

Railroad is not a Train: Saliency as Pseudo-pixel Supervision for Weakly Supervised Semantic Segmentation

Seungho Lee*Yonsei University
seungholee@yonsei.ac.kr

Minhyun Lee*
Yonsei University
lmh315@yonsei.ac.kr

Jongwuk Lee
Sungkyunkwan University
jongwuklee@skku.edu

Hyunjung Shim[†]
Yonsei University
kateshim@yonsei.ac.kr

Abstract

既存の画像レベルの弱い監視を使用した弱教師ありセマンティックセグメンテーション (WSSS) の研究には、いくつかの制限があります: オブジェクトのカバレッジが疎であること、不正確なオブジェクトの境界、非ターゲットオブジェクトからの共起ピクセル。これらの課題を克服するために、我々は新しいフレームワーク、すなわち Explicit Pseudo-pixel Supervision (EPS) を提案します。これは、ローカライゼーションマップを介してオブジェクトのアイデンティティを提供する画像レベルのラベルと、オフ□ザ□シェルフの顕著性検出モデルからの顕著性マップが豊富な境界を提供することによって、ピクセルレベルのフィードバックから学習します。我々は、両方の情報の補完的な関係を完全に活用するための共同トレーニング戦略を考案しました。我々の方法は、正確なオブジェクトの境界を取得し、共起ピクセルを破棄することができ、擬似マスクの品質を大幅に向上させます。実験結果は、提案された方法が WSSS の主要な課題を解決することにより、既存の方法

を著しく上回り、PASCAL VOC 2012 および MS COCO 2014 データセットの両方で新しい最先端のパフォーマンスを達成することを示しています。コードは<https://github.com/halbielee/EPS>で利用可能です。

1. Introduction

弱教師ありセマンティックセグメンテーション (WSSS) は、弱い監視 (e.g., 画像レベルのラベル [36, 37], スクリブル [29], またはバウンディングボックス [22]) を利用し、ピクセルレベルのラベルを必要とする完全教師ありモデルに匹敵するパフォーマンスを達成することを目指しています。ほとんどの既存の研究は、セグメンテーションモデルの弱い監視として画像レベルのラベルを採用しています。WSSS の全体的なパイプラインは 2 つのステージで構成されています。まず、画像分類器を使用してターゲットオブジェクトのための擬似マスクが生成されます。次に、セグメンテーションモデルは監視として擬似マスクを使用して訓練されます。擬似マスクを生成するための一般的な技術は、クラスアクティベーションマッピング (CAM) [52]であり、画像レベルのラベルに対応す

*indicates an equal contribution.

[†]Hyunjung Shim is a corresponding author.

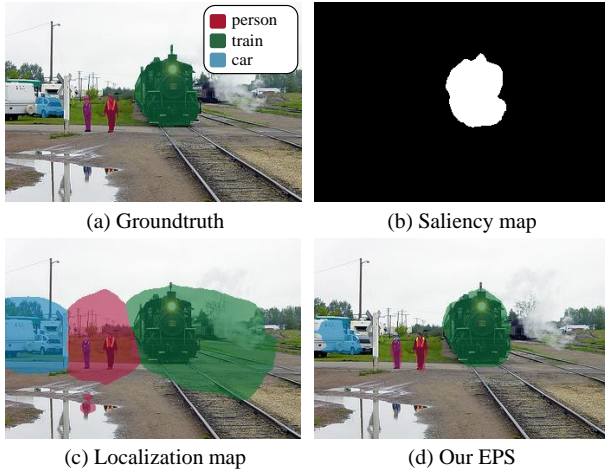


図 1. WSSS のための顕著性マップとローカライゼーションマップの両方を利用する動機付けの例。(a) グラウンドトゥールース、(b) PFAN [51]による顕著性マップ、(c) CAM [52]によるローカライゼーションマップ、(d) 顕著性マップとローカライゼーションマップの両方を利用して分類器を訓練する我々の EPS。顕著性マップは人と車を捉えることができないが、我々の結果はそれらを正しく復元でき、ローカライゼーションマップは 2 つのオブジェクトを過剰に捉えていることに注意。

るオブジェクトのローカライゼーションマップを提供します。完全 (i.e., ピクセルレベルのアノテーション) と弱 (i.e., 画像レベルのラベル) 教師ありセマンティックセグメンテーションの間の監視ギャップのために、WSSS には以下の主要な課題があります: 1) ローカライゼーションマップはターゲットオブジェクトのごく一部しかキャプチャしない [52]、2) オブジェクトの境界の不一致に苦しむ [23]、3) ターゲットオブジェクトからの共起ピクセルをほとんど分離できない (e.g., 列車からの鉄道) [25]。

これらの問題に対処するために、既存の研究は 3 つの柱に分類できます。最初のアプローチは、ピクセルを消去することによってオブジェクトの全体を捉えるためにオブジェクトのカバレッジを拡大することです [9, 23, 28]、スコアマップをアンサンブルすること [21, 27]、または自己教師あり信号を使用すること [41]です。しかし、これらの方法は、オブジェクトの形状を導く手がかりがな

いため、ターゲットオブジェクトの正確な境界を決定することができません。第二のアプローチは、擬似マスクのオブジェクト境界を改善することに焦点を当てています [13, 32]。これらは効果的にオブジェクトの境界を学習するため、自然に擬似マスクを境界まで拡張します。しかし、依然として非ターゲットオブジェクトの重複ピクセルをターゲットオブジェクトから区別することができません。これは、前景と背景の強い相関関係 (i.e., 共起) が、ターゲットオブジェクトとその重複ピクセルを観察する頻度 (i.e., 帰納的バイアス) とほとんど区別がつかないためです [10]。最後に、第三のアプローチは、追加のグラウンドトゥールースマスク [24]や顕著性マップ [35, 47]を使用して共起問題を軽減することを目的としています。しかし、[24, 28]は、弱教師あり学習パラダイムからは程遠い強いピクセルレベルのアノテーションを必要とします。[35]は顕著性マップのエラーに敏感です。また、[47]はオブジェクトの全体をカバーせず、境界の不一致に苦しんでいます。

本論文では、画像レベルのラベルで訓練された画像分類器からの CAM (i.e., ローカライゼーションマップ) と、既製の顕著性検出モデルの出力 (i.e., 顕著性マップ) を完全に活用することにより、WSSS の 3 つの課題を克服することを目指します。ローカライゼーションマップと顕著性マップの補完的な関係に焦点を当てます。図 1 に示すように、ローカライゼーションマップは異なるオブジェクトを区別できますが、その境界を効果的に分離することはできません。対照的に、顕著性マップは豊富な境界情報を提供しますが、オブジェクトの識別を明らかにしません。この意味で、2 つの補完的な情報を使用する我々の方法が WSSS のパフォーマンスボトルネックを解決できると主張します。

