一般発表 | Track 4: メディア処理・HCI・人間中心情報マネジメント

2026年3月1日(日) 15:30 ~ 17:40 | H会場

[6H] 画像分析技術

座長:松平 茅隼(NEC)　コメンテータ:陸 可鏡(山梨大学)　ジュニアコメンテータ:NGUYEN Trung Thanh(名古屋大学)

15:30 ~ 15:55

[6H-01] 表画像の補正と大規模言語モデルによる表構造解析手法の改良

*納田 朋享[1]、金澤 輝一[2]、上野 史[3]、太田 学[3] (1. 岡山大学大学院環境生命自然科学研究科、2. 国立情報学研究所コンテンツ科学研究系、3. 岡山大学学術研究院環境生命自然科学学域)

15:55 ~ 16:20

[6H-02] 日本語文書画像質問応答における参照構造分解と回答不能判定の分析

*山野 瑞月[1]、宮森 恒[1] (1. 京都産業大学先端情報学研究科先端情報学専攻)

16:20 ~ 16:45

[6H-03] 画像インペインティングを用いた展示物外観の意外性分析

*木下 真帆[1]、桑田 若菜[1]、三林 亮太[2]、大島 裕明[1] (1. 兵庫県立大学、2. 神戸大学)

16:45 ~ 17:10

[6H-04] 夜間運転時に不安感を誘発する要因の検出手法の提案とその定量的評価

*高岡 晴玖[1]、服部 峻[2]、宮城 茂幸[2] (1. 滋賀県立大学工学部電システム工学科、2. 滋賀県立大学先端工学研究院)

17:10 ~ 17:35

[6H-05] 学習済みモデルの特徴ベクトルに基づく未知個体への対応を考慮した地域猫の個体分類

*永尾 浩太[1]、服部 峻[2]、宮城 茂幸[2] (1. 滋賀県立大学大学院工学研究科電子システム工学専攻、2. 滋賀県立大学先端工学研究院)

表画像の補正と大規模言語モデルによる表構造解析手法の改良

納田　朋享† 金澤　輝一†† 上野　史††† 太田　学†††

† 岡山大学大学院環境生命自然科学研究科　〒 700–8530 岡山県岡山市北区津島中 3–1–1
†† 国立情報学研究所コンテンツ科学研究系　〒 101–8430 東京都千代田区一ツ橋 2–1–2
††† 岡山大学学術研究院環境生命自然科学学域　〒 700–8530 岡山県岡山市北区津島中 3–1–1
E-mail: †pwnk12a9@s.okayama-u.ac.jp, ††tkana@nii.ac.jp, †††{uwano, ohta}@okayama-u.ac.jp

あらまし　実環境で撮影された表画像には，撮影角度に起因する幾何学的な歪みが生じやすく，これは表構造解析の精度低下を招く要因となる．そこで本研究では，実環境での撮影を模して人工的に台形歪みを加えた表画像を生成し，生成した表画像の歪みを補正する手法を提案する．また，近年様々な分野で活用されている大規模言語モデル（LLM）に着目し，Ye らが提案した表構造解析モデルである TableMASTER の表構造解析結果を，LLM を用いて修正する手法を提案する．評価実験では，ICDAR 2021 Competition on Scientific Literature Parsing（ICDAR2021-SLP）のテストデータ 1,000 件の表画像を解析し，表構造解析精度を Tree-Edit-Distance-based Similarity（TEDS）を用いて評価した．実験の結果，台形歪みを加えたテストデータを用いた検証では，台形歪みの補正によって TEDS の値が 0.6333 から 0.7205 へ，画像の構造的類似度を示す Multi-Scale Structural Similarity Index（MS-SSIM）は元の表画像との比較において 0.3272 から 0.5648 へ改善した．また，歪みのない表画像のテストデータから得られた解析結果に対する検証では，LLM による修正によって TEDS の値が 0.9596 から 0.9599 へ向上した．

キーワード　表構造解析，表画像，大規模言語モデル（LLM）

1 はじめに

学術論文において，実験結果や統計情報は表としてまとめられることが多い．表構造を解析できれば，表から視覚的に優れたグラフへの自動変換 [2] や情報の抽出 [3]，複数の表の集約といった応用が可能となり，論文理解の効率化に大きく寄与する．しかし，表の様式は一般に著者によって異なり多様であるため，罫線の有無やマルチカラムセルなどを考慮した表構造解析が必要となる．また，表画像を解析できれば，スクリーンショットや古い文献のスキャン画像，手書きの表など，メタデータを持たない表に対しても構造解析が可能となり，その汎用性は高い．このような背景から，表画像を入力とする表構造解析の研究が活発に行われている [4] [5].

しかし，実環境での利活用を想定した場合，入力される表画像は真正面から影や歪みなく撮影された理想的な画像であるとは限らない．撮影時のカメラ角度に起因する台形歪みなどが生じている場合，既存のモデルでは正しく構造を認識できないことが多い [6].

近年，大規模言語モデル（Large Language Model，LLM）およびマルチモーダル LLM の急速な発展により，大規模事前学習で獲得された高度な言語知識と推論能力を活用した文書理解が可能となっている．これにより，従来の OCR やルールベース手法では困難であった文書の意味的理解が実現されつつある．具体的には，文脈や意味整合性を考慮した OCR 結果の修正 [7]，表やチャートからの構造化データの復元 [8]，さらに文書質問応答タスクにおいても，視覚情報と数値情報を統合的に理解する能力の有効性が示されている．これらの性能は，DocVQA [9] などの評価基盤を通じて広く検証・活用されている．これらの研究成果は，従来の解析モデルによって得られた出力結果に対し，LLM を用いて意味的な検証および修正を行う後処理が，文書および表情報の実用的な利活用において重要な役割を果たすことを示唆している．

筆者らは先行研究において，Ye らの表構造解析手法 [4] の分析および追加学習による改良について報告した [1]．そこで本研究では，Ye らの表構造解析手法 [4] を基盤モデルとして採用し，実環境での頑健性と解析精度の向上を目的として，台形歪みの補正および LLM による表構造情報の修正を提案する．提案手法は，台形歪みを加えた表画像の補正と LLM を用いた表構造解析結果の修正の 2 つから構成される．

台形歪みの補正では，人工的に台形歪みを付与した画像データセットを生成し，それを用いて画像を正規化する．本処理により，歪みによって損なわれた画像の幾何学的特徴を復元し，表構造解析モデルが想定する入力品質へ近づける．LLM を用いた修正では，表構造解析モデルが出力した HTML コードを入力とし，LLM の推論能力を活用して表構造の誤りおよびセルテキストの誤りを修正する．

評価実験では，ICDAR 2021 Competition on Scientific Literature Parsing（ICDAR2021-SLP）[10] のテストデータを解析する．評価指標は 2 つある．まず，台形歪みの補正では Multi-Scale Structural Similarity Index（MS-SSIM）を用いて画像の補正品質を評価する．Tree-Edit-Distance-based Similarity（TEDS）を用いて表構造解析精度を評価する．LLM による表構造解析結果の修正実験では，Tree-Edit-Distance-based Similarity（TEDS）を用いて表構造解析精度を評価する．

本稿の構成は以下の通りである．第 2 節では関連研究につい

て述べる．第 3 節では台形歪みを加えた表画像の補正手法，第 4 節では LLM を用いた表構造解析結果の修正手法について述べ，第 5 節では提案手法の有効性を検証するための評価実験について説明する．第 6 節でまとめる．

2 関 連 研 究

2.1 表画像を入力とする表構造解析

Ye らは，文書解析タスクである ICDAR2021-SLP において，表画像の表構造解析を「テキスト行検出」,「テキスト行認識」，および「セルへのテキスト割り当て」という 3 つのサブタスクに分割して処理するモデルである TableMASTER を提案した [4]．この手法では，表構造認識およびテキスト行認識のモデルとして，高精度な画像テキスト認識モデルである MASTER [11] を改良したモデルを採用している．特に表構造認識においては，HTML タグの予測とバウンディングボックスの予測を並列に行う構造を導入した．また，テキスト行検出には，任意の形状のテキストや近接するテキスト行を効果的に識別可能な PSENet [12] を利用した．最終的な HTML 生成に向けたボックス割り当てフェーズでは，検出されたテキストボックスとセルを関連付けるために，まず中心点ルール，次に IoU ルール，最後に距離ルールを順次適用するという，3 段階の階層的なマッチング規則を採用した．ICDAR2021-SLP [10] のテストデータセットを用いた評価実験において，提案手法は TEDS の値が 0.9632 となった．

Smock らは，物体検出手法である Detection Transformer（DETR）[13] を表構造解析問題に応用した Table Transformer を提案した [14]．Table Transformer は，表内の行，列，および見出し領域をそれぞれ独立した検出対象として直接推定し，それらの幾何的配置関係に基づいて表の論理構造を復元する手法である．彼らは，従来の表構造解析用データにおいて，行・列・セル間の対応関係に不整合が存在することが認識精度の低下を招いていると指摘し，これらの対応関係が一貫して定義された大規模表画像データセットである PubTables-1M を新たに構築し，これを用いて学習した．この枠組みにより，複数の行や列にまたがるセルを含む表や，罫線を持たない表に対しても，罫線情報に依存することなく，検出された行と列の交差関係から表構造を安定して推定できることを示した．実験の結果，Table Transformer は PubTabNet を用いた表構造解析において TEDS の値が 0.9360 となった．

2.2 歪みのある画像の補正

Bandyopadhyay らは，文書画像の幾何学的歪み補正において，畳み込みニューラルネットワークの一種である U-Net を拡張した RectiNet を提案した [15]．RectiNet では，画像内のエッジや境界線の詳細を捉えるための Gated Network と，密な歪み補正グリッドを予測する際にチャンネル間の情報の混在を防ぐための分岐型 U-Net を導入している点が特徴である．RectiNet は，約 8,000 枚の合成データによる学習でありながら，DocUNet データセット [16] を用いた評価において多重解像度での構造的類似度を測る Multi-Scale Structural Similarity Index（MS-SSIM）や局所的な歪みを測る Local Distortion（LD）といった指標で最先端の性能を達成した．

Zhu らは，歪んだ表画像では表構造解析の精度が低下するという課題に対し，表の構造的特徴を活用した U-Net ベースの新しい歪み補正モデルを提案した [17]．Zhu らは，まず，モデルが表の構造に着目できるよう，セルや表の罫線といったキー要素をセグメンテーションするモジュールを導入した．また，表の線分性を保つためには局所領域ではなく画像全体の歪みを把握する必要があるため，エンコーダに Transformer を組み込み，全体的な歪みを捉える能力を強化した．さらに，歪み補正の過程で生じるぼやけが可読性や評価指標に悪影響を与える点に着目し，軽量な鮮鋭化を後処理として適用することで最終的な画質向上を実現している．加えて，表画像の歪み補正に特化したデータセットが存在しなかったため，PubTabNet の HTML を再レンダリングし，歪みを加えることで 12,000 枚の合成データセットを独自に構築した．実験の結果，提案手法は従来の文書補正手法よりもすべての画質評価指標で優れた性能を示し，特に MS-SSIM では約 15 ポイントの大幅な改善が得られた．さらに，補正後の画像を用いて表構造を解析すると，TEDS スコアが約 6 ポイント向上した．

2.3 大規模言語モデルを用いた表構造解析結果の修正

Ren らは，表構造解析モデルが出力した結果を後処理によって修正，改善するアプローチとして，TableGLM [18] を提案した．Ren らの手法は，まず Transformer ベースの表構造解析モデルを用いて表画像から表の HTML コードを生成する．次に，この HTML コードを，表構造とテキスト内容の修正タスクに特化させてファインチューニングした LLM である TableGLM に入力する．TableGLM は，ChatGLM3-6B モデルを基盤とし，表構造解析モデルが生成した HTML コードと，それに対応する正解の HTML コードをペアにしたデータセットを用いて学習している．実験により，TableGLM による修正ステップを加えることで，PubTabNet データセットにおいて TEDS の値が平均 3.1 ポイント向上した．

Zhang らは，表構造解析を含む多様な表関連タスクに対応するための汎用モデルとして，TableLlama [19] を提案した．Zhang らの手法は，Llama 2 モデルを基盤とし，表構造の生成や修正を含む 15 種類の表タスクに対して適切に応答できるように追加学習を行う指示チューニングを行っている．TableLlama は，260 万件以上の表画像とテキストのペアを含む TableInstruct データセットを用いて学習されており，これには PubTabNet などの主要な表データセットから構築された構造解析タスクが含まれている．実験により，TableLlama は PubTabNet を含む複数のベンチマークにおいて，GPT-3.5 や GPT-4 などの汎用 LLM と同等以上の表構造理解能力を示し，7B パラメータという軽量なモデルでありながら高い TEDS の値を達成した．

3 台形歪みを加えた表画像の生成および補正

3.1 表の構成要素と表構造

本稿で扱う表画像は HTML タグによって表構造が定められる．本稿で扱う表および表に対応する HTML コードの例を図 1 に示す．

HTML 形式の表はヘッダ部分とボディ部分からなり，`<thead>...</thead>`は表の列の見出しを表すヘッダ行の部分を表しており，`<tbody>...</tbody>`は表のその下にあるボディ部分を表している．なお，他にも表のフッタ部分を表す`<tfoot>...</tfoot>`もあるが，本稿では扱わない．以下に本稿で扱う HTML タグをまとめる．

- `<thead>...</thead>`：表のヘッダ部分
- `<tbody>...</tbody>`：表のボディ部分
- `<b>...</b>`：太字表記
- `<i>...</i>`：斜体表記
- `<sup>...</sup>`：上付き文字
- `<sep>...</sep>`：行区切り文字
- `<tr>...</tr>`：表の行
- `<td>...</td>`：表のデータの各要素を表すセル
- `<td rowspan="n">...</td>`：n 個の垂直方向の結合セル
- `<td colspan="n">...</td>`：n 個の水平方向の結合セル

3.2 台形歪みを加えた表画像の生成

実環境におけるカメラ撮影では，撮影角度や遠近法の影響により，表画像が台形状に歪むことが多い．このような台形歪みは，画像中の水平性や垂直性が保持されていることを前提とする既存の表構造解析モデルにとって障害となり，解析精度の著しい低下を招く要因となる．そこで，実環境下での撮影条件を考慮した頑健な表構造解析の実現を目的として，台形歪みを加えた表画像を人工的に生成し，その補正を試みる．

本研究では，入力された表画像に対して射影変換行列を適用することで，台形歪みを有する表画像を生成する．具体的には，入力画像の四隅を変換前の基準点とし，各頂点を画像の幅および高さに対する一定割合の範囲内でランダムに変位させた座標を変換後の対応点として設定する．これらの対応点に基づいて射影変換行列を算出し，画像全体に適用することで，台形歪みを加えた表画像を生成する．

また，変換後の画像において画素が画像領域外へはみ出すことを防ぐため，変換後の四隅座標の最小値および最大値を用いて出力画像サイズを決定し，座標系を平行移動によって正規化した．この処理により，台形歪みを加えた後も表画像全体が出力画像内に収まるようにしている．以上の処理により，入力された表画像から台形歪みを加えた表画像を生成した．

歪ませた表画像の例を図 2 に示す．図 2（a）から（b）が生成される．

3.3 台形歪みを加えた表画像の補正

本節では，撮影条件や視点の影響により台形歪みが加わった表画像に対し，表構造解析の前処理として，表領域を幾何学的に正規化する補正手法を提案する．台形歪みを加えた表画像の補正処理の概要を図 3 に示す．なお，本節で述べる処理はすべて OpenCV を用いて実装した．

提案手法では，まず入力画像に対して表の罫線を強調するための前処理を行う．具体的には，局所的な輝度分布に基づいて二値化閾値を動的に決定する手法である適応的二値化を適用し，濃淡変化の影響を抑えつつ罫線を抽出する．続いて，膨張および収縮といった形態学的処理を適用し，擦れやノイズによって分断された線分の接続を促進する．次に，前処理後の画像に対して水平線分を検出する．提案手法では，画素単位よりも細かい精度で線分区間を検出できる手法である Line Segment Detector（LSD）を適用し，表罫線を高精度に検出する．しかし，ノイズの影響により，LSD では十分な線分が得られない場合がある．そのような場合には，確率的に線分を探索する手法である確率的 Hough 変換を用い，検出精度の低下を補完することで，線分検出の失敗を抑制する．検出された線分群のうち，画像内でほぼ水平方向に延びている成分のみを選択し，台形歪みに対して安定な水平線分を抽出する．抽出された水平線

(a) 表

野菜の種類	商品情報	
	個数	値段
人参	5	200
ミニトマト	20	400

(b) 表に対応する HTML コード

```
<table><thead>
<tr><td rowspan="2"><b>野菜の種類</b></td>
<td colspan="2"><b>商品情報</b></td></tr>
<tr><td><b>個数</b></td><td><b>値段</b></td></tr>
</thead><tbody>
<tr><td>人参</td><td><i>5</i></td><td>200</td></tr>
<tr><td>ミニトマト</td><td>20</td><td>400</td></tr>
</tbody></table>
```

図 1: 本稿で扱う表とその HTML コードの例

(a) 元の表画像

Parameter	Value
Self-inductance	$L = 6$ [mH]
Rated current	$I = 8$ [A] (AC, 50/60 Hz, sinus wave)
Current density	$J < 3$ [A/mm^2]
Ambient temperature	$\vartheta_a = +40$ [°C]
Self-resonance frequency range	$f_{crit} = 80 \div 140$ [kHz]

(b) 台形歪みを加えた表画像

Parameter	Value
Self-inductance	$L = 6$ [mH]
Rated current	$I = 8$ [A] (AC, 50/60 Hz, sinus wave)
Current density	$J < 3$ [A/mm^2]
Ambient temperature	$\vartheta_a = +40$ [°C]
Self-resonance frequency range	$f_{crit} = 80 \div 140$ [kHz]

図 2: 生成した台形歪みを加えた表画像の例（表画像の出典：PubTabNet）

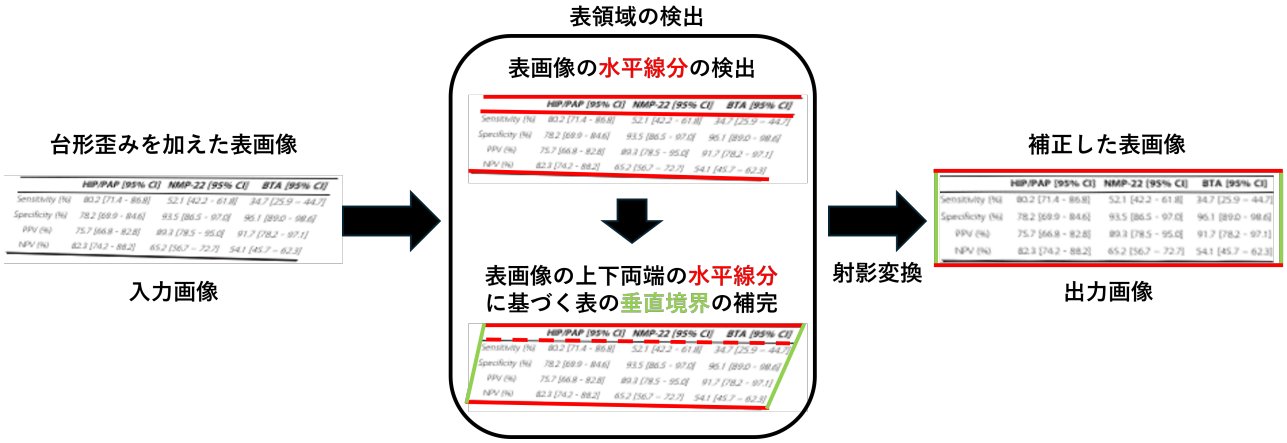


図 3: 台形歪みを加えた表画像の補正手法の概要 (表画像の出典：PubTabNet)

分は，罫線の欠損や点線化の影響により，複数の短い線分として検出される場合がある．そこで，各線分の位置および傾きの類似性に基づいて線分を統合し，同一の線分要素として扱う．続いて，各線分群の端点を点群として扱い，点群の分布特性に基づいて近似線分を推定する手法である主成分分析を適用することで，各クラスタを代表する近似線分を算出する．算出された線分群の中から，画像座標系において最上部および最下部に位置する線分をそれぞれ表領域の上端および下端と定義する．続いて，表の左右端に位置する縦方向の罫線を直接検出するのではなく，特定された上端および下端の水平線分の始点と終点を対応付け，それらを結ぶことで左右の境界を補完的に決定する．最後に，対応する 4 点から射影変換行列を算出する．得られた射影変換行列を用いて画像を幾何変換することで，台形歪みを補正した表画像を生成する．

提案手法は水平線分の検出と統合に重点を置くため，一般的な台形補正とは異なり，表の左右端に位置する縦方向の罫線が存在しない表や，罫線が部分的に欠損している表に対しても，水平方向の情報のみから頑健に表領域を推定することが可能である．

4 LLM を用いた表構造解析結果の修正

4.1 提案する修正手法の概要

本節では，表構造解析モデルが出力した HTML 形式の表構造解析結果に対し，外部の参考情報を活用した LLM による修正手法を提案する．LLM を用いた表構造解析結果の修正の概要を図 4 に示す．本稿では，Ye らによって提案された表構造解析モデルである TableMASTER [4] によって出力された表構造解析結果の HTML コードを LLM への入力として，専門用語集および過去の解析の事例を外部知識として LLM に与えることで，修正した HTML コードを出力する．その後，出力された修正案と入力 HTML コードとの幾何学的整合性や内容の一貫性に基づく比較により，修正の安定性を担保する．この修正の目的は，OCR 由来の軽微な単語誤りの訂正と一部のセルの配置ミスの改善である．

4.2 LLM に与える外部知識

提案手法では，LLM が適切な修正を行うための外部知識として，2 種類の情報を用いる．1 つ目は，PubTabNet および ICDAR2021-SLP [10] のテストデータセットに関連する生理学分野の専門用語集である Medical Subject Headings（MeSH）である．2 つ目は，TableMASTER による PubTabNet の検証データの予測結果とその正解データの HTML コードの組の情報である．

MeSH は，米国国立医学図書館が作成・管理するシソーラスであり，医学用語が階層的に定義されている．PubTabNet および ICDAR2021-SLP [10] のテストデータセットに含まれる表データは医学論文由来であるため，セル内の記述には専門的な薬剤名，疾患名，解剖学用語などが頻出する．MeSH を与えることで，OCR の誤りによる専門用語のスペルミスを LLM が検知し，正しく修正できるようになることを期待する．

TableMASTER による予測結果とその正解データの HTML コードの組は，PubTabNet の検証データセットの表 9,116 件から抽出したものである．この情報を与える目的は，表構造解析モデルが犯しやすい誤りのパターンを LLM に具体例として示すことで，入力された HTML コードに対する適切な修正を促すことである．

4.3 RAG による表構造解析結果の修正手法

4.2 節の外部知識を本稿では Retrieval Augmented Generation（RAG）[20] を利用して LLM に与える．RAG は，外部の文書集合から関連情報を検索し，その結果を LLM の入力として与える枠組みである．RAG 手法は，検索と生成を統合的に学習する end-to-end 型と，検索結果をそのままプロンプトに付与するコンテキスト注入型に大別される．前者は高い性能を示す一方で再学習が必要となる．そこで提案手法では，既存の LLM を変更することなく外部知識を導入できるコンテキスト注入型を採用する．これにより，専門用語や誤り訂正の事例を動的にプロンプトへ組み込むことが可能となる．

修正処理の具体的な手順は以下の通りである．まず，4.2 節で述べた外部知識を事前に整備する．TableMASTER による予測結果と正解データの HTML コードの組については，入力

