# 修 士 論 文 概 要 書

Master's Thesis Summary

Date of submission: 01/27/2025 (MM/DD/YYYY)

| 専攻名（専門分野）<br>Department | 情報理工・<br>情報通信専攻 | 氏 名<br>Name | 杉山 秀治 | 指 導<br>教 員<br>Advisor | 渡辺 裕 印<br>Seal |
|---|---|---|---|---|---|
| 研究指導名<br>Research guidance | オーディオビジュアル<br>情報処理研究 | 学籍番号<br>Student ID number | CD<br>5123F059-3 | | |
| 研究題目<br>Title | 顔情報の統合活用によるアピアランスベースの視線推定手法<br>Appearance-Based Gaze Estimation Method through Integrated Utilization of Facial Information | | | | |

## 1. まえがき

視線推定は，人の視線方向を推定する技術であり，AR/VR, マーケティング, 医療分野などで活用され，注目されている．視線推定手法の一種である，アピアランスベースの手法は，顔画像から得られる視覚情報を直接活用する柔軟性と，特殊なデバイスやキャリブレーションを必要としない利点から，多様な環境での応用が期待される．しかし，顔の向きや目の小ささ，照明条件，個人差などの複数の要因が複雑に絡み合うため，正確な推定は依然として難しい．このような課題に対して，アピアランスベースの手法の利便性を維持しつつ，複数の要因を一括で解決を図るアプローチで研究が進められている．

そこで，本研究では，視線推定の精度に影響を与える複数の要因を統合的に解決するために，顔全体の構造的な特徴を活用することで，推定精度を向上させる手法を提案する．さらに，顔画像のみを利用することで，従来の両目を用いる手法が抱える実世界応用時の不安定性を低減することを目指す．

## 2. 従来の視線推定手法

### 2.1. L2CS-Net

L2CS-Net[1]は，顔画像のみを用いたアピアランスベースの視線推定手法であり，Gaze360[2]において，CNN を用いた手法の中では優れた手法である．L2CS-Net の概要を図 1 に示す．L2CS-Net は，目の情報と顔全体の構造的な情報を十分に活用できていないという課題がある．

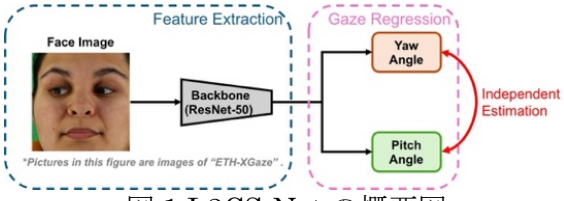


図 1 L2CS-Net の概要図

### 2.2. L2CS-Net の改善手法

L2CS-Net の改善手法[3]は，顔画像と両目の画像を用いて，人間のモノを見る動作を模倣することで，精度を向上させた手法である．L2CS-Net の改善手法の概要を図 2 に示す．この手法を含む両目を用いる手法は，目の位置を正確に検出することを前提とする．しかし，リアルタイムでの目の領域検出は，顔領域検出と比較して不安定であり，これらの従来手法は精度の低下や不作動の恐れがある．

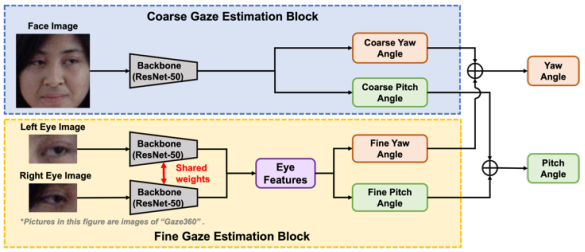


図 2 L2CS-Net の改善手法の概要図
(IWAIT2024 にて発表)

## 3. 顔領域の疎密構造を捉える手法

### 3.1. 概要

本研究では，顔領域の疎密構造を捉える手法を提案する. 提案手法の概要を図 3 に示す. 提案手法は，顔全体の大域的な形状を疎な特徴として捉え，目の周辺など局所的な詳細情報を密な特徴として捉え，補強する 2 段階構造で構成される．

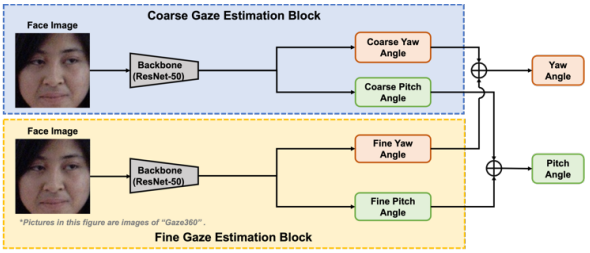


図 3 顔領域の疎密情報を捉える手法の概要図

### 3.2. 実験

従来手法と提案手法の視線推定精度の比較結果を表 1 に示す．評価実験は，データセットとして Gaze360 と RT-GENE[4]，評価指標として Mean Angular Error(MAE)を用いる．表 1 より，提案手法は L2CS-Net の精度向上を達成し，顔領域の疎密構造を捉えることの有効性が確認できる．また提案手法は，Gaze360 において，L2CS-Net の改善手法と遜色のない性能を示しており，実用性の観点で優れた手法と考えられる．一方で，RT-GENE での精度向上幅が小さい原因として，データセットの画像の目の周辺が画像修復によりぼやけていることが挙げられる．そのため，密な特徴を捉える際に影響を与えていると考えられる．

表 1 従来手法と提案手法の視線推定精度の比較

| | Gaze360 | RT-GENE |
|---|---|---|
| L2CS-Net | 10.41° | 6.59° |
| L2CS-Netの改善手法 | **10.16°** | **6.44°** |
| Ours | 10.19° | 6.55° |

## 4. 顔のキーポイント活用手法

### 4.1. 予備実験

L2CS-Net モデルに Class Activation Mapping (CAM)を適用し，モデルの判断傾向を分析する．Gaze360 で学習した L2CS-Net モデルに対して，Grad-CAM[5]と Grad-CAM++[6]の CAM 手法を適用，分析する．ヒートマップ図の例を図 4 に示す．左から元の顔画像，Grad-CAM の結果，Grad-CAM++の結果である．また，各画像の視線推定の角度誤差を表 2 に示す．これら実験結果と Gaze360 における L2CS-Net の MAE が10.41°であることから，ヒートマップ画像の Attention の傾向と推定精度に相関がある可能性が示唆される．そのため，目の領域の Attention が高くなるようにモデル設計することで視線推定精度の向上が期待されると考えられる．

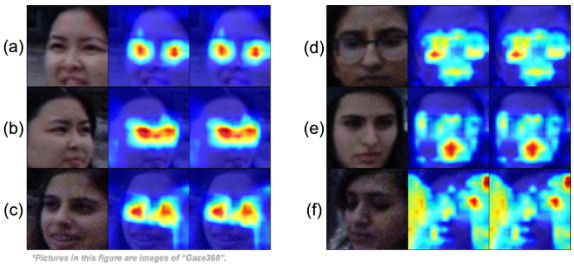


図 4 L2CS-Net モデルの可視化例
（2024 年映像情報メディア学会年次大会にて発表）

表 2 各画像の視線推定の角度誤差

| 入力画像 | (a) | (b) | (c) | (d) | (e) | (f) |
|---|---|---|---|---|---|---|
| 角度誤差 | 8.38° | 4.72° | 0.40° | 31.74° | 41.39° | 43.55° |

### 4.2. 概要

予備実験に基づき，目の領域を重視するために，顔のキーポイント情報を補助情報として活用する手法を提案する．提案手法では，目や鼻，口などの顔の重要部位を示すキーポイント情報を用いて，顔全体の構造を捉え，精度向上を図る．また，キーポイント情報を補助情報として活用する設計により，実世界応用時の不安定性を軽減する．提案手法の概要を図 5 に示す．まず，前処理として，顔のキーポイント検出手法である dlib または FaceMesh により顔のキーポイント画像を得る．そして，元の顔画像とキーポイント画像をモデルの入力とし，推定する．

