CH-007

弱教師あり領域分割のための高品質な疑似ラベルの選択
Selecting High-quality Pseudo-labels for Weakly Supervised Semantic Segmentation

藤森 和泉[1)]　大野 将樹[2)]　獅々堀 正幹[2)]
Izumi Fujimori　Masaki Oono　Masami Shishibori

1　はじめに

画像レベルのラベルを用いる弱教師あり領域分割（Weakly-supervised Semantic Segmentation: WSSS）は，アノテーションコストの低さから注目を集めている．一般的な WSSS は，画像分類モデルから得られるクラス活性化マップ（Class Activation Map: CAM）をもとに作成したピクセルレベルの疑似ラベルを用いて，領域分割モデルを学習する．しかし，疑似ラベルには，誤った領域へのクラスの割り当てといった問題が存在する．領域分割モデルは疑似ラベルに含まれるノイズを学習することで性能が低下する [1]．この課題に対処するために，ノイズに頑健な領域分割モデルの研究が進められている．一方で，近年の疑似ラベルの作成に関する研究の発展により，依然としてノイズの存在する疑似ラベルはみられるが，真値に極めて近い疑似ラベルも数多く作成することが可能となっている．図 1 に，pascal visual object classes（PASCAL VOC 2012） 訓練データ [2] における各疑似ラベルの Intersection over Union（IoU）を示す．縦軸が疑似ラベルの数であり，横軸が IoU（%）である．この疑似ラベルは，WSSS 手法の background-aware activation map optimization（BAO）[3] に基づいて作成され，segment anything model（SAM）[4] を用いた後処理 [5] が適用されている．図 1 から，数多くの疑似ラベルの IoU が 99%以上を達成していることがわかる．

さらに，半教師あり領域分割（Semi-supervised Semantic Segmentation: SSSS）は，WSSS と同様にアノテーションコストの削減が可能な領域分割手法であり，限られた量の正解データであっても，データの正確性が，領域分割精度を著しく向上させる．表 1 に，WSSS 手法 [3]，[6] と SSSS 手法 [7]，[8] の PASCAL VOC 2012 検証データにおける領域分割精度を示す．表 1 の Label type は，学習に用いるラベルの種類を示している．また，表 1 の 10,582 は WSSS 手法が用いる疑似ラベルの数を表し，366 は SSSS 手法が学習に用いる真値の数，10,216 はラベルなし画像の数を表す．評価指標として，mean IoU（mIoU）を用いている．一般に，SSSS は WSSS よりも高い領域分割精度を達成する傾向にある．これは，WSSS が作成されたすべての疑似ラベルを使用するのに対し，SSSS は少数ながら正確性の高いラベルを利用することにより，学習がより安定的かつ効果的に行われるためだと考えられる．すなわち，SSSS の優れた性能は，ラベルの数量ではなく，その正確性およびノイズの少なさが領域分割精度の向上において重要であることを示唆している．

以上の点から，WSSS によって得られる疑似ラベルの中から真値に近い疑似ラベルを選択し，選択された疑似ラベルに対応する画像はラベル付き画像として扱い，残りの画像をラベルなし画像として SSSS の学習に利用することで，WSSS の枠組みの中で，より高精度な領域分割が可能になると推察される．このとき，重要なのは真値に近い疑似ラベルを選択することである．疑似ラベルの精度は，真値であるピクセルレベルのラベルから求められる．しかし，WSSS はピクセルレベルのラベルを直接参照することはできない．そこで，本論文では高品質な疑似ラベルを選択する手法を提案する．

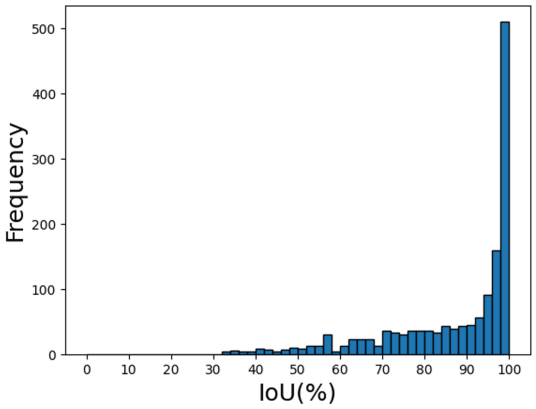

図 1: 疑似ラベルの精度の分布．

表 1: WSSS 及び，SSSS の領域分割精度の比較．

Method	Labal type	VOC-*val*
WSSS-BAO[3] JVCIR2025	Pseudo-labeled(10,582)	74.1
WSSS-MARS [6] ICCV2023	Pseudo-labeled(10,582)	77.7
SSSS-UniMatchV1[7] CVPR2023	Labeled(366), Unlabeled(10,216)	78.8
SSSS-UniMatchV2[8] PAMI2025	Labeled(366), Unlabeled(10,216)	88.9

