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

本部 勇真[,a)]　山口 廉斗[,b)]　柳井 啓司[,c)]

概要

本研究では，50 億もの画像テキストペアを学習した，拡散モデルに基づくテキストからの画像生成モデルである Stable Diffusion を利用して，追加学習せずに高速に任意のテキストに対応した領域を抽出可能なセグメンテーション手法 StableSeg を提案し，その有効性を示す．

1. はじめに

近年，深層学習の発展によりセマンティックセグメンテーションの分野は大幅に性能が向上し，ここ数年では，大規模事前学習モデルを再利用して学習コストを削減するタスクや，あらゆるクラスに対応させるタスクが注目されている．その中でもゼロショットセグメンテーションでは，大規模事前学習モデルがセグメンテーションに特化していないため多くの研究でアノテーションデータを用いた追加学習を必要とする問題がある．そのため，依然として学習コストやアノテーションデータを作成するコストの削減はできていない．

それに対して，本研究では，50 億もの画像テキストペアを学習した，大規模な視覚言語拡散モデルである Stable Diffusion を使用することで追加学習することなくセグメンテーションを可能にする手法を提案し，コスト削減の実現とその有用性を示す．

本論文の主な貢献は次の通りである．

- 学習済 Stable Diffusion の cross-attention を利用することで，追加学習が一切不要なゼロショット領域分割を実現する StableSeg の提案.
- Cross-attention map に self-attention を組み合わせる self-attention refinement と，propmpt を工夫する方法によって精度向上することの提案.
- 疑似マスクで cross-attention map の重み付けを学習し，再度，疑似マスクを生成してセグメンテーションモデルを学習する StableSeg++の提案.

2. 関連研究

2.1 ゼロショット領域分割

ゼロショット領域分割 (Zero-shot Segmentation) では，テキストデータのみで分布外データに対する領域分割を実現することを目的としている．このタスクでは，学習時とテスト時のカテゴリには共通部分が存在しないため，テスト時の入力には未知のカテゴリのクエリ画像に対して，カテゴリ名のテキストが条件として与えられることで未知のカテゴリを領域分割する．ゼロショット領域分割では，事前に同じドメインの大規模データで事前学習したバックボーンの特徴と，テキスト埋め込みに基づく関係性を学習させることで，未知カテゴリの領域分割を学習データなしで実現する．

近年, 大規模な画像テキストペアデータの類似度をニューラルネットワークに学習させた大規模視覚言語モデルが注目され，このモデルを使い様々なタスクをゼロショットで解くという傾向が深層学習ではトレンドになっている．その中の 1 つである CLIP [1] は，最初に学習済みモデルが公開されたモデルであり，様々なタスクで使用されている．CLIP を用いたゼロショット領域分割もすでに提案されており，Zhou ら [2] の MaskCLIP という手法では，学習済みの CLIP を使用し，backbone で画像から特徴を抽出し，この特徴マップに対して，ターゲットテキスト特徴を重みにした畳み込み演算によって，ゼロショット学習でピクセル単位の分類を実現した手法である．本論文での提案手法 StableSeg と同じくテキストのみからあらゆるクラスに対して領域分割を可能としている手法である．さらに MaskCLIP では，MaskCLIP で生成したマスクを疑似マスクとして DeepLabV2 [3] を学習させることでより高性能な領域分割 MaskCLIP+を実現した．

他にも CLIP を用いた手法として Lüddecke らが提案した CLIPSeg [4] と呼ばれる手法があるが，こちらのゼロショット手法ではセグメンテーションモジュールを外部アノテーションデータを用いて学習を行う必要がある．それに対して，StableSeg では自己教師あり学習を採用することで外部データを一切使用せずにゼロショット領域分割を実現する．

2.2 拡散モデル

近年，拡散モデル (DM) が画像生成タスクで大きな成果を収めている．Ho ら [5] によって提案されたノイズ除去拡散確率モデル (DDPM) では，入力画像に連続的にガウシアンノイズノイズを与え，画像から各ステップのノイズをニューラルネットワークで推定することで元画像を復元させることを何回も繰り返し徐々にノイズを除去していくことによって画像を生成する手法であり，拡散モデルを急速に発展させたモデルである．

