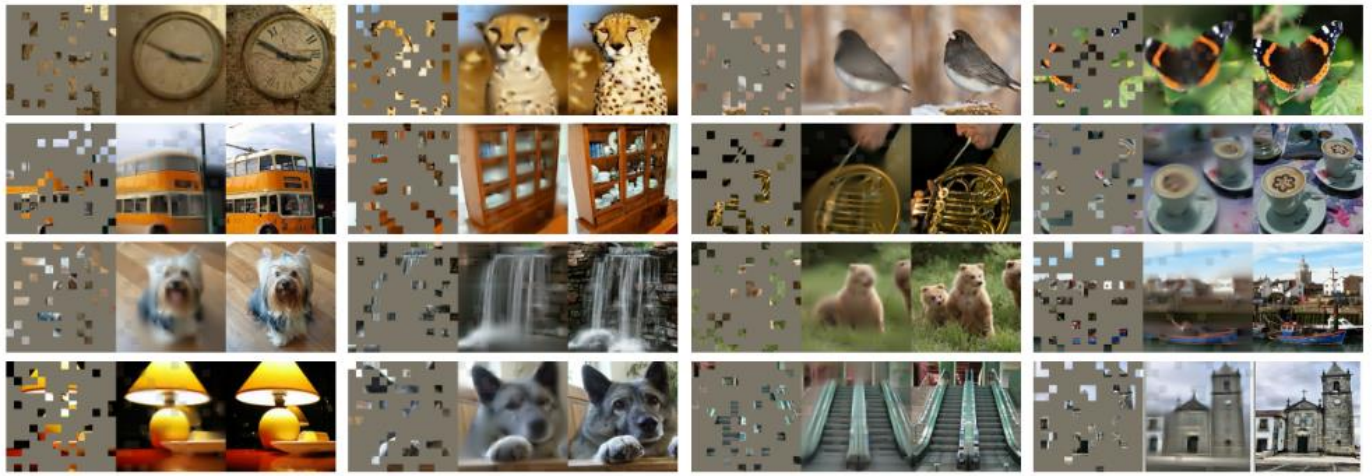


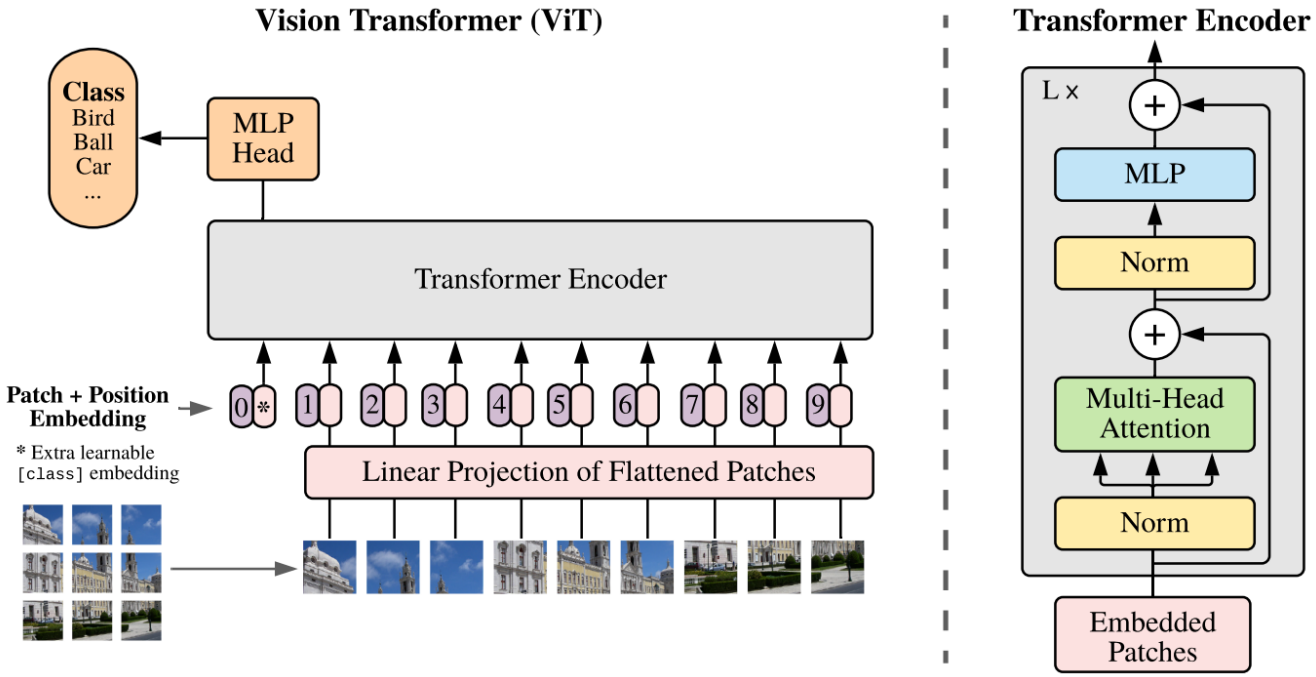
Figure 2. Example results on ImageNet *validation* images. For each triplet, we show the masked image (left), our MAE reconstruction† (middle), and the ground-truth (right). The masking ratio is 80%, leaving only 39 out of 196 patches. More examples are in the appendix.
†*As no loss is computed on visible patches, the model output on visible patches is qualitatively worse. One can simply overlay the output with the visible patches to improve visual quality. We intentionally opt not to do this, so we can more comprehensively demonstrate the method's behavior.*

Figure 3. Example results on COCO validation images, using an MAE trained on ImageNet (the same model weights as in Figure 2). Observe the reconstructions on the two right-most examples, which, although different from the ground truth, are semantically plausible.

Masked Autoencoders (MAE)

- MAE 的编码-解码架构（非对称）

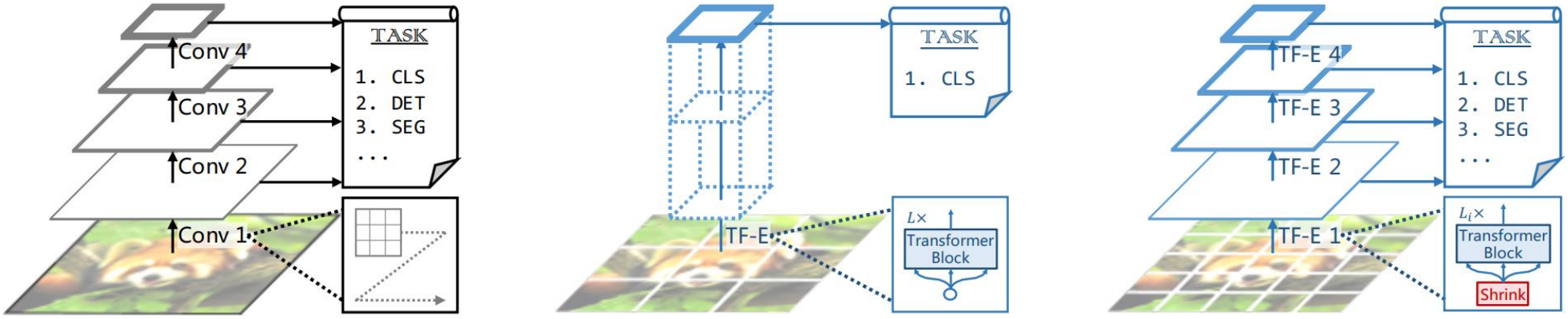
  编码器是一个标准的Vision Transformer (ViT)模型中的Transformer Encoder；解码器仅在预训练中用于图像重建，其设计可独立于编码器，且灵活和轻量级



Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. ICLR, 2020.

分层Vision Transformer

- 代表性的方法Pyramid Vision Transformer(PVT) 、Swin Transformer
- 优势：利用金字塔结构，充分挖掘视觉图像的多尺度信息

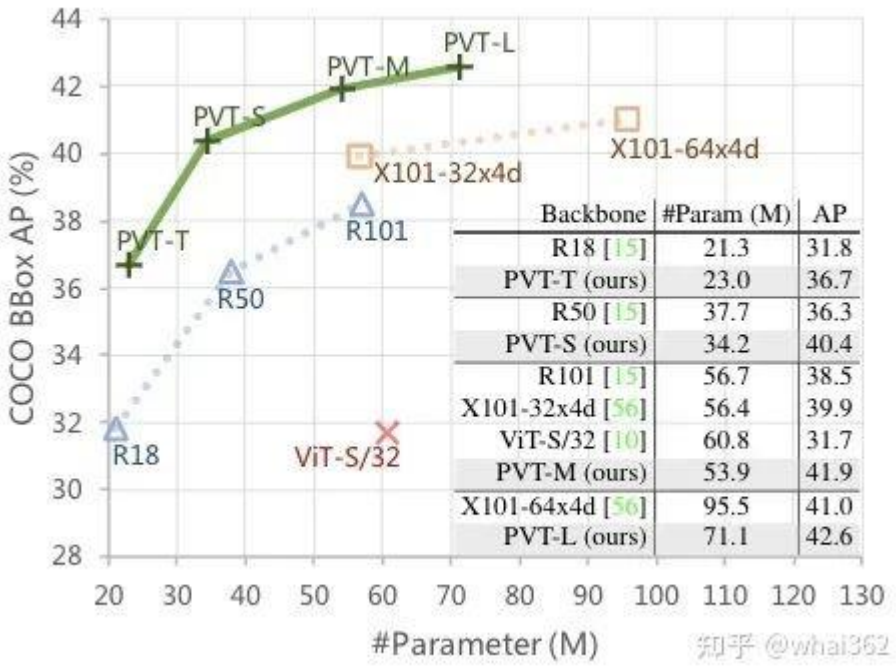


(a) CNNs: VGG [53], ResNet [21], *etc.*　(b) Vision Transformer [12]　(c) Pyramid Vision Transformer (ours)

Wang W, Xie E, Li X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]//ICCV. 2021: 568-578.

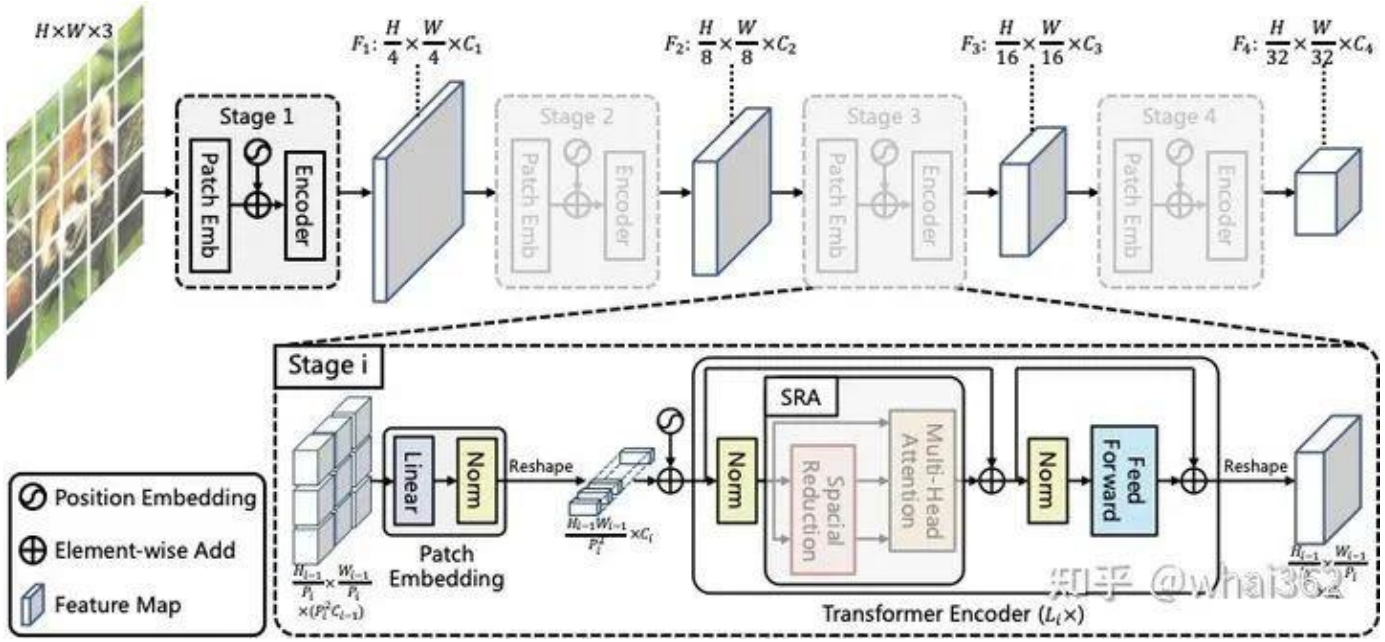
Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//ICCV. 2021: 10012-10022.

