Human-like Guidance のための Multimodal TranSalNet による 走行方向に合わせた視線情報を用いた顕著性マップ推定

難波田 雅己† 平川 翼† 山下 隆義† 藤吉 弘亘† 大久保 翔太‡ 南里 卓也‡ 王 思源‡
†中部大学 ‡日産自動車株式会社
E-mail: masaknanbt@mprg.cs.chubu.ac.jp

1 はじめに

自動車の目的地までの誘導を行う運転支援には，既にカーナビゲーションシステムが広く採用されている．既存のカーナビゲーションでは，地図データと自車の位置情報を用いた案内が主である．そのため，経路上の交差点までの距離や交差点名，地図データに含まれるランドマークを基に案内文を生成する．しかし，ドライバにとって距離や交差点名を用いた案内は，運転中にこれらの情報を確認するために，車載モニタの地図を視認する場合がある．これは，脇見運転を誘発している状態であり，事故につながる要因である．この問題を解決する次世代のナビゲーションとして，Human-like Guidance (HLG) の実現が期待されている．HLG は，人間が行うようなナビゲーションを実現することを目的としたものである．

HLG の先行研究 [1] では，ドライバの視線をシーン画像から予測し，ドライバが注視している物体を基準とした案内文を生成する．先行研究では，CNN を用いてドライバの視線を推定している．CNN の畳み込み処理では，画像内の局所的な特徴を獲得することが出来るが，画像内の文脈的特徴を獲得することが困難である．また，CNN のダウンサンプリング操作によって，特徴が喪失される可能性がある．一方で，Transformer[2] を用いた場合，Self-Attention 機構や位置埋め込み処理によって，画像内の大局的特徴や，文脈的特徴を獲得することが出来る．ここで，人間の視線情報には個人差があるため，同一の画像であっても様々な位置に点在している [3]．運転中においては，周辺環境を確認するために，近距離に存在する物体や，遠方に存在する車両や歩行者，信号機等の様々な物を注視している [4, 5]．また，ドライバは同じ周辺環境であっても，直進時や右左折時など走行方向によって注視点に違いが生じる．そこで本研究では，走行方向の情報を言語情報として視線推定モデルに画像と共に入力する，マルチモーダル Transformer モデルを提案する．これにより，走行方向毎の視線情報の特徴を獲得し，案内文生成の精度向上を図る．

また，先行研究では視線推定モデルを学習するデータセットである Driving Gaze Datasets (DGD) を提案しており，CARLA シミュレータ [6] を用いて作成したドライバ視点の走行映像，フレーム毎のシーン状況を表す走行方向情報，視線情報から構成されている．先行研究の評価では，視線推定精度が十分でないという課題を指摘している．その原因には，DGD は視線情報特有のセンターバイアスが考慮されていない点や，シミュレータの自動走行機能により走行映像データを収集しているため，ナビゲーションに不要なシーンの割合が大きい点が考えられる．そこで本研究では，先行研究で提案された DGD を分析し，DGD の持つ課題を解決するデータセット DGD-V2 を構成する．

2 関連研究

本章では，既存研究における視線推定手法と，視線推定データセットについて述べる．

2.1 視線推定手法

従来の視線推定手法には，静止画に対する画像顕著性を予測する手法と [7, 8, 9, 10]，動画像に対する映像顕著性を予測する手法に大別できる [11, 12]．動画像に対する映像顕著性を予測する手法は，時間情報を扱うためモデルサイズが大きく，計算に時間がかかる．そのため，リアルタイム性を重視するカーナビゲーションには適していない．よって本研究では，画像顕著性予測手法をベースとする．画像顕著性予測手法は，CNN ベースの手法が一般的であるが，人間の視線情報は画像上の様々な位置に点在するため，局所的な特徴を捉えることに特化した CNN は適していない．そのため，2 つの CNN モデルを並列に運用する手法 [9] や，カーネルサイズが異なるブランチを用いる手法 [10] 等，マルチスケール特徴を獲得することで精度を向上させる手法が多く提案されている．一方で，Transformer[2] を活用することで，CNN では捉えることが困難であった点在する人間の視線情報の特徴の獲得が容易となり，関連手法が提案されている [7, 8]．