[1] 電気通信大学 大学院情報理工学研究科 情報学専攻
[a)] honbu-y@mm.inf.uec.ac.jp
[b)] yamaguchi-r@mm.inf.uec.ac.jp
[c)] yanai@cs.uec.ac.jp

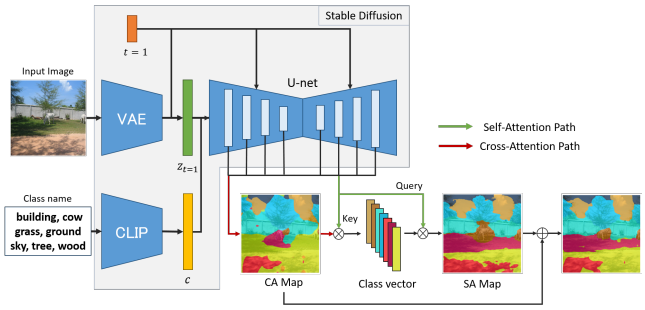

図 1 StableSeg のアーキテクチャ

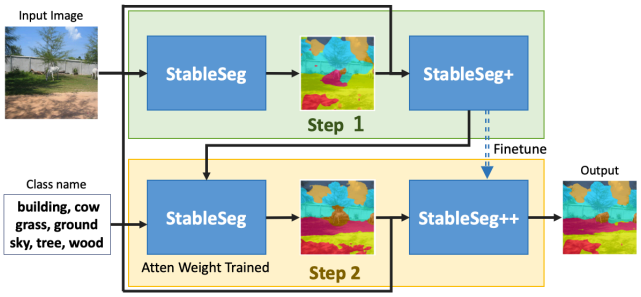

図 2 StableSeg++のアーキテクチャ図

特に近年，Stable Diffusion はテキスト入力に沿った高品質な画像を生成することができるモデルであるとして注目されている．Stable Diffusion は拡散モデルの一種である Latent Diffusion Model (LDM) [6] というモデルが使用されており，LDM では入力画像を Variational Autoencoder (VAE) [7] で潜在空間に圧縮したものに対してガウシアンノイズを付与し，様々な条件を加えることのできる U-Net アーキテクチャを使ってノイズを除去し，デコーダーを使い画像へと復元させるモデルである．特に Latent Diffusion Model の条件付けに CLIP [1] と呼ばれる大規模視覚言語モデルのテキストエンコーダーを使用してテキストで潜在空間に対して条件付けを行い，さらに LAION-5B [8] と呼ばれる 50 億枚の画像テキストペアデータセットで学習させたものを Stable Diffusion と呼ぶ．

Hertz ら [9] の提案した Prompt-to-Prompt と呼ばれる手法では，拡散モデルベースの画像生成モデルで使用されているアテンション層のクロスアテンションマップを利用して，生成される画像の空間レイアウトや形状を制御する手法を提案している．これによって，プロンプトのテキストのみを編集することで様々な画像編集を可能としている．本手法ではこの手法に触発され，精度高いアテンションマップの情報をセグメンテーションタスクに転用することができるのではないかと考えた．

Burgert らの Peekaboo [10] でも Stable Diffusion を使ったゼロショット領域分割手法を提案されているが，追加ネットワークの反復最適化による学習が必要なため 1 枚あたり 2 分程度の時間が掛かる欠点がある．一方，本手法で提案する StableSeg では拡散モデルの 1 ステップのみ，つまりノイズ推定のための UNet の評価を一度行うのみで処理時間は 1 秒未満であり，さらに追加学習や最適化は一切不要であるため，高速かつ低コストなゼロショット領域分割を実現している．

