Vol.2023-CVIM-233 No.57
2023/3/3

歩行者の注視対象データセットの構築

村上 大斗[1,a)]　出口 大輔[1]　平山 高嗣[2,1]　川西 康友[3,1]　村瀬 洋[1]

概要：本報告では，画像中の歩行者が見ている対象（以下，注視対象）を推定するタスクのためのデータセット構築，及びベースラインとなる注視対象推定手法を提案する．歩行者の注視対象の認識は，将来の歩行者の行動を予測するための手がかりとなり，自動運転車両等の実現において重要な技術となる．そこで本報告では，既存の交通環境データセットに対して新たに歩行者の注視点をアノテーションし，注視対象推定タスクにも利用可能なデータセット構築について報告する．加えて，構築したデータセットを用いた，骨格情報を利用して注視対象を推定する単純な手法を提案し，25.5%の正解率が得られることを確認した．

1. はじめに

画像中の歩行者が見ている対象（以下，注視対象）の推定は，歩行者の将来の行動を予測するための重要な手がかりである．歩行者の行動を予測することにより，歩行者のリスクに応じたドライバーへの注意喚起や，自動運転車両の走路計画，といった様々な応用に利用することができる．図 1 は車載カメラで撮影した画像中の歩行者の注目対象を示した例である．図からわかるように，青枠で囲った歩行者は，赤枠で囲った対向車両が注視対象であると推定できる．この場合，歩行者は自車両に気づいていない可能性が高く，飛び出してくる危険性が想定される．このように，歩行者の行動予測は事故の未然防止に役立てることができる重要な技術である．

一般に，機械学習ベースの注視対象推定手法の開発のためには，大規模なデータセットが必要である．Tomas ら [1] は，小売店における顧客と注視対象商品を対応付けた Gaze On Objects (GOO) データセットを構築している．これは小売店環境という限られた環境でのデータセットであり，人物と物体の距離や，物体の密集度などは交通環境と大きく異なっている．一方，交通環境理解のためのデータセットは複数公開されており，車載カメラ画像に写る車両や歩行者など，物体検出やセマンティックセグメンテーション用のアノテーションが収録されているものが多い [2,3]．しかし，我々の知る限りにおいて歩行者の注視対象がアノテーションされている交通環境データセットは存在しない．

そこで本報告では，既存の交通環境データセットを拡張し，歩行者の注視対象をアノテーションすることで注視対象推定タスクにも利用可能なデータセットの構築について報告する．具体的には，車載カメラ画像に写る歩行者それぞれに対して，3 人のアノテータが注視点を人手でアノテーションすることにより，人による判断の曖昧さを表現できるデータセットを構築した．加えて，このデータセットを用いた歩行者の注視対象推定タスクの単純な手法を提案する．車載カメラ画像に写る歩行者は，距離に応じて解像度が低下するため，顔画像を用いた視線推定は困難である．そこで，Hata ら [4] が提案した骨格情報を利用したアイコンタクト検出手法を参考にし，歩行者の骨格情報を基に視線方向を推定し，視線上から最も近い物体を選択する単純な手法を提案した．これをベースライン手法とし，今後研究を発展させる．

図 1: 歩行者注視推定タスクのイメージ

2. 関連研究

2.1 歩行者を収録したデータセット

自動運転タスク向けに，実際の交通環境で撮影されたデータセットが公開されている．Holger ら [2] は，物体検出向けに Bounding Box (BBox) と物体のクラスラベルをアノテーションした nuScenes と nuImages を公開している．これらのデータセットでは，車載カメラ画像中の自動車，自転車，歩行者などに対して BBox と 23 クラスの物体

1 名古屋大学 大学院情報学研究科
2 人間環境大学 環境科学部
3 理化学研究所情報統合本部 ガーディアンロボットプロジェクト
a) murakamih@vislab.is.i.nagoya-u.ac.jp

ラベルがアノテーションされている．しかし，歩行者が見ている方向や対象など，歩行者の状態についてはアノテーションされていない．また，Sun ら [3] は Waymo Open Dataset (Waymo) を公開しており，nuScenes や nuImages と同様に物体の BBox とクラスラベルがアノテーションされている．しかしながら，歩行者の状態についてはアノテーションされていない．

一方，歩行者の行動予測を目的として構築されたデータセットに Pedestrian Intention Estimation dataset (PIE dataset) [5] がある．このデータセットでは，車載カメラで撮影した 1,842 人の歩行者に対し，ID や BBox，道路を横断するかどうか，カメラ方向を見ているか否か，などの情報がアノテーションされている．しかし，歩行者が見ている方向や対象など，注視対象推定タスクに必要な歩行者の状態はアノテーションされていない．

