社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

Stable Diffusion によるゼロショット画像領域分割

本部　勇真†　　柳井　啓司†

E-mail: †honbu-y@mm.inf.uec.ac.jp, ††yanai@cs.uec.ac.jp

あらまし　近年研究されているセマンティックセグメンテーションでは，複数物体の領域分割によって自動運転など様々なものに活用することができると考えられている．しかし，学習の際にピクセルレベルのアノテーションを含む大量の画像が必要になり，コストがかかる問題がある．そこで本研究では，大量の画像テキストペアデータを学習した stable diffusion を活用することで，追加の学習を必要とせずに，あらゆるクラスのセグメンテーションマスクを推論するネットワークを提案する．

キーワード　Zero-shot Segmentation, Diffusion Model, Vision-Language Model

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．ピクセルレベルのアノテーションデータを使用せずにテキストのみでセグメンテーションモデルを学習させるタスクである弱教師あり学習法であったり，分布外データに関する学習データを使用することなく，未知データに対するセグメンテーションを実現するゼロショット学習などの学習法があるが，これらの手法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アノテーションデータを用いた学習を必要とする問題があり，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．

そこで本研究では，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．

- 追加学習を不要にすることで学習コスト及び，アノテーションデータ作成コストの削減．
- 大規模事前学習済みモデルを使用することによって，あらゆるクラスに対応したセグメンテーションモデルの実現

2. 関連研究

2.1 Zero-shot Segmentation

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目的としている．このタスクでは，学習時と検証時のカテゴリには共通部分が存在しないため，検証時の入力には未知のカテゴリのクエリ画像に対して，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. このネットワークでは，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の関係性を学習させることで，教師データで領域分割を可能とする手法である．

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 Zero-shot 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．Zero-shot Segmentation にも使用されており，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，Zero-shot でピクセル単位の分類を実現した手法である．本手法と同じくテキストのみからあらゆるクラスに対してセグメンテーションが可能になっている手法である．さらに MaskCLIP では，MaskCLIP で生成したマスクを疑似マスクとして DeepLabV3+ [3] を学習させることで高性能な Zero-shot Segmentation を実現した．

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 [4] によって提案されたノイズ除去拡散確率モデル (DDPM) では，図 1 のように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．

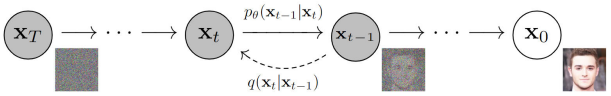

図 1: DDPM の有向グラフモデル.([4] から引用)

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) [5] 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) [6] 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 LDM 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 [7] 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．

現在では Stable Diffusion や Imagen [8] のような拡散モデルベースのテキストから画像を生成するモデルを起点に，テキストのみでの画像編集 [9]〜[12]，テキストからの動画生成 [13]〜[15]，物体検出 [16], セグメンテーション [17] など様々なタスクに拡散モデルが使用されてきている．

Hertz ら [11] 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，図 2 のような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．プロンプト内の単語の入れ替えを生成画像に反映する場合には，元画像のアテンションマップを注入し，ターゲットのアテンションマップをオーバーライドすることによって実現した．また，単語を追加する場合には，プロンプトの変更されない部分に対応するアテンションマップのみを注入することによって実現している．さらに，ある単語の意味を増幅/減衰させるために，対応するアテンションマップの重みを変更させる手法も提案している．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．

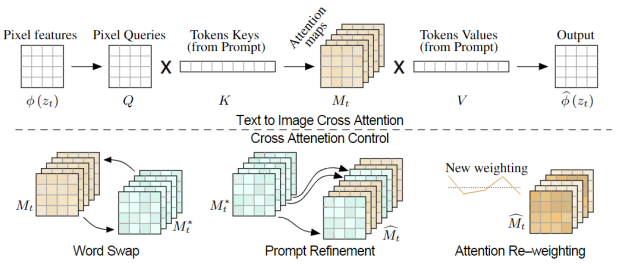

図 2: Hertz らの提案したネットワーク図 ([11] から引用)

Burgert らの Peekaboo [17] では，Stable Diffusion を使った Zero-shot セグメンテーション手法を提案している．まず，学習可能なアルファマスクをセグメンテーション画像とみなし，このマスクをニューラルネットワークによってモデル化された陰関数として表現する．そして，アルファ合成された画像とセグメンテーションされる画像領域に関連するテキストプロンプトに対して，dream loss と呼ばれるマッチングロスを使い，反復的に最適化する．最適化の結果，アルファマスクは最適なセグメンテーションマスクに収束する．このモデルでは，拡散モデルの再学習は一切行われず，陰関数表現を実現するニューラルネットワークのみを学習する．しかし，この手法では画像毎に最適化処理がされるため 1 枚当たり約 2 分ほど時間がかかってしまう問題が発生する．本手法で提案するモデルでは拡散モデルの 1 ステップのみ，かつ別のネットワークの最適化を行わないために，高速かつコスト削減を実現している．

