観測視点に着目した画像認識性能の評価及び改善に関する研究

2021年 3月

QIU　YUE

観測視点に着目した画像認識性能の評価及び改善に関する研究

QIU　YUE

システム情報工学研究科
筑波大学

2021年 3月

目 次

第 1 章　序論 **1**
1.1 研究背景 1
1.2 研究目的 1
1.3 論文の構成 2

第 2 章　視点依存性評価および多視点尤度投票型物体認識 **3**
2.1 概要 3
2.2 Deep Neural Network（DNN）を用いた画像認識技術 4
2.2.1 大規模画像認識用データセット 4
2.2.1.1 ImageNet 4
2.2.1.2 Pascal VOC 4
2.2.1.3 MSCOCO 4
2.2.2 物体認識手法 5
2.2.2.1 AlexNet 5
2.2.2.2 VGG 6
2.2.2.3 ResNet 7
2.2.3 物体検出手法 8
2.2.3.1 Faster R-CNN 8
2.2.3.2 YOLO 9
2.3 データセットの視点バイアス 9
2.4 一般物体認識用大規模画像データセットにおける視点依存性に関する考察及び定量化 11
2.4.1 概要 11
2.4.2 視点尤度ヒートマップ 11
2.4.3 視点バイアスの検証 13
2.4.3.1 ImageNet 学習済み認識モデルの分析 13
2.4.3.2 Pascal VOC, MSCOCO および検出モデルを用いた実験 . . 15
2.4.3.3 ImageNet におけるデータセット内視点バイアスの検証 . . . 16
2.4.4 視点バイアス指標の検討 17
2.4.4.1 視点バイアス指標の定義 17
2.4.4.2 視点バイアス指標を用いた評価実験 18
2.4.5 考察 19
2.5 多視点を前提とした尤度投票型物体認識 20
2.5.1 概要 20
2.5.2 多視点を前提とした尤度投票型物体認識手法 21
2.5.2.1 手法の概要 21
2.5.2.2 物体認識手法 21
2.5.2.3 カメラ位置の推定 22
2.5.2.4 観測対象の切り出し 23

2.5.2.5 3 次元空間尤度投票型物体認識 23
2.5.3 実験 24
2.5.3.1 実験の概要 24
2.5.3.2 実験の内容と結果 25
2.5.4 考察 28
2.6 本章のまとめ 28

第 3 章 多視点・複数モダリティを導入した視覚質問応答 29
3.1 概要 29
3.2 視覚質問応答 30
3.2.1 視覚質問応答タスク 30
3.2.2 視覚質問応答タスク用データセット 31
3.2.2.1 VQA v1 と VQA v2 31
3.2.2.2 CLEVR 32
3.2.3 視覚質問応答手法 33
3.2.3.1 LSTM Q+I 33
3.2.3.2 SAN 34
3.2.3.3 FiLM 35
3.3 3 次元認識 36
3.3.1 MVCNN 36
3.3.2 SRN 37
3.3.3 PointNet 38
3.4 RGB-D 画像を用いた視覚質問応答 39
3.4.1 概要 39
3.4.2 提案構造 39
3.4.3 実験 41
3.4.4 考察 42
3.5 多視点に対応した視覚質問応答 44
3.5.1 概要 44
3.5.2 提案構造 45
3.5.2.1 全体的構造 45
3.5.2.2 View pooling+FiLM 46
3.5.2.3 SRN+FiLM 46
3.5.3 実験 47
3.5.3.1 データセット 47
3.5.3.2 実験設定 49
3.5.3.3 定量的実験結果 50
3.5.3.4 入力視点数の影響 50
3.5.3.5 定性的実験結果 52
3.5.3.6 実験結果 52
3.5.4 考察 52
3.6 能動的視点選択を対応した多視点視覚質問応答 54
3.6.1 概要 54
3.6.2 Deep Q-learning Networks 55
3.6.3 提案構造 56

3.6.3.1 全体的構造 56
3.6.3.2 視点選択フレームワーク 56
3.6.3.3 実装の詳細 57
3.6.4 CG 画像を用いた実験 59
3.6.4.1 データセット 59
3.6.4.2 実験結果 59
3.6.5 リアル画像を用いた実験 61
3.6.5.1 データセット 61
3.6.5.2 実験結果 61
3.6.6 考察 64
3.7 大規模 3 次元物体モデルを用いた視覚質問応答データセット 65
3.7.1 概要 65
3.7.2 物体モデルデータセット 66
3.7.2.1 YCB 66
3.7.2.2 Bigbird 66
3.7.2.3 NEDO Items 66
3.7.3 データセット構築プロセス 67
3.7.3.1 物体モデルの収集 67
3.7.3.2 物体モデルのアノテーション 68
3.7.3.3 シーンの生成 68
3.7.3.4 質問と回答のペアの生成 69
3.7.4 実験 71
3.7.4.1 実験の設定 71
3.7.4.2 異なるスケールのデータセットでの結果 71
3.7.4.3 異なる入力画像サイズでの結果 71
3.7.5 考察 72
3.8 本章のまとめ 73

第 4 章 多視点・複数モダリティを導入したシーン変化説明文生成 74
4.1 概要 74
4.2 変化説明文生成 75
4.2.1 変化説明文生成タスク 75
4.2.2 変化説明文生成データセット 76
4.2.2.1 Spot-the-diff 76
4.2.2.2 CLEVR-Change 76
4.2.3 変化説明文手法 77
4.2.3.1 DDLA 77
4.2.3.2 DUDA 77
4.2.4 変化説明文評価指標 78
4.2.4.1 BLEU 78
4.2.4.2 ROUGE-L 78
4.2.4.3 METEOR 79
4.2.4.4 SPICE 79
4.3 多視点に対応した変化説明文生成 80
4.3.1 概要 80

4.3.2 提案構造 81
4.3.3 実験 82
4.3.3.1 データセット 82
4.3.3.2 実験設定 84
4.3.3.3 CG 物体モデルを用いた実験 84
4.3.3.4 リアル物体モデル画像を用いた実験 87
4.3.4 考察 89
4.4 多視点・複数モダリティに対応した変化説明文生成 90
4.4.1 概要 90
4.4.2 Matterport3D データセット 91
4.4.3 提案構造 92
4.4.4 実験 94
4.4.4.1 データセット 94
4.4.4.2 実験設定 96
4.4.4.3 異なる入力モダリティと画像エンコーダーでの実験 97
4.4.4.4 異なる複雑度を持ったデータセットでの実験 99
4.4.4.5 実験結果の分析 101
4.4.5 考察 101
4.5 本章のまとめ 102

第 5 章 結論 **103**

謝辞 **105**

発表論文 **106**

参考文献 **109**

図 目 次

2.1 AlexNet のフレームワーク図 (Krizhevsky et, al.[8] より) 5
2.2 VGG のフレームワーク図 (Simonyan et, al.[9] より) 6
2.3 Residual learning の図例 (He et, al.[2] より) 7
2.4 ResNet のフレームワーク図 (He et, al.[2] より) 7
2.5 Faster R-CNN のフレームワーク図 (Ren et, al.[10] より) 8
2.6 YOLO のフレームワーク図 (Redmon et, al.[11] より) 9
2.7 視点-尤度ヒートマップの概念図 11
2.8 ImageNet で学習済みの認識モデルの視点尤度ヒートマップの例 13
2.9 MSCOCO, Pascal VOC で学習済みの認識モデルの視点尤度ヒートマップの例 15
2.10 CG モデル，視点分布図，視点尤度ヒートマップ 16
2.11 球面調和関数の図例 17
2.12 視点尤度ヒートマップの平均尤度，視点バイアス指数の値 18
2.13 尤度投票型物体認識の概念図 21
2.14 実験に用いたカメラパス 24
2.15 撮影済みの bottle 画像例 25
2.16 YOLO v2 を用いた認識尤度遷移図 (bottle) 25
2.17 Faster R-CNN を用いた認識尤度遷移図 (bottle) 25
2.18 撮影済みの bowl 画像例 26
2.19 YOLO v2 を用いた認識尤度遷移図 (bowl) 27
2.20 Faster R-CNN を用いた認識尤度遷移図 (bowl) 27

3.1 視覚質問応答タスクの概念図 30
3.2 VQA v1 データセットの例 (Agrawal et, al. [25] により) 31
3.3 VQA v2 データセットの例 (Goyal et, al. [40] により) 31
3.4 CLEVR データセットの例 (Johnson et, al. [41] により) 32
3.5 LSTM Q + I のフレームワーク図 (Agrawal et, al. [25] により) 33
3.6 SAN のフレームワーク図 (Yang et, al. [42] により) 34
3.7 FiLM layer のイメージ図 (左)，FiLM を導入した視覚質問応答フレームワーク図 (右) (Perez et, al. [43] により) 35
3.8 MVCNN のフレームワーク図 (Su et, al. [44] により) 36
3.9 SRN のフレームワーク図 (Eslami et, al. [45] により) 37
3.10 PointNet のフレームワーク図 (Qi et, al. [46] により) 38
3.11 RGB-D 画像を用いた視覚質問応答の概念図 39
3.12 RGB-D 画像を用いた視覚質問応答のフレームワーク図 40
3.13 RGB-D 視覚質問応答データセットの例 41
3.14 RGB-D 視覚質問応答の結果例 42
3.15 多視点に対応した視覚質問応答の概念図 44
3.16 多視点に対応した視覚質問応答のフレームワーク図 45

3.17 仮想カメラの設定図 47
3.18 Multiview CLEVR データセットの統計情報 48
3.19 Multiview CLEVR データセットの例 49
3.20 Multiview CLEVR データセットで異なる質問タイプの定量評価 50
3.21 Multiview CLEVR データセットでの結果例 52
3.22 能動的視点選択を対応した多視点視覚質問応答の概念図 54
3.23 Experience replay を用いた deep q-learning のアルゴリズム (Volodymyr et, al. [49] により) 55
3.24 能動的視点選択を対応した多視点視覚質問応答のフレームワーク概念図 . . . 56
3.25 能動的視点選択を対応した多視点視覚質問応答のフレームワーク図 57
3.26 仮想カメラ設定とデータセットの例 59
3.27 CG データセットでの結果例 60
3.28 CG, リアル撮影，再学習用 CG データセットの例 61
3.29 リアル撮影データセットでの結果例 63
3.30 ROM_CLEVR データセット構築プロセス 67
3.31 ROM_CLEVR データセットモデルの例とクラス分布図 67
3.32 ROM_CLEVR データセットのクラス構造図 68
3.33 質問テンプレートの例 69
3.34 ROM_CLEVR データセットの例 70
3.35 ROM_CLEVR データセットでの結果例 72

4.1 シーン変化説明文生成タスクの概念図 75
4.2 Spot-the-diff データセットの例 (Jhamtani et, al. [64] により) 76
4.3 CLEVR-Change データセットの例 (Park et, al. [63] により) 76
4.4 DUDA 手法のフレームワーク図 (Park et, al. [63] により) 77
4.5 多視点に対応した変化説明文生成の概念図 80
4.6 多視点に対応した変化説明文生成のフレームワーク図 81
4.7 データセットの仮想カメラ設定 82
4.8 定性的実験結果 86
4.9 P_change_cameras データセットでの BLEU-4 評価 87
4.10 ROM_change データセットでクラスごとの BLEU-4 評価 88
4.11 多視点・複数モダリティに対応した変化説明文生成の概念図 90
4.12 Matterport3D データセットの概念図 (Chang et, al. [71] により) 91
4.13 多視点・複数モダリティに対応した変化説明文生成のフレームワーク概念図 92
4.14 多視点・複数モダリティに対応した変化説明文生成のフレームワーク図 . . . 93
4.15 仮想カメラ設定図 94
4.16 データセットでの一例 95
4.17 s15_o10 データセットで結果例 99
4.18 s15_o85, s100_o10 データセットで結果例 100

表 目 次

3.1 RGB-D CLEVR データセットでの定性評価. 41
3.2 Multi-view CLEVR データセットでの統計情報. 48
3.3 Multi-view CLEVR データセットでの定量評価 (* は CLEVR1.0 での結果). . 50
3.4 Multi-view CLEVR で異なる入力視点の組み合わせの結果. 51
3.5 データセットでの統計情報 59
3.6 Multi-view-CLEVR_12views_**CG** での定量評価 59
3.7 視点選択が精度に与える影響 60
3.8 Multi-view-CLEVR_4views_**CG** データセットでの定量評価 61
3.9 Multi-view-CLEVR_4views_**Real** データセットでの定量評価 62
3.10 ROM_CLEVR データセットでの統計情報 69
3.11 ROM_CLEVR_v0.5 と ROM_CLEVR_v1.0 データセットでの定量評価 . . 71
3.12 ROM_CLEVR_v1.0 データセットでの定量評価 71

4.1 P_change_occlusion, P_change_cameras と ROM_change データセットの統計情報 82
4.2 ROM_change データセットの物体クラス分布 83
4.3 P_change_occlusion (上 10 行) と ROM_change (ボトム 6 行) データセットでの定量評価 85
4.4 P_change_occlusion データセットでの説明文精度評価 85
4.5 ROM_change データセットでの説明文精度評価 88
4.6 データセットの物体クラスとインスタンス 94
4.7 データセット統計情報. 96
4.8 s15_o10 (上 11 行), s15_o85 (中間 6 行), と s100_o10 (ボトム 6 行) での定量評価 97
4.9 s15_o10 (上 11 行), s15_o85 (中間 6 行), と s100_o10 (ボトム 6 行) での説明文精度評価. 98

第1章　序論

1.1　研究背景

近年，グラフィックスプロセッシングユニット (GPU) の発展とディープニューラルネットワーク (DNN) の構造の研究および様々な DNN を用いた手法の進歩により，人工知能 (AI) 分野が目覚ましい発展を遂げた．大規模な画像認識用データセットの提案や DNN 手法の導入により，AI に関わる重要な研究分野の一つであるコンピュータビジョン分野においても，従来は難しいとされていた，画像中に含まれる任意の物体の種別を識別する一般物体認識問題を含む様々なタスクの性能が大幅に改善された．更に，物体検出やセマンティックセグメンテーション，画像生成などの様々な新しいタスクが提案され，研究分野が大きく拡大している．また，産業界における AI 技術のニーズについても近年急速に高まってきている．顔認証システムや医用画像診断支援システムなどは既に我々の生活の質や安全性の向上に重要な役割を果たしている．今後社会的ニーズの更なる高まりにつれて，コンピュータビジョン技術の実環境システムでの応用も更に進んでいくと考えられる．例えば自動運転システムや人と共生するロボットシステムの開発などにおいて，コンピュータビジョン技術は重要な役割を担うと考えられるが，そのような場面においては，与えられた１枚の画像に対する認識精度を向上させるような従来の取り組みだけでは不十分で，例えば観測対象がどのような姿勢で観測されるかわからないような条件下において，対象を様々な視点から能動的に観測し，実用的な認識精度を獲得するようなメカニズムについても検討していくことが必要になると考えられる．

1.2　研究目的

コンピュータビジョン分野において複数の視点を扱う問題は，3 次元復元の文脈でこれまでよく扱われてきた．一方物体認識の文脈においては，例えば物体を観測するアングルと認識精度の関係すら，関係があることは誰もが認識していながらも，これまで定量的な検討は行われてこなかった．本研究は，画像認識問題を「観測視点」という切り口で分析・考察することで，その性能の改善を図ることを目的とする．

具体的には，まず物体認識タスクをベースとした分析を行った．物体を観測するアングルと認識精度の関係について検討を行うため，任意の観測アングルと尤度の関係をヒートマップで可視化する手法を提案する．その結果，既存の複数の物体認識モデルの認識尤度が観測視点に依存する傾向が観測できた．そこにおいて発生しているバイアスの１つの原因として，ImageNet [1] などのよく用いられる学習用データセット中に観測アングルの偏りが存在していることを示す．加えて，複数の視点からの認識結果を統合することで，認識を安定させることが可能であることを示す．

次に，実環境をターゲットとした具体的なタスクとして，視覚質問応答（Visual Question Answering）およびシーン変化説明文生成（Scene Change Captioning）を取りあげ，ここに複数視点の概念を導入することで性能を大きく向上させることが可能であることを示す．これらのタスクは，人間協調型のロボットなどで特に重要であると考えられており研究が進

められてきたが，入力された画像と関係なく解答の選択肢の正解確率を単純に学習してしまうという問題点が指摘されていた．我々はこの問題の主な原因は，入力される画像情報の不足であると仮定し，(1) 処理のフレームワーク：複数視点および複数のモダリティ（RGB, depth, point cloud）を入力可能な認識フレームワーク，(2) データセットの生成：単視点且つ単純な CG に基づく画像と質問文・正解のペアを自動生成する従来法に対し，複数視点及び複数のモダリティに対応し，リアルな情景を生成可能なデータ生成フレームワークをそれぞれ構築することで，十分な画像情報を提供可能にし，従来と比較して性能が向上することを示す．また，複雑なテクスチャや遮蔽などが存在するより困難な条件設定下においても視覚質問応答やシーン変化説明文生成が可能になることを示す.

1.3 論文の構成

本論文の構成を以下に示す.

第二章では，物体認識タスクにおいて観測視点を検討する．まず，複数視点から観測された CG 物体モデルを認識対象とし，既存の物体認識モデルに認識精度が観測視点に依存する視点バイアス傾向の存在を明示する．次に，認識モデルの視点バイアスの原因としては学習用データセットに視点バイアスが存在することであると想定し，既存の物体認識用データセットに含まれる視点を手作業で集計し，データセットに視点バイアスの存在を明らかにする．また，視点バイアスを評価する指標を提案し，定量的視点バイアスの分析を可能にする．認識に視点バイアスが存在することから，多視点から物体を観測し様々な視点の認識結果を集計することにより認識の精度が向上可能であることを示す.

第三章では，第二章で得られた多視点認識の有効性を踏まえて，視覚質問応答タスクにおいて多視点および複数モダリティの影響について述べる．まず，質問回答に必要な幾何情報をデプス画像から得やすいことを想定し，RGB-D 画像を用いた視覚質問応答タスクおよび実験を行う．その結果から，デプス画像の有用性を確認した．次に多視点を対応した視覚質問応答タスクを提案し，多視点画像から得られるシーンの 3 次元情報がこのタスクにおいての有効性を示す．人間がそうであるように，観測視点を選択するフレームワークを視覚質問応答タスクに導入し，観測視点を選択することにより高効率で高い精度で質問を回答できた結果を示す．最後に，リアル撮影により得られる物体モデルを導入した多視点視覚質問応答データセットの構築を提案する.

第四章では，シーン変化説明文生成タスクにおいて多視点および複数モダリティの影響について述べる．まず第三章と同様に，多視点を変化説明文生成タスクに対応し，それによりモデル性能を向上できることを示す．次に，更に多視点・複数モダリティをこのタスクに導入し，我々が提案した室内シーン変化説明文生成データセットで多視点および複数モダリティが精度向上における効果を示す.

第五章にて本論文の結ぶとする．現在抱えている問題と本論文のまとめについて説明する．さらに今後の課題と展望について述べる.

第2章　視点依存性評価および多視点尤度投票型物体認識

2.1　概要

近年，物体認識タスクの性能が飛躍的に向上し，2次元画像の認識問題は既に概ね解決されたと見る向きすらある [2, 3, 4]．しかしながら，例えばロボットにおける物体認識を想定して，任意の視点から対象物を撮影した画像の認識を行ってみると，思いのほか良好な認識結果を得ることは難しく，良好な認識が可能な視点はむしろ限定的であることがわかる．このような現象は，これまでよく用いられてきたデータセットを構成するデータの多くが，インターネット上から収集された画像であることが原因の一つとして考えられる．インターネット上で公開されている画像の多くは，商品や個人の所有物などを紹介するという目的で，人間によって撮影された画像が支配的であると考えられる．人間が撮影を行う場合には，よほど意図的に特殊な視点を選択しない限りは，撮影対象を人間が最も認識しやすい視点を自然に選択すると考えられる．また場合よっては撮影対象を最も美しく見せるような視点が意図的に選択されることもあり得る．このようなことから，主に人間によって撮影された画像を大量に収集することで作成されたデータセットには，人間による強い視点のバイアスが与えられているという仮説を立てることができる．

従来，物体認識タスクにおける視点バイアスの存在は「カメラアングルが悪いと良好な認識結果が得られない」といった体験として，暗黙には知られていたと考えることができるだろう．しかしながら，その定量化および分析を詳細に行っている例はこれまで無かった．本章で，このような視点バイアスの仮説が実際に存在することを検証したうえで定量化・可視化し，その影響の分析を可能にすることである．これが実現すれば，将来的にはデータセットの品質評価基準の一つとしての役割も検討可能になる．

上述の検討により，認識が困難なアングルは少なからず存在し，ある視点から１ショットで認識を行うことは誤認識のリスクが高いことがわかった．そこで更に，本章で複数の視点から物体を認識し，認識結果を物体が存在する3次元領域に投票し，統合した認識結果で物体認識を行うフレームワークを検討した．実験結果により，観測に偶然な視点による誤認識が抑制され，提案手法が安定的に物体の正解ラベルを得れれることを示した．

本章の構成を以下に示す．

節2.2では画像認識技術に関する従来研究を述べる．具体的に，大規模画像認識用データセット，物体認識手法および物体検出手法を紹介した．節2.3で画像データセットの視点バイアス問題に関する関連研究を述べる．節2.4で一般物体認識用大規模データセットにおける視点依存性の検証，データセットの視点バイアス問題が物体認識手法に与える影響および視点バイアスを定量的に評価可能な評価指標を述べる．節2.5では，節2.4で観測できたデータセット視点バイアスから，多視点を導入した物体認識手法及び実験を述べる．節2.6で本章をまとめる．

2.2 Deep Neural Network（DNN）を用いた画像認識技術

2.2.1 大規模画像認識用データセット

2.2.1.1 ImageNet

ImageNet [1] データセットは 2009 年に，Fei Fei Li らによって公開された．ImageNet データセットの画像が既存のサーチエンジンにより集めた．また，ImageNet のすべてのカテゴリ名は英語の概念辞書 WordNet [5] のノードと対応付けられている．WordNet は 1985 年に George A. Miller らにより開発された概念辞書で，英単語を synset と呼ばれる同義語のグループに分類したうえでノードとして表し，単語間の関係を木構造で表す．ImageNet には現時点で 21,841 synsets，合計 14,197,122 枚の画像が含まれている．

ImageNet データセットをベースにとした大規模画像認識タスクのチャレンジ Large Scale Visual Recognition Challenge (ILSVRC) も Fei Fei Li らにより開けられ，大規模画像認識技術の急速的な発展に大きく貢献をした．ImageNet データセットが様々な画像認識タスクの学習に強いテストベッドを提供し，CNN (畳み込みニューラルネットワーク) と共に一般物体認識タスクの急激な性能向上の引き金となった．現在でも，ImageNet は，様々な深層学習タスク用ネットワークのための事前学習データセットとして頻繁に利用されている．

2.2.1.2 Pascal VOC

Pascal VOC [6] データセットは 2005 年に Mark Everingham らが公開した大規模画像データセットである．Pascal VOC データセットに含まれる画像は画像投稿 SNS である Flickr および Miscrosoft Research Cambridge (MSRC) データセットから収集した．データセットに人，動物，車および室内用品など合計 20 カテゴリの画像から合計 11,530 枚の画像から構成される．

Pascal VOC データセットは画像ごとに物体カテゴリがアノテーションされるほか，物体が画像中に存在する領域を示すバウンディングボックスがアノテーションされている．Pascal VOC データセットは画像認識の精度を競うコンテストで用いられていたほか，物体検出タスク (画像内の物体の位置を検出し，矩形領域等で示すタスク) の評価によく用いられる．

2.2.1.3 MSCOCO

MS COCO（Microsoft Common Objects in Context）[7] データセットは 2014 年に Microsoft 社が公開した大規模画像データセットである．MS COCO データセットは現在 91 物体カテゴリ合計 330K の画像から構成されている．画像ごとに密に物体の領域を示すバウンディングボクス，物体のセマンティックセグメンテーションマスクがアノテーションされている．更に，画像ごとに 5 つの画像の内容に関する説明文が付けられている．

MS COCO は提案された当初は物体検出（画像内の物体の位置を検出し，矩形領域等で示すタスク）とセマンティックセグメンテーション（画像をピクセル単位でラベルを付けるタスク）のために設計された．現在はキーポイント検出や画像キャプション生成（画像の説明文を生成するタスク）などにもよく用いられる．ImageNet データセットと同様に，MS COCO の提案が現在の物体検出，セマンティックセグメンテーション，および画像説明文生成の手法の発展に大いに貢献し，現在でもそれらのタスクの学習と評価となる重要なテストベッドである．

2.2.2 物体認識手法

2.2.2.1 AlexNet

図 2.1: AlexNet のフレームワーク図 (Krizhevsky et, al.[8] より)

AlexNet [8] は 2012 年 Krizhevsky らにより提案された画像認識用ニューラルネットワーク (NN) ベースのフレームワークである．AlexNet 前述の ImageNet データセットを用いた画像認識コンペティション ILSVRC2012 において，57.1%のトップ 1 認識率で 2 位を大幅に上回った結果で優勝を遂げた．AlexNet が提案されて以来，NN または DNN を用いた手法が画像認識タスクに主流となり，更に，画像認識からはじめ，DNN が様々なタスクに導入されるようになった．

図 2.1 に AlexNet のフレームワークを示す．AlexNet は 5 層の畳み込み層 (convolution layer) と 3 つの全結合層 (fully connection layer) から構成される．また，AlexNet に max pooling 操作，ReLU 活性化関数，バッチノーマライゼーション，ドロップアウト操作などの現在の DNN の基本となる操作も実装されている．

2.2.2.2 VGG

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

図 2.2: VGG のフレームワーク図 (Simonyan et, al.[9] より)

VGG [9] ネットワークが 2015 年に Simonyan らにより提案された．Simonyan らが DNN の層を増やすことにより性能を向上可能である仮定から，異なる深さの DNN ネットワークを構築し，性能の分析実験を行った．図 2.2 に論文に実装したネットワークの構造を示す．

VGG ネットワークが AlexNet に使用した reception field が大きい畳み込みフィルター (例：AlexNet では 11*11,7*7,5*5 サイズのフィルターが実装されている) を連続の 3*3 フィルターに調整し，ネットワークをより深く構築することを可能にした．VGG の実験でネットワークの層を増やすことによりネットワークの性能を向上できる傾向を示した．

2.2.2.3 ResNet

図 2.3: Residual learning の図例 (He et, al.[2] より)

前述の VGG ネットワークで述べたような，DNN の層を増やすことによりネットワークの性能を向上できる傾向である．しかし，単純にネットワークの層を増やす場合に，gradient vanishing や gradient exploding などの現象が起こりやすく，深層ネットワークのパラメータ更新が困難になる傾向がある．このような問題を解決するために，2015 年に He らにより residual learning [2] を提案された．図 2.3 に示しているのように，従来のネットワークではネットワークにより入力 X が $F(X)$ にマッピングされる．He らの residual learning では入力 X を $F(X) + X$ にマッピングすることにより，アイデンティティマッピングを構造上可能にした．

図 2.4: ResNet のフレームワーク図 (He et, al.[2] より)

He らの residual learning を DNN ネットワークに導入することにより，深層ネットワークのパラメータ更新が従来より容易になり，100 層を超える層を持ったネットワークの学習も安定させられた．residual learning を実装した深層ネットワーク (ResNet) の図例を 2.4 に示す．He らの residual learning が現在あらゆる DNN ネットワークの構築に広く使われている．

2.2.3 物体検出手法

2.2.3.1 Faster R-CNN

図 2.5: Faster R-CNN のフレームワーク図 (Ren et, al.[10] より)

DNN の発展および大規模画像認識データセットの発展により，物体を画像中から認識するほか，物体が画像中に占める領域を矩形のバウンディングボクスにより検出するタスクが提案された．Faster R-CNN が代表的な物体検出の手法として現在広く使用されている．

Faster R-CNN [10] のフレームワークを図 2.5 に示す．Faster R-CNN ではまず CNN により画像の特徴マップを得る．次に，画像中に物体が存在する可能性が高い領域を Region Proposal Network (RPN) により検出する．その次，RPN により得られる物体が存在する領域 proposals および画像特徴を RoI pooling に入力し，proposal ごとの特徴マップが計算される．最後に proposal の特徴マップからクラス分類を行う．

以上のプロセスにより，Faster R-CNN のフレームワークが end-to-end で一つのネットワークで画像から物体を検出できる．

2.2.3.2 YOLO

図 2.6: YOLO のフレームワーク図 (Redmon et, al.[11] より)

You Only Look Once (YOLO) [11] 手法が 2016 に Redmon らにより提案された物体検出の手法である．前述の Faster R-CNN では (1) 物体が存在する領域の識別；(2) 領域ごとに物体識別を行う；の 2 つのステップで物体検出を行っている．それと比較して，YOLO 手法で同時に物体領域および物体クラスの推定を行う．

具体的に，YOLO では画像をグリッドに分割し，図 2.6 に示しているネットワークを経由して，グリッドごとにバウンディングボクスの位置推定 (現在のグリッド領域とのオフセット)，バウンディングボクスに物体が存在する確率，および物体のクラスごとの確立を同時に予測する．以上のプロセスにより，YOLO が 1 つのステップでハイスピードおよび高精度で物体を検出可能である．

更に，2017 に Redmon らがバウンディングボクスの prior により予め定義した anchor の導入，高解像度画像の対応，バッチノーマライゼーションの導入，および多スケール画像の学習の導入などにより物体検出精度およびスピードが YOLO モデルを上回った YOLO v2 [12] モデルを提案した．現在 YOLO v2 が様々な場面で広く使われている．