2.2 注視対象推定

Wang ら [6] は，小売店環境における注視対象の推定手法である GaTector を提案している．GaTector は，顧客がどの商品を見ているかを物体レベルで推定する手法である．対象人物の頭部特徴から注視領域を推定し，推定した注視領域に重なる物体を注視対象としている．しかし，GaTector が対象とする小売店環境は，交通環境とは大きく異なる．特に，車載カメラ画像中の多くの歩行者は遠方で低解像度なため，小売店環境で撮影する歩行者より小さくなる．そのため，GaTector が注視領域の推定に用いる頭部特徴の抽出が困難であり，交通環境下でも推定可能な手法の検討が必要である．

2.3 骨格特徴点を用いた注視推定

既存の注視推定手法の多くは顔画像や頭部特徴を注視推定に用いており，対象人物が大きく鮮明に写る画像が必要となる．しかし車載カメラ画像に写る歩行者は，撮影距離やカメラ画角の関係から小さく不鮮明に写ることが多く，既存の注視推定手法を直接適用できない．そこで我々の研究グループでは，骨格情報を用いることで遠方の歩行者であってもアイコンタクト検出が可能な手法を提案した [4]．骨格情報は行動認識の研究等で広く用いられており，人物の外見の違いに対して頑健である．

同様に，Kawanishi ら [7] も骨格情報を用いて監視カメラ映像中の人物が見ている注視領域を推定する手法を提案している．これらの研究を踏まえ，提案ベースライン手法では骨格情報を注視領域推定の特徴として用いる．

3. データセットの構築

本節では，交通環境における注視対象推定タスクのためのデータセット構築について述べる．具体的には，既存の交通環境データセットに対し，新たに歩行者の注視点をアノテーションすることでデータセットを拡張する．拡張データセットは，対象歩行者の ID と BBox 座標，注視点座標，歩行者の状態（アイコンタクトや後ろ向きなど），注視対象と判断した物体の BBox 座標とそのカテゴリをアノテーションとして含む．注視対象が特定できない場合については見ている領域をアノテーションするとともに，アイコンタクトの場合や後ろ向きの場合のように見ている領域の判断が困難な場合は，それらの状況を示すフラグを付与している．なお，人による判断の曖昧さを表現するため，各歩行者に対して 3 人のアノテータがアノテーションを行なっている．以降で具体的なデータセットの詳細について述べる．

3.1 データセットの詳細

提案データセットは，既存のデータセットである nuScenes [2], nuImages [2], Waymo [3] を拡張して構築した．これらは車載カメラ画像を収録した大規模な公開データセットであり，2.1 節で述べた物体の BBox とそのクラスラベルが予めアノテーションされている．nuScenes と nuImages には 6 方向の車載カメラ（前方，右前方，右方，右後方，後方，左後方，左方，左前方）が収録されており，それぞれ 1,600×900 画素の画像で構成されている．一方 Waymo は 5 方向の車載カメラ（前方，右前方，右方，左方，左前方）を収録しており，それぞれ 1,980×1,280 画素の画像で構成されている．

データセットを拡張するにあたり，問題設定の簡単化のため，各データセットから以下の条件を満たす画像のみをアノテーション対象として使用した．

(1) 歩行者の隠蔽率が 25 %以下
(2) 歩行者 BBox の縦サイズが nuScenes, nuImages で 200 画素以上，Waymo で 300 画素以上
(3) 歩行者 BBox が元の車載カメラ画像の枠内に収まる
(4) 同じ画像内で対象歩行者の他にアノテーション済みの物体が存在
(5) nuScenes, nuImages は 6 方向のカメラのうち，前方もしくは後方に写る歩行者
(6) Waymo は 5 方向のカメラのうち，左前方，前方，右前方に写る歩行者

なお，nuScenes と Waymo は時系列情報を含むデータセットであり，同じ人物が複数フレームに亘って観測される場合がある．今回の拡張ではアノテーション対象の数を増やすため，連続する画像であっても一定時間経過後は歩行者の見えや注視対象が変化すると仮定し，5 秒間隔でフレームを選択してアノテーション対象とした．また，Waymo については左前方，前方，右前方の 3 カメラを合成した 5,940×1,280 画素のパノラマ画像を作成し，アノテーションを行なった．上記に加え，体の向きが判断できないほどブレた画像，非常に暗い画像，対象歩行者が視覚障がい者

