# DR-SPAAM: A Spatial-Attention and Auto-regressive Model for Person Detection in 2D Range Data

2022年6月16日　　20:27

# DR-SPAAM：2次元距離データにおける人物検出のための空間的注意と自己回帰のモデル

https://arxiv.org/abs/2004.14079

https://github.com/VisualComputingInstitute/2D_lidar_person_detection

## Abstract

Detecting persons using a 2D LiDAR is a challenging task due to the low information content of 2D range data. To alleviate the problem caused by the sparsity of the LiDAR points, current state-of-the-art methods fuse multiple previous scans and perform detection using the combined scans. The downside of such a backward looking fusion is that all the scans need to be aligned explicitly, and the necessary alignment operation makes the whole pipeline more expensive -- often too expensive for real-world applications. In this paper, we propose a person detection network which uses an alternative strategy to combine scans obtained at different times. Our method, Distance Robust SPatial Attention and Auto-regressive Model (DR-SPAAM), follows a forward looking paradigm. It keeps the intermediate features from the backbone network as a template and recurrently updates the template when a new scan becomes available. The updated feature template is in turn used for detecting persons currently in the scene. On the DROW dataset, our method outperforms the existing state-of-the-art, while being approximately four times faster, running at 87.2 FPS on a laptop with a dedicated GPU and at 22.6 FPS on an NVIDIA Jetson AGX embedded GPU. We release our code in PyTorch and a ROS node including pre-trained models.

2次元LiDARを用いた人物検出は、2次元レンジデータの情報量が少ないため、困難なタスクである。LiDARの点の疎密によって引き起こされる問題を軽減するために、現在の最先端

の方法は、複数の以前のスキャンを融合し、結合されたスキャンを使用して検出を実行します。このような後方視的融合法の欠点は、すべてのスキャンを明示的に位置合わせする必要があり、必要な位置合わせ操作がパイプライン全体をより高価にし、しばしば実世界のアプリケーションには高価すぎるということである。本論文では、異なる時刻に取得されたスキャンを結合する別の方法を用いた人物検出ネットワークを提案する。我々の手法であるDistance Robust SPatial Attention and Auto-regressive Model (DR-SPAAM)は、前方視のパラダイムに従うものである。DR-SPAAMは基幹ネットワークから得られた中間的な特徴をテンプレートとして保持し、新しいスキャンが利用可能になったときにテンプレートを再帰的に更新する。DR-SPAAMは、バックボーンネットワークの中間特徴量をテンプレートとして保持し、新しいスキャンが利用可能になったときにテンプレートを更新することで、現在シーンにいる人物を検出する。DROWデータセットにおいて、本手法は既存の最先端技術を凌駕し、専用GPUを搭載したノートパソコンでは87.2FPS、NVIDIA Jetson AGX組み込みGPUでは22.6FPSで動作し、約4倍高速化されました。我々は、PyTorchのコードと、事前に学習されたモデルを含むROSノードを公開します。

# I. INTRODUCTION

2022年6月16日　　20:35

## 2D LiDARの利点

- 周辺環境における人物の検出は、捜索救助、セキュリティ、ヘルスケアなど多くのロボットアプリケーションの重要な要件である。
- 現在、これは多くの場合、複数のRGB-Dカメラと、ディープラーニングに基づく物体検出器を組み合わせて実現されています[1], [2], [3]。
- しかし、そのようなカメラの限られた視野のために、検出は狭いフラストレーションに制限されます。
- さらに、遠距離での不正確な深度測定は、3D空間での正確な人物定位を困難にしています。
- その代わり、2D LiDARは、広い視野で正確な距離測定を高い取得率で提供します。そのため、人物検出のための有望なセンサーの選択肢となります。

## 2D LiDARの課題

- しかし、2次元LiDARの疎な距離測定に含まれる限られた情報は、信頼性の高い人物検出のための重要な課題となっています。
- 最近の開発では、オブジェクトを検出するために、以前のいくつかのスキャンを組み合わせることが有益であることが示されています[4], [5]。
- 特に、Beyerらは、1回のスキャンに比べ、5回のスキャンを蓄積することにより、検出結果が改善されることを報告した[4]。
- しかし、欠点は計算量の増加である。LiDARのエゴ運動とシーン内のオブジェクトの動きのために、異なる時間に記録されたスキャンは完全に整列されておらず、下流の処理のためにスキャンを融合するために高価なアライメント操作を実行する必要があります。
- [4]の場合、以前のスキャンの反復サンプリングに加え、オドメトリ情報を用いてアライメントを行う。
- このアライメントはスキャン数に対して線形な計算コストであり、5回前のスキャンを使用すると、すでに全体の検出パイプラインがモバイルプラットフォームでのリアルタイム処理には高すぎる。

## 人検出の従来手法(5回のスキャンを蓄積して検出)

- 原理的には、過去のスキャンの位置合わせと融合は、時間情報を集約するためのバックワードルッキングパラダイム(入力時に複数回スキャンを同時に入力)に従う。
- これに対し、フォワードルッキングパラダイム(ネットワーク上で過去の情報を保持：RNN型)は、現在の測定値に基づく表現を保持し、新しい測定値が利用可能になったときにこの表現を再帰的に更新するだけである。
- この表現には、過去のすべての測定値が含まれ、下流の処理に使用される。
- このようなフォワードルッキングパラダイムは、ビデオオブジェクト検出の分野で見つかり、研究者は、ビデオシーケンスの各フレームの入力特徴を再帰的に取り込み、現在のフレームで洗練された予測を出力するメモリモジュールを使用しています[6], [7], [8].

## 本論文について

- 本論文では、フォワードルッキングパラダイムに従って時間情報を集約した人物検出ネットワークを提案する。
- 本手法は、2次元スキャンを入力とし、各点に対して分類ラベルと人物の中心を指すオフセットベクトルを予測するDROW検出器[4]の既存アーキテクチャを利用する（図1）。
- 我々はDROWをで拡張し、各スキャンから基幹ネットワークからの中間特徴を集約することで、本手法が過去のすべての測定からの情報を利用できるようにする。

- また、異なるスキャンからの中間的な特徴を明示的に位置合わせする代わりに、空間的な注目メカニズムを用いて、隣接する位置からの特徴をその類似性に基づいて関連付けることで、計算コストを大幅に削減することが可能である。
- 本手法はDR-SPAAM（Distance Robust SPatial Attention and Auto-regressive Model）と呼ぶ。
- DR-SPAAMは、オリジナルのDROWデータセットで評価したところ、専用GPUを搭載したノートパソコンで87.2FPS、NVIDIA Jetson AGXで22.6FPSと、従来の手法を上回る性能を示しました。DR-SPAAMの高いフレームレートは、多くのロボットアプリケーションに適しています。

