Vol.2021-CVIM-225 No.19
2021/3/4

マルチタスク深層学習による車載単眼カメラ映像のMotion Segmentation

鷲本 昂樹[†1,a)]　吉岡 理文[†1,b)]　井上 勝文[†1,c)]

概要：自動車やロボットの自律走行技術の需要が高まる中，現在，コンピュータビジョンの分野では，深層学習を用いた車載単眼カメラ映像に対する Motion Segmentation が注目を浴びている．Motion Segmentation は，画像内における，建物や停止した車のような静的領域と，歩行者や走行車のような動的領域を識別する．本研究では，連続画像のみを入力データとし，Motion Segmentation を Optical Flow とのマルチタスクで学習させる手法を提案する．

1. はじめに

自動車やロボットの自律走行技術の需要が高まる中，センサ情報から，3 次元空間を認知する研究が盛んに行われている．コンピュータビジョンの分野では，センサ情報として，主にカメラ映像が扱われており，観測結果から，Depth[1] や Ego-Motion[1] のような 3 次元データを推定する．このような 3 次元データは，Simultaneous Localization and Mapping[2] (SLAM) に代表される，高度なリアルタイム環境復元システムを構築するために，高い推定精度が求められる．最近では，カメラ映像として，車載単眼カメラ映像がよく用いられる．これは，車載単眼カメラ映像が，収集や取扱いが容易で，自動車のような移動マシンに装着するセンサから得られる情報の中でも，汎用性が非常に高いデータと考えられているからである．ただし，カメラの視点が，単一かつ非固定であるため，フレーム間の見かけの変化が，カメラとオブジェクトのどちらの動きに依存するかについての認識が曖昧となる．これは，車載単眼カメラ映像のみを扱うモデルにおいて，3 次元データの推定精度を低下させる原因の 1 つとなっている．そこで，本研究は，画像内における，建物や停止した車のような静的領域と，歩行者や走行車のような動的領域の識別問題に注目する．この研究は，一般に，Motion Segmentation[3] と呼ばれる．Motion Segmentation は，観測環境における障害物の動きを追跡し，正確な 3 次元データを推定するための重要なタスクであると考えられている．そのようなタスクを解く手段の 1 つとして，近年では，CPU や GPU といった演算装置の発展に伴い，深層学習が注目されている．そして，本研究では，深層学習の中でも，イメージ情報を扱うことに特化した畳み込みニューラルネットワーク (CNN) を扱う．

CNN を用いた車載単眼カメラ映像に対する Motion Segmentation では，画像間の Optical Flow を利用した手法が主流となっている．代表的な先行手法の 1 つに，SMSnet[4] がある．この手法の主な特徴は，Motion Segmentation ネットワークへの入力データとして，連続画像だけではなく，動的オブジェクトに注目した Optical Flow が利用されることである．これにより，ネットワークが，画像間の差分情報を効果的に学習できると考えられ，KITTI[5] ベンチマークデータに対して，高い推定精度を達成している．ただし，動的オブジェクトに注目した Optical Flow の生成には，Ego-Motion の真値が必要である．さらに，Optical Flow や，その補正に利用される Depth の推定が必要なため，それらを推定するネットワークを事前に学習しなければならない．そして，各ネットワークの規模が大きいため，モデル全体の計算量が増大する．

これらの問題を踏まえ，本研究では，車載単眼カメラ映像に対する Motion Segmentation の新たな手法として，以下に示される特徴のモデルを設計する．

- 入力データとして，連続画像のみを用いる．
- 真値 Optical Flow と正解 Motion Mask のみを教師とし，小規模なネットワークにマルチタスクで学習させる
- Depth，及び Ego-Motion に関しては，明示的な学習を行わない．

ここでの Motion Mask とは，動的領域と静的領域を区別

†1 現在，大阪府立大学 大学院工学研究科
a) washimoto@sig.cs.osakafu-u.ac.jp
b) yoshioka@cs.osakafu-u.ac.jp
c) inoue@cs.osakafu-u.ac.jp

した 2 値画像である．

2. 関連研究

CNN を用いた車載単眼カメラ映像に対する Motion Segmentation は，画像間の Optical Flow を利用する手法が主流となっている．本節では，まず，Motion Segmentation に，Optical Flow が利用される背景について説明する．

Optical Flow は，2 枚の画像の各ピクセル間の 2 次元変位ベクトルである．そしてこれは，観測環境の変化，つまり，3 次元空間における，カメラの動きを表す Ego-Motion と，オブジェクトの動きを表す Scene Flow に起因する．観測環境の変化と Optical Flow の関係性を図 1 に示す．Ego Flow は，Ego-Motion によるピクセル間の動きを表している．図 1 で表されているように，Optical Flow は，Ego Flow と，射影された Scene flow の 2 種類の要素のベクトルに分解して考えることができる．ここで重要なことは，観測環境におけるオブジェクトが，静的な場合と動的な場合で，Optical Flow を構成する要素が異なるということである．これは，オブジェクトが静的な場合，Ego Flow のみが Optical Flow に反映されるのに対し，オブジェクトが動的な場合，Ego Flow と射影された Scene Flow の合成ベクトルが，Optical Flow に反映されるからである．特に，車載カメラ映像の場合，Ego-Motion は直線的であることが多いため，画像内における，建物や停止した車のような静的領域と，歩行者や走行車のような動的領域では，異なる Optical Flow の傾向が表れやすい．そのため，静的領域と動的領域を識別する Motion Segmentation タスクにおいて，画像間の差分情報を含む Optical Flow を考慮することは，効果的であると考えられている．本研究では，Motion Segmentation を，Optical Flow とマルチタスク学習させる手法を提案する．これにより，それぞれの学習に効果的な相互作用をもたらすことが期待できる．