2.2　ドライバの視線推定データセット

ドライバの視線推定には，ドライバの顔の向きや目の動きから視線を推定するアピアランスベースと，シーン画像のみから視線を推定するシーンベースの 2 つに大別される．アピアランスベースの手法 [13, 14, 15, 16] はドライバを撮影するカメラと自車の外側を撮影するカメラが必要であるためカメラ間のキャリブレーションのコストがかかる点や，複数のカメラを用意するコストが必要である．一方で，シーンベースの手法は [17, 18]，車両前方の映像のみからドライバが注視するであろう位置を推定するため，車両への実装コストや複雑な処理を必要としない．本研究で扱う視線推定タスクは，シーンベースでのドライバの視線推定である．これまでに，シーンベースの視線推定タスクのためのデータセットが提案されているが [1, 17]，本研究で対象とするカーナビゲーションに特化したデータセットは，先行研究で提案された Driving Gaze Datasets (DGD) のみである [1]．しかし，先行研究で提案された DGD で学習した視線推定モデルは精度が低いという問題点がある．そこで，本研究では視線推定精度の向上を図るために，DGD を分析する．また，その課題を解決するデータセットの作成を行う．

3　Driving Gaze Datasets-V2

本研究では，提案手法である Multimodal TranSalNet の学習，評価を行うデータセットを作成する．2.2 節で述べたように，本研究で対象とするカーナビゲーションに特化したデータセットは，先行研究で提案された Driving Gaze Datasets (DGD) のみである．しかし，DGD は視線情報特有のセンターバイアスが考慮されていない．また，シミュレータの自動走行機能を利用して走行映像データを収集しているため，ナビゲーションに不要なシーンのデータの割合が大きい．これらは，先行研究の視線推定精度が低い原因であると考えられる．そのため，本章では DGD について詳細に分析することでその特性を明らかにし，DGD の課題を解決した DGD-V2 の作成を行う．

3.1　DGD の分析

DGD は，走行映像，視線情報，および走行方向の情報が含まれている．ここで，視線情報には 8 人分が含まれている．本節では，走行方向ごとの視線情報の分布を分析する．先行研究では，走行方向を直進，右左折，右左折手前，信号待ちによる停止，それ以外のシーン（車線追従）の合計 7 つが定義されている．

図 1 に走行方向毎のデータ数，図 2 に，走行方向毎の視線情報の分布を可視化した結果を示す．図 1 より，走行方向毎のデータ数に大きな偏りがあり，不均衡なデータセットであることが確認できる．次に，図 2 より，交差点付近のシーンである直進シーン，左折シーン，左折手前シーン，停止シーンで注視点のバラツキが大きいことが分かる．これらのシーンは，周辺の車両や信号，進行先を確認する動作が頻繁に起きるため，バラツキが生じると考えられる．一方で，右折シーン，右折手前シーンは交差点付近のシーンであるにも関わらず，データのバラツキが小さい．これは，シミュレータ環境が右車線走行であるため，進行先を確認する際に大きな視線の移動が生じないためであると考えられる．

