複数の車両搭載カメラから鳥瞰図で意味的に分割された画像への画像変換のためのSim2Realディープラーニングアプローチ*。

Lennart Reiher[1] and Bastian Lampe[1], Lutz Eckstein[2]

概要- 自動運転には、正確な環境認識が不可欠である。単眼カメラを使用する場合、環境中の要素の距離推定は大きな課題となる。カメラの視点を鳥瞰図(BEV)に変換することで、より簡単に距離を推定することができる。平面の場合、逆パースペクティブマッピング(IPM)は画像を正確にBEVに変換できる。車両や脆弱な道路利用者のような三次元物体は、この変換によって歪んでしまい、センサーに対する相対的な位置を推定することが難しくなる。本論文では、複数の車両搭載カメラからの画像を与えて、補正された360° BEV画像を得るための方法論について述べる。補正されたBEV画像は意味クラスに分割され、オクルージョン領域の予測を含む。ニューラルネットワークのアプローチは、手動でラベル付けされたデータに依存せず、実世界のデータにうまく汎化できるように、合成データセットで学習される。意味的に分割された画像を入力として用いることで、シミュレーションデータと実世界データの間の現実のギャップを縮め、本手法が実世界でうまく適用できることを示すことができる。合成データに対して行われた広範な実験により、IPMと比較して我々のアプローチの優位性が実証された。ソースコードとデータセットは https://github.com/ika-rwth-aachen/Cam2BEV で公開されている。

I. Introduction

近年、自動運転車(AV)の開発は、研究および産業界の両方から大きな注目を集めている。自動運転の重要な要素の一つは、AVの環境を正確に認識することである。安全で効率的な行動を計画するために不可欠である。オブジェクトリストや占有グリッドなど、さまざまなタイプの環境表現を計算することができる。どちらも環境中の要素の世界座標に関する情報を必要とする。環境理解のために一般的に使用される様々なタイプのセンサーの中で、カメラは低コストで確立されたコンピュータビジョン技術により人気がある。単眼カメラは画像平面上の位置情報しか提供できないため、トップダウンや鳥瞰図(BEV)をもたらす画像に透視変換を適用することができる。それは

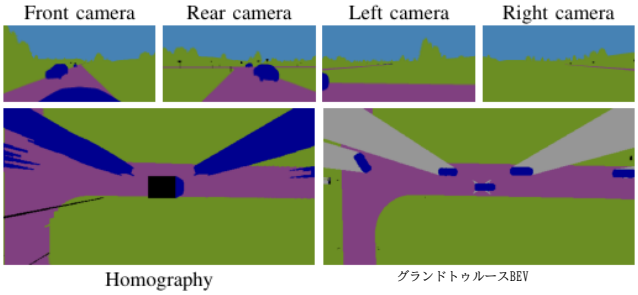

図1. 車両搭載カメラから意味的にセグメンテーションされた4つの画像にホモグラフィを適用し、BEVに変換することができる。我々のアプローチでは、視覚的な歪みなしに正確なBEV画像を計算する学習を行う。

カメラの前にある接地面と画像平面が一致する視点から見た、同じシーンの近似値。カメラ画像をBEVに変換する方法は、一般に逆視点マッピング(IPM)と呼ばれる[1]。

IPMは世界を平らに想定している。三次元物体や変化する道路の高さは、この仮定に違反する。すべてのピクセルを平面にマッピングすると、このようなオブジェクトの強い視覚的歪みが生じる。これは、他の車両や脆弱な道路利用者のような物体を、車両の環境内に正確に配置するという目標を妨げる。このため、IPMによって変換された画像は、しばしば車線検出や自由空間計算のアルゴリズムへの入力としてのみ機能し、その場合、フラットワールドの仮定はしばしば合理的である[2]。

IPMによってもたらされるエラーが修正できたとしても、BEV内の物体を検出するタスクが残されている。ディープラーニングアプローチは、画像のセマンティックセグメンテーションのようなタスクに強力であることが証明されているが、通常、膨大な量の手動でラベル付けされたデータを必要とする。シミュレーションは、BEV画像とそれに対応するラベルを提供することができるが、いわゆるリアリティギャップに悩まされる：シミュレーション環境で仮想カメラによって計算されたBEV画像は、例えば、現実世界の車両の上に撮影されたドローン画像とはかなり似ていないが、そのほとんどはシミュレーションにおける非現実的なテクスチャによるものである。したがって、シミュレーションで学習した複雑なタスクから実世界への汎化は、今のところ困難であることが証明されている。現実とのギャップを縮めるために、多くのアプローチはシミュレーションデータをより現実的なものにすることを目的としている。

*This research is accomplished within the project "UNICARagil" (FKZ 16EMO0289). We acknowledge the financial support for the project by the Federal Ministry of Education and Research of Germany (BMBF).