次に，Motion Segmentation に関する具体的な先行研究について説明する．Motion Segmentation には，動的領域についてラベル付けされた Motion Mask を利用する教師有り学習と，それを利用しない教師無し学習の 2 つのアプローチがある．

まず，教師有り学習のアプローチについては，先行研究 [4], [6], [7] が，代表的な手法として挙げられる．これらの手法の特徴は，Motion Segmentation ネットワークへの入力データとして，連続画像だけではなく，画像間の Optical Flow が利用されることである．これにより，ネットワークが画像間の差分情報を効果的に学習できると考えられている．ただし，入力データに必要な Optical Flow について，手法 [6] では，真値を利用し，手法 [4], [7] では，代表的な Optical Flow 推定モデルである，FlowNet[8] や FlowNet2[9] により生成している．さらに，[4] は，学習済みの DispNet[10] により推定した Depth と，Ego-Motion

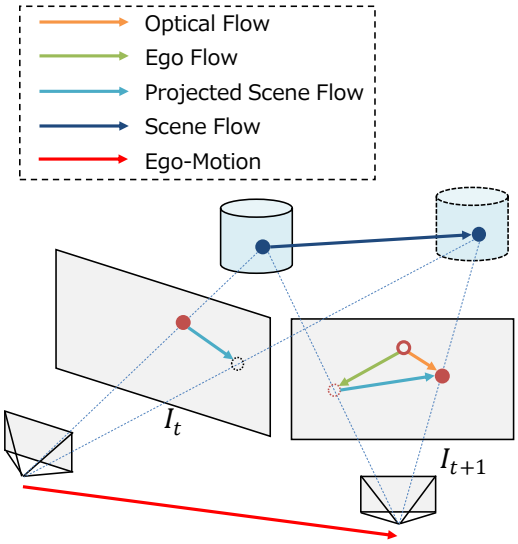

図 1 観測環境の変化と Optical Flow の関係性

の真値を利用することで，入力前の Optical Flow に対して，補正処理を加えている．具体的には，画像間の Ego Motion の影響を軽減させ，よりオブジェクトの動きを反映させた Optical Flow を新たに生成する．そのため，手法 [4], [7] では，Motion Segmentation ネットワークとは別に，複数のネットワークを事前に学習する必要がある．そして，各ネットワークの規模が大きいため，モデル全体の計算量が増大する．

一方，教師無し学習のアプローチによる Motion Segmentation 手法について，最近，注目を浴びているものに，Ranjan らの手法 [11] がある．この手法は，連続画像の入力に対して，Depth と Ego-Motion を学習する過程で生成する Ego Flow と，推定した Optical Flow の差分に注目し，Motion Segmentation を推定する手法となっている．また，Motion Segmentation だけではなく，Depth, Ego-Motion, 及び Optical Flow の全ての学習について，教師ラベルを必要としないという特徴がある．ただし，Depth と Ego-Motion を学習するためのモデル [1], [12] が，観測環境内のオブジェクトの動きに完全に対応できる手法ではないため，動的領域の Depth 推定精度が低下してしまう場合がある．そのため，動的領域の Depth の推定精度が，モデル全体の学習に大きく影響を与える手法 [11] は，Motion Segmentaion については，一定の推定精度に留まっている．本研究では，以上の 2 つのアプローチの問題を踏まえ，連続画像のみを入力データとし，Motion Segmentation を，Optical Flow のみとマルチタスク学習させる手法を提案する．さらに，小規模なネットワークモデルを目指し，Depth，及び Ego-Motion に関して，明示的な学習を行わない手法となっている．

3. 提案手法

本研究では，車載単眼カメラ映像に対する Motion Seg-

mentation の新たな手法として，連続画像のみを入力データとし，Motion Segmentation を，Optical Flow とのマルチタスクのみで学習させる手法を提案する．Optical Flow の学習については，先行研究の PWC-Net[13] モデルを参照し，提案手法に取り入れる．提案手法のモデルは，大きく分けて，(1) 特徴ピラミッドの生成，(2) Optical Flow の推定，(3) Motion Segmentation の推定，の 3 要素で構成される．まず，連続する 2 枚の入力フレームそれぞれに対して，畳み込み処理を繰り返す．このとき，中間層の出力を積み重ねることで，各入力フレームについて，複数の特徴マップから成る特徴ピラミッドを生成する．そして，これらの特徴ピラミッドを利用し，最終的に，ネットワークから Optical Flow と Motion Segmentation が推定される．本節では，モデルを構成する 3 要素の具体的な処理，及びネットワークの学習に用いられる損失関数について説明する．