3. 手　　法

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 や，図 2 の上段のように表される．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．

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

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

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 (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スク (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 3 の赤色の矢印で表される通り，CAPM を argmax によってクラスマスク (CA Map) にする．次に図 3 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，各クラスマスク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 のクラスマスク (SA Map) の元となる確率マップ (SAPM) を計算する．そして最後に CAPM と SAPM を合計しクラスマスク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=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 3 の通りである．

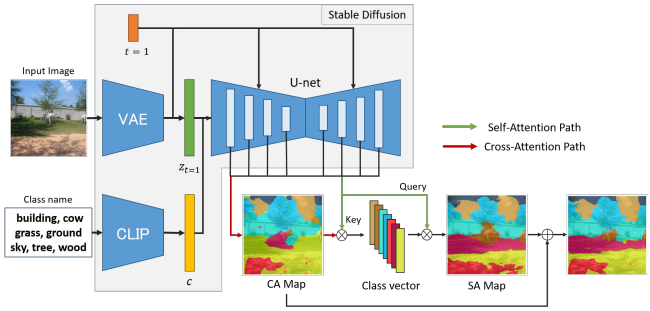

図 3: StableSeg のアーキテクチャ

また，Stable Diffusion では，プロンプトがトークンに分けられてトークンごとの別々のアテンションマップが生成されるため bedclothes などが bed と clothes に分けられてしまう問題がある．そのため StableSeg では，Prompt-engineering (p-eng) を使用している．まず，トークンごとにテキストエンコーダで特徴にし，合計することで 1 つの特徴としてクラスアテンションマップを生成した．さらに，セグメンテーションの対象は現実世界の画像であることから a photo of [prompt] とすることでより良いテキスト特徴を生成した．

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) [18]，一般物体 60 クラスで構成されている Pascal Context (PC-59) [19], 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) [20]，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) [21]，50 都市の街路景観で撮影された 18 クラスの Cityscapes (City) [22]，COCO2017 データセットに存在する 91 クラス 164K 枚の画像に対してアノテーションを施したデータセット COCO Stuff (Stuff) [23] で実験を行った．

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は混ぜずに 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．また，StableSeg では U-net の各層から 8, 16, 32, 64 スケールの 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，よりもクラス特定の場所が抽出される必要がある．そのため実験では指定がない限り，Cross-Attention にはスケールが 16 以下のすべて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，スケールが 64 以下になる Query, Key 特徴すべてを使用した．評価

表 1: 各データセットでの定量評価

	PAS-20	PC-59	A-150	Stuff	City	FoodPix	FoodSeg
StableSeg	**49.9**	37.6	23.3	28.6	14.7	**64.6**	**48.1**
MaksCLIP	47.6	**38.1**	**26.0**	**29.6**	**21.6**	33.2	37.0

表 2: SAR の違いによる各データセットでの定量評価

	PAS-20	PC-59	A-150	Stuff	City	FoodPix	FoodSeg
w/ SAR	**49.9**	**37.6**	**23.3**	**28.6**	**14.7**	**63.5**	**48.1**
w/o SAR	47.2	31.0	19.4	25.4	12.7	53.8	39.3

表 3: pronpt engineering(p-eng) の違いによる定量評価 (mIoU)

	PAS-20
w/ p-eng	**49.9**
w/o p-eng	45.6

表 4: self/cross attention map において異なるスケール使用時の定量評価 (mIoU)

	PAS-20			PC-59		
cross \ self	16	32	64	16	32	64
16	48.8	49.7	49.9	34.5	35.6	**36.0**
32	49.7	50.7	50.8	34.3	34.9	35.2
64	50.1	51.0	**51.4**	33.0	33.5	33.4
	FoodPix			FoodSeg		
16	62.9	63.4	63.5	45.9	47.4	**48.1**
32	63.5	64.1	64.2	45.4	46.9	47.7
64	64.2	**64.6**	64.5	44.7	46.0	46.5
	A-150			City		
16	22.9	**23.2**	**23.2**	**14.8**	**14.8**	14.7
32	22.3	22.6	22.6	13.1	13.2	13.0
64	21.6	21.8	21.7	13.1	13.0	12.9

