別紙４－１（課程博士（英文））

Date of Submission (month day, year) : July 5th, 2023

Department of Computer Science and Engineering	Student ID Number	D 179301	Supervisor	Yasushi Kanazawa
Applicant's name	Andi Hendra			

Abstract (Doctor)

Title of Thesis	Depth Estimation from A Single Image using Global Structure and Local Scene Information

Approx. 800 words

We propose a novel framework for accurately estimating depth information from a single image. Despite its simplicity and compactness, our framework demonstrates consistent and reliable performance by effectively integrating global and local image features.

To address the challenge, we employ two distinct deep neural network architectures. We adopt an encoder-decoder model with a two-stage strategy in the first architecture. This approach leverages multi-task loss optimization and incorporates adaptive learning rate adjustment based on the loss behaviour. The second architecture expands the conditional GAN (cGAN) model, introducing a three-player GAN (TP-GAN) framework. To enhance the reliability of the depth estimation, we include the structural similarity measure (SSIM) loss as part of this architecture. By utilizing this architecture, we aim to optimize the depth estimation performance.

Our proposed architectures utilize *1 x 1* convolution to reduce the dimensionality of the feature maps, thereby enabling the model to focus on capturing high-level semantics and global context. Conversely, local features are extracted through stacks of convolution with smaller kernels relative to the input size that can help in capturing local context and details that might be overlooked by global features alone. Combining global and local features enables the model to leverage the overall scene understanding and fine-grained local details to enhance the accuracy of depth prediction.

To evaluate the effectiveness of our approaches, we conducted comprehensive quantitative and qualitative comparisons with several state-of-the-art methods in the field. Our experiments were conducted on two well-known publicly depth datasets, the indoor NYU Depth v2 and outdoor KITTI datasets. The results consistently demonstrate that our proposed method outperforms numerous previous related monocular depth strategies. Despite the conciseness of our model architecture, it consistently delivers reliable performance when compared to the transformer-based model, further demonstrating its efficacy.

Furthermore, we investigated the generalization capabilities of our model to other datasets. To assess cross-dataset adaptation, we trained our model on one dataset and tested it on another and vice versa. Our model exhibits reliable generalization by effectively learning scene variations across indoor and outdoor datasets. Notably, when trained on indoor data and tested on the outdoor range dataset, our model achieved consistent

performance of SSIM scores, which some values close to one.

In addition, we conducted an in-depth analysis to assess the robustness of our depth estimation model under different contrast levels. To evaluate our model's performance, we generated visualizations of estimated depth and calculated the (SSIM) score using images captured under different contrast conditions. Specifically, we evaluated six random KITTI data samples containing scenes with normal, lower, and higher contrast levels. The results demonstrated that our model outperformed other methods, indicating that the SSIM metrics consistently showed superior performance across the dataset.

In future research, it is imperative to further advance the single image depth estimation field by focusing on developing models that exhibit enhanced generalization capabilities across diverse datasets. This could be achieved by designing an adaptive model that effectively discriminates between ground truth and generated depth and accurately classifies whether the input image belongs to an indoor or outdoor dataset. Such advancements would significantly contribute to the robustness and versatility of depth estimation methods.