3.1 特徴ピラミッドの生成

動画中の連続的な変化を認識するようなタスクでは，複数単位でフレームを扱う必要がある．本研究では，動画中のあるフレームに注目したとき，その直後のフレームを参照する．例えば，t 番目のフレーム I_t を基準フレームとしたとき，直後のフレーム I_{t+1} を参照フレームとして扱う．図 2 は，入力した基準フレーム I_t と参照フレーム I_{t+1} の，それぞれに対して，畳み込み処理 (conv) を繰り返し行う流れを示している．これらの処理は，特徴抽出器のような役割を果たしている．各畳み込み処理では，具体的に，2 層の convolution と 1 層の maxpooling を 1 組としたダウンサンプリング処理を行っている．入力フレームを最下層として，中間層の出力を，下から出力順に積み重ねる．すると，それぞれの入力フレームについて，複数の特徴マップから成る特徴ピラミッドを生成することができる．ピラミッドを構成する特徴マップは，レイヤが高くなるにつれて，縦横のスケールに関しては，1/2 づつ小さくなり，また，チャネル数に関しては，$[3, 16, 32, 64, 96, 128, 196]$ の順に変化する．P_t^l, P_{t+1}^l は，それぞれ，ある l 番目のレイヤにおける，基準フレームと参照フレームの特徴マップを表す．

3.2 Optical Flow の推定

提案手法では，Optical Flow の推定について，PWC-Net[13] のモデルを参照する．PWC-Net では，最も上のレイヤの特徴マップの組から順に，Optical Flow の推定を繰り返す．複数のレイヤにかけて，推定が行われるのは，レイヤごとに，特徴マップに影響を与える入力フレームの受容野が異なり，入力フレーム間の様々なスケールの差分に対応するためであると考えられる．これにより，PWC-Net は，様々なレイヤの特徴マップの組から，入力フレーム間

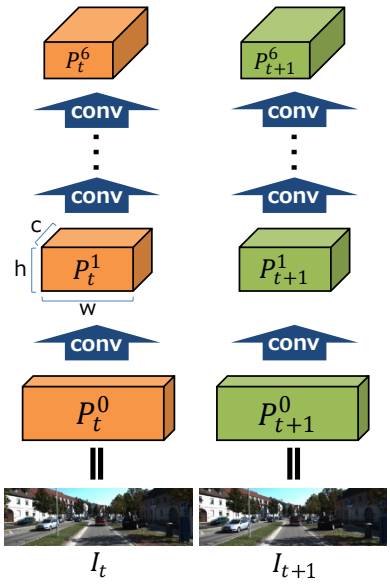

図 2 特徴ピラミッドの生成

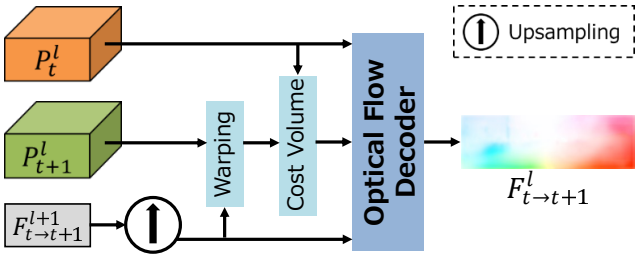

図 3 Optical Flow の推定

の大域的な変化と局所的な変化の双方を効率良く学習し，軽量でありながらも，高精度な Optical Flow の推定を実現している．本研究では，$l = 2, \ldots, 6$ の範囲でレイヤを走査する．図 3 は，走査中のある l 番目のレイヤにおいて，特徴マップの組から，Optical Flow を表す $F_{t\to t+1}^{l+1}$ を推定する流れを示している．ここで行われる処理は，(1) Warping，(2) Cost Volume，(3) Optical Flow Decoder の，3 つのモジュールで構成される．本節では，各モジュールの機能について説明する．

3.2.1 Warping モジュール

一般的に，Optical Flow は，2 つのマップ間のピクセルごとの変位ベクトルであるため，Optical Flow を利用して Warping 処理を行うことで，マップ間の差分を補間することができる．Optical Flow が高精度であればあるほど，一方のマップから，もう一方のマップの密な再構成が可能であると考えられる．このモジュールでは，参照フレームの特徴マップから，基準フレームの特徴マップを再構成する処理を行う．このとき，再構成には，1 層上のレイヤで推定した Optical Flow に対して，バイリニア補間によるアップサンプリングを行ったものが利用される．warping 処理の計算式は，以下の式 (1) のように表される．

$$\mathbf{p}_{\mathrm{warp}}^{l}(\mathbf{x}) = \mathbf{p}_{t+1}^{l}\left(\mathbf{x} + \mathrm{up}_2(\mathbf{f}^{l+1})(\mathbf{x})\right) \qquad (1)$$

l はレイヤの位置を，$\mathbf{x}$ はピクセルのインデックスを表す．$\mathbf{p}_{\text{warp}}$ と $\mathbf{p}_{t+1}$ は，それぞれ再構成された基準フレームと参照フレームの特徴マップを表す．$\text{up}_2(\mathbf{f})$ は，2 倍のスケールにアップサンプリングされた Optical Flow を表す．

