公園利用実態把握に向けた音・BLE マルチモーダル学習の検討

Multimodal Sensing Using Video, Audio, and BLE for Understanding Park Usage

寺岡 莉玖 [1*] 細川 蓮 [1]　松田 裕貴 [2]　安本 慶一 [1,3]　諏訪 博彦 [1,3]

Riku Teraoka[1] Ren Hosokawa[1] Yuki Matsuda[2] Keichi Yasumoto[1,3] Hirohiko Suwa[1,3]

[1] 奈良先端科学技術大学院大学 [1] Nara Institute of Science and Technology
[2] 岡山大学 [2] Okayama University
[3] 理化学研究所 革新知能統合研究センター [3] RIKEN Center for Advanced Intelligence Project

要旨: カメラを用いた人流推定は有効である一方，遮蔽や死角，プライバシーや設置制約により適用が難しい場合がある．本研究では，カメラが利用できない環境下での公園利用状況の把握を目的として，音センサと BLE センサのみを用いた推定手法を検討する．特に，音と BLE の特徴量設計と自己教師あり事前学習の有無が性能に与える影響を整理するため，音のみ，BLE のみ，音+BLE の条件で比較を行った．結果，人の有無推定では音+BLE 条件が最良となり，PR 曲線により閾値に依存した Precision–Recall のトレードオフが確認された．一方，音+BLE 条件における自己教師あり事前学習の有無による性能差は多くのタスクで小さかった．誤推定分析では，総人数と滞留人数で高人数帯の過小推定が生じやすいこと，通過人数で予測が 0 付近に張り付く傾向が確認され，今後の課題として残った．

1　はじめに

人流の把握は，都市開発や地域計画の立案に有用である．特に公園では，利用減少や利用形態の変化に伴い，利用状況を継続的に把握する重要性が高まっている．一方で，実環境での人手による観測やアノテーションは高コストであり，自動推定手法の確立が求められる．従来はカメラに基づく手法が主流である [1, 2]. しかし，カメラは遮蔽や死角により精度が低下しやすく，さらにプライバシーの懸念や設置制約からカメラを利用できない場所も存在する．

我々はこれまでに，公園環境で同時計測した動画，音，BLE を用い，自己教師あり学習を導入したマルチモーダル推定を検討し，タスクにより有効なモダリティが異なることを示している [3]．具体的には，人の有無推定では動画に音を加える条件が有効であり，人数推定では動画に BLE を加える条件が有効であることを示している．しかし，この知見は動画を主軸とした枠組みに基づくものであり，カメラを利用できない環境を想定した場合には，音と BLE のみでどの程度の推定が可能であり，またその際に自己教師あり事前学習の導入や特徴量設計がどのように影響するかを改めて整理する必要がある．

そこで本研究では，カメラが利用できない環境下でも適用可能な方法として，音センサと BLE センサのみを用いた推定枠組みを対象に検討を行う．具体的には，モダリティの利用数（音のみ，BLE のみ，音+BLE）と自己教師あり事前学習の有無を比較し，少量ラベル条件における人の有無推定と人数推定（総人数，通過人数，滞留人数）に対する影響を整理する．また，音についてはログメル埋め込みと音声活動検出（VAD）に基づく指標を，BLE については RSSI 統計量に加えて受信強度閾値以上の累積カウントや端末入れ替わり，rolling 統計を用いた拡張特徴を採用し，特徴量重要度と誤推定分析により，どの特徴が推定に寄与しているか，誤差がどのように生じるかを明らかにする．これにより，非映像モダリティのみで公園利用状況を推定する際の設計指針を明らかにすることを目的とする．

実験の結果，人の有無推定では音+BLE 条件が最良となり，PR 曲線から閾値に依存した Precision–Recall のトレードオフが確認された．一方，総人数，通過人数，滞留人数の回帰では，音+BLE 条件における自己教師あり事前学習の有無による性能差は多くのタスクで小さかった．さらに，特徴量重要度の分析より，VAD 由来特徴量と RSSI 統計量，受信強度閾値拡張が複数タスクで上位に現れ，提案した特徴量設計の有効性が

*連絡先：奈良先端科学技術大学院大学
〒 630-0192 奈良県生駒市高山町 8916 番地-5
E-mail: teraoka.riku.tt6@naist.ac.jp