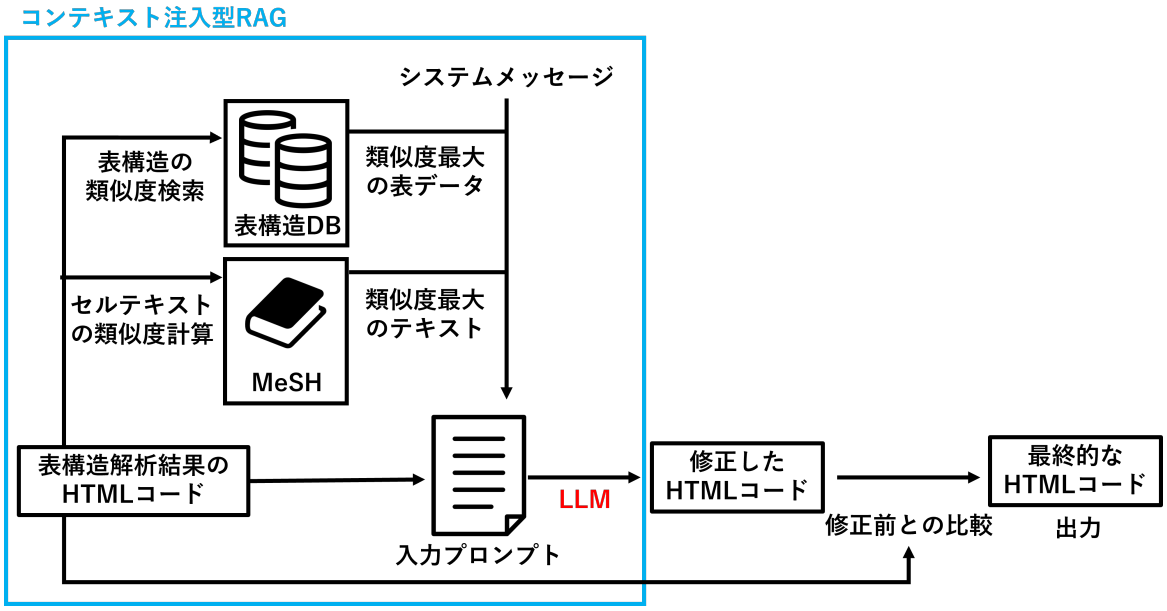


図 4: LLM を用いた表構造解析結果の修正手法の概要

HTML コードと表構造が似ている表を外部知識とするため，セル内のテキストを除去し，表構造を表すタグのみからなる HTML コードを生成する．また，MeSH については，各専門用語を収集する．これらの表構造タグのみの HTML コードおよび MeSH の各専門用語は，Sentence Transformers の文埋め込みモデルを用いてベクトル化され，データベース (DB) に格納される．

次に，修正対象となる表の HTML コードが入力されると，まず入力 HTML コードからセル内テキストを除去し，表構造を表すタグのみからなる HTML コードを生成する．この表構造のみの HTML コードをクエリとして，データベース（DB）に格納された正解の表構造 HTML ベクトルとのコサイン類似度を計算し，最も類似度の高いものを取得する．その結果として，表構造が最も類似する表の HTML コードの予測結果とその正解データの HTML コード，および行単位の類似した表の予測結果と正解の HTML コードの差分を参照情報として抽出し，LLM への入力プロンプトに付与する．一方，MeSH に基づく専門用語の参照情報については，入力 HTML コードから抽出したセルテキストをクエリとして，各専門用語とのコサイン類似度を計算し，類似度が最大となる用語を取得する．取得した用語情報は，OCR 由来の専門用語の誤り訂正を促すための参照情報として，表構造に関する参照情報とともに LLM への入力プロンプトに付与する．

その後，LLM への入力プロンプトを生成する．プロンプトは，LLM の役割と修正条件を定義したシステムメッセージ，外部知識，および修正対象の HTML コードから構成される．そして，LLM は生成されたプロンプトに基づき，外部知識を手掛かりとして入力 HTML コードを解析し，表構造に含まれる誤りを必要に応じて修正することで，構造的整合性が保たれた修正した HTML コードを出力する．

ただし，大規模言語モデル（LLM）による生成結果は常に構造的に正しいとは限らず，過度な修正による事実に基づかない出力や構成要素の不整合が発生する可能性がある． そこで，修正結果と TableMASTER による表構造の解析結果を照らし合わせ，形状の正しさや内容の一致具合に基づいて検証し，最終的な HTML 記述を選択する後処理を行う．具体的には，文字列の類似性を評価する difflib を用いた指標や行数の変動，重要な見出し情報の維持状況を確認し，内容の改ざんを防ぐ安全策を設けている．その上で，各行の横幅のばらつきを数値化した指標や空行の有無を算出し，表としての幾何学的な整合性が向上したかを定量的に評価する．この処理により，有効な修正案がない場合や，格子の整合性が修正前よりも悪化している場合は，改悪を防ぐため修正前の入力を最終結果として採用する．これにより，モデルの不安定さを抑えつつ，構造の修復や文字の微修正など，明確な改善が見込める場合のみ修正を適用することを可能にしている．

5 評 価 実 験

5.1 評価実験の概要

3 節で提案した台形歪みの補正手法および 4 節で提案した LLM による表構造の修正手法の有効性をそれぞれ検証するために，2 つの評価実験を実施する．

1 つは，台形歪みの補正に関する評価実験である．ここでは，まず補正手法による表画像の幾何学的な復元性能を検証するために，画像そのものの品質を評価する．その上で，台形歪みを付加した表画像を入力とした場合の表構造解析精度と，表構造解析の前処理として台形歪みの補正を適用した場合の解析精度を比較することで，表構造解析における補正処理の有効性を評価する．もう 1 つは，LLM による表構造解析結果の修正効果を評価する実験である．ここでは，Ye らによって提案された表構造解析モデル [4] による解析結果の精度と 4 節の LLM に

よる修正処理を加えた場合の解析結果の精度を比較する．

5.2 実験設定

5.2.1 データセットおよび使用モデル

本実験では，表構造解析モデルとして Ye らが公開している TableMASTER [4] の学習済みモデル [1]を使用する． 同モデルは，約 50 万件の学習データから成る PubTabNet [21] を用いて学習されたものである．

評価用データセットとして，ICDAR2021-SLP [10] からランダムに抽出した 1,000 件の表画像と HTML コードのペアを用いる．台形歪みの補正に関する評価実験では，この 1,000 件の表画像に対して擬似的な台形歪みを付与した画像を生成する．そしてそれを入力として用い，提案手法による補正処理が後段の表構造解析精度に与える影響等を評価する．一方，LLM による修正効果の評価実験では，台形歪みを付与していない元の表画像 1,000 件を入力とし，TableMASTER の出力結果に対して LLM による修正処理を適用した場合の解析精度の変化を評価する．

なお，解析結果の修正に用いる LLM には Llama-3.3-70B を採用した． 推論パラメータは，temperature を 0.0，repetition penalty を 1.05，max new tokens を 8,192 に設定した．

5.2.2 評価指標

台形歪みの補正における画像の品質評価では，人間の視覚特性を考慮した画像の構造的類似度指標である Multi-Scale Structural Similarity Index（MS-SSIM）[23] を用いる．MS-SSIM は，ダウンサンプリングによって段階的に解像度を下げた複数のスケール画像を用いて画質を評価する指標である．画像 x と画像 y の間の MS-SSIM は，最大スケール M における輝度比較項 l_M と，各スケール j におけるコントラスト比較項 c_j および構造比較項 s_j を統合して，次式で定義される．

$$\text{MS-SSIM}(x,y) = l_M(x,y)^{\alpha} \prod_{j=1}^{M} c_j(x,y)^{\beta} s_j(x,y)^{\gamma} \qquad (1)$$

ここで，α, β, γ は重みパラメータであり，l, c, s は SSIM において定義される輝度，コントラスト，構造の比較関数である．

また，表構造解析の評価には，表を表す正解の HTML と予測 HTML の木構造としての類似度を測る Tree-Edit-Distance-based Similarity (TEDS) [21] を用いる．TEDS は以下の式で定義される．

$$\text{TEDS}(T_a, T_b) = 1 - \frac{\text{EditDist}(T_a, T_b)}{\max(|T_a|, |T_b|)} \qquad (2)$$

ここで，T_a は正解の表構造，T_b は予測された表構造を表す．$\text{EditDist}(T_a, T_b)$ は T_a と T_b 間の木編集距離であり，$|T_a|$ および $|T_b|$ はそれぞれの木構造におけるノード数を表す．

さらに，本実験では TEDS に加え，セルの内容を無視し，表の構造を表す HTML タグの一致度のみを評価する S-TEDS (Structural-TEDS) [22] も併せて用いる．

1：https://github.com/JiaquanYe/TableMASTER-mmocr

5.3 実験結果

5.3.1 台形歪みの補正に関する実験の結果

まず，提案手法による台形歪みの補正による画像品質の改善効果について述べる．台形歪みを加えた表画像および提案手法により補正した表画像の MS-SSIM による評価結果を表 1 に示す．表 1 より，(a) の台形歪みを加えた表画像と比較して，(b) の補正後の表画像では MS-SSIM の値が 23.76 ポイント向上した．この結果から，提案手法により表画像の幾何学的な歪みが緩和され，視覚的品質が改善されていることが確認できる．

次に，表構造解析精度への影響について述べる．台形歪みの補正についての実験結果を表 2 に示す．表 2 より，(b) の台形歪みを加えた表画像を入力した場合と比較して，(c) の台形歪みを補正した表画像を入力した場合，TEDS は 8.72 ポイント，S-TEDS は 9.89 ポイント改善しており，(a) の歪みのない表画像の TEDS および S-TEDS には及ばないものの，台形歪みによる精度低下を一定程度抑制できており，提案した補正手法が表構造解析精度の改善に有効であることが確認できた．

5.3.2 表構造解析結果の修正に関する実験の結果

LLM を用いた表構造解析結果の修正に関する実験結果を表 3 に示す．

まず，総合評価指標である TEDS に着目すると，基準となる修正なしの (a) が 0.9596 であるのに対し，RAG を使用せず LLM のみで修正を行った (b) および用語集のみを付与した (c) は，共に 0.9598 と微増した．さらに，外部知識として表構造情報を与えた (d) において，精度は 0.9599 となり，僅差ながら全条件の中で最高値を示した．これに対し，用語集と表構造の双方を組み合わせた (e) は 0.9594 に留まり，基準の (a) を下回る結果となった．

次に，表の構造的な正しさを評価する S-TEDS に着目する．(a) の 0.9699 に対し，(b) および (c) が 0.9708 と最も高い値を示し，次いで (d) が 0.9705 となった．(e) を除くすべての修正手法において，S-TEDS は基準を上回っており，LLM を用いた事後修正が構造的な誤りの訂正に寄与していることが確認できる．

S-TEDS では (b) や (c) が (d) を僅かに上回るものの，総合指標である TEDS では (d) が最高値を示した．これは，(d) が構造とテキスト内容の整合性を最も高い水準で両立できたためと考えられる．

5.4 考　　察

5.4.1 台形歪みを加えた表画像の補正に関する分析

台形歪みを加えた表画像の補正に関して，補正が失敗した事例を図 5 に示す．

図 5（c）に示すように，表の下部で検出された横方向の線分が，画像中に存在する物理的な罫線の長さを超えて画像の右端まで過剰に延長されている．この誤った線分検出により，表領域下端右側の境界座標が本来の位置よりも外側に推定された．その結果，消失点の推定および射影変換行列の算出に誤差が生じ，図 5（d）に示すような不適切な補正結果が得られた．

この過剰な線分延長の要因を分析する．対象画像の最下行に

表 1: 表画像の品質評価結果

手法	評価指標
	MS-SSIM
(a) 台形歪みを加えた表画像	0.3272
(b) 台形歪みを補正した表画像	**0.5648**

表 2: 表画像の表構造解析精度

手法	評価指標	
	TEDS	S-TEDS
(a) 元の表画像	0.9596	0.9699
(b) 台形歪みを加えた表画像	0.6333	0.7716
(c) 台形歪みを補正した表画像	**0.7205**	**0.8705**

表 3: LLM を用いて修正した表構造解析精度

手法	RAG の有無	評価指標	
		TEDS	S-TEDS
(a) 修正なし	RAG なし	0.9596	0.9699
(b) システムメッセージのみ		0.9598	**0.9708**
(c) 用語集のみ	RAG あり	0.9598	**0.9708**
(d) 表構造のみ		**0.9599**	0.9705
(e) 用語集+表構造		0.9594	0.9698

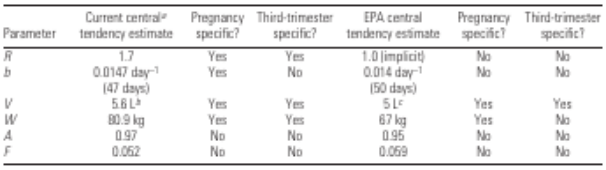
おいて，物理的な罫線は図 5 (b)に示すように表の右端付近で終端しているが，その右側には空白領域が存在する．この空白領域には，画像圧縮や輝度勾配などに起因する微細なノイズ成分が含まれており，これらがエッジ特徴として検出されていた．提案手法は，近接かつ近似した角度を持つ線分群を単一のクラスタに統合し，その両端点を結ぶ一本の線分として復元を行う．本事例では，座標系の歪みにより右端のノイズと物理的な罫線が幾何学的許容誤差内で同一線上に配置されたため，両者が誤統合された．その結果，本来の終端が無視され，画像全幅を貫通する線分が生成されたことが補正失敗の要因である．

5.4.2 LLM による表構造解析結果の修正に関する分析

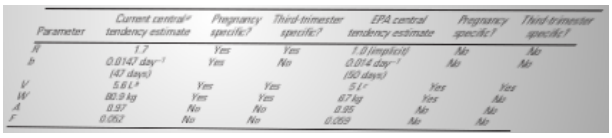
LLM を用いた修正処理が，表構造解析の評価指標である TEDS，および構造の正確性を測る S-TEDS に与える影響について，スコアが向上した事例および低下した事例を分析する．

まず，LLM を用いた修正処理により TEDS および S-TEDS が向上した事例を図 6 に示す．図 6 の事例では，修正後に TEDS スコアが 0.9154 から 0.9699 へと改善した．図 6(a) の表には，学歴区分を示す行として「<Technical」という文字列が含まれている．しかし，図 6(b) の TableMASTER による出力では，セル内の記号「<」が HTML タグの開始文字として誤って解釈され，構文エラーによって当該セル自体が構造から欠落する結果となっていた．これに対し提案手法では，LLM がセル内の文字列を文脈として再解釈した．その結果，図 6(c) の赤枠に示すように，「Technical」が教育水準を表す単一の項目であることを認識し，記号を除去した適切な文字列として再構成した．この修正により，セル内の文字列の一致度が高まったことで最終的な TEDS が向上した．また，構文エラーにより消失していたセルが表の木構造上に正しく復元されたため，構造の整合

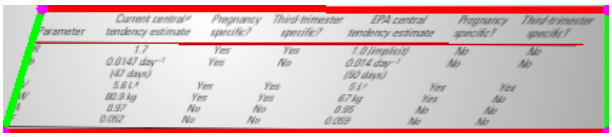
(a) 元の表画像



(b) 台形歪みを加えた表画像



(c) (b) の表画像に検出した表領域を重ねた表画像



(d) 台形歪みを補正した表画像

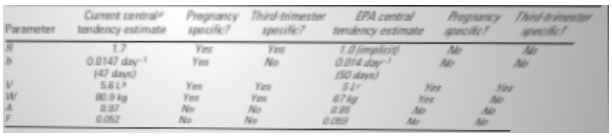


図 5: 台形歪みを加えた表画像の補正の失敗例（表画像の出典：ICDAR2021-SLP テストデータ）

性を示す S-TEDS の値も改善する結果となった．

一方，LLM を用いた修正により TEDS の値が低下した事例を図 7 に示す．図 7 の事例では，修正適用後に TEDS の値が 0.9891 から 0.8948 へと低下した．図 7(a) の元画像および図 7(b) の修正前の出力には，統計的な欠損値を表す「-」や，「$8.50E^{-01}$」のような指数表記が含まれている．提案手法において LLM は，セルの内容を一般的な数値形式へと適合させようとする過剰な正規化を行った．その結果，図 7(c) の青枠に示すように，LLM は欠損値を表す「-」を数学的な「0」へと置換し，さらに緑枠に示すように，指数表記の上付き文字タグ（<sup>）を削除した文字列へと改変した．本事例において，上付き文字タグの削除は表の行・列構成といった基本的な格子構造には影響しないため，S-TEDS の値は維持された．しかし，TEDS は正解データとの文字レベルでの厳密な一致度を評価する指標であるため，このような LLM による独断的な事実に基づかない出力や書式タグの欠落は，正解への忠実度の欠如とみなされ，TEDS スコアの大幅な低下を招いた．

以上の分析より，表構造解析における LLM を用いた修正処理に関して以下の知見が得られる．第一に，LLM は意味的文脈に基づく推論によって視覚的に曖昧な箇所のノイズを修復し，不適切なタグ解釈によるセルの欠落を防いで S-TEDS を向上させるなど，構造的な整合性を自律的に回復させる能力を有する．これは，局所的な画素情報に依存する従来の画像認識モデルの限界を，LLM の知識が補完する上で有効であることを示している．第二に，LLM の強力な推論能力は，専門的な記法

(a) 元の表画像

Social variable	Nonpregnant women	Pregnant women
Age (years)	29.250 ± 2.314	28.333 ± 1.971
Ethnicity	Han	Han
Occupation		
Housewife	0	3
Employee	20	27
Level of education		
<Technical	6	10
Bachelor	9	13
Master	5	7
Economic status	Regular	Regular

(b) 修正前の HTML コードから作成した表

Social variable	Nonpregnant women	Pregnant women
Age (years)	29.250 ± 2.314	28.333 ± 1.971
Ethnicity	Han 0.001	Han
Occupation		
Housewife	0	3
Employee	20	27
Level of education		
	6	10
Bachelor	9	13
Master	5	7
Economic status	Regular	Regular

(c) 修正後の HTML コードから作成した表

Social variable	Nonpregnant women	Pregnant women
Age (years)	29.250 ± 2.314	28.333 ± 1.971
Ethnicity	Han 0.001	Han
Occupation		
Housewife	0	3
Employee	20	27
Level of education		
Technical	6	10
Bachelor	9	13
Master	5	7
Economic status	Regular	Regular

図 6: LLM による表画像の修正で精度が向上した例（表画像の出典：ICDAR2021-SLP テストデータ）

を強制的に適合させようとする過剰な正規化のリスクを孕んでいる．今回の事例のように，S-TEDS が示す構造的な正しさは維持しつつも，文字情報の書き換えや書式タグの消去によって TEDS を低下させてしまうケースが確認された．したがって，実用化にあたっては，画像の忠実性を維持しつつ内容の改ざんを抑止するための安全策や制約条件の設計が不可欠であることを示唆している．

(a) 元の表画像（一部抜粋）

6	-	$8.50E^{-01}$
13	-	$6.80E^{-01}$
11	-	$5.90E^{-01}$
5	-	$5.40E^{-01}$

(b) 修正前の HTML コードから作成した表（一部抜粋）

6	-	$8.50E^{-01}$
13	-	$6.80E^{-01}$
11	-	$5.90E^{-01}$
5	-	$5.40E^{-01}$

(c) 修正後の HTML コードから作成した表（一部抜粋）

6	0	8.50E−01
13	0	6.80E−01
11	0	5.90E−01
5	0	5.40E−01

図 7: LLM による表画像の修正で精度が低下した例（表画像の出典：ICDAR2021-SLP テストデータ）

6 おわりに

本稿では，表画像に対する台形歪みの補正と大規模言語モデルを用いた表構造解析結果の修正を提案した．

台形歪みの補正では，人工的に台形歪みを加えた表画像データセットを構築した．その後，OpenCV を用いて生成した台形歪みを加えた表画像を補正することで，表構造解析モデルへの入力画像の品質および表構造解析精度の向上を実現した．LLM を用いた表構造解析結果の修正では，表構造解析モデルの出力として得られた表の HTML コードを専門用語集および表構造解析事例と併せて LLM に与えることで，セル中のテキストの誤りやタグの欠落や不整合などを修正した．

評価実験の結果，台形歪みの補正を適用することで，画像品質を示す MS-SSIM は 23.76 ポイント改善し 0.5648 へ，表構造解析精度を示す TEDS は 8.72 ポイント改善し 0.7205 となった．また，LLM を用いた解析結果の修正においては，外部知識として表構造情報を活用した際に TEDS が 0.03 ポイント向上し 0.9599 へ，S-TEDS が 0.06 ポイント向上し 0.9705 となった．

今後の課題としては，歪んだ表画像の補正において，台形歪みだけでなく，紙面の反りや折り目に起因する湾曲歪みや折れ歪みなどの非線形な歪みに対応することが挙げられる．また，表構造解析結果の修正に関しては，本稿では LLM として Llama-3.3-70B を用いたが，他の LLM を用いた場合の検証や，他の表構造解析モデルに対しても提案手法が有効であるかの検証などが挙げられる．さらに，台形歪みの補正時の画質劣化に起因する文字認識精度の低下を，LLM を用いた修正によって回復できるか検証することによる両手法の相乗効果の確認が挙げられる．

謝　辞

本研究の一部は，科学研究費補助金基盤研究 (B)(課題番号 23K25158) および 2025 年度国立情報学研究所公募型共同研究 (252FC-23662) の援助による．

文　献

[1] 納田 朋享, 金沢 輝一, 上野 史, 太田 学, “表画像を入力とする表構造解析手法の分析と改良,” 第 17 回データ工学と情報マネジメントに関するフォーラム (DEIM 2025), 4G-03, 2025.

[2] 田上 歩夢, 金澤 輝一, 上野 史, 太田 学, “表構造情報を利用した棒グラフの自動生成の一手法,” 第 16 回データ工学と情報マネジメントに関するフォーラム (DEIM 2024), T4-A-3-02, 2024.

[3] Hiroyuki Shindo, Yuji Matsumoto, Masashi Ishii, Hiroyuki Oka, Atsushi Yoshizawa, “Machine extraction of polymer data from tables using XML versions of scientific articles,” *Science and Technology of Advanced Materials: Methods*, Volume 1, pp. 11–23, 2021.

[4] Jiaquan Ye, Xianbiao Qi, Yelin He, Yihao Chen, Dengyi Gu, Peng Gao, Rong Xiao, “PingAn-VCGroup’s Solution for ICDAR 2021 Competition on Scientific Literature Parsing Task B: Table Recognition to HTML,” arXiv preprint arXiv:2105.01848, 2021.

[5] Nam Tuan Ly, Atsuhiro Takasu, “An End-to-End Multi-Task Learning Model for Image-based Table Recognition,” *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP*, pp. 626–634, 2023.

[6] Rujiao Long, Wen Wang, Nan Xue, Feiyu Gao, Zhibo Yang, Yongpan Wang, Gui-Song Xia, “Parsing Table Structures in the Wild,” *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Volume 1, pp. 924–932, 2021.

[7] Gavin Greif, Niclas Griesshaber, Robin Greif, “Multimodal LLMs for OCR, OCR Post-Correction, and Named Entity Recognition in Historical Documents,” arXiv preprint arXiv:2504.00414, 2025.

[8] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, Enamul Hoque, “ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning,” *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2277, 2022.

[9] Minesh Mathew, Dimosthenis Karatzas, C.V. Jawahar, “DocVQA: A Dataset for VQA on Document Images,” 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 2199–2208, 2021.

[10] Antonio Jimeno Yepes, Peter Zhong, Douglas Burdick, “ICDAR 2021 Competition on Scientific Literature Parsing,” *Proceedings of 16th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 605–617, 2021.

[11] Ning Lu, Wenwen Yu, Xianbiao Qi, Yihao Chen, Ping Gong, Rong Xiao, Xiang Bai, “MASTER: Multi-Aspect Non-local Network for Scene Text Recognition,” *Pattern Recognition*, Volume 117, Article number 107980, 2021.

[12] Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, Shuai Shao, “Shape robust text detection with progressive scale expansion network,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9336–9345, 2019.

[13] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, Sergey Zagoruyko, “End-to-End Object Detection with Transformers,” *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 213–229, 2020.

[14] Alex Smock, Rohith Anil, Mark Hasegawa-Johnson, Matthew A. Gardner, “PubTables-1M: Towards Comprehensive Table Extraction from Unstructured Documents,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4634–4642, 2022.

[15] Hmrishav Bandyopadhyay, Tanmoy Dasgupta, Nibaran Das, Mita Nasipuri, “A Gated and Bifurcated Stacked U-Net Module for Document Image Dewarping,” 2020 25th International Conference on Pattern Recognition (ICPR), pp. 10548–10554, 2021.

[16] Ke Ma, Zhixin Shu, Xue Bai, Jue Wang, Dimitris Samaras, “DocUNet: Document Image Unwarping via a Stacked U-Net,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4709, 2018.

[17] Ziyi Zhu, Zhi Tang, Liangcai Gao, “Table image dewarping with key element segmentation,” *International Journal on Document Analysis and Recognition (IJDAR)*, Volume 27, pp. 349–362, 2024.

[18] Yi Ren, Chenglong Yu, Weibin Li, Wei Li, Zixuan Zhu, Tianyi Zhang, Chenhao Qin, Wenbo Ji, Jianjun Zhang, “TableGPT: a novel table understanding method based on table recognition and large language model collaborative enhancement,” *Applied Intelligence*, Volume 55, Article number 311, 2025.

[19] Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun, “TableLlama: Towards Open Large Generalist Models for Tables,” *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 4361–4378, 2024.

[20] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wentau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 9459–9474, 2020.

[21] Xu Zhong, Elaheh ShafieiBavani, Antonio Jimeno Yepes, “Image-based table recognition: data, model, and evaluation,” *Computer Vision – ECCV 2020*, pp. 564–580, 2020.