[1]The authors contributed equally to this work. They are with the Institute for Automotive Engineering (ika), RWTH Aachen University, 52074 Aachen, Germany. {firstname.lastname}@ika.rwth-aachen.de

[2]Lutz Eckstein is head of the Institute for Automotive Engineering (ika), RWTH Aachen University, 52074 Aachen, Germany. lutz.eckstein@ika.rwth-aachen.de

本論文では、IPMの基礎となる平坦性仮定によってもたらされる誤差の影響を受けないBEV画像を得るための方法論を提案する。シミュレートされた画像をよりリアルに見せる代わりに、意味的にセグメンテーションされたカメラ画像を計算することで、実世界のデータからほとんど不要なテクスチャを削除する。我々のアルゴリズムへの入力として、ニューラルネットワークを合成データのみで学習させることができる一方で、実世界のデータで目的のタスクを成功させることができることを示す。意味的に分割された入力により、アルゴリズムはクラス情報にアクセスできるため、IPMによって生成された画像の補正にこれらを組み込むことができる。出力は、入力シーンの意味的に分割されたBEVである。物体の形状は保存されるので、出力は自由空間を決定するためだけでなく、動的な物体の位置を特定するためにも使用できる。また、意味的に分割されたBEV画像には、未知の領域に対する色分けが含まれており、元のカメラ画像ではオクルードされている。IPMによって得られた画像と、所望のグランドトゥルースBEV画像を図1に示す。

本研究の主な貢献は以下の通りである：

- 複数の車両搭載カメラの画像を、BEVで意味的にセグメンテーションされた画像に変換できる方法論を提案する。
- 我々は、異なるニューラルネットワークアーキテクチャを使用して、我々の方法論の2つのバリエーションを設計し、比較する。
- ニューラルネットワークベースのモデルを学習するために、BEV画像の手動ラベリングが不要になるように、プロセスを設計する。
- 学習済みモデルの実世界での応用に成功したことを示す。

II. Related Work

BEVへの視点変換については、数多くの文献が取り上げている。自動車の文脈では、[4]と[5]の両方が、複数のカメラ画像をトップダウンのサラウンドビューに合成変換することを扱っている。ほとんどの作品は幾何学に基づくもので、地盤レベルの正確な描写に焦点を当てている。

BEVへの変換とシーン理解のタスクを組み合わせた作品はわずかである。しかし、物体検出は物体の形状に関する手がかりを与えることができ、そこから変換が恩恵を受けることができる。最近、以下に示すディープラーニングアプローチにより、複雑なニューラルネットワークが古典的なIPM技術の改善に役立ち、環境知覚に貢献できることが示された。

6]と[7]の焦点は、IPMアプローチによってもたらされるエラーを修正することである。動的で3次元の物体は、道路シーンの理解を向上させるために、[6]によって達成された変換BEVで除去されることが求められる。これに対して、[7]で提案された方法は、動的なオブジェクトを含む正面カメラを通して見た道路シーン全体の正確なBEV表現を合成することを目的としている。基礎となるタスクの生成的性質のため、どちらの手法もGenerative Adversarial Networks [8], [9]を採用している。

Palazziら[10]は、正面カメラの画像からBEVの車両バウンディングボックスを予測する。

Roddickら[11]は、ネットワーク内の正書法特徴変換を用いて、空間の3次元離散化を行うことで、3次元バウンディングボックスの計算における高度な物体検出を実証している。

粗く静的な意味マップを導くトップダウンフレームにおける意味的な道路理解は、[12]で達成されている。6]と同様に、このアプローチは動的なトラフィック参加者を除去しようとするものである。
我々の知る限り、意味的に分割された複数の画像を直接BEVに変換するアイデアを追求する唯一の情報源は、ブログ記事[13]である。しかし、詳細なテストや実世界のデータへの応用は不足している。設計されたニューラルネットワークは完全畳み込みオートエンコーダであり、正確な物体検出の範囲が比較的狭いなど、複数の弱点がある。

III. Methodology

我々は、画像解析によく使われるディープニューラルネットワークの一種である畳み込みニューラルネットワーク(CNN)の使用に基づいて、この方法論を確立する。ほとんどの一般的なCNNは1つの入力画像のみを処理する。車両に取り付けられた複数のカメラからの画像を融合するために、1入力ネットワークは、チャネル次元に沿って連結された複数の画像を入力とすることができる。しかし、手元のタスクでは、入力画像と出力画像の間に空間的な不整合が生じる。畳み込み層は局所的に動作する。つまり、入力の特定の部分の情報は、出力のほぼ同じ部分にマッピングされる。しかし、本問題に対するエンドツーエンドの学習アプローチは、複数の視点からの画像を扱うことができる必要がある。このことは、さらなるメカニズムの必要性を示唆している。
IPMは確かに誤差をもたらすが、この技術は少なくともグランドトゥルースのBEV画像に類似した画像を生成することができる。この類似性から、入出力画像間の空間的整合性を向上させるメカニズムとしてIPMを取り入れることは合理的であると思われる。IPMから得られる画像は、[6]や[7]でも中間ガイドビューとして使用されている。以下では、IPMの適用を含む、ニューラルネットワークベースの方法論の2つのバリエーションを紹介する。2つのニューラルネットワークアーキテクチャを紹介する前に、適用したデータ前処理技術について詳しく説明する。

