Oral Presentations | Track 3: Information Retrieval, Information Recommendation, and Social Media

Thu. Feb 29, 2024 10:00 AM - 12:10 PM JST | Thu. Feb 29, 2024 1:00 AM - 3:10 AM UTC | T3-B オンライン（Zoom Events)

画像検索・生成

座長:古賀 久志(電気通信大学)
コメンテータ:片山 薫(東京都立大学)

10:00 AM - 10:05 AM JST | 1:00 AM - 1:05 AM UTC

Introduction

10:05 AM - 10:20 AM JST | 1:05 AM - 1:20 AM UTC

[T3-B-4-01] 条件付き画像検索と画像生成の統合

*澤田 一正[1]、坂地 泰紀[1]、野田 五十樹[1]、小山 聡[2] (1. 北海道大学、2. 名古屋市立大学)

10:20 AM - 10:45 AM JST | 1:20 AM - 1:45 AM UTC

[T3-B-4-02] Finding Generative Image LoRA Model by Inputting Style Sample Image

*VUTHI NGOCANH[1]、Shoji Yoshiyuki[1]、Pham HuuLong [2]、Ohshima Hiroaki[2] (1. Shizuoka University、2. University of Hyogo)

10:45 AM - 11:10 AM JST | 1:45 AM - 2:10 AM UTC

[T3-B-4-03] 生成型要約に基づくWebページのサムネイル生成

*前田 直宏[1]、山本 岳洋[1] (1. 兵庫県立大学)

11:10 AM - 11:35 AM JST | 2:10 AM - 2:35 AM UTC

[T3-B-4-04] 類似度グラフの事前再計算による拡散式画像検索の効率化

*加藤 辰弥[1,2]、駒水 孝裕[2]、井手 一郎[2] (1. 名古屋大学大学院知能システム学専攻　井手・駒水研究室、2. 名古屋大学)

11:35 AM - 12:00 PM JST | 2:35 AM - 3:00 AM UTC

[T3-B-4-05] 3Dモデルを利用したユーザーの求めるアングル画像の検索システムの開発

*永田 玄[1]、常 穹[1]、宮崎 純[1] (1. 東京工業大学情報理工学院情報工学系情報工学コース宮崎研究室)

条件付き画像検索と画像生成の統合

澤田　一正†　　坂地　泰紀††　　野田　五十樹††　　小山　　聡†††

† 北海道大学大学院情報科学院　〒 060–0814 札幌市北区北 14 条西 9 丁目
†† 北海道大学大学院情報科学研究院　〒 060–0814 札幌市北区北 14 条西 9 丁目
††† 名古屋市立大学データサイエンス学部　〒 467–8501 名古屋市瑞穂区瑞穂町字山の畑 1
E-mail: †intsawadan@eis.hokudai.ac.jp, ††{sakaji,i.noda}@ist.hokudai.ac.jp, †††oyama@ds.nagoya-cu.ac.jp

あらまし　条件付き画像検索は、クエリ画像と、クエリ画像に対する追加要求のテキストのペアから画像を検索するタスクであり、ファッションのデザインなど発想支援的な目的でも用いられる。条件付き画像検索では、クエリによっては適合する画像が十分に検索されない場合があるが、データベースに存在しない画像を生成することで、クエリの改善や発想の支援につながる可能性がある。本研究では、生成 AI を用いて条件付き画像検索と同じ設定で画像を生成し、既存の検索モデルで検索された結果と組み合わせる手法を提案する。さらに、統合した結果をクラウドソーシングを用いて評価する。

キーワード　検索モデル、検索クエリ、クラウドソーシング、LLM、評価・データセット

1　はじめに

オンラインショッピングにおいては画像を元に検索を行う手段の開発が重要となってきている。インターネットの普及により、オンラインショッピングが爆発的な成長を遂げ、従来の実店舗に代わって多くの人が商品を購入する主要な手段となっている。それに伴い、オンライン上でユーザーが望んでいる商品を検索する重要性が高まっており、様々な検索方式が提案されている。その 1 つとして、条件付き画像検索 (Conditioned Image Retrieval) という検索方式がある。条件付き画像検索は、画像と、その画像をどのように修正するかのテキストからなる検索クエリを用いて画像を検索する検索方式である。この検索方式は、衣服などの、視覚的特徴により検索を行うことができる商品に適しており、現在閲覧している商品画像に対して、修正点を入力することにより新たな商品を検索できるため、より望んだ商品を見つけやすくなることが期待できる。しかし、条件付き画像検索に関する既存の研究は、正解画像をどれだけ正確に検索できるか、といった検索の精度に焦点を当てているものが多く、検索結果の多様性などの、検索の精度以外の観点がほとんど考慮されていない。

近年、テキストや画像を生成する様々な生成 AI モデルが登場し、幅広い領域への活用が検討されている。例えば、衣服の新たなデザインの生成といった新たな商品画像の生成に活用されており、商品を検索するだけではなく、新たな商品を生成する試みがされている [1]。しかし、画像生成モデルは、高品質な画像を生成することに優れているものの、生成画像の多様性に欠けているという問題が指摘されている [2]。

また、オンライン上の商品検索において、ユーザーが最初の検索クエリとは完全に一致しないアイテムを購入する可能性があることが指摘されている [3]。例えば、最初は「黒いダウンジャケット」と検索したユーザーが、検索結果を見ている中で、ブルーのダウンジャケットが目に留まり、結果として最初の検索クエリとは異なる、ブルーのダウンジャケットを購入するといったことが考えられる。そのため、商品の検索結果は、検索クエリに完全に一致している必要はなく、検索クエリに関連する商品なども含む、検索結果の多様性が重要であると考えられる。

本研究では、条件付き画像検索というタスクにおいて、新しい商品を購入しようと検討している人が、本当に欲しい商品を見つけられるような商品画像の結果を出力する手法を提案する。具体的には、検索結果の検索クエリに対する適合性と多様性を両立する結果を得られることを目指す。既存の条件付き画像検索モデルによる検索結果と生成 AI により生成された商品画像を組み合わせることにより、生成 AI の多様性に欠けてしまう問題を解決しながら、検索クエリへの適合性を高めることを目標とする。

2　関連研究

2.1　条件付き画像検索 (Conditioned Image Retrieval)

条件付き画像検索 (Conditioned Image Retrieval) は、画像と、その画像をどのように修正するかのテキストからなる検索クエリを用いて画像を検索する画像検索タスクである。ここで、検索クエリの画像を reference image、検索クエリのテキストを relative caption、検索結果の正解画像を target image という。条件付き画像検索の一連の流れを図 1 に示す。

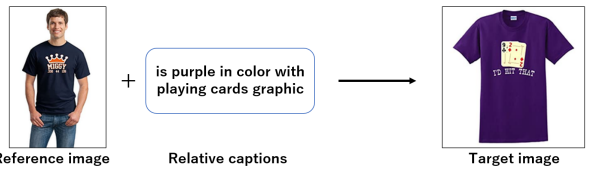


図 1　条件付き画像検索の流れ

条件付き画像検索はファッションのための対話型検索システムを実装するためのモデルとして提案された [4]。このシステムでは、条件付き画像検索を活用することにより、商品の画像と、その画像をどのように修正するかのテキストを入力することで、新たな商品を検索できるようになっている。また、検索された商品の画像に対しても、それをどのように修正するかのテキストを入力して検索することができるため、対話を通して商品を検索することができる。

条件付き画像検索のタスクにおいて、高い精度で検索を行うための様々な検索モデルが提案されてきた。Baldrati ら [5] は、CLIP [6] を活用して、条件付き画像検索を行うモデルを提案した。CLIP を用いて reference image と relative caption からそれぞれ特徴量を抽出する。これらの特徴量を組み合わせた新たな特徴量を作成するネットワークを構築し、その組み合わせた特徴量が target image の特徴量に近づくように学習を行う。Lee ら [7] は、2 つの異なるニューラルネットワークのモジュールを使用し、条件付き画像検索を行うモデルを提案した。1 つは relative caption に基づいて、reference image の特徴に対して、オブジェクトの形状やサイズ、配置などの詳細を修正するモジュールである。もう 1 つは、reference image の全体的な色調、質感などを修正するモジュールである。Chen ら [8] は、検索クエリのテキストである relative caption には、検索候補が 1 つに定まるような詳細に記述されているものと、検索候補が複数存在するような曖昧に記述されているものがある点に着目し、その両方の種類の relative caption の学習に対応した条件付き画像検索を行うモデルを提案した。

しかし、これらの既存の研究のほとんどが検索結果の精度に焦点をあてており、検索結果の多様性等の観点が考慮されていない。実際にこれらの検索システムを使用するユーザーの観点で考えると、検索結果の多様性は大切であるので、検討する必要があると考えられる。また、条件付き画像検索のタスクは基本的に、すでにあるデータベース内の画像から検索をおこなうため、データベース内に無い画像は検索できない。そのため、具体的に検索をしたいものがあったとしても、望ましい検索結果が得られない可能性がある。

2.2 画像生成を用いた衣服の画像生成

近年、画像生成を行う深層学習モデルの性能が向上しており、幅広い領域への活用が検討されている。Choi ら [9] は人間のファッションデザイナーが行う作業を模倣する AI ベースの衣服開発システムを開発した。具体的には、データ分析、トレンド抽出、デザイン生成、修正の各段階をサポートする 4 つのモジュールを備えた AI システムを提案した。デザイン生成には GAN(Generative Adversarial Network) を使用している。これらの画像生成を行う深層学習モデルを用いることで、条件付き画像検索において、検索したい画像がデータベース内に存在しない場合でも、検索クエリに適合する画像を新たに生成できる可能性がある。しかし、画像生成モデルは、高品質な画像を生成することに優れているものの、生成画像の多様性に欠けているという問題が指摘されている [2]。

そのため、条件付き画像検索と画像生成を統合することで、各々にかけているところを補うところに、本研究の狙いがある。それを検証するために使用できるデータセットを次節で紹介する。

2.3 データセット

条件付き画像検索のデータセットには、FashionIQ [4] や CIRR [10] がある。

FashionIQ はファッション画像に関するデータセットで、”Dress”、”Shirt”、”Toptee”の 3 つのカテゴリに分けられている。また、学習データ、検証データ、テストデータの 3 つに分割して提供されている。学習データには 46,609 枚の画像が含まれており、reference image と 2 件の relative caption、target image の検索セットが 18,000 件ある。検証データには 15,537 枚の画像が含まれており、検索セットが 6,117 件ある。テストデータには 15,538 枚の画像が含まれており、検索セットが 6,119 件ある。本研究の 4 章の実験にて、FashionIQ を使用する。

CIRR(Compose Image Retrieval on Real-life images) はファッションなどの特定のドメインに限定しない、一般画像に関するデータセットである。学習データ、検証データ、テストデータの 3 つに分割して提供されている。学習データには 16,939 枚の画像が含まれており、reference image と relative caption、target image の検索セットが 28,225 件ある。検証データには 2,297 枚の画像が含まれており、検索セットが 4,184 件ある。テストデータには 2,316 枚の画像が含まれており、検索セットが 4,148 件ある。

3 提案手法

3.1 検索システム

本研究の提案手法である検索システムについて説明する。検索システムの概要を図 2 に示す。この図には、検索クエリが入力されてから検索結果が出力されるまでの流れを示している。この検索システムでは、ユーザーは商品の画像とその画像をどのように修正するかのテキストからなる検索クエリを入力し、新たな商品画像の検索を行う。検索システムは、ユーザーから得られる検索クエリを入力とし、10 枚の画像の集合を検索結果として出力する。検索結果の画像枚数が 10 枚とした理由は、ユーザーにとってある程度選択肢があり、1 度にすべての商品を確認できる数であると考えたためである。画像検索モデルと画像生成モデルという別の 2 つの手法から得られた結果を組み合わせることで、各々にかけているところを補うことを期待して、提案手法を考案した。提案手法の流れとしては、まず既存の条件付き画像検索モデルを使用して、画像を検索する。次に、画像生成モデルを用いて検索と同じ条件で画像を生成する。すなわち、検索と同じ reference image と relative caption を用いて新たな画像を生成する。最後に、画像検索と画像生成で得られた画像を組み合わせることで新たな画像の集合を作成する。検索システムの各要素について次節より説明する。

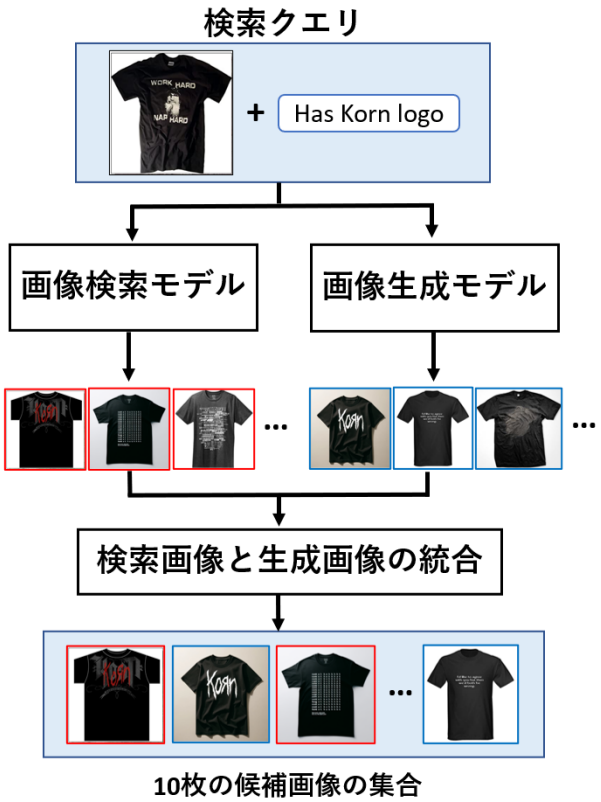


図 2　検索システムの概要

3.2　条件付き画像検索モデル

2.1 節で説明した既存の条件付き画像検索のモデル (MUR) [8] を使用し、画像データベース内から検索条件にあった画像を検索する。MUR では、reference image と relative caption の特徴量を組み合わせ、新たな特徴量を作成する。その新たに作成された特徴量と、画像データベース内のすべての画像の特徴量とのコサイン類似度を計算し、その値が大きい順に、検索結果の上位としている。本研究では、上記の手順で計算されるコサイン類似度の値を適合度と呼ぶ。本研究でも、MUR と同じように、データベース内の各画像に対して、適合度を計算する。その数値が高い順に、検索結果の上位とする。最終的に検索結果の上位 10 枚を出力する。

3.3　条件付き画像生成モデル

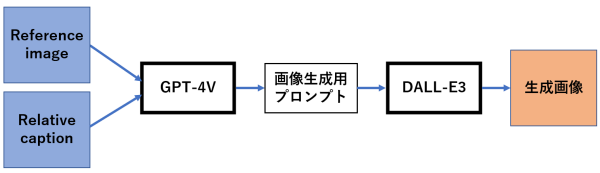


図 3　条件付き画像生成の流れ

条件付き画像検索と同じ reference image と relative caption を用いて新たな画像を生成する。条件付き画像生成の全体の流れは図 3 のようになる。まず初めに、GPT-4V(ision) [11] に reference image と relative caption を入力して、条件に合った画像を画像生成モデルで生成するためのプロンプトの作成を行う。画像生成用のプロンプトを作成する理由としては、画像生成で使用するモデルがテキストの入力しか受け付けていないためである。このとき、GPT-4V に対して以下のプロンプトを入力した。ここで、relative caption はプロンプトの最後に空行を 1 行追加し、その次の行に記述した。

> Please create a text prompt for DALL·E to generate a modified image using the following text based on the attached image.I would like to generate images from text prompts only, without inputting any images into DALL·E. Please come up with a prompt for this scenario. Try to retain as many detailed features from the provided original image as possible and make modifications only to the specified parts in the modified text.Please create a prompt that generates images of clothing only, not worn by a person. Create a prompt that generates images showing the entire piece of clothing, not just a part of it. Please output only the prompt.
>
> [relative caption]

上記のプロンプトでは、reference image の特徴を残しながら、relative caption で指定された部分のみを修正するように指示しており、今回の条件に適していると考えられる。

次に、生成された画像生成のためのプロンプトを用いて画像の生成を行う。画像生成モデルとしては DALL-E3 [12] を使用した。1 つの検索クエリに対して図 3 に示した一連の流れを、10 回繰り返し行うことにより、10 枚の画像を生成する。GPT-4V は、標準の設定が、モデルが生成するテキストの創造性と予測可能性のバランスをとるようになっているため、同じ入力を行っても、異なる出力が期待できる。

3.4　検索画像と生成画像の統合

検索モデルによって得られた画像 (検索画像)10 枚と生成モデルによって得られた画像 (生成画像)10 枚を組み合わせて、新たな 10 枚の画像の集合を作成する。このアルゴリズムを以下にまとめる。

(i)　生成画像 10 枚に対して、検索のときと同様に条件付き画像検索のモデルを使用して、検索条件にどのくらい適合しているかを表す適合度を計算する。

(ii)　検索画像 10 枚と生成画像 10 枚を合わせて 20 枚の画像の集合とし、適合度が高い順に並べる。

(iii)　もし上位 10 枚の画像の中に含まれている生成画像が、5 枚以上の場合は、その上位 10 枚の画像を新たな 10 枚の画像の集合とし、アルゴリズムを終了する。

(iv)　もし上位 10 枚の画像の中に含まれている生成画像が、5 枚未満の場合は、(検索画像の中で一番高い適合度) - (生成画像の中で一番高い適合度) - 0.0001 を計算し、その値が 0 以上の場合は、生成画像の上位 5 枚の適合度に、計算した値を加算する。

(v) 検索画像上位 5 枚と生成画像上位 5 枚を合わせて、適合度が高い順に並べ替え、その 10 枚を新たな 10 枚の画像の集合とし、アルゴリズムを終了する。

4 実　　験

提案手法が、条件付き画像検索というタスクにおいて、新しい商品を購入しようと検討している人に対して、本当に欲しい商品を見つけられるような商品画像の結果を出力することができるのかを検証するために、実験を行う。本研究では、2 つの手法を比較したときにどの程度の割合で検索クエリに適合しているかを表す「検索クエリに対する適合比」と、検索結果にどの程度多様性があるかを表す「検索結果の多様性」の 2 つの観点で評価を行う。それぞれの定義は 4.2 節で説明する。

4.1 実験設定

4.1.1 使用したデータ

本研究では 2.3 節で説明した FashionIQ [4] というファッションに関する条件付き画像検索のデータセットを使用した。本研究の実験では、データセットの 3 つのカテゴリの中から、“Shirts” のカテゴリを使用した。また、検索を行うときの検索候補画像として、FashionIQ の学習データと検証データに含まれるすべての画像を使用した。実験で使用する検索クエリは検証データのものを用いた。また、FashionIQ には 1 つのクエリに対して relative caption が 2 つ含まれているため、それらを “and” でつなげることにより、1 つの relative caption として扱った。

4.1.2 比較手法

本研究では、比較手法として、2.1 節で説明した既存の条件付き画像検索モデル [8] と 3.3 節で説明した提案手法の条件付き画像生成モデル 3.3 を用いた。条件付き画像検索モデルは、検索クエリを入力し、検索候補画像から画像を 10 件検索しそれらを出力する。条件付き画像生成モデルは、検索クエリを入力し、画像を 10 枚生成しそれらを出力する。

4.2 実験内容

各手法について、検索クエリから 10 枚の画像の集合 (画像集合) を作成する。それらの画像集合について、「検索クエリに対する適合比」と「検索結果の多様性」の 2 つの観点で評価を行う。この 2 つの観点について説明する。

4.2.1 検索クエリに対する適合比

検索クエリに対する適合比とは、2 つの手法を比較したときにどの程度の割合で検索クエリに適合しているかを表すものである。本研究では、2 つの手法を 10 件の検索クエリを用いて比較し、検索結果が検索クエリに対してより適合しているとクラウドソーシングにより判断された検索件数の比を検索クエリに対する適合比と定義する。例えば、手法 A と手法 B を比較する場合に、10 件の検索のうち、検索結果が検索クエリに対してより適合していると判断された件数が、手法 A が 2 件、手法 B が 8 件である場合、検索クエリに対する適合比は 手法 A : 手法 $B = 2 : 8$ となる。1 つの検索クエリに対して、5 人のクラウドワーカーに、2 つの検索結果のうち、どちらがより検索クエリにふさわしいかを選択してもらい、過半数である 3 人以上が選択した方の手法をその検索クエリに対してより適合していると判断する。検索クエリに対する適合比は、10 件の検索クエリそれぞれに対して 5 人のクラウドワーカーに評価を行ってもらっているため、延べ 50 件の調査により計算されている。ここで、10 件の検索クエリすべてに対して、同一の 5 人のクラウドワーカーが評価を行っているわけではなく、検索クエリごとにクラウドワーカーは異なる場合がある。クラウドソーシングのサービスはランサーズ [1] を用いた。

本研究では実験により、①画像検索モデルと提案手法の検索クエリに対する適合比、②画像検索モデルと画像生成モデルの検索クエリに対する適合比、③提案手法と画像生成モデルの検索クエリに対する適合比、の 3 つを調査した。上記の 3 つの調査を行った理由は、3 つの手法のすべての組み合わせに対して比較を行うためである。上記の 3 種類の検索クエリに対する適合比を評価する際の検索クエリ 10 件は、すべて同じものを使用した。今回の実験では、30 代後半〜60 代前半の 40 人がクラウドワーカーとして参加した。全員日本人であり、男女比は約 5:1 で男性の方が多かった。

4.2.2 検索結果の多様性

画像集合がどのくらい多様性があるかを評価する。多様性の評価指標として、情報推薦の分野で広く使用されている intra-list diversity(ILD) [13], [14] を利用した。この ILD という指標は、情報検索や推薦システムにおいて、検索結果や推薦結果のリスト内の項目間の多様性を測定するために使用される指標である。本研究においても、10 件の画像集合を ILD における検索結果のリストとみなすことができると考えたため、この指標を使用した。画像集合 I の多様性の値 $div(I)$ は以下のように計算される。

$$div(I) = \frac{1}{|I|(|I|-1)} \sum_{i \in I} \sum_{j \neq i \in I} (1 - sim(i,j)) \qquad (1)$$

ここで、$sim(i,j)$ は画像集合内の画像 i、j 間の類似度である。この類似度の値は、学習済みの CLIP [6] の Image Encoder を使用し画像の特徴量を抽出し、それらの特徴量のコサイン類似度の値とした。CLIP は画像とテキストのペアを理解するために設計されたモデルである。 インターネット上の大量の画像とテキストのペアから学習しており、関連する画像とテキストが類似した特徴表現を持つようになっている。CLIP はゼロショットで様々な種類の分類問題のデータセットを高い精度で分類できることが示されており [6]、多様な色、デザイン、形状などの衣服のビジュアル表現を理解する能力が備わっていると考えられる。以上の理由から、今回は学習済みの CLIP の Image Encoder を使用した。この多様性の値 $div(I)$ は、値が大きくなればなるほど多様性が高くなる。

実験の条件は、クエリへの一致度と同様に、①画像検索モデルと提案手法の比較、②画像検索モデルと画像生成モデルの

1 : https://www.lancers.jp/

比較、③提案手法と画像生成モデルの比較、の 3 つの条件で行った。それぞれの条件に対して、30 件の検索クエリで評価した。各条件では、2 つの手法で得られた画像集合の多様性の値 $div(I)$ を計算し、値が大きい方をより多様性の高い画像集合と判断した。

4.3 実験結果

4.3.1 検索クエリに対する適合比

クラウドソーシングにより得られた、2 つの手法を比較したときの、検索クエリに対してより適合していると判断された検索件数を表 1-3 に示す。4.2.1 節で説明したように、1 つの検索クエリに対して、5 人のクラウドワーカーに、2 つの検索結果のうち、どちらがより検索クエリにふさわしいかを選択してもらい、過半数である 3 人以上が選択した手法をその検索クエリに対してより適合していると判断する。表 1 は、画像検索モデルのと画像生成モデルを 10 件の検索クエリを用いて比較したときの、検索クエリに対してより適合していると判断された検索件数を示している。表 2 は画像検索モデルと提案手法の比較であり、表 3 は画像生成モデルと提案手法の比較である。表 1 より、10 件の検索クエリのうち、検索クエリに対してより適合していると判断された検索件数は、画像検索モデルが 2 件、画像生成モデルが 8 件であることが示されている。したがって、4.2.1 で定義した検索クエリに対する適合比は、画像検索モデル : 画像生成モデル = 2 : 8 となる。同様にして、表 2 より、画像検索モデルと提案手法の検索クエリに対する適合比は、画像検索モデル : 提案手法 = 3 : 7 となり、表 1 より、画像生成モデルと提案手法の検索クエリに対する適合比は、画像生成モデル : 提案手法 = 8 : 2 となる。

