多様なテキスト・画像モデルを活用した文字アートの自動生成

植木研究室

21J5099　日髙健士郎

目的

テキストの情報を入力するだけで自動的に文字を扱ったアート作品を生成する手法は少ない.

簡単なテキストを入力するだけで文字アートを容易に自動生成できる新しいシステムの提案.

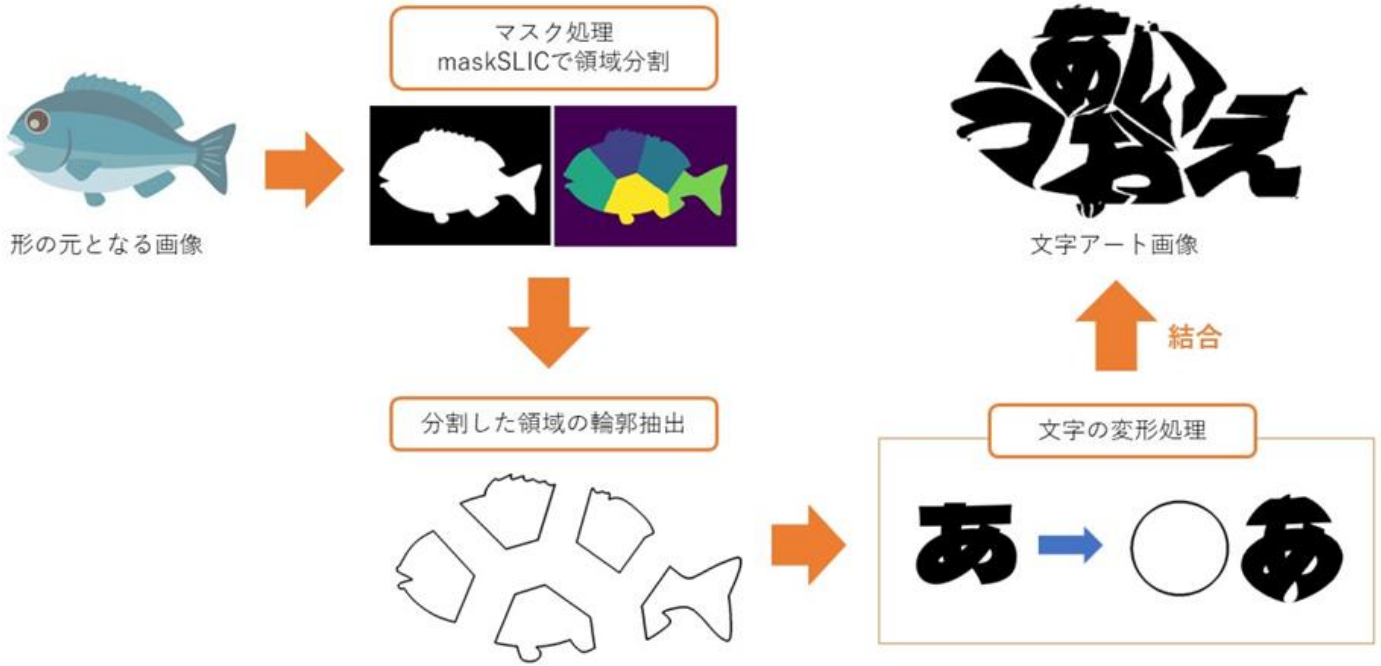
このシステムで生成される文字アートは，単なる画像として見るだけでなく，文字から連想されるイメージを通じて想像力を広げ、楽しむことができる新たな体験を提供する.