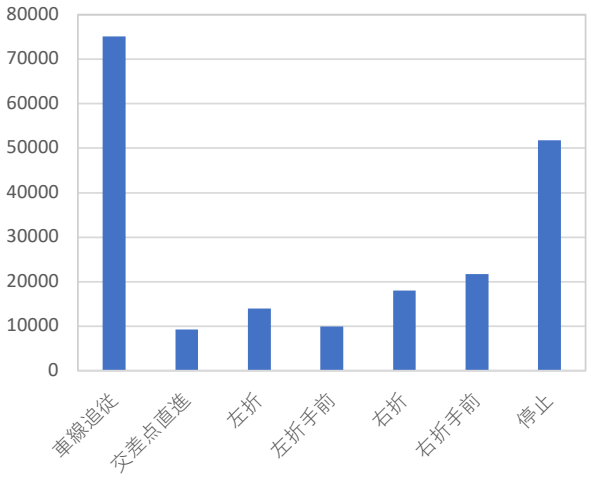

図 1　DGD における走行方向ごとのデータ数

次に，走行映像上に視線情報を重畳し，ドライバが何を注視しているのかを分析する．図 3 に，走行映像上に視線情報を重畳した際の例を示す．交差点付近のシーンでは，図 3(a) および図 3(b) のような，車両や信号等の物体を確認する視線や進行先を確認する視線が確認できる．一方で，図 3(c) および図 3(d) のような，注視先に何も情報がない視線が確認できる．これは，ドライバが周辺視野を用いて周囲の状況を視認しているシーンであると考えられる．また，図 3(c) および図 3(d) の視線は，画像中央付近に現れている．この視線の存在する領域は，図 2 に示すように，全ての走行方向において最も視線の分布が集中する領域である．これらの結果より，図 3(c) および図 3(d) のような視線は，センターバイアスの最も大きな原因であると考えられる．

次に，DGD では，走行映像 1 フレームに対して複数人の視線情報を収集している．学習時には，1 フレームに対して 1 人分の視線情報をペアとして正解データを作成している．つまり，同一の画像に対して複数パターンの真値を用いて学習していることになる．これは，先行研究 [1] ではナビゲーション対象を視線推定によって一意に定める必要があるためである．しかし，この学習方法は個人差が大きい視線情報に対してモデルの学習を困難にしている．

本節で明らかになった DGD の特性を以下に示す．

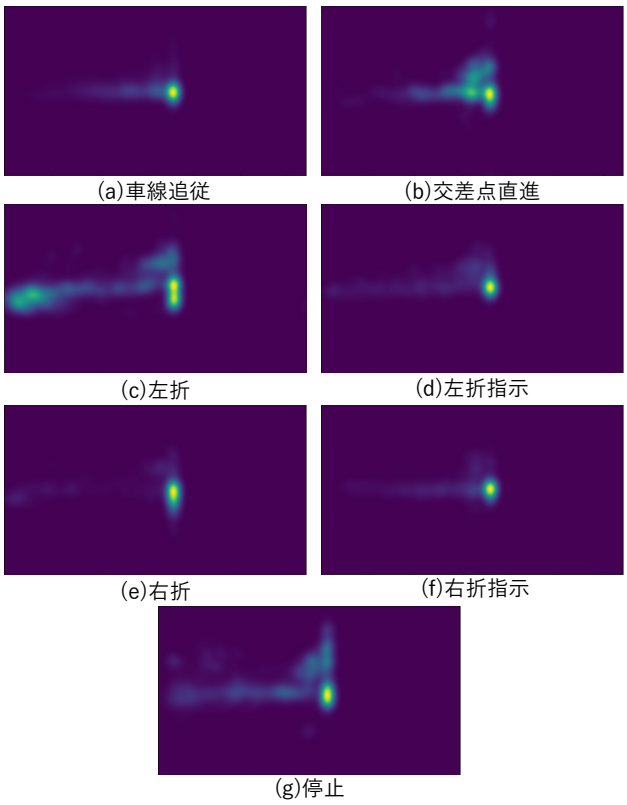

図 2　走行方向毎のデータ分布

- 走行方向毎のデータ数に大きな偏りがあり，不均衡なデータセットである．
- 右車線走行であるため，右折シーン，右折手前シーンは交差点付近のシーンであるにも関わらず視線分布のバラツキが小さい．
- 視線の分布が集中する領域があり，センターバイアスの最も大きな原因である．
- 1 フレームに対して 1 人分の視線情報のみを真値データに設定しており，学習を困難にしている．

3.2　DGD-V2 の作成

3.1 節によって明らかにした DGD の特性から，DGD の課題を解決した DGD-V2 を作成する．初めに，DGD はナビゲーションにおいて不必要な車線追従シーンのデータ数が最も多い．また，車線追従シーンは 3.1 節で明らかとなった，センターバイアスの最も大きな原因であると考えられる領域に視線情報が集中している．よって，車線追従シーンは不必要であると言える．