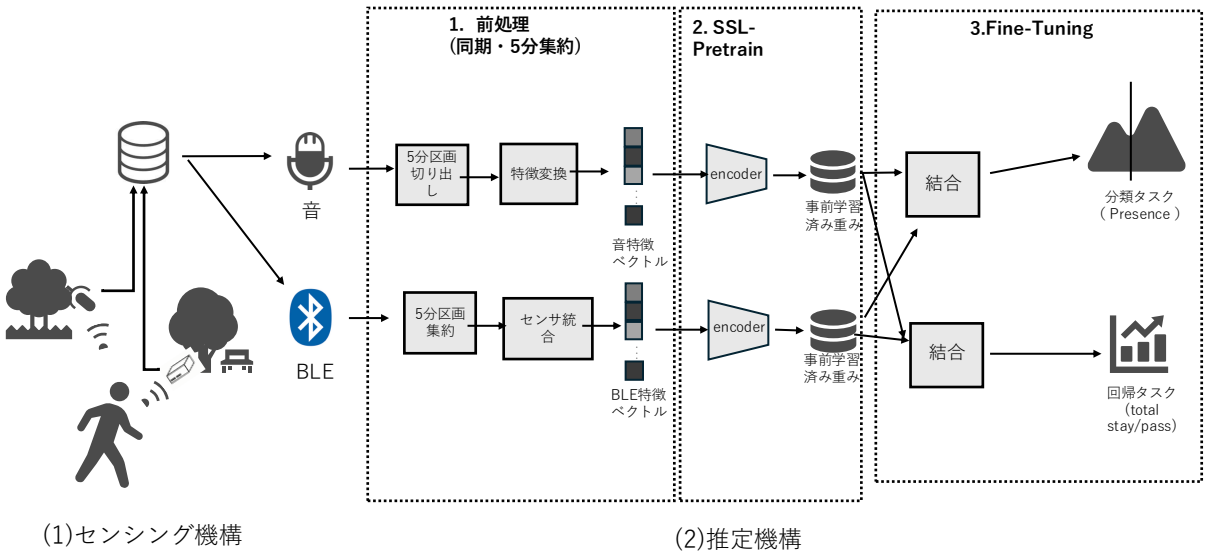


図 1: 提案手法の全体構成

示唆された．誤推定分析では，総人数と滞留人数で高人数帯の過小推定が生じやすいこと，通過人数で予測が 0 付近に張り付く傾向が確認され，今後の課題として残った．

2 関連研究

2.1 カメラを利用した人流推定

カメラ映像に基づく人流推定は，群衆カウントや混雑度推定の研究として発展してきた．静止画を対象とする群衆カウントでは，画像から密度マップを推定し，積分値として人数を得る枠組みが広く用いられている．近年は，照度変化やスケール変動，遮蔽に頑健な特徴表現や，密度推定器の学習方法が提案されている．一方，実環境ではアノテーションのコストが高く，学習データ不足が性能の制約となりやすい．そのため，ラベル付きデータに依存しない学習として自己教師あり学習を導入する試みも進んでいる．例えば動画を対象とした研究では，動画区間内で重複なく個体数を数える枠組みを想定し，ラベルなし動画から群衆の動きに関する表現を学習する方法が提案されている [4]．また，群衆カウントにおいてラベル付きデータを用いずに学習する枠組みも報告されており [5]，長期観測のようにラベル取得が困難な場面で未ラベルデータを活用できる可能性が示されている．しかし，カメラによる推定は，遮蔽や死角，逆光や夜間などの環境条件によって精度が低下しやすい．さらにプライバシーや設置制約により常設が難しい場合もある．このため，カメラに依存しない推定手法の検討が重要となる．

2.2 音を用いた人流推定

音を用いた推定は，映像に比べて個人の外見情報を含みにくく，設置条件の制約も比較的緩いことから，プライバシー制約の強い環境で有効な選択肢となる．音から得られる情報は，人数そのものよりも活動量や混雑度といった状態量に対応しやすいとされ，特徴量設計と推定タスクの設定が重要となる．例えば Hossain は，発話区間を除外した環境音を入力とし，Transformer により混雑度を推定する枠組みを提案している [6]．このように，音は視覚が破綻する状況でも取得しやすい一方，屋外環境では風や交通騒音など人数と直接関係しない音源が混在し，推定にノイズとして作用し得る．そのため，周波数分布を表すスペクトル特徴に加え，音声活動検出などに基づく有効区間の推定や，活動音の指標化を組み合わせる設計が検討されている [7]．本研究では，ログメルスペクトログラム特徴と音声活動検出に基づく指標を候補とし，少量ラベル条件での有効性を比較する．

2.3 BLE を用いた人流推定

無線信号に基づく推定は，カメラを用いずに人の存在や混雑状況を把握できる手法として注目されている．特に BLE は，受信した広告パケットの時刻や受信強度を利用して，人流や混雑度を推定する研究が進んでいる．Goto らは，道路沿いに複数の BLE スキャナを設置し，検出時刻差や信号強度差の時系列から人数に加えて移動方向まで推定する枠組みを提案している [8]．また，複数の都市環境を対象とした実践的研究 [9] や，飲食店や公共施設などの屋内空間における混雑推定 [10]，交通機関における混雑推定 [11, 12, 13] が報告されており，BLE が多様な環境で混雑状況を推定し得ることが示されている．一方で，BLE に基づく推定は端末保有率や MAC アドレスのランダム化，電波環境の変動の影響を受けやすく，観測と人数の関係が環境や時間帯により変化し得る．そのため，単純な受信数やユニーク端末数だけでなく，受信強度の統計量や分布特徴，時間変動量などを含む特徴量設計が重要となる．本研究では，受信数や受信強度の統計量に加え，受信強度分布を反映する特徴を設計候補に含め，音特徴と合わせて少量ラベル条件での推定性能を比較する．

3 提案手法

本研究では，カメラが利用できない環境を想定し，音センサと BLE センサのみを用いて公園利用状況を推定する．推定は 5 分窓を基本単位とし，各窓に対して音特徴量と BLE 特徴量を構成して学習器に入力する．提案手法の全体像を図 1 に示す．本研究の目的は，音と BLE における特徴量設計が推定性能へ与える影響を明らかにすることであり，併せて自己教師あり事前学習の有無，少量ラベル下での推定タスク構成の違いを比較する．なお，本研究で用いるモダリティ統合は単純結合による統合のみとし，統合方法の差ではなく特徴量設計と学習設定の差に焦点を当てる．

3.1 前処理

音および BLE の観測は，開始時刻を 5 分単位に丸めて整合し，同一の 5 分窓に属する特徴量を対応付ける．時刻 t における音特徴量と BLE 特徴量をそれぞれ $\mathbf{x}_t^{(a)}$，$\mathbf{x}_t^{(b)}$ とし，観測の有無を表すフラグを $m_t^{(a)} \in \{0,1\}$，$m_t^{(b)} \in \{0,1\}$ とする．欠損窓では特徴量をゼロベクトルで補完し，$m_t^{(\cdot)} = 0$ とすることで，学習器が欠損とゼロ値を区別できるようにする．

3.1.1 音の特徴量

音特徴量は，ログメルスペクトログラムに基づく埋め込みベクトルと，音声活動検出（VAD）に基づく 4 つの集約指標，および取得有無フラグを連結して構成する．ログメル埋め込みを $\mathbf{z}_t^{\mathrm{mel}}$，VAD 由来の集約指標を $\mathbf{z}_t^{\mathrm{vad}}$ とする．ログメル埋め込みおよび VAD 指標の取得有無は，欠損窓を 0 補完した場合でも学習器が未取得を識別できるようにし，取得有無フラグを m_t^{mel}，m_t^{vad} とする．本研究では，

$$\mathbf{x}_t^{(a)} = \left[\mathbf{z}_t^{\mathrm{mel}}\,;\,\mathbf{z}_t^{\mathrm{vad}}\,;\,m_t^{\mathrm{mel}}\,;\,m_t^{\mathrm{vad}}\right]$$