関連研究

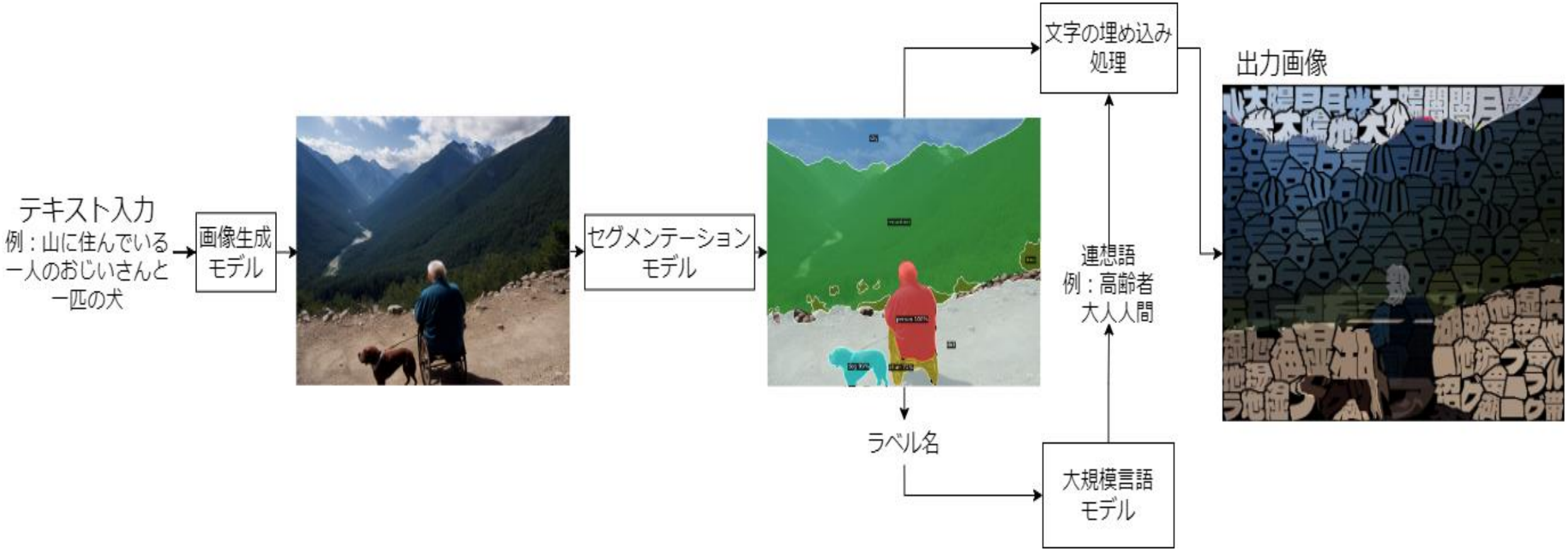
- 文字を利用したアート作品の生成

  小野 萌子，植木 一也，映情学技報，vol.47，pp.143–144，2023.

  - 形の元となる画像と入れたい文字を入力すると，画像の形に文字が変形した文字アートが生成されるシステム.



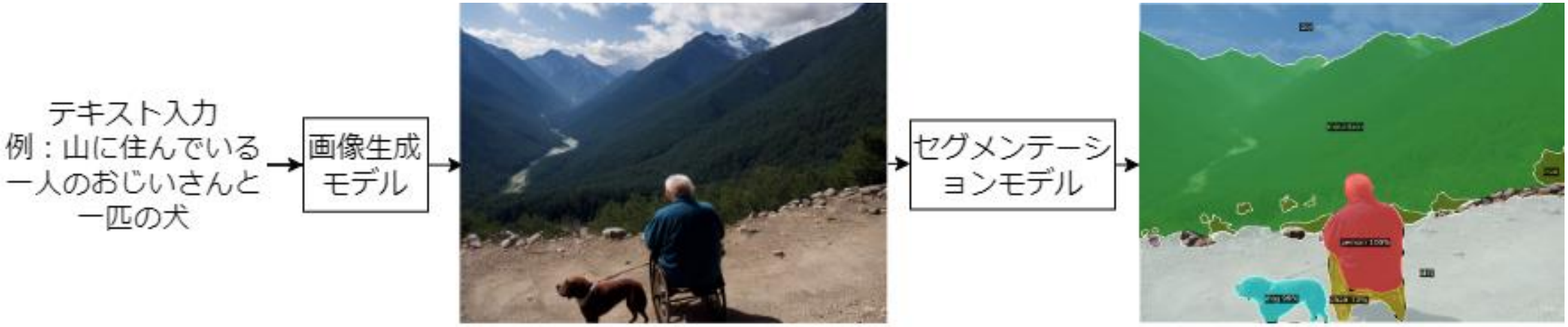
システム概要図


テキスト入力
例：山に住んでいる
一人のおじいさんと
一匹の犬
画像生成
モデル
セグメンテーション
モデル
ラベル名
大規模言語
モデル
連想語
例：高齢者
大人人間
文字の埋め込み
処理
出力画像

① 画像生成，領域分割

1. テキストを画像生成モデルに入力し，画像を出力.

2. 画像をパノプティックセグメンテーションモデルに入力し，
   各領域のマスクとそのラベル名を取得.



② 連想語出力と文字数の決定

1. 取得した各領域のラベル名を，大規模言語モデルに入力.
   ラベル名から連想される複数の単語を出力.

2. 単語の中から埋め込み処理に使用する単語を選択.

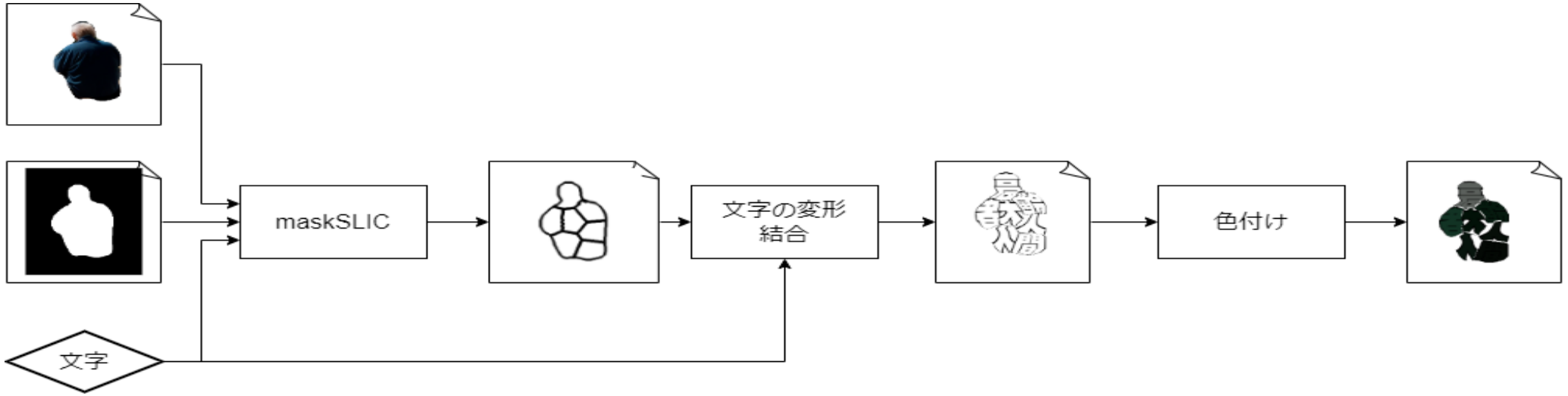
→ ラベル名：Person → 大規模言語モデル → 連想語出力：高齢者大人人間

③ 文字の埋め込み処理

1. セグメンテーションで取得した各領域を，
   連想語出力で選定された単語の文字数に応じた小領域に分割.