2.3 データセットの視点バイアス

データセットに何らかの偏った性質が観測された場合にデータセットにバイアスが存在するものとする．データセットは母集団から偏りなくサンプリングされたものであることが期待されるので，バイアスの存在には常に注意を払う必要がある．従来には，画像データセットのバイアスに着目して分析を行った例として，異なる画像データセット間における汎化性に着目する“ cross-dataset generalization ” [13] を評価するものが挙げられる．同じカテゴリを含んでいる 2 つのデータセットの片方で学習を行い，他方を認識した場合の結果が良好で，且つ両者を入れ替えた場合にも同じく良好な結果が得られれば，双方のデータセットは母集団から偏りなくサンプリングされており，バイアス性は低いと考えることができる．

また，他の研究例として，視覚と言語のマルチモーダルタスクに用いるデータセットのバイアス性に関するものも多くある．Ishan Misra [14] らは人間によるアノテーションに起因するノイズの存在を指摘した．例えば，“ street ” とアノテーションされた領域の中に高い確率で人物が含まれている場合，人物領域を“ street ” と学習してしまうことがある．し

かもこのようなバイアスはアノテーションの作成者によって異なる可能性があり問題である．Hendricks [15] らは画像説明文生成データセットにおける性別バイアスの存在を指摘した．例えば“ snowboarding ”のシーンには男性が多く，学習済みのモデルに女性による snowboarding の画像を入力すると“ A man is snowboarding ”となってしまうことがある．

他にも，Emiel van Miltenburg [16] は画像説明文データセットにおける職業，関係，人種などのバイアス性を検討した．例えば，画像説明文データセット Flickr30K において，アジア人は（他のアジアの国が選択されるケースはほとんど無く）ほぼ全てが“ Chinese ”または“ Japanese ”とアノテーションされてしまっている例が報告されている．

以上のようにデータセットにおけるバイアスは認識タスクを根幹から不確かなものにする可能性があり，従来から問題視されてきた．一方で，実用上認識精度に直接的かつ重大な影響を及ぼす「視点のバイアス」に関する検討はこれまで無く，その実態を明らかにする必要があった．

2.4 一般物体認識用大規模画像データセットにおける視点依存性に関する考察及び定量化

2.4.1 概要

近年，物体認識タスクの性能が向上し，既に人間精度を超える認識モデルが提案された．しかしながら，任意の視点から対象物を撮影した画像の認識を行ってみると，良好な認識結果を得られるとは限らない傾向がある．このような現象は，これまでよく用いられてきたデータセットを構成するデータの多くが，人間が撮影することにより集められ，撮影対象を人間が最も認識しやすい視点を自然に選択すると考えられる．このようなことから，主に人間によって撮影された画像を大量に収集することで作成されたデータセットには，人間による強い視点のバイアスが与えられているという仮説を立てることができる．

ここでは，上述のような視点バイアスの仮説が実際に存在することを検証したうえで定量化・可視化し，その影響の分析を可能にすることである．

具体的に，まず，視点と認識尤度の関係を表す視点—尤度ヒートマップを定義し，その関係を定量化・可視化すること基本とし，様々な分析を行う．次に，ヒートマップをベース，複数のデータセットを用いて学習を行ったネットワークに対して視点バイアスの可視化と分析を行う．以上により，視点バイアスが異なる学習データセットで学習済みの複数のネットワークに存在することを確認できたことから，その原因を明らかにするため，学習に用いるデータセットにおける視点の出現頻度を手作業で分析した．更に，データセットに存在する視点バイアスと学習済みネットワークの認識の視点バイアスの関係性を分析した．加えて，本来は球面上の分布として表現される視点バイアスの特徴を，近似的によく表現可能なスカラー値の指標を設計し，視点バイアスの定量化評価を行う．

2.4.2 視点尤度ヒートマップ

図 2.7: 視点-尤度ヒートマップの概念図

学習済みモデルが持つ視点バイアスを精密に評価するためには，対象カテゴリに属する複数のインスタンスに対して，偏りのない網羅的な視点から撮影したテスト画像を大量に用意する必要がある．しかし，多くのカテゴリ，インスタンスに対して実際にカメラの位置を物理的に調整しながら撮影を行うことは極めてコストが大きいという問題がある．そこで我々は，CG モデルを用い，モデルを取り囲む球面上にほぼ等間隔に設定した密な視点（1,438 視点）から仮想的にモデルを撮影し，それらの画像に対する認識尤度の値を均等にカラーマップに対応させ球面上にプロットした「視点—尤度ヒートマップ」を生成する．ここで認識尤度は，認識モデルが入力画像の正解カテゴリを示す確率と定義する．具体的には，多クラス分類を行う深層ニューラルネットワークにおける最終層の正解カテゴリに対応する出力値（Softmax 値）を用いる．また，ある「視点」からの視線ベクトルは常にモデルを取り囲む球の中心を向くこととする.（従って，視点が決まればカメラの位置姿勢が決まる.）また，尤度は，認識結果として出力される複数のカテゴリ候補のうち，正解カテゴリの尤度を用いる．

視点—尤度ヒートマップは，球面上の尤度の分布 $L(\theta, \phi)$ として表現されるが，視点 (θ, ϕ) はサンプルサイズが限定されるため，実際には離散分布となる．可視化においては，観測データを単位球面上に射影した視点を辺で結びメッシュ化した多面体の上で，尤度による彩色を行う．まず頂点に尤度に対応した色で彩色する．今回は尤度 0 を blue（#0000ff），尤度 1/3 を green（#00ff00），尤度 2/3 を yellow（#ffff00），尤度 1 を red（#ff0000）として，中間の値は RGB 色空間において線形補間することで算出したカラーマップ（図 2.7）を使用した．次に頂点間の面や辺をバイリニア補間で彩色する．この彩色したメッシュを 3D レンダラにより投影することで視点—尤度ヒートマップの可視化画像が得られる．図 2.7 に視点—尤度ヒートマップの概念図を示す．中心部に配置された物体に対し，球状に取り囲む視点から観測したときの尤度が球面上の分布として表現されていることがわかる．

以上により，視点と尤度の関係の定量化・可視化が可能になる．以降では，これを用いて視点バイアスの検証を行う．

2.4.3　視点バイアスの検証

2.4.3.1　ImageNet 学習済み認識モデルの分析

図 2.8: ImageNet で学習済みの認識モデルの視点尤度ヒートマップの例

ImageNet [1] 学習済み認識モデルに内在する視点バイアスを検証するため，ImageNet データセットで学習済みの AlexNet [8]，VGG19 [9]，及び ResNet152 [2] の認識モデルを用いて，8 カテゴリ（monitor，mug，pistol，bottle，ashcan， bowl，chair，airliner；それぞれ 2 つのインスタンス）の物体に対する視点－尤度ヒートマップを生成した（なお，このとき CG モデルへの fine tuning は行っていない）．本研究で用いる CG モデルは ShapeNet [17] のものを用い，ImageNet および ShapeNet で共通するカテゴリの中から，bowl のような比較的単純な形状から airliner のような複雑な形状までを含むように 8 カテゴリを選択した．結果を図 2.8 に示す．本来の球面上の分布を，直感的に理解しやすくするために，モデルご

とに上面図（top view），側面図（side view），正面図（front view）の 3 つの姿勢で示す．このそれぞれに対して，各認識ネットワーク（AlexNet，VGG19，ResNet152）の視点—尤度ヒートマップを 3 つ並べて示す．なおモデルおよび横に並ぶヒートマップの姿勢は一致させてある．

図 2.8 により，これまで感覚に依存していた視点と尤度の関係が明瞭に可視化できていることがわかる．また，視点によって認識尤度が大きく変化していること，および分布のランダム性は低く，それぞれ何らかの特徴的な傾向を持っていることがわかる．例えば，図 2.8 の正面図（front view；右端列）をみると，1.monitor，2.mug，4.bottle，5.ashcan，6.bowl は，それぞれ南半球（球の下半分），特に南極付近の物体を見上げる視点の尤度が低く，青色で示されていることがわかる．これはこれら 5 カテゴリの物体は通常は机や地面上に設置されるため，物体を見上げることは困難であり，主にインターネット上から画像が収集された ImageNet においてはこのような視点の画像はほとんど含まれていなかったか，他の視点と比較して相対的に少なかったのだと推測できる．これらに対して，8.airplane は空港ビルから見下ろすように撮影された画像や，駐機中に地上から見上げるように撮影された画像などがネット上にも多く存在するため，北半球のみならず，南半球も南極付近の一部を除いて高い尤度となっていることがわかる．そのほか，1.monitor は正面以外からの認識は困難であること，2.mug は取っ手が観測できない視点の認識は困難であることなどを読み取ることができる．これらも，monitor の裏面を積極的に撮影することは通常ないこと，mug を撮影する場合には取っ手が隠れている視点を積極的に選択することはないこと，などが原因と考えられる．

次に認識モデルによる性質の違いを考察する．図 2.8 の認識モデル（AlexNet，VGG19，ResNet152）による違いに着目すると，まず大局的には前述の特徴的な視点バイアスの傾向が類似して観測されていることがわかる．それぞれ大きく異なる構造を持つ認識モデルによっても同様の傾向が観測されるということは，やはり認識モデルが示す視点バイアスの支配的な原因は，データセットにおける視点バイアスにあると考えられる．なお，より詳細な違いに着目すると，提案された年代が古い順である AlexNet，VGG19，ResNet152 の順で尤度の値および視点バイアスが改善される傾向にある．これは後年のものほど多層化が進み，新たな初期化手法や正則化手法といった様々な改良が加えられたことが原因と考えられる．これは前述の性質と合致している．特に最も深いネットワーク構造を持つ ResNet152 では，（特に 4.bottle や 5.ashcan の場合において）認識の状態が大きく改善している場合があり，ネットワーク構造の工夫により，他と比較してより高い汎化性能を獲得していると推測できる．

2.4.3.2 Pascal VOC, MSCOCO および検出モデルを用いた実験

図 2.9: MSCOCO, Pascal VOC で学習済みの認識モデルの視点尤度ヒートマップの例

前節は物体認識タスク用のデータセットおよび認識モデルを用いて検証を行ったが，これと大きく異なる物体検出タスク用のデータセットおよび認識モデルを用いた場合にも同様の傾向が観測されたのかどうかを検証するために，MS COCO [7] および Pascal VOC [6] データセットを用い，ネットワークモデルとしては YOLO v2 [12]) を使用して 3 カテゴリの物体（monitor，bottle，chair；それぞれ 2 つのインスタンス）の認識を行った．CG モデルは引き続き ShapeNet のものを利用した．図 2.9 に結果を示す．1 列目はモデルを示す．2 列目は MS COCO，3 列目は Pascal VOC の結果である．1 および 2 行目，3 および 4 行目，5 および 6 行目のヒートマップを，それぞれ対応する図 2.8 の 1.monitor，4.bottle，7.chair の上面図（top view）のヒートマップと比較してみると，データセットも認識モデルも異なるにも関わらず，よく類似した視点バイアスが生じていることがわかる．このことから，当初の仮説通り，主に人間によって撮影された画像を大量に収集することで作成されたデータセットには，人間による強い視点バイアスが与えられている可能性が高いと考えられる．

2.4.3.3 ImageNet におけるデータセット内視点バイアスの検証

図 2.10: CG モデル，視点分布図，視点尤度ヒートマップ

これまでの検討により，データセット内に強い視点バイアス与えられていることが示された．これを直接的且つ確実に確かめるためには，データセット内の視点の分布を詳細に調べる必要がある．しかしながら，現時点ではこれを精度よく自動で行う方法が存在しないため容易ではない（自動で行うためには 1 枚の画像のみから撮影対象の物体の姿勢を正確に推定する必要があるが，本研究の知見を考慮すると，そのような手法においても推定が苦手な姿勢が存在しているはずである）．そこで本研究では，現時点での最も信頼できる手段として，手動で視点をアノテーションし，視点に関する統計情報を得た．視点は，球面上にほぼ均等に分布する 66 視点とし，ImageNet の 2 つのカテゴリ：monitor，mug を対象に，合計 2,600 枚の写真を一枚一枚人間が確認しアノテーションを行った．

図 2.10 に結果を示す．1 列目が認識対象のモデル（mug および monitor），2 列目が手作業によって確認した ImageNet 内の視点の分布，3，4，5 列目がそれぞれ AlexNet，VGG19，ResNet152 の視点－尤度ヒートマップである．まず，mug の 2 列目を見ると ImageNet の中には「取っ手が右向きでやや見下ろす視点」にデータが集中していることがわかる．これは人間には右利きが多いため，暗黙に取っ手を右にして撮影する傾向があることと対応しており非常に興味深い．これに対し，3，4，5 列目では左右の偏りは解消されている．これは学習時に左右鏡像化のデータ拡張を行った可能性があるほか（今回は PyTorch [18] の学習済みモデルを使用したためデータ拡張の詳細は公開されていない），認識モデルの持つ汎化性が影響した可能性も考えられる．次に，monitor の結果を見ると，2 列目と，3，4，5 列目が単純によく類似していることがわかる．この例では，データセット内の視点バイアスがそのまま認識時の視点バイアスの主因になっていると言える．

以上のように，手作業によって確認した ImageNet 内の視点の分布と，視点－尤度ヒートマップの間には強い関係性があることが明らかになった．つまり，認識モデルの視点バイアス性はデータセットの視点バイアス性に強い影響を受けるということを示している．この

ことからデータセットの品質評価を行う際には，視点バイアスも指標として考慮すべきであることがわかる．

2.4.4 視点バイアス指標の検討

視点—尤度ヒートマップは球面上の分布で表現されるため，直感的な観察・理解に適している．ここで更に，分布の特徴を良好に表現可能なスカラー値の指標を定義することができれば，様々な比較検討をグラフ上で行うことなどが可能になり，実用上メリットが大きい．

2.4.4.1 視点バイアス指標の定義

図 2.11: 球面調和関数の図例

我々は，指標を定義するにあたって分布の「急峻さ」すなわち，視点に対する尤度の移り変わりの大きさに着目した．高尤度の領域がピークとして立つようなすそ野の狭い分布である場合は急峻さが高く，逆にすそ野の広い分布である場合は急峻さが低いとする．急峻さが高い場合，僅かな視点移動で急激に尤度が変化することになり，視点バイアスが強い状態と定義する．

以下，具体的な定義について述べる．球面上の直交関数系である球面調和関数 [19] は球面上の分布を表現するために都合がよい．球面上に分布する視点の尤度ヒートマップを球面調和関数により多重極展開し，これにより得られるパワースペクトル（多重曲展開係数の二乗）の積分で球面分布の変化の度合いを表す．球面調和関数はその直交性から，球面上の分布を，空間的に異なる周期性を持った分布に分解することが可能であり，球面上のフーリエ変換と捉えることもできる．球面調和関数はフーリエ級数の周波数成分に相当し，フーリエ級数展開がある波を異なる周波数成分の波の和に展開するのと同様に，多重極展開によって球面分布を空間的に異なる周期性を持った球面調和関数に分解できる．ここで，フーリエ係数，すなわち周波数成分の振幅が高いほどに急峻な波となるように，各球面調和関数の多重曲展開係数が高いほど，その球面調和関数の示す分布の急峻さは高くなる．これらを各球面調和関数に対して（二乗して）積分することで急峻さを表現する．

具体的に，視点—尤度ヒートマップ $L(\theta, \phi)$ に対し，まずヒートマップを球面調和関数で多重曲展開し，式 (2.1) の級数を求める．これにより，球面調和関 $Y_{lm}(\theta, \phi)$ の係数 a_{lm} がそれぞれ得られる．

$$L(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} a_{lm} Y_{lm}(\theta, \phi) \tag{2.1}$$

ここで，非負整数 l は球面調和関数の周波数に相当する量（図 2.11 の縦方向）であり，l が大きいほど，細かい周期変化を持った分布に対応する．次に，式 (2.2) により尤度変化の急峻さを表す視点バイアス指標 C を定義する．視点のサンプリングの密度にしたがい，式 (2.1) が分布を近似できる程度の l の閾値 k を設定し, 閾値範囲内のパワースペクトルの積分として求める．本稿では，実験的に $k = 5$ と定めた．

$$C = \sum_{l=0}^{k} \sum_{m=-l}^{l} {a_{lm}}^2 \tag{2.2}$$

2.4.4.2 視点バイアス指標を用いた評価実験

図 2.12: 視点尤度ヒートマップの平均尤度，視点バイアス指数の値

2.4.1.2 節の実験で用いた図 2.8 の物体のうち分布の差異が大きかった 5 カテゴリ（1.monitor，2.mug，4.bottle，7.chair， 8.airliner；それぞれのカテゴリの 1 行目のインスタンス）に対し，前節で述べた視点バイアス指標 C の値を計算した結果を図 2.12 に示す．図中，視点バイアス指標 C を縦軸，認識尤度の平均値を横軸とした．各プロット点は，カテゴリの違いを色で示し，認識モデルの違いをマーク形状で示す． 2.4.1.1 節と同じ 1,438 視点を計算に用い，式 (2.1)，(2.2) における l の閾値 k を 5 に設定した．

まず，視点バイアス指標の妥当性について，図 2.8 の視点—尤度ヒートマップの定性的評価との比較により検証する．図 2.8 の 1.monitor のヒートマップは，正面視点及び周辺視点の領域が赤く（認識尤度が 1 に近い），その他の視点は青い（認識尤度が 0 に近い）ので，非常に急峻な変化を持っていると言える．これは，図 2.12 の monitor の視点バイアス指標 C が他のカテゴリと比べ最も大きい値となっていることと合致する．次に 2.mug に関しては，取っ手が見えるような両側から観測した視点の尤度が他の視点と比較して圧倒的に高

い傾向があり，視点バイアス指標 C も相対的に大きく，1.monitor に次ぐ値となっていることがわかる．4.bottle に関しては，真上と真下の領域が青く，これに対して他の視点の尤度が相対的に高いが，2.mug と比べ変化が緩やかであるため視点バイアス指標 C の値も小さくなっている．7.chair および 8.airliner は様々な視点に対して万遍なく高い尤度を示しており，視点バイアス指標 C の値も他と比較して低くなっている．以上のことから，視点バイアス指標 C により概ね狙い通りに視点バイアスの状態が表現されていると言える．

次に，データセットや認識モデルによる違いを視点バイアス指標 C によって考察する．図 2.12 中，右下に向かうほど視点の違いによる尤度変化が小さく平均尤度も高いため，物体認識において理想的な状態であることを示す．8.airliner がこの状態に最も近く，図 2.8 の視点—尤度ヒートマップが良好な状態であることと一致する．これに対して，左下に向かうほど，全ての視点で尤度が低く，認識が困難な状態であることを示す．左上方向に関しては，視点による尤度変化が大きく，平均尤度が低い状態であるので，平均的に尤度が低い中に尤度が高い部分が急峻な状態で存在する状態であると考えられる．1.monitor がこの状態に最も近く，図 2.8 の視点—尤度ヒートマップにおいても，正面のみ局所的に尤度が高い様子が観測できる．このような状態は，僅かな視点移動で急激に尤度が変化する可能性があることを示しており，視点の設定に注意が必要な状態であると言える．

次に，視点バイアス指標 C を用いて認識モデルによる視点バイアスの特性の違いを考察する．図 2.12 より概ね，より新しい認識モデルの順（AlexNet $<$ VGG19 $<$ ResNet152）で特性が改善されていることがわかる．特に，視点バイアス指 C 値が大きかった 1.monitor および 2.mug は，視点バイアス指標 C 及び平均尤度が同時に大きく改善され，データがグラフ中で右下に向かっていることがわかる．

以上のように，視点バイアス指標 C を導入することで，カテゴリ毎やインスタンス毎の視点バイアスの分析や，データセットレベルのバイアス分析を容易に行うことが可能になる．更に将来的には，データセット間のバイアス比較によるデータセットの性能評価なども可能になることが期待される．

2.4.5 考察

人間による撮影・アノテーションの過程で生ずる視点バイアスが，既存のよく用いられるデータセット中に実際に存在することを示した．このようなバイアスの存在は，誰もが漠然と気づいていながらも，これまで詳細な分析が行われていなかったという問題があった．今回はまず，標準的によく用いられるデータセットでトレーニングされた認識モデルによる認識結果に，視点バイアスが存在することを示しその傾向を分析した．また，データセットの視点分布を手動で調べることによって，データセット中に実際に視点バイアスが存在することを直接的に明らかにした．更に，球面上の分布で表現される視点バイアスの特徴を，良好に表現可能なスカラー値の指標を提案し，カテゴリ間，インスタンス間，データセット間の特性の違いの比較検討などを容易にした．視点—尤度ヒートマップおよび視点バイアス指標を手がかりとし，尤度が低い視点のデータを追加することで，データセットにおける視点バイアスの改善を図ることも今後検討可能である．

本節で問題とした視点バイアスは,（人間の傾向を反映しない）ロボットによる計測など，産業現場におけるさまざまなアプリケーションにおいて，その性能に決定的な影響を及ぼす可能性がある．今後視点バイアスがデータセットの品質を評価する際の１つの指標となり，様々なシステムの性能向上に寄与すれば幸いである．

2.5 多視点を前提とした尤度投票型物体認識

2.5.1 概要

物体認識技術は近年目覚ましい進歩を遂げている．ILSVRC2012 [1] で AlexNet [8] が圧倒的な性能で優勝して以降，いわゆるディープラーニングに基づく手法が主流となり，画像の分野では主に DNN が用いられ，その代表的なモデルには VGGNet [9] や ResNet [2] が挙げられる．近年では劇的な精度向上が見られ，2015 年には人間の精度（エラー率 5.1%）を超えるモデルが提案され，さらに 2016 年の ILSVRC ではエラー率 3.0%のモデルが提案された．

一方，例えば上記の ILSVRC においては，推定結果のうち尤度の高い上位 5 位までに正解が含まれれば，その問題に正解したとしてカウントする TOP5 方式を採用するなど，純粋な認識性能の評価にフォーカスしたものとなっている．しかし，実利用を想定すると，TOP5 までの絞り込みでは不十分なケースが多く，何らかの工夫により正解が最上位として得られるようにする必要がある．また，実環境で動作するシステムにおいては，照明が極端に変化するなどの環境条件や，対象をどのような角度から観測するかによって対象のアピアランスが大きく変化するなどの問題があり，現状の物体認識性能評価用データセットで高精度が得られている手法が必ずしも望ましい性能を発揮するとは限らない．ILSVRC においても 2 次元的に画像を識別するコンペティションは終息に向かい，実空間における物体やシーンを認識するテーマに向かっている．世界的に見ても 3 次元的な理解は共通課題として挙げられており，まさにこれから注目される分野である．

また，節 2.4 での検討により，認識モデルの視点バイアスが存在することを明らかにした．実環境アプリケーションにおいて，複数視点からの観測が可能であることから，1 視点に頼らずに多視点の認識結果をベースとした認識フレームワークが認識精度の向上や識別の安定性に有用であることが示唆された．

以上のところから，移動ロボットなどカメラが移動することによって対象を複数の視点から観測できる条件であることを前提とし，(1) まず物体を 2 次元画像の物体認識手法によって検出し，(2) カメラの移動中，物体を 3 次元的に捉えつつ 2 次元画像による物体認識の結果をターゲットの座標に投票し続ける，ことによって 3 次元空間中に物体種別を得ることを基本とする．これにより様々な視点からの認識を含む複数回の認識結果が統合され，統計的に安定かつ視点に依存しない認識結果が得られる．

以下，提案手法の詳細および実環境における実験を詳細的に述べる．

2.5.2　多視点を前提とした尤度投票型物体認識手法

2.5.2.1　手法の概要

図 2.13: 尤度投票型物体認識の概念図

図 2.13 に提案手法の概要を示す．カメラが空間中を移動しながら 2 次元画像による物体認識を繰り返し，同時に 3 次元空間中におけるカメラの位置姿勢の推定を行うことで，複数の 2 次元画像の認識結果を 3 次元空間に結びつけるフレームワークを基本とする．このフレームワークにより，様々な姿勢から観測した対象の複数の認識結果を統計的に扱い，活用することが可能になる．以降の節では，フレームワークを構築するために必要となる，物体認識手法，カメラ位置推定手法，観測対象の切り出し手法についてまず述べる．

2.5.2.2　物体認識手法

本フレームワークでは任意の物体認識手法を用いることができるが，3 次元空間に対する投票のため，物体のカテゴリのほか，対象の 2 次元画像上での領域，あるいは，バウンダリボックスが推定可能である方法が望ましい（後述する観測対象の切り出し処理があるため，画像内の近い位置に複数の物体が存在しない場合は，対象の位置が推定できない方法も利用可能である）．今回は，物体のカテゴリと 2 次元画像上での位置をバウンダリボックスで得ることができる YOLO v2 [12] および Faster R-CNN [10] を実験に用いた．これらの手法は，CNN の一種で，高速かつ高い認識精度が得られることから，近年一般物体認識の分野において非常によく用いられている．

なお，本フレームワークは投票型であるため，複数の物体認識手法を同時に用いた投票も可能であるが，今回は基本的な性質を確かめるために，一つの手法を用いて投票を行った．

2.5.2.3 カメラ位置の推定

本研究では，認識結果を 3 次元空間中の座標に結びつけるために認識時のカメラの位置姿勢を正確に推定する手法が必要となる．基本的にカメラの位置姿勢を正確に求めることができれば，手法は任意のものでよいが，今回は認識の過程でカメラの位置姿勢を求めると同時に対象の密な 3 次元点群を取得し，結果をわかりやすく 3 次元表示することを目的として，RGB-D カメラを用いた RGB-D SLAM [20] を用いた．そのアルゴリズムの概要は次の通りである．

現在時刻 T におけるフレームを $Frame_T$ とし，単位時間前のフレームを $Frame_{T-1}$ とする．次にそれぞれの RGB 画像から SIFT (Scale-Invariant Feature Transform) [21] の特徴点の集合 K_T および K_{T-1} を求め，両者の特徴点の対応付けを行う．具体的には，K_T に含まれるある特徴点 p と最も類似性が高い K_{T-1} に含まれる特徴点 q があるとき，q にとっても p がもっとも類似性が高い場合のみ，p と q は対応関係にあるとする（クロスチェック）．この結果，対応づいた特徴点のみを抽出し，更に Depth 値を用いて 3 次元点群化した集合をそれぞれ，A および B とする．

次に，3 次元点群間の関係から RANSAC (Random sample consensus) [22] と SVD（Singular Value Decomposition）を用いて，カメラの回転行列を求める．具体的には，まず RANSAC において，A および B からランダムに 3 つのペアを選択し，a および b とする．

$$a = \begin{bmatrix} A(l).x & A(m).x & A(n).x \\ A(l).y & A(m).y & A(n).y \\ A(l).z & A(m).z & A(n).z \end{bmatrix} \tag{2.3}$$

$$b = \begin{bmatrix} B(l).x & B(m).x & B(n).x \\ B(l).y & B(m).y & B(n).y \\ B(l).z & B(m).z & B(n).z \end{bmatrix} \tag{2.4}$$

式 (2.3), (2.4) において，l, m, n はそれぞれ対応付けのインデックスである． a と b から SVD を用いて仮の回転行列 R_{temp} を求め，それを用いて A を回転させたときの B の各対応点との二乗誤差を以下のように評価する．

$$\epsilon = \sum_{i=1}^{n} \|(B(i) - \bar{B}) - R_{temp}(A(i) - \bar{A})\|^2 \tag{2.5}$$

式 (2.5) において，$\bar{A}$ および $\bar{B}$ はそれぞれ A および B の重心である．このようなランダムサンプリングと二乗誤差の評価を 10,000 回繰り返し，ϵ が最小のときの回転行列を R_{best} とする．次に R_{best} によって A を回転させたとき，B の各対応点との距離 d（式 (2.6)）がしきい値を超える要素をアウトライヤとして A および B から取り除き，A' および B' を得る．

$$d = \|(B'(i) - \bar{B'}) - R_{best}(A'(i) - \bar{A'})\| \tag{2.6}$$

そして A' および B' から SVD により最終的な回転行列 R を得る．また並進行列 T を以下の式 (2.7) のように得る．

$$T = \bar{B'} - R_{best} \cdot \bar{A'} \tag{2.7}$$

2.5.2.4　観測対象の切り出し

観測した 3 次元点群から観測対象の領域を特定することを考える．2.5.2.2 節で述べたように，物体のカテゴリと同時に 2 次元画像上でのバウンダリボックスを得ることができる方法を，異なる視点から複数回用いて対象を特定していく方法も考えられるが，ここではそこで起こる誤認識の改善をテーマとしているため，問題が多重化し複雑になる．またセマンティックセグメンテーションにより，対象を切り出す方法も検討できるが，同様に誤認識や精度に関する考察が別途必要になる．従ってここでは，問題をクリアにするために単純化し，観測対象は床や机上などの平面上に置かれていると仮定して切り出す．

具体的には，画像中に観測対象と平面のみが写っていると仮定し，点群に対し前節で述べた RANSAC によって，任意の 3 点が張る平面のうち最もコンセンサスが得られるものを求めることで，点群中に存在する平面を検出することができる．検出した平面を削除（実際にはノイズを含む点群であるため推定された平面から一定の距離（今回は 10mm）にある点を削除することで，観測対象のみの点群を得る．

なお，複数の視点から観測された観測対象の点群を 2.5.2.3 節で得た回転並進行列によって合成することで，複数視点を統合した観測対象のモデルを作成することができる．モデルは表示に用いるため，2.5.2.3 節で得たパラメータを初期値として ICP (Iterative Closest Point) [23] によって更に精密な位置合わせを行い，外観の品質を向上させた．