[22] Yongshuai Huang, Ning Lu, Dapeng Chen, Yibo Li, Zecheng Xie, Shenggao Zhu, Liangcai Gao, Wei Peng, “Improving Table Structure Recognition with Visual-Alignment Sequential Coordinate Modeling,” 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11134–11143, 2023.

[23] Zhou Wang, Eero P. Simoncelli, Alan C. Bovik, “Multiscale structural similarity for image quality assessment,” *Proceedings of the 37th Asilomar Conference on Signals, Systems and Computers*, Volume 2, pp. 1398–1402, 2003.

日本語文書画像質問応答における参照構造分解と回答不能判定の分析

山野　瑞月† 宮森　恒†

† 京都産業大学先端情報学研究科　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{i2586263,miya}@cc.kyoto-su.ac.jp

あらまし　本稿では，日本語文書画像に対する質問応答 (JDocQA) において生じる誤答および回答不能判定の失敗要因を，文書内参照構造の分解という観点から体系的に分析する．日本語文書画像は，縦書きと横書きの混在，離れた箇所の参照などの非線形な読み順などの複雑なレイアウト構造を有しており，文書全体を単一の内部表現に内在化して推論を行う既存手法では，情報間の参照関係が曖昧化しやすい．本稿では，文書画像内のマルチモーダル要素をノードとする参照構造を明示的に構築し，質問応答の失敗を認識誤り，参照誤り，推論誤りに分解して評価可能な分析の枠組みを提案する．本枠組みでは，参照構造を外部から柔軟に参照可能な RAG-Anything を分析基盤として用い，モデルが実際に参照した情報を追跡可能とする．JDocQA データセットを用いた実験では，質問タイプごとに性能を比較し，さらに回答不能質問において，参照構造を外部化することで根拠のない推測回答が抑制されるかを検証する．これにより，日本語文書画像質問応答における回答不能判定の困難さが，参照構造の破綻と強く関係していることを明らかにする．

キーワード　文書画像質問応答，JDocQA，RAG-Anything，RAG，日本語文書画像

1 はじめに

与えられた自然言語入力に基づいて推論を行い，適切な回答を返す技術の実現は，知能情報処理分野における重要な課題の一つである．その中でも，文書画像質問応答 (Document Visual Question Answering; DocumentVQA) は，文書画像に含まれるテキストおよび図表やレイアウトなどの視覚的要素を理解し，質問に答えるタスクであり，近年注目を集めている [7], [11], [12]．代表的なデータセットとしては，手紙やレポート請求書などのスキャンされた紙の文書に対して質問応答を行う DocVQA [8] や，インターネット上から収集された地図やグラフなどの視覚的要素が含まれている文書画像に対して質問応答を行う InfographicVQA [7]，レイアウトやセグメント，図表が含まれているプレゼンテーションスライドに対して質問応答を行う SlideVQA [11] などがある．しかし，これらの既存研究は主に英語文書を対象としており，日本語文書画像を対象とする研究は限定的である．さらに，表やグラフなどの視覚的要素や文書の構造を適切に理解することや複数の回答根拠を参照し段階的に推論を行うマルチホップ質問に対応することは依然として困難である [8], [11]．

日本語の文書画像に対する質問応答 (JDocQA) は，グラフや表，地図，日本語の縦書き・横書き混在等の視覚的情報とテキスト情報を参照し，回答を行うタスクである [10]．JDocQA で対象としている文書画像は，レポート，スライド，パンフレット等の多様な形式であり，経済や教育，農業など様々な分野に関する質問が含まれている．日本語文書画像は，縦書き・横書きの混在や読み順の非線形性，多様な文字体系 (ひらがな，カタカナ，漢字) など，英語文書画像とは異なる特性を持つ．そのため，既存の英語文書画像質問応答モデルをそのまま適用することは困難である．

JDocQA タスクに取り組んだ従来手法では，最も良いモデルでテストデータ全体に対する評価において 29.71 %のスコアを達成している [10]．この結果は，JDocQA タスクの難易度の高さを示しており，さらなる性能向上の余地があることを示唆している．また，文書中の情報だけでは回答できない質問 (回答不可能質問と呼ぶ．) に対して，根拠のない推測回答を行ってしまう問題や [10]，文書内の図やチャート，表などにテキストが複雑に配置されていることによる情報参照の困難さなどの課題がある．しかし，JDocQA タスクに取り組んだ従来手法では，誤答や回答不可能質問に対する回答不能判定の失敗が，認識・参照・推論のいずれの段階で生じているのかの体系的な分析が十分になされていない．

そこで本稿では，日本語文書画像質問応答における誤り要因を，文書内参照構造の観点から分解・整理することを目的とする．具体的には，文書画像内のマルチモーダル要素をノードとする参照構造を明示的に構築し，それを外部参照として利用可能な枠組みを用いることで，質問応答過程における参照の成否を追跡可能とする．これにより，モデルの誤答要因を明らかにする．

そのために，文書画像内のマルチモーダル要素をノードとする参照構造を明示的に構築可能な RAG-Anything [3] を基盤とし，日本語文書画像の構造を理解できるよう拡張をする．RAG-Anything は，外部知識を柔軟に参照可能な RAG(Retrieval-Augmented Generation) の枠組みを拡張したものであり，マルチモーダルな情報をノードとして扱い，質問応答に必要な情報を動的に参照可能である．しかし，RAG-Anything は英語文書画像を主な対象としており，日本語文書画像の複雑なレイアウト構造や縦文字を理解する能力には限界がある．よって，従来の RAG-Anything で使用されていた構造パーサーである

MinerU2.5 [9] を日本語文書画像も理解可能にする．そして，日本語文書画像質問応答における誤り要因を文書内参照構造の観点から分解・整理することで，認識・参照・推論のどの段階に課題があるのかを明らかにする．

実験では，RAG-Anything を日本語文書画像を理解できるよう拡張した提案手法の有効性を検証する．また，日本語文書画像質問応答において生じる誤答および回答不能判定の誤り要因を，文書内参照構造の観点から分析することを目的とする．JDocQA データセット [10] を用いて，質問タイプ別に性能を比較し，さらに回答不可能質問において，参照構造を外部化することで根拠のない推測回答が抑制されるかを検証する．

本稿の主な貢献は以下の通りである．

- 日本語文書画像を構成するマルチモーダル要素間の参照関係に着目し，質問応答の失敗を認識誤り，参照誤り，推論誤りに分解して評価可能な分析の枠組みを提示する．
- 参照構造を外部化して追跡可能な RAG ベースの枠組みを分析基盤として用いることで，モデルが実際に参照した情報を可視化し，質問タイプごとに性能差を分析する．
- 回答不能質問に着目し，その誤答要因が，参照構造の破綻とどのように関係しているかを実験的に検証する．これらの分析を通じて，本研究は将来の日本語文書画像質問応答モデル設計に対する指針を提供する．

2 関 連 研 究

2.1 文書画像質問応答のデータセット

文書画像質問応答は，文書画像に含まれるテキストおよびレイアウト，表，図などの視覚的要素を理解し，質問に答えるタスクであり，様々なデータセットが提案されている．例えば，DocVQA [8] では，手紙やフォーム，請求書などのスキャンされた紙の文書が主に含まれており，文書構造，表などのレイアウトや，図やチェックボックスなどの非テキスト要素などを理解する必要がある．InfographicVQA [7] は，インターネット上から収集された多様なインフォグラフィック文書を対象としている．テキストだけでなく，地図やグラフなどの視覚的要素も含まれており，DocVQA と異なり，デジタル形式の文書が多く含まれている．SlideVQA [11] は，プレゼンテーションスライドを対象としている．スライド内には，表やグラフ，レイアウトやセグメント情報が豊富に含まれており，これらの要素を適切に組み合わせて理解することが求められる．WebSRC [1] は Web ページを対象とした構造読解のために構築されている．HTML ソースコードやスクリーンショット画像などの情報が含まれており，Web ページの視覚的に組織化された空間構造や意味的に組織化された論理構造を理解することが求められる．本稿で扱う JDocQA [10] は，日本語文書画像を対象ととしている．グラフや表，地図，日本語の縦書き・横書き混在等の視覚的情報とテキスト情報の両方を参照し，推論し回答する必要がある．JDocQA で対象としている文書画像は，レポート，スライド，パンフレット等の多様な形式である，本稿では，JDocQA データセットを用いて，日本語文書画像質問応答における誤答および回答不能判定の失敗要因を，文書内参照構造の分解という観点から体系的に分析する．

2.2 文書画像理解・質問応答モデル

文書画像質問応答に取り組んだ従来手法として，OCR によって抽出されたテキスト情報に，そのテキストが文書内のどの位置にあるかという空間情報を付加して学習するレイアウト認識型モデルがある．例えば，LayoutLMv3 [4] は文書理解のための多モーダル事前学習モデルであり，テキスト，レイアウト，画像の 3 つのモーダル情報を統合的に処理する．しかし，このようなモデルでは，OCR の品質に依存するため，OCR の誤りが後続のタスクに悪影響を及ぼす可能性がある．

一方，OCR を使用せず，画像から直接，構造化データを生成しようとする試みもある．代表されるモデルとして，Donut [5] がある．Donut は，文字認識，レイアウト解析，文書理解を 1 つのモデルで完結させる end-to-end の OCR-free 文書理解モデルである．しかし，Donut のようなモデルは，画像全体を低解像度のトークンとして圧縮し処理を行うため，非常に細かい文字や長文文書の処理において性能が低下する可能性がある．

近年，文書画像の対象として，紙の書類をスキャンした画像だけでなく，ウェブサイトや図表，チャートなどのスクリーンショット画像を対象とする研究も増えてきている．Pix2Struct [6] は，ウェブの膨大な構造化データを利用して，視覚的に配置された言語を理解するための事前学習法を提案している．また，DocPedia [2] は，従来のモデルを凌駕する高解像度画像の入力を可能としたモデルであり，OCR-free 大型文書理解 LMM として注目されている．

しかし，これらの従来手法は，文書全体を単一の内部表現に内在化して推論を行うため，情報間の参照関係が曖昧化しやすい．本稿では，文書画像内のマルチモーダル要素をノードとする参照構造を明示的に構築し，質問応答の課題を認識誤り，参照誤り，推論誤りに分解して評価可能な分析の枠組みを提案する．

2.3 構造・参照を明示化するタスク・枠組み

文書画像質問応答における構造・参照を明示化する試みとして，WebSRC [1] がある．WebSRC は，Web ページのテキスト内容だけでなく，その視覚的・論理的構造を明示的に扱うタスクとして，構造的読解 (SRC) を提案している．構築されたデータセットは，HTML や視覚的な空間構造であるスクリーンショットおよび各要素のメタデータをモデルの入力および参照情報として提供する．この枠組みでは，単なるテキストの羅列ではなく，HTML タグなどの包含関係や要素間の配置などといった構造を参照することで，表形式などの複雑なレイアウトから正確な情報抽出が可能となる．しかし，WebSRC は主に英語 Web ページを対象としており，日本語文書特有の縦書き・横書きの混在や、罫線が多用される複雑なレイアウトへの適用については十分に検証されていない．

2.4 マルチモーダル RAG・外部知識化

RAG(Retrieval-Augmented Generation) は，外部知識を動

的に参照しながらテキスト生成を行う枠組みである．RAG-Anything [3] は，RAG をテキストに限らず，画像・表・数式など多様なモダリティを統一的に扱えるマルチモーダル知識検索フレームワークである．RAG-Anything では，クロスモーダル関係と意味表現を統合するデュアルグラフ構造や構造的ナビゲーションと意味的マッチングを組み合わせたハイブリット検索を導入し，多様な形式の情報に跨る検索・推論が可能である．しかし，RAG-Anything は英語文書画像を主な対象としており，日本語文書画像の複雑なレイアウト構造や縦文字を理解する能力には課題があるとされている．本稿では，RAG-Anything を日本語文書画像を理解できるよう拡張し，日本語文書画像質問応答における誤り要因を文書内参照構造の観点から分解・整理することで，認識・参照・推論のどの段階に課題があるのかを明らかにする．

3 参照構造分解に基づく分析フレームワーク

3.1 問題設定

本稿で扱う JDocQA タスクは，文書 D とそれに対する質問 Q が与えられたとき，以下のように定義する．

$$f(D,Q)=\begin{cases} A & \text{（文書内に十分な情報がある場合）} \\ \text{“回答不能”} & \text{（文書内に十分な情報がない場合）} \end{cases} \tag{1}$$

ここで，A は文書内の情報に基づいて生成される回答を表す．JDocQA では，文書内の情報だけでは回答できない回答不可能質問が明示的に含まれており，モデルは根拠のない推測回答を避け，適切に回答不能判定を行うことが求められる．

従来の JDocQA タスクでは，文書全体を単一の内部表現として処理するモデルが主流であったが，日本語文書画像における誤答や誤った回答不能判定の多くは，文書内の参照構造が適切に保持・利用されないことに起因すると考えられる．特に，図表や地図などの複数の視覚的要素を横断して参照する必要がある質問や，縦書き・横書き混在により読み順が非線形となる文書において，情報参照がより困難となる．そこで本稿では，質問応答の失敗を認識誤り，参照誤り，推論誤りに分解し，参照構造の観点から体系的に分析する枠組みを提案する．

3.2 参照構造の定義

本稿では，文書画像 D 内に含まれる複数の情報要素間の参照関係を，有向ラベル付きグラフ $G=(V,E)$ として表現する．ここで，V はノード集合，E はエッジ集合を表す．

a） ノード定義

各ノード $v\in V$ は，文書中の意味的に一貫した要素を表す．ノードの属性として共通して持つものとして，ページ番号 $page(v)$，位置情報 $bbox(v)$，読み順 $order(v)$ を定義する．さらに，type ごとに異なる属性を持つ．テキストタイプのノードの場合，タイプ $type(v)$，テキスト内容 $content(v)$，テキストの方向性 $direction(v)$ を持つ．画像・図タイプのノードの場合，タイプ $type(v)$，画像のファイルパス $filepath(v)$，画像内のテキスト内容 $content(v)$ を持つ．テーブルタイプのノードの場合，タイプ $type(v)$，HTML テーブル表現 $html(v)$ を持つ．

b） エッジ定義

ノード間の参照関係は，エッジ $e=(v_i,v_j,r)\in E$ により表現され，r は関係の種別を示すラベルである．本稿では以下の関係タイプを定義する：

- **所属関係**: マルチモーダルコンテンツ（画像，表）から抽出されたエンティティと，そのコンテンツノード自体との所属関係を表す．
- **意味的関係**: LLM ベースのエンティティ抽出により，テキストおよびマルチモーダルコンテンツの記述から自動的に抽出される汎用的なエンティティ間関係である．

3.3 誤り要因の分解

本稿では，提案手法および比較手法における失敗事例を分析するため，誤り要因を以下の 3 種類に分類する．

a） 認識誤り

認識誤りは，文書画像から構造化表現を生成する段階で生じる誤りを指す．具体的には，OCR による文字認識誤りや，レイアウト解析の誤り，および図・表などの要素種別 (タイプ) の誤認識が含まれる．認識誤りは，後続の参照および推論段階に直接影響を与える．

b） 参照誤り

参照誤りとは，文書内に質問に対する正しい情報を含むノードが存在するにもかかわらず，検索または参照構造の探索に失敗することにより，適切なノードまたは参照経路を取得できない場合を指す．具体的には，必要なノードが検索対象から漏れる，誤ったノードが参照される，または参照経路が途中で断絶する場合などが含まれる．

c） 推論誤り

推論誤りは，質問に必要な参照ノードおよび参照経路が正しく取得されているにもかかわらず，それらの情報を統合した推論に失敗する場合を指す．これは主に LLM や VLM の推論能力に起因する誤りであり，数値推論や因果関係の理解において課題が生じることがある．

3.4 外部参照を用いた分析設定

参照誤りを分析するには，質問応答の過程においてモデルが実際にどの情報を参照したかを観測可能である必要がある．しかし，文書画像全体を内部表現として一括して処理する手法では，推論過程における参照経路や参照対象を明示的に追跡することが困難である．

そこで本稿では，参照構造を外部化することにより，質問応答過程における参照を観測可能にするための分析設定を採用する．具体的には，文書画像から構築された参照構造を外部参照対象として保持し，質問に応じて関連するノードを取得・参照する枠組みを用いる．この枠組みは，RAG-Anything [3] を基盤とし，日本語文書画像を理解可能に拡張したものである．

このような外部参照機構を用いることで，質問応答において参照されたノード集合や利用された参照経路を明示的に記録することが可能となる．本稿における外部参照機構は，質問応答

性能の向上を主目的とするのではなく，参照構造の成否を分析するための手段として位置づける．

4 データセット

本実験では，JDocQA [10] をデータセットとして用いる．データセットの例を図 1 に示す．JDocQA は，5,508 件の PDF ファイルに対応する 11,600 件の質問応答ペアから構成されるデータセットである．これらの PDF ファイルは，日本の官公庁が公開しているスライドやレポート，地方自治体のパンフレットなどが含まれており，収集した文書は，レポート，スライド，パンフレット，ウェブサイトに分類されている．

質問の種類として，“はい/いいえ” で回答できる “はい/いいえ形式”，本文中から事実を抜き出して回答する “事実抽出形式”，簡単な四則演算を行い回答する “数量形式”，質問に対して文章を作成して回答する “自由記述形式” の 4 つの形式が含まれている．また，JDocQA データセットには，4 つの質問形式と独立に複数ページを参照しなければ回答できないマルチホップ質問や，文中の情報だけでは回答できない質問 (回答不可能質問) がある．

本稿の実験では，JDocQA データセットのテストデータに含まれている 400 件を用いて，日本語文書画像質問応答における誤答および回答不能判定の誤り要因を，文書内参照構造の観点から分析する．

質問: 小瀬鵜飼の観覧には遊船料がかかります。「親子ふれあい鵜飼」に参加すると親子 2 人（子どもは小学生以下）で 4,200 円ですが、通常の親子 2 人（子どもは小学生以下）の乗合料金より何円安くなりますか？**回答**: 1,950 円

図 1 JDocQA データセットの例

5 実　　験

5.1 実験目的

本実験の目的は，日本語文書画像質問応答において生じる誤答および回答不能判定の誤り要因を，定量的・定性的に分析することである．そして，日本語文書画像質問応答における誤りの要因を明らかにすることである．

5.2 実験設定

本節では，定量評価で用いるモデルおよび比較手法について述べる．

比較手法として，生成ベースのモデルと表現ベースのモデルを用いる．まず，JDocQA タスクに取り組んだ従来の生成ベースのモデルとして gpt-3.5-turbo-16k [10]，gpt-4 [10]，および StableLM Base-Al.-7B [10] を用いる．また，表現ベースのモデルとして Qwen/Qwen3-VL-8B-Instruct[1] を用いる．この際，文書として数百ページに及ぶ長文文書を扱うため，文書から回答根拠のある部分をあらかじめ抽出し，その部分のみを入力する．抽出には，vidore/colqwen2-v1.0-hf[2] を用いる．

提案手法として，RAG-Anything [3] を基盤とし，日本語文書画像を理解可能に拡張したモデルを用いる．RAG-Anything では，文書の構造をパースするために MinerU2.5 [9] を用いていたが，日本語文書画像も理解可能にするため，MinerU2.5 の代わりに YomiToku[3] を用いる．YomiToku は，日本語に特化した AI 文章画像解析エンジン (Document AI) であり，文書画像内の文字の認識やレイアウト解析機能を備えている．

5.3 評価方法

5.3.1 定量評価

評価指標としては，JDocQA タスクに取り組んだ従来手法に倣い，EM スコアと BLEU スコアを用いる．“自由記述形式” の質問に対しては BLEU スコアを，その他の 3 つの質問形式に対しては EM スコアを用いる．実験では，4 種類の形式の質問タイプ別に性能を比較し，さらに回答不可能質問において，根拠のない推測回答がどの程度発生しているかを分析する．

5.3.2 定性評価と誤り要因の分析

提案する参照構造分解に基づく分析フレームワークを用いて，誤答および回答不能判定の失敗要因を，認識誤り，参照誤り，推論誤りに分解して分析する．具体的には，各推論結果が正しいかを確認し，誤答の場合は 3.3 節で述べた，認識誤り・参照誤り・推論誤りのいずれで誤りが生じたのかを特定する．本分析は定性的な誤り分析を目的としており，著者 1 名が評価

1：https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct
2：https://huggingface.co/vidore/colqwen2-v1.0-hf
3：https://github.com/kotaro-kinoshita/yomitoku

を行った．

5.4 実験結果

5.4.1 定量評価

表 1 質問タイプ別の定量評価の結果

モデル	Avg	Y/N	Fact	Num	Free
学習なし (ゼロショット)					
gpt-3.5-turbo-16k [10]	20.62	50.29	7.44	11.11	13.64
gpt-4 [10]	19.47	43.19	6.51	11.11	17.07
学習セットで学習済み					
StableLM Base-Al.-7B [10]	**29.71**	**70.41**	15.81	**22.22**	**25.51**
実験手法 (学習なし)					
Qwen3-VL-8B-Instruct	18.02	59.76	2.88	0.53	6.42
RAG-Anything	21.52	58.21	**44.00**	6.67	5.09

表 2 回答不可能質問に対する定量評価の結果

モデル	平均	件数
Qwen3-VL-8B-Instruct	0.63	145 件
RAG-Anything	0.41	51 件

表 1 に実験結果を示す．Avg は全体の平均スコア，Y/N は，はい/いいえ形式，Fact は事実抽出形式，Num は数量形式，Free は自由記述形式のスコアを表す．RAG-Anything は，400 件に対する実験結果であり，はい/いいえ形式が 67 件，事実抽出形式が 75 件，数量形式が 60 件，自由記述形式が 198 件含まれている．その他のモデルに対する結果はテストデータ全てに対する結果であり，全体の件数は 1176 件，はい/いいえ形式が 169 件，事実抽出形式が 215 件，数量形式が 171 件，自由記述形式が 621 件含まれている．

暫定の結果であるが，学習なしのゼロショット設定のモデルや，表現ベースのモデルである Qwen3-VL-8B-Instruct と比較して，RAG-Anything は全体平均スコアである Avg において，21.52 となった．依然として，学習セットで学習済みの StableLM Base-Al.-7B には及ばないものの，RAG-Anything は他の学習なしモデルを上回る結果となった．特に，事実抽出形式においては 44.00 を達成し，他のモデルを大きく上回る結果となった．

また，表 2 に，回答不可能質問に対する定量評価の結果を示す．Qwen3-VL-8B-Instruct は 0.63，RAG-Anything は 0.41 となり，やや RAG-Anything の方が低いスコアとなった．

5.4.2 定性評価と誤り要因の分析

表 3 定性評価の結果

モデル	Avg	Y/N	Fact	Num	Free
RAG-Anything	57.25	61.19	65.33	68.33	49.50

表 3 に，RAG-Anything の定性評価の結果を示す．定性評価では，400 件に対して全体平均スコアである Avg が 57.25 となった．定量評価とスコアの差がある理由としては，定量評価では正しい回答と予測回答の表層的な文字列の一致やに基づいて評価が行われるため，回答不可能質問に対する評価が不可能であることや，単位や表現が異なる場合に正答とみなされないことが挙げられる．このことから，定性評価や LLM を用いた自動評価を行うことで，より正確にモデルの性能を評価できると考えられる．

表 4 誤り要因の分析結果

認識誤り	参照誤り	推論誤り	合計
11 件	102 件	58 件	171 件

定性評価で誤答とみなした 171 件に対して，要因を認識誤り，参照誤り，推論誤りに分解して分析した結果を表 4 に示す．参照誤りが最も多い結果となり，次いで推論誤り，認識誤りの順となった．以下，各誤り要因の具体例や分析結果を示す．

a） 認識誤り

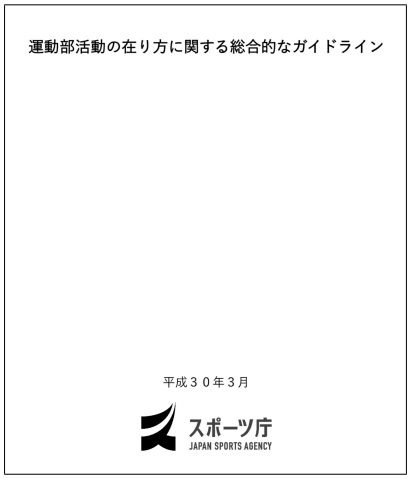
質問: 広報みずほには、様々な特集記事が掲載されていますが、その中でメインの記事として特集されているものは何ですか。解答は自由に記述してください。**回答**: 広報みずほでは、いくつかの特集記事が掲載されており、今回はメイン記事として「春、芽吹くころ、桜、咲くころ。」が特集されています。**推論結果**:「広報みずほ」には多様な特集記事が掲載されていますが、特に注目されるメインの記事として「ハッピーショット」が挙げられます。