図 2: アノテーション例

と判断される画像は手動で除外した．これらの条件より，2,836 人（1,633 枚）の歩行者をアノテーション対象として抽出した．

3.2 アノテーション作業

nuScene, nuImages, Waymo のデータセットから抽出した 2,836 人の歩行者それぞれに対し，3 人のアノテータにより注視対象をアノテーションした．具体的には以下の作業を行なった．

(1) 作業ツール上に車載カメラ画像を表示し，対象歩行者に BBox を描画する．その際，既存データセットに収録されている BBox 情報を用いる．

(2) アノテータは BBox で囲われた対象歩行者の注視対象をクリックし，その座標を記録する．この際，対象歩行者の BBox のみを表示し，注視対象となりうるその他の物体の BBox は表示しない．

上記 (2) により，アノテータは画像中から自由に注意対象を選択することが可能である．上記の作業はアノテータによって判断が異なる可能性があるため，各歩行者に対して 3 人のアノテータがアノテーション作業を行なった．これらの手順によりアノテーションした例を，図 2 に示す．対象歩行者は黄色の BBox で囲われた人物であり，赤色の点は各アノテータのクリック位置を表す．

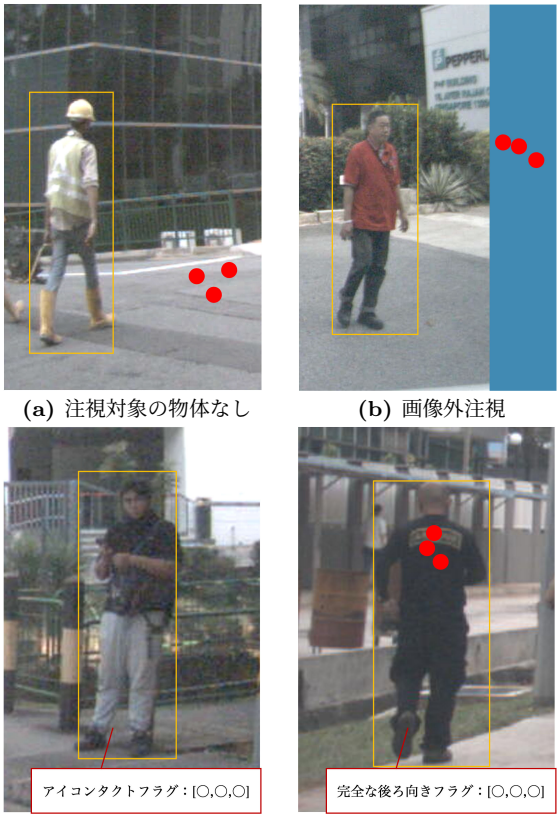
一方，歩行者によっては注視対象の物体を選択できない場面が存在する．今回のアノテーション作業においては，そのような場面として 4 種類（注視対象の物体なし，画像外注視，アイコンタクト，完全な後ろ向き）を想定し，それぞれ以下の通りアノテーションした．

注視対象の物体なし

図 3(a) に示すように，歩行者が特定の物体を注視していない場面を，注視対象の物体なしと定義する．この時，アノテータは図中の赤点で示すように，歩行者の注視領域を画像中にクリックする．

画像外注視

図 3(b) に示すように，歩行者の注視対象が画像外にある場面を，画像外注視と定義する．作業ツールで画像の周囲に 100px のクリック可能な領域を追加し，アノテータは歩行者の見ている方向を保ったまま，この領域内をクリックする．



(a) 注視対象の物体なし **(b)** 画像外注視
(c) アイコンタクト **(d)** 完全な後ろ向き

図 3: 注視対象の物体を選択できない場面

アイコンタクト

図 3(c) に示すように，歩行者が自車両を注視している場面をアイコンタクトと定義する．この時，注視対象は画像内に無いため，画像上へのクリックはせず，アイコンタクトというフラグを記録する．

完全な後ろ向き

図 3(d) に示すように，後ろ向きの歩行者で注視領域が判断しづらい場面を，完全な後ろ向きと定義する．この時，注視対象の物体なしと同様に，歩行者の人物越しに注視している場所を画像中にクリックし，完全な後ろ向きというフラグを記録する．

3.3 アノテーション結果

歩行者の注視対象のアノテーション結果を集計したものを表 1 に示す．すでに述べたように，1,633 枚の車載カメラ画像がアノテーション対象となり，述べ 2,836 人の歩行者に対してアノテーションを行なった．