2.5.2.5　3 次元空間尤度投票型物体認識

前節までの道具立てに基づき，3 次元空間尤度投票型物体認識を行う．図 2.13 の概略図の通り，視点を変えながら RGB-D カメラによって複数回の認識を行い，そのときのカメラの位置姿勢行列を 2.5.2.3 節で述べた方法により求めておく．観測された複数の 3 次元点群をカメラの位置姿勢行列に基づき合成し，2.5.2.4 節で述べた方法により対象を切り出す．切り出された対象の点群からクラスタリングにより連続領域 F_i を得る．F_i に対して，各視点での認識結果（推定された複数の候補クラスとそれらの尤度）を投票する．このとき，カメラの位置姿勢として 3 次元の回転並進パラメータが得られているため，カメラアングルの変化のみならず，対象とカメラの距離が変化した場合も投票が可能である．

これらの操作によって各認識結果を物体の 3 次元位置に投票することにより，一回の認識では正解以外との尤度が拮抗していたり，例えばペンを先端方向から観測するなど視点に問題があって認識に失敗するようなケースが含まれていたりしたとしても，観測を進めるに従って統計的に正解に近づき，対象をロバストに認識できるようにする．

2.5.3　実験

2.5.3.1　実験の概要

図 2.14: 実験に用いたカメラパス

実物体を用いた実験を行った．図 2.14　のように，認識対象を床上に配置し，カメラを動かしながら撮影した．認識対象としては，(1) 手法の基本的な性質を確かめるために，複雑かつ特殊な形状の物体ではなく，比較的単純な形状の物体であること，(2) 人間であっても観測する角度によっては形状の認識が困難になり得るものであること，という基準に基づき，bowl および bottle の２つを選定した．この選定には，(1) のようなごく普通の物体であっても (2) のような場合が頻繁に起こり得るにも関わらず，これまで物体認識問題では取り扱いが避けられる傾向にあった点に，敢えてフォーカスする意図がある．また，本手法ではシーン中に複数の物体が写っている場合でも 2.5.2.5 節で述べたクラスタリングにより，それぞれの領域に対し独立した投票を行うことが可能であるが，以下の実験では問題を単純化するためにシーン中の物体は 1 つとした．RGB-D カメラとしては，ASUS Xtion Pro LIVE [24] を用いた．撮影したデータを用いて 3 次元空間尤度投票型物体認識および視点—尤度ヒートマップを生成する実験を行った．

2.5.3.2 実験の内容と結果

図 2.15: 撮影済みの bottle 画像例

図 2.16: YOLO v2 を用いた認識尤度遷移図 (bottle)

図 2.17: Faster R-CNN を用いた認識尤度遷移図 (bottle)

図 2.14　のように床上に bottle を置き，カメラを手持ちで図中の黄色線で示すような軌道で bottle 周囲を回り込みながら 30 の視点から YOLO v2 によって認識し尤度の投票を行った．代表的な視点からの観測画像の例を図 2.15 に，投票の結果を図 2.16 に示す．当初 bottle を見下ろす視点から観測が始まったため，bottle (ground truth) の尤度は非常に低く，sportsball や bench が優勢であることがわかる．これに対し，視点を変えながら投票を続けるに従って，視点の影響でいわば偶発的に認識されていた sportsball や bench の尤度は低下し，ground truth である bottle が優勢となっていく様子が確認できる．同様の実験を Faster R-CNN を用いて行った結果を図 2.17 に示す．誤認識による投票傾向はやや異なるものの，本研究の狙いであるところの ground truth（bottle）が最終的に優勢になっていく様子は同様であることがわかる．

図 2.18: 撮影済みの bowl 画像例

図 2.19: YOLO v2 を用いた認識尤度遷移図 (bowl)

図 2.20: Faster R-CNN を用いた認識尤度遷移図 (bowl)

次に，図 2.18 のように床上に bowl を置き YOLO v2 を用いて実験を行った（48 視点）．投票の結果を図 2.19 に示す．ここでも bowl を見下ろす視点から観測が始まったため，当初 frisbee の尤度が顕著であったが，投票を続けるに従って，bowl (ground truth) が安定的に優勢となっていくことがわかる．これに対し，Faster R-CNN による結果を図 2.20 に示す．bottle の場合と同じく，誤認識による投票傾向は YOLO v2 とやや異なるものの，最終的に ground truth (bowl) が優勢になっていく様子がわかる．

以上のことから当初の仮定通り，複数視点からの尤度を 3 次元的に捕捉している対象に対して自動的に投票していくフレームワークにより，ある視点では偶発的に誤認識が発生するとしても，最終的には ground truth であるクラスの尤度を安定的に高めることが可能であることが示された．

なお，この仮定は認識が可能な視点が限定的である（大半の視点で誤認識が発生する）ようなケースでは成り立たない．そのようなケースでは学習データを改善するなどして認識性能そのものを改善する必要がある．しかしながらそもそも視点と認識性能の関係を知ることは容易でなく，学習データをどのように強化すれば良いのか戦略を立てることは難しい．したがって，節 2.4.2 で述べた視点-尤度ヒートマップにより，視点と尤度の関係が可視化されることによって，例えば苦手としている視点の学習データを増強するなどの改善策を検討することが可能になる．

2.5.4　考察

本節では，複数視点からの尤度を 3 次元的に捕捉している対象に対して自動的に投票していくフレームワークの提案について述べた．様々な視点からの認識を含む複数回の認識結果を統合することにより，統計的に安定かつ視点に依存しない認識結果が得られることを示した．本手法では 3 次元データを取得しているため，将来的には 3 次元データも物体認識に活用することが考えられるが，少なくとも現状では，特にクラス内分散が大きい一般物体認識においては，2 次元画像認識が圧倒的に性能が高く，3 次元データによる認識については今後の性能向上が期待されるところである．

2.6　本章のまとめ

本章では，視点と認識尤度の関係を表す視点—尤度ヒートマップを定義し，その関係を定量化・可視化すること目的とし，様々な分析を試みた結果，非常に興味深い知見を得ることができた．具体的には，既存の著名な大規模データセットを用いて学習を行ったネットワークを用いて認識タスクを実行した場合，明らかな視点バイアスが観測できることが明らかになったほか，学習に用いるデータセットや認識に用いるネットワークモデルによって視点バイアスが変化する様子の可視化に成功した．更に，学習に用いるデータセットにおける視点の出現頻度を手作業で分析し，認識モデルが示す視点バイアスとの強い関係性があることを明らかにした．加えて，本来は球面上の分布として表現される視点バイアスの特徴を，近似的によく表現可能なスカラー値の指標を設計した．この指標により，視点バイアスの変化をグラフで表現するなど，様々な定量的分析が可能になることを確認した．

また，上述の視点バイアス傾向の存在を考慮し，多視点を前提とした尤度投票型物体認識フレームワークを提案した．多視点から物体を認識し，認識結果を物体が存在する 3 次元領域に投票することにより，不良の視点により誤認識が抑えられ，統計的に安定した正解を得られる結果を示した．

物体認識技術が現在様々な実環境アプリケーションで応用されている．実環境アプリケーションで安定的に認識を行えるモデルを構築するため，ここで明示化した視点バイアスの存在を十分重視すべきであると考えられる．また，提案の視点バイアス評価指標が今後安定的に動作できる認識システムに有用であると考える．

この章では物体認識タスクにおいて多視点の効果を検証した．多視点認識が他のタスクにおいての有用かを検討するため，実環境をターゲットとした具体的なタスクにおいて複数視点または複数モダリティの有効性と重要性の検討を続けて第三章で述べる．

第3章　多視点・複数モダリティを導入した視覚質問応答

3.1　概要

近年，DNN の進歩により，様々な重要な技術が生まれ，ヒューマン—ロボットインタラクション (HRI) などの実環境アプリケーションへの応用も広がってきた．環境に対しての感知能力と，自然言語によりコミュニケーションする能力のそれぞれを融合した，コンピュータビジョンと自然言語処理のマルチモーダルタスクが HRI アプリケーションにおいて非常に重要である．

視覚質問応答（Visual Question Answering）[25, 26, 27, 28, 29] はコンピュータビジョンと自然言語処理のマルチモーダルタスクの一つである．このタスクでは，画像と画像に関する質問から，画像，質問文及びそれらの関係を理解し，その質問の回答を出力するタスクである．視覚質問応答タスクはさまざまな HRI アプリケーションで重要な役割を果たせると考えられる．

前章により紹介した物体認識タスクと同様に，従来の視覚質問応答手法は，単一の画像に基づいて質問に回答する．従来の視覚質問応答タスクでは，画像を無視して選択肢の正解率を単純に学習してしまう問題が広く指摘されている．このような問題は，判断に必要な画像情報の不足が原因であると想定できる．例えば，実環境では，最適な視点から連続して写真を撮るのは難しく，物体は遮蔽される場合が多く存在し，一枚の画像から質問を正しく回答できない場合がある．このような問題を解決するために，この章では HRI アプリケーションなどで複数視点から観測が可能であることを考慮し，視覚質問応答に複数視点・モダリティの認識能力を導入し，このタスクにおいての有用性・重要性を検証する．

この章の構成を以下に示す．

まず，節 3.2 で視覚質問応答タスク及びこのタスクで広く使われているデータセットおよび既存手法を紹介する．節 3.3 にでは，3 次元認識の従来手法を述べる．節 3.4 で既存の単眼 RGB 画像から行う視覚質問応答タスクの入力を拡張し，RGB 画像とデプス画像から質問を回答するデータセットおよびネットワーク構造と評価実験を述べる．節 3.4 によりデプス画像の有効性が示されたことから，節 3.5 で更に従来の入力単眼 RGB 画像を多視点画像に調整し，多視点画像から質問応答を行うデータセット，ネットワーク構造及び実験を述べる．更に，節 3.6 で観測視点を能動的選択できる仕組みを節 3.5 で述べた多視点視覚質問応答に導入し，質問応答精度を保ちながら，必要な視点を大幅に減らせるネットワーク構造を述べる．節 3.6 までの実験で単純な幾何物体を持った CG データセットをメインで実験に使用し，節 3.7 で実物体モデルを用いた視覚質問応答データセット自動生成プロセスを提案する．これにより，更なるモデルの性能評価や向上などに貢献できると考えられる．最後に，節 3.8 で本章をまとめる．

3.2 視覚質問応答

3.2.1 視覚質問応答タスク

図 3.1: 視覚質問応答タスクの概念図

近年，大規模画像認識用データセットの開発，DNN 手法の発展およびグラフ処理ユニット GPU の開発により, 様々な従来難しいとされるタスクが検討されるようになった．従来，画像認識と自然言語処理がそれぞれ別々で検討されてきたが，近年複数の画像認識と自然言語処理のマルチモーダルタスクが提案された．

視覚質問応答 (Visual Question Answering) [25] タスクが画像認識と自然言語処理のマルチモーダルタスクの一つである．図 3.1 にこのタスクの概念図を示す．画像と画像内容に関する自然言語の質問の入力から，視覚質問応答手法はその質問の回答を推定する．このタスクを行うために，画像，質問，および画像と質問間の関係性の理解が必要となる．

また，近年ビデオから質問を回答する video question answering [30, 31] タスクや，テキストが含まれる画像からテキストの内容の理解をベースとした text question answering [32, 33] や 3 次元環境中に自己ナビゲーションを行いながら，指定された質問を回答する embodied question answering [34, 35, 36, 37] などのタスクも提案された．

視覚質問応答タスクは環境に対しての感知および言語の理解を同時に行うため，画像理解と自然言語理解の単独のタスクと比べて，環境認知とコミュニケーション能力両方が重要視される Human-Robot Interaction (HRI) システムにおいて適応性が高いと思われる．また，視覚質問応答タスクが検索システムや盲人のためのアシスタントシステムの開発にも一つ鍵となる技術となる [38, 39]．

3.2.2 視覚質問応答タスク用データセット

3.2.2.1 VQA v1 と VQA v2

図 3.2: VQA v1 データセットの例 (Agrawal et, al. [25] により)

VQA v1(図 3.2) [25] は最初に提案された VQA データセットの 1 つである．VQA v1 データセットは MS COCO [7] データセットから サンプリングされた画像とシミュレーション環境で作成した CG 画像の 2 種類から構成される．VQA v1 には画像ごとに，3 ペアの質問文-回答文が AMT (Amazon Mechanical Turk) により手作業で収集された．VQA v1 には fine-grained 認識，物体認識，行動認識，知識ベース reasoning 及び commonsense reasoning などに関する質問が設けられている．しかし，近年 VQA v1 データセットに強いバイアスが存在することが指摘された．例えば，VQA v1 データセットでは ”What sport… ” で始まる質問に対し，” tennis ” という回答がデータセットに 41%程度占める．このようなデータセットで学習を行う場合，データセットのこのようなバイアスが学習されてしまう傾向がある．

図 3.3: VQA v2 データセットの例 (Goyal et, al. [40] により)

VQA v1 データセットのバイアスを解消するため，Goyal らが VQA v2 [40] データセットを提案した．図 3.3 に示しいるように，VQA v2 の質問ごとに回答が異なる類似した画像ペアが設けられている．このような設計により，質問文を回答するのに画像から何らかの特徴を利用しない限り正しく回答することが困難である．しかし，VQA v2 データセットの学習及びテストセットの回答の全体的な分布が非常に類似しているため，学習セットの回答分布を学習し，テストセットを正しく回答できる場合が多いような問題点が残っている．

3.2.2.2 CLEVR

Q: Are there an equal number of large things and metal spheres?
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
Q: How many objects are either small cylinders or metal things?

図 3.4: CLEVR データセットの例 (Johnson et, al. [41] により)

従来の VQA v1 と VQA v2 データセットにデータセットバイアス含まれることが従来問題視されている．データセットのバイアスを無くすために，Johnson らが Compositional Language and Elementary Visual Reasoning (CLEVR) [41] データセット及び自動生成プログラムを提案した．CLEVR データセットで質問の回答の分布が一様になるように調整されている．

図 3.4 に示しているように，CLEVR データセットは CG 画像と自動生成プログラムで作成した質問—回答文ペアから構成される．具体的に，CG 画像に 2 個から 10 個までのランダムの数の物体が配置される．物体ごとにランダムな属性（色，形状，サイズ，材質）が与えられる．CLEVR でデフォルトで定義されている色が 8 種類 (blue,red,green,yellow,purple,cyan,gray, brown), 形状が 3 種類 (cube, cylinder, sphere), サイズが 2 種類 (big, small)，そして材質が 2 種類となる (matte, metallic)．また，物体の属性，サイズ，色，形状を推定する query 質問や，物体の数に関する counting 問題などが設けられている．現在，VQA v1 と VQA v2 データセットと同様に，CLEVR データセットは，視覚質問応答手法のパフォーマンス評価にに広く使用されている．

3.2.3 視覚質問応答手法

3.2.3.1 LSTM Q+I

図 3.5: LSTM Q + I のフレームワーク図 (Agrawal et, al. [25] により)

VQA v1 を提案した Agrawal らが LSTM Q + I [25] 手法を提案した．図 3.5 で LSTM Q + I のフレームワーク図を示す．LSTM Q + I では RGB 画像と質問をそれぞれ CNN と LSTM により特徴抽出を行う．次に，得られた RGB 画像特徴と質問特徴をそれぞれ全結合層に経由し，point-wise multiplication により統合される．最後に，統合された特徴から，全結合層とソフトマックス関数により一つの回答が予測される．このような構造が現在あらゆる VQA 手法の基準となっている．

3.2.3.2 SAN

図 3.6: SAN のフレームワーク図 (Yang et, al. [42] により)

2016 年，Yang らが視覚質問応答手法 SAN(Stacked Attention Network) [42] を提案した．SAN でアテンションメカニズムが導入され，質問特徴によりウェイトを計算し，計算されたウェイトにより画像領域にアテンションが付けられる．このような構造により質問に関連する画像領域から情報を抽出することが可能になる．また，Yang らアテンションを付ける回数に関する実験を行い，実験結果により，図 3.6 に示しているような 2 層のアテンションを付けた SAN モデルが最も良い性能を得た．現在，このアテンションメカニズムが視覚質問応答の一つ基本的な構造として広く使われている．

3.2.3.3 FiLM

図 3.7: FiLM layer のイメージ図 (左)，FiLM を導入した視覚質問応答フレームワーク図 (右) (Perez et, al. [43] により)

Perez らが FiLM(Feature wise Linear Modulation) [43] を提案した．FiLM のコアとなる feature modulation 操作は質問文から特徴抽出を行い，抽出特徴によりアフィン変換のパラメータを計算し，RGB 画像特徴にアフィン変換を加える (図 3.7 (左))．

具体的に，まず，FiLM は入力質問 Q から，2 つのネットワーク f と h によりアフィン変換のパラメータ $\gamma_{i,c}$ と $\beta_{i,c}$ を推定する．ここで，i, c は画像特徴抽出器の i^{th} 層目の c^{th} 個目の画像特徴マップを指す．以上のプロセスは下の式により表示できる：

$$\gamma_{i,c} = f(Q),\ \beta_{i,c} = h(Q) \tag{3.1}$$

次に, 画像特徴マップを $F_{i,c}$ より表示し，予測されたパラメータを適用した feature modulation 操作は下の式により表す：

$$F_{i,c}{}' = \gamma_{i,c} \cdot F_{i,c} + \beta_{i,c} \tag{3.2}$$

このような操作を積み重ねることにより，質問文と画像の特徴の関連性が学習しやすくなり，その結果 FiLM 手法は visual reasoning において性能が抜群で，CLEVR データセットにおいて最も良い精度を達成した．

以上の feature modulation 操作を図 3.7 (右) で示す全体的フレームワークに適応される．FiLM の著者は 4 つのレイヤーから構成される ResBlock を使用して画像特徴を抽出する．また，ImageNet で事前学習した ResNet ネットワークにより入力画像から画像特徴を抽出する．画像特徴抽出ネットワークの最終出力を分類器ネットワークに適用し，答えを予測する．

3.3　3 次元認識

3.3.1　MVCNN

図 3.8: MVCNN のフレームワーク図 (Su et, al. [44] により)

Su らが 3 次元物体モデルの多視点画像から物体認識を行うフレームワーク Multi-view Convolutional Neural Networks (MVCNN) [44] を提案した．図 3.8 で示すように，MVCNN ではまず多視点画像を別々に CNN により特徴抽出する．次に，別々に抽出された多視点画像の特徴を view pooling 操作 (max pooling と average pooling 両方使える) に結合する．最後に，結合された特徴を更に CNN に入力し，最後に，全結合層とソフトマックス関数により物体のクラスを予測する．

MVCNN の構造により物体の多視点画像から統合された物体特徴を得られ，既存の 3 次元モデルデータセットで単独の視点からの認識モデルより優位な認識性能を得た．

3.3.2 SRN

図 3.9: SRN のフレームワーク図 (Eslami et, al. [45] により)

Eslami [45] らが多視点の画像および視点からシーンの 3 次元特徴表現を生成し，シーンの 3 次元特徴表現と与えられた視点に対応する画像を予測可能なネットワーク Generative Query Network (GQN) を提案した (図 3.9).

GQN は 2 つのネットワークから構成される：(1) scene representation network (SRN): 多視点の視点情報と画像からシーンの 3 次元情報を推定するネットワーク；(2) render network (RN):シーンの 3 次元情報と指定された視点情報からその視点の画像をレンダリングするネットワーク.

K 視点から観測されたシーンを s_j により表示する．式 (3.3) により，シーンの観測 o_j を定義する.

$$o_j = \{(x_j^k, v_j^k)\}_{k=0,\ldots,K-1} \tag{3.3}$$

式 (3.3) で x_j^k は視点 v_j^k(5 次元のベクトル：$(\mathbf{w}, \mathbf{y}, \mathbf{p})$, $\mathbf{w}$, $\mathbf{y}$, $\mathbf{p}$ は 3 次元の位置とカメラのヨーとピッチを表す) から観測された画像を表す.

指定された視点 m からその視点に対応する画像 x^m の推定の学習プロセスにより，SRN ネットワーク $r_j = f_{\mathrm{GQN_SRN}}(o_j)$ とレンダリングネットワーク $g_{\mathrm{GQN_RN}}(x^m|v^m, r_j)$ がジョイントで学習される.

学習済みの SRN により，多視点の画像からシーンの 3 次元情報が含まれた特徴の抽出ができる．GQN ネットワークは現在指定された視点からの画像生成や画像から観測視点を予測するなどのタスクで広く引用されている.

3.3.3 PointNet

図 3.10: PointNet のフレームワーク図 (Qi et, al. [46] により)

PointNet [46] が 2016 年に Qi らにより提案された点群データを取り扱うニューラルネットワークである．PointNet が点の 3 次元座標 (x, y, z) のリストデータの形式の点群から，点群の情報を表す特徴表現を抽出する．PointNet により抽出された点群特徴が物体認識，点群に対してのセマンティックセグメンテーション (点ごとに物体のラベルを付けるタスク) およびパーツセグメンテーション (物体のパーツごとに物体およびパーツのラベルを付けるタスク) などに活用されている．

点群が画像と比較して，3 つの特徴がある．(1) unordered: 点の順番が変化しても点群が変化しない；(2) interaction among points: 点群の性質が点と点の位置関係に決定される；(3) invariance under transformations: 点群に対して回転や並進などの操作を加えても点群が変化しない．従来画像認識のための CNN が点群に適応することが難しいと考えられている．

Qi ら以上の 3 つの特徴を対応するため，PointNet 構造を提案した (図 3.10)．PointNet で入力された点群データとネットワークの中間層の出力を T-Net に入力し，点群データおよび特徴のアライメントを行う．T-Net により以上の特徴 (3) を対応した．また，PointNet で multilayer perceptron (MLP) と global pooling，点群の集合特徴を得る．それにより上述の (1) と (2) を対応した．PointNet が 3 次元 CNN と比べて，大幅に実装のメモリーおよび時間コストを削減できる．現在，様々な点群データの学習を取り扱う応用に PointNet が広く使用される．

3.4 RGB-D 画像を用いた視覚質問応答

3.4.1 概要

図 3.11: RGB-D 画像を用いた視覚質問応答の概念図

視覚質問応答タスクにおいて，詳細な幾何情報が必要な質問や複雑な空間 reasoning が必要な質問が他の種類の質問と比べ精度が低下している傾向がある．例として，counting 問題が挙げられる．更に，実環境で動作する場合，極端な照明条件や撮影角度により，RGB 画像から十分必要な情報を得られない場合がある．一方で，デプス画像は，照明に対して相対的にロバストであり，RGB 画像と比べエッジ情報を抽出しやすい特徴がある．そのため，視覚質問応答タスクにおいてデプス画像が重要な補助情報として用いられると想定できる．

以上のことから，図 3.11 で示しているように，我々は RGB-D 視覚質問応答のフレームワークを提案した．また，従来の視覚質問応答データセット CLEVR にデプス画像自動生成の処理を追加し，RGB-D CLEVR データセットを生成した．RGB-D CLEVR データセットでの実験により，デプス画像が有効的に視覚質問応答手法の精度を改善できることを実証した．特に，counting 問題において従来のフレームワークと比べ大幅な精度向上を実現できた．更に，Attention マップの可視化分析から，デプス情報が attention の改善に効果も観測できた．実環境で動く視覚質問応答システムにおいて，幾何情報が非常に重要であることから，今後の視覚質問応答システムを構築する場合，デプス情報を利用することが性能向上において有用であることを示した．

3.4.2 提案構造

現在の視覚質問応答手法は基本的に (1). 画像特徴抽出器；(2). 質問文特徴抽出器；(3). 画像特徴と質問文特徴フュージョン器；(4). 結合した特徴により回答を予測の 4 つの構造から構成される．本稿では RGB-D 情報を使用するため，上述の (1). 画像情報抽出器の構造で

図 3.12: RGB-D 画像を用いた視覚質問応答のフレームワーク図

RGB-D の情報の統合を行う．また，画像特徴抽出器の出力特徴量の次元をコントロールすることで，RGB-D フュージョンを様々な視覚質問応答手法に適応できる．本稿では節 3.2.3 で述べた 3 つの手法：LSTM Q + I, SAN, FiLM に RGB-D 情報統合の構造を実際に導入し実験を行った．

図 3.12 で示しているように， RGB-D 情報フュージョンを VQA ベース構造の画像特徴抽出器に導入することで， RGB-D 情報をあらゆる視覚質問応答手法に適応する．具体的に，まず RGB 画像で学習済みの特徴抽出器を用いて RGB とデプス画像に対しそれぞれ別々で特徴抽出を行う．次に，質問文特徴を条件にデプス画像特徴に対し線形変換を加える (ques_c_concat)．その次，RGB 画像とデプス画像の特徴を連結する (feature-map-wise で連結)．以下，ques_c_concat 構造の詳細を述べる．

異なる質問に対し，RGB 画像特徴がより重要視すべきである場合 (例えば query color 問題) とデプス画像特徴を重要視すべきである場合 (例えば counting 問題) がある．そのため，図 3.12 で示しているように， 我々は質問文特徴を条件に RGB 画像とデプス画像特徴をコントロールできる構造を提案する．具体的に，Affine レーヤー f, h を用いて質問文特徴 $feat_Q$ から重み w とバイアス b を計算する．得られた w, b で式 (3.4) (3.5) によりデプス特徴 $feat_D$ の変換を行う (RGB 特徴に対しても同様な操作ができるが，RGB とデプスの特徴の trade-off が片方だけの変換でコントロール可能)．

$$w = f(feat_Q),\ b = h(feat_Q) \tag{3.4}$$

$$feat_D{}' = w \cdot feat_D + b \tag{3.5}$$

表 3.1: RGB-D CLEVR データセットでの定性評価.

