Lam Huynh

FROM 3D SENSING TO DENSE PREDICTION
ACTA UNIVERSITATIS OULUENSIS
C Technica 858

LAM HUYNH

FROM 3D SENSING TO DENSE PREDICTION

Academic dissertation to be presented with the assent of the Doctoral Programme Committee of Information Technology and Electrical Engineering of the University of Oulu for public defence in Auditorium IT116, Linnanmaa, on 9 December 2022, at 12 noon

UNIVERSITY OF OULU, OULU 2022
Huynh, Lam, From 3D sensing to dense prediction.
University of Oulu Graduate School; University of Oulu, Faculty of Information Technology and Electrical Engineering
Acta Univ. Oul. C 858, 2022
University of Oulu, P.O. Box 8000, FI-90014 University of Oulu, Finland

Abstract
This thesis introduces novel learning-based approaches for improving 3D sensing and dense prediction. In recent years, deep neural networks (DNNs) have thrived on various vision tasks. Nonetheless, current developments indicate a compromise between accuracy, network size, and architectural engineering cost. This work proposes accurate and lightweight DNNs by exploiting prior knowledge, integrating self-attention, leveraging multi-scale 2D-3D representations fusion, and presenting efficient neural architecture search (NAS) strategies.

Recent monocular depth estimation approaches exhibit impressive results. However, these are often achieved with bulky network architectures employing up to hundreds of millions of parameters and using massive training data. This thesis introduces architectures that exploit geometric constraints and non-local self-attention mechanisms to improve performance. Moreover, the methods achieve state-of-the-art results while using at least ten times less parameters than competing approaches.

Depth completion aims to densify sparse input depth measurements. Best performing depth completion methods only work for cases with relatively high 3D point density. This work proposes a novel multi-scale framework that operates directly on both 2D and 3D feature spaces. Unlike previous approaches, the method performs well on extremely sparse and unevenly distributed 3D points. The proposed architecture is also very compact and works with an arbitrary source of the input 3D points.

Dense prediction resolves mapping problems at the pixel level, comprising many sub-tasks such as depth estimation, semantic segmentation, optical flow prediction, and image restoration. Existing methods usually use human-engineering DNNs or focus on a single sub-task. This thesis presents a novel approach utilizing NAS towards more general dense prediction problems that enable holistic scene understanding.

Keywords: 3D sensing, dense prediction, depth completion, monocular depth estimation, neural architecture search, self-attention
Tiivistelmä


Viimeaikaisilla monokulaarisen syvyyden estimointimenetelmissä on saavutettu vaikutteita tuloksia. Niihin on kuitenkin päästynyt sekä suurilla verkoarkkitehtuureilla, jotka käyttävät jopa satoja miljoonia parametreja ja massiivista määrää opetusdataa. Tämä väitöskirjatyö esittelee arkkitehtuureita, jotka hyödyntävät geometrisia rajoituksia ja ei-paikallisia itsehuomiomekanismeja suorituskyvyn parantamiseen. Lisäksi menetelmillä saavutetaan huippuluokan tuloksia käyttämällä vähintään kymmenen kertaa vähemmän parametreja kuin kilpailevalla lähestymistavoilla.

Syvyyden täydentämisen tarkoituksena on tihentää harvat syvyyssykset. Parhaat syvyyden täydensymmenetelmät toimivat vain tapauksissa, joissa 3D-pistetietysys suhteellisen korkea. Tämä työ esittää puitteet suureille monen skaalan lähestymistavalle, joka toimii suoraan sekä 2D- että 3D-piirrevaruksissa. Toisin kuin aiemmin käytettyjen tavoitelmien, menetelmat sopivat myös harvinaisiin tilanteisiin ja epäselviin syvyydysyksiköihin.

Syvyyden täydentämisen tavoitteena on tihentää harvat syvyyssykset. Parhaat syvyyden täydensymmenetelmät toimivat vain tapauksissa, joissa 3D-pistetietysys suhteellisen korkea. Tämä työ esittää puitteet suureille monen skaalan lähestymistavalle, joka toimii suoraan sekä 2D- että 3D-piirrevaruksissa. Toisin kuin aiemmin käytettyjen tavoitelmien, menetelmat sopivat myös harvinaisiin tilanteisiin ja epäselviin syvyydysyksiköihin.


Asiasanat: 3D-havainnointi, itsehuomio, monokulaarinen syvyyden estimointi, neuroarkkitehtuurihaku, syvyyden täydentäminen, tiheän ennustus
To my family and friends.
Acknowledgements

This thesis was done in the Center for Machine Vision and Signal Analysis at the University of Oulu from 2018–2022. The work is part of the vision-based 3D perception for mixed reality applications project funded by Infotech Oulu and the Faculty of Information Technology and Electrical Engineering.

First and foremost, my deep gratitude goes to my supervisors, Professor Janne Heikkilä and Professor Esa Rahtu. Their valuable guidance, feedback, and dedication have been essential to this dissertation. Throughout my Ph.D. journey, they nurture me with inspiring ideas, challenging questions, and encouraging support while allowing me the freedom to pursue my aspiring vision.

I am incredibly grateful to Professor Jiří Matas, who has been a vital collaborator from the Center for Machine Perception at Czech Technical University in Prague from the early days. His experience and deep knowledge furnish priceless suggestions via lengthy discussions throughout my doctoral training. I want to take this opportunity to thank Phong Nguyen for working in tandem, support and for many sparked conversations over the years. I would like to thank Doctor Matteo Pedone for the fruitful discussions resulting in the second publication. My great appreciation goes to Professor Hirokazu Kato and Professor Takafumi Taketomi for hosting my research visit at the Nara Institute of Science and Technology in Japan.

I want to thank the follow-up members Doctor Pekka Sangi, Associate Professor Georgi Georgiev, and Professor Steven LaValle, for accepting the invitation on short notice and carrying out the process smoothly. My gratitude goes to Professor Theo Gevers (Amsterdam, Netherlands) and Professor Takayuki Okatani (Tohoku, Japan) for their reviews and feedback on this manuscript. I greatly appreciate Professor Michael Felsberg (Linköping, Sweden) for dedicating time and effort in being the opponent of my doctoral defense.


My deepest gratitude goes to my family, who wished to stay unnamed, for giving me life and love. Thank you for always being by my side through thick and thin, as none of this would be possible without you.

Oulu, May 2022
Lam Huynh
### List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_D$</td>
<td>Dense mapping function</td>
</tr>
<tr>
<td>$I$</td>
<td>Input color image</td>
</tr>
<tr>
<td>$V$</td>
<td>Dense estimated values</td>
</tr>
<tr>
<td>2D</td>
<td>Two-dimensional</td>
</tr>
<tr>
<td>3D</td>
<td>Three-dimensional</td>
</tr>
<tr>
<td>4D</td>
<td>Four-dimensional</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>Adaptive boosting</td>
</tr>
<tr>
<td>ASPP</td>
<td>Atrous spatial pyramid pooling</td>
</tr>
<tr>
<td>AR</td>
<td>Augmented reality</td>
</tr>
<tr>
<td>CAS</td>
<td>Customizable architecture search</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional neural network</td>
</tr>
<tr>
<td>CRFs</td>
<td>Conditional random fields</td>
</tr>
<tr>
<td>DCNF</td>
<td>Deep convolutional neural field</td>
</tr>
<tr>
<td>DNNs</td>
<td>Deep neural networks</td>
</tr>
<tr>
<td>FC</td>
<td>Fully connected</td>
</tr>
<tr>
<td>FCRN</td>
<td>Fully convolutional residual network</td>
</tr>
<tr>
<td>FLOPs</td>
<td>Floating-point operations per second</td>
</tr>
<tr>
<td>fps</td>
<td>Frames per second</td>
</tr>
<tr>
<td>GIST</td>
<td>Gistification algorithm</td>
</tr>
<tr>
<td>GPU</td>
<td>Graphics processing unit</td>
</tr>
<tr>
<td>GTX</td>
<td>Giga texel shader extreme</td>
</tr>
<tr>
<td>LiDAR</td>
<td>Light detection and ranging</td>
</tr>
<tr>
<td>MRF</td>
<td>Markov random field</td>
</tr>
<tr>
<td>NAS</td>
<td>Neural architecture search</td>
</tr>
<tr>
<td>PHOG</td>
<td>Pyramid histogram of oriented gradients</td>
</tr>
<tr>
<td>PSNR</td>
<td>Peak signal-to-noise ratio</td>
</tr>
<tr>
<td>REL</td>
<td>Mean absolute relative error</td>
</tr>
<tr>
<td>RGB</td>
<td>Red, green and blue</td>
</tr>
<tr>
<td>RGBD</td>
<td>Red, green, blue and depth</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root mean square error</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>SfM</td>
<td>Structure from motion</td>
</tr>
<tr>
<td>SIFT</td>
<td>Scale-invariant feature transform</td>
</tr>
<tr>
<td>SURF</td>
<td>Speeded up robust features</td>
</tr>
<tr>
<td>ToF</td>
<td>Time of flight</td>
</tr>
<tr>
<td>TS</td>
<td>Tabu search</td>
</tr>
</tbody>
</table>
List of original publications

This dissertation is based on the following articles, which are referred to in the text by their Roman numerals (I–V):


The author of this dissertation had the main responsibility for preparing articles I–V. This includes the implementation of the algorithms, experiments, and the writing. The ideas presented in the articles were devised in group discussions with the co-authors, during which they provided valuable suggestions and feedback.
Contents

Abstract 9
Tiivistelmä 11
Acknowledgements 13
List of abbreviations 15
List of original publications 17
Contents 19

1 Introduction 21
  1.1 Background and motivation 23
  1.2 Scope of the thesis 24
  1.3 Contributions 25
  1.4 Summary of the original articles 26
  1.5 Outline of the thesis 27

2 Single image depth estimation 29
  2.1 Depth from monocular cues 31
  2.2 Depth using hand-crafted features 33
    2.2.1 Parametric depth 34
    2.2.2 Non-parametric depth 35
  2.3 Depth using deep neural networks 36
    2.3.1 Leveraging geometric constraints 38
    2.3.2 Salient point detection with normalized Hessian loss 39
    2.3.3 Discussion 40

