CH-004

背景閾値を用いない弱教師あり領域分割のための 背景 Activation Map の最適化
BAO: Background-aware Activation Map Optimization for Weakly Supervised Semantic Segmentation without Background Threshold

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

領域分割（Semantic Segmentation）は画像をピクセルレベルでクラス分類する課題である．しかし，教師あり領域分割モデルの学習には，ピクセルレベルのクラスラベルが膨大に必要であり，アノテーションコストが高い．この課題に対処するために，弱教師あり領域分割（Weakly-Supervised Semantic Segmentation:WSSS）が提案された．WSSS では，ピクセルレベルのクラスラベルに比べ，アノテーションコストの低いラベルを用いることで，アノテーションコストを削減する試みがなされている．本論文では，WSSS が用いるアノテーションのなかで最も低コストの画像レベルのクラスラベルを用いる WSSS に焦点を当てる．画像レベルのクラスラベルを用いる WSSS の一般的な手法は，画像レベルのクラスラベルを用いて学習した分類モデルから，クラス活性化マップ（Class Activation Map:CAM）を作成し，前景領域を推定する．次に CAM から作成された疑似ラベルを訓練データとして，領域分割モデルを学習する．しかし，既存手法は CAM から疑似ラベルを作成する過程において，背景領域を定めるために，背景閾値（Background Threshold）を用いる．最適な背景閾値はピクセルレベルのクラスラベルから求められるが，WSSS においては，ピクセルレベルのクラスラベルは使用できない．本論文では，新たに背景領域を活性化する Activation Map（背景 Activation Map）を作成することで，背景閾値を定めることなく，疑似ラベルを作成する手法を提案する．まず，前景領域を活性化する CAM（前景 CAM）から背景領域を活性化した CAM（背景 CAM）を作成する．次に，前景 CAM と背景 CAM から疑似ラベルを作成する．作成した疑似ラベルを用いて，Activation Map の前景領域と背景領域がギャップを持つように学習する．Activation Map の前景領域と背景領域にギャップを持たせる方法は，AMN[1] の背景領域を活性化する特徴から着想を得た．疑似ラベルをピクセルレベルの監視として，前景領域を活性化する Activation Map（前景 Activation Map）に対して，前景領域は 1，背景領域は 0 となるように学習する．このとき，背景 Activation Map は背景領域が 1，前景領域が 0 となるように学習する．この手法により，背景 Activation Map が得られ，背景閾値を用いることなく，疑似ラベルを作成することができる．また，本手法は AMN とは異なる．AMN は，背景閾値を用いて作成される疑似ラベルをピクセルレベルの訓練データとする．また，予め作成された疑似ラベルを用いて学習を行う．本手法は，疑似ラベルの作成を学習プロセスの中で行うことにより，疑似ラベルの性能を高めながら学習することができる．しかし，単一のネットワークで学習プロセス中に作成された疑似ラベルを監視として Activation Map を学習すると学習の初期段階では，精度が向上するが学習が進むにつれて CAM の前景領域が拡大し，疑似ラベルの性能の劣化を引き起こす．これは，[2] で指摘される CAM の確証バイアスのためであると考えられる．そこで，提案手法（図 1）では，学習の初期段階では 2 つのネットワーク間で重みを共有（Share）する．その後，学習を安定的に行うために，疑似ラベルを作成するネットワークの重みを指数移動平均 [3]（EMA）で更新する．本手法は，ピクセルレベルのクラスラベルが利用できず，背景閾値のチューニングができないと想定される問題に WSSS の手法を応用する上で特に重要である．