この目的のために、我々は WSSS のための新しいフレームワーク、Explicit Pseudo-pixel Supervision (EPS) を提案します。顕著性マップ (i.e., 前景と背景の両方) を完全に活用するために、C

ターゲットクラスと背景クラスからなる $C+1$ クラスを予測する分類器を設計します。 C ローカライゼーションマップと背景ローカライゼーションマップを活用して顕著性マップを推定します。次に、顕著性損失は、顕著性マップと我々の推定顕著性マップのピクセル単位の差として定義されます。顕著性損失を導入することにより、モデルはすべてのクラスにわたる擬似ピクセルフィードバックによって監督されることができます。また、マルチラベル分類損失を使用して画像レベルのラベルを予測します。したがって、分類器を顕著性損失とマルチラベル分類損失の両方を最適化するように訓練し、背景と前景ピクセルの予測を相乗的に行います。我々の戦略が顕著性マップ（セクション 3.3 と図 3）と擬似マスク（セクション 5.1 と図 4）の両方を改善できることがわかります。

顕著性損失が擬似ピクセルフィードバックを通じて境界の不一致を罰するため、我々の方法がオブジェクトの正確な境界を学習することを強制できることを強調します。副産物として、境界までマップを拡張することによってオブジェクト全体を捉えることもできます。顕著性損失が前景（e.g.、列車）を背景から分離するのを助けるため、我々の方法は共起ピクセル（e.g.、鉄道）を背景クラスに割り当てることができません。実験結果は、我々の EPS が PASCAL VOC 2012 と MS COCO 2014 データセットで新しい最先端の精度を記録し、顕著なセグメンテーション性能を達成することを示しています。

2. 関連研究

弱教師ありセマンティックセグメンテーション。 WSSS の一般的なパイプラインは、分類ネットワークから擬似マスクを生成し、その擬似マスクを監督として使用してセグメンテーションネットワークを訓練することです。画像レベルのラベルにおける境界情報の不足により、多くの既存の手法は不正確な擬似マスクに悩まされています。この問題に対処するために、クロスイメージアフィニ

ティ [15]、知識グラフ [31]、およびコントラスト最適化 [38, 50] が擬似マスクの品質を向上させるために使用されています。[5] は、分類器が CAM を改善するように強制するために、サブカテゴリを発見する自己監督タスクを提案しています。[1, 2] は、ピクセル間のアフィニティを計算することで境界情報を暗黙的に活用しています。[49] は、信頼性のあるピクセルレベルのアノテーションを生成し、セグメンテーションマップを生成するためのエンドツーエンドネットワークを設計しています。[20, 25] は、境界損失を利用してセグメンテーションネットワークを訓練しています。最近では、[3] が自己監督型のトレーニングスキームを持つ単一のセグメンテーションベースのモデルを使用しています。[14] は、複数の不完全な擬似マスクを利用してセグメンテーションネットワークのロバスト性に焦点を当てています。

顕著性ガイド付きセマンティックセグメンテーション。 顕著性検出 (SD) 手法は、ピクセルレベルのアノテーションを持つ外部の顕著性データセット [18, 46, 51]、または画像レベルのアノテーション [39] を介して、画像内の前景と背景を区別する顕著性マップを生成します。多くの WSSS 手法 [15, 20, 27, 28, 42, 44] は、擬似マスクの背景手がかりとして顕著性マップを活用しています。[43] は、単一オブジェクト画像の完全な監督として顕著性マップを利用しています。[16] は、オブジェクトの類似性グラフを学習するためにインスタンスレベルの顕著性マップを使用しています。[6, 40, 47] は、クラス固有の注意手がかりと顕著性マップを組み合わせ、信頼性のある擬似マスクを生成しています。[48] は、単一のネットワークを使用して WSSS と SD を共同で解決し、両方のタスクのパフォーマンスを向上させています。我々の EPS は顕著性ガイド付き手法に分類されますが、以下の理由で他のすべてと明確に区別されます。既存のほとんどの手法は、擬似マスクの一部として、または分類器の中間特徴を洗練するための暗黙のガイダンスとして顕著性マップを活用しています。対照的

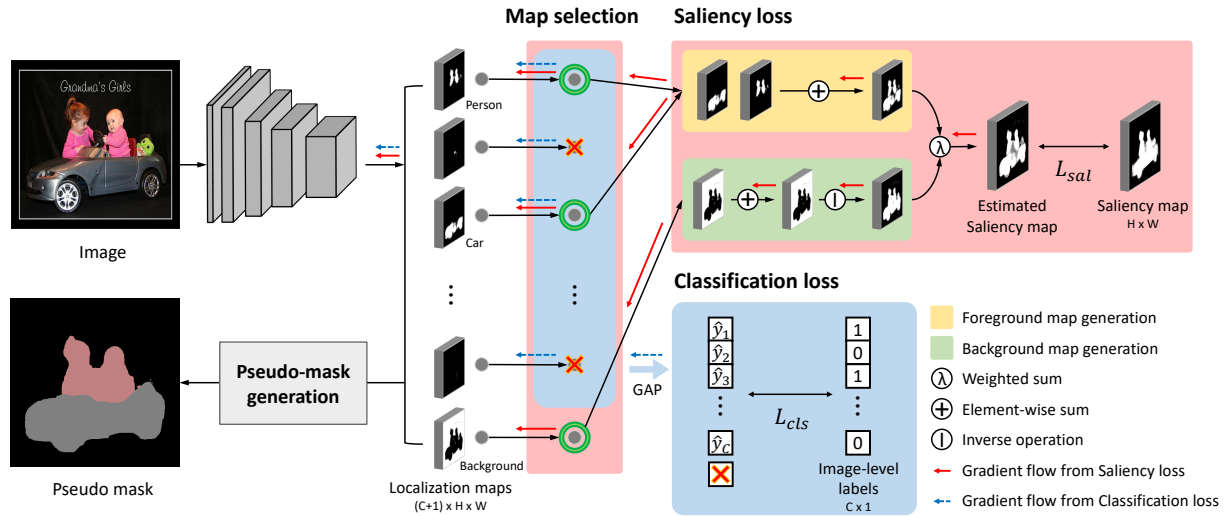


図 2. 我々の EPS の全体的なフレームワーク。 $C+1$ のローカライゼーションマップがバックボーンネットワークから生成されます。実際の顕著性マップは、既製の顕著性検出モデルから生成されます。ターゲットトラベルのためのいくつかのローカライゼーションマップは、推定顕著性マップを生成するために選択的に使用されます（セクション 3.2）。全体的なフレームワークは、顕著性損失と分類損失と共に共同で訓練されます（セクション 3.3）。

