博 士 論 文

深層生成モデルに基づく
対話型マルチメディア情報検索に関する研究

北海道大学 大学院情報科学院
情報科学専攻

柳 凜太郎

2024 年 3 月

目 次

図 目 次

表 目 次

第1章　序論

1.1　本研究の背景

現在，画像に代表されるマルチメディアコンテンツは個人の携帯端末上やWeb上に広く存在しており，そのデータ量は指数的に増加している [1]．このように増大する多量のマルチメディアコンテンツからユーザの所望する情報を正確かつ簡便に検索可能とする技術は，情報を探索する時間を削減するために有効である事から，様々な研究が行われている [2, 3]．種々の検索技術の中でも，視覚情報と言語情報の対応関係を学習することで，テキストのクエリからユーザの必要とする画像や映像を検索する技術（以降，マルチメディア情報検索）は，ユーザがクエリを簡便に用意可能であることから盛んに研究が行われている [4, 5]．従来研究によりマルチメディア情報検索の高精度化が進む一方で，テキストおよび画像・映像という異なる知覚情報を比較することの困難さやテキスト情報に含まれる曖昧性からマルチメディア情報検索は未だ十分な検索精度に到達していない．

従来のマルチメディア情報検索手法では，クエリテキストおよび検索候補を同一の特徴量空間に射影し，射影された空間において両者を比較することで，目的のマルチメディアコンテンツを検索する [6]．上記の枠組みにより，クエリテキストに関連したマルチメディアコンテンツを高精度に検索可能な枠組みが実現されている．一方，従来手法では，クエリテキストから目的のマルチメディアコンテンツを一意に特定可能であることを前提に設計されている [6, 7]．そのため，検索候補中にクエリテキストに該当するマルチメディアコンテンツが複数存在する場合，一度の検索で高精度な検索結果を得ることは困難である．一般的に，ユーザは検索候補の内容を把握しておらず，検索目的のマルチメディアコンテンツを想起しながらクエリテキストを考案するため，クエリテキストには目的のマルチ

メディアコンテンツを一意に特定可能な程に十分な情報が含まれていない可能性が高い [8,9]．そのため，マルチメディア情報検索技術の更なる高精度化には，クエリテキストに不足している情報を補完することで，検索順位を改善することが可能な技術を構築する必要がある．

クエリテキストに不足している情報を補完するためには，検索候補を絞り込むための手掛かりとなる情報をユーザと検索システムで共有する必要がある [5]．一方で，近年のマルチメディア情報検索はユーザにとって内部構造を理解することが困難な (以降，透明性の低い) 深層学習モデルを利用しており，このような透明性の低い深層学習モデルでは，ユーザと検索システムの間で正確に情報を共有することは困難となる [10, 11]．ユーザと検索システムの間で正確に情報を共有するためには，透明性の低い深層学習モデルの内部構造をそのまま取り扱うのではなく，我々が日常的に取り扱う情報であるテキストや画像などのマルチメディアコンテンツを用いることが望ましい．ここで，近年，テキストや画像などのマルチメディアコンテンツを生成可能な深層生成モデルに関する研究が盛んに行われている [12, 13]．このような深層層生成モデルを活用し，ユーザにとって理解が容易な画像やテキストの形式で検索システム側が把握する情報をユーザと共有することで，クエリテキストに不足している情報を容易に補完することが可能となり，効果的な検索順位の改善に繋がると期待される．

1.2　本研究の目的

本論文では，深層生成モデルによる情報共有を介した，新たな対話型マルチメディア情報検索技術を提案する．具体的に，深層生成モデルを用いて，検索システム側が把握する情報を画像やテキストの形式でユーザに提示することで，ユーザによるクエリテキストの修正を補助し，検索順位の改善を可能とする．本論文では，１）画像生成モデル [14] を応用することにより，ユーザと検索システムのクエリに対する解釈を共有するマルチメディア情報検索手法，２）質問文生成モデル [15] の応用により，検索候補の絞り込みに必要な情報をユーザに要求することが可能なマルチメディア情報検索手法の 2 つのアプローチにより，従来研

究に存在した問題を解決する．まず，１）では，テキストの内容を反映した画像を生成可能な画像生成モデルを用いることで，ユーザと検索システムの間でクエリテキストに対する解釈を共有することが可能な枠組みについて検討している．マルチメディア情報検索において,「クエリテキストに込められたユーザの検索意図」と「クエリテキストに対する検索システムの解釈」を一致させることは，ユーザの必要とする情報を検索するために重要である．特に，マルチメディア情報検索では，クエリがテキストであることから，クエリに多様な意味合いが含まれやすく，上述した解釈の共有は必要不可欠である．このような解釈の共有を図るため，1) ではユーザにより与えられたクエリのテキストを画像に変換し，変換後の画像を「クエリに対する検索システムの解釈」としてユーザに共有し，クエリテキストの修正を支援する手法を考案している．次に，２）では，画像や映像に関連した質問文を生成することが可能な Visual Question Generation (VQG) を用いることで，検索候補を絞り込むために必要な情報を質問応答の形式でユーザに問い合わせることが可能な枠組みについて検討している．ユーザの与えるクエリテキストに十分な情報が含まれず，検索候補中にクエリテキストに該当するマルチメディアコンテンツが複数存在する場合，一度の検索で高精度な検索結果を得ることは困難である．上記の課題を解決するために，2) では検索候補を絞り込むために必要な情報を質問応答の形式でユーザに問いかける検索手法を考案する．考案した手法により，ユーザは提示された質問に回答するだけで大幅に検索結果を改善することが可能である．以上のように，本論文では，様々な深層生成モデルをユーザと検索システムの情報の共有に活用することで，ユーザによる情報の想起を補助し，検索順位を改善することが可能な技術を実現する．

1.3　本論文の構成

本論文は，以下に示す 8 章から構成されている．第 1 章では，本論文の研究背景および目的を述べた．第 2 章では，関連研究としてマルチメディア情報検索や深層生成モデルに関する従来研究を紹介し，本論文で解決すべき課題を明らかにする．第 3 章では，画像生成モデルに基づくマルチメディア情報検索手法を提案

する．第 4 章では，画像生成モデルに基づいて，クエリテキストの入力を補助することが可能な対話型検索手法を提案し，解釈の共有が検索の高精度化に貢献することを示す．第 5 章では，上記により提案した手法が，絵画等の多様なドメインに対しても応用可能であることを確認する．第 6 章では，質問応答に基づく対話型検索の枠組みが検索精度の向上に貢献可能であることを確認するために，定型文形式の質問文を生成可能な再検索手法について検討する．第 7 章では，第 6 章における検証結果を受けて，VQG に基づいて，非定型な質問文を生成することが可能な再検索手法を提案し，検索に必要な情報を質問応答の形式でユーザに問い合わせることの有効性を示す．最後に第 8 章では，本研究の成果を要約し，論文全体のまとめとする．

第2章　関連研究

2.1　はじめに

本章においては，関連性の高いマルチメディア情報検索手法および深層生成モデルに関する従来研究に関して説明を行い，本研究との関連性を説明することで，本研究との差異を明確化する．本研究および従来研究の関連性を表すリサーチマップを図 2.1 に示す．

はじめに，**2.2** 節および **2.3** 節において，従来のマルチメディア情報検索手法について説明する．従来のマルチメディア検索手法は，ユーザにより考案されたテキストクエリから目的のマルチメディアコンテンツを検索する検索手法および検索手法により取得された検索結果に基づいて再度検索を実施することで検索結果を改善する再検索手法に分類できる．そこで，まず，**2.2** 節において検索手法の概要や関連研究について説明したのちに，**2.3** 節において，再検索手法の概要や関連研究について説明を行う．続いて，**2.4** 節および **2.5** 節では，本論文で取り扱う深層生成モデルに関して説明する．具体的に，**2.4** 節では，画像生成モデルについて説明する．続いて，**2.5** 節において，質問文や質問文に関連する回答を生成する質問応答モデルについて説明する．最後に，**2.6** 節で本研究が対象とする解決すべき課題について説明し，**2.7** 節をまとめとする．

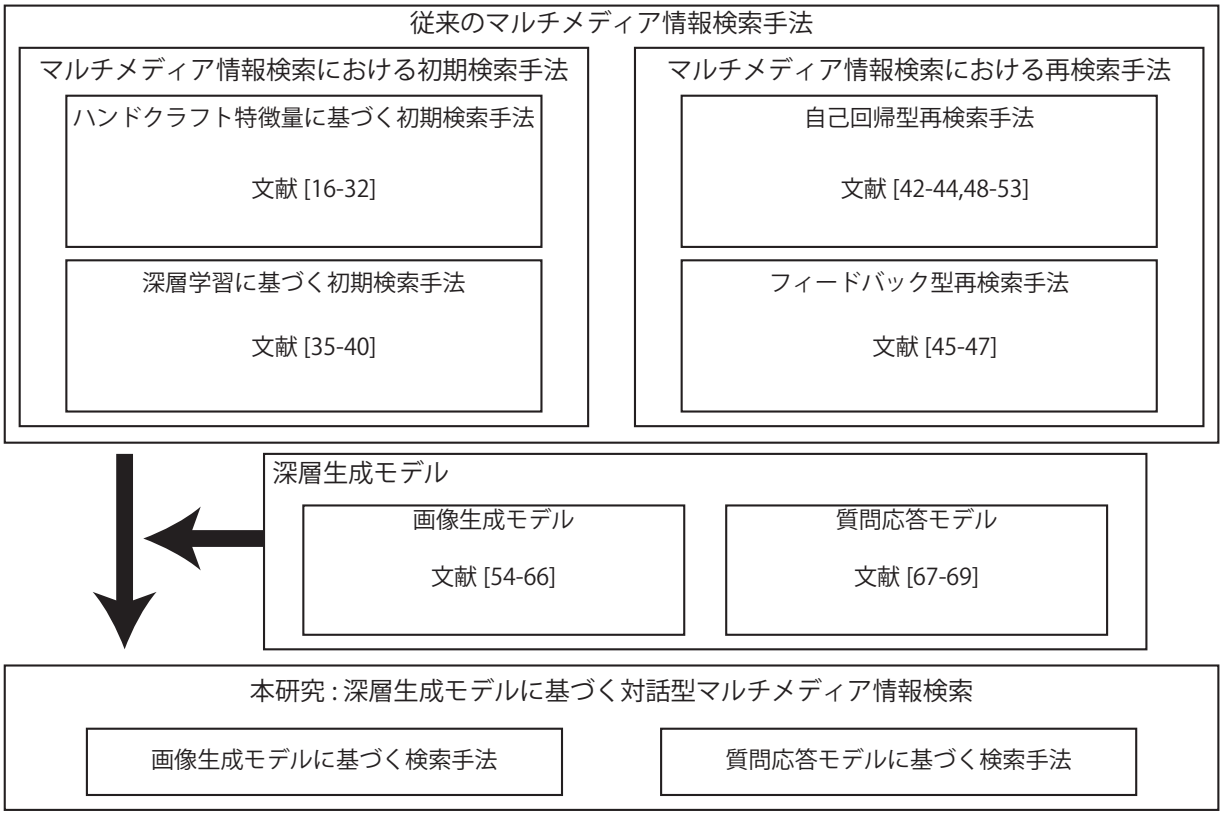

従来のマルチメディア情報検索手法
マルチメディア情報検索における初期検索手法
ハンドクラフト特徴量に基づく初期検索手法
文献 [16-32]
深層学習に基づく初期検索手法
文献 [35-40]
マルチメディア情報検索における再検索手法
自己回帰型再検索手法
文献 [42-44,48-53]
フィードバック型再検索手法
文献 [45-47]
深層生成モデル
画像生成モデル
文献 [54-66]
質問応答モデル
文献 [67-69]
本研究 : 深層生成モデルに基づく対話型マルチメディア情報検索
画像生成モデルに基づく検索手法
質問応答モデルに基づく検索手法

図 2.1: リサーチマップ

2.2　マルチメディア情報検索における検索手法

本節では，ユーザにより考案されたクエリテキストに基づいて目的のマルチメディアコンテンツを検索する検索手法について説明する．以前より，検索手法は，ユーザにより考案されたクエリテキストおよび検索候補のマルチメディアコンテンツを同一な空間に射影し，射影先の空間において両者の類似性を測定することで，目的のマルチメディアコンテンツを検索してきた．特に，従来の検索手法においては，同一な空間への射影方法の改善により，検索精度の向上が実現されている．具体的に，2015 年頃までは，Canonical Correlation Analysis (CCA) や Topic-regression Multi-modal LDA (Tr-mm LDA) などの古くより検討がなされてきた統計的な解析手法により，クエリテキストおよび検索候補を比較している [16,17]．2015 年以降は，深層学習の発展に伴い，Convolutional Neural Network(CNN) や Long-Short Term Memory(LSTM) などの深層学習モデルを用いて言語情報と視覚情報の対応関係を学習することで，ユーザにより考案されたクエリテキストと画像・映像等の検索候補を比較する手法が検討されている．

文献 [16–32] 深層学習以前の検索手法

従来より，マルチメディア情報検索は Scale Invariant Feature Transform (SIFT) や Bag-of-words などに基づいて抽出されるクエリテキストの特徴量および検索候補の特徴を CCA や Tr-mm LDA を用いて対応付けることで，実現されてきた．これらの手法により，単語の出現頻度，回転や位置変化を考慮した検索が実現される一方，対象とするデータに依らず同一な特徴量抽出を施すため，データごとの適切な特徴抽出は困難であった．また，回転や位置変化等の低次の特徴量では，クエリテキストに含有される単語や単語同士の係り付けを適切に考慮した対応付けが困難であった [33,34]．そのため，近年ではテキストやマルチメディアコンテンツに含まれる高次な特徴を取り扱う手法に関して検討がなされている．

文献 [35–42] 深層学習に基づく検索手法

近年の深層学習の発展に伴い，マルチメディア情報検索においても，大規

模なデータセットを用いて言語情報と視覚情報を深層学習モデルに学習させる手法が検討されている．文献 [35] においては CNN および LSTM に基づいて，クエリテキストから抽出されたテキスト特徴量と検索候補より抽出される視覚特徴量を同一の空間において比較することで，クエリテキストに該当するマルチメディアコンテンツの検索を可能としている．また，文献 [36] では文献 [35] の仕組みに加え，単語ごとの順序構造を考慮する枠組みについて検討することで，クエリテキストに含まれる単語の順序関係に焦点を当てた検索手法を提案している．文献 [37] ではクエリテキストの特徴量を画像特徴量の空間に射影することによって文献 [35] や [36] に比べて高精度な検索を実現している．また，文献 [38] や文献 [39] においては，識別が難しいサンプルとの距離が大きくなるように調整された損失やテキストと検索候補の分布間距離を考慮可能な KL 距離を用いた損失の導入により，従来の検索手法の高精度化を実現している．近年では，Transformer [43] と呼ばれる深層学習モデルに基づいて，約 40 億の画像-テキストペアの対応関係を学習した CLIP [40] が提案されており，マルチメディア情報検索のみではなく，多様なタスクに応用されている．検索に関する種々の検討により，クエリテキストに関連するマルチメディアコンテンツを正確に検索可能となった一方で，従来の検索手法においては，クエリテキストに目的のマルチメディアコンテンツを特定するための情報が十分に含まれていない際に検索の性能が低下するという課題が存在した．本論文では，ユーザと検索システムの間で情報共有することにより，上記課題の解決を試みる．

2.3　マルチメディア情報検索における再検索手法

本節では，検索手法により算出された初期検索結果の改善により，検索精度向上を目指す再検索手法について説明する．再検索手法に関する検討は古くから様々な検索タスクで研究が実施されており，検索精度向上に貢献することが広く知られている [5]．特に，従来の再検索手法はユーザと検索システムのインタラクションの有無により自己回帰型再検索手法 [44–46] およびフィードバック型再

検索手法 [47–49] に分別される．自己回帰型再検索手法は初期検索から取得される検索結果に基づいて，検索順位について再計算することで，検索精度の改善を図る．一方で，フィードバック型再検索手法では初期検索により取得された初期検索結果を基にユーザからフィードバックを受け取ることで，検索結果を改善する．

文献 [44–46, 50–55] 自己回帰型再検索手法

自己回帰型再検索手法は，初期検索結果において上位にあたるマルチメディアコンテンツを用いることで，検索結果の改善を図る．特に，文献 [44–46] では，初期順位付けされた各検索候補をクエリとして利用し，テキストのサンプルを検索する．その後，以上により算出されたテキストの検索順位と初期の検索順位の対応関係が正しくなるように初期の検索結果を調整することで，検索結果の改善を図る．このような自己回帰型再検索手法は，ユーザからのフィードバック情報がなくとも初期検索性能を向上可能である一方、クエリテキスト以外の追加の情報は取得困難であるため，必ずしも検索精度の向上に寄与しない可能性が存在する．

文献 [47–49, 56] フィードバック型再検索手法

フィードバック型再検索手法は，ユーザからのフィードバックを介して検索結果を改善させることを目的としている．文献 [47] では，画像間の差分を説明するテキストを学習することで，検索結果上位のマルチメディアコンテンツに対するユーザからのフィードバックコメント受け取り，そのコメントを用いることで目的のマルチメディアコンテンツを検索する．また，文献 [48] では，クエリテキストのみではなく，画像もクエリとして利用可能な枠組みについて検討することで，テキストのみならず画像によるフィードバックを再検索に利用している．文献 [49] では，複数のクエリテキストを処理可能な枠組みを考案することで，ユーザから複数回のフィードバックをテキストを介して受け取る枠組みを実現している．

2.4　画像生成モデル

本節では，関連研究として，画像生成モデルについて説明する．画像生成は大量の画像データを用いて CNN 等の深層学習モデルを学習することで，写実的な画像を生成する．以前は，ある特定集合の画像生成を目的として様々な研究が行われており，近年では，入力テキストの内容を反映した画像の生成を目的とした研究がなされている．本論文では，テキストを入力とする画像生成モデルをマルチメディア情報検索の分野に応用することにより，ユーザおよび検索システムで検索に関連した情報を共有可能な枠組みを構築する．

文献 [57–62] 条件なし画像生成モデル

近年の深層学習の発展に伴い， Variational Autoencoder， Generative Adversarial Network(GAN), Flow-based モデルなど様々な画像生成モデルが提案されている．多様な画像生成モデルの中でも，GAN は min-max ゲーム学習戦略に基づいて生成モデルの学習を行うことにより，写実的な画像を生成可能な深層学習モデルである．具体的には，生成モデルおよび識別モデルと呼ばれる二つのモデル用いて，競争的な学習を行うことで，高精度な画像生成を実現する．また，上記の競争的な学習の安定化を目的とした学習機構について検討がなされている他，生成させる画像の解像度を大幅に向上可能なモデル構造が検討されている．一方で，従来の画像生成モデルは，生成される画像の内容等は指定することが出来ず，その応用性に課題が存在していた．

文献 [63–72] テキストを条件とする画像生成モデル

画像生成モデルの応用性を向上させることを目的として，近年では，テキストを条件として画像生成可能な手法の検討がなされている．文献 [66] はテキストを入力とする画像生成モデルの初期の研究であり，テキストから抽出されたテキストの内容を表現する特徴量を画像生成モデルに入力することで，テキストの内容を反映した画像の生成を実現した．しかし，その解像度は 64×64 画素であり，生成画像の写実性は低いものであった．文

献 [66] の課題解決を目的として，複数生成器を階層的に組み合わせる画像生成モデルが提案されている [67–70]．文献 [68] では，入力されるテキストの単語ごとの詳細な特徴に着目可能な生成モデルを導入することで，入力されるテキストの詳細な情報に着目可能な画像生成を実現している．また，文献 [69] では，生成画像からテキストを生成し，入力されるテキストと生成画像を入力とすることで生成されるテキストの比較により，入力したテキストに含まれる意味内容が生成画像に反映されることを保証する枠組みを提案している．本論文では，上述したテキストを入力とする画像生成モデルを用いることで，ユーザにより入力されたクエリテキストの内容を反映した画像を生成することで，クエリテキストの内容に対する解釈をユーザおよび検索システムで視覚的に共有可能な枠組みを構築する．

2.5　質問応答モデル

本節では，本研究の関連研究として，質問応答モデルについて説明する．質問応答モデルは，大量の画像とその画像に関連する質問およびその回答がペアとなったデータを深層学習モデルに基づいて学習することで，画像に関連した質問応答のタスクを解決する．具体的に，画像および画像に関連する質問文から回答を推定するタスクや画像の内容に関連する質問文を生成するタスク等について検討がなされている．

文献 [73–76] Visual Question Answering および Visual Question Generation

近年の深層学習の発展に伴い，質問応答に関しても種々の深層学習モデルが提案されている．文献 [73] では，Visual Question Answering(VQA) と呼ばれる画像および画像に関連した質問文から回答を推定するタスクの解決を試みている．具体的に，文献 [74] において提案されるデータセットに基づいて，画像の特徴量を抽出するモデル，質問の特徴量を抽出するモデルおよび回答を推定するモデルを学習することで，画像および画像の内容から回答することが可能な質問文から回答を推定する．また，文献 [75] では，

Visual Question Generation(VQG) と呼ばれる画像から画像に関連する質問文を生成するタスクに取り組んでいる．本論文では，上記の VQA や VQG の深層生成モデルを用いることで，質問応答の形式でユーザと情報共有可能な再検索手法を構築している．

2.6　本論文で解決すべき課題と解決方法

本節では，深層生成モデルによる情報共有を介した，対話型マルチメディア情報検索の実現を目的として，本論文で解決すべき課題を明確にする．

2.2 節で説明した研究では，ユーザにより考案されたクエリテキストから目的のマルチメディアコンテンツを検索する初期検索手法について検討されている．特に，近年の初期検索手法では，CNN に代表される深層学習モデルを用いて，クエリテキストおよび検索候補を共通な空間に射影し，射影先の空間において比較することで，目的のマルチメディアコンテンツを検索する．このような初期検索手法により，クエリテキストに該当するマルチメディアコンテンツを高精度に検索可能である．しかしながら，初期検索手法では，クエリテキストに目的のマルチメディアコンテンツを特定するために必要な情報が十分に存在していない場合，目的のマルチメディアコンテンツの検索は困難であるという課題が存在していた．本論文では，上記の課題を解決するために，ユーザおよび検索システムの間で検索に有効な情報を共有することで，クエリテキストに不足している情報を補完可能なマルチメディア情報検索手法を提案する．

本研究に最も関連した研究として，**2.3** 節で説明した，再検索手法が挙げられる．再検索手法においては，ユーザに初期検索結果を提示し，様々な方法でユーザからフィードバックを受け取ることで，情報の補完を試みる．しかしながら，従来の再検索手法では，ユーザは「検索システムによるクエリテキストの解釈」や「検索候補を絞り込むために必要な情報」を把握しない状態で，情報を補完する必要があるため，必ずしも補完した情報が検索結果を改善するために有効であるとは限らなかった．また，検索システムにより提示される情報は初期の検索結果のみであり，ユーザと検索システムの間で検索精度を向上するための情報を適

切に共有可能な枠組みは構築されていない．本論文では,「検索システムによるクエリテキストの解釈」や「検索候補を絞り込むために必要な情報」などの検索精度向上に有効な情報をユーザと検索システムの間で共有可能な枠組みを構築することで，従来の検索手法よりも高精度なマルチメディア情報検索手法を構築する．

検索精度向上に有効な情報をユーザと検索システムの間で共有することにより，クエリテキストに不足していた情報を補完可能になると期待される．一方で,近年のマルチメディア情報検索はユーザにとって内部構造を理解することが困難な (以降，透明性の低い) 深層学習モデルを用いており，このような透明性の低い深層学習モデルでは，ユーザと検索システムの間で正確に情報を共有することは困難となる．そのため，検索精度の向上に必要な情報をユーザと検索システムの間で共有する枠組みを構築するためには，検索精度を向上するために有用な情報をユーザにとって理解可能な形式で表現可能な検索手法について検討する必要が存在する．そこで，本論文では，ユーザと検索システムの情報共有に **2.4** 節や **2.5** 節で説明した画像生成モデルや質問応答モデルなどの深層生成モデルを用いる．具体的に，画像生成モデルや質問応答モデルを用いることで,「検索システムによるクエリテキストの解釈」や「検索候補を絞り込むために必要な情報」を画像やテキストなどの情報で表現することで，ユーザと検索システムの間で情報を共有し，検索結果の改善を図る．

本論文では，はじめに,「検索システムによるクエリテキストの解釈」の視覚的な表現を目的として，画像生成モデルに基づくマルチメディア情報検索手法を提案する．具体的に，第 3 章では，画像生成モデルに基づくマルチメディア情報検索手法が目的のマルチメディアコンテンツを高精度に検索可能であることを確認する．次に，第 4 章では，画像生成モデルに基づいて，クエリテキストの入力を補助することが可能な対話型検索手法を提案し,「検索システムによるクエリテキストの解釈」の共有が検索の高精度化に貢献することを示す．また，第 5 章では，上記により提案した手法が，絵画等の多様なドメインに対しても応用可能であることを確認する．続いて，本論文では,「検索候補を絞り込むために必要な情報」をユーザと共有するために，質問応答モデルに基づいたマルチメディア情

報検索手法を提案する．具体的に，第 6 章では，質問応答に基づく対話型検索の枠組みが検索精度向上に寄与することを確認するために，定型文形式の質問文を生成可能な再検索手法について検討する．また，第 7 章では，第 6 章における検証結果を受けて，VQA に基づいて，非定型な質問文を生成することが可能な再検索手法を提案し，検索に必要な情報を質問応答の形式でユーザに問い合わせることの有効性を示す．

2.7　まとめ

本章では，本研究に関連するマルチメディア情報検索，マルチメディア情報再検索，画像生成モデルおよび質問応答モデルに関して説明を行った．加えて，従来研究の課題や本研究との関連性を説明することで，マルチメディア情報検索を高精度化するために必要な課題を明確にした．

第3章　画像生成モデルに基づくマルチメディア情報検索

3.1　はじめに

本章では，画像生成モデルに基づくマルチメディア情報検索手法を提案し，その検索精度を検証する．マルチメディア情報検索において,「クエリテキストに込められたユーザの検索意図」と「クエリテキストに対する検索システムの解釈」を一致させることは，ユーザの必要とする情報を検索するために重要である.「クエリテキストに対する検索システムの解釈」が「クエリテキストに込められたユーザの検索意図」と異なる場合，検索結果はユーザの望む結果とは異なるものとなり，検索精度は低下する．特に，マルチメディア情報検索では，クエリがテキストであることから，クエリに多様な意味合いが含まれやすく，上述した解釈の不一致は容易に生じる．また，近年のマルチメディア情報検索手法は，ユーザにとって内部構造を理解することが困難な深層学習モデルを採用していることが多く，検索結果がユーザの満足いく結果ではない場合に，検索システムがどのようにクエリテキストを解釈し，検索を行ったのかを把握することは困難である．そのため，ユーザの意図する検索結果が得られない場合にも，適切に目的のマルチメディアコンテンツを検索するためには,「クエリテキストに対する検索システムの解釈」をユーザに理解可能な形式で共有することが可能なマルチメディア情報検索手法を構築する必要がある．

そこで，本論文では,「クエリテキストに対する検索システムの解釈」を視覚的な情報としてユーザに共有することで，クエリテキストの修正を支援する手法を提案する．具体的に，提案手法では，深層生成モデルの一種である画像生成モデルを用いることで，ユーザにより考案されたクエリテキストを画像に変換する．

その後，変換された画像をそのままマルチメディア情報検索に利用するとともに，「クエリテキストに対する検索システムの解釈」としてユーザに共有することで，クエリテキストの修正を支援する．

本章では，上記のマルチメディア情報検索手法の前段階として，画像生成モデルに基づくマルチメディア情報検索手法が目的のマルチメディアコンテンツを正確に検索することが可能であることを確認する．これにより，画像生成モデルがマルチメディア情報検索に応用可能であることを確認する．

3.2　画像生成モデルに基づく画像検索

本節では，画像生成モデルに基づくマルチメディア情報検索手法について説明する．

3.2.1　画像生成モデルを用いたクエリ画像の生成

はじめに，クエリ画像の生成に関して画像生成モデルの一種である Attentional Generative Adversarial Netowrk (AttnGAN) [68] について説明を行う．AttnGAN は，3 つのニューラルネットワーク F_r ($r \in \{l, m, h\}$) とその出力である特徴ベクトル $\boldsymbol{s}_r$ ($r \in \{l, m, h\}$) および画像生成ネットワーク G_h から構成されている．以上の AttnGAN に対してクエリテキストより算出されたテキスト特徴量 $\boldsymbol{e}_{\text{sen}}$ およびクエリテキストに含まれる各単語より算出された特徴量により構成される単語特徴量行列 $\boldsymbol{E}_{\text{word}}$ を入力することで，クエリテキストを表現する画像を生成する．

まず，テキスト特徴量 $\boldsymbol{e}_{\text{sen}}$ およびガウス雑音 $\boldsymbol{z}$ からニューラルネットワーク F_l を用いてベクトル $\boldsymbol{s}_l$ を算出する．

$$\boldsymbol{s_l} \quad = \quad F_l(\boldsymbol{z}, F^{\text{ca}}(\boldsymbol{e}_{\text{sen}})) \tag{3.1}$$

ただし，F^{ca} は文献 [77] により提案された，学習を安定化させるための関数である．ここで，$\boldsymbol{s}_l$ はテキスト特徴量のみから算出されるためテキストの構造の情報

に着目した特徴ベクトルとなる．

次に，以上で算出された $\boldsymbol{s}_l$ および単語特徴量行列 $\boldsymbol{E}_{\text{word}}$ から F_m を用いて $\boldsymbol{s}_m$ を算出する．また，$\boldsymbol{s}_h$ についても同様に $\boldsymbol{s}_m$ および $\boldsymbol{E}_{\text{word}}$ を用いて算出する．

$$\boldsymbol{s}_m = F_m(\boldsymbol{s}_l, F_m^{\text{attn}}(\boldsymbol{E}_{\text{word}}, \boldsymbol{s}_l)) \tag{3.2}$$

$$\boldsymbol{s}_h = F_h(\boldsymbol{s}_m, F_h^{\text{attn}}(\boldsymbol{E}_{\text{word}}, \boldsymbol{s}_m)) \tag{3.3}$$

ここで，F_r^{attn} $(r \in \{m, h\})$ は $\boldsymbol{s}_r$ $(r \in \{l, m\})$ に $\boldsymbol{E}_{\text{word}}$ を組み込むニューラルネットワークである．$\boldsymbol{s}_r$ $(r \in \{m, h\})$ は一つ前のニューラルネットワークの出力および単語特徴量行列 $\boldsymbol{E}_{\text{word}}$ を用いて算出される．そのため，$\boldsymbol{s}_m$，$\boldsymbol{s}_h$ と処理が進むごとに，テキストの構造の情報のみでなく単語ごとの詳細な情報を含む特徴ベクトルが算出される．

