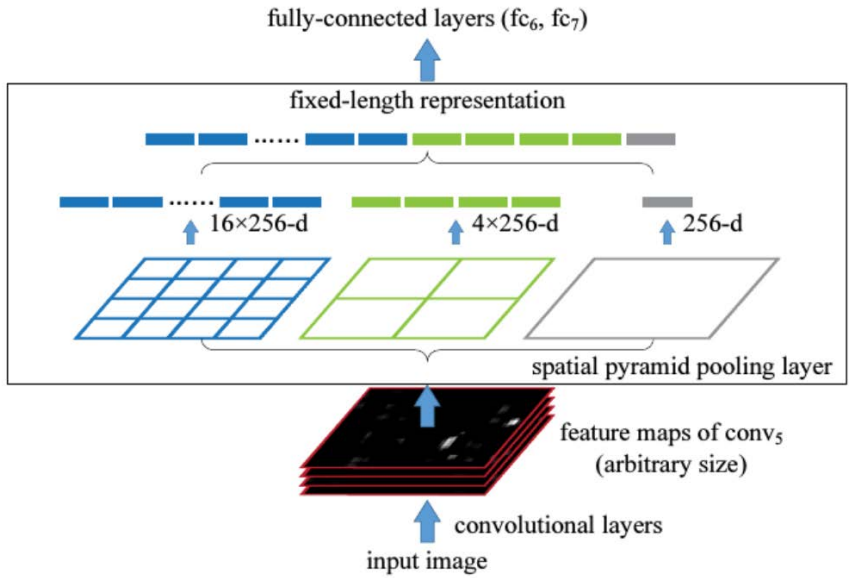
RANet: Region Attention Network for Semantic Segmentation

Lei Liu

Motivation

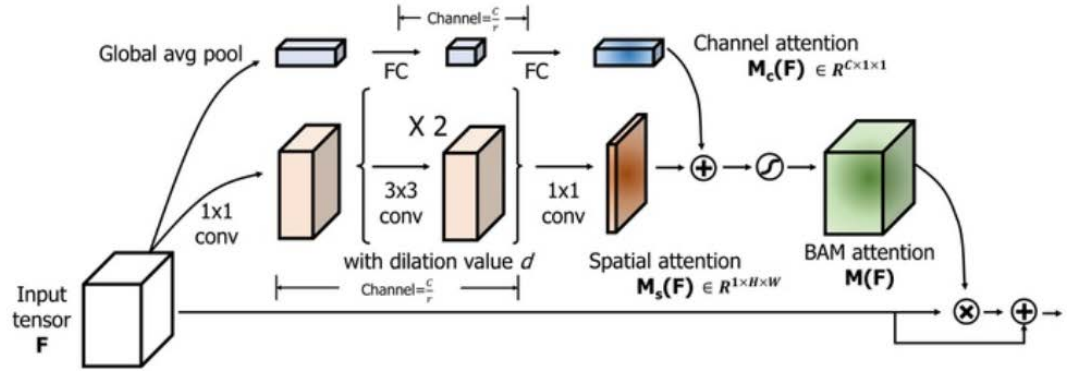
图像分割算法主流模型：

1. 空间金字塔池化 (spatial pyramid pooling)



SPP可以忽略输入尺寸并且产生固定长度的输出.