表 1　画像検索モデルと画像生成モデルの検索クエリに対してより適合している件数

	検索クエリに対してより適合している検索件数
画像検索モデル	2
画像生成モデル	8

表 2　画像検索モデルと提案手法の検索クエリに対してより適合している件数

	検索クエリに対してより適合している検索件数
画像検索モデル	3
提案手法	7

表 3　画像生成モデルと提案手法の検索クエリに対してより適合している件数

	検索クエリに対してより適合している検索件数
画像生成モデル	8
提案手法	2

4.3.2 検索結果の多様性

実験結果を表 4-6 に示す。これらの表は、2 つの手法を比較したときの、多様性の値が大きく、より多様性が高いと判断された検索クエリ数を示している。例えば、表 4 では、30 件の検索結果の多様性の値を比較したとき、画像検索モデルの多様性の値の方が高い検索結果が 26 件、画像生成モデルの多様性の値の方が高い検索結果が 4 件あったことを示している。各手法の多様性の値の平均値を表 7 に示す。この表では、各手法の 30 件すべての検索結果に対して計算された 30 個の多様性の値の平均値を表している。例えば、画像検索モデルの 30 個の多様性の値の平均値は、0.25561 となっている。

表 4　検索と生成の比較

	多様性の高い検索件数
画像検索モデル	26
画像生成モデル	4

表 5　検索と提案手法の比較

	多様性の高い検索件数
画像検索モデル	11
提案手法	19

表 6　生成と提案手法の比較

	多様性の高い検索件数
画像生成モデル	0
提案手法	30

表 7　多様性の値の平均値

	ILD の平均値
画像検索モデル	0.25561
画像生成モデル	0.16352
提案手法	0.26276

4.4 考　　察

4.4.1 検索クエリに対する適合比

4.3.1 節の実験結果から、検索クエリに対する適合比はそれぞれ、画像検索モデル : 画像生成モデル = 2 : 8、画像検索モデル : 提案手法 = 3 : 7、画像生成モデル : 提案手法 = 8 : 2 となった。画像検索モデル : 画像生成モデル = 2 : 8 と 画像生成モデル : 提案手法 = 8 : 2 の結果より、画像生成モデルが他の 2 つの手法よりも検索クエリに対する適合比の比率が高くなっていることがわかる。このことから、画像生成モデルは他の 2 つの手法と比べて、より検索クエリにふさわしい検索結果を出力すると考えられる。また、画像検索モデル : 提案手法 = 3 : 7 と 画像生成モデル : 提案手法 = 8 : 2 の結果より、提案手法は画像生成モデルよりも検索クエリに対する適合比の比率が低いが、画像検索モデルより比率が高くなっていることがわかる。このことから、提案手法は画像生成モデルよりも検索クエリにふさわしい検索結果を出力する能力は低いが、画像検索モデル

よりは検索クエリにふさわしい検索結果を出力する能力が高いと考えられる。

画像生成モデルが検索クエリに対する適合比で最も高い比率となった理由は、今回の実験条件が画像生成モデルにとって有利な条件であったためだと考えられる。なぜなら、今回使用したデータセットは、検索クエリに対する正解画像が 1 枚しかないという条件で作成されたものであるため、今回の 10 枚の検索結果を表示してその結果を比較するために十分な画像がデータベースに存在しない可能性が高いからである。また、画像生成モデルが、細かい条件に対応した画像を生成するのに優れているためだと考えられる。

4.4.2 検索結果の多様性

4.3.2 節の結果から、検索結果の多様性について考察する。

表 5 と表 6 の結果から、提案手法は他の 2 つの手法と比較して、多様性の高い検索件数が多くなっていることがわかる。また、表 7 の結果から、提案手法の多様性の値の平均値が最も高くなっていることがわかる。このことから、提案手法は他の 2 つの手法よりも、検索結果の多様性の観点では、一番優れていると考えられる。表 4 と表 5 の結果から、画像検索モデルは多様性の高い検索件数が提案手法よりも少ないが、画像生成モデルよりは多いことがわかる。また、表 7 の結果から、画像検索モデルの多様性の値の平均値は提案手法より低く、画像生成モデルより高いことがわかる。このことから、画像検索モデルは提案手法ほどではないが、画像生成モデルよりも検索結果の多様性が優れていると考えられる。表 4 と表 6 の結果から、画像生成モデルは、他の 2 つの手法と比較して、多様性の高い検索件数が少ないことがわかる。また、表 7 の結果から、画像生成モデルの多様性の値の平均値が最も低くなっていることがわかる。このことから、画像生成モデルは他の 2 つの手法よりも、検索結果の多様性の観点では、一番劣っていると考えられる。

5 おわりに

5.1 まとめ

本研究の目的は、条件付き画像検索というタスクにおいて、検索結果の検索クエリに対する適合性と多様性を両立する結果を得られることを目指すことであった。本研究では、既存の条件付き画像検索モデルによる検索結果と画像生成モデルにより生成された商品画像を組み合わせる手法を提案した。実験の結果、提案手法が既存の検索モデルよりも検索クエリへの適合性を向上させながら、高い多様性の結果が得られることが示された。したがって、提案手法は、既存の研究と比較して条件付き画像検索というタスクにおいて、新しい商品を購入しようと検討している人が、本当に欲しい商品を見つけられるような商品画像の結果を出力することが期待できる。

5.2 課題

本研究には次のような課題がある。まず、生成画像の多様性を高める手法の検討が十分ではない点である。本研究では生成画像の多様性を高めるために、検索モデルにより得られた画像と組み合わせたが、画像生成モデル自体の生成画像の多様性を高めるための手法を検討する必要がある。また、検索モデルにより得られた画像と生成画像を組み合わせる手法も、本研究では単純なアルゴリズムであるので、より高度な手法を検討する必要がある。多様性の評価方法に関しても、画像の表示される順番を考慮できていないという問題があるため、より適した評価方法を検討する必要がある。

謝辞

本研究の一部は、JST CREST JPMJCR21D1 の助成を受けたものである。

文献

[1] Liu, Linlin, et al. "Toward AI fashion design: An Attribute-GAN model for clothing match." Neurocomputing 341 (2019): 156-167.

[2] Zameshina, Mariia, Olivier Teytaud, and Laurent Najman. "Diverse Diffusion: Enhancing Image Diversity in Text-to-Image Generation." arXiv preprint arXiv:2310.12583 (2023).

[3] Manos Tsagkias, Tracy Holloway King, Surya Kallumadi, Vanessa Murdock, and Maarten de Rijke. Challenges and research opportunities in ecommerce search and recommendations. SIGIR Forum, 2020.

[4] Wu, Hui, et al. "Fashion iq: A new dataset towards retrieving images by natural language feedback." Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. 2021.

[5] Baldrati, Alberto, et al. "Effective conditioned and composed image retrieval combining clip-based features." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[6] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

[7] Lee, Seungmin, Dongwan Kim, and Bohyung Han. "Cosmo: Content-style modulation for image retrieval with text feedback." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[8] Chen, Yiyang, et al. "Composed image retrieval with text feedback via multi-grained uncertainty regularization." arXiv preprint arXiv:2211.07394 (2022).

[9] Choi, Woojin, et al. "Developing an AI-based automated fashion design system: reflecting the work process of fashion designers." Fashion and Textiles 10.1 (2023): 39.

[10] Liu, Zheyuan, et al. "Image retrieval on real-life images with pre-trained vision-and-language models." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[11] Yang, Zhengyuan, et al. "The dawn of lmms: Preliminary explorations with gpt-4v (ision)." arXiv preprint arXiv:2309.17421 9.1 (2023): 1.

[12] Li Jing Tim Brooks Jianfeng Wang Linjie Li Long Ouyang Juntang Zhuang Joyce Lee Yufei Guo Wesam Manassra Prafulla Dhariwal Casey Chu James Betker, Gabriel Goh and Yunxin Jiao. Improving image generation with better captions. In https://cdn.openai.com/papers/dall-e-3.pdf, 2023.

[13] Zhang, Mi, and Neil Hurley. "Avoiding monotony: improving the diversity of recommendation lists." Proceedings of the 2008 ACM conference on Recommender systems. 2008.

[14] Ziegler, Cai-Nicolas, et al. "Improving recommendation lists through topic diversification." Proceedings of the 14th international conference on World Wide Web. 2005.

Finding Generative Image LoRA Model by Inputting Style Sample Image

NgocAnh VUTHI†, Yoshiyuki SHOJI†, Huu-Long PHAM††, and Hiroaki OHSHIMA††

† Faculty of Informatics, Shizuoka University
Johoku, Hamamatsu-shi, Shizuoka 432 – 8011, Japan
†† Graduate School of Information Science, University of Hyogo
Gakuen-nishimachi, Nishi-ku, Kobe, Hyogo 651–2197, Japan
E-mail: †vu.thi.ngoc.anh.20@shizuoka.ac.jp, ††shojiy@inf.shizuoka.ac.jp, †††huulongpham28@gmail.com, ††††ohshima@ai.u-hyogo.ac.jp

Abstract This paper proposes a method for finding fine-tuned image-generation LoRA models suitable for generating images that have a similar style to a given sample image. Sharing trained image-generation AI models on the Internet is becoming common. However, to determine which of these models can generate the image style the user wants, it is necessary to inspect the output samples posted on the site one by one. To enable search for the model that can generate an image that has a similar style to a given image, we adopt a CNN-based multi-class classifier. The ResNet50 model was fine-tuned to classify the distinctive features of each LoRA model. We transformed pre-prepared sample images with each LoRA model using img2img as training data. Experimental results show that our customized ResNet50 can find more correct LoRA models than the original ResNet50. Future research should explore more effective methods to address the unique challenges presented in this novel approach.

Key words Finding model, Stable Diffusion model, LoRA model, CNN.

1 Introduction

Even though it has been only a short time since image-generative AI was released, people have already begun to open and share their fine-tuned models. Stable Diffusion, one of the most famous image-generative AI models, was released in 2022. This model generates images by processing input texts (*i. e.,* prompts), or a sample image. While the base Stable Diffusion model excels in various tasks, it encounters challenges in generating specific styles.

Rather than adjusting the prompt to generate specific styles, a more practical approach involves utilizing custom models that are fine-tuned models with images of the target sub-genre. These customized models, known as checkpoint models and LoRA models, offer enhanced performance. This paper focuses on custom LoRA models, demonstrating proficiency in generating specialized style images. Within one year, file-tuned LoRA models of sharing became common.

Specifically, on a website called "Civitai(注1)", many image-generative file-tuned LoRA models are shared. From anime to realistic style, they are fine-tuned with domain images and uploaded there. Users can freely download models and generate images of what they want.

Despite the proliferation of such models, subtle variations exist among those with similar styles, complicating user attempts to find a LoRA model that closely aligns with their desired output. When users want to find models that can be generated to style what they wish, they need to check the thumbnail images of each model to know whether this model can be generated to style what they want. This makes users waste a lot of time.

If thumbnail images of the model exist, the model can be found by the above methods. This method depends on existing thumbnail images. Also, adjusting hyperparameters (strength, guidance scale, etc.) of fine-tuned models when generating images can generate different style images, although using the same fine-tuned model. So that that model can generate the style the users want.

As described above, the fine-tuned image generation model can generate various styles of images by adjusting the hyperparameters and prompts. This means that even if a fine-tuned image generation model does not exist prior, if a model generated close to the style image that the user wants can be retrieved, an image in the desired style can be generated.

This paper proposes a method for obtaining an image-generative LoRA model using a single-style sample image. For example, when a user inputs a single sample image of Van Gogh, the system outputs models that can generate a painting style that is the same or close to the input image.

(注1)：Civitai: The Home of Open-Source Generative AI
https://civitai.com/

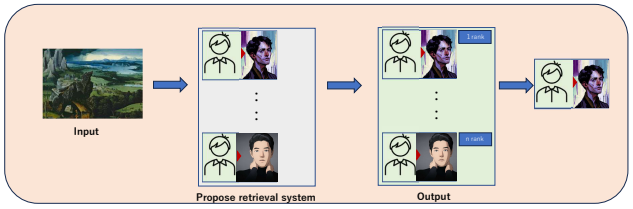


Figure 1 Overview of the retrieval system

Since multiple models may be output, model ranking is performed. The model that generates the closest match to the style of the input image is placed at the top of the list.

Figure 1 shows an overview of our system. It accepts a style image as its input, and outputs the ranking of models order by the possibility of generating a given image. To rank the models, we used a ResNet50-based classifier.

Our research aims to help users easily find fine-tuned image generative models without checking the thumbnail images of LoRA models. Moreover, it supports users in finding the alternative image generative model when the model they really want does not exist.

This system is necessary not only for users who want to generate an image, but also for artists and copyright holders. Artists and copyright holders may have to check if their original artworks were illegally used to train the model without agreement. When fine-tuning Stable Diffusion with specific domain images, it is necessary to use a lot of images in this domain. A fined-tuned model can only generate the style images used during fine-tuning. If the user is an artist with many well-known works, the painter may suspect their works are being used illegally. By using this system, the user can check for such illegal use. By inputting an image into the system, the system outputs a model that can generate a painting style similar to that of the image.

To find image-generative fine-tuned models, we used pre-trained ResNet50. To adapt pre-trained ResNet50 in our proposal, ResNet50 is fine-tuned with image classification tasks.

To create an image for fine-tuning ResNet 50, Stable Diffusion version 1.5, ControlNet - Canny, and LoRA models are used. Further, to boost ResNet50's performance with the given task, a small part of the original ResNet50 architecture is modified. By fine-tuning the pre-trained Resnet50 with the style-transformed image classification task, the fine-tuned Resnet50 obtains higher accuracy than the pre-trained ResNet50. Macro-recall, Macro-precision, and Macro-F1 scores are used to evaluate the image classification task.

Our paper is structured as follows. In this section, we describe the background and objectives of this study. Section 2 describes related studies and the position of these studies in related fields. Section 3 describes the methodology proposed in this study, and section 4 describes the evaluation method of the proposed methodology. Section 5 discusses the results obtained through experiments. Finally, we summarize this study and discuss future prospects.

2 Related Work

This study aims to help users find image-generative fine-tuned LoRA models more easily by inputting a sample image. This study relates to the image-generative AI field. Also, this study relates to the classification of images and related to the content-based image retrieval field.

2.1 Image Generative AI

In recent years, the field of artificial intelligence has witnessed remarkable advancements, especially in the domain of computer vision. One of the most intriguing developments is the emergence of generative AI image models, which have revolutionized how we perceive and interact with visual data. These sophisticated algorithms are at the forefront of the AI revolution, enabling machines to generate realistic and creative images.

First, we explain the concept of generative AI image models and their significance in the world of generative AI development. Generative AI refers to a class of artificial intelligence models that can generate data resembling authentic examples from the dataset they were trained on. In the context of images, generative AI aims to create new images that resemble the distribution of images in the training dataset, exhibiting creativity and imagination. The stable Diffusion model is a type of generative AI. It is used not only to generate images, but also to generate training data for supervised machine learning [1].

Model Agnostic Zero-Shot Classification is a method to classify real images by learning with synthetic images. Using the Stable Diffusion model improves the quality of the training dataset and solves problems relating to large-scale vision-language models [2].

Class incremental learning aims to learn new classes without forgetting previously learned classes incrementally. Several research works have shown how additional data can be used by incremental models to help mitigate catastrophic forgetting. A Stable Diffusion model [3]can generate synthetic samples belonging to the same classes as the previously encountered images.

Computer-aided surgical systems can provide surgeons with supportive information to improve procedure execution and overall outcomes. The Stable Diffusion model [4] of image-to-image task can reduce the gap between synthetic and real data. This makes computer-assisted surgical systems more accurately annotate data.

2.2 Image Classification by CNNs

CNN-based architecture is used for object detection and pattern recognition. Object detection is highly attended to in computer vision, but classification image style is not so attended to. Hence, there are few prior works that use CNNs to classify image styles.

CNN-based VGG-19 is adopted to extract object features and texture features [5]. Using a deep convolutional net to extract image features [6]. Although models were trained for object detection, these models can be used to classify aesthetics and styles.

CNNs are adapted to identify artists [7]. It has been discovered that CNNs can learn illustration styles from images. An ensemble of convolutional neural networks is adapted to learn the surface of image data, utilizing a CNN-based multiple neural network architecture [8]. The output of the CNN allows for the prediction of attribute images.

2.3 Content-Based Image Retrieval

In recent years, content-based image retrieval has become common in a variety of fields, such as product search and similar image retrieval. With the development of e-commerce sites, customers can search for and purchase just about anything on e-commerce. However, too much information hinders the search for information. Images [9] can be used to search for information easily.

Content-based image retrieval (CBIR) is a widely used technique for retrieving images from huge and unlabeled image databases. To help users more easily retrieve images [10], deep learning frameworks based on Convolutional Neural Networks (CNN) for feature extraction and Support Vector Machine (SVM) for classification are adopted. In this paper, instead of using SVM, pre-trained Resnet50 of the fully connected layer is used for classification. Fully connected layers of weights are initialized with ImageNet weights that obtain better results.

By retraining the CNN models with classification or similarity learning objective on the new domain, [11] found that the retrieval performance could be boosted significantly, which is much better than the improvements made by similarity learning.

Jin *et al.* [12] propose a method for searching images that share the same spatial semantics and enjoy visual consistency as the query image in complex scenes. They adopted CNNs to extract image features.

Adopting deep learning architecture to extract content from query images is common in the content-based image retrieval field. Instead of using text to search content, using images offers more information for the retrieval system, introducing the closest results that users expect.

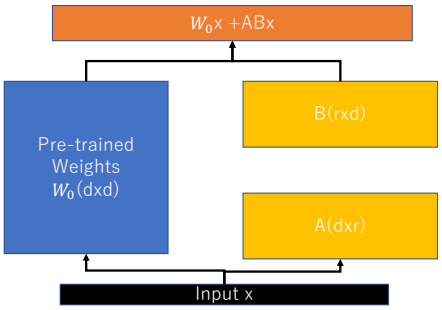


Figure 2 Low Rank Adaptation

3 Classification to Identify Model Capable to Generate Exemplified Styled Image

This section describes the proposed method. We explain the method in two sections: The image generation process, and Resnet50 fine-tuning task.

3.1 Image Generation Models

This sub-section describes the image generation process by fine-tuning image-generative AI models. Latent stable diffusion models [13], ControlNet-Canny [14], and Stable Diffusion fine-tuned by LoRA [15]are implemented. Latent stable diffusion models are exploited based on diffusion models. By training on latent space and introducing cross-attention layers into diffusion models. Latent stable diffusion models are more powerful and flexible generators for general conditioning inputs such as prompt or bounding boxes and high-resolution synthesis.

When enabling a pre-trained large model into the specific task, it is necessary to add a task-specific head or update the weights of the pre-trained large model through backpropagation during the training process. This process demands a lot of GPU memory. If small-scale data is used for full fine-tuning, the fine-tuned model can be catastrophic forgetting. This caused harmful to infer phase.

In Figure 2, instead of fine-tuning all the weights of the base model, the LoRA (low-rank adaptation) method decomposes the weights of the base model into matrix A and matrix B. The new A and B matrix weights are only updated during retraining; the weights of the base model are fixed. The LoRA method is being introduced for the first time through the fine-tuning of a large-scale language model. However, it can be applied to any dense layers in deep learning models. When fine-tuning Stable Diffusion models, the cross-attention layers in Stable Diffusion models are trained by LoRA. The cross-attention layers are the part of the model where the image and the prompt meet.

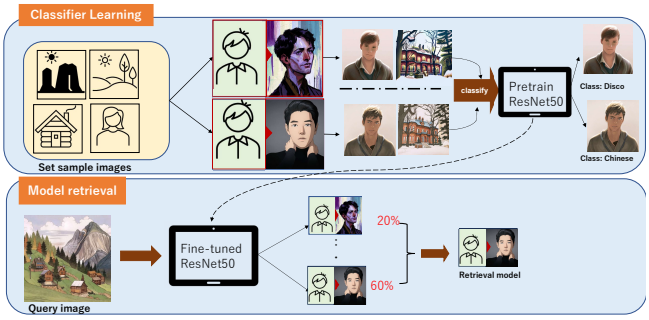


Figure 3 Image Classification and Model Retrieval Process

Using specific domain images and adopting LoRA for fine-tuning, fined-tuned Stable Diffusion models with specific styles are created; we call it the xx LoRA model. When inferring or generating, the weight of the fine-tuned LoRA model and the weight of the base model (Stable Diffusion model) are merged to implement the task. For instance, the anime LoRA model, the Chinese Ink LoRA model, etc.

With the advent of text-to-image Stable Diffusion models [13], we can create visually stunning images by inputting prompts. This paper proposes a method to classify the LoRA model style through a generated image. For this reason, image-to-image Stable Diffusion models are adopted to transfer base images into LoRA-style images. However, generating new images by image-to-image task without a prompt is still challenging.

The reason is related to the diffusion model mechanism, which uses image reconstruction. First, diffusion models noise input images by Markov chain, then reconstructing the image from the noised image via the denoising process. This means diffusion models can not retain anything about inputted images. This causes generating images through image-to-image to be stuck. ControlNet is adapted to address this challenge in the image-to-image task. The neural network architecture of ControlNet is described in Figure 4. We adopted ControlNet with Stable Diffusion model version 1.5 as the base model.

The ControlNet architecture based on the Stable Diffusion model version 1.5 works as below. ControlNet copies 12 encoder blocks and one middle block weight of the Stable Diffusion model and uses them as trainable parameters. As illustrated in Figure 4,

- x is the input features map,
- c is the internal vector calculated based on x, ControlNet of input,
- y is the output feature map.

The trainable parameters are connected to the base model with zero convolution layers, denoted $Z(\cdot;\cdot)$. Specifically, $Z(\cdot;\cdot)$ is a 1 x 1 convolution layer with both weight and bias initialized to zeros. To build up a ControlNet, two instances

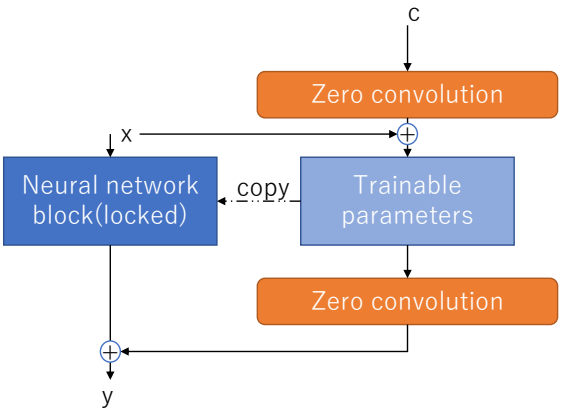


Figure 4 ControlNet neural network

of zero convolutions with parameters Θ z1 and Θ z2, respectively, are used. Although convolution layers of weight and bias are initialized to zeros, ControlNet still learns and updates parameters. We demonstrate the following formula:

$$y = Wx + b, \tag{1}$$

$$\frac{\partial y}{\partial w} = x, \tag{2}$$

$$\frac{\partial y}{\partial x} = W, \tag{3}$$

$$\frac{\partial y}{\partial b} = 1, \tag{4}$$

$$\tag{5}$$

where y denotes output, W denotes weight, x denotes input, b denotes bias. If W and b are zero and x is not zero, y is different from zero. This means that when images are input, the weight and bias of ControlNet are grown up. When training ControlNet, the weight and bias of the stable diffusion model are locked, and only trainable weights are updated by gradient descent. ControlNet offers various conditional controls in the Stable Diffusion model. In our experiment, we used the ControlNet-Canny condition. Via ControlNet-Canny conditional control, we can extract the outline of the subject in each base image, then transfer object style with the Stable Diffusion model and LoRA models.

3.2 Classifier Using Fine-tuned Resnet50

In the context of the classification problem, shallow layers would [16] extract low-level features that are almost the same as the input image. In deeper layers, more abstract features were extracted. The abstract features make the classification possible and easier. Hence, more layers are added to network architecture in the hope that the network can learn more features from input images and boost classification accuracy. However, when more layers are adapted, a degradation problem is caused. This problem makes training errors and test/validation errors go up [17]. By introducing a residual learning framework into CNN, Resnet50 [17] is easier to

optimize and can gain higher accuracy. The deep residual learning framework of ResNet50 is shown in Figure 5. As shown in Figure 5, the x identity map feature is added to network weight before activating network weight forward. Instead of output $F(x)$, the output changes into $F(x) + x$. This makes deep layers learn shallow layers of features and decreases errors in training and test/validation.