2. 各小領域の輪郭に沿って文字を変形し，元の領域に合わせて結合したのち，
   小領域内の色の平均値を用いて文字を塗りつぶす.

3. セグメンテーションで取得した各領域全てに文字を埋め込んだ後，その
   領域を結合する.



検証実験：実験条件

使用したモデル

- 拡散モデル：Stable Diffusion SDXL Turboモデル
- セグメンテーションモデル：Detectron2 Panoptic FPN R101 3xモデル
- 大規模言語モデル：Swallow 7B instruct モデル

実験方法

- 単語のみの入力，文章のみの入力，複数の単語を組み合わせた入力，
  単語と文章を組み合わせた入力

といった複数の異なる入力方法を用いてテキスト入力を実施.

•繰り返し実験を行った.

検証実験：実行結果

入力テキスト：

出力画像：

（元々の画像）

cat →

入力テキスト：

出力画像：

（元々の画像）

Person playing with dog →

検証実験：実行結果

入力テキスト：

出力画像：

（元々の画像）

Cars, futuristic, skyscrapers, planes →

入力テキスト：

出力画像：

（元々の画像）

A bird in the sky, the sea →

検証実験：実行結果＆まとめ

- 提案した内容の通りにシステムが動作し，
  文字アート画像を生成できた.

- 複数の異なる形式の入力方法に対応できた.

基本的な性能が十分に検証でき、多様な入力形式に対する適応性も実証できた.

検証実験：期待通りの出力がされなかった事例

セグメンテーションが正しく行われなかった際

入力テキスト：

big cat

出力画像：

（元々の画像）

考えられる理由

画像に対する物体が占める割合が大きすぎる場合にセグメンテーションモデルが対応できていない.

対処案

背景に収まりきれていない画像のトレーニングデータをセグメンテーションモデルに学習させ，データを拡張する.

検証実験：期待通りの出力がされなかった事例

ラベルが誤っている場合の生成結果

入力テキスト：

jellyfish

出力画像：

（元々の画像）

考えられる理由

使用したセグメンテーションモデルのDetectron2に搭載されているデータセット含まれる80個の物体ラベルの中に，「jellyfish」というラベルが存在しないため.

対処案

600カテゴリ以上のラベルを搭載しているデータセットであるOpen Images V7などの多くのラベルを搭載しているデータセットを使用する.

システムについてのアンケート評価

アンケートの実施条件

対象は10代から50代の男女21名．集計にはGoogleFormsを使用.
調査期間は，2024年12月16日～2024 年12月20日.

アンケート方法

まずアンケート回答者にシステムを使用して複数の画像を生成していただき，その後質問に回答してもらう.

質問項目

「感じない」，「あまり感じない」，「どちらでもない」，「やや感じる」，「感じる」の5段階評価を実施.