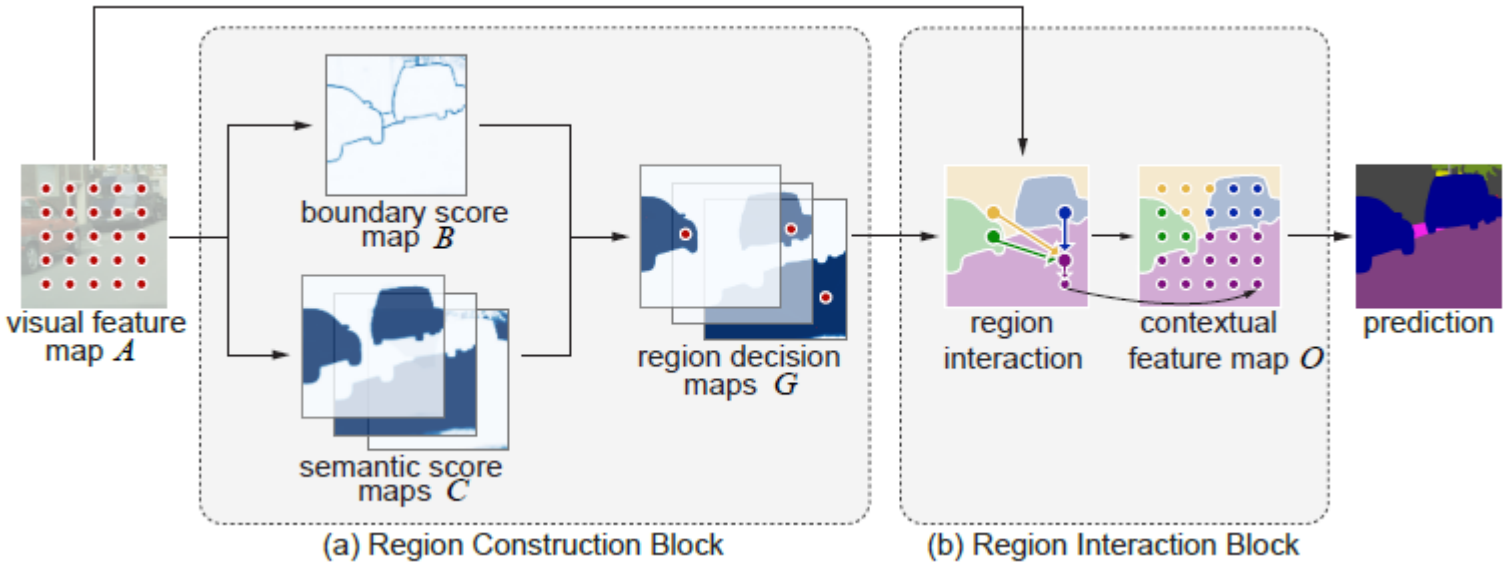
2. 注意力机制attention mechanism



可学习的权值.