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

2　関連研究

2.1　Weakly-supervised semantic segmentation

Multi-stage WSSS は一般的に 3 段階のプロセスから構成される．まず，画像分類モデルを画像レベルのクラスラベルで学習し，CAM を作成する．次に，CAM を改良した後，疑似ラベルを作成する．最後に，作成された疑似ラベルを用いて領域分割モデルを学習する．WSSS における領域分割モデルの精度は疑似ラベルの精度に依存する．そのため，CAM の作成に関する研究は高性能な疑似ラベルのための重要な取り組みとなっている．しかし，CNN（Convolutional Neural Network）に基づく CAM はオブジェクトの局所領域を活性化する問題点がある．この問題のために，CAM の品質を向上させる手法が提案されている [1][4]．AMN[1] は，疑似ラベルを用いて Activation Map の前景領域の不均衡を軽減させる．また，Activation Map の前景領域と背景領域の間でギャップをもつように学習することで，背景閾値をロバスト化する．近年は，画像認識タスクにおいてブレイクスルーを起こした ViT（Vision Transformer）は WSSS にも応用されている．TransCAM[5] は，Conformer[6] をバックボーンネットワークとして WSSS タスクに応用した先駆的な研究であり，CNN から得られる CAM の局所活性化する問題を ViT から得られる attention map で軽減する．優れた手法が提案されたにもかかわらず，これらの手法は依然として，CAM から疑似ラベルを作成する過程において，背景閾値を必要とする．本手法は，背景閾値を用いることなく疑似ラベルを作成する．

3　事前準備

3.1　Backbone Network

本手法は Conformer[6] をバックボーンネットワークとしている．Conformer は CNN-block と Transformer-block で構成される，ハイブリッドな構造のネットワークである．Conformer は FCU[6]（Feature Coupling Unit）により，CNN-block が学習する局所的な特徴量と Transformer-block が学習する大域的特徴量を結合する．

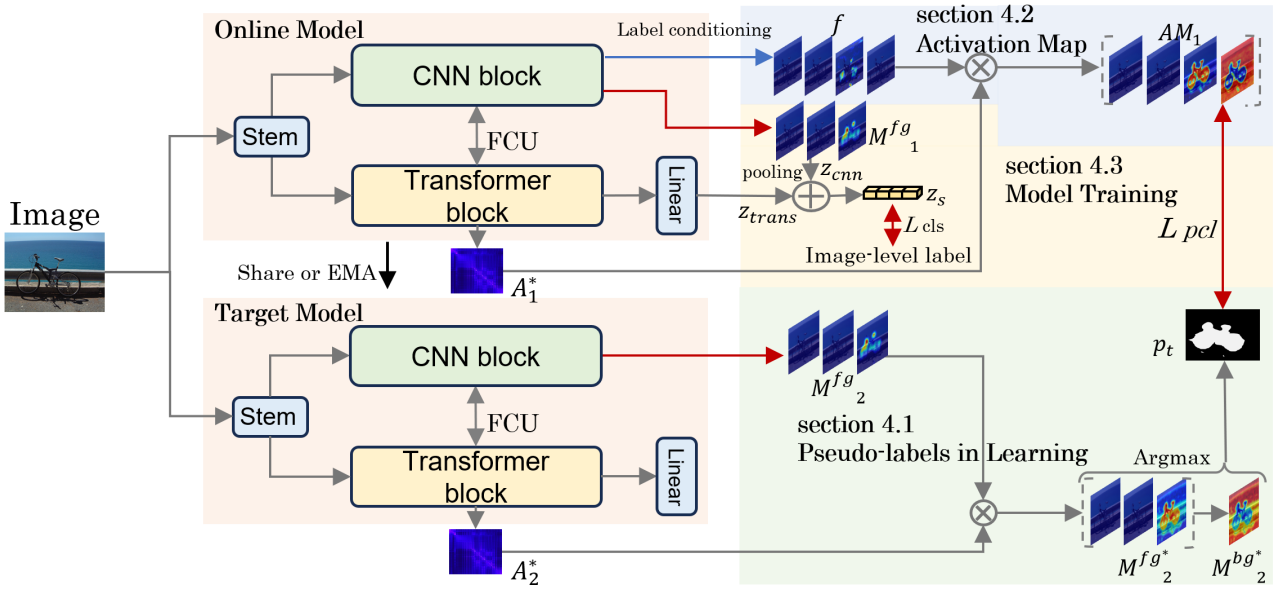

図 1: 提案手法の概要.