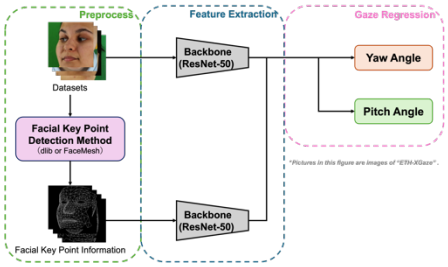


図 5 顔のキーポイント活用手法の概要図

### 4.3. 実験

L2CS-Net と提案手法の視線推定精度の比較結果を表 3 に示す．評価実験は，データセットとして ETH-XGaze[7]，評価指標として MAE を用いる．表 3 より，提案手法は L2CS-Net の精度向上を達成し，顔のキーポイント情報活用の有効性が確認できる．また，dlib と比べて，詳細なキーポイント点を取得できる FaceMesh の方が効果的であることがわかる．また，L2CS-Net と FaceMesh を用いた提案手法のモデルに Grad-CAM と Grad-CAM++を適用した際のヒートマップ図の例を図 6 に示す．図 6 より，提案手法の方が目の領域への Attention が高く，意図通りのモデル設計であることが確認できる．

表 3 従来手法と提案手法の視線推定精度の比較

| Method | ETH-XGaze |
|---|---|
| L2CS-Net | 15.82° |
| Ours(w/ dlib) | <u>15.79°</u> |
| Ours(w/ FaceMesh) | **15.72°** |

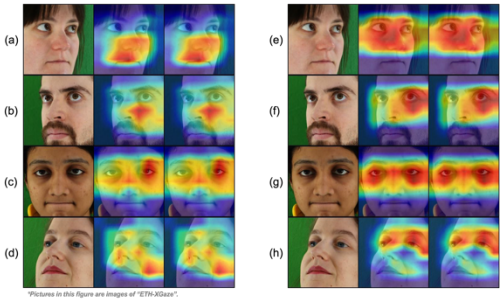


図 6 L2CS-Net と提案手法のモデルの可視化例

## 5. むすび

本研究では，アピアランスベースの視線推定手法の精度向上を目的として，実世界応用時の不安定性を考慮した，顔の疎密構造を捉える手法と顔のキーポイント情報活用手法を提案した．実験により，各手法の有効性を確認し，顔全体の構造的特徴を統合的に活用することが視線推定精度の向上に寄与することを示した．また，提案手法は視線推定手法全般に適用可能であり，さらなる活用が期待される．

## 参考文献


[1] A. A. Abdelrahman *et al.*, "L2CS-Net: Fine-Grained Gaze Estimation in Unconstrained Environments," International Conference on Frontiers of Signal Processing, pp.98-102, Oct. 2023.

[2] P. Kellnhofer *et al.*, "Gaze360: Physically Unconstrained Gaze Estimation in the Wild," IEEE International Conference on Computer Vision, pp.6911-6920, Oct. 2019.

[3] H. Sugiyama *et al.*, “Enhancing Gaze Estimation through Fine-Grained Analysis of Eye Region,” International Workshop on Advanced Image Technology, Jan. 2024.

[4] T. Fischer *et al.*, “RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments,” European Conference on Computer Vision, pp.339-357, Sep. 2018.

[5] R. R. Selvaraju *et al.*, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," IEEE International Conference on Computer Vision, pp.618-626, Oct. 2017.

[6] A. Chattopadhay, *et al.*, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," IEEE Winter Conference on Applications of Computer Vision, pp.839-847, Mar. 2018.

[7] X. Zhang *et al.*, “ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation,” European Conference on Computer Vision, pp.365-381, Aug. 2020.

2024 年度
早稲田大学大学院基幹理工学研究科情報理工・情報通信専攻 修士論文

# 顔情報の統合活用による
# アピアランスベースの視線推定手法

## Appearance-Based Gaze Estimation Method
## through Integrated Utilization of Facial Information

杉山 秀治

(5123F059-3)

提出日：2025.1.27

指導教員：渡辺 裕教授

研究指導名：オーディオビジュアル情報処理研究

# 目次

# 第１章　序論

## １．１　研究の背景

視線推定は，画像から人の視線方向を推定する技術である．人間の視線情報は，非言語コミュニケーションにおいて重要な役割を果たす．視線は，注意や意図，感情を表現するため，人間の行動理解において欠かせない要素であり，これを正確に捉えることで，より自然なインタラクションが可能となる．近年では，視線推定技術が多岐にわたる分野で活用されており，特にマーケティング，AR/VR，医療，運転支援などの産業領域で注目されている．例えば，マーケティング分野では，視線追跡技術を利用して消費者の行動を分析することができる．AR や VR では，視線を入力手段として使用することで，ユーザーとのインタラクションを向上させ，より直感的で没入感のある体験を提供する．このように，ユースケースに合わせた視線推定技術は実社会で実用化されている技術である．

視線推定手法は，大きくモデルベースの手法とアピアランスベースの手法の二つに分類される．モデルベースの手法は，目の幾何学的モデルを構築して視線方向を推定するアプローチで，高い精度が期待できる．しかし，高品質な画像やキャリブレーションが必要であるため適用範囲が限定される．一方で，アピアランスベースの手法は，画像から得られる視覚情報を活用するアプローチで，多様な環境やデバイスに対応可能である．近年，ディープラーニング技術の進展[1]により，アピアランスベースの視線推定手法の精度は大きく向上した．しかし，依然として顔の向き，表情，照明条件，目の小ささ，個人差など，さまざまな要因が視線推定に影響を与えており，これらの要因が複雑に絡み合っているため，正確な推定が難しいという課題がある．これらの要因を個別に解決するアプローチも考えられるが，アピアランスベースの手法の利点は，視線推定が容易に行えることである．したがって，深層学習を活用して複数の要因を統合的に処理するアプローチを取ることで，複雑な課題を一括で解決できる可能性が高まり，精度向上と実用性の両立が実現できると考えられる．さらに，これらの複数の要因は相互に影響し合っているため，個別のアプローチが他の要因に悪影響を与える可能性がある．そのため，深層学習を用いて複数の要因を同時に学習することにより，アピアランスベースの視線推定手法の堅牢性や汎用性を向上させることも期待され，研究が進められている．

## １．２　本研究の目的

本研究では，アピアランスベースの視線推定手法を研究対象として扱う．この手法は，広範な環境条件やデバイスの制約を受けにくく，リアルタイムでのアプリケーションへの応用に適しているという特徴を持つ．これらの柔軟性と実用性の高さを背景に，本研

究ではアピアランスベースの手法に焦点を当て，その進展を目指す．

先行研究である L2CS-Net[2]は，視線推定において約10°の誤差を示しており，人間の視線推定誤差（約4°）[3]と比較すると精度が劣っていることから，アピアランスベースの視線推定手法は改善の余地があると考えられる．そして，L2CS-Net の改善手法[4]を含む両目を用いる手法では，データセットに含まれる目の位置ラベルを基に目の画像を使用して視線推定をする．しかし，この手法は目の位置が正確に取得できる状態を前提にしており，実世界の応用環境において目の位置が安定して取得できるとは限らない．このため，両目の画像を用いる手法は実世界での適用時に精度悪化の恐れがあり，不安定性を抱えている．

そこで，本研究では，アピアランスベースの視線推定手法において，深層学習を活用して複雑な要因を統合的に解決することで精度の向上を目指す．また，安定して検出可能である顔画像のみを用いることで，実世界での応用に適した手法の実現を目指す．