最後に，$\boldsymbol{s}_h$ から画像生成ネットワーク G_h を用いて画像 Q_h を生成する．

$$Q_h = G_h(\boldsymbol{s}_h) \tag{3.4}$$

以上により生成された画像 Q_h は $\boldsymbol{s}_h$ から生成されているためテキストの構造および単語ごとの詳細な情報に着目した画像となる．

3.2.2　生成画像を用いたマルチメディア情報検索

本節では，生成画像を用いた画像検索について説明する．はじめに，検索候補の各画像を I^n $(n = 1, 2, ..., N; N$ は検索候補の画像枚数) と定義する．その後，生成画像 Q_h および検索候補 I^n から画像特徴量 $\boldsymbol{f}^{\text{g}} \in \mathbb{R}^D$ および $\boldsymbol{f}^n \in \mathbb{R}^D$ を抽出する．ここで D は画像特徴量の次元数を表す. 続いて，画像特徴量 $\boldsymbol{f}^{\text{g}}$ および $\boldsymbol{f}^n$ から以下に示すコサイン類似度 w_n を算出する．

$$w_n = \frac{\boldsymbol{f}^{\text{g}} \cdot \boldsymbol{f}^n}{|\boldsymbol{f}^{\text{g}}||\boldsymbol{f}^n|} \quad (n = 1, 2, ..., N) \tag{3.5}$$

最後に，算出された w_n の降順に整列した検索候補を検索結果とする．

3.2.3　実験: 実験設定

本節では，画像生成モデルにより生成された画像を用いて目的の画像を検索可能であることをを確認するための実験について説明する．本実験では検索候補として 11,788 枚 200 種類の鳥の画像により構成されている Caltech-UCSD Birds (CUB) データセット [78] を用いた．ただし，CUB データセットの各画像に対しては画像を説明する文 (以降，テキストラベル) が付与されている．また，本実験では，文献 [79] を参考にして，CUB データセットのうち 7,221 枚を学習用データ，4,567 枚をテストデータとした．具体的に，本実験では，学習用データを用いて画像生成モデルを学習した後に，4,567 枚のテストデータおよび各テストデータに付与されているテキストラベルをそれぞれ検索候補およびクエリテキストとして利用することで提案手法の検索精度を確認した．提案手法における画像特徴量には ImageNet [80] で学習済みの Inception-v3 [81], VGG-16 [82],ResNet-50 [83] のいづれかを用いた．ただし，本実験における評価指標として，以下の式により定義される Recall@k を用いた．

$$\text{Recall@}k = \frac{t_k}{s} \quad (k = 1, 2, ..., N). \tag{3.6}$$

ここで t_k は k 位以内に画像が正しく検索されたクエリの個数を示す．ただし，今回の実験ではクエリテキストが付与されていた画像と同種の鳥の画像が検索された場合を正しく検索されたと定義した．

3.2.4　実験: 定量評価

図 3.1 に実験の結果を示す．図 3.1 において，“Base Line” はランダムに順位を決定した場合の Recall@k を示しており，“Inceptio-v3”,“VGG-16” および “ResNet-50” はそれぞれの画像特徴量を用いた提案手法の Recall@k を示している．図 3.1

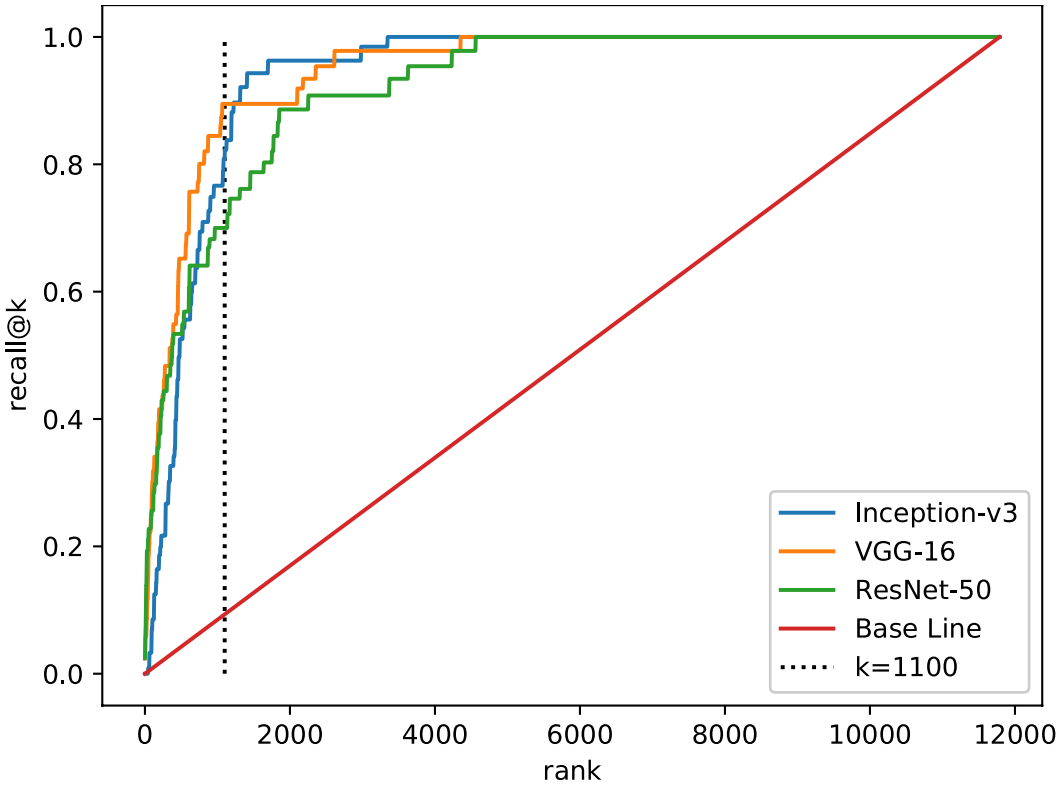


図 3.1: 生成画像を用いたマルチメディア情報検索の定量評価．縦破線はデータセットサイズの十分の一の順位を表す．

の結果より，提案手法 (“Inceptio-v3”,“VGG-16” および “ResNet-50”) がランダムに順位を決定した場合 (“Base Line”) と比較して，高精度に目的の画像を検索していることが分かる．以上の結果より画像生成モデルにより生成された画像が画像特徴量空間 $\mathcal{V}$ におけるマルチメディア情報検索に利用可能であることを確認した．

3.3　画像生成モデルに基づく映像シーン検索

本節では，提案手法である画像生成モデルを用いた映像シーン検索手法について説明する．提案手法の概要図を図 3.2 に示す．提案手法は 2 つの段階で構成されている．まず，第 1 段階では，文献 [68] に基づく画像生成モデルにより入力テキストから 3 枚の異なる解像度の画像を生成する．それぞれの画像は解像度が高くなるにつれてテキスト構造の特徴から単語それぞれの特徴へと着目される点が変化する．次に第 2 段階では，上述の 3 枚の生成画像を低解像度から高解像度へと順に用い検索対象を絞り込むことにより，目的のシーンを高精度に検索する．具体的には，それぞれの生成画像と映像のフレームとの類似度を画像特徴量を元に算出し，類似度が高いフレームを含むシーンの順に検索順位を決定する．

以降，**3.3.1** 節で敵対的生成ネットワークによる画像生成について説明し，**3.3.2** 節で生成画像を用いた階層的シーン検索について説明する．

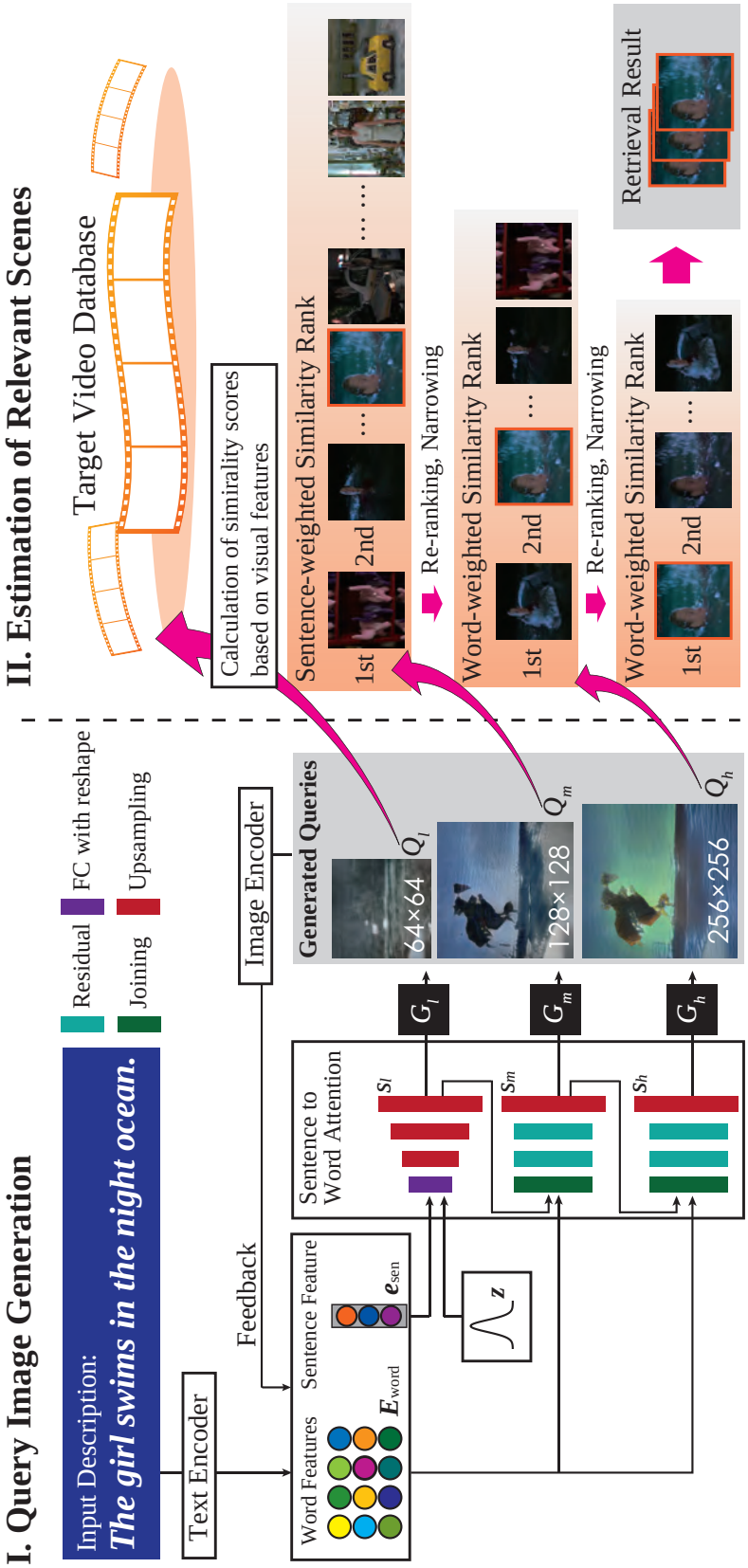


図 3.2: 生成画像を用いたマルチメディア情報検索の概要図.

3.3.1　画像生成モデルによる画像生成

本節では，文献 [68] に基づく画像生成モデルによる，画像生成について説明する．提案手法で構築する文献 [68] に基づく画像生成モデルは，3 つのニューラルネットワーク $F_r(r \in \{l, m, h\})$ とその出力である特徴ベクトル $\boldsymbol{s}_r(r \in \{l, m, h\})$ および 3 つの画像生成器 $G_r(r \in \{l, m, h\})$ から構成されている．また，クエリテキストより算出されたテキスト特徴量 $\boldsymbol{e}_{\text{sen}}$ およびクエリテキストに含まれる各単語より算出された特徴量により構成される単語特徴量行列 $\boldsymbol{E}_{\text{word}}$ を画像生成モデルの入力として用いる．

まず，テキスト特徴量 $\boldsymbol{e}_{\text{sen}}$ およびガウス雑音 $\boldsymbol{z}$ からニューラルネットワーク F_l を用いてベクトル $\boldsymbol{s}_l$ を算出する．

$$\boldsymbol{s_l} = F_l(\boldsymbol{z}, F^{\text{ca}}(\boldsymbol{e}_{\text{sen}})) \tag{3.7}$$

ただし，F^{ca} は文献 [77] により提案された，学習を安定化させるための関数である．ここで，$\boldsymbol{s}_l$ はテキスト特徴量のみから算出されるため文全体の情報のみを含む特徴ベクトルとなる．

次に，以上で算出された $\boldsymbol{s}_l$ および単語特徴量行列 $\boldsymbol{E}_{\text{word}}$ から F_m を用いて $\boldsymbol{s}_m$ を算出する．また，$\boldsymbol{s}_h$ についても同様に $\boldsymbol{s}_m$ および $\boldsymbol{E}_{\text{word}}$ を用いて算出する．

$$\boldsymbol{s}_m = F_m(\boldsymbol{s}_l, F_m^{\text{attn}}(\boldsymbol{E}_{\text{word}}, \boldsymbol{s}_l)) \tag{3.8}$$

$$\boldsymbol{s}_h = F_h(\boldsymbol{s}_m, F_h^{\text{attn}}(\boldsymbol{E}_{\text{word}}, \boldsymbol{s}_m)) \tag{3.9}$$

ここで，$F_r^{\text{attn}}(r \in \{m, h\})$ は $\boldsymbol{s}_r(r \in \{l, m\})$ に $\boldsymbol{E}_{\text{word}}$ を組み込むニューラルネットワークである．$\boldsymbol{s}_r(r \in \{m, h\})$ は一つ前のニューラルネットワークの出力および単語特徴量行列 $\boldsymbol{E}_{\text{word}}$ を用いて算出されている．そのため，$\boldsymbol{s}_m$，$\boldsymbol{s}_h$ と処理が進むごとに，文全体の特徴のみでなく単語ごとの特徴に着目した情報を含む特徴ベクトルが算出される．

最後に，$\boldsymbol{s}_l$，$\boldsymbol{s}_m$，$\boldsymbol{s}_h$ から以下の式を用いて 3 つの画像 Q_l，Q_m，Q_h を生成する．

$$Q_r = G_r(\boldsymbol{s}_r) \quad (r \in \{l, m, h\}) \tag{3.10}$$

以上により生成された Q_l は $\boldsymbol{s}_l$ により生成されているため文全体の特徴より得られる情報のみを持つ画像となり，Q_m，Q_h は $\boldsymbol{s}_m$ および $\boldsymbol{s}_h$ から生成されているため文全体および単語それぞれの特徴にも着目した情報を持つ画像となる．ただし，Q_l は 64×64，Q_m は 128×128，Q_h は 256×256 ピクセルの画像となる．提案手法では Q_l，Q_m，Q_h の 3 種の特性が異なる画像を階層的に用いることで検索精度の向上を実現する．

3.3.2　生成画像を用いた階層的シーン検索

本節では，生成画像を用いた階層的シーン検索について説明する．階層的シーン検索は 3 段階によって構成されている．

はじめに，検索対象の映像のフレームを f_{n_t} ($n_t = 1, 2, ..., N; N$ は映像のフレーム数) とする． 続いて，前節で生成した画像 Q_l から画像特徴量 $\boldsymbol{v}_l \in \mathbb{R}^D$ を，f_{n_t} から画像特徴量 $\boldsymbol{v}_{n_t} \in \mathbb{R}^D$ を算出する．ここで，D は画像特徴量の次元数を表す．ただし，提案手法では ImageNet [84] により学習済みの Inception-v3 [81] の第 3 プーリング層の出力を画像特徴量として用いる． 次に，Q_l および f_{n_t} の類似度として，$\boldsymbol{v}_l$ および $\boldsymbol{v}_{n_t}$ のコサイン類似度 w_{n_t} を算出する．

$$w_{n_t} = \frac{\boldsymbol{v}_l \cdot \boldsymbol{v}_{n_t}}{|\boldsymbol{v}_l||\boldsymbol{v}_{n_t}|} \quad (n_t = 1, 2, ..., N) \tag{3.11}$$

その後，類似度 w_{n_t} からフレームごとの順位 r_l を算出する．この順位は，文全体の特徴に着目した生成画像を用いて算出されるため文全体の特徴に着目した検索順位となる．

続いて，第 2 ステップでは前節で生成した文全体および単語の詳細な特徴に着目した画像 Q_m と r_l の上位 50 %に含まれるフレームに対して，同様に画像特徴量および類似度を算出しフレームごとの順位 r_m を求める．

最後に，第 3 ステップでは Q_h と r_m の上位 50 %に含まれるフレームに対して，同様にフレームごとの順位 r_h を算出する．これらの r_m および r_h はテキストの構造および単語の詳細な情報に着目して生成された画像 Q_m および Q_h を用いて算出されている．そのため，主に検索対象の被写体に着目した検索順位となっている．以上の検索の後に，r_h の中から最も高い順位であったフレーム I_o を算出し，その I_o が含まれるシーン S_o を検索結果として選出する．以上のように文全体の特徴に着目した情報から単語の詳細な特徴に着目した情報へと検索対象を絞り込むことで高精度なシーン検索を実現する．

3.3.3　実験: 実験概要

提案手法の有効性を確認するための定量評価および定性評価について説明しその結果を示す．

3.3.4　実験: データセット

本節では両実験で用いるデータセットについて説明する．本実験では，提案手法で用いる画像生成モデルの学習用データセットとして，33 万枚の画像で構成された Common Objects in Context (COCO) [85] データセットを用いた．ただし，COCO データセットにおいては，1 画像あたり 5 つの説明文が付与されている．また，評価用のデータセットとして 94 本の映画より抜粋された 68,000 シーンにより構成される MP2-MD データセット [86] を用いた．MP2-MD データセットに関しても同様に，それぞれのシーンに対して 1 つの説明文が付与されている

3.3.5　実験: 定量評価

本節では定量評価に関して説明し，その結果を示す．はじめに，MP2-MD データセットよりそれぞれ三本の映画 “Bad Santa” (538 シーン)，“As Good As it Gets” (430 シーン) および “Harry Potter and the Prisoner of Azkaban” (592 シーン) を選択した．本実験ではそれぞれのシーンに付与されている説明文を入力として画像を生成し，映画ごとに生成画像を用いて検索を行い，その検索精度を評価した．検索精度を評価するため k 位以上の再現率を測定することが可能な Recall@k を算出した．

$$\text{Recall@}k = \frac{t_k}{M} \quad (k = 1, 2, ..., N) \tag{3.12}$$

ここで t_k は正解が k 位以上に存在するシーンの個数を，M は検索対象の映像の総シーン数を示す．この評価により単一映像からの検索を仮定した場合の提案手法の有効性を確認する．ただし，順位はフレームごとに算出し，入力の説明文が付与されていたシーンと検索されたフレームが属するシーンが一致した場合を正解とした．また，比較手法として Convolutional Neural Network および Long-Short Term Memory を用いて画像と文より算出される特徴量をそれぞれ同一な空間に射影し比較する手法 (CM1) [35]，CM1 に加えて単語ごとの順序構造を考慮した手法 (CM2) [36]，文を視覚的特徴量の空間上に射影し比較する手法 (CM3) [37]，CM1 の学習の仕組みを検索用に改良した最新手法 (CM4) [38]，Q_h のみを用いて検索する手法 (BL) を用いた．CM1 と比較することで従来の基準となる手法との比較を行い，CM2 と比較することで提案手法がテキストの構造を考慮しているかを確認した．また，CM3 と比較することで画像を生成することの有効性を確認し，CM4 と比較することで最新の手法と比較した時の精度の差を確認した．ただし，全ての比較手法は COCO データセットで学習されている．

図 3.3-3.5 に実験結果を示す．図 3.3-3.5 において CM1，CM2，CM3 および CM4 の精度を提案手法 (PM) が上回っていることが確認できる．以上のことより，提案手法の有効性を定量的に確認した．また，同様に図 3.3-3.5 より提案手法 (PM)

が BL の検索精度を上回っていることが確認できる．以上より，画像生成モデルで生成される複数の画像を階層的に用いることの有効性が確認された．

上述の実験に加え，提案手法の頑健性を評価するために様々なシーンを含む複数の映像から目的のシーンを検索し，評価を行った．複数の映像として MP2-MD データセットより“Bad Santa”，“As got as it gets”，“Halloween”，“Rendezvous mit Joe Black”および“Harry Potter and the Prisoner of azkaban”の 5 つの映画を無作為に選択し，評価用データセットを構築した．ただし，それぞれ 430，538，676，296，592 シーンが含まれる．この評価により複数映像からの検索を仮定した場合の提案手法の頑健性を確認する．実験結果を図 3.6 に示す．図 3.6 より提案手法が他の手法の精度を上回ることが確認できる．以上より，提案手法の頑健性が確認された．

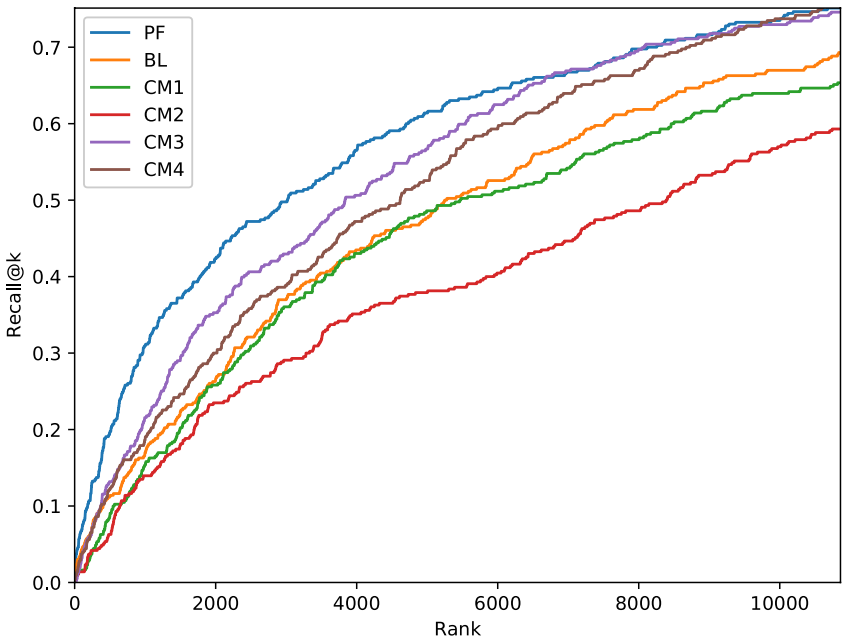


図 3.3: “Bad Santa” を検索対象とした際の Recall@k.

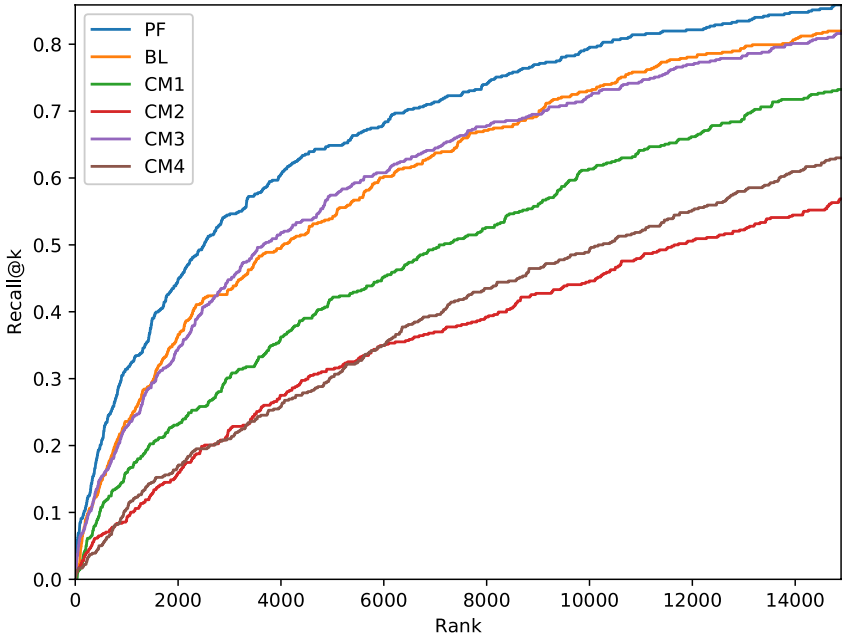


図 3.4: “As Good As it Gets” を検索対象とした際の Recall@k.

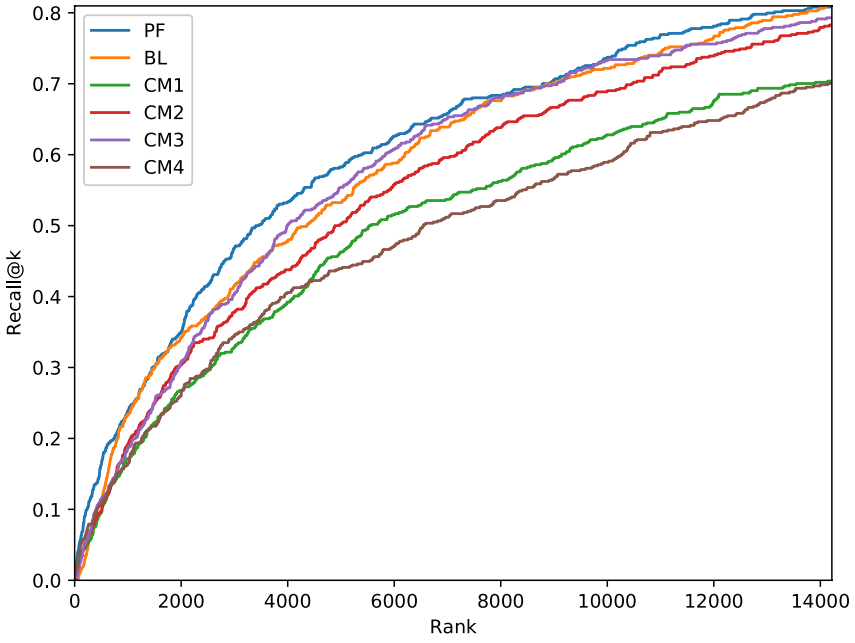


図 3.5: “Harry Potter and the Prisoner of Azkaban” を検索対象とした際の Recall@k.

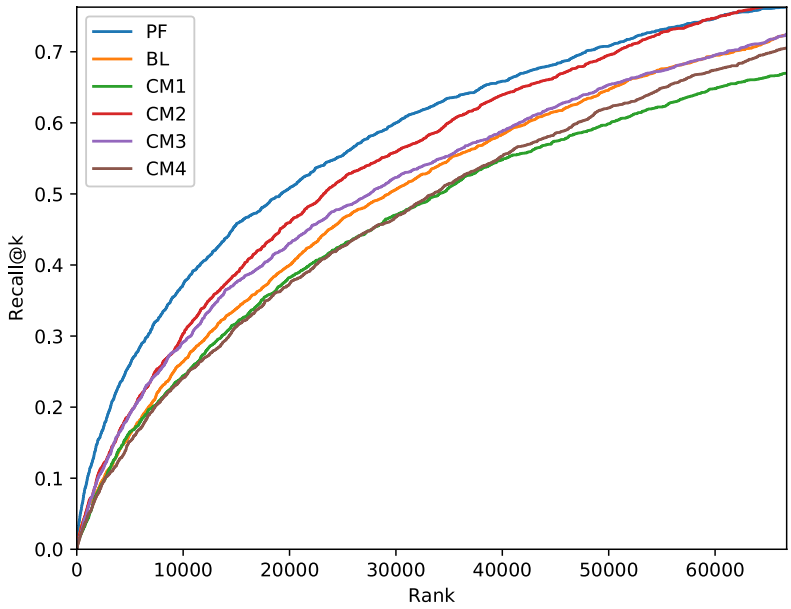


図 3.6: 複数の映像を検索対象とした際の Recall@k.

表 3.1: 画像生成モデルに基づくマルチメディア情報検索手法における被験者実験の結果.