已有方法主要关注像素级别的空间或者类别关系的建模，忽略了目标级别的上下文信息。

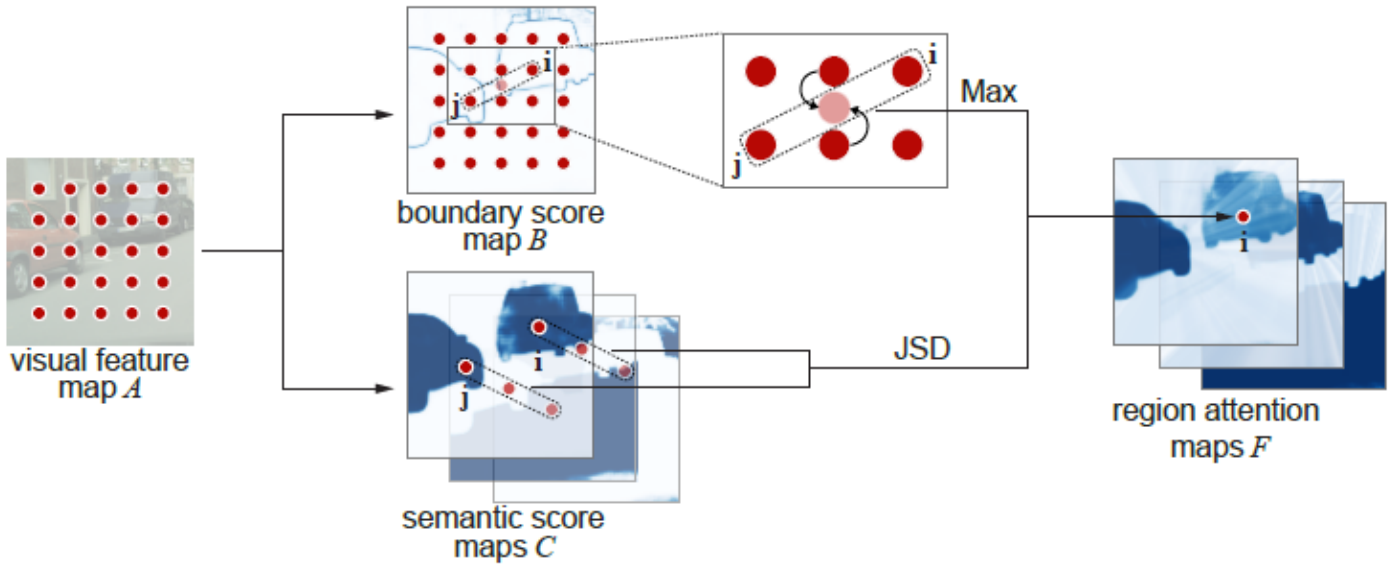
Framework



1. RCB提取目标边界信息(boundary score map)和语义区域信息(semantic score maps)，将特征图的不同像素点划分成不同目标区域；

2. RIB提取不同目标区域的代表性像素点，将不同区域进行信息交互。

Region Construction Block(RCB)



1. 寻找 i, j 像素点之间的边界　$D_{i,j} = \max(B_{i \leftrightarrow j}),$　B是 由i, j 确定的像素点集合

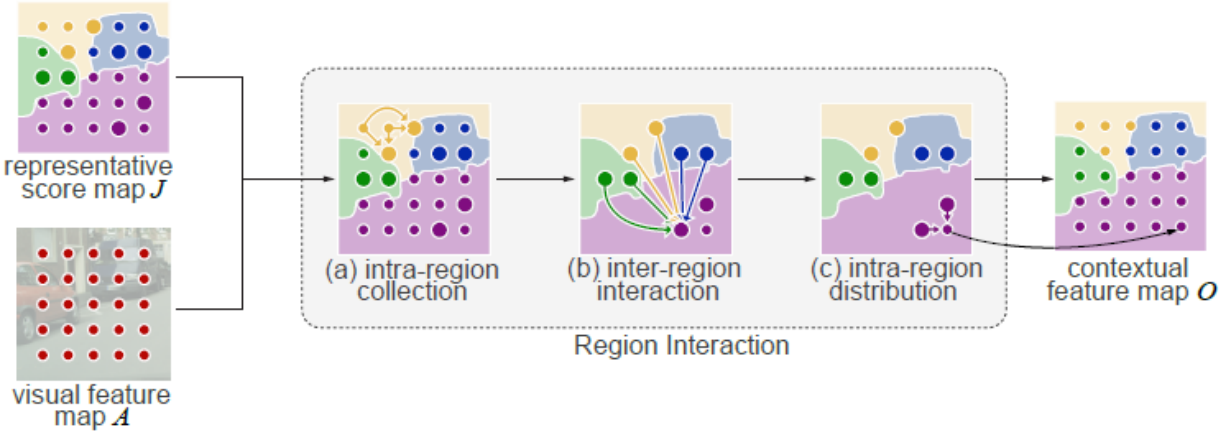
2. i, j 两点之间的语义相似度　$E_{i,j} = \sum_{n=1}^{N} \frac{C_{i,n} \log C_{i,n} + C_{j,n} \log C_{j,n}}{2U_{i,j,n}}, \quad U_{i,j,n} = \frac{C_{i,n} + C_{j,n}}{2}.$　（JS散度）

C是分类概率向量

3. 注意力分数图　$F_{i,j} = (1 - D_{i,j})(1 - E_{i,j}),$　距离越近，语义越相似的的像素点分数越大

4. 注意力决策图生成　$G_{i,j} = \frac{1}{2} sgn(\frac{F^g_{i,j} + F^g_{j,i}}{2}) + \frac{1}{2}, \quad F^g = W^g \otimes F,$　W是卷积核

Region Interaction Block(RIB)



1. 代表性计算 $J_i = \frac{1}{|R_q|} \sum_{j \in R_q} F^g_{i,j},$ R是目标区域集合

   R表示第i个像素点与第q个区域的关联性

2. 选择某个区域中代表性最高的像素点集合 $\{p_{q,k} | p_{q,k} \in \phi(R_q), k = 1, ..., K\}.$

3. 区域内局部上下文代表性计算

$$A^l_{q,k} = J(p_{q,k})A(p_{q,k}) + \sum_{i \in R_q \setminus \phi(R_q)} W^l_{q,k,i}(J_i A_i), \quad W^l_{q,k,i} = \frac{\exp\left(A(p_{q,k})(J_i A_i)^\top\right)}{\sum_{j \in R_q \setminus \phi(R_q)} \exp\left(A(p_{q,k})(J_j A_j)^\top\right)},$$