## １．３　本論文の構成

以下に本論文の構成を示す．

第１章　本研究の背景，およびその目的について述べる．

第２章　本研究における関連手法について述べる．

第３章　顔領域の疎密構造を捉える手法について述べる．

第４章　顔のキーポイント情報を補助情報とした活用する手法について述べる．

第５章　本研究の結論と今後の展望を述べる．

# 第２章　関連手法

## ２．１　まえがき

本章では，本研究における関連手法について述べる．本研究ではアピアランスベースの視線推定手法に焦点を当て，その中でも L2CS-Net と L2CS-Net の改善手法に注目する．また，提案手法で用いる Class Activation Mapping と顔のキーポイント検出手法についても述べる．

## ２．２　従来の視線推定手法

### ２．２．１　L2CS-Net

L2CS-Net は，人の顔画像を用いたアピアランスベースの視線推定手法である．ImageNet で学習済みの ResNet-50 をバックボーンに用いることで，シンプルな構造ながらも，Gaze360[5]のデータセットにおいて CNN ベースの視線推定手法の中で優れた性能を示している．L2CS-Net の視線推定ネットワークを図 2.1 に示す．

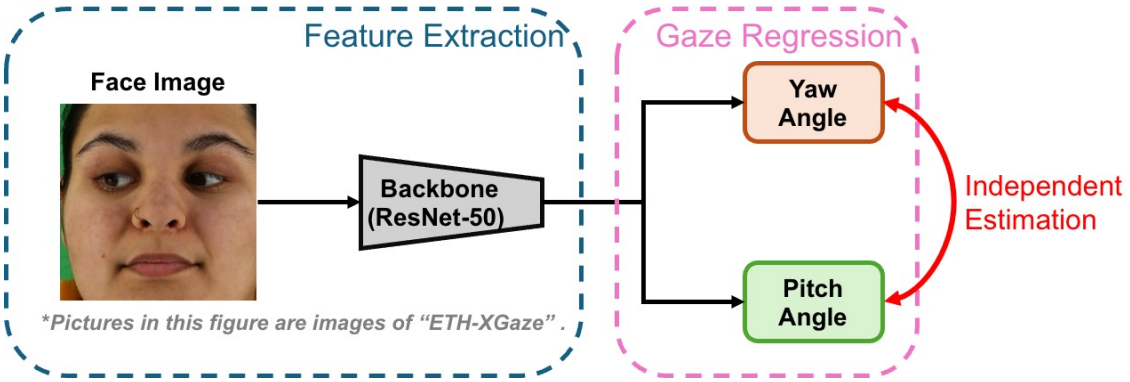


図 2.1 L2CS-Net の視線推定ネットワーク

L2CS-Net は，視線推定において一定の精度を示しているものの，依然として課題がある．具体的には，Gaze360 データセットにおいて，L2CS-Net の平均角度誤差は10.41°である．一方で，五十嵐ら[3]が自作したデータセットにおいて，人間の視線推定の平均角度誤差は約 4°である．両者はデータセットの条件が異なるため直接的な比較は難しいものの，L2CS-Net の推定精度にはさらなる改善の余地があると考えられる．また，L2CS-Net では顔画像から視線方向を推定する際，目のサイズが小さいため，視線推定における目の情報の利用が十分でない場合があると考えられる．さらに，視線推定では目の領域が重要視されるため，顔全体の情報が十分に活用されていない可能性がある．これらの要因が L2CS-Net の精度不足の一因であると考えられる．そのため，従来手法の精度を向上させるために，目の情報や顔全体の情報を十分に捉える必要がある．

### ２．２．２　L2CS-Net の改善手法

L2CS-Net の改善手法では，顔画像とそれに対応する両目の画像を用いることで，人間の視線メカニズムを模倣し，精度向上を達成している．ここで視線メカニズムとは，頭を動かし，その後瞳孔を動かし物に焦点を合わせるという人間のモノを見る動作であり，これをネットワーク構造に応用している．L2CS-Net の改善手法の視線推定ネットワークを図 2.2 に示す．

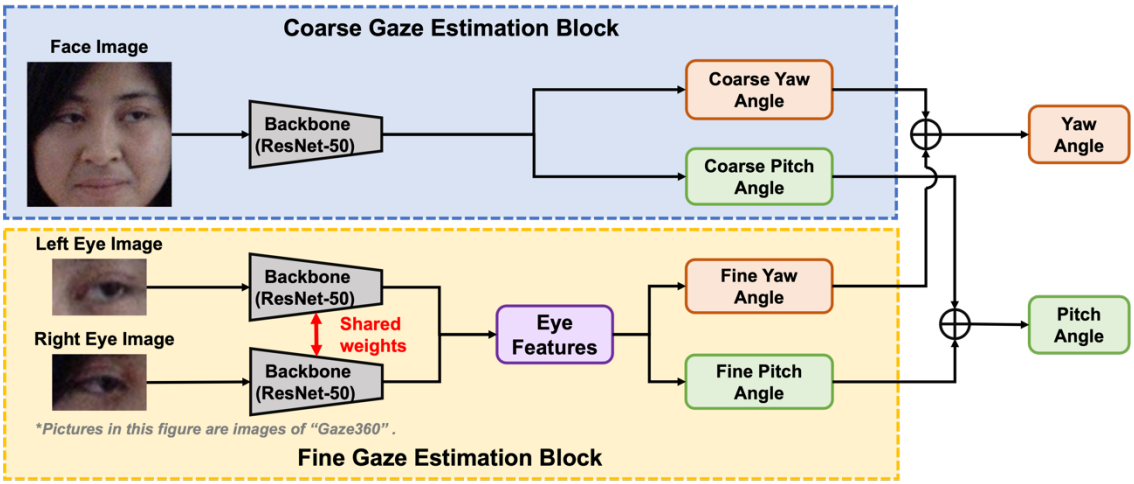


図 2.2 L2CS-Net の改善手法の視線推定ネットワーク
（IWAIT2024 にて発表）

青色の Coarse Gaze Estimation Block では，顔全体の画像から大まかな視線方向（Coarse Angle）を推定する．続く黄色の Fine Gaze Estimation Block では，両目の画像を用いて，Coarse Angle を補正する，精密な視線方向（Fine Angle）を推定する．

しかし，L2CS-Net の改善手法には課題がある．実世界でリアルタイムに動作させるには，顔領域に加えて両目の領域を検出する必要がある．実際にリアルタイムでの顔領域および目の領域の検出をした際の実行例を図 2.3 に示す．

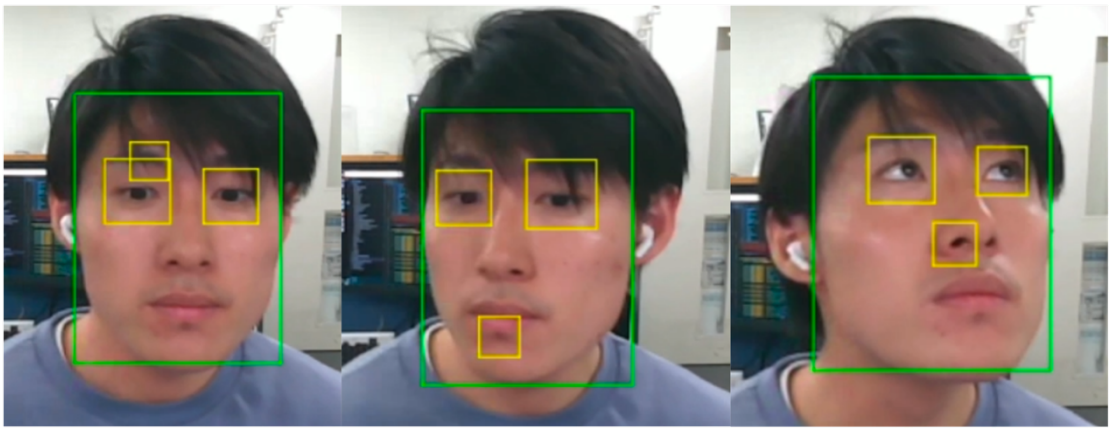

図 2.3 顔検出および目の検出結果の例

顔検出には RetinaFace[6]，目の検出には OpenCV のカスケード分類器を用いる．図中の緑色の枠が顔領域，黄色の枠が目の検出結果である．この図から明らかなように，顔領域は安定して検出できる一方，目の領域は検出結果が不安定で正確性が欠けている．そのため， L2CS-Net の改善手法や両目の画像を用いる他の手法[7, 8]は，正確な目の領域検出を前提とした手法であるため，実世界の適用時に精度は悪化したり，不動さを起こしたりする恐れがあり，不安定性を抱えている．そのため，実用化する際の安定性も考慮する必要があると考えられる．