次に，DGD では右折シーン，右折手前シーンのデータ分布が，交差点付近のシーンであるにも関わらずバラツキが小さくという問題点がある．そこで，左折シーン，左折手前シーンを左右反転させることでデータ拡張を行い，データのバラツキが大きい右折シーンを作成する．

次に，先行研究 [1] では，ナビゲーション対象を視線推定によって一意に定める必要があるため，1 フレームに対して 1 人の視線情報をペアとして正解データを作成している．この学習方法は，個人差が大きい視線情報において，モデルの学習を困難にしている．そのため，DGD-V2 では，1 フレームに対して複数人の視線情報を真値として配置する．これらの処理を行うことで，DGD-V2 を作成する．

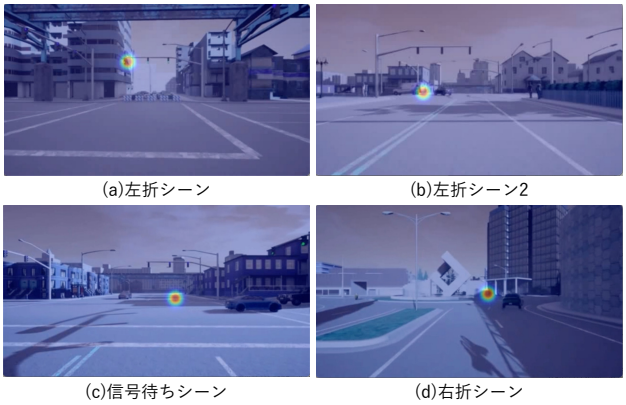

図 3　走行映像上に可視化した視線情報の例

4　提案手法

視線推定を用いた Human-like Guidance (HLG) 実現のためには，ドライバの視線を推定する精度が重要である．また，ドライバは同じ周辺環境であっても走行方向毎に注視点に違いが存在する．しかし，既存の視線推定モデルは画像やコンテキスト情報のみしか考慮しておらず，走行方向毎の特徴を獲得することは困難である．そこで，本研究では代表的な Transformer ベースの視線推定モデルである TranSalNet[8] をマルチモーダル入力に拡張した，Multimodal TranSalNet を提案する．

Multimodal TranSalNet の概要図を図 4 に示す．まず，ResNet-50 を用いて画像特徴量を抽出し，さらに Transformer エンコーダによって文脈的な特徴表現を獲得する．走行方向情報は，RoBERTa[19] を用いて言語特徴量として抽出される．抽出した画像特徴量と言語特徴量を結合し，線形層を用いて圧縮することで合成特徴量を獲得する．最後に，デコーダに合成特徴量を入力することで，走行方向に応じた視線を獲得する．以下に，各要素の詳細について述べる．

4.1　画像特徴量の獲得．

ResNet-50 と Transformer エンコーダを用いて画像の特量抽出を行う．初めに，凍結された事前学習済みの ResNet-50 を用いて，特徴抽出を行う．ResNet-50 の畳み込み層の内，最終層から 3 層分の特徴マップ (より深い層から $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, $\boldsymbol{x_3}$ とする) を利用する．これにより，マルチスケールな特徴表現を獲得する．次に，それぞれの特徴マップに 1×1 の畳み込み処理を行い，$\boldsymbol{x_1}$ お