3.2.2 Cost Volume モジュール

このモジュールでは，Warping 処理により再構成された基準フレームの特徴マップと，元の基準フレームの特徴マップに関して，2 つのマップ間の Cost Volume を計算する．ここでの Cost Volume は，先行研究 [8], [14] に基づき，マップ間の相関を表す．計算式は，以下の式 (2) のように定義される．

$$\text{cv}^l\left(\mathbf{x}_t, \mathbf{x}_{t+1}\right) = \frac{1}{N}\left(\mathbf{p}^l_{\text{warp}}\left(\mathbf{x}_t\right)\right)^\top \mathbf{p}^l_{t+1}\left(\mathbf{x}_{t+1}\right) \qquad (2)$$

$\top$ は転置演算子を，N は列ベクトル $\mathbf{p}(\mathbf{x})$ の長さを表す．

3.2.3 Optical Flow Decoder

このモジュールでは，ネットワークを用いて，走査中のレイヤにおける特徴マップ間の Optical Flow を推定する．このネットワークを，Optical Flow Decoder とする．Optical Flow Decoder の特徴としては，まず，1 層上のレイヤの Optical Flow，基準フレームの特徴マップ，そして，Cost Volume を入力データとすることが挙げられる．これにより，それまでのレイヤでは認識できなかった，より部分的な入力フレーム間の差分について，学習できると考えられる．また，ネットワーク構造は，全ての畳み込み層が他の層と直接的に結合する DenseNet[15] を利用する．本研究では，先行研究 [13] を参照し，7 層の特徴ピラミッドにおける上位 5 レイヤ分までの特徴マップの組に対して，Optical Flow を推定する．そのため，ネットワークから得られる最大スケールの出力は，元の入力フレームの 1/4 スケールの Optical Flow となる．入力フレームに対する Optical Flow は，最終的に，1/4 スケールの Optical Flow をバイリニア補間でアップサンプリングすることで得られる．

3.3 Motion Segmentation の推定

CNN による画像生成タスクでは，しばしば，U-Net[16] をベースとした構造のネットワークが利用される．U-Net は，Encoder-Decoder 型のネットワーク構造であり，特徴としては，Encoder から Decoder にかけて，スキップ接続があることが挙げられる．これにより，U-Net では，学習時の勾配消失が抑えられていると考えられており，様々な Segmentation タスクにおいて成果をあげている．提案手法では，Motion Segmentation を推定するネットワークを，U-Net ベースの構造の Decoder で設計している．このネットワークを，Motion Segmentation Decoder とする．図 4 は，特徴マップの組から，入力基準フレームの Motion Segmentation を推定するまでの流れを示している．まず，

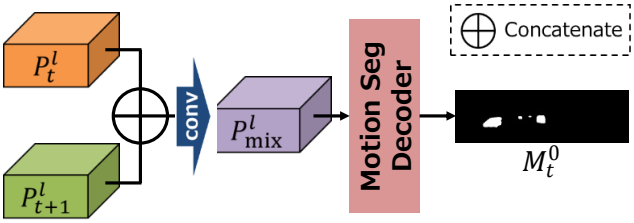

図 4 Motion Segmentation の推定

各レイヤの特徴マップの組に対して，チャネルの次元で結合した後，畳み込み処理を行う．こうすることで，全てのレイヤにおいて，混合特徴マップが得られる．次に，最も上のレイヤである 7 層目の混合特徴マップを Motion Segmentation Decoder へ入力し，畳み込み処理によるアップサンプリングを繰り返す．このとき，他のレイヤの混合特徴マップを，段階的にスキップ接続することで，U-Net ベースのネットワーク構造を実現する．最終層で出力される Motion Segmentation は，各ピクセルが，動的領域らしさを表す [0,1] の確率的な値をとる．最終的には，推定した Motion Segmentation について，0.5 以上の値のピクセルを動的領域とし，入力基準フレームの Motion Mask を表す M^0_t を生成する．

3.4 損失関数

Optical Flow に関する損失関数 $\mathcal{L}_{\text{flow}}$ を，以下の式 (3) に表す．

$$\mathcal{L}_{\text{flow}} = \sum_{l=2}^{6} \alpha_l \sum_{\mathbf{x}} \left| \mathbf{f}^l(\mathbf{x}) - \hat{\mathbf{f}}^l(\mathbf{x}) \right| \qquad (3)$$

Optical Flow について，各レイヤごとに，推定値 $\hat{\mathbf{f}}^l(\mathbf{x})$ と真値 $\mathbf{f}^l(\mathbf{x})$ の絶対誤差を計算し，それらの和を算出している．本研究では，$l = 2, \ldots, 6$ の範囲でレイヤを走査し，α_l は各レイヤにおける推定誤差の重みを表すハイパーパラメータである．

Motion Segmentation に関する損失関数 $\mathcal{L}_{\text{seg}}$ を，以下の式 (4) に表す．

$$\begin{aligned}\mathcal{L}_{\text{seg}} = \sum_{\mathbf{x}} &\left(-\frac{1}{2}\mathbf{m}_t(\mathbf{x}) \log \hat{\mathbf{m}}_t(\mathbf{x}) \right) \\ &+ \left(1 - DSC(M_t, \hat{M}_t)\right)\end{aligned} \qquad (4)$$

$$DSC(M_t, \hat{M}_t) = \frac{2|M_t \cap \hat{M}_t|}{|M_t| + |\hat{M}_t|} \qquad (5)$$

