H-045

# SAM を用いたプレゼンテーションにおける画像強調システムの作成
# Creating an image enhancement system for presentations using SAM

小林 稜生 [1)]　鈴木 海友 [2)]　松澤 智史 [3)]
Ryosei Kobayashi　Kaiyu Suzuki　Tomofumi Matsuzawa

## 1. 研究背景と目的

近年，対話型 AI の ChatGPT をはじめとする″生成 AI″が飛躍的な発展を遂げており，日常生活やビジネスにおいても幅広く活用されている．その中にはプレゼンテーション資料を自動生成する生成 AI が存在し，題目やそれに付随する情報，ページ数等の入力によって自動生成できる．しかし，現状自動生成可能な資料は画像の強調などのアニメーションのない，単調なものとなっている．

プレゼンテーションを行う際，画像の一部を自動で強調することが可能となれば，アニメーションを人の手で作る必要はなく，資料作成の時間を削減できる．本研究では状況に合わせて画像の一部を強調するシステムの作成を目的とする．

## 2. 関連研究

Alexander Kirillov らの研究 [1]では画像セグメンテーションにおいてパノプティックセグメンテーションを提唱した．パノプティックセグメンテーションは従来の画像セグメンテーションである，セマンティックセグメンテーションとインスタンスセグメンテーションを組み合わせたものである．ここで，セマンティックセグメンテーションとは領域分類，つまり画像のピクセル一つひとつに対してラベル付けを行うものであり，インスタンスセグメンテーションとは物体の領域を特定，個体ごとに領域分割・物体の種類認識を行うものである．またパノプティックセグメンテーションのネットワークは入力層に近いレイヤはセマンティックとインスタンスで共通化しており，途中で分かれて個別の結果を出力，出力層に近いレイヤでそれらをまとめ，出力している．今回用いた SAM はパノプティックセグメンテーションを行っており，全てのピクセルにラベルを振り，可算な物体に対しては個別で認識することができる．

## 3. 用語

### 3.1 SAM（Segment Anything Model）

SAM [2]は画像のセグメンテーション，つまり物体の識別や物体同士の境界を特定可能なモデルである．任意のセグメンテーション用プロンプトを入力することで有効なセグメンテーションマスクを返す．ここでいうプロンプトとは，画像内の何をセグメンテーションするかを指定するもので，例えば，セグメンテーションの対象を特定する点やボックスの座標情報などがそれにあたる．プロンプトが曖昧で，複数の物体を対象としている場合を考慮して出力されるマスクは 3 つとなっている．また，出力されるマスクそれぞれに信頼度として SAM 独自のスコアが与えられる．スコアは高いほどセグメンテーションの対象としての信頼度が高いとされる．また，SAM は 1100 万枚の画像とそれに付随した 10 億以上のマスクからなるデータセットで訓練されており，新しい画像やタスクに Zero-Shot で対応している．よって学習データにない画像に対して物体認識が可能となっている．

### 3.2 Lang SAM（Language Segment-Anything）

Lang SAM [3]は SAM と物体検出モデル GroundingDINO [4]を統合したモデルである．プロンプトとしてテキストを入力することで画像の物体認識，セグメンテーションを行う． GroundingDINO はテキストプロンプトを入力し物体検出を行うモデルで，出力は対象の物体を囲うようなボックスとなっている．このボックスを SAM に与えることで物体認識を可能としている．Lang SAM も曖昧性を考慮して出力されるマスクは複数あり，それぞれに信頼度が与えられている．

## 4. 提案手法

### 4.1 提案手法の概要

本研究では，画像強調システムの根幹となる物体認識の精度を，SAM と LangSAM を個別で用いる場合からの向上を行う．

まず，SAM と Lang SAM でマスク生成を行う．それぞれで生成されたマスクから一致度の高いマスクを選択することで物体認識の精度向上を実現する．一致しているピクセルの割合をここでの一致度とする．（図 1）

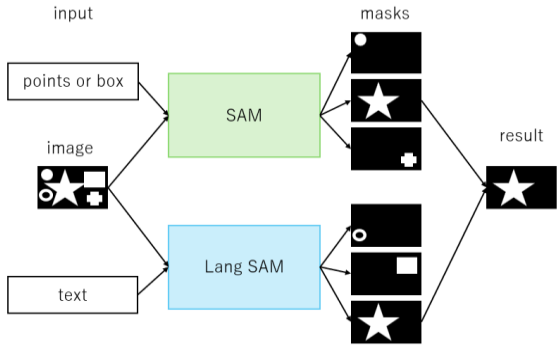


図 1：提案手法の流れ

### 4.2 マスクの選択方法

SAM と Lang SAM それぞれによって生成されたマスクを比較し，最も一致しているピクセルの割合が高いマスクの組み合わせを選択する．この組み合わせは完全に一致している場合もあるが，一致していない場合も存在する．このような場合には Lang SAM から出力されたマスクを画像の強調として実際に用いる．これは SAM のマスクを選択するよりも精度が高かったためである．