## ２．３　Class Activation Mapping 手法

### ２．３．１　Class Activation Mapping

Class Activation Mapping（CAM）[9]は，特定のクラスの分類に寄与する画像領域を可視化する手法である．初期の提案では，CNN の最終層の特徴マップと全結合層の重みを組み合わせて実現させ，全結合層を持つネットワークに限定していた．しかし，現在では勾配情報や特徴マップの操作を用いた多様な手法が開発され，可視化技術全般を CAM と総称することが一般的である．このような変遷により，CAM は画像分類や物体検出，セグメンテーションなど，幅広い応用で利用される．

### ２．３．２　Grad-CAM および Grad-CAM++

Gradient-weighted Class Activation Mapping（Grad-CAM）[10]は、CAM を拡張した手法で，CNN の特定の層の勾配情報を用いて重要な領域を可視化する手法である．この手法は，分類タスクだけでなく，物体検出やセグメンテーションなどでも有用である．

さらに，Grad-CAM++[11]は、Grad-CAM を改良した手法で，複数のクラスや複雑な構造に対してより精度の高い可視化を実現する．Grad-CAM++は、勾配の重み付けをより詳細に解析し、非線形な相互作用が多いタスクにおいても，詳細な注目領域を示すことが可能である．

ここで，ImageNet で学習済みの ResNet50 モデルに，Stanford Dogs Dataset[12]に含まれる犬の画像を入力し，これら二つの CAM 手法を適用した際の判断根拠を可視化した結果を図 2.4 に示す．

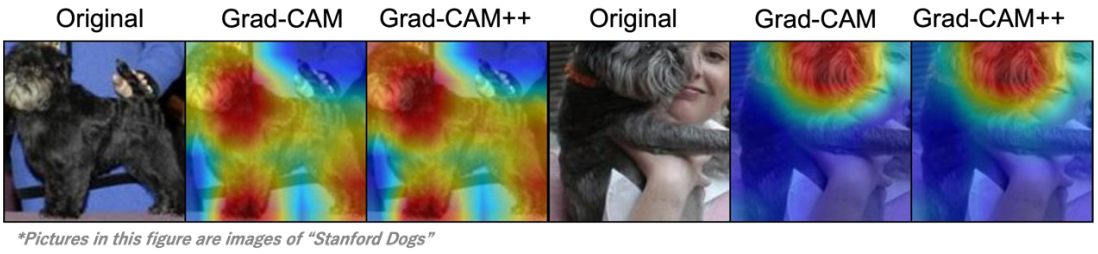


図 2.4 Grad-CAM および Grad-CAM++の可視化例

このようにクラス分類において犬と判断した際に，犬の領域に注目が集まっており，モデルの判断根拠の可視化が可能であることがわかる．そして，Grad-CAM と Grad-CAM++は，クラス分類以外のタスクの可視化も可能であることから，本研究では視線推定モデルの注目領域を可視化するために使用する．

## ２．４　顔のキーポイント検出手法

### ２．４．１　dlib

dlib[13]はオープンソースの機械学習ライブラリで，顔のランドマーク検出モデルを含む．このモデルは，顔の 68 点のキーポイントを取得できる．

### ２．４．２　OpenPose

OpenPose[14]は，人間の姿勢や顔の特徴点を推定する技術である．本研究では，顔の 70 点のキーポイントを取得する機能を利用する．

### ２．４．３　FaceMesh

FaceMesh[15]は Google が公開する機械学習モデルで，従来の 68 点のキーポイント検出に比べ，最大 468 点の詳細な特徴点を検出可能である．

## ２．５　むすび

本章では，本研究における関連手法として，L2CS-Net および L2CS-Net の改善手法をあげ，それぞれの特徴や課題を述べた．本研究では，これら従来のアピアランスベースの視線推定手法が抱える課題に対して，二つの手法を提案し，詳しくは以降の章で述べる．また，提案手法に関連する技術として，CAM と顔のキーポイント検出手法について述べた．

# 第３章　顔領域の疎密構造を捉える手法

## ３．１　まえがき

本章では，アピアランスベースの視線推定手法において，顔領域における疎密構造を捉える手法を提案する．提案手法について説明したのち，提案手法の評価実験，考察について述べる．

## ３．２　提案手法

顔領域の疎な情報と密な情報を捉えることで，顔全体の情報を統合的に活用し，精度向上を図る手法を提案する．従来のアピアランスベースの視線指定手法のうち，顔画像のみを用いる手法は，安定性に優れるが，顔全体の情報を十分に活用していない．一方で，両目の画像を用いる手法は，顔全体の構造は捉えられるものの，実世界での応用時の不安定性を抱える．そこで，提案手法は，顔画像のみを用いて安定性を維持しつつ，二つのブロック構造により，顔画像全体の疎密構造を捉えることで精度の向上を図る．これにより，目の領域を特定せずとも顔全体の情報を活用できると考えられる．提案手法の視線推定ネットワーク構造を図 3.1 に示す．

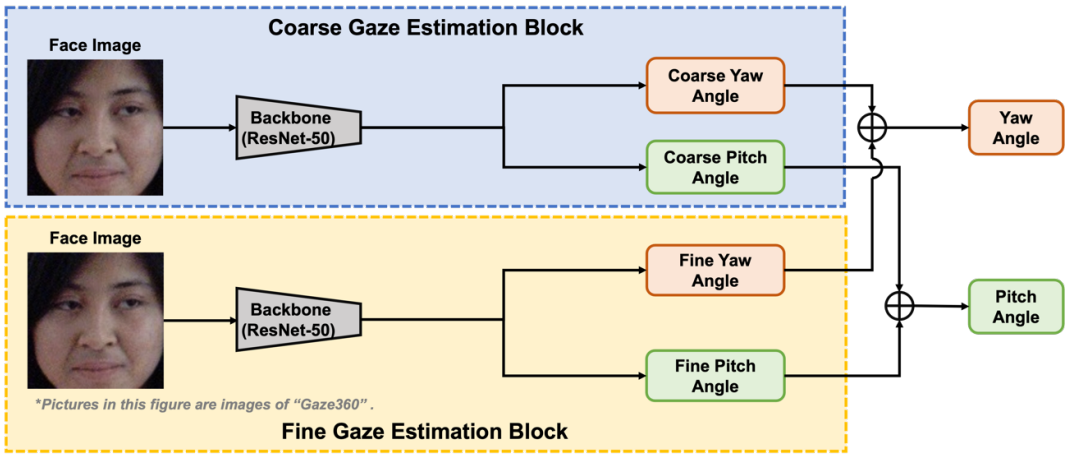


図 3.1 顔領域の疎密構造を捉える手法の視線推定ネットワーク

提案手法は以下の手順で視線方向を推定する．

- 青色で示される Coarse Gaze Estimation Block では，顔領域における顔の向きや目の位置など（疎な情報）から，大まかな視線方向である Coarse Angle を推定する．
- 黄色で示される Fine Gaze Estimation Block では，顔領域における目の微細な動きや顔の細かい特徴など（密な情報）から，Coarse Angle を正しい視線方向を補正