本手法は同一画像を対象とした異なる疑似ラベル間で IoU を計算し，IoU が高いものを高品質な疑似ラベルとする．具体的には，図 2 上段に示すように疑似ラベル A と，A に対して後処理 [5] を適用した疑似ラベル $A+$ との間で IoU を計算する．このとき，IoU が高い疑似ラベルは後処理の影響が小さい．すなわち，後処理前後で疑似ラベルの形状が大きく変化していないことを意味する．後処理 [5] は誤差のある疑似ラベルを補正し，より正確な形状へと近づけることを目的としている．そのため，後処理による修正がほとんど行われない場合，元の疑似ラベルはすでに真値に近い形状を持っていたと考えられる．したがって，後処理前後の疑似ラベル間の IoU が高い疑似ラベルは，高い精度を有するラベルであると判断できる．このように，疑似ラベル間の IoU が高いものを高品質な疑似ラベルとみなし，ラベル付き画像として扱う．一方，疑似ラベル間の IoU が低い場合はラベルなし画像として，SSSS の学習に利用する．

しかし，後処理前後の疑似ラベルの IoU により，疑似ラベルの品質を評価する手法は，評価の信頼性に課題をもたらす可能性がある．図 2 下段に示すように，疑似ラベル A の精度が低い場合，後処理が効果的に機能せず，後処理後の疑似ラベル $A+$ の形状の修正が不十分なまま

1）　徳島大学大学院 創成科学研究科 理工学専攻知能情報システムコース
2）　徳島大学大学院 社会産業理工学研究部

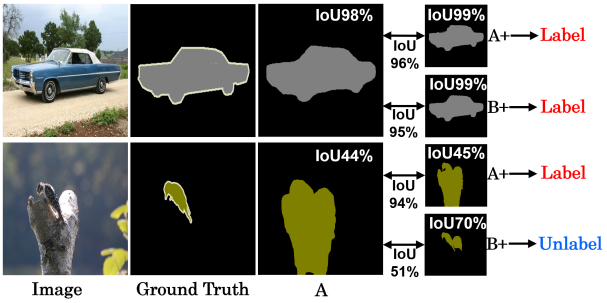

図 2: IoU に基づく，疑似ラベルの選択．

となることがある．この問題は WSSS 手法 [9] で報告されている．その結果，後処理で A の形状が変化せず，A と $A+$ 間の IoU が高くなる傾向があり，精度の低い疑似ラベルが誤って高品質な疑似ラベルであると判断される．この問題を回避するために，A とは異なる WSSS 手法によって得られた別の疑似ラベルや，その疑似ラベルに対して後処理を適用した疑似ラベルと A との間の IoU を基準として評価を行うことを提案する．図 2 の $B+$ は，A とは異なる WSSS 手法から得られた疑似ラベルに対して，後処理 [5] を適用したものである．異なる WSSS 手法の疑似ラベル間で IoU を計算することで，単一の WSSS 手法内での後処理前後の疑似ラベル間の IoU に依存するのではなく，より信頼性の高い評価と選択を行うことができる．A と $A+$ で IoU を比較した場合，図 2 上段の疑似ラベルは適切にラベル付き画像として選択されるが，図 2 下段の疑似ラベルは誤って選択される．一方，A と $B+$ の IoU を比較した場合，図 2 下段の A と $B+$ の疑似ラベルの形状は異なっているため，IoU が低くなる．そのため，精度が低い疑似ラベルに対応する画像をラベルなし画像とすることができる．

本手法の有効性を示すために，PASCAL VOC 2012 データセットを用いた実験を行った．その結果，従来の WSSS に比べて領域分割精度が向上することが確認された．

2 関連研究

疑似ラベル間の IoU に基づき，疑似ラベルの品質を評価する手法は文献 [10]，[11] に着想を得ている．文献 [10] は，diffusion model[12] を用いて作成した合成マスクと SAM[4] に合成画像とクラス名をテキストとして，プロンプトへ入力して作成されたマスク間の IoU を比較し，合成マスクの品質を評価する．文献 [10] は，品質の低い合成マスクに対応する合成画像は学習に用いない．一方で本手法は，低品質の疑似ラベルに対応する画像はラベルなし画像として，SSSS の学習に利用する．文献 [11] は，SSSS における領域分割モデルのラベルなし画像に対する領域分割結果と SAM を用いて作成した複数のマスク（SAM マスク）間の IoU に基づき，SAM マスクを選択する．SAM マスクはラベルなし画像の予測結果に対する正解データとして用いられる．しかし，領域分割モデルの予測精度が低い場合には，誤った SAM マスクが選択される可能性がある．さらに，作成された複数の SAM マスク全てに誤りが含まれている場合であっても，いずれかのマスクが正解データとして採用される．本手法は，疑似ラベル間の IoU が低い場合，その疑似ラベルに対応する画像はラベルなし画像として用いる．そのため，精度の低い疑似ラベルを排除できる．また，本手法は SAM に基づく後処理 [5] を採用しているが，文献 [10]，[11] のように SAM に依存するものではなく，他の WSSS 手法から得られた疑似ラベルを用いて，品質を評価することも可能である．重要となるのは，異なる WSSS 手法に由来する疑似ラベルを柔軟に活用できる点にあり，これにより特定の後処理や手法に縛られることなく，疑似ラベルの品質の評価が行える．

Algorithm 1 Pseudo-label selection

Input: Pseudo-labels $P^{(\mathrm{m1})}$, $P^{(\mathrm{m2+sam})}$, and the required number of samples per class R.
Output: Selected pseudo-labels HQ and the number of selected samples per class SR.