3 Depth completion 42
  3.1 High-density input 44
  3.2 Sparse input 45
    3.2.1 Direct depth completion 47
    3.2.2 Spatial propagation 48
  3.3 Extremely sparse input 50
    3.3.1 Multi-scale 2D-3D fusion network 52
    3.3.2 Discussion 54
4 Towards generalized dense prediction

4.1 Neural architecture search ........................................ 49
  4.1.1 Search space ........................................... 50
  4.1.2 Search algorithm ....................................... 51
  4.1.3 Evaluation strategy .................................... 52

4.2 Dense prediction .................................................. 53
  4.2.1 Semantic segmentation .................................... 53
  4.2.2 Image super-resolution .................................. 54
  4.2.3 Disparity estimation ...................................... 55
  4.2.4 Monocular depth estimation ............................. 56
  4.2.5 Generalized dense prediction ............................ 57
  4.2.6 Discussion ................................................ 59

5 Summary and conclusion .......................................... 61

References ............................................................. 63

Original publications ............................................... 77
1 Introduction

1.1 Background and motivation

Holistic scene understanding is essential for many applications, including augmented reality (AR), robotics, autonomous navigation, games, and real estate. Understanding the scene as a whole requires recovering many related aspects such as semantics and geometry. These can be formulated as dense predictions aiming to map every image pixel with corresponding predicted values. Depending on the sub-tasks, these output values can be discrete or continuous in which depth estimation and image super-resolution are often devised as regression, while semantic segmentation is dense classification. This thesis starts with 3D sensing and strives for more generalized dense prediction.

Most images are projections from 3D scenes in which the information on how far each point is from the camera has been lost. The inverse is an ill-posed problem because infinite solutions can exist from a single view. Many efforts have been proposed to recover the depth by utilizing a single image like the Manhattan line (Coughlan & Yuille, 1999), single-view metrology (Criminisi, Reid, & Zisserman, 2000) or multiple images as structure from motion (Ullman, 1979), multi-view geometry (Hartley & Zisserman, 2003), and simultaneous localization and mapping (Durrant-Whyte & Bailey, 2006); still, this problem persists. Using depth sensors is a potential solution, but they only provide partial observations of the world given their limited field of view, incomplete sensing, measurement noise, and they are more expensive.

Learning-based single image depth estimation demonstrates noticeable results and arise as a prominent solution. The basic idea is to train a model to infer depth maps from given input images, assuming that depth can be directly learned from the color values. High-quality predictions were achieved using large architectures (X. Chen, Chen, & Zha, 2019), integrating semantic information (Jiao, Cao, Song, & Lau, 2018) or surface normals (X. Qi, Liao, Liu, Urtasun, & Jia, 2018), recovering occluding contours (Ramamonjisoa & Lepetit, 2019) and constructing virtual normals (Yin, Liu, Shen, & Yan, 2019). Nevertheless, these approaches mostly resort to massive training data and extensive networks to gain accuracy. One solution is to exploit the coplanarity constraints and employ non-local networks (X. Wang, Girshick, Gupta, & He, 2018) to shrink the model size while yielding high-fidelity results.
Whilst monocular depth is an appealing concept, the accuracy of this approach is restricted by the lack of strong regularization. In comparison, depth completions (J. Park, Joo, Hu, Liu, & Kweon, 2020) produce more significant results by using additional depth measurements. That said, SOTA depth completions are apt for scenarios with high density and tend to suffer from extremely sparse and unevenly distributed 3D input patterns. A depth completion method based on monocular prediction and trained with sparse reconstruction points from structure-from-motion systems can give us the best of two worlds. Moreover, built upon this process, known depth values can be replaced by salient points to regularize and construct high-quality single image depth estimations.

The demand for DNNs design grows significantly with the advent of deep learning. However, this task becomes tedious as human-engineering architecture evolves to be increasingly more complex. Consequently, neural architecture search (NAS) is gaining more momentum. That said, advances in NAS research mainly target image classification (Elsken, Metzen, & Hutter, 2019; Zoph & Le, 2016). Conversely, dense prediction, formulated as per-pixel mapping problems, is more complex and requires complicated architectures to perform well. As a result, most NAS approaches for dense prediction are either computational demanding or deployed only for a single sub-task. Hence, a feasible approach requires 1) a generic backbone based on a well-defined network structure seeking flexibility and layer diversity and 2) an efficient search algorithm to reduce exploration time.

1.2 Scope of the thesis

This thesis utilizes deep learning techniques to construct the base from 3D sensing and builds towards more generalized dense predictions. The approaches presented in Papers I, II, and IV introduce novel methods to create efficient network architectures for monocular depth estimation. Paper I proposes a depth attention volume that exploits the coplanarity constraints from 2D images to improve the performance while significantly reducing the network size. Paper II introduces a self-attention mechanism recursively utilizing salient point detection and the normalized Hessian loss to further shrink the network size. Paper III presents a multi-scale 2D-3D fusion network for depth completion that works well on extremely sparse and unevenly distributed 3D points. Paper IV aims for lightweight single image depth estimation and proposes a framework to automate the architecture design using neural architecture search. Paper V extends NAS for dense prediction problems.
1.3 Contributions

The main contributions of the thesis are summarized below.

- A single image depth estimation method is proposed that achieves state-of-the-art performance. The method exploits the coplanarity constraints that are ubiquitous for indoor scenes and leverages non-local networks to learn long-range dependencies to produce high-quality depth predictions while being 2-8 times more compact than best-performing approaches. (Paper I)

- A monocular depth estimation method is introduced that performs on par with state-of-the-art approaches. Inspired by the depth completion idea, the method replaces known depth measurements with salient points to regularize and construct high-fidelity depth predictions. Additionally, the method is lightweight and uses 6.5% number of parameters compared to the best-performing baseline. (Paper II)

- A depth completion approach is presented that attains high-quality results for indoor and outdoor data. The method constructs a novel multi-scale 2D-3D point fusion neural network architecture, which is more lightweight than state-of-the-art depth completion baselines. The method works well with highly sparse and uneven distributions of input 3D points. (Paper III)

- A neural architecture search framework for lightweight monocular depth estimation is proposed that outperforms best-performing state-of-the-art methods in most cases. The method introduces an efficient search algorithm to significantly reduce exploration time. A multi-objective optimization scheme is leveraged to generate accurate and compact networks. (Paper IV)

- A fast neural architecture search framework for lightweight dense prediction is introduced that yields promising results. The method proposes a generic pre-defined backbone for computational flexibility and layer diversity and leverages a novel search algorithm for efficient neural architecture search. This will enable more generalized dense predictions. (Paper V)
1.4 Summary of the original articles

Paper I presents a single image depth estimation approach that utilizes geometric constraints. The depth attention volume is introduced to capture local and non-local dependencies from depth data based on the coplanarity assumption. The network includes the dilated residual encoder, a non-local depth attention module, and a simple decoder is trained with depth loss and attention loss. The proposed architecture is lightweight yet achieves state-of-the-art results on many indoor datasets such as NYU-Depth-v2, ScanNet, SUN-RGBD, and iBims-1.

Paper II proposes a compact monocular depth estimation targeting applications such as augmented reality, haze removal, or bokeh effect. Stemming from depth completion, the method presents a self-attention mechanism that replaces known 3D measurements with estimated salient points. Moreover, the method integrates the normalized Hessian loss to aid the learning of ambiguous structures from one color image as the term is invariant to scaling and shear along the depth direction. This leads to an efficient model that is suitable for resource-constrained devices.

Paper III boosts depth prediction accuracy by using 3D points as depth guidance. Unlike existing approaches, the method can handle a diverse source of input 3D points from multi-view stereo, structure-from-motion, depth sensors, or mobile AR frameworks. The completed depths are estimated using a multi-scale 3D point fusion network that extracts and fuses appearance and geometric features at different resolutions. Noise from input depth measurements is mitigated utilizing confidence predictors.

Paper IV introduces an automated DNNs design framework for lightweight monocular depth estimation. State-of-the-art depth predictions are mainly attained by using complex networks that are unfeasible for resource-constrained devices. Constructing accurate and lightweight depth estimation models is tedious and demands expert knowledge. The proposed method utilizes an efficient NAS scheme seeking high-performing and compact depth models.

Paper V proposes a fast NAS framework for more generalized dense predictions. The demand for DNNs design is proportional to its increasing complexity. Automating this tedious task is a promising strategy, yet, NAS studies mainly target classification. Dense prediction, in contrast, is more complex. The paper introduces a generic pre-defined backbone based on the multi-scale pyramid network structure, creating search space with computational flexibility and layer diversity. Moreover, the method utilizes the Assisted Tabu Search, enabling fast architecture explorations.
1.5 Outline of the thesis

The rest of the thesis is organized as follows. Chapter 2 describes the background for monocular depth estimation and focuses on building efficient network architectures. Chapter 3 discusses the challenges of depth completion approaches under diverse settings of input 3D points. Chapter 4 works toward a more natural expansion for dense prediction using NAS, complementing to most studies in the literature that were designed for a single sub-task. Chapter 5 presents the conclusions.
2 Single image depth estimation

Single image depth estimation (also known as monocular depth estimation problem) aims to predict a dense depth map for a given RGB image. The term *dense* means that there is a corresponding value encoding distant information in the predicted depth map for each pixel in the input image. As shown in Figure 1, the input RGB image on the left is used to estimate the dense depth map on the right. The colors in the output depth map indicate distances from which blue pixels are close to the camera while red pixels are further away.

![Input RGB Image](image1.png)  ![Predicted depth map](image2.png)

**Fig. 1.** Overview of a monocular depth estimation method.

2.1 Depth from monocular cues

Conventionally, computing depth values using vision techniques requires at least two overlapping images. Depth measurements are estimated by finding 2D correspondences, matching and triangulating to determine the 3D points. However, many different 3D objects can look the same after perspective projection. Predicting depth from just a single image is an ill-posed problem as there exists an infinite amount of 3D forms that can project to the same 2D shape. That said, there are hints such as texture gradient, perspective, line, and occlusion that make things appears 3D. In fact, research reveals that there is valuable information in 2D images that helps the monocular system perceive depth (Gibson, 1950; Kaufman, 1974). These so-called depth cues contain distance information to objects in the scene.