3.2 Class Activation Map

CAM の作成においては，Conformer をバックボーンネットワークとして WSSS に応用した TransCAM[5] を用いる．CNN-block の出力（特徴マップ）を $f \in \mathbb{R}^{fc \times fh \times fw}$．最終層の特徴マップに対応する重みを $w \in \mathbb{R}^{fc}$．CAM を $M \in \mathbb{R}^{S \times fh \times fw}$ とする．fc，fh，fw は特徴マップのチャネル数，高さ，幅を表す．S はクラス数を表す．このとき，s 番目のクラスに対応する CAM（$M_s \in \mathbb{R}^{fh \times fw}$）は以下の式で定式化できる．$s$ は特定のクラスを表す．

$$M_s = w^{\top} f, \tag{1}$$

3.3 attention map による CAM の強化

TransCAM[5] は，CNN に基づく CAM を ViT から得られる attention map で強化する．Transformer-block から得られる attention weight を $A^{b,h} \in \mathbb{R}^{(1+N)\times(1+N)}$ とする．$A^{b,h}$ は以下の式で計算される．

$$A^{b,h} = \mathrm{softmax}\left(\frac{Q^{b,h} K^{b,h\top}}{\sqrt{D/H}}\right), \tag{2}$$

b，h，1 は特定の block，特定のヘッド，class token を表す．$Q^{b,h}$，$K^{b,h}$ は b 番目の Transformer-block で得られるクエリ（Query），キー（Key）を表す．N は patch token のサイズを表す．D は token の埋め込み次元（embedding dimensions）を表す．H はヘッド数を表す．attention weight の $A^{b,h}$ を各 block ごとのヘッド方向に平均化したものを $\bar{A}^b \in \mathbb{R}^{(1+N)\times(1+N)}$ とする．各 block の $\bar{A}^b$ を合計したものを $A \in \mathbb{R}^{(1+N)\times(1+N)}$ とする．$\bar{A}^b$ 及び，A は以下の式で計算される．

$$\bar{A}^b = \frac{1}{H}\sum_h (A^{b,h}), \tag{3}$$

$$A = \sum_b (\bar{A}^b), \tag{4}$$

$A \in \mathbb{R}^{(1+N)\times(1+N)}$ において，class token に関連する attention weight を削除したものを $A^* \in \mathbb{R}^{N\times N}$ とする．s 番目のクラスに対応する CAM $M_s \in \mathbb{R}^{\sqrt{N}\times\sqrt{N}}$ を $M_s \in \mathbb{R}^{N\times 1}$ に変形する．attention map $A^* \in \mathbb{R}^{N\times N}$ により強化された CAM ${M_s}^* \in \mathbb{R}^{\sqrt{N}\times\sqrt{N}}$ は以下の式で計算される．

$${M_s}^* = A^* \cdot M_s, \tag{5}$$

このとき，${M_s}^* \in \mathbb{R}^{N\times 1}$ であり，${M_s}^* \in \mathbb{R}^{\sqrt{N}\times\sqrt{N}}$ に変形する．

4 提案手法

提案手法の概要を図 1 に示す．本手法は 2 つの Conformer による Dual 構造のネットワークである．まず，Target Model で疑似ラベル p_t を作成する．このとき，背景閾値は使用しない．Target Model の前景 CAM ${M^{fg}}^*_2$ から作成された背景 CAM ${M^{bg}}^*_2$ を背景閾値の代わりに用いる．次に，作成された疑似ラベルを監視として，Online Model から作成される Activation Map AM_1 の前景領域と背景領域にギャップをもたせるように学習する．疑似ラベル p_t は，学習プロセスの中で作成される．Online Model の重みを Share，または，EMA により Target Model の重みを更新する．推論における疑似ラベル p_i は Target Model から作成し，背景 Activation Map AM_2 を背景閾値の代わりに用いる．まとめると，提案手法は学習時における疑似ラベル作成，Activation Map の作成，モデルの学習，推論時における疑似ラベルの作成から構成される．

4.1 学習時における疑似ラベル

Target Model から前景 CAM ${M^{fg}}_2 \in \mathbb{R}^{(S-1)\times\sqrt{N}\times\sqrt{N}}$，attention map ${A^*}_2 \in \mathbb{R}^{N\times N}$ を得る．$S-1$ は，前景のクラス数を示す．${M^{fg}}_2$ と ${A^*}_2$ から得られる前景 CAM を ${M^{fg}}^*_2 \in \mathbb{R}^{(S-1)\times\sqrt{N}\times\sqrt{N}}$ とする．このとき，特定のクラス s の CAM を ${M_s}^{fg^*}_2 \in \mathbb{R}^{\sqrt{N}\times\sqrt{N}}$ とする．次に，前景 CAM から背景 CAM（${M^{bg}}^*_2 \in \mathbb{R}^{1\times\sqrt{N}\times\sqrt{N}}$）を作成する．${M^{bg}}^*_2$ は以下の式で計算される．