A. オクルージョンへの対処

入力領域とこのタスクの望ましい出力のみを考慮すると、1つの難点がすぐに明らかになる：交通参加者と静的障害物は、BEV画像内のそれらの領域の予測をほとんど不可能にする環境の一部を閉塞する可能性がある。例として、トラックの後ろを走行するときにこのようなオクルージョンが発生する。トラックの前で起こっていることは、車両搭載カメラ画像だけでは確実に判断できない。
よく提起された問題を定式化するために、カメラの視点に遮蔽されたBEVの領域に対して、追加の意味クラスを導入する必要がある。このクラスは前処理でグランドトゥルースのラベル画像に導入される。各車両カメラについて、仮想光線は、そのマウント位置から、

図2. 元のグランドトゥルース画像に、オクルージョンクラスを含む修正ラベルを重ね合わせたもの(灰色陰影)。自動車(青)とバス(暗いターコイズ)は背後の地面を塞いでいる。駐車している乗用車の後ろにある建物はまだ見えており、バスもビューの左上隅にある建物の視界を遮っている。駐車車両は部分的に互いに隠蔽しているが、完全に可視化されたままである。

意味的にセグメンテーションされたグランドトゥルースBEV画像のエッジにキャストされる。光線は、特定のカメラの視野内にあるエッジピクセルにのみキャストされる。これらの光線に沿ったすべての画素は、以下の規則に従ってオクルージョン状態を決定するために処理される:

- some semantic classes always block sight (e.g. *building*, *truck*);
- some semantic classes never block sight (e.g. *road*);
- *cars* block sight, except on taller objects behind them (e.g. *truck*, *bus*);
- 部分的に隠蔽されたオブジェクトは完全に可視のままである。オ
- ブジェクトは、すべてのカメラの視点で隠蔽されている場合にのみ、隠蔽されたオブジェクトとしてラベル付けされる。

これらのルールに従って修正されたグランドトゥルースのBEV画像を図2に示す。

B. 射影前処理

IPM技術を我々の手法に組み込む一環として、ホモグラフィ、すなわち車両カメラフレームとBEV間の射影変換を導出する。正しいホモグラフィ行列の決定には、カメラの固有パラメータと外在パラメータが含まれ、以下に簡単に説明する。

均質な世界座標$x_w \in \mathbb{R}^4$と均質な画像座標$x_i \in \mathbb{R}^3$の関係は射影行列$P \in \mathbb{R}^{3\times4}$で次のように与えられる。

$$\boldsymbol{x}_i = \mathbf{P}\boldsymbol{x}_w\,. \tag{1}$$

投影行列は、カメラの固有パラメータ(焦点距離など)を行列Kと外在(回転Rとワールドフレームに対する並進t)にエンコードする:

$$\mathbf{P} = \mathbf{K}\,[\mathbf{R}|\boldsymbol{t}]\,. \tag{2}$$

道路平面$x_r \in \mathbb{R}^3$からワールドフレームへの変換$M \in \mathbb{R}^{4\times3}$が存在すると仮定すると、s.t.

$$\boldsymbol{x}_w = \mathbf{M}\boldsymbol{x}_r\,, \tag{3}$$

を用いると、画像座標から路面への変換が得られる。

$$\boldsymbol{x}_r = (\mathbf{PM})^{\quad}\boldsymbol{x}_i\,. \tag{4}$$

なお、(1)は無限に多くの世界点が同じ画像画素に対応するため、一般に反転不可能である。Mに符号化された平面の仮定により、可逆行列(PM)を構築することが可能になる。

実世界のカメラのPを決定するために、カメラキャリブレーション法[14]を使用することができる。

本アプローチの最初のバリエーション(セクションIII-C)の前処理として、IPMを車両カメラからのすべての画像に適用する。変換は、グランドトゥルースのBEV画像と同じ視野をキャプチャするように設定されている。この領域はすべてのカメラ画像の結合によってのみカバーされるため、まずIPMによって別々に変換され、次に1つの画像(以下、ホモグラフィ画像と呼ぶ)にマージされる。重なり合う領域、すなわち2台のカメラから見える領域の画素は、変換された画像の1つから任意に選択される。

C. バリエーション1:単一入力モデル

本アプローチの最初のバリエーションとして、カメラビューとBEVの間のギャップの大部分を埋めるために、セクションIII-Bで示したようにホモグラフィ画像を事前に計算することを提案する。ここでは、ニューラルネットワークの入力と出力の間の空間的な整合性をある程度提供する。そして、ネットワークのタスクは、IPMによってもたらされるエラーを修正することである。

我々の知る限り、特に手元の問題を対象とした単一入力ニューラルネットワークアーキテクチャは存在しない。しかし、ホモグラフィ画像と目的の出力画像は同じ空間領域をカバーしているため、セマンティックセグメンテーションのような他のタスクで成功が証明されている、既存のCNNを画像処理に使用することを提案する。