として音特徴量を構成する．

VAD に基づく 4 つの集約指標は，5 分窓内における (1) 発話比率（speech ratio），(2) 発話区間数（number of speech segments），(3) 平均発話長（mean speech duration），(4) 平均無音長（mean silence duration）である．また，取得有無フラグとして has_logmel，has_vad を付加する．

3.1.2 BLE の特徴量

BLE 特徴量は，5 分窓ごとに広告パケットログを集約し，基本統計，RSSI 閾値以上の累積カウント，アドレス入れ替わり，rolling 特徴，および取得有無フラグを連結して構成する．使用した特徴量は以下の通りである．

- 取得有無フラグ
  - has_ble
- 受信数・端末数の統計（log1p を適用）
  - message_count（log1p）
  - unique_devices（log1p）
- RSSI の基本統計
  - rssi_mean，rssi_std，rssi_min，rssi_max，rssi_range
- RSSI 変動（jitter）の統計
  - rssi_jitter_std，rssi_jitter_mad
- 端末ごとの観測回数の統計
  - addr_obs_mean，addr_obs_max
- RSSI 閾値以上の累積カウント
  - rssi_ge_count（閾値:$-90, -85, -80, , , , -40$）
  - rssi_ge_unique（閾値:$-90, -85, -80, , , -40$）

- アドレス入れ替わり特徴（log1p を適用）
  - addr_new（log1p），addr_gone（log1p），addr_churn（log1p）
- rolling 特徴（移動平均と差分）
  - 15 分：message_count_ma_15m, message_count_delta_15m, unique_devices_ma_15m, unique_devices_delta_15m
  - 30 分：message_count_ma_30m, message_count_delta_30m, unique_devices_ma_30m, unique_devices_delta_30m

以上を連結したベクトルを $\mathbf{x}_t^{(b)}$ とする．

3.2 自己教師あり事前学習

長期観測データの大部分が未ラベルであることを踏まえ，本研究では自己教師あり事前学習の有無を比較する．自己教師あり事前学習ありの条件では，未ラベル全期間の特徴量から表現を学習し，少量ラベルの微調整における初期値として用いる．本研究では，時刻 t の入力から得られる表現を $\mathbf{h}_t$ とし，同一窓に対する二つのビュー $\mathbf{h}_{t,1}$，$\mathbf{h}_{t,2}$ を用いた InfoNCE 損失で学習する．バッチ内の他サンプルを負例とし，温度パラメータ τ を用いると，

$$\mathcal{L}_{\mathrm{NCE}} = -\sum_{t} \log \frac{\exp(\mathrm{sim}(\mathbf{h}_{t,1}, \mathbf{h}_{t,2})/\tau)}{\sum_{t'} \exp(\mathrm{sim}(\mathbf{h}_{t,1}, \mathbf{h}_{t',2})/\tau)}$$

を最小化する．ここで $\mathrm{sim}(\cdot,\cdot)$ はコサイン類似度である．自己教師あり事前学習なしの条件では，事前学習を行わず，下流タスクの学習のみを行う．

3.3 ファインチューニング

ファインチューニングでは，少量のラベル付きデータを用いて下流タスクを学習する．音特徴量と BLE 特徴量をそれぞれ多層パーセプトロンで埋め込みに射影し，単純結合する．すなわち，

$$\mathbf{e}_t^{(a)} = f_a(\mathbf{x}_t^{(a)}) \odot m_t^{(a)}, \quad \mathbf{e}_t^{(b)} = f_b(\mathbf{x}_t^{(b)}) \odot m_t^{(b)},$$

$$\mathbf{e}_t = [\mathbf{e}_t^{(a)}; \mathbf{e}_t^{(b)}]$$

とする．ここで f_a, f_b は多層パーセプトロン，$\odot$ は要素積である．自己教師あり事前学習ありの条件では，事前学習で得た重みを初期値として用い，自己教師あり事前学習なしの条件ではランダム初期化から学習する．

3.4 出力タスクと学習目標

本研究では公園内において，どの時間帯に，どのぐらいの人数が，どのような行動をしているかという人流を把握することを目的とするために，人の有無，総人数，通過/滞留人数の推定を目指す．

分類として人の有無 $y_t^{\mathrm{pres}} \in \{0,1\}$ を推定し，回帰として総人数 y_t^{tot}，滞留人数 y_t^{stay}，通過人数 y_t^{pass} を同時に推定する．分類ロジットを u_t とし，確率を $\hat{p}_t = \sigma(u_t)$ とする．回帰出力を $\hat{y}_t^{\mathrm{tot}}$，$\hat{y}_t^{\mathrm{stay}}$，$\hat{y}_t^{\mathrm{pass}}$ とすると，損失は

$$\begin{aligned}\mathcal{L} = {} & \lambda_{\mathrm{cls}}\, \mathcal{L}_{\mathrm{BCE}} \\ & + \lambda_{\mathrm{reg}} \Big(\big|\hat{y}_t^{\mathrm{tot}} - y_t^{\mathrm{tot}}\big| + \big|\hat{y}_t^{\mathrm{stay}} - y_t^{\mathrm{stay}}\big| + \big|\hat{y}_t^{\mathrm{pass}} - y_t^{\mathrm{pass}}\big| \Big) \\ & + \lambda_{\mathrm{cons}} \Big| \hat{y}_t^{\mathrm{tot}} - \big(\hat{y}_t^{\mathrm{stay}} + \hat{y}_t^{\mathrm{pass}}\big) \Big| .\end{aligned}$$