する Fine Angle を推定する．

- Coarse Angle と Fine Angle を合わせて視線方向である Angle を推定する．

また，提案手法の視線推定ネットワークの Loss 関数について述べる．提案手法で算出する視線方向は, オイラー角である yaw 角と pitch 角を独立して回帰する. そのため, yaw 角と pitch 角それぞれの角度について個別に Loss 関数を算出する．

まず，Coarse Angle に関する Loss 関数を式 1 に示す.

$$Loss_{Coaese} = Loss_{cross-entropy}(\widehat{\theta'}, \theta) + Loss_{MSE}(\widehat{\theta'}, \theta) \quad (1)$$

次に，Fine Angle に関する Loss 関数を式 2 に示す.

$$Loss_{Fine} = Loss_{cross-entropy}(\hat{\theta}, \theta) + Loss_{MSE}(\hat{\theta}, \theta) \quad (2)$$

最後に，ネットワーク全体の Loss 関数を式 3 に示す.

$$Loss_{total} = Loss_{Coarse} + \lambda Loss_{Fine} \quad (3)$$

なお，Coarse Angle は$\widehat{\theta'}$，Fine Angle は$\widehat{\theta''}$，最終的に推定する Angle は$\hat{\theta} = \widehat{\theta'} + \widehat{\theta''}$，正解の視線角度は$\theta$である．そして，$Loss_{cross-entropy}$は，視線方向を複数の区間に分けて分類タスクに近似することで算出するクロスエントロピー誤差，$Loss_{MSE}$は MSE ロスである．ネットワーク全体のロス関数である$Loss_{total}$における$\lambda$は，Fine Angle のロス関数の重み係数である．提案手法では，$\lambda = 2$とし，最終的に推定する Angle が正しい視線角度に近づくように設計する．

このように提案手法は, 顔領域における疎密な情報を捉え, 視線推定手法に活用する．

## ３．３　実験

### ３．３．１　実験方法

提案手法と従来手法を Gaze360 と RT-GENE[16]のデータセットを用いて比較する．

Gaze360 は，被験者 238 人，屋内外 17.2 万枚の画像を含んでおり，視線推定のデータセットの中では大規模データセットである. そして, 屋内外で撮影されていることから, 異なる照明条件や広い視線分布の視線データを含んでおり, 多様な環境での汎用性を有する．そのため，Gaze360 は現状のデータセットの中で汎用性の高い視線推定技術の検証に最適である．

RT-GENE は，被験者 15 人，12.3 万枚の画像を含む.

評価指標には，MAE（Mean Angular Error）を用いる．MAE は視線推定誤差を定量的に評価する指標であり，異なる手法間の性能比較を可能にする．MAE は式 4 で算出する．

$$MAE = \frac{1}{N}\sum_{i=1}^{N} \arccos\left(\frac{\boldsymbol{g_i}, \widehat{\boldsymbol{g_i}}}{\|\boldsymbol{g_i}\|\|\widehat{\boldsymbol{g_i}}\|}\right) \quad (4)$$

ここで，$\boldsymbol{g_i}$は正解の視線方向ベクトル，$\widehat{\boldsymbol{g_i}}$は推定した視線方向ベクトル，単位は°である．

### ３．３．２　実験結果

提案手法と従来手法である L2CS-Net および L2CS-Net の改善手法の MAE の値を表 3.1 に示す．

表 3.1 従来手法と提案手法における MAE の比較

| Method | Gaze360 | RT-GENE |
|---|---|---|
| L2CS-Net | 10.41° | 6.59° |
| L2CS-Net の改善手法 | **10.16°** | **6.44°** |
| Ours | <u>10.19</u>° | <u>6.55</u>° |

実験結果より，提案手法は，二つのデータセットにおいて，従来手法である L2CS-Net に対して精度の向上を確認できる．また，Gaze360 データセットにおいては，L2CS-Net の改善手法と比較してわずかに精度は劣るものの，顔画像のみを用いた手法ながら遜色のない精度であることが確認できる．

## ３．４　考察

評価実験より，提案手法は顔画像のみを用いた従来手法である L2CS-Net の性能向上を達成した．また，提案手法は，Gaze360 データセットにおいて，両目の画像を用いる L2CS-Net の改善手法と遜色のない性能を示した．手法を実世界で応用するという点においては，顔画像のみを用いている提案手法の方が L2CS-Net の改善手法より実用性という点において優れていると考えられる．これらのことから，手法の実用性と精度向上のトレードオフを考慮した上で提案手法は優れていると考えられる．

一方，RT-GENE データセットでは，L2CS-Net と比較した際に，提案手法の精度向上の幅が小さい．これは，RT-GENE データセットがアイトラッキンググラスを使用して取得されたデータであり，画像修復処理によってグラスをかけている領域が補正されていることが原因として考えられる．この補正処理により目の領域がぼやけているため，提案手法が目の領域を注視して視線方向を補正する際に，不自然な部分が悪影響を及ぼしていると考えられる．

## ３．５　むすび

本章では，アピアランスベースの視線推定手法において，顔領域における疎密構造を捉えることによる改善手法を提案した．実験により，提案手法が従来手法である L2CS-Net の性能向上を示し，実用性の観点でも優れていることを示した．提案手法は，アピアランスベースの手法全般に適用可能であり，さらなる応用が期待される．

# 第４章　顔のキーポイント活用手法

## ４．１　まえがき

本章では，顔のキーポイント情報を視線推定手法の補助情報として活用する手法を提案する．提案手法の背景となる予備実験について説明したのち，提案手法と提案手法の評価実験，考察について述べる．

## ４．２　予備実験

### ４．２．１　アピアランスベースの視線推定手法への CAM 適用

アピアランスベースの視線推定手法に CAM を適用することで，視線推定モデルの可視化をする．本予備実験の目的は，従来の視線推定モデルを可視化することで，モデルの特性を捉え，改善手法の検討をすることである．本予備実験の概要を図 4.1 に示す．

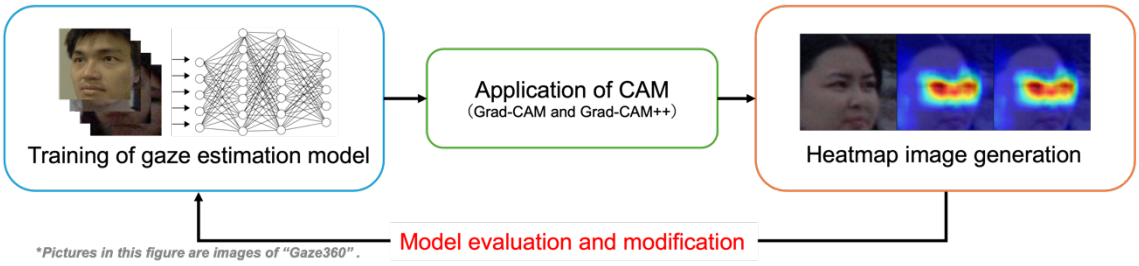


図 4.1 アピアランスベースの視線推定手法への CAM 適用の概要図

具体的には，以下の手順で視線推定モデルの特性を分析する．

- L2CS-Net モデルを Gaze360 データセットで学習する．
- 学習済みの L2CS-Net モデルに Grad-CAM および Grad-CAM++を適用し，モデルを可視化する．
- 得られたヒートマップ画像からモデルの特性を分析する．

そして，実験結果として得られたヒートマップ画像の例を図 4.2 に示す．なお，各画像は左から，入力の顔画像，Grad-CAM によるヒートマップ画像，Grad-CAM++によるヒートマップ画像である．また，各画像の角度誤差を表 4.1 に示す．