提案する単一ネットワーク入力法のアーキテクチャとして、DeepLabv3+を選択する。DeepLabv3+は、[15]で提示された、意味的な画像セグメンテーションのための最先端のCNNである。MobileNetV2[16]とXception[17]を用いて、2つの異なるネットワークバックボーンをテストした。得られたニューラルネットワークは、約2.1Mと41Mの学習可能なパラメータを持つ。

D. バリエーション2:多入力モデル

セクションIII-Cで示した最初のネットワークアーキテクチャとは対照的に、我々は、車両カメラからのすべての非変換画像を入力として処理する第2のニューラルネットワークを提案する。したがって、非変換カメラビューの特徴を抽出するため、IPMによってもたらされる誤差の影響を十分に受けることはない。空間的不整合の問題に対処する方法として、射影変換をネットワークに統合する。

複数の入力画像と1つの出力画像に対するアーキテクチャを構築するために、既存のCNNを複数の入力ストリームに拡張し、内部で前記ストリームを融合させることを提案する。その単純さと拡張性の良さから、以下に示す拡張の基礎として、一般的なセマンティックセグメンテーションアーキテクチャ U-Net [18]を選択する。

基本アーキテクチャは、逐次プーリングとアップサンプリングに基づく畳み込みエンコーダとデコーダのパスから構成される。

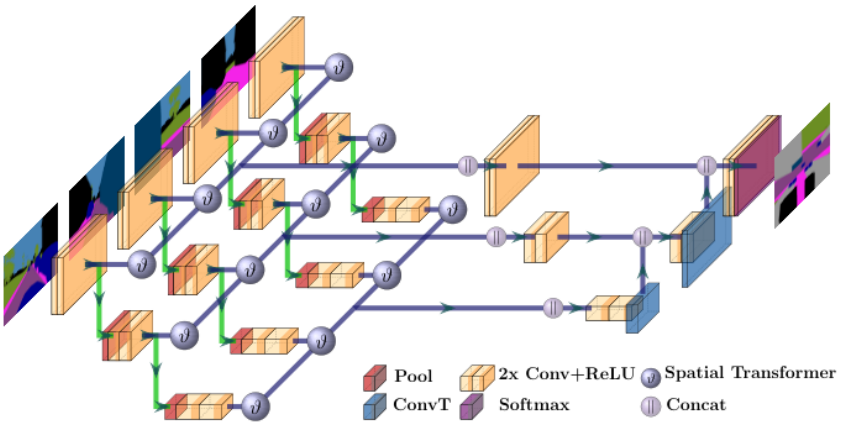

図3. uNetXSTアーキテクチャは、各入力画像に対して別々のエンコーダパスを持つ(緑色のパス)。各スケールレベルでのスキップ接続の一部として(紫色のパス)、特徴マップは射影変換され(θブロック)、他の入力ストリームと連結され(¦ブロック)、畳み込まれ、最後にデコーダパスのアップサンプル出力と連結される。この図は、2つのプーリング層と2つのアップサンプリング層のみを持つネットワークを示しており、実際の学習済みネットワークはそれぞれ4つの層を含んでいる。

さらに、エンコーダ側の高解像度特徴量は、各スケールでスキップ接続を介してデコーダ側のアップサンプリング出力と結合される。図3は、複数の入力画像を扱い、空間的な整合性を追加するために導入された2つの拡張を含むアーキテクチャを示す:

1) エンコーダの経路は、各入力画像に対して個別に複製される。各スケールについて、各入力ストリームからの特徴は連結され、単一のデコーダパスへのスキップ接続を構築するために畳み込まれる。
2) 入力ストリームを連結する前に、Spatial Transformer [19]ユニットは、IPMによって得られた固定ホモグラフィを使用して、特徴マップを射影変換する。これらの変換器について、図4で詳しく説明する。

ニューラルネットワークは、任意に多くの入力と空間変換ユニットに拡張されていることから、uNetXSTと名付けられた。約9.6Mの学習可能なパラメータを含む。

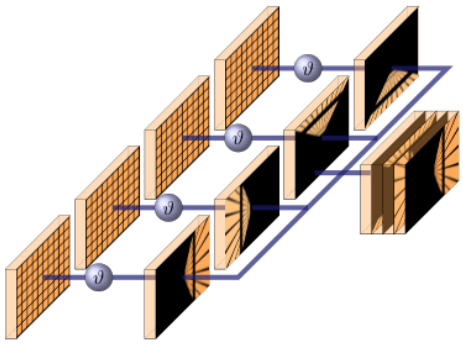

図4. θブロックは空間変換ユニットに似ている。先行する畳み込み層(オレンジ色の格子層)からの入力特徴マップは、IPMによって得られたホモグラフィによって射影変換される。異なるカメラの入力ストリーム間で変換が異なる。変換された特徴マップはすべてグランドトゥルースBEVと同じ視野を捉えているため、空間的整合性が確立される。変換された特徴マップは、1つの特徴マップに連結される(参照:¦-block)。