として最小化する．ここで $\mathcal{L}_{\mathrm{BCE}}$ は二値交差エントロピー損失である．

人の有無は不均衡となりやすいため，学習時に正例と負例が偏らないよう重み付きサンプリングを用いる．重みは正例率に基づき設定し，少数クラスが繰り返し抽出されるようにすることで，分類の学習が多数派に偏ることを抑制する．

4 評価実験

4.1 実験環境

実験は，実際の公園環境において収集したデータを用いて行った．対象とする公園は，大阪府豊能郡豊能町にある光風台中央公園（光風台二丁目公園）であり，屋外環境として照度変化，植栽や遊具による遮蔽，利用状況の時間変動（時間帯・曜日・天候）といった要因が複合的に生じる．本研究は，カメラを利用できない状況を想定して音と BLE のみで推定を行うため，視覚的な死角や遮蔽の影響を受けにくい非映像モダリティに基づく推定性能を検証する上で，実環境条件を含む本公園は適した評価環境である．

公園内には，ボイスレコーダおよび BLE スキャナを設置し，長期間にわたってデータを取得した．センサの設置位置を図 2 に，設置した各センサの外観と構成を図 3 に示す．音声はボイスレコーダにより MP3（192 kbps）で収録し，サンプリング周波数 44.1 kHz，低域カット（〜220 Hz）を有効化した．また，収録時の設定として入力レベル 90，出力レベル 4 メモリを用いた．BLE スキャナは Raspberry Pi 4 をベースに動作し，Bluetooth 4.0+EDR/LE Class1 対応 USB アダプタを使用して広告パケットを受信する．BLE は 15 秒間隔でスキャンし，取得したログを公園内の Wi-Fi 経由でクラウドに蓄積する．

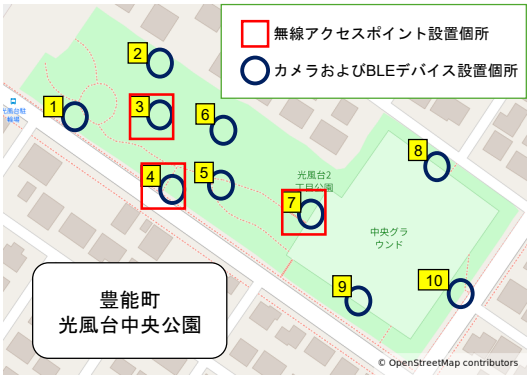


図 2: 光風台公園内に設置したボイスレコーダ・BLE スキャナの設置位置．赤枠内はグラウンドを表している．

本研究では，推定モデルの入力として映像は用いない．一方で，正解データの作成のために，グラウンドを撮影したカメラ映像を補助的に用いる．具体的には，映像を観察して 5 分ごとの時間窓に対して総人数および行動内訳として通過人数と滞留人数を付与する．ここで滞留は，当該 5 分窓のうちおおよそ 2 分以上映像内に存在した人物を滞留として扱い，それ以外を通過として扱う．

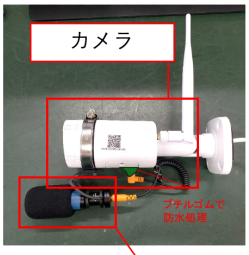
本実験では，音声の解析段階では波形そのものを用いず，ログメルスペクトログラムに基づく埋め込みや音声活動検出に基づく指標などの集約特徴量を用いることで，会話内容などの直接的情報を解析対象から外した．BLE については，5 分窓・複数センサの統計量（受信数や受信強度分布等）として集約した特徴量を用い，個々の端末を追跡することを目的としない形で利用した．

本研究では，公園内でも利用が集中しやすいグラウンド領域を対象とし，グラウンド周辺に設置したセンサ 7，8，9，10 の 4 台のみを用いて特徴量を構成した．これは対象領域を明確化することで，アノテーションおよび推定結果との対応付けを容易にし，実運用を想定した評価を行うためである．

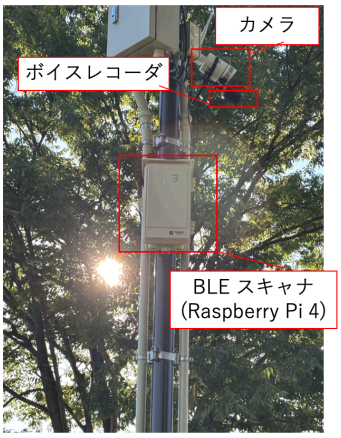
取得データの大部分は未ラベルであり，一部の時間区間についてのみ人手でアノテーションを行った．本研究では，少量のアノテーションデータとして 2 日分のデータを用意し，うち 1 日分を微調整に，もう 1 日分を評価に用いた．

4.2 比較条件

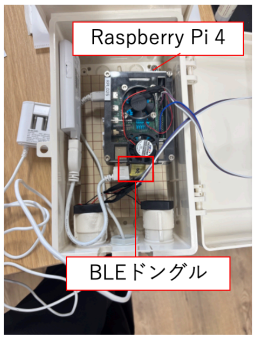
提案手法の有効性を検証するため，以下の観点で比較を行う．第一に，モダリティの利用数として，音のみ，BLE のみ，音+BLE の 3 条件を設定する．第二に，



(a) ボイスレコーダの外観と構成



(c) 設置例



(b) BLE スキャナの外観と構成

図 3: 設置したセンサ概要

自己教師あり事前学習の有無として，教師あり学習と自己教師あり事前学習を比較する．

4.3 データ分割と少量ラベル設定

本研究では，少量ラベル条件での比較を行う．具体的には，アノテーション付きデータ 2 日分を用意し，1 日分を微調整に，もう 1 日分を評価に用いる．未ラベルデータは長期観測で得られた全期間を用い，自己教師あり事前学習ありの条件では InfoNCE により事前学習を行う．

4.4 評価指標

分類は人の有無の Precision，Recall，F1，Accuracy で評価する．回帰は総人数，滞留人数，通過人数の MAE と RMSE で評価する．さらに，総人数と滞留人数と通過人数の和の差の平均値を整合性誤差として併記する．