図 2 認識誤りの例

認識誤りの例を図 2 に示す．「春、芽吹くころ、桜、咲くころ。」という特集記事がメインであるにもかかわらず，OCR でテキスト情報を正しく取得できず，誤答に繋がっている．他にも，図表内に書かれた小さな文字や，入力画像の向きが不適切である場合などに認識誤りが発生している事例が確認できた．よって，OCR の際に，テキスト認識の信頼度が一定数下回った場合，テキスト情報ではなく画像情報として情報を保持することや，画像の向きを自動で補正する仕組みが必要であると考えられる．

b） 参照誤り

参照誤りの例を図 3 に示す．文書の表紙として，「運動部活動の在り方に関する総合的なガイドライン」というタイトルとスポーツ庁のロゴが記載されたページが存在しているにもかかわらず，参照構造として図 4 のように，タイトルノードとスポーツ庁ノードが直接つながっておらず，正しい情報を参照できていない．他にも，図表とその図表のキャプションが正しく紐づ



質問: 運動部活動の在り方に関する総合的なガイドラインはスポーツ庁が作成していますか。解答は「はい」か「いいえ」で答えてください。**回答**: はい**推論結果**: いいえ

図 3　参照誤りの例

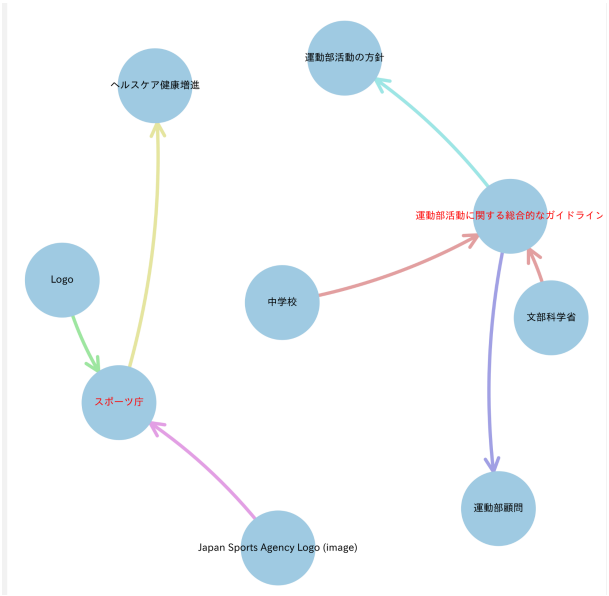


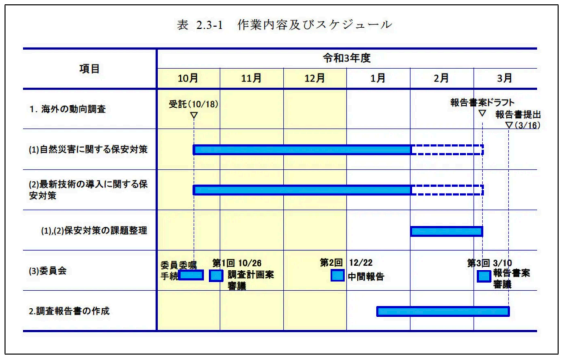
図 4　参照構造の例

いていない場合や，複数ページにまたがる情報を参照する必要がある場合などに参照誤りが発生している事例が確認できた．よって，文書の構造を理解する際に，ノードのタイプとして，タイトルやキャプションなどの特別な役割を持つノードを明示的に区別し，それらの関係性を強化する仕組みが必要であると考えられる．

c） 推 論 誤 り

推論誤りの例を図 5 に示す．スケジュール表から,「保安対策の課題整理」が令和 3 年度 2 月に開始されることを読み取る必要があるが,「委員会」の調査計画審議が 10 月 26 日に開催されることと混同して誤答している．他にも，正しい表や図が推論に利用されているにも関わらず正しく読み取ることができずに誤答している事例が確認できた．

回答不可能質問に対する定性評価の結果を表 5 に示す．51 件の回答不可能質問のうち，24 件は正しく回答不能と判定されて



質問: 石油・天然ガス開発に係る保安動向調査委員会が設置されているが、保安対策の課題整理はいつから開始されるか。解答は数量のみで答えてください。**回答**: 令和 3 年度 2 月です**推論結果**: 10 月 26 日

図 5　推論誤りの例

表 5　回答不可能質問に対する定性評価の結果

正答数	誤答数	件数
24 件	27 件	51 件

いるが，27 件は根拠のない推測回答が生成されている．根拠のない推測回答が生成される要因として，認識誤りが 1 件，参照誤りが 9 件，推論誤りが 17 件であった．推論誤りの中でも，質問に含まれているキーワードに関連した情報が文書内に多数存在し，それらを参照している際に根拠のない推測回答が生成されている事例が確認できた．よって，推論時に利用する情報の信頼度を評価しフィルタリングをすることで，根拠のない推測回答を抑制する仕組みが必要であると考えられる．

6　ま　と　め

本稿では，日本語文書画像質問応答における誤答および回答不能判定の失敗要因を，文書内参照構造の分解という観点から体系的に分析する枠組みを提案する．分析フレームワークとして，RAG-Anything を基盤とし，日本語文書画像を理解可能に拡張したモデルを用いる．JDocQA データセットを用いた実験により，提案手法は，学習なしのゼロショット設定のモデルや，表現ベースのモデルと比較して，全体平均スコアで 21.52 を達成し，他の学習なしモデルを上回る結果となった．しかし，学習セットで学習済みのモデルには及ばない結果となった．さらに，提案する参照構造分解に基づく分析フレームワークを用いて，誤答および回答不能判定の失敗要因を，認識誤り，参照誤り，推論誤りに分解して分析した結果，参照誤りが最も多い結果となった．今後の課題として，認識誤り，参照誤り，推論誤りを低減するために，ノードのタイプをより詳細に区別し，それらの関係性を強化する仕組みや，推論時に利用する情報の信頼度を評価しフィルタリングをする仕組みの導入が挙げられる．

謝　　辞

本研究の一部は科研費 23K11342 の助成を受けたものである．

文　献

[1] Xingyu Chen, Zihan Zhao, Lu Chen, Danyang Zhang, Jiabao Ji, Ao Luo, Yuxuan Xiong, and Kai Yu. Websrc: A dataset for web-based structural reading comprehension, 2021.

[2] Hao Feng, Qi Liu, Hao Liu, Jingqun Tang, Wengang Zhou, Houqiang Li, and Can Huang. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding, 2024.

[3] Zirui Guo, Xubin Ren, Lingrui Xu, Jiahao Zhang, and Chao Huang. Rag-anything: All-in-one rag framework, 2025.

[4] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking, 2022.

[5] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer, 2022.

[6] Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding, 2023.

[7] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1697–1706, January 2022.

[8] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images, 2021.

[9] Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, Zhenjiang Jin, Guang Liang, Rui Zhang, Wenzheng Zhang, Yuan Qu, Zhifei Ren, Yuefeng Sun, Yuanhong Zheng, Dongsheng Ma, Zirui Tang, Boyu Niu, Ziyang Miao, Hejun Dong, Siyi Qian, Junyuan Zhang, Jingzhou Chen, Fangdong Wang, Xiaomeng Zhao, Liqun Wei, Wei Li, Shasha Wang, Ruiliang Xu, Yuanyuan Cao, Lu Chen, Qianqian Wu, Huaiyu Gu, Lindong Lu, Keming Wang, Dechen Lin, Guanlin Shen, Xuanhe Zhou, Linfeng Zhang, Yuhang Zang, Xiaoyi Dong, Jiaqi Wang, Bo Zhang, Lei Bai, Pei Chu, Weijia Li, Jiang Wu, Lijun Wu, Zhenxiang Li, Guangyu Wang, Zhongying Tu, Chao Xu, Kai Chen, Yu Qiao, Bowen Zhou, Dahua Lin, Wentao Zhang, and Conghui He. Mineru2.5: A decoupled vision-language model for efficient high-resolution document parsing, 2025.

[10] Eri Onami, Shuhei Kurita, Taiki Miyanishi, and Taro Watanabe. JDocQA: Japanese document question answering dataset for generative language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 9503–9514, Torino, Italy, May 2024. ELRA and ICCL.

[11] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *AAAI*, 2023.

[12] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *AAAI*, 2021.

画像インペインティングを用いた展示物外観の意外性分析

木下　真帆†　　桑田　若菜†　　三林　亮太††　　大島　裕明†

† 兵庫県立大学 情報科学研究科　〒 651–2197 兵庫県神戸市西区学園西町 8–2–1
†† 神戸大学 国際文化学研究科　〒 657-8501 兵庫県神戸市灘区鶴甲 1–2–1
E-mail: †ad25n018@guh.u-hyogo.ac.jp, †af25x004@guh.u-hyogo.ac.jp, ††mibayashi@people.kobe-u.ac.jp, †ohshima@ai.u-hyogo.ac.jp

あらまし　本研究は，画像生成モデルが学習した一般的な外観を基準として，その常識から外れる部分に着目することで，展示物の外観に含まれる意外な特徴を分析する手法を提案する．展示物の装飾や模様には，全体の様式に調和しているように見えても，局所的に特定の地域や文化に固有の特徴が現れている場合がある．本研究は，そのような特徴を，生成モデルの振る舞いを通して捉えることを目的とする．対象として，国立民族学博物館が所蔵するこけし，太鼓，仮面の展示物を用いる．特定の地域の画像を除外したデータで生成モデルを学習させ，そのモデルを用いて画像の一部をインペインティングする．インペインティングされた部分と元画像との差は，生成モデルがもつ一般的な外観の枠組みでは説明しにくい特徴を反映していると考えられる．本研究では，この差を局所的な外観の意外さを捉える手がかりとして用いる．本手法を複数カテゴリの展示物に適用し，地域や様式の違いによって現れる装飾や模様の特徴を分析する．複数種類の展示物を対象とした実験の結果，典型的な特徴は比較的安定して生成される一方で，地域性の強い装飾や模様は生成されにくい傾向が確認された．これらの結果は，生成モデルが想定する一般的な外観からの逸脱が，インペインティングの困難さとして現れる可能性を示唆している．

キーワード　画像インペインティング，外観特徴，LoRA

1　はじめに

人は，予測や期待を裏切る出来事に出会うと，思わずその部分に注意を向ける傾向がある．見慣れた対象の中に，わずかな違和感や意外な要素が含まれていることに気づくと，それは強く印象に残り，記憶や理解の手がかりとなることが知られている．心理学や認知科学の分野では，このような意外性が注意喚起や学習を促進する重要な要因であることが指摘されている [6]．物における意外性の現れ方には，用途や地域，時間といった複数の側面が存在する．その中でも，人が瞬時に知覚しやすいものとして，視覚的な意外性が挙げられる．この視覚的な意外性にも，いくつかの異なる側面が存在する．たとえば，全体形状や構造が既知の対象と大きく異なることによる意外性や，形状自体は典型的である一方で，装飾模様や細部表現が期待と異なることによる意外性が挙げられる．

本研究では，装飾や模様といった局所的な視覚的特徴において発生する意外性を定量化することを目的とする．対象として，国立民族学博物館に展示されているこけし，太鼓，仮面といった種類の展示物を扱う．これらの展示物は全体形状が類似しているように見えても，表情や胴部の模様，紐による括り方などの装飾といった細部には，地域性が存在する．このような特徴は，典型的な外観の中に埋め込まれた局所的な意外性を分析する対象として適している．

たとえば，日本の伝統工芸品であるこけしは，円柱状の胴体と球状の頭部という共通した形状を持つ．しかし，国や地域によって表情や腹部の模様に違いがある．図 1 にいくつかの地域のこけしの例を示す．一見すると典型的でよく似た外観であっても，胴部の模様の配置や髪型に着目すると，地域ごとの違いが分かる場合がある．このような違いが，細部の視覚的特徴によって生じる意外性の一例であると考える．

このような意外性は，対象を単独で見た時に生じるというよりも，当該カテゴリに属する展示物としてどのような外観が期待されるかという暗黙の基準とのずれによって生じると考えられる．すなわち，意外性とは，対象そのものの性質ではなく，ある文脈や分布に基づく期待からの逸脱として捉えることができる．局所的な意外性を捉えるためには，まずそのような期待される外観を何らかの形で表現し，対象がその期待からどの程度外れているかを評価する必要がある．

この観点から，特定の地域以外の展示物群における視覚的特徴を学習したモデルは，当該地域を除いた場合に期待される典型的な外観を表現していると解釈することができる．このモデルに対して，当該地域の展示物画像のインペインティング（inpainting）が困難となった場合，その局所領域は学習された期待から逸脱した視覚的特徴を含んでいる可能性が高い．この

図 1　左から宮城県，山形県，青森県のこけし

際のインペインティングの困難さは，単なるモデルの表現能力不足ではなく，学習時に除外された地域固有の視覚的特徴が，期待される外観の分布に含まれていないことに起因すると考えられる．

そこで本研究では，画像生成モデルを用いて，展示物を部分的にインペインティングすることが可能であるかを検証する．このとき，画像生成の際には特定の地域の展示物画像を学習から意図的に除外したデータセットを用いる．インペインティングが困難となる領域を分析することで，地域に固有な視覚的特徴の可視化や局所的な意外性の定量化の実現を目指す．

本研究の主な貢献は，画像生成モデルのインペインティングに基づき，局所的な視覚的意外性を定量化する枠組みを提案した点にある．

2 関 連 研 究

2.1 意外性と注意や記憶に関する理論的背景

意外性は心理学や神経科学において，予測や期待と実際の知覚入力とのずれを検出する認知的反応として広く研究されている．人間は過去の経験に基づいて外界の状態を予測しており，その期待からの逸脱は注意を引き，理解の手がかりとなることが知られている．Itti ら [6] は，観測によって確率モデルの信念がどの程度更新されたかを，KL ダイバージェンスにより測定するベイジアンサプライズを提案し，意外性を定量的に扱う枠組みを示した．また，シャノン情報量 [14] に基づく驚きの定義は，事象の予想困難性を表す一般的指標として広く用いられている．他にも，予測処理理論では，人の知覚は外界に対する予測と実際の入力の差に基づいて形成されると考えられている．この予測誤差は注意の向け方にも関与し，予測から外れた刺激ほど目に留まりやすいとされている [2]．視覚認知研究においても，意外性とは刺激自体の顕著性ではなく，文脈に基づく期待からの逸脱として捉えられている [16]．

本研究では，これらの考え方を踏まえ，生成モデルの持つ暗黙の事前分布を，一般的な外観への期待とみなす．これによって，インペインティング時の元画像との誤差を視覚的な意外性として扱う．

2.2 文化保存や理解における AI 活用研究

Quan ら [9] は，中国貴州省のミャオ族のバティック文化の一つである蝋結染めにおいて，約 15,000 枚に及ぶ大規模な画像データセットを構築した．また画像分類・識別において，それを知識グラフと連携させることで，ユーザがバティックの模様の名称や文化的意味，あるいは使用上のタブーを理解できる仕組みを作成した．また，Goodarzi ら [4] は，貴重な実測データに基づき，歴史的景観の空間構造や特徴量パラメータを抽出，学習する識別的意思決定支援ツール（DDST）を提案した．写真測量と 3D ポイントクラウドを用いた機械学習により，遺産景観の DNA を捉え，それを維持，または改善した設計シナリオの評価と生成を可能にした．

これらにより，AI 技術が文化遺産の単なる保存にとどまらず，現代社会における理解の深化や創造における継承を促す有効な手段であることが示されている．

2.3 拡散モデルを用いたインペインティングによる解釈と異常検知

近年の拡散モデルは周囲の文脈に整合する自然な補完能力を持ち，欠陥補完や異常検知への応用が進められている．特に，画像インペインティング [1] では，マスク領域を段階的に生成することで，高品質な部分生成を行うことが可能である．この方法は再生成誤差を用いた異常検知手法にも利用されている [8]．Ancha [15] らは，拡散モデルを用いて，入力画像から学習分布外の異常を生成的に除去した編集画像を合成し，元画像との差分から異常を検出する手法を提案した．このアプローチでは，モデルが生成しやすい外観と，生成しにくい外観との差が異常の手がかりとして利用されている．このような再構成誤差に基づく異常検知の考え方は，GAN を用いた方法でも行われている．再構成誤差に基づく異常検知は拡散モデル以前にも提案されており，Schlegl [13] らの AnoGAN はその代表例である．AnoGAN では，正常データのみで学習した GAN による再構成誤差を異常スコアとし，学習分布に含まれない特徴が再構成困難性として現れることを示されている．

一方で，画像の一部をマスクしてインペインティングするという操作は，機械学習モデルの判断根拠を局所的に分析する説明可能 AI の考え方とも共通している．Ribeiro ら [10] が提案した LIME は，入力の局所的な変化に対するモデルの応答を用いて予測理由を説明する手法である．インペインティングによる生成画像と元画像との差分は，モデルがどの領域の外観を重視しているかを示す情報と解釈できる点で，この考え方と親和性が高い．

本研究では，Diffusion モデルのインペインティング能力に着目し，補完が困難な局所領域をモデルの想定からの逸脱として捉える．対象物の異常を直接検出するのではなく，インペインティング結果と入力画像との差分を用いて，画像中のどの部分がモデルの期待から外れているのかを定量的に評価することを目的とする．

2.4 生成モデルの出力に対する品質評価

生成モデルの出力した画像の品質評価には，視覚的類似度の指標が重要な役割を果たす．従来の指標としては MAE や SSIM [17] がある．MAE は画素値の平均絶対誤差を直接計算する最も単純な手法だが，人間の知覚と必ずしも一致しないという問題があり，補助的に用いられることが多い．SSIM は輝度，コントラスト，構造の三要素から画質を評価する指標で，人間の知覚に近いとされている．これらに対して，LPIPS [18] は，深層学習に基づいた特徴抽出によって，画像間の知覚的類似性を定量化する．LPIPS は人間知覚との相関が高いとされる距離尺度であり，生成モデル評価にも用いられている．本研究では，インペインティングによる生成画像と元画像の差異を LPIPS により定量化する．

図 2　特殊な造形により排除したこけしの例

3　問題定義とデータセット

3.1　問題定義

本研究の目的は，展示物画像の局所的な領域を補完する困難さを指標として，地域固有，またはその展示物特有の局所的な視覚意外性を定量化することである．具体的には，ある地域の展示物画像に対し，局所的な領域にマスク処理を施した画像を入力とする．出力として，インペインティング画像と元画像との視覚的な差異を数値化することで，その領域における視覚的な意外性スコアを得る．

3.2　データセット

本研究で使用するデータは国立民族学博物館 [1]（以下みんぱく）が所有する標本画像データを基に構築したものである．同データセットには多様な民族資料が収録されている中で，標本名こけしに文字列一致する 593 枚の画像を収集した．同じ方法で，太鼓を 929 枚，仮面を 4,629 枚収集した．各画像には産地に関するメタデータが付与されており，これは県単位であるものから地方単位，不明であるとデータなど様々である．これを用いることで，種類ごとの展示物における地域的様式差の解析が可能な構造になっている．こけしは地域ごとの系譜差が大きい展示物であるが，極端に丸味を強調した胴体や，特徴的なヘアスタイルをした頭など，こういった伝統様式とは異なるデザイン処理が加えられているものは，地域的な特徴抽出には適さない．太鼓や仮面にも同様にこういった伝統様式と大きく異なった展示物は存在する．

こうした極端な様式の展示物を省き，学習データを地域を代表するバランスの取れたスタイルに限定するため，本研究では展示物の形状の典型性を定量化する枠組みとして，VisualRank [7] を用いた．すべてのサンプルに対して典型性スコアを算出し，あらかじめ設定した閾値を下回る展示物を除外した．この処理により，特殊な造形の展示物や偶然文字列が一致してしまった別カテゴリの展示物を自動的に排除して，地域様式の分析に適した集合のみを抽出できるようにした．除外した例を図 2 に示す．この精緻化をした結果，こけし，太鼓，仮面の枚数はそれぞれ 536 枚，624 枚，2,479 枚となった．

本研究では，背景やスケールの違いが意外性推定に与える影響を抑えるため，すべての画像に対して前処理を行った．具体的には，背景除去とサイズ正規化を行い，同一サイズのキャンバス中央に配置されるようにした．

1：https://www.minpaku.ac.jp/

また，本研究でこけしカテゴリにおける地域ごとの様式差を分析するにあたって，主要なこけし産地である東北六県のこけしに着目した．各地域には多数のこけしが存在するため，VisualRank により典型性の高いサンプルを抽出し，各地域を代表するこけし画像を選定した．これにより，各地域における極端な造形や特徴の個体を排除し，各地域の様式的特徴を比較しやすい代表例を構成した．実際に選定した東北六県のこけしを図 3 に示す．

図 3　左から順に秋田，宮城，岩手，山形，青森，福島の代表的なこけし

4　LoRA 学習とインペインティングによる視覚的な意外性推定

本研究では，画像生成モデルのインペインティングした画像と元画像との誤差を用いて，画像中の局所領域における視覚的意外性を推定する手法を提案する．提案手法は，

1. Stable Diffusion の LoRA 学習
2. 局所領域のインペインティング
3. 生成された画像と元画像の誤差に基づく意外性スコアの算出

の 3 段階から構成される．図 4 に提案手法の全体概要を示す．

まず，分析対象地域の画像を除外し，それ以外の地域画像のみを用いて LoRA 学習を行う．これにより，当該カテゴリにおける一般的な外観を表現する生成モデルを構築する．次に，除外した地域の展示物画像に対し，定義した局所領域をマスクしてインペインティングを行う．この生成結果は，モデルが想定する典型的外観に基づく再構成とみなす．最後に，生成画像と元画像のマスク領域内における LPIPS 距離を算出し，この差の大きさが局所的な視覚的意外性に対応しているかを分析する．

4.1　局所領域マスク生成

本研究では，インペインティングによる生成の対象の部位と

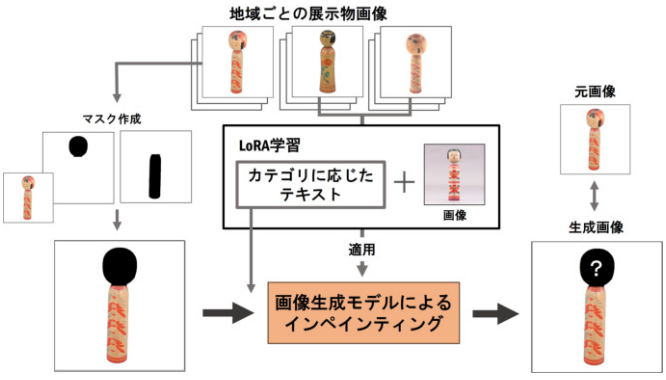


図 4　全体の流れ

して，展示物ごとに局所領域を定義した．具体的にはこけしでは頭部と胴体，太鼓では面と胴，仮面では目と口，鼻といった各部位を対象とした．既存の自動セグメンテーションモデルでは，文化財の複雑な意匠や質感に影響され，目的とする部位分割が困難であった．そのため本研究では，まず上記で対象とした部位ごとに手動アノテーションを行うことで，関心領域を設定した．その後，展示物の形態に適合するように領域を切り抜くことで，インペインティングに用いる部位ごとのマスクを作成した．これによって対象物の外観を特徴づける，大まかな部位ごとのマスクを得ることができる．実際にマスク作成の流れを図 5 に示す．

4.2 生成モデルの構築

本研究では，局所的意外性を地域の展示物ごとに評価するため，各展示物のカテゴリにおいて，分析対象とする地域を 2 つずつ選出した．以降，これらの地域集合を $\mathcal{R} = r_1, r_2$ と表記する．各地域 $r_i \in \mathcal{R}$ に対して，当該地域の展示物画像を学習データから除外し，それ以外の地域の展示物画像のみを用いて生成モデルの学習を行った．このとき，学習に用いるデータには r_i に属する画像は一切含まれない．

学習後，除外した地域 r_i に属する展示物画像を入力とし，画像中の局所領域をマスクした上でインペインティング [1] による生成を行う．この生成結果と元画像との差異を用いて，地域 r_i における局所的意外性を算出する．

この手順を，選出したすべての地域 $r_i \in \mathcal{R}$ に対して同様に適用することで，各地域の展示物が他の地域に基づく期待外観からどの程度逸脱しているかを対照的に評価する．