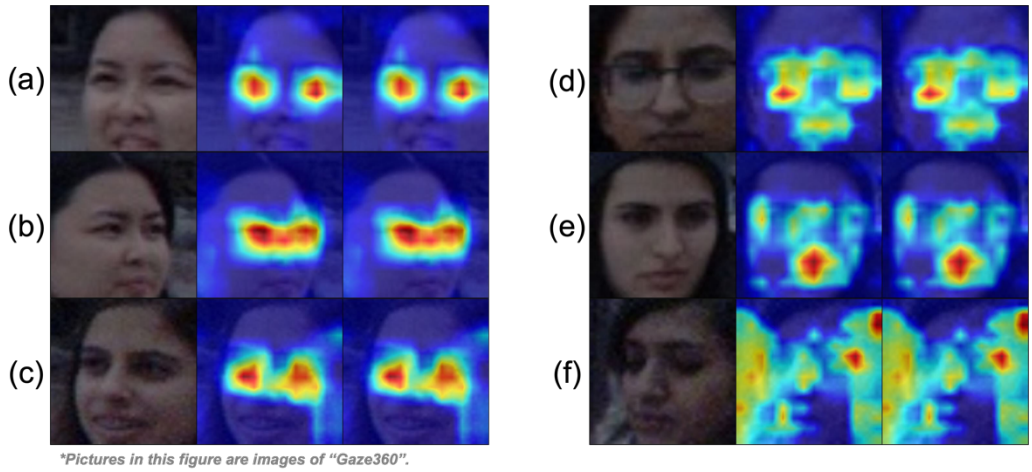


図 4.2 CAM による L2CS-Net モデルの可視化画像の例
（2024 年映像情報メディア学会年次大会にて発表）

表 4.1 各画像に対する L2CS-Net の視線推定精度

| 入力画像 | (a) | (b) | (c) | (d) | (e) | (f) |
|---|---|---|---|---|---|---|
| 角度誤差 | 8.38° | 4.72° | 0.40° | 31.74° | 41.39° | 43.55° |

これらの結果と L2CS-Net の MAE の値が10.41°であることより, 図 4.2 における(a), (b), (c) のように目の領域への Attention が高く分散が小さい場合には，推定精度が良く，(d), (e), (f) のように目以外への領域の Attention が高い場合や分散が大きい場合には，推定精度が悪いことがわかる．このことから，Attention の分布と角度誤差の間に相関がある可能性が示される．したがって，視線推定モデルにおいて，目の領域が重要であると考えられ，目の領域への Attention が高くなるようにモデルを設計することで精度向上が期待できる．

### ４．２．２　顔のキーポイント検出手法の検出精度の検証

視線推定データセットに対して，2 章で述べた三つの顔のキーポイント検出手法を適用し，検出精度を検証する．本予備実験では，Gaze360 と ETH-XGaze[17]データセットに対して，dlib と OpenPose，FaceMesh をそれぞれ適用する．

ETH-XGaze データセットは，視線推定のデータセットとしては新しく，高解像度の画像を含むという特性があり，被験者 110 人，100 万枚以上の画像を含む．

各手法の検出率と検出できた際の検出精度を表 4.2 に示す．

表 4.2 顔のキーポイント検出手法の実行結果

| 手法 | Gaze360（検出率/精度） | ETH-XGaze（検出率/精度） |
|---|---|---|
| dlib | 約 95%/約 90% | 約 60%/約 90% |
| OpenPose | 検出不可 | 約 5%/約 10% |

| | | |
|---|---|---|
| FaceMesh | 検出不可 | 約 90%/約 90% |

実験結果より，Gaze360 データセットに対して dlib 以外の検出手法では検出ができないため，提案手法の有効性確認に Gaze360 を使用せず，ETH-XGaze データセットを使用する．また，顔のキーポイント検出手法として，dlib と FaceMesh を使用し，手法間の比較をする．ETH-XGaze データセットにおいて，dlib と FaceMesh では顔のキーポイント検出できた際の精度が 9 割程度であり，検出できていない場合，顔のキーポイント情報は情報を持たないので，提案手法に用いることとする．

ここで，ETH-XGaze データセットに dlib と FaceMesh を適用して得られる顔のキーポイント画像の例を図 4.3 に示す．ただし，各画像は左から元の顔画像，dlib の検出結果，FaceMesh の検出結果である．

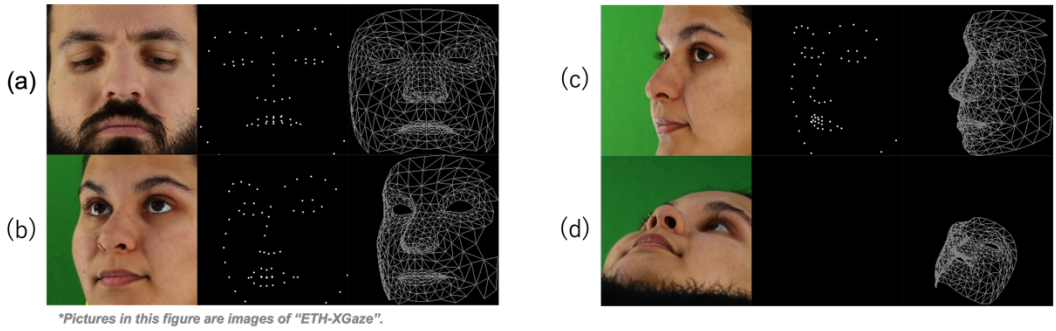


図 4.3 ETH-XGaze データセットに対するキーポイント検出結果の例

図 4.3 より，(a)，(b) はどちらの手法も検出できているのに対して，(c) では dlib が正しく検出できておらず，(d) では dlib と OpenPose が正しく検出できていないことがわかる．

## ４．３　提案手法

顔のキーポイント情報を補助情報として活用し，顔領域における重要な部位を重点的に考慮することで，精度向上を図る手法を提案する．従来のアピアランスベースの視線推定手法のうち，顔画像のみを用いる手法は，安定性に優れるが，目の領域が小さく，十分に目の情報を捉えることが難しい．そして，両目の画像を用いる手法は，目の情報を強化できるが，実世界での応用時の不安定性を抱える．そこで，提案手法は，顔画像のみを用いて安定性を維持しつつ，顔の重要部位や輪郭を示す顔のキーポイント情報により，顔全体の構造を捉えることで精度の向上を図る．また，予備実験の結果に基づき，目の領域を重視するモデル設計を採用し，顔のキーポイント情報を活用して目の情報を補完しながら，顔全体の構造的特徴もモデルに取り入れることを意図している．

提案手法の概要は以下の通りである．

- 前処理として，入力された顔画像に対して，顔のキーポイント検出手法（dlib または FaceMesh）を適用し，顔のキーポイント画像を得る．なお，キーポイント画