表 1: 人の有無推定の性能比較

条件	Prec.	Rec.	F1	Acc.
SSL・音	0.603	0.959	0.741	0.662
SSL・BLE	0.565	0.890	0.692	0.600
SSL・音+BLE	**0.644**	0.890	**0.747**	**0.697**
Sup.・音	0.588	0.959	0.729	0.641
Sup.・BLE	0.570	0.890	0.695	0.607
Sup.・音+BLE	0.602	0.932	0.731	0.655

表 2: 総人数推定の性能比較

条件	MAE	RMSE
SSL・音	1.417	2.412
SSL・BLE	1.407	2.566
SSL・音+BLE	1.340	2.338
Sup.・音	1.700	2.531
Sup.・BLE	1.374	2.506
Sup.・音+BLE	**1.314**	**2.273**

5 結果

本節では，ログメル埋め込みと音声活動検出に基づく指標および取得有無フラグを用いた音特徴量と，受信強度拡張を含む BLE 特徴量を用いた結果を示す．比較は，学習方法（自己教師あり事前学習あり，教師あり），モダリティ条件（音のみ，BLE のみ，音+BLE）の組合せで行う．表中では，自己教師あり事前学習ありを SSL，教師あり学習を Sup. と略記する．

5.1 人の有無推定

表 1 に，人の有無推定の結果を示す．F1 は SSL・音+BLE が 0.747 で最良であり，Accuracy も SSL・音+BLE が 0.697 で最良であった．またいずれの条件も Recall が 0.890 以上と高い一方で，Precision は 0.565 から 0.644 に留まっている．このことから，人の有無推定では検出漏れを抑える一方で，誤って人ありと判定する例が残る傾向が示唆される．

5.2 総人数推定

表 2 に，総人数推定の結果を示す．MAE は Sup.・音+BLE が 1.314 で最良であり，RMSE も Sup.・音+BLE が 2.273 で最良であった．次点は SSL・音+BLE であり，MAE は 1.340，RMSE は 2.338 であった．Sup.・音+BLE との差は MAE で 0.026，RMSE で 0.065 と小さく，音+BLE 条件においては自己教師あり事前学習の有無によらず，総人数推定の性能が概ね同等であり，複数モダリティの利用が有効であるといえる．

5.3 通過人数および滞留人数推定

表 3 に，通過人数および滞留人数推定の結果を示す．通過人数の MAE は Sup.・BLE が 0.492 で最良であり，次点は SSL・BLE の 0.493 であった．通過人数の RMSE は Sup.・音が 0.979 で最良であり，次点は SSL・音および Sup.・音+BLE の 0.996 であった．以上より，通過人数は条件間の差が小さいことがわかる．滞留人数の MAE は SSL・音+BLE が 0.997 で最良であり，次点は Sup.・音+BLE の 1.023 であった．滞留人数の RMSE は Sup.・音+BLE が 2.053 で最良であり，次点は SSL・音+BLE の 2.074 であった．以上より，滞留人数は自己教師あり学習の有無に関係なく，音+BLE 条件が上位となる傾向が表れている．整合性誤差は全条件で 0.004 から 0.095 の範囲にあり，いずれの条件でも比較的小さい値を示した．

6 考察

6.1 自己教師あり事前学習の寄与

自己教師あり事前学習の寄与を確認するため，同一モダリティ条件において SSL と Sup. を比較する．人の有無推定では，音+BLE 条件において SSL が F1=0.747，Sup. が F1=0.731 であり，差は 0.016 であった（表 1）．Accuracy も SSL が 0.697，Sup. が 0.655 であり，差は 0.042 であった．

総人数推定では，音+BLE 条件において SSL が MAE = 1.340，RMSE=2.338，Sup. が MAE = 1.314，RMSE = 2.273 であり，差は MAE で 0.026，RMSE で 0.065 と小さい（表 2）．滞留人数推定でも，音+BLE 条件において SSL が MAE=0.997，RMSE=2.074，Sup. が MAE = 1.023，RMSE = 2.053 であり，差は MAE で 0.026，RMSE で 0.021 であった（表 3）．通過人数推定では，音+BLE 条件において SSL が MAE = 0.494，RMSE = 1.020，Sup. が MAE = 0.508，RMSE = 0.996 であり，差は MAE で 0.014，RMSE で 0.024 であった（表 3）．

以上より，音+BLE 条件においては，分類および回帰の主要指標に関して SSL と Sup. の差は概ね 0.01 から 0.07 の範囲に収まっており，本実験設定では自己教師あり事前学習の有無による性能差は全体として大きくないといえる．

表 3: 通過人数および滞留人数推定の性能比較（整合性誤差を併記）

条件	MAE(通過)	RMSE(通過)	MAE(滞留)	RMSE(滞留)	整合性誤差
SSL・音	0.525	0.996	1.054	2.136	0.068
SSL・BLE	**0.493**	1.021	1.170	2.347	**0.004**
SSL・音+BLE	0.494	1.020	**0.997**	2.074	0.095
Sup.・音	0.552	**0.979**	1.373	2.298	0.042
Sup.・BLE	0.492	1.028	1.181	2.319	0.042
Sup.・音+BLE	0.508	0.996	1.023	**2.053**	0.030

6.2 特徴量設計の有効性

図 4a から図 4d に，SSL・音+BLE 条件における特徴量重要度（各タスク上位 20）を示す．全体として，音声活動検出（VAD）に基づく特徴量が複数タスクで上位に現れており，特に vad_segment_count は人の有無，滞留人数，総人数において最重要の特徴量として選択されている．加えて，人の有無では vad_mean_silence_len や vad_speech_ratio が上位に含まれており，音声活動の量や継続性を表す指標が存在推定に有効であることを示している．