In this paper, Resnet50 is adopted to classify LoRA style via transferred images generated from fine-tuned image generative LoRA models. To enhance classification accuracy, pre-train Resnet50 is modified and fine-tuned.

3.3 Classification-based Model Search

In this section, we explain more clearly about our method. First, we collected sample images. The sample image denotes which image we use to transform its style. Sample images are variously selected. Next, we prepared the fine-tuned image generative LoRA models. As mentioned in the previous section, to transfer the style of base images, image-to-image Stable Diffusion model version 1.5, ControlNet canny model, and LoRA models are adopted. At first, the subject in each base image of the outline is extracted and transformed into a vector. We call it a conditional control vector. Then, the Stable Diffusion model version 1.5 and the Canny model of ControlNet are implemented. In the stage of applying the LoRA models style, weight in u-net layers and text encoder layers of LoRA [15] are set with 1. Hence, enabling the LoRAs feature is clearly reflected in the transferred image. The base image and conditional control vector are used when transforming the base image style without any text input. In addition, hyperparameters such as inference steps, guidance scales, etc... are adjusted so that the transformed image better reflects LoRA features.

After transforming the style of base images, the fine-tuning phase is implemented. We fine-tune pre-trained ResNet50 with image classification tasks. Data is transformed style base images, and the label is the name of LoRA models. In our experiment, the training dataset is very tiny compared with the ImageNet dataset used for training ResNet50 [17]. For this reason, only small parts of pre-trained ResNet50 are returned, instead of returning the entire model. As mentioned in [18], deeper CNN layers can extract more abstract features of the image. These features represent a unique style of image. ResNet50 architecture has four major residual layers, which are assigned to extract the abstract features of each image. Moreover, each layer has a different amount of bottleneck [17]. The third layer of Resnet50 has the most number of bottlenecks, with six bottlenecks. Based on the above mentioned, weights between layer 3 and the last fully connected layer are set backpropagation. The primary role of the CNNs is feature extraction. Our main purpose is image classification, thus more emphasis on fully connected layers. One more fully connected layer is added.

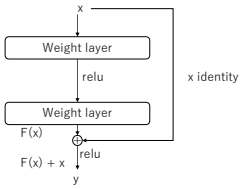


Figure 5 Residual block

4 Evaluation

We evaluate the proposed retrieval system through three factors. First, we tried to clarify whether the designed fine-tuning is effective or not. Second, we evaluated what kind of dataset can be used to obtain good performance for the proposed system. Finally, we evaluated how many style-transformed images should be included in each training data class to ensure sufficient learning.

4.1 Fine-tuning result

This section introduces the evaluation method for the proposed method. The base images are downloaded from the Internet. We call this the base image set as set B. The set B is processed to include as many different types of images as possible. Images of people, scenery, food, plants, animals, etc., are included in the set B. Each element in the set B is unique. The set B is transformed style by group models, which are the stable diffusion, canny, and LoRA models. With each image in the B set creates one image in the F set. After transforming the style, we create a transformed style image set called the F set. The size of B set and F set are the same, both set include 126 images. Such as, $(\forall x \in B | \exists x' \in F, |B| = |F|)$, where

- x : base image,
- B : base image set,
- x': transformed style image, and
- F: transformed style image set.

In the fine-tuning phase, 80 percent of the F set is trained, and 20 percent of the set is used for validation.

Table 1 shows the list of models that we used in our experiment. As the search target, 12 LoRA models are prepared. All these LoRa models are downloaded from Civitai. In the experiment, 15 epochs and a 0.001 learning rate are used. The number of nodes in the output of the fully connected layer is modified to match the number of LoRA models we wish to classify.

When returning, early stopping with five patients and the learning schedule CosineAnnealingLR are utilized to obtain the best performance fine-tuned ResNet50 model.

As shown in Figure 6, training and validating results are

Table 1 Models used in experiment

Model name
Anime Screencap Style LoRA (注2)
Anime Game Backgrounds(注3)
xiaorenshu(注4)
Chinese ink painting(注5)
Disco Elysium / Aleksander Rostov / Oil paints & Abstraction / Style(注6)
Fairy in Clouds(注7)
XSarchitectural-11Fantasyarchitecture(注8)
Phantasmal Luminous: The Radiance of Rainbow Dispersion(注9)
8bitdiffuser 64x — a perfect pixel art model(注10)
pop art(注11)
Texture illustration(注12)
Realistic Dating attire(注13)

plotted. When fine-tuning pre-trained ResNet50 with 12 LoRA models, the best result was obtained in the 6th epoch. In the loss plot, the learning loss decreases gradually as the number of learning epochs increases. Nevertheless, the validation loss decreases from the first to the third epoch, but becomes very unstable from the fourth epoch to the end. The same phenomenon is observed in the accuracy plot.

Then, to evaluate the effect of fine-tuning, we compare the classification performance of the fine-tuned ResNet50 with the classification results of the pre-trained ResNet50. Here, we use Accuracy and Macro-F1. In each experiment, shown in 2, the same set of validation data is used. Also, the number(100, 26), in second and third column, denoting amount of images included in each class.

With pre-trained ResNet50, the output of a fully connected layer is modified from 1000 to 12 to adapt with our task. When inferring, weights of pre-trained ResNet50 are frozen. The detailed result is shown in 2. The our ResNet50 performs better than the pre-trained ResNet50 in the three measures of Accuracy, Macro-F1 and MRR. We calculated the Macro-F_i score as

$$Macro - F1 = \frac{1}{N}\sum_{i=0}^{N}(\frac{2TPi}{2TPi + FPi + FNi}) \qquad (6)$$

where

- N: Amount of models(N = 12),
- TP: True positive,
- FN: False negative, and
- FP: False positive.

Via Accuracy and Macro-F1, we can evaluate classification performance of ResNet50. Via MRR, we can evaluate capability retrieval model of proposed system.

In addition, the more images included in each class, the better the system's learning performance. However, once a certain amount is reached, the change becomes small. Specifically, even if 70 or more images are included in each class, no significant change in learning accuracy occurs. In conclusion, 70 images in each class are sufficient to train the proposed retrieval system.

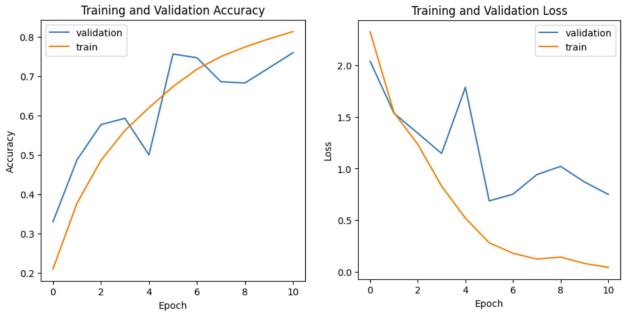


Figure 6 12 models of classification result

4.2 Experiment Model Retrieval

We prepare more than one sample image dataset to evaluate the importance of training data variety. In the above dataset, we call it a variety dataset, which includes a variety of images such as humans, animals, landscapes, *etc.*. The new one, we call it a specialized domain image dataset that contains only human face images. Each image in the domain image dataset is processed like an image in the variety dataset.

The result is shown in table 2 Fine-tuning the pre-trained ResNet50 using the same number of data, the retrieval system outperforms when using diverse data. In conclusion, diverse data images are important for building the proposed retrieval system.

A fine-tuned ResNet50 model with 12 models is also used to find LoRA models. Found models are ranked based on the output of the second fully connected layer of ResNet50. Images are used in the experiment, not included in the training or validating sets.

To enable us to evaluate objectively, these images are downloaded from the site called Civitai, and they are thumbnail images of the original LoRA models created by authors.

Table 2 Evaluation Results

	Train data	Validation data	Accuracy	Macro-F1	MRR
Fine-tuned ResNet50	100 variety images	26 variety images	79.50	74.37	60.92
Pretrained ResNet50		26 variety images	8.30	Nan	
Fine-tuned ResNet50	70 variety images	26 variety images	76.90	70.17	62.15
Fine-tuned ResNet50	50 variety images	26 variety images	67.90	65.50	55.09
Fine-tuned ResNet50	30 variety images	26 variety images	62.20	60.48	60.42
Fine-tuned ResNet50	70 human face images	26 variety images	50.30	48.33	47.92

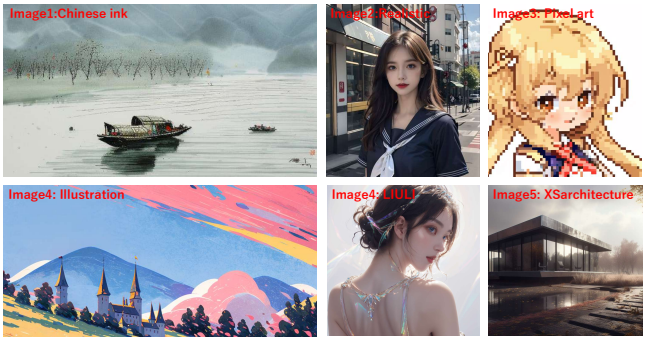


Figure 7 Experimental images

Figure 8 Image generation sample

Experimental images are shown in Figure 7. The red letters in the top left of each image are the names and correct labels of each image. A correct label indicates that the image is generated from that LoRA model.

Finding model results by a sample image is shown in Table 3. Via our method, among the six sample images, three sample images can accurately retrieve the original LoRA model. When the LoRA model is retrieved by feeding the experimental images into the proposed system, the correct results are included within the top six.

4. 3 Image generation

This section uses the retrieved LoRA models from the previous section to generate images. Following the result of Table 2, the search result is the most undesirable when using image 2 to find the LoRA model that generated image 2. The desirable search result is located in the sixth position. In the experiment of image generation, the first and sixth models are used to generate images and compare them. Figure 9 shows the four images in our experiment, such as

Depending on each LoRA model, the generated image does not exactly resemble the base image. However, they were converted to their respective painting styles. Indeed, the style of the image transformed by the Realistic LoRA model is more similar to the style of the reference style image than the style of the image transformed by the XSarchitecture LoRA model. However, when users want to generate images, they want the generated images to follow their desired style, and the generated images are beautiful. In this experiment, in the two generated images, we are not able to say that the image transformed by the Realistic LoRA model is more aesthetic than the image transformed by the XSarchitecture LoRA model.

5 Disccusion

Through this experiment, it is certainly possible for the proposed method to search for fine-tuned image-generative LoRA models. However, to retrieve the LoRA model accurately, the number of LoRA models fed in ResNet50 is limited. Furthermore, since the models are retrieved through style-transformed images, amount of models that can be retrieved may depend on the process of image style transformation. There are many hyper-parameters that need adjusting during the process of transforming the style of the image. Thus, when generating more images, the image generation process is more complicated and not sturdy.

Moreover, the number of fine-tuned LoRA models gradually increases, and in this proposal, ResNet50 will need to return when new fine-tuned LoRA models are released. This is inefficient. ResNet50 is specialized to extract image style features. Although the original ResNet50 is modified for best performance on a given task, its functionality is still limited when a number of LoRA models are added to ResNet50. We have confirmed that the capability of ResNet50 is not appropriate for our research. We need to investigate more effective methods in the future.

One of effective method is the Siamese Network, employing a loss function called contractive loss or triplet loss during

Table 3 Result of retrieving model

	1st	2nd	3rd	4th	5th	6th	7th	8th	9th
Image1	**Chinese ink**	XSarchitecture	Illustration	Anime	Pop art	Realistic	Pixel art	LIULI	Cloudy
Image2	XSarchitecture	Anime	cloudy	Illustration	Chinese ink	**Realistic**	Pixel art	LIULI	Anime background
Image3	**Pixel art**	Pop art	Illustration	Anime background	Realistic	Chinese style	XSarchitecture	Anime	LIULI
Image4	LIULI	**Illustration**	Anime background	Cloudy	Pop art	Disco	XSarchitecture	Anime	Realistic
Image5	Cloudy	XSarchitecture	Illustration	**LIULI**	Anime	Pop art	Anime background	Realistic	Chinese ink
Image6	**XSarchitecture**	Cloudy	Anime	Illustration	Chinese ink	Realistic	Pixel art	LIULI	Pop art

training. The Siamese Network takes two or three images as input values and calculates the distance between them. Based on distance, determine if they are the same image or not. When applying to our research, we assume that one image is the style the user wants and the other is a thumbnail image of the original LoRA model. The above process makes it possible to eliminate the work of image style transformation.

6 Conclusion

This paper proposes a method for retrieving fine-tuned image-generating LoRA models. Through our method, it is possible to output a LoRA model that can generate an image style similar to the style of the reference image. A dataset of various base images is prepared to implement the proposed method. The style of the sample images is transformed by using Stable Diffusion version 1.5, ControlNet-Canny, and fine-tuned image generation LoRA models. The style-transformed images are used as fine-tuning data for the pre-trained ResNet50. Several parts of the original ResNet50 architecture are modified to adapt it to our task. Experiments are conducted to compare pre-trained ResNet50 and fine-tuned ResNet50 and search fine-tuned image generation LoRA models.

Acknowledgement

This work was supported by JSPS KAKENHI Grants Numbers 21H03775, 21H03774, and 22H03905.

References

[1] Gabriele Valvano, Antonino Agostino, Giovanni De Magistris, Antonino Graziano, and Giacomo Veneri. Controllable image synthesis of industrial data using stable diffusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5354–5363, 2024.

[2] Jordan Shipard, Arnold Wiliem, Kien Nguyen Thanh, Wei Xiang, and Clinton Fookes. Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 769–778, 2023.

[3] Quentin Jodelet, Xin Liu, Yin Jun Phua, and Tsuyoshi Murata. Class-incremental learning using diffusion model for distillation and replay. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3425–3433, 2023.

[4] Joanna Kaleta, DiegoDall ' Alba, Szymon Płotka, Przemysław Korzeniowski. Minimal data requirement for realistic endoscopic image generation with stable diffusion. *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–9, 2023.

[5] Tiancheng Sun, Yulong Wang, Jian Yang, and Xiaolin Hu. Convolution neural networks with two pathways for image style recognition. *IEEE Transactions on Image Processing*, Vol. 26, No. 9, pp. 4102–4113, 2017.

[6] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. *Proc. Brit. Mach. Vis.Conf. (BMVC)*, 2014.

[7] Kazuma Kondo and Tatsuhito Hasegawa. Cnn-based criteria for classifying artists by illustration style. IVSP '20, p. 93–98, New York, NY, USA, 2020. Association for Computing Machinery.

[8] Fang Ji, Michael S McMaster, Samuel Schwab, Gundeep Singh, Lauryn Nicole Smith, Shishir Adhikari, MarcioO ' Dwyer, Farah Sayed, Antony Ingrisano, Dean Yoder ほか. Discerning the painter ' s hand: machine learning on surface topography. *Heritage Science*, Vol. 9, No. 1, pp. 1–11, 2021.

[9] Farhan Ullah, Bofeng Zhang, and Rehan Ullah Khan. Image-based service recommendation system: A jpeg-coefficient rfs approach. *IEEE Access*, Vol. 8, pp. 3308–3318, 2020.

[10] Ouhda Mohamed, El Asnaoui Khalid, Ouanan Mohammed, and Aksasse Brahim. Content-based image retrieval using convolutional neural networks. In *Lecture Notes in Real-Time Intelligent Systems*, pp. 463–476. Springer, 2019.

[11] Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, p. 157–166, New York, NY, USA, 2014. Association for Computing Machinery.

[12] Jin Ma, Shanmin Pang, Bo Yang, Jihua Zhu, and Yaochen Li. Spatial-content image search in complex scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2503–2511, 2020.

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685. IEEE, 2022.

[14] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[16] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5188–5196, 2015.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

[18] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

(注2)：civitai: https://civitai.com/api/download/models/60568?type=Model&format=SafeTensor
(注3)：civitai: https://civitai.com/api/download/models/19065?type=Model&format=SafeTensor&size=full&fp=fp16
(注4)：civitai: https://civitai.com/api/download/models/25661?type=Model&format=SafeTensor&size=full&fp=fp16
(注5)：civitai: https://civitai.com/api/download/models/42314?type=Model&format=SafeTensor
(注6)：civitai: https://civitai.com/api/download/models/169568?type=Model&format=SafeTensor
(注7)：civitai: https://civitai.com/api/download/models/269819?type=Model&format=SafeTensor
(注8)：civitai: https://civitai.com/api/download/models/30572
(注9)：civitai: https://civitai.com/api/download/models/180175
(注10)：civitai: https://civitai.com/api/download/models/231819?type=Model&format=SafeTensor
(注11)：civitai: https://civitai.com/api/download/models/188417?type=Model&format=SafeTensor
(注12)：civitai: https://civitai.com/api/download/models/113534?type=Model&format=SafeTensor
(注13)：civitai: https://civitai.com/api/download/models/228470?type=Model&format=SafeTensor

生成型要約に基づくWebページのサムネイル生成

前田　直宏†　　山本　岳洋†

† 兵庫県立大学 大学院情報科学研究科　〒651–2197 兵庫県神戸市西区学園西町 8-2-1
E-mail: †ad22p062@gsis.u-hyogo.ac.jp, ††t.yamamoto@sis.u-hyogo.ac.jp

あらまし　本研究では Text2Image モデルを用いて Web ページの主題を表したサムネイルを生成する手法を提案する．具体的には，モデルに入力するプロンプトを Web ページを要約することで複数生成し，（1）記事の主題を表している，（2）画像として自然な出力が期待できる，という 2 点の観点から適したプロンプトを分類して選択する．この条件を満たすプロンプトを取得するために本研究では，ランダムフォレストを用いて分類モデルを構築する．実験の結果，自然言語処理タスクで高い性能を示す BERT よりも高い精度で分類を行うことができた．また，構築した分類モデルを用いて CNN ニュース記事と旅行ブログなどの Web ページを対象としたサムネイル生成を行い，人手による評価を行う．サムネイル評価実験の結果，記事のタイトルをプロンプトとするベースラインに比べ，提案手法によって生成されるサムネイルの方が記事のサムネイルに合っているとして好まれた．

キーワード　情報検索，Text2Image，プロンプトエンジニアリング，画像生成，要約文生成

1　はじめに

Web ページの概要を表すサムネイルは，ユーザ所望の情報を獲得する助けになる．例えば検索結果の表示をテキストベースだけではなく，Web ページ内の画像やデザインを用いた 1 枚のサムネイルも表示することで，探している情報と関係のある情報をより正確に獲得することが可能である [4]．また，一度閲覧した Web ページの約 80% は再度閲覧することが明らかになっている [1] [3]．ブラウザに備わったブックマーク機能や閲覧履歴，記憶を辿ることで再度閲覧したページにたどり着くことができる一方，検索結果表示欄にサムネイルも表示することで，一度訪れたサイトへの再訪問が容易に行えることが報告されている [22]．このようにユーザが情報を獲得するとき，Web ページの概要を表したサムネイルは情報検索の効率を上昇させる．

Web ページの内容を表したサムネイルの生成方法としては，Web ページの上部のスクリーンショットにロゴを付与するものや，ページ内に掲載されている写真を組み合わせたもの，インターネット上からその Web ページと関連性が高い画像を用いたものなど様々なサムネイルが提案されている [9] [15] [22]．また近年の検索エンジンでは，Web ページの上部に掲載されているアイキャッチ画像とも呼ばれる画像をサムネイル代わりとして使用されることもある．このアイキャッチ画像はフリー素材サイトなどで取得する必要があるが，必ずしも記事の内容に合った画像が存在するわけではない．また内容と異なる画像を用いた場合，誤った情報の獲得を促してしまう可能性がある．そのため，Web 検索におけるサムネイルは記事の内容や主題に合っていることが重要である．

そこで本研究では Text2Image モデルを用いて，Web ページの主題を表したサムネイルの生成を行う．Text2Image モデルによる生成画像は，入力されたプロンプトと呼ばれるテキストに基づいた画像が生成される．そのため本研究では Web ページの記事テキストをもとにプロンプトを生成し，得られたプロンプトを用いてサムネイルの生成を行う．具体的には，まず記事テキストの要約文を複数生成する．その後，（1）記事の主題を表している，（2）画像として自然な出力が期待できる，という 2 点の観点からサムネイル生成に適したプロンプトを分類して取得する．このプロンプト分類モデルとして本研究ではランダムフォレストを構築する．そして得られたプロンプトを Text2Image モデルに入力することでサムネイルを生成する．

2　関連研究

2.1　サムネイルを用いた情報検索支援

Web ページのサムネイルを表したサムネイルの研究は多く行われている．Zhou らは Web ページを検索する時に使用したクエリを強調するサムネイルを提案している [23]．Teevan らは Web ページのタイトル，ロゴ画像，注目度が高い画像を用いたサムネイルを提案している [22]．また，稲垣らは Web ページに対する顕著性マップを用いたサムネイルを提案している [25]．顕著性マップとは，人が画像を認識する際に注視しやすい領域を画像で定量的に表現したものである．Web ページを閲覧する際に注視されやすい上位 10 の領域を 1 枚の画像に並べて配置した集約画像を生成することで，ページ内容の理解補助に繋げている．これらのサムネイルで使用している画像は，Web ページ内に含まれる画像を抽出し使用している．しかし内部画像が存在しない場合，サムネイルを作成することが困難であった．そこで Jiao らはインターネット上からその Web ページと関連性が高い画像を取得し，その画像を用いたサムネイルを提案している [9]．Web 検索時以外にも Web ページの概要を表したサムネイルの利用方法が提案されている．例えばサムネイルを Web ブラウザのタブ管理に使用することが提案されている [12]．Web ページの左右余白部分に閲覧した Web ページのサムネイルを配置することで，一度閲覧した Web ページへの

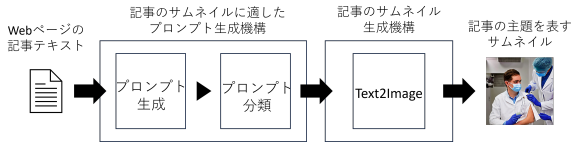


図 1　記事の主題を表したサムネイルの生成手法.

素早い切り替えを行うことができる.

2.2　画像生成モデル

画像生成モデルとしては，VAE（Variational AutoEncoder）[10] や GAN（Generative Adversarial Network）[6]，拡散モデル [8] などの手法が数多く提案されていた．VAE はエンコーダとデコーダで構成されており，エンコーダで入力データを潜在変数に落とし込み，デコーダで潜在変数から画像を生成する．GAN は敵対的生成ネットワークとも呼ばれ，生成器と識別器からなる 2 つのネットワーク構造を持つ．生成器が偽のデータを生成し，識別器が正解データと偽データを見分ける学習を行うことで生成器が本物に近く質の高い画像を生成することができる．拡散モデルは画像にノイズを付与する拡散過程と，ノイズを取り除いて画像を生成する逆拡散過程に分かれており，高い質の画像を生成することができる．また近年では，テキストから画像を生成する Text2Image モデルが多く提案されており，Stable Diffusion [20] や DALL-E2 [18] など質の高い画像を生成することができる．これらの画像生成するモデルでは，プロンプトと呼ばれるテキストを入力する必要があり，プロンプトの質が生成画像の質に大きく影響する．そのため，プロンプトの質を高めるプロンプトエンジニアリングの手法も多く提案されている.

2.3　サムネイル生成とプロンプトエンジニアリング

Text2Image モデルの登場により，テキストからサムネイル生成を行う研究が行われてきている．Liu らはニュース記事の挿絵を生成した [14]．Web 検索以外においてもサムネイル生成の研究が行われている．Shinomoto らは，ニュース記事に対してミーム画像を生成する手法を提案している [21]．また，プロンプトエンジニアリングによるサムネイル生成を行った研究もおこなわれている．Hao らは強化学習なども用いてプロンプトの最適化を行った上で画像の生成を行っている [7]．Liu らは，プロンプトの様々な条件のプロンプトで画像を生成することによってプロンプトと生成画像の関係を調査している [13].