	シーン 1	シーン 2	シーン 3	シーン 4	シーン 5	シーン 6	シーン 7	シーン 8	シーン 9	シーン 10
PM	**4.12**	**3.76**	**3.92**	**4.40**	**3.84**	**4.16**	**4.84**	4.44	**4.72**	**3.32**
BL	2.80	3.60	1.60	3.20	2.88	3.80	3.68	**4.76**	4.20	**3.32**
CM1	1.04	1.04	1.04	1.02	2.04	1.84	1.08	1.08	1.68	1.04
CM2	1.04	1.48	1.60	1.72	2.44	1.08	1.08	1.32	3.68	2.48
CM3	3.48	3.60	1.04	3.04	2.04	**4.16**	3.68	4.44	4.44	2.96
CM4	1.36	1.40	1.20	1.72	2.12	3.80	1.40	1.48	1.08	1.24

	シーン 11	シーン 12	シーン 13	シーン 14	シーン 15	シーン 16	シーン 17	シーン 18	シーン 19	シーン 20	平均
PM	**4.12**	**4.64**	**2.68**	**4.48**	**4.96**	**4.96**	**3.24**	**4.88**	**4.52**	**3.64**	**4.18**
BL	1.12	1.76	1.36	3.88	3.96	2.84	2.04	3.48	3.44	2.96	3.03
CM1	1.12	1.40	1.08	3.20	1.52	3.08	1.92	1.08	1.28	1.20	1.49
CM2	1.24	1.96	1.12	3.88	3.16	4.48	1.92	1.08	4.00	1.60	2.12
CM3	3.24	1.96	2.12	1.64	3.08	2.80	3.16	2.40	4.20	**3.64**	3.06
CM4	1.04	3.68	2.12	3.20	4.88	2.73	2.08	3.48	1.32	1.20	2.13

3.3.6　実験: 定性評価

本節では定性評価について説明し，その結果を示す．本実験では，定性的な評価を行うために 25 名の実験参加者による被験者実験を行った．まずはじめに MP2-MD データセットより無作為に選択した 20 個のシーンに対し，そのシーンを表現する文を作成した．その後，作成した文をクエリとして提案手法および **3.3.5** 節と同様の比較手法を用いて検索結果を算出した．その後，クエリテキストとそれぞれの検索結果との一致度を実験参加者に 5 段階 (1:“ 一致していない”, 2:“ 少し一致している”，3:“ どちらとも言えない”, 4:“ 少し一致している”，5:“一致している”) で評価するよう指示した．

実験結果を 3.1 に示す．表 3.1 より提案手法 (PM) の平均値が他の手法 (CM1, CM2，CM3，CM4 および BL) の平均値を上回っていることが確認できる．また，提案手法 (PM) の平均値は 4.18 を示しており，“少し一致している” を示す 4 を上回っていることを確認した．また，提案手法による検索結果を図 3.7 に示す．図 3.7 より提案手法がクエリテキストに関連するシーンを検索可能であることを確認した．以上の結果より，提案手法の定性的な有効性を確認した．

3.4　テキスト特徴量空間の導入による高精度化に関する検討

本節では，テキスト特徴量空間の導入よる高精度化について説明する．提案手法の概要図を図 3.8 に示す．提案手法では，image-to-text モデル [70] および画像生成モデル [58] に基づいて，クエリのテキストおよび検索候補をそれぞれ画像およびテキストに変換する．その後，変換された画像およびテキストに加えて，従来手法を用いることで，特徴量空間 $\mathcal{L}$，$\mathcal{V}$ および $\mathcal{E}$ を介したマルチメディア情報検索を実現する．

特徴量空間 $\mathcal{L}$ における類似度 $s_n^{\mathcal{L}}$ の算出

学習済みの image-to-text モデル [70] を用いて，検索候補の画像 $\mathrm{I}_n^{\mathrm{DB}}$ ($n = 1, 2, ..., N$; N, は検索候補の総画像数) からその画像を表現する文 $\mathrm{T}_n^{\mathrm{DB}}$ を生

成する．その後，クエリテキスト T^{Q} および生成された文 $\mathrm{T}_n^{\mathrm{DB}}$ からテキスト特徴量 $\boldsymbol{f}^{\mathcal{L}_{\mathrm{Q}}} \in \mathbb{R}^{D_{\mathcal{L}}}$ および $\boldsymbol{f}_n^{\mathcal{L}_{\mathrm{DB}}} \in \mathbb{R}^{D_{\mathcal{L}}}$ を算出する．ただし，$D_{\mathcal{L}}$ はテキスト特徴量の次元数である．その後，$\boldsymbol{f}^{\mathcal{L}_{\mathrm{Q}}}$ と $\boldsymbol{f}_n^{\mathcal{L}_{\mathrm{DB}}}$ のコサイン類似度を特徴量空間 $\mathcal{L}$ における類似度 $s_n^{\mathcal{L}}$ として算出する．

特徴量空間 $\mathcal{V}$ における類似度 $s_n^{\mathcal{V}}$ の算出

学習済みの画像生成 モデル [58] を用いて，クエリテキスト T^{Q} から画像 I^{Q} を生成する．その後，生成された画像 I^{Q} および検索候補である画像 $\mathrm{I}_n^{\mathrm{DB}}$ から画像特徴量 $\boldsymbol{f}^{\mathcal{V}_{\mathrm{Q}}} \in \mathbb{R}^{D_{\mathcal{V}}}$ および $\boldsymbol{f}_n^{\mathcal{V}_{\mathrm{DB}}} \in \mathbb{R}^{D_{\mathcal{V}}}$ を算出する．ただし，$D_{\mathcal{V}}$ は画像特徴量の次元数である．その後，$\boldsymbol{f}^{\mathcal{V}_{\mathrm{Q}}}$ と $\boldsymbol{f}_n^{\mathcal{V}_{\mathrm{DB}}}$ のコサイン類似度を特徴量空間 $\mathcal{V}$ における類似度 $s_n^{\mathcal{V}}$ として算出する．

特徴量空間 $\mathcal{E}$ における類似度 $s_n^{\mathcal{E}}$ の算出

特徴量空間 $\mathcal{E}$ における類似度 $s_n^{\mathcal{E}}$ は任意の従来手法に基づいて算出されるため，本文では最も基礎的な構造について説明する．はじめに，クエリテキスト T^{Q} および検索候補である画像 $\mathrm{I}_n^{\mathrm{DB}}$ からテキスト特徴量 $\boldsymbol{f}^{\mathcal{L}_{\mathrm{Q}}}$ および画像特徴量 $\boldsymbol{f}_n^{\mathcal{V}_{\mathrm{DB}}}$ を算出する．その後，学習済みの射影 $f : \mathcal{L} \to \mathcal{E}$ を用いて，$\boldsymbol{f}^{\mathcal{L}_{\mathrm{Q}}}$ から $\boldsymbol{f}^{\mathcal{E}_{\mathrm{Q}}} \in \mathbb{R}^{D_{\mathcal{E}}}$ を算出する．同様に，学習済みの射影 $g : \mathcal{L} \to \mathcal{E}$ を用いて，$\boldsymbol{f}_n^{\mathcal{V}_{\mathrm{DB}}}$ から $\boldsymbol{f}_n^{\mathcal{E}_{\mathrm{DB}}} \in \mathbb{R}^{D_{\mathcal{E}}}$ を算出する．ただし，$D_{\mathcal{E}}$ は射影後の特徴量の次元数である．その後，$\boldsymbol{f}^{\mathcal{E}_{\mathrm{Q}}}$ と $\boldsymbol{f}_n^{\mathcal{E}_{\mathrm{DB}}}$ のコサイン類似度を特徴量空間 $\mathcal{E}$ における類似度 $s_n^{\mathcal{E}}$ として算出する．

以上より算出された類似度 $s_n^{\mathcal{L}}$，$s_n^{\mathcal{V}}$ および $s_n^{\mathcal{E}}$ から最終的な類似度 s_n を次式により算出する．

$$s_n \quad = \quad s_n^{\mathcal{L}} + s_n^{\mathcal{V}} + s_n^{\mathcal{E}} \tag{3.13}$$

その後，s_n の値が高い順に順位を決定する．

3.4.1　実験: 定量評価

本節では，提案手法の有効性を確認するための実験について説明する．本実験では，従来のマルチメディア情報検索手法の検索精度と提案手法の検索精度を比較することで提案手法の有効性を検証する．

まず，本実験のデータセットとして，Microsoft Common Objects in Context (MSCOCO) [85] データセットを用いた．具体的には，MSCOCO データセットのうち 82,783 枚の画像を学習用データ，5,000 枚をテスト用データとして用いた．ただし，各画像には画像の内容を説明する文 (以降，キャプション) が 5 つ付与されている．本実験では，従来手法 (特徴量空間 $\mathcal{E}$ における類似度 $s_n^{\mathcal{E}}$) として，文献 [35,36,38,39,87,88] で提案されているマルチメディア情報検索手法を用いた．同様に，本実験では，類似度 $s_n^{\mathcal{E}}$ に加えて，類似度 $s_n^{\mathcal{L}}$ もしくは類似度 $s_n^{\mathcal{V}}$ のどちらかのみを用いる手法 ($\mathrm{BL}_{\mathcal{L}}$ および $\mathrm{BL}_{\mathcal{V}}$) との比較を行った．また，画像特徴量およびテキスト特徴量としてそれぞれ VGG-19 の fc2 層の出力値および BERT [89] の出力値を用いた．評価指標として，クエリテキストに用いたキャプションが付与されていた画像が検索された場合を正しく検索されたと定義した上で，画像が正しく検索された順位の平均値 (以降，平均順位)，画像が正しく検索された順位の中央値 (以降，中央順位) および次式により算出される Recall@k を用いた．

$$\mathrm{Recall@}k = \frac{p_k}{J} \quad (k = 1, 2, ..., K) \tag{3.14}$$

ここで，p_k および K は k 位以内に画像が正しく検索されたクエリの個数および検索候補の画像枚数 (=5,000) を示す．また，J はクエリの総数 (=25,000) を示す．平均順位と中央順位は値が低いほど検索精度が高いことを示し，Recall@k は値が高いほど検索精度が高いことを示す．

表 3.2 に実験結果を示す．表 3.2 より，全ての評価指標において，提案手法の基となる従来手法よりも提案手法の方が高い検索精度を示すことが確認できる．特に，提案手法の基となる従来手法と比べて，平均順位の値が低いことから，特徴量空間 $\mathcal{L}$，$\mathcal{V}$ における比較を従来手法に導入した場合，クエリの文に対する

表 3.2: 画像生成モデルに基づくマルチメディア情報検索手法および従来手法の検索精度.

	R@1	R@10	平均順位	中央順位
BL$_{\mathcal{L}}$	0.11	0.27	154.36	29
BL$_{\mathcal{V}}$	0.12	0.28	165.17	28
BL$_{\mathcal{L}}$+BL$_{\mathcal{V}}$	**0.16**	**0.39**	**109.49**	**27**
Kiros '14 [35]	0.12	0.34	166.60	30
Kiros '14+BL$_{\mathcal{L}}$	0.17	0.41	126.46	18
Kiros '14+BL$_{\mathcal{V}}$	0.14	0.34	102.52	25
Kiros '14+BL$_{\mathcal{L}}$+BL$_{\mathcal{V}}$ (PM)	**0.22**	**0.48**	**77.56**	**16**
Vendrov '16 [36]	0.13	0.34	166.66	30
Vendrov '16+BL$_{\mathcal{L}}$	0.17	0.35	155.33	24
Vendrov '16+BL$_{\mathcal{V}}$	0.17	0.32	110.09	20
Vendrov '16+BL$_{\mathcal{L}}$+BL$_{\mathcal{V}}$ (PM)	**0.23**	**0.50**	**75.44**	**13**
Faghri '17 [38]	0.15	0.38	151.95	25
Faghri '17+BL$_{\mathcal{L}}$	0.16	0.36	146.77	21
Faghri '17+BL$_{\mathcal{V}}$	0.15	0.36	140.95	19
Faghri '17+BL$_{\mathcal{L}}$+BL$_{\mathcal{V}}$ (PM)	**0.19**	**0.45**	**79.44**	**15**
Liu '17 [87]	0.16	0.40	151.95	25
Liu '17+BL$_{\mathcal{L}}$	0.18	0.40	148.55	24
Liu '17+BL$_{\mathcal{V}}$	0.17	0.40	100.04	25
Liu '17+BL$_{\mathcal{L}}$+BL$_{\mathcal{V}}$ (PM)	**0.23**	**0.48**	**69.66**	**13**
Zhang '18 [39]	0.20	0.44	142.24	23
Zhang '18+BL$_{\mathcal{L}}$	0.21	0.44	122.22	24
Zhang '18+BL$_{\mathcal{V}}$	0.18	0.51	109.98	**12**
Zhang '18+BL$_{\mathcal{L}}$+BL$_{\mathcal{V}}$ (PM)	**0.25**	**0.51**	**87.33**	**12**
Song '19 [88]	0.25	0.51	112.02	10
Song '19+BL$_{\mathcal{L}}$	**0.29**	0.52	70.55	10
Song '19+BL$_{\mathcal{V}}$	0.21	0.49	55.40	13
Song '19+BL$_{\mathcal{L}}$+BL$_{\mathcal{V}}$ (PM)	**0.29**	**0.56**	**49.38**	**9**

頑健性が向上すると考えられる．同様に，特徴量空間 $\mathcal{L}$ および $\mathcal{V}$ の両方を用いる提案手法が特徴量空間 $\mathcal{L}$ もしくは $\mathcal{V}$ を用いる手法よりも検索精度が高いことから，特徴量空間 $\mathcal{L}$ および $\mathcal{V}$ の両方が検索精度向上に寄与していることを確認した．

3.4.2　実験: テキスト特徴量空間 $\mathcal{L}$ が着目する情報に関する検証

本節では，テキスト特徴量空間 $\mathcal{L}$ が着目している情報に関して検証を行う．従来より，テキスト特徴量空間 $\mathcal{L}$ は単語の関係性を考慮可能であることが報告されている [89]．以上の報告に基づいて，本実験では提案手法により構築されたテキスト特徴量空間 $\mathcal{L}$ が従来のマルチメディア情報検索手法により構築された特徴量空間 $\mathcal{E}$ よりも単語の関係性に着目可能であることを確認する．

本実験では，はじめに，MSCOCO データセットのテスト用データに含まれるクエリテキストから単語の関係性を表現する単語を削除する．その後，通常のクエリテキストを用いた際の検索精度と単語を削除したクエリテキストを用いた際の検索精度を比較する．本実験では，従来手法 (特徴量空間 $\mathcal{E}$ における類似度 $s_n^{\mathcal{E}}$) として，文献 [35, 36, 38, 39, 87, 88] で提案されているマルチメディア情報検索手法を用いた．同様に，本実験では，類似度 $s_n^{\mathcal{L}}$ もしくは類似度 $s_n^{\mathcal{V}}$ のどちらかのみを用いる手法 ($\mathrm{BL}_{\mathcal{L}}$ および $\mathrm{BL}_{\mathcal{V}}$) の検索精度を確認した．また，画像特徴量およびテキスト特徴量としてそれぞれ VGG-19 の fc2 層の出力値および BERT [89] の出力値を用いた．評価指標として，クエリテキストに用いたキャプションが付与されていた画像が検索された場合を正しく検索されたと定義した上で，画像が正しく検索された順位の平均値 (以降，平均順位)，画像が正しく検索された順位の中央値 (以降，中央順位) を用いた．図 3.9 に実験結果を示す．図 3.9 より，単語を削除したクエリテキストを用いた場合，テキスト特徴量空間 $\mathcal{L}$ を用いる手法 ($\mathrm{BL}_{\mathcal{L}}$) の検索精度が低下していることが確認できる．以上より，テキスト特徴量空間 $\mathcal{L}$ は従来のマルチメディア情報検索により構築される特徴量空間 $\mathcal{E}$ と比較して，単語の関係性に着目していると考えられる．

図 3.7: 生成画像を用いた階層的シーン検索による検索結果.

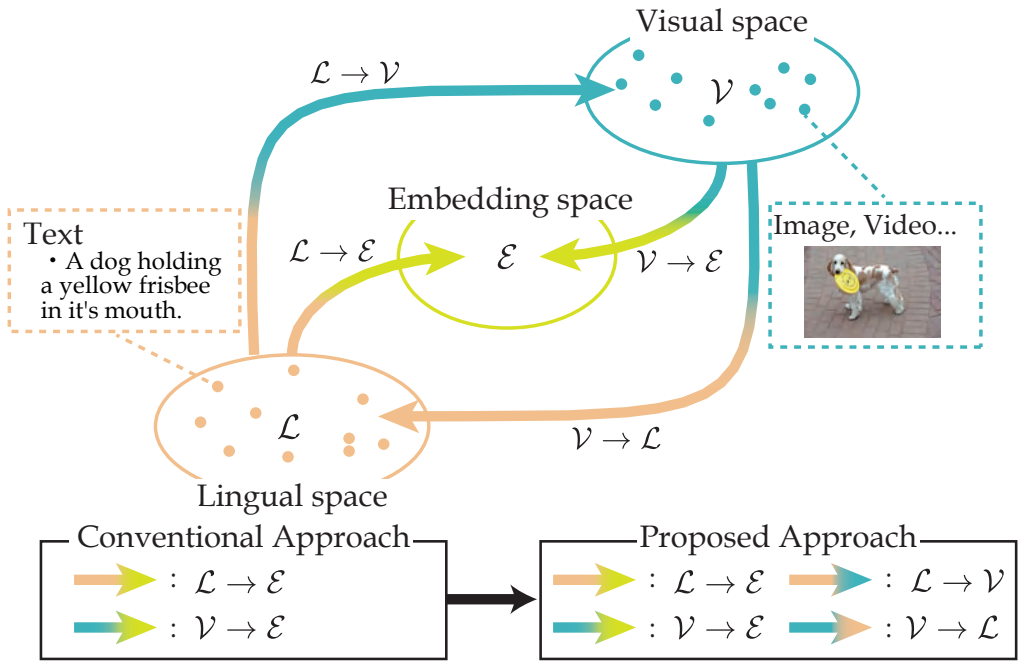


図 3.8: テキスト特徴量空間を導入したマルチメディア情報検索の概要図.

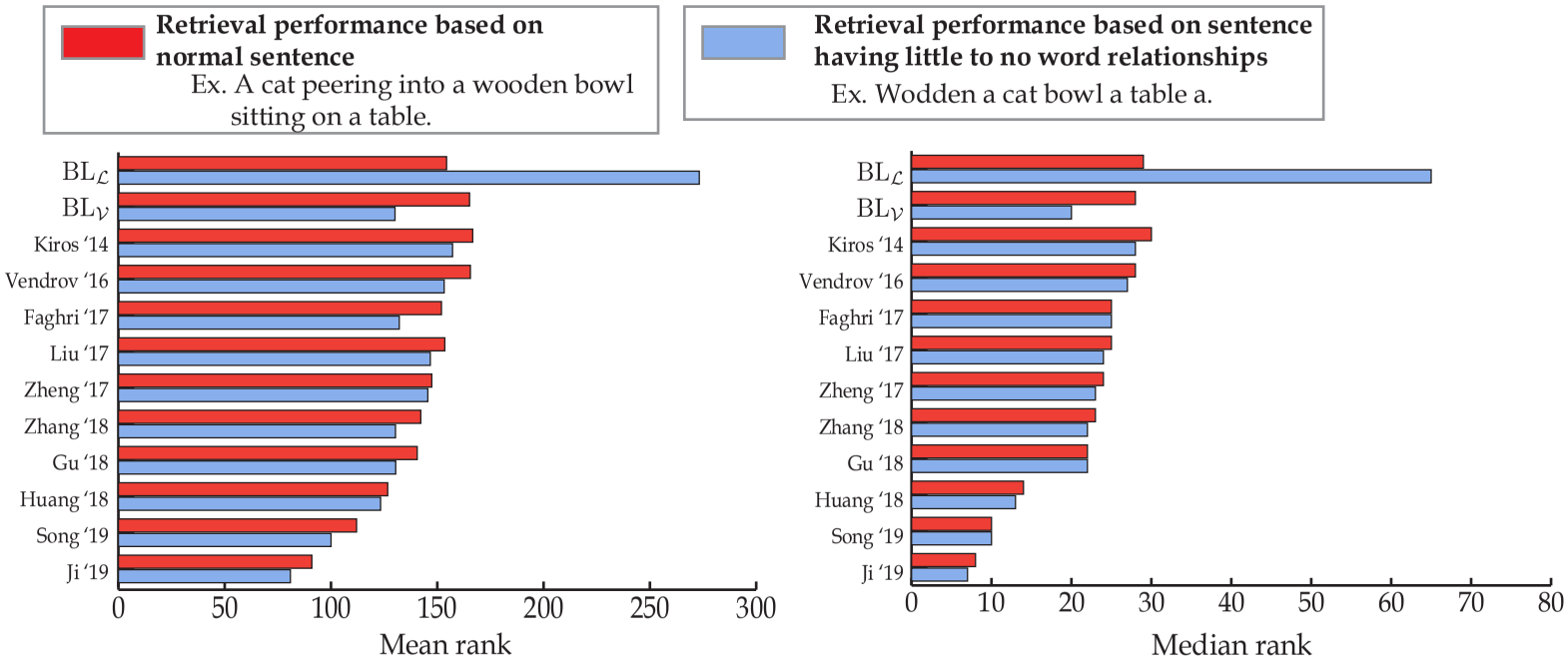


図 3.9: テキスト特徴量空間 $\mathcal{L}$ の有効性を検証するための評価に関する実験結果.

3.4.3　実験: 画像特徴量空間 $\mathcal{V}$ が着目する情報に関する検証

本節では，画像特徴量空間 $\mathcal{V}$ が着目している情報に関して検証を行う．従来より，画像特徴量空間 $\mathcal{V}$ は画像のテクスチャを考慮可能であることが報告されている [90]．以上の報告に基づいて，本実験では提案手法により構築された画像特徴量空間 $\mathcal{L}$ が従来のマルチメディア情報検索手法により構築された特徴量空間 $\mathcal{E}$ よりもテクスチャの情報に着目可能であることを確認する．

本実験では，はじめに，MSCOCO データセットのテスト用データに含まれる検索候補からテクスチャを表現する情報を削除する．その後，通常の検索候補を用いた際の検索精度とテクスチャの情報を削除した検索候補を用いた際の検索精度を比較する．本実験では，従来手法 (特徴量空間 $\mathcal{E}$ における類似度 $s_n^{\mathcal{E}}$) として，文献 [35, 36, 38, 39, 87, 88] で提案されているマルチメディア情報検索手法を用いた．同様に，本実験では，類似度 $s_n^{\mathcal{V}}$ のみを用いる手法 ($\mathrm{BL}_{\mathcal{V}}$) の検索精度を確認した．また，画像特徴量およびテキスト特徴量としてそれぞれ VGG-19 の fc2 層の出力値および BERT [89] の出力値を用いた．評価指標として，クエリテキストに用いたキャプションが付与されていた画像が検索された場合を正しく検索されたと定義した上で，画像が正しく検索された順位の平均値 (以降，平均順位)，画像が正しく検索された順位の中央値 (以降，中央順位) を用いた．図 3.10 に実験結果を示す．図 3.10 より，検索候補のテクスチャの情報を削除した場合，画像特徴量空間 $\mathcal{V}$ を用いる手法 ($\mathrm{BL}_{\mathcal{V}}$) の検索精度が大幅に低下していることが確認できる．以上より，画像特徴量空間 $\mathcal{V}$ は従来のマルチメディア情報検索により構築される特徴量空間 $\mathcal{E}$ と比較して，テクスチャの情報に着目していると考えられる．

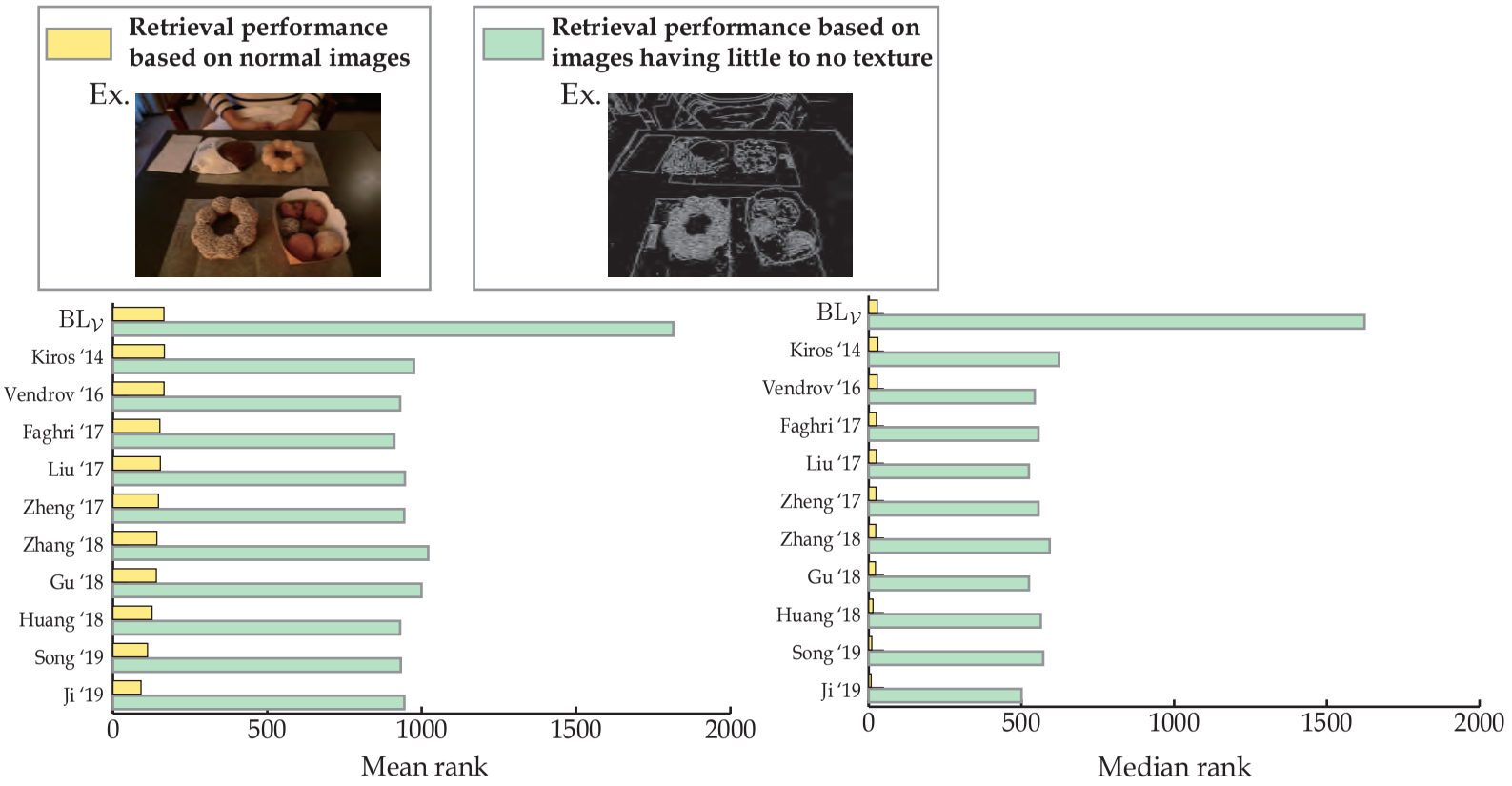


図 3.10: 画像特徴量空間 $\mathcal{V}$ の有効性を検証するための評価に関する実験結果.

3.5　まとめ

本章では，画像生成モデルを用いてクエリテキストから画像を生成し，その生成画像をマルチメディア情報検索に応用することにより，画像生成モデルに基づくマルチメディア情報検索手法が目的のマルチメディアコンテンツを正確に検索可能であることを確認した.

第4章　画像生成モデルに基づく対話型マルチメディア情報検索

4.1　はじめに

3 章では，画像生成モデルに基づくマルチメディア情報検索手法を提案し，その検索精度を検証した．また，実験により，画像生成モデルに基づくマルチメディア情報検索手法が目的のマルチメディアコンテンツを検索するために有効であることを検証した．上記の検証結果を受け，本章では，画像生成モデルに基づいて「クエリテキストに対する検索システムの解釈」をユーザに理解可能な形式で共有することが可能な対話型マルチメディア情報検索手法を構築し，その有効性を検証する．提案手法において，ユーザは初期検索結果および生成画像を閲覧した後に，クエリテキストの各単語の重要度を調整することで，再度目的のマルチメディアコンテンツを検索する．本提案により，初期の検索で目的のマルチメディアコンテンツを絞り込めない場合においても,「クエリテキストに対する検索システムの解釈」を視覚的な生成画像として閲覧しながらクエリテキストを修正し，目的のマルチメディアコンテンツを再度絞り込むことが可能である．本章では，映画のシーンにより構成されている MP2-Movie Description (MP2-MD) データセット [91] を用いた実験を行い，提案手法の有効性を確認する．

4.2　画像生成モデルに基づく対話型マルチメディア情報検索手法

本章では，提案手法である画像生成モデルに基づく対話型マルチメディア情報検索手法について説明する．提案手法の初期検索および再検索の概要図をそれぞれ図 4.1 および図 4.2 に示す．提案手法は 3 つの段階で構成されている．まず，第 1 段階では，文献 [68] に基づく画像生成モデルによりクエリテキストを表現する画像を生成する．次に第 2 段階では，上述の生成画像を用いて目的のシーンを検索することで初期検索結果を算出する．ここで，ユーザが以上により算出された初期検索結果に満足しない場合，第 3 段階として再検索を行う．第 3 段階では，ユーザは初期検索結果に反映されていなかったクエリテキストの単語を指定する．以上によりユーザから提供された情報に基づいて，提案手法では画像の再生成を行い，再生成された画像を用いて再検索を行う．

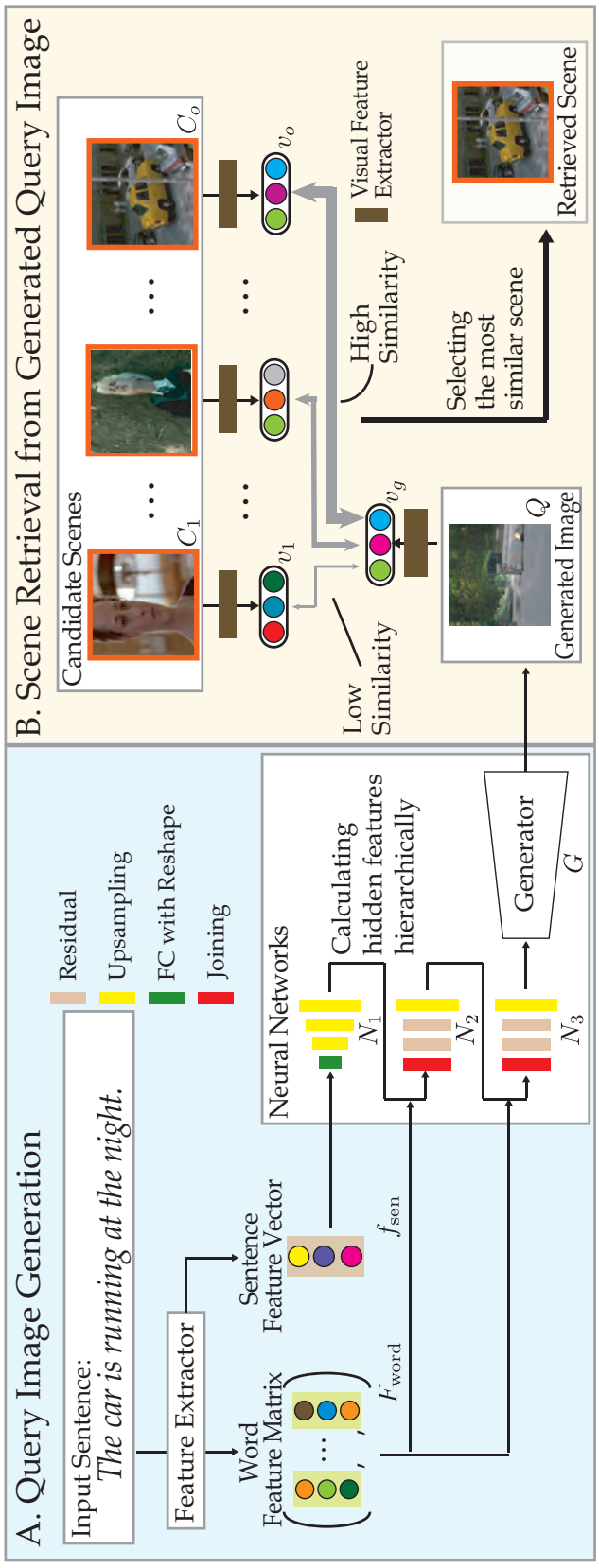


図 4.1: 画像生成モデルに基づく対話型マルチメディア情報検索手法における初期検索の概要図.

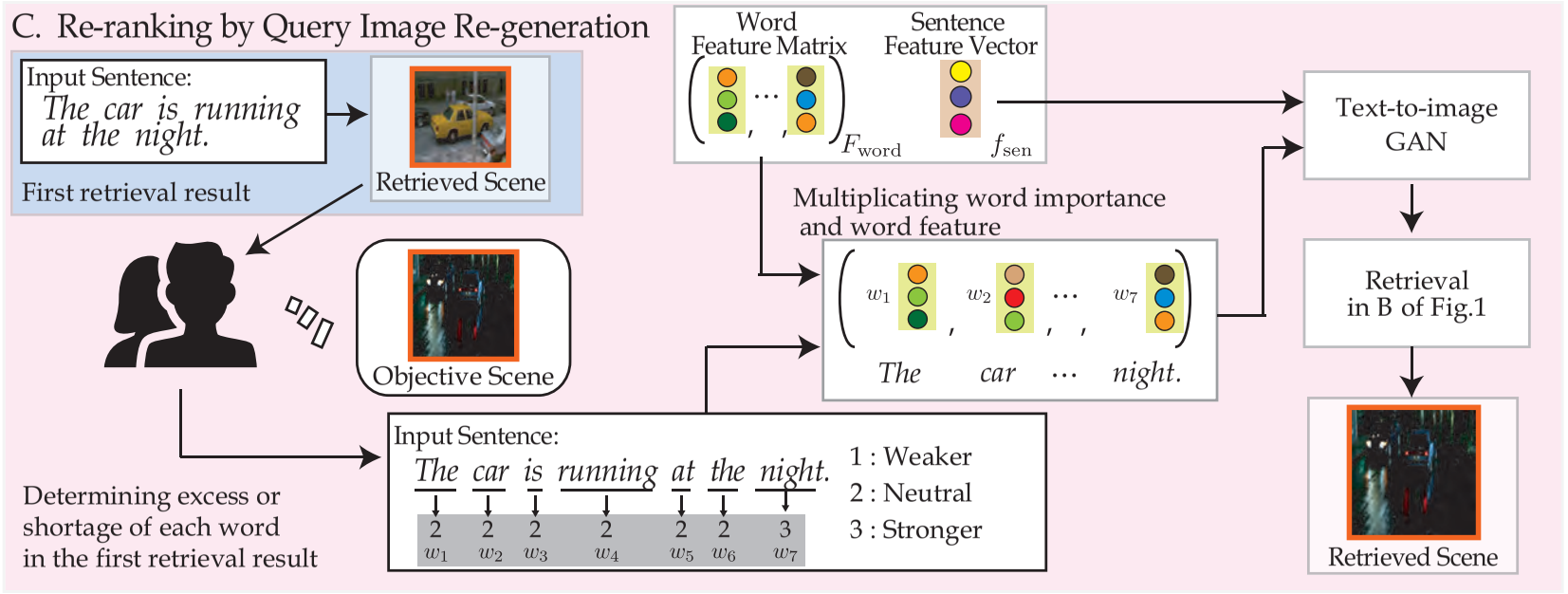


図 4.2: 画像生成モデルに基づく対話型マルチメディア情報検索手法における再検索の概要図.