Method	Overall Accuracy	Count	Query / Compare			
			Size	Color	Material	Shape
LSTM Q + I	52.43	43.05	57.18	**40.62**	55.73	49.45
LSTM Q + I-ques_c_concat-HHA	52.70	43.97	58.77	37.14	**56.82**	**50.99**
LSTM Q + I-ques_c_concat-Depth	**52.84**	**44.46**	**59.17**	38.53	55.89	48.99
SAN	67.05	48.53	78.40	71.84	77.11	73.86
SAN-ques_c_concat-HHA	**71.35**	54.16	**82.66**	**75.71**	**80.37**	**77.28**
SAN-ques_c_concat-Depth	70.30	**54.43**	80.30	73.74	78.37	76.33
FiLM	93.04	84.26	97.47	96.17	**97.35**	96.03
FiLM-ques_c_concat-HHA	**93.44**	**85.62**	98.10	95.85	96.89	**96.20**
FiLM-ques_c_concat-Depth	93.38	85.44	**98.37**	**96.32**	96.28	96.08

3.4.3 実験

図 3.13: RGB-D 視覚質問応答データセットの例

RGB-D 視覚質問応答手法を評価可能にするため，CLEVR データセットの自動生成プログラムにデプス画像生成処理を追加した．このような設定で，RGB-D CLEVR データセットを生成した．データセットのインスタンスの例を図 3.13 に示す．具体的に，学習セットは 40,000*3 枚の画像 (RGB 画像，HHA デプス画像，Colorized デプス画像) 及び 400,000 質問文-回答文ペアから構成される．テストセットに 10,000*3 枚の画像 (RGB 画像，HHA デプス画像，Colorized デプス画像) 及び 100,000 質問文-回答文ペアが含まれる．

本稿で節 3.2.3 で紹介した LSTM Q + I, SAN, FiLM をベースライン手法とする．この 3 つの手法に対し，前節で述べた RGB-D フュージョン手法 ques_c_concat を用いて RGB-D 画像を取り扱えるようにした．また，HHA と Colorized デプス画像を実験に用いた．上述の実験設定で，LSTM Q + I ベースな手法に対し 20 エポック学習し，SAN, FiLM ベースの手法に対しそれぞれ 50 エポックを学習した．

表 3.1 に RGB-D CLEVR データセットのテストセットでの全体的な精度と異なる質問タイプの精度を示す．まず全体的な精度の結果を分析する．ベースライン手法と比べ，デプス情報を用いて全体的な精度において精度向上が実現できた (最大+4.3%)．更に，FiLM 手法に対しも，デプス画像の追加での性能向上ができた．このような結果から，視覚質問応答タスクにおいてデプスの有用性が明らかになった．また，HHA デプス画像 (HHA) は

Colorized デプス画像 (Depth) と比べ少し精度が上回っている．異なる質問タイプの精度に対し，表 3.1 からデプス情報が counting 問題 (最大+5.9%) と size-related 問題において他の質問タイプと比べ大幅な精度向上が得られた．Counting 質問は現在の VQA タスクにおいて最も困難質問タイプであるため，counting 質問での精度向上でデプス画像の有用性が十分重要視すべきであると考えられる．このような結果から，デプス情報から幾何情報とエッジ情報をより抽出しやすい，counting や size-related 問題に特に有利であることが原因であると想定できる．

図 3.14: RGB-D 視覚質問応答の結果例

図 3.14 に attention マップの可視化の例を示す．ここで SAN 手法と HHA デプス画像を使用した SAN_ques_c_hha 手法の結果例を示す．図 3.14 から，デプス画像を使用した手法が物体領域に attention を集中しやすい傾向が観測できる．デプス画像の幾何情報とエッジ情報がより明らかで，デプス画像を用いたモデルが質問に密に関係する領域に集中しやすいことが想定できる．

3.4.4 考察

従来の視覚質問応答タスクでは単眼の RGB 画像から質問応答を行う．実環境アプリケーションでは，RGB 画像から十分質問に回答するための幾何情報を得られない場合がある．ここで，我々はデプス画像から質問応答タスクにおいて有用な幾何情報を抽出のに有用であると想定し，デプス情報を視覚質問応答コンテキストに導入し，RGB-D 視覚質問応答のフレームワーク及びデータセットを提案した．

提案した RGB-D 視覚質問応答データセットでの実験の結果により，デプス情報を利用することで，有効的に視覚質問応答手法の精度を向上できることが示された．特に，従来難しいとされる counting 問題に対し，RGB-D を用いることで従来の手法と比べ大幅な精度向上が実現できた．また，質問文を条件に RGB とデプス画像の特徴をコントロールする構造を用いることで，有効的に RGB-D フュージョンを行えることを示した．更に，attention マップの可視化分析により，デプス画像を用いて質問と密に関係する画像領域への attention 集中に有用性を示した．

以上により，デプス画像が視覚質問応答タスクにおいての効果を示した．このような結果から，幾何情報が視覚質問応答タスクにおいて重要であることを想定できる．その一方，実環境アプリケーションでは物体オクルージョンなどにより，一つの視点から質問応答に必

要な視覚情報が得られない場合が想定できる．以上のところから，次の節で，我々は更に多視点から観測する設定で視覚質問応答タスクの検討を行う．

3.5 多視点に対応した視覚質問応答

3.5.1 概要

図 3.15: 多視点に対応した視覚質問応答の概念図

従来の質問応答タスクでは，1 枚の RGB 画像から質問応答を行う．ただし，実環境アプリケーションでは，極端な照明条件，不適切な観測視点，又は物体オクルージョンなどにより，1 枚の RGB 画像からの質問応答に必要な視覚情報を得られない場合がある．

また，複数の観測視点からシーンを観測し，それぞれの視点で質問応答を行い，すべての視点での回答を統合するような仕組みも考えられる．しかし，実環境アプリケーションでは物体間の位置関係の推定が必要な場合が多く存在するため，多視点からシーンを 3 次元的に理解する必要がある．

以上のところから、我々は図 3.15 に示すような多視点認識能力を導入した視覚質問応答フレームワークを提案する．このフレームワークで多視点からシーンを観測し，多視点画像からシーンを 3 次元的に理解し，質問を回答する．

以上のフレームワークを実現可能に，我々は二つ方面から検討を行った．まず，既存の視覚質問応答データセットは単眼画像であるため，我々は複数視点から観測する多視点視覚質問応答データセットを提案した．また，節 3.2.2.2 で述べた CLEVR データセットをベースにし，必要に応じて調整可能な自動データセット構築が行える．

次に，既存の視覚質問応答手法は単眼画像から回答を行っている．我々は複数視点画像からシーンの 3 次元情報を得られる多視点認識手法を視覚質問応答タスクに導入し，シーン

の 3 次元情報の理解を重視した質問応答を可能にした．我々の提案フレームワークにより，密な観測視点からシーンを観測せずに任意の数の視点から質問応答が可能である．

提案データセットでの実験結果により，提案のフレームワークが多視点から回答できることが示された．提案のデータセットに物体オクルージョンが大量に含まれているにもかかわらず，提案手法が 95.4%の精度を達成した．また，従来難しいとされる空間関係の推定が必要な問題や強い物体オクルージョンが含まれるシーンに関しても，提案手法が良好的な結果を示した．実環境アプリケーションでは必ず良好な観測視点からシーンの観測を行えるとは限らないため，従来の単眼画像視覚質問応答タスクと比べて，提案の多視点観測を導入した視覚質問応答タスクが実応用への適応性が高いと考えられる．

以下の節で，提案手法と実験の詳細を述べる．

3.5.2 提案構造

図 3.16: 多視点に対応した視覚質問応答のフレームワーク図

3.5.2.1 全体的構造

従来の視覚質問応答手法は，与えられた質問と 1 枚の RGB 画像から回答を行う．ここで，多視点認識を導入し，多視点画像から得られるシーンの 3 次元情報をベースに回答を行うフレームワークを提案する．多視点画像入力 V を $V = \{v_i\}_{i=1,2,\ldots,K}$ に表す．ここで K を視点番号を示す．

さらに，入力された質問情報を Q と表記し，モデルの全体的なプロセスは次のように書くことができる：

$$a = argmax_a F_\theta(a|V;Q) \tag{3.6}$$

式 (3.6) において、a は予測される答えを示す．全体的視覚質問応答ネットワーク F は入力された V と Q から a を予測する．また，ここで θ はネットワークのパラメータを示す．

更に，F は 2 つのサブネットワークに分けることができる：(1) シーン特徴表現ネットワーク: 入力の多視点画像 $V = \{v_i\}_{i=1,2,\ldots,K}$ からシーンの 3 次元情報を表した特徴表現を得るネットワーク；(2) 視覚質問応答ネットワーク：(1) で得られる特徴表現と質問 Q から回答 a を予測する．

(1) に関して多視点情報をプーリーング操作により統合する手法が考えられる．我々は (1) に関してまず従来複数視点からの物体認識ネットワーク MVCNN を導入した．MVCNN は明示的に多視点からシーンの 3 次元情報を推定しないため，物体間の位置推定が必要となる質問に対して良い性能を得られないことが想定できる．ここで，我々は多視点画像からシーンの 3 次元情報を推定できる SRN 手法も同様に導入した．また，ここで (2) のネットワークは任意の視覚質問応答ネットワークを適用できるが，ここで節 3.2.3.3 で述べた CLEVR データセットで良い性能を達成している FiLM ネットワークを使用した．

以下の節で (1) の部分に MVCNN を導入した View pooling+FiLM と SRN を導入した SRN+FiLM ネットワークの詳細を述べる．

3.5.2.2 View pooling+FiLM

多視点画像情報を統合するには，最も自然な方法として，MVCNN のような構造により多視点画像をプーリーング操作により直接統合する操作が考えられる．ビュープーリングを使用して統合された画像特徴を FiLM に適応し，回答の推定ができる．

我々はまず図 3.16(左) に示しているような View pooling+FiLM 構造を実装した．具体的に，まず ResNet152 により画像特徴を抽出する．その後，それぞれ抽出された画像特徴を FiLM に入力し，最終の分類器に入力する前にそれぞれ特徴量を得る．次に，マックスプーリーング操作により多視点の情報を足し合わす．最後に，足し合わせた特徴を分類器に入力し，回答の予測を行う．

以上のプロセスにより，多視点の情報を統合し，回答を行うことが可能になる．しかし，プーリーング操作が明示的に多視点間の 3 次元対応を行わないため，シーンの 3 次元理解に対しての能力が不足すると想定できる．以上の原因から，我々は複数視点からシーンの 3 次元情報を推定するネットワークを更に導入したフレームワーク SRN+FiLM を提案する．

3.5.2.3 SRN+FiLM

複数視点からのシーンの観測から，シーンの 3 次元情報の推論に，節 3.3.3 で述べた SRN ネットワークが適応できる．従来，SRN は指定された視点の画像復元に広く用いられているが，ここで，SRN を多視点視覚質問応答に導入する．SRN によりシーンの 3 次元情報を得て，それをベースに質問応答を行う．

節 3.3.3 で述べたように，指定された視点 m からその視点に対応する画像 x^m の推定の学習プロセスにより，SRN ネットワーク $r_j = f_{\mathrm{GQN_SRN}}(o_j)$ とレンダリングネットワーク $g_{\mathrm{GQN_RN}}(x^m|v^m, r_j)$ がジョイントで学習される．ここで，このような学習プロセスを行い，SRN の事前学習を行う．また，学習済みのレンダリングネットワーク RN を使用しない．学習済みの SRN を用いて得られるシーンの 3 次元情報を FiLM に入力することで，質問応答を行う．

具体的に，図 3.16(右) に示しているよう，ここでまず多視点画像を SRN に別々に入力し特徴を抽出する．抽出された特徴を足し合わすことにより統合されたシーンの特徴を得る．最後に，FiLM ネットワークで用いられる画像特徴抽出器（ResNet）の代わりに統合された特徴を FiLM に入力し，回答の予測を行う．

この構造により，任意数の視点の画像からの 3D シーン情報を推定し，3 次元情報の理解をベースとした質問応答が行える．このようなフレームワークにより，観測視点が限られた場合での質問応答も可能になる．

3.5.3 実験

3.5.3.1 データセット

図 3.17: 仮想カメラの設定図

CLEVR データセットは，視覚質問応答手法のパフォーマンス評価にに広く使用されている．ここで，CLEVR データセット自動生成プログラムをベースに Multiview CLEVR データセットを構築した．CLEVR データセットに 1 つのシーンに対して 1 台の仮想カメラから撮影を行っている．この研究で，複数視点からシーンを観測するため，CLEVR のシーンに更に仮想カメラの台数を増やしシーンの多視点画像の撮影を行う．

具体的に，図 3.17 で示しているように，シーンに対して 4 台の仮想カメラを設置した．それぞれのカメラをシーンの中心に向かせ，カメラをシーンの平面から 45 度の高さまで設定した．また，4 台のカメラを同じ円周に均等的に設定した．更に，シーンにオクルージョンが含まれるように，4 枚の画像に必ず 1 枚の画像から遮蔽される物体が存在するように物体配置の調整を行っている．

表 3.2: Multi-view CLEVR データセットでの統計情報.

Splits	Image	QA pairs	Question Length
Train	40,000 (∗4)	218,888	10.10
Test	10,000 (∗4)	54,770	10.09

図 3.18: Multiview CLEVR データセットの統計情報

また，オリジナルの CLEVR データセットの設定に物体の属性の調整を行った．データセット内の各シーンに次のように属性を設定をしている：物体の色（red, yellow, blue, purple, green）, サイズ（large，small）, 素材（matte, metallic）, 形状（cube, cylinder, sphere）. また，6 つのタイプの質問を設定している：counting, 物体の存在性に関する existing 質問，および query color 質問から構成される．そのうち，半分の質問には空間的な推論が必要であり，質問の半分は空間的推論プロセスなしで答えることができる．図 3.18 にデータセットの統計情報を示す.

図 3.19: Multiview CLEVR データセットの例

上記のデータセットの設定で，50,000 の CG シーンと 273,658 の質問-回答ペアから構成する Multiview CLEVR データセットを構築した（トレーニング、テスト分割：80%, 20%）．詳細なデータセット統計を表 3.2 に示します。さらに生成された Multiview CLEVR データセットの 4 つのサンプルを図 3.19 に示す．図から分かるように，物体のオクルージョンが含まれるところからこのようなデータセットで正しく質問を回答するためにシーンの 3 次元情報の理解が重要となる．

3.5.3.2 実験設定

実装では，画像サイズを 64 * 64 に設定した．Viewpooling+ FiLM の実装では，ImageNet データセットでトレーニング済みの ResNet152 モデルを用いて画像特徴を抽出した．GQN を 240 エポック事前学習し，学習済みの SRN を使用した．Multiview CLEVR データセットの学習セットで 0.0005 の学習率及び Adam オプティマイザー [47] を使用した実装を行った．また，Viewpooling ＋ FiLM と SRN+FiLM の両方を同様に 0.0001 の学習率で 400 エポックの学習を行った．

表 3.3: Multi-view CLEVR データセットでの定量評価 (* は CLEVR1.0 での結果).

Models	Overall
view pooling + FiLM	84.3
scene representation network + FiLM	**95.4**
FiLM*	**97.7**

3.5.3.3 定量的実験結果

図 3.20: Multiview CLEVR データセットで異なる質問タイプの定量評価

まず、Viewpooling+ FiLM と SRN+ FiLM モデルの全体的精度の評価を行った．表 3.3 に Multiview CLEVR データセットのテストセットにおいての全体的精度を示す．

SRN+ FiLM モデルが Viewpooling+ FiLM モデルを大幅に上回った（+ 11.1%）．Viewpooling 操作は異なる視点間の 3 次元対応を明示的に行っていないため，3 次元情報が必要な質問に対しての性能が限られている．それと比較して，SRN ネットワークは多視点からシーンの 3 次元情報の理解を行い，それにより全体的に性能の向上を達成した. この結果から，3 次元情報が含まれるシーンの特徴表現が Multiview 環境においての有用性を示した．

質問タイプごとの精度を図 3.20 に示す．SRN+ FiLM モデルは Viewpooling ＋ FiLM と比較して，より高いパフォーマンスで空間関係が必要となる質問においての優位性を示した．また，そうでない非空間的な質問の場合，2 つのモデル間の性能の差は比較的小さい結果になった．このようなことから，当初仮定していたように，Viewpooling がシーンの 3 次元情報の推定に性能が制限され，それと比較して，SRN ネットワークが有利であることが示された．

3.5.3.4 入力視点数の影響

ここで，学習時に全周囲視点を使用して実装し，テスト時に異なる入力視点ペアから推論する実験を行った．その結果を表 3.4 に示す．ここで，1,2,3,4 が時計回りの 0, 90, 180, 270 度を回転した視点を示している．

この実験結果から，視点数が増えることにより精度が向上することを確認でき，すべての視点を統合した場合，最も良い精度を得られた．また，同じ視点数の場合に，性能が同レベルである傾向があった．一つの観測視点から行った場合でも，提案手法は一定の精度を示した．このような結果から，提案モデルが視点数が制限された場合も使用できることを示唆した．

表 3.4: Multi-view CLEVR で異なる入力視点の組み合わせの結果.

Input Viewpoint Combination	Overall Accuracy
1	65.8
2	65.0
3	68.0
4	65.4
1,2	79.2
1,3	82.8
1,4	79.0
2,3	80.0
2,4	81.8
3,4	80.2
1,2,3	89.3
1,2,4	89.8
1,3,4	88.7
2,3,4	89.3
1,2,3,4	**95.4**

また，視点“1,3”と“2,4”の視点の組み合わせが視点の組み合わせの“1,2”,“1,4”,“2,3”と“3,4”より高い性能を示した．実験に使用した Multiview CLEVR データセットの視点の設定では，視点“1,3”と“2,4”の組み合わせは，それぞれから比較的離れていた視点を使っている．そのような設定が隣接する視点の組み合わせより多くの情報を提供できることが考えられる．このような実験結果から，提案の SRN + FiLM が効率的に視点を使用している傾向にあることを示した．

3.5.3.5　定性的実験結果

3.5.3.6　実験結果

図 3.21: Multiview CLEVR データセットでの結果例

最後に，8 つのの結果の例を図 3.21 に示す．物体間の空間関係の推定が必要ない質問 (Question 1, 2) において，Viewpooling+ FiLM と SRN+FiLM 両方とも回答に成功した．これらの質問が 1 枚の画像から回答できる傾向があり，比較的に容易であることがわかる．

また，物体間の空間関係の推定が必要な質問 (Question 3, 4, 5, 6) において，Viewpooling+ FiLM が失敗し，それと比較して，SRN+ FiLM が正しく回答を予測できる傾向にある．これらの質問を正しく回答するために，多視点画像からシーンの 3 次元情報の推定が必要となり，SRN が Viewpooling より優位であることがわかった．

反射が強いシーン (Question 7, 8) では，SRN+FiLM モデルも正しく回答を予測できないことがある．これらのシーンが視覚的に複雑度が高く，反射やオブジェクトの影の理解なども必要となる．このような反射や影などが実環境で多く存在するため，反射や影が含まれていていも正しくシーンを理解できるモデルの構築を今後の課題として検討を行う．

3.5.4　考察

既存の視覚質問応答手法は 1 枚の画像と質問文から回答を行う．しかし，実環境アプリケーションでは，照明条件や物体オクルージョンなどにより，1 枚の画像から質問を回答できない場面が多く存在することが想定できる．

我々は上述のような問題点に着目し，多視点からシーンを観測し，質問応答を行う多視点視覚質問応答タスクを提案した．提案タスクのための多視点視覚質問応答データセットおよびフレームワークを提案した．提案フレームワークが多視点から観測されたシーンの 3 次元的理解するため，従来難しいとされる物体間の空間関係推定が含まれる質問に対しても高い性能を示した．更に，提案手法が強い物体オクルージョンが含まれるシーンに対しても有効である傾向が示された．

実環境アプリケーションでは単眼画像から質問応答するための情報が得られない場面が多く存在するため，既存の単眼画像視覚質問応答タスクと比べ，提案の多視点を対応できるフレームワークが実応用への適応性が高いと考えられる．

ここで提案したフレームワークはすべてのシーンを同じ視点から撮影された全周囲画像から質問応答を行う．しかし，実環境アプリケーションでは物体の配置により，全周囲からシーンを観測しにくい場合がある．また，部分観測から必要な視覚情報が十分である場合に，観測を中止し質問回答を行う仕組みがより高効率の観測システムの構築に有効であると考えられる．以上のところから，我々は能動的に観測視点を選択できるフレームワークを提案した．その詳細を次の節で述べる．

3.6 能動的視点選択を対応した多視点視覚質問応答

3.6.1 概要

図 3.22: 能動的視点選択を対応した多視点視覚質問応答の概念図

従来視覚質問応答手法では，1 枚の画像に基づいて質問に回答する．しかし，節 3.5 で述べたように，実際の環境では，最適な視点から連続して写真を撮るのは難しく，1 枚の画像から質問を回答するのに必要な視覚情報が得られない場合が多く存在する．

以上の問題点から，我々はまず節 3.5 で述べたように，視覚質問応答タスクに多視点認識を導入し，多視点から観測されたシーンに基づき，質問を回答可能にした．また，節 3.5 で述べた手法は全てのシーンを同じような全周囲視点設定から観測し，回答を行う．しかし，実環境で物体の配置などにより，全周囲からシーンを観測できない場合がある．また，質問を回答するのに必要な観測が得られた場合，観測を中止し質問に対して直ちに回答する仕組みがより効率的である．リアルタイム処理が重要視される実環境アプリケーションにおいては視点観測の効率が重要視されると想定できる．

以上のところから，図 3.22 に示す能動的に視点を選択できる視覚質問応答フレームワークを提案する．このフレームワークでは入力された質問と観測されたシーンから必要な観測視点を選択して質問を回答できる．

ここで節 3.5.2.3 で述べた SRN+FiLM フレームワークに更に視点選択フレームワークを追加し，3 つのモジュールから構成するフレームワークを提案した．提案のフレームワークを評価するために，12 視点を持った Multiview CLEVR データセットを構築した．このデータセットにおいて，この節で述べるフレームワークは節 3.5 で述べたフレームワークと同レベルな精度 (97.11%（本節）vs. 97.37% (節 3.5:SRN+FiLM)) を達成した．また，SRN+FiLM では全周囲の 12 視点から観測を行い，この節のフレームワークは平均 2.98 の視点を使用した．定性的実験結果により提案の視点選択フレームワークが有効的に観察されたシーンにより視点を効率的に選択できる傾向を示した．更に，実環境で撮影した画像データセットにおいて，提案のフレームワークは SRN+FiLM を大きく上回る精度を得られた (+11.39%)．

提案の視点選択フレームワークが精度を保ったまま，視点選択の効率を向上できる．実環

境アプリケーションにおいて観測効率性が重要視するべきであるため，提案フレームワークは多視点視覚質問応答タスクの実環境への適応性を更に向上できると期待できる．

3.6.2 Deep Q-learning Networks

Q-learning [48] は value-based な強化学習フレームワークの一種である．Q-learning で環境 Environment と環境中に行動 Action を行える Agent が定義される．Agent が環境の状態 s から，実行する行動 a を決める．行動 a により，Agent が報酬 Reward が得られる．また，Agent の行動により環境の状態が変化する．Q-learning で異なる環境の状態において，Agent が各実行可能な Action の Value が Q-table により記録される．Q-table により得られる Value と実際に Agent が行動 Action を行ったあとに得られた報酬 Reward 間の差により，Q-table が更新されるような仕組みでネットワークが学習される．しかし，環境の状態 s が膨大となる場合や連続の状態空間において，Q-learning を実行するのに必要な運行時間とメモリー消耗が巨大になる．

Algorithm 1 Deep Q-learning with Experience Replay

Initialize replay memory $\mathcal{D}$ to capacity N
Initialize action-value function Q with random weights
for episode $= 1, M$ **do**
  Initialise sequence $s_1 = \{x_1\}$ and preprocessed sequenced $\phi_1 = \phi(s_1)$
  for $t = 1, T$ **do**
    With probability ϵ select a random action a_t
    otherwise select $a_t = \max_a Q^*(\phi(s_t), a; \theta)$
    Execute action a_t in emulator and observe reward r_t and image x_{t+1}
    Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$
    Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in $\mathcal{D}$
    Sample random minibatch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from $\mathcal{D}$
    Set $y_j = \begin{cases} r_j & \text{for terminal } \phi_{j+1} \\ r_j + \gamma \max_{a'} Q(\phi_{j+1}, a'; \theta) & \text{for non-terminal } \phi_{j+1} \end{cases}$
    Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ according to equation 3
  end for
end for

図 3.23: Experience replay を用いた deep q-learning のアルゴリズム (Volodymyr et, al. [49] により)

Q-learning でメモリー上で Q-table を記録する方法と比較して，2013 年に Volodymyr らにより提案された deep q-learning (DQNs) [49] はニューラルネットワークにより，環境の状態 s から行動の value を予測する．DQNs のアルゴリズムが図 3.23 に示す．

DQNs で Q-table と対応する action-value 関数 Q を定義する．Q をニューラルネットワークで実装し，Q のパラメータが学習される．また，DQNs ではサイズが N となるリプレイメモリー D が定義され，過去の学習データを D に保存し，学習イテレーションごとに D からランダムに一つのミニバッチを取得し，学習を行っている．

DQNs では動作選択策 ϵ-greedy が実装されている ($\epsilon \in [0.0, 1.0]$). 時刻 t の学習プロセスでまず，ϵ の確率でランダム動作を選ぶ．その他の場合に，環境の状態 s_t を Q に入力し，s_t が対応する Value が最も大きい動作 a_t が選択される．エミュレータに動作 a_t を実行し，報酬 r_t と次の画像 x_{t+1} を得る．次に，s_t, a_t, x_{t+1} を Q に入力し，出力 ϕ_{t+1} を得る．学習データ $(\phi_t, a_t, r_t, \phi_{t+1})$ を D に保存する．パラメータ更新する際に，D からランダムに一つのミニバッチを取得し，一つのターンの最終動作で reward と Q の差によりパラメータを更新し，

その他の場合に減衰係数 γ で調整した reward を使用する．従来の DQNs は画像の入力から policy の学習が行える．

3.6.3 提案構造

3.6.3.1 全体的構造

図 3.24: 能動的視点選択を対応した多視点視覚質問応答のフレームワーク概念図

実際環境アプリケーションでは，全周囲から画像を取得するのが難しい場合がある．また，必要な視覚情報を得られた場合，視点を選択するのが非効率的である．上述のところから，ここで観測されたシーンの情報と質問から能動的に視点を選択できる視覚質問応答フレームワークを提案する．

提案フレームワークの概念図を図 3.24 に示す. ここで節 3.5.2.3 で述べた SRN+FiLM フレームワークに更に視点選択フレームワークを追加した．シーンの入力がデフォルト視点 v_0 と v_0 から観測された画像 x_0 および質問 q．質問情報が LSTM 構造により質問特徴 q' の抽出を行う．また，v_0 と x_0 が SRN に入力され，シーンの 3 次元情報 s_0 を得る．次に，視点選択フレームワークは s_0 と q' から，次の観測視点 (もしくは観測終了動作) を予測する．時刻 t で視点 v_t が選ばれた場合，v_t と v_t から観測された画像 x_t が SRN に入力され，v_t が対応される s_t が得られる．そして，シーンの 3 次元情報が $s_0 + s_t$ により更新される．また，時刻 t で観測終了動作が選ばれた場合，その時刻までのシーンの 3 次元情報と質問特徴 q' が視覚質問応答フレームワークに入力され，回答の予測を行う．

ここで，節 3.5.2.3 で述べた SRN+FiLM と同様な SRN フレームワークと視覚質問応答フレームワーク FiLM を用いた．以下の節で視点選択フレームワークの詳細を述べる．

3.6.3.2 視点選択フレームワーク

実環境で視覚質問応答では，質問と観測を基づいて回答を行う．例えば，“ is there a red ball?” という質問に対して，現在の観測視点ですでに “ red ball” を観測された場合に，質問を即座に答えられる．また，強い物体オクルージョンが含まれたシーンに対して，複数視点から観測する必要がある場合が想定できる．我々は以上のようなところから，節 3.6.2 で述べた強化学習フレームワーク DQN をベースに，観測されたシーンと質問から能動的に視点を選択するフレームワーク f_{VS} を提案する．

具体的に，K 個の視点から観察できるシーンをと仮定し，行動セット $A = \{a_i\}_{i=0,\ldots,K}$ を定義する．ここで，$a_0, \ldots, a_{K-1}$ は視点選択行動と対応し，a_K は観測終了動作と対応する．

環境の状態 (観測されたシーンの 3 次元情報)s_t と質問特徴 q' から，視点選択フレームワーク f_{VS} は $K+1$ 次元のベクトルを予測する．また，予測されたベクトルは $a_0, ..., a_{K-1}, a_K$ を実行した場合に得られる value の予測値を示す．時刻 t で j 番の動作 a_t が選ばれた場合，動作 a_t の予測 value 値 r_t^{eval} が下の式により定式化できる：

$$r_t^{eval} = \{f_{VS}(s_t, q')\}_j \tag{3.7}$$

また，動作 a_t の実際に得られる reward が下の式により定式化できる．ここで，r_t^{env} は環境から得られる reward と示し，γ が未来の reward に対しての減衰係数となる．

$$r_t^{real} = r_t^{env} + \gamma argmax_{a_{t+1}} f_{VS}(a_{t+1}|s_{t+1}, q') \tag{3.8}$$

視点選択フレームワークは r_t^{eval} と r_t^{real} の差を最小化するようにパラメータの更新を行う．我々は質問回答の精度と選択した視点の数により r^{env} を設計した．視点追加されたたびに，ペナルティ p_{sp} を与える．また，重複された視点が選ばれた場合に，ペナルティ p_{rt} を与える．p_{sp} と p_{rt} はハイパパラメータとなる．視覚質問応答のロスを $loss_{vqa}$ ($[-1, 1]$ にノーマライズしている) に定式し，r^{env} を下の式のように定義する：

$$r^{env} = \begin{cases} p_{sp} + p_{rt} - loss_{vqa} & (i) \\ p_{sp} + p_{rt} & (ii) \\ -loss_{vqa} & (iii) \end{cases} \tag{3.9}$$

最後の視点選択動作の reward を (i) に示す．そのたの視点選択動作の reward を (ii) に示す．また，視点選択終了動作の reward を (iii) に示す．

節 3.5.2.3 と同様に，ここで視覚質問応答フレームワークに FiLM モデルを使用した．FiLM モデルを f_{VQA} に表示し，回答 ans が下の式により予測される．

$$ans = argmax_{ans} f_{VQA}(ans|s, q') \tag{3.10}$$

視覚質問応答フレームワーク FiLM は予測回答 ans と ground truth の回答間の cross-entry ロスの最小化により更新される．

3.6.3.3 実装の詳細

図 3.25: 能動的視点選択を対応した多視点視覚質問応答のフレームワーク図

全体的なフレームワークは，SRN，視点の選択および視覚質問応答フレームワーク FiLM の 3 つから構成される．ここで，これら 3 つのモジュールの実装の詳細を紹介する．CG と実画像の両方の実験にここで紹介する設定を使用した．図 3.25 で詳細的な構造を示す．

SRN： Eslami らが提案した 12 層の tower-構造の SRN を実装した．SRN ネットワークは $64 \times 64 \times 3$ 次元の画像と $1 \times 1 \times 5$ 次元のカメラ情報を $16 \times 16 \times 256$ 次元のシーン特徴に変換する．SRN を実装する際に，5×10^{-4} の学習率で SRN ネットワークを 140 エポック事前学習した．また，Adam オプティマイザーを実装に使用した．

視点選択ネットワーク:質問情報を抽出するために，PyTorch で定義された torch.nn.Embedding を用いて，各単語のワンホットベクトルを 300 次元のベクトルに変更した．次に，2 層の隠れ層を持った LSTM を用いて，質問文を 1024 次元のベクトルにエンコードした．次に，エンコードされた質問特徴と SRN により得られた特徴を concatenation 操作により結合する．この操作により $1024 + 16 \times 16 \times 256$ 次元の特徴量が得られる．次に，その特徴量を 3 つの全結合層を介して $K + 1$ 次元のベクトルに変更する．視点選択フレームワークのハイパーパラメータ γ を 0.8 に設定した．また，ϵ-greedy をベースとした動作選択動作を実装した．初期の ϵ を 0.5 に設定し，3 エポックごとに $\epsilon = \epsilon * 1.04$ のように更新する．また，視点選択のペナルティ p_{sp} を-0.25 に設定し，重複視点選択のペナルティ p_{rt} を-0.05 に設定した．視点選択ネットワークと視覚質問応答フレームワーク FiLM がジョイントで学習される．

視覚質問応答フレームワーク：ここで FiLM 実装コードを一部変更し，実装を行った．具体的に，FiLM の画像特徴抽出器を削除し，代わりに SRN により多視点の画像特徴抽出を行っている．また，ここでの質問特徴抽出器は前述の視点選択ネットワークと同じネットワークを使用する．ネットワークのパラメータも共有している．FiLM ネットワークと視点選択ネットワークを学習率 3×10^{-4} で 40 エポックジョイントで学習を行った．また，SRN の実装と同様に，Adam オプティマイザーを実装に使用した．

3.6.4 CG 画像を用いた実験

3.6.4.1 データセット

Question 1 : Are there any large red objects right of the tiny green thing? (**no**)
Question 2 : There is a small sphere to the right of the yellow cylinder; what is its color? (**green**)
Question 3 : Are any big red cylinders visible? (**no**)
Question 4 : What color is the big cube? (**red**)