3. 手法

3.1 StableSeg

本手法では，Stable Diffusion の U-Net に使用されている Transformer に注目し，条件ベクトルが与えられる Cross Attention を使ってセグメンテーションを実現する．Stable Diffusion の Transformer Block は U-Net に複数存在し，各 Transformer で入力特徴同士の Self-Attention と条件ベクトル $\phi(C)$ との Cross-Attention が組み込まれている．Cross-Attention では時間 T のノイズベクトル z_T を U-Net(ϕ) で抽出した入力特徴 $\phi(z_T)$ に対して線形変換層 l_Q を使って Query とし，テキスト C の埋め込み $f(C)$ を 2 つの線形変換層 l_V, l_K を使用して Key，Value とする．そして Key と Query の内積を取りスケーリング ($\sqrt{d}$) した後に Softmax を計算したものが Attention Map として Value と積を取り次の層への特徴として利用されていく仕組みになっている．この計算は以下の式 1, 式 2 のように表される．そしてこの Attention Map は与えられたテキストのトークン毎に得ることができる．

$$Q = l_Q(\phi(z_T)), K = l_K(f(C)), V = l_V(f(C)) \quad (1)$$

$$\text{AttentionMap} = \text{Softmax}(\frac{QK^T}{\sqrt{d}}) \quad (2)$$

提案する手法では Stable Diffusion で使用されるすべての Transformer Block から Attention Map を抽出し Cross-Attention の確率マップ (Cross-Attention Probability Map, CAPM) として使用する．

さらに，Self-Attention 層に使用される Query と Key と Cross-Attention で生成されるクラスマップ (CA Map) を使ってセグメンテーションマスクを洗練する Self-Attention Refinement (SAR) を提案する．まず最初に，図 1 の赤色の矢印で表される通り，CAPM を argmax によってクラスマップ (CA Map) にする．次に図 1 の緑色の矢印に注目する．この部分では，すべての Self-Attention 層で利用される Key を特徴マップに変換し，各クラスマップの領域でクラスごとの平均ベクトルを計算し，画像内のそのクラスを表現する代表ベクトル (class vector) を生成する．それを新たな Key として式 2 と同様に Self-Attention の Query 特徴を使用して，クラスごとに Attention Map を計算し Self-Attention のクラスマップ (SA Map) の元となる確率マップ (Self-Attention Probability Map, SAPM) を計算する．そして最後に CAPM と SAPM を合計し最終クラスマップにすることで最終的なセグメンテーションマスクが完成する．このモデルを StableSeg とする．

また，ノイズ除去のステップは time embedding $t = 1$ の 1 ステップのみのアテンションマップを使用することで，1 枚当たり約 2 秒以下でセグメンテーションすることができる．StableSeg のアーキテクチャは図 1 の通りである．

3.2 StableSeg++

生成した疑似マスクで教師ありモデルを学習する MaskCLIP+ [2] と同様に，本研究でも疑似マスクの学

習による精度向上を図る．これを StableSeg+と呼ぶ．さらに本研究では，StableSeg+で生成した疑似マスクを使って StableSeg の推論時に使用される複数のアテンションマップの最適な重みを最適化することも行い，それで生成生成した疑似マスクでさらに教師ありモデルを学習する StableSeg++も提案する．

図 2 に示すように，最初に StableSeg を使って疑似マスクを生成し，この疑似マスクに Fully Connected CRF (DenseCRF) [11] を使用し，疑似マスクを修正する．その後 DeepLabV3+ [12] を使用してこの疑似マスクを教師データとして学習する．次に Step2 では，StableSeg+で推論したマスクを教師データとして StableSeg のアテンションマップの重みパラメータを最適化する．標準の StableSeg では U-Net の各レイヤから抽出したアテンションマップを均一重み [*1]で平均していたが，DeepLabV3+が出力した疑似マスクを擬似正解データとして，これに近づくようにすべてのアテンションマップに関して重みを学習することで，さらなる精度向上が期待できる．さらにアテンションマップの重みを学習した StableSeg で生成したマスクを用いて StableSeg+の DeepLabV3+を再学習する，

以上のように，重み推定を挟んで，StableSeg, DeepLabV3+を繰り返すことで外部のアノテーションデータを使用することなくセグメンテーションモデルの精度を向上させることができる．