4.2.1　画像生成モデルによる画像生成

本節では，文献 [68] に基づく画像生成モデルによる，画像生成について説明する．提案手法で構築する文献 [68] に基づく画像生成モデルは，3 つのニューラルネットワーク $F_r(r \in \{l, m, h\})$ とその出力である特徴ベクトル $\boldsymbol{s}_r(r \in \{l, m, h\})$ および 3 つの画像生成器 $G_r(r \in \{l, m, h\})$ から構成されている．また，クエリテキストより算出された文特徴量 $\boldsymbol{e}_{\text{sen}}$ およびクエリテキストに含まれる各単語より算出された特徴量 $\boldsymbol{e}^j_{\text{word}}$ $(j = 1, 2, ..., J; J$ はクエリテキストに含まれる単語数) により構成される単語特徴量行列 $\boldsymbol{E}_{\text{word}} = (\boldsymbol{e}^1_{\text{word}}{}^{\mathsf{T}}, \boldsymbol{e}^2_{\text{word}}{}^{\mathsf{T}}, ..., \boldsymbol{e}^J_{\text{word}}{}^{\mathsf{T}})$ を画像生成モデルの入力として用いる．

まず，文特徴量 $\boldsymbol{e}_{\text{sen}}$ およびガウス雑音 $\boldsymbol{z}$ からニューラルネットワーク F_l を用いてベクトル $\boldsymbol{s}_l$ を算出する．

$$\boldsymbol{s}_l = F_l(\boldsymbol{z}, F^{\text{ca}}(\boldsymbol{e}_{\text{sen}})) \tag{4.1}$$

ただし，F^{ca} は文献 [77] により提案された，学習を安定化させるための関数である．ここで，$\boldsymbol{s}_l$ は文特徴量のみから算出されるため文全体の情報のみを含む特徴ベクトルとなる．

次に，以上で算出された $\boldsymbol{s}_l$ および単語特徴量行列 $\boldsymbol{E}_{\text{word}}$ から F_m を用いて $\boldsymbol{s}_m$ を算出する．また，$\boldsymbol{s}_h$ についても同様に $\boldsymbol{s}_m$ および $\boldsymbol{E}_{\text{word}}$ を用いて算出する．

$$\boldsymbol{s}_m = F_m(\boldsymbol{s}_l, F^{\text{attn}}_m(\boldsymbol{E}_{\text{word}}, \boldsymbol{s}_l)) \tag{4.2}$$

$$\boldsymbol{s}_h = F_h(\boldsymbol{s}_m, F^{\text{attn}}_h(\boldsymbol{E}_{\text{word}}, \boldsymbol{s}_m)) \tag{4.3}$$

ここで，$F^{\text{attn}}_r(r \in \{m, h\})$ は $\boldsymbol{s}_r(r \in \{l, m\})$ に $\boldsymbol{E}_{\text{word}}$ を組み込むニューラルネットワークである．$\boldsymbol{s}_r(r \in \{m, h\})$ は一つ前のニューラルネットワークの出力および単語特徴量行列 $\boldsymbol{E}_{\text{word}}$ を用いて算出されている．そのため，$\boldsymbol{s}_m$，$\boldsymbol{s}_h$ と処理が進むごとに，文全体の特徴のみでなく単語ごとの特徴に着目した情報を含む特徴ベクトルが算出される．

最後に，$\boldsymbol{s}_h$ から以下の式を用いて画像 Q_h を生成する．

$$Q_h \quad = \quad G_h(\boldsymbol{s}_h) \tag{4.4}$$

以上により生成された Q_h は $\boldsymbol{s}_h$ から生成されているため文全体および単語それぞれの特徴にも着目した情報を持つ画像となる．提案手法では Q_h を用いることで目的のシーンを検索する．

4.2.2　生成画像を用いたシーン検索

本節では，生成画像を用いたシーン検索について説明する．はじめに，検索対象の映像のフレームを f_n ($n = 1, 2, ..., N; N$ は映像のフレーム数) とする． 続いて，前節で生成した画像 Q_h から画像特徴量 $\boldsymbol{v} \in \mathbb{R}^D$ を，f_n から画像特徴量 $\boldsymbol{v}_n \in \mathbb{R}^D$ を算出する．ここで，D は画像特徴量の次元数を表す．ただし，提案手法では ImageNet [84] により学習済みの Inception-v3 [81] の第 3 プーリング層の出力を画像特徴量として用いる． 次に，Q_h および f_n の類似度として，$\boldsymbol{v}$ および $\boldsymbol{v}_n$ のコサイン類似度 w_n を算出する．

$$w_n = \frac{\boldsymbol{v} \cdot \boldsymbol{v}_n}{|\boldsymbol{v}||\boldsymbol{v}_n|} \quad (n = 1, 2, ..., N) \tag{4.5}$$

その後，類似度 w_n からシーンごとの順位を決定する．

4.2.3　生成画像の再生成に基づく対話型マルチメディア情報検索

本節では，生成画像の再生成に基づく対話型マルチメディア情報検索について説明する．本節では，はじめに，初期検索を閲覧したユーザからフィードバックを受け取る．具体的に，ユーザは閲覧した検索結果を基にクエリテキストの各単語の重要度 q_j ($0 \leq q_j \leq 2$) を決定する．検索結果に含まれない場合には 1 よりも大きい値を設定し，検索結果に方に含まれる場合には 1 よりも小さい値を設定す

る．提案手法では，以上のフィードバックの後に，画像の再生成を行う．具体的には，$\boldsymbol{E}_{\text{word}} = (q_1 \boldsymbol{e}^1_{\text{word}}{}^{\mathsf{T}}, q_2 \boldsymbol{e}^2_{\text{word}}{}^{\mathsf{T}}, ..., q_3 \boldsymbol{e}^J_{\text{word}}{}^{\mathsf{T}})$ とすることで画像を再生成する．以上により再生成された画像を基に **4.2.2** 節の検索を行うことで，再度検索順位を決定する．

4.2.4　実験: 実験概要

提案手法の有効性を確認するための被験者実験について説明しその結果を示す.

4.2.5　実験: データセット

本節では実験で用いるデータセットについて説明する．本実験では，画像生成モデルの学習用データセットとして，33 万枚の画像で構成された Microsoft Common Objects in Context (MSCOCO) [85] データセットを用いた．ただし，MSCOCO データセットにおいては，1 画像あたり 5 つのテキストが付与されている．また，評価用のデータセットとして 94 本の映画より抜粋された 68,000 シーンにより構成される MP2-MD データセットを用いた．MP2-MD データセットに関しても同様に，それぞれのシーンに対して 1 つの説明文が付与されている

4.2.6　実験: 被験者実験概要

本節では被験者実験に関して説明する．はじめに，MP2-MD データセットよりそれぞれ 5 本の映画“Bad Santa”,“As got as it gets”,“Halloween”,“Rendezvous mit Joe Black ”および“ Harry Potter and the Prisoner of azkaban ”の 5 つの映画を選択し，それらの映画の中からランダムに 300 シーン抽出した．その後，15 名の実験参加者に異なる 20 のシーンを閲覧していただき，それぞれのシーンを表現する文を作成していただいた．その後，実験参加者に作成していただいた文をクエリとして提案手法により検索を行った．続いて，実験参加者に検索結果を閲覧していただき，クエリテキストに含まれる各単語に “1 Weaker”, “2 Neutral” および “3 Stronger” のいづれかのラベルを付与していただいた．そして，以上によ

り付与していただいたラベルを基に各単語の重要度を決定し (“1 Weaker”: 0.5, “2 Neutral” : 1.0, “3 Stronger” : 1.5)，再検索を行った．最後に，以上の手順により算出された検索結果に基づいて，定量的な評価および定性的な評価を行った．

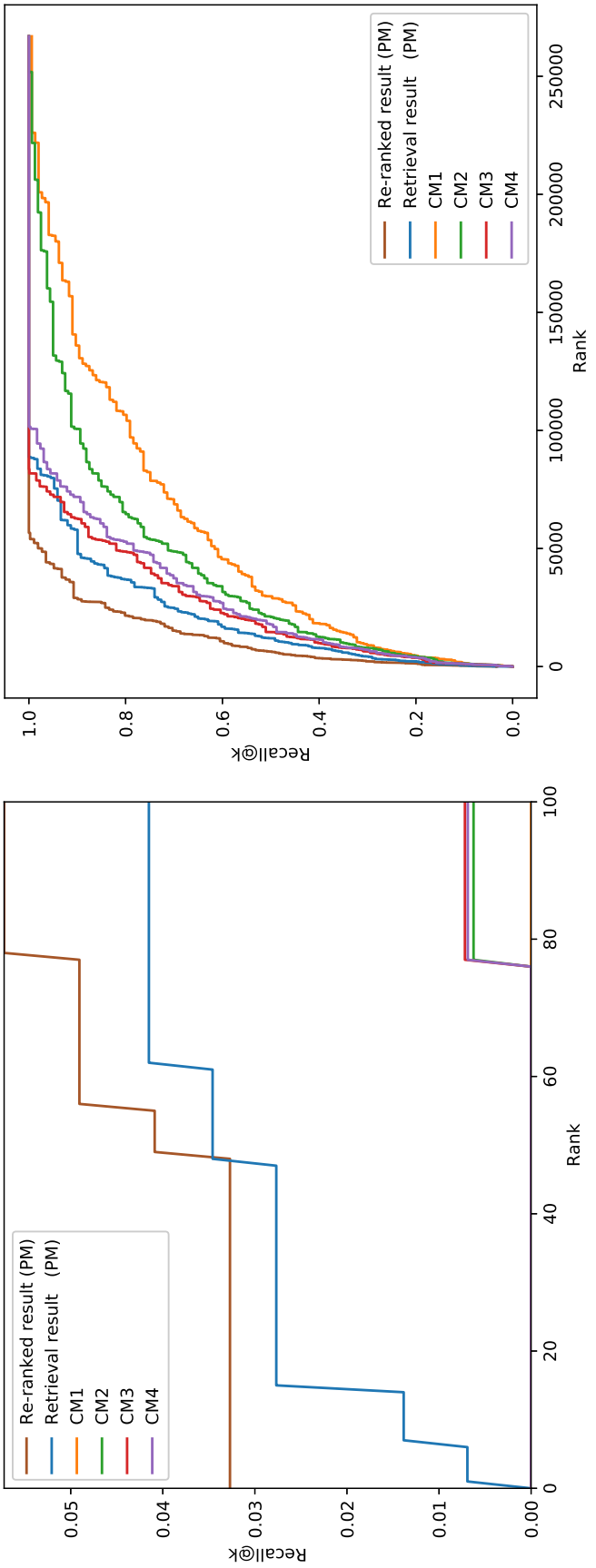
4.2.7　実験: 定量評価

まずはじめに，定量的な評価について説明する．検索精度を定量的に評価するための評価指標として，本実験では，k 位以上の再現率を測定することが可能な Recall@k を算出した．

$$\text{Recall@}k = \frac{t_k}{M} \quad (k = 1, 2, ..., N) \tag{4.6}$$

ここで t_k は正解が k 位以上に存在するシーンの個数を，M は検索対象の映像の総シーン数を示す．この評価により単一映像からの検索を仮定した場合の提案手法の有効性を確認する．ただし，順位はフレームごとに算出し，入力の説明文が付与されていたシーンと検索されたフレームが一致した場合を正解とした．また，比較手法として Convolutional Neural Network および Long-Short Term Memory を用いて画像と文より算出される特徴量をそれぞれ同一な空間に射影し比較する手法 (CM1) [35]，CM1 に加えて単語ごとの順序構造を考慮した手法 (CM2) [36]，文を視覚的特徴量の空間上に射影し比較する手法 (CM3) [37]，CM1 の学習の仕組みを検索用に改良した最新手法 (CM4) [38]，Q_h のみを用いて検索する手法 (BL) を用いた．CM1 と比較することで従来の基準となる手法との比較を行い，CM2 と比較することで提案手法が文構造を考慮しているかを確認した．また，CM3 と比較することで画像を生成することの有効性を確認し，CM4 と比較することで最新の手法と比較した時の精度の差を確認した．ただし，全ての比較手法は COCO データセットで学習されている．

図 4.3 に実験結果を示す．図 4.3 において CM1，CM2，CM3，CM4 および提案手法 (PM) の初期検索の精度を PM の再検索の精度が上回っていることが確認できる．以上により，提案手法の有効性を定量的に確認した．



(a) Recall@k (k= 1,2,⋯,100)　(b) Recall@k (k= 1,2,⋯,P;P= 267,320)

図 4.3: 画像生成モデルに基づく対話型マルチメディア情報検索手法に対する定量的な評価のための実験結果．横軸は順位を示し，縦軸は Recall@k を示す．

4.2.8　実験: 定性評価

続いて，定性的な評価について説明する．本実験では 4.2.8 節での各検索結果に関して，実験参加者に 5 段階 (1:“ 一致していない”，2:“ 少し一致している”，3:“どちらとも言えない”, 4:“ 少し一致している”，5:“一致している”) で評価を行っていただいた．

実験結果を 4.1 に示す．表 4.1 より再検索結果が初期検索結果の評価を上回っていることが確認できる．また，提案手法による検索結果を図 4.4 に示す．図 4.4 より提案手法がクエリテキストに関連するシーンを再検索可能であることを確認した．以上の結果より，提案手法の定性的な有効性を確認した．

表 4.1: 画像生成モデルに基づく対話型マルチメディア情報検索手法の被験者実験の結果. 被験者は検索結果 1:“一致していない”から 5:“一致している”までの 5 段階で評価する.

	被験者 1	被験者 2	被験者 3	被験者 4	被験者 5	被験者 6	被験者 7	被験者 8	被験者 9	被験者 10
初期検索結果	2.20	2.95	2.00	2.83	2.42	2.32	3.30	2.00	2.85	3.40
再検索結果	**3.10**	**3.53**	**2.20**	**3.63**	**3.41**	**3.25**	**3.78**	**3.25**	**3.43**	**4.10**

	被験者 11	被験者 12	被験者 13	被験者 14	被験者 15	平均
初期検索結果	3.12	2.73	2.54	3.11	2.10	2.66
再検索結果	**3.79**	**3.17**	**3.32**	**3.62**	**3.12**	**3.38**

4.3　まとめ

本章では，画像生成モデルに基づく対話型マルチメディア情報検索手法を提案した．本提案により，初期の検索で目的のマルチメディアコンテンツを絞り込めない場合にでも,「クエリテキストに対する検索システムの解釈」を視覚的な生成画像として閲覧しながらクエリテキストを修正し，目的のマルチメディアコンテンツを再度絞り込むことが可能である．また，実験により提案手法の定量的および定性的な有効性を確認した.

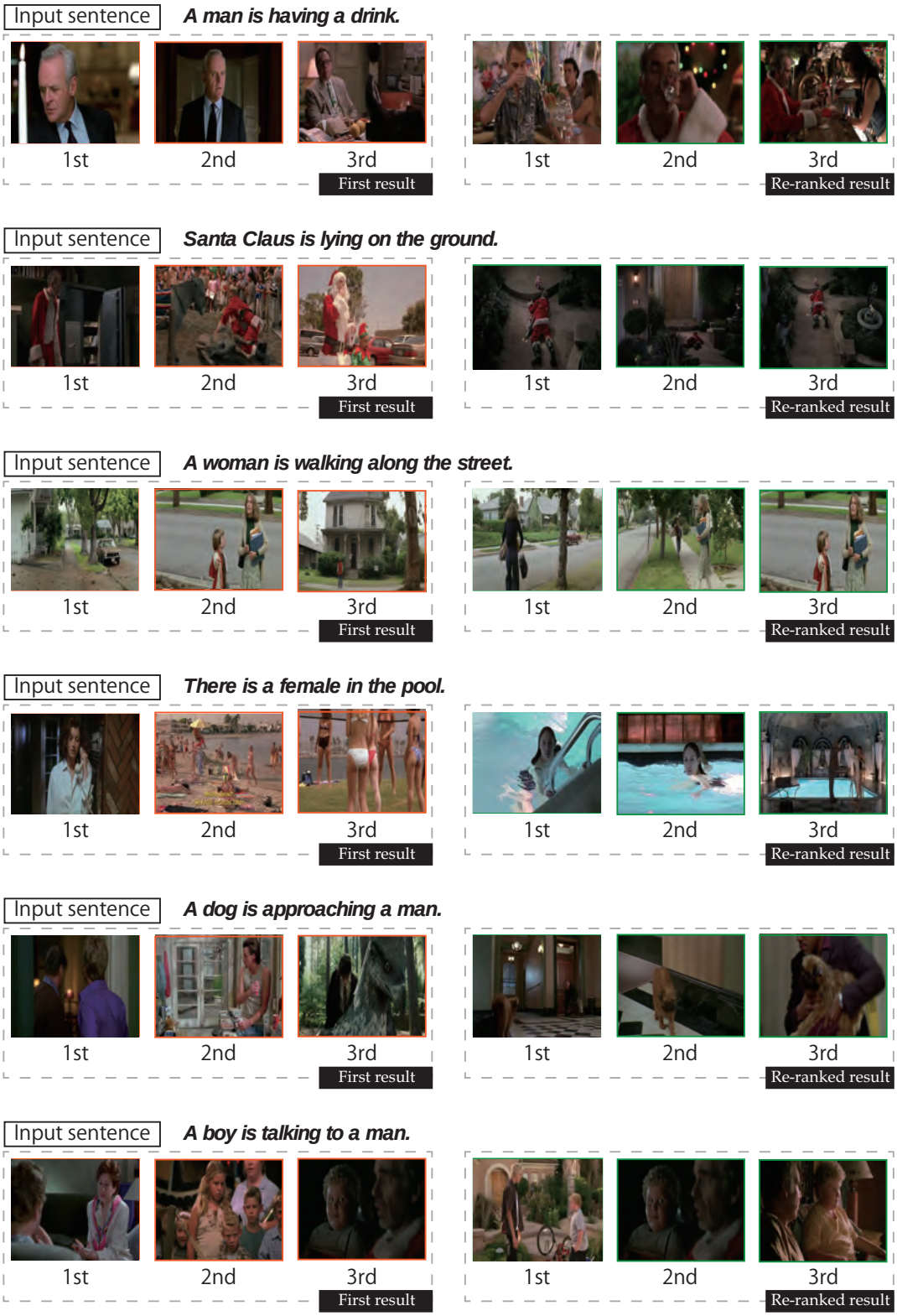


図 4.4: 画像生成モデルに基づく対話型マルチメディア情報検索手法による検索結果.

第5章　画像生成モデルに基づくドメイン適応可能なマルチメディア情報検索

5.1　はじめに

本章では絵画やアニメーション等の一般画像とは異なるドメインに対して有効な，画像生成モデルに基づくマルチメディア情報検索手法を提案する．3 章および 4 章では，画像生成モデルに基づく対話型マルチメディア情報検索が目的のマルチメディアコンテンツを検索するために有効であることを検証した．一方で，3 章および 4 章における検証では，一般的な日常風景を撮像した画像 (以降，一般画像) のみを対象としたデータセットを用いて検索手法の有効性を検証しており，絵画やアニメーション等のドメインが異なる画像に対する検索精度は検証していない．そのため，本章では絵画やアニメーション等の一般画像とは異なるドメインに対する検索精度を検証する．

3 章および 4 章で提案した検索手法を用いて絵画やアニメーション等のドメインのマルチメディアコンテンツを検索する際に課題となるのが，学習データと検索候補のドメインシフトの問題である．機械学習の分野において，学習に利用するデータのドメインとテスト時のドメインが大きく異なる際に精度が低下することが知られており，このような問題はドメインシフト問題と呼ばれている．このような，ドメインシフトの問題の解決策として，目的のドメインにおいてデータセットを構築する方法が知られているが，マルチメディア情報検索手法においては，画像およびテキストがペアとなったデータセットが必要となるため，データセットを構築することによる解決は多大な労力が必要となる．そのため，目的の

ドメインにおいてデータセットを構築することなく，目的のドメインにおいて高精度に目的のマルチメディアコンテンツを検索することが可能な検索手法が必要である．

そこで，本章では，画像生成モデルに基づくドメイン適応可能なマルチメディア情報検索手法を提案する．具体的に提案手法では，あるドメインのデータセット [85] を用いて学習された画像生成モデルを用いて，クエリテキストから画像を生成する．その後，ドメイン適応手法の一つである Style transfer モデルを用いて，生成された画像を検索候補のドメインに変換する．同様に，検索候補に関しても，生成された画像のドメインに変換する．続いて，生成画像と変換された検索候補および変換された生成画像および検索候補の類似度を算出する．最後に，算出された類似度を統合し検索結果を算出する．以上のように Style transfer モデルを用いて，画像のドメインを変換することで，学習データと検索候補のドメインが大きく異なる場合にでも，高精度に目的のマルチメディアコンテンツを検索することが可能なマルチメディア情報検索を実現する．

5.2　画像生成モデルに基づくドメイン適応可能なマルチメディア情報検索手法

本章では，**5.2.1** 節においてクエリ文からの画像生成について説明する．また，**5.2.2** 節では **5.2.2** 節で生成された画像を用いた画像検索について説明する．提案手法の概要を図 5.1 に示す．

5.2.1　画像生成モデルを用いたクエリ画像の生成

本節では，クエリ画像の生成に関して画像生成モデルの一種である Attentional Generative Adversarial Netowrk (AttnGAN) に基づいて説明を行う．AttnGAN は，3 つのニューラルネットワーク F_r ($r \in \{l, m, h\}$) とその出力である特徴ベクトル $\boldsymbol{s}_r$ ($r \in \{l, m, h\}$) および画像生成ネットワーク G_h から構成されている．以上の AttnGAN に対してクエリ文より算出された文特徴量 $\boldsymbol{e}_{\text{sen}}$ およびクエリ文に含まれ

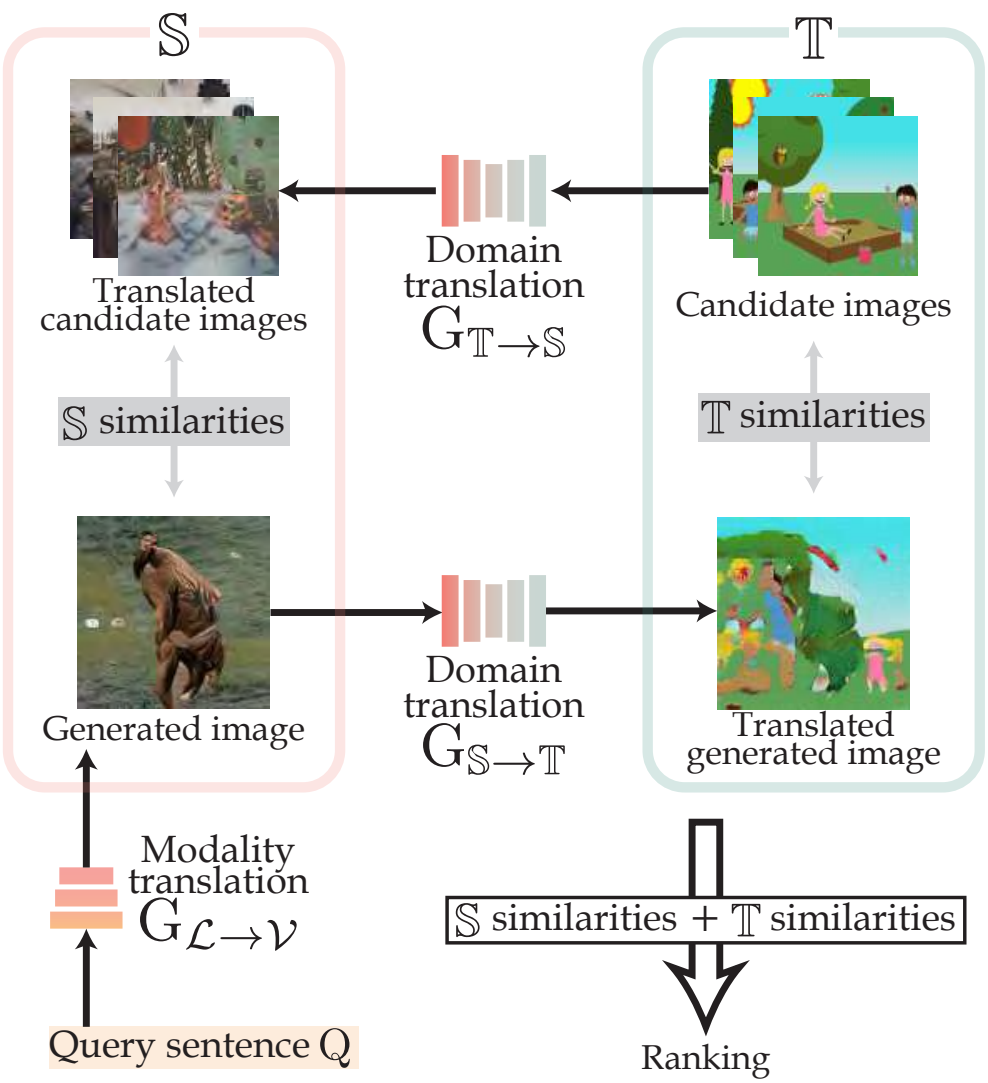


図 5.1: 画像生成モデルに基づくドメイン適応可能なマルチメディア情報検索手法の概要図.

る各単語より算出された特徴量により構成される単語特徴量行列 $\boldsymbol{E}_{\text{word}}$ を入力することで，クエリ文を表現する画像を生成する．

まず，文特徴量 $\boldsymbol{e}_{\text{sen}}$ およびガウス雑音 $\boldsymbol{z}$ からニューラルネットワーク F_l を用いてベクトル $\boldsymbol{s}_l$ を算出する．

$$\boldsymbol{s}_l = F_l(\boldsymbol{z}, F^{\text{ca}}(\boldsymbol{e}_{\text{sen}})) \tag{5.1}$$

ただし，F^{ca} は文献 [77] により提案された，学習を安定化させるための関数である．ここで，$\boldsymbol{s}_l$ は文特徴量のみから算出されるため文構造の情報に着目した特徴ベクトルとなる．

次に，以上で算出された $\boldsymbol{s}_l$ および単語特徴量行列 $\boldsymbol{E}_{\text{word}}$ から F_m を用いて $\boldsymbol{s}_m$ を算出する．また，$\boldsymbol{s}_h$ についても同様に $\boldsymbol{s}_m$ および $\boldsymbol{E}_{\text{word}}$ を用いて算出する．

$$\boldsymbol{s}_m = F_m(\boldsymbol{s}_l, F_m^{\text{attn}}(\boldsymbol{E}_{\text{word}}, \boldsymbol{s}_l)) \tag{5.2}$$

$$\boldsymbol{s}_h = F_h(\boldsymbol{s}_m, F_h^{\text{attn}}(\boldsymbol{E}_{\text{word}}, \boldsymbol{s}_m)) \tag{5.3}$$

ここで，$F_r^{\text{attn}}\ (r \in \{m, h\})$ は $\boldsymbol{s}_r\ (r \in \{l, m\})$ に $\boldsymbol{E}_{\text{word}}$ を組み込むニューラルネットワークである．$\boldsymbol{s}_r\ (r \in \{m, h\})$ は一つ前のニューラルネットワークの出力および単語特徴量行列 $\boldsymbol{E}_{\text{word}}$ を用いて算出される．そのため，$\boldsymbol{s}_m$，$\boldsymbol{s}_h$ と処理が進むごとに，文構造の情報のみでなく単語ごとの詳細な情報を含む特徴ベクトルが算出される．

最後に，$\boldsymbol{s}_h$ から画像生成ネットワーク G_h を用いて画像 Q_h を生成する．

$$Q_h = G_h(\boldsymbol{s}_h) \tag{5.4}$$

以上により生成された画像 Q_h は $\boldsymbol{s}_h$ から生成されているため文構造および単語ごとの詳細な情報に着目した画像となる．

5.2.2　ドメイン変換に基づくクロスモーダル検索

本節ではドメイン変換に基づくクロスモーダル検索について説明する．はじめに，第 **5.2.1** 節において生成された画像 Q_h および検索候補の画像 I_n ($n = 1, 2, ..., N$; N は検索候補の画像枚数) をそれぞれ検索候補のドメインおよび生成画像のドメインに変換し，変換された生成画像および変換された検索候補をそれぞれ $\hat{Q}_h$ および $\hat{I}_n$ と定義する．ここで，ドメインの変換には生成された画像 Q_h および検索候補の画像 I^n に基づいて学習済みの Style transfer モデルを用いる．各画像を変換した後に以下の式に基づいて最終的な類似度 s_n を算出する．

$$s_n = \mathrm{sim}(\mathrm{Enc}(Q_h), \mathrm{Enc}(I_n)) + \mathrm{sim}(\mathrm{Enc}(Q_h), \mathrm{Enc}(I_n)) \tag{5.5}$$

ただし，sim(·) および Enc(·) はそれぞれコサイン類似算関数および画像特徴量算出関数を示す．

最後に，算出された s_n の降順に整列した検索候補を検索結果とする．

5.2.3　実験: 実験概要

提案手法の有効性を確認するための定量評価について説明しその結果を示す．

5.2.4　実験: データセット

本節では実験で用いるデータセットについて説明する．本実験では，提案手法で用いる画像生成モデルの学習用データセットとして，33 万枚の画像で構成された Common Objects in Context (COCO) [85] データセットを用いた．ただし，COCO データセットにおいては，1 画像あたり 5 つの説明文が付与されている．また，評価用のデータセットとして 94 本の映画より抜粋された 68,000 シーンにより構成される MP2-MD データセットを用いた．MP2-MD データセットに関しても同様に，それぞれのシーンに対して 1 つの説明文が付与されている．

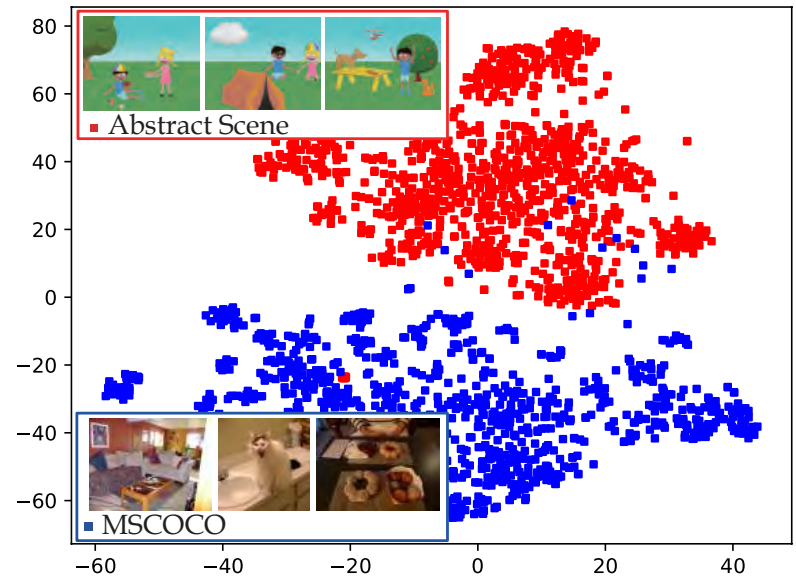
また，本実験では MSCOCO データセットは異なるドメインのデータセットとして Abstract Scene データセット [92] および Artpedia データセット [93] を用いた．Abstract Scene データセットは図 5.2 に示すように，クリップアートにより構成される画像およびその画像の説明文によって構成されている．具体的には，1,200 個の説明文が含まれており，各説明文は 10 個の画像が関連付けられている．また，Artpedia データセットは図 5.2 に示すように，2930 枚の絵画の画像により構成されており，各画像には平均で 3 つの説明文が付与されている．

