深度推定（Depth Estimation）とは

画像から深度情報を復元する技術

背景

入力画像からシーンの深度を予測することはロボットナビゲーションにおいて重要な技術。

教師あり学習

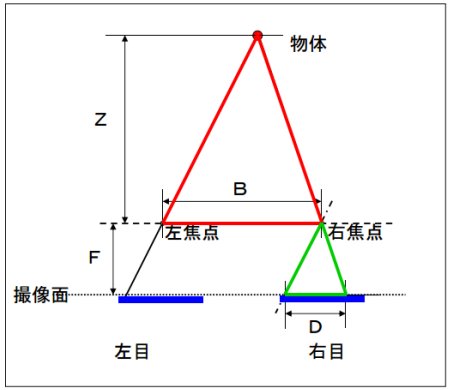
・別で収集した深度マップを教師として使う
・学習は簡単だが、深度マップを取得するのに高価な深度センサを必要

→ 1. Learning Depth from Single Monocular Images
2. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network

教師なし学習

・ステレオ画像や動画を使用する
・モデルは複雑になるが学習に利用可能なデータが格段に増加する

・視差と物体までの距離の関係



赤三角形と
緑三角形の
相似関係から

$$Z = \frac{B \times F}{D}$$

Z：距離
B：カメラ間距離
F：焦点距離
D：視差

→ 3. Unsupervised Learning of Depth and Ego-Motion from Video
4. Depth Prediction Without the Sensors: Leveraging Structure for Unsupervised Learning from Monocular Videos
5. Learning the Depths of Moving People by Watching Frozen People

日付

Learning Depth from Single Monocular Images

（2006） Ashutosh Saxena / Sung H. Chung / Andrew Y. Ng

https://arxiv.org/pdf/1502.07411.pdf

どんなもの？

単眼カメラで撮影した一枚の画像から深度を推定する課題を検討する。

議論はある？

なし

どうやって有効だと検証した？

3Dレーザースキャナーを使い、425セットの画像(森や木、建造物を含む体系化されていない屋外環境の画像)とそれに対応する深度マップを収集し、75%を訓練データに、25%をテストデータとして使用。

先行研究と比べて何がすごい？

今まで深度推定には複眼カメラからの画像を使い、複数画像を必要とするものが多かった。今回の体系化されていない単眼の画像から深度マップを学習するという課題に挑戦した。

技術の手法や肝は？

様々なスケールの局所および全体の画像の特徴を組み入れて識別学習したMRF(マルコフ確率場)を用いている。MRFで隣接するパッチに対する関係を計算する。直近以外の情報も得るため、色々な大きさのパッチを使う。

次に読むべき論文は？

これまで主流だった両眼カメラによる画像の深度推定
A taxonomy and evaluation of dense two-frame stereo correspondence algorithms
https://vision.middlebury.edu/stereo/taxonomy-IJCV.pdf

日付

Depth Map Prediction from a Single Image using a Multi-Scale Deep Network

（2014） David Eigen / Christian Puhrsch / Rob Fergus

https://arxiv.org/pdf/1406.2283.pdf

どんなもの？

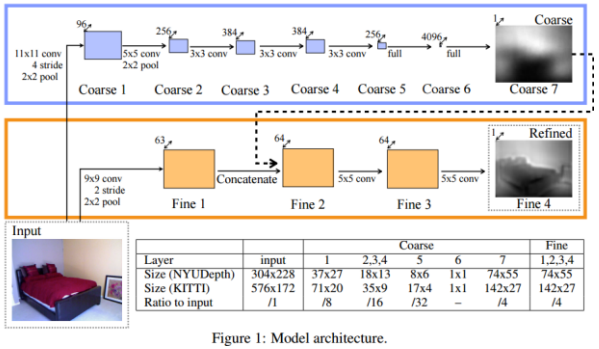
2つのディープネットワークを用いて単眼画像から深度を予測する手法。

どうやって有効だと検証した？

NYU DepthとKITTIデータセットを使用して検証。画素同士の関係も損失関数に組みこんだScale-Invariant Errorを設定。

技術の手法や肝は？

2段構成の畳み込みネットワークを用いている。1段目のCoarseネットワークでは全体の深度推定をし、2段目のFineネットワークで局所的な深度予測をする。



Layer	input	Coarse					Fine
		1	2,3,4	5	6	7	1,2,3,4
Size (NYUDepth)	304x228	37x27	18x13	8x6	1x1	74x55	74x55
Size (KITTI)	576x172	71x20	35x9	17x4	1x1	142x27	142x27
Ratio to input	/1	/8	/16	/32	–	/4	/4

Figure 1: Model architecture.