式 (4) の第 1 項では，入力基準フレームの Motion Segmentation について，[0,1] の確率的な値をとる推定値 $\hat{\mathbf{m}}_t(\mathbf{x})$ と，真値 $\mathbf{m}_t(\mathbf{x})$ 間のバイナリクロスエントロピーを計算する．また，バイナリクロスエントロピーに加え，第 2 項のような Dice 係数に関する項を設計する．Dice 係数は，集合同士の類似度を表現する係数の一つで，2 つの集合の平均要素数と共通要素数の割合を算出する．本研究では，推定した Motion Segmentation のピクセル値 0.5 を境界とし

て生成された，Motion Mask を表す $\hat{M}_t$ と，正解 Motion Mask を表す M_t 間の Dice 係数を計算する．具体的な計算方法を式 (5) に表す．Dice 係数が大きいほど 2 つの集合の類似度は高く，値は 1 に近づく．そのため，式 (4) の第 2 項のような式を設計することで，$\hat{M}_t$ と M_t の類似度が高くなるように学習を進めることができる．

最終的な損失関数 $\mathcal{L}_{\text{final}}$ を，以下の式 (6) に表す．

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{flow}} + \lambda \mathcal{L}_{\text{seg}} \tag{6}$$

λ は，式 (3) で計算した Optical Flow の推定誤差と，式 (4) で計算した Motion Segmentation の推定誤差の比重を表すハイパーパラメータである．この損失関数を最小化する方向に学習を進め，ネットワークの重みを最適化する．

4. 実験

本研究では，車載単眼カメラ映像に関する 2 種類のデータセットを用いた 2 つの実験を行う．1 つ目の実験は，CG ベースのデータを用いて，学習，及びテストを行い，Motion Mask の推定結果を評価する．2 つ目の実験は，1 つ目の実験の学習モデルに対して，実環境データによる fine-tuning を行う．そして，テストデータに対する Motion Mask の推定結果を評価する．これらの実験により，提案手法のモデルが，画像間の差分特徴を効果的に学習し，連続画像のみの入力データから，Optical Flow と Motion Segmentation のマルチタスク推定が可能であるかについて検証する．

4.1 実験条件

まず，損失関数の式 (3) のハイパーパラメータについて説明する．本研究の実験では，上位 5 レイヤ分の推定誤差を考慮する．このとき，各レイヤにおける推定誤差の重み α_l は，先行研究 [13] に倣い，最も上のレイヤから順に，$\alpha_6 = 0.32, \alpha_5 = 0.08, \alpha_4 = 0.02, \alpha_3 = 0.01, \alpha_2 = 0.005$ と設定する．損失関数の式 (6) のハイパーパラメータついては実験的に決め，$\lambda = 2$ と設定する．最適化アルゴリズムは，Adam[17] を使用する．学習率の初期値を 0.0001，重み減衰を 0.0004 と設定し，バッチサイズ 8 のミニバッチ学習を行う．

4.2 データセットと実験方法

車載カメラ映像に関する，有名なデータセットの 1 つに，KITII[5] がある．各フレームについて，RGB 画像，Optical Flow，Segmentation のような 2 次元データだけでなく，Depth や Ego-Motion のような 3 次元データも含まれており，自動運転関連の研究によく用いられるデータセットの 1 つである．先行研究 [4] では，KITTI の一部のデータに対して，走行車領域の正解 Motion Mask が作成され，公開されている．本研究では，これらのデータについて，利用可能なものを編集し，実環境シーンを扱う新たなデータセットとして用いる．以降，このデータセットを KITTI-MS とする．学習データには，2 枚 1032 組の連続画像と，それに対応する真値 Optical Flow と正解 Motion Mask が含まれる．また，テストデータには，先行研究 [4] の評価用データと同じものを扱い，2 枚 195 組の連続画像と，それに対応する走行車領域の正解 Motion Mask が含まれる．KITTI を模したデータセットとして，最近公開されたものに，vKIITI2.0 がある．vKITTI2.0 は，ゲームエンジンにより作成された CG ベースのデータセットであり，画像サイズ，及びカメラ内部パラメータ等に関して，オリジナルの KITTI と同じ設定になっている．KITTI，及び vKITTI2.0 の画像サイズは 375 × 1242 である．本研究では，vKITTI2.0 に含まれるデータのうち，2 枚 2121 組の連続画像と，それに対応する Optical Flow，そして，新たに作成した走行車領域の正解 Motion Mask を加えたものを，新たなデータセットとして用いる．以降，このデータセットを vKITTI-MS とする．

1 つ目の実験では，vKITTI-MS の 1908 組のデータを学習データとして利用する．このとき，学習データを 320 × 896 にクロッピングし，さらに，回転や拡大を行うことでデータ拡張を行う．残りの 213 組のデータをテストデータとして利用し，Motion Mask の推定結果について，定性評価，及び定量評価を行う．

2 つ目の実験では，1 つ目の実験の学習モデルに対して，KITTI-MS の学習データによる fine-tuning を行う．このとき，1 つ目の実験と同様に，学習データのデータ拡張を行う．そして，195 組のテストデータに対する Motion Mask の推定結果について，定性評価を行い，従来手法 [4] の実験結果と比較する．その後，定量評価を行う．