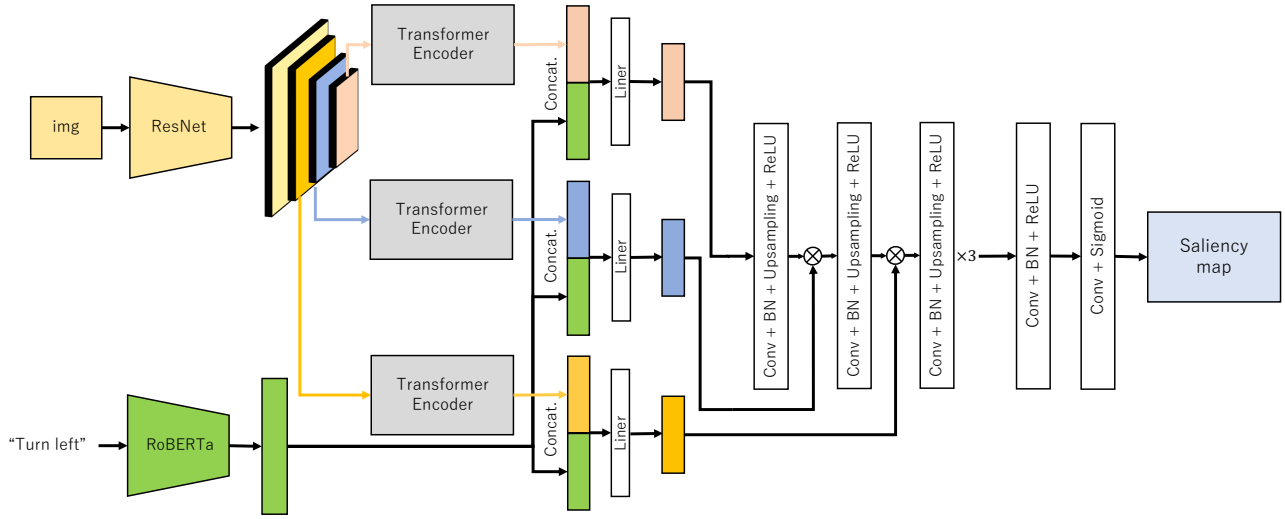

図 4　Multimodal TranSalNet の概要図

よび $\boldsymbol{x_2}$ は 768 次元，$\boldsymbol{x_3}$ は 512 次元に変形すると共に，計算量を削減する．変形したそれぞれの特徴マップを平坦化して Transformer エンコーダに入力する．この際に，特徴マップ内の位置情報を考慮するために，位置埋め込みとパッチ埋め込みを行う．これによって，相対的な位置関係や局所的な構造といった，文脈的な特徴表現を獲得する．

4.2　言語特徴量の獲得．

Multimodal TranSalNet では，走行方向情報を言語特徴量として入力する．言語情報として入力することで，“Turn right”や “Turn left”のように，“Turn”という共通する文脈的な特徴を考慮することが可能となる．また，言語情報が “Turn right at the second intersection.”のように長文の場合，より柔軟な走行方向の表現の特徴も獲得できる．DGD の走行方向データを基準とし，右折，右折手前シーンでは “Turn right”，左折，左折手前シーンでは “Turn left”，交差点直進シーンでは “Go straight”，停止シーンでは “Stop” を入力する．言語特徴量は，凍結された事前学習済みの RoBERTa[19] を用いて抽出する．RoBERTa は，Word2vec[20] や GloVe[21] のような単語埋め込みのフレームワークとは異なり，“Stop”のような単一の単語と，“Turn left”のような複数の単語で構成された物の両方を，d 次元の固定長の特徴量に変換することができる．

4.3　合成特徴量の獲得．

獲得した 3 つの画像特徴量 $(\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3})$ と言語特徴量を合成した，合成特徴量の獲得する．RoBERTa の出力次元数 d を 768 とし，言語特徴量と画像特徴量をチャンネル方向に結合する．結合した 1536 次元の特徴量を線形層によって 768 次元に圧縮し，合成特徴量 $\boldsymbol{jx_1}, \boldsymbol{jx_2}$ を獲得する．この時，$\boldsymbol{x_3}$ に対しては，言語特徴量を線形層によって 512 次元に変形し，$\boldsymbol{x_1}$ および $\boldsymbol{x_2}$ と同様の結合処理を行うことで 512 次元の合成特徴量 $\boldsymbol{jx_3}$ を獲得する．本手法では，他のマルチモーダル手法 [22, 23] と異なり，エンコーダで抽出した画像特徴量に対して言語特徴量を融合する．これによって，同一の周辺環境であっても，走行方向毎に生まれる注視点の分布の差異を学習することが期待される．