4. データセット

提案手法ではあらゆるテキストに対してセグメンテーションマスクを生成することができるため，様々なデータセットで実験する．一般物体 20 クラスで構成されている Pascal VOC (PAS-20) [13]，一般物体 60 クラスで構成されている Pascal Context (PC-59) [14], 100 クラスの食事クラスで構成されている UECFoodPix (FoodPix), 103 種類の食材で構成されている FoodSeg103 (FoodSeg) [15], 物体, 物体パーツ, もの (stuff) の 3 区分の，計 150 種類のクラスが画素単位でアノテーションされている ADE20K (A-150) [16]，50 都市の街路景観で撮影された 18 クラスの Cityscapes (City) [17] で実験を行った．なお，入力画像サイズは 512x512 に統一した．

5. 実験

StableSeg は 50 億テキスト画像ペアで学習済みの Stable Diffusion を使い，時間埋め込みは $t = 1$ を使用することで，Stable Diffusion における最後のノイズ除去過程を再現するようにした．$t = 1$ のみのノイズ除去過程で抽出される Attention を使用するため，画像一枚の処理時間は 2 秒以内で済んでいる．入力画像にはノイズを混ぜずに VAE で潜在空間に圧縮し U-net に入力することで元画像に近いアテンションマップを習得した．

StableSeg では入力画像を 512x512 とした場合，U-net の各層から縦横がそれぞれ 8, 16, 32, 64 ピクセル [*2]の self/cross アテンションマップが抽出される．SAPM は正確な CA Map を必要とするため，実験では指定がない限り，Cross-Attention にはスケールが 8, 16 のアテンションマップを平均したものを使用し，Self-Attention には，全スケールの Query, Key 特徴すべてを使用した．比較対象とするモデルには MaskCLIP [2] を使用し，それぞれのモデルへの入力には入力画像とその画像内に含まれるカテゴリ名のプロンプトを入力とした．

[*1] cross-attention に関しては 8, 16 スケールのみ利用．
[*2] 以下，スケール 8, 16, 32, 64 と呼ぶ．

表 1 各データセットでの定量評価

	PAS-20	PC-59	A-150	City	FoodPix	FoodSeg
MaskCLIP [2]	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**

表 2 SAR の違いによる各データセットでの定量評価

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
only Self	47.1	33.6	22.5	13.3	**65.4**	**50.0**

表 3 self/cross attention map において異なるスケール使用時の定量評価 (mIoU)

cross \ self	16	32	64	16	32	64
	PAS-20			PC-59		
16	49.1	50.1	50.3	35.1	35.9	**36.2**
32	50.0	51.1	51.3	34.4	35.2	35.5
64	50.3	51.3	**51.4**	33.2	33.9	33.9
	A-150			City		
16	23.0	23.5	**23.6**	14.9	15.0	**15.1**
32	22.5	22.9	22.9	13.2	13.4	13.3
64	21.8	22.1	22.0	13.2	13.3	13.1
	FoodPix			FoodSeg		
16	62.3	63.2	63.3	46.5	48.3	**49.1**
32	63.0	63.9	64.0	46.0	47.9	48.8
64	63.7	64.5	**64.5**	45.3	47.1	47.7

5.1 複数データセットにおける従来手法との比較

表 1 の定量評価の結果より，MaskCLIP [2] では，主に一般物体で構成されている PAS-20, PC-59, A-150 データセットで StableSeg と近い評価値が得られるのに対して，StableSeg では，MaskCLIP と比較して食事データセットで精度が高くなることが判明した．これは CLIP よりも Stable Diffusion が食事データなどの固有ドメインに対してもより豊富な表現力を持っていると考えられる．

また，図 3 の結果より StableSeg では MaskCLIP と比較するとノイズが少なく，様々なデータに対して正確に物体を捉えることができていることが分かる．

5.2 Self-Attention Refinement の効果検証

表 2 の結果より，Self-Attention を使ったセグメンテーションマスクの洗練は CAPM のみで作成したクラスマスクよりも改善することが分かった．要因としては，Self-Attention はピクセル同士の類似度を取るのに特化するように学習しているため，対象領域のベクトルを使うことで，より良い確率マップが得られたと考えられる．また，食事データに関しては Self-attention のみの場合が最も良い精度となった．

