NAIST-IS-MT1651035

修士論文

セマンティックマップを用いた視差マップの改善

川上 蓮也

2018 年 3 月 1 日

奈良先端科学技術大学院大学
情報科学研究科 情報科学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
修士（工学）授与の要件として提出した修士論文である．

川上 蓮也

審査委員：

小笠原 司 教授	（主指導教員）
清川 清 教授	（副指導教員）
高松 淳 准教授	（副指導教官）
丁 明 助教	（副指導教員）

セマンティックマップを用いた視差マップの改善*

川上 蓮也

内容梗概

近年，自動運転の実現に向けて様々な研究開発が行われている．中でも環境認識技術は非常に重要視されており，実現のため様々なセンサが利用されている．その一つであるステレオカメラは，軽量かつ比較的安価であり，また人間の視覚同様に色と距離の両方が取得できるという特徴を持っている．そのため，現在開発されている先進運転支援システムの多くに採用されている．しかしながら，ステレオカメラにおける距離精度は環境に依存し，特に反射や透過に対して非常に弱いという問題がある．様々な環境下でのロバスト性が求められる自動運転というタスクにおいて，この問題は解決すべきものである．

そこで，本研究ではセマンティックマップを用いた視差画像の改善手法を提案する．セマンティックマップとは画像の各ピクセルにクラスラベルを割り振ったものであり，環境を理解する上で必要な情報の一つと言える．この情報を付与することで，反射や透過がある領域に対しても人間に近い距離感を取得できることが期待される．

KITTI ベンチマークを用いた評価実験の結果，視差画像のみを用いた場合やRGB 画像を用いた場合と比べて，セマンティックマップが視差画像の改善に有効であることが確認された．

キーワード

ステレオビジョン，視差マップ，セマンティックマップ，畳み込みニューラルネットワーク，KITTI データセット

*奈良先端科学技術大学院大学 情報科学研究科 情報科学専攻 修士論文, NAIST-IS-MT1651035, 2018 年 3 月 1 日.

Refinement of the Disparity Map by Using Semantic Map*

Renya Kawakami

Abstract

In recent years, numerous researches and developments have been carried out to realize automated driving. In particular, environmental recognition is an important issue, and various sensors are used to solve it. Among them, stereo cameras have the advantage of being able to measure both color and distance, at a relatively low price. Therefore, they are used in many advanced driving assistance systems (ADAS) currently being developed. However, the accuracy of the distance calculated from stereo images depends on the environment, specifically, sterero matching weaks against reflection and transmission. This problem should be solved in order to realize automated driving requiring robustness under various environments.

In this thesis, we propose a method to refine a disparity map using a semantic map. Semantic maps assign a class label to each pixel, and it' s one of the important information to understand the environment. Hence, it is expected that the semantic map could improve the disparity map even in regions including reflection and transmission. As a result of the expectation, it was confirmed that the semantic map improves a disparity map compared to using only disparity map or RGB images. We evaluated the proposed method on the KITTI benchmark data set.

Keywords:

*Master's Thesis, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT1651035, March 1, 2018.

Stereo vision, Disparity map, Semantic map, Convolutional neural network, KITTI dataset

目次

図目次

第 1 章　はじめに

1.1　研究背景

近年，自動運転車の制御や先進運転支援システム (Advanced Driver Assistance Systems，ADAS) に関する応用技術が多数提案されている [1] [2] [3]．商品化されている ADAS は，カメラ，LIDAR，レーダー，超音波などの様々なセンサから得られる情報を統合し，利用することで実現している [4]．さらに，グローバルポジショニングシステム (GPS) や車載データネットワーク，車車間通信システムなどの外部情報を統合して利用することで，効率的かつ高精度な ADAS の実現を目指している [5]．

これらのセンサの中でも特にカメラは自動車メーカの重要な差別化要因となっている [6]．カメラベースの ADAS は，様々なコンピュータビジョン技術を用いてリアルタイムに状況分析を行い，運転者に注意を促すことができる．カメラベースの利点としては，交通標識や案内板の認識，信号機の検出 [7] [8]，車線と障害物検出 [9] [10] などの多様なアプリケーションの実現に貢献していることが挙げられる．これらに加え，車間距離なども計測可能であり，それらを利用した定速走行・車間距離制御装置 (Adaptive Cruise Control，ACC) は既に商品化されている．

ADAS や自動運転車の実現において，この距離計測は重要なタスクである．本論文では，この距離計測に対するアプローチとしてステレオカメラに着目する．実際のアプリケーションでは，使用されるセンサ数とコストから商品として実現可能か考慮する必要がある．この問題に対してステレオカメラは他のセンサと比較し，視野角と解像度に応じて豊富で大量のデータを提供できるという利点がある．さらに，ステレオカメラは比較的安価に自動車へ搭載できることもあり，ステレオカメラでの距離計測は重要なタスクであると言える．このタスクには，ADAS や自動運転車といった応用先の性質上，距離精度の他に様々な環境下でのロバスト性が求められる．

1.2 関連研究

ステレオカメラにおいて距離計測を行うには，左右の画像間で対応点探索を行い，各画素における視差を正確に推定する必要がある．しかしながら，実際の環境下では鏡面領域や透過領域が非常に多く，対応点が存在しない領域の影響を受けてしまうことが多い．このような問題を解決するため，従来は一次，もしくは二次的な平滑過程を組み合わせることで視差マップの最適化が行われてきた [11]．一方でこのような低次な制約ではなく，より人間の認識に近い制約を考えた Güney らは Fig. 1.1 に示すように，自動車に対する領域分割と物体形状の事前知識を組み合わせることでより高次の制約を考え，発表当時 KITTI2012 ベンチマーク [21] で精度一位を記録している [12]．このように，物体領域や物体形状のような高次の情報を視差マップの改善に用いることの有効性は認められるが，一方でそういった情報を扱うことは難しい．Güney らの手法では，物体ごとに代表的な 3D 形状を用意する必要があるため，物体の種類を増やすことが難しい．また物体の種類が増えた際に，それに合わせる 3D モデルを選択することがまた一つ大きなタスクとなってしまうのも課題である．

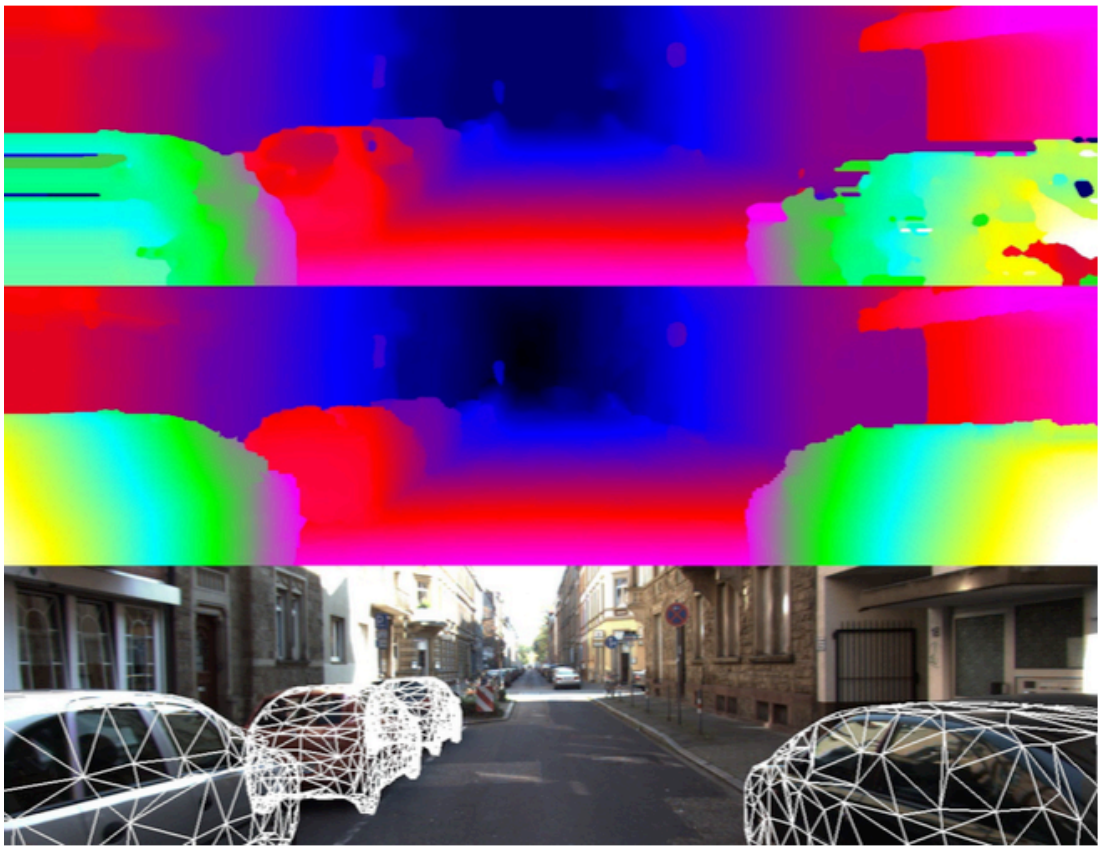


Fig. 1.1: Resolving Stereo Matching Ambiguities: Current state-of-the-art stereo methods often fail at reflecting, textureless or semi-transparent surfaces (top). By using object knowledge, we encourage disparities to agree with plausible surfaces (center). This improves results both quantitatively and qualitatively while simultaneously recovering the 3D geometry of the objects in the scene (bottom) [12].