Initialization: $HQ \leftarrow \varnothing$, $SR \leftarrow \{0, 0, \ldots, 0\}$, IoU_List $\leftarrow \varnothing$
for each $i = 1$ to N **do**
  $\mathrm{IoU}_i = \frac{|p_i^{(\mathrm{m1})} \cap p_i^{(\mathrm{m2+sam})}|}{|p_i^{(\mathrm{m1})} \cup p_i^{(\mathrm{m2+sam})}|}$
  IoU_List $\leftarrow$ IoU_List $\cup$ $(p_i^{(\mathrm{m2+sam})}, \mathrm{IoU}_i)$
end for
Sorted_Data $\leftarrow$ Sort(IoU_List, by IoU in descending order)

for each (D_i, IoU_i) in Sorted_Data **do**
  $L_i = \{l_1, l_2, \ldots, l_S\}$
  for each $l_s \in L_i$ **do**
    if $SR[l_s] < R[l_s]$ **then**
      $SR[l_s] \leftarrow SR[l_s] + 1$
      $HQ \leftarrow HQ \cup D_i$
    end if
  end for
  if all($SR[label] \geq R[label]$ for all $label \in R$) **then**
    break
  end if
end for

3 提案手法

疑似ラベルの作成に関する研究の発展により，WSSS は真値に近い疑似ラベルを作成することが可能となった．さらに，SSSS は少量の正解データから高性能な領域分割を実現している．以上を動機として，本論文では，高品質な疑似ラベルを選択し，選択された疑似ラベルに対して，SSSS 手法を適用することを提案する．まとめると，提案手法は疑似ラベルの作成（section3.1），疑似ラベルの選択（section3.2），選択された疑似ラベルによる SSSS（section3.3）から構成される．

3.1 疑似ラベルの作成

一般的な WSSS 手法は，画像分類モデルから得られる CAM を用いて疑似ラベルを作成する．まず，データセットを $D = \{(I_i, Y_i)\}_{i=1}^{N}$ と定義する．ここで，I_i は i 番目の画像，Y_i は対応する画像レベルのラベル，N は画像枚数を表す．また，各画像は $I_i \in \mathbb{R}^{3\times H\times W}$，画像レベルのラベルは $Y_i \in \mathbb{R}^{(S-1)\times 1}$ であり，3，H，W は画像のチャネル数，高さ，幅を表し，S はクラス数，$S-1$ は前景オブジェクトのクラス数を表す．encoder F に画像 I_i を入力して得られる特徴マップを $f_i \in \mathbb{R}^{fc\times fh\times fw}$ とする．fc，fh，fw は特徴マップのチャネル数，高さ，幅を表す．f_i を線形分類層に入力して得られた値を画像分類モデルの予測，画像レベルのラベル Y_i を正解データとして，F を学習する．線形分類層の特徴マップに対応する重みを $w_s \in \mathbb{R}^{fc\times 1}$ とする．前景オブジェクトの CAM を $M_i \in \mathbb{R}^{S-1\times fh\times fw}$ とする．このとき，各クラスの CAM を $M_{i,s} \in \mathbb{R}^{fh\times fw}$ とする．s は特定のクラスを表す．$M_{i,s}$ は以下の式で計算される．

$$M_{i,s} = w_s^{\mathsf{T}} f_i, \tag{1}$$

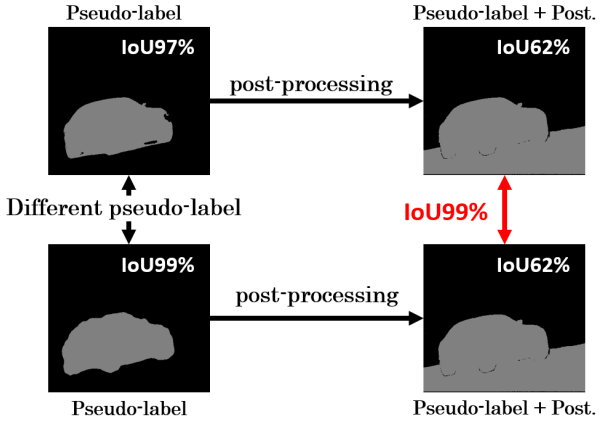

図 3: 異なる WSSS 手法の疑似ラベルに対して，同一の後処理 [5] を適用した疑似ラベル間の IoU の比較.

CAM M_i の各ピクセルにおいて，s 方向で最大値をとるピクセルのクラスを疑似ラベル p_i とする．このとき，背景閾値 τ をハイパーパラメータとして定義し，τ 未満の CAM の領域は背景クラス（$s = 0$）とする．p_i は以下の式で計算される．

$$p_i(x,y) = \begin{cases} \arg\max_{1\le s\le S-1} M_{i,s}(x,y), & \text{if } \max_{1\le s\le S-1} M_{i,s}(x,y) \ge \tau \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