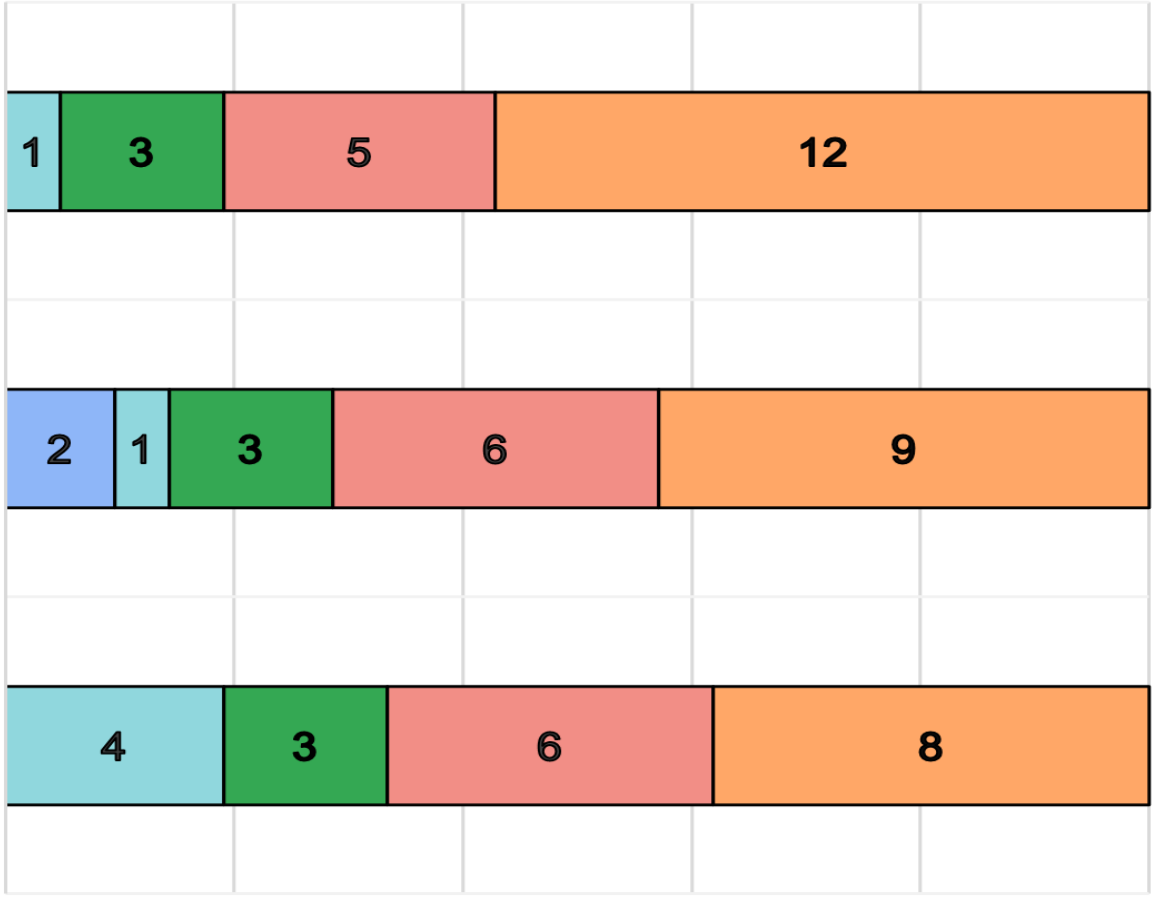
1. 簡単な入力から文字アート画像が生成されたと感じたか.
2. 入力テキストの内容に沿った画像が生成できていると感じたか.
3. 生成された画像を見て，文字から物体を連想することに楽しさを感じたか.

アンケートの回答結果

1，簡単な入力から文字アート画像が生成されたと感じますか？

2、入力テキストの内容に沿った画像が生成出来ていると感じますか？

3、生成された画像を見て、文字から物体を連想することに楽しさを感じましたか？



まとめ

簡単なテキスト入力から，文字アートを生成でき，

アンケート評価により，システムの有効性も確認できた.

- ただ，実験を行った中で，期待通りの出力がされなかった事例も見られた

今後の研究

- 実験で示した課題を解決する.
- 生成された画像を見て，文字から物体を連想することに楽しさをさらに感じてもらうため，ユーザーからもらった意見をシステムに反映し，生成画像の品質の向上や，システムの機能性を向上させるなどを行う.

発表実績

- 日髙健士郎，田島友希，松山諒大，武藤良，植木一也，
  “多様なテキスト・画像モデルを活用した文字アートの自動生成,”
  芸術科学会 (NICOGRAPH2024)，2024.

ありがとうございました。

補足資料：提案内容の用語解説

パノプティックセグメンテーション

画像内のすべての領域に対して一意のカテゴリと個体識別情報を同時に付与する手法.

画像のピクセル1つひとつに対してラベル付けしていく手法であるセマンティックセグメンテーションと，画像の中にある物体の領域を特定した後で個体ごとに領域分割し、物体の種類を認識する手法であるインスタンスセグメンテーションを組み合わせたもの.

(a) image　(b) semantic segmentation

(c) instance segmentation　(d) panoptic segmentation

補足資料：提案内容の用語解説

大規模言語モデル

大量のデータとディープラーニングによって構築された深層学習モデルで，文の生成，要約，翻訳，質問応答など，多様なタスクを行うことができる．代表的なモデルとして，GPT-4やBERTなどがある．