IV. 実験セットアップ

前述した方法論を評価するために、シミュレーションデータでニューラルネットワークを完全に訓練する。以下では、合成データセットと学習セットアップを示す。

A. データ取得

提案手法の学習と評価に使用したデータは、シミュレーション環境であるVirtual Test Drive (VTD) [20]で作成した。記録ツールチェーンは、対応するラベルを含む潜在的に任意の数のサンプル画像を生成することを可能にする。

シミュレーションでは、自車両は360°の全周囲をカバーする4台の同一の仮想広角カメラを装備している。グランドトゥルースデータは、仮想ドローンカメラによって提供される。BEVのグランドトゥルース画像は自車両の上方を中心にしており、近似視野は70m×44mである。

入力画像とグランドトゥルース画像はともに964px×604pxの解像度で記録されている。すべての仮想カメラは、現実的な画像と意味的に分割された画像の両方を生成する。セマンティックセグメンテーションのために、可視領域(道路、歩道、人、車、トラック、バス、自転車、障害物、植生)に対して9つの異なるセマンティッククラスが考慮される。
シミュレーション時間を短くすることと、データの多様性を最大化することのトレードオフとして、画像を2Hzで記録する。合計で、データセットには、トレーニング用の約33,000サンプルと検証用の約3700サンプルが含まれ、各サンプルは複数の入力画像と1つのグランドトゥルースラベルのセットである。指定された空間領域でのみ本手法を動作させる必要があるため、シミュレーション世界(道路、建物など)の静的要素は、訓練データと検証データの間で同じままである。

後に我々の手法の実際の応用をテストするために、2つ目の合成データセットを記録し、単一のフロントカメラで使用できるようにする。

このシナリオでは、可視エリア(道路、車両、占有スペース)については3つのクラスのみが考慮され、車両の前方エリアのみが注目される。このため、グランドトゥルース画像はエゴ・ビークルと左寄せされる。2つ目のデータセットには、トレーニング用の約32,000サンプルと検証用の3,200サンプルが含まれる。

B. 学習セットアップ

学習と推論の時間を比較的短くするため、ネットワークの入力画像とターゲットラベルはアスペクト比2:1にセンタートリミングされ、512px×256pxの解像度にリサイズされる。入力画像はワンホット表現に変換される。データセットのクラスの不均衡に対抗するため、損失関数は意味クラスを相対出現率の対数に従って重み付けするように修正される。学習中、学習率1e-4、パラメータ$\beta_1 = 0.9$、$\beta_2 = 0.999$のAdamオプティマイザをサイズ5のバッチに適用する。

C. 評価指標

IoU(Intersection-over-Union)スコアは、ある意味クラスを予測するタスクにおけるモデル性能の主な指標として使用される。クラスIoUスコアは、1つのMean Intersection-over-Union (MIoU)スコアに平均化される。

V. 結果と考察

このセクションでは、我々の手法のバリエーションの性能を互いに比較し、古典的なIPM手法と比較した我々の手法の全体的な改善点について議論する。IPMによって得られた標準的なホモグラフィ画像をベースラインとして評価を行う。

DeepLab XceptionとDeepLab MobileNetV2の2つの単一入力モデルと、多入力モデルuNetXSTの結果を示す。ホモグラフィを我々のアプローチに組み込むことの利点を定量化するために、IPMを用いない代替モデルの結果も示す。最初の手法のバリエーションであるDeepLabのモデルでは、セクションIIIの冒頭で説明したように、これは単純に複数の入力画像をチャンネル次元に沿って連結することを意味する。uNetXSTモデルの場合、これは空間変換ユニットを切除することを意味する。以下では、これらの簡略化したモデルをアスタリスク(*)で示す。
さらに、提案手法がシミュレーションデータから実世界のデータへ一般化できるという仮説を定性的に検証する。

TABLE I
検証セットにおけるミオユースコア(%)

Model	MIoU
uNetXST	**71.92**
DeepLab Xception	71.35
DeepLab MobileNetV2	66.60
DeepLab Xception*	60.13
DeepLab MobileNetV2*	55.09
uNetX*	45.95
Homography	30.17

A. 合成データでの結果

ベースラインと比較した我々のモデルの性能を表 I に示す。

uNetXSTモデルは検証セットで最も高いMIoUスコアを達成した。これは、uNetXSTが2番目に性能の良いネットワークであるDeepLab Xceptionよりも、学習可能なパラメータが大幅に少ないにもかかわらず、そうである。この結果は、IPMによって透視誤差が導入される前に、uNetXSTを使用するアプローチが、非変換カメラ画像から特徴を抽出することができるという仮説の証拠と見ることができる。
我々のアプローチ(*)からIPMを省略したアブレーション研究の結果は、誤ったホモグラフィビューが実際にパフォーマンスの向上に役立つことを示唆している。ホモグラフィのベースラインそのものと比較して、我々の提案するアプローチは一般的にかなり高い性能を達成する。この値は、我々の手法の両バリエーションが、環境認識のためのIPMによって得られた結果をうまく改善できることを示している。