アノテーション結果から，注視対象物ありの歩行者，画像外を注視する歩行者，それぞれを次の方法で計数した．図 4 に示すように，注視点が同じ物体の BBox 内部に 2 点以上ある場合（2 人以上のアノテータが同じ物体をクリックした場合）を注視対象物ありの歩行者として計数した．ただし，注視対象は各データセットで予めアノテーションされた物体のみを対象とし，信号や看板，スマートフォン

表 1: アノテーション結果の集計

	nuScenes	nuImages	Waymo	合計
画像数 [枚]	218	870	545	1,633
歩行者数 [人]	292	1,240	1,304	2,836
(画像 1 枚あたりの平均 ± 標準偏差)	(1.34 ± 0.61)	(1.43 ± 0.87)	(2.39 ± 2.26)	
物体数 [個]	2,882	13,979	25,950	52,811
(画像 1 枚あたりの平均 ± 標準偏差)	(13.22 ± 9.01)	(16.07 ± 10.06)	(55.96 ± 30.07)	
全注視点数 [点]	876	3,718	3,905	8,499
注視対象の物体なし [%]†	44.3	40.0	48.2	44.2
画像外注視 [%]†	22.3	28.8	16.1	22.4
アイコンタクト [%]†	16.1	19.9	5.86	13.1
完全な後ろ向き [%]†	1.71	4.60	3.00	3.57
注視対象物ありの歩行者数 [人]	87	340	372	799
画像外を見ている歩行者数 [人]	17	101	118	236

† は全注視点数に対する割合

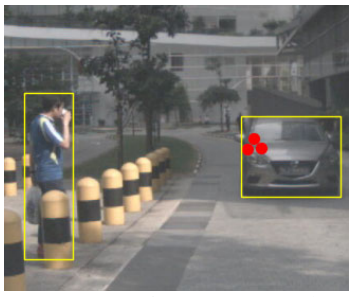


図 4: 注視対象物ありの歩行者例

図 5: 対象外の例

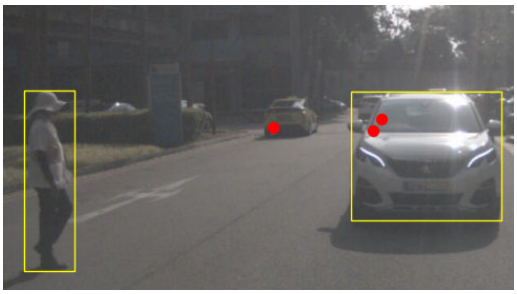


図 6: アノテータによって判断が別れた例

など，アノテーションがない物体は対象外とした．加えて，図 5 に示すように，対象歩行者自身が注視対象となる場合も対象外とした．一方，アノテーションされた注視点のうち 2 点以上が画像外である場合を画像外を注視する歩行者として計数した．例えば，図 3(b) 中の男性は，注視点が 3 点とも画像外にあるため，画像外を注視する歩行者となる．なお，完全な後ろ向きと判断された歩行者であっても，注視点のアノテーションがあれば評価の対象とした．

上記の情報を含め，今回構築した歩行者の注視対象データセットに含まれる情報は下記の通りである．

- 対象歩行者 ID
- 対象歩行者の BBox 座標
- 注視点座標（3 点）
- アイコンタクトか否か（3 件）
- 完全な後ろ向きか否か（3 件）
- 注視対象物の BBox 座標（注視対象物ありの歩行者）
- 注視対象物のカテゴリ（注視対象物ありの歩行者）

今回，各歩行者に対して 3 人のアノテータがアノテーション作業を行なった結果，図 6 に示すようにアノテータによって注視対象物の候補が異なる例を確認した．この例では，注視点が 2 点ある右側の車両が注視対象物と判断している．しかし，中央の車両もアノテータによっては注視対象物となる可能性がある．今後，歩行者の注視対象推定を行う際，各物体に注視点が何点あるかを確信度のように扱うことで，推定モデルの精度向上に貢献できると考える．

4. 注視対象推定タスクのベースライン手法

本節では，歩行者の注視推定タスクの提案ベースライン手法について述べる．ベースライン手法では，骨格情報を利用して歩行者の注視対象を推定する．以降で手法の詳細な手順と結果，及び考察について述べる．