   A是原始特征图，J是代表性分数图

4. 全局上下文代表性计算 $A^g_{q,k} = J(p_{q,k})A^l_{q,k} + \sum_{s=1, s \neq q, i \in \phi(R_s)}^{Q} W^g_{q,s,k,i}(J(p_{s,i})A^l_{s,i}),$

$$W^g_{q,s,k,i} = \frac{\exp\left(A^l_{q,k}(J(p_{s,i})A^l_{s,i})^\top\right)}{\sum_{p_{s,j} \in \phi(R_s)} \exp\left(A^l_{q,k}(J(p_{s,j})A^l_{s,j})^\top\right)}.$$

5. 将全局上下文代表性传播到区域内代表性像素点

$$O_i = J_i A_i + \sum_{p_{q,k} \in \phi(R_q)} W^d_{q,k,i}(J(p_{q,k})A^g_{q,k}), \quad W^d_{q,k,i} = \frac{\exp\left(A_i(J(p_{q,k})A^g_{q,k})^\top\right)}{\sum_{p_{q,j} \in \phi(R_q)} \exp\left(A_i(J(p_{q,j})A^g_{q,j})^\top\right)}.$$

Experimental Results

method		backbone	annotation	
			w/o coarse	w/ coarse
SPP	PSPNet[7]	ResNet-101	80.1	81.2
	Deeplabv3+[36]	Xception-71	81.0	81.9
	DPC[33]	Xception-71	82.7	-
attention	Asymmetric NL[11]	ResNet-101	81.3	-
	CCNet[10]	ResNet-101	81.4	-
	OCRNet[34]	HRNetV2-W48+ASPP	83.2	83.7
	RANet	ResNet-101	82.4	83.0
		HRNetV2-W48+ASPP	**83.4**	**84.0**

Table 5: Comparisons with other state-of-the-art methods on the Cityscapes test set.

Experimental Results

	PASCAL Context		COCO-Stuff	
	method	mIoU	method	mIoU
SPP	SVCNet[24]	53.2	DSSPN[37]	38.9
	HRNet[35]	54.0	SVCNet[24]	39.6
attention	BFP[12]	53.6	CCNet[10]	39.8
	CFNet[38]	54.0	EMANet[39]	39.9
	ACNet[40]	54.1	ACNet[40]	40.1
	RANet	**54.9**	RANet	**40.7**

Table 6: The segmentation accuracies on the PASCAL Context and COCO-Stuff validation sets.