BLE 特徴量については，RSSI の基本統計が上位に現れており，例えば rssi_mean は通過人数で最重要であり，rssi_std や rssi_max は総人数や滞留人数で上位に含まれる．また，RSSI 閾値以上の累積カウント特徴も複数タスクで上位に現れている．具体的には，rssi_ge_count_-80 や rssi_ge_count_-90 は総人数および滞留人数で上位に含まれ，人の有無でも rssi_ge_unique_-85，rssi_ge_unique_-90 や rssi_ge_count_-85，rssi_ge_count_-90 が上位に含まれる．これらは，受信強度拡張により近距離端末の存在や密度変化を捉える設計が，複数タスクに対して有効に機能していることを示唆する．

さらに，rolling 特徴や端末入れ替わり特徴も上位に現れている．例えば message_count_ma_15m や，unique_devices _delta_15m などが複数タスクで確認でき，また addr_new，addr_gone，addr_churn も人の有無や滞留人数で上位に含まれる．これらは，短期・中期の変動や端末の入れ替わりを表す特徴が，推定に対して補助的な情報として寄与していることを示している．

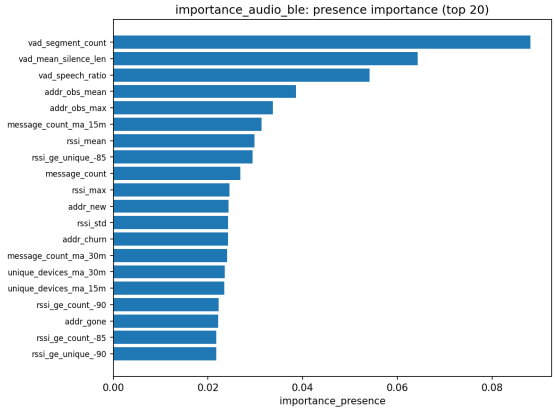
一方で，ログメル埋め込みは，通過人数では logmel_1，logmel_2，logmel_3，logmel_56 が上位に含まれる一方，人の有無，滞留人数，総人数では上位特徴として現れにくい．このことから，ログメル埋め込みの寄与はタスクにより異なり，少なくとも本設定では通過人数の推定において相対的に寄与しやすい可能性がある．以上より，提案した特徴量設計のうち，VAD に基づく指標と RSSI 統計量および受信強度閾値拡張は一貫して上位に現れており，有効な設計要素であるといえる．

6.3 誤推定分析

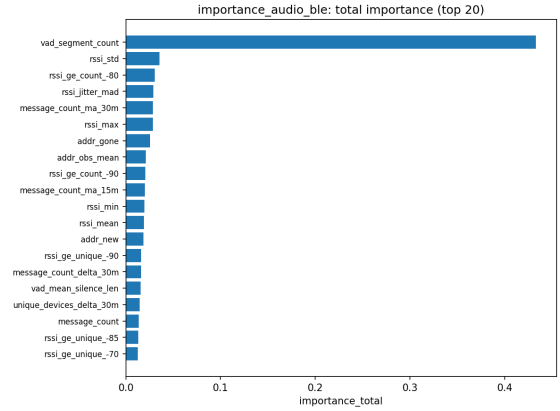
本節では，表の指標のみでは把握しにくい誤推定の傾向を明らかにすることを目的として，SSL・音+BLE 条件に対して可視化に基づく分析を行う．具体的には，人の有無推定について PR 曲線により閾値と Precision–Recall の関係を確認し，人数推定について真値と予測の散布図および時系列可視化により誤差の現れ方を整理する．図 5 に，表 1 で上位であった SSL・音+BLE 条件の PR 曲線を示す．図 5 より，Recall を高めるほど Precision が低下しており，閾値を下げて検出を増やすと誤検出が増えるトレードオフが確認できる．この傾向は，表 1 において Recall が 0.890 と高い一方で Precision が 0.644 に留まるという結果とも整合的であり，高再現率を維持する運用では誤検出が残りやすいことを示唆する．一方で，Recall を中程度に抑えた区間では Precision が相対的に高くなる領域が見られ，運用上の要件に応じた閾値調整により誤検出を抑制できる余地がある．今後は，Average Precision などの要約指標の併記や，目的とする運用点に基づく閾値選択により，誤検出と検出漏れのバランスを明確化した評価を行う必要がある

次に人数推定について真値と予測の散布図を図 6 に示す．可視化結果より，SSL・音+BLE 条件における総人数と滞留人数では低人数帯では真値に近い予測も見られる一方で，高人数帯では対角線より下側に点が集中しており，過小推定が増加する傾向が確認された（図 6a，図 6c）．また通過人数は予測が 0 付近に張り付く点が多く，通過人数の回帰が十分に機能していないことが示唆された（図 6b）．これらは，高人数帯のデータ数や損失設計，および通過人数の値域が小さく不均衡になりやすいこと等に起因する可能性がある．

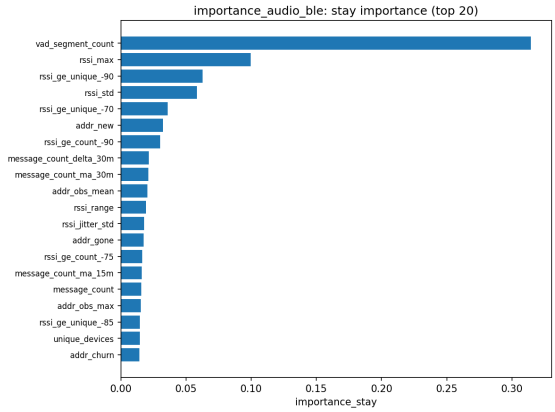
図 7 は，SSL・音+BLE 条件における総人数，通過人数，滞留人数の時系列（真値と予測）を示している．総人数では，真値が大きく上昇する区間において予測の上昇が相対的に小さく，高人数帯で過小推定が生じやすいことが時系列上でも確認できる（図 7a）．滞留人数でも同様に，真値のピークに対して予測が十分に追従しない区間が見られ，ピーク時の過小推定が示唆される（図 7c）．一方で通過人数は，真値がスパイク状