Pyramid Vision Transformer(PVT)



Backbone	#Param (M)	AP
R18 [15]	21.3	31.8
PVT-T (ours)	23.0	36.7
R50 [15]	37.7	36.3
PVT-S (ours)	34.2	40.4
R101 [15]	56.7	38.5
X101-32x4d [56]	56.4	39.9
ViT-S/32 [10]	60.8	31.7
PVT-M (ours)	53.9	41.9
X101-64x4d [56]	95.5	41.0
PVT-L (ours)	71.1	42.6

在 COCO val2017 目标检测任务上的性能比较



PVT的架构图：在每个Stage中通过Patch Embedding来逐渐降低输入的分辨率

问题

- MAE是无法直接支持层级ViT，例如PVT、Swin Transformer
- SimMIM能够支持层级ViT，但通过引入掩码向量，极大降低了预训练的效率

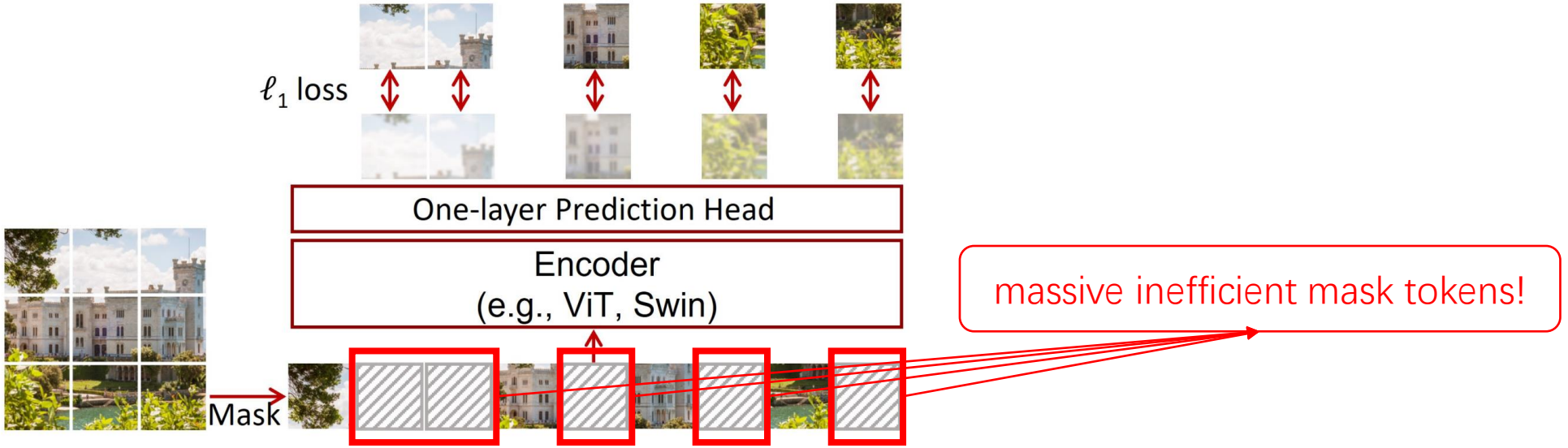
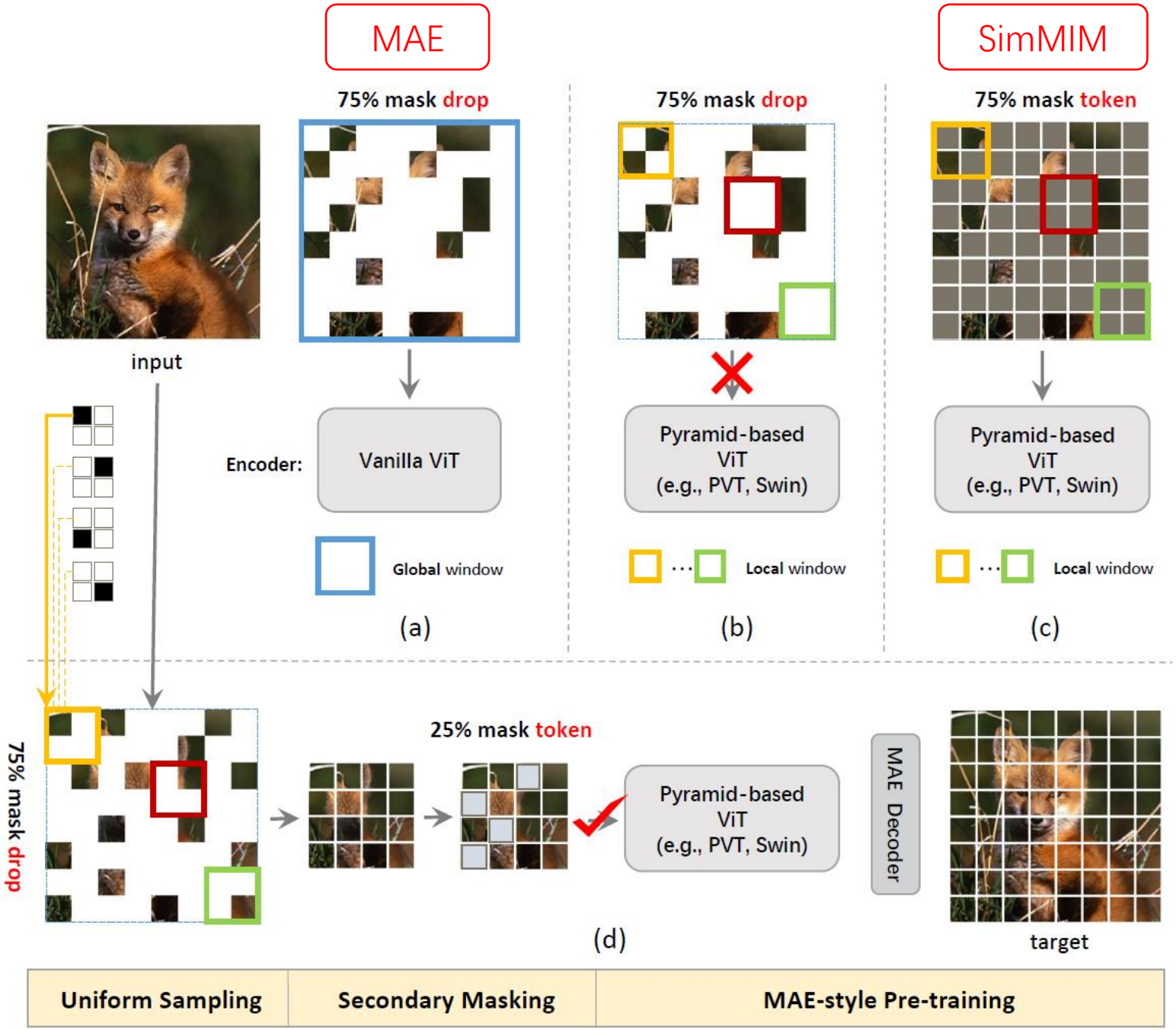


Figure 1. An illustration of our simple framework for masked language modeling, named *SimMIM*. It predicts raw pixel values of the randomly masked patches by a lightweight one-layer head, and performs learning using a simple ℓ_1 loss.

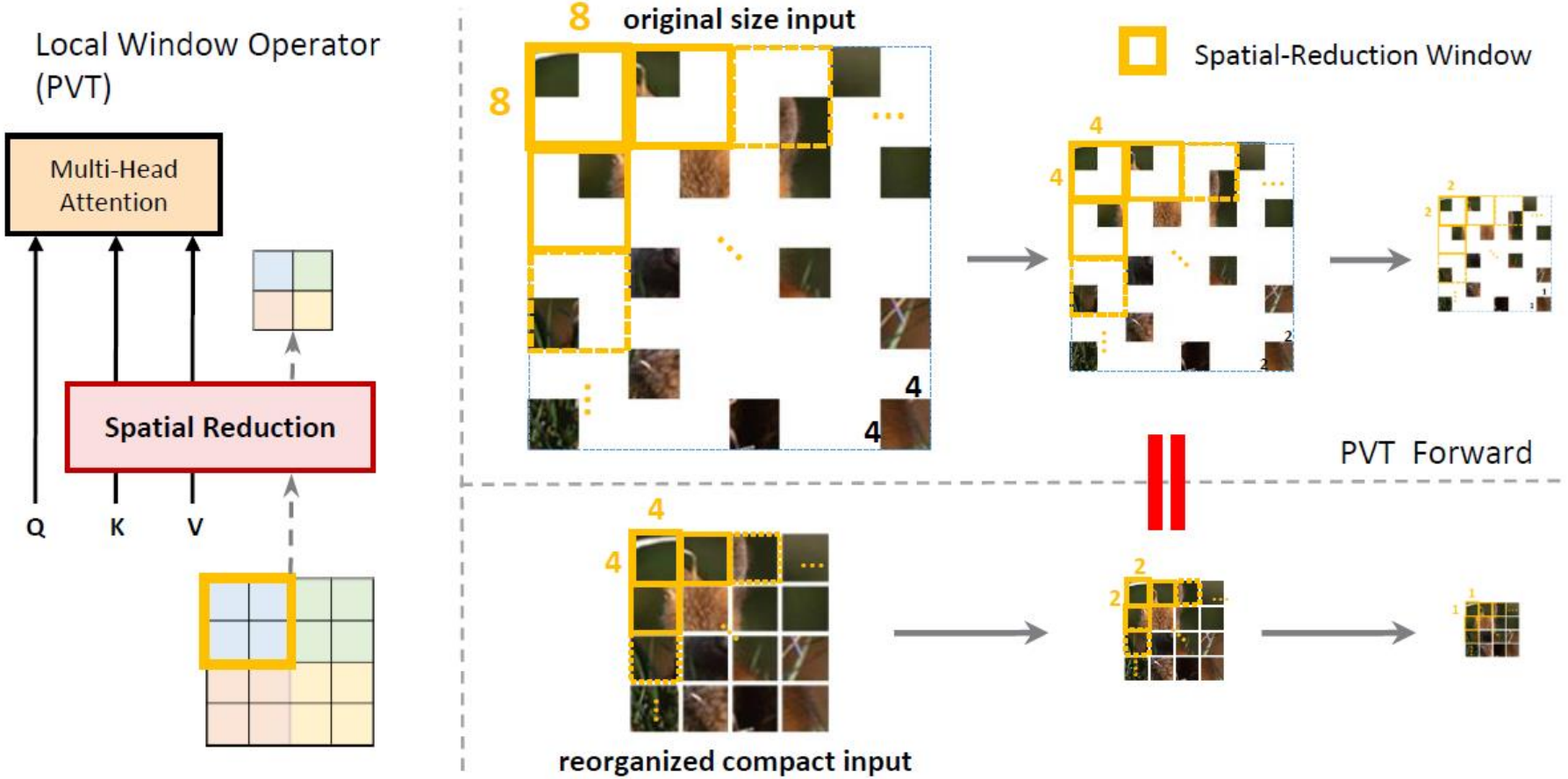
Xie Z, Zhang Z, Cao Y, et al. Simmim: A simple framework for masked image modeling[C]//CVPR. 2022: 9653-9663.