Early works in single image depth estimation revolve around the idea of recovering depth by using computational models to mimic human depth perception. Such cues are illustrated in Figure 2 in which texture gradient is a primary one. The texture gradient of
foreground objects have more clearer details. On the contrary, things that recede into the distance have less and less apparent texture. Bajcsy and Lieberman (1976) proposed a method to calculate relative distances on a ground plane using the Fourier power spectrum deriving from texture gradients. Other studies use the distortion of texture induced by perspective effects to formulate depth (Kanatani & Chou, 1989; Stevens, 1981). However, these complicated methods only work in some conditions that restrict their use in practice.

Shadows are areas that are occluded from the light source. Shading formulates the behavior of reflected light from object surfaces and shadows to determine shape and distance. Several methods aimed to recover this information using additional constraints such as smooth surface and stationary light source (Horn, 1970; Mamassian, Knill, & Kersten, 1998). However, these approaches are sensitive to lighting conditions and usually work in laboratory settings.

Another cue is the linear perspective, which refers to parallel lines that intersect at a vanishing point on the horizon line. Although being parallel, these lines converge in the projected image, suggesting which area is far away from the camera. Criminisi et al. (2000) proposed a method based on vanishing points and lines that takes object segmentation, 3D coordinates of a few points on the ground, and a known object
dimension as inputs. These values were 1) inputted by humans and 2) used to estimate depth by computing the mapping between the 2D image and the 3D ground plane. G. Wang, Hu, Wu, and Tsui (2005) introduced an approach for detecting vanishing points and using those to measure the distance on planes along the lines. However, any slight error in measuring vanishing points will result in significant fallacies in the depth estimation. Additionally, these methods only work on scenes with explicit parallel lines and orthogonal planes.

The relative and familiar size of objects in the scenes can also be used to calculate depth. Given two objects with the same dimensionalities, the further object will be smaller than the closer one. To exploit this property, H.-T. Chen and Liu (2007) categorized everyday objects with similar shapes and sizes, while Gould, Fulton, and Koller (2009) proposed decomposing image pixels into semantically and geometrically related regions for depth prediction. However, these methods need to classify the input image beforehand and often suffer from over-segmentation.

One popular cue is occlusion (also known as intersection) which occurs when an object is being partly blocked or obscured by other objects. In these cases, the occluded object is further away from the camera than the non-occluded one. Dimiccoli, Morel, and Salembier (2008) proposed a method that extracted low and mid-level cues based on occlusion to predict relative depth from a fixed viewpoint. However, their method only works in straightforward scenarios under strict constraints that require objects in the image to have sharp contours.

The aerial perspective suggests that details of distant objects become unclear by atmospheric conditions like haze and fog. In these circumstances, objects at far away distance appear blurred and tinted with blue color from the sky. He, Sun, and Tang (2009) presented a method that exploits aerial perspective to estimate the depth map for image dehazing. However, this approach is only applicable for outdoor scenes that contain haze.

### 2.2 Depth using hand-crafted features

Later works predict depth from a single image based on handcrafted features such as SURF (Bay, Tuytelaars, & Gool, 2006), SIFT (Lowe, 2004). Instead of strictly tied to some monocular cues, these features refer to image characteristics that have been used to best describe the input images. Depth estimation studies based on these features can be categorized into parametric and non-parametric approaches.
2.2.1 Parametric depth

Parametric refers to methods that estimate depth using trained models. Pioneering work proposed by Hoiem, Efros, and Hebert (2005a) reconstructs a 3D virtual scene from just a single RGB image. This method assumes that outdoor environments mainly contain some dominant subjects such as the sky, ground plane, and vertical objects pop up from the ground. This method utilized handcraft features to classify superpixel regions into one of these three classes. Afterward, using the three classes and the mentioned assumption, they automatically create the virtual environment by vertically placing the objects on the ground plane. Since the elements of the inferred scene are very simplified, like a photo pop up from a child’s book, some details are obviously missing. To this end, their approach can be described in three stages. First, an image is over-segmented into uniform regions so that each of which belongs to a particular class described by depth cues. Second, the class of each region is determined using an AdaBoost model. Finally, the 3D model is created by estimating the position of vertical objects with respect to the ground plane. However, since this method assumes a ground-vertical environment structure, it fails if none exists. Nevertheless, as recognized by the authors, the results are smooth and eye-pleasing, as shown in Figure 3.

The approach from A. Saxena, Chung, and Ng (2005) strives to estimate a metric depth map from a given single RGB image. The core idea is to use Markov Random Field (MRF) to estimate absolute and relative depths based on depth features. These features exploit texture gradients, texture variations, and haze information from the image. To achieve this, they split the image into smaller patches. Then, a single depth value is estimated for each patch based on two types of depth features. Absolute depth features are used to determine metric depth, while relative depth features are applied to estimate the depth difference between patches. A 3D laser scanner is utilized to collect
the data for training the MRF model. This model is trained in a supervised manner to estimate the depth from these depth features. The goal is to learn three different sets of parameters for each row in the image, as each row is statistically different from the other. These learned parameters are used to estimate the absolute depth, the confidence, and a smoothing term. The use of the MRF model provides flexibility for integrating several depth features.

In computer graphics, extremely complex 3D scenes are just a collection of triangular surfaces. Based on this principle, A. Saxena, Sun, and Ng (2008) later extended their work by using over-segmentation to obtain superpixels instead of uniform 2D patches. They assume that complicated scenes are comprised of many superpixel planes. Therefore, estimated parameters of these planes can be used to represent the whole scene in 3D. Similar to their previous work, they employed and trained the MRF model in a supervised manner to perform the estimation task using handcrafted features. Additionally, they also utilized features of neighboring superpixel regions to incorporate global information. By doing this, they can reconstruct rather aesthetic attractive 3D scenes from 64.9% random 2D images they downloaded from the internet.

Motivated by previous approaches, B. Liu, Gould, and Koller (2010) proposed a method using semantic information as a means to estimate depth values from images. First, they used a pre-trained semantic segmentation MRF model to label all pixels in the image. Then, they categorized the input image with classes such as sky, tree, road, grass, water, building, mountain, cars, street signs, people, animals, and more. Next, they trained a separated depth predictor for each semantic class. Outputs of these predictors are later incorporated into an MRF model using semantic priors. The goal is to use semantic labels as guidance to predict depth measurements. The final depths are estimated in both pixel and superpixel-wise manner.

2.2.2 Non-parametric depth

Non-parametric refers to methods that can estimate depth values without explicit parameter learning. Karsch, Liu, and Kang (2012) presented a pixel transfer-based approach to predict depth for single images by transferring depth values from the dataset to the input RGB image using a feature descriptor. They argued that pixels with similar representation would also have similar depth values. To this end, they created a dataset containing RGB images and corresponding depth maps using the Kinect-v1. Then, they find k candidate images for every query image. These candidates are obtained by
comparing the similarities of their representation features. Then SIFT was applied to warp the candidate images to the query image. Next, warped depth maps from candidate images are obtained using the warping function that was computed in the previous step. The final result was generated by merging all the warped depths using a global nonlinear optimization process.

Konrad, Wang, and Ishwar (2012) implemented a similar idea to estimate depth from a larger dataset containing millions of images collected from the internet. They also used photometrical similarities to find k-neighbor images but performed depth fusion using median and cross-bilateral depth filters. Moreover, they opted to replace SIFT-warping with stereo rendering to improve computational efficiency.

Later, M. Liu, Salzmann, and He (2014) introduced another non-parametric method that predicts depth by fetching similar images with known depth from a dataset without training. In this case, the nearest neighbor searching was performed using GIST, PHOG, and Object Bank features. However, their approach used both discrete and continuous variables to formulate the relations of homogeneous and adjacent regions. They then approximate the maximum a posteriori to infer the depth map using a discrete-continuous graphical model.

2.3 Depth using deep neural networks

With the advent of deep learning, vision research has been rapidly shifted to using deep neural networks (DNNs) as they are able to learn much more complex and versatile features than their hand-craft counterparts. This trend also includes monocular depth estimation, where DNNs have exhibited impressive results. Figure 4 presented an example of monocular depth estimation methods using hand-crafted features with DNNs that demonstrates the advantage of DNNs approaches.

![Fig. 4. Comparison between different monocular depth estimation strategies.](image-url)
Eigen, Puhrsch, and Fergus (2014) is one of the first works introducing deep learning-based approaches for single image depth estimation. The method employed two DNNs to perform this task. The first network predicts a global depth map from the entire image, while the second network refines the global prediction at a local level. Later on, Eigen and Fergus (2015) extended their work to multi-task learning. F. Liu, Shen, Lin, and Reid (2015) developed a deep convolutional neural field (DCNF) network. The designed DCNF combines the strength of deep CNN and continuous CRFs in a unified framework. However, both Eigen et al. (2014) and F. Liu et al. (2015) used fully connected (FC) layers, which involve a massive number of parameters resulting in expensive computation.

To mitigate the harmful effect of using FC layers, Laina, Rupprecht, Belagiannis, Tombari, and Navab (2016) proposed a fully convolutional residual network (FCRN) for monocular depth estimation. FCRN architecture has an encoder and a decoder. The ResNet-50 was used as the encoder to extract and downsample the feature maps from the input RGB image. The decoder incorporates these feature maps and produces the final depth map. They also demonstrated that the depth of the encoder network dramatically influences the accuracy of depth estimation as deeper networks have a larger receptive field. As a consequence, follow-up studies (Cao, Wu, & Shen, 2017; B. Li, Dai, & He, 2018; Ye, Chen, & Xu, 2021) employed DNNs with hundreds of layers in the encoder to enhance the accuracy of the final depth maps. Hu, Ozay, Zhang, and Okatani (2019) and X. Chen et al. (2019) also exploited multi-scale features and added a refinement module to their networks. Cao et al. (2017) formulated depth estimation as a pixel-wise classification task. In addition, fully connected CRFs are used as a post-processing operation to improve the performance further. B. Li et al. (2018) used side-outputs from the different layers of the encoder network to exploit multi-scale features for depth estimation. The methods mentioned above either build on top of the extremely deep CNNs or use post-processing operations, leading to high computational cost and unable to run in real-time without high-end GPUs.