1.3 研究目的

前述したように，物体領域や物体形状などの高次の情報は鏡面領域や透過領域における視差マップの改善に有効であると考えられるが，一方でそれをモデル的に考えると様々な物体に対応することが困難になる．そこで，本研究では物体領域をより抽象的に表現したセマンティックマップを用いることで，様々な物体に対しても適応可能な視差マップの改善手法の構築を目指す．セマンティックマップを用いる理由としては大きく分けて以下の二つであり，特に一つ目の過程は F. Güney らの手法 [12] と同様に人間の認識に近い制約となるのではないかと期待できる．

- 各画素にクラス尤度を割り当てるセマンティックマップは物体の連続性に関する情報を持っており，これが視差の連続性に対して有効である可能性が高い．
- セマンティックマップ自体が環境認識タスクと非常に相性がよく，これを使い回すことで余分な計算時間を短縮することができる．

1.4 本論文の構成

本論文の構成を簡単に示す．2 章では，セマンティックマップ用いた視差マップ改善手法を示し，3 章にてその評価を行う．最後に，4 章では本論文のまとめと今後の課題と展望について述べる．

第 2 章　提案手法

2.1　概要

本手法は Fig.2.1 に示すように初期視差マップの推定, セマンティックマップの推定，セマンティックマップを用いた視差マップの改善の 3 つに分けられる．初期視差マップ及びセマンティックマップの推定は従来研究である MC-CNN (Matching Cost CNN)[13] と PSPNet（Pyramid Scene Parsing Network）[14] を用いて実現しており，本手法ではこの二つを組み合わせるネットワークを提案する．局所的な情報である MC-CNN から得られた特徴量と大域的な情報である PSPNet から得られた特徴量を組み合わせることで，視差マップの精度向上を期待する．

2.2 節に本研究で提案するセマンティックマップを用いた視差マップの改善手法を示す．2.3 節では改善の対象となる初期視差マップの推定について，続く 2.4 節ではセマンティックマップの推定についてそれぞれ示す．

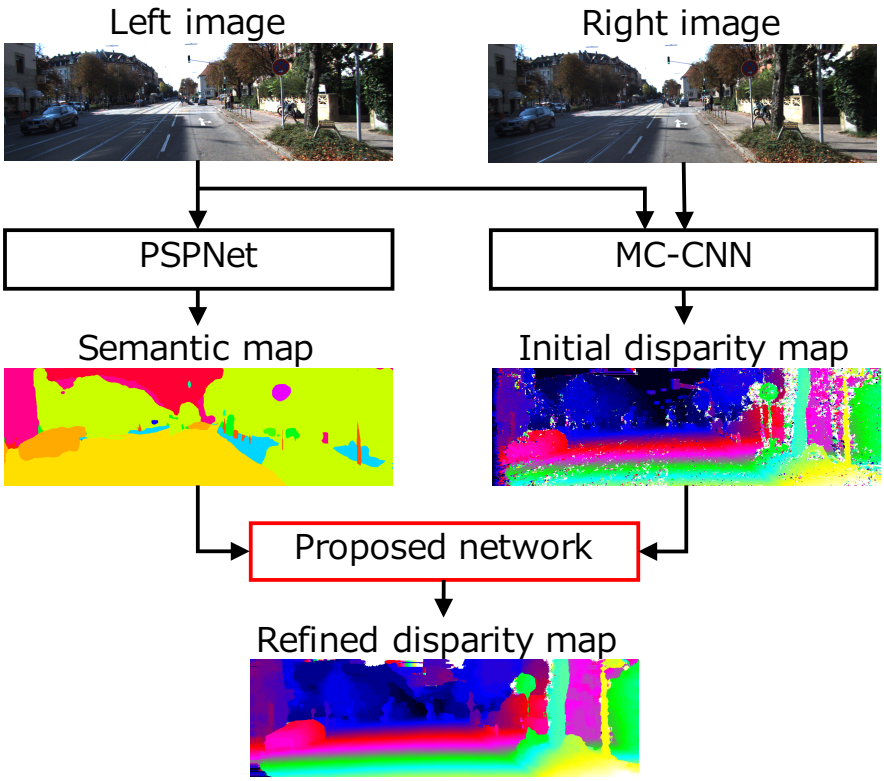


Fig. 2.1: Data flow of the proposed method

2.2 セマンティックマップを用いた視差マップの改善

2.2.1 ネットワークの定義

セマンティックマップを用いた視差マップ改善のためのネットワークを Fig. 2.2 に示す．本改善手法の位置付けはセマンティックマップを用いた局所的なフィルタリングである．入力は MC-CNN から得られた初期視差マップと PSPNet から得られたセマンティックマップの二つである．ある画素に対して，それを中心とした縦 H_{IN}，横 W_{IN} の窓で切り出した (H_{IN}, W_{IN}) をパッチとし，それぞれネットワークに入力する．

初期視差マップは CPD (Coarse Press Disparity map) モジュールにおいて，BN (Batch Nomalization) [15] で標準化した後，3D Dilated-Convlolution [16] によってスパースに $(1, 1)$ まで畳み込まれる．Dilated-Convolution は Fig. 2.3 に示すように，カーネルの間隔を空けたスパースな畳み込みを行う手法であり，計算コストを増やさずに受容野を増やすことができるといった特徴を持つ．視差方向も含めた視差マップの周辺特徴を広い範囲で畳み込むことで，局所的な誤差に対してロバストとなることを期待している．

セマンティックマップも同様に，CPS (Coarse Press Support map) モジュールにおいて BN で標準化された後，2D Dilated-Convolution によって $(1, 1)$ まで畳み込まれる．これらの特徴マップは足し合わされた後，FCN を CNN で代用した FCN モジュールを介して最終的な視差尤度を推定する．

Table. 2.1 に各モジュールの詳細を示す．CPS モジュールの入力チャンネル数は変数とし，セマンティックマップのクラス数に合わせて変更する．

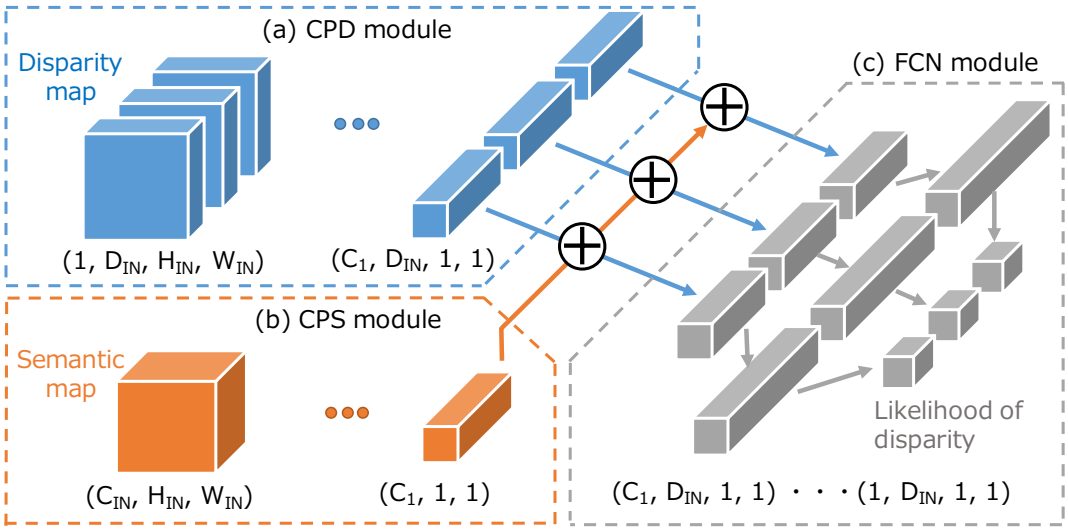
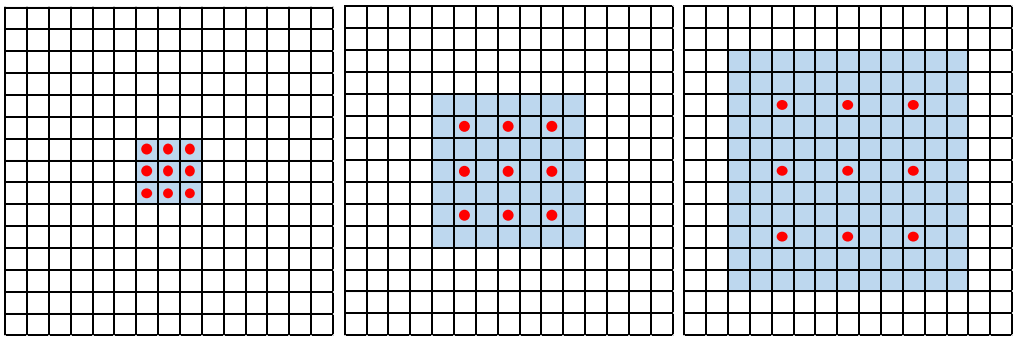


Fig. 2.2: Architecture of the proposed network. (a) CPD (Coarse Press Disparity map) module: Disparity map is pressed into $(C_1, D_{IN}, 1, 1)$ by sparse 3D convolution after BN. (b) CPS (Coarse Press Support map) module: Support map such as semantic map is pressed into $(C_1, 1, 1)$ by sparse 2D convolution after BN. $\oplus$ mean concatenating two feature map. (c) FCN (Fully Connected Network) module: BN + FCN by using 3D convolution.