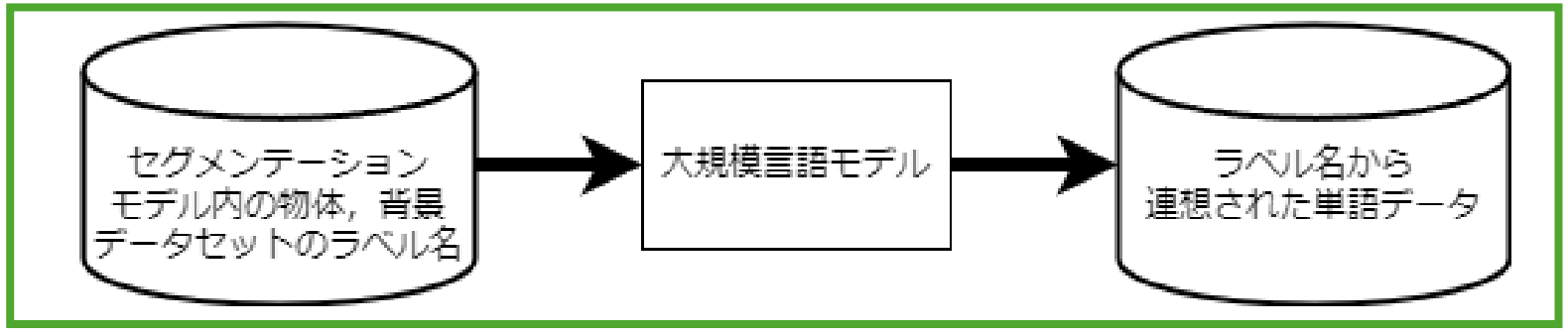
maskSLIC

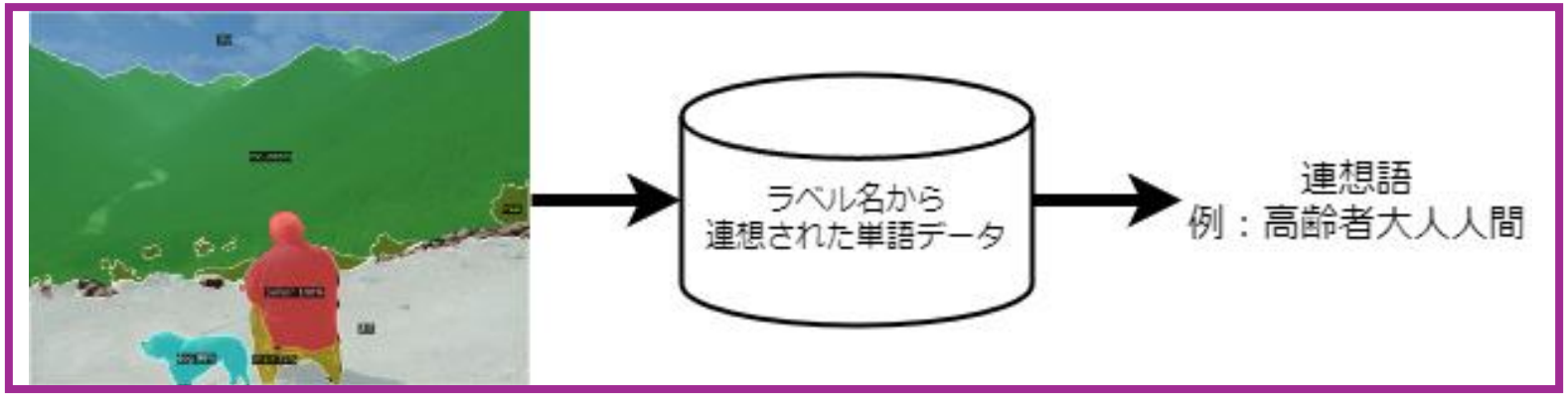
maskSLIC はSLIC を拡張したもので，色と距離に基づき画素をクラスタリングし，類似した色を持つ画素が集まる小領域を生成する手法．

maskSLIC では，マスク画像内のみを対象に領域分割を適用することが可能．

補足資料：連想語出力の詳細

本研究では，あらかじめパノプティックセグメンテーションモデルに登録されている物体，背景の領域データセットのラベル名すべてを大規模言語モデルに入力し，それらの複数の連想語を 出力してデータとして保管しておき，システムが実行された際にはそのデータから連想語を取り出すという形をとっている．こうすることで，システム実行の度に各領域のラベル 名を大規模言語モデルに入力する必要がなくなり，システムの軽量化や実行時間の短縮が可能となる．





補足資料：使用したモデルの詳細

Stable Diffusion SDXL Turboモデル

Stable Diffusion の SDXLモデルを拡張したもので，画像生成の際に，敵対的拡散蒸留（Adversarial Diffusion Distillation: ADD）という新しい蒸留技術を採用することで，生成画像の品質を保ちながらリアルタイムでの画像出力を可能にしているモデル.

本研究では，システムを動かすうえで高品質な画像を生成でき，実行時間も短くすることができるという理由からこのモデルを採用している.

Detectron2 Panoptic FPN R101 3xモデル

Facebook AI Researchによって開発された画像認識モデルで，物体検出やポーズ推定などを行うことができる.

本研究では，パノプティックセグメンテーションを行うことができるモデルであるという事から採用している.

補足資料：使用したモデルの詳細

Swallow 7B instructモデル

東京工業大学情報理工学院と，国立研究開発法人産業技術総合研究所によって開発された日本語能力に優れた生成AIの基盤である大規模言語モデル．MetaのLlama 2に日本語の文字や単語などの語彙も追加したうえ，新たに開発した日本語データを用いてモデルの構築を継続的に行う継続事前学習を行うことで構築を行っている．