5.3 アテンションマップのスケールの違いによる性能評価

表 3 に Cross/Self Attention Map のスケールの違いによる実験結果を示す．スケールにはマップサイズとして 8, 16, 32, 64 が存在するが，意味的な領域を最も捉えることのできる 8 のスケールは必ず使用しているため，表では省略している．表中の記述は，例えば，(cross,self) = (64,32)

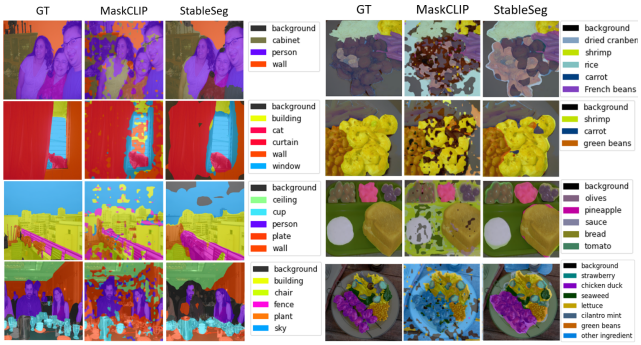

図 3 従来手法との比較例 (左：PC-59，右：FoodSeg)

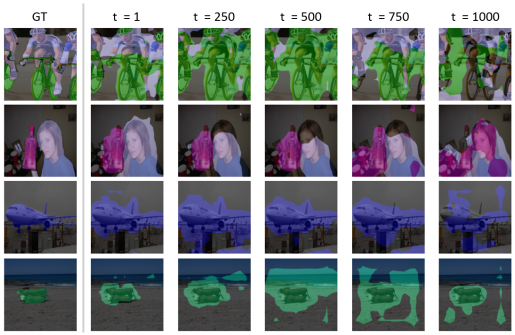

図 4 時間埋め込みを変化させたときの領域分割例

は，Cross Attention Map で使用するスケールが 64 以下のすべてのマップの合計かつ，Self Attention Map で使用するスケールが 32 以下のすべてのマップの合計という意味になる．

結果より，データセットごとに cross/self の最適なスケールが異なり，FoodPix では (cross,self) = (64,32) が最大値，PAS-20 では (64,64)，PC-59，A-150 と foodseg では (16,64)，City では (16,16) であることが分かった．提案手法では cross attention のマップをもとに Self-Attention マップを作るため，cross のスケール値に self が左右されていると考えられる．ここで cross の値に注目すると，A-150, City, PC-59, FoodSeg が 16 であることが分かる．cross のスケールが小さいほどより意味的な部分が抽出される一方で輪郭などの詳細な部分が抽出されないといった特徴があると考えられる．

5.4 異なる時間埋め込みによる性能検証

StableSeg の時間埋め込み t を変化させると図 4，表 4 のような結果となった．実際の Stable Diffusion では t が大きいほどガウスノイズに近づいた画像を復元する際に使用するため，生成する物体の位置を大まかに決めるアテンションマップが得られると考えられ，t が小さくなるにつれて実画像に近い画像を復元していくため，より実画像に存在する物体の形状に沿ったアテンションマップが推定されると考えられる．また，定量的にも $t = 1$ の時の mIoU は高くなることがわかる．これらの結果より，StableSeg では，元画像を VAE で潜在空間に圧縮した後ノイズを混ぜずに $t = 1$ で U-net に通すことで，より実画像に適したアテンションマップを得ることを可能にし，1 ステップのみでの高品質なセグメンテーションマップの推定を可能にしている．なお，複数の t の統合に関しては大規模な実験は行っておらず今後の課題である．

表 4 時間埋め込み (t) の違いによる定量評価

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

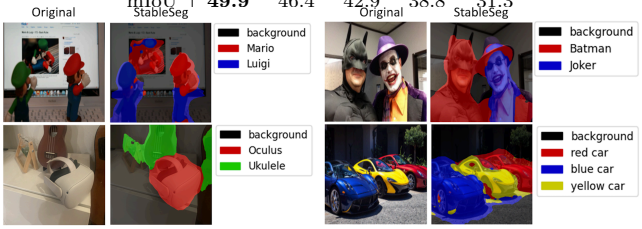

図 5 様々なクラスにおける推論結果例

表 5 StableSeg++の定量評価 (mIoU).