$${M^{bg}}^*_2 = 1 - \max_{1\le s\le S-1} {M_s}^{fg^*}_2, \tag{6}$$

$M^{bg^*}{}_2$ と $M^{fg^*}{}_2$ を結合 $(M^{bg^*}{}_2 \cup M^{fg^*}{}_2)$ したものを $M^* \in \mathbb{R}^{S\times\sqrt{N}\times\sqrt{N}}$ とする．疑似ラベル p_t は，前景 CAM と背景 CAM から作成される．M^* 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 s を疑似ラベル $p_t \in \mathbb{R}^{\sqrt{N}\times\sqrt{N}}$ とする．よって，学習時における疑似ラベルは，背景閾値を定めることなく作成することができる．

4.2 Activation Map の作成

Online Model における CNN-block の出力を $f \in \mathbb{R}^{S\times\sqrt{N}\times\sqrt{N}}$ とする．f の作成には，AMN[1] で提案された label conditioning を用いている．f と Online Model の attention map $A^*{}_1$ から得られるたものを Activation Map $AM_1 \in \mathbb{R}^{S\times\sqrt{N}\times\sqrt{N}}$ とする．

4.3 モデルの学習

本手法の学習には，画像分類のための損失関数 L_{cls} とピクセルレベルの分類のための損失関数 L_{pcl} を用いる．L_{cls} は以下の式で得られる．

$$L_{cls} = -\frac{1}{S-1}\sum_{s=1}^{S-1}[y_s \ln(\frac{1}{1+\exp(-z_s)}) + (1-y_s)\ln(\frac{\exp(-z_s)}{1+\exp(-z_s)})], \quad (7)$$

このとき，$y_s \in \{0,1\}$ は s クラスに対する真値，$z_s \in \mathbb{R}$ は Online Model の s クラスの分類予測値である．z_s は，$M^{fg}{}_1 \in \mathbb{R}^{S-1\times\sqrt{N}\times\sqrt{N}}$ に Global Average Pooling を適用した値 z_{cnn} と，class token に線形層を適用した値 z_{trans} を足し合わせたものである．L_{pcl} は，[7] で提案された損失関数であり，AMN[1] で採用されたものを用いる．L_{pcl} には，Activation Map AM_1，疑似ラベル p_t が入力される $L_{pcl}(AM_1, p_t)$．以下の L を損失として学習する．

$$L = L_{cls} + L_{pcl}, \quad (8)$$

損失 L は，Online Model の学習に用いられる．Target Model の重みの更新は Online Model の重みの Share により行われる．しかし，重みを Share する場合，CAM の確証バイアス [2] のため，前景 CAM の活性化領域が拡大する問題がある．活性化領域の拡大は疑似ラベルの品質の劣化させる．そこで，学習の初期段階では重みを Share する．その後，学習を安定化させるために Online Model の重みに対して EMA を用いることで Target Model の重みを更新する．Online Model の重みを w，Target Model の重みを θ とする．Share では，$\theta \leftarrow w$ のように更新される．EMA では，次のように更新される．

$$\theta \leftarrow \lambda\theta + (1-\lambda)w, \quad (9)$$

λ は，cosine schedule に従う数値である．

4.4 推論時における疑似ラベル

推論時における疑似ラベルの背景領域は，Target Model の背景 Activation Map $AM^{bg}{}_2 \in \mathbb{R}^{1\times\sqrt{N}\times\sqrt{N}}$ を用いる．前景領域は前景 CAM $M^{fg^*}{}_2 \in \mathbb{R}^{(S-1)\times\sqrt{N}\times\sqrt{N}}$ を用いる．$M^{fg^*}{}_2$ と $AM^{bg^*}{}_2$ を結合したものを，$M^*_{seed} \in \mathbb{R}^{S\times\sqrt{N}\times\sqrt{N}}$ とする．学習時に作成する疑似ラベルと同様に M^*_{seed} 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 s を疑似ラベル $p_i \in \mathbb{R}^{\sqrt{N}\times\sqrt{N}}$ とする．よって，推論時における疑似ラベルは，背景閾値を定めることなく作成することができる．疑似ラベル p_i を訓練データとして，領域分割モデルを学習する．