(a) 1-dilated (b) 2-dilated (c) 3-dilated

Fig. 2.3: Dilated convolution

Table 2.1: Hyper parameters of layers in each modules

Module	Layer	Input channel	Output channel	Kernel size
CPD	1st	1	256	(5, 9, 9, 5-dilated)
CPS	1st	Variable	256	(9, 9, 5-dilated)
FCN	1st	256	512	(1, 1, 1)
	2nd	512	512	(1, 1, 1)
	3rd	512	1	(1, 1, 1)

2.2.2 ネットワークの学習手法

本ネットワークにおける損失関数は，クラス分類問題に一般的に用いられる Logsoftmax と NLLLoss (Negative Log Likelihood Loss) を組み合わせたものを用いた．分類するクラス数を d_{max} とし，正解クラスを y，ネットワークからの出力をそれぞれ $x_0, x_1, ..., x_{d_{max-1}}$ とすると，損失関数は以下のように定義される．

$$L(x_d, y) = -log(exp(x_y / \sum_{d=0}^{d_{max-1}} x_d))$$

ネットワークを学習させる際は，学習データセットから選んだランダムな画像に対応する視差マップとセマンティックマップに対して，バッチの数だけランダムに (H_{IN}, W_{IN}) を切り出す．それらをまとめたテンソルをネットワークに流し，その $L(x_d, y)$ が小さくなるように最適化する．

ネットワークをテストする際は，テストデータセットのある画像に対応する視差マップとセマンティックマップに対して，バッチの数だけ作為的に (H_{IN}, W_{IN}) を切り出す．それらをまとめたテンソルをネットワークに流し，その $L(x_d, y)$ を計算する．テストデータセット内の全画像に対してこれを行うことで，学習中におけるネットワークの汎化性能を評価する．

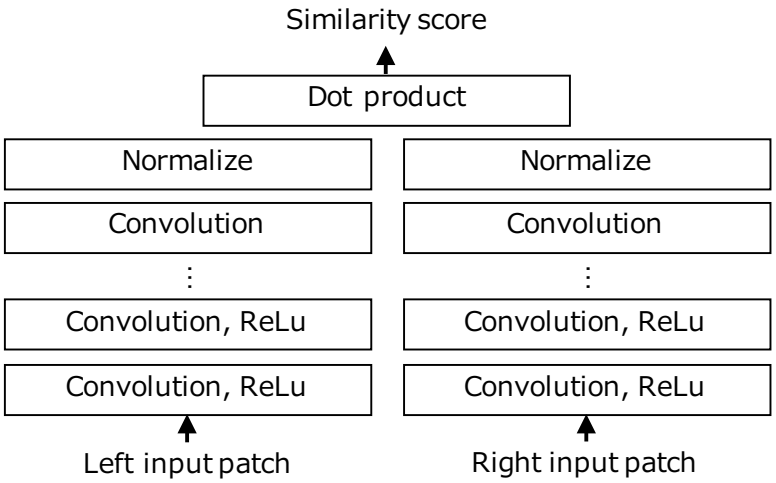
また，提案モデルを学習するための最適化手法には Adam (Adaptive moment estimation) [23] を用いており，それぞれのパラメータには Adam [23] で推奨されている値を用いた．今回使用したこれらの学習におけるパラメータを Table 2.2 に示す．

Table 2.2: Parameters of the optimization method and the batch size

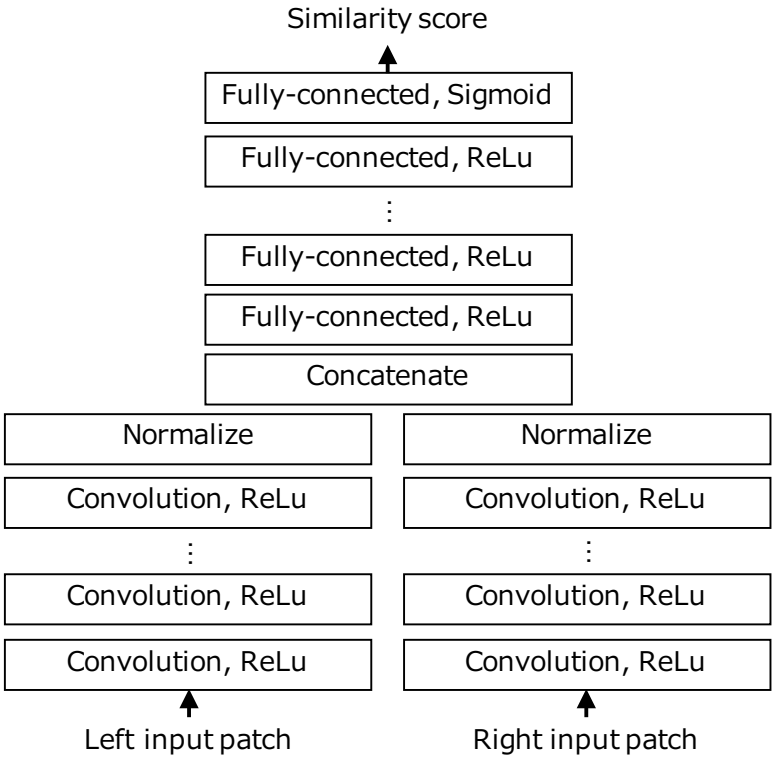
$Adam : \alpha$	0.001
$Adam : \beta_1$	0.9
$Adam : \beta_2$	0.999
$Adam : \epsilon$	10^{-8}
Batch size	256

2.3 初期視差マップの推定

初期視差マップの推定には，CNN を用いたローカルステレオマッチングの先駆けである MC-CNN [13] を用いる．Fig. 2.4 に示すように，MC-CNN ではブロックごとの特徴量を CNN を用いて抽出している．この特徴量から類似度を算出するため，特徴量の内積をとる MC-CNN-fast（Fig. 2.4(a)）と FCN を複数層かける MC-CNN-arct（Fig. 2.4(b)）が提案されている．内積をとる手法では処理が軽くなる反面，FCN を複数層かける手法に比べ精度が低下することが示されている．一般的な輝度値の SSD（Sum of Squared Difference）を類似度に使うようなブロックマッチングと比べ，CNN (Convolutional Neural Network）を用いることでより高次に特徴を掴むことができ，また平行誤差に対してロバスト性を保つことが可能である．この精度の高さはベンチマークでも示されており，Middlebury ベンチマーク [17] では発表時に精度トップであった．本研究では，初期視差マップの推定にこれら MC-CNN-fast，MC-CNN-arct をそれぞれ用いる．



(a) MC-CNN-fast



(b) MC-CNN-arct

Fig. 2.4: Difference between two types of the MC-CNN

2.4　セマンティックマップの推定

セマンティックマップの推定には，発表時に VOC2012 [20] と Cityscapes [19] のベンチマークで精度トップとなった PSPNet [14] を用いた．Fig. 2.5 に示すように，PSPNet は異なるサイズのプーリング層をかけた特徴マップを統合することで画素レベルのクラス推定を行なっている．様々なスケールでみた特徴を加味することで，従来の SegNet [18] 等のセマンティックセグメンテーションより高精度なクラス推定を実現している．

本研究では，路上で撮影された画像に対してよく観測される対象のラベルが付与された Cityscapes データセットと，一般物体を主にした VOC2012 データセットを学習データとしてそれぞれ別々に用いた．以後は，VOC2012 データセットで学習した PSPNET を PSPNET_VOC, Cityscapes データセットで学習した PSPNET を PSPNET_City と便宜上呼ぶことにする．Table 2.3 に，各 PSPNET において学習させたラベルを示す．それぞれのデータセットで学習された PSPNet で推定したセマンティックマップを視差マップの改善に用い，その差異を比較する．

KITTI2015 データセット [22] の画像に対して，それぞれ異なるデータセットで学習させたセマンティックセグメンテーションを行った結果を Fig. 2.6 に示す．この違いは学習しているクラスの差異によるものであり，Cityscapes データセットで学習されたものは人や車の他に車道や歩道，木々，建物，標識等のクラスが割り当てられている．一方で VOC2012 データセットでは車や人に限定されている．

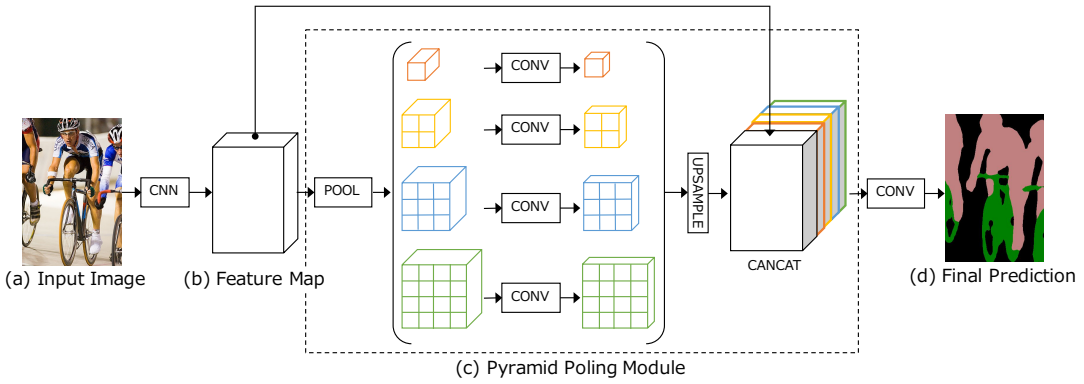


Fig. 2.5: PSPNet: Pyramid Scene Parsing Network [14]