クラスでの性能をさらに分析するために
に基づいて、それぞれのクラスIoUスコアを表IIに示す。提案された3つのネットワークは、道路や植生など、広い範囲をカバーする意味クラスの予測において、いずれも最も良い性能を示した。自動車、トラック、バスはいずれもダイナミックな交通参加者であり、良好なIoUスコアを達成している。すべてのモデルは、自転車、特に人物の正しい予測と定位に苦戦している。これは、両クラスがBEVにおいて小さなオブジェクトを表し、また学習データセットにおいて最も出現率が低いことに起因する。自転車と人物のuNetXSTの結果は、この方法が生のカメラ画像と変換されていないカメラ画像の処理から実際に利益を得ることができることを示している。クラスの不均衡に対抗するために、重み付き損失関数とは別に、さらに対策を講じることで、これら2つのクラスに関する結果を改善できる可能性がある。IPMを用いないモデル(*)は、一貫して対応するモデルよりも性能が悪い。

TABLE II
評価セットにおけるクラスI O U Sコア(%)

Model	Road	Sidewalk	Person	Car	Truck	Bus	Bike	Obstacle	Vegetation	Occluded
uNetXST	**98.10**	93.36	**13.56**	**80.90**	65.82	62.10	**32.43**	88.99	**97.27**	**86.62**
DL Xception	98.06	**94.02**	6.93	80.21	**65.94**	**65.98**	30.80	**89.05**	97.09	85.42
DL MobileNetV2	96.93	91.51	0.00	76.05	60.33	64.92	16.79	85.83	96.28	77.31
DL Xception*	96.60	88.81	0.20	68.18	53.63	32.80	2.74	84.84	95.85	77.61
DL MobileNetV2*	94.68	84.12	0.00	59.09	43.91	22.39	3.75	79.75	94.35	68.83
uNetX*	89.80	77.15	0.00	42.36	24.27	13.59	0.00	75.43	91.16	45.70
Homography	77.32	75.78	0.07	4.27	8.56	8.55	0.38	37.06	89.74	0.00

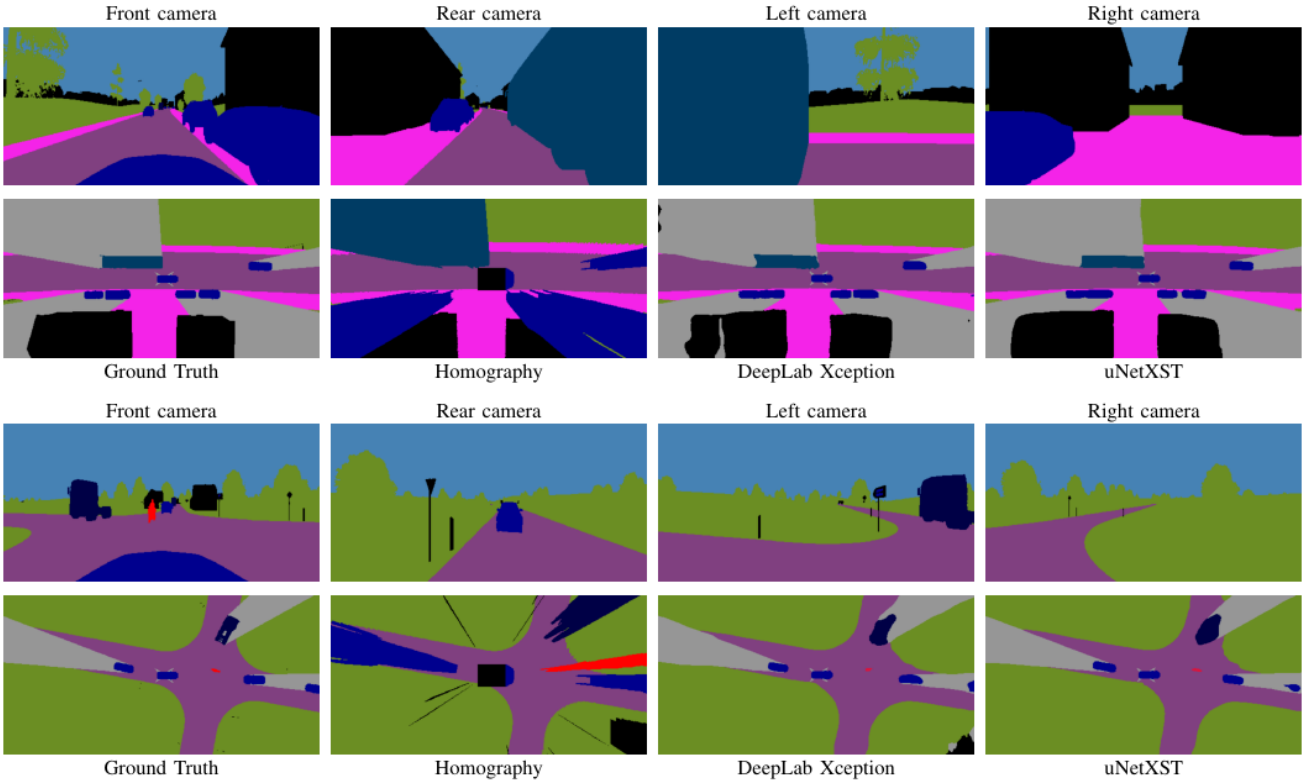

図5. 検証セットのシミュレーションデータに対する結果例

図5に描かれた例を分析することで、我々の2つの手法のバリエーションとベースラインの定性的な比較を行うことができる。両方の例示的なシーンについて、4つの車両搭載カメラの入力画像、グランドトゥルース画像、ホモグラフィ画像、および我々のDeepLab XceptionとuNetXSTアプローチからの予測値を示す。