図 3.26: 仮想カメラ設定とデータセットの例

表 3.5: データセットでの統計情報

Datasets	No. of Scenes (Train/Test)	No. of Q–A Pairs (Train/Test)	Spatial		Non-Spatial	
			Exist	Query Color	Exist	Query Color
Multi-view-CLEVR_12views_CG (**CG**)	(40,000 / 3,000)	(157,232 / 11,801)	43,000	42,993	43,000	40,040
Multi-view-CLEVR_4views_CG (**CG**)	(40,000 / 3000)	(158,860 / 11,928)	43,000	42,997	43,000	41,791
Multi-view-CLEVR_4views_3DMFP (**CG**)	(10,000 / 0)	(39,694 / 0)	10,000	9994	10,000	9,700
Multi-view-CLEVR_4views_Real (**Real**)	(0 / 500)	(0 / 1974)	500	500	500	474

ここでまず節 3.5.3.1 で述べた Multiview CLEVR データセットと類似した 12 視点の Multiview データセットを構築した．ここで図 3.26 (左) のカメラ設定をしている．具体的には，カメラをシーン平面から 30 度の高さに設定し，カメラの向きをシーンの中心に設定した．12 仮想カメラを同じ円周で 30 度おきに設定した．また，オクルージョンが発生しやすくなるように，大きい物体をシーンの中心に配置する．中心物体の周囲にランダムに 3〜5 個の物体を配置する．また，ここで existing (物体の存在に関する質問) と query color(物体の色を予測する質問) の 2 種類を設定した．更に，物体間の関係推定の必要性に関して，spatial-related と non-spatial 質問を設定した．図 3.26 (右) に以上のような設定を持った Multiview-CLEVR_12views_CG データセットの一例を示す．また，データセットの統計情報を表 3.5 に示す．

3.6.4.2 実験結果

表 3.6: Multi-view-CLEVR_12views_**CG** での定量評価

Method	Overall Accuracy	Spatial		Non-Spatial		Average no. of Used Viewpoints
		Exist	Query Color	Exist	Query Color	
SRN+FiLM	**97.37%**	94.20%	98.20%	99.03%	98.11%	12
SRN+FiLM+VS	97.11%	95.27%	96.90%	99.03%	97.25%	**2.98**

定量的結果：全体的および質問ごとの精度を表 3.6 に示す．この節で提案したフレームワーク（SRN+FiLM+VS）が，前節で討論した SRN+FiLM と同レベルの精度を達成した．

しかし，SRN+FiLM+VS は平均してはるかに少ない視点を使用した（2.98 対 12）．このような結果から，提案の視点選択フレームワークが効率的に必要な情報を取得することを学習したことがわかる．このようなフレームワークを用いて，カメラの移動コストを減らせる．実環境のアプリケーション，特に観測視点が制限されるアプリケーションの場合，提案の視点選択フレームワークがより効率的である．

図 3.27: CG データセットでの結果例

定性的結果：図 3.27 に 4 つの結果例を示す．図中，提案フレームワークにより選択された視点を赤いバウンディングボクスにより囲む．Question 1 は “tiny cub” の存在に関しての質問であり，デフォルト視点ですでに “tiny cube” が観測され，この質問において提案フレームワークもデフォルト視点だけから観測し，正しく回答を予測できた．Question 2 と Question 3 で対象となる物体がデフォルト視点で観測できずに対して，提案フレームワークは視点の追加を行い，追加された情報を使用して正しく回答を予測できた．これらの結果から，提案フレームワークが観測された情報と質問から正しく回答を予測できる傾向にあることを示した．また，Question 4 で提案フレームワークは 3 つの観測を追加し，不正解を予測した．ここで，質問の対象物がシーンの中心にある物体の影の下に配置され，このような質問が非常に難しいことが想定できる．より高い解像度を使用することや物体の影を認識できるフレームワークにより提案フレームワークの性能を向上可能であることを想定できる．そういったような場面で正しく回答を予測できる手法の検討を今後の課題とする．

表 3.7: 視点選択が精度に与える影響

Method (No. of Used Viewpoints)	Overall Accuracy
SRN+FiLM_Random (**1 view**)	66.71%
SRN+FiLM_Random (**3 views**)	81.39%
SRN+FiLM_Equal (**3 views**)	82.90%
SRN+FiLM+VS (**2.98 views**)	**97.11%**

視点選択の効果：ここで，学習済みの視点選択フレームワークの性能を評価する．比較するために，ランダムに 1 つの視点を選択して回答を行うモデル SRN+FiLM_Random (1 view) とランダムに 3 視点を選び回答する SRN+FiLM_Random (3 views) および均等的にサンプリングした視点 SRN+FiLM_Equal(3 views) の性能を評価した．実験結果を表 3.7 に示す．

SRN+FiLM_Random (1 view) は全体的に 66.71%の精度を得られた．このような結果から，提案データセットは 1 視点から回答するのが困難であることがわかる．また，3 視点を使用した SRN+FiLM_Random と SRN+FiLM_Equal はそれぞれ 81.39%と 82.90%の精度を得られた．それと比較して，提案フレームワークは 97.11%の精度を達成した．このような結果から提案フレームワークは効率的に視点選択を行い，部分観測からも高い精度で質問に回答できる傾向にあることがわかった．

3.6.5 リアル画像を用いた実験

3.6.5.1 データセット

図 3.28: CG, リアル撮影，再学習用 CG データセットの例

ここで提案フレームワークが実環境で撮影した画像においての性能評価を行う．Multiview CLEVR データセットと類似した実環境で撮影したデータセットを構築した．ここで 12 視点のカメラの設定が困難なため，3.5.3.1 節に用いた Multiview CLEVR データセットと同様な 4 視点設定を用いて，データセット Multi-CLEVR_4views_Real を作成した．物体モデルを 3D プリンターにより作成し，蛍光ペンにより色塗りした．次に，作成したモデルを通常の蛍光灯を備えた実験室のシーンテーブルに配置した．カメラをテーブルから 30 度上げ，90 度ごとに配置し，シーンの中心から 1.1m の距離でその周りの間隔で撮影を行った．また，ソニーアルファ7III ミラーレス一眼デジタルカメラ [50] を実験に使用した．

撮影する際に，カメラを少し動かして、視点ごとに平均 7 枚の写真を撮影した．全体的に合計 100 シーンを撮影し視点ごとにランダムサンプリングする操作により，合計 500 シーンから構成するテストデータを作成した．また，前述の Multiview-CLEVR_12views_CG の生成に使用したプロセスと同様に質問と回答のペアを作成した．Multi-CLEVR_4views_Real データセットの統計情報を表 3.5 に示す．また，データセットの画像例を図 3.28 (真ん中 4 列) に示す．

3.6.5.2 実験結果

表 3.8: Multi-view-CLEVR_4views_**CG** データセットでの定量評価

Method	Accuracy	No. of Used Viewpoints
SRN_FiLM	**97.67%**	4
SRN_FiLM_VS	97.64%	**2.02**

表 3.9: Multi-view-CLEVR_4views_**Real** データセットでの定量評価

Method	**Fine-Tuning**		**Accuracy**
	SRN	**FiLM**	
SRN_FiLM	-	-	67.88%
SRN_FiLM	-	✓	76.14%
SRN_FiLM	✓	-	77.30%
SRN_FiLM	✓	✓	82.62%
SRN_FiLM_VS	-	-	66.82%
SRN_FiLM_VS	-	✓	79.99%
SRN_FiLM_VS	✓	-	91.56%
SRN_FiLM_VS	✓	✓	**94.01%**

定量的結果：Multiview-CLEVR_4views_CG（図 3.28 (左 4 列)，統計情報が表 3.5 に示す）を用いて，SRN+FiLM と SRN+FiLM+VS フレームワークを 40 エポック学習した．テストセットの精度を表 3.8 に示す．SRN+FiLM+VS フレームワークは平均的に半分程度の視点で SRN+FiLM と同レベルの精度を達成した．そして，同様な学習済みモデルを用いて，Multiview-CLEVR_4views_Real で性能評価を行った．その結果から，CG 画像で学習し，実環境で撮影された画像に対して，二つのモデルともおよそ 30%性能が劣化した (3.9 の 1 行目と 5 行目).

Semi-CG データセットで再学習：以上のような性能低下は，CG と実撮影の画像の間のドメインギャップ (例：色の分布，照明，物体のテクスチャーなど) が原因であると想定できる．実環境で高い精度を達成するために，実環境データセットで再学習する方法が考えられる．しかし，物体の配置やカメラの調整などが必要であることから，実環境の Multiview 視覚質問応答データセットを構築するのに時間のコストが高い．

上記の問題を解決するために，実物体の 3 次元モデルを作成し，仮想環境に導入する Semi-CG データセットの構築プロセスを用いた．具体的に，まず Ortery 社の 3D MFP カメラ [51] で Multiview-CLEVR_4views_Real データセットを構築する際に使用した物体モデルの 3 次元モデルを作成した．次に，作成した 3 次元モデルを仮想環境に導入し，今まで用いた CLEVR ベースのデータセットと同様にシーンの撮影が可能になる．以上のプロセスにより Multiview-CLEVR_4views_3DMFP データセットを作成した．データセットの例を図 3.28 (右 4 列) に示す．また，データセットの統計情報を表 3.5 に示す．

次に，Multiview-CLEVR_4views_3DMFP データセットで再学習を行った．再学習する際に，全体的フレームワークの異なる部分の再学習実験を行った (VQA モデルを freeze し，SRN だけ再学習；SRN モデルを freeze し，VQA モデルを再学習；VQA と SRN 両方再学習の 3 つを実装した)．実験結果を表 3.9 に示す．再学習することで，モデルの精度向上が達成した．特に VQA と SRN の両方を再学習した場合，SRN+FiLM+VS モデルは最大+27.19%の性能向上を得た．再学習することで，SRN+FiLM+VS モデルが Unseen の実画像データセットに対して 94.01%の精度を達成し，SRN+FiLM モデルを 11.39%上回った．このような結果が，SRN+FiLM+VS モデルが質問回答に不要な画像のノイズにより比較的に影響せずに回答を行えることが原因であると想定する．また，今後はこのような結果の原因を詳細的に分析していく予定である．

図 3.29: リアル撮影データセットでの結果例

定性的結果：図 3.29 に結果例を示す．図中，視点選択フレームワークにより選択された視点を赤いバウンディングボックスにより囲んだ．SRN+FiLM モデルは 4 つの視点すべて使用し質問に回答した．対照的に，SRN+FiLM+VS モデルが視点を選択しながら回答を行う．Question 1，2，3 において，SRN+FiLM+VS モデルが合理的な視点の選択を行い，質問に正しく回答できた．Question 4 に対して，2 つのモデルとも不正解を予測した．CG と実撮影の画像間のドメインギャップを縮める手法を導入することで，モデルの性能を向上可能であると想定する．

3.6.6 考察

既存の視覚質問応答手法は 1 枚の RGB 画像から回答を行う．我々は節 3.5 で述べたフレームワークにより単眼画像から行う視覚質問応答タスクを多視点観測可能にした．節 3.5 で紹介した SRN+FiLM 手法では全てのシーンを全周囲から観測し，回答を行っている．しかし，実環境アプリケーションでは物体の配置により観測可能な角度が限られた場合がある．また，質問を回答するのに必要な情報を得た以上に観測を追加する手法は非効率である．

以上の問題点から，ここで節 3.5 で紹介した SRN+FiLM に更に視点選択を行うフレームワークを導入した．これにより，観測されたシーンおよび質問から能動的に視点を選択可能にした．実験より，視点選択フレームワークを導入したモデルが質問回答の精度を保ったまま，質問回答するのに必要な視点数を大幅に減らすことができた（12 から 2.98 への削減）．更に，提案のフレームワークは実画像データセットにおいても良好な精度を達成した．

提案フレームワークは観測された情報により視点選択を行えるため，部分的な視点から観測しかできない場面やよりロボットの移動の効率性が重視されるアプリケーションなどへの適応性を高めると考える．

この節までに，単純な幾何物体を持った CG データセットをメインに検討を行った．モデルの複雑な視覚情報に対しての理解を検証するため，次節で実物体のモデルを用いたデータセット自動生成プロセスを提案する．

3.7 大規模 3 次元物体モデルを用いた視覚質問応答データセット

3.7.1 概要

既存の視覚質問応答データセットは，インターネットから集められた画像を持ったデータセット (例 VQA v1 と v2) および CG で生成された画像から構成されるデータセット (例 CLEVR) の 2 種類がある．実画像の視覚質問応答データセットは，より複雑なシーンに対しての理解の能力を評価できる．その一方，データセットを構築するのに高い人件費を必要となる．さらに，人間により集めたデータセットに Human report biases [14] が存在することが広く指摘されている．それと比較して，自動生成により生成できる CG データセットのバイアスがコントロール可能である．しかし，既存の CG 視覚質問応答データセットは単純な幾何物体から構成され，視覚複雑度が制限される．そのため，CG データセットで学習済みのモデルを実環境に直接応用することが困難である．また，既存の視覚質問応答データセットは 1 枚の画像から構成されることが多く，実環境で多視点からシーンを観測することが多く想定できる．

上記の問題を解決するために，大規模の物体モデルデータセットをベースとした Multiview 視覚質問応答データセットを提案する．また，CLEVR データセット自動構築プロセスをベースデータセットを構築し，必要に応じて自動的に大規模なデータセットを作成できる．既存の 3 つの大規模物体モデルデータセットをベースにデータセットを構築した．また，物体モデルの属性と物体クラス間の階層構造を考慮したラベル付けを行った．それらにより，物体クラスの階層的理解する能力を評価できるようにした．更に，提案データセットが日用品のモデルを使用している．そのため，ロボットピッキングなどのタスクへの応用が考えられる．

提案のデータセットを用いて，節 3.5 で述べた手法の実装を行った．実験結果により，提案のデータセットは従来の CLEVR より難しい傾向が示された．今後，提案のデータセットが将来の視覚質問応答タスクの評価や，様々な実環境で事前学習用として使用されることを期待する．

3.7.2 物体モデルデータセット

3.7.2.1 YCB

YCB データセット [52] 2015 年に Berk らにより提案された物体 3 次元モデルデータセットである．YCB データセットは RGBD 画像，物体の 3 次元モデル及び物体物理属性などが定義されている．また，データセットは異なる形状，サイズ及びテクスチャーをもった実用品 (hammer, tennis ball, bowl) 合計 75 個から構成される．

YCB データセットが 5 台の RGB-D カメラおよび 5 台の RGB カメラから構成されるカメラシステムに撮影された．また，Berk らが poisson surface reconstruction アルゴリズム [55] により密なメッシュモデルを作成した．現在，YCB がロボットマニピュレーションに広く使われている．

3.7.2.2 Bigbird

Bigbird データセット [53] は YCB と類似したカメラシステムおよび構築プロセスにより作成された．YCB データセットと同様に，Bigbird データセットも日用品の物体モデル (合計 125 モデル) から構成される．また，YCB と比較して，Bigbird データセットは数多くの包装食品のモデル及びボトル形状のモデル (例：shampoo, detergent) を含めている．

Bigbird データセットもロボットマニピュレーションや 3 次元物体の認識，グラスピングなどに広く使われている．

3.7.2.3 NEDO Items

NEDO Items [54] が Araki らにより提案された Amazon Robotics Challenge (ARC)2017 用の大規模物体モデルデータセットである．現在，NEDO Items データセットには日用品や食品およびオフィス用品などの物体モデル合計 254 モデル (RGB 画像，デプス画像および物体モデルのメッシュデータ) が含まれる．

NEDO Items データセットは三菱電機社が製作した MELFA-3D Vision [56] により撮影し作成される．YCB と Bigbird と比較して，NEDO Items データセットには高精度の形状をもった物体クラスが多く含まれている．

3.7.3　データセット構築プロセス

図 3.30: ROM_CLEVR データセット構築プロセス

CLEVR で紹介されたデータセット自動生成プロセスをベースに，図 3.30 で示しているような 4 つのステップから構成されるデータセット自動生成成プロセスを提案する．データセットは異なる物体モデルが含まれるシーンの多視点画像と画像内容に関する質問から構成される．具体的に，まず前節で述べた 3 つのデータセットから合計 134 個の物体モデルを集めた．次に，WordNet をベースに物体クラス間の階層構造を考慮した物体のラベル付けと属性のアノテーションを行った．次に，物体モデルを Blender [57] シーンにランダムに配置し，4 つの仮想カメラからシーンの多視点画像を集める．最後に，予め定義したセンテンス構造にシーンの情報を記録し，シーンに関する質問を自動的に生成する．センテンス構造とは例えば，“ A (color) (shape) is added ” という文の構造で，ここで生成されたシーンの物体の属性で (color), (shape) のところを後から埋める．残りの節でデータセットの構築プロセスの詳細を述べる．

3.7.3.1　物体モデルの収集

図 3.31: ROM_CLEVR データセットモデルの例とクラス分布図

実用品のモデルを集めるため，YCB，Bigbird と NEDO Items の 3 つのデータセットから物体モデルを集めた．YCB と Bigbird は RGB-D スキャナーから物体の撮影を行っているため，一部のモデルに欠損が含まれている．それと比較して，NEDO Items データセットに含まれている物体モデルの形状が欠損が含まれていない．しかし，NEDO Items データ

セットには箱入りパッケージに梱包されているモデルが含まれており，そのため，モデルの外観からそれらのオブジェクトを認識することは困難な場合がある．上記の問題を考慮して，不完全な形状のモデル及びパッケージに詰め込まれたモデルを除き，上記の 3 つのデータセットから合計 134 個のオブジェクトモデルを得られた．その 134 モデルの一例とクラス分布を図 3.31 に示す．

3.7.3.2　物体モデルのアノテーション

図 3.32: ROM_CLEVR データセットのクラス構造図

次に，134 物体モデルのラベル付けを行った．ここで階層的にクラス定義をしている WordNet をベースに物体クラスを定義した．まず，物体の外観からクラスを選定し，1 つのクラスに対して WordNet の階層構造に定義されている 1〜3 つの上位語を使用し，木構造のアノテーションを行っている．このようなクラス定義により，物体の階層定義を考慮した質問の構造が可能になる．例えば“ Are there any foods visible? ”という質問に対して，apple が存在する場合，“ apple ”が“ food ”の下位語であるため，質問の回答が“ yes ”となる．また，各物体モデルの支配的な色により物体の色をアノテーションした．以上のプロセスにより，図 3.32 で示しているような合計 35 葉クラスと 27 上位語を含めた合計 62 クラスを定義した．また，具体的なクラスの分布を図 3.31 (右) に示す．

3.7.3.3　シーンの生成

CLEVR シーン生成エンジンに基づき，シーンの生成を行っている．シーンごとにランダムに 3〜10 個の物体モデルをランダムに配置している．また，節 3.5.3.1 で用いた 4 つの仮想カメラ設定を使用した．上記のプロセスを通じて，観測された各シーンの 4 つの画像を取得できる．また，シーンの情報 (物体の属性と位置) をシーングラフにより記録する．

表 3.10: ROM_CLEVR データセットでの統計情報

Dataset	Scenes	Obj classes	Obj instances	Obj per scene	QA pairs
ROM_CLEVR_v0.5	53,000	8	16	2-8	527,906
ROM_CLEVR_v1.0	60,000	35	134	3-10	712,328

3.7.3.4 質問と回答のペアの生成

Non-spatial exist:
1. Are there any <C> <L>?
2. Are any <H>s visible?

Spatial-related exist:
1. Are there any <C> <L2>s <R> the <L>?
2. There is a <H>; are there any <C> <H2>s <R> it?

Non-spatial query class:
1. What class is the <C> <H>?
2. The <H> has what class?

Spatial-related query class:
1. What class is the <C> <H> <R> the <L>?
2. The is a <C> <H> <R> the <L>; what is its class?

<C> : color <L> : class <H> : hypernym <R> : spatial relation ship

図 3.33: 質問テンプレートの例

シーンごとに 4 タイプの質問を生成した：existing (物体の存在に関する質問)；query color (物体の色推定の質問)；query class (物体のクラス推定の質問) と counting 質問 (物体を数える)．物体間の関係推定を含まれた spatial-related 質問と含まれない non-spatial 質問を同時に設定している．各タイプの質問のテンプレート (例：図 3.33) を予め定義した．図中，色付きの部分の単語はランダムにより埋め込まれる．次に，CLEVR で定義されている回答推定プロセスにより，各質問の回答が自動生成される．上記のプロセスにより，シーンごとに 20 の質問と回答のペアを作成した．次に，回答が一様分布になるように，データセットの調整を行った．このような調整により，画像情報なしではデータセットに答えることが困難となる．

図 3.34: ROM_CLEVR データセットの例

上記の設定で ROM CLEVR v0.5 と ROM CLEVR v1.0 を構築した．これらの詳細な統計情報を表 3.10 に示す．また，図 3.34 にデータセットの例を示す．このようなデータセットはロボットピッキングなどの実環境アプリケーションに活用可能である．更に，必要に応じて物体モデルの変更や質問タイプの設計などが容易にできる．このようなデータセット作成プログラムが実環境アプリケーションの性能向上に貢献できると考えた．

表 3.11: ROM_CLEVR_v0.5 と ROM_CLEVR_v1.0 データセットでの定量評価

Dataset	Method	Overall accuracy	S	NS
ROM_CLEVR_v0.5	View pooling+FiLM	89.74	89.24	90.36
	SRN+FiLM	**95.18**	**95.02**	**95.37**
ROM_CLEVR_v1.0	View pooling+FiLM	86.72	72.46	**95.36**
	SRN+FiLM	**87.61**	**79.48**	92.54

表 3.12: ROM_CLEVR_v1.0 データセットでの定量評価

Methods	Reso.	Overall	Exist		Query color		Query class		Counting	
			S	NS	S	NS	S	NS	S	NS
View pooling+FiLM	64	73.81	56.83	86.92	91.61	86.71	93.30	90.64	35.29	58.01
View pooling+FiLM	128	86.72	72.57	97.72	94.86	98.25	96.53	98.63	51.59	81.03
View pooling+FiLM	256	85.25	58.98	**99.50**	96.59	**99.75**	**98.29**	**99.86**	47.40	**92.68**
SRN+FiLM	64	78.57	71.10	88.66	91.05	88.79	93.19	92.52	40.28	58.01
SRN+FiLM	128	86.06	**86.16**	92.35	95.58	90.49	97.13	94.24	**56.55**	69.41
SRN+FiLM	256	**87.61**	82.23	95.06	**97.31**	96.58	97.26	97.35	54.77	73.42

3.7.4 実験

3.7.4.1 実験の設定

節 3.5 で述べた 2 つの Multiview 視覚質問応答手法 View pooling ＋ FiLM と SRN+FiLM の性能評価を行う．すべての実験で, SRN ネットワークを学習率 0.0005 で 200 エポックで事前学習を行った．また，Viewpooling ＋ FiLM と SRN+FiLM をそれぞれ学習率 0.0005 で 30 エポック学習した．また，2 種類の ROM CLEVR データセットの異なる画像解像度で性能評価を行った．

3.7.4.2 異なるスケールのデータセットでの結果

まず，ROM CLEVR v0.5 と v1.0 で View pooling+FiLM と SRN+FiLM の 2 つの手法を評価した．結果を表 3.11 に示す.2 つの手法とも，ROM CLEVR v1.0 において比較的低い精度を得た．特に空間関連の質問の場合，性能の低下が激しい．この結果から，CLEVR と比較して，提案の ROM CLEVR データセットはより複雑であることがわかる．実環境では物体のアピアランスが CLEVR で利用している幾何情報より複雑度が高い，提案の ROM CLEVR データセットを使用することでもっと実環境利用に望まれる性能の評価が可能であることが想定できる．

3.7.4.3 異なる入力画像サイズでの結果

ここで更に入力画像解像度を変更した実験を行った．実験結果を表 3.12 に示す．64*64 の入力画像解像度の場合，2 つ手法とも他の解像度での結果と比較して最も低い精度を得られた．この傾向は特に空間関連の質問に当てはまる．この結果から，提案の ROM CLEVR データセットにおいて入力画像の解像度 64 * 64 では質問を回答するに必要な視覚情報を得

られないことが想定できる．また，2 つの手法とも，128*128 と 256*256 の解像度で同レベルの精度を得た．この結果，必要な解像度が満たされた場合，入力画像の解像度を単純に上げることで良い性能を得られるとは限らないことが想定できる．

64*64 の入力画像解像度の場合，SRN+FiLM が View pooling+FiLM を大幅に上回る．特に空間関係が必要な質問に対して，SRN+FiLM が優位性を示した．また，128*128 および 256*256 の場合，2 つの手法間の性能の差が小さい傾向にある．特に SRN ネットワークに対して，単純に解像度を上げることで性能を改善できるとは限らない傾向にある．高い解像度の SRN ネットワークの検討を今後の課題とする．

図 3.35: ROM_CLEVR データセットでの結果例

図 3.35 に 4 つの結果の例を示す．existing と query color および query class 質問 (Question (1), (2), (3)) において，2 つのモデルとも正しく回答を予測できた．また，counting 質問 (Question (4)) において二つのモデルとも正しく回答を予測できなかった．

3.7.5 考察

既存のリアル撮影画像の視覚質問応答データセットは高い構築コストがかかる．また，手作業により構築されるデータセットは Human report biases が指摘されている．それと比較して，既存の CG 視覚質問応答データセットは単純な幾何形状から構成され，より複雑な視覚理解を評価する能力が制限される．以上の問題を解決するために，ここで大規模な実用品物体モデルを持った Multiview 視覚質問応答データセットを提案した．提案データセットが自動的に構築でき，データセット構築の手作業のコストを大幅に下げられる．また，既存の CG ベースデータセットと比較して，より複雑なシーンの理解の評価を可能にした．

実験より，提案する ROM CLEVR データセットは 3.5 節で述べた Multiview CLEVR より難しい傾向がわかった．視覚質問応答タスクは実環境アプリケーションにおいて重要であり，今後，ROM CLEVR データセットが実環境で発揮できるモデルの性能評価と性能向上に寄与できると期待する．

3.8 本章のまとめ

従来の視覚質問応答タスクは単眼画像および画像内容に関する質問から質問の回答を行う．実環境アプリケーションでは，シーンを観測する角度，物体オクルージョン，照明条件などにより，単眼画像から質問を回答するに必要な視覚情報を得られない場合がある．

我々は第二章の実験により，複数視点からの物体を認識し，認識結果を物体が存在する 3 次元領域に投票することにより安定的でロバストに物体を認識できる結果を示した．ここで，従来の単眼画像視覚質問応答タスクに多視点・複数モダリティ認識能力を導入し，その有用性を示した．

具体的に，まず従来の視覚質問応答では RGB 画像だけから回答を行い，RGB 画像から十分質問を回答するための 3 次元情報を得られないことを想定し，RGB-D 画像を用いた視覚質問応答手法を提案した．RGB-D により視覚質問応答の精度，特に幾何情報が必要となる質問の精度が改善された結果が得られた．RGB-D 画像の結果により 3 次元情報が視覚質問応答タスクにおいての重要性が示唆されため，次に，多視点からシーンを観測し，観測された画像からシーンの 3 次元情報を含めた特徴表現を視覚質問応答タスクに導入した．実験結果により，多視点から得られるシーンの 3 次元情報が精度を向上できることを示した．人間がシーンに対しての質問を回答する時に，能動的に必要な視点を選択して回答を行うことに注目し，我々は，能動的に視点選択フレームワークを多視点視覚質問応答に導入した．その結果，高い精度で質問を回答しながら大幅に必要な視点を減らすことができた．また，従来多視点視覚質問応答データセットでは，単純な幾何物体により合成された CG データセットがメインであったため，我々はリアル撮影により得られる物体モデルをベースとしたデータセットを提案した．従来の CG データセットの補充として，今後多視点視覚質問応答タスクの性能向上に寄与できると幸いである．

この章で，多視点・複数モダリティが視覚質問応答タスクにおいての有用性を示したため，次の章では，視覚質問応答タスクと良く類似するシーン変化説明文生成タスクに導入するための検討を行う．

第4章　多視点・複数モダリティを導入したシーン変化説明文生成

4.1　概要

天候時間の変化や人間の活動などにより，我々を取り囲む環境が常に変化する．そのため，シーンの変化の理解が様々な実環境アプリケーションにおいて重要である．変化理解に関して，従来様々な側面から検討されている．例えば，Chen らが同じシーンの異なる上空画像から変化の検出を行う ([58, 59, 60])．都市や地域の上空画像から変化を検出する研究が都市や地域の発展進行の理解において重要である．また，櫻田ら [61, 62] が津波前後の画像間の変化理解を行い，被災状況の確認や災害復旧のための計画などに活用できる．

上述の従来の変化検出の手法は，変化前後の画像間のピクセルレベルの違いを認識し，ピクセルごとに変化もしくは不変を表示したマップを推定し，変化の内容の詳細的な理解を行っていない．そういった変化検出と比較して，近年，変化を変化前後の 2 枚の画像から検出し，自然言語の形式で出力するシーン変化説明文生成タスクが提案された [63, 64, 65]．前述の変化検出手法と比べ，変化説明文生成タスクが変化の理解を自然言語の形式により表示するため，人間とコミュニケーションする能力が重要視される HRI アプリケーションに活用できる．

しかし，第三章で述べた視覚質問応答タスクと良く類似するように，従来のシーン変化説明文生成タスクが単眼画像を取り扱う．実環境アプリケーションでは単眼画像から必要な視覚情報を得られない場合が多く存在するため，既存のシーン変化説明文手法が実環境アプリケーションへの適応が制限される．

以上の問題を踏まえて，この章では，多視点および複数モダリティをシーン変化説明文生成タスクに導入し，それらの情報がシーン変化説明文生成タスクに与える影響について述べる．まず第三章と同様に，多視点を変化説明文生成タスクに対応し，それによりモデル性能を向上できることを示す．次に，多視点・複数モダリティをこのタスクに導入し，我々が提案した室内シーン変化説明文生成データセットで多視点および複数モダリティが精度向上における効果を示す．