5 実　　験

5.1 生成モデルの学習

本研究では，Diffusion モデルに基づく画像生成手法である Stable Diffusion [11] を採用し画像生成を行う．Stable Diffusion は高解像度な画像生成が可能であるだけでなく，既存モデルを基にした追加学習が比較的容易であるという利点を持つ [3], [12]．本研究では，対象地域の展示物の特徴を除外した一般的な展示物の外観を学習させるため，LoRA [5] による追加学習を実施した．LoRA は，大規模言語モデルのファインチューニングにおけるメモリ効率と計算コストの問題に対処するために導入された手法である．本研究では，この LoRA 学習を用いることで，対象地域以外の展示物画像のみから，地域固有ではないみんぱくの展示物の一般的な外観パターンを学習させた．

図 5　左から，元画像，一般的な矩形領域による部位指定例，展示物の形態に合わせて成形した最終的な部位マスク

5.1.1 LoRA 学習のためのデータセット構築

本節では，LoRA 学習に用いたデータセットの構成及び学習条件について述べる．テキストプロンプトは表 1 に示すように，対象物のカテゴリを簡潔に表す表現を用いた．テキスト情報はカテゴリ情報を補足する役割に限定し，主として，みんぱくの展示物の視覚的特徴の学習を促す設計とした．Stable Diffusion のベースモデルには `Stable-Diffusion-v1-5` [11] を使用した．学習率は 5×10^{-5}，バッチサイズは 1，LoRA ランクは 32 とした．学習ステップ数はデータ規模に応じて調整し，太鼓は 8,000step，こけしおよび仮面は 10,000step とした．

5.1.2 こけしにおける特別な対応

こけしカテゴリにおいては，他の展示物とは異なる方法でデータセットを用意して学習を行った．

太鼓及び仮面については，前述したとおりそれぞれ簡潔なテキストプロンプトと対応付けた画像を用いた LoRA 学習を行った。一方こけしについて同様に *a kokeshi doll* というプロンプトとこけし画像を用いて学習を行った場合，インペインティング時に画像中のどの領域をマスクしても，生成結果として顔が出現する問題が確認された．学習データに用いたこけし画像はいずれも類似した画角で，顔が中心付近に配置されている．これによって LoRA 学習がこけしというテキストと顔が中心に存在する構造を強く結びつけてしまったことが起因すると考えられる．

この問題を緩和するため，こけし画像を上部約 20–30%の位置で分割した．これによって上方領域を正方形化した頭部画像，下方領域を正方形化した胴体画像を作成した．また，それぞれの部位に対しては次節で記載する部位に対応したプロンプトを付与し，学習データに加えた．

以上の構成により，こけしを頭部と胴体を持つ構造的な対象として学習させ，インペインティング時にマスク位置に依存したプロンプトを用いることで，顔が不自然に生成される現象が解消された．本研究では，この手法をこけしカテゴリにのみ適用し，以降の実験を行った．

5.2 局所領域のインペインティング

本研究では，画像内の特定領域をマスクし，拡散モデルによるインペインティングを用いて当該領域の生成を行う．マスクされた領域は視覚的意外性を評価する対象であり，マスク外の領域は，生成時の文脈情報として固定される．

具体的には，Stable Diffusion に LoRA を適用することで，事前学習モデルの重みを大きく変更することなく，生成結果の傾向を調整する．その上で ControlNet を組み合わせ，インペインティング時にマスク外領域の構造および視覚的特徴を条件として生成過程に与える．ControlNet は，入力画像に基づく条件情報を拡散モデルの各ステップに反映させる拡張手法であり，生成結果を与えられた条件と整合させる役割を担う．これにより，マスク外領域の視覚的文脈を保持したまま，マスク領域内部を周囲と自然に整合する外観として生成することができる．本研究では，インペインティングのための ControlNet として

表 1　学習およびインペインティングに使用したプロンプト

用途	カテゴリ（部位）	プロンプト
LoRA 学習時	こけし（頭部）	kokeshi doll's head with simple face
	こけし（胴体）	kokeshi doll's torso with some pattern, cylindrical body
	こけし（全体）	a kokeshi doll with a round head and a cylindrical body, painted with a simple face and some patterns
	太鼓	a drum
	仮面	a mask
インペインティング時	全対象	上記プロンプトに加えて centered, isolated on a plain white background, studio lighting

`control_v11p_sd15_inpaint` を使用し，マスクには 4.1 章で作成したものを用いる．

本研究では，このような生成結果と元画像との差異を，局所的な視覚的意外性の手がかりとして用いる．この差異の定量化方法および局所意外性スコアの定義については，次節で詳述する．

5.3　インペインティング画像と元画像の視覚的距離計算

本研究では，各局所領域に対して複数回インペインティングを行い，得られた生成画像群と元画像との間で LPIPS 距離 [18] を算出する．LPIPS の計算は，画像全体ではなく，マスクにより指定した局所領域内部の画素のみに限定して行った．具体的には，元画像および生成画像のマスク領域に対応する部分を抽出し，その部分画像同士の LPIPS 距離を計算した．インペインティングではマスク領域のみが生成対象となるため，評価もその生成部分に対応する領域に限定している．

6　結果と考察

本章では，まずインペインティング結果を示し，続いて各局所領域における LPIPS 平均値を示す．これらの差を，局所的な外観の意外さを捉える手がかりとして解釈し，複数カテゴリの展示物に適用した結果を分析する．各カテゴリの定量評価結果は，表 2 にまとめて示す．

6.1　こけしにおける局所外観特徴のインペインティング分析

こけしを対象として，提案手法による局所領域のインペインティング結果を示し，定性的な観点から生成結果と意外性の傾向を分析する．

まず，VisualRank により抽出したみんぱく所蔵のこけし代表例 10 枚を図 6 に示す．これらのこけしの胴体部分には，赤系を基調とした花柄模様が用いられることが多く，同一の柄が縦方向に連続して描かれている．また，首元及び胴体最下部には境界を強調するような横縞模様が描かれている例が多い．頭部に関しては，前髪及び顔の両側に限定して髪が描かれており，簡略化された髪表現が主流であることが分かる．本研究では，これらを生成モデルが学習する一般的なこけし外観の基準とみなす．

次に，意外性が低い展示物として直感的に判断される例として，山形県のこけしを示す．胴体のインペインティング結果では，首元及び胴体下部の網模様が保持されており，胴体全体には花柄に類似した模様の連続が生成されている．また，頭部のインペインティングにおいては，顔の向きが少し変化しているが，前髪と両側のみに限定された簡略的な髪の表現といった共通した特徴が確認できる．生成結果はこけしの典型例とある程度特徴が一致しており，本例は一般的なこけしの外観に近いと解釈できる．LPIPS の平均値は頭部 0.561，胴体 0.434 であった．

一方で，全体的に意外性が高いと判断される例として，秋田県のこけしを示す．このこけしでは，胴体において花柄が連続的に配置されるのではなく，一枝の花が独立して描かれている点が特徴的である．また，頭部は全体が塗りつぶされたようなおかっぱ状の髪型となっている点も特徴として挙げられる．インペインティングを行った結果，頭部では前髪及び顔両側の髪のみが描かれ，元画像に見られた全体的な髪表現や装飾的な模様は再現されなかった．胴体においても，元画像に見られる独立した花枝の構成は失われ，一般的な花柄模様へと書き換えられている．このような結果は，例として示した秋田県のこけしが生成モデルの学習した外観分布から逸脱しており，モデルにとってインペインティングが困難な視覚的特徴を含んでいることを示唆している．すなわち，本例では，装飾様式や髪表現といった局所的特徴が，典型的なこけし外観に対して意外性の高い要素として機能していると考えられる．LPIPS の平均値も秋田県のこけしでは，頭部が 0.638，胴体が 0.553 となり，いずれの部位でも山形県より高い値を示した．これによって，生成結果の元画像からのずれが数値的にも確認することができたと言える．

6.2　太鼓における局所外観特徴のインペインティング分析

太鼓を対象として，提案手法による局所領域のインペインティング結果を示し，定性的な観点から生成結果と意外性の傾向を分析する．太鼓は面の模様，胴体における紐や布による装飾が地域ごとに異なる点が特徴であり，本研究の枠組みにおい

図 6　みんぱくに所蔵されている中でも典型的な外観のこけしの例

て局所的意外性を検証する対象として適している．

まずパキスタンの太鼓を用いて分析する．本例では，面部分をマスクして生成した結果と，胴体部分をマスクして生成した結果をそれぞれ示す．面部分をインペインティングによって生成した場合，元画像に存在する中央の灰色の円形模様は再現されなかったが，単一の素材からなる比較的単純な面構造が生成された．また，胴体部分をインペインティングした結果では，面を引っ張る紐の構造がある程度自然に生成されている．各部位の LPIPS 平均値は 0.51 から 0.56 程度を示した．紐の材質や細部の種類は元画像と一致していないものの，太鼓としての構造的整合性は保たれており，生成結果が元画像から大きくずれることはなかった．このことから，胴体における紐構造そのものは，モデルが学習した一般的な太鼓の外観に含まれている一方で，その細かな様式差は十分に再現されていないと考えられる．

次に，ビルマ連邦社会主義共和国の太鼓の結果を示す．この太鼓では，胴体の両端に赤い布の装飾が施されている点が大きな特徴である．まず，赤い装飾部分をマスクしてインペインティングした場合，装飾要素は一切生成されず，太鼓の両面を同一素材で単純に接続するような胴体構造が生成された．次に，面部分をマスクしてインペインティングした場合，パキスタンの太鼓と同様に，面中央に見られる円形模様は再現されず，比較的均質な面構造が生成された．さらに，両端の赤い装飾を除いた胴体中央部分のみをマスクしてインペインティングした結果では，両側の赤い装飾と同一の素材や，面に似た単一素材による胴が生成された．各部位の LPIPS の平均値は 0.62 から 0.77 とパキスタンの例よりも明確に大きい値が観察されている．この結果は，マスク外領域の強い文脈情報により，生成モデルが周囲の装飾を接続する形で補完を行ったことを示しており，局所領域の設定がインペインティング結果に大きな影響を与えることを示している．

6.3 仮面における局所外観特徴のインペインティング分析

仮面を対象として，提案手法による局所領域のインペインティング結果を示し，生成の成否と視覚的意外性との関係について定性的に分析する．仮面は，目，鼻，口といった顔部位の配置や素材，さらには人型か人外といった表現様式の幅が広く，生成モデルが想定する一般的な外観とのずれが浮き彫りになりやすい対象である．

まず，パプアニューギニアの仮面を分析する．目領域をマスクしてインペインティングを行った結果，貝殻を用いた装飾的な目の表現は再現されず，一般的な眼球状の形状や，単純な暗部として補完される傾向が見られた．また，頭部全体をマスクした場合，インペインティング結果では高い頻度で目が出現した．本仮面は縦方向に長く，頭部領域の占める面積が大きい．そのため，通常の顔構造において目が配置されやすい位置を想定して，生成モデルが人型の顔構造を補完した可能性がある．なお，口元では元画像に見られる口角の上がった表情は再現されなかったものの，口に相当する形状自体は生成されており，素材感も大きくは変化していなかった．そのため，視覚的印象としては差異が認められる一方で，LPIPS 距離は 0.392 と比較的小さい値にとどまったと考えられる．この仮面において，各部位で LPIPS の平均値は 0.392 から 0.538 を示した．

次に，ブータン王国の人外的外観を持つ仮面をもとに分析する．この仮面は，全体として人の顔構造から大きく逸脱した造形を持ち，角状の突起，側方に配置された眼，強調された顎縁装飾などが特徴である．この仮面は，いずれの局所領域においても，インペインティングによる生成は困難であった．頭部の角状構造を含む領域をマスクした場合は，インペインティング画像ではそれらの突起が消失し，平坦な頭部形状として補完される傾向が顕著に見られた．また，口元についても，元画像に見られる歯や舌は再現されなかった．周囲の緑色の装飾を文脈として解釈した結果，髭のような表現へと置き換えられる例が見られた．目元領域に関しては，黒い二点状の要素が生成されることはあったものの，それが目として明確に再現されているとは言い難い結果となった．また，元画像に見られる顔の側面にあるような目の配置は再現されなかった．とりわけブータン王国の仮面では，LPIPS が各部位で 0.60 以上を示し，最大で 0.724 に達した．これらの結果は，本仮面が特定の部位のみならず，仮面カテゴリ全体において非典型的な外観を有していることを示していると考えられる．すなわち，本例では局所的な意外性というよりも，生成モデルが参照可能な一般的文脈自体が存在しないために，全体的に生成結果が破綻するケースであると解釈することができる．この結果から，本手法が局所的な様式差とカテゴリ全体からの逸脱を異なる挙動として検出する可能性があると示唆された．

6.4 複数地域のこけしにおける局所外観特徴のインペインティング分析

さらに，本研究では東北六県の代表的なこけしを対象として，地域ごとの局所外観特徴の差異を定量的に比較した．各地域に対して頭部および胴体領域にマスクを設定し，それぞれインペインティングを行う．また，各マスク領域において，生成画像と元画像の LPIPS 距離を算出する．インペインティングの結果を図 7 に，LPIPS 値の結果を表 3 に示す。頭部領域では山形県のこけしが最も大きな LPIPS 距離を示し，胴体領域では秋田県のこけしが最も大きな値となった．具体的な生成結果を確認すると，各地域のこけしに見られる特徴的な外観要素が，インペインティングによってより一般的なこけしの表現へと置き換えられる傾向が確認された．例えば山形県のこけしでは，前髪の短い真っ黒なおかっぱ状の髪型が特徴として観察されたが，頭部をインペインティングした結果では，典型的なこけしに見られる，前髪と顔の両側のみを簡略化して表現した髪型へと置き換えられる傾向が見られた．また秋田県のこけしでは，胴体に大きな一輪の花のようなモチーフが描かれていたが，胴体領域をインペインティングした場合，この独立した花の構成は再現されず，複数の花が縦方向に連続する典型的な花柄模様へと変化する結果が観察された．

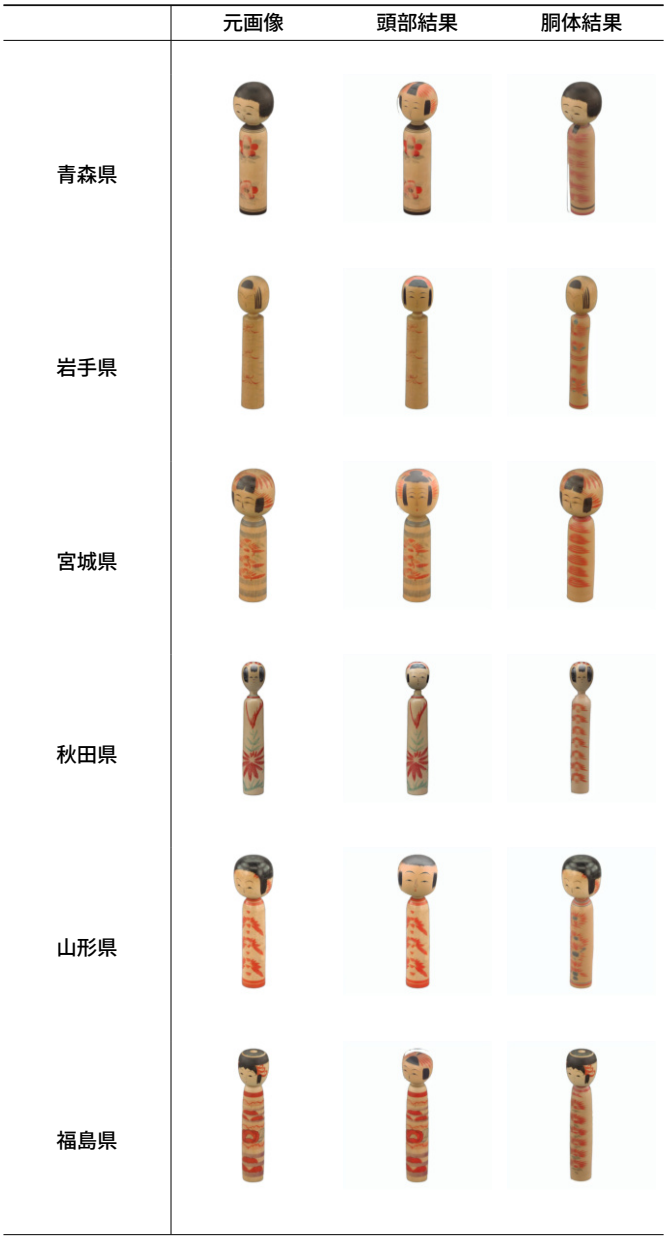


図 7　東北六県の代表的なこけしにおける各部位インペインティングの結果例

表 3　東北六県の代表的なこけしにおける部位ごとの LPIPS 結果

	秋田	宮城	岩手	山形	青森	福島
頭部	0.56	0.53	0.55	**0.64**	0.58	0.55
胴体	**0.70**	0.54	0.61	0.50	0.54	0.50

6.5　考　　察

本研究では，画像生成モデルによる局所領域のインペインティング結果と元画像との差異を用いて，展示物外観における局所的な視覚的意外性を分析する方法を提案した．その結果，対象とする展示物の種類や地域によって，インペインティングによる生成結果と元画像の差に一定の傾向を確認することができた．典型的な様式に近い展示物では，生成結果は元画像と概ね一致しており，LPIPS の値も比較的低い値が確認された．一方，地域固有の装飾や典型的ではない表現を含む場合には，生成結果がより一般的な様式へ置き換えられる傾向がみられ，LPIPS 値も高くなった．これらの結果から，本手法では様式の違いに由来する局所的な意外性を捉えることができていると解釈できる．

しかし，仮面のようにカテゴリ内での外観の多様性が大きい展示物では，局所的な非典型性は，一般的な外観として生成され発見することができるものの，対象全体が生成モデルの想定する外観から大きく逸脱している場合には，生成結果が全体的に崩れることがあった．これは，本手法が装飾や模様などの局所的な外観特徴の違いに対しては有効に反応することを示している．一方で，対象全体がカテゴリの典型的外観から大きく逸脱している場合には，インペインティングによる局所的な比較が適用しにくくなる可能性がある．

7　ま　と　め

本研究では，画像生成モデルによる局所領域のインペインティング結果と元画像の差異を用いて，展示物外観における局所的な視覚的意外性を分析する方法を提案した．実験では，こけし，太鼓，仮面といった展示物を対象とし，インペインティングによる生成結果と元画像との LPIPS 距離を比較した．その結果，典型的な様式に近い展示物では，生成結果と元画像の差は小さく，LPIPS 値も低くなる傾向が確認された．一方で，地域固有の装飾や典型的でない表現を含む場合には，生成結果がより一般的な様式へ置き換えられる傾向が見られ，LPIPS 値が高くなることが確認された．これらの結果から，本手法により，展示物の装飾や模様などの局所的な外観特徴に基づく視覚的意外性を，定量的に捉えられる可能性が示された．

7.1　今後の課題

本研究における最大の課題は，生成モデルにおけるインペインティングの困難性と，人間の感じる主観的な意外性との対応関係を定量的に検証することである．意外性は人間の認知に依存する概念であり，人手評価実験を通じた検証が不可欠である．インペインティングによる元画像との誤差に基づく指標が，人間の直感とどの程度一致するのかを明らかにすることにより，本手法の妥当性をより明確に示すことが可能となる．また，部位やカテゴリに応じた生成モデルの学習方法を検討するべきであるという課題が挙げられる．本研究では LoRA を用いてカテゴリ単位で外観を学習したが，同一空間において部位に応じた局所的な特徴が十分に捉えられていない可能性が示唆された．これを解決するために，今後は本研究のこけしのように部位毎に異なる LoRA 学習を行うことを考える．あるいは部位情報を条件として明示的に与えることで，局所意外性の推定精度を向上させることができると考える．

謝　　辞

本研究は，JSPS 科研費 JP25K03229，JP25K03228，JP24K03228 の助成を受けたものです．また，本研究の実施にあたっては，国立民族学博物館より提供いただいた展示物デー

タベースを利用しました．ここに記して謝意を表します．

文　献

[1] Nicolas Cherel, Andrés Almansa, Yann Gousseau, and Alasdair Newson. Diffusion-based image inpainting with internal learning. In *Proceedings of the 32nd European Signal Processing Conference (EUSIPCO 2024)*, pp. 446–450, 2024.

[2] Karl Friston. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, Vol. 360, No. 2456, pp. 825–836, 2005.

[3] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *Proceedings of the 2023 Eleventh International Conference on Learning Representations (ICLR 2023)*, pp. 1–31, 2023.

[4] Parichehr Goodarzi, Mojtaba Ansari, Farzad Pour Rahimian, Mohammadjavad Mahdavinejad, and Chansik Park. Incorporating sparse model machine learning in designing cultural heritage landscapes. *Automation in Construction*, Vol. 155, , 2023.

[5] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 2022 International Conference on Learning Representations (ICLR 2022)*, 2022.

[6] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. In *Proceedings of the 19th Advances in Neural Information Processing Systems (NIPS 2005)*, pp. 547–554, 2005.

[7] Yushi Jing and Shumeet Baluja. Visualrank: Applying pagerank to large-scale image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30, No. 11, pp. 1877–1890, 2008.

[8] Jing Liu, Zhenchao Ma, Zepu Wang, Yang Liu, Zehua Wang, Peng Sun, Liang Song, Bo Hu, Azzedine Boukerche, and Victor Leung. A survey on diffusion models for anomaly detection. In *Proceedings of the 2025 International Joint Conferences on Artificial Intelligence (IJCAI 2025)*, pp. 1–9, 2025.

[9] Huafeng Quan, Yiting Li, Dashuai Liu, and Yue Zhou. Protection of guizhou miao batik culture based on knowledge graph and deep learning. *Heritage Science*, Vol. 12, , 2024.

[10] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)*, pp. 1135–1144, 2016.

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pp. 10684–10695, 2022.

[12] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023)*, pp. 22500–22510, 2023.

[13] Thomas Schlegl, Philipp Seeböck, Sebastian Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *Proceedings of the 2017 International Conference on Information Processing in Medical Imaging (IPMI 2017)*, pp. 146–157, 2017.

[14] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, Vol. 27, No. 3, pp. 379–423, 1948.

[15] Ancha Siddharth, Jiang Sunshine, s Manderson Travi, Brandt Laura, Du Yilun, R. Osteen Philip, and Roy Nicholas. Anomaly detection using generative diffusion models for off-road navigation. In *Proceedings of the 2025 IEEE International Conference on Robotics and Automation (ICRA 2025)*, pp. 1–13, 2025.

[16] Christopher Summerfield and Tobias Egner. Expectation and attention in visual cognition, trends in cognitive sciences. *Trends in Cognitive Sciences*, Vol. 13, No. 9, pp. 403–409, 2009.

[17] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, Vol. 13, No. 4, pp. 600–612, 2004.

[18] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, pp. 586–595, 2018.

表 2　各展示物におけるマスク画像とそれを用いたインペインティングの結果

展示物画像	展示物画像＋マスク画像	インペインティング結果	LPIPS 平均
 山形県のこけし			0.561
			0.434
 秋田県のこけし			0.638
			0.553
 パキスタンの太鼓			0.513
			0.562
 ビルマ連邦社会主義共和国の太鼓			0.620
			0.691
			0.768
 パプアニューギニアの仮面			0.538
			0.513
			0.392
			0.490
 ブータン王国の仮面			0.607
			0.665
			0.724
			0.638

夜間運転時に不安感を誘発する要因の検出手法の提案とその定量的評価

高岡　晴玖† 　服部　峻†† 　宮城　茂幸††

† 滋賀県立大学工学部電子システム工学科　〒 522–8533 滋賀県彦根市八坂町 2500
†† 滋賀県立大学先端工学研究院　〒 522–8533 滋賀県彦根市八坂町 2500
E-mail: †tt23htakaoka@ec.usp.ac.jp, ††{hattori.s,miyagi.s}@e.usp.ac.jp

あらまし　夜間の運転は視界不良により，昼間に比べて事故の危険性が高い．このため多くの運転者は不安を感じやすい．本稿では，夜間路上で撮影された画像から運転時に運転者が抱く不安要素を検出する手法を提案する．提案手法では，YOLOv8 を用いて「車」,「人」,「自転車」を検出し単眼深度推定モデルの MiDaS を組み合わせて不安要因であるかどうかを決定する．また「見通しの悪い場所」についても対象とし，検出されたカーブミラーの画像中での相対的な位置関係から不安要素となる見通しの悪い地点の推定も行う．これらの結果と，運転者へのアンケート調査を照合し提案手法が夜間運転時の不安状態をどの程度客観的に捉えられるかを検証する．

キーワード　画像処理，物体検出，深度推定，感情，夜間運転

1　はじめに

夜間は昼間よりも非常に危険であり，多くの人が夜間運転に対して怖い・不安と感じている．このように感じてしまう理由として，歩行者等の危険となりうる対象の発見が遅れてしまうこと，ドライバーから見えている範囲が狭まってしまうことが原因と考えられている [1].