Table 2.3: Detail of class labels

Model name	Labels
PSPNET_VOC	background, aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorcycle, person, potted plant, sheep, sofa, train, TV monitor, void
PSPNET_City	road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, rider, car, truck, bus, train, motorcycle, bicycle

(a) PSPNET_VOC

(b) PSPNET_City

Fig. 2.6: Difference of semantic maps between PSPNET_VOC and PSPNET_City

第 3 章　評価実験

3.1　評価方法

本提案ネットワークにおいてセマンティックマップを用いることの優位性を評価するため，入力を以下の 4 つのタイプに分けてそれぞれ学習させた．便宜上，視差マップのみを入力に持つネットワークを Refnet，視差マップと RGB 画像を入力に持つネットワークを RefnetImg，視差マップと VOC2012 データセットで学習済みの PSPNet から得られるセマンティックマップを入力に持つネットワークを RefnetSem_VOC，視差マップと Cityscapes データセットで学習済みの PSPNet から得られるセマンティックマップを入力に持つネットワークを RefnetSem_City と以後呼ぶことにする．改善対象となる初期視差マップとしては，MC-CNN-fast と MC-CNN-arct で生成されたものをそれぞれ用いる．評価については，KITTI ベンチマークを用いて改善前と改善後の視差マップを比較することで行う．その具体的な評価指標については 3.4 節で述べる．

- Refnet: 視差マップのみ
- RefnetImg: 視差マップ ＋ RGB 画像
- RefnetSem_VOC: 視差マップ ＋ セマンティックマップ (PSPNet_VOC)
- RefnetSem_City: 視差マップ ＋ セマンティックマップ (PSPNet_City)

3.2 ネットワークの学習

3.2.1 学習データセット

学習データセットには KITTI2012 [21] と KITTI2015 [22] をそれぞれ用いた．それぞれのデータセットには学習用とテスト用のデータが含まれているが，視差の真値が含まれているのは学習用のみである．そこで，Table 3.1 に記すように学習用のデータを学習用とテスト用に分割し，ネットワークの学習を行った．

3.3 学習結果

Fig. 3.1，3.2 に KITTI2012，KITTI2015 データセットでそれぞれ学習した時の結果を示す．学習能力と汎化能力それぞれを比較，検討するため，学習データに対する損失とテストデータに対する損失を以下に示している．また，グラフのプロット間隔は 100 ステップであり，100 ステップの平均値を代表値としてプロットしている．さらに，入力の違いによる差異をわかりやすくするため 15 ステップで移動平均をとっている．

Table 3.1: Data number used for training and testing the network

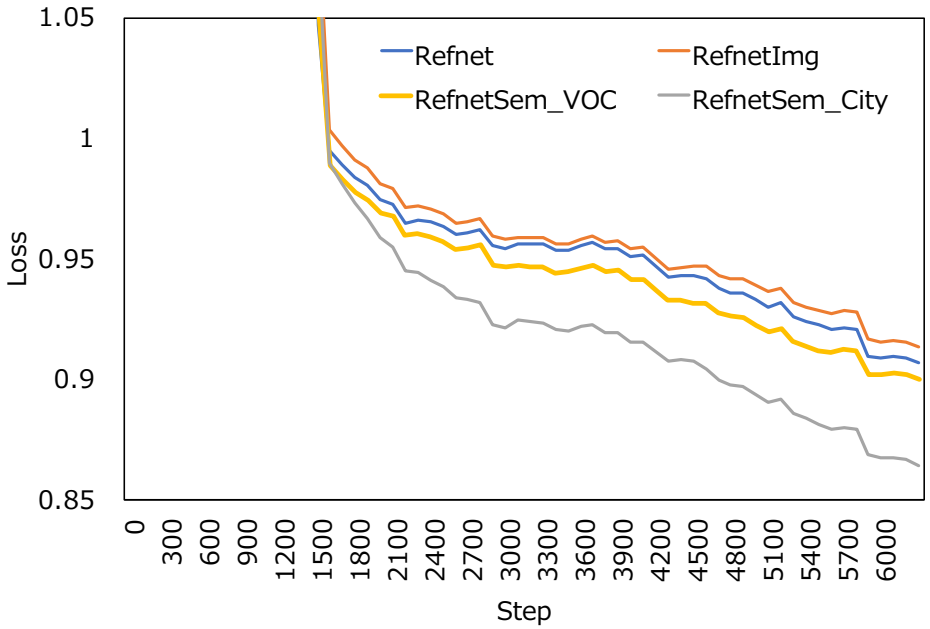
Dataset	Training images	Testing images
KITTI2012	train/0〜154	train/155〜193
KITTI2015	train/0〜159	train/160〜199

3.3.1 KITTI2012 データセットでの学習結果

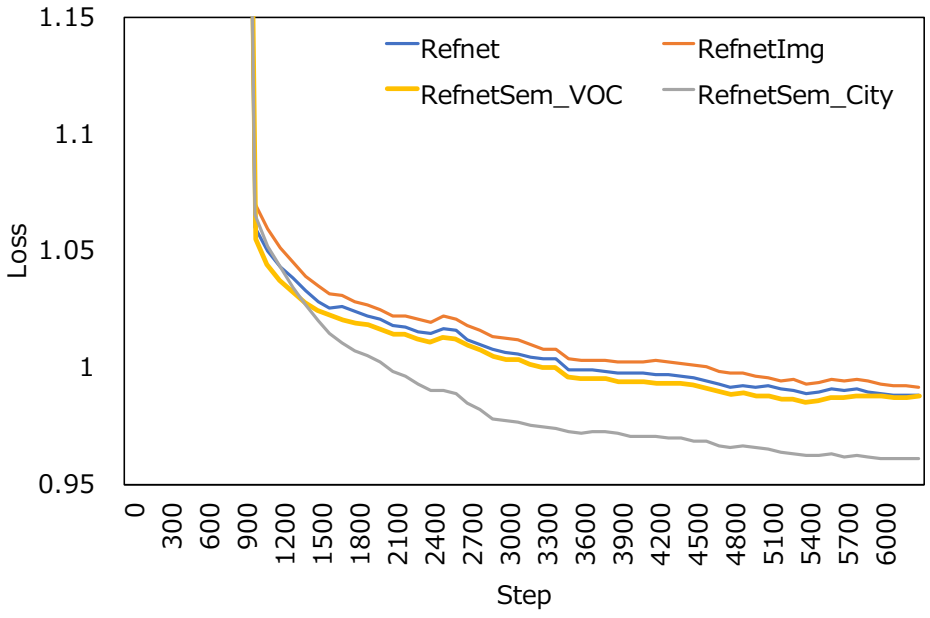
KITTI2012 データセットにおいて提案ネットワークの学習能力を比較する．Fig. 3.1(a)，3.2(a) から初期視差マップの種類によらず，RefnetSem_City が最も損失が小さくなっていることがわかる．さらに，RefnetImg より Refnet の損失が少ないことも共通して見て取れる．つまり，提案ネットワークの学習能力を向上させる上で RGB 画像を直接用いることは有効ではないが，セマンティックマップを学習させた深層ネットワークを通すことで有効になることがわかった．

次に提案ネットワークの汎化能力を比較する．Fig. 3.1(b)，3.2(b) から，先ほどと同様に RefnetSem_City が最も損失が少なくなっている．さらに，RefnetImg より Refnet の損失が少ないことも共通しているため，汎化能力に対しても RGB 画像を直接用いることは有効ではないが，深層ネットワークを通すことで有効に働くことがわかる．

最後に，Fig. 3.1，3.2 において Refnet_VOC と RefnetSem_City を比較すると，常に RefnetSem_City の方が損失が少ないことが見て取れる．つまり，学習能力や汎化能力を向上させるにはセマンティックマップの精度も重要であることがわかる．