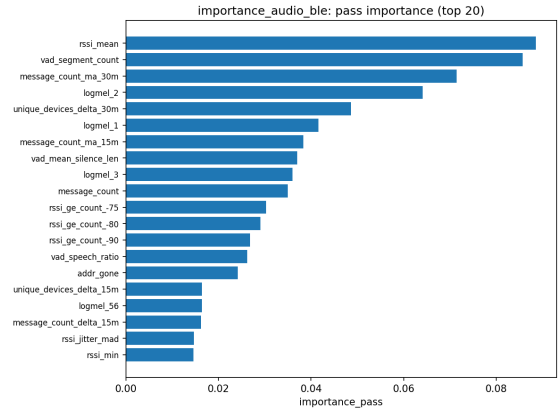
(a) 人の有無（presence）の特徴量重要度（上位 20）



(b) 総人数（total）の特徴量重要度（上位 20）



(c) 滞留人数（stay）の特徴量重要度（上位 20）



(d) 通過人数（pass）の特徴量重要度（上位 20）

図 4: SSL・音+BLE 条件における特徴量重要度（各タスク上位 20）.

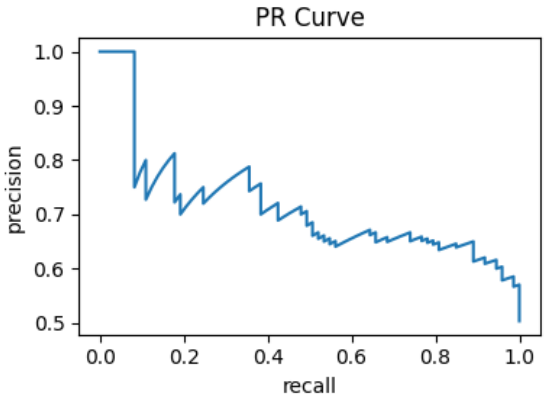


図 5: SSL・音+BLE 条件における人の有無推定の PR 曲線.

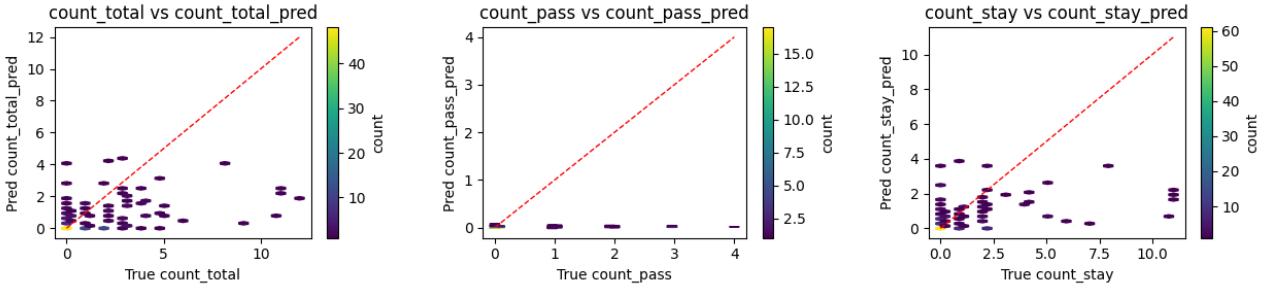
に増加する区間が存在するにもかかわらず，予測はほぼ 0 付近に留まる区間が多く，通過人数の回帰が機能していないことが時系列上でも確認できる．

6.4 改善に向けた示唆

本実験設定では，音+BLE 条件における SSL と教師あり学習の性能差は概ね小さいことから，性能改善に向けては自己教師あり事前学習の有無の調整よりも，閾値設計，特徴量設計，および学習目標の設計を優先して検討することが有効であると考えられる．

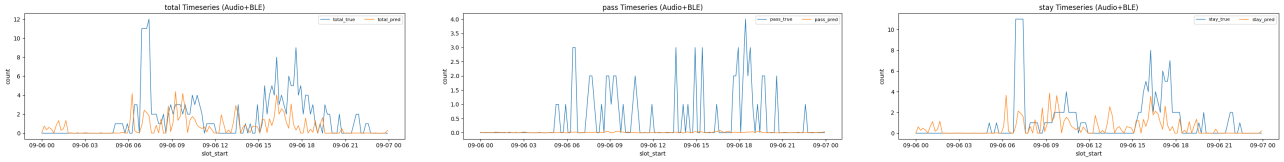
まず，人の有無推定では，PR 曲線（図 5）が示す通り，Recall を高めるほど Precision が低下するため，運用要件に応じた閾値最適化が重要である．特に，高再現率運用では誤検出が残りやすいことから，閾値の選択に加え，偽陽性を抑制するための負例強化や，損失の重み付けなど学習設計の検討が必要である．

次に，特徴量設計の観点では，特徴量重要度（図 4a–4d）より，vad_segment_count を中心とする VAD 特徴量や，rssi_mean，rssi_std，rssi_max といった RSSI 統計量，および rssi_ge_count や rssi_ge_unique に代表される受信強度閾値拡張が複数タスクで一貫して上位に現れている．したがって，これらの情報をより安定に取得できる前処理や集約方法の改善が有効と考えられる．また，rolling 特徴や端末入れ替わり特徴も上位に含まれるため，窓幅や集約方法を調整し，短期変動と中期変



(a) 総人数（真値と予測）　(b) 通過人数（真値と予測）　(c) 滞留人数（真値と予測）

図 6: SSL・音+BLE 条件における回帰結果の散布図（真値と予測）．赤破線は $y = x$ を示す．



(a) 総人数の時系列（真値と予測）　(b) 通過人数の時系列（真値と予測）　(c) 滞留人数の時系列（真値と予測）

図 7: SSL・音+BLE 条件における人数推定の時系列可視化．

動を適切に捉える設計を検討する余地がある．

回帰タスクについては，散布図および時系列（図 6，図 7）から，総人数と滞留人数では高人数帯で過小推定が生じやすく，通過人数では予測が 0 付近に張り付く傾向が確認された．このため，高人数帯の誤差を重視する損失設計や，総人数と内訳の整合性項の重み調整が有効と考えられる．さらに，通過人数については，値域が小さく不均衡になりやすいことから，回帰のまま扱う場合には専用の重み付けを導入すること，あるいは通過の発生を段階的に扱う定式化など，学習目標の見直しを検討する必要がある．