- 要約すると、我々は以下のような重要な貢献をする。
  - ○ 我々は、明示的な位置合わせ操作を必要とせずに、以前のLiDARスキャンからの情報を融合する空間的な注意と自己回帰モデルを提案する。
  - ○ 空間的注意と自己回帰モデルを用いた高速な2D LiDARベースの人物検出器であるDR-SPAAMを提案する。提案手法は、2次元レンジデータに基づく人物検出において、速度と検出性能の両方で、従来の最先端技術を凌駕している。
  - ○ 我々は、ロボットプロジェクトに容易に展開できるように、事前に学習されたモデルを持つROSノードを含むPyTorchでの実装を公開します1。

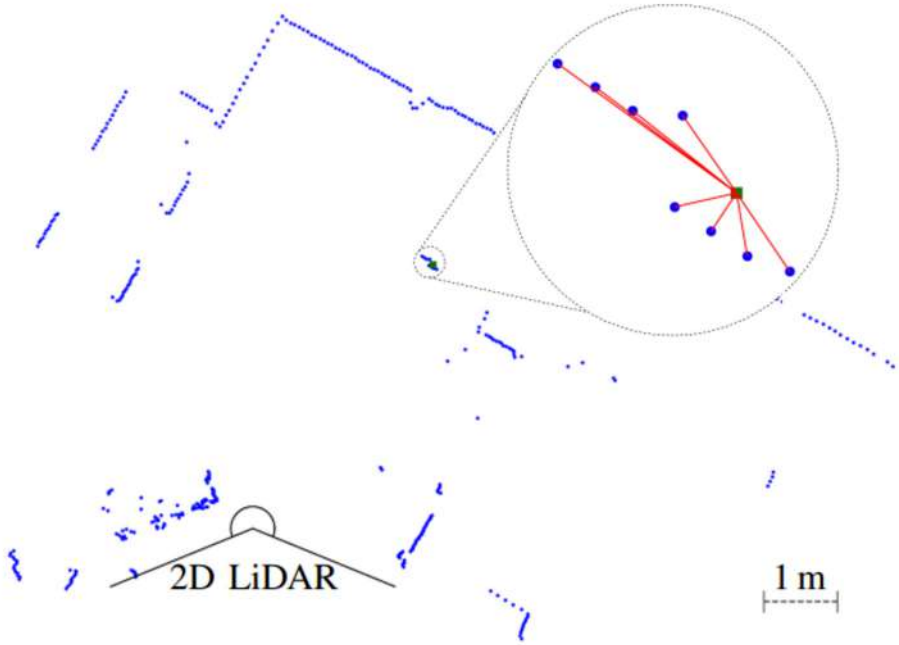


図1：2次元LiDARスキャンの鳥瞰図（青い点）、シーン内の人物（緑の四角）。
DR-SPAAMは各点に対して、分類ラベルと人物の中心までのオフセット（赤線）を出力し、
後処理によって検出値にグループ化される。

# II. RELATED WORKS

2022年6月16日　　20:35

## A.　2次元距離データからの人物検出

- 2次元レンジデータからの人物検出は、ロボット工学の分野では長い歴史を持っている
- 初期の手法は主にレンジデータから特定の形状を見つけるヒューリスティックに基づいていましたが、最も一般的なパラダイムは、スキャンを連結したセグメントに分割し、これらのそれぞれについて手作業で作成した特徴のセットを計算し、最後にそれらを分類して検出を作成するものです[9], [10], [11].
- 一般的な手法は，人物を検出するために個々の脚を検出し，追跡するアプローチ[10]，[11]と，1つのセグメントに両脚を含めることを直接目的とするアプローチ[9]に分類される.
- 理想的には、このような難しい設計上の選択を避けるために、データから直接人物の表現を学習することである。
- DROW検出器[12]は、2Dレンジデータで動作する最初の深層学習ベースの歩行補助検出器で、後に追加で人物を検出するように拡張されました[4]。
- 人物検出の結果を改善した重要な点の1つは、時間情報の統合であった。
- しかし、これは実行時間を大幅に増加させ、オンライン利用には不向きである。
- 我々は、DROW検出器の既存のアーキテクチャにフォワードルッキングパラダイムモジュールを組み合わせた新しい人物検出器を提案し、速度と検出品質の両方でオリジナルのDROWバージョンを凌駕しています。

- 他の者は、2Dレンジデータで深層学習に基づく方法を使用している。Ondruskaら[13]は、静止した2D LiDARから占有グリッドを作成し、各グリッドセルについてRNNに基づいてクラスラベルと将来のグリッド構成を予測する。
- 後のバージョンで、彼らはこのアプローチを移動するLiDARで動作するように拡張しました[14]。
- しかし、どちらの場合も、物体の検出や軌跡を作成しないため、人物検出器と比較することはできない。

## B. 3次元点群における物体検出

- 多くの研究は、3D LiDARから得られる点群中のオブジェクトの検出に焦点を当てている。
- このようなタスクは、自律走行アプリケーションにおいて重要な役割を果たすからである。
- 点群はデータ表現として、本質的に構造と近傍性の定義がないため、一般的なCNNアーキテクチャの使用を禁じている。

- この問題を解決するために、以前の研究では、点群を画像平面に投影するか、または既知の外部キャリブレーションを用いてRGB画像上で行われた2D検出をポップアップすることにより、画像ベースのオブジェクト検出方法を利用してきた[15], [16], [17], [18].
- しかし、これらの方法は、使用する2次元物体検出器によって実行時間や精度がボトルネックとなる。
- その後，現在の最先端手法を含め，この分野では投影を行わず，点群全体を利用する手法が開発された.
- VoxelNet [19]やSECOND [20]のような手法は、点群から変換された構造化3Dボクセルグリッド上で（スパース）コンボリューション [21], [22] を実行し、PointRCNN [23] などは PointNet [24] から着想を得たバックボーン [25], [26], [27] を使用して非構造化点群 を直接処理します。
- また、両者を組み合わせて2段階の検出器とするものもある [28], [29], [30]。

- 特に興味深いのは、Qiらによって提案されたVoteNetである[31]。
    - ○ VoteNetは、3次元点群が与えられると、DROW検出器と同様に、各点に対して物体中心へのオフセットベクトルを回帰させる（投票）が、後処理段階を用いる代わりに、別のサブネットワークを用いて、点ごとの予測を境界ボックス提案にグループ化する。
    - ○ このように、ネットワーク全体がエンドツーエンドで学習可能である。
    - ○ また，3D-MPA[32]では，点レベルのインスタンスマスクにグルー プ化されたインスタンス提案に対して投票を行う．この微分可能な投票集計は、2D LiDARのための興味深いアプローチである可能性もあります。
    - ○ しかし、我々は、一連のスキャンの時間的統合に焦点を当て、これら2つのアプローチは直交している。