議論はある？

より微細なスケールのローカルネットワークを反復して適用することで深度マップを入力のオリジナルの解像度にも広げられるようにすることが次の課題である。(今回は最終的に入力画像の1/4のサイズの深度マップが推定されている)

先行研究と比べて何がすごい？

保存しやすいパラメータの学習を行い、リアルタイムで画像に適用できる。また、学習時に使う教師データを取るために、深度を計るセンサーを利用するが、テスト時においては完全にソフトウェアベースで、RGB画像から深度を予測する。

次に読むべき論文は？

この手法の改良バージョン
Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture
https://arxiv.org/pdf/1411.4734v4.pdf

日付

Unsupervised Learning of Depth and Ego-Motion from Video

（2017） Tinghui Zhou / Matthew Brown / Noah Snavely / David G. Lowe

https://arxiv.org/pdf/1901.00979.pdf

どんなもの？

SfMLearnerと呼ばれる手法で、単眼（カメラが1つ）で深度を推定する手法、それも単眼で撮影された動画を元に、教師なしで学習できるという手法。

議論はある？

Explainability Maskは、原理的にソース画像からの合成が困難な場合に有用である一方で、オクルージョン（手前にある物体が背後にある物体を隠す状態）が存在するという物体の前後関係を示す重要な幾何的情報を無視してしまうという欠点がある。

どうやって有効だと検証した？

損失関数として、合成画像の損失と深度マップの平滑化損失と正則化損失を足し合わせたものを設定し、教師データとしてKTTIを用いる手法比較を行う。

先行研究と比べて何がすごい？

一般的によく知られている画像を元にした深度推定の手法は、ステレオカメラなど複数台のカメラを利用した手法だった。本手法は、このような考え方とは異なっており、いわばセグメンテーションの連続値版のようなことをしている。つまり、1枚の入力画像を元に、深度マップ画像を出力するという手法をとっている。

技術の手法や肝は？

ターゲットとなる深度マップの正解データが存在していないため、一般的な教師ありのセグメンテーション手法とは訓練プロセスの勝手が違う。本手法では、深度マップの推定を間接的に最適化するための仕組みを作っている。

次に読むべき論文は？

今回のSfMLearnerをベースにしたMonoDepth2と呼ばれる単眼深度推定モデルの論文
Digging Into Self-Supervised Monocular Depth Estimation
https://arxiv.org/pdf/1806.01260.pdf

日付

Depth Prediction Without the Sensors: Leveraging Structure for Unsupervised Learning from Monocular Videos

（2018） Vincent Casser/ Soeren Pirk / Reza Mahjourian / Anelia Angelova

https://arxiv.org/pdf/1811.06152.pdf

どんなもの？

Struct2depthと呼ばれるモデルで、単眼カメラから撮影したRGB画像を入力に、深度とエゴモーションを推定する教師なし学習を用いた手法を提案している。

議論はある？

なし

どうやって有効だと検証した？

KITTIデータセット、Cityscapesデータセット、Fetch Indoor Navigationデータセットで実験。

先行研究と比べて何がすごい？

センサを利用せずRGB画像のみを利用して(正解となる深度情報なしで)深度を推定。さらに深度だけでなくエゴモーション(カメラ自身の動き)も推定している。

技術の手法や肝は？

ニューラルネットワークを用いて直接深度を学習するのではなく、個々のオブジェクトに分解するアプローチにより安価な単眼カメラのみで対処できることを実証している。
個々のオブジェクトとモーションを独立して3Dモデリングすることで、深度とエゴモーションを推定する。また、オンザフライで学習を適応する。

次に読むべき論文は？

今回の手法のstruct2depthをベースにしつつも、内部パラメータが未知のときでもうまく訓練できるようにした手法の論文
Depth from Videos in the Wild: Unsupervised Monocular Depth Learning from Unknown Cameras
https://arxiv.org/pdf/1904.04998.pdf

日付

Learning the Depths of Moving People by Watching Frozen People

（2019）Zhengqi Li / Tali Dekel / Forrester Cole / Richard Tucker / Noah Snavely / Ce Liu / William T. Freeman

https://arxiv.org/pdf/1904.11111.pdf

どんなもの？

カメラとカメラが映している人物が同時に移動していても奥行き情報を推定可能な手法。

どうやって有効だと検証した？

ロスとして、教師との二乗誤差、教師とのL1誤差、平滑化項を足し合わせたものを設定。NYUとRGBDを教師データとして利用したモデルと比較し、ロスが小さくなっていることを確認している。

技術の手法や肝は？

ディープラーニングに人間のポーズや形状に関して事前知識を学習させることで実現している。YouTubeにあるマネキンチャレンジを教師として使用している。撮影風景内の物体は全て静止しているため、MVSのような三角測量ベースの手法を適用することが可能になり、風景内の人物を含む全体の正確な奥行き、つまり深度情報を取得できる。

議論はある？

ほとんどのシーンを動くものが占めているとき、推測するのが難しくなる。また、予測された深度は、車や影などの人間以外が動いている領域では不正確になる場合がある。今回の手法は2つのビューのみを使用しているため、時間的に一貫性のない深度推定につながることがある。

先行研究と比べて何がすごい？

教師データを作るには深度センサが必要であったため、コストが高かった。そこでマネキンチャレンジと呼ばれる動画を利用して人に関する深度の正解値をSfMとMVSを利用して作った。

次に読むべき論文は？

Google Researchによる深度推定の最新論文
Unsupervised monocular depth and ego-motion learning with structure and semantics
https://openaccess.thecvf.com/content_CVPRW_2019/papers/VOCVALC/Casser_Unsupervised_Monocular_Depth_and_Ego-Motion_Learning_With_Structure_and_Semantics_CVPRW_2019_paper.pdf

日付