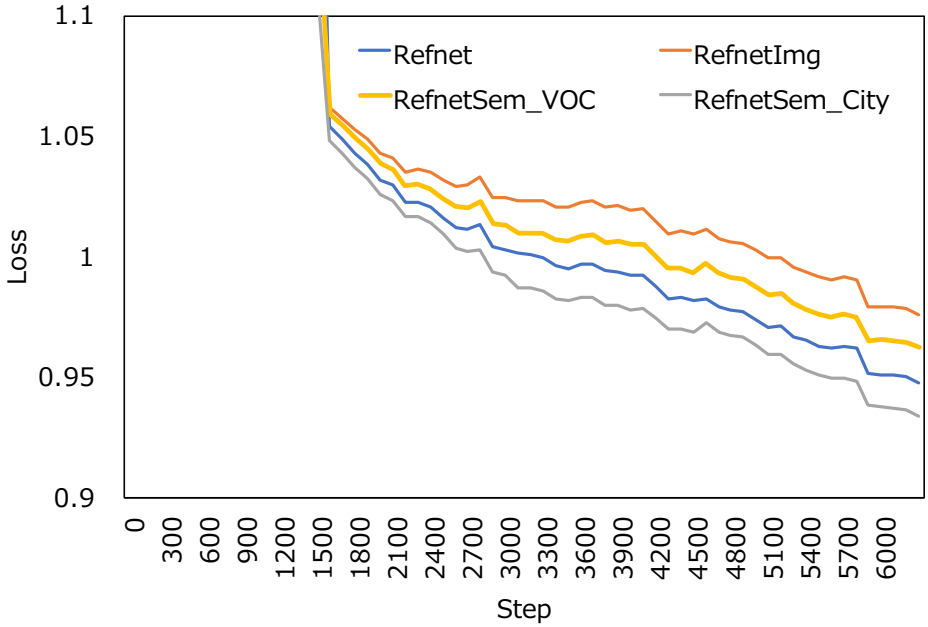


(a) Training

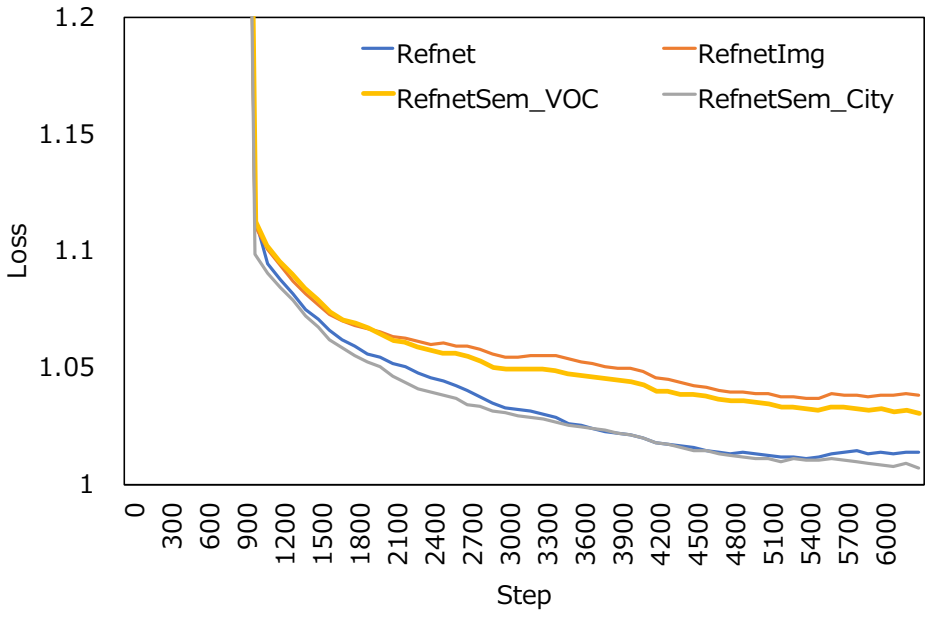


(b) Testing

Fig. 3.1: Loss during training model with the KITTI2012 dataset. Initial disparity map: MC-CNN-fast



(a) Training



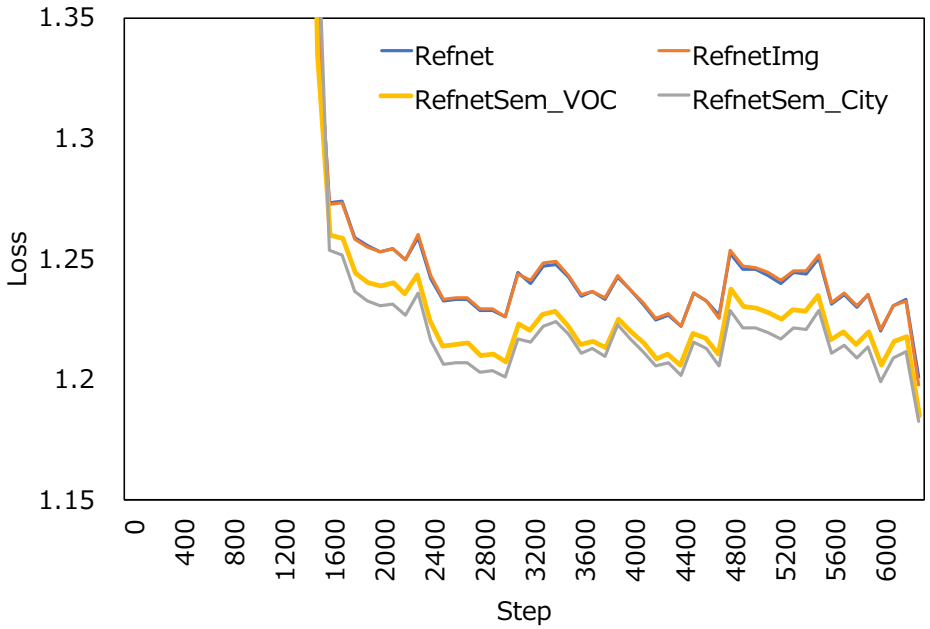
(b) Testing

Fig. 3.2: Loss during training model with the KITTI2012 dataset. Initial disparity map: MC-CNN-arct

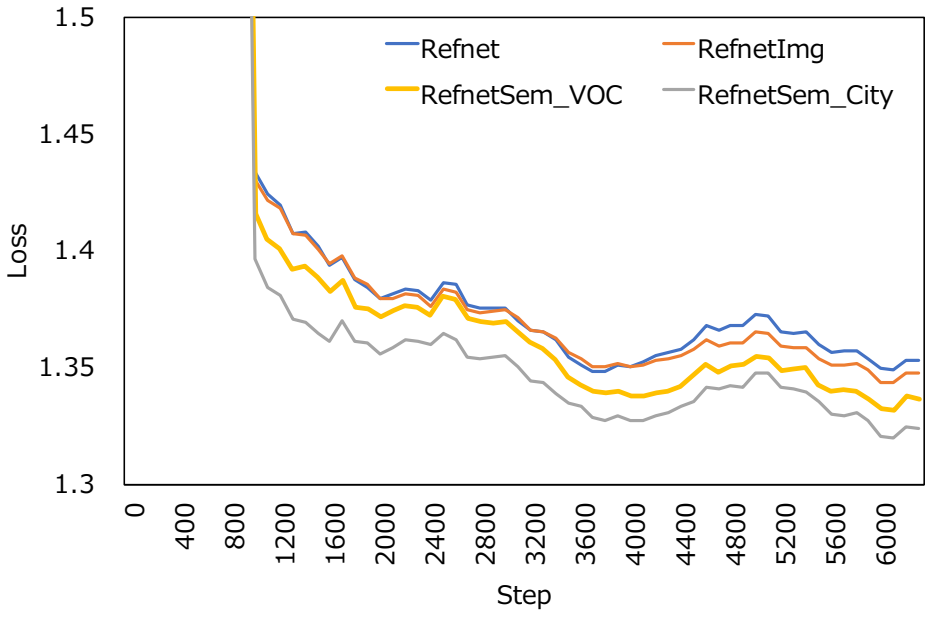
3.3.2 KITTI2015 データセットでの学習結果

KITTI2015 データセットにおいても同様に提案ネットワークの学習能力を比較する．Fig. 3.3(a)，3.4(a) から，KITTI2012 データセットの場合と同様に初期視差マップの種類によらず，RefnetSem_City が最も損失が小さくなっていることがわかる．ただ損失の順序が変化しており，KITTI2015 データセットにおいては Refnet より RefnetImg が優れていることが見て取れる．つまり，RGB 画像が視差マップの改善に有効であるかどうかはデータセットにより定まらないが，深層ネットワークを通すことで有効に定まることがわかる．汎化能力に対しても，Fig. 3.3(b)，3.4(b) から KITTI2012 データセットでの結果と同様に，RGB 画像を直接用いるよりも深層ネットワークを通す方が有効であることが見て取れる．

また，Fig. 3.3，3.4 において RefnetSem_VOC と RefnetSem_City を比較すると，KITTI2012 データセットの結果と同様に RefnetSem_City が常に優れており，この結果からも学習能力や汎化能力の向上にセマンティックマップの精度が重要であることがわかる．