- ハードウェア的には、2次元LiDARは3次元LiDARと動作原理が似ていますが、3次元点群に適用される手法は、2次元LiDARから得られるレンジデータに素直に適用することはできません。
- 私たちの知る限り、既存の3D手法を2Dレンジデータに直接適用した研究はなく、2Dレンジスキャンの情報の希少性を考えると、素朴な適応が可能かどうかはまだ分からない。

## C. 映像オブジェクトの検出

- ビデオオブジェクトの検出は、複数のフレームに同じオブジェクトが表示されたビデオシーケンスを入力とする特殊なオブジェクト検出タスクである。
- このため、あるフレームの情報を利用して、別のフレームでの検出を支援することが可能である。
- この情報の時間的伝播は、オブジェクトが速い動き、オクルージョン、またはカメラ

アングルの変更によって大きな外観変化を起こすことがあるため、重要である。

- 初期のアプローチ[33], [34], [35]では、各フレームのオブジェクトを独立して検出し、得られたバウンディングボックスに対してシーケンスレベルの後処理を適用している。
- このようなアプローチはエンド・ツー・エンドで最適化することができません.
- 後者のアプローチは、オプティカルフローを用いた明示的な特徴の位置合わせ[36]、[37]、[38]、またはメモリネットワークを用いた特徴の集約[6]、[7]、[8]により、フレーム間で直接特徴を集約することに着目しています。

- 本手法では、これらのアプローチと同様に、ネットワークを用いて連続するスキャン間で特徴量を集約する。
- 本手法では，メモリモジュールの代わりに，指数関数的に減衰する重みで前スキャンから次スキャンに情報を伝達する自己回帰モデルを用いる.
- 映像の物体検出では、隣接するフレームが類似していることが多く、新しい情報をほとんど導入しないため、長期的な特徴を集約できることが重要である。
- そのため、メモリネットワークがよく利用される。
- その代わりに、我々は連続したスキャンから短期的な特徴を集約することに着目し、検出に利用可能な情報を充実させることを目的とする。
- したがって、より複雑なメモリモジュールと比較すると、自己回帰モデルがより適している。
- さらに、我々は類似性に基づく空間的注意モデル[39]、[40]を用いて、時間的集約の前にまず近傍の空間情報を融合させることを提案する。

# III. METHOD

2022年6月16日　20:36

- 本節では、我々が提案する人物検出手法について詳細に説明する。
- まず、我々の基本アーキテクチャであるDROW検出器[4]について説明し、次に、時間情報を集約する様々なパラダイムについて議論する。
- 最後に、[4]とは対照的に、フォワードルッキングパラダイムアプローチで時間情報を集約する、我々の提案するDR-SPAAM 検出器を紹介する。

## A. DROW検出器

- DROW検出器[12], [4]は、2次元レンジデータから人物を検出する、ディープラーニングに基づく最初のアプローチであった。
- これは 3 つのステージから構成される.
- まず、異なる深度にわたる外観を正規化するために、生のスキャンは切り出しと呼ばれるポイントごとの小さなウィンドウに前処理される。
- これらの切り出しはネットワークによって物体か背景のどちらかに分類され、各切り出しに対して物体中心が回帰される。
- 最後に、回帰されたすべての物体中心（投票と呼ぶ）を収集し、最終的な検出の集合に集約する。

### 処理手順

- 前処理において、$N$個の点 $\{S_n^t\}_{n=1}^N$からなるスキャン$S^t \in \mathbb{R}^N_{>0}$が与えられるとN 個の切り出し $\{C_n^t\}_{n=1}^N$ が生成される。
- それぞれがLiDAR点$S_n^t$の周りのユークリッド空間における固定サイズのウィンドウに対応する。
- これは、各点 $S_n^t$の角度開度 $a_n^t$を次のように計算することで行われる。

$$a_n^t = 2 \cdot \tan^{-1} \frac{0.5 \cdot \bar{W}}{S_n^t} \qquad (1)$$

- $\bar{W}$ は切り出し幅を指定するハイパーパラメータである.
- この角度近傍内の点は、次に固定数$M$の点に再サンプリングされ、中心点の距離$S_n^t$を減算することによって中心化される。
- 深度$\pm d$の範囲外の背景点および前景点は切り取られ、閾値dに基づく一定値に置き換えられ、最後に切り出し$C_n^t$のすべての値は$[-1, 1]$に正規化される。
- 正規化された切り出し値 $\{C_n^t\}_{n=1}^N$ は分類と回帰のためにネットワークに渡される。
- 後処理では、投票が投票グリッドに蓄積され、NMS(非最大抑制)ステップが適用され、検出のセットが得られる。
- 検出は、検出に属する投票のクラス分布を集約することによってさらに精緻化される。
- DROW検出器の結果から、切り出しの前処理が有効であることがわかる。
- 特に、異なる距離での不均等なサンプリング密度（LiDARポイントは遠距離でまばらになる）による問題を軽減し、また、DROW検出器は再トレーニングを必要とせずに異なる角度分解能のLiDARで動作することができます。
- さらに、クリッピング操作により背景情報が除去されるため、ネットワークは同じ距離領域にある隣接点に焦点を当てることができます。
- 詳細については、[12]を参照されたい。

## B. 時間的な情報集約

- LiDARセンサから得られる測定値は情報量が少ないため（特に遠距離）、いくつかの検出器は空間の豊かな表現を得るために異なる時間に行われた測定値を集約し、この時間的集約が下流のタスクのパフォーマンスを向上させることが観察されています[4], [5].
- 時間情報を蓄積するための多くの一般的な技術は、過去数ステップ内の測定値を一緒に結合する、いわゆる後ろ向きのパラダイムに従っています。
- これらの計測値の間には、センサーの自走や動的な物体によって空間的なずれが生じることが多く、このずれはオドメトリや点群登録に基づいて補正されなければならない。
- 同様に、DROW検出器の第二バージョンは、後方視により時間情報を蓄積する[4]。
- これは、過去 $T$回の走査の切り出し$\{C_n^{t-T}, \cdots, C_n^T\}$を計算し、ネットワークの中間段階から得られた特徴量し$\{F_n^{t-T}, \cdots, F_n^T\}$を単純な総和で融合させるものである。
- 融合された特徴は、分類と回帰のために後段のネットワークに供給される。
- センサーの自走により、同じ角度指標$n$で行われた2つの距離測定$S_n^t$と$S_n^{t-1}$は、世界における単一の整列した点に対応しないため、DROW検出器は、特徴量$F_n^T$と$F_n^{t-T}$を融合する前にロボットオドメトリーを用いてズレを補正している。
- しかし、オドメトリだけでは、シーン内の動的な物体による位置ずれを補正することはできない。
- 人物の場合、これは特に重要で、同じ $n$番目の角度指数での LiDAR光線は、時間t-1では人物の脚に当たり、時間$t$では脚の間を通過して遠くの背景構造物に当たり、著しく異なる特徴量となる可能性があるためです。
- そこで、Beyer らは、切り出しの中心となるサンプリング位置を、現在の点$S_n^t$の位置に固定することを提案している。
- しかし、この場合、過去の走査の切り出しは、時間ステップごとに再計算する必要がある。
- DROWでは、オドメトリによるアライメントと固定位置サンプリングにより、過去5回のスキャンを組み合わせることで、1回のスキャンのみを使用した場合に比べて検出精度を向上させることができる。