ここで，MSCOCO データセットとその他のデータセットの分布が分離していることを確認するために、各データセットの視覚特徴量から t-sne を算出した．t-sne の結果を図 5.2 に示す．図 5.2 から，MSCOCO データセットの視覚特徴量とその他のデータセットの視覚特徴量の分布が分離されていることが確認された．

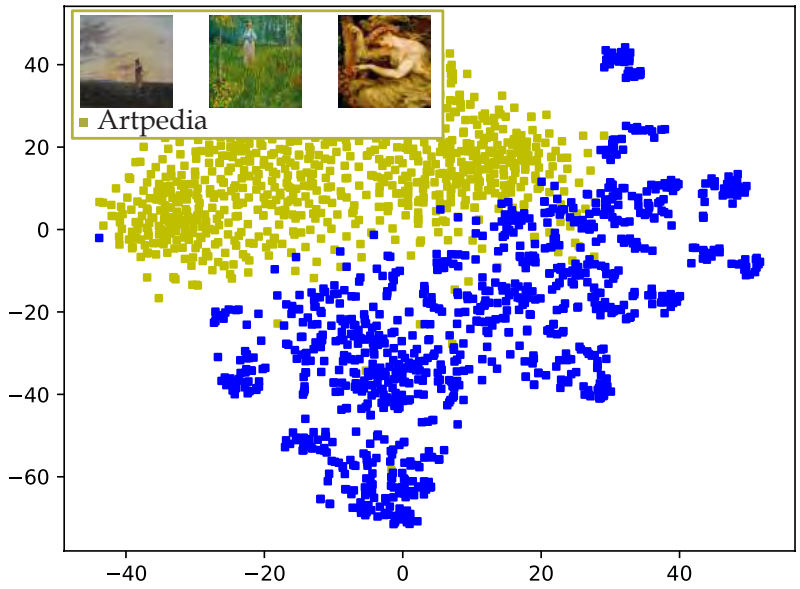
5.2.5　実験:MSCOCO データセットおよび Abstract Scene データセットを用いた定量評価

本実験では，はじめに MSCOCO データセットおよび Abstract Scene データセットをそれぞれ学習データセットおよび検索候補として利用した実験を行うことで，クリップアートを検索対象とした際の有効性を確認する．

本実験では，従来手法として，文献 [35,36,38,39,87,88,94–96] で提案されているクロスモーダル検索手法を用いた．また，生成画像 Q_h のみを検索に利用する手法 (BL1)，生成画像のドメインにおける類似度のみを検索に利用する手法 (BL2) および検索候補のドメインにおける類似度のみを検索に利用する手法 (BL3) との比較を行った．画像特徴量として ImageNet で学習済みの DenseNet121 モデル [97] の出力値を用いており，Text-to-image モデルや Style transfer モデルにはそれぞれ Dynamic Memory GAN [70] および Cycle GAN [98] を用いた．評価指標として，クエリ文に用いたキャプションが付与されていた画像が検索された場合を正しく検索されたと定義した上で，画像が正しく検索された順位の平均値 (以降，平均順位)，画像が正しく検索された順位の中央値 (以降，中央順位) および次式によ



(a) MSCOCO and Abstract Scene



(b) MSCOCO and Artpedia

図 5.2: 実験で用いる各データセットの画像の一例および各データセットの分布の差分．(a) は MSCOCO データセット (青) と Abstract Scene データセット (赤) の分布、(b) は MSCOCO データセット (青) と Artpedia データセット (黄) の分布を示す．本実験では，これらの分布を計算するために、ImageNet で学習済みの DenseNet-121 モデルの出力を画像特徴量として用いた．

り算出される Recall@k を用いた．

$$\text{Recall@}k = \frac{p_k}{J} \quad (k = 1, 2, ..., K) \tag{5.6}$$

ここで，p_k および K は k 位以内に画像が正しく検索されたクエリの個数および検索候補の画像枚数を示す．また，J はクエリの総数を示す．平均順位と中央順位は値が低いほど検索精度が高いことを示し，Recall@k は値が高いほど検索精度が高いことを示す．

表 5.1 に実験結果を示す．表 5.1 において PM は提案手法を示す．表 5.1 より，提案手法の検索精度が従来手法の検索精度よりも高い値を示すことが確認できる．以上より，提案手法が従来手法と比較してクリップアートを検索対象としたクロスモーダル検索に有効であることを確認した．また，提案手法は BL1，BL2，BL3 と比較して検索精度が高い値であることが確認できる．BL1 が Style transfer モデルを利用しない検索であることから Style transfer モデルをクロスモーダル検索に利用することの有効性を確認した．さらに，提案手法の検索精度が生成画像のドメイン，もしくは，検索候補のドメインのどちらかしか利用していない BL2 および BL3 の検索精度よりも高い値を示すことから，提案手法のように両者のドメインを利用することの有効性を確認した．同様に，図 5.3 に検索結果の一例を示す．図 5.3 より，提案手法が定性的にもクエリに関連する画像を検索可能である事を確認した．

表 5.1: MSCOCO データセットおよび Abstract Scene データセットをそれぞれ学習データセットおよび検索候補として利用した実験.

	R@1	R@10	R@100	Mean	Median
Kiros ’14	0.0	0.023	0.031	1371.01	881
Vendrov ’16	0.0	0.021	0.044	1354.43	850
Zheng ’17	0.0	0.014	0.054	1332.55	815.5
Faghri ’17	0.012	0.022	0.051	1291.11	793
Liu ’17	0.015	0.042	0.063	1341.65	810
Zhang ’18	0.014	0.032	0.078	1301.21	789.5
Huang ’18	0.022	0.044	0.084	1286.66	713
Gu ’18	.025	0.031	0.094	1254.18	768.5
Song ’19	.032	0.042	0.11	1234.61	732
Ji ’19	0.044	0.048	0.12	1205.39	711.5
BL1	0.0	0.013	0.080	1301.13	803.5
BL2	0.10	0.11	0.20	1130.02	609
BL3	0.10	0.12	0.21	878.91	450
PM	**0.20**	**0.21**	**0.29**	**811.98**	**437.5**

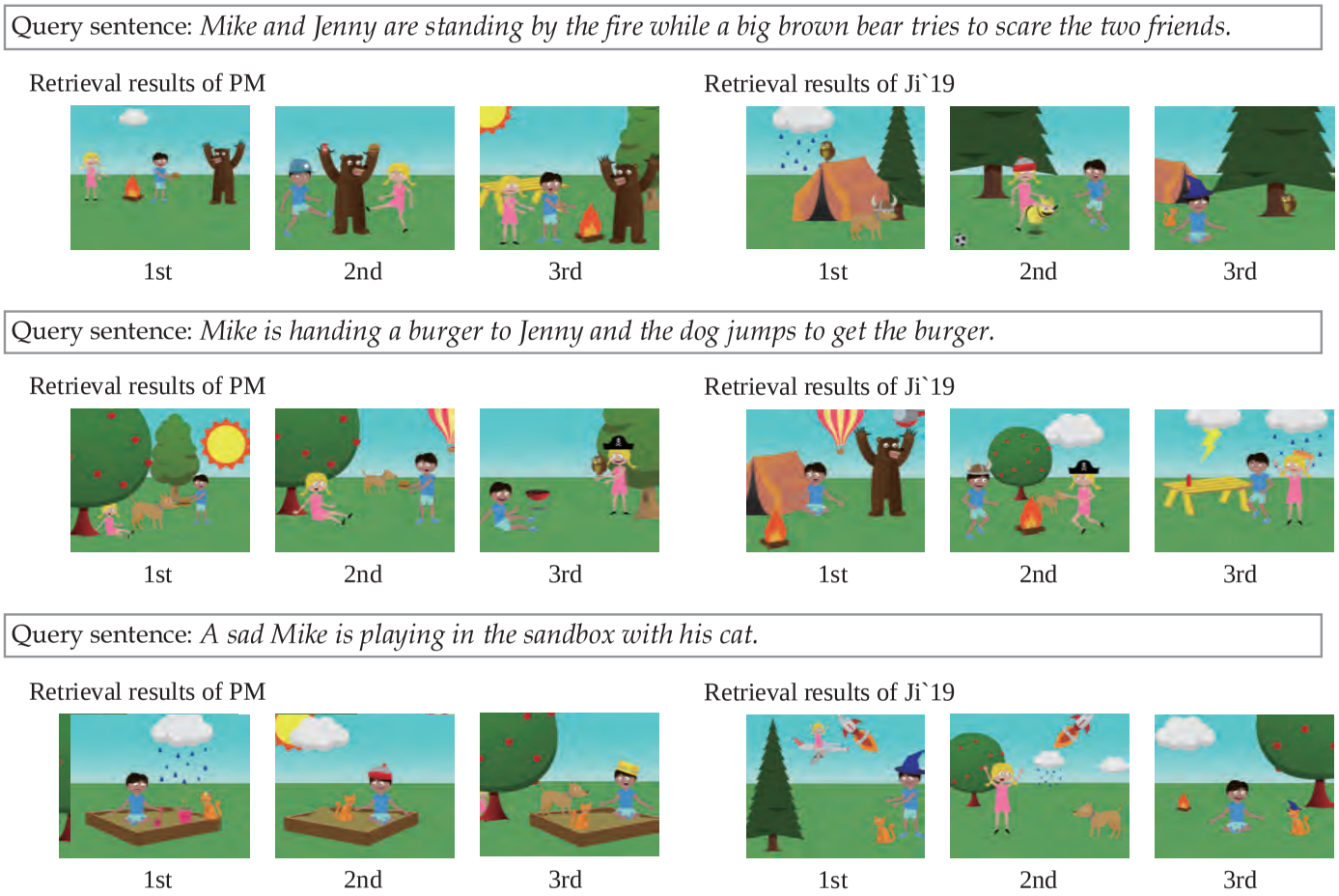


図 5.3: MSCOCO データセットおよび Abstract Scene データセットをそれぞれ学習データセットおよび検索候補として利用した際の検索結果の一例

5.2.6　実験: MSCOCO データセットおよび Artpedia データセットを用いた定量評価

本実験では，はじめに MSCOCO データセットおよび Artpedia データセットをそれぞれ学習データセットおよび検索候補として利用した実験を行うことで，クリップアートを検索対象とした際の有効性を確認する．

本実験では，**5.2.5** 節と同様に，従来手法として，文献 [35,36,38,39,87,88,94–96] で提案されているクロスモーダル検索手法を用いた．また，生成画像 Q_h のみを検索に利用する手法 (BL1)，生成画像のドメインにおける類似度のみを検索に利用する手法 (BL2) および検索候補のドメインにおける類似度のみを検索に利用する手法 (BL3) との比較を行った．画像特徴量として ImageNet で学習済みの DenseNet121 モデル [97] の出力値を用いており，Text-to-image モデルや Style transfer モデルにはそれぞれ Dynamic Memory GAN [70] および Cycle GAN [98] を用いた．評価指標として，クエリ文に用いたキャプションが付与されていた画像が検索された場合を正しく検索されたと定義した上で，画像が正しく検索された順位の平均値 (以降，平均順位)，画像が正しく検索された順位の中央値 (以降，中央順位) および次式により算出される Recall@k を用いた．

$$\text{Recall@}k = \frac{p_k}{J} \quad (k = 1, 2, ..., K) \tag{5.7}$$

ここで，p_k および K は k 位以内に画像が正しく検索されたクエリの個数および検索候補の画像枚数を示す．また，J はクエリの総数を示す．平均順位と中央順位は値が低いほど検索精度が高いことを示し，Recall@k は値が高いほど検索精度が高いことを示す．

表 5.2 に実験結果を示す．表 5.2 において PM は提案手法を示す．表 5.2 より，絵画を検索対象とした際もクリップアートを検索対象とした際と同様の結論が得られた．また，図 5.4 に検索結果の一例を示す．図 5.4 より，提案手法が定性的にもクエリに関連する画像を検索可能である事を確認した．

表 5.2: MSCOCO データセットおよび Artpedia データセットをそれぞれ学習データセットおよび検索候補として利用した実験

	R@1	R@10	R@100	Mean	Median
Kiros ’14	0.033	0.070	0.10	1400.63	1421
Vendrov ’16	0.064	0.073	0.098	1356.70	1371
Zheng ’17	0.071	0.082	0.10	1456.32	1401
Faghri ’17	0.075	0.084	0.10	1311.09	1351
Liu ’17	0.089	0.094	0.13	1321.12	1300.5
Zhang ’18	0.085	0.096	0.12	1301.91	1296.5
Huang ’18	0.091	0.099	0.13	1278.45	1233
Gu ’18	0.098	0.10	0.14	1243.33	1199.5
Song ’19	0.093	0.11	0.15	1275.66	1205
Ji ’19	0.10	0.14	0.19	1200.02	1161.5
BL1	0.035	0.078	0.11	1371.33	1406
BL2	0.18	0.22	0.27	900.34	894
BL3	0.22	0.26	0.31	849.70	822
PM	**0.25**	**0.29**	**0.34**	**795.52**	**751**

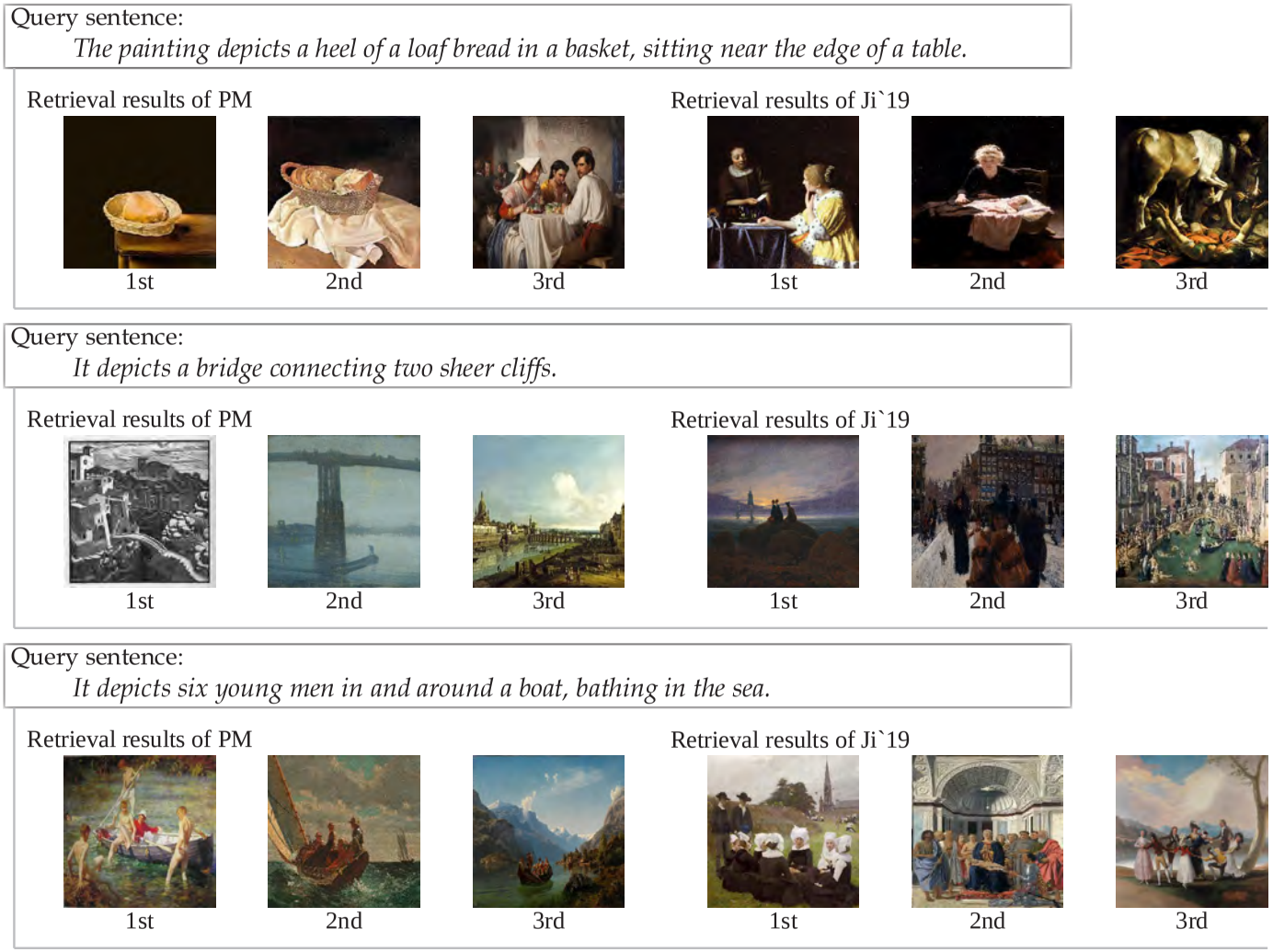


図 5.4: MSCOCO データセットおよび Artpedia データセットをそれぞれ学習データセットおよび検索候補として利用した際の検索結果の一例

5.3　まとめ

本章では学習データと検索候補のドメインが大きく異なる場合にでも，高精度に目的のマルチメディアを検索することが可能なクロスモーダル検索手法を提案した．クリップアートおよび絵画を検索対象とした実験により提案手法の有効性を確認した．

第6章　定型文を用いた質問応答に基づく対話型マルチメディア情報検索

6.1　はじめに

本章では，定型文を用いた質問応答に基づく対話型マルチメディア情報検索を提案する．従来のマルチメディア情報検索手法により，クエリテキストに関連したマルチメディアコンテンツを高精度に検索可能な枠組みが実現されている．一方，従来の手法は，クエリテキストから目的のマルチメディアコンテンツを一意に特定可能であることを前提に設計が行われているそのため，ユーザの与えるクエリテキストに十分な情報が含まれず，検索候補中にクエリテキストに該当するマルチメディアコンテンツが複数存在する場合，一度の検索で高精度な検索結果を得ることは困難である．上記の課題を解決するために，本章では検索候補を絞り込むために必要な情報を質問応答の形式でユーザに問いかけるマルチメディア情報検索手法を提案する．考案した手法により，ユーザは提示された質問に回答するだけで大幅に検索結果を改善することが可能である．

6.2　ユーザとの質問応答に基づく画像再検索

本章では、ユーザとの質問応答に基づく画像再検索について説明する．提案手法の概念図および概要図を図 6.1 および図 6.2 に示す．提案手法では，“ある物体が検索目的の画像に含まれるかどうか?” という質問をユーザに行い，従来の初期検索手法により順位付けされた検索候補 I_n ($n = 1, 2, ..., N; N$ は検索候補の総画

像数) を再決定する．ただし，I_n は n 位の画像を表す．また，I_n には一切のラベルが付与されていない．

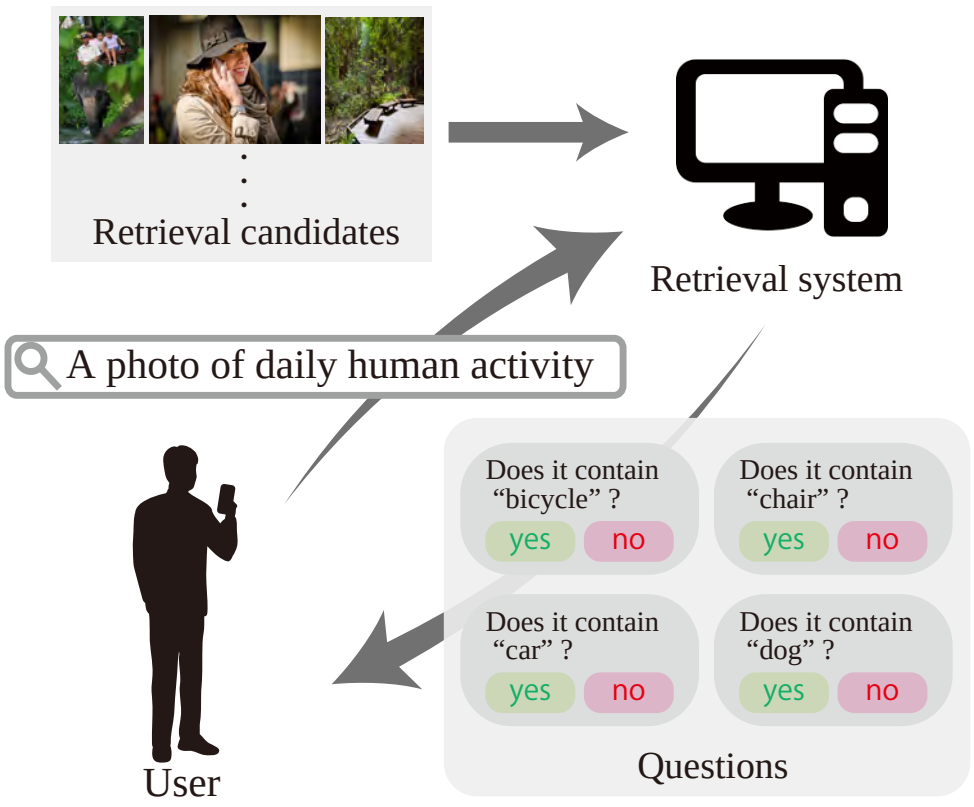


図 6.1: 定型文を用いた質問応答に基づく対話型マルチメディア情報検索手法の概念図.

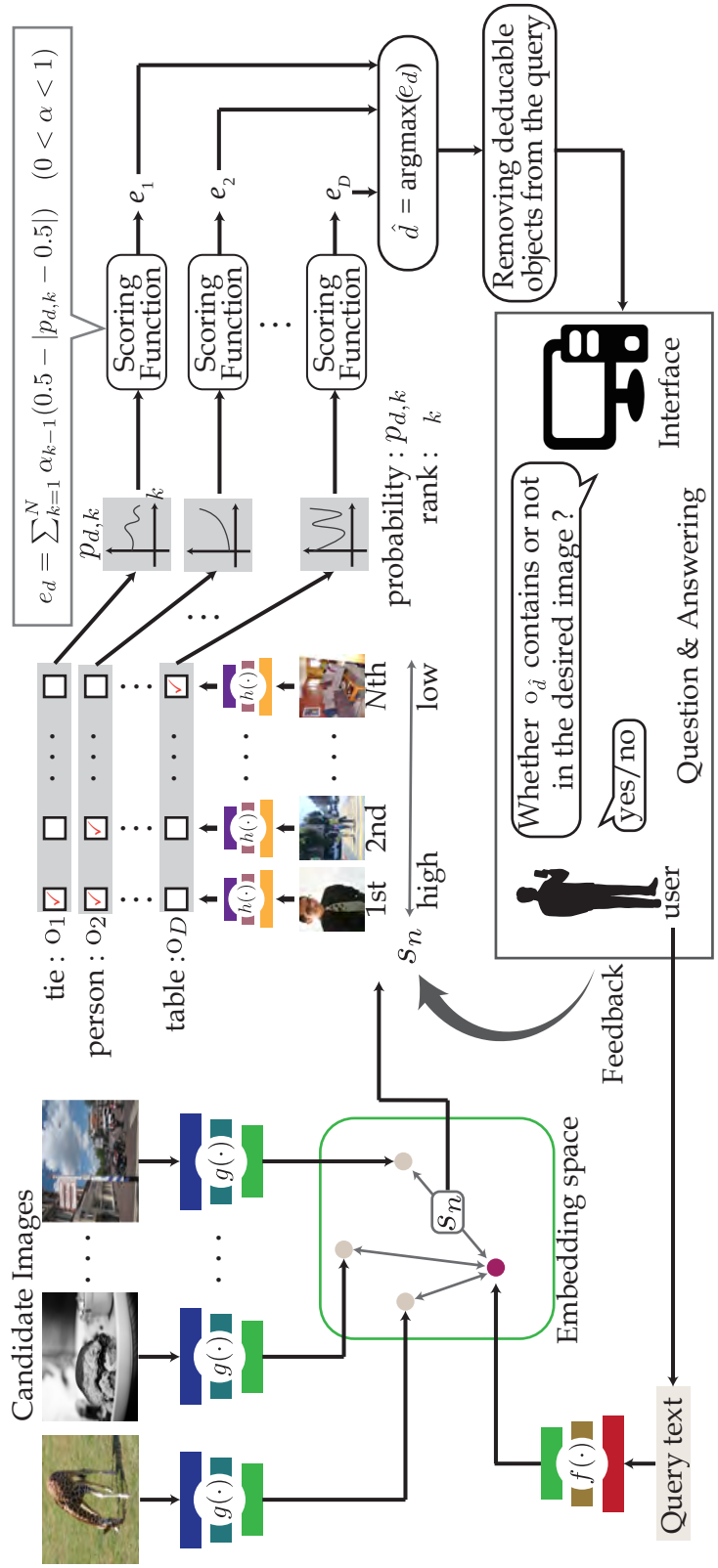


図 6.2: 定型文を用いた質問応答に基づく対話型マルチメディア情報検索手法の概要図.

6.2.1　ユーザへの質問応答に最適な物体の探索

本節ではユーザへの質問応答に最適な物体の探索について説明する．本探索では上位の画像を効率よく絞り込むことが可能な質問応答を行うために，検索上位の画像において情報量が多い物体を探索する．はじめに，学習済みの物体検出手法を用いて，物体 obj_d ($d = 1, 2, ..., D$; D は検出され得る物体の総数) が I_n に含まれる確率 $l_{d,n}$ を算出する．その後，obj_d が I_n に含まれるかを表す二値表現 $x_{d,n}$ を以下の通り算出する．

$$x_{d,n} = \begin{cases} 1 & (l_{d,n} \geq \mathrm{th}_{\mathrm{obj}}) \\ 0 & (l_{d,n} < \mathrm{th}_{\mathrm{obj}}) \end{cases} \tag{6.1}$$

ただし，$\mathrm{th}_{\mathrm{obj}}$ は物体検出手法ごとに設定される最適な閾値である．以上により算出された $x_{d,n}$ に基づいて，物体 obj_d を含む k ($k = 1, 2, ..., N$) 位以上の検索候補の割合 $p_{d,k}$ を以下の通り算出する．

$$p_{d,k} = \sum_{n=1}^{k} \frac{x_{d,n}}{k} \tag{6.2}$$

最後に，以下の式により，各物体の評価値 s_d を算出する．

$$s_d = \sum_{k=1}^{N} \alpha^{k-2}(0.5 - |p_{d,k} - 0.5|) \quad (0 < \alpha < 1) \tag{6.3}$$

ただし，α はハイパーパラメータである．以上により算出された評価値 s_d が高い物体は検索上位の画像において情報量の多い物体であると考えられる．そのため，評価値 s_d が高い物体を質問文に用いることで，検索上位の画像を効率よく絞り込む質問応答が可能となる．

Algorithm 1 検索順位の再決定

```
c ⇐ 1
for m = M ... 0 do
  for n = 1 ... N do
    if Σ_{i=1}^{m} a_{i,n} = m then
      I_c^out ⇐ I_n
      c ⇐ c + 1
    end if
  end for
end for
```

6.2.2　質問応答および検索順位の再決定

本節ではユーザとの質問応答およびユーザの回答に基づく検索順位の再決定について説明する．簡単のため，本節ではユーザは提示された質問全てに回答すると仮定する．はじめに，s_d が高い順に物体 $\mathrm{obj}_m^{\mathrm{Q}}$ ($m = 1, 2, ..., M$; M は質問を行う最大回数) を抽出する．ただし，クエリテキストに含まれる単語との類似度が閾値 $\mathrm{th}_{\mathrm{word}}$ 以上の物体は抽出されないようにする．その後，“$\mathrm{obj}_m^{\mathrm{Q}}$ が検索目的の画像に含まれるかどうか?” という質問文をユーザに提示する．続いて，“ユーザによる m 個目の質問に対する回答” と “n 位の検索候補における物体 $\mathrm{obj}_m^{\mathrm{Q}}$ の有無” が一致する場合 1 をとり，一致しない場合は 0 をとる $a_{m,n}$ を算出する．ただし，検索候補における物体の有無は $x_{d,n}$ に基づいて決定する．最後に，$a_{m,n}$ を用いて，最終的な検索順位 $\mathrm{I}_n^{\mathrm{out}}$ を Algorithm 1 の通りに再決定する．

6.2.3　実験: 初期検索手法との比較

本実験では，従来の初期検索手法の検索精度と提案手法の検索精度を比較することで提案手法の定量的な有効性を検証する．同時に，質問応答を複数回行うことの有効性について検証する．

まず，本実験のデータセットとして，Microsoft Common Objects in Context (MSCOCO) [85] データセットおよび Visual Genome データセット [99] を用いた．MSCOCO データセットは学習用データとして 82,783 枚の画像が含まれてお

り，テスト用データとして 5,000 枚の画像が含まれている．また，学習用の画像に対しては画像を説明する文 (以降，テキストラベル) が 5 つ付与されており，テスト用の画像に対してはテキストラベルが 1 つ付与されている．Visual Genome データセットは学習用データとして 75,578 枚の画像が含まれており，テスト用データとして 32,433 枚の画像が含まれている．さらに，ユーザによる質問文への回答には各データセットに付与されている物体ラベルを用いた．物体検出手法には Yolov3 [100] を使用し，クエリテキストに含まれる単語との類似度には word2vec [101] により抽出される特徴量に基づくコサイン類似度を用いた．また，$\mathrm{th}_{\mathrm{obj}}$，$\alpha$ および $\mathrm{th}_{\mathrm{word}}$ にはそれぞれ 0.5，0.99 および 0.6 を用いた．さらに，従来手法として，文献 [35, 36, 38, 39, 87, 88, 94–96, 102] で提案されている検索手法を使用し，評価指標として次式により算出される Recall@k を用いた．

$$\mathrm{Recall@}k = \frac{r_k}{J} \tag{6.4}$$

ここで，r_k は正解が k 位以上に存在する入力の個数を示す．ただし，今回の実験ではクエリであるテキストラベルが付与されていた画像が検索された場合を正解とした．また，J は入力の総数 (=5,000) を示す．

表 6.1 および表 6.2 に実験結果を示す．また，図 6.3 および図 6.4 に提案手法により検索された検索結果のサンプルを示す．表 6.1 および表 6.2 において，PM (Kiros ’14 + 1r) は Kiros ’14 らにより提案された手法 [35] に対して提案手法を用いることで 1 度再検索を行った際の実験結果を示す．表 6.1 および表 6.2 より，質問応答を一度だけ行う PM の評価値が基となる初期検索手法の検索精度を上回っていることが確認できる．以上より，提案手法を用いて順位を再決定することの有効性を確認した．同様に，複数回再検索を行った際の評価値が他の手法の評価値を上回っていることが確認できる．このことから質問応答を複数回行うことの有効性を確認した．

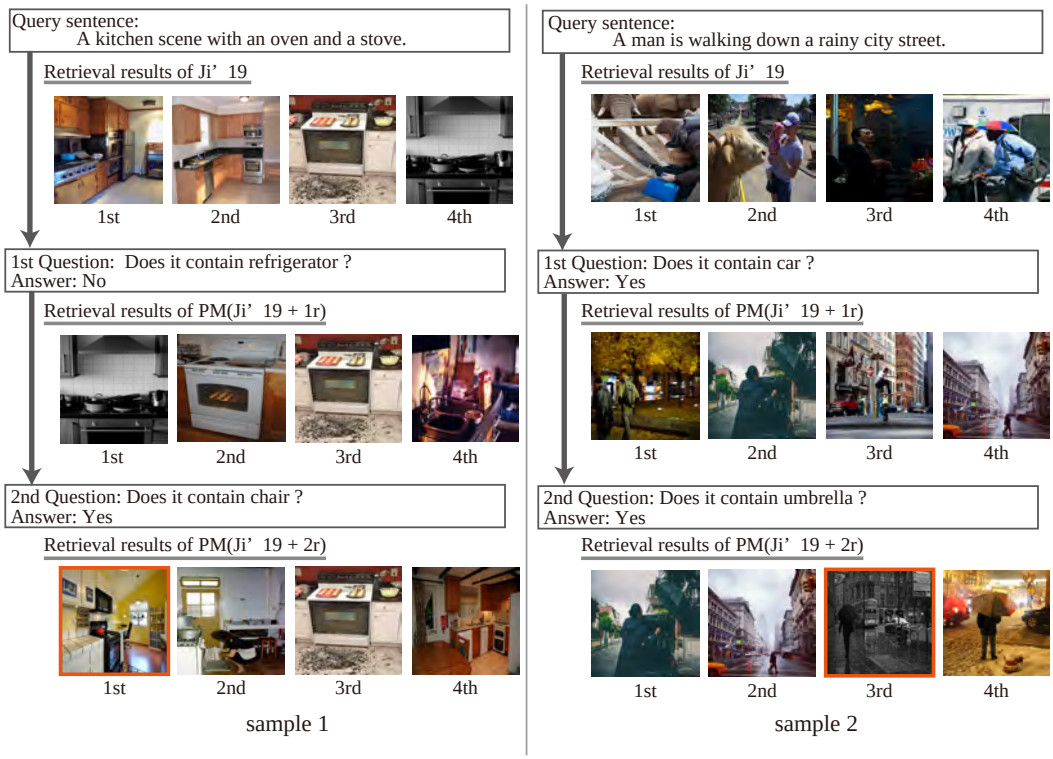


図 6.3: MSCOCO データセットを用いた際の検索結果のサンプル.