动机



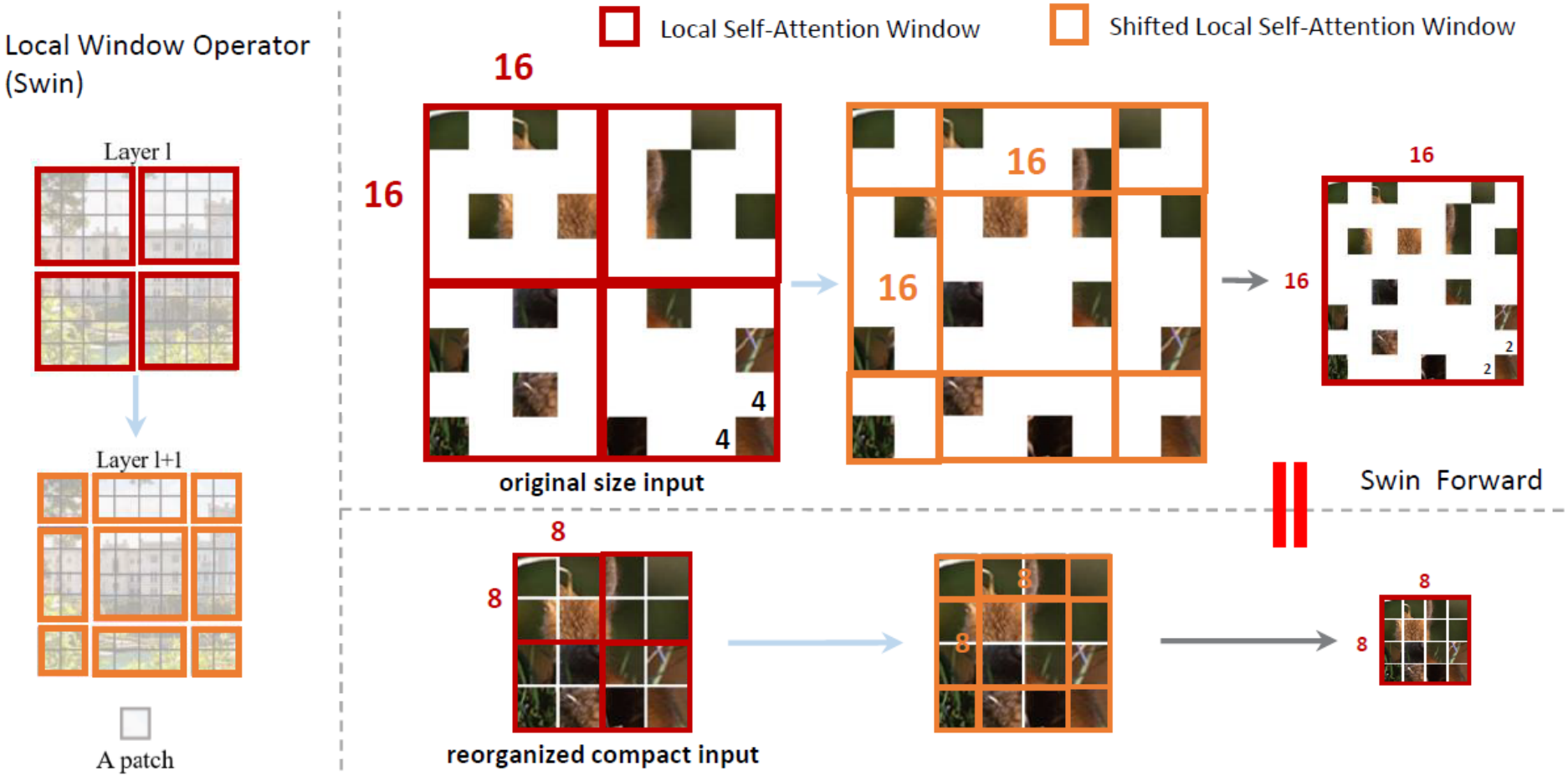
Li, X., Wang, W., Yang, L., & Yang, J. (2022). Uniform Masking: Enabling MAE Pre-training for Pyramid-based Vision Transformers with Locality. *arXiv preprint arXiv:2205.10063*.

Compatibility to PVT



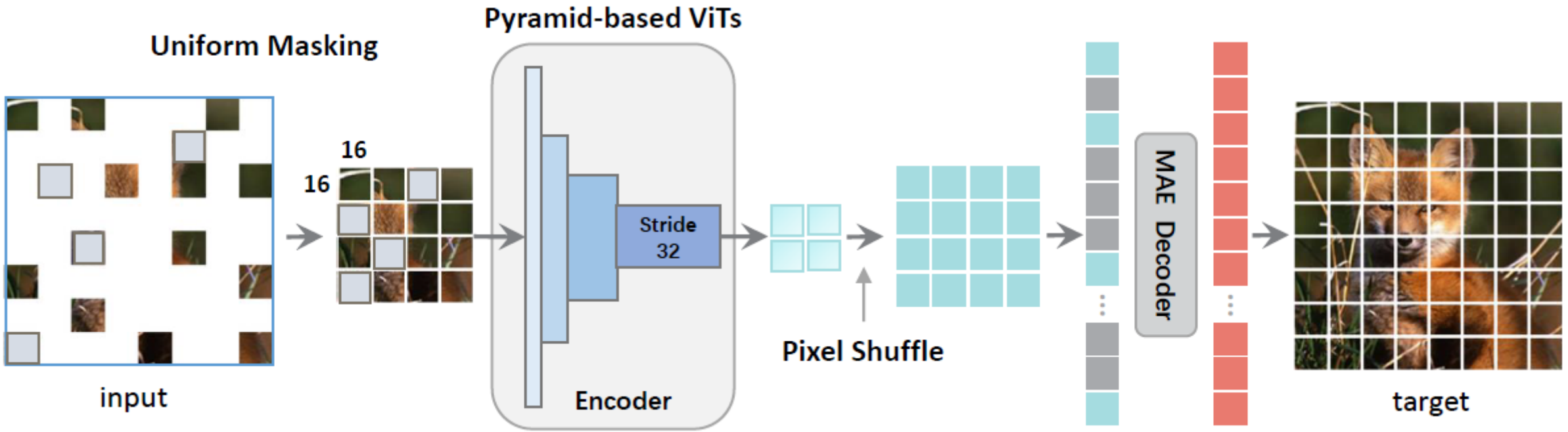
Wang W, Xie E, Li X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]//ICCV. 2021: 568-578.

Compatibility to Swin



Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//ICCV. 2021: 10012-10022.

UM-MAE Pipeline

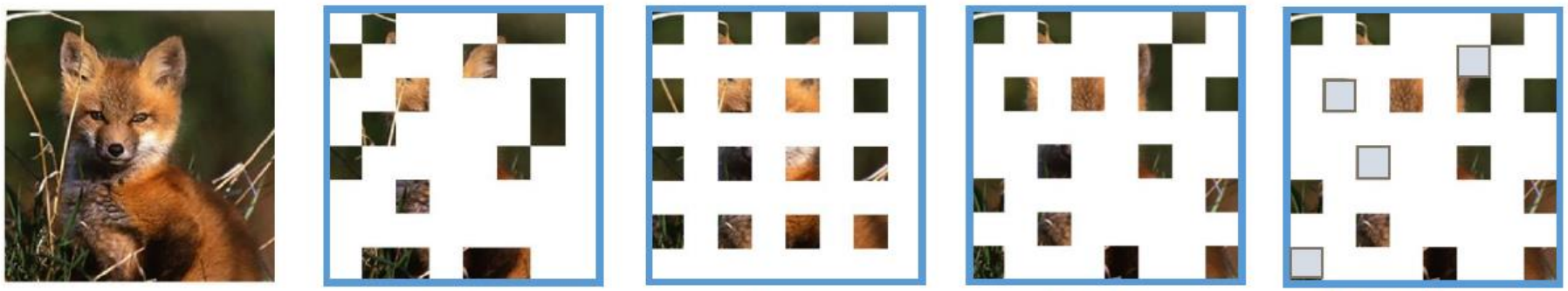


首先采用提出的均匀掩码策略对输入图像进行处理，将缩小后的图像输入层级ViT架构，例如Swin或PVT；接下来，利用亚像素卷积（即Pytorch中的Pixel Shuffle算子）恢复表征尺寸；最后采用MAE中的解码器，恢复出原始图像

Li, X., Wang, W., Yang, L., & Yang, J. (2022). Uniform Masking: Enabling MAE Pre-training for Pyramid-based Vision Transformers with Locality. *arXiv preprint arXiv:2205.10063*.

Effectiveness of UM

a）随机采样，　（b）网格采样，　（c）均匀采样，　（d）均匀掩码



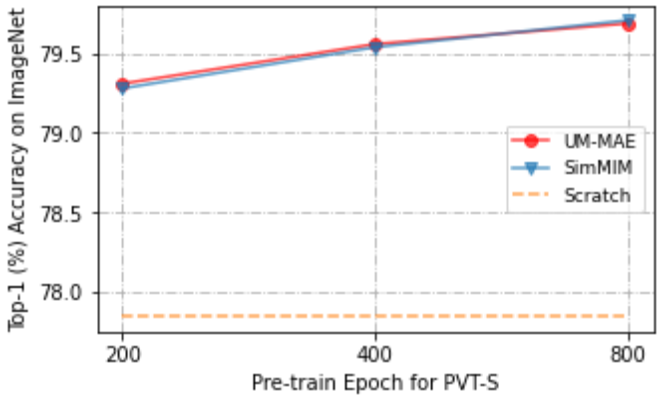
(a) RS　**(b)** GS　**(c)** US　**(d)** UM(US + SM)

Sampling Strategy (75% drop tokens)	Pyramid Support	SM Ratio	Pre-train Loss	ImageNet-1K	ADE20K		COCO		
				Top-1 Acc	mIoU	aAcc	AP	AP_{50}	AP_{75}
(a) RS (MAE [20] Baseline)	×	–	0.4256	**82.88**	**42.54**	**80.85**	**46.0**	**64.7**	**49.8**
(b) GS	✓	–	0.3682	82.48	38.79	79.16	44.4	63.2	48.6
(c) US (i.e., UM with 0% SM Ratio)	✓	–	0.3858	82.74	41.55	80.48	45.5	64.2	49.6
(d) UM (Ours)	✓	15%	0.4171	82.75	41.68	80.54	45.8	**64.6**	49.8
	✓	25%	0.4395	**82.88**	**42.59**	**80.80**	**45.9**	64.5	**50.2**
	✓	35%	0.4645	82.68	42.02	80.72	**45.9**	**64.6**	50.1

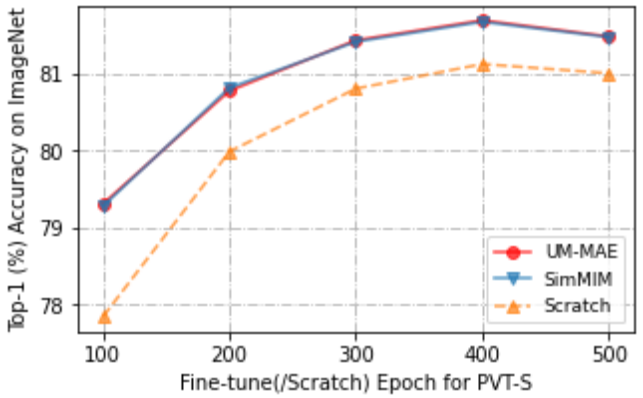
Li, X., Wang, W., Yang, L., & Yang, J. (2022). Uniform Masking: Enabling MAE Pre-training for Pyramid-based Vision Transformers with Locality. *arXiv preprint arXiv:2205.10063*.