- このようなバックワードルッキングアプローチによる性能向上は、計算時間の増加という代償を払うことになる。
- 現在の測定値と以前の測定値の間のずれを補正する必要があり、その結果、集計ウィンドウ内のフレーム数に対して計算時間が線形的に増加する。
- DROW検出器では、わずか5回のスキャンを使用するだけで、検出パイプライン全体のコストが高くなり、モバイルプラットフォームでのリアルタイムアプリケーションには適していません。
- 時間情報を集約する別のアプローチとして、フォワードルッキングパラダイムに従うことが挙げられます。
- フォワードルッキングアプローチでは、過去の複数の測定値を明示的に整列・結合する代わりに、現在の測定値に基づく表現を単純に保持し、新しい測定値ごとに表現を再帰的に更新します。
- 理想的には、この更新ステップはわずかな計算オーバーヘッドしか発生しません。
- その結果、フォワードルッキングアプローチは、時間窓のサイズに関する好ましくない実行時スケーリング動作なしに、過去からの情報を集約することができる。

## C. DR-SPAAM 検出器

- 我々は、時間情報を集約するためにフォワードルッキングパラダイムに従ったDistance Robust Spatial-Attention and Auto-regressive Model (DR-SPAAM)を提案する。
- 過去のスキャンについて空間的に整列した切り出しを計算する代わりに、類似性に基づく空間的注意モジュール[40]を使用し、
  ネットワークが空間的近傍からずれた特徴を関連付けることを学習するようにしたものである。
- さらに、自己回帰モデルを用いて表現を更新し、時間を通して前方に情報を集約する。
- 提案する検出器はDROW検出器よりも性能が高く、約4倍高速である。
- 図2にDR-SPAAMの構成図、図3に提案する空間注目と自己回帰モデルの構成図を示す。

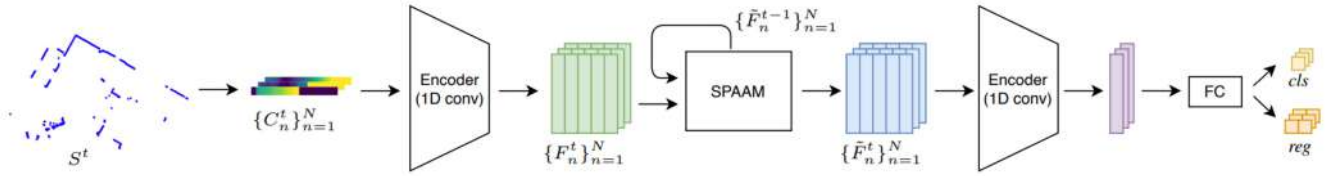


図2：DR-SPAAMのアーキテクチャの概要

現在のLiDARスキャンから各ポイントの切り出し$C_n^t$を作成し、

そこからネットワークコンピュータが中間表現$F_n^T$を作成する。

SPAAMモジュールを用いて、過去のスキャンから時間情報を集約する（図3および第III-C章参照）。

ネットワークは統合された表現$\tilde{F}_n^t$に基づき、分類ラベルを出力し、各点の相対的な物体中心を予測する。

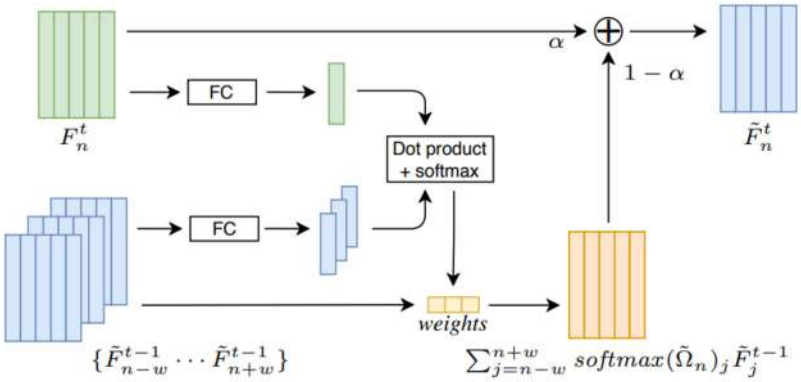


図3：SPAAMモジュールは、類似性に基づく空間的注意を用いて、過去のスキャンから特徴を関連付ける。

類似性で重み付けされた特徴は、その後、自動回帰的に現在の特徴と結合される。

- 位置ずれにより、2つの時間ステップで計算された特徴量$F_n^T$と$F_n^{t-T}$を素直に結合することはできない。
- [4]のように整列を明示的にモデル化するのではなく、類似性に基づく注意メカニズムを用いて、ネットワークに特徴の関連付けを学習させることを提案する。
- ある点 $S_n^t$に対して、前の時間 $t-1$ でその空間的な隣接点 $\{S_{n-w}^{t-1}, \cdots, S_{n+w}^{t-1}\}$ を調べ、前の各点 $S_j^{t-1}$で抽出した特徴と現在の点 $S_n^t$ からの特徴の対の類似度を算出する。
- ここで、$w$は近傍の大きさを定義するパラメータ、$F_n^t$は$S_n^t$の中間特徴量、$\psi$はニューラルネットワークで実現された、特徴を埋め込み空間に写像する汎用的な写像関数である。

$$\Omega_{nj} = \psi(F_j^{t-1})^T \cdot \psi(F_n^t) \qquad (2)$$

- そして、ソフトマックス関数を用いて類似度を重み付け係数に変換し、前フレームからの融合特徴量$\tilde{F}_n^t$を生成する。
- このモデルは、$S_n^t$の近傍の領域の情報を含む可能性が高い類似性スコアをより重視し、他の無関係な点の特徴を抑制するものである。
- そして、前フレームの融合特徴量$\tilde{F}_n^{t-1}$を現在の特徴量$F_n^t$と結合し、さらなる処理に利用することができる。
- しかし、このモデルは、連続する2つの走査からの情報しか結合しない。
- さらに過去のスキャンからの情報を集約するために、式（3）と自己回帰的なアプローチを組み合わせることを提案する。

$$\tilde{F}_n^{t-1} = \sum_{j=n-w}^{n+w} softmax(\Omega_n)_j F_j^{t-1} \qquad (3)$$