		PAS-20	PC-59
MaskCLIP [2]	init	44.7	37.9
	+	53.4	40.5
StableSeg ours	init	51.4	33.9
	+	55.6	34.6
	++	59.1	36.6

5.5 多様なクラスにおける定性分析

図 5 では，ユーザーが指定したクラスを StableSeg によってセグメンテーションした結果である．Stable Diffusion では，50 億画像テキストペアの関係性を学習しているため，あらゆる単語に対して対象としているアテンションマップが生成される．固有名詞の Mario, Batman, Oculus などや，red car などの形容詞を付与した場合にも条件に従った領域が分割されることが判明した．

5.6 StableSeg++の定量評価

この実験では StableSeg++と比較する手法として MaskCLIP+ [2] のを使用する．このとき MaskCLIP+は Annotation-Free 設定にし，MaskCLIP は画像中のクラスを指定して疑似マスクを作成し，DenseCRF [11] で処理したものを学習データとして DeepLabV3+をバックボーンとした MaskCLIP+を学習したものとする．表 5 に示す実験結果より，StableSeg の結果よりも，+, ++と順次，精度向上が得られることが示され，2 度の DeepLabV3+を使った疑似マスク生成は効果的であったと考えられる．MaskCLIP+との比較では，PAS-20 に関しては提案手法が良い結果，PC-59 に関しては上回ることはできなかった．これは，表 3 に示す (cross, self)=(64,64) で実験を行ったため，PC-59 の初期結果 (init) が MaskCLIP に対して大きく劣っていたことが理由として考えられる．

6. おわりに

StableSeg では，Stable Diffusion に使われる Attention Map と大規模データで学習した事前学習済みの知識を有効活用した手法を提案した．実験では，様々なデータセットによる評価を行い，あらゆるテキストをセグメンテーションできる可能性を見出した．食事データなどの固有ドメインにも汎用的な性質があることが判明し，様々なドメインに強い頑健性があると考えられる．また，追加の学習データを必要としないことでコスト削減も実現した．さらに StableSeg のアテンションマップの重みを擬似マスクで学習させたもので学習データを作り，StableSeg+を再学習させるモデルである StableSeg++を提案し，追加学習データを使用しない手法を実現するとともにセグメンテーション品質の向上させた．

参考文献

[1] A. Radford, J. Kim, C. Hallacy, Ramesh, G. A. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and Sutskever I. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[2] C. Zhou, C C. Loy, and B. Dai. Extract free dense labels from clip. In *Proc. of European Conference on Computer Vision (ECCV)*, 2022.

[3] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. In *Proc.of IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.

[4] T. Lüddecke and A S. Ecker. Image segmentation using text and image prompts. In *Proc.of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7086–7096, June 2022.

[5] J. Ho, A. Jain, and P. Abbeel. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[6] D. Lorenz P. Esser R. Rombach, A.Blattmann and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proc.of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

[7] P. Kingma and M. Welling. Auto-encoding variational bayes. In *Proc.of International Conference on Machine Learning (ICML)*, 2014.

[8] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[9] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[10] R. Burgert, K. Ranasinghe, X. Li, and M. S. Ryoo. Peekaboo: Text to image diffusion models are zero-shot segmentors. In *Proc. of arXiv:2211.13224*, 2022.

[11] P Krähenbühl and V Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Proc.of Advances in Neural Information Processing Systems*, Vol. 24. Curran Associates, Inc., 2011.

[12] L Chen, Y Zhu, G Papandreou, F Schroff, and H Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[13] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, Vol. 111, No. 1, pp. 98–136, January 2015.

[14] R. Mottaghi, X. Chen, X. Liu, N. Cho, S. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proc.of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[15] W. Xiongwei, F. Xin, L. Ying, L. Ee-Peng, H. Steven, and S. Qianru. A large-scale benchmark for food image segmentation. *arXiv preprint arXiv:2105.05409*, 2021.

[16] B. Zhou, H. Zhao, X. Puig, S. Fidler, and A. Barriuso. Scene parsing through ade20k dataset. In *Proc.of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc.of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.