IPMのフラットワールドの仮定によってもたらされる誤差は、ホモグラフィ画像にはっきりと現れている。我々の2つのモデルは、シーンの正しいBEVを計算するのに良い性能を発揮する。

最初の例では、移動車両と駐車車両は特によくローカライズされ、予測されたオブジェクトの寸法はグランドトゥルースデータと密接に一致する。オクルージョンの影は合理的にキャストされ、2つの建物の検出によって遮られる。uNetXSTとは対照的に、DeepLab Xceptionモデルはホモグラフィ画像から建物の寸法を確実に推測できないことに注意してください。2つ目の例は、自動車、トラック、オートバイとの4通りの交差点で、もうひとつの困難なシーンを描いたものである。この結果は、交通参加者のローカライズが良好であることを示している。物体の寸法の推定は、最初の例と比較して悪いようである。しかし、交差点により、車両はわずかに回転しており、ほとんどの学習サンプルではそうではない。右端の車はほとんど完全に隠されているため、正しく検出されないことに注意してください。

ホモグラフィ画像と比較して、我々のアプローチの両バリエーションは、IPMによってもたらされるエラーをうまく除去している。さらに、ビークルカメラの観点からオクルードされたBEVの領域を合理的に予測する。

B. 実世界での応用

実世界のデータで我々の手法をテストするためには、我々のアプローチの入力として、意味的にセグメンテーションされたカメラ画像を得る方法が必要である。この目的のために、我々はセマンティックセグメンテーションのために、内部ラベル付けされたテストデータセットで79.56%のMIoUスコアを達成する追加のCNNを採用する。

DeepLab XceptionとuNetXSTモデルによって計算された2つの実世界シーンのBEVを図6に示す。どちらも他の交通参加者の位置と次元に対して妥当な予測を行うが、uNetXSTモデルはより滑らかで定性的に良い結果を生成する。

最初の例では、どちらのネットワークも、左側の駐車車両と前方の車の位置と寸法を合理的に予測する。2番目の例では、5台の可視車両、たとえ部分的に隠蔽された車両であっても、両方のモデルによって検出される。ここで、uNetXSTは一般的に、特に右側のより遠くの車両に対して、より合理的なオブジェクト寸法を生成する。

車両ダイナミクスのため、現実には、シミュレーションデータの場合のように、道路面に対する車両カメラの姿勢は一定ではないことに注意してください。このように、両モデルで使用された固定IPM変換は、図6に描かれたシーンで誤較正される可能性がある。したがって、車両ダイナミクスを測定し、動的な変換変化をネットワーク推論に組み込むことで、実世界での結果を改善することができる。

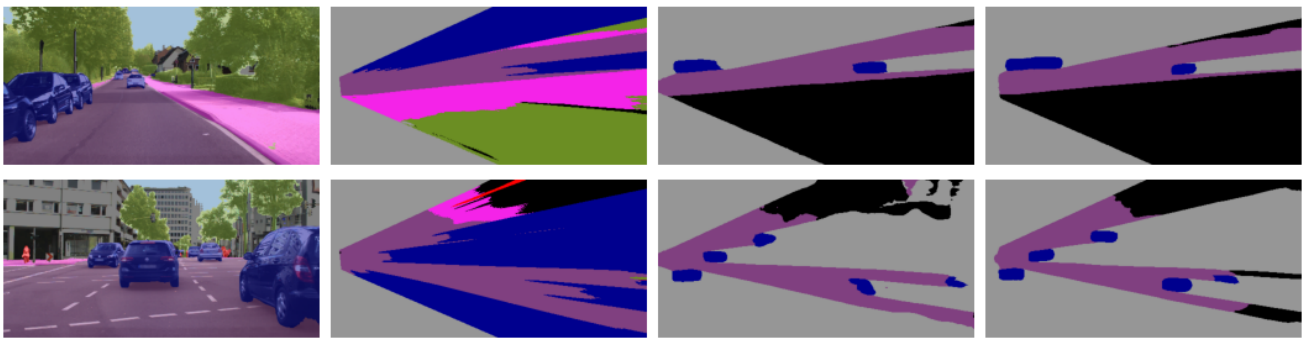

図6. 実世界でのアプリケーションの結果例

VI. Conclusion

複数の車両搭載カメラの画像を、鳥瞰図において意味的にセグメンテーションされた画像に変換できる手法を提案した。この過程で、逆パースペクティブマッピングの基礎となる不正確な平坦性の仮定に起因するエラーが除去される。合成データセットの使用と、カメラ画像の意味的にセグメンテーションされた表現への入力抽象化により、BEV画像の手動ラベリングなしで実世界のデータへの適用が可能になる。さらに、本手法はBEV画像中のオクルージョン領域を正確に予測することができる。複数の入力を処理し、ネットワーク内変換を行うニューラルネットワーク uNetXST を設計した。このように、このタスクにおいて、ネットワークはDeepLab Xceptionのような一般的なアーキテクチャを凌駕することができる。我々のアプローチで学習した全てのモデルは、Inverse Perspective Mappingのみを適用した結果を定量的、定性的に上回る。