- 時刻 $t-1$ の融合特徴量$\tilde{F}_n^{t-1}$をテンプレートとして扱い、時刻 $t$ の新しい特徴量$F_n^t$が利用可能になると、テンプレートを更新して計算する。
- ここで、$\alpha \in [0,1]$は更新速度を制御するパラメータである。
- ここで、第 1 項は保存されたテンプレートに対する 我々の更新であり、第 2 項は過去からの情報を要約したものである。
- 式2とは異なり、ここでの項$\tilde{\Omega}_{nj}$は、現在の特徴$F_n^t$と、前のスキャンからではなく、前のテンプレートからの隣接特徴$\{\tilde{F}_{n-w}^{t-1}, \cdots, \tilde{F}_{n+w}^{t-1}\}$との間の類似性を表すことに注意する、すなわち、以下の通りである。

$$\tilde{F}_n^t = \alpha F_n^t + (1-\alpha) \sum_{j=n-w}^{n+w} softmax(\Omega_n)_j F_j^{t-1} \qquad (4)$$

- 更新されたテンプレート$\tilde{F}_n^t$は、最終的な分類とオフセット回帰のためにネットワークの後段へ渡される。
- オリジナルのDROW検出器と比較して、我々のDR-SPAAM検出器は計算複雑度が著しく低く、ロボットのオドメトリや以前のスキャンのカットアウトを再計算する必要がありません。
- また、自己回帰モデルにより、過去のスキャンを複数保存することなく、角度インデックスごとに1つのテンプレートだけを保持し、より大きな時間窓で情報を蓄積することができます。

# IV. EVALUATION

2022年6月16日　　20:36

## データセットについて

- DROWデータセット[12]、[4]は、SICK S300スキャナを使用して、屋内のリハビリテーション施設で記録されたもので、我々の手法を評価する。
- このデータセットには24,012のアノテーションされたスキャンが含まれ、トレーニング（17,665）、検証（3,919）、テスト（2,428）セットに分割されている。
- このデータセットには、車椅子、歩行器、人物の3つのクラスのオブジェクトの位置が含まれている。
- この研究では、特に人物の検出に焦点を当て、他の2つのカテゴリのアノテーションを無視するが、我々の手法は調整されたハイパーパラメータで他のクラスを扱うのに十分一般的であるはずである。
- 物体検出のコミュニティにおける標準に従って、我々は主な評価指標として異なる関連距離における平均精度（AP）を用いる。
- APdは、予測された位置の半径d m以内にマッチしないグランドトゥルースが存在する場合、検出が正とみなされることを意味する。
- [12]、[4]は精度-想起曲線下面積（AUC）を報告しており、これは定義上平均精度と同等であることに注目されたい。
- さらに、我々は、異なる精度と再現率の値の最大調和平均であるピークF1スコア（0.5mの関連距離を使用）、及び、精度と再現率が等しくなる値である等エラー率（EER）を報告する。
- すべてのモデルは、バッチサイズ8スキャン、40エポックの訓練セットで学習されます。
- DR-SPAAM 検出器では、指数関数的な減衰のため、過去に遡ったスキャンはあまり意味がないため、訓練中に過去に 10 フレーム読み込みます。
- アダム・オプティマイザーを使用し、初期学習率は10-3、完全学習中は10-6まで指数関数的に減衰する（各反復の後）。
- 分類には2値クロスエントロピー損失、回帰には回帰誤差のL1-ノルムを用いる。ネットワークの出力を検出に変換するために、[4]で紹介したのと同じ後処理スキームを使用する。
- 我々はHyperopt[41]を使用して、各モデルの検証セットにおいてAP0.5を個別に最大化することにより、投票ステップのハイパーパラメータを最適化する。
- 評価時には、学習時と同様に、各テストスキャンの過去10フレームの時間的コンテキ

ストを提供する。しかし、我々のアプローチは完全なシーケンスで実行するように容易に一般化できる。

## A. 定量的な結果

- テストセットを用いて提案手法を評価し、関連性閾値0.5 mにおける提案手法の平均精度、ピークF1スコア、等誤差率を表Iに報告する。
- また、ベースラインとして、1スキャンと5スキャンを用いた2つの再トレーニングDROWモデルの性能も報告する。オリジナルの実装と比較して、我々の再トレーニングされたモデルは、より小さな切り出しウィンドウと各切り出し内のより多くのサンプリングポイントを使用している。
- これは、検証セットでのより良いパフォーマンスに基づいて選択したものである（参照：Sec. IV-C）。比較に意味を持たせるために、再トレーニングされたDROWベースラインとDR-SPAAMの両方に同じ切り出しパラメータを使用し、オドメトリ情報は使用しません。
- また、Beyerらが[4]で報告した人物クラスに関するオリジナルのDROWスコアと、DROWデータセットで再トレーニングしたLeigh[11]とArras検出器[9]のスコアも表Iに記載しています。

表1: 0.5mの関連性閾値を用いたテストセットでの検出精度。
DR-SPAAMと我々の再学習ベースラインDROWは、オドメトリ情報を使用していないことに注意。

| Method | $AP_{0.5}$ | peak-F1 | EER |
|---|---|---|---|
| ROS leg detector [10] | 23.2 | 41.7 | 41.0 |
| Arras (re-trained) [9] | 47.6 | 50.3 | 50.1 |
| Leigh (re-trained) [11] | 57.2 | 64.3 | 62.3 |
| DROW ($T = 1$) in [4] | 59.4 | 61.5 | 61.4 |
| DROW ($T = 5$) in [4] | 67.0 | 65.9 | 64.9 |
| DROW ($T = 5$, + odom.) in [4] | 68.1 | 68.1 | **67.2** |
| DROW ($T = 1$) baseline | 66.6 | 66.1 | 65.2 |
| DROW ($T = 5$) baseline | 67.9 | 65.1 | 63.8 |
| DR-AM (w/o spatial attention) | 66.3 | 65.2 | 64.0 |
| DR-SPA (w/o auto-regression) | 68.0 | 67.0 | 66.1 |
| DR-SPAAM | **70.3** | **68.5** | **67.2** |