To accelerate the running speed of monocular depth model, Wofk, Ma, Yang, Karaman, and Sze (2019) designed an encoder-decoder network by using a lightweight CNN (A. G. Howard et al., 2017) as the encoder. Moreover, they utilize network pruning techniques (T.-J. Yang et al., 2018) to further reduce the model size. Although the obtained network has achieved great improvement in speed, the accuracy of depth estimation decreased significantly. Recently, K. Zhou, Wang, and Yang (2020) designed
PADENet for panoramic monocular depth estimation. They trained the PADENet to overcome the lack of an omnidirectional image dataset with ground-truth depth.


Recent depth estimation methods shifted the focus back to learning monocular priors. However, instead of naively imitating human depth perception like prior arts, DNN are employed to effectively exploit cues such as occlusion Ramamonjisoa and Lepetit (2019), planarity constraints explicitly J. H. Lee, Han, Ko, and Suh (2019); C. Liu, Kim, Gu, Furukawa, and Kautz (2019); C. Liu, Yang, Ceylan, Yumer, and Furukawa (2018); Yin et al. (2019). GonzalezBello and Kim (2020) proposed to synthesize the right view from the left view for training from stereo images. G. Yang, Tang, Ding, Sebe, and Ricci (2021) and Ranftl, Bochkovskiy, and Koltun (2021) utilize transformer modules to estimate high-quality depth maps.

Despite recent success, deep neural networks for monocular depth estimation tend to use very deep encoder such as VGG-16 (Eigen & Fergus, 2015), ResNet-50 (Laina et al., 2016; X. Qi et al., 2018; Ramamonjisoa & Lepetit, 2019), ResNet-101 (Fu et al., 2018), ResNext-101 (Yin et al., 2019), SeNet-154 (X. Chen et al., 2019; Hu et al., 2019) followed by some upsampling and fusion strategy including the up-projection module (Laina et al., 2016), multi-scale feature fusion (Hu et al., 2019) or adaptive dense feature fusion (X. Chen et al., 2019) that all result in bulky networks with a large number of parameters. Because high computational complexity and memory requirements limit the use of these networks in practical applications, also fast monocular depth estimation models such as Wofk et al. (2019) have been proposed. However, this further emphasize the trade-off between accuracy and model size as models with great speed comes with the price of reduced accuracy. Moreover, despite promising results achieved with standard benchmark datasets, it still remains questionable if these networks can generalize well to unseen scenes and poses not present in the training data. Instead, this thesis proposes two alternative approaches to effectively address this problem.
2.3.1 Leveraging geometric constraints

Paper I proposed a compact monocular depth estimation method utilizing an attention mechanism called depth-attention volume (DAV) driven by planar structures in the scene. The DAV is integrated to implicitly encourage the network to learn the non-local coplanarity constraints. The network takes an RGB image as input, feeds it to the encoder, the non-local depth attention module, and the decoder to produce a final depth map. The outline of the network architecture is presented in Figure 5.

The core idea of the method is using the coplanarity constraint to guide the depth estimation model. This is because 3D points belonging to the same planar structures are linearly dependent. Therefore, they are good depth predictors of each other. The
DAV is a 4D tensor that utilizes to exploit the coplanarity constraint. It aggregates spatial information of points in both fronto and non-fronto-parallel planes. As shown in Figure 6, the DAV integrates non-local relationships of every point in the image as a set of 2D depth-attention maps.

The architecture of the proposed monocular depth estimation network contains an encoder, a non-local depth attention module, and a decoder. Unlike prior arts that seek to employ increasingly larger DNN for features extraction, the method utilizes more simplified dilated residual networks with 22 layers as the encoder. The dilated residual model learns expressive representations yet only downsample the input image eight times. This is crucial as highly detailed features are directly used by the decoder to produce the final depth map as the spatial resolution of the non-local depth attention module remains the same.

The non-local depth attention module is located between the encoder and the decoder. It is used to learn and add the non-local context from the planar constraints. Its structure consists of 1) the DAV-predictor that is expanded from encoder’s output by implementing cross-denormalization embedding to learn the DAV, and 2) the transformation branch and residual connection to amplify and moderate representation learning. Besides the regular depth losses, the attention loss that formulates the least absolute and angular error between predicted and ground-truth DAV is introduced to aid the training.

The proposed method was trained and evaluated using NYU-Depth-v2 (Silberman, Hoiem, Kohli, & Fergus, 2012) and Scan-Net (Dai et al., 2017) datasets. These are

![Fig. 7. Reconstructed point clouds for randomly selected samples from the SUN-RGBD. Adapted by permission, paper I 2020 © Springer.](image-url)
large-scale indoor datasets in which the former contains more than 120 K RGB+depth images while the latter (from Robot Vision Challenge) consists ∼ 20 K pairs. The ground-truth depth maps were collected using depth sensors (Kinect-v2 and Occipital Structure). The method outperforms state-of-the-art approaches while employing the least amount of parameters. The best performing baselines such as Yin et al. (2019), Hu et al. (2019), and X. Chen et al. (2019) use 4.5, 6.2, and 8.3 times more parameters compared to the proposed approach, respectively. In addition, the method produces clear geometric structures and fine details depth maps.

The generalization ability of the model was also evaluated. The model trained with NYU-Depth-v2 was used to perform the cross-dataset evaluation on the SUN-RGBD (Janoch et al., 2013; S. Song, Lichtenberg, & Xiao, 2015; Xiao, Owens, & Torralba, 2013) dataset without fine-tuning. The method produces well-estimated geometric structures and details depth maps despite the discrepancy in data distributions between the training and testing set. Figure 7 presents the reconstructed point cloud from unseen data on SUN-RGBD. Moreover, extensive planarity analysis on the iBims-1 (Koch, Liebel, Fraundorfer, & Körner, 2018) benchmark shows that the proposed method also outperforms recent works (C. Liu et al., 2018; Ramamonjisoa & Lepetit, 2019).

2.3.2 Salient point detection with normalized Hessian loss

Paper II proposed a high-performance and lightweight monocular depth estimation method called FuSaNet. It has been shown that although achieving great results, deep learning-based single image depth estimation approaches exhibit an apparent compromise between accuracy and network size. Best-performing methods rely heavily on expensive network architectures and massive training data, while lightweight approaches struggle to produce accurate predictions. This issue is more prominent in applications for resource-constrained platforms such as mobile phones or embedded devices.

![Fig. 8. Framework that utilizes self-attention auto-regressive mechanism to improve depth estimation. Adapted by permission, paper II 2021 © IEEE.](image-url)
An alternative to 3D sensing is depth completion, which aims to densify a sparse depth map. These sparse measurements regularize nearby depth values, enabling high accuracy depth prediction with considerably smaller networks. However, depth completion requires additional data obtained with active sensors such as LiDARs or ToF cameras. Inspired by the depth completion hypothesis, the method replaces known depth measurements with salient points to regularize the estimated depth map. A major benefit compared to depth completion is that these salient points are determined from monocular RGB images while still providing similar advantages as the additional depth data. In this context, salient points are assumed to learn depth cues containing local and global structures, assisted by the provided ground truth depth measurements during training. Thus, it can also be considered to be a self-attention mechanism. For this purpose, the method trains confidence predictors to highlight important keypoint positions from an RGB image as a confidence map. This map is then used to generate salient points where predicted depth values tend to be more accurate. These points are utilized to enhance the performance of our network, similar to depth completion methods.

As shown in Figure 8, the overall architecture of the FuSaNet model consists of the Fusion-Net and the Saliency-Net. The Fusion-Net is a fully convolutional framework. Its inner structure is assembled by Fusion-Blocks that is used to extract and fuse 2D and 3D features. Each Fusion-Block can be considered as a mini-network operating at a particular spatial scale. It contains a feature fusion encoder, a decoder, a refinement, and a confidence predictor module. The Fusion-Net uses the RGB image and sparse 3D points (only during training) to produce a dense depth map, a confidence map, and a fusion features tensor.

The Saliency-Net utilizes the outputs of the Fusion-Net to detect a set of salient points that iterates through the Fusion-Net one more time to produce the final depth prediction.

Fig. 9. The effect of the salient points. Points are enhanced for visualization. Adapted by permission, paper II 2021 © IEEE.
The main component of the Saliency-Net is the multi-scale feature extraction layer that contains two 2D convolutional layers and seven different 2D atrous convolutions layers. The goal is to capture features at various spatial resolutions and utilize those to best capture salient points in versatile contexts. Figure 9 provides examples of the predicted confidence map, detected salient points, and their effects on the final depth maps. In addition, the method introduced the use of the normalized Hessian loss to deal with the generalized bas-relief transformation during training. This invariance is motivated by the inherent depth scale ambiguity, which is not penalized by the normalized Hessian loss function, unlike the other loss terms.

The proposed method was evaluated on NYU-Depth-v2, KITTI (Geiger, Lenz, Stiller, & Urtasun, 2013), and SUN-RGBD datasets. In the case of NYU-Depth-v2, the method performs on par with state-of-the-art methods while utilizing 38.4, 25.9, 15.3 times fewer parameters than G. Yang et al. (2021), X. Chen et al. (2019), Ranftl et al. (2021), respectively. For KITTI, the method outperforms best-performing methods while being at least 13.5 times more compact. Moreover, the method seems to generalize well to SUN-RGBD and in-the-wild data.

2.3.3 Discussion

The monocular depth estimation methods presented in Paper I and Paper II have many applications, including occlusion handling in augmented reality and haze removal or bokeh effect in mobile photography where dense depth is needed. Unlike state-of-the-art methods (X. Chen et al., 2019; Fu et al., 2018; Hu et al., 2019; Ranftl et al., 2021; G. Yang et al., 2021) that demand large network architectures, the proposed methods yield high-quality depth maps while being much more compact than the best-performing approaches. As a consequence, the methods are more suitable for resource-constrained hardware such as mobile phones or embedded devices. Moreover, monocular methods require only a single camera. Hence, it is more appealing with a broader range of applications compared to stereo approaches.