に、我々の手法は、ローカライゼーションマップの擬似ピクセルフィールドバックとして顕著性マップを利用します。[48]は、2つの補完的な情報を利用するという点で我々の手法に最も類似していますが、共起問題に対処せず、ノイズの多い顕著性マップの問題を扱っていません。

3. 提案手法

このセクションでは、弱教師ありセマンティックセグメンテーション（WSSS）のための新しいフレームワークである Explicit Pseudo-pixel Supervision (EPS) を提案します。WSSS の2つのステージを考慮すると、最初のステージは擬似マスクを生成し、2番目のステージはセグメンテーションモデルを訓練することです。ここでの我々の主な貢献は、正確な擬似マスクを生成することです。WSSS の慣例に従って [13, 21, 27, 28, 41, 42]、最初のステージで生成された擬似マスクを監督として使用してセグメンテーションモデルを訓練します。

3.1. 動機

EPS の重要な洞察は、ローカライゼーションマップからのオブジェクトアイデンティティと顕著

性マップからの境界情報という2つの補完的な情報を完全に活用することです。この目的のために、ターゲットトラベルと背景の両方に対して、ローカライゼーションマップへの擬似ピクセルフィールドバックとして顕著性マップを利用します。我々は、追加の背景クラスを持つ分類器を考案し、図 2 に示すように、合計 $C+1$ クラスを予測するようにします。この分類器を使用して、ターゲットトラベル用の C 個のローカライゼーションマップと背景ローカライゼーションマップの合計 $C+1$ 個のローカライゼーションマップを学習できます。次に、EPS が WSSS における境界の不一致と共起問題の両方にどのように対処できるかを説明します。境界の不一致問題を管理するために、 C のローカライゼーションマップから前景マップを推定し、それを顕著性マップの前景と一致させます。この方法により、ターゲットトラベルのローカライゼーションマップは顕著性マップから擬似ピクセルフィールドバックを受け取り、オブジェクトの境界を改善することができます。非ターゲットオブジェクトの共起ピクセルを軽減するために、背景のローカライゼーションマップも顕著性マップと一致させます。背景のローカライゼーションマップも顕著性マッ

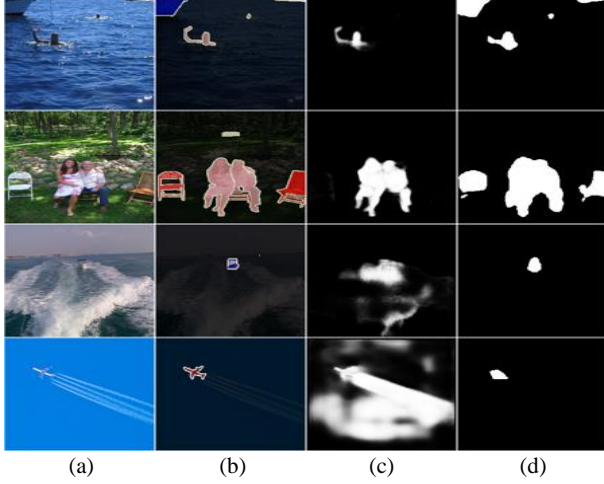


図 3. PASCAL VOC 2012 における推定された顕著性マップの定性的な例。(a) 入力画像、(b) グラウンドトゥルース、(c) [51]からの顕著性マップ、(d) 我々の推定した顕著性マップ。

プから擬似ピクセルフィードバックを受け取るため、共起ピクセルは背景にうまく割り当てられます。非ターゲットオブジェクトの共起ピクセルは主に背景と重なっているためです。これが、私たちの方法がターゲットオブジェクトから共起ピクセルを分離できる理由です。最後に、EPS の目的関数は 2 つの部分で構成されています：顕著性マップを介した顕著性損失 \mathcal{L}_{sal} (図 2 の赤いボックス/矢印で示されています) と、画像レベルのラベルを介したマルチラベル分類損失 \mathcal{L}_{cls} (図 2 の青いボックス/矢印で示されています)。2 つの目的を共同で訓練することにより、ローカライゼーションマップと顕著性マップを補完的な情報でシナジーさせることができます。図 3 に示されているように、互いのノイズや欠落情報が私たちの共同訓練戦略を通じて補完されることを観察します。例えば、市販のモデル [18, 34, 51] から得られた元の顕著性マップには欠落やノイズ情報があります。一方、私たちの結果は欠落したオブジェクト (例: ボートや椅子) をうまく復元し、ノイズ (例: 水泡や飛行機雲) を除去し、元の顕著性マップよりも明らかに優れています。その結果、EPS はより正確なオブジェクトの境界をキャプチャし、ターゲッ

トオブジェクトから共起ピクセルを分離することができます。これらの利点は、驚くべき性能向上をもたらします。表 6 は、EPS が既存のモデルを最大 3.8–10.6% のセグメンテーション精度の向上で著しく上回っていることを報告しています。

3.2. 明示的な擬似ピクセル監督

顕著性マップを擬似ピクセル監督にどのように利用するかを説明します。顕著性マップの主な利点は、オブジェクトのシルエットを提供し、オブジェクトの境界をよりよく明らかにすることです。この特性を利用するために、顕著性マップを前景と背景の 2 つのケースと一致させます。クラスごとのローカライゼーションマップを顕著性マップと比較可能にするために、ターゲットラベルのローカライゼーションマップをマージし、前景マップ $M_{fg} \in \mathbb{R}^{H \times W}$ を生成します。また、背景ラベルのローカライゼーションマップである背景マップの反転を行うことで前景を表現することもできます (後で、ノイズのある顕著性マップに対処するために前景マップをどのように洗練するかを説明します)。

具体的には、次のようにして M_{fg} と M_{bg} を使用して顕著性マップ \hat{M}_s を推定します：

$$\hat{M}_s = \lambda M_{fg} + (1 - \lambda)(1 - M_{bg}), \quad (1)$$

ここで、 $\lambda \in [0, 1]$ は前景マップと背景マップの反転の重み付き和を調整するハイパーパラメータです (デフォルトでは、実験で λ を 0.5 に設定し、 λ の追加のアブレーション研究は補足資料に記載されています)。次に、推定された顕著性マップと実際の顕著性マップのピクセルごとの差の合計として顕著性損失 \mathcal{L}_{sal} を定義します (\mathcal{L}_{sal} の正式な定義はセクション 3.3 に示されています)。