また，夜間走行時の快適性評価の関連研究として，保田ら [2] が実施した夜間走行性に関するアンケート調査がある．その研究では,「夜間の道路走行は昼間に比べて快適と感じるか」という質問において，快適でないと回答した人が全体の 8 割ほどであった．また，夜間走行時に快適でないと感じる要素を記述する質問では,「見通しが悪い」,「明るさが十分でない」といった回答が全体の 8 割以上を占めていた．つまり視界不良により視認範囲が狭まり，危険となり得る対象や場所の発見および状況予測が困難となり，多くの運転者は細心の注意を払いながら運転しなければならない．この精神的な負担が運転に対する快適性を妨げ，不安感情を誘発されやすくなる．

本稿では，運転者の主観的な不安感情と提案手法がどれくらい一致しているかを検証することを目的として，運転者が不安を抱く要因を検出する手法を提案する．具体的には，夜間道路画像を収集し，その道路画像に含まれる不安となり得る要因を検出する．また要因は不安を誘発する物体と場所に限定する．そこで，物体に関しては「車」,「人」,「自転車」とする．これらの物体を YOLOv8 で検出し，さらに深度推定モデル MiDaS を用いて物体の相対的な奥行きから不安要因かどうかを決定する．また,「カーブミラー」を検出する新たな学習済みモデルを用いて，カーブミラーとの相対的な位置関係から見通しの悪い場所についても推測する．そして，それらの提案手法のパラメータを変化させ，手法の性能がどう変化しどのような結果となるかを調査し，最良となるパラメータの条件を探索することで定量的評価を行った．

本稿の構成として，2 章では関連研究を紹介する．3 章では提案手法を「車」,「人」,「自転車」に対する不安要因の検出方法，カーブミラー周辺の見通しの悪い地点の推定方法の 2 通りに分けて示す．4 章では，アンケート調査および提案手法を用いた実験方法，5 章では，アンケート調査結果をもとに評価実験の結果や考察を記述する．最後に 6 章で本稿のまとめと今後の課題や将来について説明する．

2　関連研究

本稿では関連研究を紹介する．2.1 節では YOLOv8 に関する説明，2.2 節では MiDaS に関する説明，2.3 節では実際に物体検出を用いた先行研究を 2 つ紹介する．その後，本研究の新規性を紹介する．

2.1　YOLOv8 について

YOLOv8 は，2023 年 1 月に Ultralytics 社から発表されたモデルである．Muhammad Hussain [3] は，産業用製造の観点から初代 YOLO から YOLO v8 までの技術的進化に関する詳細なレビューを提供した．Muhammad によると，YOLOv5 や画像サイズが 640 の解像度で学習させた YOLOv6 と同等のパラメータ数で比較した場合，YOLOv8 の方がより高いスループットを出力することが示された．また，この結果からハードウェアの高効率性とアーキテクチャの改良が示されたと述べた．これは YOLOv8 が今までの YOLO シリーズの中でも特に高精度かつ迅速な物体検出ができ，計算資源が限られた制約のあるエッジデバイスへの展開にも可能であることを示唆している．さらに，YOLOv8 には，アンカーフリーという技術が組み込まれている．従来の YOLO では，画像の中の物体を検出する際にあらかじめ用意された予測用のバウンディングボックスと呼ばれるアンカーボックスを用いてその画像と同じサイズのアンカーボックスを選び検出している．アンカーフリーでは，このアンカーボックスの数を減らすことで，全体的な処理スピー

ドを向上できる．

2.2 MiDaS について

Rene Ranftl ら [4] によって提案された MiDaS は，画像上の全画素分の視差を一度に推定（出力）することができるモデルである．また，従来の単眼深度推定の課題であるデータセットバイアスや異なる環境やシーンへのドメイン汎用性の低さという従来のモデルが抱えた課題を克服し，高いロバスト性と汎用性をも実現した．Rene らは，大量かつ多様な複数の深度データセットを統合させ，それらのデータセット間の学習バランスを最適にするため，マルチオブジェクト最適化を採用することで，データセットバイアスを打ち消し，異なるデータセットに対するロバスト性を向上させた．しかし，データセットが多様かつ多数であるために，各データセットの深度のスケール（奥行きの単位）とシフト（カメラの基準点）の曖昧さが目立ってしまい，正確な絶対深度（カメラを基準とした物体の距離）を計ることが困難であった．そこでこれらの曖昧さを解決するために，スケールとシフトの曖昧さを調整し，最適に推測できるように損失関数を導入した．これにより，MiDaS は絶対的な物理距離ではなく，相対的な「視差」を推定するモデルとして構築された．視差は値が大きいほど物体は画像から近いことを表す．この関係性は逆深度に対応したものであり，実際に Rene らの論文でも視差と逆深度の関係性を示している．したがって，MiDaS が推測する出力値はスケールやシフトの影響を受けないデータの数学的解釈では「逆深度」に対応した視差である．本稿では，単眼深度推定モデルとして MiDaS を使用する．

2.3 物体認識を用いた予測に関する研究

徳丸ら [5] は，自転車事故における脇見運転や漫然運転によって引き起こされる前方不注意を防ぐために，物体認識・物体追跡を用いて障害物を認識し，移動先予測を経て衝突を予測するシステムを構築した．徳丸らの研究では，深度カメラに RealSense，物体認識のモデルで YOLO，物体追跡を可能とするモデルに DeepSORT を用いた．この研究の主なシステムは，まず RealSense により目の前の風景の RGB 画像と深度画像を取得する．次に小型 PC である Jeston を用いて YOLO を動作させ，YOLO の物体検出機能で障害物を検出する．その後 DeepSORT により障害物を追跡する．その中で，追跡している障害物がどこに移動するのかを予測する．最後に，障害物の深度（奥行き）および移動先予測に基づき，あらかじめ設定された危険ゾーン内に障害物が存在する場合に警告する．この研究では MiDaS は用いられていないが，Realsense による深度画像と YOLO を用いた障害物認識という点では本稿と類似している．一方で，違いについては，研究対象や手法の違いである．徳丸らは自転車を対象として衝突予測システムを提案し，本稿では自動車，人，自転車を研究対象としている．また徳丸らはカメラを用いてリアルタイムに予測を行うが，本稿では，リアルタイムな処理を前提とせず，1 枚の夜間画像から運転者が不安を感じる物体を検出する．

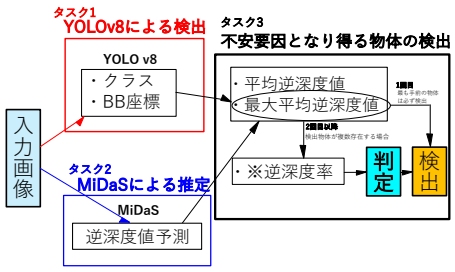
小野ら [6] は，視界が悪い信号のない交差点での安全運転を実現するために，車載カメラ映像からカーブミラー先の死角における交通状況を把握し，危険を予知する研究を行った．カーブミラーの検出器には AutoML，追従には IoUTracking，ミラー内の物体の検出には efficient-d7x を用いた．この研究の主な提案手法は，車載カメラ映像内のミラーを AutoML で検出し，efficient-7dx でミラー内の物体の特徴点を抽出し物体の有無を確認し，IoUTracking で物体を追従して危険の有無を確認した．小野らの研究と比較すると，カーブミラーを検出対象とする点で類似している．一方で，本稿ではカーブミラーを間接的に検出することで運転者が感じる不安要因を推定することができる点で異なる．これらの先行研究から，危険となる物体を検出して衝突や事故のリスクを予測するといった研究は多数存在する．しかし，不安感といった主観的感情と画像を用いた検出を組み合わせた研究は少ない．これを踏まえ，本稿の新規性は，危険となる物体や場所の検出ではなく，不安感という主観的な感情を誘発させる要因を客観的な物体検出で見出す点である．

3 提 案 手 法

本章では，YOLOv8 と MiDaS を用いた「車」,「人」,「自転車」に対する不安要因となり得る物体の検出手法とデータセットで学習させた新たなモデルとカーブミラーの相対的な位置を用いたカーブミラー周辺に存在する見通しの悪い地点の推定手法の 2 通りに分けて記述する．

3.1 「車」,「人」,「自転車」に対する不安要因となり得る物体の検出手法

まず,「車」,「人」,「自転車」に対する不安要因となり得る物体の検出手法の概要図を図 1 に示す．



※逆深度率 = 平均逆深度値/最大平均逆深度値

図 1: 「車」,「人」,「自転車」に対する不安要因となり得る物体の検出手法

次に，YOLOv8 と MiDaS から得られた情報で不安要因を検出するタスクを 3 つに分けて説明する．また，提案手法の最終的な表示例として用いた入力画像のサンプルを図 2 に示す．

3.1.1 タスク 1:YOLOv8 による検出

本稿では，物体検出モデルとして YOLOv8 を用いた．また，画像のみを使用するため，リアルタイム性よりも視界不良による検出精度を最優先に考慮し，精度と負荷のバランスを考えモデルサイズについては 2 番目に大きい m サイズをベースとし

た．実験方法について，まず YOLOv8m から，出力結果から検出した物体のクラスとバウンディングボックス座標（BB 座標）を取得する．実際に YOLOv8m を用いた際の出力結果を図 3 に示す．図 3 では，確信度が 0.94 の車，確信度が 0.90 の人の検出に成功したので，車と人の BB 座標を取得する．

3.1.2 タスク 2:MiDaS による推定

関連研究から単眼深度推定モデル MiDaS の推測する視差とはスケールとシフトの影響を受けない逆深度に対応した相対的な値である．本稿では物体の相対的な距離間（画像からの物体の近さ）を重視するため，本稿では特記がない限り，MiDaS から得られた相対的な値を逆深度値として解釈し扱うこととする．よって提案手法や評価実験では，評価指標として MiDaS の出力値は逆深度値と表現する．そこで，タスク 2 では，1 枚の入力画像から MiDaS を用いて推測を行い，逆深度値を出力する．その際に使う画像は図 2 のような入力画像である．実際に図 2 から MiDaS により推測した際の出力結果を図 4（深度マップ）に示す．この深度マップは予測された出力値（逆深度値）を 0 から 255 の範囲で正規化しグレースケール画像として可視化されたものである．図 4 から，出力値が大きいほど明るい色で示され，値が小さいほど暗い色で示されることとなる．これは「値が大きいほど物体が近くなる」という逆深度の性質を表し，この性質を生かし次節の不安要因の検出を行う．

3.1.3 タスク 3: 不安要因となり得る物体の検出

最後に，タスク 1 とタスク 2 で得られた情報から，サンプル画像中の不安要因検出を行う．

まずタスク 2 で得られた逆深度値と図 3 の車と人の座標 (位置) を使用する．具体的には，それぞれの車や人の BB 座標内に存在するすべての逆深度値を平均化した値を取得する．これを平均逆深度値と呼び，各平均逆深度値を持つ物体の中から最も手前に存在する物体が持つ平均逆深度値を最大平均逆深度値と呼ぶ．ただし「車」,「人」,「自転車」それぞれのクラスで独立して最大逆深度値を持っていることとする．

例えば図 2 なら，同じ走行方向に路上駐車している車と目の前に人が存在している．これらの平均逆深度値をとるとき，値は画像から近ければ近いほど値が大きくなるので，最大平均逆深度値を持つ車は路上駐車した車と目の前にいる人となる．

最後に，最大平均逆深度値と平均逆深度値を用いて検出を行う．具体的に，車を対象として不安要因となり得る物体の検出を行うとする．このとき最大平均逆深度値を持つ車の検出を行う際，最も手前にある物体は不安要因であると仮定したうえで必ず検出するものとする．さらに，その他の同じクラスに所属する物体（車）が存在すれば検出方法は先ほどとは異なり，逆深度率を用いて検出を行う．逆深度率とは，各平均逆深度値を最大平均逆深度値を除した値であり，各物体が最大平均逆深度値をもつ物体からどの程度離れているかを表す．逆深度率が高いほど物体は近いと判定し，低いほど遠いと判定する．さらに，判定する際手法では不安要因となり得る物体の検出に限定するために逆深度率の閾値を設ける．よって逆深度率がその閾値以上であれば検出し，閾値未満であれば検出しないものとする．車の検出が終了し，人や自転車についても検出する必要があれば，次にそれらの検出も行う．今回はサンプル画像には人と車それぞれ 1 人，1 台ずつしか映っていないため，それぞれのクラス別で最大平均逆深度値を用いて検出するだけで終了した．検出のつけ方について，車は赤色，人は青色，自転車は緑色で囲んだ．そのときの最終的な出力結果を図 5 に示す．また，サンプル画像とは異なる画像を用いて，複数の人と複数台の車を検出した際の最終出力結果を図 6 に示す．

図 2: サンプル画像

図 3: 出力結果

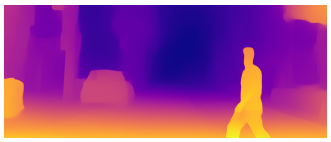


図 4: MiDaS による出力結果（深度マップ）

図 5: 最終出力結果

図 6: 最終出力結果（複数検出）

3.2 カーブミラー周辺の見通しの悪い地点の推定

カーブミラー周辺に存在する見通しの悪い地点（遮蔽領域）の推定手法の概要図を図 7 に示す．図 7 を用いてデータセット作成，学習，推定の 3 つのタスクに分けて説明する．

3.2.1 タスク 1: データセット作成

まず，最初のタスクで既存の物体検出モデルをベースとした新モデルを生成するためのデータセットを作成する．本稿で使用するデータセットの中身はカーブミラーを映した画像データセット 241 枚とそれらの画像を拡張させたデータセットとの合計約 600 枚である．また本稿では，物体検出モデルの学習に用いる画像データセットとして，自身で独自に収集・アノテーションを行った画像群に加え，Roboflow 上で公開されている

他者作成のデータセット [7,8] を併用した．これにより，データの多様性と量を確保し，モデルの汎化性能向上を図った．

3.2.2 タスク 2: 学習

次に，タスク 1 で作成されたデータセットを用いて学習を行う．学習対象はカーブミラーとして，学習用，評価用，検証用のデータセットの比率を 8:1:1 とし，エポック数は 50，バッチは 16，画像サイズは 1280 × 1280 と設定する．

3.2.3 タスク 3: 見通しの悪い地点の推定

最後に，タスク 2 で学習させた新モデルを使いカーブミラー周辺に存在する見通しの悪い地点の推定を行う．

まず，入力画像の中心点を取得する．なお，本稿では画像の原点を左上と定義しているため，下方向へ移動するほど y 座標の値は大きくなるとしている．次に，画像内における真ん中の範囲を中央領域として定義する．定義方法として，画像の左端を基準（0%）としたうえで右方向への水平移動距離を割合で指定した際に L（左境界）と R（右境界）により囲まれた領域を設定する．これが中央領域である．そして，カーブミラーの BB 座標 $(x_1, y_1), (x_2, y_2)$ に基づき，BB 座標の左右端である x_1 と x_2 が同時に中央領域内に含まれたときの推定方法と x_1 と x_2 のいづれかおよび両方が中央領域外に外れたときの推定方法の 2 通りに分けて説明する．x_1 と x_2 の両方が中央領域内に存在する場合，カーブミラーは画像中央付近に位置しているとみなし，画像の左右両端の場所を見通しの悪い場所として推定する．具体的には，画像幅を W，画像高さを H としたとき，左右それぞれの推定位置の中心座標を $(0.9W, 0.6H)$（左寄り）および $(0.1W, 0.6H)$（右寄り）とし，BB 座標の横（$x2 - x1$）の長さの 3 倍を半径とした円で囲む．円の中心の y 座標について，円の長さを考慮すると路面付近の死角となる場所を重点的に推定することができるため，$0.6H$ と設定した．

一方，x_1 または x_2 のいずれかもしくは両方の座標が中央領域外に存在する場合は，カーブミラーの位置関係に基づいて見通しの悪い場所を推定する．具体的には，BB 内の中心座標を取得する．そして BB 内の中心点の y 座標が画像中心点の y 座標より小さい場合，取得した座標を用いて画像中心点を基準とした点対称移動を行う．逆に，画像中心点の y 座標以上である場合には，画像中心点の x 座標を軸とした線対称移動を行う．これらの対称移動によって得られた座標を，見通しの悪い場所の中心座標として設定する．その際，見通しの悪い場所の座標を中心点とし BB 座標の横（$x2 - x1$）の長さの 3 倍を半径とした円で囲む．また円のサイズについて，3 倍未満の場合画像によっては推定範囲が小さくなる可能性があり，3 倍より大きくなると推定範囲が大きすぎて視覚的に見にくくなる可能性があるため 3 倍とした．

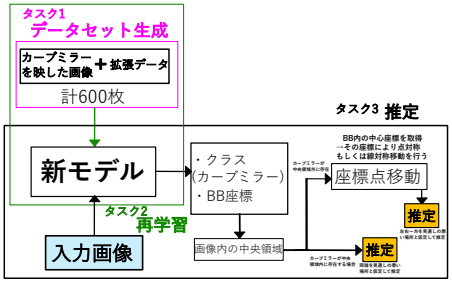


図 7: カーブミラー周辺の見通しの悪い地点の推定手法

4 評価実験

本章では，4.1 節で実際に人が不安だと感じる要因を明らかにすることを目的に実施したアンケート調査の概要とその結果，さらに 4.2 節では人が実際に不安だと感じた要因を提案手法がどの程度捉えることができるかを検証し，そのうえでパラメータの最良条件を求めるために実施した評価実験を説明する．

4.1 アンケート調査

まず，アンケート調査の概要と実際の調査結果を示す．

4.1.1 目　　的

夜間に運転をする人が夜間運転を想定して画像からどのような要因を不安に感じるかについて明らかにし，提案手法の正解データセットを作成させるための土台を作ることを目的とする．

4.1.2 概　　要

次に，アンケート調査の概要について説明する．今回の調査では，運転者を対象とした研究であるため，運転免許証を所持した大学生および社会人の計 25 名を対象とした．次に，調査方法について説明する．25 名の協力者に 17 枚の夜間路上で撮影された画像を提示し，不安に感じる場所や物体を指摘し，マークをしていただいた．その後回答済みの結果を収集し，マークされた「車」,「人」,「自転車」,「見通しの悪い地点」に限定し，番号を振り分け，それぞれの番号で分けられた対象の不安認識率（アンケート回答者のうち指摘した人の割合）を求めた．実際にアンケート調査で用いた画像を図 8 に示す．

4.1.3 調 査 結 果

アンケートの調査結果を図 9 に示す．

実際に分析した調査結果の集計を表 1 に示す．

4.2 実 験 方 法

本稿では 2 種類の評価実験方法を説明する．

4.2.1 評価実験 1

評価実験 1 では,「車」,「人」,「自転車」に対する不安要因となり得る物体の検出手法の評価実験を行う．扱う画像は図 8 の画像 1〜11 である．また,「不安認識率の閾値」，パラメータとして「逆深度率の閾値」の 2 種類を使用し，これらの閾値を組み合わせて評価する．

- 不安認識率の閾値：100 %，75 %，50 %の 3 通り
- 逆深度率の閾値：40 %，50 %，60 %の 3 通り

表 1: アンケート調査結果

画像番号	項目番号	参加人数 (人)	指摘人数 (人)	認識率 (%)
画像 1	1	25	25	100
	2	25	2	8.00
画像 2	3	25	5	20.0
	4	25	13	52.0
	5	25	7	28.0
	6	25	10	40.0
	7	25	5	20.0
画像 3	8	25	24	96.0
	9	25	2	8.00
	10	25	8	32.0
画像 4	11	25	12	48.0
	12	25	22	88.0
画像 5	13	25	21	84.0
	14	25	18	72.0
	15	25	9	36.0
画像 6	16	25	23	92.0
	17	25	3	12.0
	18	25	11	44.0
画像 7	19	25	25	100
画像 8	20	25	17	68.0
	21	25	20	80.0
	22	25	5	20.0
画像 9	23	25	7	28.0
	24	25	22	88.0
	25	25	2	8.00
	26	25	21	84.0
	27	25	5	20.0
画像 10	28	25	2	8.00
	29	25	24	96.0
	30	25	2	8.00
画像 11	31	25	25	100
画像 12	32	25	18	72.0
	33	25	19	76.0
画像 13	34	25	4	16.0
	35	25	21	84.0
画像 14	36	25	14	56.0
	37	25	16	64.0
画像 15	38	25	22	88.0
	39	25	3	12.0
画像 16	40	25	16	64.0
	41	25	19	76.0
画像 17	42	25	12	48.0
	43	25	23	92.0

図 8: アンケート調査で用いた画像群

計 9 通りの 2 種類の閾値の組み合わせを作成し，各条件下での検出性能を評価する．評価を行うにあたり，アンケート調査結果と実験結果を照合して以下の 5 種類の指標を用いる．

- 検出成功率 (Recall / 再現率)：不安認識率が閾値以上の要因を正しく検出できた割合.
- 検出失敗率：不安認識率が閾値以上の要因を検出できなかった割合.
- 過剰検出率：不安認識率が閾値未満，または対象外の要因を過剰に検出した割合.
- 適合率 (Precision)：システムが検出した全要素のうち，実際に不安認識率が閾値以上であった要因の割合.
- F1 値：適合率と再現率により検出性能を総合的に評価する指標.