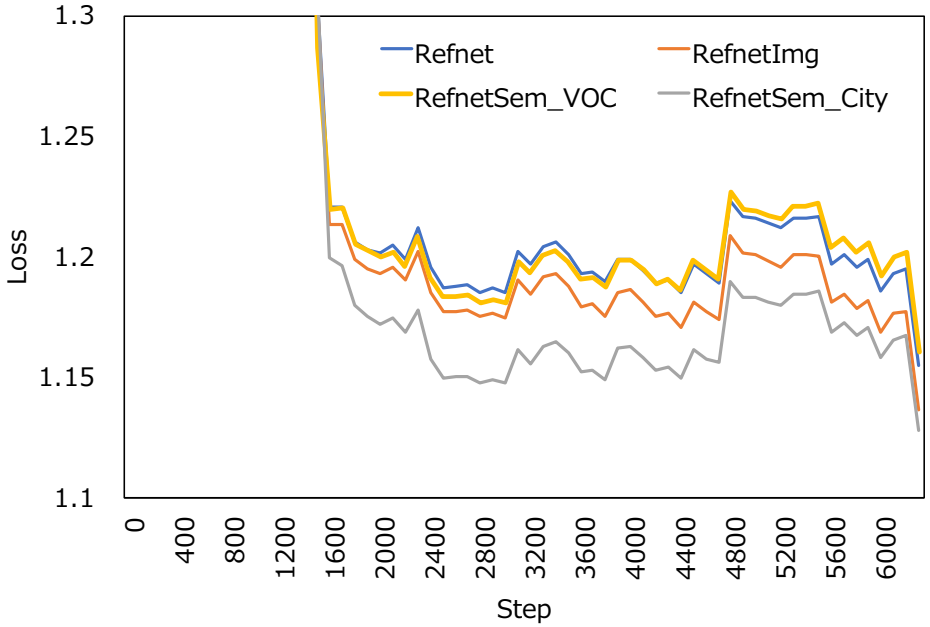


(a) Training

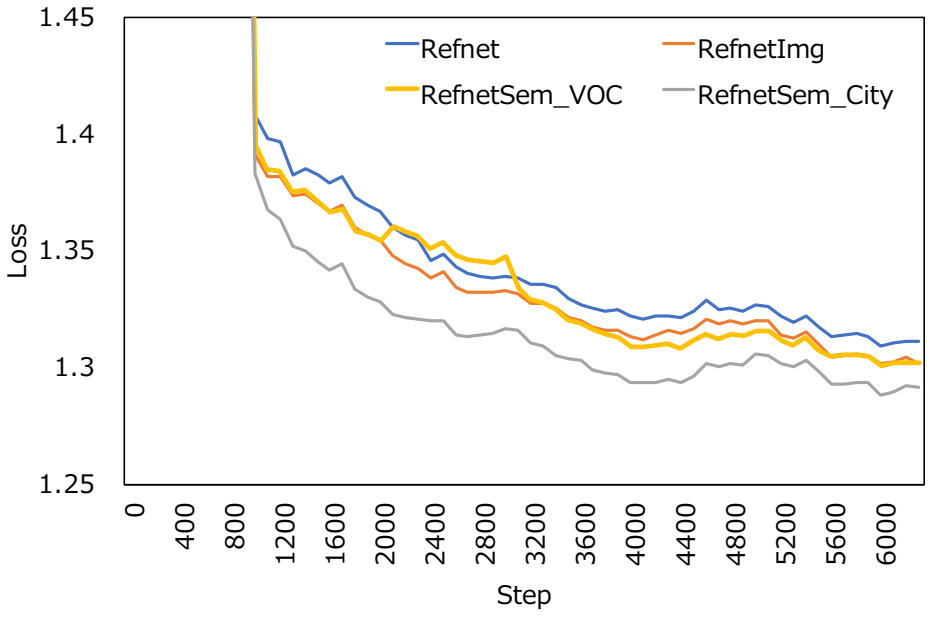


(b) Testing

Fig. 3.3: Loss during training model with the KITTI2015 dataset. Initial disparity map: MC-CNN-fast



(a) Training



(b) Testing

Fig. 3.4: Loss during training model with the KITTI2015 dataset. Initial disparity map: MC-CNN-arct

3.3.3 考察

上記の学習結果から，提案したネットワークにおいて視差の改善に最も有益な情報となったのは Cityscapes データセット学習済みのセマンティックマップであることがわかった．RGB 画像や対象画像に対してクラス分割数の少ない VOC2012 データセット学習済みのセマンティックマップが有益な情報となるかどうかはシーンにより異なり，逆に精度を下げる結果になる場合もあることがわかった．

提案ネットワークにおける視差マップの改善手法はあくまで局所的な 9×9 のフィルタ処理である．さらにカーネルの間隔が空いているため，重要な情報となるのは前景と背景を分割可能な情報であると考えられる，この情報として RGB 画像等は適切な場合もあるが不適切な場合（前景と背景が酷似しているなど）もあり，これが提案ネットワークの学習に影響したと考えられる．

3.4 KITTI ベンチマークでの評価

次に，KITTI2012 データセットと KITTI2015 データセットのテスト画像に対して視差マップの改善を行った．視差マップの改善には提案した Refnet, RefnetImg, RefnetSem_VOC，RefnetSem_City の 4 つに加え，比較評価のため MC-CNN の論文で用いられていた古典的な手法である SGM やメディアンフィルタ，ガウシアンフィルタをそれぞれ用いることにする．

評価には，KITTI ベンチマークで定められている指標を用いる．KITTI2012 と KITTI2015 ではそれぞれエラーの定義が異なり，KITTI2012 では誤差が 3 画素以上，KITTI2015 では誤差が 3 画素以上かつ 5% 以上であればエラーとなる．評価範囲内においてエラーと判定される視差の割合をエラー率として最終的に評価する．この評価範囲については隠れ領域を含まない領域に限定した．これは本手法が局所的なフィルタリングであるため，左右の画像間で対応点が広く存在しない隠れ領域の視差改善に有効ではないことが明らかなためである．また，推定する最大視差は真値の最大値と同じ 227 と規定した．

3.4.1 KITTI2012 データセットでの評価結果

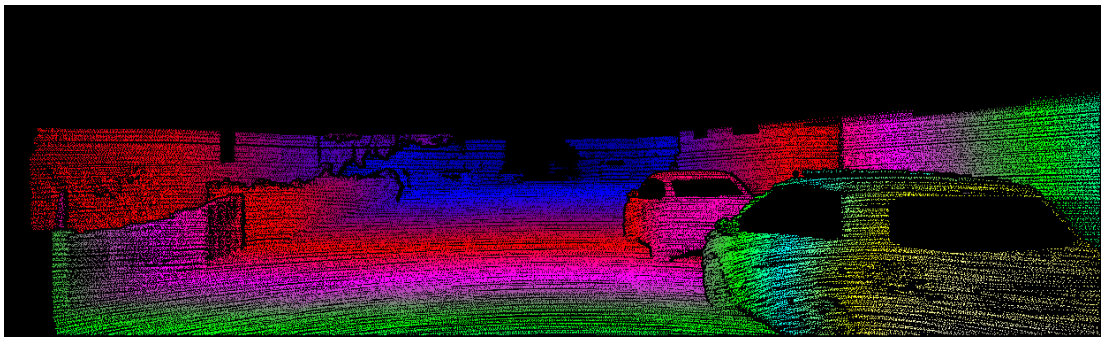
KITTI2012 データセットのあるテスト画像に対して，上記した 5 つの手法を用いて視差マップの改善を行なった．入力となるステレオ画像，視差の真値，セマンティックマップを Fig. 3.5 に示し，改善前と改善後の視差マップとそのエラーマップ，エラー率を Fig. 3.6 に示す．改善例においては SGM とフィルタ処理（メディアンフィルタ，ガウシアンフィルタ）と比べ，RefnetSem_City が優れた結果を出している．それぞれのエラーマップを比較すると，エラーが改善されているのは主に物体の境界付近やエラーが広く存在している領域であることがわかる．改善例において SGM より優れていた理由としては，用いるカーネルが前述した様にスパースであるため大域的なエラーの影響を受け難いこと，またセマンティックマップを用いることで上手く前景の視差を強調できたためだと考えられる．

次に，全テスト画像に対して同様のことを行い，そのエラー率の平均を計算した．この結果を Table 3.2 に示す．Table 3.2 から，平均的には SGM に及んでいないことがわかる．ただし，あくまでローカルなフィルタ処理としては本手法は有効で