6.2.4　実験: 再検索手法との比較

6.2.3 節では，初期検索手法と比較することで，提案手法が初期検索手法の検索精度を向上させることが可能であることを確認した．次に，本節では，他の再検索手法と検索精度を比較することで，提案手法が従来の再検索手法 [103–107] よりも高精度に初期の検索結果を改善することが可能であることを示す．本実験では，文献 [96] により算出された検索結果を初期検索結果として利用した．ここで，従来の再検索手法のハイパーパラメータは従来手法の検索精度が最大となる値を採用した．

実験結果を表 6.3 および 6.4 に示す．表 6.3 および表 6.4 より，提案手法の評価値が他の再検索手法の評価値よりも高い値であることが分かる．以上の実験結果より，提案手法の有効性を確認した．

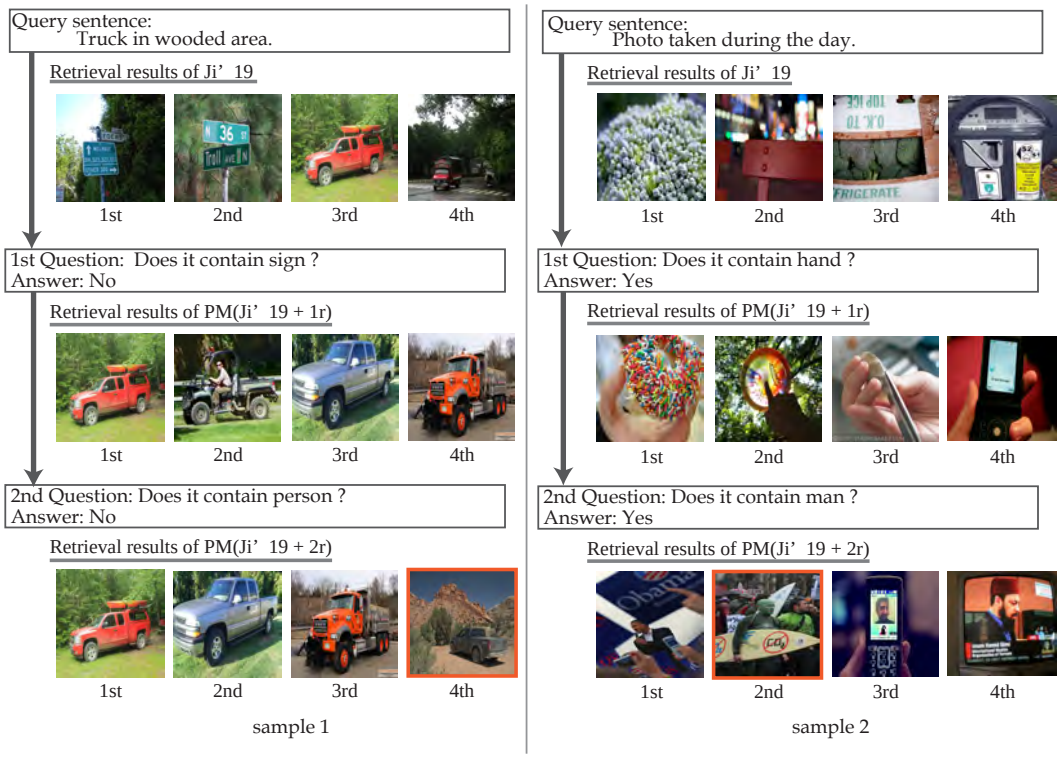


図 6.4: Visual Genome データセットを用いた際の検索結果のサンプル.

表 6.1: MSCOCO データセットを用いた際の定型文を用いた質問応答に基づく対話型マルチメディア情報検索手法と初期検索手法との比較.

	Recall@1	Recall@10	Mean	Median
Kiros '14 [35]	0.154	0.504	70.464	10
PM (Kiros '14 + 1r)	0.205	0.523	65.243	8
PM (Kiros '14 + 2r)	0.224	0.541	61.305	7
PM (Kiros '14 + 3r)	**0.255**	**0.588**	**58.242**	**7**
Vendrov '16 [36]	0.187	0.576	64.592	9
PM (Vendrov '16 + 1r)	0.224	0.589	62.534	7
PM (Vendrov '16 + 2r)	0.256	0.604	61.442	7
PM (Vendrov '16 + 3r)	**0.275**	**0.625**	**57.634**	**6**
Zheng '17 [102]	0.288	0.703	53.540	9
PM (Zheng '17 + 1r)	0.305	0.715	50.735	6
PM (Zheng '17 + 2r)	0.313	0.745	47.634	5
PM (Zheng '17 + 3r)	**0.342**	**0.779**	**43.756**	**4**
Faghri '17 [38]	0.297	0.724	52.440	9
PM (Faghri '17 + 1r)	0.313	0.741	47.645	7
PM (Faghri '17 + 2r)	0.335	0.764	44.692	5
PM (Faghri '17 + 3r)	**0.351**	**0.790**	**40.645**	**4**
Liu '17 [87]	0.294	0.709	47.300	9
PM (Liu '17 + 1r)	0.316	0.719	45.656	7
PM (Liu '17 + 2r)	0.336	0.736	42.412	6
PM (Liu '17 + 3r)	**0.362**	**0.766**	**37.051**	**4**
Zhang '18 [39]	0.303	0.752	30.223	7
PM (Zhang '18 + 1r)	0.317	0.768	29.846	5
PM (Zhang '18 + 2r)	0.334	0.789	27.632	3
PM (Zhang '18 + 3r)	**0.360**	**0.815**	**25.735**	**2**
Huang '18 [94]	0.301	0.755	28.480	5
PM (Huang '18 + 1r)	0.322	0.769	26.624	3
PM (Huang '18 + 2r)	0.346	0.781	24.723	**2**
PM (Huang '18 + 3r)	**0.369**	**0.803**	**20.724**	**2**
Gu '18 [95]	0.307	0.746	26.731	3
PM(Gu '18 + 1r)	0.326	0.768	22.267	2
PM(Gu '18 + 2r)	0.355	0.791	20.874	2
PM(Gu '18 + 3r)	0.378	0.810	18.615	**1**
Song '19 [88]	0.317	0.759	20.463	2
PM (Song '19 + 1r)	0.349	0.776	19.315	**1**
PM (Song '19 + 2r)	0.375	0.805	18.624	**1**
PM (Song '19 + 3r)	**0.391**	**0.819**	**17.637**	**1**
Ji '19 [96]	0.317	0.777	23.143	2
PM (Ji '19 + 1r)	0.345	0.793	22.972	**1**
PM (Ji '19 + 2r)	0.362	0.808	20.624	**1**
PM (Ji '19 + 3r)	**0.392**	**0.821**	**17.905**	**1**

表 6.2: Visual Genome データセットを用いた際の定型文を用いた質問応答に基づく対話型マルチメディア情報検索手法と初期検索手法との比較.

	Recall@1	Recall@10	Mean	Median
Kiros '14 [35]	0.0218	0.107	5842.051	588.5
PM (Kiros '14 + 1r)	0.0255	0.123	4758.590	517
PM (Kiros '14 + 2r)	0.0266	0.141	4333.004	486
PM (Kiros '14 + 3r)	**0.0274**	**0.155**	**4004.018**	**442**
Vendrov '16 [36]	0.0248	0.110	4324.212	531.5
PM (Vendrov '16 + 1r)	0.0268	0.133	3783.305	400
PM (Vendrov '16 + 2r)	0.0279	0.145	3599.032	356
PM (Vendrov '16 + 3r)	**0.0292**	**0.156**	**3329.069**	**288**
Zheng '17 [102]	0.0263	0.118	3659.612	475
PM (Zheng '17 + 1r)	0.0272	0.132	3544.433	355
PM (Zheng '17 + 2r)	0.0284	0.144	3432.694	320
PM (Zheng '17 + 3r)	**0.0295**	**0.150**	**3320.202**	**278**
Faghri '17 [38]	0.0266	0.118	3646.613	365
PM (Faghri '17 + 1r)	0.0278	0.144	3469.609	295
PM (Faghri '17 + 2r)	0.0281	0.156	3296.960	278
PM (Faghri '17 + 3r)	**0.0287**	**0.169**	**2803.332**	**266**
Liu '17 [87]	0.0268	0.119	3588.127	318
PM (Liu '17 + 1r)	0.0270	0.132	3361.629	278
PM (Liu '17 + 2r)	0.0283	0.145	3255.130	268
PM (Liu '17 + 3r)	**0.0299**	**0.158**	**3059.640**	**255**
Zhang '18 [39]	0.0270	0.122	3530.614	274
PM (Zhang '18 + 1r)	0.0274	0.146	3329.502	268
PM (Zhang '18 + 2r)	0.0282	0.150	3059.403	252
PM (Zhang '18 + 3r)	**0.0291**	**0.159**	**2758.643**	**231**
Huang '18 [94]	0.0275	0.121	3266.327	266
PM (Huang '18 + 1r)	0.0281	0.145	3098.205	251
PM (Huang '18 + 2r)	0.0292	0.161	2911.590	241
PM (Huang '18 + 3r)	**0.0301**	**0.172**	**2877.593**	**230**
Gu '18 [95]	0.0275	0.127	3054.524	260
PM(Gu '18 + 1r)	0.0291	0.134	2960.430	255
PM(Gu '18 + 2r)	0.0309	0.151	2877.519	241
PM(Gu '18 + 3r)	**0.0325**	**0.177**	**2759.523**	**232**
Song '19 [88]	0.0280	0.129	2943.513	258
PM (Song '19 + 1r)	0.0313	0.139	2885.490	240
PM (Song '19 + 2r)	0.0336	0.162	2809.539	233
PM (Song '19 + 3r)	**0.0343**	**0.178**	**2755.493**	**210**
Ji '19 [96]	0.0286	0.130	2861.242	248
PM (Ji '19 + 1r)	0.0304	0.155	2855.645	243
PM (Ji '19 + 2r)	0.0312	0.162	2810.443	235
PM (Ji '19 + 3r)	**0.0324**	**0.171**	**2790.539**	**222**

表 6.3: MSCOCO データセットを用いた際の定型文を用いた質問応答に基づく対話型マルチメディア情報検索手法と再検索手法との比較.

	Recall@1	Recall@10	Mean	Median
BL(Ji ’19 [96])	0.317	0.777	23.143	2
Giacinto ’07 [103]	0.316	0.655	20.052	2
Liang ’08 [104]	0.275	0.605	39.701	3
Xu ’13 [105]	0.317	0.765	20.534	2
Lin ’15 [106]	0.315	0.650	19.110	2
putzu ’20 [107]	0.278	0.614	35.435	3
PM	**0.345**	**0.793**	**22.972**	**1**

表 6.4: Visual Genome データセットを用いた際の定型文を用いた質問応答に基づく対話型マルチメディア情報検索手法と再検索手法との比較.

	Recall@1	Recall@10	Mean	Median
BL (Ji ’19 [96])	0.0286	0.130	2861.242	248
Giacinto ’07 [103]	0.0283	0.129	2862.621	255
Liang ’08 [104]	0.0205	0.0952	4109.735	610
Xu ’13 [105]	0.0287	0.132	2899.920	245
Lin ’15 [106]	0.0282	0.130	2892.026	247
putzu ’20 [107]	0.0208	0.0966	4001.001	548
PM	**0.0304**	**0.155**	**2855.645**	**243**

6.3　想起のしやすさを考慮した質問応答に基づく対話型マルチメディア情報検索

本節では、想起のしやすさを考慮した質問応答に基づく対話型マルチメディア情報再検索手法について説明する．提案手法の概念図および概要図を図 6.5 および図 6.6 に示す．提案手法では，質問応答に最適な物体を探索した後に，“その物体が検索目的の画像に含まれるか?” という質問をユーザに行う．その後，ユーザは “含まれる” および “含まれない” という 2 つの選択肢の中から検索目的の画像に最も当てはまる選択肢を選択する．最後に，以上により選択された選択肢に基づいて，従来のマルチメディア情報検索手法により順位付けされた検索候補 I_n ($n = 1, 2, ..., N$; N は検索候補の総画像数) の順位を再決定する．ただし，I_n は n 位の画像を表す.

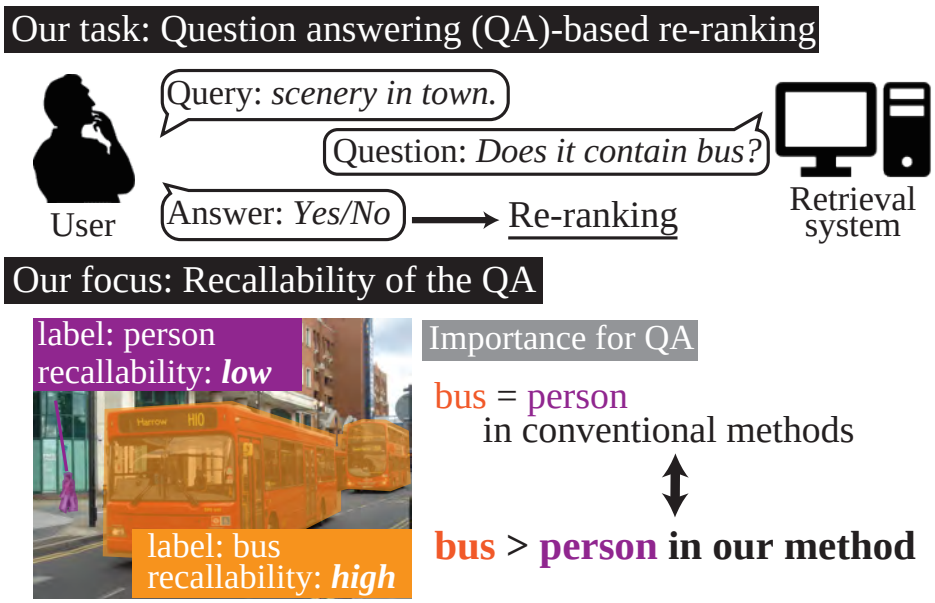


図 6.5: 想起のしやすさを考慮した質問応答に基づく対話型マルチメディア情報検索の概念図.

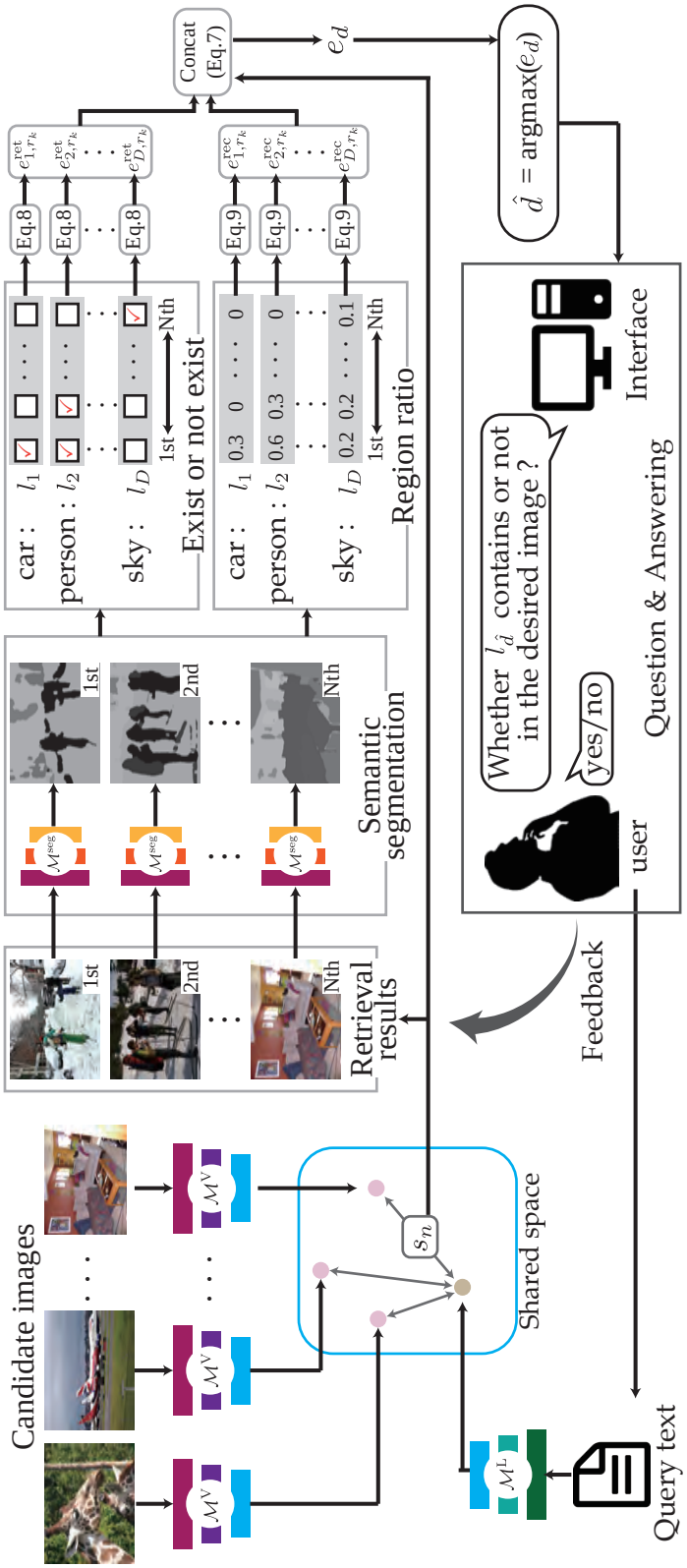


図 6.6: 想起のしやすさを考慮した質問応答に基づく対話型マルチメディア情報検索の概要図.

6.3.1　質問応答に最適な物体の探索

本節では質問応答に最適な物体の探索について説明する．本探索では上位の画像を効率よく絞り込むことが可能であり，かつ，想起のしやすい質問応答を行うために，検索上位の画像において情報量が多く，画像内における領域のサイズが大きい物体を探索する．はじめに，学習済みのセマンティックセグメンテーション手法を用いて，I_n に含まれるラベル obj_d ($d = 1, 2, ..., D$; D は検出され得る物体の総数) のピクセル数 $l_{d,n}$ を算出する．その後，obj_d が I_n に含まれるかを表す 2 値表現 $x_{d,n}$ を以下の通り算出する．

$$x_{d,n} = \begin{cases} 1 & (l_{d,n} \geq 1) \\ 0 & (l_{d,n} < 1) \end{cases} \tag{6.5}$$

以上により算出された $x_{d,n}$ に基づいて，物体 obj_d を含む k ($k = 1, 2, ..., N$) 位以上の検索候補の割合 $p_{d,k}$ を以下の通り算出する．

$$p_{d,k} = \sum_{n=1}^{k} \frac{x_{d,n}}{k} \tag{6.6}$$

最後に，以下の式により，各物体の評価値 s_d を算出する．

$$s_d = \sum_{k=1}^{N} s_{r_k}(\beta(0.5 - |p_{d,k} - 0.5|) + (1 - \beta)\frac{l_{d,k}}{l_k^{\mathrm{sum}}}) \quad (0 < \alpha < 1) \tag{6.7}$$

ただし，β はハイパーパラメータであり，s_n はテキストと画像の類似度，l_k^{sum} は I_k の総ピクセル数である．以上により算出された評価値 s_d が高い物体は検索上位の画像において情報量が多く，かつ，画像内に占める領域のサイズが大きい物体であると考えられる．そのため，評価値 s_d が高い物体を質問文に用いることで，検索上位の画像を効率よく絞り込むことが可能であり，また，ユーザにとって想起のしやすい質問応答が可能になると期待される．

Algorithm 2 検索順位の再決定

```
c ⇐ 1
for n = 1 ... N do
  if a_n = 1 then
    I_c^out ⇐ I_n
    c ⇐ c + 1
  end if
end for
for n = 1 ... N do
  if a_n = 0 then
    I_c^out ⇐ I_n
    c ⇐ c + 1
  end if
end for
```

6.3.2　質問応答および検索順位の再決定

本節ではユーザとの質問応答に基づく検索順位の再決定について説明する．はじめに，s_d が最も高い物体 $\mathrm{obj}^{\mathrm{Q}}$ を抽出する．ただし，クエリ文に含まれる単語との類似度が閾値 $\mathrm{th}_{\mathrm{word}}$ 以上の物体は抽出されないようにする．その後，“$\mathrm{obj}^{\mathrm{Q}}$ が検索目的の画像に含まれるかどうか?” という質問文をユーザに提示する．続いて，“ユーザによる質問に対する回答” と “n 位の検索候補における物体 $\mathrm{obj}^{\mathrm{Q}}$ の有無” が一致する場合 1 をとり，一致しない場合は 0 をとる a_n を算出する．ただし，検索候補における物体の有無は $x_{d,n}$ に基づいて決定する．最後に，a_n を用いて，最終的な検索順位 $\mathrm{I}_n^{\mathrm{out}}$ を Algorithm 1 の通りに再決定する．

6.3.3　実験: 類似した検索候補に対する検索精度の検証

一般的に，従来研究において，ユーザは検索候補の全容を確認することなくクエリテキストを考案する．そのため，クエリテキストの内容が複数の画像に該当する可能性が存在する．また，このようなクエリテキストが与えられた際には，従来研究では，検索候補を絞り込むことが困難となり，検索精度が低下する可能性が高い．特に，検索候補の内容が類似している場合 (以降，偏りのある状況) に

は，クエリの内容が複数の画像に該当する可能性が高くなることから，上記の課題はより顕著になると考えられる．そこで，本節では，データベースに類似した検索候補が複数含まれる場合の従来のマルチメディア情報検索手法の検索精度を検証することで，上記の仮説を検証する．本実験では，文献 [85] より提供される MSCOCO データセットを基に偏りのある評価用データセット (以降，バイアス DB) および偏りのない評価用データセット (以降，ランダム DB) を構築した．まず，バイアス DB として MSCOCO データセットから物体ラベル “person” を含むサンプルのみを抽出した．また，バイアス DB と同じサンプル数になるように MSCOCO データセットからランダムにサンプルを抽出することでランダム DB を構築した．評価指標としては，Recall@k(R@k)= $\frac{w_k}{J}$ を用いた．ここで，J および w_k はそれぞれクエリの総数 (=5,000) および k 位以内に画像が正しく検索されたクエリの個数を示す．ただし，今回の実験ではクエリが付与されていた画像が検索された場合を正しく検索されたと定義した．

表 6.5 に実験結果を示す．表 6.5 より，バイアス DB に対する検索精度がランダム DB に対する検索精度を下回っていることが確認できる．以上より，従来のマルチメディア情報検索手法では偏りのある状況における検索が困難であることを確認した．

6.3.4　実験: 初期検索手法との比較

本実験では，想起のしやすさを考慮した質問応答に基づく対話型マルチメディア情報再検索手法の定量的な有効性を検証する．

本実験のテスト用データとして，MSCOCO データセット [85] から 5,000 枚の画像を用いた．また，質問文への回答には MSCOCO データセットに付与されているラベルを用いた．セマンティックセグメンテーション手法には文献 [113] より提供されるモデルを使用し，クエリテキストに含まれる単語との類似度には word2vec [101] により抽出される特徴量に基づくコサイン類似度を用いた．また，α，β および $\mathrm{th}_{\mathrm{word}}$ にはそれぞれ 0.99，0.5 および 0.6 を用いた．提案手法の初期検索には文献 [35, 39, 87, 88, 94–96, 108–112] で提案されているマルチメディア情

表 6.5: 偏りのない評価用データセット (ランダム DB) および偏りのある評価用データセット (バイアス DB) に対する想起のしやすさを考慮した質問応答に基づく対話型マルチメディア情報検索とマルチメディア情報検索手法の検索精度.

	R@1		R@10	
	ランダム DB	バイアス DB	ランダム DB	バイアス DB
UVS [35]	0.272	0.161	0.546	0.471
RRF-Net [87]	0.381	0.314	0.721	0.635
CMPM [39]	0.398	0.326	0.802	0.687
SISM [94]	0.422	0.335	0.819	0.716
GXN [95]	0.432	0.332	0.831	0.741
PVSE [88]	0.425	0.343	0.830	0.761
SAN [96]	0.437	0.347	0.821	0.768
VSRN [108]	0.441	0.379	0.838	0.793
PCME [109]	0.402	0.349	0.803	0.751
SGRAF [110]	0.447	0.375	0.849	0.788
CGMN [111]	0.426	0.336	0.841	0.740
IMC [112]	0.416	0.329	0.820	0.709

報検索手法を使用した．検索精度に関する評価指標として，次式により算出される Recall@k を用いた．

$$\text{Recall@}k = \frac{r_k}{J} \tag{6.8}$$

ここで，r_k および J は正解が k 位以内に存在するクエリの個数およびクエリテキストの総数を示す．ただし，今回の実験ではクエリテキストに対応する画像が検索された場合を正解とした．また，想起のしやすさに関する評価においては，画像内の物体の大きさと想起のしやすさには相関があるという知見に基づき，検索目的画像における物体の大きさを評価指標として用いた．ただし，本評価では質問応答に利用された物体の大きさを評価した．

表 6.6 および表 6.7 に実験結果を示す．表 6.6 および表 6.7 において，UVS+Ours は従来のマルチメディア情報検索手法 [35] に提案手法を導入した際の検索精度

表 6.6: 偏りのない評価用データセットを対象とした際の想起のしやすさを考慮した質問応答に基づく対話型マルチメディア情報検索と初期検索手法との比較.

	R@1	R@10	平均順位	中央順位
UVS [35]	0.154	0.504	70.464	10
UVS+Ours	**0.278**	**0.688**	**43.523**	**8**
RRF-Net [87]	0.294	0.709	47.300	9
RRF-Net+Ours	**0.378**	**0.751**	**31.112**	**5**
CMPM [39]	0.303	0.752	30.223	7
CMPM+Ours	**0.387**	**0.804**	**17.453**	**4**
SISM [94]	0.301	0.755	28.480	5
SISM+Ours	**0.402**	**0.808**	**19.442**	**3**
GXN [95]	0.307	0.746	26.731	3
GXN+Ours	**0.409**	**0.804**	**14.532**	**2**
PVSE [88]	0.324	0.759	20.463	3
PVSE+Ours	**0.465**	**0.830**	**10.421**	**2**
SAN [96]	0.337	0.777	23.143	2
SAN+Ours	**0.475**	**0.841**	**12.453**	**1**
VSRN [108]	0.403	0.701	15.323	**1**
VSRN+Ours	**0.481**	**0.802**	**10.332**	**1**
PCME [109]	0.379	0.735	20.632	3
PCME+Ours	**0.474**	**0.841**	**10.452**	**2**
SGRAF [110]	0.401	0.802	14.301	3
SGRAF+Ours	**0.451**	**0.842**	**7.429**	**2**
CGMN [111]	0.410	0.811	14.294	3
CGMN+Ours	**0.464**	**0.836**	**10.344**	**2**
IMC [112]	0.306	0.617	27.537	6
IMC+Ours	**0.396**	**0.791**	**16.001**	**4**

を表す．表 6.6 および表 6.7 より，提案手法を導入した際のマルチメディア情報検索手法の検索精度が元となるマルチメディア情報検索手法の検索精度よりも高いことが確認できる．以上より，提案手法を導入することの有効性を確認した．

表 6.7: 偏りのある評価用データセットを対象とした際の想起のしやすさを考慮した質問応答に基づく対話型マルチメディア情報検索と初期検索手法との比較.

	R@1	R@10	平均順位	中央順位
UVS [35]	0.161	0.471	57.684	12
UVS+Ours	**0.296**	**0.688**	**32.753**	**8**
RRF-Net [87]	0.314	0.635	19.324	8
RRF-Net+Ours	**0.396**	**0.705**	**16.556**	**5**
CMPM [39]	0.326	0.687	18.567	6
CMPM+Ours	**0.406**	**0.756**	**14.132**	**4**
SISM [94]	0.335	0.716	15.126	4
SISM+Ours	**0.398**	**0.785**	**12.453**	**2**
GXN [95]	0.332	0.741	15.441	4
GXN+Ours	**0.401**	**0.788**	**12.421**	**3**
PVSE [88]	0.343	0.761	15.810	2
PVSE+Ours	**0.405**	**0.833**	**11.538**	**1**
SAN [96]	0.347	0.768	15.110	2
SAN+Ours	**0.436**	**0.803**	**12.453**	**1**
VSRN [108]	0.379	0.793	14.332	**2**
VSRN+Ours	**0.458**	**0.843**	**11.334**	**2**
PCME [109]	0.349	0.751	17.422	3
PCME+Ours	**0.427**	**0.828**	**12.453**	**2**
SGRAF [110]	0.375	0.788	17.112	3
SGRAF+Ours	**0.433**	**0.824**	**11.421**	**2**
CGMN [111]	0.336	0.740	13.711	**2**
CGMN+Ours	**0.446**	**0.841**	**11.581**	**2**
IMC [112]	0.329	0.709	16.472	5
IMC+Ours	**0.411**	**0.760**	**14.261**	**3**

6.3.5　実験: 再検索手法との比較

6.3.4 節では，初期検索手法と比較することで，提案手法が初期検索手法の検索精度を向上させることが可能であることを確認した．次に，本節では，他の再検索手法と検索精度を比較することで，提案手法が従来の再検索手法 [103–107] よりも高精度に初期の検索結果を改善することが可能であることを示す．本実験

表 6.8: 偏りのない評価用データセットを対象とした際の定型文を用いた質問応答に基づく対話型マルチメディア情報検索手法と再検索手法との比較.