これらの指標を用いて，検出性能の変化を調べ，各不安認識率の閾値に対して最も検出性能が良い時の逆深度率の閾値を求める．さらに，全 9 通りの中から最も性能が高いときの組み合わせを決定し，その組み合わせで固定された際の画像 1 枚ごとの F1 値を算出する．

4.2.2 評価実験 2

評価実験 2 では，カーブミラー周辺に存在する見通しの悪い地点の推定手法についての評価実験を行う．扱う画像は図 8 の画像 12〜17 である．また,「不安認識率の閾値」，パラメータとして「中央領域の閾値」の 2 種類を使用し，これらの閾値を組み合わせて評価する．

図 9: 調査結果

- 中央領域の閾値 (L, R)：(35, 65)，(40, 60)，(45, 55)，(50 , 50) の 4 パターン
- 不安認識率の閾値：100 %，75 %，50 %の 3 通り

計 12 通りの 2 種類の閾値の組み合わせを作成し，各条件下での推定性能を評価するまた，評価を行うにあたりアンケート調査結果と実験結果を照合して 5 種類の指標を用いる．

- 推定成功率 (Recall / 再現率)：不安認識率が閾値以上の地点を正しく推定できた割合．
- 推定失敗率：不安認識率が閾値以上の地点を推定できなかった割合．
- 過剰推定率：不安認識率が閾値未満，または対象外の地点を過剰に推定した割合．
- 適合率 (Precision)：システムが推定した全要素のうち，実際に不安認識率が閾値以上であった地点の割合．
- F1 値：適合率と再現率により推定性能を総合的に評価する指標．

これらの指標を用いて，推定性能の変化を調べ，各不安認識率の閾値に対して最も推定性能が良いときの中央領域の閾値を求める．さらに，全 12 通りの中から最も性能が高い時の閾値の組み合わせを決定し，その組み合わせで固定された際の画像 1 枚ごとの F1 値を算出する．

ここで評価実験 1，2 に共通して，不安認識率が閾値以上である要素を検出した回数 TP，不安認識率が閾値未満であるまたは振り分けられた番号以外の要素を検出した回数 FP，不安認識率が閾値以上である要素を検出できなかった回数 FN と定義し，適合率と再現率は式 (1)，(2) で表される．

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

さらに，F1 値は式 (3) で表される．

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{3}$$

また本稿では，不安要因でない物体（アンケートで指摘されたが閾値未満の不安認識率を持つ物体と指摘されなかったが実際に検出された物体）を正解データと定義しないこととする．

5 実験結果と考察

本章では評価実験 1，2 で得られた実験結果と考察を示す．

5.1 評価実験 1

5.1.1 実 験 結 果

逆深度率と不安認識率の閾値の組み合わせを変えたときの性能結果を表 2，逆深度率の閾値を変化させた際の出力結果を図 10 に示す．表 2 の実験結果から，100 %，75 %，50 %のときのすべての不安認識率の閾値に対して逆深度率を変化させた際，100 %のときは検出成功率は 100 %のまま変化しなかったが，75 %と 50 %のときは検出成功率が下がることが確認できた．同時に過剰検出率と F1 値については，過剰検出率は逆深度率を上げると減少し，F1 値は増加することが示された．この結果から，100 %，75 %，50 %のときのどの不安認識率に対しても，パラメータの最良条件は 60 %であることが確認できた．さらに表 2 から，不安認識率と逆深度率をセットした全 9 通りの組み合わせの中から，最も性能が良いとき（検出成功率が高く，過剰検出率を抑え，さらに F1 値が最大のとき）の閾値の組み合わせは不安認識率が 50 %，逆深度率が 60 %のときであることが示された．そこで，不安認識率を 50 %，逆深度率を 60 %で固定し，画像 1 から 11 までのそれぞれの画像に対する適合率，再現率，F1 値を算出した．その結果を表 3 に示す．表 3 の結果から，画像 1，7，11 に関しては値は 1.00 となり，逆に画像 2，9 に関しては F1 値が顕著に低くなることが確認できた．表 3 から，適合率の低さが F1 値を大きく下げる原因となった．

5.1.2 考　　察

表 2 の結果から，それぞれの不安認識率ごとに結果を見てみると，逆深度率を上げるほど検出成功率は下がるが過剰検出率も下がることが分かり，F1 値は増加することが確認できた．また逆深度率が 60 %のときがパラメータの最良条件であることも示された．このような結果から，どの不安認識率の閾値に対しても，画像から距離が近い物体に絞ったときの検出方法が検出性能が良くなると考えられる．しかし，どの不安認識率の閾値でも最良条件を含め逆深度率を変化させたところで，過剰検出率は依然として高いことが確認できた．原因としては，モデ

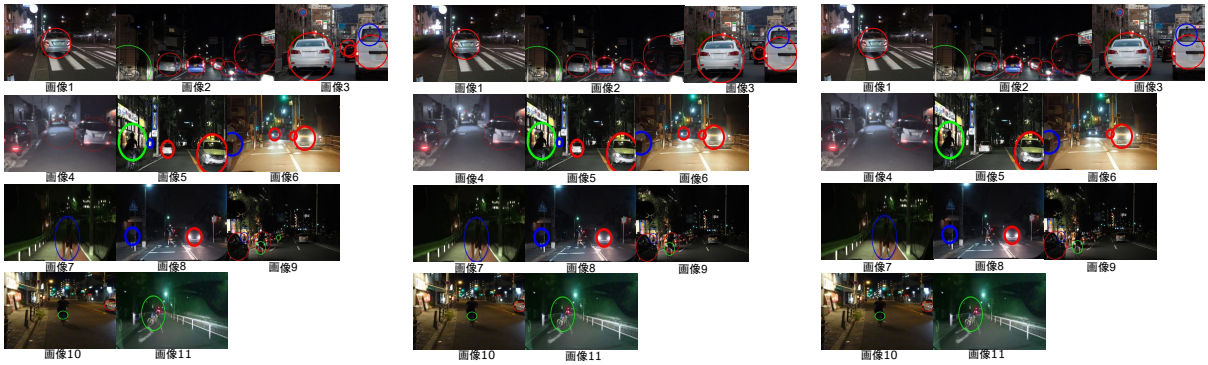


図 10: 逆深度率変化に伴う検出結果の比較（左：40%，中央：50%，右：60%）

図 11: 中央領域変化に伴う推定結果（左上：(35, 65)，右上：(40, 60)，左下：(45, 55)，右下：(50, 50)）

表 2: 実験結果（評価実験 1）

不安認識率 (%)	逆深度率 (%)	検出成功率（再現率）	検出失敗率 (%)	過剰検出率 (%)	適合率 (%)	F1 値
100	40	100	0	89	9	0.16
	50	100	0	85	10	0.18
	60	100	0	72	12	0.21
75	40	91	9	89	28	0.43
	50	82	18	88	29	0.43
	60	82	18	71	35	0.49
50	40	93	7	88	37	0.53
	50	86	14	86	39	0.53
	60	79	21	71	42	0.55

表 3: 画像 1〜11 に対する適合率，再現率，F1 値の結果

画像番号	適合率 (%)	再現率 (%)	F1 値
画像 1	100	100	1.00
画像 2	17	100	0.29
画像 3	33	100	0.50
画像 4	50	100	0.67
画像 5	50	50	0.50
画像 6	33	100	0.50
画像 7	100	100	1.00
画像 8	50	50	0.50
画像 9	33	50	0.40
画像 10	50	100	0.67
画像 11	100	100	1.00

ルが不必要な物体（不安認識率が閾値未満の物体と番号以外の物体）を多数検出してしまっているためであると考える．実際に図 10 の検出結果から，画像 2 ではほとんどの物体は不安認識率が 50 %未満なのに関わらず検出し，番号が振り分けられていない右の自転車や画像 3 における右の白い車に乗車している人を（a），(b)，(c）の全パターンで検出した．このような過剰に検出した結果が検出性能を阻害する原因となったと推察する．また表 3 から，図 10 の画像 1，7，11 のような情報量が少ないもしくは視覚的占有率が強い画像の場合，不安要因の候補となる物体を比較的容易に絞り込むことができるため，F1 値が高くなる一方，逆に画像 2，9 のような情報量が多い画像の場合は不安要因の候補となる物体が 1 枚の画像に多数存在するため，手法が不安要因となる物体を絞り込むことが難しく過剰に検出するケースが多くなる．そのため，F1 値が低くなると考える．したがって，これらの結果および考察から，本稿の提案手法は不安要因でない物体の過剰検出や情報量が多く複雑な画像に対する検出精度の低さが課題であり，これらの課題に対する対策を打つ必要がある．具体的な対策として，最大平均逆

深度値もしくは逆深度率から検出された物体について，その物体が真の不安要因であるかどうかを判断する機械学習により生成されたモデルを利用する．具体的な学習としては，アンケート調査で得られた各物体の「不安認識率」をラベルとし，その物体の種類や位置，状況そして距離感をデータ化したものを特徴量としてデータセットで扱う．つまり，提案手法により不安要因となり得る物体を 1 度検出し，さらに学習済みのモデルにより真の不安要因かどうかを決める 2 段階検出で検出精度を上げることが可能であると考える．また今回は「車」,「人」,「自転車」のすべてを検出するような手法だったため，夜間でも比較的見やすい物体でそこまで不安要因にならない物体までも検出していたことが実験結果で明確になった．したがって他の対策として，例えば「黒色の服を着た人」や「黒い車」のような夜間では発見しづらい物体に限定して検出するといった手法を用いて比較実験を行うことで検出精度が改善する傾向がみられると考える．さらに，今回は不安認識率の閾値の下限を 50 %に，逆深度率の閾値の上限は 60 %に限定して実験を行ったため，それぞれの閾値を 50 %未満及び 60 %以上にしたときの結果についても考察する．実際に不安認識率の閾値を 50 %未満に下げると，それまで不安要因でなかった物体が正解データとして扱われるため，そのデータ数が増える．つまり，不安認識率の閾値を 50 %未満に下げたときに TP が大きくなる．一方で，逆深度率の閾値を 60 %以上に引き上げることで過剰検出率を下げることが期待できる．これらの根拠から，不安認識率及び逆深度率の閾値を変えた場合，検出成功率や過剰検出率，さらに F1 値といった各検出性能，パラメータの最良条件が変化すると考えられる．

5.2 評価実験 2

5.2.1 実 験 結 果

中央領域と不安認識率の閾値の組み合わせを変えたときの性能結果を表 4，中央領域を変化させた際の出力結果を図 11 に示す．表 4 から，不安認識率の閾値が 100 %の場合は，すべての画像において不安認識率 100 %の場所が存在しなかったため中央領域を変化させても推定成功率が 0 %であることが確認できた．そのため，最良条件といえるパラメータは存在しない．次に不安認識率の閾値が 75 %と 50 %のとき，中央領域の範囲を広めると，過剰検出率は増加するが値は抑えつつ，推定成功率や F1 値は増えていくことが確認できた．また F1 値については，不安認識率が 75 %のときの最大値は 0.77，50 %のときの最大値 0.83 であるため，不安認識率の閾値が 75 %，50 %のときの最良条件となるパラメータは $(L, R) = (35, 65)$ のときであると示された．

そして，不安認識率の閾値と中央領域の閾値の全 12 通りの組み合わせの中で，F1 値が最大であり，過剰推定率が低く，推定成功率を高く維持している組み合わせは不安認識率が 50 %，$(L, R) = (35, 65)$ のときであった．そこでこれらの組み合わせを用いて，画像 12〜17 までのそれぞれの画像に対する適合率，再現率，F1 値を算出した．その結果を表 5 に示す．表 5 から，それぞれの画像に対して，F1 値は高い場合が多いことが示された．

表 4: 実験結果（評価実験 2）

不安認識率 (%)	中央領域 (L,R)	推定成功率（再現率）	推定失敗率 (%)	過剰推定率 (%)	適合率 (%)	F1 値
100	35,65	0	0	67	0	0.00
	40,60	0	0	58	0	0.00
	45,55	0	0	50	0	0.00
	50,50	0	0	33	0	0.00
75	35,65	100	0	43	63	0.77
	40,60	80	20	43	57	0.67
	45,55	80	20	29	67	0.73
	50,50	60	40	33	60	0.55
50	35,65	78	22	33	88	0.83
	40,60	67	33	33	86	0.75
	45,55	67	22	0	100	0.80
	50,50	44	56	40	67	0.53

表 5: 画像 12〜17 に対する適合率，再現率，F1 値の結果

画像番号	適合率 (%)	再現率 (%)	F1 値
画像 12	100	50	0.67
画像 13	100	100	1.00
画像 14	100	50	0.67
画像 15	100	100	1.00
画像 16	100	100	1.00
画像 17	50	100	0.67

5.2.2 考　　察

提案手法では，運転者が不安に感じている見通しの悪い場所を比較的推定できたと考えられる．その中で $(L, R) = (50, 50)$ は中央領域を設定しないのと同義である．中央領域の有無に関して比較すると，領域なしより領域ありのほうが過剰検出率を抑えつつ，推定成功率と F1 値を比較的高く維持できる．したがって，100 %以外のどの不安認識率の閾値に対しても，中央領域は範囲を広く設定することで性能が良くなると考えられる．実際に図 11 の画像 16，17 に関して，他の中央領域（(a)，(b)，(c)）では両端もしくはいづれかの端の場所を推定したのに対し，(d) の時の中央領域では失敗しており，何もない路上を推定したことが分かる．これが推定成功率を下げ，失敗率を上げている原因であると考える．評価実験 2 に対して，75 %，50 %の時点で F1 値が比較的高いため，不安認識率の閾値を 50 %未満に下げた際，正解データは少し増えると考えられるため，F1 値は微増すると考える．

次に表 5 から，適合率と再現率が所々50 %と低い箇所が見られた．提案手法はカーブミラーが左右どちらかに存在する場合その反対側の見通しの悪い場所のみしか推定できない．このような提案手法の限界により，図 9 の画像 12，13，14，15 の番号 32，34，36，39 のようなカーブミラーの位置関係に依存しない場所が推定できず，適合率または再現率を下げる原因となったと考える．つまり，カーブミラーの位置関係に依存した場所しか推定できないことが評価実験 2 の結果で得られた課題れたため大きな課題点は特に見られなかった．しかし，所々適合率や再現率が 50 %になるところは確認できた．

である．この課題について，カーブミラーの位置関係に依存しない場所を推定することができれば手法の性能や F1 値は上がると考える．しかし，それぞれの場所がもつ不安認識率は異なり定められた閾値により不安要因かどうかが変わるため，この変動性にもしっかり対応した手法を考えなければならない．具体的な対策として，セグメンテーションと機械学習により生成されるモデルの 2 段階判定を用いることで性能が高まると考えられる．例えば，表 1 の画像番号 12 から項目番号 32 については，不安認識率が 72 %と比較的不安認識率が高い場所を現在の手法では推定できない．そのため，今の手法では推定できない場所についてはセグメンテーションにより抽出し，その抽出した位置が真の不安要因かどうかを学習済みモデルで判定することで課題を克服できると考える．具体的に，見通しの悪い場所に関して運転者がその場所に関してどう感じており，どういう影響を及ぼすのかといった特徴量をデータセットして機械学習させ，モデルを生成する．そのモデルの判断で最終的な不安要因となる見通しの悪い場所のみの検出を行う．

6 おわりに

本稿では,「車」,「人」,「自転車」に対する不安要因となり得る物体の検出とカーブミラーの相対的な位置を利用した見通しの悪い地点の推定手法を提案した．これらの手法を用いて，不安認識率の閾値ごとの性能評価を行いながらパラメータの最良条件を探索した．さらに，最良な不安認識率の閾値とパラメータの組み合わせを一組選び，画像別に手法の精度についても検証を行った．まず，物体の検出手法については，どの閾値でもパラメータ（逆深度率）を上げるほど徐々に検出性能が上がることが確認できた．よって，画像から近い物体に絞り検出する方法が検出性能が高くなることが示された．しかし，画像別の精度に関しては複雑な画像ほど適合率が悪くなり結果として F1 値が悪化することから，過剰検出が多いことが本手法の課題となった．次に，見通しの悪い地点の推定手法については，画像の中央領域を設定するほど推定性能が向上することが確認できた．しかし，本稿の推定方法はカーブミラーの位置関係を利用した推定のみに留まっており，位置関係に依存しない地点やカーブミラーが近くにない地点の推定は困難であるため今後の課題として残ることとなった．

したがって,「車」,「人」,「自転車」に対する不安要因となり得る物体の検出手法，カーブミラーとの相対的な位置関係を利用した見通しの悪い場所の推定手法は運転者が感じる不安要因を検出することが可能であると示された．今後の課題として，不安認識率の閾値やパラメータの範囲を変化させ，アンケート調査の規模を拡大させたときの性能変化やパラメータの最良条件の探索を検討する．特にアンケート調査に関しては，画像から不安要因となる物体に丸を付けるだけの作業だったため，今後は追加で物体に丸を付けた理由も回答してもらい，5 段階評価にして不安の度合いをもっと細かく明確にしていく．これにより物体の特徴量を抽出することができ，物体が不安要因かの有無に対して高性能なモデルを生成できると考える．そして「車」,「人」,「自転車」を対象とした検出手法では，単眼深度推定モデルは MiDaS を用いたが，ほかにも Deep-Anything, Deep Pro などが存在する．そこで MiDaS 以外のモデルも採用し，実際に評価実験を行い，それぞれのモデルを用いることで提案手法の性能変化がどのようになるのかという比較実験も検討していく．

さらに,「車」,「人」,「自転車」に関する不安要因の検出手法はさらなる性能の改善が必要であると考えた．具体的に，検出方法を「最大平均逆深度値」・「逆深度率」および機械学習により生成された真の不安要因を決めるモデルで 2 段階検出をすることで提案手法の課題を解決し，より信頼性のある検出手法に改善していく．またその他に「黒色の服を着た人」や「黒い車」のように夜間では見にくくなる物体に限定して検出する手法を提案し，本稿の提案手法と比較実験を行い検出性能の改善見込みがあるかを評価していく．また，カーブミラーの位置関係を利用した推定手法については，カーブミラーの位置関係に依存しない場所をセグメンテーションにより抽出し，学習済みモデルで 2 段階判定をする手法も検討していく予定である．最後に，本稿の成果を社会に活用する具体的な技術的活用として，実際の走行環境で不安要素を検出し，その要素に対して運転者に適切な注意喚起を行う支援システムが構築可能となる．このシステムにより運転者の不安・怖いといった精神的な負担を軽減できると考えられる．このような技術的活用は，夜間走行に不安を抱く人や運転に慣れていないペーパードライバー等の夜間運転に対する苦手意識の克服に寄与する．その結果，ドライバー人口の維持や増加が期待でき，ひいては自動車産業の持続的な発展も期待できる．

文　献

[1] CLUT, “夜の運転が怖いあなたへ！夜の危険を回避する安全運転の 4 つのポイントを解説,” https://clutch-s.jp/p000833/ (参照 2025-11-23).
[2] 保田 敬一, 白木 渡, 井面 仁志, “夜間車両走行時の快適性評価,” 特集「道路走行時の快適性に影響する要因と快適性向上策」, Vol.22, No.3, pp.138-143 (2024).
[3] Muhammad Hussain, “YOLO-v1 to YOLO-v8, the Rise of YOLO and Its Complementary Nature toward Digital Manufacturing and Industrial Defect Detection,” Machines, Vol.11, No.7, 677 (2023).
[4] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, Vladlen Koltun, “Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer,” arXiv:1907.01341 (2020).
[5] 徳丸 拓希, 山口 暢彦, 福田 修, 奥村 浩, “物体認識を用いた自転車運転時の衝突予測システム,” 日本知能情報ファジィ学会 ファジィシステムシンポジウム 講演論文集, Vol.37, No.0, pp.211-214 (2021).
[6] 小野 晋太郎, 日野 裕介, 須田 義大, 板垣 紀章, “走行中の車載カメラとカーブミラーによる死角の危険予知,” 生産研究（Journal of Institute of Industrial Science, University of Tokyo, Vol.74, No.1, pp.123-128 (2022).
[7] iniad, curve-mirror_20241129 Computer Vision Model, Roboflow, https://universe.roboflow.com/iniad/curve-mirror_20241129 (参照 2025-10-15).
[8] Murata Eri, mirror Computer Vision Dataset, Roboflow, https://universe.roboflow.com/murata-eri/mirror-yl9fz (参照 2025-10-15).

学習済みモデルの特徴ベクトルに基づく 未知個体への対応を考慮した地域猫の個体分類

永尾 浩太† 服部 峻†† 宮城 茂幸††

† 滋賀県立大学大学院工学研究科電子システム工学専攻 〒 522–8533 滋賀県彦根市八坂町 2500
†† 滋賀県立大学先端工学研究院 〒 522–8533 滋賀県彦根市八坂町 2500
E-mail: †to23knagao@ec.usp.ac.jp, ††{hattori.s,miyagi.s}@e.usp.ac.jp

あらまし 不幸な猫の発生を防ぐ取り組みとして，地域猫活動が行われている．活動を円滑に進めるためには各個体を正確に把握する必要があるが，首輪やタグの装着が難しい場合も多い．そこで地域住民が日常的に撮影する画像を用いて個体識別ができれば，負担の少ない有効な把握手法となる．機械学習による猫の個体識別の研究の多くは学習時に想定した既知個体のみを識別の対象としている．しかし，実際の地域環境においては未知個体が存在する可能性を考慮する必要がある．本稿では，地域猫を対象とし，既知個体の識別に加えて未知個体の判別を同時に行うことを目指す．まず既知個体を学習したモデルを作成し，それを特徴抽出器として用いることにより既知個体画像および未知個体画像の特徴ベクトルを抽出する．その後，特徴ベクトル間の類似度を算出し，設定した閾値により既知個体と未知個体の分類を行う．

キーワード 地域猫，機械学習，個体識別，分類，オープンセット認識

1 はじめに

現代社会において，引き取り手が見つからず行き場がなくなり殺処分されてしまう猫が存在する．環境省の統計 [1] によると，日本国内における猫の殺処分数は平成 16 年度には 238,929 匹，令和 5 年度には 6,899 匹が殺処分されており，年々減っているものの依然として多くの命が失われている．

そのような不幸な猫を生まない・減らすための活動として地域猫活動が行われている．地域猫活動とは，地域住民が主体となって不妊去勢手術（TNR）や適切な餌やり，糞尿の管理，見守りなどを行うことで，飼い主のいない猫を増やさないことや，人と猫が共生できる社会をつくることを目的とする活動である．一方で猫によるトラブルも発生しており，地域猫活動を快く思わない住民もいる．地域猫活動を円滑に進めることや，地域住民の理解を得るためには猫の適切な管理が必要である．

猫の管理には各個体の把握が重要であるが，首輪やタグのなどの物理的な装置の装着が難しい場合もある．そこで，地域住民がスマートフォン等のカメラを用いて日常的に撮影するような画像を用いて個体識別ができれば，負担の少ない有効な管理手法となる．

野生の猫や飼い猫を対象とした画像を用いた機械学習による個体識別の研究は行われているが，多くは学習時に想定した個体のみを対象として再識別を行なっており，学習していない猫への対応は十分に検討されていない．地域においてはこのような未知の猫がいるという可能性が考えられる．

そこで本研究では，地域猫を対象として猫の個体分類を行い，既知の猫の個体識別を行うと同時に，未知の猫を未知として判別することを目指す．アプローチとしては，学習済みモデルによって抽出された特徴ベクトル間の類似度に基づいて分類を行う．具体的には，まず初めに既知の猫についての学習済みモデルを作成する．次にそのモデルを特徴抽出器として用いて，学習で使用した画像の特徴ベクトルを得た後，各個体を表すベクトルとして個体ごとに平均ベクトルを求める．識別対象の画像も同様に，特徴抽出器を用いて特徴ベクトルを求める．そして，そのベクトル間で類似度を求める．類似度について閾値を設定し，それにより既知個体と未知個体の分類を行う．なお，対象とする画像は，スマートフォン等のカメラで撮った画像に対し物体検出モデルを用いて猫を検出したものとし，可能ならば背景の削除を行ったものを想定とする．

以下，本稿の構成について述べる．2 章では本研究に関連する研究や技術について述べる．次に，3 章で提案手法の詳細について述べる．そして，4 章において提案手法の評価実験を行い，最後に 5 章で結論および本研究の今後の展望について述べる．

2 関連研究・技術

本章では，本研究の関連研究と関連技術について述べる．

2.1 関連研究

Trein らの研究 [2] では，中国の都市部に生息する野良猫を対象とした個体識別の自動化を目的に，Siamese Networks を用いて個体識別を行った．Siamese Networks は，2 つの同一のネットワークを用いて入力データの類似性を学習するネットワークで，この研究では EfficientNetB0, MobileNet, VGG16 などをバックボーンとして利用し比較検討している．またライブ配信映像から個体画像を抽出し，猫の正面と上部の 2 つの視点を含むデータセットを構築している．

Li らの研究 [3] では，ペットショップでの猫の管理や野生の猫

の監視の改善のため，深層学習を用いた猫の個体識別における様々なニューラルネットワークモデルの比較を行った．この研究ではImageNet学習済みのResNetや，DenseNet, EfficientNet, ConvNeXt, Siamese Networksといったモデルに対して体系的に比較しており，転移学習を行ったモデルとの比較も行っている．

Yangらの研究[4]では，オーストラリアの外来種である野生猫のモニタリングや生態系保護のため，YOLOv5を用いてカメラトラップ画像から得られた不均衡なデータを使用して個体識別を行った．この研究では個体ごとの画像が不均衡な問題に対し，データ増強や転移学習を用いることでモデルの汎化性を高めている．

Caquilpanの研究[5]では，不規則な姿勢や向きの変化などが伴うカメラトラップ画像において，多様な条件下においても野良猫を正確に識別するため姿勢情報を活用し，猫の体の部位パーツに基づいて個体識別を行った．この研究では，Liuらの研究[6]で提案されたアムールトラの再識別のためのアムールトラの身体パーツに基づいて作られたPPGNetと呼ばれるネットワークを参考にしており，猫の画像の特徴に合わせて改良したPPGNet-Catという猫の再識別のためのネットワークモデルを提案している．

これらの研究は，猫の個体識別のため様々なニューラルネットワークモデルの比較や検証，追求などを行っている．しかしこれらの研究を含めた多くの研究は，モデルに対して学習を行った猫のみを想定しており，学習を行っていない猫への対応は十分に考慮できていない．本研究では学習を行った既知の猫に加え，学習を行っていない未知の猫にも対応するため，未知の猫を既知の猫に分類せず未知として判別する．

2.2 関連技術

2.2.1 YOLO

YOLO (You Only Look Once) [8] はRedmonらによって提案された，リアルタイム性に優れた物体検出モデルである．YOLOは画像全体に対して，物体の検出とクラス予想を同時に行うことができるため高速に動作する．Ultralytics社[7]がライブラリとして提供しており，YOLOv5，YOLOv8，YOLO11は同社が開発したバージョンである．本研究では，この中で最も新しく性能が向上されたモデルであるYOLO11を，猫の検出やセグメンテーションによる背景削除などのデータ作成の際に用いた．

2.2.2 ResNet

ResNet [9] はHeらによって提案された，深い層を持つCNN（畳み込みニューラルネットワーク）の一種である．従来のCNNモデルが抱えていた，層を深くしても学習が進まなくなる勾配消失問題を解決することで，ネットワークの深層化を可能にした．ResNetは標準的なCNNモデルで，コンピュータビジョンのタスクに広く使用されており，他のモデルや技術と比較する際によく用いられる．また，ResNetには層の数に応じてResNet-18, 34, 50, 101, 152など様々なバリエーションが存在する．本研究では学習のためのCNNとして速度と精度のバランスのとれたResNet-50を用いた．

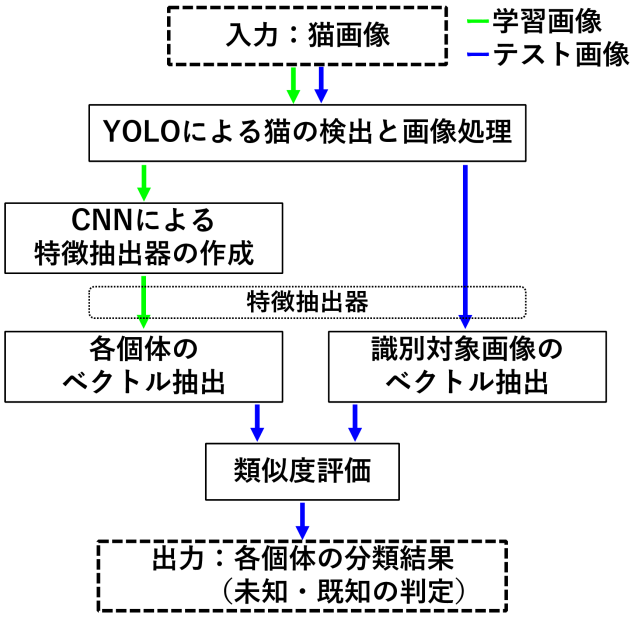


図1　提案手法のシステム概略図

2.2.3 オープンセット認識

オープンセット認識 (OSR: Open Set Recognition) は機械学習において，モデルが学習時に与えられた既知クラスを正しく分類するとともに，学習時には存在しなかった未知クラスのデータを「未知」と検知し識別・拒否することを目的とする．従来のクローズドな集合 (closed-set) では，推論時に与えられるデータは全て既知クラスに属すると仮定されるが，オープンセット認識では未知のクラスが混在する状況を想定する．

未知クラスを検知する手法の1つとして，閾値を設定し既知クラスに対する予測確率や信頼度が，その閾値以下の場合未知と判断する方法が挙げられる．OpenMax [10] はBendaleらによって提案された手法で，未知クラスへの対応を目的として，ネットワークの最終層のSoftMaxを拡張している．Weibull分布を用いた処理を行い，入力を単に既知クラスの中で予測するのではなく，未知クラスに属する可能性を考慮して評価を行う．

3 提案手法

本章では，本研究の提案手法について述べる．図1にシステムの流れを示す．システムは主に，猫の検出と画像処理，特徴抽出器の作成，画像のベクトル抽出，類似度評価の段階に分かれている．各節でその詳細について述べる．

3.1 猫の検出と画像処理

入力された猫画像に対して，物体検出モデルYOLO11を用いてバウンディングボックスを作成し猫の検出を行う．検出後，バウンディングボックスが画像の中心に来るように配置．背景情報を少しでも減らすためバウンディングボックス部分以外を黒くなるようにし，画像全体を正方形に加工する．その後，可能な画像は更に背景の影響を減らすため，セグメンテーションにより猫以外の部分が黒くなるように加工する．図2にその処

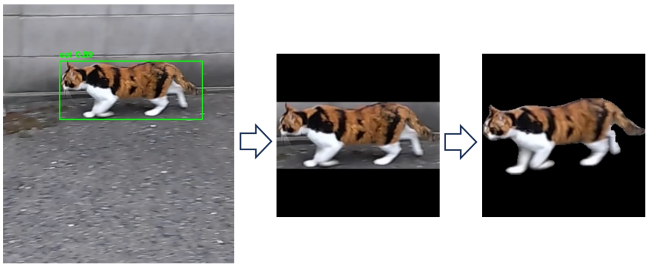


図 2　猫の検出と画像処理の例

理例を示す.