表 1: 初期疑似ラベルと後処理後の疑似ラベル及び，領域分割結果の mIoU に基づく比較．

Method	Post.	Seed	Mask	*val*	*test*
TransCAM[5] JVCIR23	PSA[4] + CRF[8]	64.9	70.2	69.3	69.6
SeCo[9] CVPR2024	CRF[8]	74.8	76.5	74.0	73.8
Ours	CRF[8]	**75.5**	**77.4**	**74.1**	**74.6**

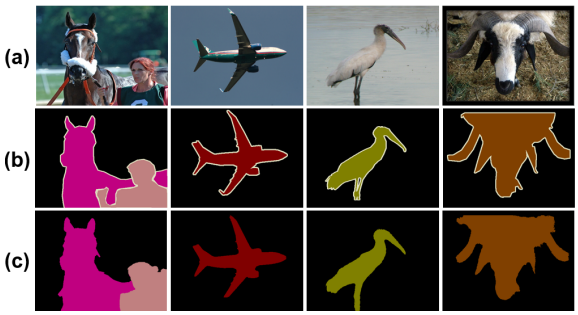

図 2: 領域分割結果．

5 実験

5.1 データセットと評価指標

評価用データセットとして，PASCAL VOC 2012[10] を用いた．PASCAL VOC 2012 データセットでは背景クラスと 20 の前景クラスを含む 21 のクラスで領域分割する．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[11] 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．評価指標として mIoU（mean Intersection over Union）を用いる．実験において，TransCAM[5](Baseline) と同じハイパーパラメータの値を用いる．

5.2 実験結果

初期疑似ラベル（Seed）の付与精度を評価する．表 1 に PASCAL VOC 2012 訓練データで行った実験結果を示す．Baseline との精度比較を行う．表 1 中の Post. は，初期疑似ラベルに用いられる後処理を示す．PASCAL VOC 2012 訓練データにおいて，提案手法は mIoU が 75.5%であった．次に CRF[8] による後処理後の疑似ラベル（Mask）の性能を評価する．PASCAL VOC 2012 訓練データの後処理後の疑似ラベルにおいて，提案手法は mIoU が 77.4%であり，TransCAM[5](Baseline) を上回る疑似ラベルの精度を得た．本手法は，[4] のような追加の学習が必要な後処理手法を用いていない．疑似ラベルを訓練データとして，教師あり学習 [12] を行った．表 1 に領域分割精度を示す．PASCAL VOC 2012 検証データ（val）において，提案手法は mIoU が 74.1%であった．さらに，PASCAL VOC 2012 テストデータ（test）において，提案手法は mIoU が 74.6%であった．また，本手法は最先端の WSSS 手法である [9] を上回る精度を得た．図 2 に領域分割モデルの推論結果を示す．上から，(a) 元画像，(b) 真値，(c) 提案手法の領域分割の結果である．図 3 に提案手法で得られる背景 Activation Map を示す．上から，(a) 元画像，(b) 真値，(c) 背景 Activation Map である．本手法により背景領域を活性化する Activation Map が得られた．

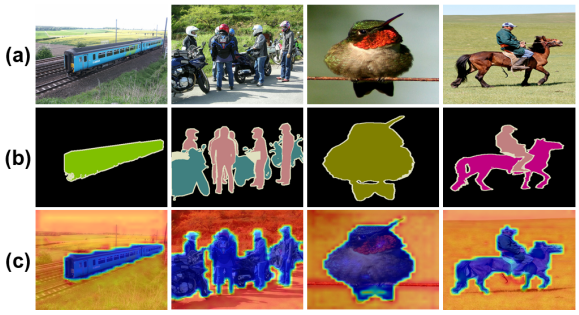

図 3: 提案手法の背景 Activation Map.

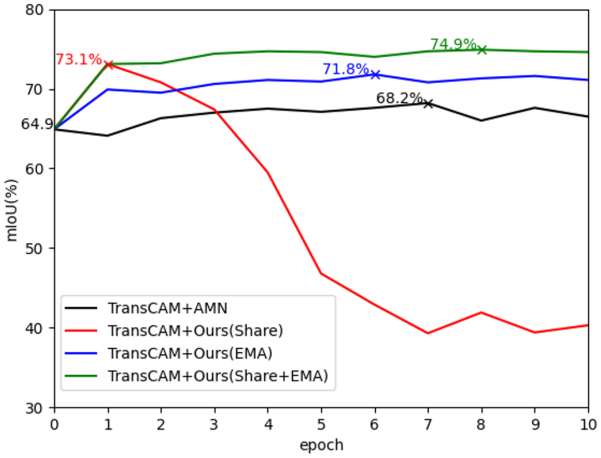

図 4: epoch 毎の疑似ラベルの mIoU の変移.