Efficiency for Pyramid based ViTs

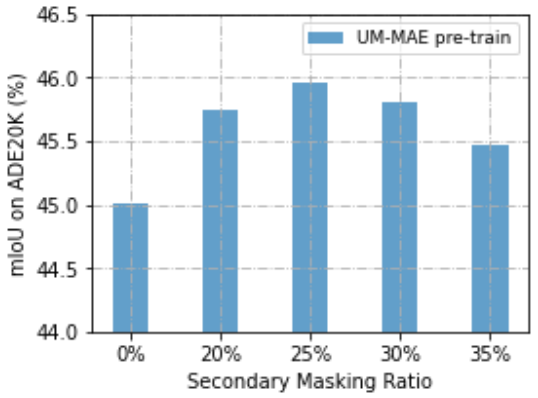
Architecture	Method	Pre-train (200 epoch)		Fine-tune (/Scratch) Performance		
		Time	Memory	ImageNet-1K	ADE20K	COCO
PVT-S [40]	Supervised from Scratch (Baseline)			77.84	40.38	42.3
	SimMIM [45]	38.0 h	20.6 GB	79.28 (+1.44)	**43.04 (+2.66)**	44.8 (+2.5)
	UM-MAE **(ours)**	**21.3 h**	**11.6 GB**	**79.31 (+1.47)**	43.01 (+2.63)	**45.1 (+2.8)**
Swin-T [31]	Supervised from Scratch (Baseline)			81.82	44.51	47.2
	SimMIM [45]	49.3 h	37.4 GB*	**82.20 (+0.38)**	45.35 (+0.84)	47.6 (+0.4)
	UM-MAE **(ours)**	**25.0 h**	**13.4 GB**	82.04 (+0.22)	**45.96 (+1.45)**	**47.7 (+0.5)**



(a)



(b)



(c)

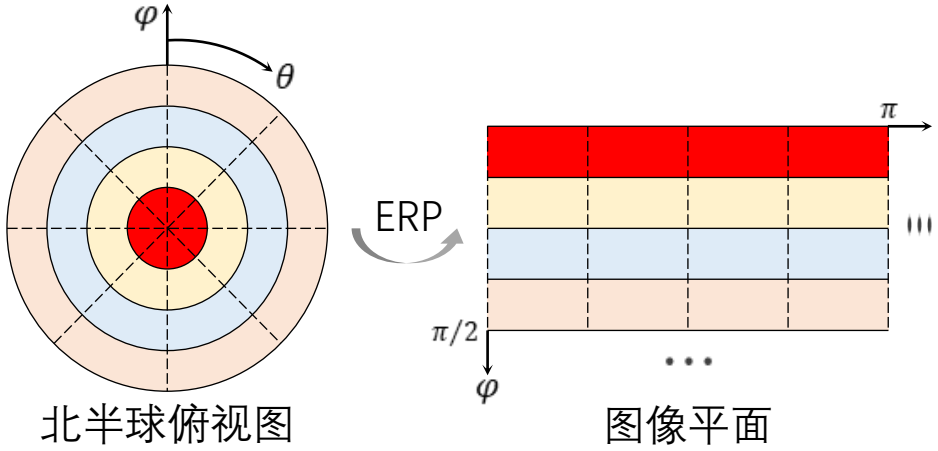
基于多模态掩码预训练的全景深度补全

- 面向实际应用需求，我们定义了全景深度补全的新任务（问题）；
- 不同于现有基于**单模态**和**Transformer**框架的预训练方法，我们提出了面向**多模态**和**CNN模型的预训练**策略

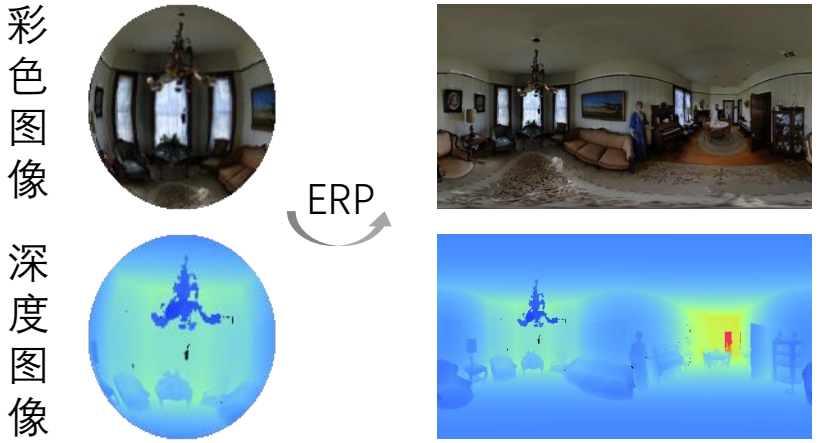
Z. Yan, X. Li, K. Wang, Z. Zhang, J. Li, J. Yang. Multi-Modal Masked Pre-Training for Monocular Panoramic Depth Completion, *ECCV 2022.*

研究背景

- **全景图像**



(a) 等量矩形投影 (ERP)



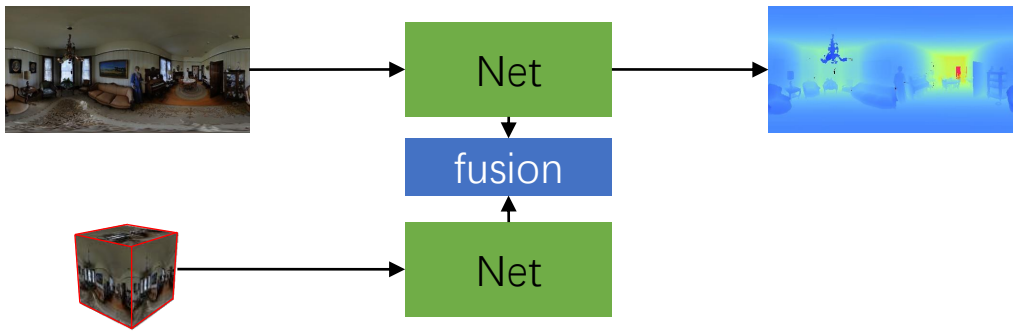
(b) 全景图像投影示例

[1]. A. Chang, A. Dai, T. Funkhouser, et al. Matterport3D: Learning from RGB-D Data in Indoor Environments, *3DV 2017*.

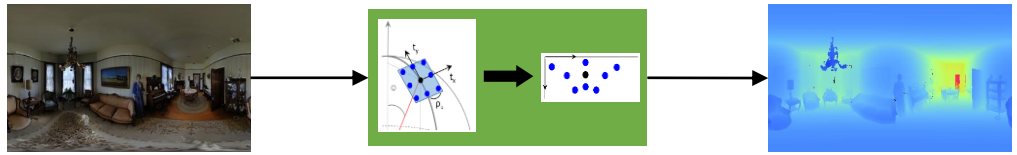
[2]. N. Zioulis, A. Karakottas, D. Zarpalas, et al. OmniDepth: Dense Depth Estimation for Indoors Spherical Panoramas, *ECCV 2018*.

研究背景

- 全景深度估计任务：从全景彩色图像预测深度图



(a) 常规视角辅助



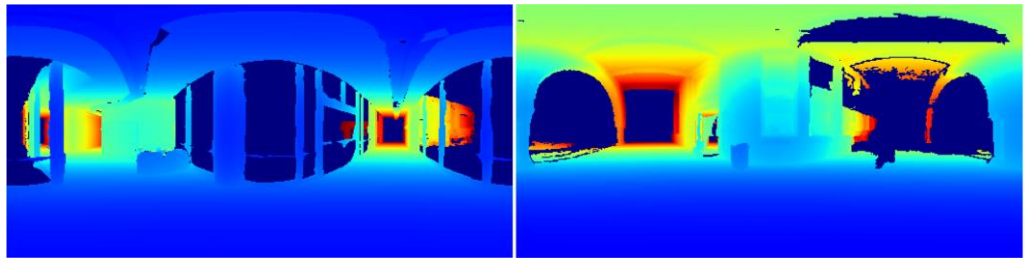
(b) 畸变感知卷积

[1]. F. Wang, Y. Yeh, M. Sun. BiFuse: Monocular 360° Depth Estimation via Bi-Projection Fusion, *CVPR 2020*.

[2]. K. Tateno, N. Navab, F. Tombari. Distortion-Aware Convolutional Filters for Dense Prediction in Panoramic Images, *ECCV 2018*.

研究任务

- **全景深度补全任务：给定从全景彩色图像，补全对应的稀疏深度图**



(a) 全景深度图存在深度值缺失（深蓝色部分）

(b)从全景相机获取的深度图比较稀疏

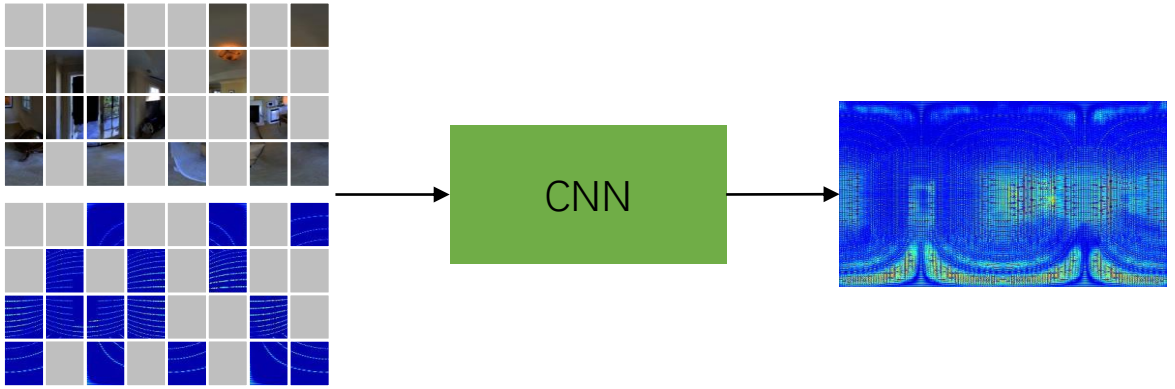
[1]. J. Uhrig, N. Schneider, L. Schneider, et al. Sparsity Invariant Cnns. *3DV 2017*.

[2]. S. Nathan, H. Derek, K. Pushmeet, et al. Indoor Segmentation and Support Inference from Rgbd Images, *ECCV 2012*.

M³PT: Multi-Modal Masked Pre-Training

- **多模态（RGB-D）预训练**
- **CNN作为骨架网络的可行性**

(a) MAE范式 (单模态+transformer)



(b) M³PT范式 (多模态+CNN)

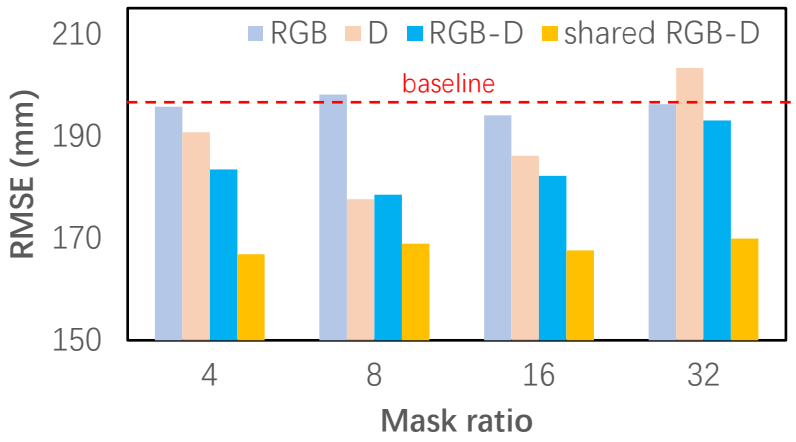
[1]. K. He, X. Chen, X. Xie, et al, Masked Autoencoders Are Scalable Vision Learners. *CVPR 2022*.

[2]. Z. Xie, Z. Zhang, Y. Cao, et al. Simmim: A Simple Framework for Masked Image Modeling, *CVPR 2022*.