Experimental Results

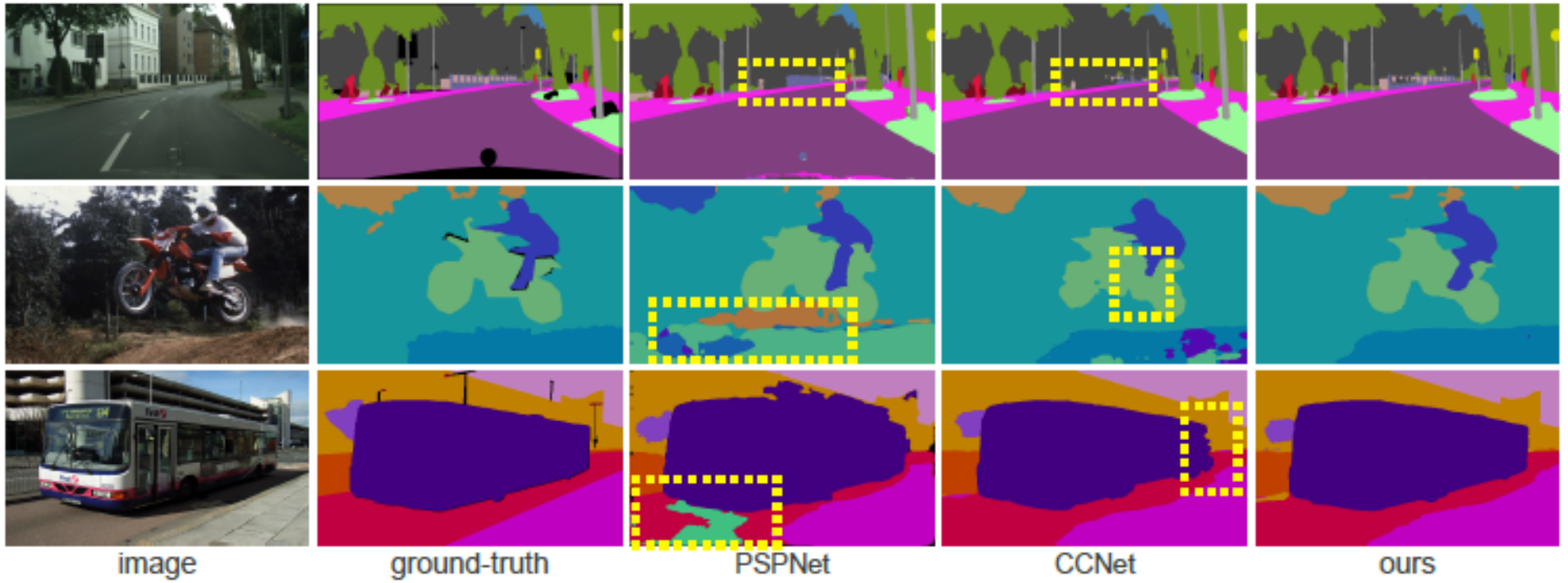


Figure 6: Segmentation results on the Cityscapes, PASCAL Context and COCO-Stuff validation sets.

Learning to Recommend Frame for Interactive Video Object Segmentation in the Wild