5.3 本手法の課題

本手法は，TransCAM[5] の精度を向上させることができるが，限界も存在する．図 4 に，epoch 毎の疑似ラベルの mIoU の変移を示す．Online Model から得られた疑似ラベルの mIoU で比較を行っている．黒線が TransCAM[5] に AMN[1] を適用した場合，赤線が TransCAM に本手法（Share）を適用した場合，青線が TransCAM に本手法（EMA）を適用した場合，緑線が Share と EMA を組み合わせて TransCAM に本手法を適用した場合の epoch 毎の mIoU を示している．それぞれ，TransCAM の学習済みの重みを用いて初期化している．緑線の Share と EMA を組み合わせた手法では，1epoch のみ重みを Share で更新している．その後は，EMA により重みを更新している．Share のみで学習した場合は，1epoch 目で mIoU が向上しているが，epoch を重ねると mIoU が低下している．学習初期段階では，学習プロセス中に疑似ラベルを更新することで，mIoU を向上させるが，1epoch 目以降は CAM の確証バイアス [2] の影響により mIoU の低下が著しい．一方，EMA のみで学習した場合は，epoch ごとに mIoU が向上していることがわかる．Target Model の重み θ は誤差逆伝搬では更新されず，Online Model の重み w に EMA を適用することで更新する．また，λ は 1 に近い値（本手法は $\lambda = 0.9995$）に設定することで w に比べ，安定的に更新される．よって，CAM の疑似ラベルに対する過学習が軽減されたと考察できる．本手法では，1epoch のみ Share を行い，その後の epoch は EMA を適用することで Share と EMA の両方の利点を活用する．しかし，Share から EMA への変更の epoch 数は人手で決めており，本手法のボトルネックとなっている．今後の課題として，Share から EMA への変更の自動化，また，Share 及び，EMA のみの学習による精度向上が挙げられる．

6 結論

本稿の目的は，背景閾値を用いない WSSS の可能性を示すことにある．学習時は前景 CAM を反転した背景 CAM を背景閾値の代わりに用い，推論時は学習により得られる背景 Activation Map を背景閾値の代わりに用いる．本手法は，PASCAL VOC 2012 データセットにおいて，既存手法の精度を上回った．

参考文献

[1] Lee, M., Kim, D. and Shim, H.: Threshold matters in wsss: Manipulating the activation for the robust and accurate segmentation model against thresholds, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4330–4339 (2022).

[2] Arazo, E., Ortego, D., Albert, P., O' Connor, N. E. and McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning, *2020 International joint conference on neural networks (IJCNN)*, IEEE, pp. 1–8 (2020).

[3] He, K., Fan, H., Wu, Y., Xie, S. and Girshick, R.: Momentum contrast for unsupervised visual representation learning, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738 (2020).

[4] Ahn, J. and Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4981–4990 (2018).

[5] Li, R., Mai, Z., Zhang, Z., Jang, J. and Sanner, S.: Transcam: Transformer attention-based cam refinement for weakly supervised semantic segmentation, *Journal of Visual Communication and Image Representation*, Vol. 92, p. 103800 (2023).

[6] Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J. and Ye, Q.: Conformer: Local features coupling global representations for visual recognition, *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 367–376 (2021).

[7] Huang, Z., Wang, X., Wang, J., Liu, W. and Wang, J.: Weakly-supervised semantic segmentation network with deep seeded region growing, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7014–7023 (2018).

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

[9] Yang, Z., Fu, K., Duan, M., Qu, L., Wang, S. and Song, Z.: Separate and Conquer: Decoupling Co-occurrence via Decomposition and Representation for Weakly Supervised Semantic Segmentation, *arXiv preprint arXiv:2402.18467* (2024).

[10] 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 (2010).

[11] Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S. and Malik, J.: Semantic contours from inverse detectors, *2011 international conference on computer vision*, IEEE, pp. 991–998 (2011).

[12] Liang-Chieh, C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A.: Semantic image segmentation with deep convolutional nets and fully connected crfs, *International conference on learning representations* (2015).