Accurate, Explainable, and Robust Depth Estimation from Monocular Images

氏名	HU JUNJIE 胡 君傑
学位の種類	博 士（情報科学）
学位記番号	情 博 第 707 号
学位授与年月日	令和 2年 3月25日
学位授与の要件	学位規則第4条第1項該当
研究科、専攻	東北大学大学院情報科学研究科（博士課程） システム情報科学専攻
学位論文題目	Accurate, Explainable, and Robust Depth Estimation from Monocular Images（高精度で説明可能かつ頑健な単眼画像からの奥行推定）
論文審査委員	（主査）東北大学教 授 岡谷 貴之 東北大学教 授 橋本 浩一　東北大学教 授 滝沢 寛之 東北大学准教授 鏡 慎吾

論文内容の要旨

Chapter 1
We introduce the background knowledge of this thesis. To be specific, we first introduce the concept of depth map and deep neural networks. We then show how to estimate depth map from monocular images. We show previous methods built on deep neural networks and analyze their limitations. We then discuss the motivations of our works. In order to make a significant step forward towards real-world deployment of methods built on neural networks, we argue that there are three issues that exist to be solved and investigated including accuracy improvement, interpretability and robustness against adversarial attacks.

Chapter 2
We propose to improve the accuracy of depth estimation from two aspects: 1) We introduce a multi-scale feature fusion module that fuses features of different scales extracted by an encoder, the module is inserted into an encoder-decoder network. Our network consists of four modules: an encoder, a decoder, a multi-scale feature fusion module, and a refinement module. 2) We propose a more comprehensive loss function that penalizes errors on depth, gradients and normals. Based on experimental results, the two improvements contribute to improving the accuracy of depth map estimation.

Chapter 3
To understand how depths are predicted by CNNs, we are interested in the question that what's the cues that CNNs use to infer depths. To this end, we propose to learn a mask that reveals the important features for depth inference when given an input image. We formulate it as an optimization problem that CNNs can still estimate accurate depths while using pixels as small as possible. We then analyze the remained pixels to understand the cues employed by CNNs for depth estimation. We experimentally show the effectiveness of our approach.

Chapter 4
We then attempt to understand what kind of features are learned inside CNNs.
We first carefully designed a method that automatically clusters features learned inside CNNs, we find that there is a high redundancy inside CNNs. Many neurons repeatedly learn similar features which also explains why applying compression to CNNs does not cause damage to the accuracy of models.
We then quantify those features to calculate the importance of each feature for depth inference. Experiment results show similar to image recognition, removing individual unit does not cause a drop for overall accuracy, but the drop is significant for some images.

Chapter 5
We investigate the performance of previous studies on adversarial images, we find that those methods could be easily fooled. We consider it's necessary to develop a defense method for monocular depth estimation since it's usually used in real-world applications. We propose to defend adversarial attacks with the prediction of saliency maps. Our method boosts the robustness of CNNs while simultaneously contributes to better understanding to black-boxs. Extensive experimental results validate the effectiveness of our method.

Chapter 6
In this chapter, we conclude our works in this thesis.

論文審査結果の要旨

単眼深度推定，すなわち 1 枚の画像からシーンの立体的な奥行きを計算する問題は，ロボットビジョンの基本的な問題である。近年，深層畳み込みニューラルネットワーク（以下 CNN）を用いた機械学習の適用によって大幅な性能向上が果たされ，現在，自動運転を始めとする多様な実世界応用が検討されている。本論文は，そこで課題となっている推定精度の一層の向上と，説明性や安全性の確保に，それぞれ取り組むものであり，全編 6 章からなる。

第 1 章は序論であり，本研究の目的と背景を述べている。

第 2 章では，CNN による単眼深度推定の推定精度を向上させる方法を提案している。ネットワーク構造並びに学習時の損失関数の改良により，従来方法と比べ空間解像度が高く，より高精度な奥行きマップの推定を可能にしている。2 つの標準的なデータセットを用いた評価実験において，従来の評価尺度で世界最高水準の精度を達成するとともに，新たに導入した解像度の高い領域での評価尺度において，世界最高精度を達成しており，重要な成果である。

第 3 章では，CNN が行う単眼深度推定を可視化する方法を提案している。 1 枚のシーンの画像に対して，深度推定に不可欠な，ただしなるべく少数の画素を，もう一つの CNN の学習によって特定し，提示する方法である。この方法によって，入力画像内であまり目立たない濃淡のエッジ構造やシーンの消失点，すなわち無限遠方の画像上の像が，深度推定に不可欠であることが明らかになるなど，重要な結果を与えている。

第 4 章では，単眼深度推定を行う CNN 内部の計算機構を明らかにすることを目標に，画像入力時の各層の出力を分析している。CNN の各層，各チャネルの出力をクラスタ分析することで，チャネル単位での役割を特定している。シーンの点までの遠近や，奥行きの構造の空間周波数の高低に応じ，担当するチャネルが異なるという未知の知見が得られており，これは重要な結果である。

第 5 章では，単眼深度推定を行う CNN に対する敵対的攻撃への防御の方法を提案している。3 章で提案した可視化方法を用い，深度推定に重要でない画素をマスクして取り除くことにより，敵対的攻撃の防御が可能となることを明らかにしている。単眼深度推定のための CNN を対象とする最初の防御方法の提案であり，加えてそもそも敵対的攻撃がなぜ実行可能なのか，また CNN 内部の計算機構に対する示唆を同時に与えており，重要な成果である。

第 6 章は結論である。

以上要するに本論文は，CNN を用いた単眼深度推定に現在望まれている，推定精度，説明性，並びに安全性の各要素の向上について，これを達成する複数の方法を提案している。この成果は，深層学習による単眼深度推定を実世界での応用に一層近づけるものであり，システム情報科学ならびにロボットビジョンの発展に寄与するところが少なくない。

よって，本論文は博士（情報科学）の学位論文として合格と認める。