4.3 定性評価

本研究の 2 つの実験について，推定した Motion Mask の定性評価を行う．図 5 と図 6 に，各実験の推定結果の例を示す．まず，図 5 を確認する．1 例目と 2 例目について，推定 Motion Mask は，正解 Motion Mask と概ね一致していることが確認できる．そして，1 例目について，特に，動的領域と推定した左側車線の向かってくる黒い車に注目したとき，周辺の静的領域と明らかに異なる Optical Flow の傾向が表れている．一方，2 例目について，特に，静的領域と推定した左側に駐車する車に注目したとき，周辺の静的領域と Optical Flow の傾向が類似している．これらの結果から，提案モデルが画像間の差分情報，すなわち Optical Flow と，動的領域らしさの因果関係を学習し，適切に Motion Segmentaion を推定していることが考えられる．また，3 例目から，横方向の動きにも対応していることが確認できる．ただし，右端の車に注目したとき，オクルージョン領域については，動的領域の見落としが発生している．理由としては，オクルージョン領域では，画像間

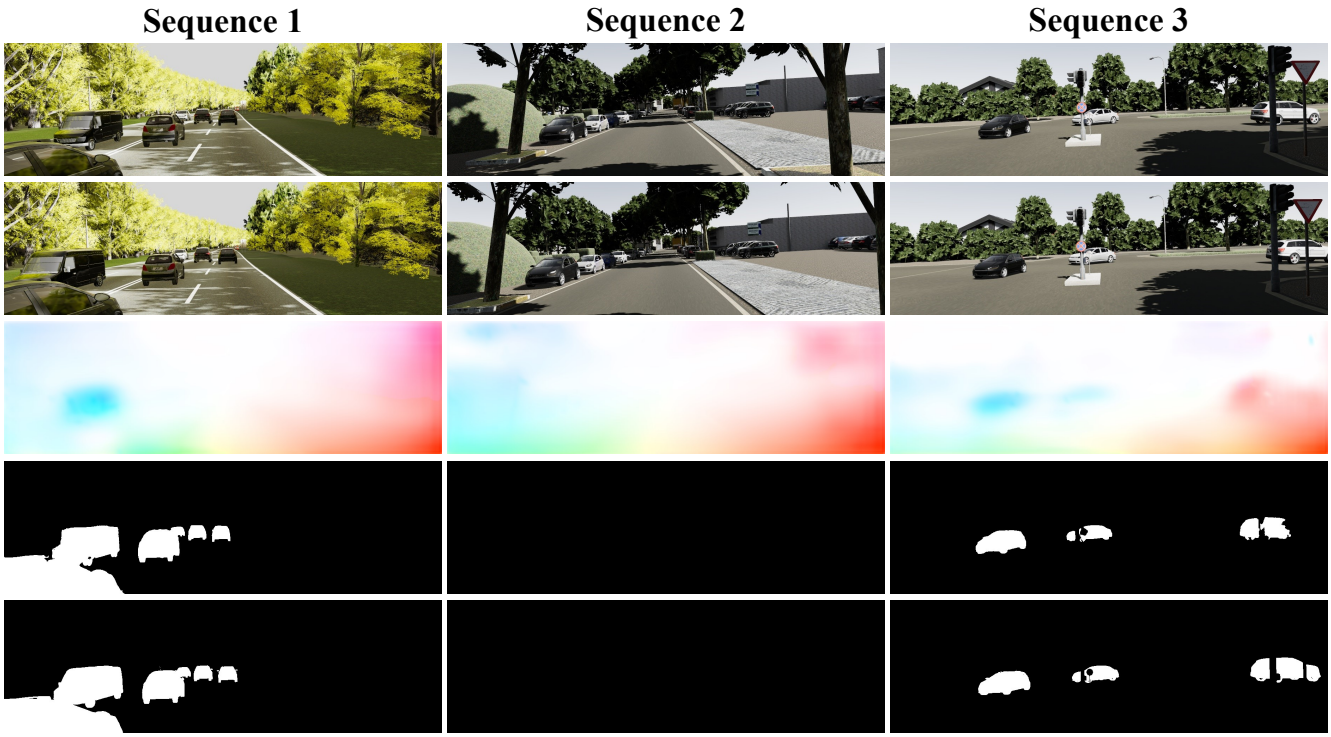

図 5 vKITTI-MS のテストデータに対する推定例. 上から順に，基準フレーム，参照フレーム，可視化した推定 Optical Flow，推定 Motion Mask，正解 Motion Mask を示す.