1) 東京理科大学 創域理工研究科 情報計算科学専攻
2) 東京理科大学 創域理工学部電気電子情報工学科
3) 東京理科大学 創域理工学部 情報計算科学科

### 4.3 マスク部分の表示

本研究では図２のようにマスク部分以外の透明度を下げることで物体の強調を行った．図２では車の前輪を強調している．

図 ２：マスク部分の強調

## 5. 実験

### 5.1 実験の概要

本実験では 14 枚の画像に対して言語入力，座標入力を変更しつつ，36 パターンの入力において物体認識を行った．

SAM や Lang SAM のみを用いた場合の物体認識の精度とこれらを組み合わせた今回の手法の物体認識の精度を求め，比較を行った．SAM や Lang SAM を個別で用いる場合は出力される複数のマスクのうち，最もスコアの高いものを使用し精度の計測を行った．

### 5.2 実験結果

今回の実験結果は表１のようになった．

| | SAM | Lang SAM | 今回の手法 |
|---|---|---|---|
| 割合（%） | 52.8 | 61.1 | 72.2 |

表１：実験結果

SAM や Lang SAM を個別で用いた場合より今回の手法を用いた場合の方が物体認識の精度が高いことがわかる．

## 6. 評価と考察

### 6.1 精度向上についての考察

今回の精度が向上した理由として，今回の手法でのみ目的通りに認識を行うことができたパターンの存在があげられる．このようなパターンの原因として，セグメンテーションの対象となる物体が，ある物体の一部となっている場合に SAM における目的のマスクのスコアが低い傾向にあることがあげられる．このような場合，SAM の入力座標を変更する必要がある．

また，そもそも SAM と Lang SAM それぞれの出力結果に目的のマスクが含まれていない場合も存在した．この場合においては各モデル自体の精度向上を図る必要がある．

### 6.2 各モデルの精度についての考察

SAM を用いたとき，人工物など境界がはっきりした物体の場合は精度高く，自然物など境界があいまいな物体の場合は精度が低くなる傾向にあることがわかった．

逆に Lang SAM を用いたとき，自然物など境界があいまいな物体の場合は精度高く，人工物など境界がはっきりした物体の場合は精度が低くなる傾向にあることがわかった．また，Lang SAM では対象となる物体が画像内に多数存在するとき，１つずつ認識されるのではなくまとめて認識されてしまう場合も確認された．この場合，集合体の中の１つのみを認識するということが困難となる．

### 6.3 総評

今回の実験結果から SAM と Lang SAM 両方を用いることで物体認識の精度が向上した．しかし，SAM もしくは Lang SAM では目的通り結果が得られたが今回の手法で得られなかった場合があるため，さらなる改善の余地がある．

改善策として，入力の再調整，認識する物体に応じてマスクの選択方法の変更，SAM や Lang SAM 自体の認識精度の向上などがあげられる．

また，今回はプレゼンテーション時における画像強調システムであるため，強調したい部分を含むような形で認識することで，目的通りでなくても利用することが可能である．よって，マスクをより包含するような形で出力することでシステムとして最悪の事態を回避することができる．

## 7. 今後の課題

今回の課題は主に 3 つある．一つ目は実行時間だ．今回は Google Colab の GPU，T4 を用いて 2 分程度の時間を要している．これは実際のプレゼンテーション時に使用するには厳しいと考えられる．よって，複数ある SAM のモデルのうち，今回用いた精度重視ではなく速度重視のモデルにおける実装も試し，実行時間の短縮を図る．二つ目は物体認識の精度である．さらなる物体認識の精度向上を図るため，入力方法の追加，入力の自動調整，物体に応じたマスクの選択方法の調整を行い，精度を調査する．三つ目は他要素の実装だ．今回はシステムのうち物体認識の精度向上を行ったが，実際に利用するにあたって音声をテキスト形式に変換する方法や座標入力の方法を検討する必要がある．

## 8. おわりに

本研究ではプレゼンテーション時における画像の強調システムの一部を作成した．また，物体認識においてより高い精度で認識することに成功した．しかし，実用するにあたって実行時間の短縮やシステムの他要素の実装を行う必要があり，今後，改善する予定である．

### 参考文献

[1] Alexander Kirillov and Kaiming He and Ross Girshick and Carsten Rother and Piotr Dollár，” Panoptic Segmentation”， Computer Vision and Pattern Recognition， 2019

[2] Alexander Kirillov and Eric Mintun and Nikhila Ravi and Hanzi Mao and Chloe Rolland and Laura Gustafson and Tete Xiao and Spencer Whitehead and Alexander C. Berg and Wan-Yen Lo and Piotr Dollár and Ross Girshick，``Segment Anything"，Computer Vision and Pattern Recognition， 2023

[3] <https://github.com/paulguerrero/lang-sam/>，2023 年 1 月 27 日閲覧

[4] Shilong Liu and Zhaoyang Zeng and Tianhe Ren and Feng Li and Hao Zhang and Jie Yang and Chunyuan Li and Jianwei Yang and Hang Su and Jun Zhu and Lei Zhang，” GroundingDINO”，Computer Vision and Pattern Recognition， 2023