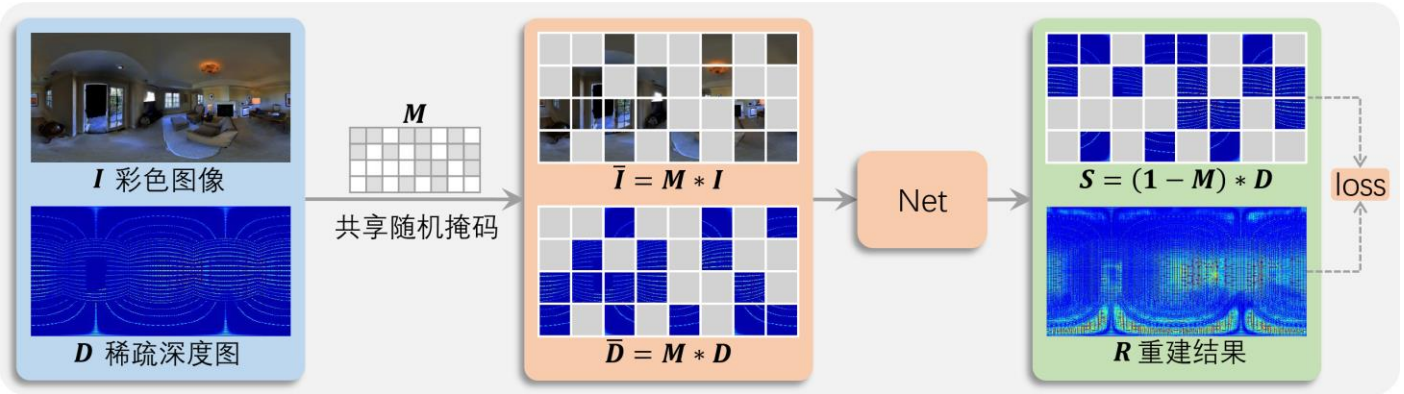
M³PT: Multi-Modal Masked Pre-Training

- **掩码策略选择**

（1）仅掩码RGB图，（2）仅掩码深度图D，（3）同时掩码RGB-D，但不共享掩码器，以及（4）用共享的掩码器对RGB-D进行掩码



随机掩码对比实验

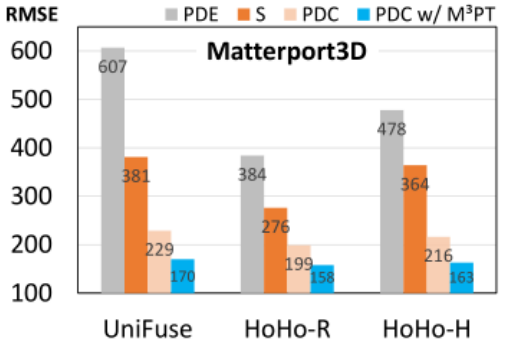


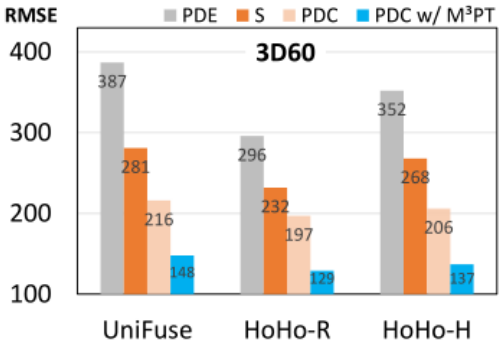
实验结果

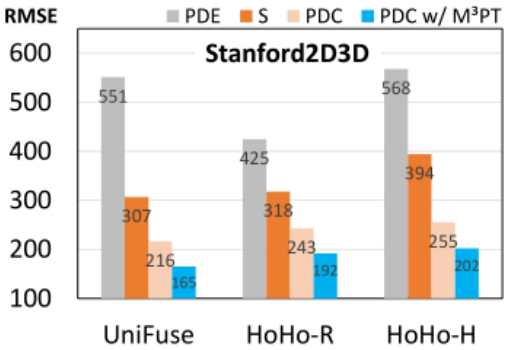
- 对比实验

Dataset	Method	Error Metric ↓				Accuracy Metric ↑		
		MRE	MAE	RMSE	RMSElog	$\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$
Matterport3D	UniFuse [23]	0.0475	95.2	229.1	0.0381	0.9710	0.9924	0.9970
	HoHo-R [46]	0.0355	75.0	199.2	0.0311	0.9806	0.9945	0.9977
	HoHo-H [46]	0.0406	85.7	215.5	0.0337	0.9772	0.9938	0.9975
	PENet [21]	0.0493	91.5	248.0	0.0350	0.9728	0.9935	0.9970
	GuideNet [47]	0.0438	87.2	192.9	0.0327	0.9806	0.9948	0.9981
	M^{3}PT	**0.0164**	**36.2**	**138.9**	**0.0193**	**0.9927**	**0.9976**	**0.9990**
Stanford2D3D	UniFuse [23]	0.0489	93.4	216.2	0.0392	0.9661	0.9919	0.9973
	HoHo-R [46]	0.0677	123.9	242.5	0.0478	0.9463	0.9862	0.9959
	HoHo-H [46]	0.0695	127.9	254.8	0.0497	0.9434	0.9852	0.9957
	PENet [21]	0.0530	95.9	200.6	0.0404	0.9694	0.9934	0.9981
	GuideNet [21]	0.0506	92.1	196.7	0.0380	0.9689	0.9926	0.9978
	M^{3}PT	**0.0274**	**52.9**	**149.0**	**0.0263**	**0.9859**	**0.9963**	**0.9988**
3D60	UniFuse [23]	0.0446	94.1	215.6	0.0342	0.9749	0.9947	0.9984
	HoHo-R [46]	0.0338	75.6	196.9	0.0294	0.9818	0.9954	0.9983
	HoHo-H [46]	0.0376	81.9	205.8	0.0317	0.9788	0.9947	0.9981
	PENet [21]	0.0680	120.3	233.9	0.0321	0.9743	0.9926	0.9980
	GuideNet [21]	0.0689	144.2	239.3	0.0418	0.9711	0.9954	0.9987
	M^{3}PT	**0.0144**	**34.1**	**127.2**	**0.0165**	**0.9944**	**0.9985**	**0.9995**

- 消融实验





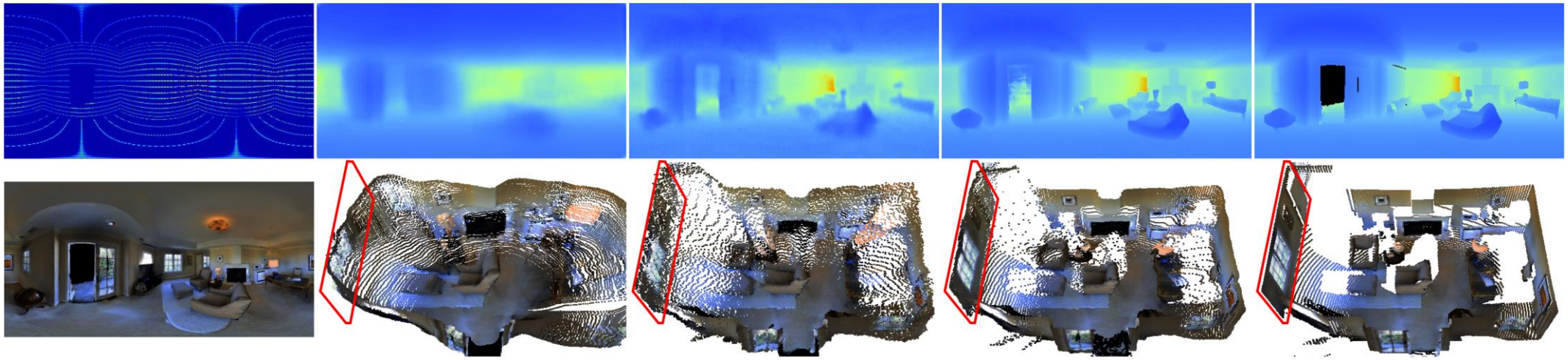


Method	RMSE	MAE	iRMSE	iMAE
S2D [35]	858.02	311.47	3.07	1.67
+M^{3}PT	**844.16**	**267.64**	**3.01**	**1.51**
GuideNet [47]	777.78	221.59	2.39	**1.00**
+M^{3}PT	**761.57**	**217.68**	**2.26**	**1.00**
ACMNet [64]	789.72	216.65	2.32	0.96
+M^{3}PT	**774.63**	**209.31**	**2.25**	**0.93**

常规视角深度补全数据集KITTI的泛化实验

引入多模态掩码预训练不仅可以降低全景深度图的预测误差，还能有效提升常规深度补全的性能

可视化结果



我们的方法提升了三维重建的结果

主要内容

（1）对比范式

Contrastive Learning

（2）掩码范式

Masked Autoencoders

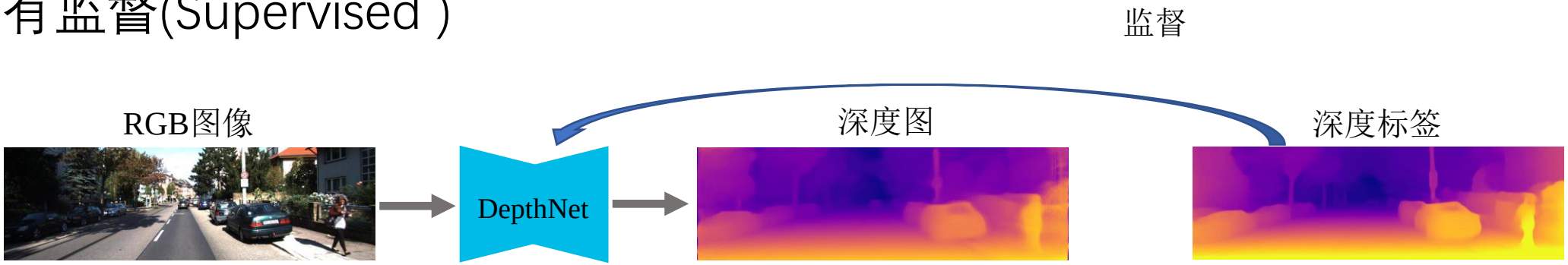
（3）内隐范式

内隐范式

- 隐含了监督信号，比如
- 读书百遍，其义自见
- 从试卷中找到答案
- 空间几何结构中内隐了距离信息
- **单目深度估计**

单目深度估计

- 有监督(Supervised)