事前訓練されたモデルを使用することは弱教師あり学習と見なされるため、顕著性マップを利用することは WSSS で一般的な慣行として広く受け入れられています。その人気にもかかわらず、完全に教師ありの顕著性検出モデルを採用することは、異なるデータセットからのピクセルレベルの

アノテーションを使用するため、議論の余地があります。本論文では、異なる顕著性検出方法の効果を調査し、1) 教師なしおよび 2) 完全に教師ありの顕著性検出モデル（セクション 5.3 を参照）を使用し、いずれかを使用する私たちの方法が完全に教師ありの顕著性モデルを使用する他のすべての方法 [13, 21, 40, 43, 47] を上回ることを実証的に示します。既存の方法が顕著性マップを完全に活用することに制限されている一方で、私たちの方法は顕著性マップを擬似ピクセル監督として組み込み、境界と共起ピクセルの手がかりとして活用します。

顕著性バイアス进行处理するためのマップ選択。以前は、前景マップはターゲットラベルのローカライゼーションマップの合併であり、背景マップは背景ラベルのローカライゼーションマップであると仮定していました。しかし、そのような単純な選択ルールは、市販のモデルによって計算された顕著性マップと互換性がないかもしれません。例えば、[51] の顕著性マップは、しばしばいくつかの物体を顕著な物体として無視します（例：図 1 の列車の近くにいる小さな人々）。この体系的な誤差は、顕著性モデルが異なるデータセットの統計を学習するため、避けられません。この誤差を考慮しない限り、同じ誤差が我々のモデルに伝播し、性能の低下を引き起こす可能性があります。

この体系的な誤差に対処するために、ローカライゼーションマップと顕著性マップの重なり率を使用した効果的な戦略を開発しました。具体的には、 i 番目のローカライゼーションマップ M_i は、顕著性マップと $\tau\%$ 以上重なっている場合に前景に割り当てられ、それ以外の場合は背景に割り当てられます。形式的には、前景と背景のマップは次のように計算されます：

$$\begin{aligned} M_{fg} &= \sum_{i=1}^C y_i \cdot M_i \cdot \mathbb{1}[\mathcal{O}(M_i, M_s) > \tau], \\ M_{bg} &= \sum_{i=1}^C y_i \cdot M_i \cdot \mathbb{1}[\mathcal{O}(M_i, M_s) \leq \tau] + M_{C+1}, \end{aligned} \quad (2)$$

ここで、 $y \in \mathbb{R}^C$ はバイナリの画像レベルラベルで

あり、 $\mathcal{O}(M_i, M_s)$ は M_i と M_s の重なり率を計算する関数です。そのために、まずローカライゼーションマップと顕著性マップをバイナライズし、ピクセル p に対して、 $B_k(p) = 1$ もし $M_k(p) > 0.5$ ならば； $B_k(p) = 0$ 、それ以外の場合。 B_i と B_s はそれぞれ M_i と M_s に対応するバイナライズされたマップです。次に、 M_i と M_s の重なり率を計算します、すなわち、 $\mathcal{O}(M_i, M_s) = |B_i \cap B_s| / |B_i|$ 。データセットやバックボーンモデルに関係なく、 $\tau = 0.4$ に設定します。補足資料では、 τ の選択に対して我々の方法が頑健であることを示しています（すなわち、 τ が $[0.3, 0.5]$ の範囲であれば、同等の性能を示します）。

背景ラベルの単一のローカライゼーションマップの代わりに、背景ラベルのローカライゼーションマップを前景として選択されなかったローカライゼーションマップと組み合わせます。単純ではありますが、顕著性マップの誤差を回避し、顕著性マップから無視されたいくつかの物体を効果的に訓練することができます。（表 3 では、顕著性マップの誤差を克服するための提案された戦略の有効性を報告しています。）

3.3. 共同訓練手順

顕著性マップと画像レベルのラベルを使用して、EPS の全体的な訓練目標は、顕著性損失 \mathcal{L}_{sal} と分類損失 \mathcal{L}_{cls} の 2 つの部分で構成されています。まず、顕著性損失 \mathcal{L}_{sal} は、実際の顕著性マップ M_s と推定された顕著性マップ \hat{M}_s の間の平均ピクセルレベルの距離を測定することによって定式化されます。

$$\mathcal{L}_{sal} = \frac{1}{H \cdot W} \|M_s - \hat{M}_s\|^2, \quad (3)$$

ここで、 M_s は DUTS データセット [39] で訓練された市販の顕著性検出モデル PFAN [51] から取得されます。我々の方法は、顕著性検出モデルに関係なく、すべての以前の技術を一貫して上回ることに注意してください。

次に、分類損失は、画像レベルのラベル y とその予測 $\hat{y} \in \mathbb{R}^C$ の間のマルチラベルソフトマー

ジンの損失によって計算されます。これは、各ターゲットクラスのローカライゼーションマップに対するグローバル平均プーリングの結果です。

$$\mathcal{L}_{cls} = -\frac{1}{C} \sum_{i=1}^C y_i \log \sigma(\hat{y}_i) + (1 - y_i) \log (1 - \sigma(\hat{y}_i)), \quad (4)$$

ここで、 $\sigma(\cdot)$ はシグモイド関数です。最後に、総訓練損失はマルチラベル分類損失と顕著性損失の合計です、すなわち、 $\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{sal}$ 。図 2 に示すように、 \mathcal{L}_{sal} は、ターゲットオブジェクトと背景を含む $C + 1$ クラスのパラメータを更新するのに関与しています。一方、 \mathcal{L}_{cls} は背景クラスを除く C クラスのラベル予測のみを評価し、 \mathcal{L}_{cls} からの勾配は背景クラスに流れません。しかし、背景クラスの予測は、分類器のトレーニングを監督するため、 \mathcal{L}_{cls} によって暗黙的に影響を受ける可能性があります。

4. 実験設定

データセット。我々は、PASCAL VOC 2012 [12] と MS COCO 2014 [30] という 2 つの人気のあるベンチマークデータセットで実証研究を行います。PASCAL VOC 2012 は 21 クラス（すなわち、20 のオブジェクトと背景）で構成されており、トレーニング、検証、テストセットにはそれぞれ 1,464、1,449、1,456 枚の画像が含まれています。セマンティックセグメンテーションの一般的な手法に従い、10,582 枚の画像を含む拡張トレーニングセットを使用します [17]。次に、COCO 2014 は背景を含む 81 クラスで構成されており、トレーニングと検証にはそれぞれ 82,081 枚と 40,137 枚の画像が含まれています。ターゲットクラスがない画像は [9] で行われたように除外されます。一部のオブジェクトのグラウンドトゥルースセグメンテーションラベルが重なっているため、同じ COCO データセットで重なりの問題を解決する COCO-Stuff [4] からのグラウンドトゥルースセグメンテーションラベルを採用します。