- その結果、DR-SPAAMはAP0.5の最高値である70.3%を達成し、ベースラインモデルを2.4%、[4]のオリジナルDROWを2.2%上回りました。
- このように、DR-SPAAMは、オドメトリ情報を用いないにもかかわらず、新たな最先端技術を確立しています。

- また、再学習したDROWとオリジナルのDROWを比較することで、特にシングルスキャンの場合に、提案する調整の有効性を観察することができます。

- 図4は、すべてのモデルの精度-再現性曲線を示しています。

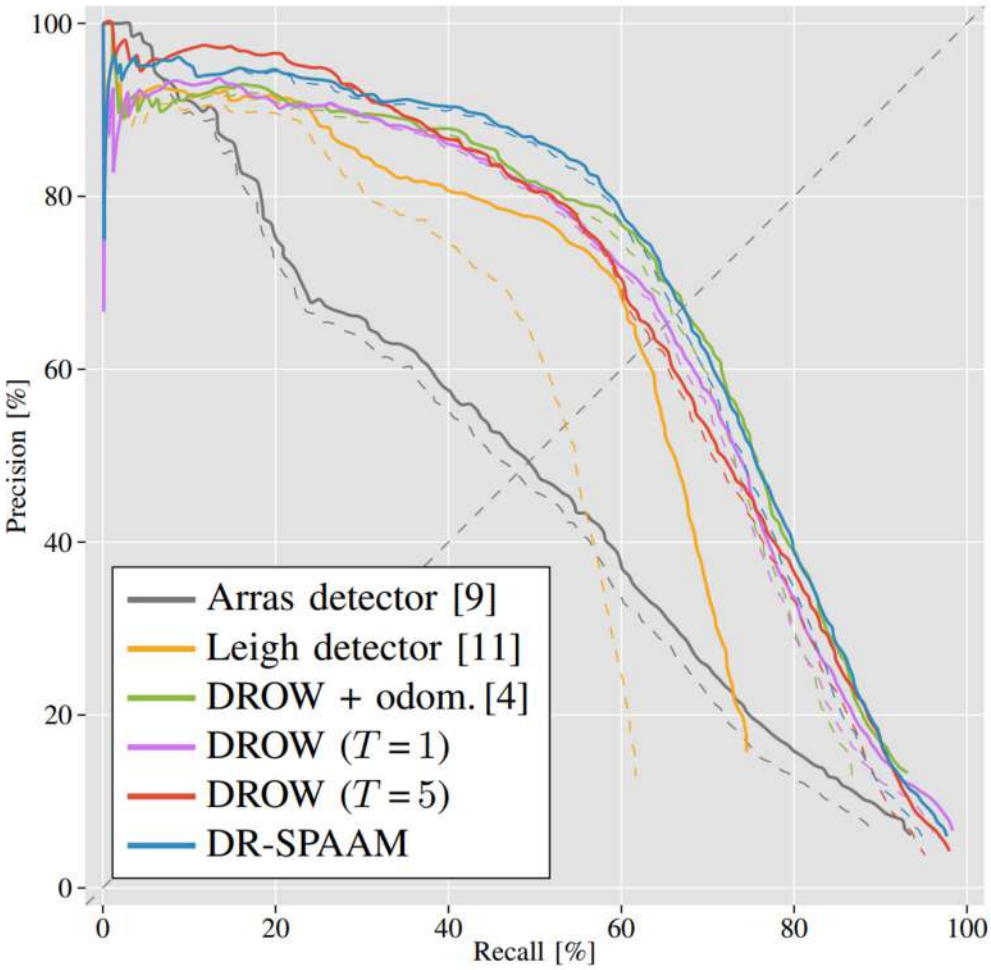


図4：ベースラインとDR-SPAAM検出器の精度-再現性曲線
（距離0.5m（実線）と0.3m（破線）で評価した場合）

- DROW (T=5)ベースラインが高精度領域で高いスコアを示した小さな領域を除き、DR-SPAAMが他のすべてのセットアップを上回っていることが分かります。
- 表Iは、時間集約モジュールの異なるコンポーネントの寄与を強調するアブレーション研究の結果も示しています。
- DR-AMは、空間的注意メカニズムからの加重和を使用せずに、自己回帰モデルが新しい特徴で直接更新されるネットワークに相当する。
- DR-AMは、1スキャンDROWのベースラインと比較して若干性能が悪く、ずれた特徴を素朴に組み合わせることは有益でないことを示しています。
- 一方、DR-SPAは、自動回帰モデルから得られる累積特徴テンプレートを用いず、空

間的注意を用いて現スキャンと前スキャンからの特徴を組み合わせるのみである。

- この2スキャンのアプローチは、5スキャンのDROWベースラインをすでに上回っており、学習ベースのアプローチを使って以前の測定値から特徴を取り込むことの利点を示している。
- DR-SPAAMは、2スキャンDR-SPAを上回り、より大きな時間窓で情報を集約することの利点を示している。

## B. 推論時間

- DR-SPAAMはより良い検出を達成するだけでなく、過去の切り出しの高価な再計算を行う必要がないため、計算量が大幅に削減されます。
- 我々は、TensorRTなどの推論時間高速化フレームワークを使用せずに、すべてのネットワークをPythonとPyTorchで実装し、完全なパイプラインの異なるコンポーネントの実行時間のプロファイルを作成しました。
- 2つのモバイルプラットフォームでのタイミング結果を表IIに報告する。ここでは、モバイルNVIDIA GeForce RTX 2080 Max-Q GPUとIntel-i7 9750H CPUを搭載したノートPCと、Jetson AGX Xavierを使用しています。

表 II: 2 つの異なるモバイルプラットフォームでの
異なるセットアップの計算時間（ミリ秒）とフレームレート。

| | | Laptop (RTX 2080) | | | | Jetson AGX | | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | $AP_{0.5}$ | cutout | net | vote | FPS | cutout | net | vote | FPS |
| DROW ($T=1$) | 66.6 | 7.0 | 1.4 | 6.1 | 68.6 | 63.3 | 4.8 | 29.3 | 10.4 |
| DROW ($T=5$) | 67.9 | 34.3 | 1.5 | 19.2 | 18.2 | 306.3 | 5.1 | 78.1 | 2.6 |
| DR-SPAAM | **70.3** | 7.0 | 2.0 | 7.7 | 59.8 | 62.0 | 6.9 | 33.6 | 9.7 |
| DR-SPAAM* | **71.8** | 1.1 | 1.9 | 8.5 | 87.2 | 4.2 | 7.7 | 32.4 | 22.6 |