- 无监督或自监督（Unsupervised, Self-supervised）

  训练阶段无需任何深度标签数据，从图像本身的几何特性中获取监督信号

自监督（无监督）深度估计的基本思想

- 自监督（无监督）单目深度估计方法的基本思路

  将深度估计任务转化成一个图像重建问题，利用图像重建损失来训练网络

- 主要分为两类：

  1）基于双目立体对的自监督单目深度估计

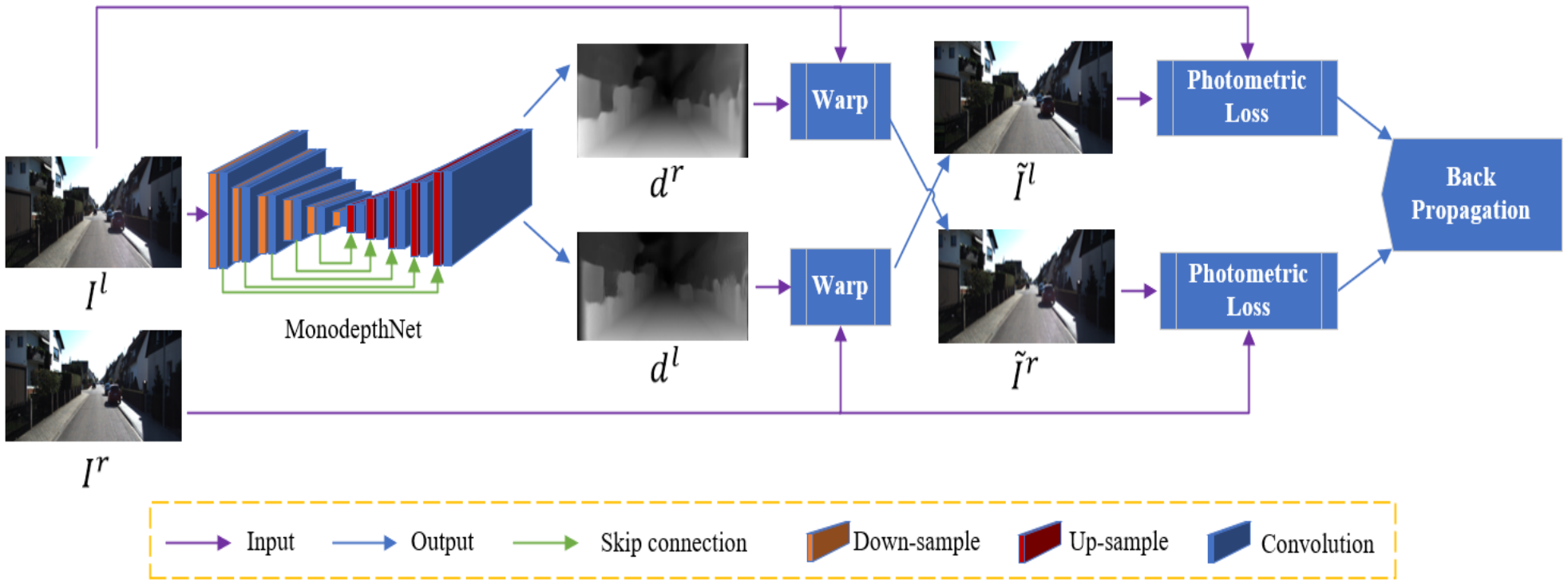
  代表性工作： Monodepth[1]

  2）基于单目视频流的自监督单目深度估计

  代表性工作：SfM-Learner[2]

[1] Godard C, Mac Aodha O, Brostow G J. Unsupervised monocular depth estimation with left-right consistency[C]//CVPR. 2017: 270-279.
[2] Zhou T, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motion from video[C]//CVPR. 2017: 1851-1858.

基于双目立体对的自监督单目深度估计：Monodepth



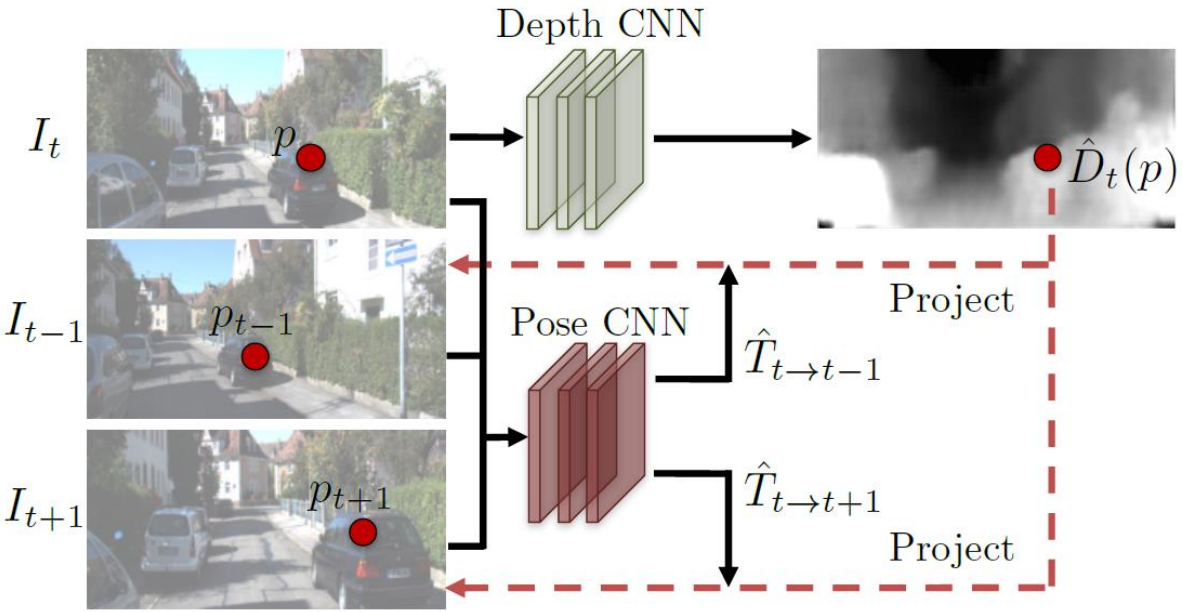
➢ 训练阶段

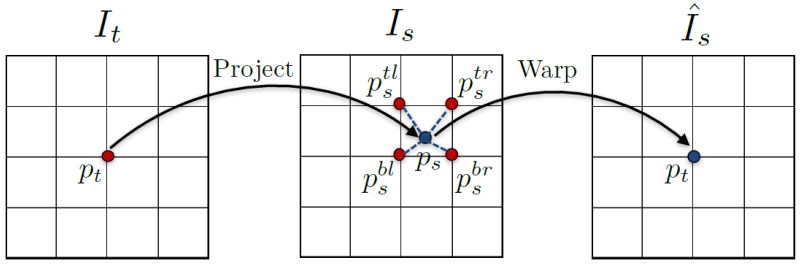
深度估计网络MonodepthNet**预测双目图像的视差图**，然后**通过视差图进行图像重建**，最后将图像重建损失作为目标函数来训练网络

➢ 测试阶段

通过公式将视差图d^l转化成深度图D：$D = b * f / d^l$

基于单目视频流的自监督单目深度估计：SfM-Learner





$$p_s \sim K\hat{T}_{t\to s}\hat{D}_t(p_t)K^{-1}p_t$$

: target像素坐标系 ---> target相机坐标系

: target相机坐标系 ---> source相机坐标系

: source相机坐标系 ---> source像素坐标系

➢ 训练阶段

同时训练两个网络：DepthNet和PoseNet，分别用来预测深度图和相机的相对位姿。通过预测出的深度图和相机位姿进行图像重建，将图像重建损失作为目标函数来同时训练两个网络

➢ 测试阶段

测试阶段DepthNet和PoseNet分开单独测试，输入单张图像到DepthNet中得到对应的深度图

自监督的单目深度估计: 我们的工作

- 针对夜晚环境：**正则化的自监督单目深度估计方法**
- 面向开放环境：**防止遗忘的单目视频在线深度估计方法**

Kun Wang, Zhiqiang Yan, Xiang Li, Zhenyu Zhang, Jun Li*, Baobei Xu, Jian Yang*, Regularizing Nighttime Weirdness: Efficient Self-supervised Monocular Depth Estimation in the Dark, ICCV2021.
Zhenyu Zhang, Yan Yan, Stéphane Lathuilière, Elisa Ricci, Jian Yang*, Nicu Sebe, Online Depth Learning against Forgetting in Monocular Videos, CVPR 2020