x，y は各ピクセルの位置を表す．各画像 I_i に対応する各疑似ラベル p_i の集合を P とし，以下のように定義する．

$$P = \{p_1, p_2, \dots, p_N\}. \tag{3}$$

3.2 疑似ラベルの選択

本手法は疑似ラベル間の IoU を基に疑似ラベルの品質を評価する．具体的には，疑似ラベルと，その疑似ラベルに対して後処理 [5] を適用した疑似ラベルとの間で IoU を計算する．このとき，IoU が高い疑似ラベルは，後処理前後で疑似ラベルの形状が大きく変化していないことを意味する．後処理 [5] は，誤差のある疑似ラベルを補正し，より正確な形状に近づけることを目的としているため，後処理による修正がほとんど行われない場合，元の疑似ラベルはすでに真値に近い形状を持っていたと考えられる．しかし，疑似ラベルの精度が低い場合には後処理が機能せず，修正が不十分なままとなることがある [9]．このような場合，後処理前後の疑似ラベルの IoU が高くなる傾向がみられる．この問題に対処するため，本手法では異なる WSSS 手法によって作成された疑似ラベル間で IoU を計算する．異なる WSSS 手法による疑似ラベル間で IoU を評価することにより，後処理前後の疑似ラベル間の IoU に依存することなく，より信頼性の高い評価と選択が可能となる．異なる WSSS 手法により得られた疑似ラベルの集合をそれぞれ $P^{(\mathrm{m1})}$，$P^{(\mathrm{m2})}$ とする．このとき，$P^{(\mathrm{m2})}$ に後処理 [5] を適用した疑似ラベルを $P^{(\mathrm{m2+sam})}$ とする．$P^{(\mathrm{m1})}$，$P^{(\mathrm{m2+sam})}$ を以下のように定義する．

$$P^{(\mathrm{m1})} = \{p_1^{(\mathrm{m1})}, p_2^{(\mathrm{m1})}, \dots, p_N^{(\mathrm{m1})}\}. \tag{4}$$

$$P^{(\mathrm{m2+sam})} = \{p_1^{(\mathrm{m2+sam})}, p_2^{(\mathrm{m2+sam})}, \dots, p_N^{(\mathrm{m2+sam})}\}, \tag{5}$$

$P^{(\mathrm{m1})}$ と $P^{(\mathrm{m2+sam})}$ の間の IoU を比較し，IoU が高いものを高品質な疑似ラベルとして選択する．

しかし，IoU のみに基づいて疑似ラベルを選択すると，特定のクラスに偏った疑似ラベルが選ばれる．この偏りを抑制するため，本手法では，データセットが有するクラス分布を維持しつつ，IoU の平均が最大となるように疑似ラベルを選択する．PASCAL VOC 2012 訓練データにおけるクラス分布を表 2 に示す．また，クラス分布を考慮した疑似ラベルの選択手順を Algorithm 1 に示す．Algorithm 1 は，2 つの疑似ラベルの集合 $P^{(\mathrm{m1})}$，$P^{(\mathrm{m2+sam})}$ 及び，クラス分布 R を入力とし，選択された疑似ラベルの集合 HQ と，選択された疑似ラベルのクラスごとのサンプリング数 SR を出力する．Algorithm 1 において，SR は各クラスに対して選択された疑似ラベルの累積数を格納するベクトルであり，事前に定められたクラス分布 R を満たしているかを判定するために用いられる．Algorithm 1 は，まず，$P^{(\mathrm{m1})}$ と $P^{(\mathrm{m2+sam})}$ の疑似ラベル間で IoU を計算し，IoU が高い順に疑似ラベルをソートする．次に，IoU の高い順に疑似ラベルを 1 つずつ評価し，その疑似ラベルが含む各クラスに対して，SR が R を満たしていない場合に，SR を 1 つ増加させ，疑似ラベルを HQ に追加する．この処理は，各クラスに対して独立に行われるため，ある疑似ラベルがすでに目標数に達したクラスを含んでいても，未達成のクラスが含まれていればその疑似ラベルは選択される．つまり，クラス $C1$ はクラス分布を満たしているが，クラス $C2$ はクラス分布を満たしていないとき，クラス $C1$ と $C2$ の両方を含む疑似ラベルは $C2$ のために選択される．疑似ラベル間の IoU とクラス分布に基づいて選択された疑似ラベルの集合 HQ は以下のように定義される．

$$HQ = \left\{ p_i^{(\mathrm{m2+sam})} \mid i \in Q^l \right\}. \tag{6}$$

Q^l は，選択された疑似ラベルのインデックスを示す．これにより，クラス分布の偏りを防ぎつつ，高品質な疑似ラベルを選択することが可能となる．

また，本手法では，異なる WSSS 手法から得られた疑似ラベルのうち，一方の疑似ラベルのみに後処理 [5] を適用している．これは，それぞれの疑似ラベルに同じ後処理 [5] を適用すると，図 3 に示すように類似した疑似ラベルが作成される傾向がみられたためである．後処理により精度が向上する場合は問題にならないが，後処理後に著しく疑似ラベルの精度が低下する場合においても疑似ラベル間の IoU が高くなるため，誤って精度の低い疑似ラベルが選択される．この問題を回避するために，一方の疑似ラベルのみに後処理を適用している．疑似ラベルの選択における適切な疑似ラベルの組み合わせについては，section4.4 に記す．

3.3 選択された疑似ラベルによる SSSS

選択された疑似ラベルに対応する画像はラベル付き画像として扱い，残りの画像はラベルなし画像として SSSS の学習に利用する．領域分割モデルの学習においては，SSSS を用いるが，正解データとなるピクセルレベルのラベルは WSSS から得られる疑似ラベルを用いるため，WSSS の枠組みの中で領域分割モデルの学習が可能である．SSSS の目的はラベル付き画像と，ラベルなし画像から領域分割モデルを学習することである．このとき，ラベル付きデータセットを $D^l = \{(I_i^l, p_i^{(\mathrm{m2+sam})}) \mid i \in Q^l\}$ とし，ラベルなしデータセットを $D^u = \{I_i^u \mid i \in Q^u\}$ とする．Q^u は，ラベルなし画像のインデックスを示す．

表 2: PASCAL VOC 2012 訓練データにおける，各クラスのサンプリング数.