4.1 ベースライン手法

本手法では，歩行者の骨格情報を基に視線を推定し，視線と各物体の位置から，注視対象を推定する．視線上に物体がある場合は，その位置が歩行者の目の位置に最も近い物体を注視対象と推定し，視線上に物体がない場合は，画像外を注視していると推定する．3.3 節に示した注視対象物ありの歩行者（図 4）と画像外を見ている歩行者（図 3(b)）の合計 1,035 人を対象に，以下の 3 段階の処理を行なった．

(1) 既存の骨格推定手法で歩行者の骨格特徴点を推定
(2) 推定した骨格特徴点と歩行者の注視点を多層パーセプトロンに入力し，視線を推定
(3) 各物体から推定した視線への距離を計算し，注視対象を推定

まず骨格推定には，骨格推定ツールボックスの mmpose [8] 上で公開されている事前学習済みのトップダウン型手法 [9–12] を用いた．この骨格推定手法に，3.1 節で述

図 7: 骨格特徴点の推定結果

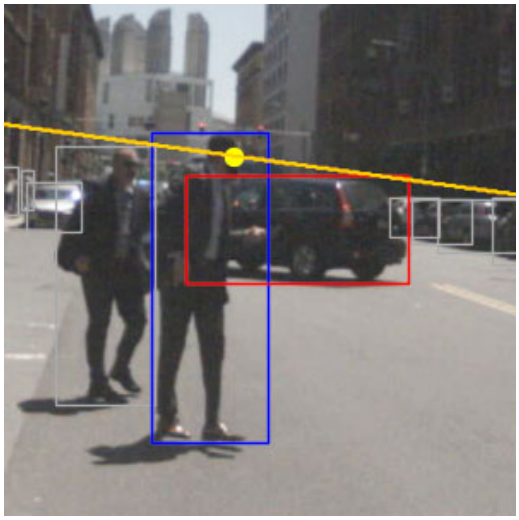
べた既存のデータセットに収録済みの車載カメラ画像と対象歩行者の BBox 座標を入力し，図 7 に示す 17 の骨格特徴点を得た．

次に，推定した骨格特徴点と 3 章で収録した歩行者の注視点を多層パーセプトロンに入力し，歩行者の視線を推定した．骨格特徴点は (1) で得た 17 点を用いた．歩行者の注視点は，注視対象物ありの歩行者の場合には，注視対象物の BBox 内にある注視点の平均，画像外を見ている歩行者の場合には，画像外にある注視点の平均を用いた．推定する視線は歩行者の目の位置を通過する直線とし，その傾きを，歩行者の目の位置と注視点の 2 点を通る直線から学習した．歩行者の目の位置は，推定した骨格特徴点の左目，右目，鼻の 3 点から求めた重心を用いた．多層パーセプトロンの学習には，注視対象物ありの歩行者と画像外を見ている歩行者をそれぞれ訓練データ 8 割，テストデータ 2 割に分割して用いた．

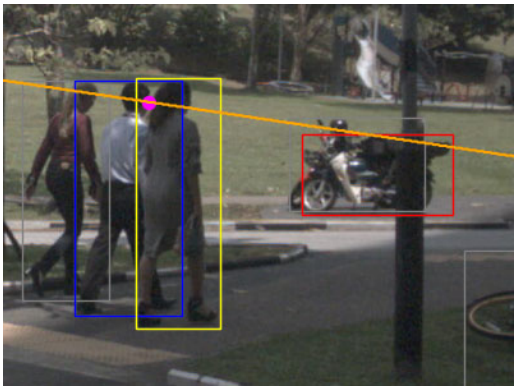
最後に，各物体から推定した視線への距離を計算し，歩行者の注視対象を推定した．推定対象の歩行者は，多層パーセプトロンによる視線推定でテストデータに用いた 208 人である．まず，車載カメラ画像中の全ての物体に対し，その BBox 内の全画素から視線への距離を計算し，最も値が小さい画素から視線への距離を，その物体から視線への距離とした．視線までの距離が同じ画素が複数ある場合は，より歩行者の目の位置に近い画素を選択した．次に，視線上（物体から視線への距離が 0）の物体のうち，最も歩行者の目の位置に近い物体を注視対象とし，視線上に物体がない場合は，画像外を注視していると推定した．なお，3.3 節と同様，各データセットで予めアノテーションされた物体のみを対象とした．