3　提 案 手 法

本研究では Web ページの記事から，その記事の主題を表したサムネイルを生成する．提案手法の全体像を図 1 に示す．サムネイルの生成手法としては，テキストから画像を生成する Text2Image モデルを用いる．この Text2Image モデルには出力したい画像に対するプロンプトを入力する必要がある．そこで記事本文を要約することで Text2Image モデルに入力するプロンプトを複数生成する．複数生成されたプロンプトの内，サムネイル生成に適したプロンプトを分類し，最も適したプロンプトを Text2Image モデルに入力することで記事の主題を表したサムネイルを生成する．次節では Text2Image モデルに入力するプロンプトを生成するプロンプト生成機構と，記事のサムネイル生成機構の 2 つについて説明する.

3.1　記事本文の要約によるプロンプト生成

サムネイル生成を行うために必要であるプロンプトの生成について説明する．Web ページの記事本文には多くの情報が含まれているため，Text2Image モデルに全ての文を入力することによって記事の主題を表した画像を生成することは難しい．そこで本研究ではタイトルを除いた記事本文を T5 [17] を用いて要約することで，記事の要点を短くまとめたプロンプトを生成する．しかし，Web ページの記事は章立てられていることが多く要点が多く存在すると考えられる．そのため，全文をまとめて要約するのではなく全文をいくつかの文章に分け，各文章を要約することでプロンプトを生成する．本文を区切って入力する理由としては 2 つある．1 つ目は T5 の入力トークン数には上限があり，本文全文を入力することが困難な場合があるからである．2 つ目は Web ページの文章は章立てられていることが多く，文の位置によって述べられているトピックが異なるからである．あまりにも多くのトピックが含まれた文章の要約を行うと情報量が減少してしまう恐れがあるため，数文に区切って入力する.

記事本文 article を文 s_i の系列として以下のように表現する.

$$\text{article} = s_1 s_2 \dots s_n \tag{1}$$

ここで n は記事本文 article の文数を表す．ただし，1 文はピリオド，エクスクラメーションマーク，クエスチョンマークで区切る．また，記事本文から系列の一部を取り出す関数である窓関数を以下に定義する.

$$\text{window}(\text{article}, i, j) = s_i s_{i+1} \dots s_{i+j-1} \tag{2}$$

この窓関数を用いて j 文で構成される文章 sentence_i を以下に定義する.

$$\text{sentence}_\text{i} = \text{window}(\text{article}, i, j) \tag{3}$$

ただし，i は取り出したい部分の開始位置を表す.

こうして得られた各文章 sentence_i を要約することでプロンプトを生成する.

3.2　サムネイル生成に適したプロンプトの分類

3.2.1　問 題 定 義

Text2Image モデルによって生成される画像の質は，入力するプロンプトの質に大きく依存する．そのため，画像で表現することが難しいプロンプトを入力すると質の低い画像が生成される．例えば「方法」や「安い」といった単語を画像で表現することは難しい．そのため入力として適するプロンプトを選ぶ必要がある．また，3.1 節で生成された複数のプロンプト内には，記事の主題を含んでいるプロンプトと含んでいないものが

表 1　ランダムフォレストで使用する特徴量.

特徴量変数名	説明
interrogative_words	疑問詞を含むか
word_length	単語数
title_prompt_sim	タイトルとプロンプトの類似度
LexRank	LexRank
adjective_count	形容詞の数
noun_count	名詞の数
proper_noun_count	固有名詞の数
verb_count	動詞の数
proportion_adjective	形容詞の割合
proportion_noun	名詞の割合
proportion_proper_noun	固有名詞の割合
proportion_verb	動詞の割合
concreteness	各単語の具体度の平均

あると考えられる．本研究の目的は記事の主題を表したサムネイルの生成であるため，記事の主題を含んでいないものはサムネイル生成に適していないプロンプトである．そこで，プロンプトの良し悪しを分類する分類モデルを構築することで，記事の主題を表したサムネイル生成に適したプロンプトを選ぶ．ただし，ここでの良いプロンプトとは以下の 2 つを十分に満たすものである．

- 記事の主題を含んだプロンプトである．
- 画像で表現することが可能なプロンプトである．

本研究ではプロンプトの良さを 3 段階に分割し，以下の様にラベルを定義する．

$$\text{label} = \begin{cases} l_1 & \text{（サムネイル生成に適さないプロンプト）} \\ l_2 & \text{（どちらでもないプロンプト）} \\ l_3 & \text{（サムネイル生成に適したプロンプト）} \end{cases} \quad (4)$$

プロンプト分類モデルに入力された複数のプロンプトから，最もサムネイル生成に適したプロンプト l_3 を 1 つ取得することが目標である．

3.2.2　プロンプト分類モデル

本研究では，プロンプトの分類モデルとしてランダムフォレストを用いる．ランダムフォレストは複数の決定木を用いたアンサンブル学習法の 1 つである．数値やカテゴリといった様々な種類の特徴量を扱うことができるため，良いプロンプトの条件を種類の異なる複数の特徴量で表現することができると考えられる．ランダムフォレストで使用する特徴量を表 1 に示す．特徴量の選択は主に 3 つの軸に基づき決定した．1 つ目は記事の主題を表すかである．特徴量として，LexRank [5] とタイトルとプロンプトの類似度を選択した．LexRank とは，文書から各文をノードとみなしたグラフ構造を作り出し，各ノード間の類似度を計算することでそれぞれの文の重要度を算出するアルゴリズムである．本研究では 1 つのノードに 1 文を割り当てるのではなく，プロンプト生成に用いた要約元の文章を 1 つのノードとして類似度を計算する．要約元の文章の重要度を計算することで，各プロンプトの文書内での重要度を算出する．これらの文章間の類似度を計算するには Sentence-BERT[1] [19] を用いてベクトル化を行い，コサイン類似度を計算する．Web ページのタイトルは文書の内容を簡潔に表していることがある．特にニュース記事などは読者に記事の情報を素早く理解してもらうために，タイトルが内容を簡潔に表現していると考えられる．そこでタイトルとプロンプトの類似度を特徴量に用いた．類似度は，タイトルとプロンプトのコサイン類似度を計算することで算出する．2 つ目は画像として自然な出力が期待できるかである．画像で表現することが容易な単語は具体度が高いと考えられるため，特徴量としてプロンプトに含まれる単語の具体度を用いる．具体度の計算には，様々な単語に具体度が付与された辞書 [2] を使用する．これは 40,000 単語に対して 1 から 5 の具体度を付与したものである．各プロンプトに対する具体度を以下の様に定義する．

$$\text{concreteness}_i = \frac{\sum_i^n \text{conc}(w_i)}{n} \quad (5)$$

ここで w_i はプロンプトを構成する単語を表し，$\text{conc}(w_i)$ は単語 w_i の具体度を返す関数である．ただし，具体度は 1 から 5 の間の実数であり，最も具体度が高い場合 5 を取る．3 つ目は，1 つ目と 2 つ目の両方を含むもしくはそれ以外に当てはまるものである．特徴量として，疑問詞の有無，プロンプトの単語数，品詞の数と割合を選択した．プロンプトに疑問詞が含まれる場合，例えばプロンプトに「何」,「なぜ」,「どこで」といった疑問詞を含む場合，表現することが困難であると考えられる．そのため疑問詞の有無を特徴量に用いた．またプロンプトの長さは生成画像の質に影響することが報告されている [24]．そのため特徴量にプロンプトの長さを用いた．ただしプロンプトの長さとはトークン数を表す．また，プロンプト中に名詞や固有名詞が含まれていなければ抽象度が高く，うまく画像を生成できないと考えられる．同様に形容詞や動詞が含まれていない場合でも抽象度が高まると考える．そのため，品詞の数とトークン数に対する割合を特徴量に選択した．

3.3　Web ページの主題を表したサムネイル生成

分類モデルによって得られたサムネイル生成に適したプロンプトを用いて，Web ページの内容に合ったサムネイルの生成を行う．サムネイルの生成手法としては，テキストから画像を生成する Text2Image モデルを用いる．

4　評 価 実 験

4.1　サムネイル生成に適したプロンプト分類の評価

構築したプロンプト分類モデルの性能評価を行う．分類モデルとしてはランダムフォレスト，XGBoost，BERT を用いる．

4.1.1　データセット

プロンプトを生成するためのデータセットとして CNN ニュース記事を用いる．CNN はアメリカ合衆国の国際的なケーブルテレビニュースネットワークであり，全世界にニュースを伝え

1：https://huggingface.co/sentence-transformers/all-mpnet-base-v2

表 2　ランダムフォレストの学習に用いたハイパーパラメータ.

パラメータ	値
決定木の深さ	20
葉に必要な最小サンプルサイズ	1
分割後ノードの最小サンプルサイズ	2
決定木の数	300

ている．そのため，記事に対するサムネイルやタイトルなどの質は高いと考えられる．CNN ニュースサイトからニュース記事 577 件を収集し，各ニュース記事のページから URL，タイトル，記事本文，サムネイルを取得した．これらを取得する際には Python のライブラリである Newspaper3k[2]を使用した．

4.1.2　実 験 条 件

収集したニュース記事の本文を用いてプロンプトのデータセット構築を行う．プロンプトの生成を行うための言語生成モデル T5 は，t5-one-line-summary[3]を使用する．T5 へ本文を 3 文ずつに区切って入力することで記事についてのプロンプトを複数生成する．生成された全プロンプト数は 7,255 である．生成された各プロンプトに対してプロンプトの良さを表すラベルを付与した．記事に掲載されているサムネイルとプロンプトから生成される画像との類似度を計算し，その値によってラベルを付与する．類似度が 0.0〜0.4 であればサムネイル生成に適さないプロンプト，0.4〜0.7 であればどちらでもないプロンプト，0.7〜1.0 であればサムネイル生成に適したプロンプトとラベルを付与する．ただし，プロンプトから画像を生成するモデルとして Stable Diffusion [20] を使用した．また，類似度を計算するときには 1 つのプロンプトから 5 枚画像を生成し，類似度の平均をとった．複数画像を生成することで，生成のたびに画像の質が異なることを考慮した．画像同士の類似度を計算するために，CLIP [16] を使用する．

4.1.3　ランダムフォレスト

プロンプトを分類する 1 つ目のモデルは，ランダムフォレストである．使用する特徴量として，表 1 を用いた．学習データ，テストデータを 8:2 の比率で分割し，さらに学習データを 8:2 の比率で学習データと検証データに分割した．学習データと検証データを用いてグリッドサーチと交差検証を行い，ハイパーパラメータのチューニングを行った．求まったハイパーパラメータを表 2 に示す．チューニングを行った後，学習データと検証データを用いて再学習を行った．

4.1.4　XGBoost

2 つ目のモデルは XGBoost である．使用した特徴量はランダムフォレストと同様に表 1 を用いた．データについてもランダムフォレストと同様に学習データ，テストデータを 8:2 の比率で分割し，さらに学習データを 8:2 の比率で学習データと検証データに分割した．学習データと検証データを用いてグリッドサーチと交差検証を行い，ハイパーパラメータのチューニングを行った．求まったハイパーパラメータを表 3 に示す．チューニングを行った後，学習データと検証データを用いて再学習を行った．

表 3　XGBoost の学習に用いたハイパーパラメータ.

パラメータ	値
学習率	0.2
決定木の深さ	5
決定木の数	300
各決定木のサンプル抽出比	0.9

表 4　BERT のファインチューニングに用いたハイパーパラメータ.

パラメータ	値
バッチサイズ	16
学習率	2×10^{-5}
早期終了	3 epoch
最適化手法	Adam
損失関数	交差エントロピー

4.1.5　BERT

3 つ目のモデルは BERT である．事前学習済みモデルとして，インターネット上の大規模なテキストコーパスを用いて学習された bert-base-uncased[4]を使用し，この事前学習済みモデルをファインチューニングして使用する．ファインチューニングのデータとしては，4.1.1 節で述べたデータセットを用いる．ファインチューニングに用いたハイパーパラメータを表 4 に示す．BERT への入力は，プロンプトである．先頭には特殊トークンの [CLS] を付与し，プロンプトの末尾には [SEP] を付与して入力する．ただし入力トークン長が最大入力長に満たない場合，特殊トークンの [PAD] で埋める．

4.1.6　評 価 尺 度

プロンプトをクラス l_k と分類するタスクにおいて，実際にクラス l_k である場合を正例，そうでない場合を負例とする．またクラスが l_k であるプロンプトを実際に正例とモデルが予測できたときを真陽性（TP），負例と予測したときを偽陰性（FN）とし，クラスが l_k でないプロンプトを正例とモデルが予測した時を偽陽性（FP），負例と予測したときを真陰性（TN）とする．評価尺度としては適合率（Precision），再現率（Recall），F 値のマクロ平均を用いる．適合率 P，再現率 R，F 値 F は以下の式で求められる．

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{7}$$

$$F = \frac{2\mathrm{PR}}{\mathrm{P} + \mathrm{R}} \tag{8}$$

マクロ平均とは，各クラスの評価指標の平均である．マクロ平均適合率 P_{macro}，マクロ平均再現率 R_{macro}，マクロ平均 F 値 F_{macro} は以下の式で求められる．

2：https://newspaper.readthedocs.io/en/latest/

3：https://huggingface.co/snrspeaks/t5-one-line-summary

4：https://huggingface.co/bert-base-uncased

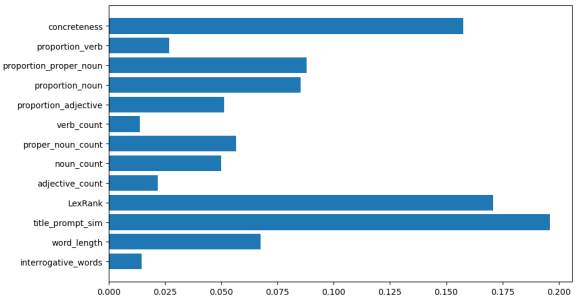


図 2　ランダムフォレストにおける各特徴量の重要度.

$$P_{\text{macro}} = \frac{1}{K}\sum_{i}^{K} P_i \tag{9}$$

$$R_{\text{macro}} = \frac{1}{K}\sum_{i}^{K} R_i \tag{10}$$

$$F_{\text{macro}} = \frac{1}{K}\sum_{i}^{K} F_i \tag{11}$$

ここで K は分類するクラス数であり，$K = 3$ である.

4.2　サムネイル生成に適したプロンプトの分類実験結果

各モデルによるプロンプトの分類結果を表 5 に示す．実験の結果，ランダムフォレストはマクロ平均適合率，マクロ平均再現率，マクロ平均 F 値全てにおいて最も高い値であった．また，良いプロンプトであるラベル l_3 の適合率，再現率，F 値も最も高い値であった．反対に自然言語処理タスクにおいて高い性能を示す BERT は最も低い値であった．深層学習に基づく分類手法よりも特徴量による分類の方が高い精度で分類を行えたことがわかる．また，最も分類精度が高かったランダムフォレストにおける各特徴量の重要度を図 2 に示す．図 2 より，タイトルとプロンプト類似度，LexRank，各単語の具体度の平均が特に重要な特徴量であったことがわかる.

4.3　Web ページの主題を表したサムネイル生成の評価

構築したプロンプト分類モデルを用いて選ばれたプロンプトを使用し，記事に対するサムネイルの生成を行う．使用するプロンプト分類モデルとしては，分類精度が最も高かったランダムフォレストを使用する．モデルがサムネイル生成に適したプロンプトと最も高い確率で予測した 1 つのプロンプトを画像生成モデルに入力する．もしプロンプトの中にサムネイル生成に適したプロンプトと予測したものが無い場合，最も高い確率でどちらでもないプロンプトと予測した 1 つのプロンプトを画像生成モデルに入力することで記事に対する画像を生成する．また比較手法として記事のタイトルをプロンプトとし，タイトルを画像生成モデルに入力することで記事に対する画像を生成する．本評価実験では，提案手法で生成された画像と比較手法で生成された画像について評価を行う.

4.3.1　データセット

CNN ニュース記事と旅行記事が掲載されているブログなどのサイトを用いる．CNN ニュース記事からは，Business，Entertainment，Health，Sport，Travel のジャンルから各 5 記事ずつの計 25 記事を使用する．ただし，CNN ニュース記事は分類モデルの学習，評価に使用していないデータを用いる．また，ブログ等の旅行記事サイトは 14 サイトからランダムに計 25 記事収集し使用する．ただし，ページのタイトルが「人気観光地 10 選」のようなまとめ記事は使用しない．CNN ニュース記事からは Newspaper3k ライブラリを用いて本文を取得し，旅行記事サイトからは手作業で本文を取得した.

4.3.2　実験条件

画像生成モデルとして Stable Diffusion と DALL-E2 を使用する．DALL-E2 は Stable Diffusion と同様にテキストから画像を生成することができるモデルである．1 つの Web ページ記事に対し，Stable Diffusion と DALL-E2 を用いて 1 枚ずつ画像を生成する．すなわち評価者は，1 つの記事に対して（提案手法とベースライン）×（Stable Diffusion と DALL-E2）による生成の 4 枚の画像を評価する．生成された画像について，同研究室の学生 2 名と著者によるアンケートを通じて評価を行う．評価アンケートでは以下の点について評価を行ってもらう.

（1）　画像が記事の主題を表していますか.

（2）　画像として自然ですか.

（3）　記事のサムネイルとしてどちらの画像が適していますか.

評価 1 では，各サムネイルについて記事の主題を表しているかを評価する．評価者には 5 段階評価で回答を求めた．ここでは 5 が最高点とした．評価 2 では，各サムネイルについて，画像としての自然であるかを評価する．不自然な画像とは，画像が表している風景，状態を読み取ることができない画像のことを指す．評価 1 と同様に評価者には 5 段階評価で回答を求めた．ここでも最高点は 5 点とした．評価 3 では，提案手法とベースラインどちらで生成された画像が記事のサムネイルとして適しているかを回答してもらう．なお，評価者にはどの画像がどの手法によって生成されたかなどの情報を公開せずに評価を行う.

4.3.3　評価尺度

評価 1 と評価 2 では，各生成モデルについて評価者の点数の平均を用いて手法の比較評価を行う．Web ページのジャンル g について，評価で用いる点数 Score_g は各記事 Article_i の評価点の平均であり，以下の式で求められる.

$$\text{Score}_g = \frac{\sum_i^n \text{Article}_i}{n} \tag{12}$$

ただし，n はジャンル g に含まれる記事数である.

評価 3 では，各生成モデルについて提案手法とベースラインで好まれた記事数の平均を用いて評価を行う．Web ページのジャンル g について，評価で用いる点数 Score_g は以下の式で求められる.

$$\text{Score}_g = \frac{\sum_i^n f(\text{Article}_i)}{n} \tag{13}$$

$$f = \begin{cases} 1 & (\text{Article}_i\text{のサムネイルとして適している}) \\ 0 & (\text{Article}_i\text{のサムネイルとして適していない}) \end{cases} \tag{14}$$

ここで f はサムネイルが Article_i のサムネイルとして適していれば 1 を返し，適していなければ 0 を返す関数である.

表 5　プロンプト分類モデルの評価結果.

分類モデル	ラベル l_3 の適合率	ラベル l_3 の再現率	ラベル l_3 の F 値	マクロ平均適合率	マクロ平均再現率	マクロ平均 F 値
ランダムフォレスト	**0.756**	**0.523**	**0.617**	**0.796**	**0.683**	**0.725**
XGBoost	0.704	0.478	0.570	0.760	0.659	0.696
BERT	0.593	0.467	0.523	0.670	0.648	0.656

4.4　Web ページの主題を表したサムネイルの評価結果

評価 1 のサムネイルが記事の主題を表しているかの評価結果を表 6，評価 2 のサムネイルが画像として自然であるかの評価結果を表 7，評価 3 の記事に適したサムネイルはどちらであるかの評価結果を表 8 に示す．表 6 より，全体的にはベースラインの方が記事の主題を表すことができていた．しかし Entertainment と Health のジャンルでは，提案手法のサムネイルの方が記事の主題を表すことができていたことがわかる．また Health のジャンルでは，Stable Diffusion と DALL-E2 両方において提案手法のサムネイルが記事の主題を表すことができた．表 7 より，全体的には提案手法とベースラインで画像としての自然さに違いはなかった．しかし Health のジャンルでは Stable Diffusion と DALL-E2 両方において提案手法の方がスコアが高い結果となった．表 8 より，全体的にはベースラインの方が記事のサムネイルとして選ばれる数が多かった．しかし Health のジャンルでは Stable Diffusion と DALL-E2 の両方において提案手法の方が高いスコアになった．また評価 1，評価 2，評価 3 について一致度を算出するために Krippendorff の α 係数 [11] を計算した．求まった α 係数の値を表 9 に示す．

5　議　　論

5.1　サムネイル生成に適したプロンプト分類

サムネイル生成に適したプロンプトの分類では，自然言語処理タスクで高い性能を示す BERT よりも，決定木をベースとしたアンサンブル学習手法であるランダムフォレストや XGBoost の方が高い分類精度を示した．その理由として，プロンプト分類に使用した情報量の差ではないかと考える．BERT の入力はプロンプトのみである．そのため BERT はプロンプトの単語のみの情報を分類に使用している．一方でランダムフォレストや XGBoost はプロンプトのみではなく，プロンプトの元となった文章の LexRank やタイトルといった記事中の情報も用いている．サムネイル生成に適したプロンプトとは，記事の主題を含んでいる且つ画像で表現することが可能なプロンプトである．プロンプトの単語情報だけでなく記事中の情報も用いていることから，サムネイル生成に適した 2 つの条件を考慮することができたため，ランダムフォレストの分類精度の方が高かったのではないかと考える．

5.2　Web ページの主題を表したサムネイル生成

プロンプトの分類を行うにあたり，良いプロンプトとして（1）記事の主題を表している,（2）画像として自然な出力が期待できる，の 2 つの観点を考慮した．そのため（1）を満たす特徴量として LexRank，タイトルとプロンプトの類似度,（2）の特徴量としてプロンプトの具体度が影響すると考えた．サムネイルに対する評価の点数とプロンプトの特徴量について相関係数を計算することで，特徴量の影響を分析した．相関係数の値を表 10 と表 11 に示す．記事の全ジャンルについて相関係数を計算した結果，記事の主題を表すかの点数と LexRank，タイトルとプロンプトの類似度にそれぞれ正の相関がみられた．画像として自然であるかの点数とプロンプトの具体度にも正の相関がみられた．また，Business と Entertainment 以外のジャンルに絞ると全ジャンルに比べて強い相関がみられた．以上のことから，LexRank とタイトルとプロンプトの 2 つの特徴量は類似度は記事の主題を表すかに影響を与え，プロンプトの具体度は画像の自然さに影響を与えることがわかった．

評価 1 と評価 2 において，Entertainment と Health のジャンルでは提案手法が好まれた．この 2 つのジャンルの中でも特に Health ジャンルはベースラインとの差が大きかった．その理由として，プロンプトの具体度が影響していると考える．Text2Image モデルに入力したベースラインの記事タイトルに比べ，提案手法によるプロンプトの方が具体度の平均が約 0.5 高く，これは他ジャンルには見られない傾向であった．以上の理由から Health ジャンルは他のジャンルと異なり，ベースラインよりも提案手法の方がスコアが高くなったと考える．

評価 3 においては，提案手法によるサムネイルよりもベースラインによるサムネイルの方が好まれる結果であった．CNN ニュース記事では大きく差はないが，CNN ニュースの Travel と旅行記事サイトにおいてはベースラインが大きく好まれる傾向がみられた．旅行系記事には様々な風景についての情報などが含まれていた．そのため主題とは少し異なる内容を含むプロンプトも多く生成されたためにプロンプトの選択で誤ったものが選ばれ，ベースラインに勝ることができなかったのではないかと考える．また，Stable Diffusion では提案手法によるサムネイルが多く好まれ，DALL-E2 ではベースラインによるサムネイルが多く好まれた．その理由として，プロンプトのラベル付けを行ったときに用いた画像生成モデルが Stable Diffusion であったためだと考える．プロンプトに対するラベルは，プロンプトから生成された画像と記事に掲載されているサムネイルとの類似度を計算することで 3 段階のラベルを付与した．このときの画像生成モデルとして本研究では Stable Diffuision を使用した．そのため各プロンプトに付与されたラベルは Stable Diffusion を用いて画像生成を行うときに対するプロンプトの良さであると考えられるため，画像生成モデルによって提案手法と比較手法の評価に差が生じたと考える．

5.3　限　界　点

本研究ではサムネイルを生成するにあたり，T5 によって生

表 6　評価 1：サムネイルが記事の主題を表しているか.