Estimated depth maps from the methods can also be implemented for a real-time reconstruction application for mobile devices. This is feasible as modern frameworks on mobile phones such as ARKit (Apple, Release 01 June 2017), ARCore (Google, Released 01 March 2018), AREngine (Huawei, Released 30 June 2019) can provide camera poses for the reconstruction process. Moreover, compared to SfM (Schonberger & Frahm, 2016; C. Wu, 2011) results that do not provide the scale or other monocular
methods (J.-H. Lee & Kim, 2019; B. Liu et al., 2010) that can estimate only relative distances, Paper I and Paper II methods provide metric units that can be easily combined with the reconstructed models.

Planar structures are ubiquitous, especially for indoor and man-made scenes. The current implementation of Paper I only implicitly exploits the coplanarity constraints for depth prediction. However, further analysis utilizing the attention volume to extract information such as surface normal at inference time can benefit many practical applications.

Relying on ever-expanding deep neural architectures and massive training data for accuracy is a comfortable yet unsustainable practice. Paper II demonstrates the regularizing power of input 3D points. They give valuable guidance to train the salient point detection. The method provides a potential approach for creating accurate and compact monocular depth estimation models without the need for using active depth sensors or multiple view geometry.
3 Depth completion

Depth completion is a part of the 3D sensing problem that aims to fill in missing pixels of a given incomplete depth image, in which only a subset of pixels contains depth measurements. The remaining pixels without valid depth measurements have values of zeros. Typically, these incomplete depth images are outputs from depth sensors, simultaneous localization and mapping or structure-from-motion systems. Figure 10 illustrates a common depth completion approach that uses an incomplete depth map, and optionally an RGB image, to produce a dense depth map.

Depth completion has diverse input settings, leading to other terminologies such as depth inpainting or depth enhancement. Nevertheless, they share the same goal and are used interchangeably in some contexts. In fact, there are many ways to categorize this problem. On the one hand, a depth completion method can perform with or without color images for guidance based on its input modalities. On the other hand, depending on the input depth patterns, it could operate using a randomly uniform or spatially unevenly distributed set of 3D points. Finally, it could take in a variety of input depth densities from dense to extremely sparse 3D measurements. However, since the main challenge of the depth completion problem depends on the density of the input 3D points, this chapter is organized following the latest categorical order.

Fig. 10. Depth completion methods take an incomplete depth image and RGB image (optional) as inputs.
3.1 High-density input

High-density refers to cases where the incompletely input depth to the system having a tangible amount of pixels has already been measured where typically $\sim 20\%$ or fewer pixels are missing. These missing values are usually caused by hardware limitations, especially on transparent or specular surfaces from commodity-grade depth sensors such as Kinect-v1 (active stereo with infrared structure light), Kinect-v2, Kinect-Azure, various smartphone depth sensors (time-of-flight). Approaches toward this problem act as a post-processing task in 3D scanning, where holes and noisy measurement signals from 3D sensors need to be corrected.

As the number of available distance measurements is relatively large, Gaussian kernels could be applied to fill in the missing holes for these input depths. However, such simple methods usually produce flawed predictions when the filter fails to captivate the statistic of the input signal. Instead, Berdnikov and Vatolin (2011) dug deeper into the signal statistics and discovered that missing values are typically caused by edges of foreground objects and reflective surfaces. They then proposed a real-time algorithm that separated and filled the holes based on the types of artifacts. Although this method performs well to some extent, it forfeits valuable information such as RGB and temporal consistency between frames.

Milani and Calvagno (2012) introduced a depth completion method utilizing both RGB and depth information in which these inputs were first re-aligned by matching the object borders. Next, object segmentation was applied to the color image for partitioning depth regions where, later on, holes and gaps were filled under the assumption that depth values in these areas were smooth. However, the proposed approach is sensitive to segmentation outputs and tends to suffer where object textures are cluttered and vibrant color. Following this line of work, Schmeing and Jiang (2012) suggested using RGB and depth images to fix corrupted edges in the depth maps by partitioning color data using superpixel-segmentation approaches. The algorithm then corrected the edges of input depths based on this information.

Camplani and Salgado (2012) and Barron and Poole (2016) proposed approaches using a joint-bilateral filter that takes into account spatial as well as temporal information. The bilateral filter is a non-linear filter capable of retaining sharp edges while smoothing noise from the input measurements in which a weighted average of neighboring pixels supersedes the value at a pixel. The weights of the filters are determined using the color.
Fig. 11. An example case for high density input. Predicted depth is obtained using Senushkin, Romanov, Belikov, Patakin, and Konushin (2020) method.

image, the input depth map, and the temporal consistency map. The depth information is iteratively updated by processing the RGBD video stream.

The inpainting technique proposed by F. Qi, Han, Wang, Shi, and Li (2013) aggregates color structure information to fill and correct depth values around discontinuity areas. Unlike previous approaches, they introduced a fusion scheme using a non-local means filtering strategy based on image textures to enhance depth completion accuracy. Moreover, as the color and depth cameras are calibrated for pixel alignment, the method can provide good predictions even around textureless regions.

Lu, Ren, and Liu (2014) also argued that because depth discontinuities are strongly co-aligned with their color counterparts, combining intensity and depth maps is the sensible solution. They proposed a method to construct a matrix from equivalent RGBD patches and then use its low rank to enhance the depth completion accuracy. However, the proposed algorithm encounters difficulties when recovering large regions as local limitations of the patch-based technique.

Senushkin et al. (2020), introduced a method called saic, that is one of the recent depth completion approaches using deep neural networks. They employed a standard encoder structure and integrated the spatially-adaptive denormalization (T. Park, Liu, Wang, & Zhu, 2019) into the decoder to obtain sharp object edges and smooth details. This approach yields relatively good performance in many scenarios, especially for indoor scenes as shown in Figure 11.

3.2 Sparse input

The depth completion problem is considered as sparse when the number of input depth measurements is 7% – 4% of pixels that typically comes from Velodyne LiDAR or similar sensors. In these scenarios, the amount of input 3D points is small, rendering a
significantly more challenging task compared to high-density cases. Consequently, earlier methods developed for high-density input tend to produce worse results leading to a new set of approaches that can be categorized as direct depth completion and spatial propagation techniques.

3.2.1 Direct depth completion

In general, direct depth completion methods for sparse input employ modalities such as depth and color images to predict the dense depth map in a single pass. Hawe, Kleinsteuber, and Diepold (2011) proposed an approach for estimating dense disparity maps from sparse measurements using wavelet analysis. Nevertheless, learning-based methods often have the advantage in this difficult circumstance as they can effectively fill in large chunks of missing data from prior knowledge. As one of the pioneering studies, Diebel and Thrun (2005) suggested exploiting the color stream from RGB cameras. Similar to previous works, this method is also based on the assumption that depth and RGB images tend to correlate. However, they leveraged much higher resolution color images and the multi-scales Markov Random Field to densify the low-resolution input depth.

The 3D information is sparse and irregular by nature. Despite the success in image representation learning tasks such as classification or object detection, convolutional neural networks are designed to function on a regular grid. To overcome this problem, Uhrig et al. (2017) proposed a sparse convolutional layer that takes into account the validity of pixels during the convolving operation. The proposed approach is simple, invariant to the sparsity level of input depth measurements, yet produces promising results. On the contrary, Jaritz, De Charette, Wirbel, Perrotton, and Nashashibi (2018) claimed the same goal could be achieved without the validity mask. Their analysis pointed out that validity masks are saturated after just a couple of first layers, and more importantly, the network can still learn the sparsity invariant without the masks. They proposed a method that processes the RGB image and sparse depth map separately and later concatenated these output volumes before feeding them to the decoder. In addition, this method can also produce semantic segmentation maps.

F. Ma and Karaman (2018) took a more straightforward approach where they concatenated the color and sparse depth images before putting them through different backbone networks for KITTI (ResNet-18) and NYU-Depth-v2 (ResNet-50). They experimented with several types of decoders and also did not employ the validity mask.
in their system. In this work, the effect of the number of depth samples on performance was also measured to some extent.

Applying geometric constraints is a promising approach for enhancing depth estimation accuracy. Such a method, proposed by Xu et al. (2019), utilizes depth and surface normal to construct an anisotropic diffusion block to exploit the piece-wise planar assumption of 3D scenes. Moreover, they incorporated a confidence estimation stream to mitigate the effect of noisy measurements. This method gives consistent predictions in large and small areas compared to prior arts.

Accurately predicting missing depth values between objects is a more complex problem, especially around depth discontinuity regions, as these pixels neither belong to the foreground nor background. They are also called mixed-depth pixels. Although appearing insignificant, resolving this produces high-quality depth estimation with well-rounded object boundaries and potentially enhances the accuracy of downstream tasks such as pose estimation. To achieve this, Imran, Long, Liu, and Morris (2019) divided depth into bins and incorporated a probabilistic depth loss demonstrating the efficacy for indoor and outdoor data.

Despite the progress, depth completion for sparse input remains unsolved mostly because capturing 3D geometric information from sparse data using standard 2D convolution is problematic. The connectivities between appearance features and 3D features are largely different. Y. Chen, Yang, Liang, and Urtasun (2019) proposed a method that extracts and fuses these features resulting in a high-performing and compact network architecture.

### 3.2.2 Spatial propagation

Unlike direct depth completion methods, spatial propagation techniques iteratively densify the sparse input utilizing learned guidance from the color stream. The original idea (S. Liu et al., 2017) is to use the spatial propagation networks to learn the relationships of pixel pairs for vision tasks such as semantic segmentation and depth completion. These learned affinities obtained from the RGB image are linearly propagated to refine the coarse prediction. However, the propagation directions were limited to bi-directional vertical and horizontal neighbors.

Inspired by this idea, Cheng, Wang, and Yang (2018, 2019) proposed the convolutional spatial propagation networks to diffuse and learn the affinity values for all local neighboring pixels. Learned affinities were attained by 1) scan-line and
scan-column pixels from the convolutional context and 2) other neighbors using the
recurrent convolution scheme. These procedures are operated simultaneously, resulting
in a more efficient and accurate depth completion approach. Nonetheless, the network
tends to produce mixed-depth predictions as the learning is carried out only in the local
vicinity.