夜间自监督深度估计的主要挑战

低光照

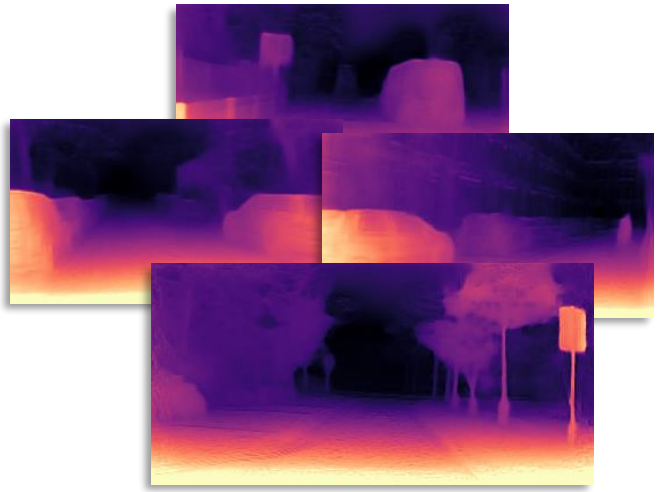
变化的照明

导致现有的方法无法很好的预测深度，生成带有空洞和不连续区域的深度图像

研究动机

日间场景

自监督训练



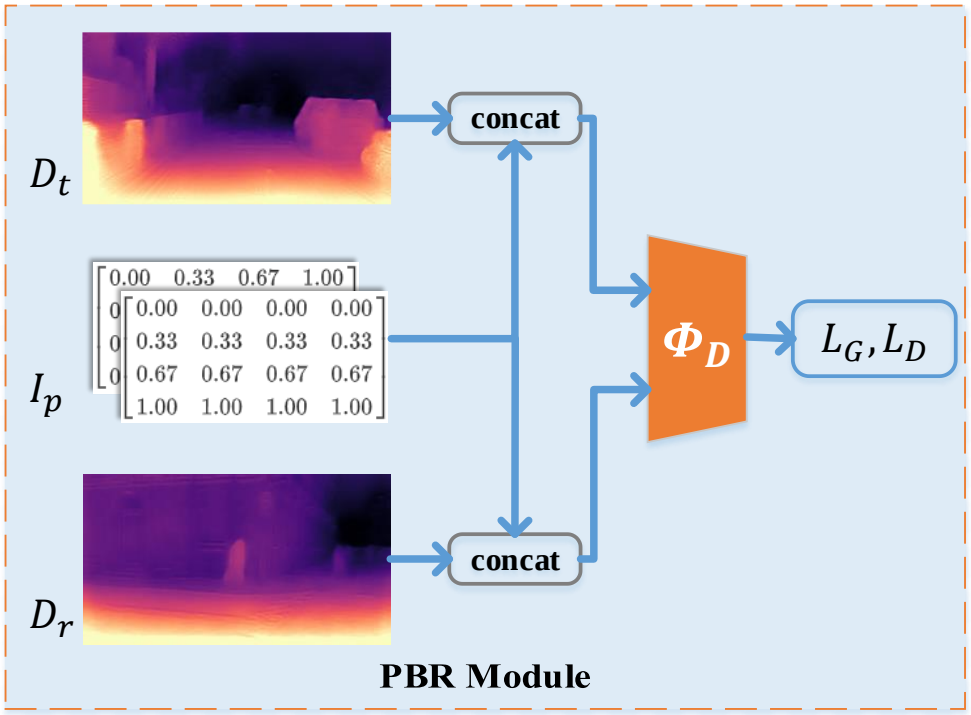
预测的深度图

现有的自监督深度估计算法可以在白天环境下产生良好的深度预测

白天环境下的深度图与夜间环境下的并无差异，却更容易获得

方法思想：基于先验深度分布的正则化（PBR）

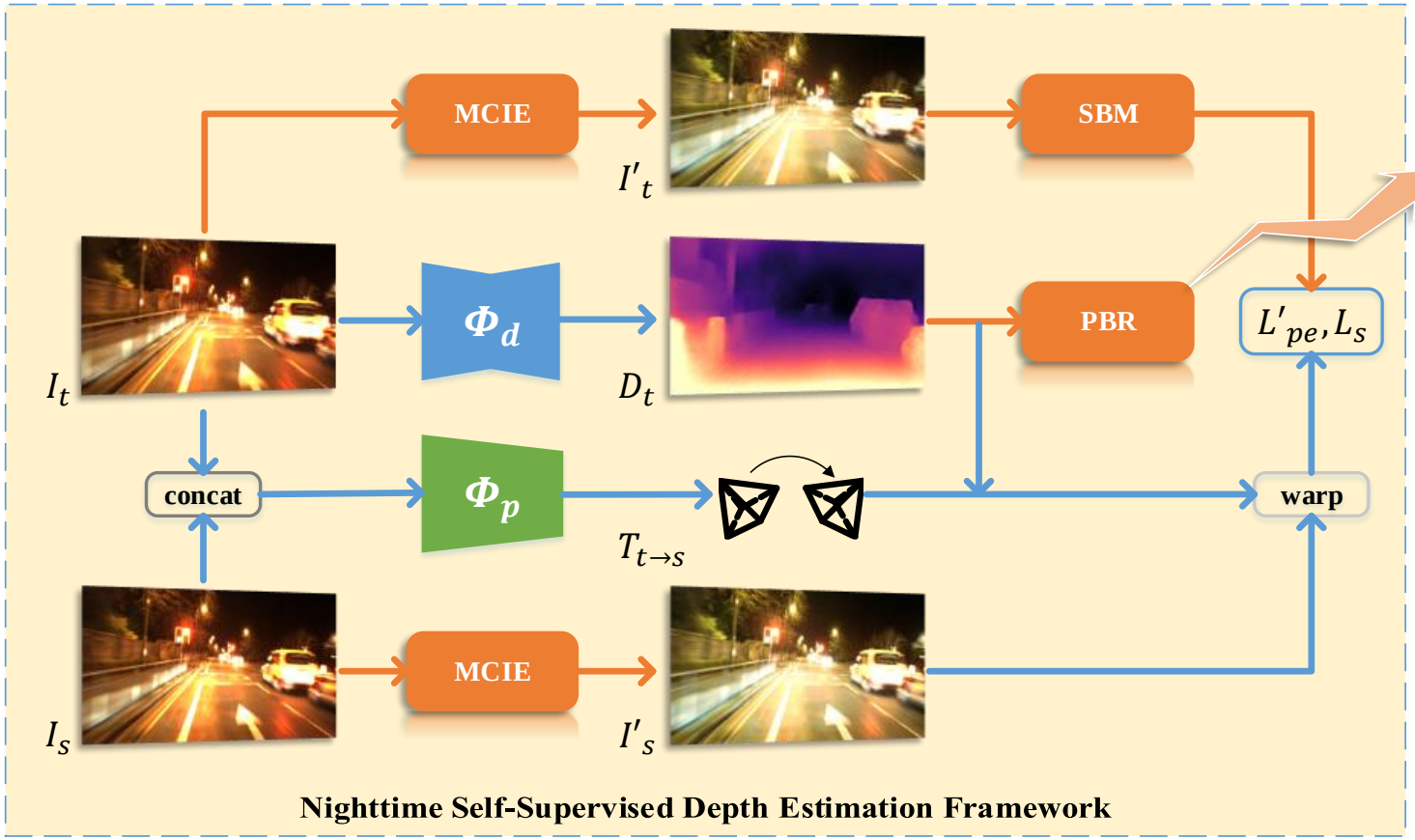
通过对抗学习的方式，约束模型在**夜间的预测结果**Dt与**白天的参考深度图**Dr在**分布上保持一致**，从而降低重建损失的歧义性，产生更加合理的深度估计结果

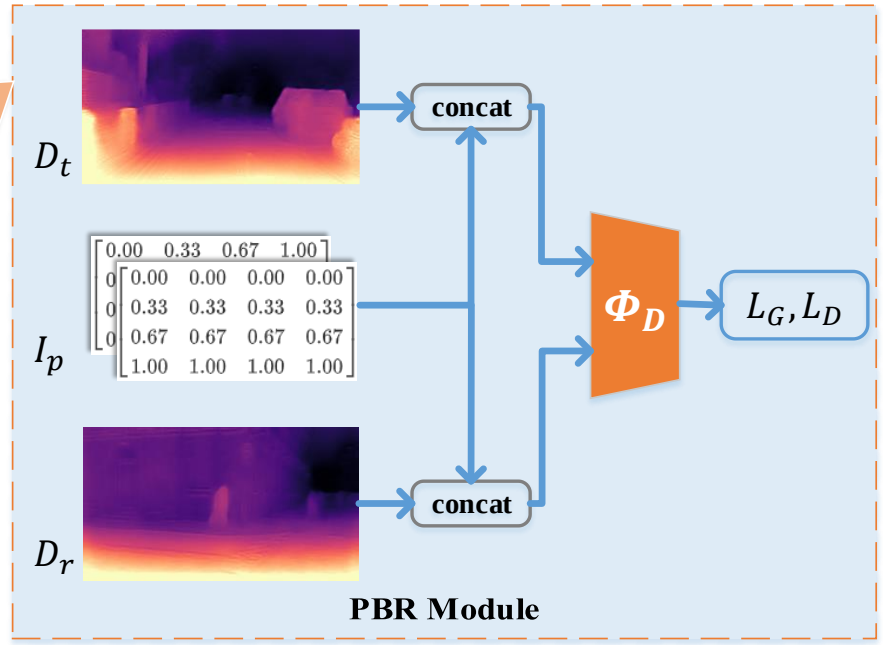


$$\min_{\omega_D} L_D = \frac{1}{2}\mathbb{E}_{D_r \in \{D_r\}}[(\Phi_D(cat(I_p, \mu(D_r))) - 1)^2] + \frac{1}{2}\mathbb{E}_{D_t \in \{D_t\}}[\Phi_D(cat(I_p, \mu(D_t)))^2]$$

$$\min_{\omega_d} L_G = \frac{1}{2}\mathbb{E}_{D_t \in \{D_t\}}[(\Phi_D(cat(I_p, \mu(D_t))) - 1)^2],$$

总体框架

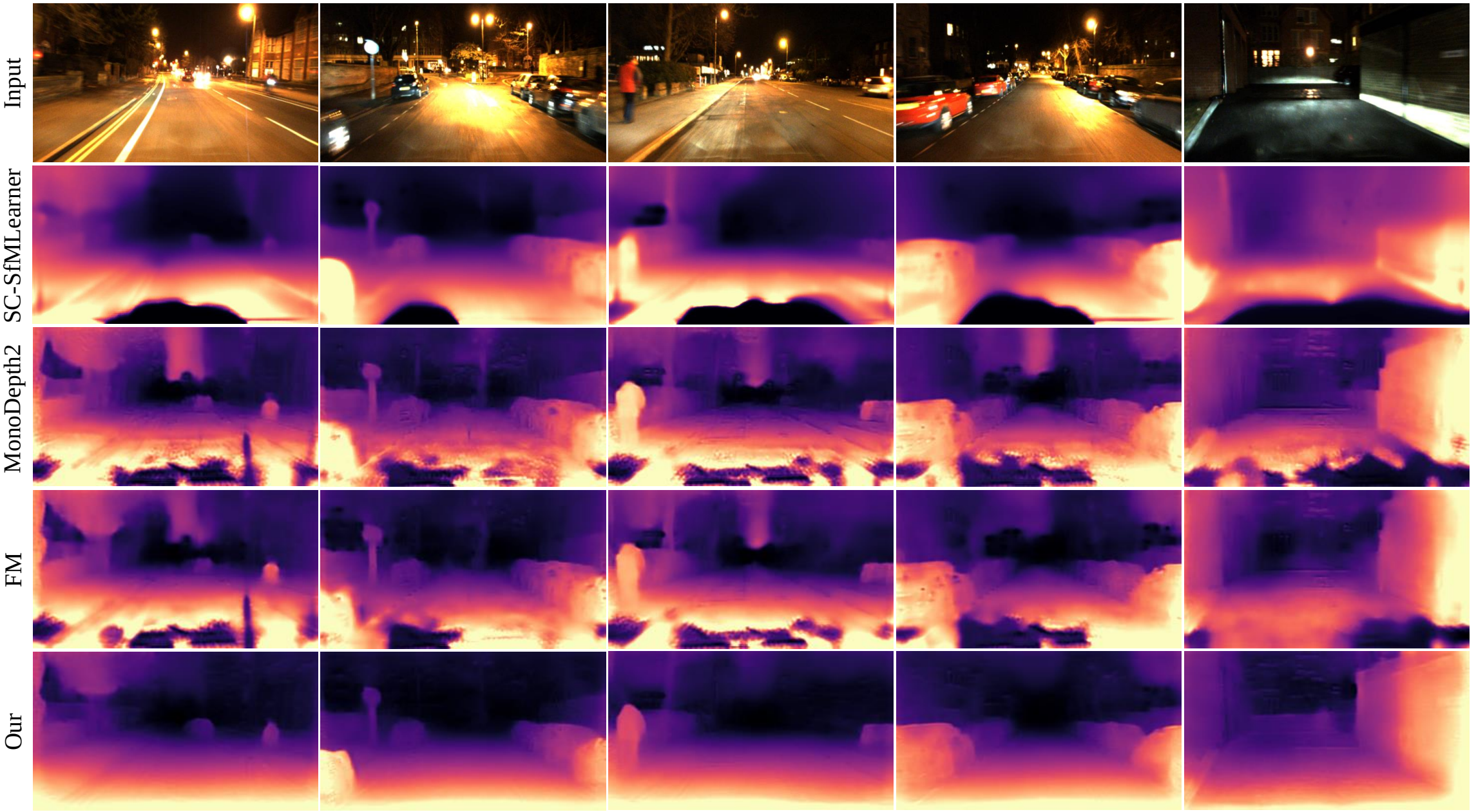

MCIE
SBM
I'_t
I_t
Φ_d
D_t
PBR
L'_{pe}, L_s
concat
Φ_p
warp
$T_{t\to s}$
I_s
MCIE
I'_s
Nighttime Self-Supervised Depth Estimation Framework


D_t
concat
I_p
0.00 0.33 0.67 1.00
0.00 0.00 0.00 0.00
0.33 0.33 0.33 0.33
0.67 0.67 0.67 0.67
1.00 1.00 1.00 1.00
Φ_D
L_G, L_D
D_r
concat
PBR Module

与现有方法的定量比较

Method	Abs Rel	Sq Rel	RMSE	RMSE log	δ_1	δ_2	δ_3
RobotCar-Night							
MonoDepth2 [15]	0.3999	7.4511	6.6416	0.4429	0.7444	0.8921	0.9280
SfMLearner [54]	0.6754	15.4334	9.4324	0.6046	0.5465	0.8003	0.8733
SC-SfMLearner [2]	0.6029	16.0173	9.2453	0.5620	0.7185	0.8722	0.9091
PackNet [18]	0.2836	4.0257	5.3864	0.3351	0.7425	0.9143	0.9560
FM [41]	0.3953	7.5579	6.7002	0.4391	0.7605	0.8943	0.9299
DeFeat-Net [42]	0.3929	4.8955	6.3429	0.4236	0.6256	0.8290	0.8992
MonoDepth2 (Day)	0.3211	1.8672	4.9818	0.3568	0.4446	0.7813	0.9353
FM (Day)	0.2928	1.5380	4.5951	0.3337	0.4888	0.8054	0.9497
Reg Only	0.5006	3.7608	6.6351	0.7518	0.2841	0.5643	0.8156
Our	**0.1205**	**0.5204**	**2.9015**	**0.1633**	**0.8794**	**0.9688**	**0.9896**
nuScenes-Night							
MonoDepth2 [15]	1.1848	42.3059	21.6129	1.5699	0.1842	0.3598	0.5044
SfMLearner [54]	0.6004	8.6346	15.4351	0.7522	0.2145	0.4166	0.5961
SC-SfMLearner [2]	1.0508	30.5865	19.6004	0.8854	0.1823	0.3673	0.5422
PackNet [18]	1.5675	61.5101	25.8318	1.3717	0.1387	0.2980	0.4313
FM [41]	1.1383	41.6166	20.8481	1.1483	0.2376	0.4252	0.5650
Our	**0.3150**	**3.7926**	**9.6408**	**0.4026**	**0.5081**	**0.7776**	**0.8959**

与现有方法在RobotCar数据集上的可视化比较



防止遗忘的单目视频在线深度估计方法

- **开放环境下的在线深度估计**

利用源域自动驾驶场景数据训练出的模型，无法很好地适应开放环境下不断变化的目标域场景

- **在线深度估计面临的问题**

模型容易过拟合至当前环境而遗忘先前学习得到的经验

  - 尺度混淆
  - 无监督深度估计框架的脆弱性

Zhenyu Zhang, Stephane Lathuiliere, Elisa Ricci, Nicu Sebe, Yan Yan, Jian Yang, Online Depth Learning Against Forgetting in Monocular Videos, in CVPR 2020