生成モデル	手法	Business	Entertainment	Health	Sport	Travel	旅行記事サイト	全体の平均
Stable Diffusion	提案手法	2.87	**4.13**	**3.20**	4.53	2.20	3.93	3.47
	ベースライン	**3.47**	3.93	2.47	**4.60**	**3.93**	**4.10**	**3.89**
DALL-E2	提案手法	**2.6**	3.07	**3.40**	3.87	3.80	3.65	3.02
	ベースライン	2.4	**4.00**	2.87	**4.13**	**4.07**	**3.80**	**3.65**

表 7　評価 2：サムネイルが画像として自然であるか.

生成モデル	手法	Business	Entertainment	Health	Sport	Travel	旅行記事サイト	全体の平均
Stable Diffusion	提案手法	3.67	**4.93**	**4.27**	3.57	4.53	4.51	4.41
	ベースライン	3.67	4.73	3.53	**4.73**	**4.73**	**4.53**	4.41
DALL-E2	提案手法	**3.53**	4.27	**3.93**	3.38	4.33	3.84	4.17
	ベースライン	3.20	**4.73**	3.60	**4.40**	**4.80**	**4.20**	4.17

表 8　評価 3：記事に適したサムネイルはどちらか.

生成モデル	手法	Business	Entertainment	Health	Sport	Travel	旅行記事サイト	全体の平均
Stable Diffusion	提案手法	**2.67**	**2.67**	**2.67**	**2.67**	0.67	8.33	19.7
	ベースライン	2.33	2.33	2.33	2.33	**4.33**	**16.7**	**30.3**
DALL-E2	提案手法	2.33	1.33	**3.33**	1.67	0.67	8.00	17.3
	ベースライン	**2.67**	**3.67**	1.67	**3.33**	**4.33**	**17.0**	**32.6**

表 9　各評価における Krippendorff の α 係数.

生成モデル	評価 1	評価 2	評価 3
Stable Diffusion	0.59	0.24	0.44
DALL-E2	0.50	0.28	0.33

表 10　評価 1 についてのサムネイルに対する評価点と各特徴量間の相関係数.

記事のジャンル	LexRank	タイトルとプロンプトの類似度
全ジャンル	0.306	0.749
Business と Entertainment 以外	0.921	0.929

表 11　評価 2 についてのサムネイルに対する評価点と各特徴量間の相関係数.

記事のジャンル	プロンプトの具体度
全ジャンル	0.310
Business と Entertainment 以外	0.373

成されたプロンプトの中から記事の主題を表している，画像で表現が可能なプロンプトを分類モデルを用いて選ぶという手法を提案した．そのため T5 によって生成されたプロンプトの中に，サムネイルのための優れたプロンプトが存在しているということが前提条件の手法である．したがって T5 によるプロンプトの生成がうまく行われなかった場合，提案手法によるサムネイルの質は悪くなる．本研究の実験と評価では，T5 によって生成された複数のプロンプトの中に優れたプロンプトが存在するという仮定の下で行った．そのため T5 の性能向上，あるいは他の言語モデルの性能向上によって提案手法によるサムネイルの評価結果が改善される可能性がある．

6　まとめと今後の課題

本研究では，サムネイル生成に適したプロンプトの分類を行うことによって画像生成モデルへの入力を選別し，Text2Image モデルを用いて Web ページの主題を表したサムネイル生成に取り組んだ．提案手法のプロンプト分類モデルとして，ランダムフォレストを構築した．プロンプト分類の結果，提案手法の分類精度が最も高く，自然言語処理タスクで高い性能を示す BERT よりも高い分類性能を示した．また，この構築したプロンプト分類モデルによって選ばれたプロンプトを用いて，Web ページのサムネイル生成を行った．CNN ニュースと旅行記事サイトに対して生成をし評価を行った結果，CNN ニュースの内 Travel のジャンル以外では提案手法によるサムネイルが好まれ，Travel と旅行記事サイトではベースラインであるタイトルをプロンプトとしたサムネイルが好まれた．ベースラインが勝った理由として，タイトルが記事の内容を端的に表現しており強力であったことや，プロンプトの誤分類が提案手法で起きていたことなどが考えられる．今後は，生成するプロンプトの質を高める必要があることやプロンプト分類の精度をさらに向上させる必要がある．

謝辞　本研究は JSPS 科学研究費助成事業 JP21H03774, JP21H03775, JP22H03905, による助成を受けたものです．ここに記して謝意を表します．

文　献

[1] Eytan Adar, Jaime Teevan, and Susan T Dumais. Large scale analysis of web revisitation patterns. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 1197–1206, 2008.

[2] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, Vol. 46,

pp. 904–911, 2014.

[3] Andy Cockburn and Bruce McKenzie. What do web users do? an empirical analysis of web use. *International Journal of human-computer studies*, Vol. 54, No. 6, pp. 903–922, 2001.

[4] Susan Dziadosz and Raman Chandrasekar. Do thumbnail previews help users make better relevance decisions about web search results? In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 365–366, 2002.

[5] Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, Vol. 22, pp. 457–479, 2004.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, Vol. 27, , 2014.

[7] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *arXiv preprint arXiv:2212.09611*, 2022.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, Vol. 33, pp. 6840–6851, 2020.

[9] Binxing Jiao, Linjun Yang, Jizheng Xu, and Feng Wu. Visual summarization of web pages. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 499–506, 2010.

[10] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[11] Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011.

[12] Shenwei Liu and Keishi Tajima. Wildthumb: a web browser supporting efficient task management on wide displays. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pp. 159–168, 2010.

[13] Vivian Liu and Lydia B Chilton. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–23, 2022.

[14] Vivian Liu, Han Qiao, and Lydia Chilton. Opal: Multimodal image generation for news illustration. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–17, 2022.

[15] A Porselvi and S Gunasundari. Survey on web page visual summarization. *International Journal of Emerging Technology and Advanced Engineering*, Vol. 3, No. 1, pp. 26–32, 2013.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

[17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, Vol. 21, No. 1, pp. 5485–5551, 2020.

[18] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, Vol. 1, No. 2, p. 3, 2022.

[19] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

[21] Erica K Shimomoto, Lincon S Souza, Bernardo B Gatto, and Kazuhiro Fukui. News2meme: An automatic content generator from news based on word subspaces from text and image. In *2019 16th International Conference on Machine Vision Applications (MVA)*, pp. 1–6. IEEE, 2019.

[22] Jaime Teevan, Edward Cutrell, Danyel Fisher, Steven M Drucker, Gonzalo Ramos, Paul André, and Chang Hu. Visual snippets: summarizing web pages for search and revisitation. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 2023–2032, 2009.

[23] Allison Woodruff, Andrew Faulring, Ruth Rosenholtz, Julie Morrsion, and Peter Pirolli. Using thumbnails to search the web. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 198–205, 2001.

[24] Yutong Xie, Zhaoying Pan, Jinge Ma, Luo Jie, and Qiaozhu Mei. A prompt log analysis of text-to-image generation systems. In *Proceedings of the ACM Web Conference 2023*, pp. 3892–3902, 2023.

[25] 稲垣有哉, 岩田一, 白銀純子, 深澤良彰. ウェブページの構造と顕著性マップの組み合わせによる重要領域の視覚化手法. Web インテリジェンスとインタラクション研究会 予稿集 第 14 回研究会, pp. 45–50. Web インテリジェンスとインタラクション研究会, 2019.

表 12　ベースラインとして用いた記事タイトル例.

記事のジャンル	記事タイトル
Business	These popular fast food drive-thrus are getting faster
Entertainment	Late-night hosts return to air post-writers' strike
Health	mRNA vaccine: 5 things to know
Sport	Solheim Cup: Europe retains title after stunning comeback against USA
Travel	The top superyachts at Monaco Yacht Show 2023
旅行記事サイト	Hershey ' s Chocolate World Announces New Experience That Will Immerse Visitors In A Delicious Train Ride

表 13　提案手法によって選ばれたプロンプト例.

記事のジャンル	選ばれたプロンプト
Business	Artificial Intelligence Drive-Thru and Walk-Up Locations at Chick-fil-A
Entertainment	Late-Night Network Talks Revisited
Health	Essence RNA-Based Vaccine Technology
Sport	First blood vs. Ireland in the Solheim Cup
Travel	The peak of the super yachting sector
旅行記事サイト	The Chocolate World Experience

表 14　提案手法が好まれたサムネイル例.

記事のジャンル	ベースライン	提案手法
Business		
Entertainment		
Health		
Sport		
Travel		
旅行記事サイト		

類似度グラフの事前再計算による拡散式画像検索の効率化

加藤　辰弥†　　駒水　孝裕†　　井手　一郎†

† 名古屋大学　〒 464–8601 愛知県名古屋市千種区不老町
E-mail: †katot@cs.is.i.nagoya-u.ac.jp, ††taka-coma@acm.org, †††ide@i.nagoya-u.ac.jp

あらまし　画像検索タスクにおいて，事前計算による高速な手法とリランキングに基づく高精度な手法がそれぞれ提案されているが，速度と精度を両立した手法の開発が課題になっている．我々はこれらの手法を統合することで，この課題に対処した画像検索のフレームワーク R-DiP （Re-ranking based Diffusion Pre-computation）を提案する．具体的には，事前計算手法において類似度を計算する際に，リランキングで用いられる高精度な類似度再計算を導入することで，高速で高精度な検索を実現する．ベンチマークを用いた実験により，提案手法が最先端の手法と同等な検索精度を維持しつつ，高速な検索を実現できることを示す．提案手法は特に大規模なデータセットにおいて，mAP スコアを 2.0% 程度向上し，クエリごとの検索速度を平均で 75% 程度の削減をし，最先端の高精度手法を上回る検索精度を示した．提案する枠組みは，任意のリランキング手法を取り込むことが可能であり，今後出現するであろう高精度なリランキング手法に対しても効果を発揮し，画像検索技術の進展に貢献することが期待される．

キーワード　画像検索，Content-based Image Retrieval，Diffusion，Re-ranking，効率化

1　はじめに

ディジタルカメラの低廉化やスマートフォンなどの携帯端末に搭載される高性能カメラの普及，画像編集ソフトウェアの大衆化により，個人がディジタル画像を作成することが容易になった．また，Web 上のソーシャルメディアなどのコンテンツ共有プラットフォームの普及に伴い，作成された画像を手軽に公開することができるようになり，我々がアクセスできる画像データの量が増大している．この状況に対して，画像データを管理・活用するために，画像検索は重要な技術のひとつである．画像は実世界を記録するメディアとして言葉だけでは表現することができない事象を記録することに向いており，画像検索は与えられた検索要求に基づいて，過去の事象の記録や関連する事象を探すための主要な手段の一つである．画像検索の中でも，画像を検索要求とする「Content-Based Image Retrieval（CBIR）」があり，キーワード検索やメタデータ検索では捉えきれない画像自身がもつ情報を活用するという特徴がある．

CBIR は 1990 年代 [17] から長い間研究されてきているものの，その性能向上は依然として研究課題である．その主な課題は，画像からの特徴抽出である．近年，ニューラルネットワーク技術の進歩に伴い，従来の人手による特徴量の設計から，大量のデータから自動的に適確な画像特徴を抽出する枠組みが主流となってきた．例えば，畳み込みニューラルネットワーク（Convolutional Neural Network；CNN）[1], [9], [31] や Vision Transformer（ViT）[6], [7] など，数多くの手法が提案されている．これらは検索の目的に合わせて構築されてきたわけではないため，必ずしも検索に適した特徴とは限らない．これに対して，より良い検索を実現するために，検索に適した結果の並び替えを行なうリランキングを利用する手法 [2], [22], [26], [35], [37] やランダムウォークに基づく拡散 [4], [11], [14], [46], [47] などが提案されている．この中でも，Correlation Verification Networks（CVNet）[22] のような，局所特徴をリランキングに用いる手法が高い検索精度を示している．一方で，このようなニューラルネットワークモデルを用いた類似度の推定は低コストとは言えず，検索のオーバヘッドが増し，検索速度が低下する．それに対応するために，大域特徴のみをリランキングに用いる手法 [34] も提案されているが，依然として十分に効率的とは言えない．

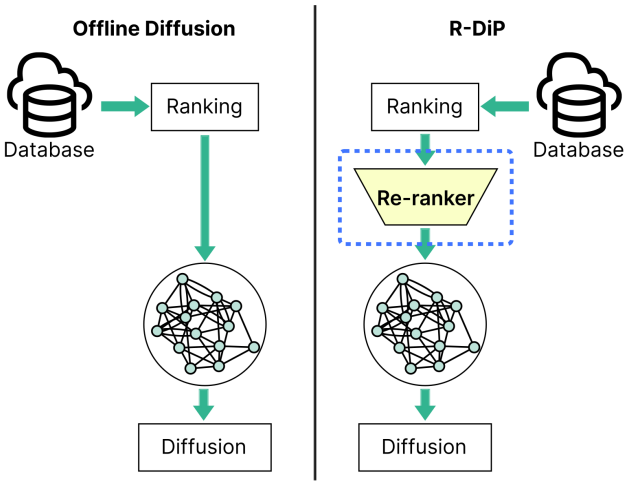


図 1: R-DiP の枠組み．

これに対して，高速な検索を実現するために，データベース画像間の類似性をランダムウォークにより事前計算する Offline Diffusion [44] と呼ばれる手法が提案されている．その基礎となっている拡散（Diffusion）モデルは，擬似適合性フィードバック [47] の考え方を用いることで，同一物体を異なる視点から写した画像や，遮蔽物により物体の一部が隠れている画像を検

索することができる手法である．具体的には，検索要求とデータベース画像から k 近傍グラフを構築し，その上でランダムウォークを行なうことで，データの大域的な構造を考慮した効果的な検索を実現している．Offline Diffusion は，k 近傍検索によるグラフ構築および画像間の類似度計算をデータベース側で事前に（つまりオフラインで）実行する．これにより，検索時に k 近傍検索の結果との単純な線型結合によって検索を行なうことが可能になり，高速な検索を実現している．しかし，その検索精度は事前処理時のグラフ構築に大きく依存し，リランキングを利用する高精度な手法 [22], [34] と比較して劣っている．

本研究では，以上のような速度と精度の間のトレードオフを解決するための手法として，事前計算に基づく高速な手法である Offline Diffusion とリランキングを利用する高精度な手法を融合するための新しいフレームワーク R-DiP （Re-ranking based Diffusion Pre-computation）を提案する．R-DiP の枠組み（図 1）は，Offline Diffusion の近傍検索や関連する類似度検索に高性能なリランキング手法を取り込むことで精度の向上を図る．これにより，検索における関連度を計算する任意のリランキング手法を取り込むことができる柔軟性がある．本研究では，ベンチマークを用いた実験により，特に大規模データセットにおいて，提案手法がリランキングに基づく SuperGlobal [34] に匹敵する検索性能を維持しつつ，Offline Diffusion と同等に高速な検索を実現できることを示す．

本研究の主な貢献は以下の通りである．

- **フレームワーク R-DiP の提案：**Offline Diffusion を基礎とし，リランキング手法の利点を取り込むフレームワークを提案する．提案フレームワークは検索における関連度を計算する任意のリランキング手法を取り込める柔軟性があるため，今後より検索精度が高い高計算コストなリランキング手法が出現した場合にも適用可能である．
- **高速・高性能な検索の実現：**R-DiP は，リランキングに基づく最先端手法である SuperGlobal に匹敵する検索精度を維持しつつ，Offline Diffusion と同等に高速な検索を達成する．特に，大規模なデータセットにおいて，mAP スコアを Offline Diffusion の特性により，最先端の手法より 2.0% 程度向上したうえで，クエリごとの検索速度を平均で 75% 程度の削減をし，最先端の手法を上回る検索性能を示した．

2 関 連 研 究

画像検索における重要な課題は，画像特徴の設計である．機械学習の進歩に伴い，コンピュータビジョンに関する研究として，CNN [19], [21], Regional Maximum Activation of Convolutions（R-MAC）[42]，ViT [5]，GAN [8] などのニューラルネットワークによる手法が登場してきた．これらの手法は画像検索にも採用され [1], [6], [7], [9], [10], [18], [23], [25], [31], [32], [36], [38]，従来の問題であったセマンティックギャップを克服し，画像検索の精度を大幅に向上させている．

画像特徴の設計に続き，画像検索の特性を考慮した検索の最適化に向けた手法が提案されている．例えば，k 近傍検索に基づく検索 [3], [24]，上位に順位付けられた検索結果の類似度を再計算するリランキング，ランダムウォークに基づく拡散 [4], [11], [14], [46], [47] などがある．その中で，リランキングを利用した手法は検索精度の面で優れており，拡散の中でも事前計算に基づく手法 [44] や k 近傍検索に基づく手法は速度の面で優れていることが明らかにされている．以下に，リランキングおよび拡散に基づく手法について詳細に述べる．

a） **画像検索におけるリランキング**

リランキングは，2 段階の検索手法であり，始めに軽量な手法を用いて検索結果を大まかにフィルタリングした後，高精度な類似度計算モデルによって順序付けを行なう．リランキングを利用した CBIR 手法では，最初の段階で大域的特徴に基づいて検索し，物体の局所的な関係性を捉える幾何検証 (Geometric Verification) により類似度を再計算する手法が主流である．幾何検証に関する研究としては，特徴点検出に焦点をあてた研究 [20], [27], [33], [45] や特徴抽出に焦点をあてた研究 [13], [16], [26], [33], [35], [40], [41], [45]，特徴点検出と抽出の順番を入れ替えた研究 [39] など，様々な手法が提案されている．

さらに，幾何検証の代替手法として，ニューラルネットワークを用いて抽出した局所特徴マップを用いる手法が登場した．Transformer を利用した Re-Ranking Transformer（RRT）[37]，局所特徴と大域特徴を効果的に統合した Super-Feature を提案する研究 [43]，局所特徴を用いて複数の異なる大きさで画像間の類似度を計算する CVNet [22] などが提案されている．これらの手法は高精度な検索結果を示すが，速度に関するオーバヘッドが大きいことが問題である．より低オーバヘッドな手法として，リランキングに大域特徴のみを用いる SuperGlobal [34] が提案されているが，事前計算に基づく高速な手法 [44] などと比較すると速度の面で改善の余地が残っている．

b） **拡散（Diffusion）/ Offline Diffusion**

拡散 [14] の核となる考え方は，クエリ画像とデータベース内の画像の類似性をデータベース内の類似画像グラフを通して拡散させることである．拡散の主要な利点は，k 近傍検索のように単に Euclidean 空間上での画像特徴の距離に依存するのではなく，近接性に基づくネットワーク構造を考慮することによって，間接的に類似している画像を発見できる点にある．このような違いにより，例えば，同一物体を異なる視点から写した画像や，遮蔽物により物体の一部が隠れている画像の検索が可能になる．しかし，特に大規模なデータセットに適用する際には，ランダムウォークの計算コストが高いため，検索に時間を要するという欠点がある．

この問題に対処するために，Offline Diffusion [44] が提案されている．Offline Diffusion では，拡散過程がデータベース側で事前に実行されることで，検索時の計算負荷を大幅に軽減できる．具体的には，拡散を行ない事前計算した到達確率と $C = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n : \mathbf{c}_i \in \mathbb{R}^n\}$（ここで，$n$ はデータベース画像枚数）と k 近傍検索で求める y の線型結合として，以下のようにして画像間の類似度を高速に取得する．

（1） k_q 近傍検索：k_q は初期の k 近傍検索における検索枚数である．クエリ画像 q に対して，類似画像のインデックス

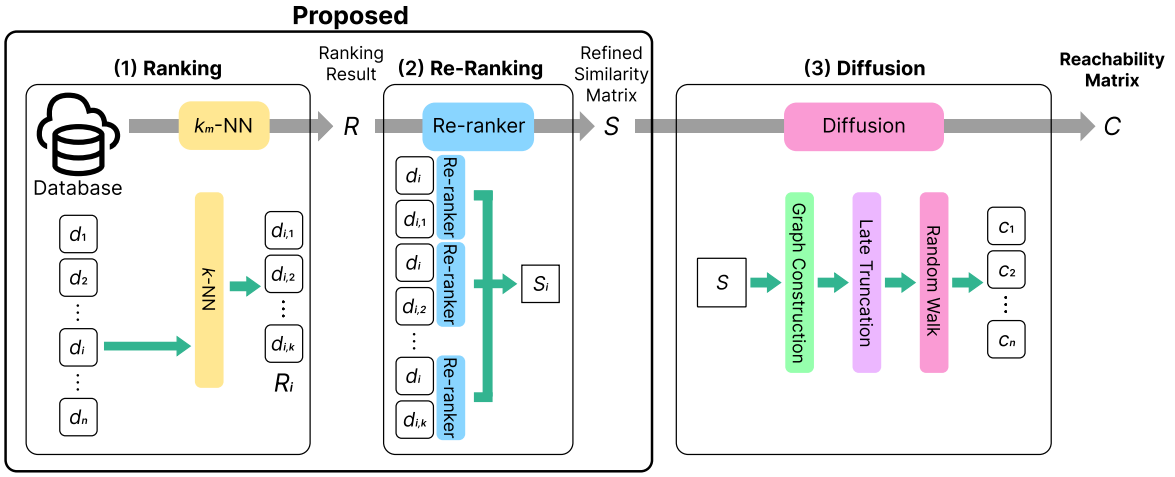


図 2: 提案手法におけるオフライン過程の概要.

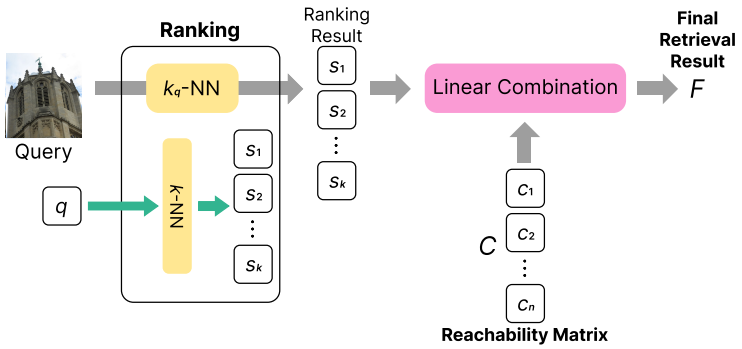


図 3: 提案手法におけるオンライン過程の概要.