するモデルには MaskCLIP 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プロンプトを入力とした．MaskCLIP と各データセットで比較した結果は表 1，Self-Attention Refinement (SAR), Prompt-engineering (p-eng) のアブレーションスタディは表 4 は U-net の異なる層の 16,32,64 スケールから cross/self attention 層でどの特徴を使うのかを実験した結果である．

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, Stuff 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．

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

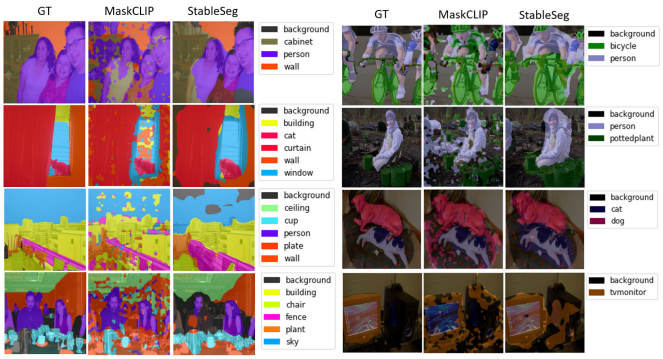

図 4: 一般物体データを使った従来手法との比較例

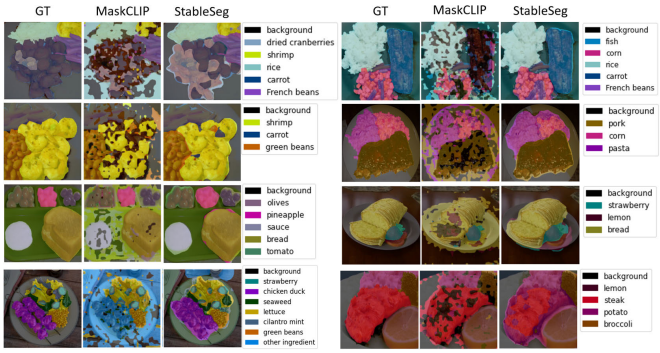

図 5: 食事データを使った従来手法との比較例

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．

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 4 では Cross/Self Attention Map のスケールの違いによる実験結果の表である．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．そのため，表には存在しない．また，(cross,self) = (64,32) であった場合は，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．図 6 のように，A-150, City, PC-59, FoodSeg の 3 つのデータセットでは 1 枚当たりの画像にクラスが複数含まれることが多く，個々の意味的な領域の判断が重要視されるからであると考えられる．一方で PAS-20，FoodPix では 1 枚当たりのクラス数が少ないため，cross で弱い意味的な部分を抽出しても精度が下がることがなく，かつ詳細な部分を抽出して精度を向上させることができていると考えられる．この考察より，画像 1 枚当たりのクラス数が多い画像では cross のスケールを小さい値に設定し，少ない場合は cross のスケールの値を小さくするのが最適であると考えられる．

また，画像 1 枚当たりのクラス数が大きいデータセット程に精度が下がる傾向があるが，1 枚当たりのクラス数が近いの FoodPix，PAS-20 を比較すると食事データセットである FoodPix のほうが大きく精度が高くなることが分かった．また，PAS-20 のクラス数の約 2 倍もある FoodSeg と比較すると競争力のある精度になっていることから，StableSeg では特定のドメインに強い頑健性があると考えられる．

5.4 MaskCLIP との比較

図 4 は従来手法の MaskCLIP [2] と比較した例になる．MaskCLIP では，物体の位置は捉えられているもののノイズが多く誤った部分が多いように見ることができる．一方で StableSeg では，MaskCLIP と比較するとノイズが少なく物体を捉えることができていることが分かる．しかし詳細な部分は MaskCLIP のほうが捉えることができているため，一般物体データにおいて各手法は，一長一短であると考えられる．図 5 の食事データセットにおいては，MaskCLIP と比較すると StableSeg のほうが高品質なマスクを生成していることが分かる．MaskCLIP では，一般物体データとは異なり，捉えることができないクラスが出現している．一方で StableSeg では様々なクラスに対して適切に対象領域を捉えることができており，さらに料理内の複数クラスに対しても正しい場所を推定していることが分かった．これは Stable Diffuison において CLIP [1] と比較すると，学習データの多様性が高いことから StableSeg は食事ドメインにも強い汎化性能を示していると考えられる．

5.5 self/cross/self+cross のアテンションマップの比較

図 7 のように，StableSeg では CAPM で作成したマスク (Cross) と SAPM で作成したマスク (Self) とそれらを足し合わせたセグメンテーションマスク (Cross+Self) が作成される．この 3 つの結果例を示している．結果より，Cross では対象領域が抽出できているが詳細な部分が抽出できていないことが多く，対象領域からはみ出していることや，誤った領域として認識していることが多い場合があることが分かる．Self では，一方で Cross の