Labeled images	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv
1/16(92)	5	4	2	6	7	5	9	5	10	4	7	15	4	10	32	8	5	5	2	2
1/8(183)	7	15	14	6	14	15	13	23	17	6	7	17	11	11	52	10	8	10	8	9
1/4(366)	17	13	26	19	23	21	28	39	34	15	26	26	16	23	117	19	12	18	22	26
1/2(732)	46	32	64	40	39	42	68	63	72	31	39	66	35	29	235	34	30	50	38	43
Full(1,464)	88	65	105	78	87	78	128	131	148	64	82	121	68	81	442	82	63	93	83	83

表 3: PASCAL VOC 2012 訓練データにおける，疑似ラベル間の IoU の比較により，選択された疑似ラベルの各クラスのサンプリング数.

Labeled images	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv
1/16 (92) → 147/10,582	5	4	2	6	8	5	9	6	10	4	7	15	4	10	43	8	5	5	2	2
1/8 (183) → 273/10,582	7	15	14	6	15	15	13	25	17	6	7	17	11	11	69	10	8	10	8	9
1/4 (366) → 539/10,582	17	13	26	19	26	21	28	41	40	15	26	29	16	23	170	20	13	19	22	26
1/2 (732) → 1,082/10,582	46	32	64	40	45	42	69	72	77	31	39	80	35	30	334	35	33	54	38	43
Full (1,464) → 2,095/10,582	88	67	110	78	102	78	141	150	159	65	82	175	73	86	658	83	67	109	83	83

表 4: 後処理前の疑似ラベルと後処理後の疑似ラベルの精度の比較.

Method	Post.	Seed	Mask
BAO[3] JVCIR2025	CRF	75.5	77.4
MARS[6] ICCV2023	CRF	-	81.8
BAO[3] + SAM[5]	SAM	-	81.3*
MARS[6] + SAM[5]	SAM	-	83.6*

表 5: 領域分割精度の比較.

Method	Sup.	VOC-*val*	VOC-*test*
BAO [3] JVCIR2025	I	74.1	74.6
MoRe [13] AAAI2025	I	76.4	75.0
MARS [6] ICCV2023	I	77.7	77.2
S2C [9] CVPR2024	I+A	78.2	77.5
VPL [14] AAAI2025	I+C	79.3	79.0
FMA-WSSS [15] WACV2024	I+C+A	82.6	81.6
SemPLeS[16] WACV2025	I+C+A	83.4	82.9
Ours	I+A	**84.4**	**83.3**

D^l，D^u を入力として，領域分割モデルを学習する．本手法では SSSS 手法として，UniMatchV2[8] を用いる．

4 実験

4.1 データセット及び評価方法

本手法の有効性を評価するために，PASCAL VOC 2012 データセット [2] を用いて実験を行う．PASCAL VOC 2012 データセットは背景クラスを含む 21 クラスが存在する．訓練データが 1,464 枚，検証データが 1,449 枚，テストデータが 1,456 枚存在するが，semantic boundary dataset[17] による 10,582 枚に拡張された訓練データを用いるのが一般的である．SSSS では，訓練データ 1,464 枚の中から，それぞれ 1/16（92 枚），1/8（183 枚），1/4（366 枚），1/2（732 枚），Full（1,464 枚）をラベル付き画像として用い，それ以外の画像をラベルなし画像としたときの結果を報告している．つまり，1,464 枚のラベル付き画像を用いる場合は，残りの 9,118 枚はラベルなし画像となる．Algorithm 1 で使用するクラス分布 R は，これらのラベル付き画像における各クラスのサンプリング数で構成されている．PASCAL VOC 2012 訓練データにおける，クラス分布を表 2 に示す．

表 6: 疑似ラベル間の IoU が 90%以上の疑似ラベルの mIoU（%）の比較.

Pseudo-label selection(IoU>90.0)	Pseudo-label pool	1,464 images	mIoU(%)
BAO, MARS+SAM	MARS+SAM	609	91.9
MARS, MARS+SAM	MARS+SAM	1,072	86.5
BAO+SAM, MARS+SAM	MARS+SAM	898	89.8
BAO, MARS	MARS	555	90.1

表 7: 選択された疑似ラベルを用いたときの領域分割精度の比較.