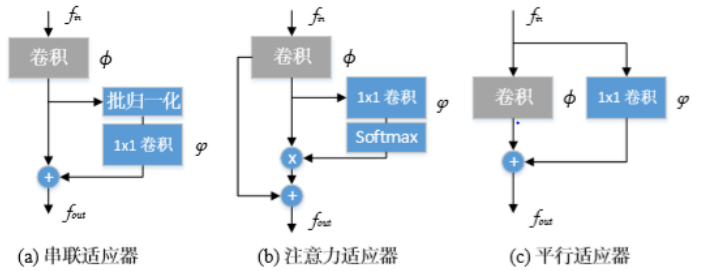
防止遗忘的单目视频在线深度估计方法

研究思路

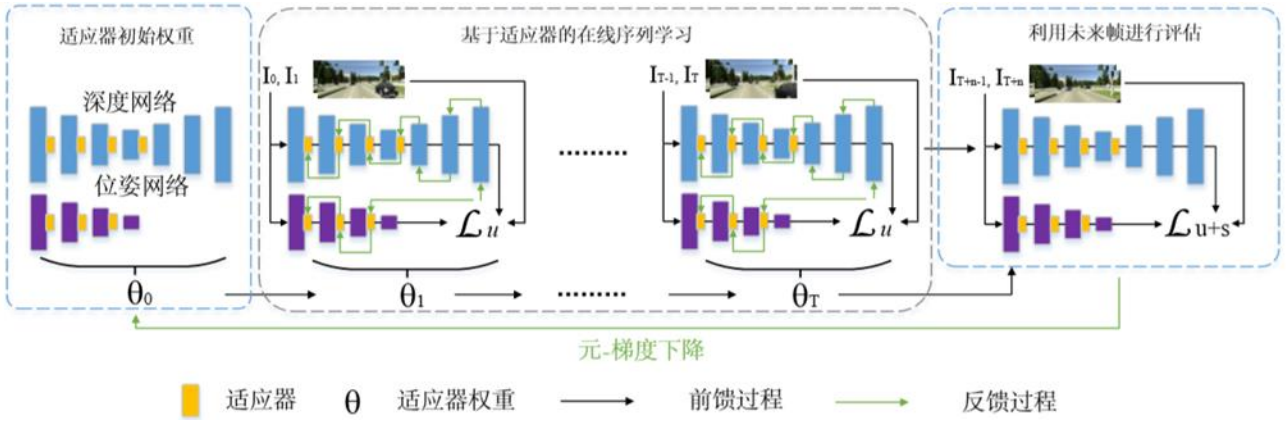
- 提出了一种新的防止遗忘的无监督单目深度估计框架
- 不同于通常采用的更新模型参数的方法，我们设计了一系列的**适应器**以有效的调整特征分布与表达，避免模型经验的丢失

$$\begin{aligned}\hat{\mu}^{t|t} = \mu^t &= (1-a^t)\hat{\mu}^{t|t-1} + a^t\bar{u}^t, \\ &= (1-a^t)\mu^{t-1} + a^t\tilde{\mu}^t \\ \hat{\Sigma}^{t|t} = \Sigma^t &= (1-a^t)\hat{\Sigma}^{t|t-1} + a^t S^t \\ &= (1-a^t)\Sigma^{t-1} + a^t\tilde{\Sigma}^t\end{aligned}$$

$$\hat{x} = \omega\frac{x-\mu^t}{\sqrt{\Sigma^{t^2}+\epsilon}} + \rho.$$



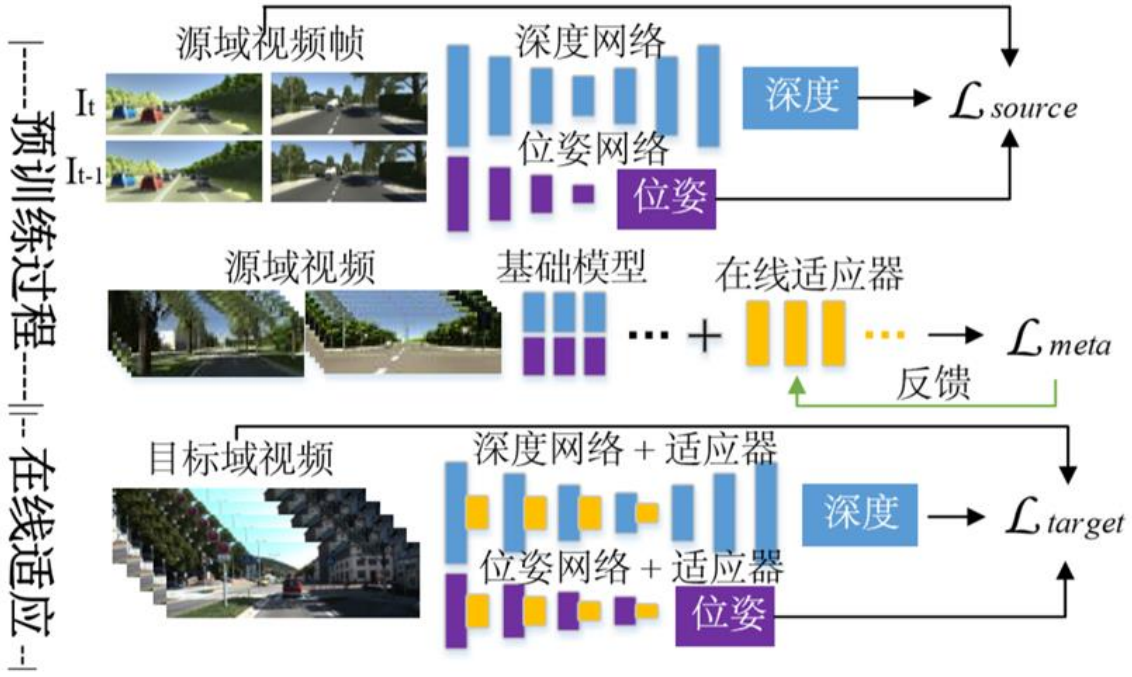
(a) 串联适应器　(b) 注意力适应器　(c) 平行适应器



防止遗忘的单目视频在线深度估计方法

模型与方法

- 利用源域视频帧进行预训练，使模型获得深度估计能力
- 采用基于元学习的预训练方法，在源域视频流上学习适应器
- 在目标域视频流上进行在线适应，不断更新适应器



防止遗忘的单目视频在线深度估计方法

- 优于当前最好的算法，与理想模型（无在线学习）可比

表 6.5: 与理想方法和当前最优方法的对比

		在线评估效果				最后20%帧的在线评估效果			
方法	训练集	Abs Rel ↓	RMSE ↓	< 1.25 ↑	$< 1.25^2$ ↑	Abs Rel ↓	RMSE ↓	< 1.25 ↑	$< 1.25^2$ ↑
SfM-Learner [29]									
理想模型（无在线学习）	Kitti	**0.2024**	6.5597	0.7180	0.8935	0.2091	6.5403	**0.7111**	0.8879
理想模型+ 一般方法	Kitti	0.2032	**6.5080**	**0.7220**	**0.8989**	**0.2113**	**6.4522**	0.7075	**0.8962**
一般方法	vKitti	0.2242	7.1179	0.6558	0.8709	0.2195	6.9022	0.6683	0.8798
L2A [47]	vKitti	0.2171	6.8024	0.6759	0.8762	0.2103	6.7256	0.6783	0.8791
LPF (Ours)	vKitti	**0.2033**	**6.5613**	**0.6835**	**0.8965**	**0.1962**	**6.3887**	**0.7147**	**0.8996**
SC-SfM-Learner [33]									
理想模型（无在线学习）	Kitti	0.1537	5.6295	0.8086	0.9338	0.1535	5.6412	0.8027	0.9335
理想模型+ 一般方法	Kitti	**0.1468**	**5.2768**	**0.8203**	**0.9431**	**0.1399**	**5.1413**	**0.8320**	**0.9479**
一般方法	vKitti	0.1782	5.9408	0.7473	0.9021	0.1662	5.7583	0.7596	0.9198
L2A [47]	vKitti	0.1708	5.8764	0.7548	0.9157	0.1615	5.6831	0.7728	0.9213
LPF (Ours)	vKitti	**0.1628**	**5.6581**	**0.7762**	**0.9234**	**0.1495**	**5.4327**	**0.7936**	**0.9301**

防止遗忘的单目视频在线深度估计方法

- 具有更稳定的在线学习能力

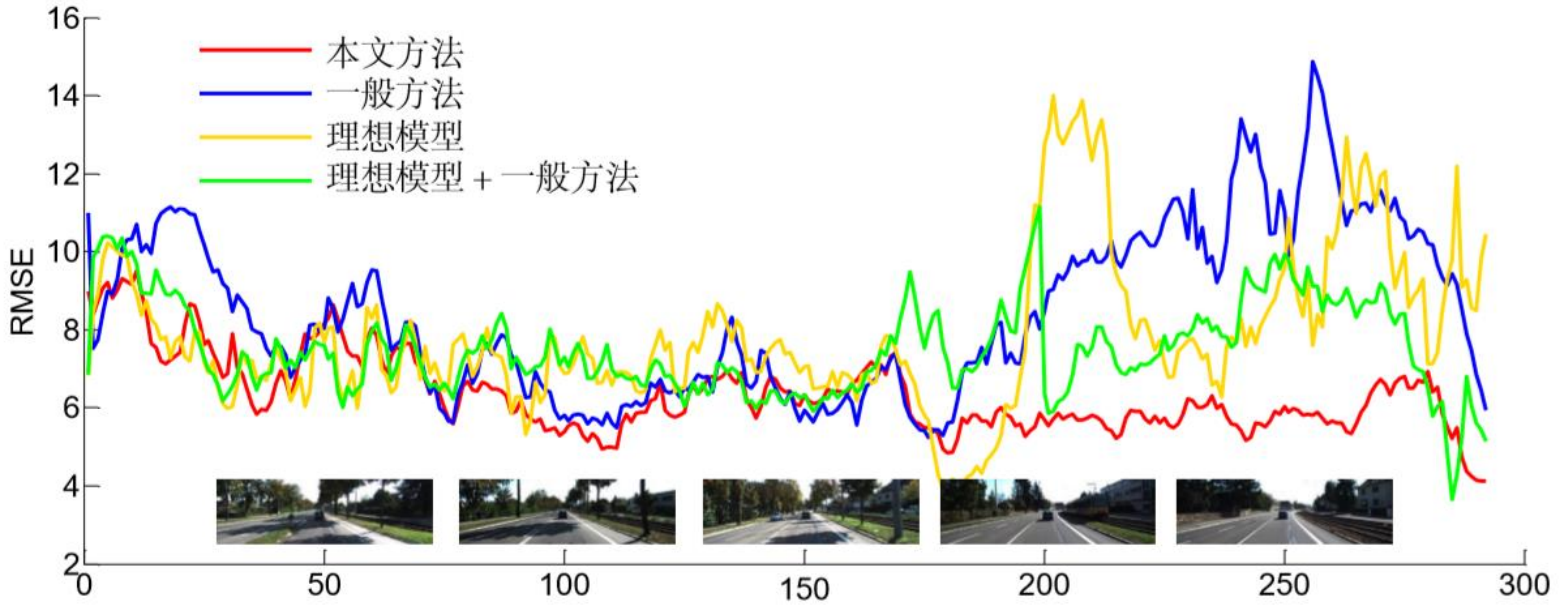


图 6.4: 不同模型在同一段KITTI视频序列中的在线学习效果。根据视频图像和理想模型预测效果的震荡，可以得知在$t = 200$处环境发生了突变。一般的方法无法处理该问题，使得模型收敛变慢。而LPF方法则表现出稳定和鲁棒的在线学习效果。

谢谢大家！

敬请批评指正！