本研究では，埋め込む文字に平仮名，カタカナ，漢字などの日本語の文字を使用したいという理由からこのモデルを採用している．

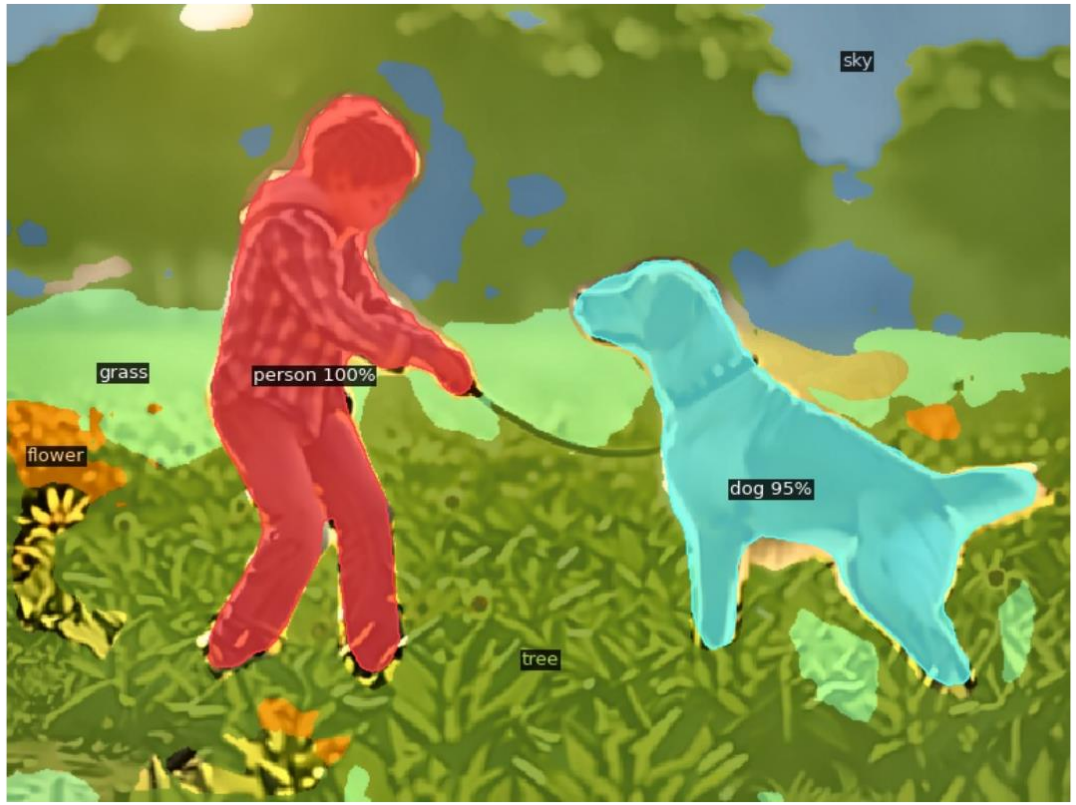
補足資料：実験の実行結果の詳細

セグメンテーション時の画像

ラベル名	連想語
cat	猫，犬，イヌ科，哺乳類，獣，ネコ科，動物，アニマル
wall	壁，板，厚い，薄い，レンガ，家，建物，建築，石膏，コンクリート，セメント，鋼鉄，鉄，木造建築

補足資料：実験の実行結果の詳細

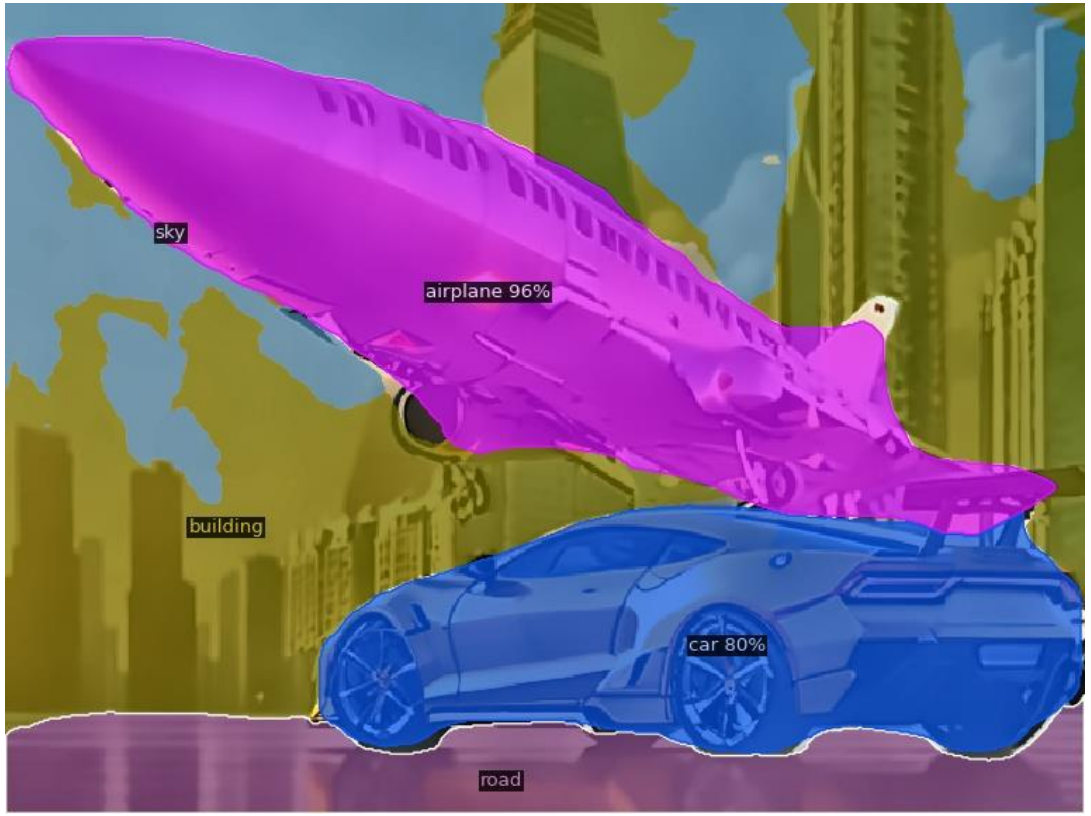
セグメンテーション時の画像



ラベル名	連想語
person	人間、人種、人類、個人、人道、人類学、humanity、human、people、persons、sapiens
dog	イヌ、犬、ドッグ、コギー、ピットブル、ポインター、レトリーバー、スパニエル、ブルドッグ、ラブラドール・レトリーバー
flower	桜、桃、梅、菊、バラ、ひまわり、チューリップ、ゆり、あさがお、カーネーション、ポピー、ダリア、マーガレット、コスモス、デイジー、ラン、グラジオラス、ボタン、フリージア、スイートピー、ユリ
tree	木、葉、樹皮、枝、葉の葉、森の葉、森の木、森
sky	水、大地、星、月、太陽、大地、光、木、闇
grass	草、花、草むら、 ハーブ、牧場、草原、芝、芝刈、牧場