評価プロトコル。我々の手法を PASCAL VOC 2012 の検証セットとテストセット、および COCO 2014

の検証セットで検証します。PASCAL VOC 2012 のテストセットでの評価結果は、公式の PASCAL VOC 評価サーバーから取得されます。また、セグメンテーションモデルの精度を測定するために平均交差オーバーユニオン (mIoU) を採用します。

実装の詳細。我々の手法のバックボーンネットワークとして ResNet38 [45] を選択し、出力ストライドは 8 です。すべてのバックボーンモデルは ImageNet [11] で事前トレーニングされています。バッチサイズ 8 の SGD オプティマイザーを使用します。我々の手法は学習率 0.01 (最後の畳み込み層は 0.1) で 20k イテレーションまでトレーニングされます。データ拡張のために、ランダムスケーリング、ランダムフリッピング、およびランダムクロップを 448×448 に使用します。セグメンテーションネットワークには、DeepLab-LargeFOV (V1) [7] と DeepLab-ASPP (V2) [8]、およびそのバックボーンネットワークとして VGG16 と ResNet101 を採用します。具体的には、VGG16 ベースの DeepLab-V1 と DeepLab-V2、ResNet101 ベースの DeepLab-V1 と DeepLab-V2 の 4 つのセグメンテーションネットワークを使用します。詳細な設定は補足資料にあります。

5. 実験結果

5.1. 境界と共起の処理

境界不一致問題。擬似マスクの境界を検証するために、最先端の手法 [32, 41, 52] と境界の品質を比較します。PASCAL VOC 2011 の境界アノテーションと境界ベンチマークを提供する SBD [17] を利用します。[32] で行われたように、擬似マスクのエッジをラプラシアンエッジ検出器から計算することにより、クラス非依存の方法で境界の品質を評価します。次に、予測された境界とグラウンドトゥルース境界を比較して、リコール、精度、および F1 スコアを測定することにより、境界の品質を評価します。表 1 は、我々の手法がすべての 3 つの指標で他の手法を大幅に上回っていることを報告

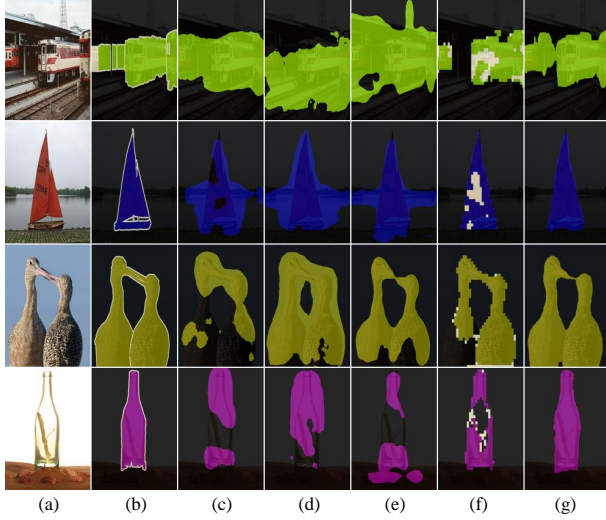


図 4. PASCAL VOC 2012 における擬似マスクの定性的比較。(a) 入力画像、(b) 正解データ、(c) CAM、(d) SEAM、(e) ICD、(f) SGAN、(g) 我々の EPS。

Method	Recall (%)	Precision (%)	F1-score (%)
CAM [52] _{CVPR'16}	22.3	35.8	27.5
SEAM [41] _{CVPR'20}	40.2	45.0	42.5
BES [32] _{ECCV'20}	45.5	46.4	45.9
Our EPS	60.0	73.1	65.9

表 1. SBD trainval セットで評価された境界精度。BES の結果は、[32] で提案された境界予測ネットワークから測定されたことに注意してください。

しています。図 4 の定性的な例は、我々の手法が他のすべての手法よりも正確な境界を捉えることができることを示しています。

共起問題。いくつかの研究 [20, 25, 28, 35] で議論されているように、PASCAL VOC 2012 では、いくつかの背景クラスがターゲットオブジェクトと頻繁に共起することを観察します。PASCAL-CONTEXT データセット [33] を使用して、共起オブジェクトの頻度を定量的に分析します。このデータセットは、シーン全体のピクセルレベルのアノテーションを提供します (例: 水と鉄道)。3 つの共起ペアを選択します; ボートと水、列車と鉄道、および 列車とプラットフォーム。ターゲットクラスの IoU とターゲットクラスとその共起クラスの間の 混同率を比較します。混同率は、共起クラス

方法	ボート w/ 水	列車 w/ 鉄道	列車 w/ プラットフォーム
CAM [52] _{CVPR'16}	0.74 (33.1)	0.11 (52.9)	0.09 (49.6)
SEAM [41] _{CVPR'20}	1.13 (30.7)	0.24 (48.6)	0.20 (45.5)
ICD [13] _{CVPR'20}	0.47 (41.4)	0.11 (56.7)	0.09 (49.2)
SGAN [47] _{ACCESS'20}	0.10 (42.3)	0.02 (48.8)	0.01 (36.3)
我々の EPS	0.10 (55.0)	0.02 (78.1)	0.01 (73.0)

表 2. 共起問題を扱う代表的な既存の方法との比較。各エントリは青で示された $m_{k,c}$ (低いほど良い) と括弧内の IoU (高いほど良い) です。

	Baseline	Na"ive	Pre-defined	Our adaptive
mIoU	66.1	66.5	67.9	69.4

表 3. マップ選択戦略の効果。異なるマップ選択戦略を使用した疑似マスクの精度は、PASCAL VOC 2012 トレインセ

ットで評価されます。
がターゲットクラスとして誤って予測される程度を測定します。混同率 $m_{k,c}$ は、 $m_{k,c} = FP_{k,c} / TP_c$ で計算され、 $FP_{k,c}$ は共起クラス k のターゲットクラス c として誤分類されたピクセル数であり、 TP_c はターゲットクラス c の真陽性ピクセル数です。共起問題に関する詳細な分析は補足資料にあります。