7　まとめ

本研究では，カメラが利用できない環境を想定し，音センサと BLE センサのみを用いて公園利用状況を推定する枠組みを提案した．特徴量設計として，音はログメル埋め込みと音声活動検出（VAD）に基づく指標および取得有無フラグを組み合わせ，BLE は RSSI 統計量に加えて受信強度閾値以上の累積カウント，端末入れ替わり，rolling 統計を含む拡張特徴を導入した．

少量ラベル条件で自己教師あり事前学習の有無を比較した結果，音+BLE 条件における SSL と教師あり学習の性能差は多くのタスクで小さく，自己教師あり事前学習の有無による差は本実験設定では全体として大きくないことが確認された．一方，人の有無推定では PR 曲線により閾値に依存した Precision–Recall のトレードオフが確認され，運用要件に応じた閾値設計の重要性が示された．

また特徴量重要度の分析より，VAD 由来特徴量と RSSI 統計量，受信強度閾値拡張が複数タスクで上位に現れ，提案した特徴量設計が有効に機能していることが示唆された．誤推定分析では，総人数と滞留人数で高人数帯の過小推定が生じやすいこと，通過人数で予測が 0 付近に張り付く傾向が確認され，これらが主要な課題であることが明らかとなった．

今後は，誤検出を抑制するための閾値最適化と学習設計の検討に加え，高人数帯の誤差を重視する損失設計や整合性項の重み調整，ならびに通過人数に対する学習目標の見直しを通じて，非映像モダリティのみでの実用的推定精度の確立を目指す．

謝辞

本研究は，JST さきがけ（JPMJPR2465）および JST 共創の場形成支援プログラム (JPMJPF2115) の支援を受けたものである．

参考文献

[1] Guangshuai Gao, Junyu Gao, Qingjie Liu, Qi Wang, and Yunhong Wang. Cnn-based density estimation and crowd counting: A survey. *arXiv*, 2020.

[2] Preksha Pareek and Ankit Thakkar. A survey on video-based human action recognition: recent up-

dates, datasets, challenges, and applications. *Artificial Intelligence Review*, Vol. 54, pp. 2259–2322, 2021. Published: 25 Sep 2020.

[3] 寺岡莉玖, 細川蓮, 松田裕貴, 安本慶一, 諏訪博彦. 公園利用実態把握のための動画・音・ble を用いたマルチモーダルセンシング. 第 32 回社会情報システム学シンポジウム, 2026.

[4] Feng-Kai Huang, Bo-Lun Huang, Li-Wu Tsao, Jiun-Cheng Wu, Hong-Han Shuai, and Wen-Huang Cheng. Flowing crowd to count flows: A self-supervised framework for video individual counting. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*, pp. 8234–8243, 2025.

[5] Deepak Babu Sam, Abhinav Agarwalla, Jimmy Joseph, Vishwanath A. Sindagi, R. Venkatesh Babu, and Vishal M. Patel. Completely self-supervised crowd counting via distribution matching. In *Computer Vision – ECCV 2022 (LNCS 13691)*, pp. 186–204. Springer, 2022.

[6] Forsad Al Hossain, M. Tanjid Hasan Tonmoy, Andrew A. Lover, George A. Corey, Mohammad Arif Ul Alam, and Tauhidur Rahman. Crowdotic: A privacy-preserving hospital waiting room crowd density estimation with non-speech audio. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications (HotMobile '24)*, pp. 79–85. ACM, 2024.

[7] Forsad Al Hossain, Andrew A. Lover, George A. Corey, et al. Flusense: A contactless syndromic surveillance platform for influenza-like illness in hospital waiting areas. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 2020. Full text available on PubMed Central.

[8] Ippei Goto, Kentaro Ueda, Yuki Matsuda, Hirohiko Suwa, and Keiichi Yasumoto. Bless: Ble based street sensing for people counting and flow direction estimation. In *IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops 2024)*, pp. 76–81, 2024.

[9] Yuki Matsuda, Hirohiko Suwa, Kotaro Hayashi, Taito Yoshimura, Arata Yoshihara, and Ismail Arai. Estimating people flow and crowdedness for various urban environments based on ble signal sensing: Practical studies. *IEICE Transactions on Communications*, pp. 1–11, 2025.

[10] Yuki Matsuda, Kentaro Ueda, Eigo Taya, Hirohiko Suwa, and Keiichi Yasumoto. BLECE: BLE-based crowdedness estimation method for restaurants and public facilities. In *Fourteenth International Conference on Mobile Computing and Ubiquitous Network, ICMU 2023, Kyoto, Japan, November 29 - Dec. 1, 2023*, pp. 1–6. IEEE, 2023.

[11] Takumi Ikenaga, Yuki Matsuda, Ippei Goto, Kentaro Ueda, Hirohiko Suwa, and Keiichi Yasumoto. Using BLE signals to estimate objective and subjective crowdedness levels on fixed-route buses. *IEEE Access*, Vol. 13, pp. 67488–67499, 2025.

[12] Yuji Kanamitsu, Eigo Taya, Koki Tachibana, Yugo Nakamura, Yuki Matsuda, Hirohiko Suwa, and Keiichi Yasumoto. Estimating congestion in a fixed-route bus by using ble signals. *Sensors*, Vol. 22, No. 3, pp. 1–15, 2022.

[13] Eigo Taya, Yuji Kanamitsu, Koki Tachibana, Yugo Nakamura, Yuki Matsuda, Hirohiko Suwa, and Keiichi Yasumoto. Estimating congestion in train cars by using ble signals. In *The 2nd Workshop on Data-Driven and Intelligent Cyber-Physical Systems for Smart Cities (DI-CPS '22)*, pp. 1–7, 2022.