像に関しては，正確に検出されなかった画像や，キーポイントが全く検出できず，情報を含まない画像も含まれている．

- 元の顔画像と得られた顔のキーポイント画像をモデルの入力とし，視線方向の推定をおこなう．

提案手法のネットワーク構造を図 4.4 に示す．このネットワークは，顔全体の画像とキーポイント情報を統合的に処理することで，精度向上を図る．

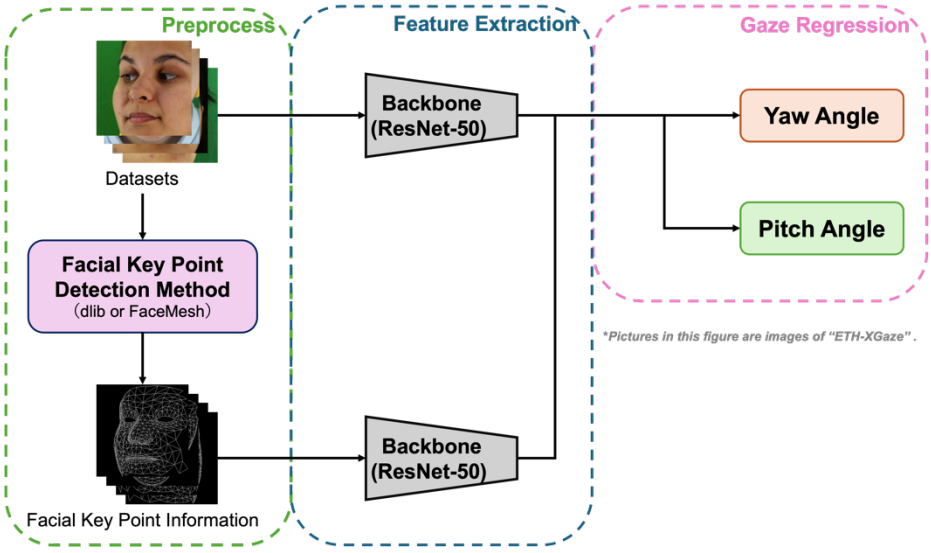


図 4.4 顔のキーポイント情報を用いた手法の視線推定ネットワーク

ここで，提案手法は，顔のキーポイント検出手法を用いるため，実世界に応用した際には，その検出手法の性能に影響を受ける可能性がある．しかし，提案手法では，キーポイント情報はあくまで補助情報であり，キーポイントの検出有無に関わらず，全てのキーポイント画像を学習データとして使用する設計である．そのため，検出精度や欠損の影響を最小限に抑えつつ，安定した視線推定が可能である．さらに，両目の画像を用いる従来手法のように目の位置を正確に検出する厳しい前提条件を必要としないため，実世界でのロバスト性が向上されると考えられる．

## ４．４　実験

### ４．４．１　実験方法

提案手法と従来手法を ETH-XGaze データセットを用いて比較する．評価指標には，式 1 で算出される MAE を用いる．そして，ETH-XGaze で学習した，L2CS-Net モデルと FaceMesh を用いた提案手法のモデルに対して，Grad-CAM と Grad-CAM++を適用し，得られたヒートマップ図を比較する．

### ４．４．２　実験結果

まず，定量的評価として，従来手法である L2CS-Net と提案手法の MAE の値を表 4.3 に示す.

表 4.3 従来手法と提案手法の視線推定結果の比較

| 手法 | MAE |
|---|---|
| L2CS-Net | 15.82° |
| Ours（w/ dlib） | 15.79° |
| Ours（w/ FaceMesh） | **15.73°** |

実験結果より，dlib と FaceMesh のどちらを用いる提案手法においても精度の向上を確認できる.

次に定性的評価として，ETH-XGaze で学習した，L2CS-Net モデルと FaceMesh を用いた提案手法のモデルに対する，Grad-CAM と Grad-CAM++によるヒートマップ図の例を図 4.5 に示す．ここで，(a) ～ (d) は L2CS-Net モデルのヒートマップ図，(e) ～ (h) は提案手法のモデルのヒートマップ図であり，各画像は左から元の顔画像，Grad-CAM による可視化画像，Grad-CAM++による可視化画像である.

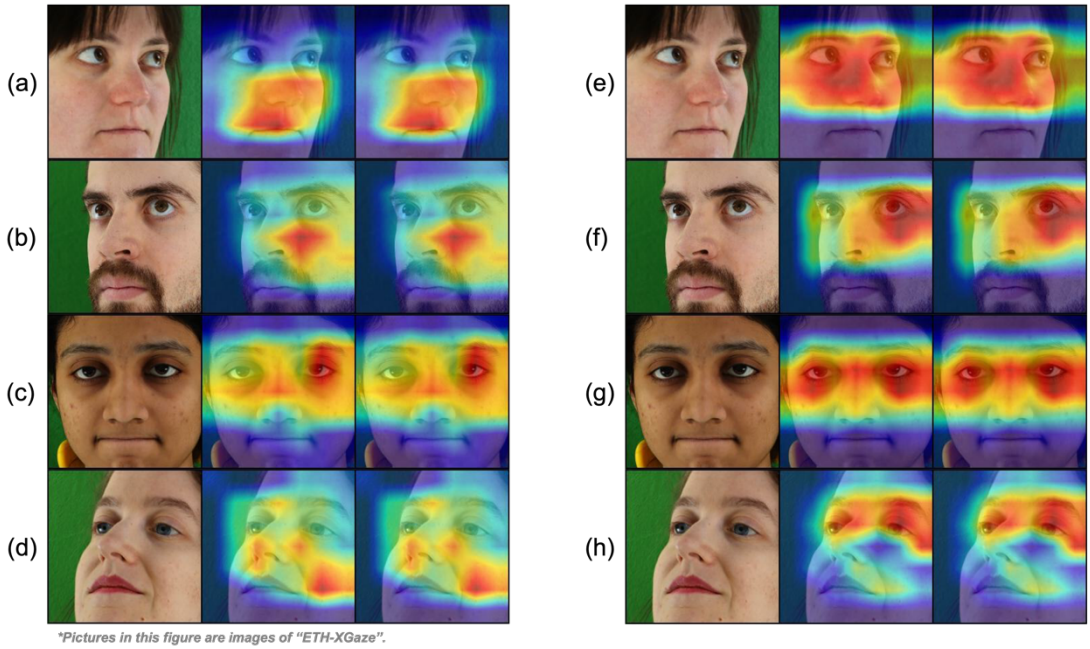


図 4.5 CAM による L2CS-Net モデルと提案手法のモデルの可視化画像の例

実験結果より，提案手法のモデルは，L2CS-Net モデルに対して，目の領域を中心として高い Attention が集まっていることがわかる.

## ４．５　考察

評価実験における定量的評価より，提案手法が従来手法である L2CS-Net の性能向上を示すことを確認した．この結果より，提案手法は，安定性を向上させることで両目の画像を用いる手法の課題を克服するアプローチを図り，実用的な視線推定手法としての可能性を示したと考えられる．さらに，提案手法で用いる二つの顔のキーポイント検出手法に関しては，詳細なキーポイント検出が可能な FaceMesh の方が有効であるとわかる．dlib は約 6 割の入力画像のみにしかキーポイント情報を付与できていないことや検出可能なキーポイント点が少ないことから改善値が小さいと考えられる．

一方で，向上した値が最大0.09°とわずかであり，不安定性の低減を図る設計が，性能向上に対しては一定の制約を与える結果となったと考えられる．このことより，安定性の向上と性能向上の間にトレードオフが存在することが示唆される．今後は，キーポイント情報の活用方法の改善を図り，このトレードオフを解消できるアプローチが必要であると考えられる．

評価実験における定性的評価より，予備実験の結果に基づく，目の領域への Attention が高いようにモデル設計することで精度向上が期待できるという仮説に対して，提案手法が意図通り設計されていることを確認した．このことより，顔のキーポイント情報を補助情報として活用することで，元の顔画像において目の領域を重視することができると考えられる．また，MAE の値より提案手法の精度向上を確認しており，目の領域の Attention が高くなるようにモデル設計することで精度向上が期待されるという仮説の妥当性を確認した．一方で，今回使用した ETH-XGaze データセットの評価用データは，正解の視線方向がラベリングされていないため，各画像に対する角度誤差を評価することができない．したがって，CAM の結果と視線推定誤差を比較できるようなデータセットを用意し，さらなる分析をする必要があると考えられる．

## ４．６　むすび

本章では，顔のキーポイント情報を活用した視線推定手法を提案した．実験より，提案手法が従来手法である L2CS-Net よりも性能向上を示し，実用的な視線推定手法に向けた提案として有効であることを確認した．顔のキーポイント情報をより有効的に活用することで精度の向上が期待される．