Method	Pseudo-label selection(Full(1,464))	Pseudo-label pool	Labeled images	mIoU(%)
UniMatchV2	-	MARS+SAM	Full(1,464)	83.2
	BAO, MARS+SAM	MARS+SAM	2,095	84.4
	MARS, MARS+SAM	MARS+SAM	2,021	82.5
	BAO+SAM, MARS+SAM	MARS+SAM	2,061	82.7
	BAO, MARS	MARS	2,068	82.5

評価指標として，mIoU を用いる．

4.2 実験の詳細

実験では，NVIDIA RTX A6000（VRAM 48GB）を使用した．WSSS 手法として，BAO[3] 及び，MARS[6] を採用し，これらを用いて疑似ラベルを作成する．疑似ラベルの後処理として SAM に基づく手法 [5] 及び，conditional random fields（CRF）[18] を用いる．実験に用いる疑似ラベルの精度を表 4 に示す．表 4 中の * は実験により再現した結果である．また，表 4 の Post. は，疑似ラベルに適用した後処理，Seed は後処理前の疑似ラベルの精度，Mask は後処理後の疑似ラベルの精度を示している．PASCAL VOC 2012 訓練データにおいて，BAO[3] の後処理前の疑似ラベルの精度は 75.5%であり，CRF[18] を適用した後の疑似ラベルの精度は 77.4%である．BAO の後処理前の疑似ラベルに対して，後処理 [5] を適用した場合，疑似ラベルの精度（BAO+SAM）は 81.3%となる．また，MARS によって作成された疑似ラベルの精度は 81.8%である．この疑似ラベルに対して，後処理 [5] を適用したときの疑似ラベル（MARS+SAM）の精度は 83.6%となる．BAO 及び，MARS+SAM の疑似ラベル間の IoU を基に高品質な疑似ラベルを選択し，データセットを構築する．このとき，高品質な疑似ラベルは Algorithm 1 に従い，MARS+SAM の疑似ラベルの中から選択される．SSSS 手法として，UniMatchV2[8] を採用した．UniMatchV2

表 8: PASCAL VOC 2012 検証データにおける，ラベル数ごとの領域分割精度の比較.

Method	Pseudo-label selection	Pseudo-label pool	1/16(92)	1/8(183)	1/4(366)	1/2(732)	Full(1,464)
Ours	-	MARS+SAM	81.2	82.7	82.1	82.1	83.2
Ours	BAO, MARS+SAM	MARS+SAM	82.7(147)	82.9(273)	82.6(539)	83.2(1,082)	84.4(2,095)

が用いる領域分割モデルの encoder，decoder はそれぞれ，DINOv2-B[19]，DPT[20] で構成されている．学習における更新回数は 80epoch とした．本手法においては，凍結（frozen）された DINOv2-B を用いる．その他のパラメータにおいては，UniMatchV2 に従う．領域分割結果は，UniMatchV2 における生徒モデルの推論結果を用いて評価する．

4.3 既存の WSSS 手法と本手法の精度の比較

この section では，本手法の領域分割精度を state-of-the-art モデルと比較し，有効性を検証する．表 3 に BAO，MARS+SAM の疑似ラベル間の IoU を基に，表 2 のそれぞれのクラス分布を満たすように疑似ラベルを選択したときの，各クラスのサンプリング数を示す．表 2 のクラス分布の Full（1,464 枚）を満たすように選択された 2,095 枚の疑似ラベルに対応する画像をラベル付き画像とし，その他の 8,487 枚をラベルなし画像としてデータセットを構築する．構築したデータセットを用いて，UniMatchV2[8] を用いて領域分割モデルの学習を行う．

表 5 に領域分割結果を示す．表 5 の Sup. は学習に用いたアノテーションであり，I は画像レベルのラベル，S は saliency map，C は contrastive language-image pre-training[21]，A は SAM[4] を表す．また，VOC-*val*，VOC-*test* は，それぞれ，PASCAL VOC 2012 検証データ，テストデータを示している．提案手法は PASCAL VOC 2012 検証データにおいて，mIoU が 84.4%を達成した．PASCAL VOC 2012 テストデータでは，mIoU が 83.3%であった．本手法の領域分割精度は従来の WSSS 手法を上回る結果を示した．

4.4 擬似ラベルの選択に関する分析

この section では，IoU に基づく疑似ラベルの選択が有効か検証する．実験結果は，1,464 枚の PASCAL VOC 2012 訓練データから得られたものである．表 6 に BAO と MARS+SAM，MARS と MARS+SAM，BAO+SAM と MARS+SAM，BAO と MARS のそれぞれの疑似ラベル間の IoU を基に選択された疑似ラベルの真値との mIoU を示す．表 6 の Pseudo-label selection は疑似ラベルの比較対象，Pseudo-label pool は疑似ラベルの選択元，1,464 images は 1,464 枚の中から選択された疑似ラベルの枚数を示す．疑似ラベル間の IoU が 90%以上のものを選択したときの精度を示す．BAO と MARS+SAM の疑似ラベル間の IoU を基に選択された疑似ラベルの精度は，BAO+SAM と MARS+SAM，MARS と MARS+SAM の疑似ラベル間の IoU を基に得られた精度を上回る．BAO+SAM と MARS+SAM により選択された疑似ラベルの精度が低下した原因は，同一の後処理手法を用いることに起因すると推察される．同一の後処理により，後処理後の疑似ラベルの精度の向上，低下に限らず，類似する疑似ラベルが作成されるため，疑似ラベル間の IoU が高くなり精度の低い疑似ラベルに対しても，高品質な疑似ラベルとして選択される．また，MARS と MARS+SAM によって選択された疑似ラベルの精度が低下した要因として，精度の低い疑似ラベルに対して後処理が十分に機能せず，後処理前後の疑似ラベルが類似したままとなるため，疑似ラベル間の IoU が高くなることが挙げられる．さらに，BAO と MARS により選択された疑似ラベルの精度は，疑似ラベルの選択元が MARS であるにもかかわらず，MARS と MARS+SAM により選択された疑似ラベルの精度を上回る．これらの結果は，異なる WSSS 手法から得られた疑似ラベル間の IoU を比較することで，頑健性の高い疑似ラベルの選択が可能となることを示唆している．