4.4　デコーダ．

獲得した合成特徴量をデコーダに入力し，入力画像と同じサイズの顕著性マップを獲得する．デコーダは 7 つの畳み込みブロックからなる CNN デコーダである．1~6 層目までは，3×3 の畳み込み処理とバッチ正規化 (BN) 処理があり，前者は非線形な特徴表現の獲得，後者は収束の促進に貢献している．7 層目では畳み込み処理を用いて特徴マップを 1 次元に変換し，Sigmoid 関数を適用することで，ピクセル単位の確率分布としての損失計算を可能としている．3 つの合成特徴量は，最初の 2 層のそれぞれの出力に対してスキップコネクションで接続されており，乗算することで融合される．これらによって，マルチスケール特徴や文脈的特徴を考慮した顕著性マップを獲得する．

5　評価実験

本章では，DGD-V2 を用いて，提案手法の有効性を評価する実験を行う．従来の TranSalNet による実験結果と，Multimodal TranSalNet よる実験結果を比較することで，本手法の有効性を検証する．また，実環境データに対する実験を行い，本手法の実用性を検証する．評価指標には，視線推定タスクにおいて一般的に用いられる，Correlation Coefficient (CC) と Normalized Scanpath Saliency (NSS) を用いる．CC は，出力された顕著性マップを確立分布として捉え，真値の分布との線形相関を評価する評価指標である．NSS は，顕著性マップと真値の座標を評価する評価指標である．CC は，