- 表 II から、DR-SPAAMの計算時間はシングルスキャンのDROW法と同程度であることがわかります。
- これは、各タイムステップで現在のスキャンのみを処理する必要があり、DR-SPAAMのより高価なフュージョンがわずかなオーバーヘッドを追加するため、予想されることです。
- しかし、速度が同じでも、1スキャンDROW法の検出精度は圧倒的に低い。
- 一方、時間的統合を行うDROWは、過去のすべてのスキャンに対する切り出し再計算が必要なため、スキャン数に比例して実行時間が増加し、5スキャンを使用した時点ですでに大幅に遅くなることが分かっています。
- Jetson AGXのプラットフォームを考えると、2.6FPSというフレームレートはリアル

タイムアプリケーションとしては遅すぎるのですが、今回の9.7FPSはまだ十分に使用可能な範囲内です。

- ノートパソコンでは、DROWデータセットが約13FPSのフレームレートで記録されていることから、実際にはすべてのモデルをリアルタイムよりも高速に実行することができます。
- しかし、DR-SPAAMは、必要に応じて、より低いレイテンシーで検出を行い、モバイルプラットフォームに関連する低消費電力で、大幅に高いフレームレートで動作させることができます。

## C. ハイパーパラメータの選択

### 切り出し操作

- 切り出し操作は、窓の幅（W）と深さ（D）、および、再サンプリングに使用する点の数（N）によってパラメータ化される。
- [4]では，大きな歩行補助器クラスに対応するため，48点の大きな窓（1.66m×2.0m）を使用した.
- しかし，本研究では人物を検出することを目的としているため，人物の足跡にフィットするような小さな窓を用いることを提案する.
- DROWネットワークを用いて、1回のスキャンで切り出す窓の大きさについて、検証セットで実験を行った。
- その結果を表IIIに示す。

表 III: DROW (T = 1)検出器の検証セットのスコアと異なる切り出しパラメータ。

| $\overline{W}$ | $D$ | $N$ | $AP_{0.3}$ | $AP_{0.5}$ | peak-F1 | EER |
|---|---|---|---|---|---|---|
| 1.66 | 2.0 | 48 | 41.9 | 43.0 | 48.1 | 47.6 |
| 1.66 | 1.0 | 48 | 42.6 | 43.4 | 49.2 | 48.6 |
| 1.0 | 2.0 | 48 | 43.6 | 44.8 | 50.7 | 50.4 |
| 1.0 | 1.0 | 48 | 44.0 | 45.0 | 50.3 | 50.2 |
| 1.0 | 1.0 | 32 | 42.0 | 43.0 | 49.1 | 48.8 |
| 1.0 | 1.0 | 40 | 43.1 | 44.1 | 50.0 | 49.6 |
| 1.0 | 1.0 | 48 | 44.0 | 45.0 | 50.3 | 50.2 |
| 1.0 | 1.0 | 56 | **45.1** | **46.3** | **50.9** | **50.8** |
| 1.0 | 1.0 | 64 | 43.8 | 45.1 | 50.7 | 50.4 |

- この結果をもとに、切り出し窓を(1.0 m×1.0 m)、56点に設定し、学習するすべて

のモデルにこの切り出しパラメータを使用することにした。

- 検証セットのスコアはテストセットのスコアより大幅に低いことが観察される。
- オリジナルのDROWデータセットは、車いす、歩行者、人の3つのクラスを検出するために作成されています。
- このデータセットでは、人物クラスのアノテーションのみを残しており、検証セットでは、遠距離の人物のアノテーションが多く含まれていることが確認されました。
- 距離ロバストな前処理を用いても、距離が遠くなると情報が疎になり、検出の信頼性が低下するため、検証セットの難易度は高くなります。

### 空間的アテンションと自己回帰モデル

- 空間的注意と自己回帰モデルは、更新率αと探索窓のサイズWによってパラメータ化される。
- この2つのパラメータの意味は直感的に明らかであるが、適切な組み合わせを選択することは些細な作業ではない。
- 検証セットの結果（表IV）から、W = 11、α = 0.5を最終モデルとして選択した。
- これらの結果には明確なパターンは見られず、より大きな探索窓も特定の更新レートも一貫してより良く機能しない。
- パラメータ空間をより徹底的に探索することで、より優れたモデルが得られる可能性がある。

表 IV: DR-SPAAMのウィンドウサイズと更新レートを変えた場合の検証セットのスコア。

| $W$ | $\alpha$ | $AP_{0.3}$ | $AP_{0.5}$ | peak-F1 | EER |
|---|---|---|---|---|---|
| 7 | 0.3 | 45.0 | 46.2 | 52.5 | 52.5 |
| 7 | 0.5 | 49.5 | 50.9 | 54.6 | 53.6 |
| 7 | 0.8 | 46.8 | 48.3 | 54.1 | 54.0 |
| 11 | 0.3 | 51.5 | 53.0 | 56.8 | 56.4 |
| 11 | 0.5 | **52.7** | **53.9** | **57.3** | **57.3** |
| 11 | 0.8 | 47.4 | 48.7 | 53.6 | 53.2 |
| 15 | 0.3 | 51.5 | 52.8 | 56.1 | 55.3 |
| 15 | 0.5 | 50.7 | 52.1 | 55.0 | 54.7 |
| 15 | 0.8 | 47.0 | 48.2 | 53.1 | 53.0 |

## D. サンプリングレート

- 異なるLiDARはしばしば異なるサンプリングレートを持つ。

- 検出ネットワークは、異なるLiDARセンサーに展開される可能性が高いため、センサー仕様の違いに対するロバスト性を検証する必要があります。
- そこで、DROW（T = 5）とDR-SPAAMの2つのネットワークを取り上げ、異なるサンプリングレートを模擬した時間的にサブサンプリングされたシーケンスを用いて、テストセットで評価しました。
- 表Vは、異なる時間ストライドにおける検出精度を示している（ストライドがnの場合、n番目のスキャンのみを保持することを意味する）。

表 V: 時間軸を変えた場合のテストセットの結果。

| | DROW ($T=5$) | | | | DR-SPAAM | | | |
|---|---|---|---|---|---|---|---|---|
| Stride | $AP_{0.3}$ | $AP_{0.5}$ | p-F1 | EER | $AP_{0.3}$ | $AP_{0.5}$ | p-F1 | EER |
| 1 | **66.6** | **67.9** | **65.1** | **63.8** | 68.5 | 70.3 | 68.5 | 67.2 |
| 2 | 59.3 | 60.5 | 60.1 | 59.3 | 69.3 | 70.8 | **68.8** | **67.6** |
| 3 | 54.3 | 55.8 | 56.8 | 56.7 | **69.4** | **70.9** | 68.1 | 66.5 |
| 4 | 53.6 | 55.1 | 56.0 | 55.7 | 67.7 | 69.1 | 66.4 | 64.9 |
| 5 | 51.5 | 53.4 | 54.6 | 54.3 | 66.4 | 67.7 | 65.5 | 64.5 |