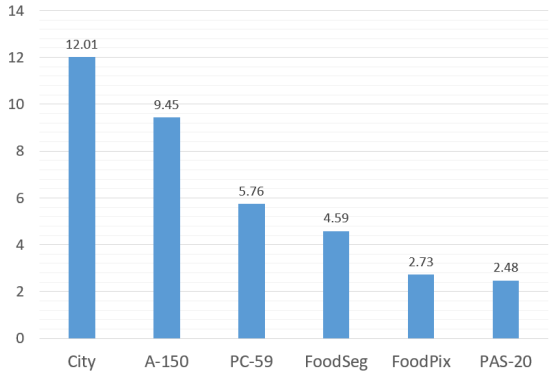

図 6: 各データセットでの 1 枚当たりのクラス数

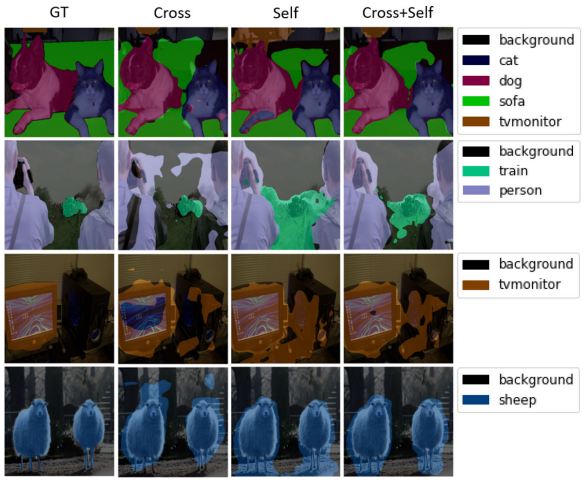

図 7: cross/self/cross+self のアテンションマップで作った領域分割の良い例

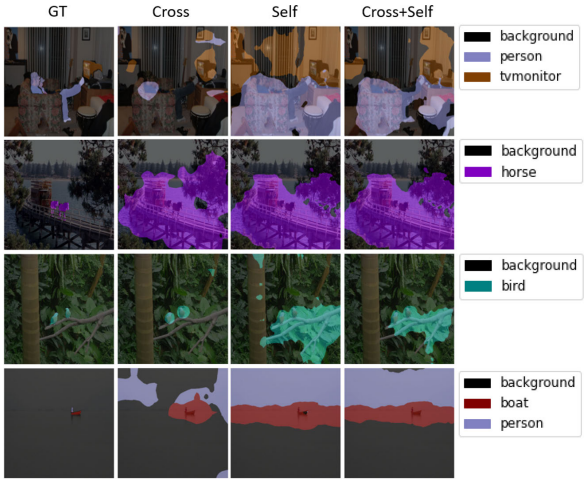

図 8: cross/self/cross+self のアテンションマップで作った領域分割の悪い例

領域特徴と似ている部分を抽出するため人の形や，動物の形が Cross と比較して輪郭に沿って抽出されている．しかし，Cross で誤った領域が含まれるため，対象領域とその外側のノイズの部分が抽出される問題がある．Cross+Self では，Cross の特徴である対象領域の高い部分の抽出と，Self の特徴である対象領域の輪郭を抽出する部分の抽出を足し合わせることで，より対象領域に近い領域を抽出することができると考えられる．

一方で，うまくいかない例としては図 8 のような例がある．Cross で正しい対象領域が抽出できない場合や，対象領域が小さすぎる場合には，Cross でノイズ部分の割合を大きく抽出してしまう問題があるため，正しい領域が領域分割できなくなってしまい，さらにノイズ部分の特徴が際立ってしまうため Self で大きく間違った領域が抽出されてしまうと考えられる．

5.6 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 9，表 5 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 の埋め込みと 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．

表 5: 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

5.7 Cross Attention のスケールの違いによる定性分析

図 10 では，U-net 内の異なるスケールで作られる Cross Attention でどのようなセグメンテーションマスクが生成されているのかを実験した結果を示したものである．64, 32, 16, 8 のすべてのスケールで抽出されるアテンションマップを Down, Middle, Up 毎に平均を取り可視化した．Down64, Down32, Up64, Up32 のようなスケールが大きい場合はより詳細な部分が抽出され，誤った部分も多いが対象領域の輪郭に沿った綺麗なセグメンテーションマスクが生成されていることが分かる．一方で Middle8