It is well known that vanilla 2D convolutions fall short when capturing long-
rangle dependences. To overcome this, X. Wang et al. (2018) proposed a deep neural
architecture using the non-local block that can extract both local and non-local affinities.
Combining the idea from X. Wang et al. (2018) and Cheng et al. (2018), J. Park et al.
(2020) introduced a non-local spatial propagation network capable of handling mixed
depth estimation as well as producing high-quality depth maps.

3.3 Extremely sparse input

The depth completion is deemed extremely sparse if the amount of input depth pixels is
less than 0.5%. Compared to high-density or sparse, these cases are inextricable yet
invaluable because such inputs can be easily acquired using low-cost depth sensors,
structure from motion systems, or simultaneous localization and mapping pipelines.
Additionally, their performances are superior to monocular depth methods. In fact, they
can be rendered as general depth completion approaches that perform well in diverse
sparsity settings.

One such method addressing this problem is Qiu et al. (2019), in which they proposed
a deep neural network to predict dense depth from the RGB image and sparse LiDAR
measurements. Their network produces the intermediate confidence mask, surface
normal, and attention map from these inputs. The confidence mask was employed to
handle noise and occlusion from LiDAR signals. The surface normal is easier to predict,
that later is incorporated with the confidence mask to refine the final depth map. The
attention map is utilized to integrate estimated depth from the color and surface normal
pathways. Although generating good results, this method only worked with uniformly
distributed measurements from commodity depth sensors and was not designed for other
types of 3D data coming from multi-view stereo or structure from motion.
3.3.1 Multi-scale 2D-3D fusion network

Paper III introduced a framework called Point-Fusion that, starting from single image depth estimation, utilizing a set of 3D points as constraints to obtain high-quality depth maps. Point-Fusion is a fully convolutional framework that uses an RGB image and sparse 3D points as inputs to estimate a dense depth map. Its structure is presented in Figure 12. The 3D points are first projected to the image plane, and their $z-$coordinates are used to create a sparse depth map. The core component of Point-Fusion contains five Fusion-Nets that extract and fuse 2D and 3D features at multiple scales before predicting the final depth map at the highest spatial resolution. The network was trained by optimizing the predicted depth map at multiple scales.

Each Fusion-Net has a feature fusion encoder, a confidence predictor, a decoder, and a refinement module. The feature fusion encoder includes 2D convolutions in 2D branches and the feature-kernel alignment convolution in the 3D branch. The encoder generates a 3D volume by fusing the features from the 2D and 3D branches. The decoder then transforms the fused features before feeding them to the refinement module.
The confidence predictor finally modifies learned representations from the decoder to mitigate the effect of input noise.

The proposed method was evaluated on the NYU-depth-v2 and KITTI datasets. It achieves state-of-the-art results while being more compact than baseline methods regarding the number of parameters. Observations show that the model performs well while baseline methods tend to suffer in the case of extremely sparse and unevenly distributed 3D point inputs.

Figure 13 presents an example using 32 input 3D points on the KITTI dataset. Error maps on the right side show that Point-Fusion produces less error compared to state-of-the-art approaches. Additionally, its predicted depth maps are also more consistent and less noisy than the best-performing methods.

Figure 14 depicts the predicted depths and reconstructed point clouds from NYU-Depth-v2 data. The sparse 3D points are extracted using a real-time stereo system. The dense depth maps were obtained with this setting for all methods. Typically, depth completion approaches are sensitive in extreme sparsity cases, while deep multi-view stereo performance degrades with fewer input views. On the other hand, Point-Fusion can generate high-quality depth predictions with much less distortion.
3.3.2 Discussion

The main advantage of Point-Fusion is that it is agnostic to the sparsity, distribution patterns, and the source of the input 3D points. Unlike previous depth completion approaches, the method proposed in Paper III works well with extremely sparse and unevenly sampled point clouds. Moreover, the model can take in the 3D points that are generated from either conventional multi-view stereo, ranging sensors such as LiDARs, or points from SLAM systems such as ARKit (Apple, Release 01 June 2017), ARCore (Google, Released 01 March 2018), and AREngine (Huawei, Released 30 June 2019).
Assessing the performance of the proposed model on unseen data shows some promising results. The ARCore framework is utilized to collect the RGB frames and extract the 3D points that were fed as inputs to models. The results in Figure 15 show that the predicted depth maps from Point-Fusion yield higher quality and are more consistent than state-of-the-art approach.

Point-Fusion is considerably more lightweight than most depth completion approaches, especially for indoor scenarios. The proposed method utilizes only 33% number of parameters compared to the best-performing baseline. Using a simple stereo reconstruction system, the whole pipeline runs at 197 fps for input frames of 640 × 480 pixels using one GTX-1080 GPU. With further optimization, Point-Fusion can operate in real-time on handheld devices.

As the approach introduced in Paper III functions well with just a few input 3D points, it is a natural extension to examine its performance using even fewer input depth, for example, utilizing Point-Fusion in settings where input data is obtained from a system with a very cheap laser depth measurement tool and an RGB camera.
This is intriguing as such frameworks can potentially be applied for various practical applications.

Additionally, experiments suggest that Point-Fusion performs reasonably well as a monocular depth estimation method without the input 3D points. This property can be utilized to significantly shrink the network size of monocular approaches while maintaining desirable accuracy. On top of this, better designs for the encoder and confidence predictor in the Fusion-Net module can potentially lead to state-of-the-art performance.
4 **Towards generalized dense prediction**

Recent years have witnessed the thriving success of deep learning across many disciplines in which features are no longer hand-crafted but instead learned end-to-end from data. As a result, the demand for neural network architectural design has skyrocketed. Recently, neural architecture search (NAS) has accumulated significant momentum as human-engineering of deep networks becomes increasingly complex. Possessing such ability propels NAS as a promising generalized approach. Nonetheless, advances in NAS research have mainly been proposed for image classification. However, dense prediction, formulated as per-pixel mapping problems, is more complex and requires complicated architectures to perform well. Consequently, most NAS approaches for dense prediction are either computational demanding or deployed only for a single subtask. Hence, solutions toward generalized dense prediction prove challenging and remain largely unsolved. In this chapter, section 4.1 gives an overview of the current development of neural architecture search, while section 4.2 provides NAS context for the dense prediction problem.

### 4.1 Neural architecture search

Neural architecture search is a part of automated machine learning that aims to replace the manual architectural design process with learned deep neural networks from the input data. As shown in Figure 16, NAS can be seen as a recursive optimization approach that utilizes the search algorithm to find an optimized architecture from the search space by applying the evaluation strategy to predict the performance of a network candidate. Generally, NAS methods can be constructed or categorized based on three

![Fig. 16. Overview of neural architecture search approach. The search algorithm finds an optimized architecture from the search space by utilizing the evaluation strategy to predict the performance of a network candidate.](image-url)
main components: the search space, search algorithm, and evaluation strategy (Elsken et al., 2019, 2022).

### 4.1.1 Search space

The search space contains all feasible architectures that can be generated from searchable elements such as types of network building blocks, types of the layer in a block, number of input and output channels, kernel sizes, types of connections, number of blocks inside a module, and network hyperparameters. Moreover, a neural architecture search method can also search for high-level network structures, including encoder and decoder anatomy, number of resolution scales, and long-range connections.

There are two common strategies to create the search space, emphasizing the balance between efficiency and diversity. One can construct a highly generalized exploration space by searching for basic mathematical operations to build completely new architecture from scratch (Real, Liang, So, & Le, 2020). However, this immensely increases the search space size hindering real-world practicality.

On the other hand, priors knowledge about the problem can help determine a more effective search space. For image classification, this type of network only requires the encoder that consists of some primal operations such as convolutional and pooling layers. In fact, methods proposed by Zoph, Vasudevan, Shlens, and Le (2018) and Zhong, Yan, Wu, Shao, and Liu (2018) only search for the intricate cells that later were repeatedly stacked up to construct the whole architecture. Although this simplifies the search process, it also dramatically restricts layer diversity because searching for new connectivities on the higher level was forfeited. As a result, such methods tend to fall into the narrow pitfall where architectural and performance differences between network candidates are insignificant (A. Yang, Esperança, & Carlucci, 2019).

In contrast, because NAS solutions for dense prediction tasks tend to be more expensive, standard approaches systematically scale down the exploration space and search for optimized hyperparameters such as the number of channels and kernel sizes. The structure of the encoder and decoder is borrowed from best-performing methods, including MobileNet-v3 (A. Howard et al., 2019), ShuffleNet-v2 (N. Ma, Zhang, Zheng, & Sun, 2018), FPN (T.-Y. Lin et al., 2017), PSPNet (Zhao, Shi, Qi, Wang, & Jia, 2017), and PANet (S. Liu, Qi, Qin, Shi, & Jia, 2018).
4.1.2 Search algorithm

A neural architecture search algorithm is utilized to manage the exploration process and spawn a set of candidate architectures. These models, after some form of training, will be evaluated to obtain the reward signals for optimizing and selecting the best-performing architectures. There are several common search algorithms, in which random search (Bergstra & Bengio, 2012) being the most straightforward one. Albeit being simple, this algorithm can achieve good results in a well-defined search space.

Another search strategy is reinforcement learning Baker, Gupta, Naik, and Raskar (2016); Zhong et al. (2018); Zoph and Le (2016); Zoph et al. (2018), which typically requires a controller to generate new child architectures. This controller is a recurrent neural network inducing configuration sequences that use to define the network structure.

Stemming from the genetic algorithm, neuron evolution is utilized to evolve network architecture. The evolution algorithm (Elsken, Metzen, & Hutter, 2018; Real et al., 2017; Stanley & Miikkulainen, 2002) mutates a model by: 1) modifying network structure and connection weights and 2) performing cross-over between parent nodes.