$J = \{j_1, j_2, \ldots, j_{k_q}\}$ と類似度 $S = \{s_1, s_2, \ldots, s_{k_q}\}$ を得る．

（2） 線型結合：k_q 近傍検索の結果を用いて，最終的な類似度を $F = \sum_{\ell=1}^{k_q} \mathbf{c}_{j_\ell} s_\ell^\gamma$ として計算する．ここで，$\mathbf{c}_i \in C$ は事前に計算した到達確率ベクトル，γ は画像特徴に基づく類似度の重要度を調整する重み変数である．

この手法は高速な反面，検索精度はリランキングを利用する高精度な手法には劣るため，検索精度の向上が課題である．

3 提案手法：R-DiP

本節では，提案する R-DiP の概要について述べる．R-DiP は，Offline Diffusion における事前計算に，リランキングで用いられる高精度な類似度再計算モデルを組み込むことで，高速で高精度な画像検索の実現を図る．R-DiP の考え方の特徴は以下の二つである．

- **拡散過程の改善**：拡散による類似度計算は，k 近傍グラフの質に大きく依存する．単純な k 近傍検索による構築は画像特徴の設計に依存しており，検索を目的として設計されていない画像特徴の場合に，最適な近傍グラフを構築できるとは限らない．そこで，画像検索におけるリランキングで用いられる類似度計算モデルを用いることでより適した近傍グラフを再構築し，検索精度の向上を図る．
- **リランキングの push-down**：リランキングに基づく手法は精度の面で優れているものの，検索時にリランキングを適用する際のオーバーヘッドは依然として大きい．Offline Diffusion のオフライン過程にリランキングのプロセスを移行することにより，Offline Diffusion の精度を補いつつ，リランキングに必要な計算時間を削減する．

R-DiP は，事前計算を行なうオフライン過程と，実際に検索を行なうオンライン過程の二つから構成される．オフライン過程とオンライン過程の概要をそれぞれ図 2 と図 3 に示す．

ここで，データベース画像の集合を $D = \{d_1, d_2, \ldots, d_n\}$（ただし，$n$ は画像の総数である）とし，画像検索タスクを以下のように定義する．

- **入力**: 一つのオブジェクトを含むクエリ画像 q
- **出力**: q に含まれるオブジェクトを含む画像集合 $D' \subseteq D$

なお，以下では，データベース中の画像をデータベース画像と呼ぶ．本研究では，クエリ画像 q とデータベース画像 $d_i \in D$ は，事前に抽出された画像特徴とする．

3.1 オフライン過程

オフライン過程では，以下のようにして到達確率を計算する．

（1） k 近傍検索による類似画像の取得（k_m-NN）

（2） 類似度再計算モデルによる類似度再計算（Re-ranking）

（3） 拡散に基づく到達確率計算（Diffusion）

これらは，図 2 の（1）～（3）に対応している．Offline Dif-

fusion [44] では，k 近傍検索による画像間類似度計算後に，拡散に基づく類似度計算を行なう．しかし，この類似度は，データベース画像の特徴量の品質に大きく影響を受けるため，その後に計算される類似度の精度を大きく損なう可能性がある．これに対処するため，本研究では，(1) の k 近傍検索の後に (2) の類似度再計算モデルによる類似度再計算を行なうことで，(3) の拡散に基づいて計算される到達確率の精度向上を図る．

ここで，類似度再計算モデルを全てのデータベース画像に適用せず，k 近傍検索の後に適用する理由は二つある．一つ目は，リランキングの概念を取り入れることで事前計算コストを削減できるからである．k 近傍検索により類似度が低いとされる画像が検索結果に与える影響は少ないため，類似度が高い画像のみを考慮することで類似度再計算に要する計算コストを削減することができる．二つ目は，拡散の特性により，ランダムウォークを通じて類似度が高い画像群を特定することができるため，類似度再計算された画像が少ない場合でも効果的に画像間の類似度を求めることができるからである．

3.1.1 k 近傍検索による類似画像の取得

この過程では，各データベース画像 $d_i \in D$ を入力として，$k = k_m$ とする k 近傍検索を行ない，データベース全体から最も類似した k_m の画像集合 $R_i = \{d_{i,1}, d_{i,2}, \ldots, d_{i,k_m}\} \subseteq D$ を取得する．ここで，$d_{i,j}$ は d_i と j 番目に類似度の高い画像を示す．また，すべてのデータベース画像 $d_i (1 \leqq i \leqq n)$ の R_i からなる集合を $R = \{R_1, R_2, \ldots, R_n\}$ とする．なお，画像間の類似度は余弦類似度を用いて計算する．

3.1.2 類似度再計算モデルによる類似度再計算

k 近傍検索により類似画像を取得したのち，各 $R_i \in R$ について，類似度再計算モデル $s : D \times D \to \mathbb{R}$ を用いて d_i と $d \in R_i$ の類似度を再計算する．この過程は近傍グラフの質を向上させることを目的としている．これにより，検索により適した類似度行列 $\mathbf{S} \in \mathbb{R}^{n \times n}$ を得る．ここで，検索結果に含まれない画像 $d_j \notin R_i$ との類似度 $\mathbf{S}_{ij}$ は 0 とする．

3.1.3 拡散に基づく到達確率計算

まず，再計算された類似度行列 $\mathbf{S}$ を用いて，拡散過程のための類似度グラフを構築する．ここでは，再計算された類似度の上位 k_d 件を用いて類似度グラフ $G = (V, E, \delta)$ を作る．ここで，$V = D$ はノード集合，$E \subseteq V \times V$ は $\mathbf{S}_{ij} > 0$ である場合に $E_{ij} = (d_i, d_j)$ で示されるエッジ集合，写像 $\delta : E \to \mathbb{R}$ は $\delta(E_{ij}) = \mathbf{S}_{ij}$ とする．次に，Offline Diffusion [44] 同様に拡散を行ない，到達確率を計算する．これにより，データベース内画像間類似度 $C = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n\}$ が求まり，オンライン過程でこれに基づいて検索を行なう．ここで，$\mathbf{c}_i \in \mathbb{R}^n$ は i 番目の画像とデータベースの他の画像間の類似度のベクトルを表す．

3.2 オンライン過程

オンライン過程では，効率的な検索を実現するために，Offline Diffusion と同じ方法を採用する．これにより，最終的な検索結果は，$F = \sum_{\ell=1}^{k_q} \mathbf{c}_{j_\ell} s_\ell^\gamma$ として計算する．ここで，j_ℓ と s_ℓ は，クエリ画像 q に対する k 近傍検索 $(k = k_q)$ の結果のうち，ℓ 番目に類似度が高い画像のインデックスと類似度であり，γ は画像特徴に基づく類似度の重要度を調整する重み変数である．

4 実　　験

R-DiP を検索精度と速度の両面で評価するために，提案手法が基礎とした Offline Diffusion [44] と最先端の検索性能を達成している SuperGlobal [34] と比較する実験を行なった．

4.1 比 較 手 法

各手法の詳細と比較手法とした目的は以下の通りである．

- **Offline Diffusion：**拡散に基づく画像検索の速度を改善するために設計された手法である．拡散過程の一部を事前に計算することで，検索時の計算コストを削減して検索を高速化している．提案手法と比較することで，提案する類似度再計算の効果を評価する．
- **SuperGlobal：**効果的な大域特徴の取得とそれを用いたリランキングを行なう手法である．この手法は，SuperGlobal Pooling と SuperGlobal Reranking の二つで構成される．SuperGlobal Pooling では Generalized Mean（GeM） Pooling [32] に L_p pooling [12] を取り入れた Regional GeM [34] を利用することで，局所的な特徴を考慮した大域特徴を出力する．一方，SuperGlobal Reranking では，k 近傍検索で取得した初期検索結果の各画像と他の画像のうち類似するものの特徴量を組み合わせることで新たな大域特徴を作成し，類似度を再計算する．この手法は最先端の画像検索手法であり，リランキングに基づく他の手法よりも高速かつ高精度な検索を実現しているが，依然としてリランキング時のオーバヘッドにより Offline Diffusion に比べて検索速度で大きく劣っている．提案手法と比較することで，提案手法の検索速度と精度をともに評価する．

4.2 ベンチマーク

本研究では，提案手法の性能評価のために，画像検索でベンチマークとして用いられる $\mathcal{R}$Oxford5k データセット [30] と $\mathcal{R}$Paris6k データセット [30] を用いる．これらは，旧来のベンチマークである Oxford 5k データセット [28] と Paris 6k データセット [29] のアノテーションや評価難易度を改良したベンチマークである．以下に，$\mathcal{R}$Oxford5k データセットと $\mathcal{R}$Paris6k データセットの特徴を述べる．

- **$\mathcal{R}$Oxford5k データセット：**英国 Oxford 市内のランドマークや建築物の画像，Oxford 大学内の歴史的建造物や，同大学周辺の橋や教会などの画像で構成される．建築物の視覚的特徴を，異なる条件下（異なる視点や時間帯，気候など）において識別する能力の評価に適している．
- **$\mathcal{R}$Paris6k データセット：**フランス Paris 市内の著名なランドマークや観光名所の画像，Eiffel 塔や Louvre 美術館，Notre-Dame 大聖堂など象徴的な場所を写した画像で構成される．都市景観や観光名所を，異なる条件下（異なる視点や時間帯，気候など）において識別する能力の評価に適している．

これまでの研究 [2], [22], [34] では，$\mathcal{R}$Oxford5k データセットは $\mathcal{R}$Paris6k データセットと比較して，検索精度が低くなる傾向が見られる．

表 1: 提案手法と比較手法の精度評価（mAP）.

Method	Medium				Hard			
	$\mathcal{R}$Oxf	+$\mathcal{R}$1M	$\mathcal{R}$Par	+$\mathcal{R}$1M	$\mathcal{R}$Oxf	+$\mathcal{R}$1M	$\mathcal{R}$Par	+$\mathcal{R}$1M
Offline Diffusion [44] (k_q=100)	89.12	85.78	95.22	92.09	75.95	71.11	88.99	84.63
Offline Diffusion [44] (k_q=400)	89.88	85.91	95.37	92.75	77.06	**72.10**	89.29	85.20
SuperGlobal [34] (k_q=100)	88.26	81.76	92.20	83.24	76.67	67.74	83.79	67.66
SuperGlobal [34] (k_q=400)	**90.79**	84.08	93.36	85.53	**80.05**	70.93	86.77	72.52
R-DiP(k_q=100)	89.71	**86.55**	95.39	92.94	76.94	71.97	90.68	**85.61**
R-DiP(k_q=400)	90.40	**86.55**	**95.79**	**93.15**	77.67	71.80	**91.30**	85.46

表 2: 最適なパラメータ.

Method	Medium															
	$\mathcal{R}$Oxf				+$\mathcal{R}$1M				$\mathcal{R}$Par				+$\mathcal{R}$1M			
	k_m	k_d	k_t	γ	k_m	k_d	k_t	γ	k_m	k_d	k_t	γ	k_m	k_d	k_t	γ
Offline Diffusion [44] (k_q=100)	—	200	400	15	—	60	400	15	—	200	800	5	—	200	800	5
Offline Diffusion [44] (k_q=400)	—	200	200	10	—	40	100	20	—	200	800	10	—	100	800	10
R-DiP(k_q=100)	100	60	800	10	400	100	800	10	400	200	800	5	400	200	800	5
R-DiP(k_q=400)	100	60	800	10	400	60	800	20	400	200	800	10	400	200	800	10

Method	Hard															
	$\mathcal{R}$Oxf				+$\mathcal{R}$1M				$\mathcal{R}$Par				+$\mathcal{R}$1M			
	k_m	k_d	k_t	γ	k_m	k_d	k_t	γ	k_m	k_d	k_t	γ	k_m	k_d	k_t	γ
Offline Diffusion [44] (k_q=100)	—	200	400	15	—	100	400	10	—	200	800	5	—	200	800	5
Offline Diffusion [44] (k_q=400)	—	200	200	10	—	40	100	20	—	80	400	15	—	100	800	10
R-DiP(k_q=100)	400	400	400	10	400	100	800	10	400	200	800	10	400	80	800	5
R-DiP(k_q=400)	400	400	200	10	400	40	100	20	400	200	800	10	400	80	800	10

これらのデータセットはそれぞれ 70 個のクエリがあり，それぞれに異なる数の検索対象画像が割り当てられている．データベースに含まれる画像の枚数は，それぞれ 4,933 枚と 6,322 枚である．各クエリに対して，対応するデータベース画像には，検索の難易度に応じて *unclear*，*easy*，*hard* の三つのラベルが付与されている．そして，使用するラベルに基づいて三つ（Easy，Medium，Hard）の評価プロトコルに分かれる．

- Easy プロトコル： *easy* ラベルのみ
- Medium プロトコル： *easy* と *hard* ラベル
- Hard プロトコル： *hard* ラベルのみ

これまでの研究では，Easy プロトコルは既に十二分な性能が達成されており，検索手法の評価に適さないと言われているため，本研究では，Medium プロトコルと Hard プロトコルで実験を行なう．加えて，大規模なデータセットにおける検索性能を評価するために，クエリ画像に含まれない建造物や風景などの約 100 万枚の類似ドメインの画像からなる Distractor セット（$\mathcal{R}$1M）も用意されている．本実験では，Distractor セットを追加した実験も行なう．

評価指標には mAP（mean Average Precision）を用いる．

4.3 実装の詳細

本実験では，提案手法と全ての比較手法において，SuperGlobal Pooling によって抽出された 2,048 次元の大域特徴を用いる．また，提案手法における類似度再計算モデルには SuperGlobal Reranking を用いる．k 近傍検索には，Offline Diffusion 同様に Facebook AI Similarity Search（FAISS）ツールキット [15] を用いる．

各手法について，実験で用いた変数を以下に示す.

- **R-DiP**

提案手法の変数 k_m，k_d，k_t, k_q，γ は次のように設定した.

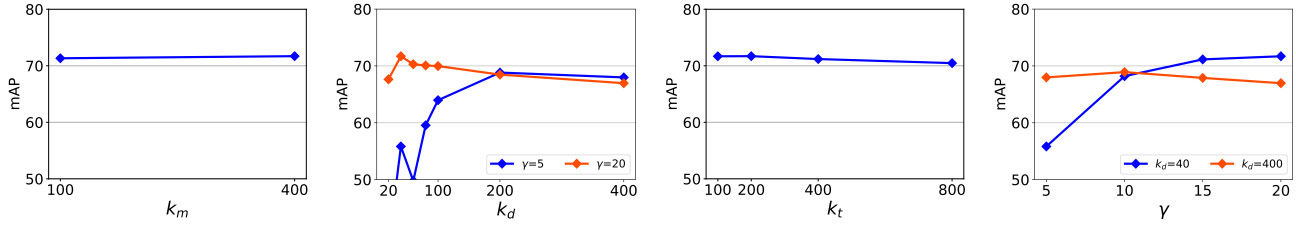
– オフライン過程における k 近傍検索による初期類似画像検索で取得する件数（Offline k-NN Size）： $k_m \in \{100, 400\}$

– オフライン過程における近傍グラフ構築時の近傍数（Offline Graph Degree）： $k_d \in \{20, 40, 60, 80, 100, 200, 400\}$

– オフライン過程における Late Truncation のための k 近傍検索での画像インデックス取得件数（Offline Truncation Size）： $k_t \in \{100, 200, 400, 800\}$

– オンライン検索時の k 近傍検索により取得する画像件数（Online k-NN Size）： $k_q \in \{100, 400\}$

– オンライン検索時の類似度に対する重み変数（Online



(a) k_m (b) k_d (c) k_t (d) γ

図 4: 各パラメータが与える影響.

表 3: クエリごとのオンライン検索時間の平均［ms］.

手法	k_q=100		k_q=400	
	$\mathcal{R}$Oxf	+$\mathcal{R}$1M	$\mathcal{R}$Oxf	+$\mathcal{R}$1M
Offline Diffusion [44]	**1.18**	37.21	1.31	40.18
SuperGlobal [34]	17.74	41.43	139.34	162.86
R-DiP	1.19	**37.14**	**1.30**	**40.01**

Weight)： $\gamma \in \{5, 10, 15, 20\}$

- **Offline Diffusion**

Offline Diffusion では，k_d，k_t，k_q，γ に関して提案手法と同じ変数集合を用いる．なお，Offline Diffusion において類似度再計算過程は存在しないため，k_m は存在しない．

- **SuperGlobal**

SuperGlobal では， k_q に関して提案手法と同じ変数集合を用いる．提案手法におけるその他の変数は，オフライン過程における拡散のための変数であるため，存在しない．

4.4 実験結果

提案手法の性能を評価するための実験結果を示す．

4.4.1 検索性能 (mAP) の比較

表 1 に比較手法と R-DiP の検索精度（mAP）を示す．また，各評価プロトコルにおいて，各手法で最良の結果を示した変数を表 2 に示す．提案手法は，$\mathcal{R}$Oxford5k データセットでは SuperGlobal に匹敵する性能を，$\mathcal{R}$Paris6k データセットでは SuperGlobal を上回る精度を示した．また，Offline Diffusion との比較では，ほとんどの結果において提案手法が優れた精度を示した．これは，提案手法における類似度再計算が Offline Diffusion における近傍グラフ構築を改善したことを示す．また，大規模データセットにおける実験では，提案手法はほとんどの場合で他の比較手法を上回る精度を示した．特に，$\mathcal{R}$Oxford5k+$\mathcal{R}$1M データセットにおける Hard プロトコルでは，2.0%程度の mAP の向上が見られた．Offline Diffusion もまた，大規模データセットにおける実験では，SuperGlobal を上回る精度を示しており，Offline Diffusion の良い特性を取り入れることができていることが示唆される．

4.4.2 オンライン検索時間

次にオンライン検索の実行時間を評価する．$\mathcal{R}$Oxford5k データセットと+$\mathcal{R}$1M セットのクエリあたりの平均オンライン検索実行時間（単位 ms）を表 3 に示す．提案手法は SuperGlobal と比較して実行時間を大幅に削減し，Offline Diffusion と同等の実行時間で検索できることが示された．

4.4.3 パラメータ感度分析

実験で用いる変数が提案手法に与える影響を調べるために，$\mathcal{R}$Oxford5k+$\mathcal{R}$1M データセットの Hard プロトコルにおいて，各変数に対する感度分析を行なった．

- **Offline k-NN Size（k_m）**：図 4(a) に示すように，$k_m = 100$ に比べて $k_m = 400$ の時にわずかに高い mAP を示した．理由として，k_m の値が大きいほど構築される類似度グラフの質が向上し，拡散による類似度計算の精度が向上したためであると考えられる．ただし，事前計算で用いる類似度再計算モデルによる計算コストは k_m の値に比例することをふまえると，実運用では $k_m = 100$ で十分効果的であると考えられる．
- **Offline Graph Degree（k_d）**：$k_d = 20$ の時のみほとんどの場合で最低の mAP を示した．また，k_d の変化が mAP に与える影響は，γ の値によって変わる傾向が見られた．例として，図 4(b) に $\gamma \in \{5, 20\}$ の場合の mAP を示す．γ の値が小さいほど，k_d の値が大きい場合に高い mAP を示し，γ の値が大きいほど，k_d の値が小さい場合に高い mAP を示した．
- **Offline Truncation Size（k_t）**：他の変数と比較して k_t の値の変化は mAP の大きな影響を与えなかった．また，mAP が最良となる k_t の値に対しても，特に規則性は見られなかった．例として，図 4(c) に実験で最良の mAP を示した際に k_t を変化させたグラフを示す．
- **Online Weight（γ）**：前述の通り，mAP が高い変数集合では γ の値が大きいものが多かった．また，γ の値が mAP に与える影響として，k_d の分析同様に，k_d の値が大きいほど，γ の値が小さい場合に高い mAP を示し，k_d の値が小さいほど，γ の値が大きい場合に高い mAP を示した．例として，図 4(d) に $k_d \in \{40, 400\}$ の場合の γ と mAP を示す．

4.4.4 考　　察

精度評価において，提案手法は概ね良好な結果を示し，ほとんどの場合において提案手法は Offline Diffusion を上回る精度を示した．このことは，Offline Diffusion に対して類似度再計算モデルを組み込むことの有効性を示している．

また，提案手法は大規模データセットにおいて SuperGlobal の精度を大きく上回った．SuperGlobal のようなリランキングに基づく手法は，初期の k 近傍検索における検索結果画像集合に大きく依存するのに対し，拡散に基づく手法はデータの大域

的な構造を考慮するため，画像枚数が多い場合に特にその有効性を示すという特性によるためと考えられる．

一方で，提案手法は特に $\mathcal{R}$Oxford5k データセットの Hard プロトコルにおいて， Offline Diffusion と同様に SuperGlobal に劣る精度を示した．この結果は，この状況では Offline Diffusion の有効性が低いことを示し，提案手法もその特性に影響を受けているためであると考えられる．この結果に対し，具体的に二つの理由が考えられる．一つ目は，$\mathcal{R}$Oxford5k データセットにおいて，データの枚数が少ないことにより，大域的構造がもつ情報量の少なさである．拡散は，その特性により大域的構造がもつ情報の質や量に対して大きく影響を受ける．このことにより，特に $\mathcal{R}$Oxford5k データセットの Hard プロトコルのような画像間類似度推定の難易度が高い場合において，画像間の類似度を直接計算するリランキングに基づく手法に比べて，提案手法のような拡散に基づく手法は有効性が低下したと考えられる．二つ目の理由として，Offline Diffusion の特性である，検索結果の初期の k 近傍検索結果における画像間類似度への依存性に影響を受けていることが考えられる．Offline Diffusion は事前計算した類似度に対して初期の k 近傍検索結果における画像間類似度を線型結合する．初期の k 近傍検索結果における画像間類似度により，ランダムウォークが遷移する画像群は大きく変化し，検索結果に大きな影響を与える．これにより，特に高難易度な場面において，オンライン検索で使用される k 近傍検索がボトルネックとなり，提案手法のような拡散に基づく手法の有効性が低下すると考えられる．

5 ま と め

本論文では，画像検索タスクにおける速度と精度のトレードオフに対処する，新しい画像検索フレームワークを提案した．これは，リランキング手法で用いられる高精度な類似度再計算を Offline Diffusion [44] に組み込むことで達成された．実験の結果，提案手法は $\mathcal{R}$Oxford5k [30] と $\mathcal{R}$Paris6k [30] の両データセットにおいて，Offline Diffusion と同等の検索速度を維持しながら，最先端の手法よりも優れている，もしくは同等であることが示された．大規模データセットにおける実験では，提案手法は全てのベースラインと比較して優れた結果を残し，これは，現実世界で大規模データを扱うシナリオに適しており，画像検索技術の進展に貢献することが期待される．

謝　　辞

本研究の一部は JSPS 科研費 21H03555，NII との共同研究による．

文　　献

[1] A. Babenko, A. Slesarev, A. Chigorin, and V. S. Lempitsky. Neural codes for image retrieval. In *Computer Vision —ECCV 2014— 13th European Conf., Zurich, Switzerland, September 6–12, 2014, Procs. Part I, Lecture Notes in Computer Science*, volume 8689, pages 584–599. Springer, 2014.

[2] B. Cao, A. Araujo, and J. Sim. Unifying deep local and global features for image search. In *Computer Vision — ECCV 2020— 16th European Conf., Glasgow, UK, August 23–28, 2020, Procs. Part XX, Lecture Notes in Computer Science*, volume 12365, pages 726–743. Springer, 2020.

[3] T. Dharani and I. L. Aroquiaraj. Content based image retrieval system using feature classification with modified kNN algorithm. *Computing Research Repository arXiv Preprint*, arXiv:1307.4717, 2013.

[4] M. Donoser and H. Bischof. Diffusion processes for retrieval revisited. In *Proc. 2013 IEEE Conf. on Computer Vision and Pattern Recognition, Portland, OR, USA*, pages 1320–1327, 2013.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. 9th Int. Conf. on Learning Representations, Online*, 22 pages, 2021.

[6] S. R. Dubey, S. K. Singh, and W. Chu. Vision transformer hashing for image retrieval. In *Proc. 2022 IEEE Int. Conf. on Multimedia and Expo, Taipei, Taiwan*, 6 pages, 2022.

[7] A. El-Nouby, N. Neverova, I. Laptev, and H. Jégou. Training vision transformers for image retrieval. *Computing Research Repository arXiv Preprint*, arXiv:2102.05644, 2021.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27:2672–2680, 2014.

[9] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. Deep image retrieval: Learning global representations for image search. In *Computer Vision —ECCV 2016— 14th European Conf., Amsterdam, the Netherlands, October 11–14, 2016, Procs. Part VI, Lecture Notes in Computer Science*, volume 9910, pages 241–257. Springer, 2016.

[10] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. End-to-end learning of deep visual representations for image retrieval. *Int. J. Computer Vision*, 124(2):237–254, 2017.

[11] L. J. Grady. Random walks for image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(11):1768–1783, 2006.

[12] Ç. Gülçehre, K. Cho, R. Pascanu, and Y. Bengio. Learned-norm pooling for deep feedforward and recurrent neural networks. In *Machine Learning and Knowledge Discovery in Databases —European Conf., ECML PKDD 2014, Nancy, France, September 15–19, 2014. Proceedings, Part I, Lecture Notes in Computer Science*, volume 8724, pages 530–546. Springer, 2014.

[13] K. He, Y. Lu, and S. Sclaroff. Local descriptors optimized for average precision. In *Proc. 2018 IEEE Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA*, pages 596–605, 2018.

[14] A. Iscen, G. Tolias, Y. Avrithis, T. Furon, and O. Chum. Efficient diffusion on region manifolds: Recovering small objects with compact CNN representations. In *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition, Honolulu, HI, USA*, pages 926–935, 2017.

[15] J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with GPUs. *IEEE Trans. Big Data*, 7(3):535–547, 2021.

[16] H. Jun, B. Ko, Y. Kim, I. Kim, and J. Kim. Combination of multiple global descriptors for image retrieval. *Computing Research Repository arXiv Preprint*, arXiv:1903.10663, 2019.

[17] T. Kato. Database architecture for content-based image retrieval. In *Image Storage and Retrieval Systems, Proc. SPIE*, volume 1662, pages 112–123, 1992.

[18] J. Kim and S. Yoon. Regional attention based deep feature for image retrieval. In *Proc. British Machine Vision*

Conf. 2018, Newcastle-upon-Tyne, England, UK, number 209, pages 1–13, 2018.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1106–1114, 2012.

[20] A. B. Laguna, E. Riba, D. Ponsa, and K. Mikolajczyk. Key.Net: Keypoint detection by handcrafted and learned CNN filters. In *Proc. 17th IEEE/CVF Int. Conf. on Computer Vision, Seoul, Korea*, pages 5835–5843, 2019.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

[22] S. Lee, H. Seong, S. Lee, and E. Kim. Correlation verification for image retrieval. In *Proc. 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, New Orleans, LA, USA*, pages 5364–5374, 2022.