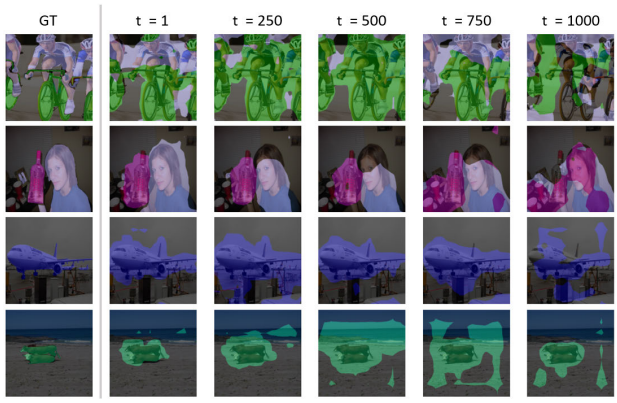

図 9: 時間埋め込みを変化させたときの領域分割例

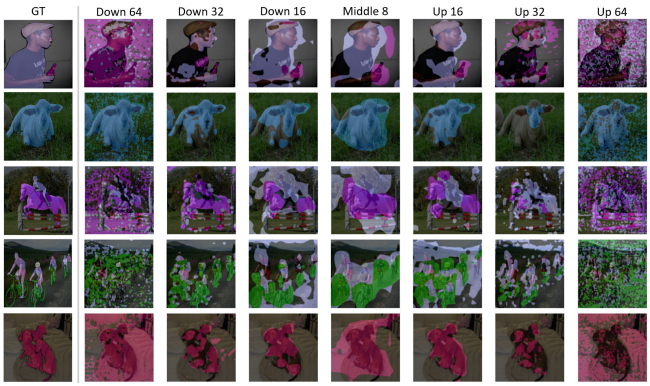

図 10: スケールの異なる Cross Attention Map を使用した領域分割例 (Down,Middle,Up は Unet の位置を示し，64, 32, 16, 8 はアテンションマップの一辺のサイズを示している)

のようにスケールが小さい場合は，アテンションマップを作成する Key，Query の特徴チャネル数が多く，表現力があるため対象領域の意味的な部分が抽出されると考えられる．これにより誤った部分が抽出されていることが減少するが，対象領域の形状を無視した領域を抽出してしまうことが分かった．

5.8 多様なクラスにおける定性分析

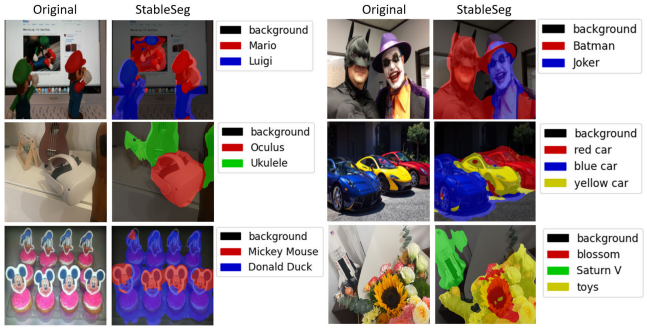

図 11: 様々なクラスにおける推論結果例

図 11 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 Mickey Mouse や Mario, Batman, Oculus, Saturn V などにも適切な領域が分割されていることが分かる．また，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．これにより結果のような特殊なクラスであっても StableSeg によってセグメンテーションが可能になると考えられる．

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．今後の課題としては，Stable Diffuison を使ったさらなる Zero-shot Segmentation の改良及び，その活用法をさらなる分析とともに取り組む必要がある．

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] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Proc.of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[5] 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.

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

[7] 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.

[8] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. Kamyar, B. Karagol, S. Sara, R. Gontijo, T. Salimans, J. Ho, D. J, and M. Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[9] O. Avrahami, D. Lischinski, and O. Fried. Blended diffusion for text-driven editing of natural images. In *Proc.of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18208–18218, June 2022.

[10] G. Kim, T. Kwon, and J. Chul. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proc.of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2426–2435, June 2022.

[11] 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.

[12] T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proc. of arXiv*, 2022.

[13] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P., B. Poole, M. Norouzi, D. J., and T. Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[14] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

[15] D. Zhou, W. Wang, H. Yan, W. Lv, Y. Zhu, and J. Feng. Magicvideo: Efficient video generation with latent diffusion model. *arXiv preprint arXiv:2211.11018*.

[16] . S. Chen, P. Sun, Y. Song, and P. Luo. Diffusiondet: Diffusion model for object detection. In *Proc. of arXiv:2211.09788*, 2022.

[17] 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.

[18] 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, 2015.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] H. Caesar, J. Uijlings, and V. Ferrari. COCO-Stuff: Thing and stuff classes in context. In *Proc.of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.