4.2 実験結果と考察

歩行者の視線を推定し，その視線上にある物体のうち，



(a) 正しく推定できた例



(b) 異なる対象を推定した例

図 8: ベースライン手法の推定結果

歩行者に最近傍の物体を注視対象物と推定，視線上に物体がない場合は画像外を注視していると推定した結果，正解率は 25.5％であった．

図 8 にベースライン手法の推定結果例を示す．対象歩行者は青色の BBox，正解の注視対象は赤色の BBox，推定した注視対象が正解と異なる場合は黄色の BBox で示す．図 8(a) は正しい注視対象を推定できた例で，視線上にある歩行者の目から最近傍の物体が，正解に一致した．これより，骨格情報を利用して注視対象を推定することができる．

しかしこの手法では，歩行者から離れた物体を注視対象に選択できない．例えば，図 8(b) に示すように，推定対象は歩行者に近い物体を選択し，歩行者から離れた正解の物体は選択できなかった．特に，図 8(b) の例のように，歩行者の目の位置と物体が重なっている場合，どの方向に視線を推定しても，重なる物体が選択されるため，正解の物体を選択できない．よって，歩行者の目の位置や視線までの距離を基準に注視対象を推定する場合，精度の向上が困難だと考える．

今後の課題として，シーンコンテキストを考慮した手法の検討が考えられる．既存の注視対象推定手法では，シー

ン中における歩行者や物体の位置や向きなど，シーンコンテキストを推定に用いている [4,6]．骨格特徴に加え，3 章で収録した歩行者の注視点とその曖昧さ，シーンコンテキストを用いて推定することで，より精度が改善できると考える．

5. むすび

本報告では，歩行者の注視対象推定タスクのためのデータセットの構築，及びベースライン手法を提案した．歩行者の注視対象推定の研究開発のためには大規模なデータセットが必要不可欠である．しかし，交通環境を対象として歩行者と注視対象を対応付けた公開データセットは存在しない．そこで，既存の交通環境データセットに収録された車載カメラ画像 1,633 枚に写る 2,836 人の歩行者に対し，その注視対象を対応付けるアノテーション作業を行なった．アノテーションの結果，799 人の歩行者に対して注視対象物体との対応付けを行なった．

また，歩行者の骨格情報から視線を推定し，その視線から最も近い物体を注視対象として選択するベースライン手法を提案し，今回構築したデータセットを用いてその評価を行なった．その結果，注視対象の正解率は 25.5 %であった．

今後の課題として，アノテータによる判断の曖昧さを考慮したモデルの構築，交通環境以外のデータセットを用いた注視対象推定モデルとの比較，シーンコンテキストを考慮した手法の検討，などが挙げられる．

謝辞 本研究の一部は，JST，CREST，JPMJCR22D1 の支援を受けたものである．

参考文献

[1] Tomas, H., Reyes, M., Dionido, R., Ty, M., Mirando, J., Casimiro, J., Atienza, R. and Guinto, R.: GOO: A Dataset for Gaze Object Prediction in Retail Environments, *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3119–3127 (2021).

[2] Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G. and Beijbom, O.: nuScenes: A multimodal dataset for autonomous driving, *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11618–11628 (2020).

[3] Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z. and Anguelov, D.: Scalability in Perception for Autonomous Driving: Waymo Open Dataset, *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2443–2451 (2020).

[4] Hata, R., Deguchi, D., Hirayama, T., Kawanishi, Y. and Murase, H.: Detection of distant eye-contact using spatio-temporal pedestrian skeletons, *In Proceedings of the IEEE 25th International Conference on Intelligent Transportation Systems*, pp. 2730–2737 (2022).

[5] Rasouli, A., Kotseruba, I., Kunic, T. and Tsotsos, J.: PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction, *2019 IEEE/CVF International Conference on Computer Vision*, pp. 6261–6270 (2019).

[6] Wang, B., Hu, T., Li, B., Chen, X. and Zhang, Z.: GaTector: A Unified Framework for Gaze Object Prediction, *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19588–19597 (2022).

[7] Kawanishi, Y., Murase, H., Xu, J., Tasaka, K. and Yanagihara, H.: Which Content in a Booklet is he/she Reading? Reading Content Estimation using an Indoor Surveillance Camera, *In Proceedings of the IEEE 24th International Conference on Pattern Recognition*, pp. 1731–1736 (2018).

[8] Contributors, M.: OpenMMLab Pose Estimation Toolbox and Benchmark, `https://github.com/open-mmlab/mmpose` (2020).

[9] Xiao, B., Wu, H. and Wei, Y.: Simple baselines for human pose estimation and tracking, *In Proceedings of the European conference on computer vision*, pp. 466–481 (2018).

[10] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows, *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022 (2021).

[11] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S.: Feature pyramid networks for object detection, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125 (2017).

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