Table 2 は、EPS が他の手法よりも一貫して低い混同行列を示すことを報告しています。SGAN [47] は我々の手法と非常に似た混同行列を持っていますが、我々の手法は IoU の観点でターゲットクラスをより正確に捉えています。興味深いことに、SEAM は高い混同行列を示し、CAM よりも悪い結果を示しています。これは、SEAM [41] が自己教師あり学習を適用してターゲットオブジェクトの全範囲をカバーすることを学習するためであり、ターゲットオブジェクトの一致するピクセルに簡単に騙されるからです。一方、CAM はターゲットオブジェクトの最も識別的な領域のみを捉え、あまり識別的でない部分、e.g.、一致するクラスをカバーしません。この現象は Figure 4 でも観察できます。

Method	w/o refinement	w/ CRF [26]	w/ AffinityNet [26]
CAM [52] _{CVPR'16}	48.0	-	58.1
SEAM [41] _{CVPR'20}	55.4	56.8	63.6
ICD [32] _{CVPR'20*}	59.9	62.2	-
SGAN [47] _{ACCESS'20*}	62.8	-	-
Our EPS	69.4	71.4	71.6

表 4. PASCAL VOC 2012 のトレインセットで評価された擬似マスクの精度 (mIoU)。* は低信頼度のピクセルが無視されることを示します。他の方法はすべてのピクセルを評価に使用します。

5.2. マップ選択戦略の効果

我々は、サリエンスマップの誤差を軽減するためのマップ選択戦略の有効性を評価します。我々は、マップ選択モジュールを使用しないベースラインと 3 つの異なるマップ選択戦略を比較します。単純な戦略として、前景マップはすべてのオブジェクトローカリゼーションマップの和集合であり、背景マップは背景クラスのローカリゼーションマップに等しい (i.e.、単純戦略)。次に、以下の例外を除いて単純戦略に従います。いくつかの事前に決定されたクラス (e.g.、ソファ、椅子、ダイニングテーブル) のローカリゼーションマップは背景マップに割り当てられます (i.e.、事前定義クラス戦略)。最後に、提案された選択方法は、ローカリゼーションマップとサリエンスマップの重なり率を利用し、Section 3.2 で説明されているように (i.e.、我々の適応戦略)。

Table 3 は、我々の適応戦略がサリエンスマップの体系的なバイアスを効果的に処理できることを示しています。単純戦略は、ローカリゼーションマップから推定されたサリエンスマップを生成する際にバイアスを考慮しないことを意味します。この場合、擬似マスクの性能は特にソファ、椅子、ダイニングテーブルクラスで低下します。事前定義クラスを使用することによる性能は、サリエンスマップで欠落しているクラスを無視することでバイアスを軽減できることを示しています。

しかし、人間の観察者による手動選択が必要であるため、実用的ではなく、画像ごとに最適な決定を下すことはできません。一方、我々の適応戦略はバイアスを自動的に処理し、与えられたサリエンスマップに対してより効果的な決定を下します。

5.3. 最先端技術との比較

擬似マスクの精度。我々は、異なるスケールの画像からの予測結果を集約することでマルチスケール推論を採用し、これは [2, 41] で利用される一般的な手法です。その後、我々は、トレインセットでの擬似マスクの精度を、ベースラインの CAM [52] および 3 つの最先端手法、i.e.、SEAM [41]、ICD [13]、SGAN [47] と比較して評価します。ここで、トレインセットでの擬似マスクの精度を測定することは、WSSS における一般的なプロトコルです。なぜなら、トレインセットの擬似マスクはセグメンテーションモデルを監督するために使用されるからです。Table 4 は擬似マスクの精度を要約し、我々の手法がすべての既存手法を大きな差 (i.e.、7-21% のギャップ) で明確に上回っていることを示しています。Figure 4 は擬似マスクの定性的な例を視覚化し、我々の手法がオブジェクトの境界を著しく改善し、擬似マスクの品質において 3 つの最先端手法を大幅に上回っていることを確認しています。我々の手法はオブジェクトの正確な境界を捉えることができ (2 行目)、したがって自然にオブジェクトの全範囲をカバーし (3 行目)、一致するピクセルを軽減します (1 行目)。我々の手法のさらなる例と失敗例は補足資料で提供されています。

セグメンテーションマップの精度。以前の手法 [2, 13, 41] は擬似マスクを生成し、CRF 後処理アルゴリズム [26] またはアフィニティネットワーク [2] でそれらを洗練します。一方、Table 4 に示されているように、我々が生成した擬似マスクは十分に正確であるため、擬似マスクの追加の洗練なしにセグメンテーションネットワークを訓練します。我々は、Pascal VOC 2012 データセットの 4 つのセグメンテーションネットワークで我々の手法を他



図 5. PASCAL VOC 2012 におけるセグメンテーション結果の定性的な例。(a) 入力画像、(b) グラウンドトゥルース、(c) 我々の EPS。

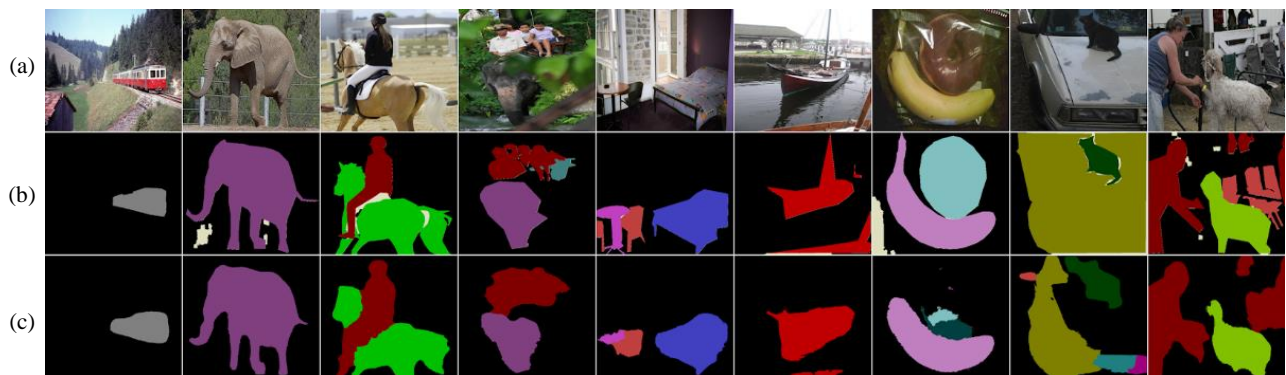


図 6. MS COCO 2014 におけるセグメンテーション結果の定性的な例。(a) 入力画像、(b) グラウンドトゥルース、(c) 我々の EPS。

の手法と広範に評価し、正確に比較します。

私たちの手法は、セグメンテーションネットワークに関係なく、他の手法よりも著しく優れた性能を発揮します。表 5 は、同じ VGG16 バックボーンを使用した場合、私たちの手法が他の手法よりも正確であることを報告しています。さらに、VGG16 での私たちの結果は、より強力なバックボーン (i.e. 表 6 の ResNet101) に基づく他の既存の手法と比較しても同等かそれ以上です。私たちの手法は、既存の手法に対して明確な改善を示しています。最後に、表 6 は、ResNet101 ベースの DeepLab-V1 とサリエンスマップを使用した私たちの手法が、PASCAL VOC 2012 データセットで新しい最先端の性能 (検証セットで 71.0、テストセットで 71.8) を達成したことを示しています。既存の最先端モデルによって達成された向上は約 1% でしたが、私たちの手法は前回の最高記録