[23] F. Magliani and A. Prati. An accurate retrieval through R-MAC+ descriptors for landmark recognition. In *Proc. 12th Int. Conf. on Distributed Smart Cameras, Eindhoven, the Netherlands*, number 6, pages 1–6, 2018.

[24] H. Nezamabadi-pour and E. Kabir. Concept learning by fuzzy k-NN classification and relevance feedback for efficient image retrieval. *Expert Systems and Application*, 36(3):5948–5954, 2009.

[25] T. Ng, V. Balntas, Y. Tian, and K. Mikolajczyk. SOLAR: Second-Order Loss and Attention for image Retrieval. In *Computer Vision —ECCV 2020— 16th European Conf., Glasgow, UK, August 23–28, 2020, Procs. Part XXV, Lecture Notes in Computer Science*, volume 12370, pages 253–270, 2020.

[26] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han. Large-scale image retrieval with attentive deep local features. In *Proc. 16th IEEE Int. Conf. on Computer Vision, Venice, Veneto, Italy*, pages 3476–3485, 2017.

[27] Y. Ono, E. Trulls, P. Fua, and K. M. Yi. LF-Net: Learning local features from images. *Advances in Neural Information Processing Systems*, 31:6237–6247, 2018.

[28] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. 2007 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, Minneapolis, MI, USA*, 8 pages, 2007.

[29] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proc. 2008 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, Anchorage, AK, USA*, 8 pages, 2008.

[30] F. Radenovic, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. Revisiting Oxford and Paris: Large-scale image retrieval benchmarking. In *Proc. 2018 IEEE Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA*, pages 5706–5715, 2018.

[31] F. Radenovic, G. Tolias, and O. Chum. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In *Computer Vision —ECCV 2016— 14th European Conf., Amsterdam, the Netherlands, October 11–14, 2016, Procs. Part I, Lecture Notes in Computer Science*, volume 9905, pages 3–20. Springer, 2016.

[32] F. Radenovic, G. Tolias, and O. Chum. Fine-tuning CNN image retrieval with no human annotation. *IEEE Trans. Pattern Analysis and Machine Intelligence.*, 41(7):1655–1668, 2019.

[33] J. Revaud, C. R. de Souza, M. Humenberger, and P. Weinzaepfel. R2D2: Reliable and Repeatable Detector and Descriptor. *Advances in Neural Information Processing Systems*, 32:12405–12415, 2019.

[34] S. Shao, K. Chen, A. Karpur, Q. Cui, A. Araujo, and B. Cao. Global features are all you need for image retrieval and reranking. In *Proc. 19th IEEE/CVF Int. Conf. on Computer Vision, Paris, France*, pages 11036–11046, 2023.

[35] O. Siméoni, Y. Avrithis, and O. Chum. Local features and visual words emerge in activations. In *Proc. 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, Long Beach, CA, USA*, pages 11651–11660, 2019.

[36] J. Song, T. He, L. Gao, X. Xu, A. Hanjalic, and H. T. Shen. Binary generative adversarial networks for image retrieval. In *Proc. 32nd AAAI Conf. on Artificial Intelligence, New Orleans, LA, USA*, pages 394–401, 2018.

[37] F. Tan, J. Yuan, and V. Ordonez. Instance-level image retrieval using reranking transformers. In *Proc. 18th IEEE/CVF Int. Conf. on Computer Vision, Montreal, QC, Canada*, pages 12085–12095, 2021.

[38] M. Teichmann, A. Araujo, M. Zhu, and J. Sim. Detect-to-retrieve: Efficient regional aggregation for image search. In *Proc. 2019 IEEE Conf. on Computer Vision and Pattern Recognition, Long Beach, CA, USA*, pages 5109–5118, 2019.

[39] Y. Tian, V. Balntas, T. Ng, A. B. Laguna, Y. Demiris, and K. Mikolajczyk. D2D: Keypoint extraction with Describe to Detect approach. In *Computer Vision —ACCV 2020— 15th Asian Conf. on Computer Vision, Kyoto, Japan, November 30–December 4, 2020, Revised Selected Papers, Part III, Lecture Notes in Computer Science*, volume 12624, pages 223–240. Springer, 2020.

[40] Y. Tian, X. Yu, B. Fan, F. Wu, H. Heijnen, and V. Balntas. SOSNet: Second Order Similarity regularization for local descriptor learning. In *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition, Long Beach, CA, USA*, pages 11016–11025, 2019.

[41] G. Tolias, T. Jenícek, and O. Chum. Learning and aggregating deep local descriptors for instance-level recognition. In *Computer Vision —ECCV 2020— 16th European Conf., Glasgow, UK, August 23–28, 2020, Procs. Part I, Lecture Notes in Computer Science*, volume 12346, pages 460–477. Springer, 2020.

[42] G. Tolias, R. Sicre, and H. Jégou. Particular object retrieval with integral max-pooling of CNN activations. In *Proc. 4th Int. Conf. on Learning Representations, San Juan, Puerto Rico*, 12 pages, 2016.

[43] P. Weinzaepfel, T. Lucas, D. Larlus, and Y. Kalantidis. Learning super-features for image retrieval. In *Proc. 10th Int. Conf. on Learning Representations*, 19 pages, 2022.

[44] F. Yang, R. Hinami, Y. Matsui, S. Ly, and S. Satoh. Efficient image retrieval via decoupling diffusion into online and offline processing. In *Proc. 33rd AAAI Conf. on Artificial Intelligence, Honolulu, HI, USA*, pages 9087–9094, 2019.

[45] T. Yang, D. Nguyen, H. Heijnen, and V. Balntas. UR2KiD: Unifying Retrieval, Keypoint Detection, and Keypoint Description without local correspondence supervision. *Computing Research Repository arXiv Preprint*, arXiv:2001.07252, 2020.

[46] S. Zhang, M. Yang, T. Cour, K. Yu, and D. N. Metaxas. Query specific rank fusion for image retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 37(4):803–815, 2015.

[47] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf. Ranking on data manifolds. *Advances in Neural Information Processing Systems*, 16:169–176, 2003.

3D モデルを利用したユーザーの求めるアングル画像の検索システムの開発

永田　玄† 常　窨†† 宮崎　純††

† 東京工業大学情報理工学院情報工学系情報工学コース　〒 152-8550 東京都目黒区大岡山 2-12-1
†† 宮崎研究室　〒 152-8550 東京都目黒区大岡山 2-12-1
E-mail: †nagata@lsc.c.titech.ac.jp, ††{q.chang,miyazaki}@c.titech.jp

あらまし　現在画像を入力として画像を検索する画像検索ツールに関する研究が多くされている. しかし, 既存の研究の主な目的は入力した画像と色合いが同じものやカテゴリが同じものを検索することで, 角度違いのものを検索するという目的で使用することはできない. そこで, 本研究ではユーザーが入力した画像の 3D モデルを推定し, 回転させることで任意の角度からの画像を検索可能とした. また、データベース上の画像とのマッチングにおいても, 検索対象の画像の 3D モデルを推定することで角度によって色やシルエットが異なるものも検索可能とした. こうして、構築されたシステムの精度を実画像データを用いた評価実験により評価し，手法や入力画像，アングル毎に結果を確認し，提案手法について考察した．その結果, 先行研究では困難だった, 角度によって輪郭線や色は類似するが全く別のオブジェクトのペアの判別や, 逆に角度によって輪郭線や色が大きく異なるオブジェクトに対する検索も可能とした.

キーワード　画像検索エンジン, 3D モデル, Image Search Engine, 3D model

1 はじめに

情報検索システムの起源は 1970 年代にまで遡り, 大規模に蓄積される学術文献や論文等の管理をコンピュータ上で行うために, データの管理と検索が行われるようになったことで, 需要に応えるべく情報検索システムの開発が進んだ. そして今日では，インターネットの発達とスマートフォンなどの画像撮影デバイスの普及に伴い，生活者は役に立つ情報を見つけることから最寄り店舗の検索, あるいは何か新しいものを学習するためなど, あらゆる目的に情報検索を行う．そのため, 情報検索システムの利用形態の多様化に伴って, テキストによる検索だけでなく画像による検索システムも必要とされている. このような背景のもと, Google 画像検索などの画像を結果として提示する検索エンジンが普及し，多くの画像検索システムに関する研究，中でも, CBIR(content based image retrieval) と呼ばれる画像を入力とする画像検索システムの開発が盛んに行われている [4].CBIR では一般に入力された画像とデータベース上の画像から色, テクスチャ, 輪郭線などの特徴量を抽出し, 比較することで色合いが同じものやカテゴリが同じものを検索することができる. しかし, このような CBIR 手法の多くは, オブジェクトの角度に関する情報を十分に抽出することができず, 角度の違うものを検索するという目的で使用することは難しい. 従来のコンテンツベースの画像検索の例を図 1 に示す．そこで本研究では，2D 画像から 3D モデルを推定することによりこの問題を解決することを目標とする．近年, 機械学習を利用して 2D 画像から 3D モデルを推定する手法が多く開発され, 3D モデルの精度を向上している [9] [16] [7]. 本研究の目的は，画像を入力として，その画像中のオブジェクトをユーザーによって指定された別アングルから見た画像を提示することができる検索システムの構築方法を提案することである．提案手法での画像検索の例を図 2 に示す．この目的を達成，発展させることで, 主にオンラインショッピングにおいて事前に商品を様々な角度から閲覧することが可能となるため, 想像していた商品と実際の商品との差異を減らすことが可能であると考える.

雛形ら [17] は同様の目的をもった研究をしている．雛形らの研究では，通常のカメラで撮ることのできない距離情報を持った画像であるデプス画像を入力画像に用いる．そこで本研究では，通常の画像データからなる写真を入力とし，より汎用性の高い検索アルゴリズムを提案する．次に, 画像中の物体を他の角度から見た画像を探すため, その物体の形状の情報，すなわちその物体の 3D モデルを推定する．また, 前川ら [18] も同様の目的をもった研究をしている．前川らの研究では, 推定された 3D モデルの形状が不正確で, 画像の曖昧な輪郭線のマッチングを図っている．そのため, 角度によって形状や色が著しく異なるものや 逆に角度によって色と形状が似ているが全くの別ものは判別が難しく, 検索結果の精度に悪影響を及ぼすことがある．そこで本研究では, 機械学習を用いたより精度の高い 3D モデル推定手法を利用して, 3D モデル間でのマッチングとランキング化を行い, 角度による形状変化に影響されないシステムを提案する．この提案手法の精度を，実画像データを用いて入力画像やアングル毎に検証する.

2 関連研究

2.1 画像検索に関する研究

一般的な画像検索手法は 2 種類ある．第一の方法は, 画像に注釈をつけるためのキーワードに基づくもので, テキストベースの画像検索として知られている．しかし, この方法には多くの欠点があり, 1) 大規模なデータベースに手作業で注釈を付け

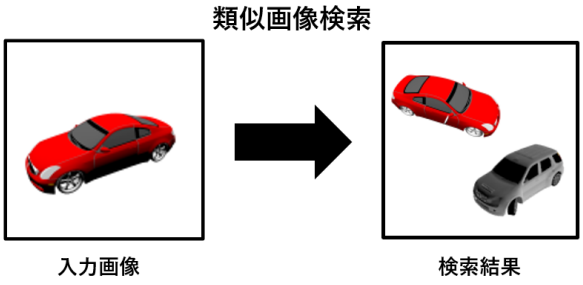


図 1　従来の検索システムのイメージ

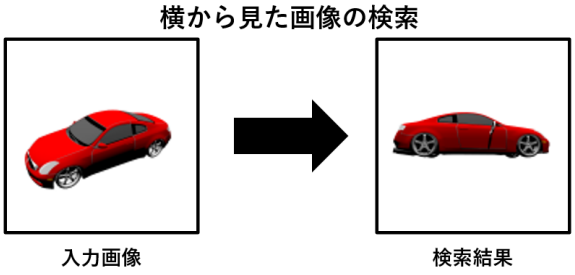


図 2　本研究の検索システムのイメージ

るのは現実的ではない, 2) エンドユーザーが注釈を付けなければならず, そのためこの方法は人間の知覚に左右される, 3) これらの注釈は 1 つの言語にしか適用できないといったものが代表される. そこで第二の方法としてコンテンツベースの画像検索 (CBIR) が提案され, テキストベースの画像検索方法の欠点を克服するために強く推奨されている [4]. CBIR で最も重要な段階は画像の特徴を抽出する段階である. 一般に, 特徴量は大域的特徴量と局所的特徴量に分類される. 大域的特徴 (色, テクスチャ, 形状, 空間情報), は画像全体の表現を与える特徴量であり, 特徴抽出や類似度計算が速いという利点がある一方, 画像内の背景と物体を区別することができないといったデメリットもある. そのため, 複雑なシーンでの検索や物体認識には適さないが, 物体の分類や検出には適している. そして, 大域的特徴と比較して局所的特徴 (コーナー, エッジ) はスケール, 回転, 平行移動, 背景の変化, 部分的なオクルージョンに対してロバストであるため, 画像検索, マッチングタスク, 認識に適している. 最近の CBIR では 2 つの特徴量を組み合わせたり, 抽出された特徴を機械学習アルゴリズムに与えることで精度の向上を図っている. Srivastava と Khare [15] は, 複数のレベルで画像を分析し, 1 つのレベルがスキップした情報を他のレベルが捕捉する, 新しいマルチ解像度分析アルゴリズムを開発した. このアプローチは、テクスチャ特徴を抽出するために LBP(local binary pattern) 記述子 [10] を使用し, テクスチャ特徴から形状特徴を抽出するために Legendre モーメントを多解像度レベルで使用する 2 つの特徴量抽出手法をベースとしたアルゴリズムを提案している. Sarwar ら [14] は, LBPV(local binary pattern variance) [3] と LIOP(local intensity order pattern) を用いて, セマンティックギャップを減らすことで CBIR 性能を向上させるシステムを提案した. このシステムではこれら 2 つの特徴記述子を 2 つの小さな視覚辞書を形成するために使用し, その後, 1 つの大きな視覚辞書を形成するために連結し, サイズを小さくするために主成分分析 (PCA) を用いてヒストグラムが計算することで, LBPV と LIOP に基づいて SVM(support vector machine) [2] を学習する.

しかし，これらのような最先端の CBIR 手法であっても，入力画像内のオブジェクトに隠れている部分が存在する場合，指定した別のアングルから見た画像を提示することは難しい．本研究は，その目的を達成することができる手法を提案することが目標である.

2.2　2D 画像からの 3D モデル推定に関する研究

深層学習の発展により, 2D から 3D モデルを再構成するための学習ベースの手法に関する研究が盛んに行われている.

Mescheder ら [9] は, 3D 再構成手法として, Occupancy Network を提案した．このネットワークでは, 3D の表面情報を，ディープニューラルネットワーク分類器の連続性決定性境界として暗黙的に表している．単一の画像からだけではなく，ノイズの多い点群や粗いボクセル状のグリッドからでも, 3D 構造をエンコードできる．また，他の既存手法よりも低いメモリ使用量での 3D モデル推定を実現させた．学習されているモデルのカテゴリは，主に家具や乗り物である.

Yao ら [16] は対称性をもつものに対する 3D 再構成手法として, Front2Back を提案した．このシステムでは, 入力画像から CNN を利用してデプス画像, 法線マップ, シルエット画像 (以降これらの画像を前面図とする) を予測する．また, 前面図より入力画像中の物体の対称面を探索し, 得られた対称面の情報と前面図より条件付き生成的敵対的ネットワークの変種 [5] を使用して, 物体を入力画像と反対の視点から見たデプス画像と法線マップ (以降これらの画像を背面図とする) を生成し, これらの前面図と背面図を用いて 3D メッシュを再構成する．この手法は対称性をもつものに対しては Occupancy Network よりも高い精度をもつ.

Long ら [7] は多視点画像から 3D 再構成する手法として, Wonder3D を提案した．このシステムでは, 入力画像からマルチビューのカラー画像と法線マップを生成し, SDF(Signed Distance Field) と呼ばれる 3D 空間内の点に対して, その点が物体の表面からどれだけ離れているかを示す関数を深層学習モデルで学習し, 3D メッシュを再構成する．このアプローチは 2D 画像から 3D モデルを推定に当たり問題となる忠実性, 一貫性, 一般化可能性, 効率性に対処するように設計されている．実際に Wonder3D により推定した 3D モデルを図 3,4 に示す．図 3 に示したモデルから Wonder3D により推定されたモデルは同じカテゴリのものでもその違いを表すことができるほど精度は高い．また, 自動車のドアミラーや椅子, 机の空洞部など細かい部分も表現することができる．しかし, 入力画像からマルチビューのカラー画像と法線マップを生成するため, 図 4 のように奥行が分かりにくい画像や画像中にオクルージョン部分が存在すると上手くモデルを生成することは難しい.

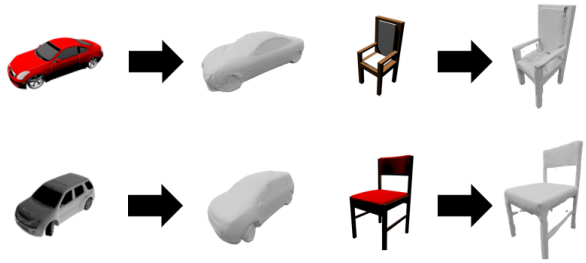


図 3　Wonder3D で推定した 3D モデルの例

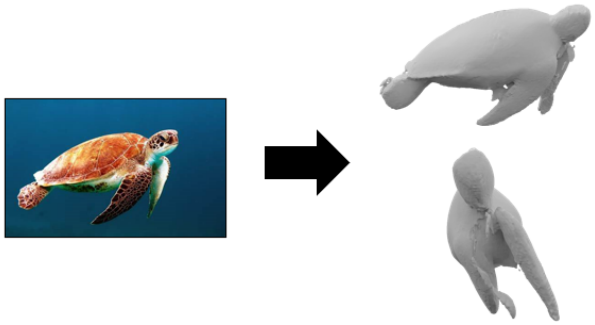


図 4　Wonder3D で推定した 3D モデルの失敗例

2.3　3D モデルの特徴量に関する研究

2 つの 3D モデルの類似度を調べる代表的な手法として, 特徴量によるマッチングが挙げられる. 3 次元データの特徴量に関する研究は盛んに行われている. [8]

Kazhdan ら [6] は大域的特徴量として Spherical Harmonics を提案した. この手法は, 任意の物体の 3 次元形状を球面上のある関数と見なし, これを有限個の基底ベクトルの線形結合で近似するものである.

Rusu ら [12] は FPFH(fast point feature histograms) と呼ばれる PFH(point feature histograms) [13] を高速に改良した局所特徴量を提案した. ある点 $\boldsymbol{P}$ を表す PFH は, その点を中心とした小球領域に含まれる k 個の近傍点を求め, それらの点から 2 点を選ぶすべての組み合わせ $(\boldsymbol{P}_i, \boldsymbol{P}_j)(i \neq j, 1 \leqq i, j \leqq k)$ に対して計算される 3 つのパラメータのヒストグラムを計算する. このヒストグラムの計算量は $O(k^2)$ となり, サンプリング数によって計算量が指数関数的に増加する. 一方, FPFH の計算量は O(k) に改良されており, PFH に比べてかなり高速な処理となっている.

しかしこれらの特徴量は 3D モデルの回転に対して不変性を示すため, 本研究には適用することができない. そこで本研究では 3D モデルの姿勢に依存する類似度測定手法として, CD(Chamfer Distance) を利用する.

CD は一方の点が他方の点にどれだけ近いかを示す指標であり, 各点がもう一方の点集合の最も近い点までの距離を計算し, その距離の合計を取ることで定義される. そのため, 3D モデルの回転も考慮した姿勢と形状の類似度を比較することができ, 本研究に適している. 具体的に点群 X と点群 Y 間の CD を計算する式は以下のように定義される.

$$CD(X,Y) = \frac{1}{|X|}\sum_{x\in X}\min_{y\in Y}\|x-y\|_2^2 + \frac{1}{|Y|}\sum_{y\in Y}\min_{x\in X}\|x-y\|_2^2 \tag{1}$$

具体的な処理は以下のようになる.

1. 点群 X 内の点 x と, 点群 Y の全ての点のユークリッド距離を算出し, 2 乗する.
2. 1 で求めた 2 乗ユークリッド距離の最小値を, 加算する.
3. 加算した最小距離の平均を計算する.
4. 同様の処理を点群 Y で行う (こちらの場合は点 y と点群 X の全ての点を比較する).
5. 算出した点群 X における最小距離と点群 Y における最小距離を加算する.

2.4　先 行 研 究

雛形ら [17] や前川ら [18] は，本研究と同様に，画像を入力として，その画像中のオブジェクトをユーザーによって指定された別アングルから見た画像を提示することができる検索システムの構築方法を提案している．雛形らは入力画像として，デプス画像と呼ばれる物体までの距離情報を持った画像を用いて，そのデプス画像から復元した部分モデルと完全な 3D モデルのマッチング精度を評価している．本研究では，デプス画像を普通のカメラで撮影できず，今日社会の中で普及していないことに着目し，同様の目的を普通の画像をクエリとして達成可能にすることを目標としている．前川らは入力した画像から 3D モデルを推定し，それをユーザーの任意の方向に回転させることで，テキスト等の追加情報なしでアングルを指定して，そのアングルから見たオブジェクトを検索可能にしている．しかし, 推定された 3D モデルは形状が不正確であり，色も単色で使用できず，データベース上の画像と比較する際は画像の輪郭線と回転させた 3D モデルの曖昧な輪郭線を比較している．前川らの使用した 3D モデルの例 [18] を図 5 に示す．前川らの研究では, 輪郭線を比較するにあたって, Oliva ら [11] が提案した GIST 特徴量を用いてを用いて, 入力画像とデータベース内の全ての画像の類似度を比較し, 同じ物体が写っている画像を事前に抽出することを考えている．しかし, GIST 特徴量は画像の雰囲気を表現し, シーン認識に特化した特徴量であるため, 角度によって形状や色が著しく異なるものは抽出することができず, 逆に角度によって色と形状が似ているものは全く違うものであっても抽出されてしまい, 検索結果の精度に悪影響を及ぼすことがある．前川らの提案したシステムの苦手とする検索例 [18] を図 6 に示す．そこで, 本研究ではより精度の高い 3D モデル推定を行うことができる手法を利用して, 入力画像とデータベース内の全ての画像の 3D モデルを推定し, 3D モデル同士で特徴量マッチングをすることで, 角度による形状の違いが検索結果へ及ぼす悪影響を抑制する.