4.2 変化説明文生成

4.2.1 変化説明文生成タスク

図 4.1: シーン変化説明文生成タスクの概念図

従来，静止な画像や 3 次元のシーンをベースとしたシーン理解の研究が行われている．しかし，人間の活動により，シーンが常に変化するため，シーンの変化を理解する能力が様々な実環境アプリケーションにおいて重要である．

従来，静止シーンの理解と比較して，シーンの変化の理解に関する研究が少ない現状である．ロボットアプリケーションにおいて，変化前後の点群やサーフェイスデータから変化した領域を検出する研究がある．また，street view 画像や上空画像などの室外シーンの変化前後の画像から変化領域のマスクの検出などのタスクが提案された [58, 59, 60, 61, 62]．

2018 に Jhamtani ら [64] により，変化前後の画像から画像中に含まれるシーンの変化を自然言語のテキストの形式で出力するシーン変化説明文生成タスクが提案された．図 4.1 にシーン変化説明文生成タスクの概念図を示す．変化説明文生成タスクで変化前後の画像及び画像間の対応関係の理解を自然言語の形式により出力する．

従来の変化検出タスクにより，変化説明文生成タスクは言語の形式により理解を表すため，人とコミュニケーションする能力が期待されるロボットアプリケーションにおいて非常に実用性が高いと想定できる．

4.2.2　変化説明文生成データセット

4.2.2.1　Spot-the-diff

図 4.2: Spot-the-diff データセットの例 (Jhamtani et, al. [64] により)

2018 に Jhamtani らにより Spot-the-diff データセットが提案された [64]．図 4.3 にデータセットの例を示す．Jhamtani らが VIRAT surveillance ビデオデータセットから，ビデオフレームワークを抽出し，変化前後の画像ペアを作成した．また，変化前後の画像ペアを AMT により手動で変化説明文のアノテーションを付けた．ここで，車や歩行者などの移動に関する変化が多く含まれている．

Spot-the-diff データセットが合計 13,192 ペア変化前後の画像およびペアごとに平均 1.86 の変化説明文から構成される．Spot-the-diff データセットの視覚および言語的に複雑度が高いが，データセット構築の人件費コストが高い傾向になる．

4.2.2.2　CLEVR-Change

図 4.3: CLEVR-Change データセットの例 (Park et, al. [63] により)

Park ら [63] が CLEVR データセットをベースに CLEVR-Change データを提案した．CLEVR-Change データセットに変化前後に観測された 2 枚のシーンの画像およびシーンの変化のテキストの説明文から構成される．

CLEVR-Change データセットのシーンが CLEVR データセットをベースとし，シーンの中にランダムに物体を配置することにより自動構築される．節 3.2.2.2 の CLEVR データセット [41] と同様に，CLEVR-Change では物体にそれぞれランダムな形状，色，形状及びテクスチャーが与えられている．CLEVR-Change で add (一つの物体を追加)，delete (一つの物体を削除)，move (一つの物体の位置を変化)，texture (一つの物体のテクスチャーを変化) および color (一つの物体の色を変化) の 5 つの変化タイプが定義されている．また，変化前後の画像間に微小なカメラの位置変更およびイルミネーションの変更が与えている．また，カメラとイルミネーションだけが変更する distractor も定義されている．

CLEVR-Change データセットに 1 種類の変化または distractor に 2〜3 つのセンテンステンプレートが定義され，記録されたシーンの変化内容により変化の説明文が自動生成される．Spot-the-diff データセットと比べて，CLEVR-Change データセットのシーンの複数度および変化説明文の複数度が低い傾向にあるが，データセットが自動構築可能であり，大幅にデータセット構築の人件費コストを減らせる．

4.2.3 変化説明文手法

4.2.3.1 DDLA

Spot-the-diff データセットを提案した Jhamtani らが Difference Description with Latent Alignment (DDLA) 手法を提案した．DDLA 手法では変化前後の画像のピクセルレベルの差をベースに変化の説明文を生成する．

具体的に，まず変化前後の 2 枚の画像間のピクセルレベルの差を表す pixel-difference マスクを計算する．次に，前述のマスクをクラスタリングし，object-level differences を求める．そして，得られた物体領域の画像特徴を ResNet モデルにより抽出する．最後に，抽出された特徴量を LSTM 構造の説明文生成器に入力し，変化の説明文が生成される．

4.2.3.2 DUDA

図 4.4: DUDA 手法のフレームワーク図 (Park et, al. [63] により)

CLEVR-Change データセットを提案した Park らが図 4.4 に示す DUal attention Dynamic Attention (DUDA) 手法を提案した．

前節で述べた DDLA 手法が画像のピクセルレベルの差により変化の認識を行っている．それと比較して，DUDA 手法が変化前後の画像の抽出特徴量の差をベースに変化を認識する．具体的に，まず変化前後の画像を ResNet により変化前後の画像特徴 X_{bef} と X_{aft} を抽出する．次に，X_{aft} と X_{bef} の引き算により特徴差 X_{diff} を求める．次に，変化前後の画像特

徴量 X_{aft} と X_{bef} にそれぞれ別々に X_{diff} と concatenation する．次に，変化前後の画像特徴の spatial attention を計算し，attention 操作により $l_{\text{bef}}, l_{\text{aff}}$ 及びそれらの差 l_{diff} を得る．

$l_{\text{bef}}, l_{\text{aff}}$ 及び l_{diff} が LSTM 構造の説明文生成器に入力し，説明文が生成される．説明文生成器に dynamic attention 仕組みが導入され [66]，それにより変化説明文の各々の単語の生成ステップに $l_{\text{bef}}, l_{\text{aff}}$ 及び l_{diff} に対しての attention が計算され，それらの重み付き和が生成器に入力される．

DUDA 手法が CLEVR-Change データセットにおいて，高い精度を得られた．更に，定性的実験結果により，DUDA 手法の Dual attention メカニズムが有効的に変化が発生した領域を画像アテンションにより検出できる傾向が示せられた．

4.2.4 変化説明文評価指標

4.2.4.1 BLEU

BiLingual Evaluation Understudy (BLEU) [67] が Papineni らが 2002 年に提案した機械翻訳を評価する指標である．BLEU では翻訳された文章と参考文の N-gram (N 文字が連続した文字列) が共に出現する程度を評価．BLEU が下の式により計算される．

$$BLEU - N = BP \cdot exp(\sum_{n=1}^{N} w_n longp_n) \tag{4.1}$$

上式に，BP (Brevity Penalty) が短い文 (長さ c) が高いスコアを得るのを防ぐ係数であり，下の式により計算される．r が最も短い参考文の長さを表す．

$$BP = min(1, exp(1 - \frac{r}{c})) \tag{4.2}$$

式 (4.1) に w_n が $n-gram$ のウェイトを示す．p_n が出力の $n-gram$ の中の正解 $n-gram$ の割合を示し，下の式により計算する：

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C' \in \{Candidates\}} \sum_{n-gram \in C'} Count(n-gram')} \tag{4.3}$$

$$Count_{clip} = min(Count, Max_Ref_Count) \tag{4.4}$$

現在，BLEU が機械翻訳だけではなく，画像説明文の評価指標としても広く使われている．

4.2.4.2 ROUGE-L

BLEU と類似するように，Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [68] も $n-gram$ をベースとした機械翻訳用評価指標である．ROUGE が生成された説明文の中に含まれる ground truth 説明文の recall 率を評価する．また，画像説明文の評価指標として ROUGE-L(Longest common subsequence) が広く使われる．参考文を X(長さ m), 生成文を Y(長さ n), $LCS(X,Y)$ が X と Y の最大サブ文の長さで，ROUGE-L は式 (4.5) により計算される：

$$F_{LCS} = \frac{(1+(\beta)^2)R_{LCS}P_{LCS}}{R_{LCS} + (\beta)^2 P_L CS} \tag{4.5}$$

R_LCS と P_LCS が下の式により計算される：

$$R_{LCS} = \frac{LCS(X,Y)}{m} \tag{4.6}$$

$$P_{LCS} = \frac{LCS(X,Y)}{n} \tag{4.7}$$

また，式 (4.5) のパラメータ β が通常 ∞ に設定されるため，F_{LCS} の値が R_{LCS} に近似できる．

4.2.4.3 METEOR

Metric for Evaluation of Translation with Explicit ORdering (METEOR) [69] がワードベースの機械翻訳用評価指標である．METEOR で WordNet を導入し，類似語や同義語を考慮した評価が行える．METEOR が式 (4.8) により計算される：

$$METEOR = (1 - pen) \times F_{means} \tag{4.8}$$

w_t で生成文の長さ，w_r で参考文の長さ，m を生成文と参考文の共通単語数 (ここで同義語が共通単語として取り扱える) を表示し，P と R が下の式により計算される：

$$P = \frac{m}{w_t}, R = \frac{m}{w_r} \tag{4.9}$$

F_{means} を下の式により計算される (α が調整可能なパラメータであ)：

$$F_{means} = \frac{PR}{\alpha P + (1-\alpha)R} \tag{4.10}$$

また，式 (4.8) の pen が生成文と参考文の単語の順序の違いに対してのペナルティ係数で，$chunks$ で参考文と生成文の両方に存在するサブ文の個数を表し，以下の式により pen を計算する：

$$pen = \frac{chunks}{m} \tag{4.11}$$

4.2.4.4 SPICE

Semantic Propositional Image Caption Evaluation (SPICE) [70] が画像説明文の評価指標として提案され，現在あらゆる画像説明文の手法で重要な評価指標として用いられている．

以上の BLEU，ROUGE-L, METEOR などの評価指標が生成文と参考文の $n-gram$ の類似性を評価する．それらと比較して，SPICE では SPICE が説明文のセマンティック構造を抽出する．抽出されたセマンティック構造をベースとした類似度評価を行っている．

具体的に，SPICE の計算ではまず Probabilistic Context-Free Grammer (PCFG) dependency parser により，生成文と参考文の syntatic dependencies trees を求める．次に，ルールベースの手法で得られた dependencies trees をシーングラフに変更する．シーングラフをベースに生成文の中に含まれる物体，物体属性およびそれらの関係性の正解さを評価する．

以上の BLEU，ROUGE-L, METEOR などと異なって，SPICE がセンテンスの意味構造を考慮した生成文評価ができる．

4.3 多視点に対応した変化説明文生成

4.3.1 概要

図 4.5: 多視点に対応した変化説明文生成の概念図

シーンで起こる変化を理解する能力がさまざまな HRI アプリケーションにおいて重要である．そのような能力を実装したホームロボットが部屋の再スキャンと再理解することを回避でき，自動的にシーンの変化を理解して部屋に対しての理解をを常に最新状態に更新できる．また，シーンの変化を理解する能力が監視カメラシステムにおいても重要な役割を果たせる．近年，DNN と大規模計算ユニット GPU の発展により，3 次元認識技術が目覚ましい発展を遂げた．しかし，様々な実環境アプリケーションにおいて非常に重要であることにも関わらず，3 次元シーンの変化に関する研究が比較的に少ない現状である．

最近複数のシーンの変化を理解し変化の内容を自然言語の文章により出力するシーン変化説明文生成に関する研究が提案された．しかし，既存の手法が変化前後のシーンの 2 枚の画像から変化の認識を行っている．そこで，節 3.5 で述べた問題点と同様に，実環境アプリケーションで既存の単語画像シーン変化説明文生成手法を実装すると，単眼画像により必要な情報が得られない場合が多く存在することが想定できる．

そこで，節 3.5 で述べた取り組みと類似したように，ここで図 4.5 に示すように，ここで多視点認識能力をシーン変化説明文生成タスクに導入し，多視点からシーンを観測しシーンの変化を理解することを可能にする．

具体的に，まず既存シーン変化説明文生成タスクの評価用データセットが単眼画像から構成されることから，我々は今まで用いてきた CLEVR をベースに，多視点シーン変化説明文データセットを自動構築した．また，ここも節 3.7 の述べたプロセスによりシーンにリ

アル撮影により生成された物体モデルの導入も可能にした．また，従来のシーン変化説明文生成の手法が単眼画像を取り扱うため，ここで多視点からシーンを 3 次元的に認識する SRN を導入し実装を行った．更に，ここで変化前後のシーンの特徴のアライメントが必要となるため，我々は異なる変化前後のシーンの特徴を結合する手法を実験した．提案データセットにおいて，我々の SRN を導入したモデルは従来の強いベースライン手法を大幅に上回った性能を達成した．更に，既存手法よりオクルージョンが含まれる環境やカメラの動きにロバストである傾向を示した．また，提案手法が実物体モデルに対しても既存手法より良い性能を得られ，実環境アプリケーションへの適応の可能性を示した．

4.3.2 提案構造

図 4.6: 多視点に対応した変化説明文生成のフレームワーク図

図 4.6 に全体的なフレームワークを示す．ここで，変化前のシーンを s_{bef} で表示し，変化後のシーンを s_{aft} に表示する．提案フレームワークは s_{bef} と s_{aft} の多視点の入力から，シーンの変化を表すテキストの説明文を出力する．全体的フレームワークは 2 つのモジュールから構成する：(1) 変化前後のシーンの 3 次元的理解を行う scene representation モジュール; (2) 前述の (1) から変化の説明文を生成する caption モジュール．

scene representation モジュール: まず，scene representation モジュールに関して，節 3.5 及び節 3.6 と同様に Eslami らが提案した GQN [45] を用いた．事前学習のプロセスで多視点画像からシーンの特徴抽出器 SRN と指定の視点から画像をレンダリングする RN をジョイントで学習する．また，変化説明文生成の学習で事前学習済みの SRN を使用して，変化前後のシーンのそれぞれの多視点画像から別々に変化全のシーンの 3 次元特徴 r_{bef} と変化後のシーンの 3 次元特徴表示 r_{aft} を得る．

caption モジュール: ここで scene representation モジュールにより抽出された変化前後のシーンの特徴 r_{bef} と r_{aft} から，シーンの変化を表すテキスト変化説明文を生成する．Park らが提案したモデルをベースに，ここで下の式により，まず変化前後のシーンの特徴差 r_{diff} を求める：

$$r_{\text{diff}} = r_{\text{aft}} - r_{\text{bef}} \tag{4.12}$$

表 4.1: P_change_occlusion, P_change_cameras と ROM_change データセットの統計情報

Dataset	Scenes (train/test)	Captions (train/test)	Change types	Viewpoints	Object classes	Unique objects	Objects / scene
P_change_occlusion	(24,000 / 6,000)	(720,000 / 180,000)	5	4	-	30	2-6
P_change_cameras	(24,000 / 6,000)	(720,000 / 180,000)	5	7	-	30	2-7
ROM_change	(24,000 / 6,000)	(720,000 / 180,000)	5	4	13	72	3-8

Park ら [63] が上の式により，変化前後のシーンを ResNet モデルにより抽出された画像特徴間の差を求めるのに対し，我々は SRN により抽出された特徴差を求める．また，下の式により上式に得られる r_{bef}，r_{aft} と r_{diff} に対して別々に畳み込み操作を加える．

$$\begin{cases} l_{\text{bef}} = Conv1(r_{\text{bef}}) \\ l_{\text{aft}} = Conv2(r_{\text{aft}}) \\ l_{\text{diff}} = Conv3(r_{\text{diff}}) \end{cases} \tag{4.13}$$

ネットワークの残りの部分は Park らにより提案されたモデルを使用した．具体的に，Park らが提案した LSTM 構造の LSTM_d (dynamic attention) と LSTM_c (caption generator) を使用して，$l_{\text{bef}}, l_{\text{aft}}, l_{\text{diff}}$ から変化の説明文を生成した．

4.3.3 実験

ここで，提案のフレームワークと Park らが提案した DUDA 手法との比較実験を行う．具体的に，まず変化前後に異なる視点の組み合わせから観測して評価実験を行った．ここで，画像特徴に対してのアテンションの付け方，異なる画像特徴のアンサンブル及び説明文生成器のアテンションの付け方に対して実験を行った．次に，カメラ変動に対してのロバスト性評価実験を行った．最後に，実物体モデルを使用したデータセットでの評価実験を行った．

4.3.3.1 データセット

図 4.7: データセットの仮想カメラ設定

上記の実験用に 3 つのデータセットを作成した．表 4.1 にデータセットの統計情報を示す．各データセットは変化前後のシーンの多視点画像とシーンの間に発生した変化の説明文から構成される．また，ここでも CLEVR データセットの構築プロセスをベースにしている．

P_change_occlusion データセット：このデータセットは CLEVR に使用されている幾何物体を使用した．また，シーンを 4 つの仮想カメラから観測を行う (図 4.7 (左))．シーン

表 4.2: ROM_change データセットの物体クラス分布

Classes	Object instances	Source dataset
snack	12	NEDO:2; YCB:5; Bigbird:5
cup	11	NEDO:6; YCB:4; Bigbird:1
shampoo	10	NEDO:1; Bigbird:9
dishwasher	9	NEDO:8; Bigbird:1
minicar	7	NEDO:7
bar soap	6	NEDO:3; Bigbird:3
bowl	4	NEDO:3; YCB:1
sponge	4	NEDO:3; YCB:1
soft drink	3	Bigbird:3
weight	2	NEDO:2
water bottle	2	NEDO:2
soccerball	1	YCB:1
teddy bear	1	NEDO:1

ごとにランダムに 2 個から 6 個のランダム属性を持った物体 (5 種類の色，3 種類の形状と 2 種類のサイズ) を配置する．物体オクルージョンが含まれるシーンでのモデル性能を評価するために，ここで変化前のシーンに対してオクルージョンの制約を加えた．具体的に，4 つの視点から 2 視点以上のシーンに遮蔽された物体が存在するように物体の配置を行った (画像解像度：320*240；遮蔽物体のピクセル数：100 ピクセル以下)．このような設定で高い精度で変化説明文を生成するために，多視点観測からシーンを 3 次元的に認識する能力及びオクルージョンに対してのロバスト性が必要となる．

P_change_cameras データセット：このデータセットによりモデルのカメラの動きに対してのロバスト性を評価する．P_change_cameras データセットで変化前後のシーンをそれぞれ 1 視点から観測される．また，変化前後のシーンの仮想カメラ間の回転角度の調整を行った．具体的に，図 4.7 (右) の設定のように，変化前のシーンを Viewpoint 1 から観測し，変化後のシーンを Viewpoint 1 から Viewpoint 7 までの 7 つの設定で，7 種類の変化前後のシーン観測ペアを持ったデータセットを作成した．

ROM_change データセット：このデータセットは節 3.7 で述べたデータセットと類似した実物体モデルをシーンの構築に使用した．ここでも，節 3.7 で述べたデータセットと同様に NEDO Items，YCB および Bigbird の 3 つのデータセットから物体のモデルを集めた．具体的に実験に使用したモデルのクラスとクラスごとのモデル数を表 4.2 に示す．

説明文の自動生成：変更前と変更後のシーンを生成するために，各々のシーンを Blender 上に 2 回撮影を行った．また，2 回の撮影の間に物体の配置の変更を加えた．具体的に，ここで 5 種類の変化タイプを定義している：add (シーンに一つの物体を追加)；delete (シーンから一つの物体を削除)；move (一つの物体の位置を変更)；swap (2 つの物体の位置を交換)；distract (変化がなし)．また，Johnson らに提案されたシーングラフによりシーンごとの物体の属性と位置を記録した．記録したシーングラフおよび予め定義したセンテンスの構造から自動的にシーンの変化説明文を生成できる．ここで，一つの物体変化に対して 5 つ

のセンテンス構造を定義した．以上のプロセスにより，一つのシーンに対して 5 種類の変化および 30 変化説明文が自動的に生成された．

4.3.3.2 実験設定

すべての実験で、SRN ネットワークを 0.0001 の学習率で 200 エポック事前学習をした．DUDA モデルの実装に対して，変化前後のシーンの画像をそれぞれ ImageNet で学習済みの ResNet モデルにより特徴抽出を行った．また，提案フレームワークと従来手法の DUDA を変化説明文データセットで 0.0001 の学習率で 40 エポック学習した．また，全ての実験で Adam オプティマイザーを使用した．

4.3.3.3 CG 物体モデルを用いた実験

ここで，まず P_change_occlusion データセットを用いて，提案フレームワークと DUDA の性能評価を行った．また，異なるサブ構造や変化前後のシーンの観測視点の組み合わせに対して実験した．

図 4.7 (左) で示しているように，まず変化前のシーン s_{bef} を視点 1, 2 から観測し，変化後のシーン s_{aft} を視点 3 と 4 から観測するような設定で実験を行った．ここで，変化前後のシーンを異なる視点から観測を行っているため，正しく変化説明文を生成するために，多視点画像からシーンの 3 次元的理解が必要となる．また，ここで異なるサブ構造の実装を行った．モデルを $E/F/A$ のような形式で表示する．ここで E がアンサンブルされた画像特徴の種類を示し，F で特徴抽出構造，A をデコーダーのアテンションの仕組みを表示する．以下，実装したモデルの詳細を述べる．

no-diff/-/spa-att:この構造は [21] に基づいている．この構造でシーン変化前後の特徴量 r_{bef} と r_{aft} を入力とする．その二つの特徴量を Concatenation 操作により結合する．また，デコーダーの部分で結合された特徴量に対して spatial attention 操作を加えている．

no-diff/-/no-att: この構造は前述の no-diff/-/spa-att の spatial attention を除くことにより構築されている．

no-diff/conv/no-att: 前述の no-diff/-/no-att と比較して，この構造で量 r_{bef} と r_{aft} に対して別々の畳み込み操作を加えた．

diff/conv/no-att: no-diff/conv/no-att と比較して，この構造で r_{bef} と r_{aft} のほか，二つの特徴量の差 $r_{\mathrm{aft}} - r_{\mathrm{bef}}$ も入力に入れた．また，そこで得られた r_{diff} および r_{bef} と r_{aft} に別々に畳み込み操作を加えた．

diff/conv/dyn-att: diff/conv/no-att と比較して，この構造でデコーダーの部分で Park らにより提案された dynamic attention 操作を加えた．この操作により説明文生成のプロセスで単語ごとに画像特徴に対して異なるアテンションがかけられている．

定量的結果: P_change_occlusion データセットのテストセットでの評価結果を表 4.3 に示す．ここで節 4.3.3 で紹介した 4 つの評価指標：BLEU-4, ROUGE_L, SPICE, および METEOR を使用した．また，異なる変化タイプの BLEU-4 の精度も計算した．

表 4.3 の 1，2 と 3 行目から spatial attention を加えることにより，モデルの性能が大幅に劣化する傾向にあることがわかった．ここで no-diff/-/spa-att が最も低い精度を得て，spatial attention を除くことにより性能が向上できた．このような傾向から，現状用いている SRN のシーンの抽出特徴が spatial attention により破壊されてしまう可能性があることを想定した．また，2 と 3 行目の結果の比較から，畳み込み層を加えることで性能を改善できる傾向にあることがわかった．3 と 4 行目の結果の比較から，変化前後の特徴間の差 $r_{\mathrm{aft}} - r_{\mathrm{bef}}$

表 4.3: P_change_occlusion (上 10 行) と ROM_change (ボトム 6 行) データセットでの定量評価

Approaches, (viewpoint pair)	ROUGE_L [68]	SPICE [70]	METEOR [69]	BLEU-4 [67]						
				overall	add	delete	move	swap	replace	distractor
no-diff/-/spa-att, (1,2-3,4)	73.4	40.5	45.3	58.8	66.9	84.5	64.5	48.6	79.8	11.7
no-diff/-/no-att, (1,2-3,4)	76.3	41.0	46.9	63.2	67.4	84.6	68.2	50.6	82.3	24.5
no-diff/conv/no-att, (1,2-3,4)	87.8	47.8	54.1	77.9	79.0	90.2	85.8	57.4	81.1	78.6
diff/conv/no-att, (1,2-3,4)	87.9	49.3	54.1	78.5	80.5	88.9	83.9	55.9	83.1	81.5
diff/conv/dyn-att, (1,2-3,4)	88.4	49.4	54.7	79.4	81.2	89.7	85.2	56.3	83.1	84.4
diff/conv/dyn-att, (1,2,3,4-1,2,3,4)	**96.4**	**56.9**	**64.3**	**92.3**	**99.8**	**99.8**	**97.6**	**70.3**	**97.1**	**100.0**
diff/conv/dyn-att, (1-3)	79.4	42.6	48.1	67.9	64.0	75.9	80.8	53.4	71.2	60.1
diff/conv/dyn-att, (1-1)	84.6	45.9	51.5	74.2	71.7	74.5	75.8	50.8	72.8	99.7
DUDA, (1-3)	66.6	34.3	37.6	50.1	50.7	58.1	52.0	36.1	61.2	33.4
DUDA, (1-1)	80.7	43.6	48.7	68.6	63.8	71.1	62.2	47.5	67.5	99.4
Ours, (1-1)	73.2	33.1	40.8	58.6	51.0	57.0	58.2	28.1	52.4	**100.0**
Ours, (1-3)	64.0	25.7	33.7	47.6	36.9	54.7	54.3	25.5	48.9	56.3
Ours, (1,2-3,4)	73.4	32.8	40.0	58.1	54.5	68.2	59.4	30.1	63.3	65.6
Ours, (1,2,3,4-1,2,3,4)	**94.6**	**46.9**	**59.5**	**87.1**	**96.1**	**97.3**	**89.4**	**50.2**	**91.9**	**100.0**
DUDA, (1-1)	66.7	27.9	35.4	51.9	47.0	52.2	39.8	24.1	47.8	**100.0**
DUDA, (1-3)	46.3	14.6	21.9	26.1	22.4	45.3	32.2	12.4	27.9	10.4

表 4.4: P_change_occlusion データセットでの説明文精度評価

Approaches	Accuracy				
	change type	Object	Color	Shape	Size
Ours (1-1)	87.1	74.6	89.4	84.9	91.7
Ours (1-3)	77.8	58.6	78.0	69.5	85.4
Ours (1,2-3,4)	90.4	70.1	87.4	78.8	92.3
Ours (1,2,3,4-1,2,3,4)	**99.4**	**91.5**	**98.0**	**96.6**	**99.2**
DUDA (1-1)	76.9	59.8	78.8	74.9	84.0
DUDA (1-3)	60.1	43.4	60.0	59.0	73.8

を用いることで性能を改善できることがわかった．また，4 と 5 行目の結果から，dynamic attention を用いることですこし性能の向上に効果があることがわかった．以上の結果から，この節の残りの実験では，diff/conv/dyn-att 構造を使用した．

diff/conv/dyn-att 構造に対して，異なる変化前後のカメラ視点の実験も行った．

diff/conv/dyn-att に対して変化前後の視点ペアを (1,2,3,4-1,2,3,4) (before-after) の設定をした．また，既存手法 DUDA が変化前後のシーンを単眼画像から観測を行うため，ここで比較に DUDA と提案モデルに対して変化前後のシーンを (1-3) と (1-1) の視点ペアから観測する実験を行った．

(1-1) 視点ペア以外に，提案のモデルが多くの視点が使うほど高い精度を得られている傾向を示した．また，既存手法 DUDA と比較して，提案手法が (1-3) と (1-1) の両方においても既存手法 DUDA を上回った．提案手法は変化前後のシーンの多視点観測からシーンの理解を行い，それをベースに変換の理解の説明文生成に成功した．このような結果から，提案の多視点変化説明文生成タスクにシーンの 3 次元的な理解が重要であると想定できる．

また，提案手法と DUDA 両方においても，(1-1) 視点ペアと比較して，(1-3) の視点ペアの性能が劣化した傾向にあった．その原因としては変化前後のシーンを (1-1) 視点ペアのような同じ視点から観測することがこのタスクにおいて有利であると想定できる．

キャプションの正確性の評価: 従来の画像説明文の評価指標 (例：BLEU-4) が精密に生成された説明文と Ground Truth 説明文の類似性を評価できない問題点がある．ここで，生成された説明文と Ground Truth 説明文から変化のタイプ，物体の属性 (クラス，色と形状など)，物体 (クラスと色と形状) などの単語を抽出し，accuracy の計算を行った．P_change_occlusion データセットにおいてのキャプションの正確性の評価を表 4.4 に示す．ここで，我々の最も高い性能のモデルに対して (1-1), (1-3), (1,2-3,4), (1,2,3,4-1,2,3,4) の視点ペアを実装し，DUDA モデルに対して (1-1), (1-3) の観測視点ペアの評価を行った．

DUDA モデルと比較して，提案手法が変化タイプ，物体の属性推定において高い精度を示した．また，(1,2,3,4-1,2,3,4) の視点ペアを実装したモデルが全ての推定に 90%以上の精度を達成した．このような結果から，提案手法が正しく詳細な変化情報を推定できることがわかった．

図 4.8: 定性的実験結果

定性的な結果: P_change_occlusion データセットの結果例図 4.8 (a,b,c) に示す．物体オクルージョンが比較的に軽い傾向にある例 (a) においてすべてのモデルが正しい変化説明文を生成できた．そして，比較的に強い物体オクルージョンが含まれる傾向にある例 (b,c) において，提案手法すべて正しく変化タイプを予測した．それと比較して，DUDA モデルが不正確な変化タイプと物体の属性を予測した．

DUDA 手法と比較して，提案モデル変化前後のシーンを 3 次元的に対応し，変化前後のシーンの潜在的な 3 次元の認識をベースに変化の理解を行える．以上のところから，DUDA と比べて，提案モデルがオクルージョンが含まれるシーンにおいて高い性能を示した．実環境アプリケーションでは物体のオクルージョンが多く存在し，ここでの実験結果から提案フレームワークが実環境アプリケーションへの適応に一つ可能なアプローチを示した．

