OS1A-L1

StableSeg: StableDiffusionによるゼロショット領域分割

本部勇真, 山口廉斗, 柳井啓司 (電通大)

はじめに

領域分割タスクでは、モデルの学習や
アノテーションデータは高コスト

大規模画像生成モデルStable Diffusionを活用して

学習や外部データを必要としない領域分割を実現

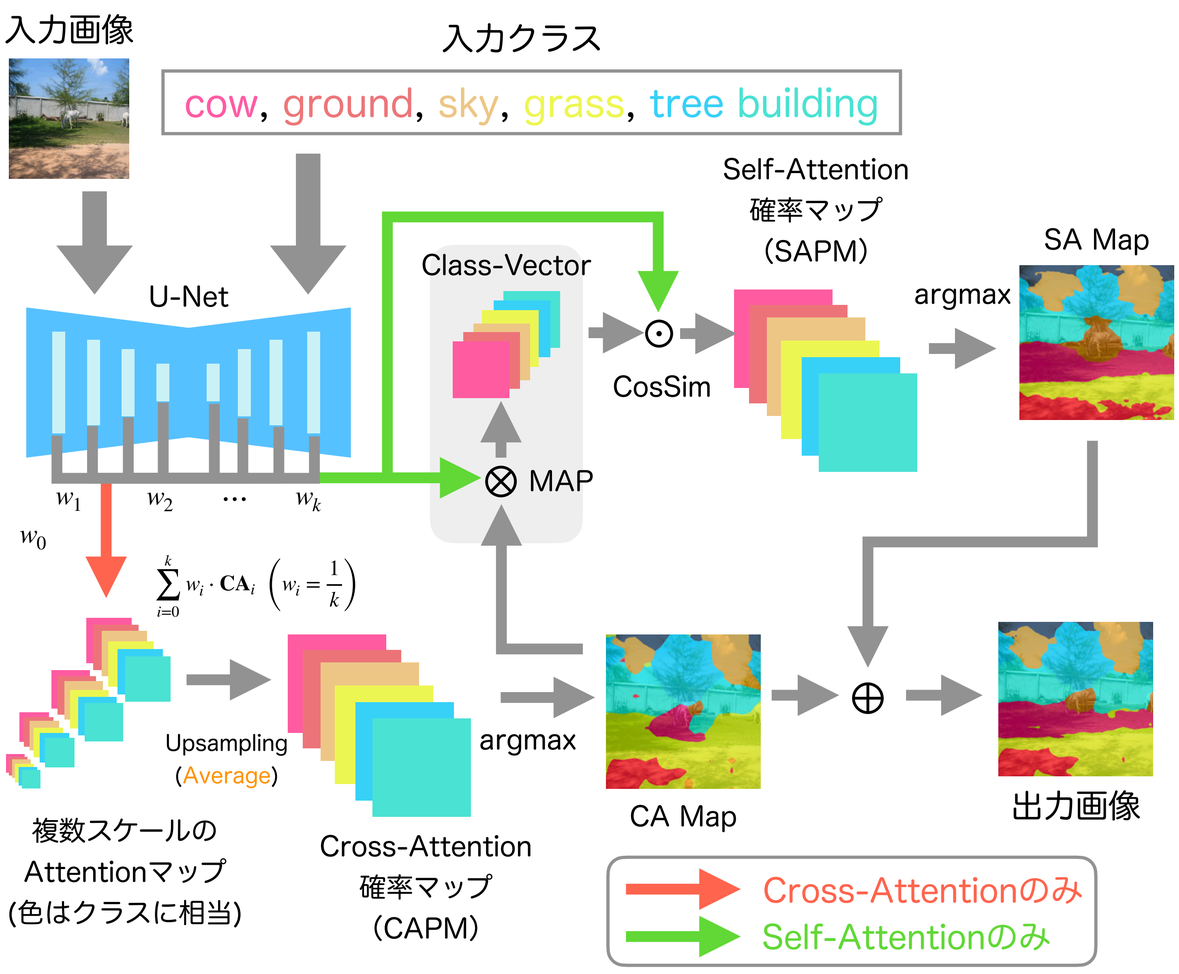
手法

StableSeg

Stable Diffusionに含まれるAttentionに注目

Cross/Self-Attentionを用いて領域分割

タイムステップは $t = 1$ で固定



StableSeg+

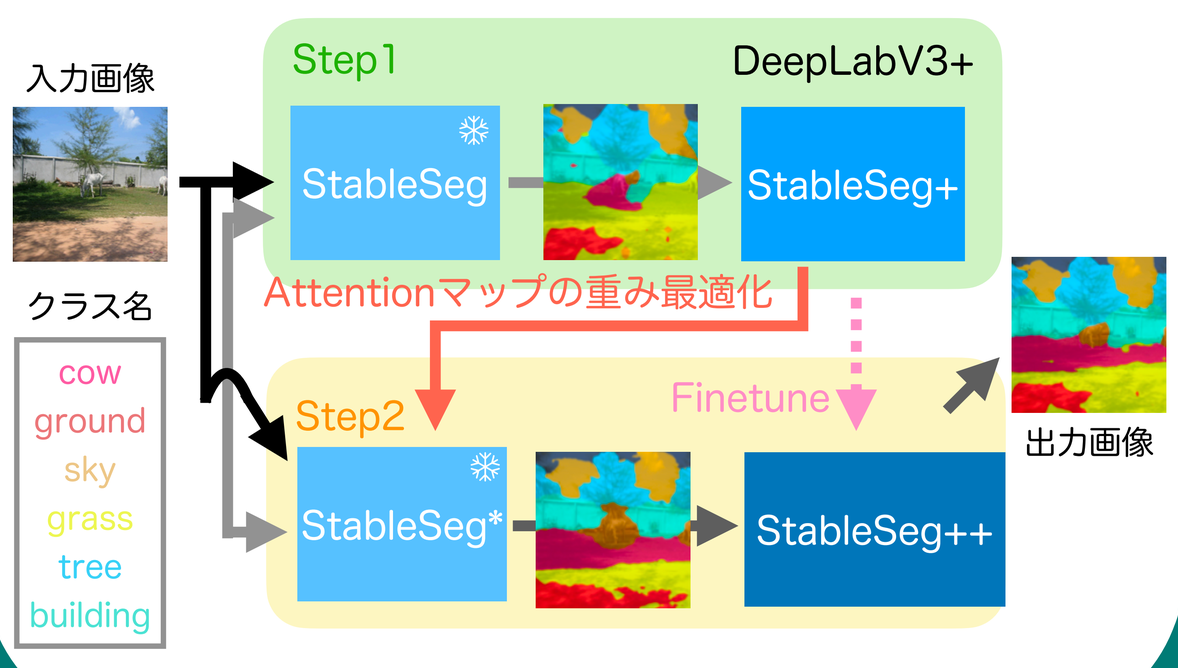
StableSegで生成した擬似マスクでDeepLabV3+を学習

StableSeg++

StableSeg+で生成した擬似マスクを用いて、
各層のCross-Attentionマップの最適な重み w_i を学習

StableSegで平均していた部分を修正して精度向上させる