図 5　前川らの使用した 3D モデル ([18] より抜粋)

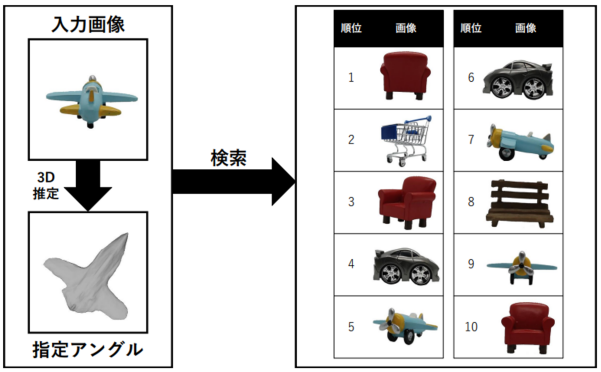


図 6　前川らの提案したシステムで飛行機を上から見た画像を検索した結果 ([18] より抜粋)

3　提 案 手 法

3.1　システムのアーキテクチャ

ここでは，本研究で提案するシステムのアーキテクチャについて包括的に述べる．本研究で提案する検索システムの全体像を図 7 に示す．このシステムのフローは, 大きく以下の 4 つに分けられる．

1. 事前にデータベースの画像の 3D モデルを推定
2. 入力画像から 3D モデルを推定し，3D モデルをユーザーが任意のアングルへ回転
3. 回転した 3D モデルとデータベースの 3D モデルで特徴量マッチング
4. 特徴量マッチングに基づいてランキング化

あらかじめデータベース内のすべての画像から 3D モデルを推定し, 3D モデルデータベースとして保存しておく．入力は, JPG や PNG などの一般的な画像データ 1 枚のみであり, 入力された 1 枚の画像から 3D モデルを推定し, ユーザーが要求する角度へと回転させる．次に, 3D データベース上の 3D モデルを 1 つずつ選択し, ユーザーによって回転させておいた 3D モデルとの類似度を比較する．最後に類似度を昇順に並べ, データベース上の画像をランキング化して提示する．それぞれの内容に関して詳しく，以降の節で述べる．

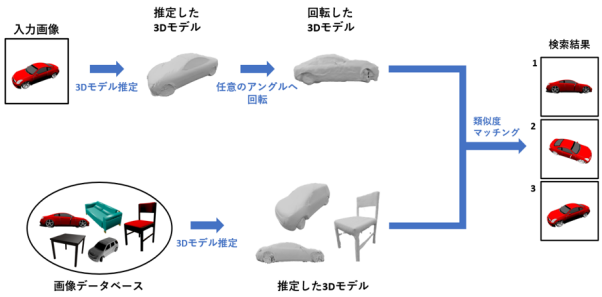


図 7　提案するシステムのアーキテクチャ

3.2　2D 画像からの 3D モデル推定

本手法では，入力した画像とデータベース上の画像から 3D モデルを推定し, 入力画像の 3D モデルはユーザーの任意の方向に回転させることで，テキスト等の追加情報なしでアングルを指定して，そのアングルから見たオブジェクトを検索可能にする．

深層学習の発展により, 2D から 3D モデルを再構成するための学習ベースの手法に関する研究が盛んに行われている．Mescheder ら [9] の Occupancy Network は事前にオブジェクトを学習する必要があり, また, Yao ら [16] の Front2Back は対称性をもたないオブジェクトに対する 3D モデルの精度が低い．本研究では, スマートフォンを利用する人が老若男女問わず, また情報検索を必要とするユーザーは特定の分野に限らないことから両者のように特定の条件でのみ使用できる 3D モデル推定手法を利用することはできない．そのため, 事前に学習する必要がなくモデルの生成範囲に制限がない, Long [7] の提案する Wonder3D を採用している．

3.3　特徴量マッチング

提案する検索システムでは, 入力した画像やデータベース上の画像から 3D モデルを推定し, 入力画像の 3D モデルをユーザーの任意に回転させることで検索したいアングルを指定する．その後, 各 3D モデルの姿勢も考慮した特徴量マッチングをし比較する．

3D モデルの特徴量には Spherical Harmonics [6], FPFH [12] など様々な種類がある．これらは全て 3D モデルの回転に依存しない特徴量である．本研究ではオブジェクトの回転を含めた姿勢に対して類似度を計算する必要があるため, これらの特徴量は本研究には適さない．よって本研究では, Chamfer Distance(CD) と呼ばれる 2 つの 3D 点群間の位置誤差を測る指標を用いる．CD の値は小さいほど 2 つの 3D モデルの類似度が高いことを示す．本研究では 3.2 節で推定したデータベース上の画像の 3D モデルから 1 つずつ選択し, ユーザーの要求する角度へと回転させた入力画像の 3D モデルとの CD を計算する．最終的にデータベース上の画像のすべての 3D モデル間で計算された CD を昇順に並び替え, 対応するデータベース上の画像をランキングとして出力する．

4 評価実験

4.1 実験用データと前処理

本手法の実験では，複数のアングルから見たオブジェクトの画像が必要になる．よって，事前に用意した 13 個のオブジェクトについてそれぞれ 4 アングルから撮影し，データセットとして 52 枚の画像を用意した．撮影したオブジェクトと，実験のため付けた名前を図 8 に示す．この時，撮影した 4 アングルは以下の通りである．なお, オブジェクトを正面から見たときに左方向を x 軸, 上方向を z 軸, 手前方向を y 軸の正の方向とし, 各軸を正の方向から見た場合の反時計回りを回転の正の方向とする．

- 前：正面から見た画像 (以降「f」とする)
- 横：正面から見てオブジェクトを z 軸周り-90 回転させた画像 (以降「s」とする)
- 後ろ：正面から見てオブジェクトを z 軸周り 180 回転させた画像 (以降「b」とする)
- 上：正面から見てオブジェクトを x 軸周り-90 回転させた画像 (以降「u」とする)

また, 図 9 に示すような角度によって色や輪郭線が 52 枚の画像の一部と類似するオブジェクトの画像も画像データベースに加えた．それぞれのオブジェクトについて，スマートフォンで写真を撮影し，背景を取り除いた．それぞれの手法の効果を正確に比較できるよう，様々なジャンルのものや，色や形状，ジャンルが似ているものなど幅広く用意した．さらに, これらの 59 枚の画像すべての 3D モデルを事前に推定し, 3D データベースとして格納した．

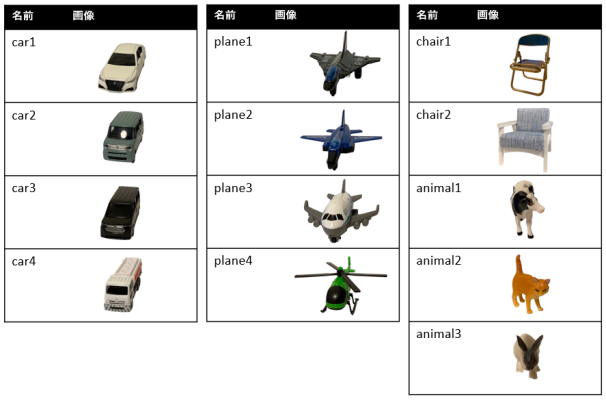


図 8 実験のために撮影したオブジェクト

4.2 実験方法

4.2.1 3D モデル推定

本手法では，入力した画像から 3D モデルを推定し，それをユーザーの任意の方向に回転させることで，テキスト等の追加情報なしでアングルを指定して，そのアングルから見たオブジェクトを検索可能にする．ここでは，実際に 3D モデルを推定した方法を述べる．入力は撮影した画像から背景を取り除いた PNG 画像 1 枚のみである．Long ら [7] の提案する

図 9 角度によって色や輪郭線が類似するオブジェクト

Wonder3D で，3D モデル化した．この時，Wonder3D では奥行がオクルージョン部分となり, 適切に 3D モデル化できないオブジェクトがあった．そのうちの一部を図 10 に示す．また, 各オブジェクトの前アングルから推定した 3D モデルを図 11 に示す．実験では，これらの 3D モデルを入力として用いた．

図 10 奥行が適切に 3D モデル化できていないオブジェクトの例

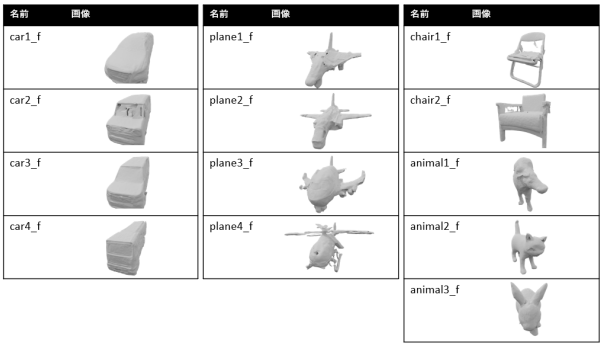


図 11 実験の入力に用いた正面画像の 3D モデル

4.2.2 指定アングル

アングルの指定には，図 11 に示した各オブジェクトの正面の 3D モデルを入力として用いた．アングルには, 1 つのオブジェクトに対して，z 軸方向-30 刻みにアングルを指定し ，実験した．

4.2.3 特徴量マッチング

本手法では推定した 3D モデルを任意のアングルに回転し，そのアングルから見た画像を検索する．その際，回転させた 3D モデルとデータベースの画像の 3D モデルの特徴量を比較し，近いものをランキングの上位に出力する．特徴量マッチングで

は，CD による特徴量マッチングを用いた．データベース内の画像 59 件に対してマッチングを行い類似度の高かった画像の上位 10 件をランキングとして出力し，評価した．

4.3 評価方法

検索結果の評価には，nDCG (normalized Discounted Cumulative Gain) を用いた．nDCG は，予測されたランキングの評価指標である．nDCG は 0 から 1 の値を取り，1 に近いほど正しいランキングであることを示す．nDCG は，DCG (Discounted Cumulative Gain) を，予測ランキングの DCG を正しいランキングの DCG で割ることにより，正規化した値である．本研究では Burges ら [1] が提案した DCG を用いた．Burges らが提案した DCG は以下の式で定義されている．

$$DCG(k) = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{2}$$

そして，nDCG は，DCG を考えられる最大の DCG($DCG_{perfect}$) で正規化した値で，0 から 1 の間の値をとる．

$$nDCG = \frac{DCG}{DCG_{perfect}} \tag{3}$$

rel_i は，ランキングの i 番目の要素の適合度を表しており，今回の実験では i 番目の画像が以下の条件を満たす数を rel_i とする．

- 入力した画像のジャンルと同じジャンル (例: car1_f を入力とした場合，「car」であるどうか)
- ジャンルが一致している条件のもと，番号が一致する (例: car1_f を入力とした場合，「car1」であるかどうか．chair1 や animal1 は無効とする)
- ユーザーの要求する角度に対して ±30 度以内の画像かどうか (例: car1_f を入力とし，-30 度の z 軸回転をした場合，0 度の画像であるかどうか)

また，k は評価に用いる要素数を表すため，今回はランキングの上位 10 件を評価しているので，k は 10 である．

4.4 評価結果

13 個の 3D モデルをそれぞれ 7 つのアングルで，59 件の画像から検索した結果の nDCG を図 12 に示す．具体的なランキングの結果をいくつか図に示す．図中の赤丸は正解となる画像を示す．

図 15 は car1_f を入力として，z 軸の回転角を-180 度とした場合の検索結果であり，検索対象は car1_b である．図 15 は正解となる画像が 3 位となっている．

図 17 は plane3_f を入力として，z 軸の回転角を-90 度とした場合の検索結果であり，検索対象は plane3_s である．図 17 は正解となる画像が 1 位となっている．

4.5 考　　察

4.5.1 入力画像ごとの検索結果に関する考察

表より入力画像によって nDCG の値にばらつきがあることが分かる．car1～car4 では nDCG の平均値が 0.64～0.82, chair1

入力オブジェクト ＼ z軸の回転角(度)	0	-30	-60	-90	-120	-150	-180	平均
car1_f	0.92	0.77	0.71	0.64	0.6	0.56	0.79	0.71
car2_f	0.89	0.7	0.81	0.78	0.81	0.86	0.91	0.82
car3_f	0.88	0.65	0.77	0.7	0.56	0.66	0.66	0.7
car4_f	0.9	0.59	0.58	0.69	0.49	0.43	0.77	0.64
plane1_f	0.81	0.72	0.22	0.3	0.34	0.19	0.35	0.42
plane2_f	0.89	0.69	0.17	0.35	0.37	0.12	0.35	0.42
plane3_f	0.8	0.34	0.17	0.76	0.51	0.15	0.32	0.44
plane4_f	0.91	0.71	0.61	0.29	0.34	0.26	0.47	0.51
chair1_f	0.79	0.77	0.58	0.73	0.82	0.57	0.64	0.7
chair2_f	0.76	0.76	0.51	0.61	0.6	0.76	0.71	0.67
animal1_f	0.69	0.77	0.27	0.15	0.24	0.2	0.35	0.38
animal2_f	0.75	0.5	0.04	0.61	0.76	0.12	0.64	0.49
animal3_f	0.68	0.69	0.39	0.39	0.29	0.42	0.43	0.47
平均	0.82	0.67	0.45	0.54	0.52	0.41	0.57	

図 12　nDCG による評価結果

と chair2 の nDCG の平均値は 0.685 と両者の nDCG の値は比較的高いことが分かる．しかし，plane1～plane4 では nDCG の平均値が 0.42～0.51, animal1～animal3 では nDCG の平均値が 0.38～0.49 であり，car や chair と比べると低いことが分かる．これは，Wonder3D によって生成された 3D モデルのアングルによる差にあると考えられる．car や chair で生成された 3D モデルは正面，横，後ろからそれぞれ生成された 3D モデルの形状が類似しており，animal や plane ではそれぞれの角度から生成された 3D の形状に差がある．実際に car2 と animal1 の正面，横，後ろから生成された 3D モデルを図 13 に示す．car2 から生成された 3D モデルはすべての角度でほぼ正確に car2 が表現されているのに対して，animal1 から生成された 3D モデルは特に正面から見た 3D モデルで animal1 の体部分の長さが実際のものとは異なっていることが分かる．

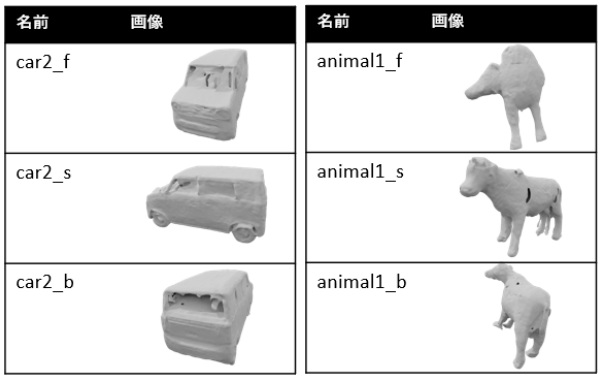


図 13　car2 と animal1 を正面，横，後ろから見た画像から生成された 3D モデル

4.5.2 各アングルごとの検索結果に関する考察

表より z 軸の回転角ごとに nDCG の値にばらつきがあることが分かる．図 14 に z 軸の回転角と nDCG の平均値グラフを示す．0 度や-30 度で比較的高い数値を記録していることが分かる．これは単に入力した画像と検索対象となる画像が同じであるため，3D モデルで比較する際にも差が小さいためであると考えられる．次に数値の高い回転角は-180 度，-90 度，-120 度の順である．-90 度や-180 度はそれぞれその角度から生成された 3D モデルが存在するためだと考えられる．-120 度は-60 度と条件がほぼ同様であるにも関わらず，差が出ている．しかし，入力画像によっては-60 度のほうが-120 度のほうが高いものもある．-60 度と-120 度に差が出る原因は検索対象の 3D モデルよ

りも, 入力画像をそれぞれの角度分回転させた 3D モデルに近い 3D モデルの数に差があるためである. そのため, 今回使用したデータベースでは-60 度よりも-120 度回転させた 3D モデルに近い 3D モデルが全体的に多かったのだと考えられる. しかし, 別のデータセットを使用した場合, この結果は逆転する可能性があるため, 明確に提案するシステムが-60 度よりも-120 度で検索したのほうが優れているとは言い切れない.

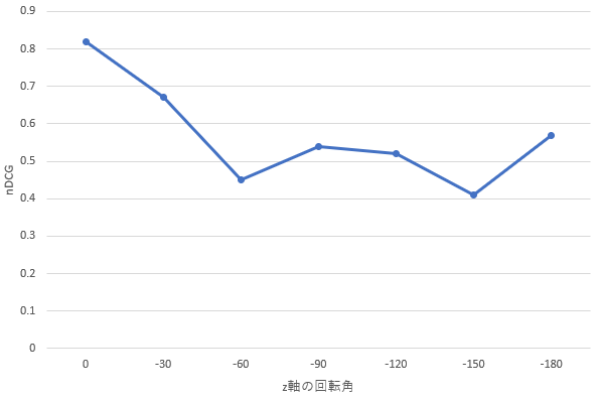


図 14　z 軸の回転角ごとの nDCG の平均値

4.5.3　角度によって輪郭が類似するオブジェクトに関する考察

今回のデータセットには図 8 の 13 個のオブジェクトに加えて, 図 9 に示す 7 個のオブジェクトを加えた. これらのオブジェクトの輪郭線や色に注目すると, black1〜black3 は car3 に, white1 から white4 は car1 に類似していることが分かる. このように角度によって輪郭線や色は類似するが全く別のオブジェクトのペアがデータセットに存在する場合, 前川らの提案する輪郭線によるマッチングでは上手くランキング化することができない場合がある. しかし, 本研究では図 15 に示すように, このような画像がデータベースに存在する場合でも正確にランキング化することができることが分かる. 図 15 は car1_f を入力として, z 軸の回転角を-180 度とした場合の検索結果であり, 検索対象は car1_b である. white2 や white3 は car1_b と色や輪郭線が類似しているが図 15 は正解となる画像が 3 位となっており, white2 や white3 はランク外となっている. このように, 角度によって輪郭線や色は類似するが全く別のオブジェクトのペアがデータセットに存在する場合, 前川らの提案するシステムよりも本研究のほうが優れていることが分かる.

4.5.4　角度によって形が大きく異なるオブジェクトに関する考察

図 6 のように前川らの研究では角度によって形が大きく異なるオブジェクトに対する検索は上手くいかないことがある. 本研究で使用した plane3 の正面, 横, 後ろから撮影した画像をそれぞれ図 16 に示す. 図 16 に示すように plane3 は正面や後ろから見た場合には胴体と羽で十字のような輪郭線となっているが, 横から見た場合には羽部分が輪郭線には現れず, 胴体部分のみであるため car を横から見た画像と同じような輪郭線となる. 図 17 に plane3_f を入力として, z 軸の回転角を-90 度とした場

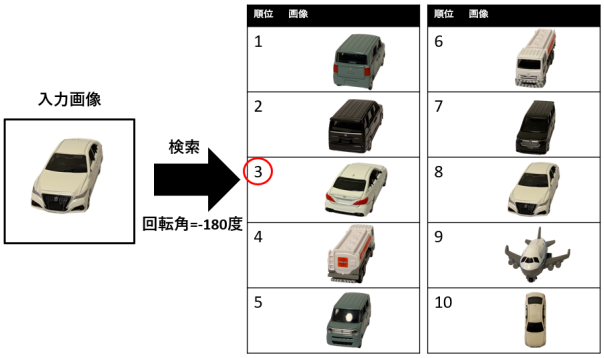


図 15　入力: car1_f, z 軸の回転角: -180 度としたときの検索結果

合の検索結果を示す. この場合検索対象の画像は plane3_s である. 図 17 に示すように正解となる画像が 1 位となっている. このことから, 本研究では前川らのシステムが苦手としている角度によって形が大きく異なるオブジェクトに対する検索が可能だということが分かる.

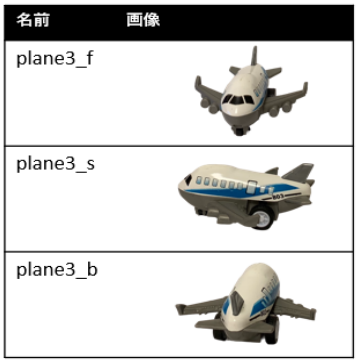


図 16　plane3 を正面, 横, 後ろから撮影した画像

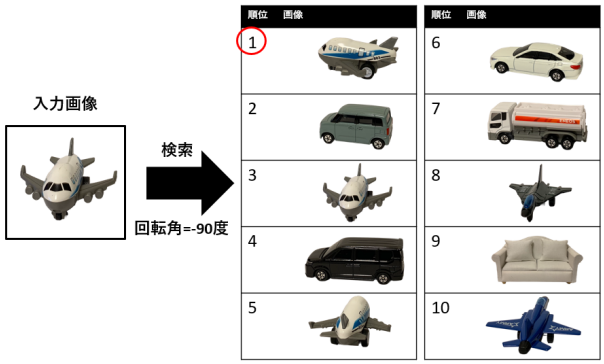


図 17　入力: plane3_f, z 軸の回転角: -90 度としたときの検索結果

5　おわりに

本研究では，画像を入力として，その画像中のオブジェクトをユーザーによって指定された別アングルから見た画像を提示することができる検索システムの構築方法を提案した. 従来の画像検索システムでは，画像を入力して，その画像内のオブジェクトをユーザーの要求するアングルから見た画像をピンポ

イントで検索するといった使い方が困難である．同様の目的をもった研究 [17] [18] では，一般的ではないデプスカメラの利用や、精度の低い 3D モデルの推定により、検索精度の低さが問題点となっていた。そこで本研究では，通常の写真画像を入力とし，より精度の高い Wonder3D [7] を用いて, 入力画像から 3D モデルを推定した．そして, 入力画像の 3D モデルを回転させることで，検索したいアングルを指定することに成功した．さらにデータベース上の全ての画像の 3D モデルも推定し, 回転させた入力画像の 3D モデルと 3D モデルによる特徴量マッチングをすることで, 要求されたアングルの写真を検索結果として提示するシステムを構築した．こうすることで, 輪郭線マッチングでは判定することが困難だった, 角度によって輪郭線や色は類似するが全く別のオブジェクトのペアも判別や, 逆に角度によって輪郭線や色が大きく異なるオブジェクトに対する検索も可能とした．この提案手法を，実画像データを用いた評価実験により検証した．その際，z 軸の回転角を 30 度ごとに変更して実験することで各角度ごとの得意不得意を比較した．実験の結果，生成した 3D モデルを正確に検索する角度の場合精度が向上することがわかった．また, それぞれのオブジェクトについて各角度から生成された 3D モデルの差異が小さいほど検索結果の精度が向上することも分かった．

評価の結果生成された 3D モデルの精度が検索結果に大きな影響を及ぼすことが分かった．本研究で用いた, 2D 画像から 3D モデルを推定する手法 Wonder3D は, 精度が高く, 事前に学習する必要がないため, 広範囲の 2D 画像を 3D モデル化することが可能であったが, 入力された画像からそのオブジェクトの多視点画像を推測するため, 入力画像の角度によっては推定される 3D モデルの奥行がうまく表現されないことがあった．そのため, 実験で用いたオブジェクトの一部は正しく 3D モデル化できずに検索結果が落ちることとなった．今後は 3D モデルに適した軸でスケーリングをする, または同様の精度で角度による影響を受けない 3D モデル手法を検討する必要がある．また, 本手法ではデータベース上の画像をすべて 3D モデル化する必要があるため, 計算コストが膨大になってしまった．そのため, 今後本研究をウェブサービスとして実現するためには, より早く 3D モデルを推定する手法を検討する, そして並列して 3D モデルを推定することができる環境が必要であると考えられる．

謝　　辞

本研究は, JSPS 科研費 JP23H03401, JP23H03694 の助成を受けたものである．

文　　献

[1] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96, 2005.

[2] Corinna Cortes and Vladimir Naumovich Vapnik. Support-vector networks. *Machine Learning*, Vol. 20, pp. 273–297, 1995.

[3] Zhenhua Guo, Lei Zhang, and David Dian Zhang. Rotation invariant texture classification using lbp variance (lbpv) with global matching. *Pattern Recognit.*, Vol. 43, pp. 706–719, 2010.

[4] Ibtihaal M. Hameed, Sadiq H. Abdulhussain, and Basheera M. Mahmmod. Content-based image retrieval: A review of recent trends. *Cogent Engineering*, Vol. 8, , 2021.

[5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

[6] Michael M. Kazhdan, Thomas A. Funkhouser, and Szymon Rusinkiewicz. Rotation invariant spherical harmonic representation of 3d shape descriptors. In *Eurographics Symposium on Geometry Processing*, 2003.

[7] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, 2023.

[8] Graciela Lara López, Adriana Peña Pérez Negrón, Angélica de Antonio Jiménez, Jaime Ramírez Rodríguez, and Ricardo Imbert. Comparative analysis of shape descriptors for 3d objects. *Multimedia Tools and Applications*, Vol. 76, pp. 6993–7040, 2017.

[9] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4460–4470, 2019.

[10] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 24, pp. 971–987, 2002.

[11] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, Vol. 42, No. 3, p. 145–175, may 2001.

[12] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. *2009 IEEE International Conference on Robotics and Automation*, pp. 3212–3217, 2009.

[13] Radu Bogdan Rusu, Nico Blodow, Zoltán-Csaba Márton, and Michael Beetz. Aligning point cloud views using persistent feature histograms. *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3384–3391, 2008.

[14] Amna Sarwar, Zahid Mehmood, Tanzila Saba, Khurram Ashfaq Qazi, Ahmed Adnan, and Habibullah Jamal. A novel method for content-based image retrieval to improve the effectiveness of the bag-of-words model using a support vector machine. *Journal of Information Science*, Vol. 45, pp. 117 – 135, 2018.

[15] Prashant Srivastava and Ashish Khare. Integration of wavelet transform, local binary patterns and moments for content-based image retrieval. *J. Vis. Comun. Image Represent.*, Vol. 42, No. C, p. 78–103, jan 2017.

[16] Yuan Yao, Nico Schertler, Enrique Rosales, Helge Rhodin, Leonid Sigal, and Alla Sheffer. Front2back: Single view 3d shape reconstruction via front to back prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 531–540, 2020.

[17] 雛形奎祐, 欅惇志, 宮崎純. 他視点のオブジェクトを検索可能な画像検索システムの構築. DEIM Forum 2018 論文集, 2018.

[18] 前川由依, 宮崎純. クエリ画像から別アングル画像を検索可能なシステムの構築. 東京工業大学修士論文, 2022.