図 4.9: P_change_cameras データセットでの BLEU-4 評価

ここで変化前後のカメラ視点間に回転の変動を加えて実験を行った (P_change_cameras データセット)．図 4.7 (右) で示しているように，ここで変化前のシーンを視点 1 から観測し，変化後のシーンを視点 1 から視点 7 の 7 つの位置で観測を行っている．提案モデルと DUDA を以上のような 7 種類の視点ペアの設定で実装した．BLEU-4 の結果を図 4.9 に示す．

変化前後のシーンを同じカメラ視点から観測する場合 ((1-1) ペア)，DUDA と提案モデル両方とも最も良い精度を達成した．変化後のカメラを回転することにより，DUDA 手法が大幅な性能劣化傾向にあることがわかった．それと比較して，提案モデルが比較的に安定した精度を達成し，DUDA 手法よりカメラ回転に対してのロバスト性を示した．実環境アプリケーションでは変化前後のシーンを同じ視点から観測できないことが多く存在することが想定できるため，DUDA と比較して，提案モデルが実環境アプリケーションへの適応性が高いことが想定できる．

4.3.3.4 リアル物体モデル画像を用いた実験

ここで ROM_change データセットを用いた実験の結果を述べる．

定量的な結果: ROM_change データセットで P_change_occlusion データセットと同じカメラの設定を用いた．異なる画像説明文の評価指標の結果を表 4.3 に示す．

P_change_occlusion データセットでの実験結果と比較して，提案モデルと DUDA 両方とも ROM_change で低い精度を得た．前述の 2 つのデータセットと比較して，ROM_change データセットはより視覚複雑度が高いことが性能劣化の原因であると想定した．また，ここでも P_change_occlusion と同様に提案モデルが DUDA 手法を大きく上回った．また，変化前後のシーンをそれぞれ 4 視点から観測する提案モデルが全体的 87.1 の BLEU-4 の精度を達成した．このような結果から，提案モデルが実物体への応用可能性を示した．

表 4.5: ROM_change データセットでの説明文精度評価

Approaches	Accuracy			
	change type	Object	Color	Class
Ours (1-1)	74.1	42.6	56.4	54.3
Ours (1-3)	63.2	31.7	45.5	43.0
Ours (1,2-3,4)	75.4	40.1	53.0	50.9
Ours (1,2,3,4-1,2,3,4)	**98.2**	**81.6**	**88.1**	**86.9**
DUDA (1-1)	64.2	37.8	50.9	46.8
DUDA (1-3)	37.0	19.8	31.5	28.2

図 4.10: ROM_change データセットでクラスごとの BLEU-4 評価

データセットのクラスインバランスに対しての敏感性評価: ROM_change データセットのクラス分布がインバランス的である．ここでこのようなインバランスに対しての敏感性評価を行った．ここで物体クラスごとの BLEU-4 値を計算した．single-object 変化の add, delete, move が対応する ground truth の物体クラスに生成されたセンテンスを記録し，また multi-object 変化の swap と replace に係る 2 つの物体クラスに生成されたセンテンスを 2 回記録するように BLEU-4 の計算を行った．ここで，DUDA 手法と提案モデルの (1-3) 視点ペアの実装の評価を行った．BLEU-4 の値を図 4.10 に示す．2 つのモデル両方ともそれぞれのクラスに対して異なった BLEU-4 の値を得た．また，提案モデルが DUDA より高いクラス間の分散を示した．提案モデルが大きい物体 (例：soccer ball) に対して BLEU-4 の値が高い傾向になり，小さい物体 (例：bar soap) に対して BLEU-4 の値が低い傾向にある．

キャプションの正確性の評価: 表 4.5 にキャプションの正確性評価の結果を示す．

P_change_occlusion での結果と比較して，2 つのモデルとも性能劣化した．変化前後のシーンをそれぞれ 4 視点から観測する提案モデルが比較的に最も良い性能を得た．また，同じ視点の設定をした場合，P_change_occlusion データセットと同様に提案モデルの方が高い精度を示した．提案モデルが特に変化タイプの推定に優位である傾向にある．

定量的な結果: 3 つの結果例を図 4.8 に示す．物体オクルージョンが比較的に軽い傾向にある例 (d) においてすべてのモデルが正しい変化説明文を生成できた．変化が含まれない distractor の例 (e) において，すべての提案モデルが正しく変化を理解した．それと比較して，DUDA が不正確な変化説明文を生成した．また，例 (f) において，すべての提案モデルが正しく変化のタイプを推定し，DUDA が変化タイプと物体属性の両方とも不正確な推定

を行った．

4.3.4 考察

3 次元のシーンにおける変化を理解する能力が様々な実環境アプリケーションにおいて重要である．既存のシーン変化説明文生成手法が変化前後のシーンをそれぞれ単眼画像から観測を行っている．実環境において，単眼画像の情報が不足する場面が多いため，既存の単眼画像ベースの手法の応用場面が制限される．以上のところから，我々は 3 章で行った試みと同様に，ここで多視点認識をシーン変化説明文生成タスクに導入した．また，多視点からシーンを 3 次元的に理解し，それをベースとしたシーンの変化の理解を行っている．

ここで既存の単眼画像ベースの手法と提案モデルを単純な幾何物体モデルと実物体モデルを用いたデータセットで性能評価を行った．従来の画像説明文の評価指標と詳細的変化説明文の内容を評価するキャプション正確さの指標の両方において，提案モデルが従来手法を大きく上回る精度を達成した．更に，提案モデルが物体オクルージョンが含まれるシーンやカメラの回転変換が発生する場合でも良好な性能を示した．

シーン変化説明文生成タスクがホームロボットなどの実環境アプリケーションにおいて重要であると想定できる．従来のロボティクスのアプリケーションでは RGB-D センサーが広く使われているため，次の節によりデプス画像や RGB-D 画像から得られる点群データがこのタスクでの性能における影響の分析を行う．

4.4 多視点・複数モダリティに対応した変化説明文生成

4.4.1 概要

図 4.11: 多視点・複数モダリティに対応した変化説明文生成の概念図

節 4.3 にも述べたように，3 次元シーンの変化を理解する能力が様々な実環境アプリケーションにおいて重要である．しかし，既存のシーン変化説明文生成手法は変化前後のシーンの単眼画像から変化の認識を行っている．視覚情報がより複雑な実環境において，単眼画像からの情報が不足である場面が多く，従来の単眼画像変化説明文生成手法の実用性が制限される．以上の問題点から，我々はまず節 4.3 により，多視点認識を変化説明文生成タスクに導入し，多視点からのシーン変化理解を可能にした．

しかし，節 4.3 では背景が一色となるより単純なシミュレーションのテーブルシーンでしか検討を行っていない．また，変化説明文生成タスクがロボットアプリケーションにおいての有用性が高いと思われる．従来，ロボットアプリケーションにおいて RGB センサーのほか，デプス値を取得できるデプスセンサーも広く使われている．

上記の問題点から，ここで大規模な質問シーンデータセットを導入し，より複雑な 3 次元背景を持ったシミュレーション環境のシーン変化説明文のデータセット自動構築プロセスを提案する．更に，図 4.11 に示しているように，仮想環境に RGB-D カメラを設定し，多視点 RGB 画像のほか，多視点デプス画像および点群を取り扱うフレームワークを提案する．

我々は提案した異なる入力モダリティおよび異なる複雑度を持ったデータセットにおいて実験を行った．実験結果により，提案のマルチモダリティをアンサンブルしたモデルが視覚複雑度が高いデータセットにおいても高い精度で変化の説明文を生成できた．

4.4.2 Matterport3D データセット

図 4.12: Matterport3D データセットの概念図 (Chang et, al. [71] により)

Matterport3D [71] が 2017 年 Chang らにより提案された大規模室内シーンデータセットである．図 4.12 に示しているように，Matterport3D データセットに多層ビルのテクスチャー付き 3 次元メッシュモデル，RGB 画像，デプス画像，セマンティックセグメンテーションマスクが密にアノテーションされている．

Matterport3D データセットは Matterport カメラにより撮影し Matterport 社によりテクスチャー付き 3 次元メッシュの生成を行った．データセットには室内部屋 90 ビル合計 2,056 部屋から構成される．また，データセットに 50,811 物体インスタンス及び 1,659 物体クラスがアノテーションされている．

Matterport3D データセットが 3 次元シーンの認識タスクに大きく貢献し，現在，3 次元シーンの識別，物体検出，セグメンテーション，更にコンピュータビジョンと自然言語処理のマルチモーダルタスクの embodied question answering [34, 35, 36, 37] や vision-language navigation [72, 73, 74], vision-and-dialog navigation [75, 76, 77] などに広く使われている．

4.4.3 提案構造

RGB-D センサーがロボットアプリケーションで広く使われているため，ここで節 4.3 で述べた多視点 RGB 画像の入力に加えて，RGB-D センサーにより得られる多視点デプス画像および点群を取り扱えるフレームワークを提案する．提案したフレームワークが多視点 RGB とデプス画像および点群から end-to-end で変化前後のシーンに含まれる変化の説明文を生成できる．また，提案したフレームワークは任意数 (1 つ以上) の入力モダリティから行える．

図 4.13: 多視点・複数モダリティに対応した変化説明文生成のフレームワーク概念図

図 4.13 に全体的フレームワークのイメージ図を示す．提案フレームワークが変化前後のシーンの複数モダリティの観測から変化の内容を説明文として出力する．また，ここで一つシーンの観測が多視点 RGB 画像，多視点デプス画像および点群から構成される．

全体的フレームワークが 3 つのサブ構造から構成される：(1) シーンエンコーダー：入力のモダリティを別々のエンコーダーにより特徴抽出する；(2) シーンフュージョン構造：変化前後のシーン特徴をフュージョンする構造；(3) 説明文生成器：(2) によりフュージョンされた特徴から変化の説明文を生成する構造．このフレームワークは更なるモダリティの追加が可能である．また，異なる構造 (例：シーンエンコーダー，説明文生成器) が他のモデルにより精度を高めることが可能となる．

シーンエンコーダー：この構造で入力の多視点 RGB 画像，デプス画像及び点群を 1 つの特徴量に変更する．具体的に，図 4.13 に示しているように，まず入力された多視点 RGB 画像，デプス画像及び点群をそれぞれ別々のエンコーダーにより特徴抽出を行う．次に，concatenation 操作により 3 つの特徴量を結合する．RGB 画像のエンコーダーに対して，MVCNN [44] と SRN [45] を用いた．デプス画像を MVCNN により特徴抽出を行った．また，Qi らにより提案された PointNet [46] により点群の特徴抽出を行った．次に，これらのエンコーダーの実装の詳細を述べる．

図 4.14: 多視点・複数モダリティに対応した変化説明文生成のフレームワーク図

RGB/デプス画像エンコーダー (MVCNN)：ここで Su らにより提案された MVCNN 構造に調整を加えた．図 4.14 (a) に示しているように，我々の実装ではまず ImageNet で学習済みの ResNet101 により異なる視点の画像の特徴を抽出する．また，デプス画像を事前に RGB 画像に変換した (OpenCV に実装されている applyColorMap 関数を使用し，COLORMAP_JET の色対応でデプス画像を RGB 画像に変更した)．次に，それぞれの多視点画像特徴に別々の畳み込み層により処理する．次に，全結合層とソフトマックス関数により畳み込んだ多視点画像特徴から重みベクトルを求める．次に，ResNet101 により抽出された画像特徴を畳み込み層に処理し，重みベクトルと掛け算を行う．以上の操作により，最終的に $4 \times 4 \times 128 \times 8$ 次元の特徴量を得られる．

RGB 画像エンコーダー (SRN)：ここで，節 3.5, 3.6 および 4.3 で用いた Eslami らにより提案された SRN を使用した．ここでも tower-構造の SRN を使用した．我々の実装が図 4.14 (b) に示す．まず GQN の全体的フレームワークを事前学習する．学習済みの SRN により多視点画像からシーンの特徴量を抽出する．

点群エンコーダー (PointNet)：ここで Qi らにより提案された PointNet を実装した．我々の実装の詳細を図 4.14 (c) に示す．

以上の三つのエンコーダーにより，異なる入力モダリティを $1 \times 1 \times k$ 次元の特徴に変更可能である (異なる入力モダリティの k が異なる)．3 つの特徴を concatenation 操作により結合する．このような構造で，任意数の入力モダリティを同様にプロセスできる．

変化前後の特徴の結合：ここでも節 4.3 で述べたような仕組みで変化前後の特徴量の結合を行う．具体的に，変化前後のシーンの特徴を r_{bef} と r_{aft} に表示し，式 (4.1) と同様に r_{diff} を求める．次に，r_{bef}，r_{aft} および r_{diff} を concatenation 操作により結合し，説明文生成器に入力する．

説明文生成器：説明文生成器が前述の r_{bef}，r_{aft} および r_{diff} を結合した特徴量から変化の内容を表すセンテンスを生成する．ここで図 4.12 で示しているような 2 層の LSTM 構造を持った生成器を実装した．生成された変化説明文を $\mathbf{x}$ に表示し，ground truth 説明文を $\mathbf{y}$ に示し，全体的フレームワークは以下の式で表すロス L を最小化するように学習される．

$$L = -\log(\mathrm{P}(\mathbf{y}|\mathbf{x})) \tag{4.14}$$

4.4.4 実験

4.4.4.1 データセット

室内環境のシーンの視覚複雑度が高いため，室内環境の理解をターゲットした認識モデルを学習するのに膨大な学習データが必要となる．また，手動で室内環境のデータセットを構築するために非常に労力がかかる．しかし，大規模な室内環境のシーン変化説明文データセットが提案されていない．

上記の問題を解決するため，大規模な室内シーンの 3 次元データセット Matterport3D をベースとした室内シーン変化説明文生成用データセットの自動構築プロセスを提案する．Matterport3D シーンに物体モデルをランダム配置および配置変更により変化前後のシーンの作成を行っている．また，ここでも節 3.7 で用いた物体 3 次元モデルデータセット NEDO Items，YCB から物体モデルを集めた．ここで節 4.3 と類似したように 4 タイプの変化：add; delete; move および replace を定義した．また，ここも節 4.3 と同様に distractor を設置した．distractor では変化前後のシーンの間にカメラの微小な位置変更を設定している．

表 4.6: データセットの物体クラスとインスタンス

Setups	Class (Number of Instances)
10-object setup	barsoap (2); cup (2); dishwasher (2); minicar (2); snack (2)
85-object setup	snack (12); dishwasher (11); tv dinner (11); minicar (7); barsoap (5); cup (5); plate (3); soft drink (3); sponge (3); air brush (2); baseball (2); bowl (2); facial tissue (2); magic marker (2); sauce (2); water bottle (2); weight (2); glass (1); glue (1); shampoo (1); soccer ball (1); tape (1); teddy bear (1); timer (1); toilet tissue (1); tooth paste (1)

シーンと物体モデル：前述したように，我々は 3 次元シーンに物体をランダムに配置することで変化前後のシーンを作成している．Matterport3D データセットが 90 建物の 3 次元メッシュから構成される．ここで，Matterport3D データセットから 115 個の立方体に近似した形状をする部屋をサンプリングした．また，NEDO と YCB データセットから集めた物体モデルのクラスとクラスごとに物体モデル数を表 4.6 に示す．

図 4.15: 仮想カメラ設定図

仮想カメラ設定：節 4.3 と類似するように，ここでも変化前後のシーンを多視点から観測する．変化前後のシーンが部屋になるため，我々は図 4.15 のように部屋ごとに 8 台の仮想カメラから観測を行う．天井の 4 つの角と天井の四つのエッジの真ん中に合計 8 台のカメラを設定した．また，全てのカメラを部屋の中心に向くように調整した．また，カメラの位置変動に対してロバストにするために，シーンを撮影するたびに，仮想カメラごと 3 つの軸に対してに [–10.0 cm, +10.0 cm] のランダム位置変動を設定した．

図 4.16: データセットでの一例

データセット自動構築プロセス：AI Habitat シミュレーター [78] を用いて，シーンと物体の 3 次元メッシュデータから同時に RGB-D 画像および点群を取得している．AI Habitat シミュレーターでは指定された視点から RGB-D 画像と点群を得られる．ここで 4 つのステップで変化前後のシーンの観測データおよび変化説明文の生成を行っている．まず，前述した 115 個の部屋と物体モデルからランダムに一つの部屋と物体モデル (3 から 5 個) を選定する．次に，物体モデルをランダムにシーンの床やテーブルまたはベッド上に配置する．図 4.15 で示す仮想カメラ位置から 8 枚の RGB 画像及びデプス画像を撮影する．また，AI Habitat により同時に部屋と物体モデルの三角形メッシュを得て，メッシュの頂点をサンプリングすることにより点群データを同時に集めた．次に，4 つの変化 (add, delete, move, replace) および distractor を生成し撮影する．4 つの変化は物体モデルの位置の再配置により生成できる．また，distractor を作成する際に，カメラの位置をランダムに変更して撮影を行った．変化と distractor を生成する際に，同時にシーンの変化を記録する．最後に，記録したシーン変化および予め定義したセンテンスの構造により変化の説明文を自動生成する．以上のプロセスを用いて生成したデータセットの 1 例を図 4.16 に示す．以上のようなプロセスにより，シーンの数および物体の数などを調整することで，異なる複雑度を持ったデータセットの生成が可能になる．

表 4.7: データセット統計情報.

Dataset	No. of Scenes (Train/Test)	No. of Captions (Train/Test)	Change Types	Viewpoints	Scene Types	Object Classes	Objects Per Scene
s15_o10	(9000/3000)	(225,000/75,000)	5	8	15	10	2–6
s15_o85	(9000/3000)	(225,000/75,000)	5	8	15	85	2–6
s100_o10	(9000/3000)	(225,000/75,000)	5	8	100	10	2–6

データセットの統計情報：まず 15 個のシーンおよび 10 個の物体モデルから構成するデータセット s15_o10 を生成した．物体のクラス分布は表 4.6 に示す．s15_o10 データセットを用いて異なる入力モダリティ，入力モダリティのエンコーダーを用いたモデルの性能評価を行った．

更に，異なる複雑度を持ったデータセットに対しの性能評価を行うために，15 シーンおよび 85 個の物体 (物体のクラス分布は表 4.6 に示す) から構成されるデータセット s15_o85 および 100 シーン 10 物体を持ったデータセット s100_o10 を生成した．これらの 3 つのデータセットの統計情報を表 4.7 に示す．

4.4.4.2 実験設定

データセット s15_o10，s15_o85 および s100_o10 を使用してモデルの学習と性能評価を行った．具体的に，s15_o10 を使用して異なる入力モダリティおよびエンコーダーを用いたモデルの性能評価を行った．また，s15_o85 データセットにより，異なる物体数がモデルの性能に与える影響と s100_o10 を使用して異なるシーンの数の影響の分析実験を行った．

全ての実験で，RGB およびデプス画像のサイズを 64*64 に設定した．また，点群のポイントの数をランダムサンプリングにより 5000 に調整した．0.0001 の学習率で，SRN の事前学習を 10 エポックを行った．また，全ての実験において全体的説明文生成のフレームワークを 0.001 の学習率で 40 エポック学習した．PointNet エンコーダーの学習率を 0.0001 に設定し，全ての実験で Adam オプティマイザーを使用した．

4.4.4.3 異なる入力モダリティと画像エンコーダーでの実験

表 4.8: s15_o10 (上 11 行), s15_o85 (中間 6 行), と s100_o10 (ボトム 6 行) での定量評価

Modality (Encoder)	ROUGE [?]	SPICE [?]	METEOR [?]	BLEU-4 [?]					
				Overall	Add	Delete	Move	Replace	Distractor
Depth (MVCNN)	58.36	18.46	26.50	36.93	21.30	39.15	42.22	32.90	45.11
RGB (MVCNN)	65.53	26.52	34.80	49.43	25.73	49.83	54.18	51.83	60.60
RGB (GQN)	81.98	36.02	48.49	71.01	59.59	77.31	59.52	68.52	85.80
PCD (PointNet)	89.75	35.32	49.34	76.17	62.89	79.19	77.98	71.73	96.19
Depth,RGB (MVCNN,MVCNN)	80.74	34.44	46.30	67.58	55.04	71.82	69.25	63.41	75.80
Depth,RGB (MVCNN,GQN)	84.52	36.51	50.71	74.46	63.99	77.12	65.49	71.48	90.22
Depth,PCD (MVCNN,PointNet)	84.41	27.74	43.03	65.91	54.84	68.73	59.37	58.00	98.85
RGB,PCD (MVCNN,PointNet)	89.31	36.40	50.50	75.65	65.97	80.02	75.76	70.93	91.00
RGB,PCD (GQN,PointNet)	92.36	41.36	57.00	84.74	81.65	**89.87**	70.67	80.77	**99.41**
Depth,RGB,PCD (MVCNN,MVCNN,PointNet)	89.87	35.99	52.11	79.02	64.18	82.38	77.06	73.49	96.51
Depth,RGB,PCD (MVCNN,GQN,PointNet)	**93.38**	**44.29**	**58.46**	**86.33**	**83.32**	89.69	**80.21**	**82.84**	98.55
Depth (MVCNN)	57.56	15.27	24.19	36.46	12.44	33.70	30.21	19.83	74.73
RGB (MVCNN)	57.31	15.17	24.80	35.67	18.08	31.78	32.46	29.78	59.91
RGB (GQN)	63.90	18.94	29.05	43.75	23.83	42.51	34.60	34.96	77.00
PCD (PointNet)	80.44	21.04	34.98	52.02	33.26	52.79	**55.59**	53.63	52.66
Depth,RGB,PCD (MVCNN,MVCNN,PointNet)	**81.26**	23.76	**38.72**	**59.79**	**43.52**	**60.90**	43.64	**55.96**	95.98
Depth,RGB,PCD (MVCNN,GQN,PointNet)	78.83	**24.49**	36.86	57.11	39.74	59.96	41.75	43.27	**97.69**
Depth (MVCNN)	56.06	16.51	24.56	33.46	16.45	31.00	35.64	29.72	48.53
RGB (MVCNN)	65.72	25.02	33.80	47.01	31.74	43.49	51.31	51.69	50.32
RGB (GQN)	65.23	23.91	32.24	47.10	35.31	46.79	47.36	39.92	58.85
PCD (PointNet)	82.93	26.29	41.48	63.30	47.22	66.04	49.33	62.50	**99.60**
Depth,RGB,PCD (MVCNN,MVCNN,PointNet)	**87.99**	35.72	**49.68**	**75.31**	**60.49**	**77.45**	**73.79**	**68.44**	96.72
Depth,RGB,PCD (MVCNN,GQN,PointNet)	86.07	**35.97**	47.23	73.10	60.46	74.30	63.08	67.83	99.02

まずデータセット s15_o10 を使用して，異なる入力モダリティ，モダリティの特徴を抽出するエンコーダーおよび異なる入力モダリティのアンサンブルを用いたモデルの性能評価を行った．ここで 4 種類の single-モダリティのモデル：デプス画像に MVCNN エンコーダーを対応したモデル；RGB 画像に MVCNN エンコーダーを対応したモデル；RGB 画像に SRN エンコーダーを対応したモデル；および点群に PointNet エンコーダーを対応したモデル；を実装した．また，5 種類の two-モダリティモデル (RGB, デプスと点群から任意 2 個を選定し，RGB において 2 種類のエンコーダーを適応した) および 2 種類の three-モダリティモデルを実装した．2 種類の three-モダリティモデルに異なる RGB エンコーダーを使用した (MVCNN と SRN).

定量的評価：表 4.8 (上 11 行) に s15_o10 のテストセットにおいての結果を示す．4 つの single-モダリティのモデルにおいて，PointNet エンコーダーを対応した点群を用いたモデルが最も高い精度を達成し，MVCNN エンコーダーを対応したデプス画像を使ったモデルが比較的に低い精度を得た．また，SRN エンコーダーを対応した RGB 画像を用いたモデルが MVCNN エンコーダーを使ったモデルより高い精度を得られた．点群から幾何情報および物体のエッジ情報を抽出しやすく，このような性質が提案のデータセットに有利であることを想定できる．また，SRN により多視点画像からシーンの 3 次元情報を含めた特徴を抽出可能であり，MVCNN エンコーダーより多視点の画像の対応に強い傾向にあることを想定できる．デプス画像に物体の色情報が含まれず，また MVCNN エンコーダーによりデプス画像から幾何情報の抽出が困難である可能性があり，デプス画像単独では良い性能を得ることが困難な傾向にある．また，デプス画像単独では良い性能を得られないが，デプス画像と RGB 画像を使ったモデルが RGB 画像単独の場合より高い精度を示した．これがデプス画像に含まれる幾何情報が RGB 画像から抽出しにくい傾向があり，デプス画像の幾何情報により RGB 画像単独のモデルの性能向上に貢献したと考えた．また，2 つの three-

モダリティモデルが全てのモデルにおいて最も良い精度を得られた．このような結果から，提案データセットにおいてモダリティをアンサンブルするのとによりモデルの性能を向上できると示唆された．

表 4.9: s15_o10 (上 11 行), s15_o85 (中間 6 行), と s100_o10 (ボトム 6 行) での説明文精度評価.

Modality (Encoder)	**Accuracy (%)**			
	Change Type	**Object**	**Color**	**Class**
Depth (MVCNN)	44.54	26.82	35.11	36.78
RGB (MVCNN)	49.26	49.52	62.16	53.18
RGB (GQN)	74.48	69.41	79.54	71.88
PCD (PointNet)	**97.24**	53.65	59.20	65.44
Depth,RGB (MVCNN,MVCNN)	73.59	62.53	73.74	66.38
Depth,RGB (MVCNN,GQN)	78.44	72.46	81.84	74.91
Depth,PCD (MVCNN,PointNet)	93.03	39.64	48.75	52.06
RGB,PCD (MVCNN,PointNet)	93.37	58.52	70.68	65.29
RGB,PCD (GQN,PointNet)	93.94	74.39	83.85	77.39
Depth,RGB,PCD (MVCNN,MVCNN,PointNet)	93.09	63.02	73.32	69.32
Depth,RGB,PCD (MVCNN,GQN,PointNet)	94.96	**75.35**	**84.05**	**78.30**
Depth (MVCNN)	52.42	9.13	22.10	19.73
RGB (MVCNN)	46.38	17.23	38.99	23.34
RGB (GQN)	59.68	17.33	34.45	26.77
PCD (PointNet)	**96.17**	14.62	27.53	27.28
Depth,RGB,PCD (MVCNN,MVCNN,PointNet)	92.55	**23.51**	**44.92**	**32.18**
Depth,RGB,PCD (MVCNN,GQN,PointNet)	90.75	20.29	37.65	**32.18**
Depth (MVCNN)	42.88	19.85	29.70	31.29
RGB (MVCNN)	50.66	43.58	58.18	47.73
RGB (GQN)	52.00	40.28	52.13	45.72
PCD (PointNet)	90.49	33.73	44.60	47.13
Depth,RGB,PCD (MVCNN,MVCNN,PointNet)	**91.54**	**55.28**	**69.60**	**61.64**
Depth,RGB,PCD (MVCNN,GQN,PointNet)	90.07	48.20	59.93	55.77

説明文正確さ評価：表 4.9 (上 11 行) に s15_o10 のテストセットにおいての説明文正確さ評価の結果を示す．ここで SRN エンコーダーを対応した RGB 画像を用いたモデルが点群単独のモデルより高い精度で物体と物体の属性の推定を行っている．それに対して，点群を使用したモデルが高い精度で変化のタイプを推定できた．点群から幾何情報および物体のエッジ情報を抽出しやすいところが点群が変化のタイプ推定において有利であることの原因であると想定した．また，RGB 画像と点群の両方を用いたモデルが single-モダリティのモデルより高い精度で物体の属性を推定した．このような結果から，点群に含まれる幾何情報および RGB 画像に含まれる物体の属性情報両方が提案データセットにおいて重要であると考えた．

図 4.17: s15_o10 データセットで結果例

図 4.17 に 2 つの結果例を示す．1 番目の例 (物体移動) において，すべての three-モダリティアンサンブルモデルが正しく説明文を生成できた．また，デプス画像と MVCNN エンコーダーを適応した RGB 画像を使用したモデルが正しく物体の属性 (cyan minicar) を正しく推定できたが，変化のタイプの推定に不正解を出力した．それと比較して，点群を使用した single-モダリティを使用したモデルが正しく変化のタイプを推定したが，物体の色の推定に不正解を予測した．2 番目の例では，すべての点群を使用したモデル及び SRN エンコーダーを対応した RGB 画像とデプス画像を使ったアンサンブルのモデルが正しい変化説明文を生成した．このような結果から，アンサンブルにより性能向上できることがわかり，点群が提案データセットの変化タイプの推定において有利である傾向が示唆された．

4.4.4.4 異なる複雑度を持ったデータセットでの実験

ここで s15_o10 の物体のクラス数を 85 個に変更し，より複雑な設定 s15_o85 でモデルの性能を評価した．この実験で 4 種類の single-モダリティモデルと 2 種類の three-モダリティの性能を評価した．

まず，従来の評価指標の結果を表 4.8 (中間 6 行) に示す．s15_o10 データセットでの精度と比較して，すべてのモデルの性能が劣化した．また，SRN を適応した RGB 画像を使った single-モダリティモデルが大幅に性能が劣化した．このような結果から，SRN モデルが高視覚複数度のシーンの適応性が低下することが原因であると想定した．ここでも s15_o10