- この評価結果から、DR-SPAAMはサンプリングレートの変化に対して非常に強いことがわかる。
- サンプリング周波数を5倍（約2Hz）に下げても、AP0.3は2.1%しか低下しない。
- この結果は、学習された空間注意モジュールが、固定された時間文脈窓に依存することなく、外観の類似性に基づく情報を結合することの利点を示している。
- したがって、DR-SPAAMは、幅広いサンプリングレートのLiDARに搭載することができ、計算容量に制限がある場合は、サンプリングレートを下げて動作させることも可能である。
- 一方、DROW検出器の性能は、時間ストライドが大きくなると急激に劣化します。
- 連続したスキャンの時間差が大きいと、運動による位置ずれが大きくなり、DROW検出器の精度が低下します。

## E. 時間的な関連性

- DR-SPAAMは、時間的な集約の段階で、集約されたテンプレートと最新のスキャン特徴との間の類似度を計算する（式5）。
- この類似性は、異なるスキャンにまたがる点の関連付けにさらに利用することができる。

- ここで、予備的な例を示す。
- まず、連続した200フレーム（約15秒）のシーケンスから、DR-SPAAMを用いて各フレームの人物を検出する。
- 各検出点$D_i^t$に対して、類似度の高い点を選択することで、前スキャンでの対応点を見つけることができる。
- これらの対応する点が検出$D_j^{t-1}$にグループ化され、2つの検出の間の距離が閾値（0.5m）より小さい場合、両方の検出を1つのトラックレット(軌跡情報)にグループ化する。
- そうでなければ，$D_i^t$ を用いて新しいトラックレット(軌跡情報)を開始する。
- 200フレーム後に各トラックレット(軌跡情報)の信頼度を全検出数の平均値として計算する。
- 図5では、信頼度が0.35以上であり、少なくとも5つの検出からなるトラックレット(軌跡情報)をプロットしています。

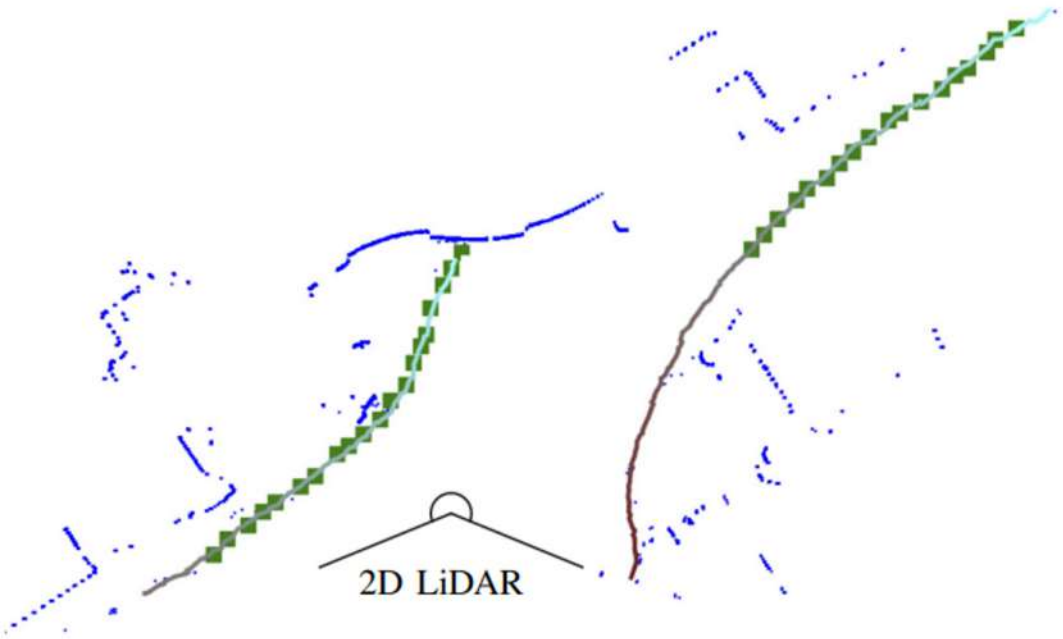


図5: DR-SPAAMで生成した200枚の連続スキャンのトラックレット。
青い点は重ね合わせたスキャン画像、
緑の四角はグランドトゥルースアノテーションである。
わかりやすくするために、人物と判定された点はプロットから除外していることに注意。
色のついた線はトラックレット(軌跡情報)であり、
色付けは検出時刻を符号化したものである。
検証セットには静止したLiDARで記録されたシーケンス（重ね合わせたスキャンをプロットするのに必要）がないため、
シーケンスはトレーニングセットから取得したものである。

しかしながら、ほとんどのスキャンはアノテーションされておらず
（トラックレットに沿ったアノテーションの欠落として示される）、
トレーニング中にネットワークにさらされることはなかった。

- シーン内の人物の軌跡がはっきりと見える。
- これらの関連付けは、トラッキングアルゴリズムに有益な情報を提供することができます。
- 今後の研究により、その可能性はさらに広がっていくでしょう。
- また、人物の速度と移動方向は、関連付けられた検出ペアから導き出すことができ、この情報は運動計画に役立つ可能性があります。

# V. CONCLUSION

2022年6月16日　　20:36

- 我々は、DROW検出器の距離に頑健な検出方式と、強力な空間的注意と自己回帰的な時間統合モデルを組み合わせたDR-SPAAM人物検出器を提案する。
- 空間的注意は、異なるフレームからのずれた特徴をその外観の類似性を用いて関連付けることができ、一方、自己回帰モデルは時間を通して前方の時間情報を統合する。
- DR-SPAAMは、従来の最先端手法と比較して、高い検出精度を達成すると同時に、大幅に高速化し、低消費電力のモバイルプラットフォームでもリアルタイムに動作させることが可能である。
- DRSPAAMは、異なる時間サンプリングレートのLiDARにうまく一般化できることが実験で示されており、提供されるコードとROSノードにより、本モデルが多くのロボットアプリケーションに有用であることが期待されます。

## 謝辞


- Francis Engelmann の貴重なフィードバックに感謝する。
- このプロジェクトは、EU H2020プロジェクト「CROWDBOT」（779942）およびBMBFプロジェクトFRAME（16SV7830）から部分的に資金提供を受けたものである。
- ほとんどの実験は、RWTH Aachen University CLAIX 2018 GPU Cluster (rwth0485)で行われました。