補足資料：実験の実行結果の詳細

セグメンテーション時の画像



ラベル名	連想語
airplane	飛行機、ホテル、予約、宿泊、ツアー、旅行、空港、航空券、チケット、パスポート、スーツケース、国際線、観光、出張、空港ラウンジ、飛行機事故、旅行会社、空港ホテル、旅行会社
car	車、汽車、自動車、オートバイ、バイク、車輪、タイヤ、ホイール、自転車、 車両、クルマ、くるま
road	道、街、店、街路、ロード、ストリート、ハイウェイ、アヴェニュー、高速道路、自動車、都市、山道、田舎道、峠道、道路、ドライブ
sky	水、大地、星、月、太陽、大地、光、木、闇
building	建物の、ビル、ビルディング、建築物、家屋、住居、ハウス、建物、家、構造物

補足資料：実験の実行結果の詳細

セグメンテーション時の画像

ラベル名	連想語
bird	鳥、鶏、コウモリ、ワシ、フクロウ、ペンギン、オウム、シギ、ツバメ
person	人間、人種、人類、個人、人道、人類学、humanity、human、people、persons、sapiens
person	上と同じ
sea	水、地球、ビーチ、波、表面、海洋、地中海、プール、海洋、オーシャン
sky	水、大地、星、月、太陽、大地、光、木、闇

補足資料：その他の実行結果

入力テキスト：

出力画像：

（元々の画像）

woman

入力テキスト：

出力画像：

（元々の画像）

Tree

補足資料：その他の実行結果

入力テキスト：

Boat next to a house with sea view

出力画像：

（元々の画像）

入力テキスト：

People getting off at bus stops in town

出力画像：

（元々の画像）

補足資料：その他の実行結果

入力テキスト：

Horse,
mountain,
winter, person

出力画像：

（元々の画像）

入力テキスト：

pizza, table,
Wine, glass

出力画像：

（元々の画像）

補足資料：その他の実行結果

入力テキスト：

Bright colors painted the garden in beauty, flower

出力画像：

（元々の画像）

入力テキスト：

Running swiftly down the road, motorcycle

出力画像：

（元々の画像）

補足資料：日本語入力での実験

検証したこと

日本語の入力でも，提案した内容の通りにシステムが動作し，文字アート画像を生成できるか.

実験方法

検証実験で使用した入力を日本語に変換して行った.

補足資料：日本語入力での実験

入力テキスト：

猫

出力画像：

（元々の画像）

入力テキスト：

犬と遊ぶ人

出力画像：

（元々の画像）

補足資料：日本語入力での実験

入力テキスト：

車、未来、
高層ビル、
飛行機

出力画像：

（元々の画像）

入力テキスト：

空に飛ぶ鳥、
海

出力画像：

（元々の画像）

補足資料：日本語入力での実験

- 入力テキストにおいて単一の単語に基づく生成はできているが，文章全体の内容や複数の単語に基づく生成においては不十分.

考えられる理由

SDXL Turboモデルが日本語入力からの画像生成に適しておらず，文章などの内容は十分に理解できていない.

対処案

- 日本語を英語に翻訳して，英語を画像生成モデルに入力できるようにする.

- JapaneseStableDiffusion XLなどの日本語に特化したモデルを導入.

補足資料：期待通りの出力がされなかった事例

セグメンテーションが正しく行われなかった際，セグメンテーション時の画像

補足資料：期待通りの出力がされなかった事例

ラベルが誤っている場合の生成結果，セグメンテーション時の画像

補足資料：期待通りの出力がされなかった事例

一貫性のない文章，複数の単語を入力した際

入力テキスト：

horse, car,
smartphone,
apple, curry

出力画像：

（元々の画像）

考えられる理由

SDXL Turboモデルが入力テキストを完全には理解できておらず，入力テキストを部分的に解釈した上で画像を生成している.

対処案

入力を，自然言語処理モデルに渡し，文章を解析，モデルが解釈しやすい形に整理してから画像生成モデルに入力するなど.

補足資料：期待通りの出力がされなかった事例

セグメンテーションが正しく行われなかった際

入力テキスト：

superman

出力画像：

（元々の画像）

考えられる理由

Detectron2のPanoptic FPN R101 3xモデルが，アニメ，漫画風の画像のセグメンテーションには適していない.