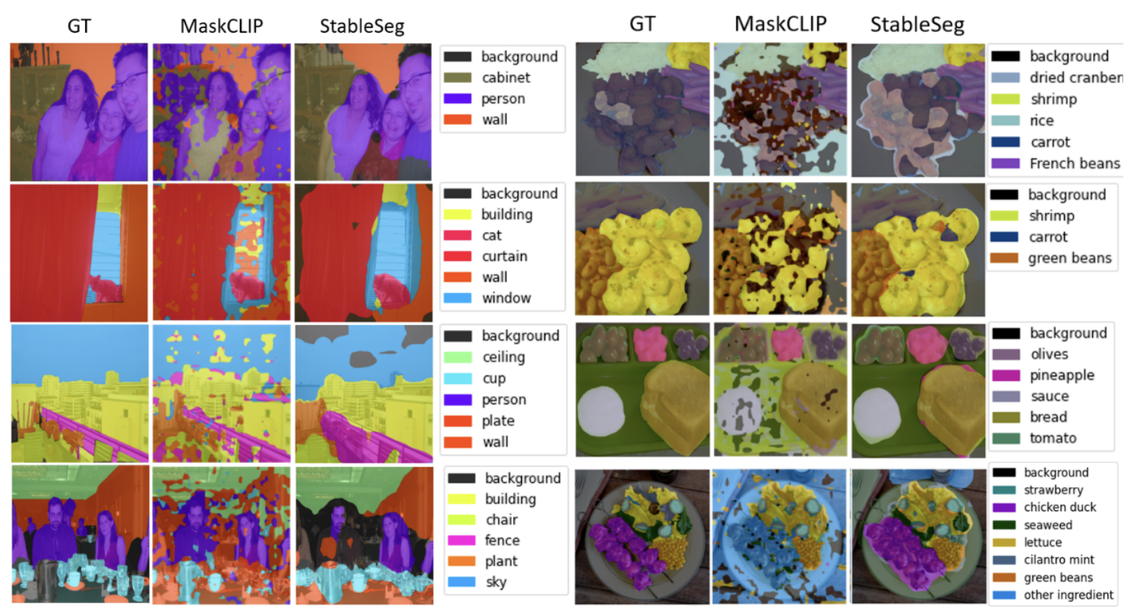
➡ その生成結果を擬似マスクとし更にStableSeg+を学習



実験

各データセットでの定量評価 (mIoU)

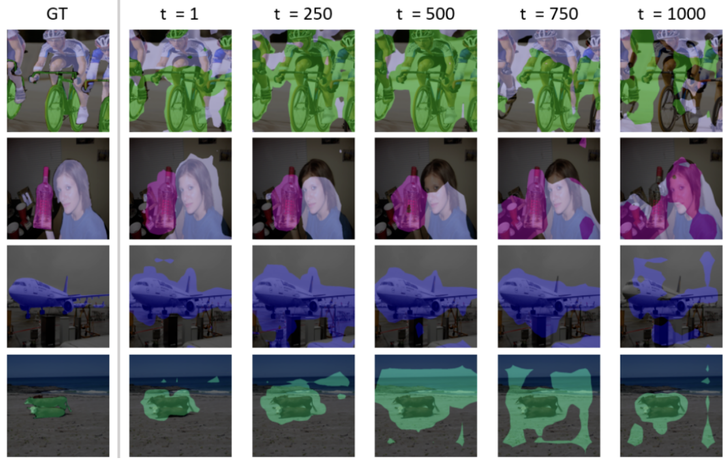
	PAS-20	PC-59	A-150	City	FoodPix	FoodSeg
MaskCLIP	44.7	**37.9**	**26.0**	**21.6**	33.2	37.0
StableSeg(Ours)	**50.3**	36.2	23.6	15.1	**63.3**	**49.1**



従来手法との比較例 (左: PC-59、右: FoodSeg)

SARの違いによる各データセットでの定量評価 (mIoU)

	PAS-20	PC-59	A-150	City	FoodPix	FoodSeg
w/ SAR	**50.3**	**36.2**	**23.6**	**15.1**	63.3	49.1
w/o SAR	47.2	31.0	19.4	12.7	53.8	39.3
SAMap	47.1	33.6	22.5	13.3	**65.4**	**50.0**