	R@1	R@10	平均順位	中央順位
BL [108]	0.403	0.701	15.323	**1**
NN + BQS [103]	0.387	0.701	17.521	2
SVM [104]	0.400	0.693	18.458	2
EMR [105]	0.397	0.691	16.532	2
PRF [106]	0.403	0.704	15.441	2
RFNet [107]	0.411	0.705	15.661	2
Ours	**0.481**	**0.802**	**10.332**	**1**

表 6.9: 偏りのある評価用データセットを対象とした際の想起のしやすさを考慮した質問応答に基づく対話型マルチメディア情報検索と再検索手法との比較.

	R@1	R@10	平均順位	中央順位
BL [108]	0.379	0.793	14.332	**2**
NN + BQS [103]	0.372	0.783	15.329	**2**
SVM [104]	0.170	0.388	162.734	28
EMR [105]	0.373	0.785	15.442	3
PRF [106]	0.370	0.781	15.128	3
RFNet [107]	0.112	0.387	132.14	26
Ours	**0.458**	**0.843**	**11.334**	**2**

では，文献 [108] により算出された検索結果を初期検索結果として利用した．ここで，従来の再検索手法のハイパーパラメータは従来手法の検索精度が最大となる値を採用した.

実験結果を表 6.8 および表 6.9 に示す．表 6.8 および表 6.9 より，提案手法の評価値が他の再検索手法の評価値よりも高い値であることが分かる．以上の実験結果より，提案手法の有効性を確認した.

表 6.10: 比較手法および提案手法により生成された質問文の想起のしやすさ．

	想起のしやすさ
比較手法	0.068
提案手法	**0.266**

6.3.6　実験: 生成された質問文の想起のしやすさに関する検証

本実験では，想起しやすい質問文を生成可能であるかどうかを定量的に検証する．本実験のテスト用データとして，Microsoft Common Objects in Context (MSCOCO) [85] データセットから 5,000 枚の画像を用いた．また，想起のしやすさに関する評価においては，画像内の物体の大きさと想起のしやすさには相関があるという知見 [114] に基づき，検索目的画像における物体の大きさを評価指標として用いた．ただし，本評価では質問応答に利用された物体の大きさを評価した．比較手法として，**6.2** 節にて提案している手法を用いた．

表 6.10 に実験結果を示す．表 6.10 より提案手法により提示された質問文が比較手法により提示された質問文よりも想起しやすいことが確認できる．これらの結果から，提案手法は想起のしやすい質問文を生成可能であることを確認した．

6.4　まとめ

本章では，定型文を用いた質問応答に基づく対話型マルチメディア情報検索を提案した．提案手法により，ユーザの与えるクエリテキストに十分な情報が含まれず，検索候補中にクエリテキストに該当するマルチメディアコンテンツが複数存在する場合にも，目的のマルチメディアコンテンツを高精度に検索することが可能であることを確認した．

第7章　質問応答モデルに基づく対話型マルチメディア情報検索

7.1　はじめに

6章では，定型文を用いた質問応答に基づく対話型マルチメディア情報検索を提案した．また，6章で提案した手法により，検索候補中にクエリテキストに該当するマルチメディアコンテンツが複数存在する場合にも，目的のマルチメディアコンテンツを高精度に検索することが可能あることを確認した．しかしながら，6章で提案した手法は，物体の有無に関する定型的な質問応答しか行うことが出来ず，物体の有無以外の情報が検索候補を絞り込むために有効である際に目的のマルチメディアコンテンツを効果的に絞り込むことが困難であるという課題が存在した．

本章では，物体以外の多様な情報に着目した質問文を生成することが可能なVisual Question Generation(VQG) モデルに基づく対話型マルチメディア情報検索手法を提案する．提案手法では，画像に関連する質問文を生成可能なVQGモデルを検索候補を絞り込むことが可能な質問文を生成するように再学習する．また，学習されたVQGモデルを用いて，検索候補を絞り込むことが可能な質問文をユーザに問いかけることで，クエリテキストに不足していた情報を補完し，再検索を行う．以上により，クエリテキストに十分な情報が含まれていない場合にも，高精度に目的のマルチメディアコンテンツを検索することが可能なマルチメディア情報検索を実現する．

7.2　VQG モデルに基づく対話型マルチメディア情報検索手法

本章では、提案手法である，VQG モデルに基づく対話型マルチメディア情報検索手法について説明する．提案手法の概要図を図 7.1 に示す．以降，**7.2.1** 節，**7.2.2** 節および **7.2.3** 節でそれぞれ識別器の事前学習，提案手法の学習および学習後の検索について説明する．

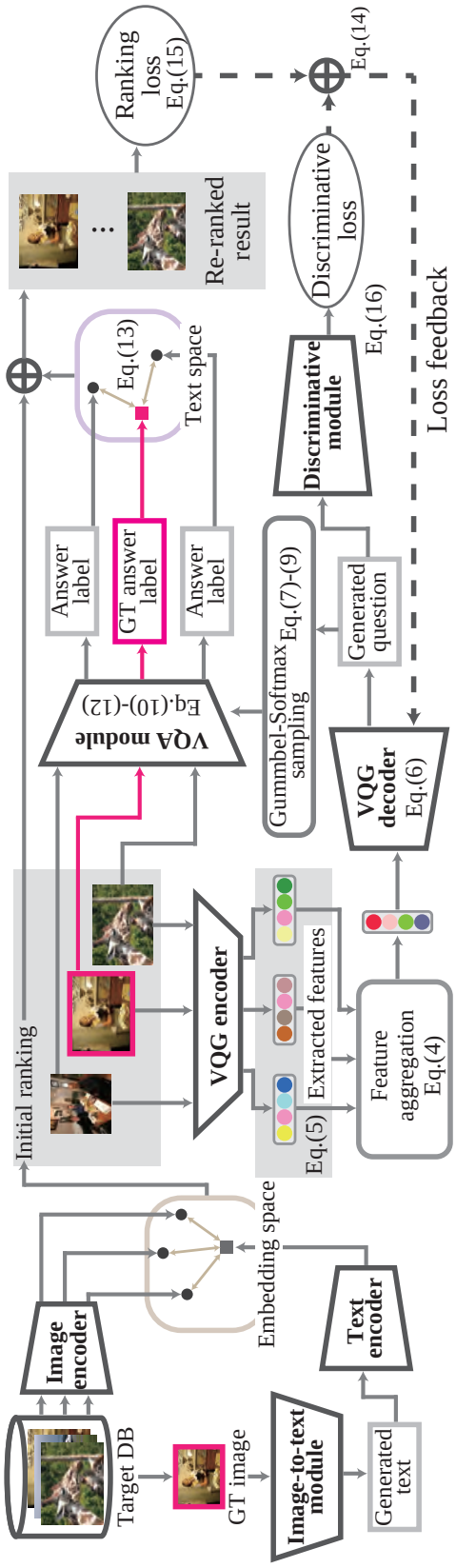


図 7.1: 質問応答モデルに基づく対話型マルチメディア情報検索手法の概要図.

7.2.1　識別器の事前学習

本節では，識別器の事前学習について説明する．提案手法では，はじめに，VQA 用のデータセットに含まれる質問文 $\mathrm{Q}_i^{\mathrm{train}}$ ($i = 1, \ldots, I^{\mathrm{train}}$; I^{train} は VQA データの総質問数) に対して，単語の順序の変更，単語の消去および単語の複製のいずれかを無作為に行うことで，$\hat{\mathrm{Q}}_i^{\mathrm{train}}$ を生成する．その後，以下により算出される損失 $L_{\mathrm{D}}^{\mathrm{pre}}$ が最小となるように識別器 $D(\cdot)$ を学習する．

$$L_{\mathrm{D}}^{\mathrm{pre}} = -\frac{1}{I^{\mathrm{train}}} \sum_i (\log D(\mathrm{Q}_i^{\mathrm{train}}) + \log(1 - D(\hat{\mathrm{Q}}_i^{\mathrm{train}}))) \tag{7.1}$$

以上により，学習された識別器 $D^*(\cdot)$ は入力された質問文の文法構造が正確であるかどうかを判別可能となる．

7.2.2　VQG モデルの finetuning

本節では，VQG モデルの finetuning について説明する．具体的に，提案手法では検索候補 I_n ($n = 1, \ldots, N$; N は検索候補の総画像枚数)) を利用して事前学習済み VQG モデルの finetuning を行う．

はじめに，検索候補 I_n を学習済みの Image captioning モデル [58] に入力することで各検索候補を説明する文 T_n を生成する．その後，従来のマルチメディア情報検索手法 [88] を用いて，T_n と I_m ($m = 1, \ldots, N$) の類似度 $s_{n,m}$ を算出する．ここで，検索候補 I_m を $s_{n,m}$ が降順となるように並び替えた際の k 位の画像の添字を $r_{n,k}$ ($k = 1, \ldots, N$) と定義する．続いて，事前学習済み VQG モデルの特徴量抽出器 $\mathrm{VQG}_{\mathrm{en}}(\cdot)$ を用いて検索順位を考慮した特徴量 $\boldsymbol{v}_n$ を算出する．

$$\boldsymbol{v}_n = \sum_k \alpha^{k-1} \boldsymbol{v}_{n,k} \tag{7.2}$$

$$\boldsymbol{v}_{n,k} = \mathrm{VQG}_{\mathrm{en}}(\mathrm{I}_{r_{n,k}}) \tag{7.3}$$

ただし，α は検索順位ごとの重要度を調節するハイパーパラメータである．以上により算出された特徴量 $\boldsymbol{v}_n$ を事前学習済み VQG モデルの質問文生成器 $\mathrm{VQG_{de}}(\cdot)$ に入力することで，質問文 Q_n を生成する．

$$\mathrm{Q}_n = \mathrm{VQG_{de}}(\boldsymbol{v}_n) \tag{7.4}$$

また，学習済み VQA モデル $\mathrm{VQA}(\cdot)$ を用いて，質問文 Q_n と検索候補 I_m を入力した際の回答 $\mathrm{A}_{n,m}$ を算出する．

$$\mathrm{A}_{n,m} = \mathrm{VQA}(\mathrm{Q}_n, \mathrm{I}_m) \tag{7.5}$$

続いて，以上により算出された $\mathrm{A}_{n,m}$ および類似度 $s_{n,m}$ を基に再順位付けのための類似度 $\hat{s}_{n,m}$ を以下の通り算出する．

$$\hat{s}_{n,m} = \beta s_{n,m} + (1-\beta)\mathrm{sim^{txt}}(\mathrm{A}_{n,n}, \mathrm{A}_{n,m}) \tag{7.6}$$

ここで，$\mathrm{sim^{txt}}(\cdot)$ および β はそれぞれテキスト類似度算出関数および初期検索と再検索のバランスを調整するハイパーパラメータを表す．最後に，以上により算出された類似度 $\hat{s}_{n,m}$，質問文 Q_n および学習済み識別器 $D^*(\cdot)$ から損失 L を計算する．

$$L = \frac{1}{(N(M-1))} L_{\mathrm{rank}} + \frac{1}{N} L_{\mathrm{D}} \tag{7.7}$$

$$L_{\mathrm{rank}} = \sum_n \sum_m \begin{cases} \max\{0, \gamma - \hat{s}_{n,n} + \hat{s}_{n,m}\} & (n \neq m) \\ 0 & (n = m) \end{cases} \tag{7.8}$$

$$L_{\mathrm{D}} = -\sum_n \log D^*(\mathrm{Q}_n) \tag{7.9}$$

ここで，γ は $\hat{s}_{n,n}$ と $\hat{s}_{n,m}$ の差分をどの程度許容するかを調整するハイパーパラメータである．以上により算出される損失 L が最小となるように $\mathrm{VQG_{de}}(\cdot)$ を finetuning

することで，$\mathrm{VQG}_{\mathrm{de}}(\cdot)$ は文法構造が正確でありながらも，検索順位が向上するような質問文を生成可能になると考えられる．

7.2.3　提案手法における学習後の検索

本節では，学習後の VQG モデルを用いた再検索について説明する．はじめに，従来のマルチメディア情報検索手法を用いて，クエリテキスト T と I_m ($m = 1, \ldots, N$) の類似度 s_m を算出する．ここで，検索候補 I_m を s_m が降順となるように並び替えた際の k 位の画像の添字を r_k ($k = 1, \ldots, N$) と定義する．続いて，事前学習済み VQG モデルの特徴量抽出器 $\mathrm{VQG}_{\mathrm{en}}(\cdot)$ を用いて検索順位を考慮した特徴量 $\boldsymbol{v}$ を算出する．

$$\boldsymbol{v} = \sum_k \alpha^{k-1} \boldsymbol{v}_k \tag{7.10}$$

$$\boldsymbol{v}_k = \mathrm{VQG}_{\mathrm{en}}(\mathrm{I}_{r_k}) \tag{7.11}$$

ただし，α は検索順位ごとの重要度を調節するハイパーパラメータである．算出された特徴量 $\boldsymbol{v}$ を事前学習済み VQG モデルの質問文生成器 $\mathrm{VQG}_{\mathrm{de}}(\cdot)$ に入力することで，質問文 Q を生成する．

$$\mathrm{Q} = \mathrm{VQG}_{\mathrm{de}}(\boldsymbol{v}) \tag{7.12}$$

以上により生成された質問文 Q をユーザに提示することで，ユーザからの回答 $\mathrm{A}^{\mathrm{user}}$ を受け取る．また，学習済み VQA モデル $\mathrm{VQA}(\cdot)$ を用いて，質問文 Q と検索候補 I_m を入力した際の回答 A_m を算出する．

$$\mathrm{A}_m = \mathrm{VQA}(\mathrm{Q}, \mathrm{I}_m) \tag{7.13}$$

続いて，以上により算出された A_m および類似度 s_m を基に再順位付けのための類似度 $\hat{s}_m$ を以下の通り算出する．

$$\hat{s}_m = \beta s_{n,m} + (1-\beta)\mathrm{sim}^{\mathrm{txt}}(\mathrm{A}^{\mathrm{user}}, \mathrm{A}_m) \tag{7.14}$$

最後に，類似度 $\hat{s}_n^{\mathrm{test}}$ が降順となるように検索順位を再決定する．

7.2.4　実験: 初期検索手法との比較

本実験では，従来の初期検索手法の検索精度と提案手法の検索精度を比較することで提案手法の定量的な有効性を検証する．また，提案手法における再学習および損失 $L^{\mathcal{D}}$ の有効性を検証するための実験を行う．

本実験では VQG モデル [115] および VQA モデル [73] の事前学習および学習に文献 [74] より提供されるデータセットを用いた．評価用のデータセットには文献 [85] より提供されるデータセットおよび文献 [99] から 5,000 枚および 32,433 枚を用いた．ただし，各画像に対して付与されている画像を説明するテキストをクエリとして利用した．ここで，提案手法において再検索を行うためには，ユーザからの回答 ($\mathrm{A}^{\mathrm{user}}$) を用意する必要がある．しかしながら，VQG により生成される質問文は多種多様であり，人手で全ての質問文に対応する回答を用意することは困難である．そこで，本実験では学習済みの VQA モデルをユーザであると仮定して実験を行う．ただし，実験の際に用いる VQA モデルには学習時とは異なる VQA モデルを利用している．また，α，β および γ にはそれぞれ 0.99，0.3 および 0.15 を用いた．さらに，初期検索結果を算出するためのマルチメディア情報検索手法として SISM [94]，GXN [95]，PVSE [88]，SAN [96]，VSRN [108]，PCME [109]，SDE [42]，CLIP [40] および BLIP [41] を用いた．また，本実験では，提案手法の再学習および損失 $L^{\mathcal{D}}$ の導入の有効性を検証するために，比較手法として再学習行わない場合の提案手法 (Ours w/o opt) および損失 $L^{\mathcal{D}}$ を導入しない場合の提案手法 (Ours w/o $L^{\mathcal{D}}$) を用いた．評価指標としては，従来手法に従

い，次式により算出される Recall@k(R@k) を用いた．

$$\text{Recall@}k = \frac{w_k}{J} \tag{7.15}$$

ここで，w_k は k 位以内に画像が正しく検索されたクエリの個数を示す．ただし，今回の実験ではクエリが付与されていた画像が検索された場合を正しく検索されたと定義した．また，J はクエリの総数を示す．

表 7.1 および表 7.2 に実験結果を示す．また，図 7.2 に検索結果のサンプルを示す．表 6.1 および表 7.2 において，PCME+Ours は PCME [109] に対して提案手法を用いることで再検索を行った際の実験結果を示す．表 6.1 および表 6.2 より，提案手法による再検索を行った際の評価値が基となる初期検索手法の検索精度を上回っていることが確認できる．以上より，提案手法を用いて順位を再決定することの有効性を確認した．また，提案手法による再検索が，Ours w/o opt および Ours w/o $L^{\mathcal{D}}$ の評価値を上回っていることが確認できる．また，提案手法においては，NVIDIA GeForce RTX 2080 Ti GPU および Intel Core i9-10980XE CPU を用いた際に，MSCOCO データセットおよび Visual Genome データセットそれぞれで，学習が収束するまでに 1 時間および 6 時間必要となる．SDE [42] 等の学習に必要となる時間が同様な環境で 40 時間および 50 時間程度であることを考慮すると，提案手法は比較的短時間で学習が収束することが分かる．以上より，提案手法による再学習および損失 $L^{\mathcal{D}}$ を導入することの有効性を確認した．

表 7.1: MSCOCO データセットを用いた際の質問応答モデルに基づく対話型マルチメディア情報検索手法と初期検索手法との比較.

	R@1	R@10	Mean	Med
SISM [94]	0.301	0.755	28.480	5
SISM+Ours w/o opt	0.354	0.819	20.115	3
SISM+Ours w/o $L^{\mathcal{D}}$	0.420	0.833	14.682	**2**
SISM+Ours	**0.433**	**0.849**	**12.011**	**2**
GXN [95]	0.307	0.746	26.731	3
GXN+Ours w/o opt	0.374	0.792	19.931	3
GXN+Ours w/o $L^{\mathcal{D}}$	0.432	0.827	11.042	**2**
GXN+Ours	**0.455**	**0.830**	**10.615**	**2**
PVSE [88]	0.324	0.759	20.463	3
PVSE+Ours w/o opt	0.393	0.809	15.195	2
PVSE+Ours w/o $L^{\mathcal{D}}$	0.479	0.860	9.771	**1**
PVSE+Ours	**0.481**	**0.864**	**9.515**	**1**
SAN [96]	0.337	0.777	23.143	2
SAN+Ours w/o opt	0.397	0.814	14.293	2
SAN+Ours w/o $L^{\mathcal{D}}$	0.452	0.836	12.663	2
SAN+Ours	**0.499**	**0.865**	**9.193**	**1**
VSRN [108]	0.403	0.701	15.323	**1**
VSRN+Ours w/o opt	0.455	0.825	9.995	**1**
VSRN+Ours w/o $L^{\mathcal{D}}$	0.503	0.854	7.524	**1**
VSRN+Ours	**0.521**	**0.868**	**6.777**	**1**
PCME [109]	0.352	0.765	25.322	3
PCME+Ours w/o opt	0.376	0.786	20.421	2
PCME+Ours w/o $L^{\mathcal{D}}$	0.413	0.823	13.422	2
PCME+Ours	**0.441**	**0.851**	**9.551**	**1**
SDE [42]	0.379	0.735	20.632	3
SDE+Ours w/o opt	0.389	0.800	14.158	2
SDE+Ours w/o $L^{\mathcal{D}}$	0.427	0.837	11.223	2
SDE+Ours	**0.475**	**0.858**	**8.339**	**1**
CLIP [40]	0.378	0.722	26.421	3
CLIP+Ours w/o opt	0.392	0.764	21.011	2
CLIP+Ours w/o $L^{\mathcal{D}}$	0.402	0.814	14.502	2
CLIP+Ours	**0.436**	**0.846**	**9.744**	**1**
BLIP [41]	0.402	0.753	18.773	2
BLIP+Ours w/o opt	0.423	0.792	16.532	2
BLIP+Ours w/o $L^{\mathcal{D}}$	0.498	0.841	12.331	1
BLIP+Ours	**0.531**	**0.867**	**8.631**	**1**

表 7.2: Visual Genome データセットを用いた際の質問応答モデルに基づく対話型マルチメディア情報検索手法と初期検索手法との比較.

	R@1	R@10	Mean	Med
SISM [94]	0.0275	0.121	3266.327	266
SISM+Ours w/o opt	0.0398	0.170	2901.392	212
SISM+Ours w/o $L^{\mathcal{D}}$	0.0951	0.201	2503.441	201
SISM+Ours	**0.110**	**0.218**	**2478.531**	**171**
GXN [95]	0.0278	0.125	3185.392	271
GXN+Ours w/o opt	0.0409	0.174	2885.332	201
GXN+Ours w/o $L^{\mathcal{D}}$	0.110	0.219	2603.225	183
GXN+Ours	**0.118**	**0.225**	**2500.032**	**165**
PVSE [88]	0.0280	0.129	2943.513	258
PVSE+Ours w/o opt	0.0431	0.179	2742.562	190
PVSE+Ours w/o $L^{\mathcal{D}}$	0.0985	0.221	2504.593	176
PVSE+Ours	**0.104**	**0.234**	**2485.331**	**163**
SAN [96]	0.0286	0.130	2861.242	248
SAN+Ours w/o opt	0.0442	0.180	27694.322	185
SAN+Ours w/o $L^{\mathcal{D}}$	0.114	0.225	2634.507	164
SAN+Ours	**0.122**	**0.242**	**2433.492**	**153**
VSRN [108]	0.0390	0.130	2861.242	248
VSRN+Ours w/o opt	0.0472	0.185	2800.293	176
VSRN+Ours w/o $L^{\mathcal{D}}$	0.120	0.233	2421.063	152
VSRN+Ours	**0.124**	**0.248**	**2394.391**	**147**
PCME [109]	0.0341	0.131	2822.341	253
PCME+Ours w/o opt	0.0432	0.162	2704.321	221
PCME+Ours w/o $L^{\mathcal{D}}$	0.0823	0.201	2563.432	181
PCME+Ours	**0.102**	**0.229**	**2499.32**	**164**
SDE [42]	0.0304	0.142	2755.495	249
SDE+Ours w/o opt	0.0451	0.182	2694.291	182
SDE+Ours w/o $L^{\mathcal{D}}$	0.0895	0.209	2554.322	167
SDE+Ours	**0.115**	**0.236**	**2423.391**	**154**
CLIP [40]	0.0405	0.151	2742.491	231
CLIP+Ours w/o opt	0.0574	0.176	2698.752	215
CLIP+Ours w/o $L^{\mathcal{D}}$	0.0968	0.205	2477.231	175
CLIP+Ours	**0.112**	**0.247**	**2313.441**	**146**
BLIP [41]	0.0454	0.173	2652.311	227
BLIP+Ours w/o opt	0.0603	0.181	2534.123	211
BLIP+Ours w/o $L^{\mathcal{D}}$	0.0994	0.212	2405.123	165
BLIP+Ours	**0.121**	**0.264**	**2223.486**	**128**

表 7.3: MSCOCO データセットを用いた際の再検索手法との比較.

	R@1	R@10	Mean	Med
BLIP(Baseline)	0.402	0.753	18.773	1
NN + BQS [103]	0.401	0.747	20.052	2
SVM [104]	0.379	0.721	39.701	3
EMR [105]	0.401	0.743	20.534	2
PRF [106]	0.397	0.739	19.110	2
RFNet [107]	0.387	0.750	35.435	3
Ours	**0.531**	**0.867**	**8.631**	**1**

7.2.5　実験: 再検索手法との比較

7.2.4 節では，初期検索手法と比較することで，提案手法が初期検索手法の検索精度を向上させることが可能であることを確認した．次に，本節では，他の再検索手法と検索精度を比較することで，提案手法が従来の再検索手法 (NN+BQS [103], SVM [104], EMR [105], PRF [106], RFNet [107]) よりも高精度に初期の検索結果を改善することが可能であることを示す．本実験では，文献 [41] により算出された検索結果を初期検索結果として利用した．ここで，従来の再検索手法のハイパーパラメータは従来手法の検索精度が最大となる値を採用した．

実験結果を表 7.3 および表 7.4 に示す．表 7.3 および表 7.4 より，提案手法の評価値が他の再検索手法の評価値よりも高い値であることが分かる．以上の実験結果より，提案手法の有効性を確認した．

表 7.4: Visual Genome データセットを用いた際の質問応答モデルに基づく対話型マルチメディア情報検索手法と再検索手法との比較.

	R@1	R@10	Mean	Med
BLIP(Baseline)	0.0454	0.173	2652.311	227
NN + BQS [103]	0.0455	0.173	2652.311	235
SVM [104]	0.0449	0.154	4109.735	610
EMR [105]	0.0452	0.171	2699.920	225
PRF [106]	0.0449	0.170	2625.026	227
RFNet [107]	0.0454	0.167	4001.001	548
Ours	**0.121**	**0.264**	**2223.486**	**128**

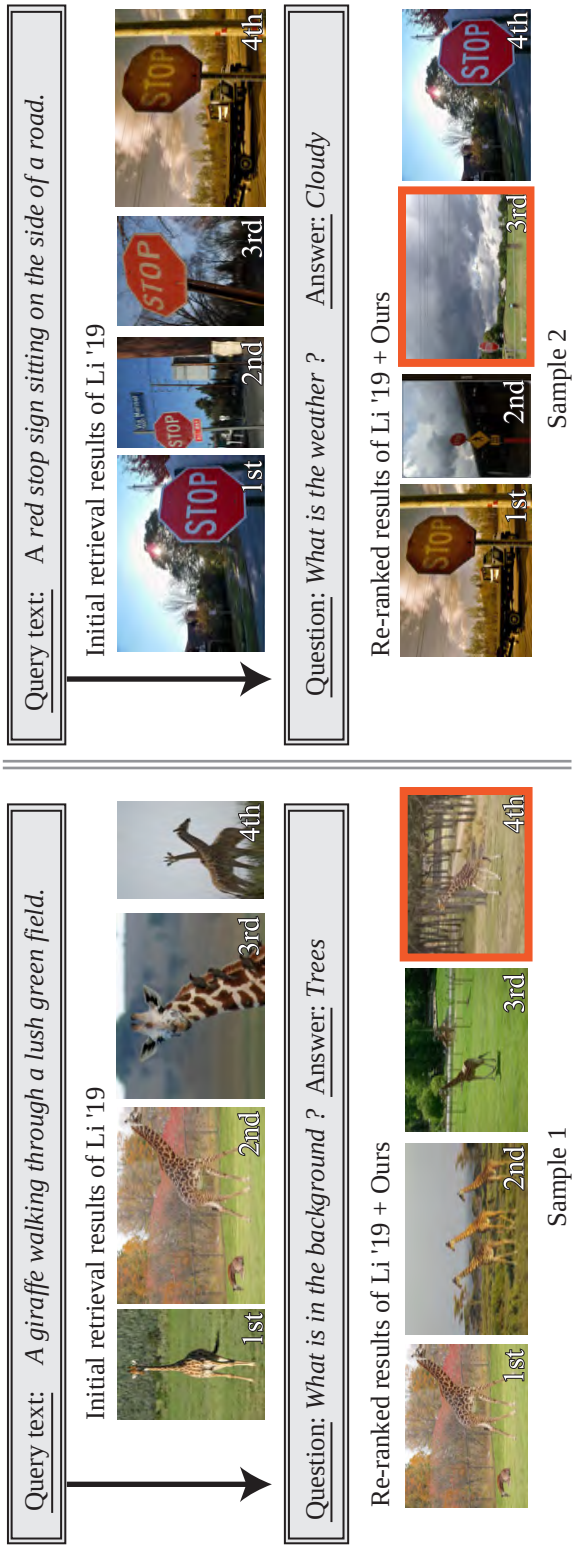


図 7.2: 質問応答モデルに基づく対話型マルチメディア情報検索手法による検索結果のサンプル.

7.3　まとめ

本章では，物体以外の多様な情報に着目した質問文を生成することが可能な Visual Question Generation(VQG) モデルに基づく対話型マルチメディア情報検索手法を提案を提案した．提案手法により，ユーザの与えるクエリテキストに十分な情報が含まれず，検索候補中にクエリテキストに該当するマルチメディアコンテンツが複数存在する場合にも，目的のマルチメディアコンテンツを高精度に検索することが可能であることを確認した．

第8章　結論

8.1　総括

本論文では，深層生成モデルによる情報共有を介した，対話型マルチメディア情報検索手法を新たに提案した．以降では，本論文の各章の概要を示す．

第 2 章では，関連研究としてマルチメディア情報検索や深層生成モデルに関する従来研究を紹介し，本論文で解決すべき課題を明らかにした．第 3 章では，画像生成モデルがマルチメディア情報検索に応用可能であることを検証するために，画像生成モデルに基づくマルチメディア情報検索手法を提案し，検索精度を確認した．第 4 章では，画像生成モデルに基づいて，クエリテキストの入力を補助することが可能な対話型検索手法を提案し，解釈の共有が検索の高精度化に貢献することを示した．第 5 章では，上記により提案した手法が，絵画等の多様なドメインに対しても応用可能であることを確認した．第 6 章では，質問応答に基づく対話型検索の枠組みが検索精度の向上に貢献可能であることを確認するために，定型文形式の質問文を生成可能な再検索手法について検討した．第 7 章では，第 6 章における検証結果を受けて，VQG に基づいて，非定型な質問文を生成することが可能な再検索手法を提案し，検索に必要な情報を質問応答の形式でユーザに問い合わせることの有効性を示した．以上により，深層生成モデルによる情報共有を介した，対話型マルチメディア情報検索手法を実現した．

8.2　今後の課題

本論文では，深層生成モデルによる情報共有を介した，対話型マルチメディア情報検索手法を実現した．本論文において提案した手法により，クエリテキスト

に不足している情報を補完することが可能となり，マルチメディア情報検索の検索精度は向上する．一方，本研究では，マルチメディアコンテンツを説明するテキストをクエリテキストとして用いるデータセットのみで検証が行われており，そのようなクエリテキストは実際のアプリケーションで入力されるクエリテキストとは乖離が存在する可能性が高い．そのため，今後は実際のアプリケーションにおいて収集されたデータを基にデータセットを構築し，検索精度を検証する必要があると考えられる．

また，本研究では，画像および映像の二種類のマルチメディアコンテンツのみに着目しており，三次元点群データや音声データなどの多様なマルチメディアコンテンツに対する有効性は検証していない．そのため，今後は，三次元点群データや音声データなどのより多様なマルチメディアコンテンツに対して，提案手法の有効性を検証する必要性があると考えられる．

以上の 2 点が本研究における今後の課題としてまとめられる．

謝辞

本研究は，著者が北海道大学および北海道大学大学院に在学した期間，約 5 年間にわたって行ったものである．

本研究に関して，研究遂行のみならず，終始御指導および御鞭撻を頂きました長谷山美紀教授に心より深謝申し上げます. 加えて，多くの国内・国外学会への参加，論文執筆，および教育活動等，様々な有益な機会を頂けたことに対しても，深くお礼申し上げます.