と同様にすべてのの single-モダリティモデルにおいて点群を使用したモデルが最も良い性能を得られた．さらに，アンサンブルにより性能を向上できた．

説明文の正確さ評価の結果を表 4.9 (中間 6 行) に示す．ここでも従来の評価指標と類似するように，すべてのモデルの物体の属性の推定の精度が劣化した．また，ここで点群を使用したモデルが変化タイプの推定に s15_o10 と類似した性能を得られた．このような結果から，点群から抽出できる幾何情報およびエッジ情報が変化タイプの推定のロバスト性に貢献できると想定できる．

図 4.18 (トップ) に一つの結果例を示す．ここですべての single-モダリティモデルが間違えた変化説明文を生成した．二つの three-モダリティモデルが正しく変化説明文を推定した．デプス画像と RGB 画像を使った single-モダリティが間違えた変化のタイプを予測した．それと比較して，点群を使用したモデルが正しく変化のタイプを予測した．より多くのバリエーションの物体が含まれているため，s15_o85 データセットが s15_o10 より難しいことが想定できる．s15_o85 での結果から，複雑度が高いデータセットに対して，異なるモダリティをアンサンブルすることが高い性能を得るに有効であることが示唆された．

図 4.18: s15_o85, s100_o10 データセットで結果例

次に，シーンの数を増やしたデータセット s100_o10 を用いた評価の結果を述べる．ここで，100 シーンを使用した．

表 4.8 (下 6 行) に通常の画像説明文の評価指標の結果を示す．ここでも s15_o10 データセットでの精度と比較してすべてのモデルの性能が劣化した．また，s15_o85 データセットでの結果と類似したように，SRN エンコーダーを対応した RGB 画像を用いたモデルが最も性能を低下した．ここでも同様に，SRN ネットワークが複雑度が高いシーンへの対応性が不足することが示唆された．また，single-モダリティにおいて点群を使用したモデルが最も良い性能を示し，デプス画像が比較的に低い精度を得た．また，モデルをアンサンブルすることにより性能の改善が観測できた．表 4.9 (下 6 行) に示している説明文正確さの評

価結果から，各モデルが物体の属性の推定において性能を劣化し，点群を使用したモデルが比較的に変化タイプの推定にロバストな傾向があった．

図 4.18 (ボトム) に一つの結果例を示す．ここでは，点群を使用した single-モダリティモデルと three-モダリティモデル (MVCNN エンコーダーを対応した RGB 画像を使用したモデル) が正しい変化説明文を生成した．ここで SRN を使用したモデルが大幅に性能を劣化した傾向があった．

4.4.4.5 実験結果の分析

s15_o10, s15_o85 および s100_o10 を使用した実験で，single-モダリティモデルにおいて点群を使用したモデルが多数の評価指標に最も良い精度を達成し，デプス画像を使用したモデルが比較的に精度が低い傾向にある．また，説明文正確さの評価結果から，RGB 画像の使用が物体の属性の推定に有利であり，点群が変化のタイプの推定に有利であることが示唆された．アンサンブルによりモデルの性能を向上できる傾向が多くの実験に見られた．デプス画像単独の場合で性能が比較的に低いが，デプス画像が RGB 画像の補充情報として有効である傾向があった．

物体の数およびシーンの数を増加した設定の s15_o85 および s100_o10 において，s15_o10 での結果と比較して，すべてのモデルの性能が劣化した．しかし，点群を使用したモデルが変化タイプの推定に比較的強いロバスト性を示した．また，2 つの RGB 画像エンコーダーの MVCNN と SRN に対して，s15_o10 データセットにおいて SRN の性能が MVCNN エンコーダーを上回ったが，s15_o85 と s100_o10 において SRN を使用したモデルが比較的大きく性能が劣化した．このような結果から，MVCNN と比較して，SRN 構造が複雑度が高いシーンに対しての適応性が低いことが示された．

ここで述べた実験結果が，今後のシーン変化説明文生成に様々な知見をもたらした．シーンの変化を理解するために，幾何情報と色情報両方が良い性能を得るために重要であることがわかった．また，ここで異なるモダリティの特徴を単純な concatenation 操作により結合しているため，シーンおよびシーン変化に応じて選択的にモダリティの情報を使用するアテンション構造を導入することでモデルの性能を向上できると期待できる．ここで s15_o85 および s100_o10 データセットによりシーンの複雑度に対して分析実験を行った．物体の数およびシーンの数を更に様々な変更をしてもっと詳細的にシーンの複数度に対しての対応性を分析可能になると考える．

4.4.5 考察

人間の日常活動により，室内環境が常に変化する．3 次元シーンの変化を理解する能力が様々な実環境アプリケーションにおいて重要である．しかし，従来のシーン変化説明文生成手法が単眼画像からシーンの変化を理解するため，実環境への適応性が制限される．この問題に対して，節 4.3 により我々は多視点認識をシーン変化説明文生成タスクに導入した．4.3 節でよりシンプルなテーブルシーンだけに対して検討を行った．従来ロボットアプリケーションで RGB-D センサーが広く使用されることから，RGB-D の入力からシーン変化の理解を行うモデルがより望ましいと考えられる．

以上の問題点を踏まえて，ここでまず大規模な室内シーンの 3 次元データセットを導入し，より高い視覚複雑度を持ったシーン変化説明文生成データセットを提案した．更に，多視点 RGB 画像，デプス画像および点群を入力として取り扱えるフレームワークを提案した．異なる複雑度を持ったデータセットにより，提案の RGB 画像，デプス画像および点群をア

ンサンブルしたモデルが高い精度で変化説明文の生成に成功した．更に，提案モデルがより複雑なデータセットにおいても有効性を示した．

この章では，最初の大規模室内シーン変化説明文生成データセットおよび多視点・複数モダリティを用いたモデルを提案した．このデータセットおよび実験に得られた知見が今後のシーン変化理解分野の更なる研究に寄与できることを期待する．

4.5 本章のまとめ

シーン変化の理解が都市や地域の発展進行状況の把握，被災地状況の確認や災害復旧，部品管理システムの部品リストの更新などの様々な実環境応用において重要である．その一方，我々が第二章および第三章で述べた物体認識タスクと視覚質問応答タスクと同様に，重要のシーン変化説明文生成タスクでは単眼画像がメインで取り扱われている．

実環境アプリケーションでは単眼画像だけから必要な視覚情報が得られない場面が多く存在するため，この章でシーン変化説明文生成タスクに多視点・複数モダリティを導入し，それらの情報がこのタスクに与える影響について検討を行った．その結果，多視点・複数モダリティの観測により大幅に変化説明文生成の精度を向上可能である結果を示した．更に，提案モデルが物体のオクルージョンが含まれるシーンや変化前後のカメラ間にカメラ変動がある設定においても従来手法より有利である結果が得られた．シーン変化説明文生成タスクがロボットアプリケーションにおいて重要であり，そういった分野で RGB-D センサーが広く使用されている．そこで，我々は更に多視点 RGB 画像以外に，デプス画像と点群を取り扱えるモデルを導入した．我々の RGB，デプス画像および点群をアンサンブルしたモデルが視覚複雑度が高い質問シーンの変化説明文の生成においても高い精度を示せた．

この章では，物体認識，視覚質問応答およびシーン変化説明文生成タスクにおいて多視点・複数モダリティ認識能力の導入を検討し，それぞれのタスクにおいての有効性を示した．今後，AI が我々の日常生活や生産活動でさらに活用されることが想定できるため，今後の AI システムの構築に多視点や複数モダリティの使用することでより高精度を実現可能であることを示唆した．

第5章　結論

本論文について次のようなことが示された．

第二章では，物体認識タスクにおいて，従来の単眼画像から認識を行う物体認識タスクに視点バイアス現象が起こることを明示化した．また，このような視点バイアス傾向が画像認識用データセット ImageNet データセットだけではなく，複数の物体が同時にアノテーションされている物体検出データセットにおいても存在することを示した．次に，そのような視点バイアスの原因を調べるために，手作業により ImageNet データセットに含まれるクラスの視点を集計し，データセットのクラスの画像において視点バイアスが存在することを明らかにした．実環境で物体認識を行う際に，継続的に良好的な視点から物体を観測することが困難であると想定できる．そこで我々は複数視点から物体を認識し，認識の結果を物体が存在する 3 次元領域に投票する仕組みを提案し，実験で提案の仕組みで単一視点により誤認識する傾向が抑えられ，最終的に統計的に安定的な認識が行えることを示した．

第二章より，多視点観測が物体認識タスクにおいての有効性を示したことから，第三章では，多視点および複数モダリティを視覚質問応答タスクに導入した．従来の視覚質問応答では RGB 画像だけから回答を行い，RGB 画像から十分質問を回答するための 3 次元情報を得られないことを想定し，RGB-D 画像を用いた視覚質問応答手法を提案した．RGB-D 画像により視覚質問応答の精度，特に幾何情報が必要となる質問の精度が改善された結果が得られた．RGB-D 画像の結果により 3 次元情報が視覚質問応答タスクにおいての重要性が示唆されため，次に，多視点からシーンを観測し，観測された画像からシーンの 3 次元情報を含めた特徴表示を視覚質問応答タスクに導入した．実験結果により，多視点から得られるシーンの 3 次元情報が精度を向上できることを示した．人間がシーンに対しての質問を回答する時に，能動的に必要な視点を選択し回答を行うため，我々は次に，能動的に視点選択フレームワークを多視点視覚質問応答に導入し，高い精度で質問を回答しながら大幅に必要な視点を減らせた結果が得られた．また，従来多視点視覚質問応答データセットで単純な幾何物体により合成された CG データセットがメインで，我々はリアル撮影により得られる物体モデルをベースとしたデータセットを提案した．従来の CG データセットの補充として，今後多視点視覚質問応答タスクの性能向上に寄与できると幸いである．

第三章より，多視点およびデプス画像が視覚質問応答タスクにおいての有用性を示したため，我々は次に視覚質問応答タスクと良く類似するシーン変化説明文生成タスクに多視点認識を導入した．ここでも同様に，多視点の観測により大幅に変化説明文生成の精度を向上可能である結果を示した．更に，提案モデルが物体のオクルージョンが含まれるシーンや変化前後のカメラ間にカメラ変動がある設定においても従来手法より有利である結果が得られた．シーン変化説明文生成タスクがロボットアプリケーションにおいて重要であり，そういったアプリケーションで RGB-D センサーが広く使用される．そこで，我々は更に多視点 RGB 画像以外に，デプス画像と点群を取り扱えるモデルを導入した．我々の RGB，デプス画像および点群をアンサンブルしたモデルが視覚複雑度が高い質問シーンの変化説明文の生成においても高い精度を示せた．

以上，従来の単眼物体認識タスクに認識精度が観測視点に依存することを示した．更に，認識モデルの視点バイアスを直感的に評価できる視点—尤度ヒートマップおよび視点バイ

アスの定量化指標を提案した．工業現場などの実環境アプリケーションで物体認識を行う場合，視点バイアスが存在することへの意識向上や，安定的に物体認識を行うためのシステムの構築に寄与できると期待できる．また，実環境アプリケーションにおいて複数の視点またはから観測が行えることから，我々は物体認識，視覚質問応答およびシーン変化説明文生成の 3 つのタスクに多視点または複数モダリティを導入し，それぞれのタスクにおいて多視点および複数モダリティが性能向上に有効であることを示した．本研究で得られる知見が，今後高性能のコンピュータビジョンの実環境アプリケーションの構築に応用できると幸いである．

本研究で行った実験で，多視点および複数モダリティが複数のコンピュータビジョンタスクの性能向上に有効であることを示し，更に能動的視点選択の効率性を示せた．今後は，実環境で人間と共生しつつ，実用レベルの作業を行うロボットの実現のために，より積極的な視点選択法の検討や，画像以外のセンサの活用などについても検討を行っていく予定である．

謝辞

本研究を進めるにあたり，多くのご指導をいただきました産業技術総合研究所／筑波大学システム情報工学研究科佐藤雄隆教授 (連携大学院) に御礼を申し上げます．また，副指導教員および本論文の副査として支えていただきました福井和広教授に御礼を申し上げます．

2016 年の 4 月から，佐藤先生の研究室で研究させていただきました．これまでの 5 年間で，研究課題設定の方法，研究の進め方，論文調査のやり方，日本語や英語論文の作成方法，研究プレゼンテーション資料の作り方や発表の仕方など，研究に関する様々な方面で佐藤先生にご指導を賜りました．ここに心から感謝の意を表します．

本論文の副査として，ご助言を賜りました筑波大学システム情報工学研究科亀山啓輔教授，滝沢穂高准教授，北原格教授に厚く御礼申し上げます．

産業技術総合研究所知能システム研究部門（2020 年 4 月から人工知能研究センター）コンピュータビジョン研究グループにおきまして，連携大学院生およびリサーチアシスタントとして最先端の研究に携わるという大変 貴重な経験をさせていただきました．ミーティングをはじめ，様々な場面で適切なご指導を多数くださった岩田健司様，片岡裕雄様，鈴木亮太様，石塚光様，佐藤智実様に心より感謝いたします．アシスタントの尾形仁子様には研究生活に関わる様々なサポートで充実した 5 年間を過ごさせていただきました．

研究室の後輩であり，研究のアドバイス，博論の日本語チェック，そして私生活に至るまで多大なアドバイスをくださった中嶋航大君，中野真理子さん，松本晨人君，海老澤颯君に心より感謝いたします．

最後になりましたが，今まで支えてくれた両親に感謝いたします．

発表論文

[1] Yue Qiu, Yutaka Satoh, Ryota Suzuki, Kenji Iwata, Hirokatsu Kataoka. (2020). Indoor Scene Change Captioning Based on Multimodality Data. Sensors, 20(20), 4761.

[2] Yue Qiu, Yutaka Satoh, Ryota Suzuki, Kenji Iwata, Hirokatsu Kataoka. (2020). 3D-Aware Scene Change Captioning from Multiview Images. IEEE Robotics and Automation Letters (RA-L), 2020.

[3] Yue Qiu, Yutaka Satoh, Ryota Suzuki, Kenji Iwata, Hirokatsu Kataoka. (2020). 3D-Aware Scene Change Captioning From Multiview Images. International Conference on Intelligent Robots and Systems (IROS), 2020.

[4] Yue Qiu, Yutaka Satoh, Ryota Suzuki, Kenji Iwata, Hirokatsu Kataoka. (2020). Multi-View Visual Question Answering with Active Viewpoint Selection. Sensors, 20(8), 2281.

[5] Yue Qiu, Yutaka Satoh, Ryota Suzuki and Kenji Iwata, "Multi-view Visual Question Answering Dataset for Real Environment Applications," 22nd International Conference on Human-computer Interaction (HCI), 2020.

[6] Kenji Iwata, Ryota Suzuki, Yue Qiu and Yutaka Satoh, "Single Image Contrast Enhancement by Training the HDR Camera data," 22nd International Conference on Human-computer Interaction (HCI), 2020.

[7] 邱 YUE，佐藤 雄隆，鈴木 亮太，片岡 裕雄，岩田健司，“一般物体認識用大規模画像データセットにおける視点依存性に関する考察及び定量化 ”，精密工学会論文誌，Vol.85，No.12，pp.1087-1093，2019.

[8] Yue Qiu, Yutaka Satoh, Kazuma Asano, Kenji Iwata, Ryota Suzuki and Hirokatsu Kataoka, "Incorporating 3D Information Into Visual Question Answering," ICCV 2019 3rd Workshop (ICCVW) on Closing the Loop Between Vision and Language, 2019.

[9] Yue Qiu, Yutaka Satoh, Hirokatsu Kataoka and Ryota Suzuki, "Incorporating 3D Information Into Visual Question Answering," 2019 International Conference on 3D Vision (3DV), Québec City, QC, Canada, pp. 756-765, 2019.

[10] Yue Qiu, Hirokatsu Kataoka, Ryota Suzuki, "Incorporating Depth into Visual Question Answering," CVPR 2019 Visual Question Answering and Dialog Workshop (CVPRW VQA), 2019.

[11] Yue Qiu, Hirokatsu Kataoka, "Visual Question Answering with RGB-D Images," CVPR 2019 Women in Computer Vision Workshop (CVPRW WiCV), 2019.

[12] QIU YUE，佐藤 雄隆，片岡 裕雄，鈴木 亮太，“ RGB-D 画像を用いた視覚質問応答 (VQA) に関する考察 ”，画像センシングシンポジウム (SSII2019), 2019.

[13] 飯樋 遼輝，鈴木 亮太，邱 YUE，岩田 健司，佐藤 雄隆，“ 任意の屋内環境での VR 地震シミュレーションにおける客観・主観的な危険度可視化およびその効果の調査 ”，インタラクション 2019 論文集, pp.221-226, 2019.

[14] Ryota Suzuki, Ryoki Iitoi, Yue Qiu, Kenji Iwata, Yutaka Satoh, "AI-Based VR Earthquake Simulator," 20th International Conference on Human-computer Interaction (HCI), pp. 213-222, 2018.

[15] Kaori Abe, Munekata Minoguchi, Teppei Suzuki, Tomoyuki Suzuki, Naofumi Akimoto, Yue Qiu, Ryota Suzuki, Kenji Iwata, Yutaka Satoh, Hirokatsu Kataoka, Fashion Culture Database: Construction of Database for World-wide Fashion Analysis, 15th International Conference on Control, Automation, Robotics and Vision (ICARCV), 2018.

[16] Yue Qiu, Hirokatsu Kataoka, "Image Generation Associated With Music Data, " CVPR 2018 Sight and Sound Workshop (CVPRW), 2018.

[17] 飯樋 遼輝，鈴木 亮太，QIU YUE，佐藤 雄隆，“ 物理演算に基づく屋内空間地震三次元ハザードマップの生成 ”，ビジョン技術の実利用ワークショップ (ViEW2018) 講演論文集，pp.350-355，2018.

[18] QIU YUE，佐藤 雄隆，鈴木 亮太，片岡 裕雄，“ 一般物体認識用大規模画像データセットにおける視点依存性に関する考察および定量化 ”，ビジョン技術の実利用ワークショップ (ViEW2018) 講演論文集，pp.356-362，2018.

[19] 鈴木 亮太，飯樋 遼輝，邱 YUE，岩田 健司，佐藤 雄隆，“ AI による三次元モデル自動構成に基づく VR 地震シミュレーションシステム ”，画像センシングシンポジウム (SSII2018), 2018.

[20] QIU YUE，佐藤 雄隆，鈴木 亮太，片岡 裕雄，“ 観測視点および物体の外観が物体識別器の性能に与える影響に関する考察 ”，画像センシングシンポジウム (SSII2018), 2018.

[21] 鈴木 亮太，飯樋 遼輝，邱 YUE，岩田 健司，佐藤 雄隆，“ AI 技術を用いた地震 VR シミュレーションシステム ”，インタラクション 2018 論文集, pp.670-673, 2018.

[22] (4) 邱 YUE，佐藤 雄隆，片岡 裕雄，鈴木 亮太，“ 多視点を前提とした三次元空間尤度投票型物体認識 ”，精密工学会論文誌，Vol.83，No.12，pp.1125-1130，2017.

[23] Yue Qiu, Yutaka Satoh, Ryota Suzuki, Hirokatsu Kataoka, "Sensing and recognition of typical indoor family scenes using an RGB-D camera," CVPR 2017 Women in Computer Vision Workshop (CVPRW WiCV), 2017.

[24] QIU YUE，陳 放歌，大喜 周平，片岡 裕雄，“ 音楽から連想される画像の自動生成 ”，ビジョン技術の実利用ワークショップ (ViEW2017) 講演論文集，pp.209-213，2017.

[25] QIU YUE，佐藤 雄隆，鈴木 亮太，片岡 裕雄，“ 複数視点の観測による三次元空間尤度投票型物体認識 ”，ビジョン技術の実利用ワークショップ (ViEW2017) 講演論文集，pp.386-392，2017.

[26] QIU YUE，佐藤 雄隆，鈴木 亮太，片岡 裕雄，“ RGB-D カメラを用いた生活空間センシング及び機能解析に関する研究 ”，画像センシングシンポジウム (SSII2017), 2017.

[27] 鈴木 亮太，QIU YUE,　金崎 朝子，岩田 健司，佐藤 雄隆,“ 物体認識技術を活用したインタラクティブ VR 空間の構成 ”，画像センシングシンポジウム (SSII2017), 2017.

参考文献

[1] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., and Fei-Fei, L. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

[2] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[3] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. (2016). Inception-v4, inception-resnet and the impact of residual connections on learning. arXiv preprint arXiv:1602.07261.

[4] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[5] Fellbaum, Christiane. "WordNet." The encyclopedia of applied linguistics (2012).

[6] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2), 303-338.

[7] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... and Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.

[8] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Communications of the ACM 60.6 (2017): 84-90.

[9] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[10] Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence, 39(6), 1137-1149.

[11] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[12] Redmon, J., and Farhadi, A. (2017). YOLO9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263-7271).

[13] Torralba, A., and Efros, A. A. (2011, June). Unbiased look at dataset bias. In CVPR 2011 (pp. 1521-1528). IEEE.

[14] Misra, I., Lawrence Zitnick, C., Mitchell, M., and Girshick, R. (2016). Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2930-2939).

[15] Hendricks, L. A., Burns, K., Saenko, K., Darrell, T., and Rohrbach, A. (2018, September). Women also snowboard: Overcoming bias in captioning models. In European Conference on Computer Vision (pp. 793-811). Springer, Cham.

[16] Van Miltenburg, E. (2016). Stereotyping and bias in the flickr30k dataset. arXiv preprint arXiv:1605.06083.

[17] Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., ... and Xiao, J. (2015). Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012.

[18] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... and Desmaison, A. (2019). Pytorch: An imperative style, high-performance deep learning library. In Advances in neural information processing systems (pp. 8026-8037).

[19] Seeley, R. T. (1966). Spherical harmonics. The American Mathematical Monthly, 73(4P2), 115-121.

[20] Sturm, J., Engelhard, N., Endres, F., Burgard, W.,and Cremers, D. (2012, October). A benchmark for the evaluation of RGB-D SLAM systems. In 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 573-580). IEEE.

[21] Ng, P. C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research, 31(13), 3812-3814.

[22] M. A. Fischler and R. C. Bolles. Random Sample Consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6):381–395, June 1981.

[23] Rusinkiewicz, S., and Levoy, M. (2001, May). Efficient variants of the ICP algorithm. In Proceedings third international conference on 3-D digital imaging and modeling (pp. 145-152). IEEE.

[24] https://www.asus.com/jp/

[25] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. (2015). Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision (pp. 2425-2433).

[26] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6077-6086).

[27] Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., and Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847.

[28] Kim, J. H., Jun, J., and Zhang, B. T. (2018). Bilinear attention networks. Advances in Neural Information Processing Systems, 31, 1564-1574.

[29] Jiang, H., Misra, I., Rohrbach, M., Learned-Miller, E., and Chen, X. (2020). In Defense of Grid Features for Visual Question Answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10267-10276).

[30] Lei, J., Yu, L., Bansal, M., and Berg, T. L. (2018). Tvqa: Localized, compositional video question answering. arXiv preprint arXiv:1809.01696.

[31] Zhu, L., Xu, Z., Yang, Y., and Hauptmann, A. G. (2017). Uncovering the temporal context for video question answering. International Journal of Computer Vision, 124(3), 409-421.

[32] Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., ... and Rohrbach, M. (2019). Towards vqa models that can read. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 8317-8326).

[33] Biten, A. F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., ... and Karatzas, D. (2019). Scene text visual question answering. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4291-4301).

[34] Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., and Batra, D. (2018). Embodied question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 2054-2063).

[35] Yu, L., Chen, X., Gkioxari, G., Bansal, M., Berg, T. L.,and Batra, D. (2019). Multi-target embodied question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6309-6318).

[36] Das, A., Gkioxari, G., Lee, S., Parikh, D., and Batra, D. (2018). Neural modular control for embodied question answering. arXiv preprint arXiv:1810.11181.

[37] Wijmans, E., Datta, S., Maksymets, O., Das, A., Gkioxari, G., Lee, S., ... and Batra, D. (2019). Embodied question answering in photorealistic environments with point cloud perception. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6659-6668).

[38] Gurari, D., Li, Q., Lin, C., Zhao, Y., Guo, A., Stangl, A., and Bigham, J. P. (2019). VizWiz-Priv: a dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 939-948).

[39] Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., ... and Bigham, J. P. (2018). Vizwiz grand challenge: Answering visual questions from blind people. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3608-3617).

[40] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2017). Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6904-6913).

[41] Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2901-2910).

[42] Yang, Z., He, X., Gao, J., Deng, L., and Smola, A. (2016). Stacked attention networks for image question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 21-29).

[43] Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. C. (2018, January). FiLM: Visual Reasoning with a General Conditioning Layer. In AAAI.

[44] Su, H., Maji, S., Kalogerakis, E., and Learned-Miller, E. (2015). Multi-view convolutional neural networks for 3d shape recognition. In Proceedings of the IEEE international conference on computer vision (pp. 945-953).

[45] Eslami, S. A., Rezende, D. J., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., ... and Reichert, D. P. (2018). Neural scene representation and rendering. Science, 360(6394), 1204-1210.

[46] Qi, C. R., Su, H., Mo, K., and Guibas, L. J. (2017). Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 652-660).

[47] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 12(Jul):2121–2159, 2011.

[48] Watkins, C. J., Dayan, P. (1992). Q-learning. Machine learning, 8(3-4), 279-292.

[49] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.

[50] https://www.sony.jp/ichigan/

[51] https://afm.oxinst.cn/products/mfp-3d-afm-systems/

[52] Calli, B., Singh, A., Bruce, J., Walsman, A., Konolige, K., Srinivasa, S., ... and Dollar, A. M. (2017). Yale-CMU-Berkeley dataset for robotic manipulation research. The International Journal of Robotics Research, 36(3), 261-268.

[53] Singh, A., Sha, J., Narayan, K. S., Achim, T., and Abbeel, P. (2014, May). Bigbird: A large-scale 3d database of object instances. In 2014 IEEE international conference on robotics and automation (ICRA) (pp. 509-516). IEEE.

[54] Araki, R., Yamashita, T., and Fujiyoshi, H. (2018). Arc2017 rgb-d dataset for object detection and segmentation. In Late Breaking Results Poster on International Conference on Robotics and Automation.

[55] Kazhdan, M., Bolitho, M., and Hoppe, H. (2006, June). Poisson surface reconstruction. In Proceedings of the fourth Eurographics symposium on Geometry processing (Vol. 7).

[56] https://us.mitsubishielectric.com/fa/en/products/industrial-robots-melfa/intelligent-options/3d-vision/melfa3dvision

[57] https://www.blender.org/

[58] Chen, B., Chen, Z., Deng, L., Duan, Y., Zhou, J. (2016). Building change detection with RGB-D map generated from UAV images. Neurocomputing, 208, 350-364.

[59] Palandro, D., Andréfouët, S., Dustan, P., Muller-Karger, F. E. (2003). Change detection in coral reef communities using Ikonos satellite sensor imagery and historic aerial photographs. International Journal of Remote Sensing, 24(4), 873-878.

[60] Im, J., Jensen, J. R., Tullis, J. A. (2008). Object - based change detection using correlation image analysis and image segmentation. International Journal of Remote Sensing, 29(2), 399-423.

[61] Sakurada, K., and Okatani, T. (2015, September). Change Detection from a Street Image Pair using CNN Features and Superpixel Segmentation. In BMVC (Vol. 61, pp. 1-12).

[62] Sakurada, K., Okatani, T., and Deguchi, K. (2013). Detecting changes in 3D structure of a scene from multi-view images captured by a vehicle-mounted camera. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 137-144).

[63] Park, D. H., Darrell, T., and Rohrbach, A. (2019). Robust change captioning. In Proceedings of the IEEE international conference on computer vision (pp. 4624-4633).

[64] Jhamtani, H., and Berg-Kirkpatrick, T. (2018). Learning to describe differences between pairs of similar images. arXiv preprint arXiv:1808.10584.

[65] Shi, X., Yang, X., Gu, J., Joty, S., and Cai, J. (2020, August). Finding It at Another Side: A Viewpoint-Adapted Matching Encoder for Change Captioning. In European Conference on Computer Vision (pp. 574-590). Springer, Cham.

[66] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... and Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057).

[67] Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).

[68] Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).

[69] Banerjee, S., and Lavie, A. (2005, June). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization (pp. 65-72).

[70] Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2016, October). Spice: Semantic propositional image caption evaluation. In European Conference on Computer Vision (pp. 382-398). Springer, Cham.

[71] Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., ... and Zhang, Y. (2017). Matterport3d: Learning from rgb-d data in indoor environments. arXiv preprint arXiv:1709.06158.

[72] Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., ... and van den Hengel, A. (2018). Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3674-3683).

[73] Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L. P., ... and Darrell, T. (2018). Speaker-follower models for vision-and-language navigation. In Advances in Neural Information Processing Systems (pp. 3314-3325).

[74] Wang, X., Huang, Q., Celikyilmaz, A., Gao, J., Shen, D., Wang, Y. F., ... and Zhang, L. (2019). Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6629-6638).

[75] Thomason, J., Murray, M., Cakmak, M., and Zettlemoyer, L. (2020, May). Vision-and-dialog navigation. In Conference on Robot Learning (pp. 394-406). PMLR.

[76] Zhu, Y., Zhu, F., Zhan, Z., Lin, B., Jiao, J., Chang, X., and Liang, X. (2020). Vision-Dialog Navigation by Exploring Cross-modal Memory. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10730-10739).

[77] Roman, H. R., Bisk, Y., Thomason, J., Celikyilmaz, A., and Gao, J. (2020). RMM: A Recursive Mental Model for Dialog Navigation. arXiv preprint arXiv:2005.00728.

[78] Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., ... and Parikh, D. (2019). Habitat: A platform for embodied ai research. In Proceedings of the IEEE International Conference on Computer Vision (pp. 9339-9347).