対処案

画風を判定し，アニメーション用に特化したセグメンテーションモデルを適用する.
※https://github.com/SkyTNT/anime-segmentation

補足資料：アンケート結果の検定

検定方法

- 「感じる」「やや感じる」を肯定的な回答，「どちらでもない」を中立的な回答，「感じない」「あまり感じない」を否定的な回答とする．
- これら3種類の回答についてカイ二乗検定を行い，それぞれの回答に有意な差があるのかライアンの名義水準を用いて多重比較を実施．

検定結果

1. 簡単な入力から文字アート画像が生成されたと感じたか
   - 「感じる」「やや感じる」の肯定的な回答と，「どちらでもない」の中立的な回答の間に有意な差が確認できた．（p値＝0.00365，名義水準α'：0.03333），
   - 「感じる」「やや感じる」の肯定的な回答と，「感じない」「あまり感じない」の否定的な回答の間にも有意な差が確認できた．（p値＝p=0.00041，名義水準α'：0.01667）

補足資料：アンケート結果の検定

検定結果

2. 検定結果 2：入力テキストの内容に沿った画像が生成できていると感じたか
   - 「感じる」「やや感じる」の肯定的な回答と，「どちらでもない」の中立的な回答の間に有意な差が確認できた．（p値＝0.00952，名義水準α'：0.03333）
   - 「感じる」，「やや感じる」の肯定的な回答と，「感じない」，「あまり感じない」の否定的な回答の間にも有意な差が確認できた．（p値＝0.00952，名義水準α'：0.01667）
3. 検定結果3：生成された画像を見て，文字から物体を連想することに楽しさを感じたか
   - 「感じる」「やや感じる」の肯定的な回答と，「どちらでもない」 の中立的な回答の間に有意な差が確認できた（p値＝0.01529，名義水準α'：0.01667）
   - 「感じる」「やや感じる」の肯定的な回答と，「感じない」「あまり感じない」の否定的な回答の間に は有意な差がわずかに確認できなかった（p値＝0.03389，名義水準α'：0.03333）．

補足資料：回答者が生成した画像の結果

結果のまとめ

- 簡単な単語での入力が多く，文章での入力は少なかった.
- 固有名詞での入力も多く見られた.

固有名詞を入力した際の問題について考えられる理由

Detectron2のPanoptic FPN R101 3xモデルに搭載されているデータセット含まれる80個の物体ラベルの中に，ラベルが存在しないため.

対処案

オープンボキャブラリのセグメンテーションであるTAGや，EOV-Segを導入.

さらに物体の分類を細分化したラベルから連想される単語の文字を使用できることが期待できる.

補足資料：回答者が生成した画像の結果

入力テキスト：

出力画像：

（元々の画像）

Joker

入力テキスト：

出力画像：

（元々の画像）

godzilla

補足資料：回答者が生成した画像の結果

入力テキスト：

出力画像：

（元々の画像）

the king of sandland

入力テキスト：

出力画像：

（元々の画像）

America

補足資料：アンケートでもらった意見

必要だと思う機能について

- 日本語入力
- 画像内の物体に埋め込まれている文字を表示する機能
- 文字を編集できる機能

活用方法

- クイズへの活用
- 小中学生の美術の時間など想像力を使った教育への活用
- 商品のパッケージデザインへの活用

参考文献

- [関連研究]小野萌子，植木一也，“文字を利用したアート作品の生成,” 映情学技報，vol.47，pp.143–144， 2023.
- [パノプティックセグメンテーション]A. Kirillov，K. He，R. Girshick，C. Rother，and P. Dollar，“Panoptic segmentation,” Computer Vision and Pattern Recognition，2019.
- [maskSLIC]三宅克典，Benjamin Irving，Iulia A. Popescu，Russell Bates，P. Danny Allen，Ana L. Gomes，Pavitra Kannan，Paul Kinchesh，Stuart Gilchrist，Veerle Kersemans，Sean Smart，Julia A. Schnabel，Sir J. Michael Brady and Michael A. Chappell “maskSLIC: Regional Superpixel Generation with Application to Local Pathology Characterisation in Medical Images,” arXiv:1606.09518v2, 2017.
- [SLIC]Radhakrishna Achanta，Appu Shaji，Kevin Smith，Aurelien Lucchi，Pascal Fua，and Sabine Susstrunk，“SLIC Superpixels Compared to State-of-the-art Superpixel Methods,” 2011 dec.

参考文献

- [stable diffuison]Robin Rombach，Andreas Blattmann，Dominik Lorenz，Patrick Esser，Björn Ommer，“High-Resolution Image Synthesis with Latent Diffusion Models,” In Proc. Of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
- [SDXL] Dustin Podell，Zion English，Kyle Lacey，Andreas Blattmann，Tim Dockhorn，Jonas Müller，Joe Penna，Robin Rombach，“SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis,” Computer Vision and Pattern Recognition，2024.
- [swallow]藤井一喜，中村泰士，Mengsay Loem，飯田大貴，大井聖也，服部翔，平井翔太，水木栄，横田理央，岡崎直観，“継続事前学習による日本語に強い大規模言語モデルの構築,” 言語処理学会第 30 回年次大会発表論文集，2024.
- [Llama 2] Hugo Touvron et al.，“Llama 2: Open Foundation and Fine-Tuned Chat Models,” arXiv，2023 Jul

参考文献

- [TAG] Yasufumi Kawano, Yoshimitsu Aoki, “TAG: Guidance-free Open-Vocabulary Semantic Segmentation,” Meeting on Image Recognition and Understanding, 2024.
- [EOVSeg] Hongwei Niu, Jie Hu, Jianghang Lin, Guannan Jiang, Shengchuan Zhang, “EOVSeg: Efficient Open-Vocabulary Panoptic Segmentation,” Computer Vision and Pattern Recognition, 2024.