3.2　特徴抽出器の作成

各猫画像をその特徴を捉えたベクトルへと変換するため，既知の猫とした画像を用いて CNN を学習させ特徴抽出器を作成する．なお，未知の猫とする画像は CNN には学習させずテスト時にのみ用いる．ここで学習を行った CNN は直接画像を分類するためには用いず，特徴抽出のために用いる．

3.3　画像のベクトル抽出

3.2 節で作成した特徴抽出器を用いて，各個体のベクトル抽出と識別対象画像のベクトル抽出を行う．各個体のベクトル抽出では，まず初めに学習に用いた全画像を特徴抽出器に再度入力し画像のベクトル化を行う．次に各個体ごとに特徴ベクトルの平均を求め，そのベクトルを各個体の代表ベクトルとする．識別対象画像も同様に，識別対象としたい画像を作成した特徴抽出器に 1 度入力し画像のベクトル化を行う．

3.4　類似度評価

3.3 節で求めた各個体のベクトルと識別対象画像のベクトルを用いて類似度を求め，既知個体や未知個体の分類を行う．まず初めに，既知猫である各個体のベクトルと，判別を行いたい識別対象画像のベクトル間でそれぞれ cos 類似度を計算する．次に，求めた各 cos 類似度の中での最大値を求める．その後，既知・未知の判別のための閾値を設定し，これを用いて以下のように分類や行う．

- 最大 cos 類似度が閾値以上の場合，既知猫として cos 類似度をとったベクトルに該当する猫へと分類する．
- 最大 cos 類似度が閾値未満の場合，既知猫へ分類せず未知猫と分類する．

4　評 価 実 験

本章では，本研究の評価実験における実験方法と実験結果について述べる．

4.1　実 験 方 法

本実験の準備としてまず初めにデータセットの作成を行った．画像は野外の猫をカメラで撮ったものを想定し，今回はデータ収集の効率のため，猫の動画を撮った後，動画から画像を切り出し集めた．その後，3 章で述べた猫の検出と画像処理を行った．最終的にデータセットとして猫画像を 6 個体（各個体約 200 枚）計 1228 枚用意した．内訳を表 1 に示す．この 6 個体を既知 5 体未知 1 体に分け，学習: 検証: テスト＝ 6:2:2 に分割して使用した．使用した猫画像の例を図 3 に示す．

表 1　使用した画像の内訳

個体名	画像数
mike1	206
saba	204
mike2	200
tora1	207
tora2	209
siro	202
合計	1228

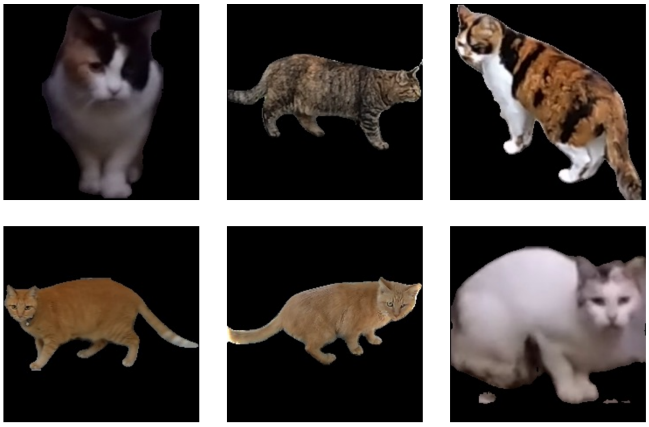


図 3　使用した猫画像の例

学習では CNN として ImageNet 学習済みの ResNet-50 を用いた．CNN への入力の際，画像の画素は 224 × 224 ピクセルとした．バッチサイズ 32，エポック数 10，学習率 0.001，損失関数は交差エントロピー誤差を用いた．また，特徴抽出器として用いる際の特徴ベクトルの次元は 2048 次元とした．

実験は，初めに未知個体を想定せずに既知猫 5 体で学習した時の分類，既知猫 6 体で学習した時の分類を行った．次に既知猫 5 体を学習したモデルを特徴抽出器として用い，提案手法に基づき，cos 類似度の閾値を変化させ既知猫の分類と未知猫の判定を行った．

また，提案手法以外に，SoftMax，OpenMax，マハラノビス距離を用いて，各々閾値を活用し同様に実験を行った．なおマハラノビス距離は，ユークリッド距離とは異なり各変数の相関を考慮した距離のことを言う．

4.2　実 験 結 果

4.2.1　提案手法の実験結果

使用した画像のうち，siro を未知個体として扱い実験を行った．未知個体を想定せずに既知猫 5 体で学習した時の検証時の正解率は 99.51%，テスト時の正解率は 97.10%であり，分類結果は表 2 に示す通りとなった．同様に，既知猫 6 体で学習した時の検証時の正解率は 99.18%，テスト時の正解率は 97.60%であり，分類結果は表 3 に示す通りとなった．正解率はどちらも 97%以上となり高い正解率を出した．これは事前の処理により，モデルが画像の特徴表現を上手く捉えることができたから

だと考えた．なお，tora1 や tora2 など誤分類された場合もあり，これは対象画像の猫の毛色が似ていたためだと考えた．ただ，モデルの正解率は高く，特徴抽出器として用いるには十分な性能だと考えた．

次に既知猫 5 体を学習したモデルを用いて，提案手法により閾値を様々変えて既知猫の分類と未知猫の判定を行った時の結果を表 4 に示す．最大 cos 類似度の閾値が 0.60 の時，既知個体正解率は 97.61%を記録したが未知個体正解率は 0%であり，全体正解率は 81.60%となった．閾値を大きくすると，閾値 0.90 の時，既知個体正解率が 97.13%，未知個体正解率が 100%，全体正解率が 97.60%となった．既知個体正解率は低くなったが，未知個体正解率は高くなり，全体正解率が最も高くなった．更に閾値を大きくし，閾値 0.95 の時，既知個体正解率が 94.26%，未知個体正解率が 100%，全体正解率が 95.20%となった．未知個体正解率は 100%のまま変わらなかったが既知個体正解率が低くなり，全体正解率も低くなった．これより，閾値を大きくすると既知個体正解率は低くなるが，未知個体正解率は高くなり，どちらかの正解率を高くしようとすると，もう一方の値が低くなるため，閾値を上手く設定することが重要だと考えた．

全体正解率が最も高い，閾値 0.90 の時の分類結果を表 5 に示す．これを見ると，本来既知個体だった画像が一部未知個体へと分類されてしまっていることが確認できる．ただ全体正解率は 97.60%を記録しており，誤分類はあるものの精度が高く分類することができた．これは未知個体が他個体と類似しておらず，類似度の分布が閾値の前後でよく分かれていたからだと考えた．

4.2.2 他手法の実験結果

提案手法以外に，SoftMax と閾値，OpenMax，マハラノビス距離と閾値を用いた方法で実験を行った．各手法で最良時の結果を表 6 に示す．提案手法を用いた場合，全体正解率 97.60%, SoftMax と閾値を用いた場合，全体正解率 88.00%, OpenMax を用いた場合，全体正解率 93.20%, マハラノビス距離と閾値を用いた場合，全体正解率 94.80%となった．SoftMax を用いた時に正解率が最も低くなったが，データが既知クラスにも当てはまらない場合でも確率の合計は 1.0 であるため，未知の場合でも高い確率が割り当てられることがあるからだと考えた．他手法について，提案手法では特徴ベクトル同士で類似度を求め分類していたが，OpenMax は Weibull 分布，マハラノビス距離は変数の相関を用いており，これらの手法は統計分布を考慮している．今回の地域猫データはデータ数が多くなく，単純なモデルでは表せなかったことや，既知個体に未知個体に類似した個体がいなかったため，統計に基づく手法より類似度に基づく提案手法の方が安定したのではないかと考えた．

4.2.3 各個体を未知として扱った時の実験結果

次に，siro 以外の個体も同様に未知個体として扱い，全個体に対し実験を行った．各個体を未知として扱った時の全体正解率の結果を表 7 に示す．提案手法の全体正解率は，saba,siro を未知個体とした時，それぞれ 96.40%,97.60%と他個体，他手法と比べ高い値を記録した．一方，mike1,mike2,tora1,tora2 を未知個体とした時，提案手法の正解率はそれぞれ 88.80%,92.40%,84.40%,86.80%となり，saba や siro といった個体に比べ低い値となった．また，他手法と比べるとこれらの個体の場合，SoftMax やマハラノビス距離を用いた時の値が提案手法以上の値となった．

表 2　既知猫 5 体の分類結果 (siro 以外)

正解＼予測	mike1	saba	mike2	tora1	tora2	正解率
mike1	42	0	0	0	0	1.0000
saba	0	42	0	0	0	1.0000
mike2	0	0	40	0	0	1.0000
tora1	0	0	0	39	3	0.9290
tora2	0	2	0	1	40	0.9300
全体正解率	–	–	–	–	–	0.9710

表 3　既知猫 6 体の分類結果

正解＼予測	mike1	saba	mike2	tora1	tora2	siro	正解率
mike1	42	0	0	0	0	0	1.0000
saba	0	42	0	0	0	0	1.0000
mike2	0	0	38	2	0	0	0.9500
tora1	0	0	0	41	1	0	0.9762
tora2	0	0	0	3	40	0	0.9302
siro	0	0	0	0	0	41	1.0000
全体正解率	–	–	–	–	–	–	0.9760

表 4　提案手法の閾値ごとの分類結果 (siro)

閾値	既知個体正解率	未知個体正解率	全体正解率
0.60	0.9761	0.0000	0.8160
0.65	0.9761	0.0244	0.8200
0.70	0.9761	0.1463	0.8400
0.75	0.9761	0.2439	0.8560
0.80	0.9713	0.3415	0.8680
0.85	0.9713	0.5366	0.9000
0.95	0.9426	1.0000	0.9520
1.00	0.0000	1.0000	0.1640

これらより，saba や siro といった未知個体がユニークな特徴を持つ場合，特徴空間上で未知個体が独立しているため，類似度を用いる提案手法が優位となったと考えた．一方，mike1,mike2 や tora1,tora2 といった未知個体が既知個体に類似している場合，特徴空間上で未知個体が既知個体と近い位置になるため，境界や分布を重視する他手法の方が優位となったと考えた．

また，saba のマハラノビス距離による正解率は 84.00%,siro の SoftMax による正解率は 88.00%であり，各個体の中で他の全手法と比べ 5%以上の差をつけて低くなった．提案手法や OpenMax を用いたときの正解率は，各個体の中で他の全手法と比べここまでの差はなく，提案手法や OpenMax は平均を用いているため，安定した正解率を記録したと考えた．

5　結論および今後の展望

本稿では，地域猫を対象として猫の個体分類を行い，既知の猫の個体識別を行うと同時に，未知の猫を未知として判別することを目的とした．手法としては，既知個体を学習したモデルを作成し，それを特徴抽出器として用いることにより既知個体

表 5　最良閾値の時の分類結果 (siro)

正解＼予測	mike1	saba	mike2	tora1	tora2	unknown	正解率
mike1	42	0	0	0	0	0	1.0000
saba	0	42	0	0	0	0	1.0000
mike2	0	0	40	0	0	0	1.0000
tora1	0	0	0	39	2	1	0.9286
tora2	0	0	0	0	40	3	0.9302
unknown	0	0	0	0	0	41	1.0000
全体正解率	–	–	–	–	–	–	0.9760

表 6　他手法との比較 (siro)

方法	既知個体正解率	未知個体正解率	全体正解率
提案手法	0.9713	1.0000	0.9760
SoftMax	0.9330	0.6098	0.8800
OpenMax	0.9187	1.0000	0.9320
マハラノビス距離	0.9474	0.9512	0.9480

表 7　各個体を未知として扱った時の全体正解率

方法	mike1	saba	mike2	tora1	tora2	siro
提案手法	0.8880	0.9640	0.9240	0.8440	0.8680	0.9760
SoftMax	0.8880	0.9560	0.9320	0.8440	0.8640	0.8800
OpenMax	0.9160	0.9560	0.9240	0.8240	0.8640	0.9320
マハラノビス距離	0.9320	0.8400	0.9280	0.8440	0.8960	0.9480

画像および未知個体画像の特徴ベクトルを抽出した後，特徴ベクトル間の類似度を求め，設定した閾値により既知個体と未知個体の分類を行った．実験の結果，ユニークな特徴を持つ個体の場合，提案手法は高い正解率を記録し，最大で 97.60%となった．一方，類似個体を持つ場合，提案手法に比べ他手法の方が優れた結果を記録した．以上より，他手法の方が優れる場合はあるものの，本手法により概ね既知個体と未知個体の分類を行うことができることが分かった．

課題としては，個体が類似している場合明確な分類が難しく，既知個体に類似した未知個体をいかに精度良く未知と分類するかが，性能改善の上で今後重要となる．また，このような閾値を活用する手法を用いる場合，既知個体正解率か未知個体正解率どちらかを上げようとすると，もう一方が下がるため，全体正解率を高くするために手動で適切に閾値を設定しないといけないという問題がある．

今後の展望としては，各個体をより識別できるように特定の体の特徴や猫の向きといった情報を用いることや，特徴抽出器の改善の為，Triplet loss や ArcFace 等の距離学習手法を検討したい．また，現在はデータセットの中で特定の 1 個体のみ未知個体として扱っているため，未知個体を複数個体に設定しての実験を行いたい．加えて，実用化としてリアルタイムで動くアプリケーションの開発を目指したい．

文　献

[1] 環境省統計資料「犬・猫の引取り及び負傷動物等の収容並びに処分の状況」，https://www.env.go.jp/nature/dobutsu/aigo/2_data/statistics/dog-cat.html （参照 2025/12/9）

[2] Tobias Trein, Luan Fonseca Garcia, “Siamese Networks for Cat Re-Identification: Exploring Neural Models for Cat Instance Recognition,” arXiv:2501.02112 (2025)

[3] Mingxuan Li, Kai Zhou, “The Comparison of Individual Cat Recognition Using Neural Networks,” arXiv:2410.02305 (2024)

[4] Zihan Yang, Richard Sinnott, Qiuhong Ke, James Bailey, “Individual Feral Cat Identification through Deep Learning,” BDCAT'21: 2021 IEEE/ACM 8th International Conference on Big Data Computing, Applications and Technologies, pp.101-110 (2021)

[5] Victor Caquilpan, “What cat is that? A re-id model for feral cats,” arXiv:2507.11575 (2025)

[6] Cen Liu, Rong Zhang, Lijun Guo, “Part-Pose Guided Amur Tiger Re-Identification,” 2019 IEEE/CVF International Conference on Computer Vision Workshop(ICCVW) (2019)

[7] Ultralytics YOLO Docs, https://docs.ultralytics.com （参照 2025-12-30）

[8] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp.779–788 (2016)

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Deep Residual Learning for Image Recognition,” 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp.770-778 (2016)

[10] Abhijit Bendale, Terrance E. Boult, “Towards Open Set Deep Networks,” 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp.1563-1572 (2016)

表 8　提案手法の閾値ごとの分類結果 (mike1)

閾値	既知個体正解率	未知個体正解率	全体正解率
0.60	0.9615	0.0000	0.8000
0.65	0.9615	0.0000	0.8000
0.70	0.9615	0.0000	0.8000
0.75	0.9615	0.0952	0.8160
0.80	0.9615	0.1905	0.8320
0.85	0.9567	0.3571	0.8560
0.88	0.9471	0.5952	0.8880
0.90	0.9327	0.6667	0.8880
0.95	0.8894	0.8095	0.8760
1.00	0.0000	1.0000	0.1680

表 9　最良閾値の時の分類結果 (mike1)

正解＼予測	unknown	saba	mike2	tora1	tora2	siro	正解率
unknown	25	0	6	0	0	11	0.5952
saba	0	42	0	0	0	0	1.0000
mike2	0	0	40	0	0	0	1.0000
tora1	5	0	1	36	0	0	0.8571
tora2	3	0	0	1	39	0	0.9070
siro	1	0	0	0	0	40	0.9756
全体正解率	–	–	–	–	–	–	0.8880

表 10　他手法との比較 (mike1)

方法	既知個体正解率	未知個体正解率	全体正解率
提案手法	0.9471	0.5952	0.8880
SoftMax	0.8990	0.8333	0.8880
OpenMax	0.9087	0.9524	0.9160
マハラノビス距離	0.9471	0.8571	0.9320

表 11　提案手法の閾値ごとの分類結果 (saba)

閾値	既知個体正解率	未知個体正解率	全体正解率
0.60	0.9904	0.0000	0.8240
0.65	0.9904	0.0000	0.8240
0.70	0.9904	0.0238	0.8280
0.75	0.9904	0.1190	0.8440
0.80	0.9904	0.2619	0.8680
0.85	0.9808	0.5952	0.9160
0.90	0.9760	0.7857	0.9440
0.93	0.9615	0.9762	0.9640
0.95	0.9519	1.0000	0.9600
1.00	0.0000	1.0000	0.1680

表 12　最良閾値の時の分類結果 (saba)

予測 / 正解	mike1	un-known	mike2	tora1	tora2	siro	正解率
mike1	42	0	0	0	0	0	1.0000
unknown	0	41	0	0	1	0	0.9762
mike2	0	1	39	0	0	0	0.9750
tora1	0	4	0	38	1	0	0.8837
tora2	0	3	0	0	40	0	0.9302
siro	0	0	0	0	0	41	1.0000
全体正解率	–	–	–	–	–	–	0.9602

表 13　他手法との比較 (saba)

方法	既知個体正解率	未知個体正解率	全体正解率
提案手法	0.9615	0.9762	0.9640
SoftMax	0.9615	0.9286	0.9560
OpenMax	0.9471	1.0000	0.9560
マハラノビス距離	0.9615	0.2381	0.8400

表 14　提案手法の閾値ごとの分類結果 (mike2)

閾値	既知個体正解率	未知個体正解率	全体正解率
0.60	0.9905	0.0000	0.8320
0.65	0.9905	0.0000	0.8320
0.70	0.9905	0.0000	0.8320
0.75	0.9905	0.0250	0.8360
0.80	0.9857	0.1000	0.8440
0.85	0.9762	0.2750	0.8640
0.90	0.9619	0.5500	0.8960
0.95	0.9524	0.7500	0.9200
0.97	0.9238	0.9250	0.9240
1.00	0.0000	1.0000	0.1600

表 15　最良閾値の時の分類結果 (mike2)

予測 / 正解	mike1	saba	un-known	tora1	tora2	siro	正解率
mike1	42	0	0	0	0	0	1.0000
saba	0	42	0	0	0	0	1.0000
unknown	0	0	37	2	0	1	0.9250
tora1	0	0	4	38	0	0	0.9048
tora2	0	0	10	0	33	0	0.7674
siro	0	0	2	0	0	39	0.9512
全体正解率	–	–	–	–	–	–	0.9240

表 16　他手法との比較 (mike2)

方法	既知個体正解率	未知個体正解率	全体正解率
提案手法	0.9238	0.9250	0.9240
SoftMax	0.9571	0.8000	0.9320
OpenMax	0.9095	1.0000	0.9240
マハラノビス距離	0.9714	0.7000	0.9280

表 17　提案手法の閾値ごとの分類結果 (tora1)

閾値	既知個体正解率	未知個体正解率	全体正解率
0.60	1.0000	0.0000	0.8320
0.65	1.0000	0.0000	0.8320
0.70	1.0000	0.0000	0.8320
0.75	1.0000	0.0000	0.8320
0.80	1.0000	0.0238	0.8360
0.82	1.0000	0.0714	0.8440
0.85	1.0000	0.0714	0.8440
0.90	1.0000	0.0714	0.8440
0.95	0.9808	0.1190	0.8360
1.00	0.0000	1.0000	0.1680

表 18　最良閾値の時の分類結果 (tora1)

予測 / 正解	mike1	saba	mike2	un-known	tora2	siro	正解率
mike1	42	0	0	0	0	0	1.0000
saba	0	42	0	0	0	0	1.0000
mike2	0	0	40	0	1	0	0.9756
unknown	0	0	2	3	37	0	0.0714
tora2	0	0	0	0	43	0	1.0000
siro	0	0	0	0	0	41	1.0000
全体正解率	–	–	–	–	–	–	0.8406

表 19　他手法との比較 (tora1)

方法	既知個体正解率	未知個体正解率	全体正解率
提案手法	1.0000	0.0714	0.8440
SoftMax	1.0000	0.0714	0.8440
OpenMax	0.9327	0.2857	0.8240
マハラノビス距離	0.9760	0.1905	0.8440

表 20　提案手法の閾値ごとの分類結果 (tora2)

閾値	既知個体正解率	未知個体正解率	全体正解率
0.60	0.9952	0.0000	0.8240
0.65	0.9952	0.0698	0.8360
0.70	0.9952	0.0698	0.8360
0.75	0.9952	0.1163	0.8440
0.80	0.9903	0.1860	0.8520
0.85	0.9855	0.2326	0.8560
0.90	0.9807	0.2558	0.8560
0.94	0.9662	0.3953	0.8680
0.95	0.9614	0.4186	0.8680
1.00	0.0000	1.0000	0.1720

表 21　最良閾値の時の分類結果 (tora2)

予測 / 正解	mike1	saba	mike2	tora1	un-known	siro	正解率
mike1	42	0	0	0	0	0	1.0000
saba	0	42	0	0	0	0	1.0000
mike2	0	0	36	4	1	0	0.8780
tora1	0	0	0	41	1	0	0.9762
unknown	0	0	1	25	17	0	0.3953
siro	0	0	0	2	0	39	0.9512
全体正解率	–	–	–	–	–	–	0.8645

表 22　他手法との比較 (tora2)

方法	既知個体正解率	未知個体正解率	全体正解率
提案手法	0.9662	0.3953	0.8680
SoftMax	0.9662	0.3721	0.8640
OpenMax	0.9130	0.6279	0.8640
マハラノビス距離	0.9614	0.5814	0.8960