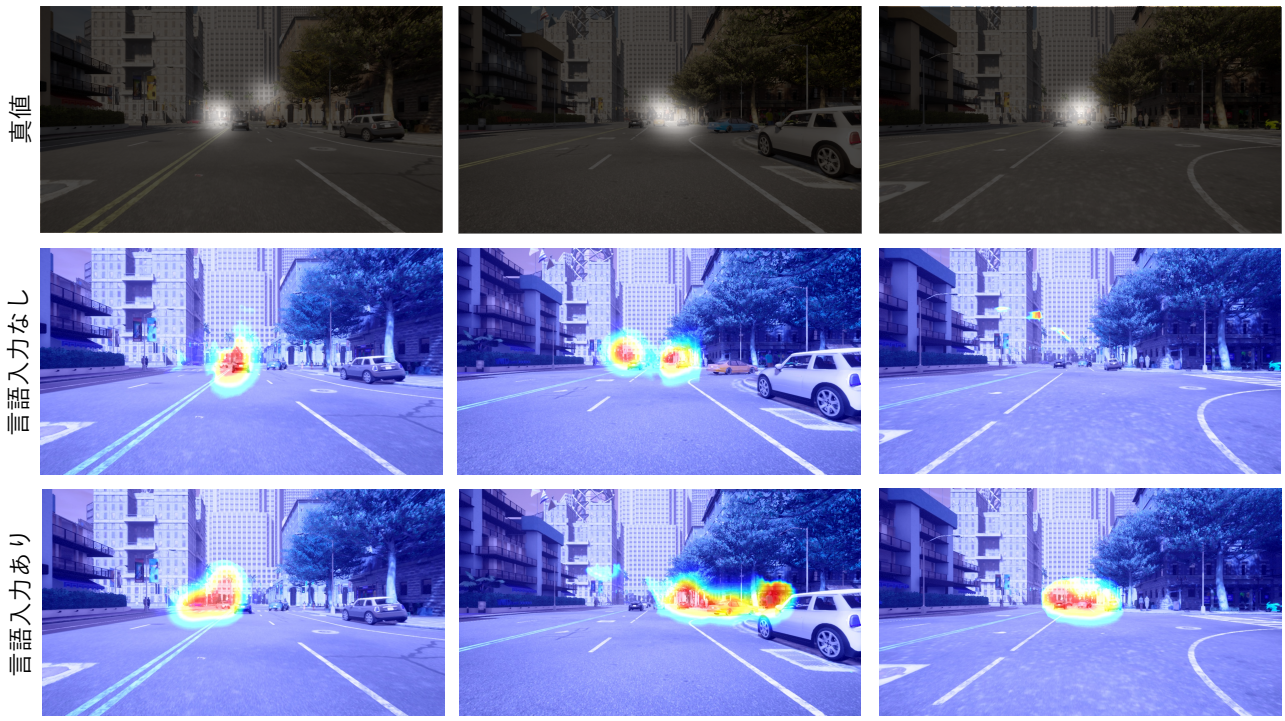

図 5　言語入力の有無による精度変化の定性的比較

表 1　言語入力の有無による精度変化

	CC	NSS
言語入力なし	0.7290	1.4679
言語入力あり	0.7554	1.5136

ピクセル単位での画素値の類似度を評価できるが，偽陽性に対する評価がしづらい特性を持つ．一方で，NSS は出力結果に対して標準化処理を行うため，偽陽性に対して敏感であるが，ピクセル単位での画素値の類似度の評価ができない特性を持つ．

5.1　Multimodal TranSalNet の検証

本実験では，従来の TranSalNet と Multimodal TranSalNet を用いて，言語入力の有無による精度変化を定量的，定性的に比較し，本手法の有効性を検証する．

表 1 に f 定量的評価結果，図 5 に定性的評価結果を示す．表 1 より，言語情報を入力して学習することで，精度が向上することを確認した．また，図 5 より，言語情報を入力することで，推定結果が真値の視線分布に近似することを確認した．それぞれのシーンにおいて，言語情報を入力する事で，左折シーンでは左前方，右折シーンでは右前方の物体，直進シーンでは正面の物体を注視する傾向が確認できる．これらの結果は，言語情報を入力することで走行方向毎の視線の特徴を学習できているためであると考えられる．よって，走行方向情報を言語情報として入力することは有効である

図 6　実環境データに対する推定結果

と言える．

5.2　実環境データに対する評価実験

本研究の目的は，カーナビゲーションシステムである Human-like Guidance (HLG) の実現である．そのため，本章では本研究で作成したモデルを用いて実環境データに対する評価実験を行う．

図 6 に実環境データに対する実験結果を示す．実験結果より，シミュレータ環境下での実験と同様の傾向を確認した．左折シーンである図 6(a) では，左前方に存在する左折中の車両を注視している．右折シーンである図 6(b) では，自車両の進行先である右前方に存在する車両を注視している．交差点付近に存在する車両を注視していることから，これらの結果は HLG に活用可能な結果であると言える．

5.3　DGD-V2 の有効性の検証

本研究で作成した DGD-V2 は，先行研究で提案された DGD を基に，その課題を解決したデータセットである．本章では，先行研究で用いられていた DGD で学

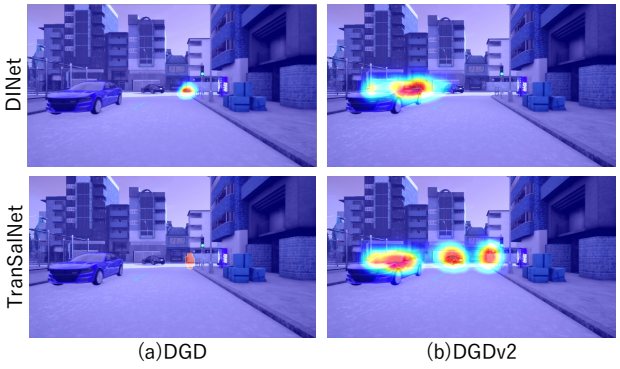

図 7　DGD-V2 による精度変化の定性的比較

表 2　DGD-V2 による精度変化の定量的比較

	DGD	DGD-V2	CC	NSS
DINet	✓		0.4283	1.2752
		✓	0.7004	1.3655
TranSalNet	✓		0.4122	1.2877
		✓	0.7290	1.4679

習したモデルと，DGD-V2 で学習したモデルの精度を比較し，DGD-V2 の有効性を明らかにする．視線推定モデルには，先行研究で用いられた DINet[10] と，提案手法のベースとなったモデルである TranSalNet[8] を用いる．

表 2 に，定量的評価結果，図 7 に定性的評価結果を示す．表 2 より，DGD-V2 を用いた場合に大幅に精度が向上していることが確認できる．図 7 より，DGD で学習したモデルはどちらもセンターバイアスの影響を受け，中央付近かつ物体が無い空間を注視している．一方で，DGD-V2 で学習したモデルは，同じシーンでも周辺車両を注視していることが確認できる．これらの結果より，DGD-V2 は有効であると言える．