よりも 3% 以上の向上を達成しています。図 5 は、PASCAL VOC 2012 での私たちのセグメンテーション結果の定性的な例を視覚化しています。これらの結果は、私たちの手法が正確な境界を提供し、共起問題を成功裏に解決することを確認しています。

表 7 では、COCO 2014 データセットで私たちの手法をさらに評価しています。VGG16 ベースの DeepLab-V2 をセグメンテーションネットワークとして使用し、COCO データセットでの最先端の WSSS モデルである SGAN [47] と比較します。私たちの手法は、検証セットで 35.7 mIoU を達成し、SGAN [47] よりも 1.9% 高いです。結果として、COCO 2014 データセットで新しい最先端の精度を達成しました。両方のデータセットでの既存の最先端を超えるこれらの優れた性能は、私たちの手法の有効性を確認しています。ローカライゼー

Method	Seg.	Sup.	val	test
SEC [25] _{ECCV'16}	V1	I.	50.7	51.7
AffinityNet [2] _{CVPR'18}	V1	I.	58.4	60.5
ICD [13] _{CVPR'20}	V1	I.	61.2	60.9
BES [32] _{ECCV'20}	V1	I.	60.1	61.1
GAIN [28] _{CVPR'18}	V1	I.+S.	55.3	56.8
MCOF [40] _{CVPR'18}	V1	I.+S.	56.2	57.6
SSNet [48] _{ICCV'19}	V1	I.+S.	57.1	58.6
DSRG [20] _{CVPR'18}	V2	I.+S.	59.0	60.4
SeeNet [19] _{NeurIPS'18}	V1	I.+S.	61.1	60.7
MDC [44] _{CVPR'18}	V1	I.+S.	60.4	60.8
FickleNet [27] _{CVPR'18}	V2	I.+S.	61.2	61.9
OAA [21] _{ICCV'19}	V1	I.+S.	63.1	62.8
ICD [13] _{CVPR'20}	V1	I.+S.	64.0	63.9
Multi-Est. [14] _{ECCV'20}	V1	I.+S.	64.6	64.2
Split. & Merge. [50] _{ECCV'20}	V2	I.+S.	63.7	64.5
SGAN [47] _{ACCESS'20}	V2	I.+S.	64.2	65.0
Our EPS	V1	I.+S.	66.6	67.9
	V2	I.+S.	67.0	67.3

表 5. PASCAL VOC 2012 におけるセグメンテーション結果 (mIoU)。すべての結果は VGG16 に基づいています。すべての実験で最高のスコアは太字で示されています。

Method	Seg.	Sup.	val	test
ICD [13] _{CVPR'20}	V1	I.	64.1	64.3
SC-CAM [5] _{CVPR'20}	V1	I.	66.1	65.9
BES [32] _{ECCV'20}	V2	I.	65.7	66.6
LIID [31] _{TPAMI'20}	V2	I.	66.5	67.5
MCOF [40] _{CVPR'18}	V1	I.+S.	60.3	61.2
SeeNet [19] _{NeurIPS'18}	V1	I.+S.	63.1	62.8
DSRG [20] _{CVPR'18}	V2	I.+S.	61.4	63.2
FickleNet [27] _{CVPR'18}	V2	I.+S.	64.9	65.3
OAA [21] _{ICCV'19}	V1	I.+S.	65.2	66.4
Multi-Est. [14] _{ECCV'19}	V1	I.+S.	67.2	66.7
MCIS [38] _{ECCV'20}	V1	I.+S.	66.2	66.9
SGAN [47] _{ACCESS'20}	V2	I.+S.	67.1	67.2
ICD [13] _{CVPR'20}	V1	I.+S.	67.8	68.0
Our EPS	V1	I.+S.	71.0	71.8
	V2	I.+S.	70.9	70.8

表 6. PASCAL VOC 2012 におけるセグメンテーション結果 (mIoU)。すべての結果は ResNet101 に基づいています。

Method	Seg.	Sup.	val
SEC [25] _{ECCV'16}	V1	I.	22.4
DSRG [20] _{CVPR'18}	V2	I.+S.	26.0
ADL [9] _{TPAMI'20}	V1	I.+S.	30.8
SGAN [47] _{ACCESS'20}	V2	I.+S.	33.6
Our EPS	V2	I.+S.	35.7

表 7. MS COCO 2014 におけるセグメンテーション結果 (mIoU)。すべての結果は VGG16 に基づいています。

ションマップとサリエンスマップの両方を完全に活用することで、ターゲットオブジェクトの全体を正確に捉え、既存モデルの欠点を補います。図 6 は、COCO 2014 データセットでのセグメンテーション結果の定性的な例を示しています。私たちの手法は、いくつかのオブジェクトが遮蔽されずに現れる場合にうまく機能しますが、多くの小さなオブジェクトを扱うのには効果が低いです。私たちの手法のさらなる例と失敗例は、補足資料で提供されています。

サリエンス検出モデルの効果。異なるサリエンス検出モデルの効果を調査するために、3 つのサリエンスモデルを採用しました。PFAN [51] (私たちのデフォルト)、OAA [21] と ICD [13] で使用される DSS [18]、および USPS [34] (i.e.、教師なし検出モデル) です。Resnet101 ベースの DeepLab-V1 でのセグメンテーション結果 (mIoU) は、PFAN で 71.0/71.8、DSS で 70.0/70.1、USPS で 68.8/69.9 (検証セットとテストセット) です。これらのスコアは、3 つの異なるサリエンスモデルのいずれかを使用しても、私たちの EPS が表 6 の他のすべての手法よりも正確であることを支持しています。特に、教師なしサリエンスモデルを使用した私たちの EPS は、教師ありサリエンスモデルを使用した既存のすべての手法を上回っています。