Lei Liu

Motivation

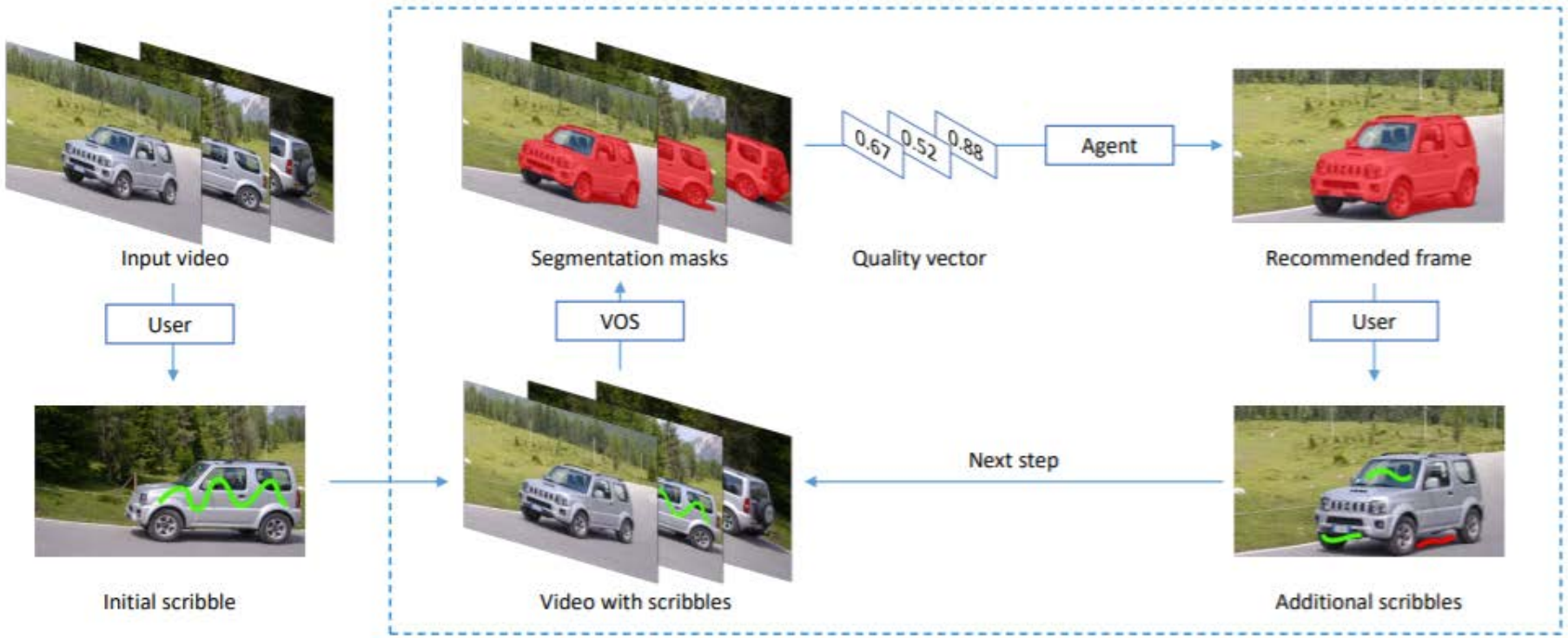
1. 视频目标分割数据集标注代价昂贵

2. 传统视频目标分割人机交互系统依赖于图像分割质量评价

3. 传统视频目标分割人机交互系统依赖于分隔质量最差的样本

如何衡量视频帧的重要性？

Framework



1. 初始状态：视频帧及对应的初始标注

2. VOS算法进行视频目标分割以及分割质量估计

3. 将分割质量作为输入，利用MDP进行视频帧推荐，用户进行额外标注

Framework

State. $s_t = \text{CONCAT}(q_t, h_t),$

$q_t \in [0,1]^N$ N个视频帧的分割质量分数

$h_t \in \{0,1,...,T\}^N$ N个视频帧的被推荐次数

Action. State为输入，LSTM为agent，输出为推荐的视频帧ID

Reward. 衡量推荐的帧对性能提升的重要性

1. 以随机推荐作为参考：运行30次来估计随机推荐的性能提升的方差和均值

$r_t^{\text{goal}}(P) = \frac{P - \hat{\mu}}{\hat{\sigma}}.$ P为最终的性能，若性能高于随机推荐，reward为正，否则为负

2. 远远高于随机推荐 $r_t^{\text{goal}}(P) = \frac{P - (\hat{\mu} + \hat{\sigma})}{\hat{\sigma}}.$

Framework

Reward. 衡量推荐的帧对性能提升的重要性

3. 考虑视频的运动特性和视角变化特性，推荐应该考虑具有多样性的帧

$$r_t^{\text{aux}} = \begin{cases} 1, & a_t = \arg\min h_t, \\ -1, & \text{otherwise.} \end{cases}$$

4. 结合以上两种策略

$$Q_t^* = \begin{cases} \delta \cdot r_t^{\text{goal}}, & t = T, \\ \delta \cdot r_t^{\text{aux}} + \gamma \cdot Q^T(s_{t+1}, a_{t+1}), & t < T, \end{cases}$$

5. Policy Network $a_{t+1} = \text{argmax}_a Q^P(s_{t+1}, a)$

   Value of the action $Q^T(s_{t+1}, a_{t+1})$

6. update

$$\mathcal{L}_{\text{agent}} = \text{MSE}(Q_t, Q_t^*).$$

Framework

Setting	Strategy	DAVIS			YouTube-VOS		
		IPN [21]	MANet [19]	ATNet [12]	IPN [21]	MANet [19]	ATNet [12]
Oracle	Worst	48.02	70.85	73.68	**44.67**	66.03	74.89
	Ours	**48.25**	**71.11**	**74.01**	43.86	**66.90**	**75.37**
Wild	Random	47.52(4)	69.81(1)	72.99(3)	43.22(5)	64.97(8)	74.11(8)
	Linspace	46.97	70.10	72.93	42.75	64.75	73.47
	Worst	47.26	69.32	73.33	43.29	65.98	74.69
	Ours	**47.99**	**70.82**	**74.10**	**43.69**	**66.85**	**75.33**

Table 1. Quantitative results (AUC) of the interactive VOS on DAVIS and YouTube-VOS dataset.