# 第５章　結論と今後の展望

## ５．１　結論

本研究では, 実世界での応用を考慮したアピアランスベースの視線推定手法の精度向上を目的として，大域的・局所的特徴を考慮した手法と顔のキーポイント情報を活用した手法を提案した．いずれの手法も顔画像のみを用いることで，実世界応用においての不安定性を解消し，顔の構造情報を活用することで，視線推定精度を向上させ，リアルタイムで動作可能な視線推定を実現した．

顔領域の疎密構造を捉える視線推定手法は，従来の L2CS-Net やその改善手法の課題に着目し，顔領域の粗い情報と細かい情報を二段階で捉える構造により，顔画像のみを用いながら，視線推定精度の改善を実現した．

顔のキーポイント情報を活用した視線推定手法は, 顔のキーポイント情報を補助情報として活用する手法である．CAM を用いた予備実験により目の領域への Attention を高める設計が視線推定精度を向上させる可能性を示唆した．この仮説に基に，詳細な顔の特徴点を捉えられる顔のキーポイント手法である FaceMesh を用いることで，従来手法 L2CS-Net の視線推定精度の改善を実現した．さらに，提案手法と L2CS-Net のモデルへの CAM 適用によって，仮説の妥当性を確認した．

## ５．２　今後の展望

今後の展望として，以下の三つの方向性が挙げられる．

一つ目は，本研究で提案した二つの手法が独立しているため，これらを包括的に組み込める視線推定手法の検討が必要である．これにより，モデルが扱える顔領域の情報の幅を広げることができ，精度向上につながると考えられる．

二つ目は，現状の改善手法では，わずかな改善は達成できているものの，精度の向上に限界が見られるため，それに対するアプローチを図る必要がある．さまざまな手法を検討した結果，理論的には精度向上が見込める可能性があるものの，実際にはそれを実現することができないという現状に直面している．このことから，アピアランスベースの手法における改善の限界を示唆していると考えられる．そのため，今後の研究では，複雑に絡み合った課題を個別に解決する方法の検討も必要である．例えば，目の検出精度や照明条件に特化した手法を開発し, それらを統合することで全体的な精度向上を目指すことが考えられる．

三つ目は, データセットの多様化や実世界のシナリオを考慮した訓練データを拡充することである．これにより，モデルの汎用性と精度の向上に有効であると考えられる．

# 謝辞

本研究を進めるにあたり，温かく，かつ専門的な指導をいただき，快適な研究環境を整えてくださった渡辺教授に，心から感謝申し上げます.

また，貴重なアドバイスや励ましを日々くださり，研究室で充実した環境を共に築いてくれた同期や先輩，後輩に，感謝の意を表します．皆さんと一緒に過ごした時間が，私にとってかけがえのないものとなりました.

そして，これまで支えてくださり，いつも励ましと安心を与えてくれた家族に心より感謝いたします.

# 参考文献


[1] X. Zhang, Y. Sugano, M. Fritz and A. Bulling, "Appearance-based gaze estimation in the wild," IEEE Conference on Computer Vision and Pattern Recognition, pp.511-4520, Jun. 2015.

[2] A. A. Abdelrahman, T. Hempel, A. Khalifa, A. Al-Hamadi and L. Dinges, "L2CS-Net: Fine-Grained Gaze Estimation in Unconstrained Environments," International Conference on Frontiers of Signal Processing, pp.98-102, Oct. 2023.

[3] 五十嵐裕史, 荒井直也, 熊本達也, 町田和彦, 清水雅夫, “人間は他人の視線方向をどれくらいの精度で推定できるのか,” 第 56 回日本大学理工学部学術講演会予稿集, pp.339-340, Nov. 2012.

[4] H. Sugiyama and H. Watanabe, “Enhancing Gaze Estimation through Fine-Grained Analysis of Eye Region,” International Workshop on Advanced Image Technology, Jan. 2024.

[5] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik and A. Torralba, "Gaze360: Physically Unconstrained Gaze Estimation in the Wild," IEEE International Conference on Computer Vision, pp.6911-6920, Oct. 2019.

[6] J. Deng, J. Guo, E. Ververas, I. Kotsia and S. Zafeiriou, "RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild," IEEE Conference on Computer Vision and Pattern Recognition, pp.5202-5211, Jun. 2020.

[7] Z. Chen and B. E. Shi, “Appearance-Based Gaze Estimation Using Dilated-Convolutions,” Asian Conference on Computer Vision, pp.309-324, Dec. 2018.

[8] Y. Cheng, S. Huang, F. Wang, C. Qian and F. Lu, “A Coarse-to-Fine Adaptive Network for Appearance-Based Gaze Estimation,” Proceedings of the AAAI Conference on Artificial Intelligence, pp.10623-10630, Feb. 2020.

[9] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba, "Learning Deep Features for Discriminative Localization," IEEE Conference on Computer Vision and Pattern Recognition, pp.2921-2929, Jun. 2016.

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," IEEE International Conference on Computer Vision, pp.618-626, Oct. 2017.

[11] A. Chattopadhay, A. Sarkar, P. Howlader and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," IEEE Winter Conference on Applications of Computer Vision, pp.839-847, Mar. 2018.

[12] A. Khosla, N. Jayadevaprakash, B. Yao and L. Fei-Fei, “Novel Dataset for Fine-Grained Image Categorization: Stanford Dogs,” Proceeding IEEE Conference on Computer Vision and Pattern Recognition Workshop on Fine-Grained Visual Categorization, Jun. 2011.

[13] “dlib C++ Library,”(最終閲覧日：2025 年 1 月 10 日）, http://dlib.net

[14] Z. Cao, T. Simon, S. -E. Wei and Y. Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields," IEEE Conference on Computer Vision and Pattern Recognition, pp.1302-1310, Jul. 2017.

[15] Google, “Face mesh detection | ML Kit,” （最終閲覧日：2025 年 1 月 10 日）, https://developers.google.com/ml-kit/vision/face-mesh-detection

[16] T. Fischer, H. J. Chang and Y. Demiris, “RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments,” European Conference on Computer Vision, pp.339-357, Sep. 20 18.

[17] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang and O. Hilliges, “ETH-XGaze: A La rge Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation,” European Conference on Computer Vision, pp.365-381, Aug. 2020.

# 図一覧

# 表一覧

# 研究業績


[1] 杉山秀治, 渡辺裕, “目の特徴量を付与したアピアランスベースの視線推定モデルの検討”, 電子情報通信総合大会, D-12-52, Mar. 2023.

[2] H. Sugiyama and H. Watanabe, “Enhancing Gaze Estimation through Fine-Grained Analysis of Eye Region,” International Workshop on Advanced Image Technology, No.11, Jan. 2024.

[3] 杉山秀治, 福田大翔, 渡辺裕, ”アピアランスベースの視線推定における実現可能性を考慮した精度向上”, 情報処理学会全国大会, 5S-04, Mar.2024.

[4] 福田大翔, 杉山秀治, 渡辺裕, ”ホモグラフィ変換に基づくホームベースの大きさ情報を活用したストライクゾーン取得精度の向上”, 情報処理学会全国大会, 7U-02, Mar. 2024.

[5] 杉山秀治, 福田大翔, 中山光典, 渡辺裕, “Class Activation Mapping を用いた視線推定手法の評価および有効性の確認”, 映像情報メディア学会年次大会, 11A-4, Aug. 2024.

[6] 中山光典, 杉山秀治, 福田大翔, 渡辺裕, “主成分分析を用いたエッジの移動抑制による点群平滑化の改善手法”, 映像情報メディア学会年次大会, 13-B, Aug. 2024.

[7] 福田大翔, 中山光典, 杉山秀治, 渡辺裕, “自動ボールストライク判定におけるフロートラッキングを用いた処理量削減法”, 映像情報メディア学会年次大会, 11A-5, Aug. 2024.