本論文をまとめるにあたり，副査をお引き受けいただいた北海道大学大学院情報科学研究院 言語メディア学研究室 荒木健治教授，北海道大学大学院情報科学研究院 メディア創生学研究室 坂本雄児教授，北海道大学大学院情報科学研究院 情報メディア環境学研究室 土橋宜典教授，ならびに北海道大学大学院情報科学研究院 メディアダイナミクス研究室 小川貴弘教授に深謝の意を表します.

本研究の遂行において，多大なる御助力を賜りました北海道大学 大学院情報科学研究院，藤後廉特任助教に心よりお礼申し上げます．研究活動のみならず進学や日々の学生生活に関するご助言もいただけたこと，ご多忙の中においても真剣に対応していただきましたこと，深謝申し上げます. また，北海道大学 大学院情報科学研究院，前田 圭介特任准教授，ならびに北海道大学 総合 IR 本部 斉藤直輝 助教に深謝申し上げます.

著者の研究室の所属期間中，多くの御協力を賜りました北海道大学大学院情報科学院情報科学専攻メディアネットワークコースメディアダイナミクス研究室の先輩，同輩ならびに後輩学生の皆様に感謝申し上げます．皆様と共に高め合うことで，私は約 5 年間で研究を成し遂げることができました.

最後に，自分の進路に対して，日々温かく見守りおよび支援してくださった家族に深謝申し上げ謝辞とさせていただきます.

参考文献

[1] M. Chen, S. Mao, and Y. Liu, “Big data: A survey,” Mobile Networks and Applications, vol.19, no.2, pp.171–209, 2014.

[2] R. Datta, D. Joshi, J. Li, and J.Z. Wang, “Image retrieval: Ideas, influences, and trends of the new age,” ACM Computing Surveys (Csur), vol.40, no.2, pp.1–60, 2008.

[3] S.R. Dubey, “A decade survey of content based image retrieval using deep learning,” IEEE Transactions on Circuits and Systems for Video Technology, vol.32, no.5, pp.2687–2704, 2021.

[4] L. Zhang and Y. Rui, “Image search—from thousands to billions in 20 years,” ACM Transactions on Multimedia Computing, Communications, and Applications, vol.9, no.1, 2013.

[5] T. Mei, Y. Rui, S. Li, and Q. Tian, “Multimedia search reranking: A literature survey,” ACM Computing Surveys, vol.46, no.3, pp.1–38, 2014.

[6] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, “A comprehensive survey on cross-modal retrieval,” arXiv:1607.06215, 2016.

[7] P. Kaur, H.S. Pannu, and A.K. Malhi, “Comparative analysis on cross-modal information retrieval: A review,” Computer Science Review, vol.39, p.100336, 2021.

[8] X. Li, J. Yang, and J. Ma, “Recent developments of content-based image retrieval (cbir),” Neurocomputing, vol.452, pp.675–689, 2021.

[9] D.C.G. Pedronette, J. Almeida, and R.d.S. Torres, “A scalable re-ranking method for content-based image retrieval,” Information Sciences, vol.265, pp.91–104, 2014.

[10] N. Burkart and M.F. Huber, “A survey on the explainability of supervised machine learning,” Journal of Artificial Intelligence Research, vol.70, pp.245–317, 2021.

[11] F.K. Došilović, M. Brčić, and N. Hlupić, “Explainable artificial intelligence: A survey,” 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO), pp.0210–0215, IEEE, 2018.

[12] M. Suzuki and Y. Matsuo, “A survey of multimodal deep generative models,” Advanced Robotics, vol.36, no.5-6, pp.261–278, 2022.

[13] G. Harshvardhan, M.K. Gourisaria, M. Pandey, and S.S. Rautaray, “A comprehensive survey and analysis of generative models in machine learning,” Computer Science Review, vol.38, p.100285, 2020.

[14] A. Oussidi and A. Elhassouny, “Deep generative models: Survey,” 2018 International conference on intelligent systems and computer vision (ISCV), pp.1–8, IEEE, 2018.

[15] W. Guo, Y. Zhang, J. Yang, and X. Yuan, “Re-attention for visual question answering,” IEEE Transactions on Image Processing, vol.30, pp.6730–6743, 2021.

[16] D. Putthividhy, H.T. Attias, and S.S. Nagarajan, “Topic regression multi-modal latent dirichlet allocation for image annotation,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.3408–3415, 2010.

[17] H. Hotelling, “Relations between two sets of variates,” in Breakthroughs in statistics, pp.162–190, 1992.

[18] Y. Jia, M. Salzmann, and T. Darrell, “Learning cross-modality similarity for multinomial data,” 2011 International Conference on Computer Vision, pp.2407–2414, IEEE, 2011.

[19] B. Bai, J. Weston, D. Grangier, R. Collobert, K. Sadamasa, Y. Qi, O. Chapelle, and K. Weinberger, “Learning to rank with (a lot of) word features,” Information retrieval, vol.13, pp.291–314, 2010.

[20] D. Grangier and S. Bengio, “A discriminative kernel-based approach to rank images from text queries,” IEEE transactions on pattern analysis and machine intelligence, vol.30, no.8, pp.1371–1384, 2008.

[21] X. Lu, F. Wu, S. Tang, Z. Zhang, X. He, and Y. Zhuang, “A low rank structural large margin method for cross-modal ranking,” Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp.433–442, 2013.

[22] F. Wu, X. Lu, Z. Zhang, S. Yan, Y. Rui, and Y. Zhuang, “Cross-media semantic representation via bi-directional learning to rank,” Proceedings of the 21st ACM international conference on Multimedia, pp.877–886, 2013.

[23] J. Weston, S. Bengio, and N. Usunier, “Wsabie: Scaling up to large vocabulary image annotation,” 2011.

[24] T. Yao, T. Mei, and C.W. Ngo, “Learning query and image similarities with ranking canonical correlation analysis,” Proceedings of the IEEE International Conference on Computer Vision, pp.28–36, 2015.

[25] R. Rosipal and N. Krämer, “Overview and recent advances in partial least squares,” International Statistical and Optimization Perspectives Workshop” Subspace, Latent Structure and Feature Selection”, pp.34–51, Springer, 2005.

[26] A. Sharma, A. Kumar, H. Daume, and D.W. Jacobs, “Generalized multiview analysis: A discriminative latent space,” 2012 IEEE conference on computer vision and pattern recognition, pp.2160–2167, IEEE, 2012.

[27] J.B. Tenenbaum and W.T. Freeman, “Separating style and content with bilinear models,” Neural computation, vol.12, no.6, pp.1247–1283, 2000.

[28] D. Li, N. Dimitrova, M. Li, and I.K. Sethi, “Multimedia content processing through cross-modal association,” Proceedings of the eleventh ACM international conference on Multimedia, pp.604–611, 2003.

[29] V. Mahadevan, C. Wong, J. Pereira, T. Liu, N. Vasconcelos, and L. Saul, “Maximum covariance unfolding: Manifold learning for bimodal data,” Advances in Neural Information Processing Systems, vol.24, 2011.

[30] X. Shi and P. Yu, “Dimensionality reduction on heterogeneous feature space,” 2012 IEEE 12th International Conference on Data Mining, pp.635–644, IEEE, 2012.

[31] F. Zhu, L. Shao, and M. Yu, “Cross-modality submodular dictionary learning for information retrieval,” Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp.1479–1488, 2014.

[32] D.M. Blei and M.I. Jordan, “Modeling annotated data,” Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp.127–134, 2003.

[33] L. Zheng, Y. Yang, and Q. Tian, “SIFT meets CNN : A decade survey of instance retrieval,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.40, no.5, pp.1224–1244, 2018.

[34] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, “A multimedia retrieval framework based on semi-supervised ranking and relevance feedback,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.34, no.4, pp.723–742, 2011.

[35] R. Kiros, R. Salakhutdinov, and R.S. Zemel, “Unifying visual-semantic embeddings with multimodal neural language models,” arXiv:1411.2539, 2014.

[36] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, “Order-embeddings of images and language,” Proceedings of the International Conference on Learning Representations, pp.1–12, 2016.

[37] J. Dong, X. Li, and C.G.M. Snoek, “Word2VisualVec: Image and video to sentence matching by visual feature prediction,” arXiv:1604.06838, 2016.

[38] F. Faghri, D.J. Fleet, J.R. Kiros, G.B. Toronto, and S. Fidler, “VSE ++ : Improving visual-semantic embeddings with hard negatives,” arXiv:1707.05612, 2017.

[39] Y. Zhang and H. Lu, “Deep cross-modal projection learning for image-text matching,” Proceedings of the IEEE European Conference on Computer Vision, pp.686–701, 2018.

[40] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, “Learning transferable visual models from natural language supervision,” International conference on machine learning, pp.8748–8763, PMLR, 2021.

[41] J. Li, D. Li, C. Xiong, and S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” International Conference on Machine Learning, pp.12888–12900, PMLR, 2022.

[42] D. Kim, N. Kim, and S. Kwak, “Improving cross-modal retrieval with set of diverse embeddings,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.23422–23431, 2023.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol.30, 2017.

[44] T. Wang, X. Xu, Y. Yang, A. Hanjalic, H.T. Shen, and J. Song, “Matching images and text with multi-modal tensor fusion and re-ranking,” Proceedings of the 27th ACM international conference on multimedia, pp.12–20, 2019.

[45] X. Yu, T. Chen, Y. Yang, M. Mugo, and Z. Wang, “Cross-modal person search: A coarse-to-fine framework using bi-directional text-image matching,” Proceedings of the IEEE International Conference on Computer Vision Workshops, pp.0–0, 2019.

[46] W. Wei, M. Jiang, X. Zhang, H. Liu, and C. Tian, “Boosting cross-modal retrieval with mvse++ and reciprocal neighbors,” IEEE Access, vol.8, pp.84642–84651, 2020.

[47] X. Guo, H. Wu, Y. Cheng, S. Rennie, G. Tesauro, and R. Feris, “Dialog-based interactive image retrieval,” Advances in neural information processing systems, pp.678–688, 2018.

[48] N. Vo, L. Jiang, C. Sun, K. Murphy, L.J. Li, L. Fei-Fei, and J. Hays, “Composing text and image for image retrieval-an empirical odyssey,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.6439–6448, 2019.

[49] F. Tan, P. Cascante-Bonilla, X. Guo, H. Wu, S. Feng, and V. Ordonez, “Drill-down: Interactive retrieval of complex scenes using natural language queries,” Advances in Neural Information Processing Systems, pp.2651–2661, 2019.

[50] R. Yan, A. Hauptmann, and R. Jin, “Multimedia search with pseudo-relevance feedback,” Image and Video Retrieval: Second International Conference, CIVR 2003 Urbana-Champaign, IL, USA, July 24–25, 2003 Proceedings 2, pp.238–247, Springer, 2003.

[51] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, “Learning object categories from google’s image search,” Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1, pp.1816–1823, IEEE, 2005.

[52] W.H. Hsu, L.S. Kennedy, and S.F. Chang, “Video search reranking via information bottleneck principle,” Proceedings of the 14th ACM international conference on multimedia, pp.35–44, 2006.

[53] W.H. Hsu, L.S. Kennedy, and S.F. Chang, “Video search reranking through random walk over document-level context graph,” Proceedings of the 15th ACM international conference on Multimedia, pp.971–980, 2007.

[54] Y. Jing and S. Baluja, “Pagerank for product image search,” Proceedings of the 17th international conference on World Wide Web, pp.307–316, 2008.

[55] K. Wnuk and S. Soatto, “Filtering internet image search results towards keyword based category recognition,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.1–8, IEEE, 2008.

[56] T.T. Nguyen-Dang, X.D. Thai, G.H. Vuong, V.S. Ho, M.T. Tran, V.T. Ninh, M.K. Pham, T.K. Le, and G. Healy, “Lifeinsight: An interactive lifelog retrieval system with comprehensive spatial insights and query assistance,” in Proceedings of the 6th Annual ACM Lifelog Search Challenge, pp.59–64, 2023.

[57] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.3156–3164, 2015.

[58] K. Xu, J.L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R.S. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” Proceedings of the IEEE International Conference on Machine Learning, pp.2048–2057, 2015.

[59] K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston, “Engaging image captioning via personality,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.12516–12526, 2019.

[60] T. Yao, Y. Pan, Y. Li, and T. Mei, “Hierarchy parsing for image captioning,” Proceedings of the IEEE International Conference on Computer Vision, pp.2621–2629, 2019.

[61] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei, “Pointing novel objects in image captioning,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.12497–12506, 2019.

[62] F. Zhan, Y. Yu, R. Wu, J. Zhang, S. Lu, L. Liu, A. Kortylewski, C. Theobalt, and E. Xing, “Multimodal image synthesis and editing: A survey and taxonomy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[63] E. Zavesky and S.F. Chang, “Cuzero: Embracing the frontier of interactive visual search for informed users,” Proceedings of the 1st ACM international conference on Multimedia information retrieval, pp.237–244, 2008.

[64] C. Snoek, K. Sande, O. Rooij, B. Huurnink, J. Uijlings, M.v. Liempt, M. Bugalhoy, I. Trancosoy, F. Yan, M. Tahir, *et al.*, “The mediamill trecvid 2009 semantic video search engine,” TRECVID workshop, 2009.

[65] J. Wang and X.S. Hua, “Interactive image search by color map,” ACM Transactions on Intelligent Systems and Technology (TIST), vol.3, no.1, pp.1–23, 2011.

[66] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis,” arXiv:1605.05396, 2016.

[67] Z. Zhang, Y. Xie, and L. Yang, “Photographic text-to-image synthesis with a hierarchically-nested adversarial network,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.6199–6208, 2018.

[68] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, “AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.1316–1324, 2018.

[69] T. Qiao, J. Zhang, D. Xu, and D. Tao, “MirrorGAN: Learning Text-to-image Generation by Redescription,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.1505–1514, 2019.

[70] M. Zhu, P. Pan, W. Chen, and Y. Yang, “DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.5802–5810, 2019.

[71] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y.J. Lee, “Gligen: Open-set grounded text-to-image generation,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.22511–22521, 2023.

[72] C. Zhang, C. Zhang, M. Zhang, and I.S. Kweon, “Text-to-image diffusion model in generative ai: A survey,” arXiv preprint arXiv:2303.07909, 2023.

[73] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C.L. Zitnick, and D. Parikh, "Vqa: Visual question answering," Proceedings of the IEEE International Conference on Computer Vision, pp.2425–2433, 2015.

[74] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[75] N. Vedd, Z. Wang, M. Rei, Y. Miao, and L. Specia, "Guiding visual question generation," arXiv preprint arXiv:2110.08226, 2021.

[76] Y. Srivastava, V. Murali, S.R. Dubey, and S. Mukherjee, "Visual question answering using deep learning: A survey and performance analysis," Computer Vision and Image Processing: 5th International Conference, CVIP 2020, Prayagraj, India, December 4-6, 2020, Revised Selected Papers, Part II 5, pp.75–86, Springer, 2021.

[77] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "Stack-GAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," Proceedings of the IEEE Conference on Computer Vision, pp.5907–5915, 2017.

[78] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-UCSD birds-200-2011 dataset," Technical Report CNS-TR-2011-001, 2011.

[79] S. Reed, Z. Akata, B. Schiele, and H. Lee, "Learning deep representations of fine-grained visual descriptions," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.49–58, 2016.

[80] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "ImageNet classification with deep convolutional neural networks," Proceedings of the Advances in Neural Information Processing Systems (NIPS), pp.1097–1105, 2012.

[81] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.2818–2826, 2015.

[82] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” Proceedings of the IEEE International Conference on Learning Representations, pp.1–14, 2015.

[83] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.770–778, 2016.

[84] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.248–255, 2009.

[85] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick, “Microsoft COCO: Common objects in context,” Proceedings of the IEEE European Conference on Computer Vision, pp.740–755, 2014.

[86] A. Rohrbach, M. Rohrbach, S. Tang, S. Joon Oh, and B. Schiele, “Generating descriptions with grounded and co-referenced people,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.4979–4989, 2017.

[87] Y. Liu, Y. Guo, E.M. Bakker, and M.S. Lew, “Learning a recurrent residual fusion network for multimodal matching,” Proceedings of the IEEE International Conference on Computer Vision, pp.4107–4116, 2017.

[88] Y. Song and M. Soleymani, “Polysemous visual-semantic embedding for cross-modal retrieval,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.1979–1978, 2019.

[89] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” arXiv:1908.10084, 2019.

[90] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F.A. Wichmann, and W. Brendel, “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness,” Proceeding of the IEEE International Conference on Learning Representations, pp.1–22, 2019.

[91] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, “A dataset for movie description,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.3202–3212, 2015.

[92] C.L. Zitnick and D. Parikh, “Bringing semantics into focus using visual abstraction,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.3009–3016, 2013.

[93] M. Stefanini, M. Cornia, L. Baraldi, M. Corsini, and R. Cucchiara, “Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain,” International Conference on Image Analysis and Processing, pp.729–740, 2019.

[94] Y. Huang, Q. Wu, C. Song, and L. Wang, “Learning semantic concepts and order for image and sentence matching,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.6163–6171, 2018.

[95] J. Gu, J. Cai, S.R. Joty, L. Niu, and G. Wang, “Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.7181–7189, 2018.

[96] Z. Ji, H. Wang, J. Han, and Y. Pang, “Saliency-guided attention network for image-sentence matching,” Proceedings of the IEEE International Conference on Computer Vision, pp.5754–5763, 2019.

[97] G. Huang, Z. Liu, L. Van Der Maaten, and K.Q. Weinberger, "Densely connected convolutional networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.4700–4708, 2017.

[98] J.Y. Zhu, T. Park, P. Isola, and A.A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," Proceedings of the IEEE international conference on computer vision, pp.2223–2232, 2017.

[99] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.J. Li, D.A. Shamma, *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," International journal of computer vision, vol.123, no.1, pp.32–73, 2017.

[100] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[101] C.K.C.G. Mikolov, T. and J. Dean, "Efficient estimation of word representations in vector space," arXiv:1301.3781, pp.1–12, 2013.

[102] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, and Y.D. Shen, "Dual-path convolutional image-text embedding with instance loss," ACM Transactions on Multimedia Computing, Communications, and Applications, 2020.

[103] G. Giacinto, "A nearest-neighbor approach to relevance feedback in content based image retrieval," Proceedings of the ACM International Conference on Image and Video Retrieval, pp.456–463, 2007.

[104] S. Liang and Z. Sun, "Sketch retrieval and relevance feedback with biased svm classification," Pattern Recognition Letters, vol.29, no.12, pp.1733–1741, 2008.

[105] B. Xu, J. Bu, C. Chen, C. Wang, D. Cai, and X. He, "Emr: A scalable graph-based ranking model for content-based image retrieval," IEEE Transactions on Knowledge and Data Engineering, vol.27, no.1, pp.102–114, 2013.

[106] W.C. Lin, Z.Y. Chen, S.W. Ke, C.F. Tsai, and W.Y. Lin, “The effect of low-level image features on pseudo relevance feedback,” Neurocomputing, vol.166, pp.26–37, 2015.

[107] L. Putzu, L. Piras, and G. Giacinto, “Convolutional neural networks for relevance feedback in content based image retrieval,” Multimedia Tools and Applications, vol.79, no.37, pp.26995–27021, 2020.

[108] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, “Visual semantic reasoning for image-text matching,” Proceedings of the IEEE International Conference on Computer Vision, pp.4654–4662, 2019.

[109] S. Chun, S.J. Oh, R.S. de Rezende, Y. Kalantidis, and D. Larlus, “Probabilistic embeddings for cross-modal retrieval,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.8415–8424, 2021.

[110] H. Diao, Y. Zhang, L. Ma, and H. Lu, “Similarity reasoning and filtration for image-text matching,” Proceedings of the AAAI Conference on Artificial Intelligence, pp.1218–1226, 2021.

[111] Y. Cheng, X. Zhu, J. Qian, F. Wen, and P. Liu, “Cross-modal graph matching network for image-text retrieval,” ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol.18, no.4, pp.1–23, 2022.

[112] J. Chen, L. Zhang, Q. Wang, C. Bai, and K. Kpalma, “Intra-modal constraint loss for image-text retrieval,” 2022 IEEE International Conference on Image Processing (ICIP), pp.4023–4027, IEEE, 2022.

[113] W. Bousselham, G. Thibault, L. Pagano, A. Machireddy, J. Gray, Y.H. Chang, and X. Song, “Efficient self-ensemble framework for semantic segmentation,” arXiv preprint arXiv:2111.13280, 2021.

[114] R. Dubey, J. Peterson, A. Khosla, M.H. Yang, and B. Ghanem, “What makes an object memorable?,” Proceedings of the IEEE International Conference on Computer Vision, pp.1089–1097, 2015.

[115] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende, “Generating natural questions about an image,” arXiv preprint arXiv:1603.06059, 2016.

著者の研究業績

(A) 学術論文

[**A-1**] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama, “Query is GAN: Scene retrieval with attentional text-to-image generative adversarial network,” IEEE Access, DOI: 10.1109/ACCESS.2019.2947409, vol. 7, pp. 153183-153193, 2019.

[**A-2**] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama, “Text-to-image GANbased scene retrieval and re-ranking considering word importance,” IEEE Access, DOI: 10.1109/ACCESS.2019.2952676, vol. 7, pp. 169920-169930, 2019.

[**A-3**] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama, “Enhancing cross-modal retrieval based on modality-specific and embedding spaces,” IEEE Access, DOI: 10.1109/ACCESS.2020.2995815, vol. 8, pp. 96777-96786, 2020.

[**A-4**] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama, “Domain adaptive crossmodal image retrieval via modality and domain translations,” IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, DOI: 10.1587/transfun. 2020IMP0011, vol. E104-A, no. 6, pp. 866-875, 2020.

[**A-5**] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama, “Interactive re-ranking via object entropy-guided question answering for cross-modal image retrieval,” ACM Transactions on Multimedia Computing, Communications, and Applications, DOI: 10.1145/3485042, vol.18, issue 3, 2021.

[**A-6**] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama, “Recallable question answering-based re-ranking considering semantic region for cross-modal retrieval,”IEEE Open Journal of Signal Processing, DOI: 10.1109/OJSP.2023.3238280, vol. 4, pp. 1-11, 2023.

[**A-7**] Huaying Zhang, Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama, “Cross-modal Image Retrieval Considering Semantic Relationships with many-to-many Correspondence Loss,” IEEE Access, DOI: 10.1109/ACCESS.2023.3239858, vol. 11, pp. 10675-10686, 2023.

[**A-8**] Rintaro Yanagi, Ren Togo, Keisuke Maeda, Takahiro Ogawa, Miki Haseyama, “ Material compound-property retrieval using electron microscope images for rubber material development,”IEEE Access, DOI: 10.1109/ACCESS.2023.3304341, vol. 11, pp. 88258-88264, 2023.

(B) 国際会議 (査読あり)

[**B-1**] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama, “ Image retrieval from vague description based on AttnGAN,” in Proceedings of the IEEE Global Conference on Consumer Electronics (IEEE GCCE), pp. 167-168, Nara, Japan, 2018.

[**B-2**] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama, “ Scene retrieval from multiple resolution generated images based on text-to-image GAN, ” in Proceedings of the IEEE International Symposium on Circuits and Systems (IEEE ISCAS), pp. 1-5, Sapporo, Japan, 2019.

[**B-3**] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama, “ Scene retrieval for video summarization based on text-to-image GAN, ” in Proceedings of the IEEE International Conference on Image Processing (IEEE ICIP), pp. 1825-1829, Taipei, Taiwan, 2019.

[**B-4**] R. Yanagi, R. Togo, T. Ogawa, and M. Haseyama, "Scene Retrieval Using Text-to-image GAN-based Visual Similarities and Image-to-text Model-based Textual Similarities," in *Proc. 2019 IEEE 8th Global Conference on Consumer Electronics (GCCE)*, Oct. 2019, pp. 13–14.

[**B-5**] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama, "Voice-input multimedia information retrieval system based on text-to-image GAN," in Proceedings of the IEEE Global Conference on Consumer Electronics (IEEE GCCE), pp. 967-968, Osaka, Japan, 2019.

[**B-6**] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama, " Scene retrieval using text-to- image GAN-based visual similarities and image-to-text model-based textual similarities, " in Proceedings of the IEEE Global Conference on Consumer Electronics (IEEE GCCE), pp. 13-14, Osaka, Japan, 2019.

[**B-7**] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama, " Image retrieval with data augmentation of sentence labels based on paraphrasing, " in Proceedings of the IEEE International Conference on Consumer Electronics – Taiwan (IEEE ICCE-TW), pp. 1-2, Taoyuan, Taiwan (Online Participation), 2020.

[**B-8**] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama, " Image retrieval with lingual and visual paraphrasing via generative models, " in Proceedings of the IEEE International Conference on Image Processing (IEEE ICIP), pp. 2431-2435, Abu Dhabi, United Arab Emirates (Online Participation), 2020.

[**B-9**] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama, " Interactive re-ranking for cross-modal retrieval based on object-wise question answering, " in Proceedings of the ACM International Conference on Multimedia in Asia (ACM MM Asia), pp. 1-7, Singapore (Online Participation), 2020.

[**B-10**] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama, " IR Questioner: QA-based interactive retrieval system, " in Proceedings of the ACM Interna-

tional Conference on Multimedia Retrieval (ACM ICMR), pp. 967-968, Taipei, Taiwan (Online Participation), 2021.

[**B-11**] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama, “Database-adaptive reranking for enhancing cross-modal image retrieval,” in Proceedings of the ACM International Conference on Multimedia (ACM MM), pp. 3816-3825, Chengdu, China (Online Participation), 2021.

[**B-12**] Masato Kawai, Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama, “Free-viewpoint sports video generation based on dynamic NeRF considering time series,” in Proceedings of the IEEE Global Conference on Consumer Electronics (IEEE GCCE), Osaka, Japan, pp. 408- 409, 2022.

[**B-13**] Huaying Zhang, Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama, “Cross-modal image retrieval considering semantic relationships with object information,” in Proceedings of the IEEE Global Conference on Consumer Electronics (IEEE GCCE), Osaka, Japan, pp. 775-776, 2022.

[**B-14**] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama, “Rubber material retrieval system using electron microscope images for rubber material development,” in Proceedings of the ACM International Conference on Multimedia Asia (ACM MM Asia), pp. 1-3, Tokyo, Japan, 2022.

[**B-15**] Huaying Zhang, Rintaro Yanagi, Ren Togo, Takahiro Ogawa, Miki Haseyama, “Parameter-efficient tuning of a pre-trained model via prompt learning in cross-modal retrieval” in Proceedings of the IEEE International Conference on Consumer Electronics – Taiwan (ICCE-TW), PingTung, Taiwan, pp. 811-812, 2023.

[**B-16**] Yuya Moroto*, Rintaro Yanagi*, Naoki Ogawa, Kyohei Kamikawa, Keigo Sakurai, Ren Togo, Keisuke Maeda, Takahiro Ogawa, Miki Haseyama, “Personalized content recommender system via non-verbal interaction using face

mesh and facial expression, ” in Proceedings of the ACM Multimedia (ACM MM), pp. 9399-9401, Ottawa, Canada, 2023. *: Equal contributions

(C) 国内学会 (査読なし)

[**C-1**] 柳凜太郎, 藤後廉, 小川貴弘, 長谷山美紀, “ AttnGAN を用いたシーン検索に関する検討ー再検索の導入による高精度化ー, ” 平成 30 年度電気・情報関係学会北海道支部連合大会, pp.12-13, 札幌市, 2018 年.

[**C-2**] 柳凜太郎, 藤後廉, 小川貴弘, 長谷山美紀, “ 敵対的生成ネットワークにより文から生成される画像の意味的評価に関する検討, ” イメージ・メディア・クオリティ研究会 (IMQ), pp.21-24, 札幌市, 2019 年.

[**C-3**] 柳凜太郎, 藤後廉, 小川貴弘, 長谷山美紀, “ 敵対的生成ネットワークに基づくドメイン適応可能な文をクエリとする画像・映像検索手法に関する検討, ” 第 22 回画像の認識・理解シンポジウム (MIRU), pp. 1-4, 大阪市, 2019 年.

[**C-4**] 柳凜太郎, 藤後廉, 小川貴弘, 長谷山美紀, “ 画像内の物体に着目した画像検索に関する検討-RetinaNet を用いた物体認識に基づく高精度化-, ” 映像情報メディア学会技術報告, vol.44, no. 6, pp. 377-381, 札幌市, 2020 年.

[**C-5**] 柳凜太郎, 藤後廉, 小川貴弘, 長谷山美紀, “ ゴム材料開発のための conditional StyleGAN に基づく配合量からの電子顕微鏡画像の生成に関する検討, ” 映像情報メディア学会技術報告, vol. 45, no. 4, pp. 171-175, 札幌市, 2021 年.

[**C-6**] 柳凜太郎, 藤後廉, 小川貴弘, 長谷山美紀, “ ゴム材料開発のための generative adversarial network に基づく配合量および物性値からの電子顕微鏡画像の生成に関する一検討, ” 映像情報メディア学会技術報告, vol. 46, no. 6, pp. 187-191, 札幌市, 2022 年.

[**C-7**] 張華瀛, 柳凜太郎, 藤後廉, 小川貴弘, 長谷山美紀, “ データベース特化型クロスモーダル画像検索のためのテキストプロンプトチューニングに関する

検討, ”映像情報メディア学会技術報告, vol. 47, no. 6, pp. 217-221, 札幌市, 2022 年.

[**C-8**] 河合雅斗, 柳凜太郎, 藤後廉, 小川貴弘, 長谷山美紀, “ フーリエ振幅成分を考慮した neural radiance fields のノンリファレンス評価指標に関する検討, ” 令和 4 年度電気・情報関係学会北海道支部連合大会, pp. 192-193, オンライン, 2022 年.

[**C-9**] 柳凜太郎, 藤後廉, 小川貴弘, 長谷山美紀, “Database-adaptive transfer learning for question answering-based re-ranking in cross-modal retrieval, ” 第 25 回画像の認識・理解シンポジウム (MIRU), pp. 1-4, 浜松市, 2023 年.

[**C-10**] 張華瀛, 柳凜太郎, 藤後廉, 小川貴弘, 長谷山美紀, “ 画像とテキストの関係性を考慮した textual inversion に基づく zero-shot composed image retrieval 手法に関する検討, ” 令和 5 年度電気・情報関係学会北海道支部連合大会, pp. 150-151, 函館市, 2022 年.

(D) 受賞

[**D-1**] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, and Miki Haseyama, “Outstanding Prize IEEE GCCE2019 Excellent Demo! Award,” 2019.

[**D-2**] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, and Miki Haseyama, “ACM MM Asia 2020 Best Paper Runner-Up Award,” 2020.

[**D-3**] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, and Miki Haseyama, “The 2020 IEEE Sapporo Section Best Paper Award,” 2020.

[**D-4**] Rintaro Yanagi, Ren Togo, Takahiro Ogawa, and Miki Haseyama, “The 2021 IEEE Sapporo Section Encouragement Award,” 2021.

[**D-5**] Rintaro Yanagi, “MIRU2022 学生奨励賞,” 2022.

[**D-6**] Masato Kawai, Rintaro Yanagi, Ren Togo, Takahiro Ogawa, and Miki Haseyama, “IEEE GCCE 2022 Excellent Poster Award Silver Prize,” 2022.

[**D-7**] Masato Kawai, Rintaro Yanagi, Ren Togo, Takahiro Ogawa, and Miki Haseyama, “The 2022 IEEE Sapporo Section Student Paper Contest Encouraging Prize,” 2022.