4.5 SSSS における擬似ラベルの選択の効果

この section では，IoU とクラス分布に基づいて選択された疑似ラベルから構築されたデータセットが領域分割モデルの精度に与える影響について調査する．表 7 に疑似ラベルを選択しなかった場合の精度と選択した場合の精度の比較を示す．疑似ラベルの選択においては，BAO と MARS+SAM，MARS と MARS+SAM，BAO+SAM と MARS+SAM，BAO と MARS のそれぞれの組み合わせで，表 2 中の Full（1,464 枚）のクラス分布を満たすように選択された疑似ラベルを用いて，UniMatchV2 で領域分割モデルを学習した際の精度を示す．表 7 の Labeled images は選択された疑似ラベルの枚数を示す．このとき，10,582 枚の疑似ラベルの中から選択されており，選択された疑似ラベルに対応する画像はラベル付き画像とし，その他の画像をラベルなし画像とする．SSSS の慣例に従い，疑似ラベルを選択せず，1,464 枚の疑似ラベルをラベル付き画像とし，9,118 枚をラベルなし画像として用いた場合の精度は，PASCAL VOC 2012 検証データにおいて，83.2%であった．MARS と MARS+SAM，BAO+SAM と MARS+SAM により選択された疑似ラベルを用いたときの精度は，それぞれ 82.5%，82.7%であるのに対し，BAO と MARS+SAM により選択された疑似ラベルを用いたときの精度は 84.4%であった．さらに，BAO と MARS により選択された疑似ラベルを用いて学習した場合の精度は 82.5%であり，MARS+SAM よりも精度の劣る MARS の疑似ラベルを正解データとして，学習に用いたにもかかわらず，MARS と MARS+SAM，BAO+SAM と MARS+SAM と同等の精度を達成した．実験の結果，異なる WSSS 手法に由来する疑似ラベル間の IoU を基準として選択された疑似ラベルが，領域分割モデルの性能向上に有効であることが示された．

4.6 ラベル付き画像の絶対数に関する分析

この section では，選択した疑似ラベルの数が領域分割モデルの精度に与える影響について調査する．PASCAL VOC 2012 データセットを用いて実験を行う．表 8 にクラス分布を満たすように選択された疑似ラベルを用いて学習したときの，それぞれの領域分割精度を示す．BAO，MARS+SAM の比較により選択された疑似ラベルを用いて学習を行った．このとき，各クラスに

対するサンプリング数は，表 2 を満たすように設定されている．選択された各クラスのサンプリング数を表 3 に示す．疑似ラベルの選択を行わず，SSSS と同様にそれぞれ 1/16（92 枚），1/8（183 枚），1/4（366 枚），1/2（732 枚），Full（1,464 枚）をラベル付き画像として学習したときの精度と比べて，疑似ラベルの選択を行ったときの精度は，いずれの条件においても優位な結果を示した．1/16（92 枚），1/8（183 枚），1/4（366 枚），1/2（732 枚），Full（1,464 枚）のラベル付き画像に対して，選択されたラベル付き画像の数が多い（例えば，Full（1,464 枚）→ 2,095 枚）原因は，IoU が高い疑似ラベルは単一クラスのみを有する傾向があったためである．そのため，複数クラスを有する疑似ラベルが選択されず，ラベル付き画像の数が多くなった．

5 本研究の課題

本手法では，疑似ラベル間の IoU が高いものを高品質な疑似ラベルとみなす．しかし，一方の疑似ラベルの精度が高く，もう一方の疑似ラベルの精度が低い場合，この 2 つの疑似ラベル間の IoU は低くなるため，高品質な疑似ラベルとして選択されない可能性がある．このように，疑似ラベル間の IoU に基づく疑似ラベルの選択には限界がある．また，IoU に基づく選択は，単一クラスを含む疑似ラベルが優先的に選ばれる傾向を示した．これは，単一クラスの疑似ラベルが比較的高い精度を有しているためである．一方で，複数クラスを含む疑似ラベルは，単一クラスの疑似ラベルよりも精度が低くなりやすいため，結果として選択から除外されやすい．このことは，1 枚の画像に対して，複数クラスを有するデータセットを対象とした学習において，モデルの汎化性能を制限する要因となりえる．今後は，複数クラスを有する疑似ラベルの選択や，一方の疑似ラベルの精度が高い場合の選択漏れなどを軽減する手法の開発が求められる．