さらに研究を進めるのは、提示された方法論がカメラによる環境認識に貢献できる可能性があるからである。1つの有望なアイデアは、深度情報のようなさらなる入力を取り入れることである。奥行き情報は、ステレオカメラから計算されたもの、単眼カメラの奥行き推定のためのアプローチによって推定されたもの、またはLiDARのようなセンサーから得られたものである。実世界のアプリケーションに関しては、360*マルチカメラセットアップでこのアプローチをテストする必要があり、フロントカメラ画像だけでなく、セマンティックセグメンテーションの性能も高くなければならない。

References

[1] H. A. Mallot, H. H. Bülthoff, J. J. Little, and S. Bohrer, "Inverse perspective mapping simplifies optical flow computation and obstacle detection," *Biological Cybernetics*, vol. 64, pp. 177–185, 1991.

[2] A. Bar Hillel, R. Lerner, D. Levi, and G. Raz, "Recent progress in road and lane detection: a survey," *Machine Vision and Applications*, vol. 25, no. 3, pp. 727–745, 2014.

[3] S. Zhao, B. Li, X. Yue, Y. Gu, P. Xu, R. Hu, H. Chai, and K. Keutzer, "Multi-source Domain Adaptation for Semantic Segmentation," in *Advances in Neural Information Processing Systems*, 2019, pp. 7285–7298.

[4] K. Sung, J. Lee, J. An, and E. Chang, "Development of Image Synthesis Algorithm with Multi-Camera," in *2012 IEEE 75th Vehicular Technology Conference (VTC Spring)*. IEEE, 2012, pp. 1–5.

[5] B. Zhang, V. Appia, I. Pekkucuksen, Y. Liu, A. U. Batur, P. Shastry, S. Liu, S. Sivasankaran, and K. Chitnis, "A Surround View Camera Solution for Embedded Systems," in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2014, pp. 676–681.

[6] T. Bruls, H. Porav, L. Kunze, and P. Newman, "The Right (Angled) Perspective: Improving the Understanding of Road Scenes Using Boosted Inverse Perspective Mapping," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, 2019, pp. 302–309.

[7] X. Zhu, Z. Yin, J. Shi, H. Li, and D. Lin, "Generative Adversarial Frontal View to Bird View Synthesis," *arXiv:1808.00327 [cs]*, 2019.

[8] Jürgen Schmidhuber, "Making the World Differentiable: On Using Self-Supervised Fully Recurrent Neural Networks for Dynamic Reinforcement Learning and Planning in Non-Stationary Environments," Tech. Rep., 1990.

[9] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems – NIPS'14*, vol. 2. MIT Press, 2014, pp. 2672–2680.

[10] A. Palazzi, G. Borghi, D. Abati, S. Calderara, and R. Cucchiara, "Learning to Map Vehicles into Bird's Eye View," in *Image Analysis and Processing - ICIAP 2017*, vol. 10484. Cham: Springer International Publishing, 2017, pp. 233–243.

[11] T. Roddick, A. Kendall, and R. Cipolla, "Orthographic Feature Transform for Monocular 3D Object Detection," in *British Machine Vision Conference (BMVC)*, 2019.

[12] S. Sengupta, P. Sturgess, L. Ladicky, and P. H. S. Torr, "Automatic dense visual semantic mapping from street-level imagery," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 857–862.

[13] M. Dziubiński. (2019, 05) From semantic segmentation to semantic bird's-eye view in the CARLA simulator. [Online]. Available: https://medium.com/asap-report/from-semantic-segmentation-to-semantic-birds-eye-view-in-the-carla-simulator-1e636741af3f

[14] A. Kaehler and G. R. Bradski, *Learning OpenCV 3: Computer Vision in C++ with the OpenCV Library*, 1st ed. O'Reilly Media, 2017.

[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *Computer Vision – ECCV 2018*, vol. 11211. Springer International Publishing, 2018, pp. 833–851.

[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 4510–4520.

[17] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 1800–1807.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351. Springer International Publishing, 2015, pp. 234–241.

[19] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial Transformer Networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, vol. 2. MIT Press, 2015, pp. 2017–2025.

[20] K. von Neumann-Cosel, M. Dupuis, and C. Weiss, "Virtual Test Drive – Provision of a Consistent Tool-Set for [D,H,S,V]-in-the-Loop," in *Proceedings of Driving Simulation Conference Europe, 2009*, 2009.