あり，中でも PSPNET_City で推定されたセマンティックマップを用いたものは最も改善率が高いことがわかる．

(a) Left image

(b) Right image

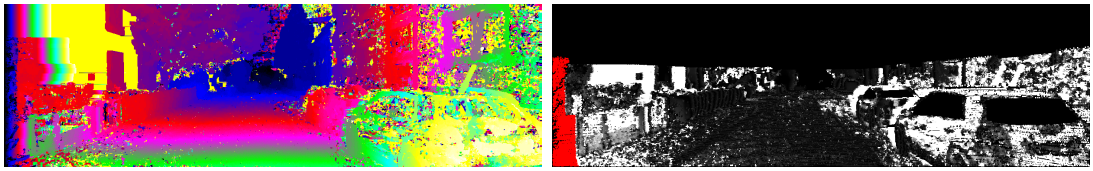


(c) Ground truth

(d) Semantic map: VOC2012

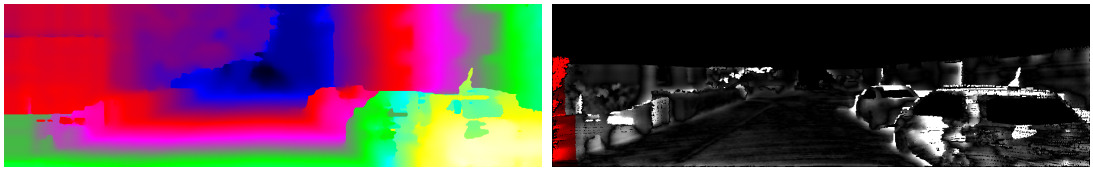
(e) Semantic map: Cityscapes

Fig. 3.5: KITTI2012: Example of input data and ground truth



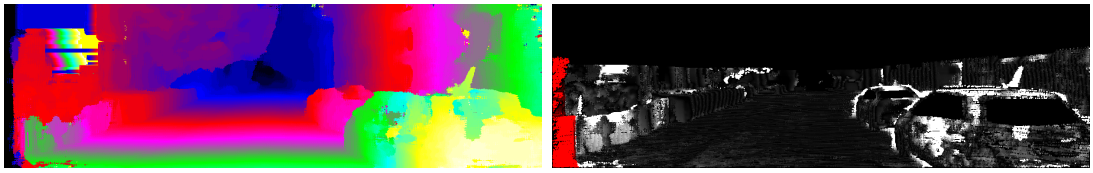
(a) Non refinement

Error: 25.6 %



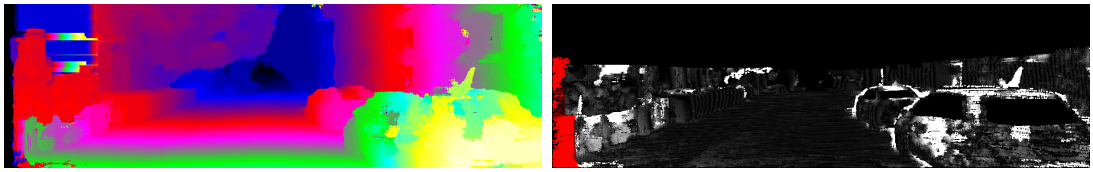
(b) SGM, Median filter, Gausian filter

Error: 5.63 %



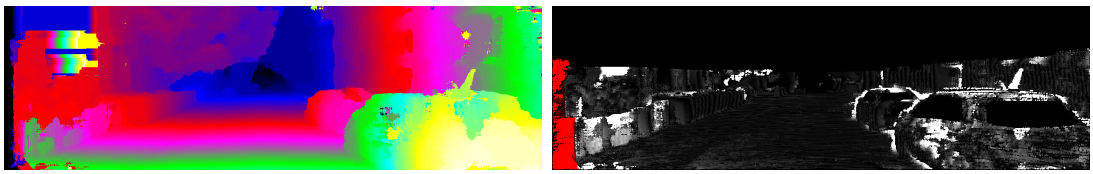
(c) Refnet

Error: 7.33 %



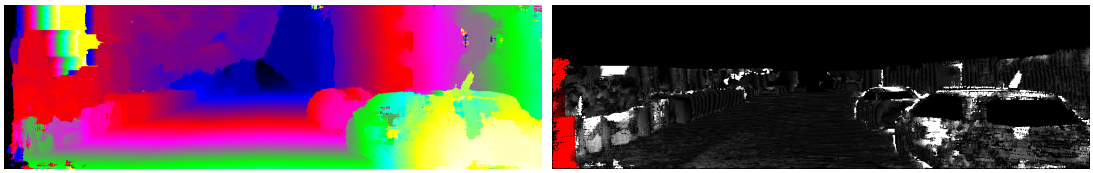
(d) RefnetImg

Error: 7.61 %



(e) RefnetSem_VOC

Error: 6.51 %



(f) RefnetSem_City

Error: 5.22 %

Fig. 3.6: KITTI2012: Example of the evaluation result

Table 3.2: Average error rate in the testing images of KITTI2012

Initial disparity map	Refinement method	Error [%]
MC-CNN-fast	Non-refinement	17.2
	Default[a]	**3.60**
	Refnet	5.07
	RefnetImg	5.17
	RefnetSem_VOC	4.95
	RefnetSem_City	4.57
MC-CNN-arct	Non-refinement	15.0
	Default[b]	**2.84**
	Refnet	4.37
	RefnetImg	4.46
	RefnetSem_VOC	4.26
	RefnetSem_City	4.02

[a]SGM, Median filter, Gausian filter

[b]Cross-based cost aggregation, SGM, Median filter, Gausian filter

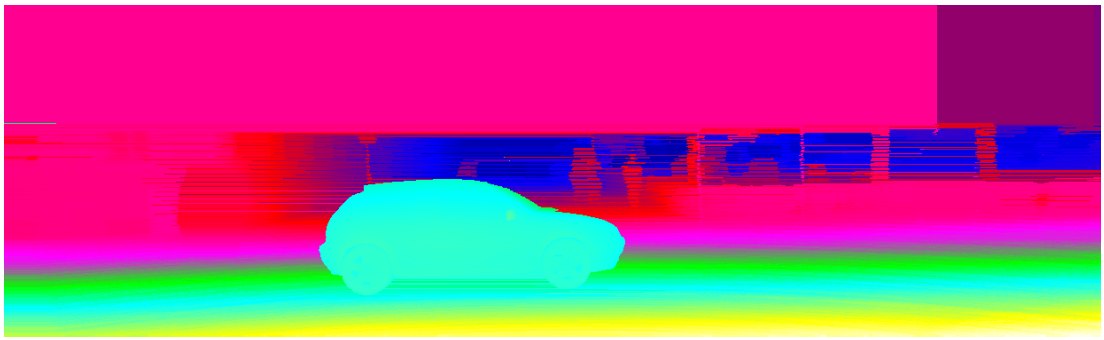
3.4.2 KITTI2015 データセットでの評価結果

KITTI2015 データセットのあるテスト画像に対して，上記した 5 つの手法を用いて視差マップの改善を行なった．入力となるステレオ画像，視差の真値，セマンティックマップを Fig. 3.5 に示し，Fig. 3.6 の左列に改善前と改善後の視差マップを，右列にそれぞれのエラーマップ，エラー率を示す．改善例においては，前述した KITTI2012 データセット同様に Refnet_City が最も誤差率が少ない結果となっている．Fig. 3.8(c) と 3.8(f) を比較すると，RefnetSem_City では自動車境界周りや窓ガラス部分におけるエラーが低減されていることがわかる．また，Fig. 3.8(b) と 3.8(f) を比較すると，やはり透過領域である窓ガラス部分のエラーが SGM と比べ低減できていることがわかる．

テスト画像全体に対するエラー率を Table 3.3 に示す．初期視差マップが MC-CNN-fast である場合においては，SGM とメディアンフィルタ，ガウシアンフィルタによる改善を超えるエラー率を達成できた．一方で，初期視差マップが MC-CNN-arct である場合においては本来の改善手法を上回ることができていない．しかしながら，KITTI2012 データセットでの評価同様に，提案ネットワークにおいて最も視差マップ改善に有益なのは PSPNET_City で推定されたセマンティックマップであることが確認できた．