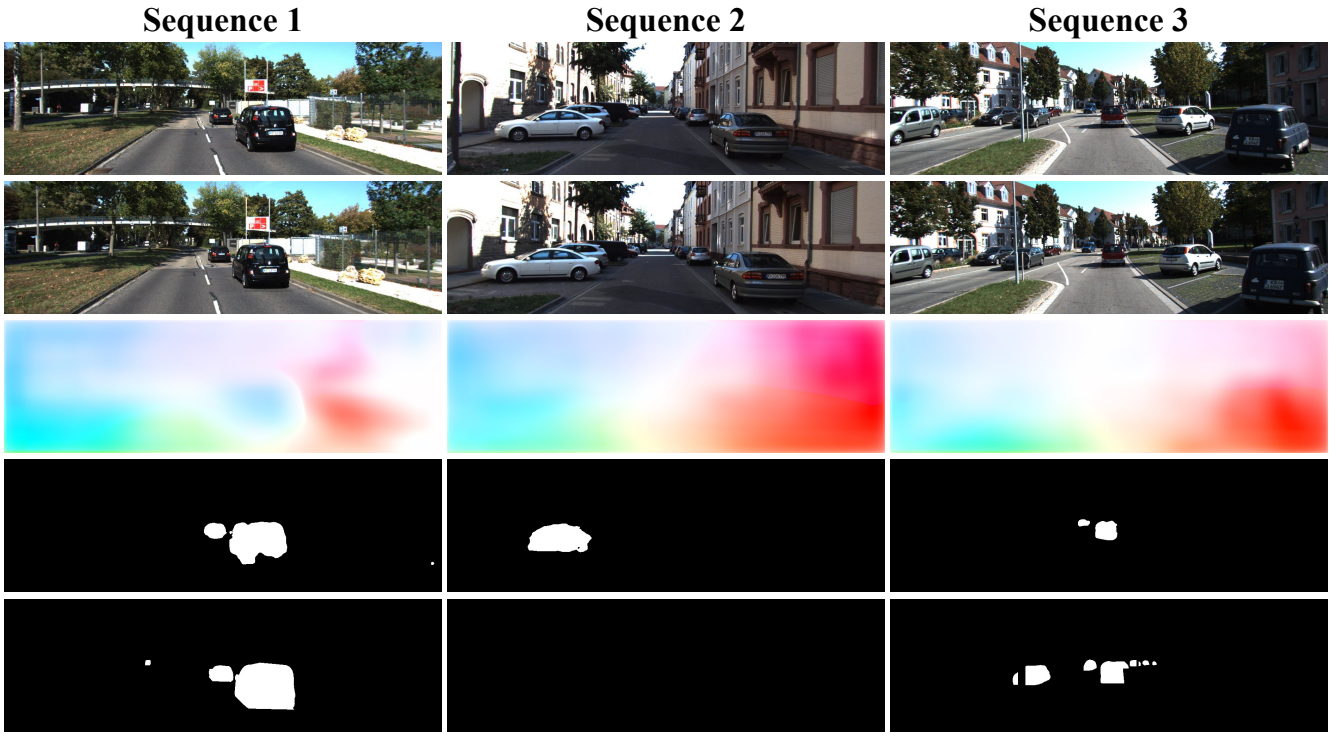

図 6 KITTI-MS のテストデータに対する推定例 . 上から順に，基準フレーム，参照フレーム，可視化した推定 Optical Flow，推定 Motion Mask，正解 Motion Mask を示す.

の対応関係が曖昧となるため，Optical Flow の推定が難しく，動的領域らしさの判定に影響を及ぼしてしまうことが考えられる．次に，図 6 を確認する．1 例目について，推定 Motion Mask は，正解 Motion Mask と概ね一致していることが確認できる．特に，動的領域と推定した前方を走行する黒い車に注目したとき，周辺の静的領域と明らかに異なる Optical Flow の傾向が表れていることが確認できる．しかし，2 例目ついて，特に，左側に駐車している車に注目したとき，動的領域と誤検出してしまっている．そして，周辺領域と異なる Optical Flow の傾向は見られない.

表 1　Moton Segmentation の定量評価

Method	Training data	Test data	IOU[%]
Ours	vKITTI-MS	vKITTI-MS	**78.28**
SMSnet [4]	KITTI-Motion	KITTI-MS	**70.32**
Ours	vKITTI-MS + KITTI-MS	KITTI-MS	51.50

表 2　Motion Segmentation の推定時間の比較

Method	GPU	TFLOPS	Time[ms]
SMSnet [4]	NVIDIA TITAN X	11.0	313
Ours	NVIDIA GeForce GTX 1080 Ti	11.3	**27.5**

理由としては，左側で向かってくるように見える車は，実際に走行しているというデータが多く，学習モデルが，車を認識した時点で，移動領域と推定するといった過学習を行っている可能性が考えられる．ただし，テストデータ全体では，このような誤検出は多く見られなかった．最後に，3 例目について，左側の車線の遠方から向かってくる車に注目したとき，動的領域の見落としが発生している．理由としては，この領域の推定 Optical Flow を確認したとき，周辺の静的領域のそれと傾向が類似していることから，静的領域である可能性が高いと推定してしまったことが考えられる．ただし，テストデータ全体では，このような見落としはほとんど見られなかった．