6　おわりに

本研究では，Human-like Guidance (HLG) における視線推定精度の向上を目標とし，視線推定モデルである Multimodal TranSalNet を提案した．また，先行研究で提案されたデータセットの課題を解決した，Driving Gaze Datasets V2 (DGD-V2) の作成を行なった．DGD-V2 を用いた実験から，走行方向情報を言語情報として入力する Multimodal TranSalNet により，従来手法より精度が向上することを確認した．また，本研究で作成したモデルを用いて，実環境データに対する評価実験を行った．実験により，実環境データに対しても前方の車両に対する視線を獲得し，HLG への有効性を確認した．今後は，モデルに入力する言語表現の拡張や，LLM を活用した，視線推定結果を元に案内文を生成する手法についての検討を行う．

参考文献

[1] Masaki Nambata, Kota Shimomura, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Human-like guidance with gaze estimation and classification-based text generation. *26th IEEE International Conference on Intelligent Transportation Systems*, 2023.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *In Advances in Neural Information Processing Systems (NIPS)*, 30, 2017.

[3] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. *Computer Vision and Pattern Recognition Conference (CVPR)*, 2015.

[4] Toru Hagiwara and Terutoshi Kaku. Measurement and evaluation of driver's eye-movement. *Cinfrastructure plannung review*, 6:121–128, 1988.

[5] Geoffrey Underwood, Peter Chapman, Neil Brocklehurst, Jean Underwood, and David Crundall. Visual attention while driving:sequences of eye fixations made by experiencedand novice drivers. *Ergonomics*, 46:629–646, 2003.

[6] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. *Proceedings of Machine Learning Research (PMLR)*, 78:1–16, 2017.

[7] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. *International Conference on Computer Vision (ICCV)*, 2021.

[8] Jianxun Lou, Hanhe Lin, David Marshall, Dietmar Saupe, and Hantao Liu. Transalnet: Towards perceptually relevant visual saliency prediction. *Neurocomputing*, 494, 2022.

[9] Sen Jia and Neil D. B. Bruce. Eml-net:an expandable multi-layer network for saliency prediction.

[10] Sheng Yang, Lin, Guosheng, Jiang, Qiuping, Lin, and Weisi. A dilated inception network for visual saliency prediction. *IEEE Transactions on Multimedia*, 22(8):2163–2176, 2019.

[11] Ziqiang Wang, Zhi Liu, Gongyang Li, Yang Wang, Tianhong Zhang, Lihua Xu, and Jijun

Wang. Spatio-temporal self-attention network for video saliency prediction. *arXiv:2108.10696*, 2022.

[12] Shiping Zhu Qinyao Chang. Temporal-spatial feature pyramid for video saliency detection. *arXiv:2105.04213*, 2021.

[13] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. *Computer Vision and Pattern Recognition Conference (CVPR)*, 2018.

[14] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. *Computer Vision and Pattern Recognition Conference (CVPR)*, 2016.

[15] Xucong Zhanga, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. *Computer Vision and Pattern Recognition Conference (CVPR)*, 2015.

[16] Isaac Kasahara, Simon Stent, , and Hyun Soo Park. Look both ways: Self-supervising driver gaze estimation and road scene saliency. *European Conference on Computer Vision (ECCV)*, pages 126–142, 2022.

[17] Andrea Palazzi, Davide Abati, Simone Calderara, Francesco Solera, and Rita Cucchiara. Improving driver gaze prediction with reinforced attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7), 2019.

[18] Kai Lv, Hao Sheng, Zhang Xiong, Wei Li, and Liang Zheng. Improving driver gaze prediction with reinforced attention. *IEEE Transactions on MultiMedia (TMM)*, 23, 2021.

[19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692*, 2019.

[20] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *International Conference on Learning Representations (ICLR)*, 2013.

[21] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014.

[22] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr - modulated detection for end-to-end multi-modal understanding. *International Conference on Computer Vision (ICCV)*, 2021.

[23] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.