Another favorable search technique is Bayesian optimization (Kandasamy, Neiswanger, Schneider, Poczos, & Xing, 2018; Mendoza, Klein, Feurer, Springenberg, & Hutter, 2016; Ru, Wan, Dong, & Osborne, 2020; Swersky, Duvenaud, Snoek, Hutter, & Osborne, 2014; White, Neiswanger, & Savani, 2021) which operates based on a prior distribution function like Gaussian processes (Seeger, 2004). As the search procedure is bound to some approximations, the evaluation cost is reduced, leading to a more efficient search method.

Recently, H. Liu, Simonyan, and Yang (2019) introduced a continuous relaxation condition to update both architecture parameters and weights using gradient-based optimization. Instead of making discrete choices from the operations pool, they relax (or weighted) the ordinal decision of operations using the softmax. This, in turn, dramatically decreases the validation cost.
4.1.3 Evaluation strategy

An evaluation strategy is implemented to compute or estimate the performance of candidate architectures. These metrics are vital response signals, enabling the search algorithm to update and find optimal network designs. The standard protocol is obtaining the evaluation results from the fully trained candidates (Zoph & Le, 2016), but the expense is excessive. To mitigate this problem, Mishkin, Sergievskiy, and Matas (2017) and Zela, Klein, Falkner, and Hutter (2018) reduced the number of training epochs while Vodrahalli, Li, and Malik (2018) suggested performing the evaluation on a selective subsample of the original dataset. Chrabaszcz, Loshchilov, and Hutter (2017) proposed to perform NAS on ImageNet variations where samples are downsized to $16 \times 16$, $32 \times 32$, and $64 \times 64$. Zoph et al. (2018) validated down-scale networks in the search stage. D. Zhou et al. (2020) introduced a method to carefully rank and design an adaptive proxy to decrease the exploration time.

Another common strategy is parameter sharing, in which instead of training thousand of models separately, the weights are re-used during the search process. S. Saxena and Verbeek (2016) proposed an embedding fabric to cover a large amount of candidate architectures. An individual model can be considered a fabric path, and the weights are shared via overlapping routes. Cai, Chen, Zhang, Yu, and Wang (2018) used a meta-controller to predict efficient network transformation and build new architectures on top of inherited weights of descendants networks. Pham, Guan, Zoph, Le, and Dean (2018) introduced an efficient NAS approach by applying weight sharing among the child architecture in which the search space and candidate model are defined as super-graph and sub-graph.

Recent NAS studies strive to predict the performance of network architectures to further reduce the exploration expense. Baker, Gupta, Raskar, and Naik (2017) utilized the architectures, hyperparameters, and validation accuracies of partially trained networks to predict their learning curves. Brock, Lim, Ritchie, and Weston (2017) trained a buffer network (Ha, Dai, & Le, 2016) and utilized it to generate the weights for the offspring architectures. By doing so, generated models can be evaluated directly. Wen et al. (2020) proposed the idea of randomly generating, training, and evaluating a set of architectures, then utilizing this data to train a performance predictor. Next, the trained predictor is applied to immediately estimate the accuracy of new architectures. Mellor, Turner, Storkey, and Crowley (2021) introduced a scoring technique based on the initial weights and input data to predict network performance without training.
4.2 Dense prediction

Dense prediction is a long-standing problem in computer vision that can be defined as a mapping function

\[ f_D : I \rightarrow V, \]  

where \( f_D \) is the function that map every pixel in the input image (\( I \)) with some estimated values (\( V \)). These values can be continuous or discrete depending on the subtasks. For instance, semantic segmentation is a dense classification discrete task, whilst monocular depth estimation, surface normal prediction, disparity estimation and image super-resolution are often formulated as regression. Dense prediction is a prevalent research topic with various applications such as augmented reality, scene understanding, robotics, medical and surveillance imaging, and image restoration.

4.2.1 Semantic segmentation

Semantic segmentation is a dense prediction task aiming to assign class labels for every pixel from a given image. Over the years, substantial efforts have been concentrated on tackling this problem as this is a crucial component for visual scene understanding–recent NAS-based approaches for semantic segmentation yield noticeable results.

One pioneering study introduced by C. Liu, Chen, et al. (2019), namely Auto-DeepLab, modifies DARTS framework (H. Liu et al., 2019) for semantic segmentation. Unlike DARTS, they not only search for an optimized cell but also seek for the optimal spatial resolution of the feature maps. The proposed method applied the continuous relaxation technique to select the most suitable spatial resolution as well as components inside the cells. They then feed these multi-scales outputs through ASPP modules (L.-C. Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2017) resulting in architectures that can operate at multiple resolutions. Inspired by this, W. Chen et al. (2019), Shaw, Hunter, Landola, and Sidhu (2019), and P. Lin et al. (2020) search for lightweight architectures by enforcing the hardware constraints. W. Chen et al. (2019) suggested updating the search space with channel expansion ratios and multiple branches network. Shaw et al. (2019) replaced vanilla convolutional layers with inverted residual blocks. Favoring proxyless search, X. Zhang et al. (2021) proposed a densely connected search space with path-level and channel-level sampling to reduce memory consumption. Weng, Zhou, Li, and Qiu (2019) introduced a method called NAS-Unet that utilized ProxylessNAS (Cai,
Zhu, & Han, 2018) to search for down-sampling and up-sampling cells based on a pre-defined Unet backbone. Q. Yu et al. (2020) leveraged the Auto-DeepLab search space with an evolution algorithm looking for macro modules to generate the architecture. That being said, cell structures inside a macro are also randomly sampled following L. Li and Talwalkar (2020). Zhu, Liu, Yang, Yuille, and Xu (2019) employed DARTS to search for optimal encoder-decoder architectures.

Built upon a fixed network structure, L.-C. Chen et al. (2018) introduced a dense prediction cell to learn multi-scale features by searching for optimized dilated separable convolution and average spatial pyramid pooling modules. The dense prediction cell, consisting of many branches, is searched using an integrating random exploration with various proxies. Despite these efforts, the method still demand thousand of GPU hours to perform the task. Following this line of work, Nekrasov, Chen, Shen, and Reid (2019) employed a fixed encoder and long-range connections to search for the best possible decoder. They opt to use a recurrent controller to 1) seek for optimal and auxiliary cells to construct the decoder and 2) apply proxies such as early-stopping, freeze encoder weights, and knowledge distillation. They later expanded this work for semantic video segmentation using an LSTM controller to search for dynamic cells that can learn temporal information from neighboring frames (Nekrasov, Chen, Shen, & Reid, 2020).

On the other hand, Y. Zhang et al. (2019) proposed a Customizable Architecture Search (CAS) method aiming for both encoder and decoder modules. First, they search for the best-performing normal cells and reduction cells. Then, CAS searches for a multi-scale cell to carefully fuse learned features from multiple resolutions, taking into account accuracy and network expenditure. Considering fixed encoder and decoder structures, H. Wu, Zhang, and Huang (2019) focus on searching for the optimized connectivity between these modules. Starting from architecture with densely weighted connections, the proposed method seeks for the sparsest connectivity patterns that yield the best performance.

4.2.2 Image super-resolution

Single image super-resolution (SISR) is a low-level vision task that generates a high-resolution image from its low-resolution counterpart. NAS-based approaches for this problem are typically based on studies proposed for image classification. D. Song et al. (2020) suggested searching for well-defined residual dense blocks for stacking the new networks. Instead, the method proposed by Guo, Luo, He, Huang, and Chen (2020)
operates at a micro-level to search for normal and upsampling cells. Chu, Zhang, Ma, Xu, and Li (2021) applies reinforcement learning and evolutionary methods with an elastic search space to find lightweight SISR architectures considering PSNR, FLOPs and memory in the optimization process. R. Lee et al. (2020) introduced a platform-agnostic to search for tiny architectures, with optimizing the perceptual quality, that can be fitted on embedded devices. Esmaeilzehi, Ahmad, and Swamy (2021) proposed constructing a super-block by stacking up dense residual blocks and integrating the squeeze and excitation module for lightweight single image super-resolution.

4.2.3 Disparity estimation

Disparity estimation refers to getting the apparent pixel motion between frames. Although this requires at least two images during training, the inference can usually work with a single image. Saikia, Marrakchi, Zela, Hutter, and Brox (2019) is one of the first studies using NAS for disparity prediction. Similar to previous subtasks, this method, namely AutoDisp-Net, expands the search space by adding the upsampling cell together with the normal and reduction cell to build encoder-decoder-like networks. Moreover, they use the first-order DARTS and Bayesian optimization (Falkner, Klein, & Hutter, 2018) for efficient exploration with possible extensions to depth estimation. AutoDisp-Net still consumes thousands of GPU hours to search for the optimized architectures despite these measures.

![Fig. 17. The search space of our LiDNAS framework. Models are constructed from a pre-defined backbone network containing encoder, decoder, refine, downsample and upsample blocks (green). A block is formed by several identical layers (orange) that are generated from a pool of operations and connections. Layers within a block are the same while layers of different blocks can be different. Adapted by permission, paper IV 2022 © IEEE.](image-url)
4.2.4 Monocular depth estimation

Paper IV proposed a framework for lightweight monocular depth using neural architecture search called LiDNAS. The proposed method uses a multi-objective exploration framework to search for both accurate and lightweight monocular depth estimation architectures. Paper IV also introduces a novel search algorithm, namely Assisted Tabu Search, for fast neural architecture search. Additionally, the method builds upon a well-defined search space that allows computational flexibility and layer diversity.

As shown in Figure 17, the search space was built based on a pre-defined backbone in which the green blocks are searched for. The pre-defined network was divided into multi-scale pyramid networks operating at different spatial resolutions. Each network scale consists of two encoder blocks, two decoder blocks, a refine block, a downsample and a upsample block. Each block is constructed from a set of identical layers, as shown

Fig. 18. Comparison on KITTI (left) and NYU-Depth-v2 (right). (a) input image, (b) FastDepth, (c) DSNet, (d) PyD-Net, (e) LiDNAS-K, (f) Ef+FBNet, (g) ground truth, and (h) LiDNAS-S. Images in the right column presented zoom-in view for better visualization. Adapted by permission, paper IV 2022 © IEEE.
in orange here. The kernel size and number of output channels are balanced to keep the complexity in control.