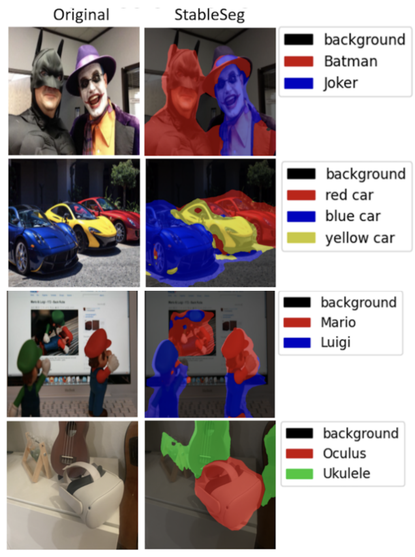
時間埋め込み t を変化させた時の領域分割の例

時間埋め込みtの実験

	mIoU
1	**49.9**
250	46.4
500	42.9
750	38.8
1000	31.3

MaskCLIP+との比較

	PAS-20	PC-59
MaskCLIP	44.7	37.9
MaskCLIP+	53.4	**40.5**
StableSeg	51.4	33.9
StableSeg+	55.6	34.6
StableSeg++	**59.1**	36.6



MaskCLIP+との比較

まとめと今後の課題

- Stable DiffusionのAttentionを用いて学習なしで
  低コストに領域分割を行うStableSegを提案
- 弱教師あり学習を用いて、注釈データを必要とせずに
  StableSegの精度を向上させたStableSeg+とStableSeg++の提案
- Referring Image Segmentation での有効性も検証する
- Instance Segmentation への拡張を検討する