4.4　定量評価

本研究の 2 つの実験ついて，推定した Motion Mask の定量評価を，表 1 にまとめて示す．1 つ目の実験の定量評価を 1 段目に示し，2 つ目の実験の定量評価を，従来手法である SMSnet[4] と比較し，2 段目以降に示す．定量評価には，Intersection over Union (IoU) を用いる．IoU とは，ある 2 つの領域が存在するとき，領域の共通部分を領域の和集合で割った値であり，本研究で算出する IoU は，動的領域における推定領域と正解領域の重なり率を表す．提案手法の学習モデルは，1 つ目の実験結果から，vKITTI-MS のテストデータに対して，高精度な Motion Segmenatation の推定を行っていると考えられる．IoU は非常に厳密な評価手法の 1 つとされており，互いの領域同士が僅かでもズレると値が大きく下がる．IOU の値を 80 %近く出していることは，推定結果が概ね正解していることを表している．次に，2 つ目の実験結果から，KITTI-MS のテストデータに対する推定精度について，SMSnet が提案手法を上回っていることが分かる．SMSnet の学習データが一部公開されておらず，学習データを揃えることができなかったため，単純な比較は難しいが，提案手法を上回った大きな理由の 1 つとしては，SMSnet が，Motion Segmentation ネットワークへの入力データとして，連続画像だけではなく，動的オブジェクトに注目した Optical Flow を利用していることが考えられる．また，この Optical Flow の生成に，Ego Motion の真値を扱っている．ただし，Optical Flow や，その補正に利用される Depth の推定が必要なため，それらを推定するネットワークを事前に学習しなければならない．そして，各ネットワークの規模が大きいため，モデル全体の計算量が増大する．表 2 に，従来手法 [4] と提案手法の推定時間を示す．TFLOPS は，GPU 等の処理性能を表す単位の 1 つで、浮動小数点演算を 1 秒間に 1 兆回行うことを表す単位のことである。提案手法では，同程度の性能の GPU を用いながらも，大幅に推定時間が削減されていることが分かる．提案手法では，Optical Flow の学習については，軽量モデルである PWC-Net[13] を参照し，別のネットワークをによる，Depth や Ego-Motion の明示的な学習を行っていないため，モデル全体の計算量は小さく抑えられている．

以上，本研究の 2 つの実験について，Motion Mask の推定結果の評価から，提案手法は，Depth，及び Ego-Motion に関して，真値を扱うことや明示的な学習を行わなくとも，比較的小規模なモデルでありながら，画像間の差分情報を効果的に学習し，vKITTT-MS のテストデータに対して，高い推定精度を達成したと考えられる．また，従来手法の SMSnet から推定時間を大幅に削減し，KITTI-MS のテストデータに対して，推定精度では SMSnet に及ばないながらも，一定の推定精度を達成したと考えられる．

5.　おわりに

本研究は，車載単眼カメラ映像から 3 次元空間を認知する研究の第一歩として，画像内における，建物や停止した車のような静的領域と，特に，走行車のような動的領域の識別問題に注目した．そして，連続画像のみを入力データとし，Motion Segmentation を，Optical Flow のみとマルチタスク学習させる手法を提案した．提案手法は，Depth，及び Ego-Motion に関して，真値を扱うことや明示的な学習を行わなくとも，比較的小規模なモデルでありながら，画像間の差分特徴を効果的に学習し，2 種類のデータセットに対して，一定の推定精度を達成した．今後の課題としては，画像間の変化が小さい領域や，オクルージョン領域へ対応が考えられる．本研究の提案手法では，参照フレームの枚数を 1 枚に設定しているが，参照フレームの枚数を増やすことで，これらの領域への認識性が高まると考え

られる．また，本研究で扱ったデータセットは，一般道路上の観測データであり，さらに，動的オブジェクトが走行車のみに限定されていた．しかし，実世界では，より特殊な環境での観測や，動的オブジェクトとして，人や自転車などが多く含まれることが想定される．そのため，自動運転の実用化に向けて，より大規模な車載カメラ映像データセットを扱うことが必要になってくる．

参考文献

[1] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6612–6619, 2017.

[2] R. Mur-Artal and J. D. Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, pp. 1255–1262, 2017.

[3] F. Arrigoni and T. Pajdla. Robust motion segmentation from pairwise matches. In *International Conference on Computer Vision (ICCV)*, pp. 671–681, 2019.

[4] J. Vertens, A. Valada, and W. Burgard. SMSnet: Semantic motion segmentation using deep convolutional neural networks. In *International Conference on Intelligent Robots and Systems (IROS)*, pp. 582–589, 2017.

[5] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3354–3361, 2012.

[6] M. Siam, H. Mahgoub, M. Zahran, S. Yogamani, M. Jagersand, and A. El-Sallab. MODNet: Motion and appearance based moving object detection network for autonomous driving. In *International Conference on Intelligent Transportation Systems (ITSC)*, pp. 2859–2864, 2018.

[7] N. Haque, D. Reddy, and K. M. Krishna. Joint semantic and motion segmentation for dynamic scenes using deep convolutional networks. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, pp. 75–85, 2017.

[8] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. v. d. Smagt, D. Cremers, and T. Brox. FlowNet: Learning optical flow with convolutional networks. In *International Conference on Computer Vision (ICCV)*, pp. 2758–2766, 2015.

[9] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1647–1655, 2017.

[10] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4040–4048, 2016.

[11] A. Ranjan, V. Jampani, L. Balles, K. Kim, D. Sun, J. Wulff, and M. J. Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12232–12241, 2019.

[12] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5667–5675, 2018.

[13] D. Sun, X. Yang, M. Liu, and J. Kautz. PWC-Net: Cnns for optical flow using pyramid, warping, and cost volume. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8934–8943, 2018.

[14] J. Xu, R. Ranftl, and V. Koltun. Accurate optical flow via direct cost volume processing. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5807–5815, 2017.

[15] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017.

[16] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

[17] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, pp. 1–15, 2015.