Compared to state-of-the-art NAS method (Saikia et al., 2019), LiDNAS is $\times 9$ times faster while generated network (LiDNAS-N) improves REL and RMSE by 22.3% and 18.7% using only 3% of AutoDisp-Net’s model parameters. Moreover, examples in Figure 18 show that the predicted depth maps generated by LiDNAS models are more accurate and contain much less artifacts.

Experiments with different searching scenarios found that alpha close to 0.6 tends to generate both lightweight and good accuracy architectures. Further comparison between TS and ATS shows that ATS runs $\sim 7$ times faster while producing similar results to the original TS algorithm.

4.2.5 Generalized dense prediction

Extended from LiDNAS, paper V introduces a novel framework for fast neural architecture search for lightweight dense prediction networks referred to as LDP. The idea is to employ a well-defined generic backbone, and apply the novel Assisted Tabu Search for efficient architecture exploration. Unlike previous NAS approaches that are either excessively costly or work only for a single subtask, LDP is fast and suitable for a diverse range of dense prediction problems. Figure 19 presents a comparison between LDP models with best-performing baselines for single image depth estimation, semantic segmentation, and image super-resolution.

Fig. 19. From left to right, absolute relative error, mean intersection-over-union and peak signal-to-noise ratio vs. the number of parameters for recent dense prediction methods on NYU-Depth-v2 (left), Cityscapes (middle) and Urban100 (right) – the LDP models outperforms the lightweight baselines (black), while using substantially less parameters than the current state-of-the-art methods (blue). Compared to the recent NAS-based approaches (red), LDP improves in both performance and compactness. Adapted by permission, paper V 2022 © IEEE.
To achieve this, LDP searches for architectures with maximal validation grade $G$, as shown in Figure 20. This grade is formulated from the validation accuracy and the compactness of model $m$. Additionally, LDP leverages a method by Mellor et al. (2021) to predict network performance before training and integrate it into the mutation exploration reward to significantly reduce the search time.

To search for good candidate architectures, LDP first randomly generate 60 K parent models and rank them based on their initial scores. Next, it selects six architectures in which three are the highest-ranked while the other three have the highest score of the networks with the size closest to the target compactness. Finally, the Assisted Tabu Search is applied to speed up the mutation process.

Table 1. Comparison of average execution time of the proposed method and state-of-the-art lightweight approaches. Runtime are obtained utilizing the Pixel 3a phone with the input image resolution of $640 \times 480$. Adapted by permission, paper V 2022 © IEEE.

<table>
<thead>
<tr>
<th>Task</th>
<th>Architecture</th>
<th>CPU(ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth Estimation</td>
<td>FBNet Tu et al. (2020); B. Wu et al. (2019)</td>
<td>852</td>
</tr>
<tr>
<td></td>
<td>FastDepth Wolf et al. (2019)</td>
<td>458</td>
</tr>
<tr>
<td></td>
<td>LDP-Depth-N</td>
<td>287</td>
</tr>
<tr>
<td></td>
<td>LiDNAS-K</td>
<td>205</td>
</tr>
<tr>
<td></td>
<td>LDP-Depth-K</td>
<td>204</td>
</tr>
<tr>
<td>Semantic Segmentation</td>
<td>Lite-HRNet C. Yu et al. (2021)</td>
<td>217</td>
</tr>
<tr>
<td></td>
<td>LDP-Seg-Ci</td>
<td>207</td>
</tr>
<tr>
<td></td>
<td>VDSR Kim, Lee, and Lee (2016)</td>
<td>425</td>
</tr>
<tr>
<td></td>
<td>LDP-Sup-x2</td>
<td>203</td>
</tr>
</tbody>
</table>
Experiments on datasets such as NYU-Depth-v2, KITTI, Cityscapes, COCO-stuff, DIV2K, Set5, Set14, BSD100, and Urban100 demonstrate that the proposed method achieves state-of-the-art results while being more compact than baseline methods regarding the number of parameters. Table 1 shows that LDP models run faster on handheld devices, suggesting real-time possibility with further optimization.

4.2.6 Discussion

Paper IV and Paper V introduced novel approaches towards a more general dense prediction problem. They provide several advantages over prior arts, including 1) tractable exploration time that is usually within the budget at most research units and 2) good performance with great flexibility thanks to a well-defined generic backbone.

The proposed framework can be seen as a hybrid approach that inherits the advantages of automated machine learning and human-engineering architectural structures. The search space is easily expanded using more optimized operations design either by machine or human experts. As demonstrated in Paper V, adding the micro-blocks (Y. Li et al., 2021) to the operation pools tends to generate better performing models.

The proposed framework seems to obey the Pareto optimality suggesting a trade-off between network performance and model compactness. Moreover, experiments indicate the importance of picking the right balance coefficient. Therefore, conducting some form of coefficient search before using the framework is a good practice.

There is still a gap when employing generated models on resource constraints devices as the current flow first aims for specific hardware, then determines target compactness before searching for a lightweight architecture. Future work based on Tan et al. (2019), where target devices are integrated during the exploration, will significantly reduce the deployment time.
Summary and conclusion

This thesis has presented novel deep-learning approaches for improving 3D sensing and dense prediction. This includes using prior knowledge with non-local networks (Paper I), and salient points along with normalized Hessian loss (Paper II) to enhance monocular depth accuracy. Acquiring sparse 3D points becomes more manageable. This enables the development of high-performing and low-cost depth completion utilizing 2D-3D features fusion. Manually designing DNNs becomes increasingly intricate, especially for dense prediction. Efficient neural architecture search frameworks are valuable for single image depth estimation (Paper IV) and dense prediction (Paper V).

High-performance depth estimation methods using DNNs typically resort to massive training data and large network architectures. Paper I proposed a monocular depth prediction method that is both accurate and lightweight. The coplanarity constraint and the depth attention volume are incorporated to assist the learning of non-local depth dependencies from data. The method achieves state-of-the-art results on various indoor datasets and outperforms the best-performing baseline (Yin et al., 2019) while using 78% fewer model parameters.

Most accurate depth estimation models are large, while small networks tend to produce inadequate predictions. Equipped with additional 3D points, depth completion is superior to the monocular approach but requires input from depth sensors. Built upon this, Paper II introduced a single image depth estimation method by replacing known depth measurements with predicted salient points. Moreover, the normalized Hessian loss was applied to aid the network learning. The method performs on par with prior art (Ranftl et al., 2021) while employing a 15 times smaller model enabling depth estimation for resource-constrained devices.

Depth completion methods struggle with highly sparse data and are mostly limited to input coming from depth sensors. Paper III presented a depth completion approach that is agnostic to the data source as it performs well with extremely sparse and unevenly distributed 3D points. Unlike previous approaches, the proposed method directly extracts and fuses geometric features in the 3D space with appearance features at multiple resolutions. The approach outperforms prior arts in various settings. It is also lightweight and capable of running in real-time on handheld devices.
Manually designing accurate and lightweight depth estimation networks is laborious, while the state-of-the-art neural architecture search for this task is expensive. Paper IV introduced a fast NAS framework that aims for high-performing and compact depth prediction models. The method leverages recent NAS advances (Mellor et al., 2021) that score network performance at the initial phase to decrease exploration time. The proposed approach outperforms the NAS-based method (Saikia et al., 2019) in all metrics as well as uses less searching time. Generated networks are lightweight and capable of deploying on mobile hardware.

The increasing demand for novel DNNs designs makes this task more and more complex. Neural architecture search emerges as an appealing solution. However, NAS research mainly focuses on classification, while dense prediction is a more complicated problem (Elsken et al., 2022). On top of that, NAS-based methods are either computationally expensive or built only for a single sub-task. Paper V presented an efficient NAS method for more generalized dense predictions. The method constructs flexibility and diversified search space based on a well-defined backbone. Additionally, a novel search algorithm is utilized to further reduce the exploration time. The proposed approach attains competitive results for various tasks such as depth estimation, semantic segmentation, and image super-resolution.
References


1437–1446).


Gould, S., Fulton, R., & Koller, D. (2009). Decomposing a scene into geometric and semantically consistent regions. In 2009 ieee 12th international conference on computer vision (pp. 1–8).


opaque object from one view. *PhD thesis.*


computer vision and pattern recognition (pp. 2117–2125).


intelligent robots and systems (iros) (pp. 5848–5854).


reconstructed using sfm and object labels. In *Proceedings of the ieee international conference on computer vision* (pp. 1625–1632).


Original publications


Reprinted with permission from Springer (I) and IEEE (II, III and IV).

Original publications are not included in the electronic version of the dissertation.
841. Suukko, Laura (2022) Managing integration capabilities in collaborative inter-organizational projects
842. Tiensuu, Henna (2022) Modelling the quality of the steel products under challenging measurement conditions
843. Avsievič, Tatiana (2022) Red blood cells and novel nanomaterials: towards nanosafety and nanomedicine
844. Wu, Xiaoting (2022) Machine learning for audio-visual kinship verification
845. Pérez Centeno, Víctor (2022) Merging neuroscience technologies in entrepreneurship research
846. Cajander, Niko (2022) Temporary agency work and worker well-being at restaurants: insights into socially sustainable work
847. Nouri, Parisa (2022) Cooperative diversity mechanisms for critical machine-type communications
848. Afonin, Nikita (2022) Development of passive seismic interferometry to study shallow subsurface structure
850. Tampio, Kari-Pekka (2022) Enhancing value creation at the front-end of a collaborative hospital construction project
851. Annunen, Petters (2022) Industrialization in construction: a process model for capability creation
852. Omodara, Linda (2022) Developing and testing sustainability assessment tools for chemical processes and products: case study on critical rare earth elements
854. Kallio, Rita (2022) Multidisciplinary study of the beneficiation potential of the Kiviniemi ferrous scandium deposit
855. Kiani, Sepideh (2022) Development of passive bioreactors treating different types of low carbon wastewater in cold climate conditions
857. Sone, Su Pyae (2022) Proactive resource allocation in spectrum sharing with radar systems via ML-based wireless network time series forecasting

Book orders:
Virtual book store
http://verkkokauppa.juvenesprint.fi
Lam Huynh

FROM 3D SENSING TO DENSE PREDICTION