(a) Left image

(b) Right image

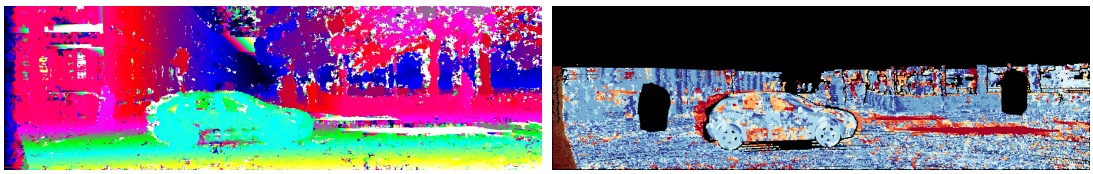


(c) Ground truth

(d) Semantic map: VOC2015

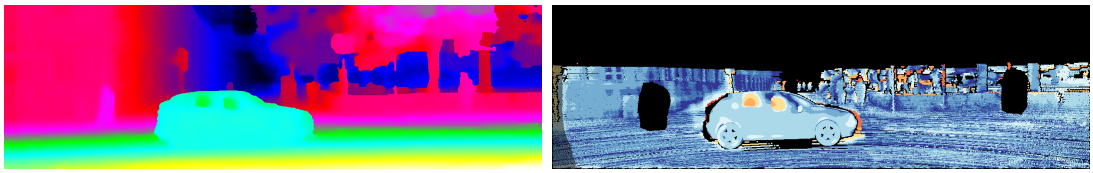
(e) Semantic map: Cityscapes

Fig. 3.7: KITTI2015: Example of input data and ground truth



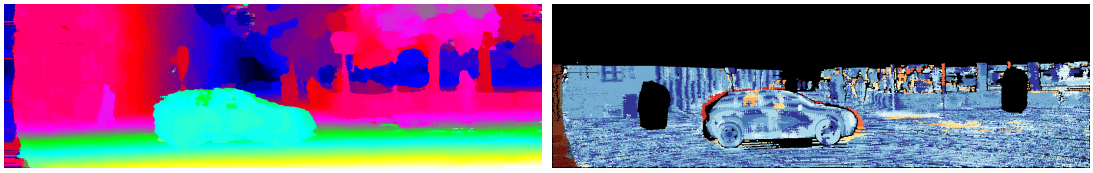
(a) Non refinement

Error: 20.0 %



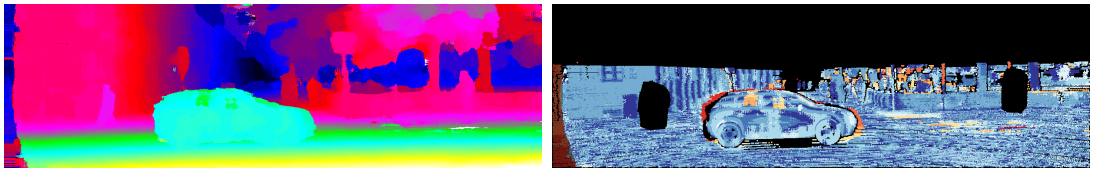
(b) SGM, Median filter, Gausian filter

Error: 3.00 %



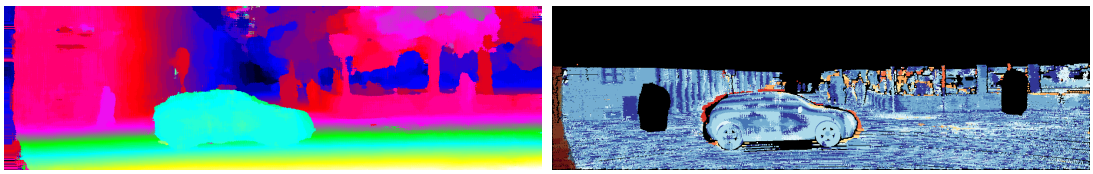
(c) Refnet

Error: 3.87 %



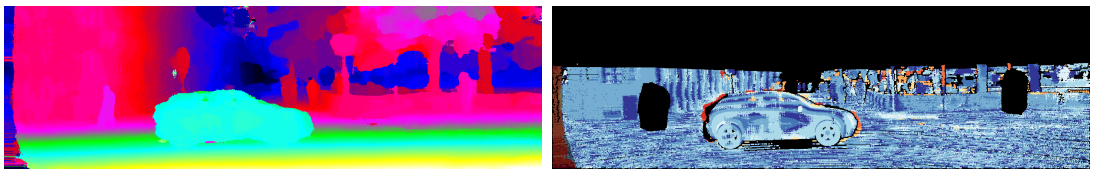
(d) RefnetImg

Error: 3.95 %



(e) RefnetSem_VOC

Error: 2.43 %



(f) RefnetSem_City

Error: 2.13 %

Fig. 3.8: KITTI2015: Example of the evaluation result

Table 3.3: Average error rate in the testing images of KITTI2015

Initial disparity map	Refinement method	Error [%]
MC-CNN-fast	Non-refinement	16.6
	Default[a]	4.25
	Refnet	4.74
	RefnetImg	4.66
	RefnetSem_VOC	4.26
	RefnetSem_City	**4.14**
MC-CNN-arct	Non-refinement	14.3
	Default[b]	**3.06**
	Refnet	3.94
	RefnetImg	3.78
	RefnetSem_VOC	3.65
	RefnetSem_City	3.50

[a]SGM, Median filter, Gausian filter

[b]Cross-based cost aggregation, SGM, Median filter, Gausian filter

3.5 実行環境と実行時間

本評価実験で用いた実行環境を Table 3.5 に示す．また，この実行環境における提案手法の平均実行時間を Table 3.4 に示す．なお，これは提案手法の実行時間であり，視差マップやセマンティックマップ等の入力データを生成する時間は含めていない．メモリに制限があるため，本実行環境では一つの視差マップを改善する際に 50×50 画素に分割して実行する必要があった．

Table 3.4: Processing time

Method	Time [s]
Refnet	52.4
RefnetImg	56.0
RefnetSem_VOC	56.3
RefnetSem_City	56.2

Table 3.5: Computer specification

OS	Ubuntu 16.04 (64bit)
CPU	Intel Core i7-6700HQ (2.6GHz × 8 Cores)
GPU	NVIDIA GeForce GTX 1070 8GB
Memory	16GB DDR4

第 4 章　おわりに

4.1　まとめ

本研究では，ステレオビジョンにおいて，対応点探索が困難な透過領域や鏡面領域に対してロバストな距離計測を目的とし，セマンティックマップを用いたローカルフィルタリング手法を提案した．評価実験では，KITTI ベンチマークを用いて，視差マップの改善にセマンティックマップが RGB 画像よりも優位に働くことを示した．また，同じ画像に対するセマンティックマップに対しても，より前景と背景を分けることのできているをセマンティックマップを用いた方が視差マップの改善に役立つことがわかった．さらに，隠れ領域を除けば SGM とほぼ同程度の改善精度を得ることができる場合もあった．

本研究を通して，セマンティックマップが視差マップの改善に有益であるということが示せた．これは，人間が普段感じている距離感の認識に近いものではないかと考えられる．セマンティックセグメンテーションの精度は年々向上しており，今後も向上することが期待される．それに伴い，本手法の様なセマンティックマップを活用する手法もより実用的なものになることが考えられる．本研究ではこの一例を示すことができた．

4.2　今後の課題と展望

今後の課題としては，改善精度と実行時間が挙げられる．本手法はあくまで 9×9 の局所的なフィルタリングであるため，SGM の様な大域的な最適化が期待できない．また，各ピクセルに対する全ての視差の確率を考慮する必要がある本手法は計算コストが大きく，まだまだ実用的ではないことも課題の一つである．

今後の展望としては，視差マップを用いたセマンティックマップの改善が考えられる．本研究では，視差マップの改善に対してセマンティックマップが有効に働くことを示した．これを踏まえ，今後は反対にセマンティックマップを視差マップで改善する手法を考えることで，相互に改善できるネットワークが実現できるのではないかと考えられる．

謝辞

本研究は，奈良先端科学技術大学院大学 情報科学研究科 情報科学専攻 ロボティクス研究室 小笠原司教授のご指導の下で行いました．本研究の遂行にあたり，多岐にわたるご指導やご助言を賜りました小笠原司教授に深く感謝いたします．

研究方針や論文執筆にあたり，専門分野から様々な御指導や御助言をいただきました同研究科の高松淳准教授に深く感謝いたします．

同研究科の丁明助教には本研究室の研究会などで多くの御指導や御助言を頂きました．深く感謝いたします．

本研究室博士研究員の GARCIA RICARDEZ Gustavo Alfonso 氏には本研究についての様々な議論や，論文執筆，英文添削など多くの点において御指導や御助言を頂きました．深く感謝いたします同研究科秘書の大脇美千代氏，池田美輝氏には学生生活を送る上で様々なサポートをして頂きました．深く感謝いたします．

研究方針や論文執筆にあたり，厳しくも的確な御意見を頂きました本研究室博士後期課程の築地原里樹氏並びに山﨑亘氏, 趙崇貴氏, 湯口彰重氏, VOL DRIGALSKI Felix 氏，Lotfi El Hafi 氏に深く感謝いたします．

先輩，同輩ならびに後輩達には研究に対してのご助言や様々な形での支援を頂きました．深く感謝いたします．

最後に，長い学生生活を支えてくださった両親，家族ならびに友人達に心から感謝いたします．

参考文献

[1] M. Aly, “Real time detection of lane markers in urban streets,” in Proc. of the IEEE Intelligent Vehicles Symposium (IV), pp. 7–12, 2008.

[2] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, “Survey of pedestrian detection for advanced driver assistance systems,” IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 32, no. 7, pp. 1239–1258, 2010.

[3] R. Okuda, Y. Kajiwara, and K. Terashima, “A survey of technical trend of adas and autonomous driving,” in Proc. of the International Symposium on VLSI Technology, Systems and Application (VLSI-TSA), pp. 1–4, 2014.

[4] 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.

[5] M. R. Hafner, D. Cunningham, L. Caminiti, and D. D. Vecchio, “Cooperative collision avoidance at intersections: Algorithms and experiments,” IEEE Transactions on Intelligent Transportation Systems, vol. 14, no. 3, pp. 1162–1175, 2013.

[6] P. Viswanath, K. Chitnis, P. Swami, M. Mody, S. Shivalingappa, S. Nagori, M. Mathew, K. Desappan, S. Jagannathan, D. Poddar, A. Jain, H. Garud, V. Appia, M. Mangla, and S. Dabral, “A diverse low cost high performance platform for advanced driver assistance system (adas) applications,” in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 1-9, 2016.

[7] C. Bahlmann, Y. Zhu, V. Ramesh, M. Pellkofer, and T. Koehler, “A system for traffic sign detection, tracking, and recognition using color, shape, and motion information,” in Proc. of the IEEE Intelligent Vehicles Symposium (IV), pp. 255–260, 2005.

[8] J. Greenhalgh and M. Mirmehdi, “Recognizing text-based traffic signs,” IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 3,

pp. 1360–1369, 2015.

[9] K.-Y. Chiu and S.-F. Lin, “Lane detection using color-based segmentation,” in Proc. of the IEEE Intelligent Vehicles Symposium (IV), pp. 706–711, 2005.

[10] A. Broggi, C. Caraffi, P. P. Porta, and P. Zani, “The single frame stereo vision system for reliable obstacle detection used during the 2005 darpa grand challenge on terramax,” in Proc. of the IEEE Intelligent Transportation Systems Conference (ITSC), pp. 745–752, 2006.

[11] H. Hirschmuller, “Accurate and efficient stereo processing by semi-global matching and mutual information,” in Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 807 - 814, 2005.

[12] F. Güney, A. Geiger, “Displets: Resolving Stereo Ambiguities using Object Knowledge,” in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[13] J. Žbontar, Y. LeCun, “Stereo matching by training a convolutional neural network to compare image patches,” Journal of Machine Learning Research (JMLR), vol. 17, pp. 2287-2318, 2016.

[14] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, “Pyramid Scene Parsing Network,” in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[15] S. Ioffe, C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” in Proc. of the International Conference on Machine Learning (ICML), 2015.

[16] F. Yu, V. Koltun, “Multi-Scale Context Aggregation by Dilated Convolutions,” in Proc. of the International Conference on Learning Representations (ICLR), 2016.

[17] D. Scharstein, R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” International Journal of Computer Vision (IJCV), vol. 47, no. 1-3, pp. 7-42, 2002.

[18] V. Badrinarayanan, A. Handa, R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015

[19] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[20] M. Everingham, S. A. Eslami, L. V. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A ret- rospective," International Journal of Computer Vision, vol. 111, no. 1, pp. 98–136.

[21] A. Geiger, P. Lenz, R. Urtasun, "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[22] M. Menze, A. Geiger, "Object Scene Flow for Autonomous Vehicles," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[23] K. Diederik P, B. Jimmy, "Adam: A Method for Stochastic Optimization," in Proc. of the International Conference on Learning Representations (ICLR), 2015.