6 結論

WSSS は真値に近い疑似ラベルを作成できる．さらに，SSSS は少数の真値から優れた領域分割精度を達成する．以上の点に着目し，本論文では，IoU に基づいて選択された高品質な疑似ラベルを SSSS に適応することを提案した．実験の結果，異なる WSSS 手法の疑似ラベル間で IoU を計算することで頑健な疑似ラベルの選択を行えることが示された．また，選択された疑似ラベルを用いて学習された領域分割モデルは，PASCAL VOC 2012 データセットにおいて従来手法を上回る精度を達成した．

参考文献

[1] Liu, S., Liu, K., Zhu, W., Shen, Y. and Fernandez– Granda, C.: Adaptive Early-Learning Correction for Segmentation from Noisy Annotations, *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2596–2606 (online), doi:10.1109/CVPR52688.2022.00263 (2022).

[2] Everingham, M., Van Gool, L., Williams, C. K., Winn, J. and Zisserman, A.: The pascal visual object classes (voc) challenge, *International journal of computer vision*, Vol. 88, pp. 303–338 (online), doi:10.1007/s11263-009-0275-4 (2010).

[3] Fujimori, I., Oono, M. and Shishibori, M.: BAO: Background-aware activation map optimization for weakly supervised semantic segmentation without background threshold, *Journal of Visual Communication and Image Representation*, Vol. 107, p. 104404 (online), doi:https://doi.org/10.1016/j.jvcir.2025.104404 (2025).

[4] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P. and Girshick, R.: Segment Anything, *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3992–4003 (online), doi:10.1109/ICCV51070.2023.00371 (2023).

[5] Chen, T., Mai, Z., Li, R. and Chao, W.-l.: Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation, *arXiv preprint arXiv:2305.05803* (2023).

[6] Jo, S., Yu, I.-J. and Kim, K.: MARS: Model-agnostic Biased Object Removal without Additional Supervision for Weakly-Supervised Semantic Segmentation, *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 614–623 (online), doi:10.1109/ICCV51070.2023.00063 (2023).

[7] Yang, L., Qi, L., Feng, L., Zhang, W. and Shi, Y.: Revisiting Weak-to-Strong Consistency in Semi-Supervised Semantic Segmentation, *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7236–7246 (online), doi:10.1109/CVPR52729.2023.00699 (2023).

[8] Yang, L., Zhao, Z. and Zhao, H.: UniMatch V2: Pushing the Limit of Semi-Supervised Semantic Segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 47, No. 4, pp. 3031–3048 (online), doi:10.1109/TPAMI.2025.3528453 (2025).

[9] Kweon, H. and Yoon, K.-J.: From SAM to CAMs: Exploring Segment Anything Model for Weakly Supervised Semantic Segmentation, *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19499–19509 (online), doi:10.1109/CVPR52733.2024.01844 (2024).

[10] Mae, S., Yamada, R. and Kataoka, H.: Keeping Segment Mask Quality with Self-generated Masks.

[11] Yang, D., Ji, J., Ma, Y., Guo, T., Wang, H., Sun, X. and Ji, R.: Sam as the guide: mastering pseudo-label refinement in semi-supervised referring expression segmentation, *arXiv preprint arXiv:2406.01451* (2024).

[12] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models, *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685 (online), doi:10.1109/CVPR52688.2022.01042 (2022).

[13] Yang, Z., Meng, Y., Fu, K., Wang, S. and Song, Z.: More: Class patch attention needs regularization for weakly supervised semantic segmentation, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39, No. 9, pp. 9400–9408 (online), doi:10.1609/aaai.v39i9.33018 (2025).

[14] Xu, Z., Tang, F., Chen, Z., Su, Y., Zhao, Z., Zhang, G., Su, J. and Ge, Z.: Toward modality gap: Vision prototype learning for weakly-supervised semantic segmentation with clip, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39, No. 9, pp. 9023–9031 (online), doi:10.1609/aaai.v39i9.32976 (2025).

[15] Yang, X. and Gong, X.: Foundation Model Assisted Weakly Supervised Semantic Segmentation, *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 512–521 (online), doi:10.1109/WACV57701.2024.00058 (2024).

[16] Lin, C.-S., Wang, C.-Y., Wang, Y.-C. F. and Chen, M.-H.: Semantic Prompt Learning for Weakly-Supervised Semantic Segmentation, *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8764–8774 (online), doi:10.1109/WACV61041.2025.00849 (2025).

[17] Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S. and Malik, J.: Semantic contours from inverse detectors, *2011 International Conference on Computer Vision*, pp. 991–998 (online), doi:10.1109/ICCV.2011.6126343 (2011).

[18] Krähenbühl, P. and Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials, *Advances in neural information processing systems*, Vol. 24 (online), doi:10.5555/2986459.2986472 (2011).

[19] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A. et al.: Dinov2: Learning robust visual features without supervision, *arXiv preprint arXiv:2304.07193* (2023).

[20] Ranftl, R., Bochkovskiy, A. and Koltun, V.: Vision Transformers for Dense Prediction, *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12159–12168 (online), doi:10.1109/ICCV48922.2021.01196 (2021).

[21] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al.: Learning transferable visual models from natural language supervision, *International conference on machine learning*, PmLR, pp. 8748–8763 (2021).