6. 結論

私たちは、新しい弱教師ありセグメンテーションフレームワーク、すなわち明示的擬似ピクセル監督 (EPS) を提案します。ローカライゼーシ

ョンマップとサリエンシーマップの補完的な関係に動機付けられ、私たちの EPS は、サリエンシーマップとローカライゼーションマップを組み合わせた擬似ピクセルフィールドバックから学習します。私たちの共同トレーニングスキームのおかげで、両側のノイズや欠落情報をうまく補完します。その結果、私たちの EPS は正確なオブジェクト境界を捉え、非ターゲットオブジェクトの共起ピクセルを排除し、擬似マスクの品質を著しく向上させます。広範な評価とさまざまなケーススタディは、私たちの EPS の有効性と、PASCAL VOC 2012 および MS COCO 2014 データセットの両方での WSSS の新しい最先端精度を示しています。

謝辞。 Duhyeon Bang と Junsuk Choe にフィールドバックを感謝します。この研究は、MSIP (NRF-2019R1A2C2006123、2020R1A4A1016619) によって資金提供された NRF 韓国を通じた基礎科学研究プログラム、MSIT (2020-0-01361、人工知能大学院プログラム (YONSEI UNIVERSITY)) によって資金提供された IITP 助成金、および韓国政府によって資金提供された韓国医療機器開発基金助成金 (プロジェクト番号: 202011D06) によって支援されました。

参考文献

- [1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2209–2218, 2019. [3](#)
- [2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4981–4990, 2018. [3](#), [9](#), [11](#)
- [3] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4253–4262, 2020. [3](#)
- [4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1209–1218, 2018. [7](#)
- [5] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8991–9000, 2020. [3](#), [11](#)
- [6] Arslan Chaudhry, Puneet K. Dokania, and Philip H. S. Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. In Proceedings of the British Machine Vision Conference, 2017. [3](#)
- [7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In International Conference on Learning Representations, 2015. [7](#)
- [8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4):834–848, 2017. [7](#)
- [9] Junsuk Choe, Seungho Lee, and Hyunjung Shim. Attention-based dropout layer for weakly supervised single object localization and semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020. [2](#), [7](#), [11](#)
- [10] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3133–3142, 2020. [2](#)
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255. IEEE, 2009. [7](#)
- [12] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision, 111(1):98–136, 2015. [7](#)

- [13] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4283–4292, 2020. [2](#), [4](#), [6](#), [8](#), [9](#), [11](#)
- [14] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Employing multi-estimations for weakly-supervised semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, 2020. [3](#), [11](#)
- [15] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. Cian: Cross-image affinity net for weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10762–10769, 2020. [3](#)
- [16] Ruochen Fan, Qibin Hou, Ming-Ming Cheng, Gang Yu, Ralph R Martin, and Shi-Min Hu. Associating inter-image salient instances for weakly supervised semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 367–383, 2018. [3](#)
- [17] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998. IEEE, 2011. [7](#)
- [18] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3203–3212, 2017. [3](#), [5](#), [11](#)
- [19] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *Advances in Neural Information Processing Systems*, pages 549–559, 2018. [11](#)
- [20] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7014–7023, 2018. [3](#), [8](#), [11](#)
- [21] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. Integral object mining via online attention accumulation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2070–2079, 2019. [2](#), [4](#), [6](#), [11](#)
- [22] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 876–885, 2017. [1](#)
- [23] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Two-phase learning for weakly supervised object localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3534–3543, 2017. [2](#)
- [24] Alexander Kolesnikov and Christoph Lampert. Improving weakly-supervised object localization by micro-annotation. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference*, pages 92.1–92.12. BMVA Press, September 2016. [2](#)
- [25] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 695–711. Springer, 2016. [2](#), [3](#), [8](#), [11](#)
- [26] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in Neural Information Processing Systems*, pages 109–117, 2011. [9](#)
- [27] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5267–5276, 2019. [2](#), [3](#), [4](#), [11](#)
- [28] Kunpeng Li, Ziyang Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9215–9223, 2018. [2](#), [3](#), [4](#), [8](#), [11](#)
- [29] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3159–3167, 2016. [1](#)

- [30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014. [7](#)
- [31] Yun Liu, Yu-Huan Wu, Pei-Song Wen, Yu-Jun Shi, Yu Qiu, and Ming-Ming Cheng. Leveraging instance-, image-and dataset-level information for weakly supervised instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. [3](#), [11](#)
- [32] Chen Liyi, Wu Weiwei, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In *Proceedings of the European Conference on Computer Vision*, 2020. [2](#), [7](#), [8](#), [9](#), [11](#)
- [33] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014. [8](#)
- [34] Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. Deepusps: Deep robust unsupervised saliency prediction via self-supervision. In *Advances in Neural Information Processing Systems*, pages 204–214, 2019. [5](#), [11](#)
- [35] Seong Joon Oh, Rodrigo Benenson, Anna Khoreva, Zeynep Akata, Mario Fritz, and Bernt Schiele. Exploiting saliency for object segmentation from image level labels. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5038–5047. IEEE, 2017. [2](#), [8](#)
- [36] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1796–1804, 2015. [1](#)
- [37] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1713–1721, 2015. [1](#)
- [38] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, 2020. [3](#), [11](#)
- [39] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 136–145, 2017. [3](#), [6](#)
- [40] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1354–1362, 2018. [3](#), [6](#), [11](#)
- [41] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020. [2](#), [4](#), [7](#), [8](#), [9](#)
- [42] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1568–1576, 2017. [3](#), [4](#)
- [43] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2314–2320, 2016. [3](#), [6](#)
- [44] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7268–7277, 2018. [3](#), [11](#)
- [45] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for

visual recognition. *Pattern Recognition*, 90:119–133, 2019. 7

- [46] Huaxin Xiao, Jiashi Feng, Yunchao Wei, Maojun Zhang, and Shuicheng Yan. Deep salient object detection with dense connections and distraction diagnosis. *IEEE Transactions on Multimedia*, 20(12):3239–3251, 2018. 3
- [47] Qi Yao and Xiaojin Gong. Saliency guided self-attention network for weakly and semi-supervised semantic segmentation. *IEEE Access*, 8:14413–14423, 2020. 2, 3, 6, 8, 9, 10, 11
- [48] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7223–7233, 2019. 3, 4, 11
- [49] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12765–12772. AAAI Press, 2020. 3
- [50] Tianyi Zhang, Guosheng Lin, Weide Liu, Jianfei Cai, and Alex Kot. Splitting vs. merging: Mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, 2020. 3, 11
- [51] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3085–3094, 2019. 2, 3, 5, 6, 11
- [52] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 1, 2, 7, 8, 9